# A genomic analysis using RNA-Seq to investigate the adaptation of the psychrophilic diatom *Fragilariopsis cylindrus* to the polar environment

Jan Strauss

A thesis submitted for the degree of Doctor of Philosophy

University of East Anglia, Norwich, UK

School of Environmental Sciences

September 2012

*Für meine Familie*

*Success in research needs four Gs: Glück, Geduld, Geschick und Geld.*
*(Luck, patience, skill and money)*

*—Paul Ehrlich (Nobel Prize for Physiology or Medicine 1908)*
*Quoted in M. Perutz, "Rita and the Four Gs", Nature, 1988,* ***332****, 791*

# Abstract

Diatoms are unicellular photosynthetic eukaryotes with a silicate cell wall. They often dominate polar marine ecosystems, driving the major biogeochemical cycles in these areas. The obligate psychrophilic diatom *Fragilariopsis cylindrus* is a keystone species in the Southern Ocean. It thrives both in open waters and sea ice and has become a model for studying eukaryotic microalgal adaptations to polar marine conditions. The aim of this thesis was to identify how the genome of *F. cylindrus* has evolved to cope with marine environmental conditions of the Southern Ocean. To identify key genes, comparative genomics, high-throughput transcriptome sequencing and reverse genetics were applied. Comparative genomics with the sequenced mesophilic diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* was combined with genome-wide RNA-Seq transcriptome analysis, leading to the discovery a new bacteria-like rhodopsin not present in other sequenced diatoms. The characterisation of a bacteria-like rhodopsin in *F. cylindrus* was conducted by applying reverse genetics tools.

The genome was characterised by a low G+C content, which affected codon usage. High sequence polymorphism resulted in pronounced unequal expression of putative heterozygous allelic gene copies in response to six different conditions. RNA-Seq detected transcriptional activity for 95% of the 27,137 predicted genes and > 4 fold expression changes between 55% of putative alleles. The most significant transcriptional changes were detected during prolonged darkness affecting 70% of genes and 30% of RNA-Seq reads mapped to unannotated regions of the genome. Two rhodopsin alleles showed unequal bi-allelic expression in response to iron starvation and heterologous expression in *Xenopus laevis* oocytes experimentally confirmed light-driven proton pumping for the iron-induced rhodopsin allele, suggesting significance for the adaptation of *F. cylindrus* to environmental conditions of the Southern Ocean.

These data show how the polar environment can shape the genome of a eukaryotic phytoplankton in unprecedented detail. High numbers of species-specific genes resulting in expansion of gene and protein families, low G+C likely enabling efficient translation at low temperatures and a high degree of heterozygosity combined with unequal bi-allelic expression, may provide an adaptive strategy to polar conditions by conferring metabolic flexibility and capacity to adapt to a rapidly changing environment.

# Contents

# List of tables

# List of figures

# Preface

This statement certifies that the work presented in this thesis was conceived, conducted, written and disseminated by Jan Strauss. I was responsible for the planning, execution and analysis of the experimental work presented and I have written the six chapters contained in this thesis. As my primary supervisor, Dr. Thomas Mock was involved in all aspects of this work including conceptualisation and critical reviews of earlier draft versions of this thesis. Additional contributions are explained below.

In **chapter 2**, I have summarised the material and methods, which have been used to generate and analyse the data presented in this work. This includes not only materials and methods used by myself, but also materials and methods used by project collaborators. In the following, the individual contributions of collaborators are explained in more detail for each of the three result chapters.

**Chapter 3** is based on data, which was generated within the framework of the *Fragilariopsis cylindrus* CCMP 1102 genome sequencing project, an international collaborative effort initiated and led by Dr. Thomas Mock, who also extracted the DNA and RNA for sequencing. *F. cylindrus* genome and EST sequences were generated at the US Department of Energy Joint Genome Institute (JGI). The *F. cylindrus* sequence assembly was built by Jeremy Schmutz and an 8-fold draft assembly and annotation was made publicly available in October 2009. Igor V. Grigoriev and Robert P. Ottilar provided the nuclear genome sequence annotation for the draft assembly from the JGI Genome Annotation Pipeline and were responsible for implementing genomic data into the JGI Genome Portal. The *F. cylindrus* genome portal provided access to all JGI genomic databases and analytical tools and allowed to analyse these data in different contexts over the web. Custom analyses were performed by the whole *F. cylindrus* genome consortium consisting of Thomas Mock[1], Jan Strauss[1], Andrew Toseland[2], Rachel Hipkin[1], Barbara R. Lyon[1,10], Mark McMullan[1], Cock van Oosterhout[1], Andrew E. Allen[3], Ruben E. Valas[3], Christopher L. Dupont[3], Beverley R. Green[4], Ansgar Gruber[5], Peter G. Kroth[5], Christoph Mayer[6], Florian Leese[6], Michal R. Sussmann[7],

[1] School of Environmental Sciences, University of East Anglia, Norwich, NR4 7TJ, United Kingdom
[2] School of Computational Sciences, University of East Anglia, Norwich, NR4 7TJ, United Kingdom
[3] J. Craig Venter Institute, San Diego, California 92121, USA
[4] Department of Botany, University of British Columbia, 3529-6270 University Boulevard, Vancouver, British Columbia V6T 1Z4, Canada
[5] Fachbereich Biologie, University of Konstanz, 78457 Konstanz, Germany

Alexandra Z. Worden[8], James A. Raymond[9], Michael G. Janech[10], Angela Falciatore[11], Nils Kroeger[12], Nicole Poulsen[12], Remo Sanges[13], Stephan Frickenhaus[14], Christiane Uhlig[14], Gernot Gloeckner[15], Antonio E. Fortunato[11], Robert P. Otillar[16], Igor V. Grigoriev[16], Chris Bowler[17], Alaguraj Veluchamary[17], E. Virginia Armbrust[18], Micaela Schnitzler Parker[18], Klaus Valentin[14], Florian Maumus[19] and Jeremy Schmutz[16]. I manually annotated about 500 genes and performed custom comparative analyses on metabolic pathways including carbohydrate metabolism, lipid metabolism, chlorophyll metabolism and *F. cylindrus*-specific genes. Additionally, to compile a synopsis of the *F. cylindrus* genome, I interpreted and analysed data provided by other members of the *F. cylindrus* genome consortium, whose individual contributions are specified as follows. The identification, annotation and classification of genomic repeats were performed by Florian Maumus and Hadi Quesneville. F. Maumus also performed analysis codon usage analysis. Christoph Mayer and Florian Leese performed tandem repeat analysis. Gene promoter analyses were performed by Remo Sanges and Stephan Frickenhaus identified tRNA genes. Andrew E. Allen and Ruben E. Valas carried out comparative analysis with available phytoplankton genomes and identified the diatom core genome. Christopher L. Dupont analysed metal-binding protein families and identified the overrepresentation of zinc-binding protein domains. The role of natural selection on expanded zinc-binding protein families was further investigated by Mark McMullan and Cock van Oosterhout, who also assisted in the analysis of allelic divergence in *F. cylindrus*. Andrew Toseland assisted with the analysis of allelic nucleotide identity and with bioinformatics analysis of allelic divergence in *F. cylindrus*. Light-harvesting complex proteins were annotated by Beverley E. Green and anti-freeze proteins in *F.*

[6] Department of Evolutionary Ecology and Biodiversity of Animals, Ruhr University Bochum, 44780 Bochum, Germany
[7] University of Wisconsin Biotechnology Center, 425 Henry Mall, Madison, Wisconsin 53706, USA
[8] Monterey Bay Aquarium Research Institute, Moss Landing, CA 95039 USA
[9] School of Life Sciences, University of Nevada, Las Vegas, Nevada 89154, USA
[10] Medical University of South Carolina, Charleston, South Carolina 29425, USA
[11] Universite Pierre et Marie Curie, Paris 06, Centre National de la Recherche Scientifique, UMR7238, Laboratoire de Genomique des Microorganismes, 75006 Paris, France
[12] ZIK B CUBE, Technical University Dresden, Arnoldstrasse 18, 01307 Dresden, Germany
[13] Animal Physiology and Evolution, Stazione Zoologica Anton Dohrn, Villa Comunale I, 80121 Napoli, Italy
[14] Alfred Wegener Institute for Polar and Marine Research, Am Handelshafen 12, 27570 Bremerhaven, Germany
[15] Leibniz Institute for Age Research - Fritz-Lipmann-Institute e.V., Beutenbergstrasse 11, 07745 Jena, Germany
[16] Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598, USA
[17] CNRS UMR8197 INSERM U1024, Environmental and Evolutionary Genomics Section, Institute of Biology, Ecole Normale Superieure, 46 rue d'Ulm, 75005 Paris, France
[18] School of Oceanography, University of Washington, Seattle, Washington 98195, USA
[19] Unité de Recherche en Génomique-Info UR1164, INRA de Versailles-Grignon, Route de Saint Cyr, Versailles, 78026, France

*cylindrus* were annotated and analysed by James Raymond and Christiane Uhlig. A phylogenetic analysis of anti-freeze proteins contributed by C. Uhlig was also analysed in this work.

In **Chapter 4**, I planned the experiments together with Thomas Mock. I was responsible for the execution and analysis of the experimental work. I performed growth experiments with *F. cylindrus* and extracted RNA for RNA-Sequencing. RNA-Seq libraries were constructed at The Gene Pool (Ashworth Laboratories, University of Edinburgh) under supervision of Dr. Karim Gharbi and RNA-Seq reads were mapped to the *F. cylindrus* genome by Gaganjot Kaur (The Gene Pool, University of Edinburgh). I interpreted the data with technical support from Andrew Toseland (PhD student at School of Computational Sciences, University of East Anglia, Norwich), who assisted in building and maintaining the bioinformatics pipeline for transcriptome analysis. Additionally, I developed and performed relative RT-qPCR assays from extracted RNA including allele-specific RT-qPCR assays to confirm RNA-Seq data.

The idea for **Chapter 5** was developed by me and Thomas Mock. I conducted growth experiments with *F. cylindrus*, extracted RNA and performed cDNA synthesis. Computational *in silico* analyses of the *Fragilariopsis* rhodopsin were performed by me and I cloned a full-length cDNA sequence from a *Fragilariopsis* rhodopsin allele. I also initiated collaboration with the laboratory of Prof. Dr. Georg Nagel (University of Würzburg, Germany), who contributed to the functional characterisation of the *Fragilariopsis* rhodopsin. The heterologous expression of polar microbial rhodopsins from *F. cylindrus* and the dinoflagellate *Polarella glacialis* in various expression systems was performed by me with contributions from Sabrina Förster and Shiqiang Gao (both PhD students in the Nagel Lab). Shiqiang Gao cloned a full-length cDNA sequence of a second *Fragilariopsis* rhodopsin allele and Sabrina Förster cloned a full-length cDNA sequence of a *Polarella* rhodopsin. I subcloned both *Fragilariopsis* rhodopsin gene copies and made different sequence constructs for genetic transformation in *Phaeodactylum tricornutum*. While I performed transformation and heterologous expression of different *Fragilariopsis* rhodopsin sequence constructs in *P. tricornutum*, the heterologous expression of rhodopsins from *F. cylindrus* and *P. glacialis* in *Xenopus laevis* oocytes was carried out by Sabrina Förster and Shiqiang Gao under supervision of Georg Nagel. Shiqiang Gao also measured *Fragilariopsis* rhodopsin photocurrents. I performed the gene expression analysis of *Fragilariopsis*

rhodopsin gene copies and developed relative and absolute RT-qPCR assays of the *Fragilariopsis* rhodopsin including allele-specific RT-qPCR.

# Acknowledgements

This work would not have been possible without the help of many people and Paul Ehrlich's four Gs for success in research (p. III). The fourth G (i.e., funding) for genome sequencing was provided by the US Department of Energy's (DOE) Office of Science, Biological and Environmental Research Program, the University of California, Lawrence Berkeley National Laboratory (contract no. DE-AC02-05CH11231), Lawrence Livermore National Laboratory (contract no. DE-AC52-07NA27344) and Los Alamos National Laboratory (contract no. DE-AC02-06NA25396). Diatom genome sequencing was performed at the DOE Joint Genome Institute (JGI, Walnut Creek, CA, USA). Additional funding for functional genomics and marine genomics work was provided by The Royal Society of London, UK (two grants for diatom functional genomics and general marine genomics given to T. Mock in 2008 and 2009) and the Natural Environment Research Council UK (NERC; grant no. NE/I001751/1). Diatom transcriptome sequencing (RNA-Seq) was performed at the GenePool hosted by the NERC Biomolecular Analysis Facility, Edinburgh, UK. I also would like to thank the University of East Anglia (UEA, Norwich, UK) for supporting me with a stipendship as well as the American Society for Limnology and Oceanography (ASLO), the Challenger Society for Marine Science, the Society for General Microbiology (SGM) and the British Phycological Society (BPS) for financial support for travels to scientific conferences and practical workshops.

I would like to thank my principal supervisor **Dr. Thomas Mock** for his endless enthusiasm, constructive feedback and professional guidance. He got the best out of me. Also I would like to thank my co-supervisor **Dr. Gill Malin** for her scientific advice, the proofreading of earlier drafts of this thesis and her words of encouragement for me.

This extensive research would not have been possible without the help of research collaborators. Especially, I wish to express my sincere gratitude for their intellectual contributions, data and pleasant collaboration. A great thanks goes to all members of the ***Fragilariopsis cylindrus*** **genome consortium**, and in particular **Andrew E. Allen**, **Christoph L. Dupont**, **Stephan Frickenhaus**, **Beverley E. Green**, **Florian Leese**, **Florian Maumus**, **Christoph Mayer**, **Robert P. Otillar**, **James A. Raymond**, **Remo Sanges**, **Jeremy Schmutz**, **Andrew Toseland**, **Christiane Uhlig** and **Ruben E. Valas**, who contributed genomic data analysed in this work. I also would like to thank **Ansgar Gruber**, for the gift of the pPha-T1 expression vector and its

derivative StuI-GFP pPha-T1. Additionally, I am thankful to **Cock van Oosterhout** and **Mark McMullan** for fruitful discussions on evolutionary biology and genomic heterozygosity, as well as critical comments on earlier drafts of this thesis.

Moreover, I am very grateful to **Prof. Dr. Georg Nagel** for his collaboration on phytoplankton rhodopsins and for inviting me to visit his lab as well as to a German Research Foundation (DFG) hosted workshop on "Specific light driven reactions in unicellular model algae". I am also very grateful to his PhD students **Shiqiang Gao** and **Sabrina Förster**, who were of invaluable help in the functional analysis of phytoplankton rhodopsins.

Special thanks also to all current and former lab members of the Mock Lab, who created a pleasant working environment throughout my PhD. Especially, I like to thank **Rachel Hipkin** and **Amy Kirkham** for their friendly support and assistance with lab issues as well as for lightening up the long days in the lab. Additional thanks goes to A. K. for her critical comments on earlier versions of this thesis. Furthermore, a very special thanks to **Andrew Toseland** for his invaluable help and assistance with computational analysis in this work. Many thanks also go to **Rob Utting** for his great technical support in the lab.

Thanks also to all my colleagues and friends in Norwich, back home and in the rest of the world, for sharing their time with me and supporting me along the way. A very special thanks to **Anne Velenturf** for her loving support and her constant encouragement throughout the years.

Finally, I would like to give special thanks to my parents **Regina and Martin Strauss**, my brother **Jens Strauss**, who contributed Figure 1 of this work, and the rest of my family for their moral and financial support as well as their endless encouragement, which kept me afloat in the rough sea of doing my PhD.

Thanks! Bedankt! Danke!

# Chapter 1

# General introduction

## 1.1 The polar oceans and sea ice

The oceans cover approximately 70% of the Earth's total surface area and the polar oceans contribute about 10% to the total ocean area. The polar oceans have significant influence on climate and global cycles through the formation of cold nutrient-rich deep water, sea ice and physical and biological carbon sequestration. However, the two polar oceans, Arctic Ocean and Southern Ocean, are very different with regard to their genesis and environmental conditions.

**Geographic and oceanographic overview.** The Arctic Ocean is a pole-centred intercontinental mediterranean sea, which is enclosed by Eurasia, North America and Greenland (Figure 1). This marine system covers an area of about $14 \times 10^6$ km$^2$ in area and $20 \times 10^6$ km$^3$ in volume (Fahrbach et al., 2009) and is the smallest of the world oceans. The Fram Strait between Greenland and Spitsbergen provides the only deep water passage (sill depth about 2600m) through which the North Atlantic Drift ("Gulf Stream") streams northbound into the Arctic Ocean. The opening of the Fram Strait ~10-17.5 Myr ago (Jakobsson et al., 2007; Engen et al., 2008) allowed critical water mass exchange between Arctic Ocean and North Atlantic and marked the onset of the modern Arctic Ocean. In addition, the Bering Strait between North America and Eurasia provides only a shallow passage (sill depth about 50 m) to the Pacific Ocean with only surface water mass exchange. The North Atlantic Drift, Transpolar Drift Stream and Beaufort Gyre create a complex ocean stream pattern within the Arctic Ocean. Due to high freshwater input from Siberia and Canada as well as repeated annual melting and freezing, the Arctic Ocean possesses a stable 200 – 300 m thick surface layer of low salinity water (Polar surface water).

In comparison, the Southern Ocean is a deep (3000 – 4000 m) circumpolar ring ocean (Figure 1), delimited at the poleward edge by the Antarctic continent (65° – 70° S) and by the oceanographic feature of the Antarctic Convergence (Antarctic Polar Front) at 50 – 60° S, which represent a boundary with steep physical (e.g. temperarture change of 2 °C) (Rintoul and Bullister, 1999) and chemical gradients (e.g. salinity, nutrients) (Zentara and Kamykowski, 1981; Deacon, 1982). This boundary effectively

isolates phytoplankton in the surface layer of the Southern Ocean from the warmer waters found further north, but allows free biotic exchanges with the world's oceans (Atlantic Ocean, Pacific Ocean, Indian Ocean) in deep zones below 500 – 1000 m (Hempel, 1987).

The Southern Ocean is approximately twice as big as the Arctic Ocean covering an area of $36 \times 10^6$ $km^2$ and water mass exchange with the world's oceans through the Antarctic Convergence is much stronger than in the Arctic Ocean. The tectonic opening of the ocean gateways between Antarctica and Australia (Tasmanian Passage), and Antarctica and South America (Drake Passage) ~34 Myr ago resulted in the isolation of Antarctica by the Antarctic Circumpolar Current and caused the extreme polar conditions of the modern Southern Ocean (Kennett, 1977; Exon et al., 2002). These circumstances allowed for the evolution of numerous highly specialised endemic species in the Southern Ocean (Rogers, 2007). The Antarctic Circumpolar Current is driven by strong westerly winds, which dominate the movement of surface water around the Antarctic continent, with more complex circulation patterns close to the continent, including the Antarctic Coastal Current and the Weddell and Ross Sea Gyres. In contrast to the Arctic Ocean, the Southern Ocean is lacking a low salinity surface layer due to little freshwater input from tabular icebergs and the underside of the ice shelves of the Antarctic continent. Thus, the thermohaline stratification in the Southern Ocean is low and vertical water circulation is more pronounced (Hempel and Piepenburg, 2010).

Annual water temperatures throughout the water column are relatively constant, ranging between −1.9 °C and −1.7 °C (Littlepage, 1965). Both polar oceans are characterised as oxygen-rich, because oxygen solubility is inversely related to temperature. In the Southern Ocean oxygen levels are approximately 1.6-fold higher than seawater with a temperature of 20 °C (Littlepage, 1965). High oxygen concentrations promote the formation of free oxygen radicals, which can damage biological macromolecules including DNA, lipids and proteins. Thus, Antarctic organisms protect themselves against oxidative damage using antioxidants (Schriek, 2000; Yamamoto et al., 2001; Abele and Puntarulo, 2004; Ha et al., 2006; Regoli et al., 2011; Um et al., 2012; O'Brien and Crockett, 2013). Additionally, compared to temperate species, Antarctic organisms comprise higher concentrations of proteins that mediate iron metabolism, because iron also promotes free radical production (Chen et al., 2008; Clark et al., 2011). Furthermore, trace elements play an important role in the Southern Ocean, because phytoplankton productivity is predominantly limited by iron

(Martin and Fitzwater, 1988), whereas the concentrations of the macronutrients nitrate, phosphate and silicic acid are high all year round, making regions within the Southern Ocean the largest High Nutrient Low Chlorophyll (HNLC) areas.

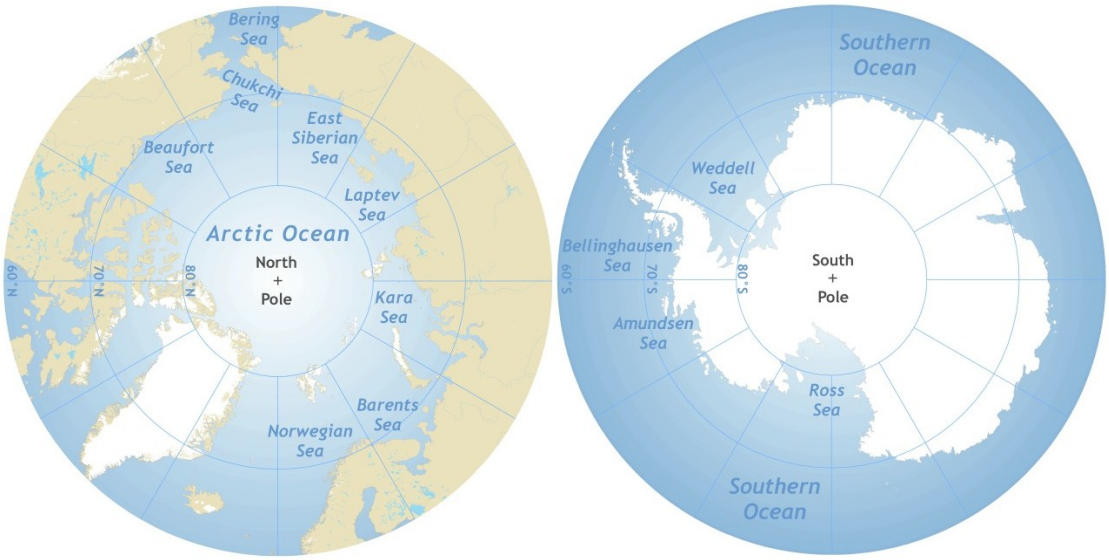



**Figure 1. Comparison of the Arctic Ocean and Southern Ocean (courtesy of Jens Strauss, Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research (AWI), Potsdam, Germany. Made with Natural Earth. Free vector and raster map data @ naturalearthdata.com).**

**Sea ice.** A distinct feature of polar oceans is their coverage with perennial (multi-year ice) or seasonal (annual ice) sea ice, which first appeared during the greenhouse-icehouse climate transition (Zachos et al., 2001) of the Eocene epoch (55 – 34 Myr) (Shackleton and Kennett, 1975; Tripati et al., 2005; Moran et al., 2006; DeConto et al., 2007; DeConto et al., 2008; Stickley et al., 2009). Thus, polar sea ice habitats represent geologically new habitats and their geological age affects the evolutionary legacy of the Polar Regions (Crame, 1997).

Sea ice is frozen seawater, which forms as a result of the prevailing cold air temperatures in Polar Regions when the ocean surface water cools down to its freezing temperature of −1.8 °C. Ice and its snow cover cause a high albedo, reflecting most of the radiation from the sun back to space. This results in an albedo of 70 – 80% in the sea ice zone, whereas the albedo of open waters is only 10-15% (Turner and Marshall, 2011). Sea ice can cover up to $35 \times 10^6$ km$^2$ (13% of earth's surface) at its maximum extent (Parkinson and Gloersen, 1993). The Arctic Sea Ice can almost completely cover the Arctic Ocean during winter, resulting in a sea ice extent of $14 \times 10^6$ km$^2$. In

comparison, the Antarctic sea ice covers about $19 \times 10^6$ km$^2$ during winter, and the northern parts of the Southern Ocean stay permanently ice free. The amount of multi-year sea ice (approximately 3 m thickness) differs greatly between both polar oceans, which covers ~30% of the Arctic Ocean, but only 10% of the Southern Ocean (Hempel and Piepenburg, 2010).

Although sea ice appears to be hostile to life, it serves as habitat for a community of highly adapted microorganisms including unicellular algae (e.g. diatoms), bacteria, viruses, protists, flatworms, and small crustaceans, which live in the network of sea ice brine channels and pores that develop during sea ice formation (Arrigo and Thomas, 2004; Mock and Thomas, 2005). The network of sea ice brine channels and pores is formed during the freezing process of sea water, when the dissolved compounds of seawater, especially salts, are not included into the crystal structure of ice but are concentrated in sea ice brine. During the process of ice formation, planktonic organisms are scavenged by ice crystals and get concentrated in sea ice (Eicken, 1992). With decreasing temperatures and ice growth in winter, the network of pores and channels becomes narrower and the volume of sea ice brine decreases further, increasing its salinity. Thus, organisms within the sea ice are not only subjected to changing temperatures, but also to varying salinity and available space. Generally, strong vertical gradients of physical and chemical conditions throughout the ice column characterise the narrow sea ice habitat (Figure 2) and the pushing and rafting of ice floes due to waves and swell causes additional mechanical stress.

The conditions in sea ice are extreme with respect to radiation (extreme light conditions, high UV), temperatures (between −2 °C and −40 °C), salinity (between 30 and 150 $S_A$) and high pH (up to pH 10). Moreover, photosynthetic activity of unicellular algae leads to depletion of dissolved inorganic carbon, shifting the pH to high values and causing hyperoxic conditions. The latter can exceed oxygen saturation (Mock et al., 2002; Trenerry et al., 2002) and facilitate the production of reactive oxygen radicals (Thomas and Dieckmann, 2002). As a result sea ice organisms have to cope with multiple stresses, whereas physiological stresses are more moderate in the open waters of polar oceans.

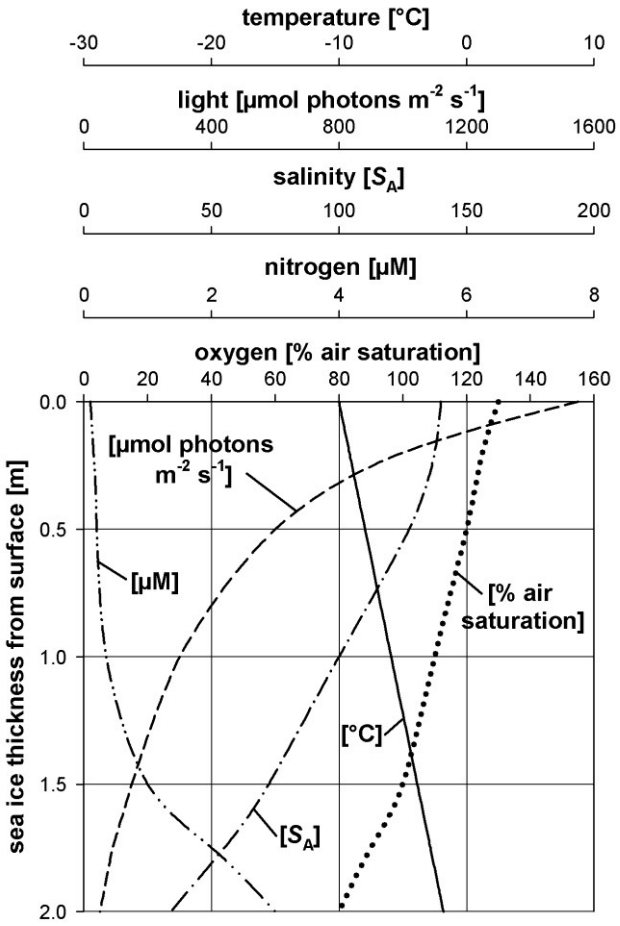



**Figure 2. Typical gradients of temperature, light, salinity, nutrients (e.g. nitrogen) and oxygen across a sea ice column [extended and modified after McGrath Grossi, S. and C. W. Sullivan (1985), *J. Phycol.* 21: 401-409 in Schrieck (2002), *Rep. Polar Res.* 349, 1-130].**

## 1.2 Phytoplankton and primary production in polar oceans

The polar oceans contain a high diversity of pico- (< 2 µm) and nanophytoplankton (2 – 20 µm) (Kang et al., 2001; Smetacek et al., 2004; Knox, 2007; Lovejoy et al., 2011) and the most prominent phytoplankton groups are diatoms, autotrophic nanoflagellates and dinoflagellates. Strikingly, cyanobacteria are almost absent in polar oceans (Vincent, 2002). Significant taxonomic overlap exists between phytoplankton and sea ice algae assemblages due to their tight coupling (Smith and Sakshaug, 1990). Diatoms make up the most biomass in polar oceans and sea ice, and can form major phytoplankton blooms. In addition to diatom blooms, autotrophic nanoflagellates including the brown-yellow prymnesiophyte *Phaeocystis antarctica* (Southern Ocean) and *Phaeocystis pouchetii* (Arctic Ocean) can produce large blooms and high biomass in polar oceans. During blooms polar phytoplankton produces significant amounts of the sulphur compounds dimethylsulfide (DMS) and its metabolic

precursor dimethylsulfoniopropionate (DMSP), contributing to the global biogeochemical cycle of sulphur (Kettle et al., 1999). Generally, the growth and development of polar phytoplankton is determined by environmental factors temperature, light, nutrients, vertical mixing and grazing (Sakshaug and Holm-Hansen, 1984) as in other world oceans. Moreover, viral lysis (Brussaard, 2004; Suttle, 2005) and genetically programmed cell death (Kirchman, 1999; Bidle and Falkowski, 2004; Franklin et al., 2006; von Dassow and Montresor, 2011) have recently been recognised as a significant phytoplankton loss process. In polar oceans some of these factors, including temperature, light, nutrients, vertical mixing and grazing, strongly depend directly or indirectly on sea ice coverage.

**Temperature.** The water temperatures in polar oceans are relatively constant and vary much less than in temperate and tropical oceans. In the Arctic Ocean, annual water temperatures range between −1.8 °C to 0 °C, whereas in the Southern Ocean temperatures they vary between −1.8 °C to +4 °C (Knox, 2007; Locarnini et al., 2010). It is estimated that low sea surface temperatures in the Southern Ocean were established during the Eocene-Oligocene transition (34 Myr) (Zachos et al., 1994). Temperatures in sea ice, however, can vary between −10 °C and −1.8 °C across the sea ice column (Figure 2). They are lowest at the top of the ice and may vary between −4 to −20 °C, depending on the ambient air temperature.

Although, early workers anticipated that oceanic phytoplankton is fully adapted to low temperatures of polar oceans including the Southern Ocean (Hart, 1934), recent research has shown that phytoplankton of polar oceans are psychrotolerant, rather than psychrophilic, with temperature optima for growth and photosynthesis in excess of that prevailing in polar oceans (Knox, 2007). Various studies showed that polar phytoplankton in general do not have specific ecophysiological adaptations that are different to temperate or tropical phytoplankton (Jacques, 1983; Tilzer et al., 1986; Smith and Harrison, 1991). Maximum growth rates at 0 °C are approximately half that at +10 °C ($Q_{10}$ rule), which corresponds to the known temperature dependencies of general biological processes, including photosynthesis and respiration. It has been shown that the $Q_{10}$ value for respiration is higher (Tilzer and Dubinsky, 1987) than for photosynthesis (Neori and Holm-Hansen, 1982; Tilzer and Dubinsky, 1987) in Antarctic phytoplankton. Thus, light-saturated photosynthesis is apparently more temperature-sensitive than respiration, which has been ascribed to temperature dependency of maximum photosynthetic quantum yields (Tilzer et al., 1986; Tilzer and Dubinsky,

1987). As a result net growth of polar phytoplankton may occur even under low temperature and short day length during winter (Tilzer and Dubinsky, 1987). In Arctic phytoplankton temperature optima for photosynthesis were found to be up to 10 °C higher than *in situ* temperatures (Li, 1985), and activation energies for photosynthesis and the carboxylating enzyme ribulose-1,5-bisphosphate carboxylase (RubisCO) were in close agreement, suggesting that RubisCO is the rate-limiting step of photosynthesis in polar phytoplankton (Li et al., 1984; Li, 1985). Furthermore, Toseland et al. (under peer review (2013)) suggested that additional phytoplankton core metabolism, including protein translation, is strongly affected by temperature and that initiation of translation at the ribosomes might be the rate-limiting step for protein synthesis under low temperatures in polar oceans. Low temperatures resulted in higher cellular contents of ribosomal RNA and ribosomal proteins in comparison to temperate and tropical oceans (Toseland et al., under peer review (2013)). Nevertheless, it appears that optimum temperatures for growth of polar phytoplankton are higher than the mean annual temperatures in polar oceans (Jacques, 1983; Thomas et al., 2012). Thus, temperature sets an upper limit on phytoplankton growth rates and at a given temperature the growth of phytoplankton is determined by the supply of light and nutrients (Smith and Sakshaug, 1990).

**Light and vertical mixing.** Light intensities are generally lower in marine environments compared to terrestrial environments (Depauw et al., 2012). On average polar regions receive five times less solar radiation than tropical regions (Bertler and Barrett, 2010). In marine environments, light varies due to incident solar radiation, time of day and year, concentration of suspended particles and organic matter, as well as absorption and scattering by water. Solar radiation or irradiance in the wavelength range of 400 – 700 nm of the electromagnetic spectrum is essential for marine phytoplankton.

The amount of incident solar radiation is controlled by solar activity, time of day and season as well as atmospheric conditions, including cloudiness, ozone content, turbidity, and humidity. Cloudiness, fog and snowfall can greatly limit the light available in the polar oceans and cause light attenuation in the range of 40 – 90% (Smith and Sakshaug, 1990). Additionally, a significant proportion of incident light in polar regions is reflected at the sea surface depending on the solar angle and the roughness of the sea (Powell and Clarke, 1936). Although the roughness of the sea increases the average angle between light direction and point of entry reducing the reflectance (Kirk, 2011), heavy storms, as frequently occurring in the Southern Ocean,

can produce air bubbles, which increase surface reflection (Powell and Clarke, 1936). Whilst wind speed determines the roughness of the sea and formation of air bubbles, the time of day and year determines the solar angle and thus the amount of light entering the surface waters of polar oceans. Low solar angles cause a higher reflection at the sea surface and a greater attenuation of shorter wavelengths in the atmosphere. Thus, light attenuation is more pronounced with increasing latitudes. As the Arctic Ocean lies north of 70° N, whereas the Southern Ocean is limited by the Antarctic continent at 65 – 70° S (Figure 1), the annual light cycle is more extreme in the Arctic Ocean than in the Southern Ocean. However, polar phytoplankton are well adapted to survive periods of darkness and laboratory experiments demonstrated that polar phytoplankton can endure > 3 months of darkness (Peters and Thomas, 1996). Nevertheless, light in the Arctic Ocean is only sufficient for a single phytoplankton bloom during the year (Heimdal, 1989) as opposed to other oceans where two blooms can occur (Longhurst, 1995).

Light that enters the water is quickly scattered and absorbed in a wavelength dependent manner. The light absorption characteristics of sea water cause a dominance of blue-green wavelengths with increasing water depth (Austin and Petzold, 1986). Moreover, the penetration depth of light also depends on the concentration of suspended particles and organic matter (Kirk, 2011). While, in contrast to the Arctic Ocean, the Southern Ocean is largely free of terrigenous matter and coloured soluble material (Mitchell, 1992), the opacity of the Southern Ocean waters is largely dependent on phytoplankton concentrations and light penetration is generally high (Sakshaug and Holm-Hansen, 1984). However, high phytoplankton concentrations, occuring during blooms, can be found in polar oceans and may cause self-shading, because phytoplankton absorbs strongly in the blue, blue-green, and red parts of the spectrum of photosynthetic active radiation (Kirk, 2011).

Additionally, light and vertical mixing in polar oceans can be greatly attenuated, when the ocean surface is covered by sea ice. The degree of light attenuation depends on its properties and includes sea ice thickness, snow cover, and presence of sea ice algae, brine pockets and air bubbles. Sea ice itself is more transparent (Maykut and Grenfell, 1975; Palmisano et al., 1987) than snow, which is highly opaque. While sea ice of 1 m thickness will reduce incident light to 20% of incident irradiance (Sullivan et al., 1984), a 50 cm thick layer of snow will reduce incident light to 0.01 – 3% of the surface irradiance (Palmisano et al., 1987). Nonetheless, massive under-ice phytoplankton blooms have been observed under Arctic first-year sea ice (Gradinger,

1996; Arrigo et al., 2012) and polar phytoplankton and sea ice algae are well adapted to low light levels (Cota, 1985; Kirst and Wiencke, 1995). Arctic diatoms can grow at irradiances as low as 10 µmol photons $m^{-2}$ $s^{-1}$ (Hegseth, 1989) and maximum growth rates were obtained at ~50 µmol photons $m^{-2}$ $s^{-1}$, after which the growth rates were independent of irradiance (Gilstad and Sakshaug, 1990).

As sea ice coverage decreases the light availability, it also decreases water turbulences and vertical mixing, which controls the mean light availability to polar phytoplankton (Sakshaug et al., 1991). Strong wind-driven vertical mixing in the Southern Ocean causes average mixing depth ranging between 60 – 100 m (Boyd et al., 2001), being deeper compared to other oceans, and may cause light limitation of phytoplankton growth (Mitchell et al., 1991). As a result, light availability is highly variable in the Southern Ocean, ranging between 0 – 800 µmol photons $m^{-2}$ $s^{-1}$ in summer (Hoogstraten et al., 2012), and causing light limitation to phytoplankton growth in the Southern Ocean in combination with low concentrations of the trace metal iron (Mitchell et al., 1991; de Baar et al., 2005; Alderkamp et al., 2010; Alderkamp et al., 2011) (see below).

Of growing relevance, regarding solar irradiance in polar oceans, is the anthropogenic thinning of the stratospheric ozone layer (ozone hole) in spring over the Antarctic continent. A pronounced ozone hole causes high doses of damaging ultraviolet radiation (100 – 400 nm) at the surface of the Southern Ocean and may result in limitation of phytoplankton production (Cullen et al., 1992; Smith et al., 1992; Neale et al., 1998). In comparison, the thinning of the ozone layer is less pronounced in the Arctic Ocean (Brune et al., 1991; von der Gathen et al., 1995; Turner and Marshall, 2011) but significant ozone loss has been observed (Newman et al., 1997).

**Nutrients and trace metals.** There are major differences in the nutrient (Levitus et al., 1993) and trace metal regimes in the Arctic and Southern Ocean, which affect polar phytoplankton productivity in addition to light availability.

In general, nutrient concentrations in the Arctic Ocean are lower than in the Southern Ocean and are maintained by its physical oceanographic features, which produce strong stratification that limits the supply of new nutrients from upwelling deep water masses. In contrast, nutrient levels in the Southern Ocean are higher due to the large scale Antarctic divergence which supplies new nutrients to the surface waters.

While nutrients become depleted after a phytoplankton bloom in the Arctic, nutrients are rarely depleted in the Southern Ocean, because scarcity of the essential trace metal iron limits phytoplankton processes (Martin, 1990). Iron limitation severely affects photosynthetic activity due to the strong dependency of the photosynthetic apparatus on iron as cofactor (Merchant and Dreyfuss, 1998; Strzepek and Harrison, 2004). Generally, trace metals like iron, zinc and cobalt, which are almost insoluble in sea water (Thuróczy et al., 2010), are largely supplied by dust (Jickells et al., 2005), river input and sediments (Johnson et al., 1997). Thus, in contrast to the high terrigenous freshwater input from surrounding continents into the Arctic Ocean, the input of trace metals from the Antarctic continent is low, because it is sealed off by its extensive ice sheets. As a result, most of the trace metals are transported to the Southern Ocean via dust from adjacent continents and icebergs, which discharge their dust when melting but maintain only low concentrations. As iron follows this pattern (Martin, 1990), relatively high concentrations of zinc (Fitzwater et al., 2000; M. Franck et al., 2000; Croot et al., 2011) and cobalt (Bown et al., 2011; Bown et al., 2012) in the Southern Ocean do not.

Zinc and cobalt play important physiological roles in phytoplankton metabolism and growth (Morel et al., 2006; Thuróczy et al., 2010). While zinc has been suggested to be a key micronutrient in the Southern Ocean (Saito et al., 2010), cobalt is involved in the biosynthesis of the vitamin cobalamin ($B_{12}$) (Kobayashi and Shimizu, 1999), and together with zinc it can serve as co-factor in metalloenzymes including carbonic anhydrases and hydrolytic enzymes (Morel et al., 2006). Although vitamin $B_{12}$ is not an essential requirement for growth of some phytoplankton (Croft et al., 2005; Croft et al., 2006; Helliwell et al., 2011), low concentrations of cobalt and vitamin $B_{12}$ may effect phytoplankton growth in the world ocean overall (Sañudo-Wilhelmy et al., 2006; Panzeca et al., 2008; Sañudo-Wilhelmy et al., 2012). While vitamin $B_{12}$ appears to be sufficient for sea ice microalgae growth in Antarctic sea ice communities (Taylor and Sullivan, 2008), low concentrations of cobalt in combination with low iron may limit phytoplankton growth in the Southern Ocean (Bertrand et al., 2007).

In addition to trace metal limitations, silicifying phytoplankton like diatoms and some chrysophytes may become limited by low silicic acid concentrations in the Arctic Ocean and the Southern Ocean (Nelson et al., 2001) during the summer growth season (Figure 3). While annual silica concentrations in the Arctic Ocean are generally low and high silica concentrations are restricted to Arctic river deltas (Garcia et al., 2010), silicic acid limitation in the Southern Ocean appears to be complex and often occurs in

combination with iron limitation (Figure 3). Generally, it has been recognised in recent years that phytoplankton processes are controlled by an interplay of environmental factors (Cullen, 1991; Lehman, 1991; Saito et al., 2008), including not only simultaneous limitation by iron and silicic acid (Hutchins et al., 2001) but also iron and irradiance (Boyd et al., 2001) (Figure 3).

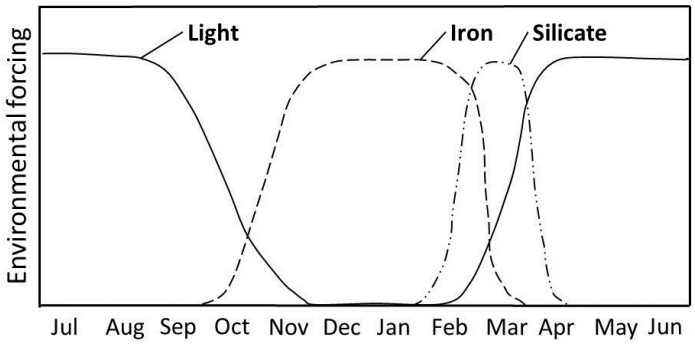



**Figure 3. Schematic diagram of putative seasonal progression of major environmental factors limiting or simultaneously limiting Southern Ocean diatoms. Note that the time period of limitation will vary with geographical location and from year-to-year and that other factors may exert environmental control [after Boyd, P. W. (2002), *J. Phycol.* 38, 844-61].**

**Grazing, viral lysis and programmed cell death.** In addition to the bottom-up controlling factors temperature, light and nutrients, phytoplankton are controlled by grazing (Frost, 1991) and viral lysis (Brussaard, 2004). The main consumers of pico- and nanophytoplankton are microzooplankton including flagellates, ciliates, heterotrophic dinoflagellates and crustaceans (Tsuda and Kawaguchi, 1997; Calbet and Landry, 2004).

Generally, the grazing pressure on phytoplankton is considered to be high in both polar oceans (Bathmann et al., 1990; Wheeler et al., 1996; Dubischar and Bathmann, 1997; Tsuda and Kawaguchi, 1997; Calbet and Landry, 2004). Although bacterial grazing mortality is similar in both polar oceans (Anderson and Rivkin, 2001), to my knowledge, the grazing mortality of phytoplankton in polar oceans has not been studied in a comparative manner yet. A general comparison, however, is difficult, because grazing pressure is a function of available food and can vary widely in the Arctic (Rysgaard et al., 1999) and Southern Ocean (Smetacek et al., 2004). Moreover, the food webs of the polar oceans differ, showing varying topologies with differences in the shape of the trophic linkages, including the ratio of basal and top species, and the

ratio of prey and predator species (de Santana et al., 2013). While in both polar oceans copepods are the most abundant zooplankton (Conover and Huntley, 1991), euphausiids, including the Antarctic krill, and salps play an important role in the Southern Ocean (Dubischar and Bathmann, 1997; Voronina, 1998). The role of copepod grazing in the Arctic Ocean may be more pronounced due to advection from the North Atlantic (Olli et al. (2007); Philip Assmy, Norwegian Polar Institute, Tromsø, *personal communication* 22/01/2013). Although swarms of Antarctic krill can excert high local grazing pressure on phytoplankton in the Southern Ocean, their distribution is patchy and thus their effect on phytoplankton grazing may also be patchy (Smetacek et al., 2004). In comparison, high abundances of copepods in the Arctic Ocean can control phytoplankton (Bathmann et al., 1990; Rysgaard et al., 1999) and support planktivorous fish and their predators including polar cod (Frank et al., 2005). In contrast, planktivorous fish are largely absent in the Southern Ocean (Rodhouse and White, 1995), still, phytoplankton stocks support largely copepods and krill, the forage of baleen whales and squid (Rodhouse and White, 1995; Smetacek et al., 2004).

Grazing and viral lysis have also been recognised as controlling factor of phytoplankton (Brussaard, 2004; Suttle, 2005). The abundances of viruses in the oceans are estimated to $10^5$ – $10^8$ viruses $mL^{-1}$ (Suttle, 2005) and it is suggested that viral abundances in polar oceans fall in the same range (Smith and Steward, 1992; Maranger et al., 1994; Payet and Suttle, 2008). Viruses may be enriched during sea ice formation with 10 – 100 fold higher abundances in sea ice than in the water column (Gowing et al., 2002; Collins and Deming, 2011), however, opposite patterns have also been observed with higher viral abundances in the water column compared to sea ice (Paterson and Laybourn-Parry, 2012).

Viruses of a size likely to infect eukaryotes rather than bacteria (capsid diameter > 100 nm) have been found in polar oceans including sea ice and constitute approximately 10% – 20% of the total viral abundance (Maranger et al., 1994; Gowing, 2003; Payet and Suttle, 2008). Interestingly, no viral infections were found in the major bloom forming phytoplankton in the Southern Ocean, diatoms and mucilaginous colonies of non-flagellated *Phaeocystis* (Gowing et al., 2002; Gowing, 2003). Furthermore, viral lysis has been found to play only a minor role in comparison to microzooplankton grazing in the Southern Ocean (Brussaard et al., 2008; Evans and Brussaard, 2012). Potential eukaryotic phytoplankton viruses were only detected at very low concentrations in the Southern Ocean and viral lysis was found to be only a minor

loss factor in comparison to microzooplankton grazing, contributing 6% compared to 45% of the total loss of phytoplankton standing stock (Brussaard et al., 2008). Yet, in the Arctic Ocean, phycodnaviruses infecting eukaryotic phytoplankton, have been identified (Payet, 2012), but their contribution to phytoplankton loss remains uncertain. Overall, the loss of primary production in aquatic microbial communities through viral lysis has been estimated to 2 – 3% (Suttle, 1994).

Compared to our knowledge of grazing and viral lysis, very little is known about the quantitative significance of phytoplankton losses due to genetically programmed cell death (Kirchman, 1999; Franklin et al., 2006). Programmed cell death in phytoplankton has been investigated as a response to abiotic or biotic stressors (Ferroni et al., 2007; Timmermans et al., 2007; Bidle and Bender, 2008; Franklin et al., 2012), but it may also be an intrinsic outcome of life cycle stages (von Dassow and Montresor, 2011). In a laboratory study of the Antarctic marine chlorophyte *Koliella antarctica*, programmed cell death was found in 1 – 2% of the cells after prolonged darkness (Ferroni et al., 2007). Nevertheless, the role of programmed cell death in polar oceans remains elusive.

In summary, the environmental factors temperature, grazing, viral lysis and programmed cell death are likely to be relatively constant in both polar oceans. However, while the extremely short summer growing season, combined with low nutrients and sea-ice cover, restrict phytoplankton in the Arctic Ocean, iron availability and deep vertical mixing restrict primary production of phytoplankton in the Southern Ocean (Noethig et al., 2009). Although primary production can be high, especially on continental shelves of the Arctic Ocean (Wheeler et al., 1996) and the Southern Ocean (Arrigo et al., 2008), the annual primary production of polar phytoplankton is generally low. The annual productivities of the Arctic and Southern Ocean are similar with pelagic primary production in the Arctic Ocean estimated at 44 g C $m^{-2}$ $yr^{-1}$ (Pabi et al., 2008), with 57 g C $m^{-2}$ $yr^{-1}$ for the Southern Ocean (Arrigo et al., 2008). In comparison, the annual oceanic primary production is estimated to 140 g C $m^{-2}$ $yr^{-1}$ (Field et al., 1998) with maximum oceanic primary production rates of 1000 – 1500 g C $m^{-2}$ $yr^{-1}$ in highly productive marine upwelling and estuarine systems (Walsh, 1981; Field et al., 1998). Additionally, the annual Antarctic marine primary production of sea ice is estimated to 63 – 70 Tg C $yr^{-1}$ (Lizotte, 2001; Arrigo et al., 2010a), contributing ~4% to the total Antarctic marine primary production of ~1,949 Tg C $yr^{-1}$ (Arrigo et al., 2008), while the annual Arctic marine primary production of sea ice is estimated to contribute 15 – 20%

(Pabi et al., 2008; Arrigo et al., 2010a) to the total primary production in the Arctic Ocean of ~419 Tg C $yr^{-1}$ (Pabi et al., 2008). Generally, diatoms tend to dominate phytoplankton communities in polar oceans, when sufficient light and nutrients are available (Figure 3) to sustain their growth (Armbrust, 2009).

## 1.3 Marine diatoms in the polar environment

Diatoms are photosynthetic unicellular, eukaryotic microalgae with a cell wall made of silica (Round et al., 1990), hence they require dissolved silicic acid. They are highly divers with an estimate of around 100,000 different species (Mann and Droop, 1996; Norton et al., 1996) in about 250 genera (Round et al., 1990). Two groups are distinguished: bilateral symmetric, elongate "pennate" and radial symmetric, round "centric" diatoms. While the first fossil deposits for centric diatoms appeared in deposits from the Jurassic (~180 Myr), pennate diatoms are younger and fossil evidence dates back to the Late Cretaceous (~90 Myr) (Sims et al., 2006; Kooistra et al., 2007).

Both, pennate and centric diatoms can occur as solitary cells or chain-forming colonies and they inhabit nearly all aquatic environments on earth. Pennate diatoms are abundant in benthic epiphytic communities (e.g. seafloor habitats of coastal seas) (Kooistra et al., 2007) or polar sea ice (Horner, 1990). However, several pennate lineages have adapted to a pelagic lifestyle (Kooistra et al., 2007). In comparison, centric diatoms are usually more successful in the open water column and contribute the most successful group of planktonic diatoms (Kooistra et al., 2007). Generally, diatoms often dominate phytoplankton communities of polar ecosystems including the ice edge zone (Smetacek et al., 2002), making them key players in Arctic and Southern Ocean (Armbrust, 2009). They take part in major marine biogeochemical cycles of silicate (Treguer et al., 1995) and carbon (Smetacek, 1999), and serve as the basis of the polar food chain (Smetacek et al., 2004). In the Southern Ocean, diatoms contribute as much as two-thirds of the total ocean silica export (Treguer et al., 1995; Falkowski et al., 1998). Additionally, it is estimated that diatoms contribute ~40% of the annual marine primary production (Nelson et al., 1995; Bowler et al., 2010). However, due to their dominance in the polar environment, diatoms can be responsible for > 90% of primary production during blooms in the ice edge zone of the Ross Sea, Antarctica (Smith and Nelson, 1985; Wilson et al., 1986; Tsuda et al., 2003). Moreover, they outnumber other taxa in the extremely cold sea ice ecosystem in numbers and biomass (Wilhelm et al., 2006; Arrigo et al., 2010a) and can tolerate extreme changes in radiation, temperature

and salinity. They constitute > 90% of the photosynthetic diversity in sea ice, which is likely to exceed 500 species (Arrigo et al., 2010a). Generally, sea ice diatoms are dominated by pennate diatoms (Horner, 1990; Poulin, 1990; Lizotte, 2001; Thomas and Dieckmann, 2002; Smetacek and Nicol, 2005; Arrigo et al., 2010a). Pennate diatoms can contribute 80 – 90% of sea ice assemblages in the Arctic Ocean (Poulin, 1990) and Southern Ocean (Ligowski et al., 1992), though centric diatoms can also be common in polar sea ice (Arrigo et al., 2010a). Blooms of pennate diatoms have been reported in all parts of the ice column (Arrigo et al., 2010a) and the diatom biomass in sea ice can reach concentrations of up to 1000 µg of chlorophyll per litre (Thomas and Dieckmann, 2002; Arrigo et al., 2010a).

The dominant diatom genera do not differ much between both polar seas and dominant genera include e.g. *Thalassiosira*, *Chaetoceros*, *Eucampia*, *Fragilariopsis* and *Rhizosolenia*. Nevertheless, the species composition in both polar oceans is different and only few diatom species have a bipolar distribution (i.e., species occurring exclusively in both polar oceans and nowhere else) (Hasle, 1976; Lundholm and Hasle, 2008; Noethig et al., 2009). Due to its geological history and oceanographic conditions, most of the phytoplankton species in the Arctic can be found in other oceans and only 10 species are endemic to the Arctic Ocean (Heimdal, 1989). In contrast, the Southern Ocean has the largest percentage of endemic diatom species of any ocean region (Priddle and Fryxell, 1985), including at least six endemic planktonic diatoms (Zielinski and Gersonde, 1997). Two prominent large Southern Ocean diatoms *Corethron pennatum* and *Fragilariopsis kerguelensis* do not occur in the Arctic, while *Melosira arctica*, a species forming large filaments on the underside of Arctic sea ice, does not occur in the Southern Ocean (Noethig et al., 2009). Polar diatom assemblages are often characterised by high abundances of very large species, which is particularly relevant though not exclusive to the Southern Ocean (Smetacek et al., 2004; Noethig et al., 2009) and blooms of large chain-forming diatoms have been observed in the Arctic Ocean (Heimdal, 1989). In the iron-limited areas of the Southern Ocean the heavily silicifying large diatom species *F. kerguelensis*, *Thalassiothrix antarctica* and *Thalassiosira lentiginosa* are among the most prominent species. During iron-replete phytoplankton blooms in the Southern Ocean, the diatom species, *Thalassiosira antarctica*, *T. gravida*, *Chaetoceros socialis*, *C. curvisetus*, *C. debilis*, *C. neglectus*, *Rhiszosolenia hebetata*, *Proboscia alata*, *Corethron pennatum*, *Fragilariopsis curta* and *F. cylindrus* contribute most of the biomass (Smetacek et al., 2004). The latter *F.*

*cylindrus* can contribute 35% of total diatom abundances in water column assemblages and sea ice zones (Kang and Fryxell, 1992).

## 1.4 The species under investigation: *Fragilariopsis cylindrus*

The diatom *Fragilariopsis cylindrus* (Grunow) Krieger is a nanoplanktonic (2 – 20 µm) pennate diatom of ~20 µm in its longest dimension (Figure 4) and is common in the pack ice and near the ice edge of the Arctic and Antarctic Ocean (Grunow, 1884; Hasle, 1965; Garrison and Buck, 1989; Ligowski et al., 1992; Lundholm and Hasle, 2008). However, its bipolar distribution is uncertain (Lundholm and Hasle, 2008). *F. cylindrus* belongs to the group of raphid pennate diatoms, which contain a slit opening in the cell wall called raphe. The raphe permits raphid pennate diatoms to move actively by exudation of polysaccharide-rich mucilages and allowed this group to colonise unstable environments (e.g. sea ice), resulting in a rapid and massive radiation. They represent the largest group in terms of genera and species number within contemporary diatoms (Round et al., 1990; Sims et al., 2006; Kooistra et al., 2007). Although it has a raphe, the genus *Fragilariopsis*, including *F. cylindrus*, is nonmotile (Sims et al., 2006).

While it is estimated that the first raphid pennate diatoms appeared at ~50 Myr (Strelnikova and Simola, 1990), the genus *Fragilariopsis* emerged during the Oligocene (~30 Myr) (Sims et al., 2006; Kooistra et al., 2007) and adaptation of the taxonomic group of Bacillariaceae (again including *F. cylindrus*) to a pelagic existence is suggested to have occurred by the Miocene (22 – 5 Myr) (Sims et al., 2006). However, as for the evolutionary transition from centric to pennate diatoms, there is no substantial fossil record of the evolutionary transition from araphid (without raphe) to raphid diatoms and accurate evolutionary dating remains difficult (Sims et al., 2006). According to phylogenetic analysis (Lundholm et al., 2002) the genera *Fragilariopsis* and *Phaeodactylum*, which includes the model diatom *P. tricornutum*, diverged early during the radiation of raphid pennate diatoms in the Eocene (55 – 35 Myr) (Round et al., 1990; Strelnikova and Simola, 1990; Sims et al., 2006). In contemporary oceans, the genus *Fragilariopsis* consists of ~20 extant species (Cefarelli et al., 2010), and its representatives (including *F. cylindrus*) possess two chloroplasts (Hasle and Syvertsen, 1997).

*F. cylindrus* is ubiquitously found in phytoplankton counts of Antarctic water samples (Kopczynska, 2008) and is the most abundant diatom in phytoplankton

assemblages of Antarctic sea ice zones (Kang and Fryxell, 1992; Ligowski et al., 1992). It contributes up to 30% of total diatom abundances in surface ocean sediments of the Southern Ocean (Zielinski and Gersonde, 1997; Gersonde and Zielinski, 2000) and is proposed as an indicator species to reconstruct palaeooceanographic conditions (Gersonde and Zielinski, 2000; Quillfeldt, 2004). Due to its environmental importance, *F. cylindrus* has become a model for algal adaptation to polar marine conditions and diverse physiological, biochemical and molecular studies have been conducted.

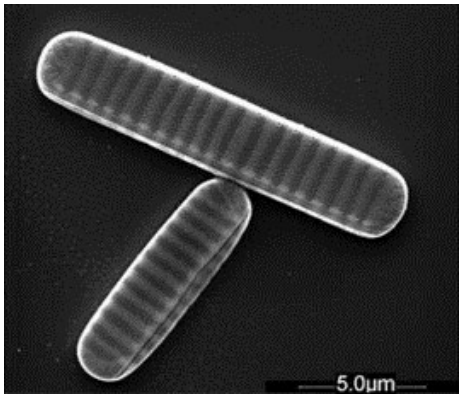



**Figure 4. Micrograph of *F. cylindrus* cells visualised by scanning electron microscopy (courtesy of Henrik Lange and Friedel Hinz, Alfred Wegener Institute for Polar and Marine Research, Bremerhaven, Germany).**

*F. cylindrus* is an obligate psychrophilic (cold-loving) organism and has an optimum growth temperature of +4 to +5 °C and an upper temperature limit of ≤ +10 °C (Fiala and Oriol, 1990). It is capable to grow in salinities of 100 $S_A$ (Bartsch, 1989), and accumulates osmoprotectants with increasing salinities including proline, betaine, homarine, and dimethylsulfoniopropionate (DMSP) as well as free amino acids (Plettner, 2002). In the presence of light, *F. cylindrus* shows photophysiological plasticity and is able to respond rapidly to changes in temperature and salinity that can occur during the shift between sea ice and pelagic conditions (Petrou and Ralph, 2011). Furthermore, it grows at low light intensities with minimum saturating irradiances of 60 µmol photons $m^{-2}$ $s^{-1}$ and optimal photosynthetic irradiances of 120 µmol photons $m^{-2}$ $s^{-1}$ (Pankowski and McMinn, 2009; Petrou and Ralph, 2011), but also at constant high irradiances, typical for shallow mixed layer depths and shows high levels of photoprotection (Kropuenske et al., 2009; Arrigo et al., 2010b; Mills et al., 2010; Alderkamp et al., 2012). However, growth and photosynthesis in *F. cylindrus* is challenged by hyperoxia (McMinn et al., 2005), which may be caused by high dissolved

oxygen concentrations and oxygen oversaturation in sea ice (Figure 2) (Mock et al., 2002; Trenerry et al., 2002). In the absence of light, as occurs in overwintering *F. cylindrus*, vegetative cells showed dark survival times of up to 60 days (Reeves et al., 2011). In summary, these physiological studies showed that *F. cylindrus* is well-adapted to sea ice and open water conditions (Quillfeldt, 2004; Petrou and Ralph, 2011). Thus, its survival and prolific growth under the extreme conditions of sea ice and the more moderate conditions of open waters require a complex suite of biochemical and molecular adaptations.

Molecular studies over the last 10 years have made *F. cylindrus* a popular model for polar eukaryotic genomics. *F. cylindrus* can acclimate photosynthesis over a wide range of polar water temperatures (Mock and Valentin, 2004; Mock and Hoch, 2005) and expression of genes encoding for Photosystem II as well carbon-fixing Rubisco subunits were similar at −1 °C and +7 °C after 4 month of acclimation (Mock and Hoch, 2005). Additionally, cold adaptation of *F. cylindrus* was investigated using an expressed sequence tag (EST) approach (Mock et al., 2005). It was found that most EST sequences were similar to genes encoding for proteins involved in translation, ribosomal structure, biogenesis, amino acid metabolism and post-translational modifications (Mock et al., 2005). In particular, DNA/RNA helicases, peptidases, ABC transporter protein domains were highly abundant (Mock et al., 2005) and these protein domains were suggested to be involved in various functions relevant for cold adaptation, including minimization of secondary structures and duplexes of mRNA to initiate protein translation as well as repair of photodamaged protein under freezing temperatures by activity of peptidases (Mock et al., 2005). However, more than half of EST sequences showed no similarity to known proteins (Mock et al., 2005). Similarly, in a salt stress-induced cDNA library of *F. cylindrus* > 30% of EST sequences produced no significant hit against any sequence database (Krell et al., 2008). Nevertheless, from analysis of the salt stress-induced cDNA library it was shown that genes encoding for proteins involved in proline synthesis, light-harvesting complexes (LHCs), protection against oxidative damage and antifreeze proteins (AFP) were expressed in *F. cylindrus* during salt stress (Krell et al., 2008). Additionally, the upregulation of specific key genes encoding for proteins involved in proline synthesis and accumulation of this osmoprotectant in *F. cylindrus* was also shown during multiple stresses combining salt stress and cold stress (Krell et al., 2007). Moreover, genes and proteins involved in DMSP synthesis were upregulated during salt stress, leading to the accumulation of the

osmoprotectant DMSP in *F. cylindrus* (Lyon et al., 2011). Due to the upregulation of LHC family proteins in a salt-stress induced cDNA library (Krell et al., 2008) and LHC family proteins including fucoxanthin chlorophyll a/c-binding proteins showed altered expression during cold stress (Mock and Valentin, 2004), hence a role of LHC proteins in stress acclimation in *F. cylindrus* was suggested (Krell et al., 2008). Additionally, the expression of AFP during salt stress helped to identify a new class of AFP, formerly unknown in photosynthetic eukaryotes (Janech et al., 2006; Krell et al., 2008).

Further gene-specific studies of genes encoding for AFP showed strong regulation under freezing temperatures and high salinities (Bayer-Giraldi et al., 2010) and AFP were shown to accumulate in the cells (Bayer-Giraldi et al., 2011). AFP in *F. cylindrus* belong to a multigene family of about 56 genes and isoforms (Krell et al., 2008) and likely consist of non-secretory as well as secretory AFP isoforms (Bayer-Giraldi et al., 2011; Uhlig et al., 2011). While non-secretory AFP might function as intracellular or cell-wall associated antifreeze (Uhlig et al., 2011), it has been suggested that secretory AFP may accumulate in a sheath of extracellular polymeric substances (EPS) and shape the microstructure of sea ice around the cell (Krembs et al., 2002; Bayer-Giraldi et al., 2011; Uhlig et al., 2011). Indeed, *F. cylindrus* increases yields of EPS under high salinities and ice formation was inhibited down to −12 °C by *F. cylindrus* cells, free EPS, and enhanced EPS content (Aslam et al., 2012).

Additionally, under the iron-replete conditions of sea ice (Sedwick and DiTullio, 1997; Edwards and Sedwick, 2001; Thomas, 2003; Lannuzel et al., 2007), *F. cylindrus* expresses the iron-sulfur protein ferredoxin (Pankowski and McMinn, 2009), which is involved in electron-transport systems in the cell including respiratory and photosynthetic electron transport. When iron availability decreases, *F. cylindrus* replaces the iron-binding ferredoxin with its functional analogue flavodoxin (Pankowski and McMinn, 2009), which instead relies on riboflavin 5'-phosphate as a cofactor (Roche et al., 1996). Furthermore, *F. cylindrus* uses the iron-storage protein ferritin to safely concentrate and store iron (Marchetti et al., 2009), thereby minimizing potential cell damage from reactive oxygen species and oxidative stress via iron-mediated Fenton chemistry (Imlay, 2008). As a result, *F. cylindrus* has a low half-saturation constant for iron with $0.51 \times 10^{-12}$ M (= 0.51 pM) total inorganic Fe (Pankowski and McMinn, 2009) and is highly competitive in high nutrient low chlorophyll areas of the Southern Ocean (Arrigo et al., 2010b; Mills et al., 2010; Alderkamp et al., 2012).

In summary, these physiological and molecular studies give insight into the adaption and responses of *F. cylindrus* to extreme conditions of sea ice, such as low temperatures, high salinities and fluctuating light conditions as well as open water conditions including high irradiances in shallow mixed water layers and low iron concentrations. Key findings from these studies were that *F. cylindrus* uses antifreeze proteins to cope with cold and salt stress (Janech et al., 2006; Krell et al., 2008; Bayer-Giraldi et al., 2010) and these serve different functions including antifreeze activity, inhibition of ice crystallisation, attachment to ice and retention of a liquid environment (Raymond, 2011). Moreover, under salt stress *F. cylindrus* accumulates the organic osmolytes proline (Plettner, 2002; Krell et al., 2007), betaine (Plettner, 2002) and dimethylsulfoniopropionate (Lyon et al., 2011) and recently high levels of the metabolite isethionic acid in *F. cylindrus* have also been suggested to contribute to the osmotic balance of the cells (Boroujerdi et al., 2012). Furthermore, during fluctuating light and temperature conditions *F. cylindrus* significantly regulates photosynthetic genes (Mock and Valentin, 2004; Mock and Hoch, 2005; Mock et al., 2005). Additionally, *F. cylindrus* encounters decreases in iron availability, as may occur during the melting of sea ice and transition to open water conditions, by using the iron-storage protein ferritin (Marchetti et al., 2009) and the replacement of the iron-demanding electron transport protein ferredoxin with flavodoxin (Pankowski and McMinn, 2009).

These studies reveal the high metabolic flexibility of *F. cylindrus* in acclimating to a wide range of environmental conditions. However, molecular studies have also revealed many unknown genes (Mock et al., 2005; Krell et al., 2008) and the molecular basis of adaptation of *F. cylindrus* and other polar eukaryotes remains largely unknown. Moreover, a large fraction of diatom genes are still functionally uncharacterised (Krell et al., 2008; Bowler et al., 2010). As described in this thesis, we sequenced the genome and transcriptome of *F. cylindrus* to get further insights into evolution and adaptation of a diatom to conditions of polar oceans.

## 1.5 Diatom genomics and transcriptomics

Genomics or genome science is the study of all nucleotides within a genome (the complete set of genes encoded in a cell) including genome structure, content and evolution (Gibson and Muse, 2004). More broadly defined, genome science also encompasses the analysis of gene expression (Gibson and Muse, 2004) represented in the transcriptome (the complete set of transcripts in a cell for a specific physiological

condition). Technological advances have revolutionised the biological sciences (Schuster, 2008) and genomics-enabled technology has recently also been applied in diatom research, providing a better understanding of diatom evolution, biology and ecology (Armbrust et al., 2004; Montsant et al., 2007; Oudot-Le Secq et al., 2007; Bowler et al., 2008; Oudot-Le Secq and Green, 2011; Lommer et al., 2012).

The evolution of diatoms is based on secondary endosymbiosis (dated to ~200 Myr at the Permian-Triassic boundary, (Medlin et al., 2000)), in which a non-photosynthetic heterotrophic eukaryotic host cell acquired a red algal plastid (Falkowski et al., 2004). Although the analysis of the two sequenced diatom genomes *T. pseudonana* and *P. tricornutum* provided support for a red algal secondary endosymbiont (Armbrust et al., 2004; Bowler et al., 2008), subsequent comparative genome analysis revealed many more genes from a green algal endosymbiont suggesting a cryptic plastid endosymbiosis in diatoms (Moustafa et al., 2009). Although the contribution of genes with green algal origin in diatom genomes has recently been challenged (Deschamps and Moreira, 2012), there is strong evidence that secondary endosymbiosis happened multiple times (Baurain et al., 2010) and that diatoms are derived from a serial secondary endosymbiosis (Sanchez-Puerta and Delwiche, 2008). Thus, diatoms may have had a plastid from a green alga, which was later replaced by a red algal plastid, leading to a complex evolution of diatom plastids (Petersen et al., 2006; Frommolt et al., 2008; Moustafa et al., 2009). As a result, diatom plastids consist of four distinct membranes, the inner two being derived from the red algal symbiont and outer two from the host cell. Subsequently, not only gene loss, but also gene transfer from the red algal symbiont (nucleus, mitochondria and plastid) to the host nucleus took place, leading to a reduction of the plastid genome as well as loss of the nucleus of the former photosynthetic organism. In consequence many transferred genes from the chloroplast or nucleus of the endosymbiont can be found in diatom genomes (Oudot-Le Secq et al., 2007).

In addition to endosymbiontic gene transfer from red and green algal plastids, horizontal gene transfer from bacteria appears to have also played a significant role in the evolution of diatom genomes, contributing up to 5% of the total gene content (Bowler et al., 2008; Lommer et al., 2012). Overall, genes from different partners of secondary endosymbiosis combined with bacterial genes acquired by horizontal gene transfer provided diatoms with novel metabolisms never found together before and included the coexistence of plant-like photosynthesis and animal-like mitochondrial

fatty acid oxidation as well as the urea cycle (Armbrust et al., 2004; Allen et al., 2006; Bowler et al., 2008). Additionally, identification of many genes important for prolific growth under particular conditions in the world oceans (e.g. genes involved in nutrient uptake and storage) provide insights into metabolic adaptations to marine environments. The discovery of entirely unexpected metabolic pathways (like the urea cycle) from whole genome sequences, despite intense research on the same metabolism in preceding years, underlines the importance of genome sequences as "hypothesis-generating machines" (Moran and Armbrust, 2007). However, only about 50% of genes in diatom genomes, especially diatom-specific genes, can be assigned with putative functions based on sequence homologies to model organisms and experimental data (Bowler et al., 2010).

To address this discrepancy, genome-wide gene expression studies have been conducted for diatoms in response to different environmental conditions, including nutrient limitation with nitrogen, silicon, iron and cobalamin (vitamin $B_{12}$) as well as limitations induced by carbon dioxide (high pH) and low temperatures (Allen et al., 2008; Mock et al., 2008; Bertrand et al., 2012; Lommer et al., 2012). Novel insights into genes of unknown function can be gained by analysis of specific responses to different environmental conditions and associations with known genes (Allen et al., 2008; Mock et al., 2008; Bertrand et al., 2012; Lommer et al., 2012). A study of iron-limited *P. tricornutum*, using a combination of gene expression profiling and metabolomic analysis, showed strong down regulation of carbon metabolism and photosynthesis (Allen et al., 2008). Genes encoding for plastid beta-carbonic anhydrase and phosphoribulokinase, both providing substrate to RubisCO, were strongly down regulated, suggesting that they play key roles in the regulation of diatom carbon metabolism. In contrast, gene clusters involved in iron uptake, including ferric reductase and a putative ferrichrome-binding protein, were strongly upregulated (Allen et al., 2008). In comparison, a genome-enabled microarray study of *T. pseudonana* revealed considerable overlap in the transcriptional responses to insufficient iron and silicic acid (Mock et al., 2008). Moreover, a recent genomic analysis of *T. oceanica*, assisted by massive parallel pyrosequencing of cDNA libraries from iron-deplete and iron-replete cultures, showed that *T. oceanica* down regulates genes involved in chlorophyll biosynthesis and photosynthetic carbon fixation, the Calvin cycle, photosynthetic subunit proteins and light harvesting proteins in response to iron starvation (Lommer et al., 2012). Conversely, cytochrome c oxidase, cytochrome b and subunits of the NADH

dehydrogenase, proteins associated with the mitochondrial respiratory chain were upregulated (Lommer et al., 2012). This transcriptome-derived knowledge supported the interpretation of the functional elements of the *T. oceanica* genome in response to chronic iron limitation typical of the open ocean (Lommer et al., 2012). Furthermore, RNA-Sequencing transcriptome analysis combined with metabolic profiling of *T. pseudonana* and *P. tricornutum* under limited cobalamin availability revealed three distinct strategies used by diatoms to cope with low cobalamin ($B_{12}$) availability, including upregulation of a novel cobalamin acquisition protein (Bertrand et al., 2012).

In summary, the analysis of the expression of genes encoding for proteins with known function provided novel mechanistic insights into metabolic responses to relevant environmental conditions in *P. tricornutum*, *T. pseudonana* and *T. oceanica*. Moreover, as shown by these genome-enabled studies, the association of genes with unknown function to environmental conditions in which they are expressed provides a useful mean to explore gene function and assist genome annotation (Maheswari et al., 2009). Finally, the genome-wide analysis of gene expression in the mesophilic diatoms *P. tricornutum*, *T. pseudonana* and *T. oceanica* provided insights into the molecular adaptation and acclimatisation to important environmental conditions, however this level of analysis is lacking for psychrophilic polar diatoms.

## 1.6 Aims of thesis

In general, little is known about genomic adaptations in polar eukaryotes, making the obligate psychrophilic diatom *Fragilariopsis cylindrus* an ideal candidate for a genome sequencing project. It was hypothesised that *F. cylindrus* contains significantly different genomic adaptations to thrive under polar ocean conditions in comparison to sequenced non-polar diatoms and that its genome is differentially regulated under different conditions. Thus, the initial core aims of the *F. cylindrus* sequencing project were to

1. generate and order genomic and expressed sequence tag (EST) sequences,
2. identify and annotate the complete set of genes,
3. characterise DNA sequence diversity,
4. establish an integrated web-based genome browser and research interface, and
5. to provide resource for comparative genomics.

Using comparative analysis with sequenced mesophilic diatoms, it was aimed to identify genes and structural changes of DNA that are necessary for eukaryotic photosynthetic organisms to live under polar conditions. Moreover, it was intended to assemble metabolic pathways and obtain new insights into how a polar diatom interacts with its environment. A core aim of this thesis research was to analyse metabolic pathways and physiological responses of *F. cylindrus* to important environmental changes using genome-wide expression profiling. High-throughput sequencing of cDNA (RNA-Seq) from *F. cylindrus* grown under six experimental conditions was used to compile comprehensive atlases of gene expression. Thereby, the aims were to analyse how the genomic information is used to acclimatise to important environmental conditions and to also obtain novel insights into genes of unknown function, identify novel genes involved in polar adaptation and improve genome annotation. Finally, the intention was to obtain functional data, including the biochemical properties of specific genes in *F. cylindrus* that emerged as adaptive key genes from genomic and transcriptomic analyses. The characterisation of a newly discovered proton-pumping bacteria-like rhodopsin, which showed strong gene expression under low iron, was chosen as an example of a molecular adaptation to the low iron conditions phytoplankton encounter in the high nutrient low chlorophyll regions of the Southern Ocean.

The scientific questions to answer were:

1. What genes and genomic features in the genome of a photosynthetic eukaryote are necessary to live under polar conditions?
2. What are the molecular responses of an obligate psychrophilic eukaryote to important environmental conditions and how flexible are these responses to global changes?
3. Can we use high-throughput transcriptome sequencing of *Fragilariopsis cylindrus* to quantify genome-wide expression, identify novel protein coding genes and refine gene boundaries?
4. What are the physiological roles and functions of specific key genes in adaptation to polar conditions including low iron?

## 1.7 Outline of thesis

The work presented in this thesis is divided into three major parts: (1) Sequence analysis and annotation of the draft genome sequence of *F. cylindrus*, (2) high-throughput transcriptome sequencing of *F. cylindrus* transcriptomes under different environmental conditions and (3) characterisation of a new bacteria-like rhodopsin identified in the *F. cylindrus* genome.

In **chapter 2** I describe the laboratory and computational methods used in this work. The aim is to provide an overview of how the genomic data of *F. cylindrus* compiled in this work has been generated and analysed. Therefore, I first describe methods used for sequencing and annotation of the *F. cylindrus* genome, followed by a description of material and methods, which are based on the experimental work performed in the framework of this thesis.

**Chapter 3** reports the genome structure, gene content and deduced metabolic capacity of *F. cylindrus*. The main objectives were to identify and annotate the complete set of genes, assemble metabolic pathways and identify genes and structural changes of DNA that are necessary to live under polar conditions using comparative genomic analysis. For this purpose, I describe general characteristics of the *F. cylindrus* genome, as revealed from whole genome sequencing and a comparative analysis with other sequenced diatoms and phytoplankton. I describe putative adaptations to polar marine conditions and set the stage for the following chapters.

In **chapter 4** I present results from a genome-wide expression analysis of *F. cylindrus* using RNA-Seq. The objective was to determine and quantify the molecular responses of *F. cylindrus* to important environmental conditions, which is essential for interpreting the functional elements of the genome. For this purpose, the transcriptomes of *F. cylindrus* grown under optimal polar summer growth conditions (nutrient replete, +4 °C, 35 µmol photons $m^{-2}$ $s^{-1}$), freezing temperatures (−2 °C), elevated temperatures (+10 °C), elevated carbon dioxide (1000 ppm $CO_2$), iron starvation (−Fe) and prolonged darkness (one week darkness) were analysed. I give an overview of the *F. cylindrus* transcriptome providing further understanding of molecular mechanisms of acclimatisation to environmental stresses and highlight selected metabolic pathways and genes involved in acclimation to prolonged darkness.

In **chapter 5** the functional characterisation of a bacteria-like rhodopsin from *F. cylindrus* is documented. A bacteria-like rhodopsin in *F. cylindrus* was a novel discovery during the genome analysis and it was found to be strongly expressed under low iron conditions using RNA-Seq transcriptome sequencing. Although bacteria-like rhodopsins have recently been identified to be very abundant in eukaryotic marine phytoplankton, the function and role of putative light-driven proton-pumping rhodopsins in the presence of a proton gradient-generating chlorophyll-based photosynthetic apparatus remains speculative. Thus, the *Fragilariopsis* rhodopsin provided an ideal model to study molecular adaptations to low iron conditions of the Southern Ocean and was chosen over other putative "ice-specific proteins". The gene expression of the *Fragilariopsis* rhodopsin was studied under silicate limitation as well as red and blue light conditions (in addition to the above described six transcriptome conditions) using RT-qPCR. Additionally, functional analysis using heterologous expression systems including the diatom *P. tricornum* and *Xenopus laevis* oocytes are described. Possible hypotheses for its physiological role are provided.

Finally, in **chapter 6** I conclude with a summary of the major findings of the research and a general discussion of the main results. I draw on all data generated in this work and attempt to give an integrative view on the ecological and evolutionary implications together with future research perspectives.

# Chapter 2

# Materials and methods

## 2.1 Genome sequencing and computational analysis

Genome sequencing and computational analysis of the *F. cylindrus* genome was performed in an international collaborative effort by the *Fragilariopsis* genome consortium (see p. XI). Genomic DNA for sequencing was extracted by Thomas Mock using a cationic surfactant cetyltrimethylammonium bromide (CTAB)-based extraction protocol modified from Friedl (1995) (Supplementary protocol S1). The genome sequencing data presented in this work has been obtained by my own custom analyses as well as significant contribution by other consortium members, in particular Andrew E. Allen, Christoph L. Dupont, Stephan Frickenhaus, Beverley E. Green, Florian Leese, Florian Maumus, Christoph Mayer, Robert P. Otillar, James A. Raymond, Remo Sanges, Jeremy Schmutz, Andrew Toseland, Christiane Uhlig and Ruben E. Valas as indicated below.

### 2.1.1 Genome sequencing and assembly

The general process eukaryotic genome sequencing and annotation is outlined in Yandell & Ence (2012). The *Fragilariopsis cylindrus* Krieger (CCMP 1102) draft genome was sequenced with 8-fold average coverage at the U.S. Department of Energy Joint Genome Institute (JGI, http://www.jgi.doe.gov/, Walnut Creek, CA, USA) using a whole-genome shotgun approach. During the shotgun sequencing approach the genome was fragmented into small, sequenceable units, which were assembled into contigs derived from overlapping sequences using computer algorithms. The genome sequence assembly v1.0 was built by Jeremy Schmutz (JGI) using Arachne assembler from whole genome shotgun and paired end sequencing reads. The *F. cylindrus* genome was annotated using the JGI Genome Annotation Pipeline by Jeremy Schmutz, Robert P. Otillar and Igor V. Grigoriev. The Mauve Genome Alignment Software (available at http://gel.ahabs.wisc.edu/mauve/download.php) was used by Robert P. Otillar to estimate SNP/polymorphism rates between two aligned regions of putative homologous chromosomes. Additionally, polymorphism was determined by Jeremy Schmutz using read depth coverage analysis of the assembled genome sequence. Finally, custom

analyses were performed by the *Fragilariopsis* genome consortium (see p. XI), including my own analyses.

### 2.1.2 Sequence analysis

To perform my own customised sequence analysis, the current JGI (http://www.jgi.doe.gov/) diatom sequencing project for the polar diatom *Fragilariopsis cylindrus* v1.0 (http://genome.jgi-psf.org/Fracy1/Fracy1.home.html) was searched using the BLAST algorithm (Altschul et al., 1997). Comparison with the genome sequences of the diatoms *Thalassiosira pseudonana* (http://genome.jgi.doe.gov/Thaps3/Thaps3.home.html) (Armbrust et al., 2004) and *Phaeodactylum tricornutum* (http://genome.jgi-psf.org/Phatr2/Phatr2.home.html) (Bowler et al., 2008) as well as with other publicly available sequences helped to delimit gene models.

In many cases an N-terminal targeting domain can direct proteins into the endoplasmic reticulum (ER), mitochondria, plastids, peroxisomes and the extracellular space or to other cell compartments. The subcellular location of proteins was predicted using the program TargetP (http://www.cbs.dtu.dk/services/TargetP/) (Emanuelsson et al., 2000) using non-plant networks (Nielsen et al., 1997) and mitochondrial transit peptides were identified. Signal peptides of the endoplasmic reticulum (ER) proteins were identified using the web tool SignalP 4.0 (http://www.cbs.dtu.dk/services/SignalP/) (Petersen et al., 2011). In addition it was checked manually if protein sequences contain a C-terminal retention signal (KDEL, HDEL, DDEL or DEEL) (Pagny et al., 1999). Plastid proteins of diatoms possess bipartite targeting signals consisting of a signal peptide and a transit peptide domain with a conserved sequence motif at the signal peptide cleavage site (Kilian and Kroth, 2005; Gruber et al., 2007). Sequences were screened for signal peptides using SignalP 4.0 (http://www.cbs.dtu.dk/services/SignalP/) and cleavage site predictions were performed using a eukaryotic neuronal-network based method (Petersen et al., 2011). For prediction of chloroplast transit peptide-like domains the program ChloroP (http://www.cbs.dtu.dk/services/ChloroP/) (Emanuelsson et al., 1999) was used. In some cases transit peptide-like domains of plastid proteins are also recognised as mitochondrial transit peptides by the program TargetP. Proteins were considered to be plastid targeted if they possess a signal peptide but no ER retention signal, possess a N-terminal extension longer than the signal peptide with transit peptide features and contain the amino acids F, W, Y or L at the signal peptide cleavage site,

because mutational analysis showed that only the amino acids phenylalanine (F), tryptophan (W), tyrosine (Y) and leucine (L) at the +1 position of the predicted signal peptidase cleavage site allow plastid import (Gruber et al., 2007). Peroxisomal proteins possess conserved peroxisome targeting signals (PTS) (Lanyon-Hogg et al., 2010). Sequences were screened for the consensus sequences PTS1 and PTS2 using the web based application PTS1Prowler (http://pprowler.itee.uq.edu.au) (Hawkins et al., 2007). Putative enzymes without recognisable targeting sequences were considered cytosolic although it cannot be excluded that they might be co-translocated by association with targeting sequence-containing proteins or be targeted to further cellular compartments. For a detailed description of tools for protein localisation prediction see also (Emanuelsson et al., 2007).

Gene identification and functional analysis was facilitated by transcriptome sequencing of *F. cylindrus* cells grown under six different growth conditions using RNA-Sequencing at the British Natural Environment Research Council (NERC) Biomolecular Analysis Facility (NBAF GenePool, Edinburgh, UK) (see Chapter 4).

### 2.1.3 Identification, annotation and classification of repeats

Florian Maumus and Hadi Quesneville identified and annotated repeats in the *F. cylindrus* genome. Additionally, Christoph Mayer and Florian Leese performed specific analyses of the tandem repeat content in *F. cylindrus* and comparative analyses of tandem repeats.

The repeated sequences present in *F. cylindrus* were identified and annotated using the TEdenovo pipeline from the REPET package (http://urgi.versailles.inra.fr/Tools/REPET) that integrates a combination of de novo and similarity-based approaches (Quesneville et al., 2005). At first, high-scoring segment pairs (HSPs) were identified by comparing the whole *F. cylindrus* genome to itself using the program BLASTER. HSPs were clustered using the GROUPER, RECON, and PILER programs, and groups comprising at least three sequences (n = 1,421) were retained for further analysis. Clusters of sequences were then aligned using the MAP algorithm and multiple sequence alignments were used to derive a consensus sequence for each cluster. In a second step, the set of consensus sequences were aligned on the *F. cylindrus* genome using TEannot pipeline from the REPET package which combines the RepeatMasker, BLASTER, and CENSOR programs. MATCHER was used to handle

overlapping HSPs and to make connections (also called defragmentation). In addition, low-copy and degenerate TEs were searched in the whole genome by comparison with the Repbase database using BLASTER with tBLASTx and BLASTx. Furthermore, the whole genome was screened for SSRs using Tandem Repeats Finder (TRF), Mreps, and RepeatMasker (using Repbase SSR library), and results from the three programs were merged (the same method was used to estimate microsatellite coverage in the *P. tricornutum* and *T. pseudonana* genomes). SSRs were removed when included into TE annotations and TE doublons were purged from annotation files. Finally, locally co-linear annotations of the same consensus were recovered and joined using the 'long join' procedure if the fragments were of similar age and interrupted by younger TE insertions. A GFF3 annotation file comprising all the repeats detected during this analysis is available for insertion in the JGI genome browser. Each repeat consensus was analyzed using a tool called PASTEC designed to support the process of automatic repeat classification. PASTEC combines three complementary approaches to detect a variety of features in the consensus sequences: i) screen for structural features characteristic of transposable elements (TEs) such as long terminal repeats (LTRs), terminal inverted repeats (TIRs), and polyA tails, as well as for the presence of simple sequence repeats (SSRs) using TRF; ii) search for similarity with known nucleic and amino acid TE sequences deposited in Repbase (http://www.girinst.org/) using BLASTx, tBLASTx, and BLASTn; iii) probe for virtually all hidden Markov models (HMMs) from Pfam annotation database using HMMER. The bank of HMMs was modified to distinguish between two classes of Pfam annotations: TE-specific or not (host gene-specific). According to the features detected, PASTEC proposes an automated classification of the input sequences. In an effort to improve TE classification, it was attempted to manually construct a library of *F. cylindrus*-specific TEs. Thereby, LTR FINDER was used with whole genome as input in order to identify full length LTR-retrotransposons sequences in the genome. Consensus sequences from REPET output were screened for similarity with TEs referenced in the Repbase and home-made databases using BLASTx and tBLASTx. The results were manually curated and worked to compile a library of *F. cylindrus* reference TEs comprising Class1 and Class 2 elements including sequences classified as Ty1/copia, Ty3/gypsy, DIRS, and LINE (including Ambals), and PiggyBac, Harbinger, and MuDR, respectively. These nucleotide sequences were appended to the Repbase library to launch PASTEC. In addition, transfer and ribosomal RNA genes were searched in the *F. cylindrus* genome sequence using the tRNAscan-SE and RNAmmer programs, respectively, and compared

to the consensus sequences using BLASTn. The features collected from each consensus sequence were subsequently examined and used as a support for the manual curation of the results obtained from automated classification with PASTEC.

Additionally, tandem repeats were detected using the software Phobos (v.3.3.12) and results were analysed with the sat-stat software (v.1.3.12). The Phobos search parameters allowed for moderate gegree of imperfection in the repeats and searches were carried out for a unit size of 1 – 50 bp, which spans microsatellites as well as a large part of the size range of minisatellites (7 bp – 100 bp). Phobos search parameters were as follows: match score 1, mismatch and indel score −5, N score 0. The first unit was not scored and a maximum of two successive Ns were allowed in a tandem repeat. Tandem repeats were required to have a minimum score of 12 or the unit length.

### 2.1.4 General annotation of protein families

The general annotation of protein families in *F. cylindrus* was performed by Andrew E. Allen and Ruben E. Valas. To annotate protein families, a BLASTP search of all 27,137 predicted gene models (Filtered Models1) was performed using an *E* value cut-off of $1\times10^{-9}$ and results were compared to searches performed with filtered gene model sets from *T. pseudonana* and *P. tricornutum* to define genes of a diatom core genome. Subsequently, the OrthoMCL Database pipeline (http://www.orthomcl.org/cgi-bin/OrthoMclWeb.cgi) was used to define orthologs and paralogs. A four section venn diagram was created after clustering BLASTP results of all three diatoms with results obtained for all currently available red and green algae genomes. Finally, a three section venn diagram was constructed by performing a BLAST search of the diatom cluster against the phylodb_1.04 database.

### 2.1.5 Annotation of metal-binding protein families

In addition to the general annotation of protein families, specific metal-binding protein families were annotated by Christoph L. Dupont in a separate analysis. Therefore the Structural Classification of Proteins (SCOP) data base was utilised. SCOP version 1.75 included 38,221 Protein Data Bank (PDB) three dimensional structures sorted into a hierarchy encompassing class, folds, fold superfamilies (FSF), fold families (FF), and domains. The Superfamily database provided hidden Markov models for each of the FSFs and FFs within the SCOP database that could be used to annotate protein sequences according to structural domain composition. The collection of

phytoplankton and *Phytophthora* genomes within PHYTAX were analysed using the Superfamily HMMs and HMMER3. Metal annotations of the SCOP database were built upon those of Dupont *et al*. (2006; 2010). 600 new FFs and FSFs had been added to SCOP since this annotation; these were manual curated according to metal binding. This manual curation involved an examination of the structure, and where possible, the literature associated with it. Particular attention was paid to the domain boundaries in proteins; many metal FFs recombine with non-metal FFs. An automated annotation of metal binding by SCOP FFs from the Procognate database was compared to the manual annotations. Generally the two annotations agreed (30 mismatches out of 2602 FFs). Occassionally, Procognate assigned Mg as a ligand when the actual element was Mn. In select cases, a non-native metal was used in the crystallization matrix; here, literature surveys resolved the disagreement. Assignments made by KEGG or GO contained upwards of 50% false positives/negatives. In total, 602 FFs were designated as metal binding. If all structures within a FF bound the same metal, it was described as X-metal binding, with X being Fe, Zn, Cu, Co, Ni, Mo, Ca, or Mn. In the circumstances where the structures within a FF bind different metals, that FF was categorized cambialistic.

## 2.2 Experimental work

### 2.2.1 Phytoplankton strains, media and growth conditions

*Fragilariopsis cylindrus* (Grunow) Krieger CCMP 1102 and the polar dinoflagellate *Polarella glacialis* Montresor, Procaccini et Stoecker CCMP 2088 were obtained from the Provasoli-Guillard National Centre for Marine Algae and Microbiota (NCMA, https://ncma.bigelow.org/, West Boothbay Harbor, ME, USA, formerly CCMP). *F. cylindrus* was grown and maintained in filter-sterilised (0.2 µm pore size) Aquil artificial seawater medium (Morel et al., 1979; Price et al., 1988/89), which had been adjusted to pH 8.1 – 8.4 prior to use, while *P. glacialis* was grown and maintained in filter-sterilised (0.2 µm pore size) L1 artificial seawater medium (Guillard and Hargraves, 1993) as modified from f/2 medium (Guillard and Ryther, 1962; Guillard, 1975), instead. Polar phytoplankton cultures were grown at 4 °C under continuous illumination at a photon flux density of approximately 35 µmol photons $m^{-2}$ $s^{-1}$ (QSL 2101, Biospherical Instruments Inc., San Diego, CA, USA) from cool white fluorescent tubes. Cell cultures were handled under strict sterile conditions and potential bacterial contamination was eliminated as stock cultures were subjected to a multi-antibiotic treatment with ampicillin (50 µg $mL^{-1}$), gentamycin (1 µg $mL^{-1}$), streptomycin (25 µg

$mL^{-1}$), chloramphenicol (1 µg $mL^{-1}$) and ciprofloxacin (10 µg $mL^{-1}$) (Jaeckisch et al., 2011). Fluorescence microscopy combined with 4',6-diamidino-2-phenylindole (DAPI) fluorescent nucleic acid staining was used to confirm axenic cultures before the beginning of culture experiments. Therefore, 1 – 5 mL of cell cultures were fixed with 0.2 µm-filtered solutions of 3 µL $mL^{-1}$ Lugol's iodine (aqueous KI 10% w/v and iodine 5% w/v), 50 µL $mL^{-1}$ neutralised formalin (20% aqueous formaldehyde with 100 g $L^{-1}$ hexamine), followed by destaining of fixed cell mixtures with 6 µL $mL^{-1}$ 3% w/v sodium thiosulfate. DAPI staining was performed by adding 10 µL $mL^{-1}$ DAPI solutions (1 mg $mL^{-1}$) and incubation in the dark for 15 min. For visulalisation of DAPI-stained cells, samples were vacuum-filtered onto 0.2 µm pore black polycarbonate filter (Millipore), backed with 0.45 µm cellulose nitrate filter. DAPI filters were examined for axenity under UV light using an epifluorescence microscope (Olympus BX40-F equipped with Olympus U-RFL-T-200 high pressure mercury burner, Olympus Corp., Tokyo, Japan).

*P. glacialis* was grown in batch cultures to stationary phase under nutrient-replete and continuous light conditions before sampling for RNA preparations. *F. cylindrus* experimental batch cultures were grown in three biological replicates in chemically defined Aquil artificial seawater media using a temperature and light controllable incubator (RUMED light thermostate type 1301, Rubarth Apparate GmbH, Laatzen, Germany). *F. cylindrus* cultures were subjected to nine different experimental treatments including (1) optimal growth (+4 °C, nutrient replete, 24 h light at 35 µmol photons $m^{-2}$ $s^{-1}$), (2) freezing temperatures (−3 °C, nutrient replete, 24 h light at 35 µmol photons $m^{-2}$ $s^{-1}$), (3) elevated temperatures (+11 °C, nutrient replete, 24 h light at 35 µmol photons $m^{-2}$ $s^{-1}$), (4) elevated carbon dioxide (+4 °C, 1000 ppm CO2, 24h light at 35 µmol photons $m^{-2}$ $s^{-1}$), (5) iron starvation (+4 °C, −Fe, 24 h light at 35 µmol photons $m^{-2}$ $s^{-1}$), (6) prolonged darkness (+4 °C, nutrient replete, 7 d darkness), (7) half-saturation with silicate (+4 °C, 0.3 µM silicate, 24h light at 35 µmol photons $m^{-2}$ $s^{-1}$) as well as (8) red (+4 °C, nutrient replete, 24 h light at 35 µmol photons $m^{-2}$ $s^{-1}$, 550 – 700 nm colour filter) and (9) blue light illumination (+4 °C, nutrient replete, 24h light at 35 µmol photons $m^{-2}$ $s^{-1}$, 480 – 540 nm colour filter). *F. cylindrus* stock cultures from exponential growth phase were used to inculate three replicates of 2 L experimental batch cultures with an initial cell count of 50,000 cells $mL^{-1}$. During experimental treatments (except elevated $CO_2$ treatment), cultures were bubbled with filtered ambient air (Swinnex unit equipped with 25 mm Whatman GF/F filter) passed through milliQ-

$H_2O$ and manually shaken before subsampling to ensure sufficient $CO_2$ supply and mixing. Subsamples were taken on a daily basis throughout the experiments to determine physiological parameters including specific growth rate and maximum quantum yield of photosystem II ($F_v/F_m$) as a proxy for cell fitness (Parkhill et al., 2001). Cell counts were determined using automated cell counting with a Multisizer 3 particle counter (Beckman Coulter, Brea, CA, USA) equipped with a 100 µm aperture capillary. Specific growth rates per day (µ) were calculated from the linear regression of the natural log of cell counts versus time during the exponential growth phase or (when using only two sampling points) according to the following formula:

$$\mu\ (d^{-1}) = \frac{\ln(C_1) - \ln(C_0)}{t_1 - t_0},$$

where $C_1$ denotes the cell concentration at time $t_1$, $C_0$ the cell concentration at time $t_0$, and $t_1 - t_0$ the time difference in days between sampling intervals. The maximum quantum yield of photosystem II ($F_v/F_m$) was measured using pulse-amplitude-modulated (PAM) fluorometry, using a Phyto-PAM fluorometer equipped with a Phyto-ED measuring head (Walz GmbH, Effeltrich, Germany). The *in vivo* quantum yields were determined in each culture and calculated using PhytoWin software (v2.00a; Walz GmbH) from fluorescence readings of dark acclimated samples as follows:

$$F_v/F_m = (F_m - F_o)/F_m\ ,$$

where $F_m$ and $F_o$ denote the maximum and minimum fluorescence level (Maxwell and Johnson, 2000). Additionally, pH was measured in each sample using a conventional pH meter (Jenway 3150, Bibby Scientific Ltd., Staffordshire, UK).

Whilst experimental teatments of *F. cylindrus* with elevated carbon dioxide, red light illumination and blue light illumination were instantly applied to *F. cylindrus* cultures, cultures grown under prolonged darkness, freezing temperatures and elevated temperatures were first grown to early-exponential phase at optimal growth conditions before sudden shifts to the final experimental condition (i.e., darkness, +11 °C and −3 °C). These experimental treatments were initiated during early exponential phase when cultures had cell density of approximately 300,000 cells per mL. Different blue and red light spectra were created by wrapping culture vessels in commercial colour filters (172 Lagoon Blue/025 Sunset Red, LEE Filters Worldwide, Andover, UK). Light spectra

were confirmed using a spectroradiometer (SR9910, Macam Photometrics Ltd., Livingston, UK). For low iron treatments, *F. cylindrus* was grown from iron-replete conditions in iron-free Aquil media that had been passed through a Chelex cation exchange column (Chelex 100 Resin, biotechnology grade sodium form, 100–200 dry mesh size, 150–300 µm wet bead size, Bio-Rad Laboratories, Hercules, CA, USA). Therefore, cells from iron-replete stock cultures were transferred into iron-free Aquil media and allowed to grow for several days prior to experimentation to ensure iron limitation as performed previously (De La Rocha et al., 2000). Preparation of iron iron-free Aquil media and handling of low iron cultures were carried out using standard trace metal clean techniques as described for trace metal studies (Fitzwater et al., 1982; Price et al., 1988/89; Sunda et al., 2005). Accordingly, 2 L aliquots of Aquil seawater were supplemented with macronutrients ($NO_3$, $PO_4$ and $Si(OH)_4$ in accordance with Aquil medium concentrations), passed through a Chelex cation exchange column, filter-sterilised (nitrocellulose membrane filter, 47 mm 0.22 µm GSWP, Millipore, MA, USA) and placed into 10% hydrochloric acid-cleaned, milli-Q $H_2O$-rinsed 2.5 L polycarbonate bottles. Trace metal concentrations were buffered using 100 µmol $L^{-1}$ of ethylenediaminetetraacetic acid (EDTA), which reacts with metal ions (including $Fe^{3+}$) to metal chelates that are not directly available to phytoplankton, rendering potential iron contaminations insignificant (Supplementary note S1). Dispensed chelexed and filter-sterilised Aquil seawater was supplemented with filter-sterilised (25 mm 0.2 µm syringe filter) EDTA-trace metals (minus iron) and vitamins ($B_{12}$, thiamine and biotin), and allowed to equilibrate chemically overnight at final growth conditions before inoculation of cells. For batch culture growth of *F. cylindrus* under half saturation with silicate, silicate was added back to a final concentration of 0.3 µmol $L^{-1}$ to the cultures on a regular basis during the experiment. The half-saturation constant $K_m$ of *F. cylindrus* for silicate was determined in a preliminary experiment, growing cells over a concentration range of 0.01 – 100 µmol $L^{-1}$ silicate (Supplementary Figure S1).

Experimental *F. cylindrus* cultures were sampled for RNA preparations during mid-exponential phase (approximately 500,000 cells $mL^{-1}$) after several days of acclimation to the experimental treatment by gentle filtration of cultures (~300 psi vacuum pressure) onto 1.2 µm membrane filters (Isopore membrane, Millipore, MA, USA), placement in 2 mL cryogenic centrifuge tubes and flash-freezing in liquid nitrogen. Finally, the limiting effect of experimental treatments on *F. cylindrus* was confirmed according to La Roche et al. (1993), which is based on addition of the

experimental factor to reconstitute optimal growth conditions leading to the recovery of physiological parameters that are depressed by the experimental treatment.

### 2.2.2 RNA preparation

Total RNA was extracted using guanidinium thiocyanate-phenol-chloroform extraction according to Chomczynski & Sacchi (1987) and TRI Reagent (Sigma-Aldrich, St. Louis, MO, USA) (Supplementary protocol S2), followed by DNase I (Quiagen, Hilden, Germany) treatment (1 h, 37 °C) and purification using RNeasy MiniElute Cleanup Kits (Quiagen, Hilden, Germany) according to the manufacturer's instructions. Purity of RNA was checked on a NanoDrop (Thermo Fisher Scientific, Waltham, MA, USA) and integrity using 2% denaturating formaldehyde gels or an Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA), respectively. RNA concentrations were determined in duplicate readings using a NanoDrop.

### 2.2.3 Transcriptome sequencing and computational analysis

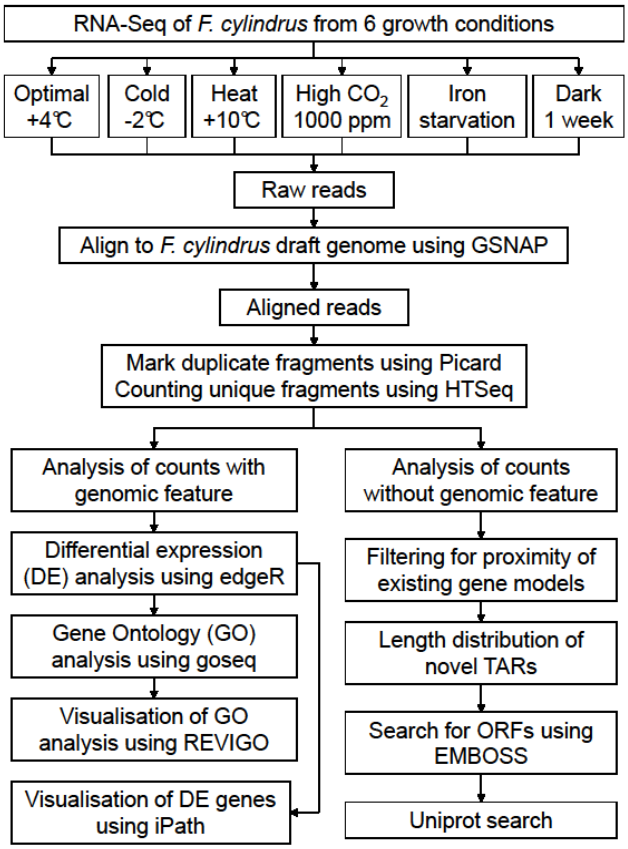



**Figure 5. Overview of the RNA-Seq analysis steps of *Fragilariopsis cylindrus*.**

#### 2.2.3.1 Library preparation and Illumina sequencing

Library preparation and Illumina sequencing was performed at the National Environmental Research Council (NERC) Sequencing Facility "The GenePool" (University of Edinburgh, UK) by technical staff. Triplicate samples of *F. cylindrus* grown under six different experimental conditions (4.2.1) were sequenced on an Illumina HiSeq 2000 platform. All samples were run in a single lane of a flowcell using multiplex DNA barcodes, generating paired-end reads of 101 bases length. Sequencing was conducted according to the Illumina TruSeq RNA Sequencing protocol. RNA-Seq libraries were prepared using the RNA-Seq Sample Prep Kit (Illumina). First strand cDNA synthesis was performed with random hexamers and reverse transcriptase. Following library construction, each molecule was sequenced in high-throughput manner to obtain short sequence reads. The number of reads over a genomic feature was a measure of its level of expression.

#### 2.2.3.2 RNA-Seq read mapping

RNA-Seq read mapping was performed in collaboration with bioinformatics support from "The GenePool" sequencing facility (University of Edinburgh, UK), and particularly Gaganjot Kaur, who performed most of the bioinformatics analyses. In a first step, all sequenced reads were aligned to the *F. cylindrus* genome assembly (http://genome.jgi-psf.org/Fracy1/download/portalData/Fracy1_assembly_scaffolds.fasta.gz) using the Genomic Short-read Nucleotide Alignment Program (GSNAP, version 2011-03-28) (Wu and Nacu, 2010), which supports alignment of spliced reads. To assist mapping of spliced reads, splice sites were extracted from the Filtered Models 1 annotation file (http://genome.jgi-psf.org/Fracy1/download/portalData/Fracy1_GeneModels_FilteredModels1.gff.gz) and given to GSNAP. The Java-based command-line utilitiy tool Picard (v1.55) was used to mark duplicate fragments using its "MarkDuplicates" function. A digital gene expression analysis was carried out on the uniquely and concordantly mapped reads with fragments mapping uniquely and both reads mapping properly in pair. Reads that failed these criteria were excluded. The Python package HTSeq (v0.5.3p1, http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html) was used to count unique fragments mapping in each genomic feature using the intersection-nonempty mode. Briefly, intersection-nonempty mode determines how to deal with reads overlapping

more than one feature. In intersection-nonempty mode reads that map to the region of overlap between two or more features are not counted. Uniquely and concordantly mapping fragments were used to generate a count table.

#### 2.2.3.3 Differential expression analysis of RNA-Seq data

Differential gene expression analysis was performed on original digital expression counts using the R Bioconductor package EdgeR (Robinson et al., 2010). Prior to analysis, weakly expressed genes were filtered from the data set and analysis was performed on transcriptionally active regions. A gene model was considered transriptionally active if its sum of counts (row) that mapped to the model in one or more libraries (columns) was greater than 0. The default "trimmed mean of M values" (TMM) normalisation method was used to calculate the effective library size and to avoid RNA composition biases, which are caused by highly abundant transcripts. Calling of differentially expressed genes was performed using EdgeR's generalized linear models (glm) functionality and a pair-wise multiple comparison between treatment conditions was made.

#### 2.2.3.4 Gene Ontology analysis of RNA-Seq data

A Gene Ontology (GO) analysis was performed on the differentially expressed genes using the R Bioconductor package goseq (Young et al., 2010). The gene length and GO term mappings were extracted from the *F. cylindrus* filtered model1 annotation file using customised Perl scripts (courtesy of A. Toseland) and given to goseq. The GO term analysis was performed individually on the following data sets: (1) all differentially expressed genes, (2) upregulated genes and (3) down regulated genes. Testing for overrepresented GO terms was performed using the default Wallenius approximation method and overrepresented GO terms were selected using a 0.05 false discovery rate (FDR) cut-off (Benjamini and Hochberg, 1995). The produced lists of overrepresented GO terms were summarized by removing redundant GO terms and visualised in semantic similarity-based scatterplots using the webserver tool Revigo (Supek et al., 2011) (available at http://revigo.irb.hr/).

#### 2.2.3.5 Identification of putative novel protein coding genes

Identification of putative novel protein coding genes was performed in collaboration with bioinformatics support from "The Gene Pool" sequencing facility

(University of Edinburgh, UK) and Andrew Toseland (University of East Anglia, Norwich, UK), who assisted with the bioinformatics analyses. In a first step, the genomic coordinates of regions which were covered by reads but did not contain genomic features in the GFF annotation file were extracted. Subsequently, these regions were filtered for proximity of adjacent gene models (cut-off ≤ 250 nt from existing gene model), which were considered to belong to 5' and 3' untranslated regions of inaccurately predicted gene models. Genomic regions meeting our cut-off criteria were searched for novel protein-coding genes using the ORF detection tool from the EMBOSS package (available at http://emboss.sourceforge.net/apps/cvs/emboss/apps/getorf.html). Parameters were set to report the translations of any ORFs ≥ 100 nt between stop codons, checking both forward and reverse strands. A BLASTP search against the Swiss-Prot database (available at http://www.uniprot.org, BLASTP *E*-value ≤ $1e^{-3}$, and allowing up to 10 hits per sequence) was performed for detected ORFs.

#### 2.2.3.6 Interactive pathway analysis of RNA-Seq data

A cellular pathway analysis was performed using the interactive pathway explorer iPath2.0 (available at http://pathways.embl.de) (Yamada et al., 2011) to visualise cellular pathways represented by differentially expressed genes. The Enzyme European Commission (EC) number annotation and KEGG pathway mapping were extracted from the *F. cylindrus* functional annotation file (Fracy1_ecpathwayinfo_FilteredModels1.tab.gz) available at http://genome.jgi-psf.org/Fracy1/Fracy1.download.ftp.html, using a customised Perl script (courtesy of A. Toseland) and assigned to differentially expressed genes in each individual treatment (likelihood ratio test $P < 0.001$, log2 fold change ≤ −2 or ≥ +2). The mean FPKM expression values for each treatment were calculated from a general FPKM count table using Microsoft Office Excel 2007 (Microsoft, Redmond, WA, USA) and a vertical lookup table (vlookup function) was applied to generate a table of differentially expressed genes with associated mean FPKM values from each treatment, which was given to iPath. The iPath map was scaled to mean FPKM expression values using different line thickness and colour shading.

### 2.2.4 Real time quantitative polymerase chain reaction

Real time quantitative polymerase chain reaction (RT-qPCR) was performed for selected genes and to corroborate RNA-Seq data (Chapter 4) using a two-tube RT-qPCR protocol according to Nolan *et al*. (2006).

#### 2.2.4.1 Reverse transcriptase reaction

First strand cDNA synthesis was performed using Superscript II reverse transcriptase (Invitrogen, Carlsbad, CA, USA) utilising Anchored Oligo$(dT)_{20}$ Primer (Invitrogen, Carlsbad, CA, USA) or Oligo$(dT)_{20}$ Primer (Invitrogen, Carsbad, CA, USA), respectively. Reverse transcription of 500 ng of total RNA was carried out in 50 µL reactions at 42 °C for 50 minutes, followed by inactivation at 70 °C for 15 minutes. As a control for DNA contamination, RNA was pooled from each biological replicate and first strand synthesis reaction mix was added omitting reverse transcriptase.

#### 2.2.4.2 Primer design and quantitative polymerase chain reaction

Oligonucleotides (Table 1) were designed towards the 3' end of the gene of interest using the web-based *RealTimeDesign* Software (available at http://www.biosearchtech.com/realtimedesign, Biosearch Technologies, Novato, CA, USA) aiming for an amplicon length of 80 – 150 bp (optimum 115 bp), a GC content of amplicon and primer of 30 – 80%, a primer length of 18 – 30 bp and a primer melting temperature ™ of 63 – 68 °C. BLAST searches of the primer sequences against the *F. cylindrus* genome sequence (http://genome.jgi-psf.org/Fracy1/Fracy1.home.html) were performed and if necessary primer sequences were modified manually to ensure maximum specificity. Oligonucleotides were assessed for $T_M$, hairpins, and primer dimers using the web-based tool *OligoAnalyzer 3.1* (available at http://eu.idtdna.com/analyzer/Applications/OligoAnalyzer; Integrated DNA Technologies, Coralville, IA, USA) parametrised with concentrations for oligos of 0.4 µM, $Na^+$ of 50 mM, $Mg^{++}$ of 5.5 mM and dNTPs of 0.5 mM. Primers were synthesised by Eurofins MWG Operon (Ebersberg, Germany).

**Table 1. Genes investigated during this study and sequences of the primers used to amplify target genes by qPCR.**

| Gene Target/protein ID | Primer sequence (5' - 3') | Amplicon size (bp) |
|---|---|---|
| Actin-like protein (ACTIN_LIKE)/228346 | Fwd: TGACACGTACTCCGTTGGTC<br>Rev: TTGGTGCCTGATACCGTTCTG | 111 |
| Beta Tubulin (TUBB_2)/274017 | Fwd: GCAATGATGTTCCGTGGAAG<br>Rev: GATGCCTTCACGTTGTTGG | 116 |
| Hypoxanthine phosphoribosyl transferase (HPRT)/184309 | Fwd: TCAACCCAGCATCATTGGAAG<br>Rev: TGTAGTCGAGACCATACCCTAC | 129 |
| Importin alpha subunit (IPO)/259093 | Fwd: TTGCAGCAACACTCGAACAATG<br>Rev: CGCAAGTGCAGCCATCTC | 99 |
| Large ribosomal protein (L27)/269038 | Fwd: GTCCGTCATATCTTCCCAACAC<br>Rev: TACTCGACGTTCCGCATCAAC | 93 |
| Large ribosomal protein (L22)/270383 | Fwd: TGCACATGGTCGAATTGGTA<br>Rev: GTTTGGCGGCCATCTTTCTG | 131 |
| Large ribosomal protein (L14)/271911 | Fwd: TTGCCCTAACGGATTTAACTGTG<br>Rev: AGACGTGTCTTCTTGGATTGC | 142 |
| Large ribosomal protein (L14b/L23e)/269874 | Fwd: GCCTGGAATTGTGGTTCG<br>Rev: ACATTCCTTTGCAACAGGTC | 145 |
| Major allergen (MA) spike control* | Fwd: TCGGTTGACAGATACCTTAAAGGAA<br>Rev: TCAAAGGTGACGTTCGAGTTCAT | 100 |
| nitrile-specifier protein (NSP) spike control* | Fwd: ACGATGCCTTCAGAGCTACCTT<br>Rev: TACGCATCAAGCGTTTGGAA | 100 |
| Peptidylprolyl isomerase A (PPIA)/271442 | Fwd: ATGGCAAGCACGTTGTCTTC<br>Rev: TGGTTGTTCCAGATTGTGATCC | 90 |
| RNA polymerase II (RNAP II)/183218 | Fwd: TCGGAGCTGCTTCCTTTTCTC<br>Rev: TTGTGGACTGGATGGGTTGTAAC | 128 |
| Small ribosomal protein (S1)/274976 | Fwd: GATTCCCTCGATGGATTAGGTGA<br>Rev: GAATCAAGAGAATCAGAAACATCCG | 89 |
| Small ribosomal protein (RPS11)/268264 | Fwd: TACTGCCTTACACATCAAAGTTC<br>Rev: AGAGGGGATTGGTGTGACATC | 142 |
| TATA-box binding protein (TBP)/143154 | Fwd: GCATTTGCCTCCTATGAACCAGA<br>Rev: CTTTGCACCTGTTATCACAACCTTC | 114 |
| *Fragilariopsis* rhodopsin (FR)/267528† | Fwd: GTTACCGTTCCTCTACATTGTCC<br>Rev: GTCCACCATTGAACACCCTTA | 111 |
| *Fragilariopsis* rhodopsin (FR)/267528† | Fwd: GTGGTCGTTGGGTCTATTGGA<br>Rev: GACTGAGTGGCATCGTTAAGTC | 91 |

*spike-in controls of artificial RNA of genes from *Pieris rapae* (cabbage white butterfly).
†RHO primers amplify different regions of the same gene.

For qPCR reactions and second strand amplification, 5 µL of a 10-fold diluted reverse transcriptase reaction mix was supplemented with 20 µL 2× SensiMix SYBR Green NoROX Master Mix (Bioline, London, UK). Forward and reverse primers were added at a concentration of 200 nM. Amplifications were performed in white 96-well plates on a CFX96 Real Time System (Bio-Rad, Hercules, CA, USA) using the

following conditions: initial denaturation 95 °C, 10 minutes, followed by 40 amplification and quantification cycles of 15 seconds at 95 °C, 15 seconds at 59 °C, 10 seconds at 72 °C. Finally, a melting curve analysis (65 °C to 95 °C, increments of 0.5 °C, dwelling time 5 seconds) was carried out to check for primer dimers and non-specific amplification. For each primer pair the reliability of qPCR was demonstrated by five to six point standard curves made by amplification from 1:10 serial dilutions of reverse transcription reactions. Standards for absolute qRT-PCR gene expression analysis were generated as follows. Target sequences were amplified using conventional PCR from cDNA or plasmid templates, separated by agarose gel electrophoresis and purified (illustra GFX PCR DNA and Gel Band Purification Kit, GE Healthcare UK Ltd., Little Chalfont, UK). The concentration of agarose gel-purified target sequences was determined in duplicate readings using a NanoDrop and diluted 1:10,000. Subsequently, six point standard curves were determined for specific target sequences by qPCR amplification from 1:10 serial dilutions of the initial 10,000× dilution. Finally, the absolute amount of cDNA in the samples was calculated based on the equation obtained for logarithmic regression lines for standard curves.

#### 2.2.4.3 Quantitative polymerase chain reaction data analysis

For qPCR data analysis, cycle thresholds ($C_t$) were automatically determined using the CFX Manager Software Version 1.1 (Bio-Rad, Hercules, CA, USA). The relative expressions software tool (Pfaffl et al., 2002) REST-MCS© (available at http://rest.gene-quantification.info/) was used to test the expression of target genes under different experimental conditions. Data was normalised to the exogenous reference gene MA and/or the endogenous reference genes TBP and RNAP II, which both were determined to be most stable expressed in *F. cylindrus* across experimental treatments (Supplementary Figure S6) using the BestKeeper software (Pfaffl et al., 2004)(available at http://rest.gene-quantification.info/). Efficiancies of the qPCR reactions were calculated with REST from the slope of the standard curves, according to the established equation (Bustin, 2000; Rasmussen, 2001):

$$E = 10^{[-1/slope]},$$

where E is PCR efficiency ranging from 1 (minimum value) to 2 (theoretical maximum and optimum) and slope is determined from the linear regression of log(target concentration) versus $C_t$. If no PCR efficiencies were determined, optimal efficiency of

E = 2.0 were assumed. Finally, statistical significances were tested in REST by a pair-wise fixed reallocation randomisation test using 2000 iterations.

#### 2.2.4.4 Allele-specific quantitative polymerase chain reaction

Allele-specific qPCR to discriminate between expression of heterozygous allelic gene copies in *F. cylindrus* was performed according to Germer et al. (2000). Briefly, the specificity of the PCR amplification was conferred by placing the 3'-end of a forward or the reverse allele-specific primer directly over a SNP but matching one or the other variant of the heterozygous allele. Then allele-specific qPCR was performed in two separate reactions, using a common primer and either an allele$_1$-specific primer or an allele$_2$-specific primer. Although, in theory, only completely matching primers should be extended, and only the matching heterozygous allele should get amplified, there will be amplification of the mismatched allele but with lesser efficiency (Germer et al., 2000). The more frequent allele will reach the cycle threshold ($C_t$) at an earlier qPCR amplification cycle (i.e., having a smaller $C_t$) and the difference in $C_t$ values between the two seperate qPCR reactions, the $\Delta C_t$, provides a measure of the allele frequency (Germer et al., 2000). The allele frequency was calculated according to the following equation (Germer et al., 2000):

$$frequency\ of\ allele_1 = 1/(2^{\Delta Ct} + 1),$$

where $\Delta C_t$ = ($C_t$ of allele$_1$-specific qPCR) – ($C_t$ of allele$_2$-specific qPCR) describes the difference in $C_t$ values between the two qPCR reactions.

Generally, allele-specific primers were designed as described above (2.2.4.2). However, in addition to their specific design to match only one of the allele sequences at its 3'-terminal nucleotide, additional nucleotide mismatches located three bases from the 3'-end of the allele-specific primer were incorporated to improve amplification specificity as performed previously (Newton et al., 1989; Okimoto and Dodgson, 1996; Gupta et al., 2005; Wilkening et al., 2005). The allele-specific primers used in this study are described in Table 2.

**Table 2. Primers for allele-specific qPCR.**

| Locus | Primer | Sequence (5' - 3') | Amplicon size (bp) |
|---|---|---|---|
| RHO | RHO1_271123-C2fw: | GGGTGGTCGTTGGGTC**A**AC | |
| | RHO1_271123-C2re: | GGTGGCGTCATTGAGT**G**CA | 88 |
| | RHO2_267528-C2fw: | AGGTGGTCGTTGGGTC**A**AT | |
| | RHO2_267528C2re1: | AGTGGCATCGTTAAGT**T**CC | 88 |
| L27 | L27e269038_C2-fw: | TCGAGTAGATGTTAAGAAGAATTTGAAGA**C**AC | 95 |
| | L27e273430_C2-fw: | TCGAGTAGATGTTAAGAAGAATTTGAAGA**C**AG | 102 |
| | L27e_common_rev2: | TTACTTCCCTGTGCTTTCTTCTC | |

*Note*. Locus describes the genetic loci for *Fragilariopsis* rhodopsin (RHO) and the large ribosomal protein L27 (L27e). Primer IDs consist of gene name, Protein ID, code for additional mismatch at 3'-end (e.g., C0: no mismatch, C2: mismatch three nucleotides from 3'-end), abbreviation fw: forward primer, and re/rev: reverse primer). Underlined nucleotides indicate the sites of polymorphisms to the other allelic variant. Bold nucleotides indicate additionally introduced nucleotide mismatches three bases from the 3'-terminus to increase primer specificity.

To show that the applied method of allele-specific qPCR was valid, a standard mix consisting of predetermined, different ratios of plasmid DNAs containing cloned full length sequences either of the two *Fragilariopsis* rhodopsin alleles was generated and measured for allele frequencies. Therefore, pPha-T1 plasmids (pPha-T1 FR::GFP and pPha-T1 FRext::GFP) were linearised (KpnI restriction digest, 37 °C, 3 h) to avoid strong biases by circular (supercoiled) plasmid standards (Hou et al., 2010). DNA concentrations of purified linearised plasmid (illustra GFX PCR DNA and Gel Band Purification Kit, GE Healthcare UK Ltd., Little Chalfont, UK) were determined in three technical replicates using a NanoDrop. Subsequently, 10 µL of purified linearised plasmids were concentrated until dry using a centrifugal evaporator (miVac DNA concentrator, Genevac Ltd., Ipswich, UK) and 10 nM standard solutions of each plasmid were set up with molecular grade water. Subsequently, 0.001 nM (1 pM) plasmid standards were made using 1:10 serial dilutions of 10 nM standards. Standard mixtures 100 µL were made up from both 1 pM plasmid standards to contain *Fragilariopsis* alleles with known copy frequencies of 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 0.95, which were used as templates for qPCR but with an annealing temperature of 60 °C as described above (2.2.4.2). Finally, allele frequencies were determined according to Germer et al. (2000) as described above.

### 2.2.5 Heterologous expression of rhodopsins from *F. cylindrus* and the Antarctic dinoflagellate *Polarella glacialis*

After cloning of full-length rhodopsin sequences from *F. cylindrus* and the Antarctic dinoflagellate *Polarella glacialis*, both rhodopsins were heterologously expressed in *Xenopus laevis* oocytes to provide direct experimental evidence for their function as light-driven proton pump (Nagel et al., 1995). Expression of phytoplankton rhodopsins was performed in collaboration with the group of Georg Nagel at the University of Würzburg, Germany. Additionally, the *Fragilariopsis* rhodopsin (FR) was cloned and expressed in the diatom *Phaeodactylum tricornutum* to study subcellular targeting in diatoms (Kroth, 2007) and obtain insights into its physiological role based on specific subcellular localisation.

#### 2.2.5.1 Cloning of full-length rhodopsin sequences

The full length FR1/287459 (FRext) allele (see 5.2) was amplified and subloned into *Xenopus laevis* expression vectors by Shiqiang Gao (Nagel Lab, University of Würzburg, Germany). A full length product of FR1 was amplified from cDNA, which was synthesised from RNA of iron-limited *F. cylindrus* cells using touchdown PCR amplification with the forward primer 5'-CCT TTT ACC GTA CAA TGC GAG AG-3' and reverse primer 5'-CAA AAT CTG ACA CTA GGC CCT ACC-3' and successive annealing temperatures of 72 °C (5 cycles), 70 °C (5 cycles) and 62 °C (30 cycles). Touchdown PCR reactions were performed with Phusion DNA polymerase (Finnzyme) according to the manufacturer's recommendations but using a 1:1 mixture of HF and GC buffer as well as addition of 3% DMSO and 1 µg/µL BSA. After purification of agarose gel fragments, gene-specific primers containing BamHI and HindIII sites were used for directed cloning into the *Xenopus laevis* expression vector pGEMHE (Supplementary Figure S2). The full length FR2/274098 allele (see 5.2) was amplified from ~100 ng cDNA template with proofreading DNA Polymerase (Pfu, Fermentas), using the forward primer 5'-ATG ATC AGC GGA ACT CAA TTC AC-3' and reversere primer 5'-AAG GAG AGG AGT TTC TTC GTT TC-3'. A 50 µL reaction contained 0.4 pM of each primer, 0.2 pM of each dNTP, 1× Pfu buffer and 10 mM $MgSO_4$. The amplification profile was as follows: 4 min initial denaturation at 95 °C, followed by 35 cycles of 95 °C for 45 s, 55 °C for 45 s, 72 °C for 90 s extension, and final extension at 72 °C for 5 min. Amplified products were purified (illustra GFX PCR DNA and Gel Band Purification Kit, GE Healthcare UK Ltd., Little Chalfont, UK) from 1.2% TAE

agarose gels (40 mM Tris acetate, 1 mM EDTA, 0.5 µg $mL^{-1}$ Ethidium bromide) and ligated into the *Phaeodactylum tricornutum* transformation vector StuI-GFP pPha-T1 (Gruber et al., 2007) (Supplementary Figure S4) using blunt-ended non-directional ligation with StuI (Eco147I) restriction enzyme. Restriction digest with StuI and ligation was performed in a single tube. Therefore, a 20 µL reaction mix (1 µL vector, 5 µL insert, 1 µL StuI restriction enzyme, 2 µL PEG 4000, 2 µL ATP/DTT (10 mM/100 mM), 2 µL restriction enzyme buffer, 1 µl T4 ligase; adjusted to 20 µL with molecular grade water) was incubated overnight at room temperature until the reaction was inactivated by heating to 65 °C for 20 min. After cooling down to room temperature, empty vector molecules were digested by adding 1 µL StuI restriction enzyme to the inactivated reaction mix and incubation for 1.5 h at 37 °C. The digest of empty StuI-GFP pPha-T1 vector molecules was inactivated by heating to 65 °C for 20 min. After cooling down to room temperature, 7 µL of ligation reaction mix was transformed in $CaCl_2$-competent *E. coli* DH5α cells and selected on ampicillin (100 µg $mL^{-1}$). Colony PCR using a combination of insert and vector primers were used to screen for plasmids with correct orientation of insert. Finally, the full-length *Polarella* rhodopsin was amplified and cloned by Sabrina Förster (Nagel Lab, University of Würzburg, Germany) from cDNA of nutrient-replete *P. glacialis* cultures. Different rhodopsin gene constructs were subcloned into heterologous expression vectors (Supplementary Figures S2-S4: Vectors used in this study) using customised gene-specific primers. Expression vectors were transformed and maintained in *E. coli*. Plasmids from *E. coli* were isolated by the method of Birnboim & Doly (1979) using commercial plasmid prep kits (e.g. Promega, Madison, WI, USA). The orientation and accuracy of cloned rhodopsin sequences was verified by small scale capillary sequencing (e.g. Genome Enterprise Ltd., Norwich, UK).

#### 2.2.5.2 Heterologous expression of *Fragilariopsis* rhodopsin and *Polarella* rhodopsin in *Xenopus* oocytes

Heterologous expression of phytoplankton rhodopsins from *F. cylindrus* and *P. glacialis* in *Xenopus laevis* oocytes was performed in collaboration with the group of Georg Nagel at the University of Würzburg, Germany, and in particular Shiqiang Gao and Sabrina Förster, who performed all of the oocyte expression experiments and analyses according to published procedures (Nagel et al., 1995; Nagel et al., 1998; Nagel et al., 2002). Briefly, full-length rhodopsin sequences were subcloned into pGEMHE (Liman et al., 1992), a derivative of pGEM3z (Promega, Madison, WI,

USA). The pGEMHE plasmid (Supplementary Figure S2) is a high expression oocyte vector for *in vitro* transcription and expression in *Xenopus* oocytes and contains 3' and 5' untranslated regions (UTRs) of a *Xenopus* β-globin gene (Liman et al., 1992). Different rhodopsin gene constructs containing truncated N-terminal ends were tested to optimise plasma membrane expression and electrophysiological measurements in oocytes. Plasmid DNA linearised with NheI restriction enzyme was used for *in vitro* transcription of cRNA (Ambion, Life Technologies, Carlsbad, CA, USA). The oocytes were injected with 20 – 30 ng of cRNA, and incubated in a modified oocyte Ringer's solution (110 mM NaCl, 5 mM KCl, 1 mM $MgCl_2$, 2 mM $CaCl_2$, 5 mM HEPES, pH 7.6) supplemented with *all-trans* retinal (1 µM) two to five days at 16 – 18 °C (Nagel et al., 1995). Alternatively, oocytes were incubated in ND96 (96 nM NaCl, 2 mM KCl, 1 mM $MgCl_2$, 2 mM $CaCl_2$, 5 mM HEPES, pH 7.6). The antibioticum gentamycin (100 µg $mL^{-1}$) was added to solutions for storage of oocytes at 16 °C. Finally, oocytes were examined using two electrode voltage-clamp (TEVC) experiments as described previously (Nagel et al., 1995; Nagel et al., 1998). Therefore, two glass micropipettes (i.e., pulled capillary glass pipettes filled with 3 M KCl into which Ag/AgCl electrodes were inserted) were penetrated into a single oocyte cell, allowing the control of membrane voltage with a current electrode and measuring the transmembrane current with a potential electrode. The potential electrode was used to register the membrane voltage against a reference electrode (3 M KCl electrode), which provided the reference ground of the system. The photocurrents of phytoplankton rhodopsins were recorded by voltage clamping of oocytes at a predetermined holding potential and application of light-pulses close to the oocyte membrane using light fibres (1.5 $mm^2$ diameter).

#### 2.2.5.3 Heterologous expression of *Fragilariopsis* rhodopsin in *Phaeodactylum tricornutum*

The *Fragilariopsis* rhodopsin (FR) gene was cloned into the *P. tricornutum* expression vector pPha-T1 (Zaslavskaia et al., 2000) for analysis of subcellular targeting using green fluorescent protein (GFP) labelling (Kroth, 2007), protein purification using His-Tag (Joshi-Deo et al., 2010) and functional studies on *P. tricornutum* mutants complemented with FR. Nuclear transformation of *Phaeodactylum tricornutum* was performed using a Biolistic PDS-1000/He Particle Delivery System (Bio-Rad, Hercules, CA, USA) fitted with 1350 psi rupture disks as described previously (Kroth, 2007). For the selection and cultivation of *P. tricornutum* transformants 75 µg $ml^{-1}$ Zeocin (InvivoGen, San Diego, CA, USA) was added to the

solid 1.2% agar medium. To detect green fluorescence signals from GFP-transformed *P. tricornutum*, flow cytometry (FACScalibur, BD, Franklin Lakes, NJ, USA) with the standard optical filter configuration was used. Therefore, green fluorescence was measured in the FL1 channel using a 515 – 545 nm emission filter. FL1 histograms were used to identify transformed (peak about $10^3$) and non-transformed (peak about $10^1$) cells. Milli-Q water was used as a sheath fluid and all analyses were perfomed using a low flow rate (~20 µL $min^{-1}$). Triggered on green fluorescence, 10,000 events were collected. An event rate between 100 and 400 cells $s^{-1}$ was used to avoid coincidence, and when needed samples were diluted in 0.2 µm filtered artificial seawater prior to analysis. To confirm presence of GFP and analyse the morphology of the cells, upright widefield fluorescence microsocopy (Axioplan 2 IE imaging microscope equipped with CCD Axiocam camera, Carl Zeiss, Germany) was performed. Chloroplasts were identified by red autofluorescence of chlorophyll *a/c* during excitation at 562 ± 20 nm (Alexa568 filter set). The excitation and emission of filters used during microscopical analyses are liste in Table 3.

**Table 3. Optical filter sets used in fluorescence microscopy**

| Filter | Excitation (nm) | Emission (nm) |
|---|---|---|
| UV | 365 ± 30 | 445 ± 30 |
| GFP | 469 ± 17.5 | 525 ± 19.5 |
| Alexa568 | 562 ± 20 | 624 ± 20 |

Additionally, to screen for FR in non-GFP labelled cell lines, a combination of colony PCR and RT-qPCR was applied. For colony PCR, a *P. tricornutum* colony was picked from selective plates and transferred into 20 µL lysis buffer (10% Triton X-100, 20 mM Tris HCl pH 8, 2 mM EDTA). Cells were solubilised by > 30 s vortexing, followed by 15 min incubation on ice, 10 min incubation at 95 °C (thermal cycler) and final storage at room temperature. PCR amplification (4 min initial denaturation at 95 °C, followed by 35 cycles of 95 °C for 45 s, 55 °C for 45 s, 72 °C for 50 s, and final extension at 72 °C for 5 min) of FR with gene-specific primers was performed using 5 µL of a 1:5 diluted cell lysate as template. Moreover, to screen for FR transcript in non-GFP labelled *P. tricornutum* transformants, RNA was extracted as described above (2.2.2) and absolute RT-qPCR expression analysis (2.2.4) with custom primers (Table 1) was performed.

# Chapter 3

# The draft genome of the psychrophilic diatom *Fragilariopsis cylindrus*

## 3.1 Introduction

Diatoms are the most successful group of eukaryotic phytoplankton and dominate the permanently cold environment sea ice. The obligate psychrophilic pennate diatom *Fragilariopsis cylindrus* is a keystone species in the Arctic and Antarctic Ocean (Lundholm and Hasle, 2008) and forms large populations in sea ice brine channels and wider sea ice zone (Kang and Fryxell, 1992).

Although to date more than 30 psychrophilic prokaryotic genomes from sea ice and other permanently cold environments have been sequenced providing insights into molecular adaptations to psychrophily (Casanueva et al., 2010), only a single polar eukaryotic genome has been sequenced for the psychrotolerant terrestrial green alga *Coccomyxa subellipsoidea* (Blanc et al., 2012) and a genome sequence for a obligate psychrophilic marine eukaryote is lacking.

*F. cylindrus* has become a model for algal adaptation to polar marine conditions and diverse physiological and biochemical studies have been conducted (Mock and Hoch, 2005; Janech et al., 2006; Bayer-Giraldi et al., 2010; Lyon et al., 2011). Furthermore, different expressed sequence tag (EST) libraries (Mock et al., 2005; Krell et al., 2008), a macroarray study (Mock and Valentin, 2004) and specific gene expression studies (Krell et al., 2007; Bayer-Giraldi et al., 2010) enabled first genomic insights into genetic adaptation of *F. cylindrus* to polar conditions. Thus, we choose to sequence *F. cylindrus* to reveal its metabolic potential on a genomic scale. Furthermore, *F. cylindrus* is an ideal candidate for comparative genomic analysis with sequenced mesophilic diatoms *Thalassiosira pseudonana* (Armbrust et al., 2004) and *Phaeodactylum tricornutum* (Bowler et al., 2008), not only to obtain genomic insights into the evolution and adaptation of diatoms to permanently cold environments, but also to allow further exploration of the ecological success of diatoms and particularly their success in extreme conditions of brine channels in polar sea ice. Last but not least, additional genome sequences are required to identify diatom specific evolutionary

innovations and to anchor environmental sequences to known genomes (Bowler et al., 2010).

Available diatom genome sequences have provided the blueprint for understanding their evolutionary origin and extraordinary ecological success (Armbrust et al., 2004; Bowler et al., 2008). The analysis of diatom genomes confirmed the secondary endosymbiotic origin of their plastids from red alga (Armbrust et al., 2004) and amino acid sequence comparisons showed the presence of genes originating from the ancestral heterotrophic host and the photosynthetic symbiont via endosymbiotic gene transfer (Armbrust et al., 2004) as well as from bacteria via horizontal gene transfer (Bowler et al., 2008). Genes from different partners of secondary endosymbiosis together with large numbers of bacterial genes acquired by horizontal gene transfer permitted novel metabolisms never previously found together and includes coexistence of plant-like photosynthesis together with animal-like mitochondrial fatty acid oxidation and urea cycle (Armbrust et al., 2004; Bowler et al., 2008). Additionally, identification of genes important for survival in particular conditions, such as genes involved in high-affinity iron-uptake and survival in low iron oceanic conditions, provide insights into metabolic adaptations to specific marine environments (Armbrust et al., 2004).

Some adaptive strategies for survival in permanently cold environments have been revealed by psychrophilic prokaryotic genomes and include identification of cold shock proteins, antifreeze proteins and proteins involved in unsaturated fatty-acid synthesis to maintain membrane fluidity at low temperatures (D'Amico et al., 2006; Casanueva et al., 2010). Additionally, comparative analysis of prokaryotic genomes showed amino acid modifications in protein-coding genes to confer molecular flexibility to increase catalytic efficiency and prevent cold denaturation (Saunders et al., 2003; Medigue et al., 2005; Methe et al., 2005; Ayala-del-Río et al., 2010; Zhao et al., 2010). However, little is known for adaptation of polar eukaryotes and only since the first macroarray study of *F. cylindrus* in 2004 (Mock and Valentin, 2004) modern molecular tools been used to discover the molecular bases of the adaptation and gene composition of sea ice algae. In addition to the construction of EST libraries from *F. cylindrus* grown under freezing temperatures (cold stress) and high salt (salt stress) (Mock et al., 2005; Krell et al., 2008), an EST library has been constructed for the polar diatom *Chaetoceros neogracile* under polar summer growth conditions (+4 °C, continuous light) (Jung et al., 2007) and two cDNA microarray studies of *C. neogracile*

have been conducted under thermal stress (+10 °C) (Hwang et al., 2008) and high light stress (600 µmol photons $m^{-2}$ $s^{-1}$) (Park et al., 2010). The array studies with *F. cylindrus* (Mock and Valentin, 2004) and *C. neogracile* (Hwang et al., 2008; Park et al., 2010) showed that the acclimation to lower temperatures seems to put less stress on these algae compared to high-temperature (+10 °C) and that photosynthesis is negatively affected under high temperatures and high light. However, the molecular basis of these phenomena remains largely unknown and it is only recently that a polar eukaryotic genome has become available for the psychrotolerant green alga *C. subellipsoidea* (Blanc et al., 2012). Although *C. subellipsoidea* can survive extremely low temperatures in Antarctic soil (−50 °C), it is not fully specialised to grow in a permanently cold environment and its optimal growth temperature is ~20 °C (Blanc et al., 2012). Furthermore, there are significant metabolic differences between green algae and diatoms (Wilhelm et al., 2006). Thus, we sequenced the genome of *F. cylindrus* to provide the first obligate psychrophilic eukaryotic genome from a marine phytoplankton species and to gain first insights into evolution and adaptation of a diatom to conditions of polar oceans.

Here, I report the genome structure, gene content and deduced metabolic capacity of *F. cylindrus* in comparison to other sequenced diatoms and explain putative adaptations to extreme environments.

## 3.2 Results

The *F. cylindrus* draft genome sequence has been made available at http://genome.jgi-psf.org/Fracy1/Fracy1.home.html and has been annotated by an international consortium led by Thomas Mock (see Preface). In the following, I report genome structure, gene content and metabolic capacity of *F. cylindrus*. To compile a coherent synopsis of the *F. cylindrus* genome, I also interpreted and analysed data provided by other members of the *F. cylindrus* genome consortium, whose individual contributions are specified above (see p. XI). Data contributed by other consortium members are indicated in the text and figure legends throughout this chapter. Personally, I manually annotated about 500 genes and performed custom comparative analyses on metabolic pathways including carbohydrate metabolism, lipid metabolism, chlorophyll metabolism and *F. cylindrus*-specific genes.

### 3.2.1 Genome structure, assembly and gene content

The nuclear genome assembly of *F. cylindrus* was determined to be ~80.5 Mb (Table 4) with a sequencing read coverage depth of 7.25× and was assigned to 271 genomic scaffolds (118 scaffolds > 50 kb). Each genomic scaffold represents a portion of the genome sequence reconstructed from end-sequenced whole-genome shotgun clones (fragments) and is composed of contigs and gaps. In total, the *F. cylindrus* nuclear genome assembly contained 4602 contigs and 5.4% sequence gaps. The contigs are defined as contiguous genomic sequences in which the order of bases is known to a high confidence level and gaps occur where reads (i.e., known sequences) from the two sequenced ends of shotgun clones overlap with other reads on a different contig. Since the average fragment lengths were known, the number of bases between contigs and thus gap size could be estimated. There are small gaps of 1 – 5 kb in the current *F. cylindrus* genome assembly, which interrupt genes and affect gene annotations. The available *F. cylindrus* genome sequence reads allowed only incomplete assembly falling short in the aim of whole-genome shotgun assembly to represent each genomic sequence in one scaffold. Thus, one chromosome may be represented by many scaffolds and the relative locations of scaffolds in the genome remain unknown.

Interestingly, the *F. cylindrus* genome was found to be heterozygous with the effect of punctuated high nucleotide polymorphism estimated to ~6% between selected syntenic scaffolds that represent putative homologous chromosomes with the same order of genes (Robert P. Otillar, JGI, *personal communication* 21/06/2012). A high degree of nucleotide polymorphism prevented heterozygous haplotypes, which contained different alleles at genetic loci, to be collapsed into a single haplotype (i.e., the genotype of linked genomic loci on a chromosome) and caused a diffuse haplotype structure. This diffuse haplotype structure contains both highly heterozygous genomic regions, which were too different for the assembly algorithm to be combined into a single contig (and thus scaffold), as well as consensus sequences of merged haplotypes, which represent DNA from two homologous chromosomes. As a result of the diffuse haplotype structure, syntenic scaffolds in *F. cylindrus* that represent putative homologous chromosomes overlap in regions with a high degree of heterozygous allelic differences between haplotypes and split into separate sets of scaffolds, each representing one heterozygous allele. Thus, a heterozygous allele, which is commonly defined as a polymorphic DNA sequence that exists in only one location of the genome (genomic locus), appears on more than one scaffold. The high degree of heterozygousity

in the *F. cylindrus* genome affected ~30% (7,966) of the 27,137 predicted gene models and if not specified differently, the following analyses refer to the set of 27,137 predicted genes. Assuming synteny (i.e., similar blocks of genes in the same relative position of the genome) representing heterozygous regions of putative homologous chromosomes (heterozygous polymorphism), we predict 18,077 genes after filtering highly heterozygous gene copy pairs (Table 4). To explore the hypothesis that heterozygous gene copy pairs were gene duplications (paralogs) as oppose to highly diverged alleles, the sequencing depth coverage of the genome was investigated for heterozygous gene copy pairs. With the ideal assumption that alleles represent the same (divergent) genomic loci, we might expect to observe a twofold lower coverage in comparison to paralogous genes, which would contain reads from two assembled haplotypes. It was observed that the primary gene copy variant, defined as the copy variant on the larger of the two scaffolds, had a higher average coverage of ~6-fold in comparison to ~3-fold for secondary alleles (Supplementary Figure S7, courtesy of Robert P. Otillar, unpublished data). Furthermore, under the assumption that paralogs are more divergent (> 2% dissimilarity of nucleotide sequence) than alleles, the sequence polymorphism of gene copy variants was analysed and it was found that sequence polymorphism of the gene copy pairs ranged from single nucleotide polymorphisms (SNPs), insertions/deletions (InDels) to larger structural variants involving larger DNA fragments. Gene copy pairs showed a high nucleotide sequence similarity of > 98% for most pairs (Figure 8), which corresponded to amino acid sequence similarity > 99% (not shown, courtesy of Andrew Toseland, unpublished data), indicating the presence of putative heterozygous alleles. The final functional analysis of the gene copy variant pairs showed that biological process gene ontology (GO) terms “metabolic process” (GO:0008152), “lipid metabolism” (GO:0006629) and “intracellular protein transport” (GO:0006886) were significantly overrepresented compared to the set of genes not present as variant pairs (Fisher exact test, $P < 0.05$; courtesy of Remo Sanges, unpublished data).

The annotation of *F. cylindrus* showed that repeats represent ~38% of the assembly including transposable elements (TEs), unclassified repeats and simple sequence repeats (SSRs) (Table 4). TEs represented 5.9 Mb (7.3%) and unclassified repeats contributed 7.3 Mb (9.1%) of sequences (Florian Maumus, unpublished data). SSRs constituted 17.4 Mb (21.6%) of the *F. cylindrus* genome including low-complexity DNA and tandem repeats (Florian Maumus, unpublished data). Tandem

repeats were most abundant in introns (~7%), followed by intergenic regions (~3.5%) and specific analysis of showed that ACT and ATC repeat patterns were particularly common constituting ~800 bp/Mbp (0.1% coverage) and ~1600 bp/Mbp (0.2% coverage), respectively (Christoph Mayer & Florian Leese, unpublished data, not shown). Additionally, CAA tandem repeat patterns were enriched in gene promoter regions of *F. cylindrus* and the CAACAA motif was found most significant (motif finding analysis, $P = 3.33E^{-16}$) (Remo Sanges, unpublished data, not shown). Last, not least the Leucine Rich Repeat (LRR) and Pentatricopeptide repeat (PPR) protein domains were enriched in *F. cylindrus* in comparison to the diatom core genome (Figure 10, Andrew E. Allen & Ruben E. Valas, unpublished data).

**Table 4. General features of sequenced diatom genomes**

| | *F. cylindrus* | *P. tricornutum* | *T. pseudonana* |
|---|---|---|---|
| Genome size | 80.5 Mb | 27.4 Mb | 32.4 Mb |
| G+C content (coding %) | 39.8% | 50.6% | 47.8% |
| Predicted genes | 18,077 | 10,402 | 11,776 |
| Species-specific genes | 6,913 | 1,404 | 2,450 |
| Species-specific paralogs | 2,859 | 366 | 891 |
| Average gene length | 1566 bp | 1621 bp | 1745 bp |
| Average no. introns per spliced gene | 2 | 2 | 3 |
| Average intron length | 246 bp | 137 bp | 125 bp |
| Repeat content (overall %) | 38% | ~17% | ~17% |
| Transposable elements (overall %) | 5.9 Mb (7.3%) | 2 Mb (7.3%) | ~0.9 Mb (2.9%) |
| Simple sequence repeats (overall %) | 17.4 Mb (21.6%) | 2 Mb (7.3%) | 4 Mb (12.3%) |
| Unclassified repeats (overall %) | 7.3 Mb (9.1%) | 0.4 Mb (1.5%) | 0.55 Mb (1.7%) |

As shown in Table 4, the coding G+C content of the *F. cylindrus* genome was found to be 39.8%. It was shown that the G+C content variation affected codon and tRNA anticodon usage causing a bias towards adenine (A) and thymidine (T) bases (Figure 6; Figure 7; Stephan Frickenhaus, unpublished data). Analysis of relative frequencies of codon usage for all sequenced diatoms showed that A/T rich third codon bases were found more abundant in *F. cylindrus* than in the other two sequenced diatoms (Figure 6). Complementary, the analysis of relative frequencies of A or T at anticodon position 1 ($AT_1$) for 330 tRNA sequences showed an A/T preference for codons with a genetic code degeneracy greater than one. For 12 amino acids there was a chance of ≥ 50% that A or T was at position 1 of the tRNA anticodon (Figure 7).

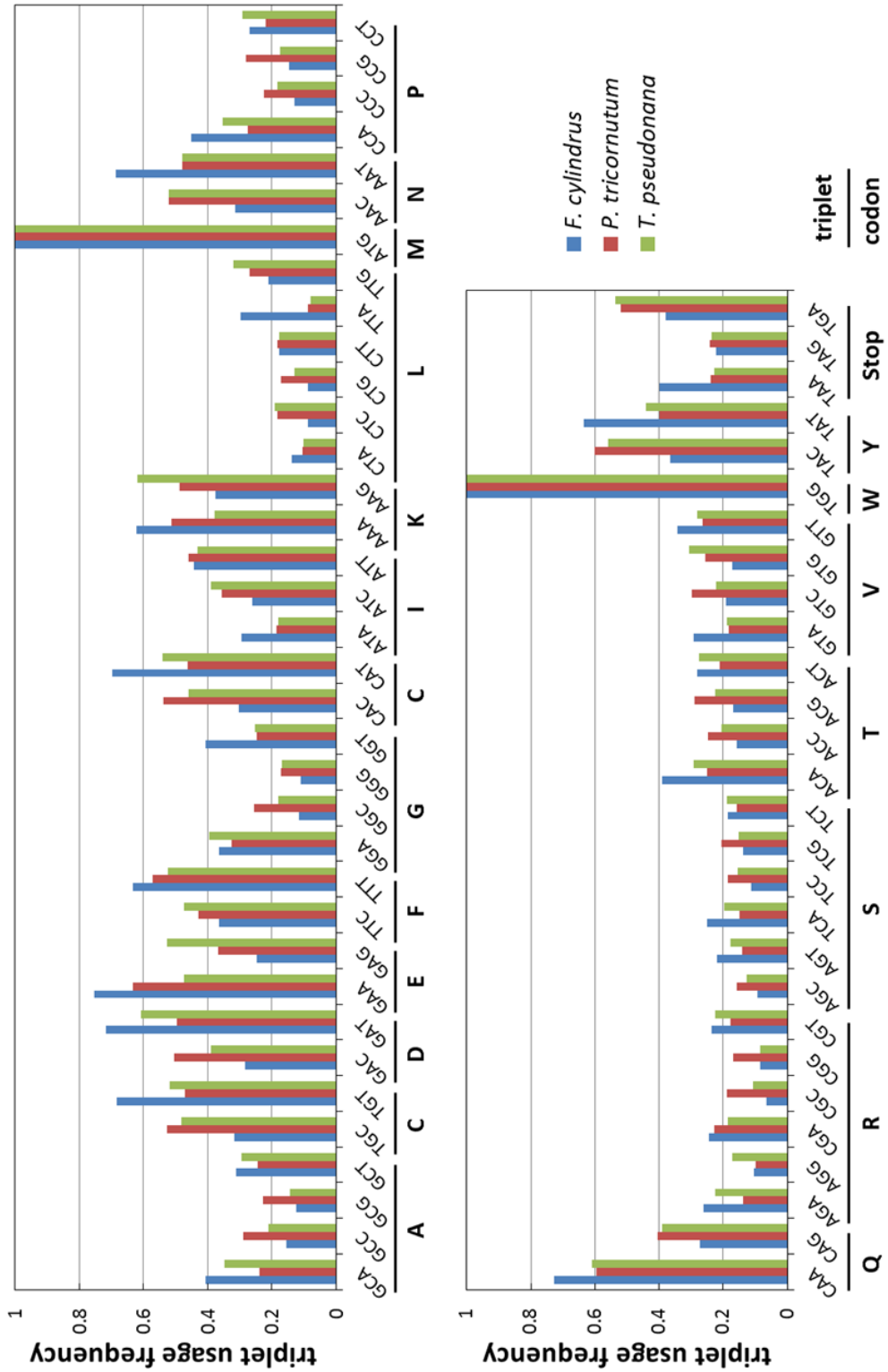



**Figure 6. Codon usage analysis of diatom genes . The codon usage frequency of protein coding genes is shown for *F. cylindrus* (blue), *P. tricornutum* (red) and *T. pseudonana* (green) (Florian Maumus, unpublished data).**

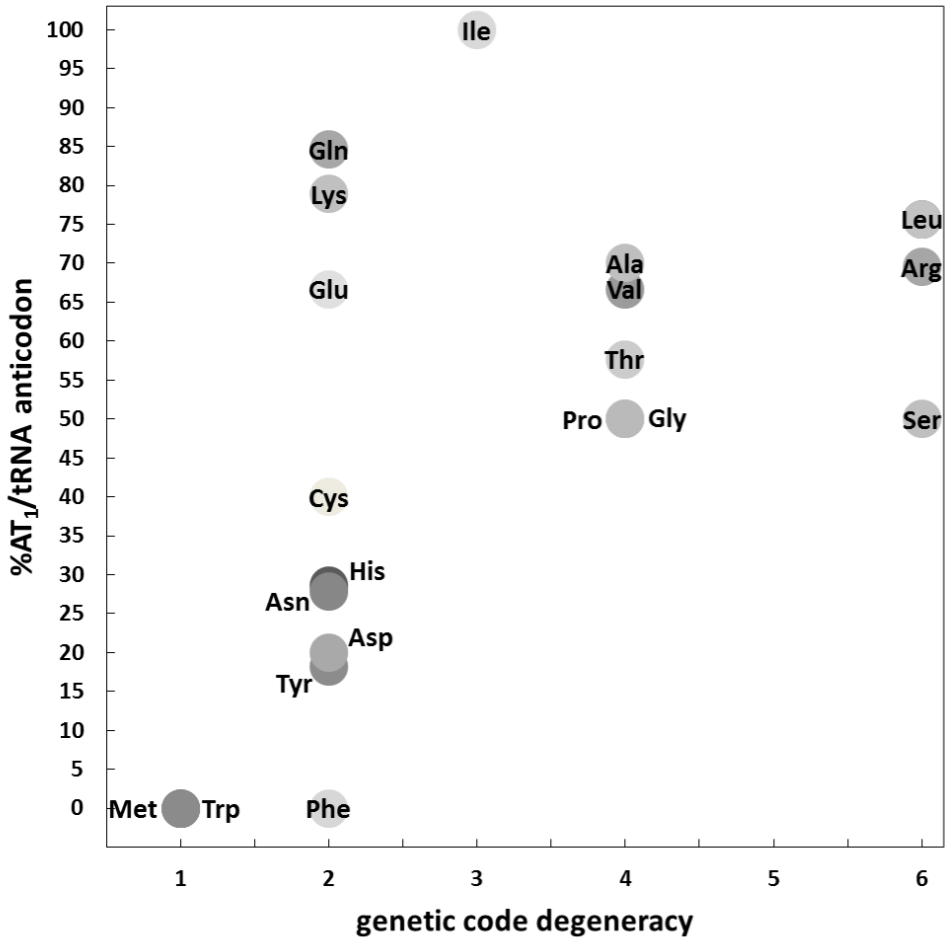



**Figure 7. Relative adenine/thymidine nucleobase frequency at anti-codon position 1 in 330 tRNA sequences of *F. cylindrus* plotted against the genetic code degeneracy (modified from Stephan Frickenhaus, unpublished data).**

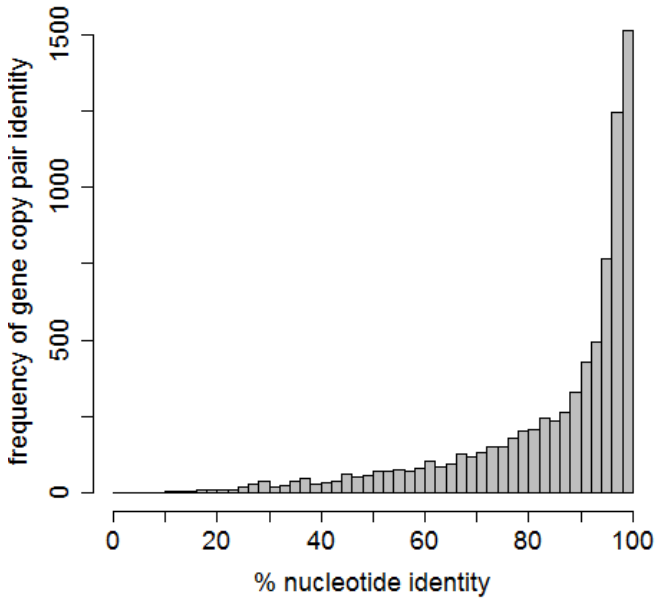



**Figure 8. Histogram of allelic nucleotide identity in *F. cylindrus* as function of frequency**

Finally, to identify the gene repertoire contributing to adaptation of *F. cylindrus* a comparative analysis of the gene family content of all three sequenced diatoms with other eukaryotes and prokaryotes was carried out. It showed that based on the diffuse haplotype (27,137 genes), 6659 genes were shared by all three diatoms (reciprocal best BLAST(P), *E*-value < $1e^{-9}$; Andrew E. Allen & Ruben E. Valas, unpublished data) and represented the diatom core genome. Moreover, 1260 genes (19%) from the diatom core were specific to diatoms, whereas 5399 genes (81%) had orthologs in other organisms with the largest number of 247 genes shared with Plantae including Rhodophyceae and Unikonts (Figure 9; Andrew E. Allen & Ruben E. Valas). As shown by comparison of all three diatom genomes with genomes from Prasinophycea and Rhodophyceae, *F. cylindrus* shared 1657 orthologs with *P. tricornutm*, 971 orthologs with *T. pseudonana*, 357 orthologs with Prasinophycea and 131 orthologs with Rhodophyta (Figure 9). 11,511 genes (6913 for the single haplotype) were identified as *F. cylindrus*-specific.

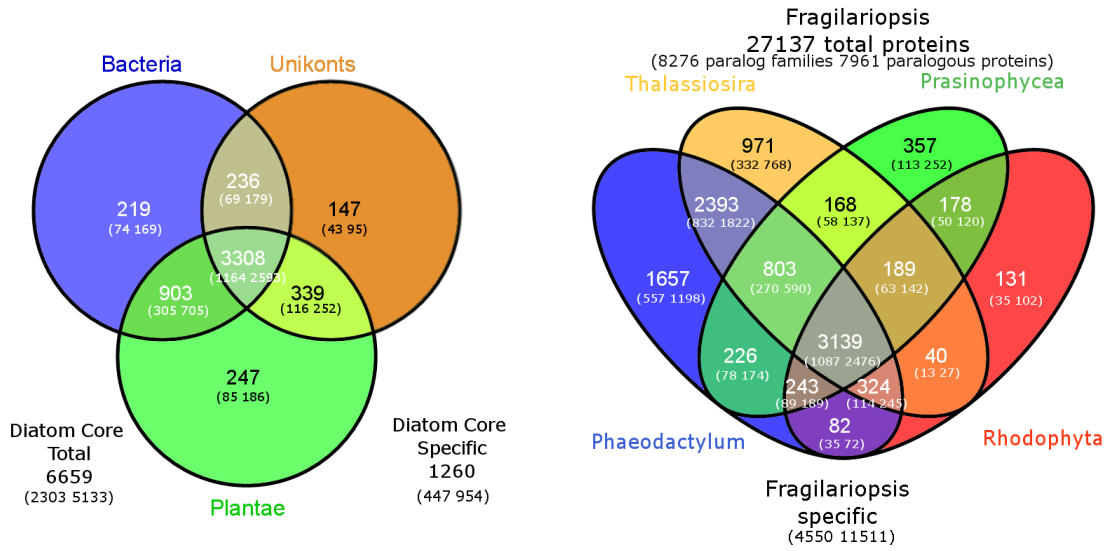



**Figure 9. Venn diagrams of diatom core genome and *F. cylindrus*-specific gene families. Left panel: The diatom core genome composed of genes present in all three sequenced diatom genomes. Venn diagram representing genes of the core diatom genome shared with bacteria, unikonts and plantae (including red algae). Left number in brackets shows number of paralogous families and right number the family abundance. Right panel: Venn diagram showing shared and unique gene families in *F. cylindrus* (Fragilariopsis), *P. tricornutum* (Phaeodactylum), *T. pseudonana* (Thalassiosira), Prasinophycea and Rhodophyta (Andrew E. Allen & Ruben E. Valas, unpublished data).**

### 3.2.2 Protein family and metabolic pathway expansions

Annotated proteins of *F. cylindrus* and the two sequenced diatoms *T. pseudonana* and *P. tricornutum* were organised into 7,972 protein families (Pfam) and 1,200 Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic pathways based on shared sequence similarity and referred to as diatom core genome (Andrew E. Allen

& Ruben E. Valas, unpublished data). The assignment of Pfam domains and KEGG pathways to proteins identified several protein families and metabolic pathways which had a significantly higher (hypergeometric test, $P < 0.05$) number of proteins in *F. cylindrus* in comparison to the diatom core genome (Figure 10; Figure 11). The expansions of protein families and metabolic pathways in *F. cylindrus* was evaluated in the context of evolution and adaptation to environmental constrains of the Southern Ocean including low temperatures, trace metal availability and low light conditions. Based on the rationale that temperature strongly affects core metabolism including protein translation, temperature-related protein families and transcription factors were investigated. Additionally, since an increase in the synthesis of unsaturated fatty acids to maintain membrane fluidity at low temperatures is hypothesised to be a molecular adaptation to psychrophilic lifestyle, the expansion of lipid metabolism of *F. cylindrus* was investigated. Moreover, as trace metals, including iron, may cause major constraints on phytoplankton in the Southern Ocean, a comparative analysis of metal-binding protein families was performed. Furthermore, based on the rationale that *F. cylindrus* is constrained by light availability, which requires quantitative regulation of photosynthetic pigments, the expanded genetic repertoire of *F. cylindrus* in respect to biosynthesis of photosynthetic pigments and light-harvesting proteins was investigated. Finally, invidual *F. cylindrus*-specific proteins that are involved in adaptations to life in the polar environments are presented.

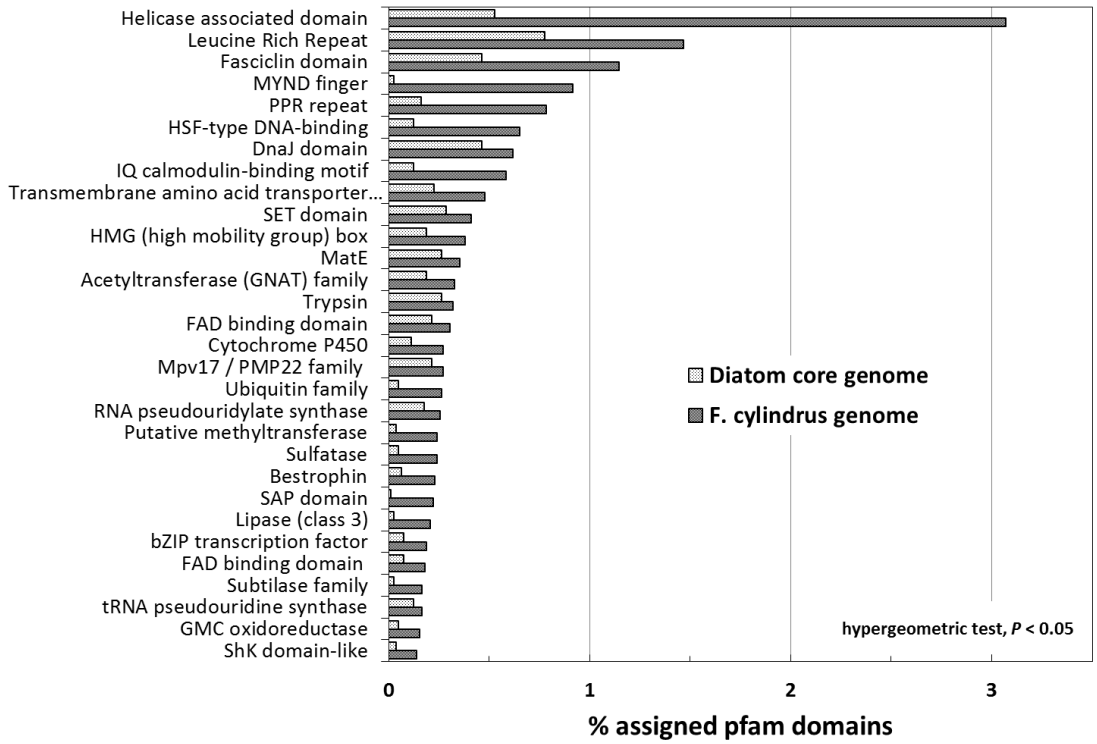



**Figure 10. Enrichment of protein domains in the *F. cylindrus* genome in comparison to the diatom core genome shown as total percentage of annotated Pfam domains (based on data from Andrew E. Allen & Ruben E. Valas).**

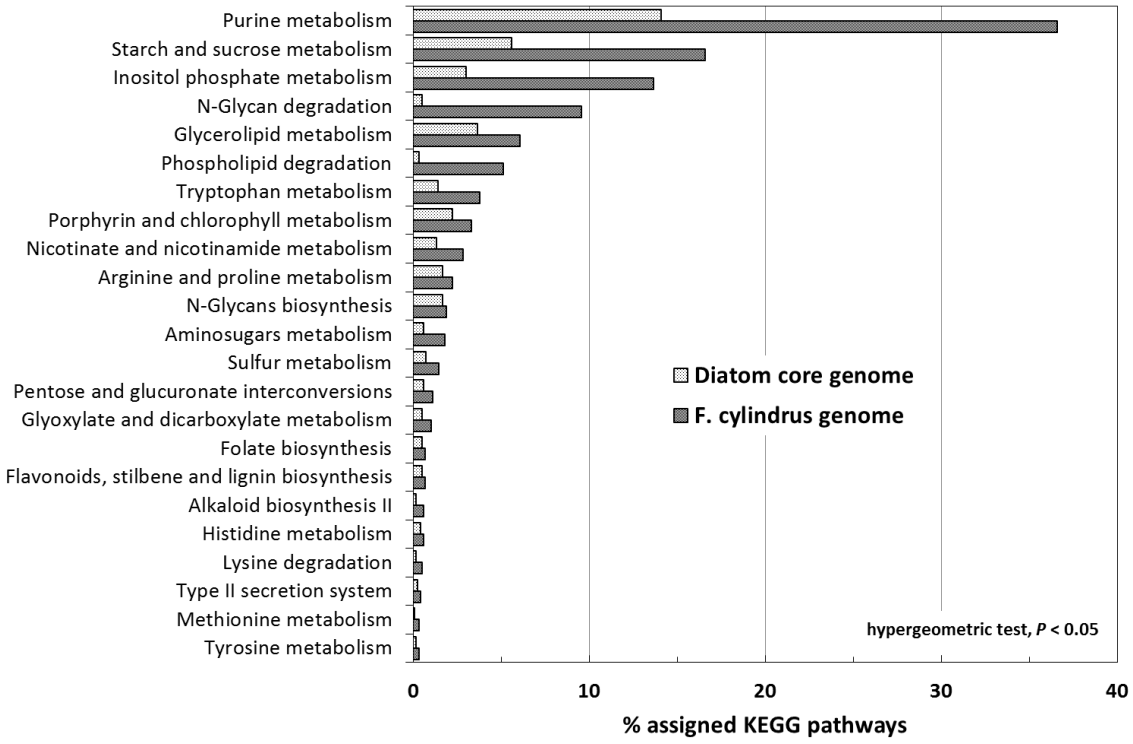



**Figure 11. Enrichment of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways in the *F. cylindrus* genome in comparison to the diatom core genome shown as total percentage of annotated KEGG pathway annotations (based on data from Andrew E. Allen & Ruben E. Valas).**

### 3.2.2.1 Temperature-related protein families and transription

Over-represented protein families in comparison to the diatom core genome included heat and cold shock factors involved in transcriptional response. The HSF-type DNA binding domain and DnaJ domain, which both exhibit chaperone activity, were among the top 10 most enriched Pfam domains in comparison to the diatom core (Figure 10). Additionally, a cold shock DNA binding domain was significantly enriched in *F. cylindrus* and constituted 0.08% of the total assigned Pfam domains and a bZIP transcription factor domain constituted 0.2% (Figure 10). Furthermore, a novel domain fusion protein combining an N-terminal Myb-like transcription factor domain with a C-terminal silicon transporter domain could be identified in *F. cylindrus* (Protein ID 233781). Overall, the Gene Ontology (GO) terms "nucleic acid binding" (GO:0003676) and the broad term "transcription factor activity" (includes GO:0000988, GO:0001070-71 and GO:0003700) from the major ontology branch "Molecular Function" ranked among the top 30 over-represented terms in *F. cylindrus* compared to the diatom core (Figure 12).

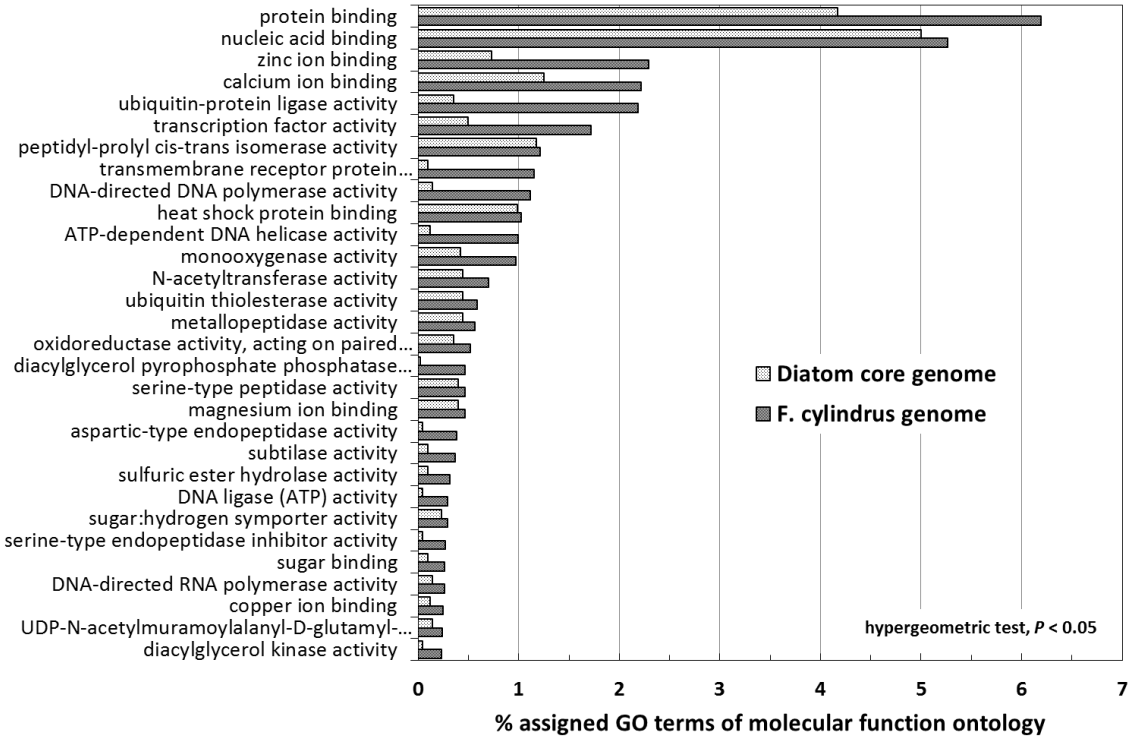



**Figure 12. Enrichment of molecular function gene ontology (GO) annotations in the *F. cylindrus* genome in comparison to the diatom core genome shown as percentage of total ontology annotations (based on data from Andrew E. Allen & Ruben E. Valas).**

#### 3.2.2.2 Metal-binding protein families

The comparative analysis of copper and iron-binding domains showed that phytoplankton genomes including *F. cylindrus* were enriched in Fe binding domains relative to the related non-photosynthetic *Phytophtora* species (Figure 13; Christoph L. Dupont, unpublished data). Both, green and red lineages of phytoplankton contained a similar number of Fe-binding protein domains.

Copper binding domains in all investigated phytoplankton genomes scaled in abundance to genome size (Figure 13). A power law slope of 1.8 indicated that a doubling of phytoplankton genome size results in nearly a quadrupling in the number of copper binding domains (Figure 14; Christoph L. Dupont, unpublished data). Even in consideration of the shared scaling of Cu-binding domains, the genomes of *Chlamydomonas reinhardtii*, *F. cylindrus* and *Aureococcus anophagefferens* deviated from the global trend (Figure 13). In the case of *F. cylindrus*, its genome contained disproportionate abundant domains including the plastocyanin/azurin-like fold family (Christoph L. Dupont, unpublished data). Manual examination of putative plastocyanin proteins identified a clear plastocyanin with putative N-terminal targeting sequence (protein ID 272258), whereas the remaining 10 contained conserved histidine and cysteine residues, providing a known Cu binding site (Christoph L. Dupont, unpublished data). In the same context, the molecular function GO term “copper ion-binding” (GO:0005507) was enriched in *F. cylindrus* compared to the diatom core genome (Figure 12), multicopper oxidase domains were found amplified in *F. cylindrus* (Andrew E. Allen & Ruben E. Valas, unpublished data, not shown) and Fe-binding Cytochrome P450 domains were enriched in comparison to the diatom core genome (Figure 10). Additionally, five Fe-binding hemoproteins containing globin-like domains were identified in *F. cylindrus* and included isoenzymes for neuroglobin (235866, 246319, 241443) which were not detected in *P. tricornutum* and *T. pseudonana*, a flavohemoglobin (249631) detected also in *P. tricornutum* but not in *T. pseudonana* and a bacteria-like haemoglobin (241146) detected also in both sequenced diatoms. Additionally, two putative Fe-binding hemopexin domain-containing proteins (protein IDs 261622 and 196981) were identified in *F. cylindrus* but were absent in *T. pseudonana* and *P. tricornutum*. In the context of Fe acquisition, protein-coding genes involved in high affinity iron uptake systems including five isoenzymes for ferric-chelate reductase (protein IDs 232972, 227601, 238487, 246292 and 259423), a Fe permease (243554) and a ferroportin (223989) could be identified in *F. cylindrus*.

Moreover, comparative analysis of clusters of orthologous groups of proteins (COG) in *F. cylindrus* and the other two sequenced diatoms showed that the COG group "Ferric reductase, NADH/NADPH oxidase and related proteins" was enriched in *F. cylindrus* in comparison to the diatom core genome (not shown). Furthermore, putative proteins serving as Fe siderophore were present in *F. cylindrus* including ferritin (291658) and genes involved in enterobactin biosynthesis. An isochorismatase Pfam domain involved in biosynthesis of enterobactin was more than four-fold over-represented in *F. cylindrus* and constituted 0.23% of the assigned Pfam domains in comparison to 0.05% assigned Pfam domains in the diatom core genome (Andrew E. Allen & Ruben E. Valas, unpublished data, not shown).

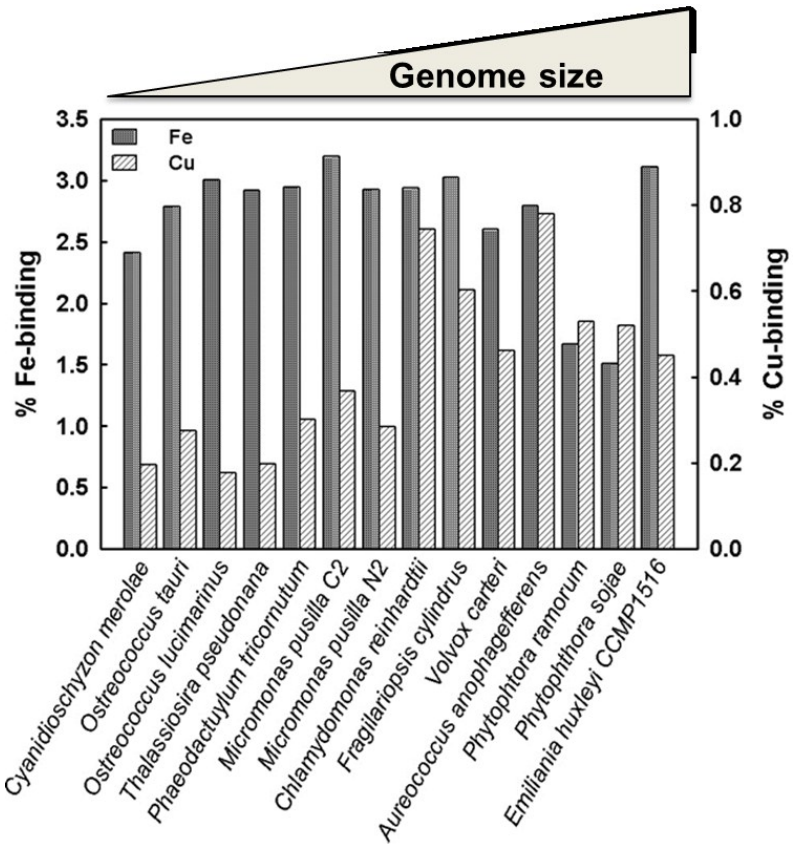



**Figure 13. Relative abundance of iron (Fe) and copper (Cu) binding proteins in selected eukaryotic genomes. Genomes are arranged according to genome size (modified from Christopher L. Dupont, unpublished data).**

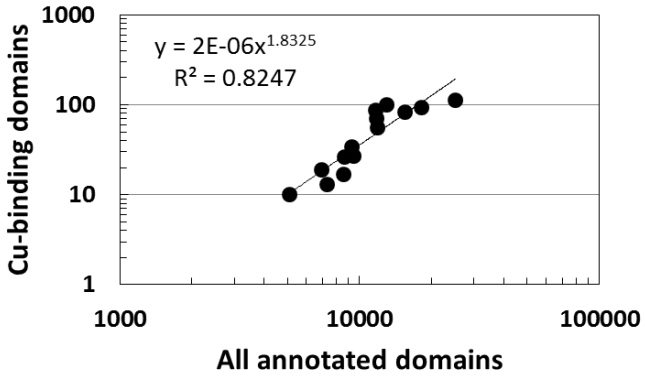



**Figure 14. Scaling of copper (Cu)-binding domains according to genome size. Shown are annotated Cu-binding domains as function of all annotated domains for selected eukaryotic genomes (modified from Christopher L. Dupont, unpublished data).**

In addition to Fe and Cu-binding, zinc (Zn)-binding domains appeared to play an important role in *F. cylindrus* as indicated by significant enrichment of the GO Molecular Function term "zinc ion binding", which constituted 2.3% of assigned terms in *F. cylindrus* compared to 0.7% in the diatom core genome (Figure 12). While the total number of zinc-binding protein domains in F. cylindrus was comparable to other phytoplankton genomes, the conserved Zn-binding myeloid-Nervy-DEAF-1 (MYND) domain (named after myeloid translocation protein 8, Nervy and DEAF-1) was greatly expanded (Figure 15; Christoph L. Dupont, unpublished data). MYND domains in *F. cylindrus* contained seven conserved cysteine and histidine residues, which formed two Zn binding sites and were always found in combinations with DNA-binding or protein-protein binding domains, such as ankyrin repeats, HCP domains, F box domains, RING domains and tetracopeptide repeats (Christoph L. Dupont, unpublished data). Two MYND domains were found associated with the Fe-containing Hypoxia Induction Factor prolyl hydroxylase (HIF) domain involved in the cellular response to changing oxygen in Eukarya but > 75% (98/128) of MYND-associated domains were not annotated by Superfamily Hidden Markov Models (Christopher L. Dupont, unpublished data). A phylogenetic analysis of MYND domains in *F. cylindrus* showed a high nucleotide divergence resulting in a functional divergence of binding sites, which was likely to have occurred within the last 30 Myr (Mark McMullan & Cock van Oosterhout, unpublished data, not shown). Additionally, a BLAST based analysis found that most of the MYND-containing proteins in *F. cylindrus* appeared to be diatom-lineage specific (Figure 9) (Andrew E. Allen & Ruben E. Valas, unpublished data).

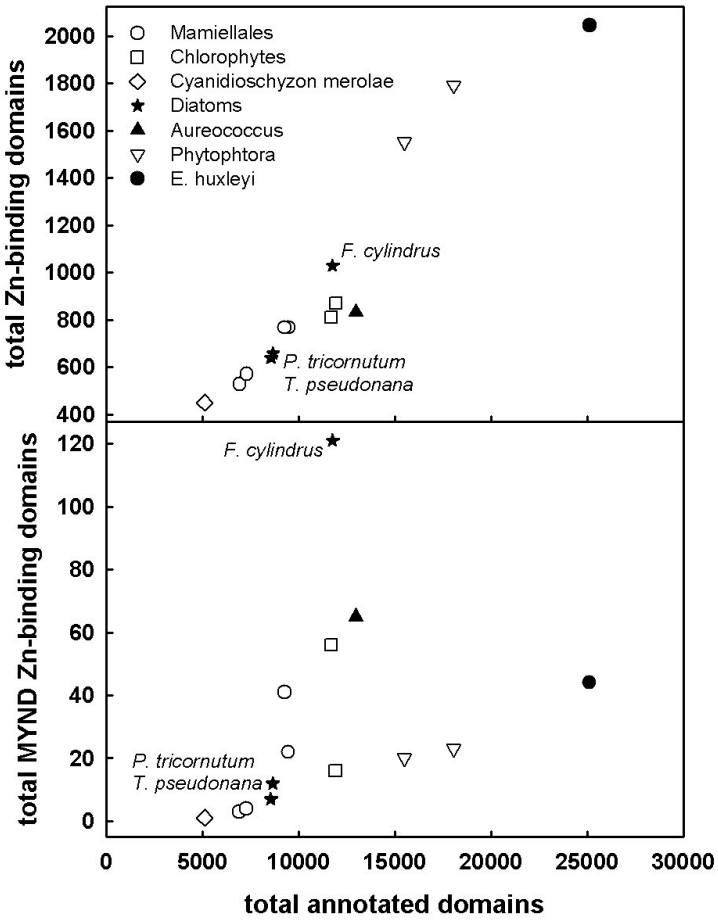



**Figure 15. Amplification of Zinc (Zn)-binding domains in *F. cylindrus*. Top panel shows number of total Zn-binding domains as function of total annotated domains. Bottom panel shows specific MYND Zn-binding domains as function of total annotated domains (modified from Christopher L. Dupont, unpublished data).**

### 3.2.2.3 Carbohydrate metabolism

Carbohydrate metabolism encompasses the photosynthesis-driven synthesis of carbohydrates from atmospheric carbon dioxide, followed by synthesis and degradation of storage products. As shown in Figure 11 carbohydrate metabolism was represented in *F. cylindrus* by several enriched KEGG pathway annotations in comparison to the diatom core genome, which included “Starch and sucrose metabolism”, “Inositol phosphate metabolism” and “Glyoxylate and dicarboxylate metabolism”.

Additionally, complete pathways for glycolysis and gluconeogenesis could be annotated and variation of the common glycolytic Embden-Meyerhof-Parnas pathway and putative presence of mitochondrial Entner-Dourdoff glycolysis was indicated by identification of a 6-phosphogluconate dehydratase (EDD1, 274061) and 2-keto-3-deoxyphosphogluconate aldolase (EDA1, 267632) in *F. cylindrus*. Noteworthy, in

contrast to EDA1, EDD1 did not contain a mitochondrial targeting sequence but contained an N-terminal chloroplast targeting motif suggesting localisation in the chloroplast and similar targeting prediction was obtained for EDD in *T. pseudonana*. Additionally, in the context of glycolytic pathways no ortholog for a phosphoketolase (XFP) identified in *P. tricornutum* could be found in the genome of *F. cylindrus* indicating a lack of the catabolic phosphoketolase pathway.

The analysis of KEGG pathway annotations showed almost three-fold amplification of the "Starch and sucrose metabolism" pathway in *F. cylindrus* constituting 16.6% of all assigned KEGG pathway annotations in comparison to 5.6% in the diatom core genome (Figure 11). Additionally, in the same annotation category, the "pentose and glucuronate interconversion" pathway was amplified in *F. cylindrus* (Figure 11). Glucuronate is also a primary breakdown product in inositol metabolism, which was the third most over-represented KEGG pathway in *F. cylindrus* in comparison to the diatom core genome (Figure 11). Key enzymes involved in inositol metabolism could be identified including a methylmalonate-semialdehyde dehydrogenase (MMSDH1, 291598) as well as several isoenzymes for myo-inositol dehydrogenase (InDH1-3, 181045, 181882, 207854), triosophosphate isomerase (TIM1-4, 275634, 269372, 191478, 170563), inositol phosphate synthase (INPS1-2, 185965, 157585) and inositol monophosphatase (IMP1-3, 264100, 268347, 268346).

Last, not least the KEGG pathway "Glyoxylate and dicarboxylate metabolism" was amplified in *F. cylindrus* compared to the diatom core genome and the "Purine metabolism" pathway, which provides glyoxylate was the top most over-represented KEGG pathway in *F. cylindrus* compared to the diatom core genome (Figure 11).

#### 3.2.2.4 Lipid metabolism

Lipid metabolism was found enriched in *F. cylindrus* in comparison to the diatom core genome as indicated by significant enrichment of the biological process GO term "lipid metabolic process" (Figure 16) as well as the KEGG pathway annotations "Glycerolipid metabolism" and "Phospholipid degradation" (Figure 11). Furthermore, comparative analysis of annotations for European Commission number for enzymes (EC) identified the lipid metabolic enzymes exo-alpha-sialidase (EC 3.2.1.18), phospholipase D (EC 3.1.4.4) and diacylglycerol kinase (EC 2.7.1.107) as the top most amplified EC number annotations in *F. cylindrus* compared to the diatom core genome

showing between ~20-fold and ~25-fold amplification in the *F. cylindrus* genome (Andrew E. Allen & Ruben E. Valas, unpublished data, not shown), which was indicated by the enriched molecular function GO term "diacylglycerol kinase activity" (Figure 12).

Similar to other sequenced diatoms, a complete pathway for *de novo* fatty acid biosynthesis via a type II fatty acid biosynthesis pathway within chloroplasts could be identified in *F. cylindrus* and included acyl carrier protein (ACP1, 263929), malonyl-CoA ACP transacylase (MCAT1, 259810), beta-ketoacyl ACP synthase III (KASIII/FabH, 291592), and beta-hydroxyacyl ACP dehydratases (FabA/Z, 232938) (Table 5). Additionally, similar to the other two sequenced diatoms, at least seven microsomal desaturases involved in polyunsaturated fatty acid biosynthesis could be identified in *F. cylindrus* and two complete mitochondrial and peroxisomal pathways for beta-oxidation of fatty acids were annotated in *F. cylindrus*.

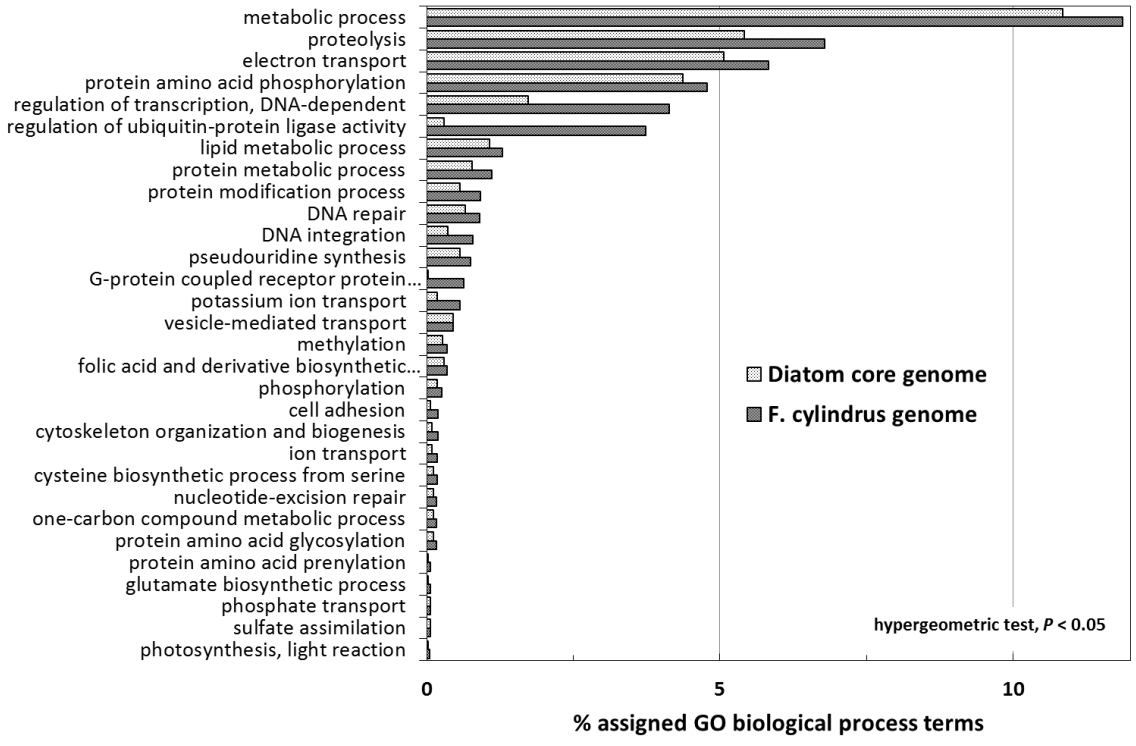



**Figure 16. Enrichment of biological process gene ontology (GO) annotations in the *F. cylindrus* genome in comparison to the diatom core genome shown as percentage of total ontology functional annotations (based on data from Andrew E. Allen & Ruben E. Valas, unpublished).**

**Table 5. Manually annotated gene models in *F. cylindrus* involved in lipid metabolism.**

| Enzyme Name | Gene Name in *F. cylindrus* | JGI protein identifier |
|---|---|---|
| malonyl-CoA decarboxylase | *MCD1* | 291591 |
| Malonyl-CoA ACP transacylase | *MCAT1* | 259810 |
| 3-oxoacyl-synthase III | *FabH* | 291592 |
| acyl carrier protein | *ACP1* | 263929 |
| acyl carrier protein | *ACP2* | 173760 |
| fatty acid desaturase | *FAD1* | 268672 |
| fatty acid desaturase | *FAD2* | 228533 |
| fatty acid desaturase | *FAD3* | 241272 |
| fatty acid desaturase | *FAD4* | 226788 |
| fatty acid desaturase | *FAD5* | 238138 |
| fatty acid desaturase | *FAD6* | 267824 |
| fatty acid desaturase | *FAD7* | 208053 |
| beta-hydroxyacyl ACP dehydratases | *FabA/Z* | 232938 |
| acyl-CoA dehydrogenase | *ACD1* | 209569 |
| acyl-CoA dehydrogenase | *ACD2* | 187779 |
| acyl-CoA dehydrogenase | *ACD3* | 291593 |
| acyl-CoA dehydrogenase | *ACD4* | 264525 |
| acyl-CoA dehydrogenase | *ACD5* | 233211 |
| long-chain acyl-CoA synthetase | *ACSL1* | 262994 |
| acyl-CoA oxidase | *ACOX1* | 210789 |
| acetyl-CoA acetyltransferase | *ACAT1* | 274265 |
| alcohol dehydrogenase | *ADH1* | 146601 |
| alcohol dehydrogenase | *ADH2* | 277191 |
| enoyl-CoA hydratase | *ECH1* | 180456 |
| enoyl-CoA hydratase | *ECH2* | 159942 |
| enoyl-CoA hydratase | *ECH3* | 273959 |
| enoyl-CoA hydratase | *ECH4* | 235018 |
| enoyl-CoA hydratase | *ECH5* | 202663 |
| enoyl-CoA hydratase | *ECH6* | 193150 |
| bifunctional enoyl-CoA hydratase/ 3-hydroxyacyl-CoA dehydrogenase | *ECH_HADH1* | 207194 |
| bifunctional enoyl-CoA hydratase/ 3-hydroxyacyl-CoA dehydrogenase | *ECH_HADH2* | 183437 |
| 3-hydroxyacyl-CoA dehydrogenase | *HADH1* | 270026 |

#### 3.2.2.5 Light harvesting, photoprotection

The "Porphyrin and chlorophyll synthesis" KEGG pathway annotation was found enriched in the *F. cylindrus* genome compared to the diatom core constituting ~3.3% and ~2.3% of the total pathway annotations (Figure 11). Additionally, the GO

term “photosynthesis, light reaction” of the biological process ontology was weakly enriched in *F. cylindrus*, constituting 0.05% of functional annotations compared to 0.03% contribution in the diatom core genome (Figure 16).

Most enzymes required for a plastid-localised methylerythritol phosphate/deoxyxylulose phosphate (MEP/DOXP) pathway and a cytosolic mevalonate (MVA) pathway contributing to carotenoid, chlorophyll and tocopherol synthesis could be identified in *F. cylindrus* by manual annotation (Table 6). Similar to other sequenced diatoms, no homologs for CHL27, a subunit of the Mg-protoporphyrin IX monomethyl ester (MPE) cyclase, and subunits of the light-independent protochlorophyllide oxidoreductase (DPOR) could be detected in the nuclear and plastid genome of *F. cylindrus*. However, the *F. cylindrus* plastid genome encoded for a protoporphyrin IX Mg-chelatase subunit (Table 6). Putative isoenzymes involved carotenoid biosynthesis were identified for isopentenyl diphosphate isomerase (IDI), phytoene synthase (PSY), phytoene dehydrogenase (PDH), lycopene cyclase (LCYB), zeaxanthin epoxidase (ZEP), violaxanthin de-epoxidase (VDE). Noteworthy, two proximate gene models encoded for cytosolic IDI (226527) and PSY (209449) and could be merged to a putative gene model encoding for a bi-functional fusion protein. Like in other sequences diatoms, no homolog with lycopene epsilon-cyclase (LCYE) could be identified in *F. cylindrus*. However, in contrast to *T. pseudonana,* no homolog for beta-carotene hydroxylase (CHYB) was found in *F. cylindrus* and neither in *P. tricornutum*. Last, not least a putative beta-carotene 15,15'-monooxygenase (BCMO, protein ID 291475), catalysing the central cleavage of beta-carotene to yield two molecules of retinal, was identified in *F. cylindrus* and other sequenced diatoms.

The genome of *F. cylindrus* contained approximately 64 gene models encoding for light-harvesting complexes (LHC; Beverly E. Green, unpublished data) and 55 models were supported by EST sequences. In comparison, the diatom genomes of *T. pseudonana* and *P. tricornutum* encoded for about 40 LHC genes (Beverly E. Green, unpublished data) and further comparison with other sequenced phytoplankton showed that content of total LHC domains in a genome scaled according to genome size (Figure 17). In comparison to that a similar scaling was not observed for the LHC family of LHCX proteins (Figure 17), which is involved in photoprotection. The *F. cylindrus* genome contained 11 gene models encoding LHCX proteins including a LHCX1 homolog (218498), which was demonstrated to be a regulator of photoprotection via non-photochemical quenching (NPQ) in *P. tricornutum*. In comparison four and six

gene models were found encoding for LHCX proteins in the other sequenced diatoms and 9 LHCX protein coding genes were found in *Ostreococcus* spp (Figure 17). Additionally, a complete xanthophyll cycle could be identified in *F. cylindrus* and similar to other sequenced diatoms, no homolog encoding for the PS II protein PsbS, which is involved in sensing the thylakoid lumen pH and photoprotection by onset of NPQ in green algae, could be detected.

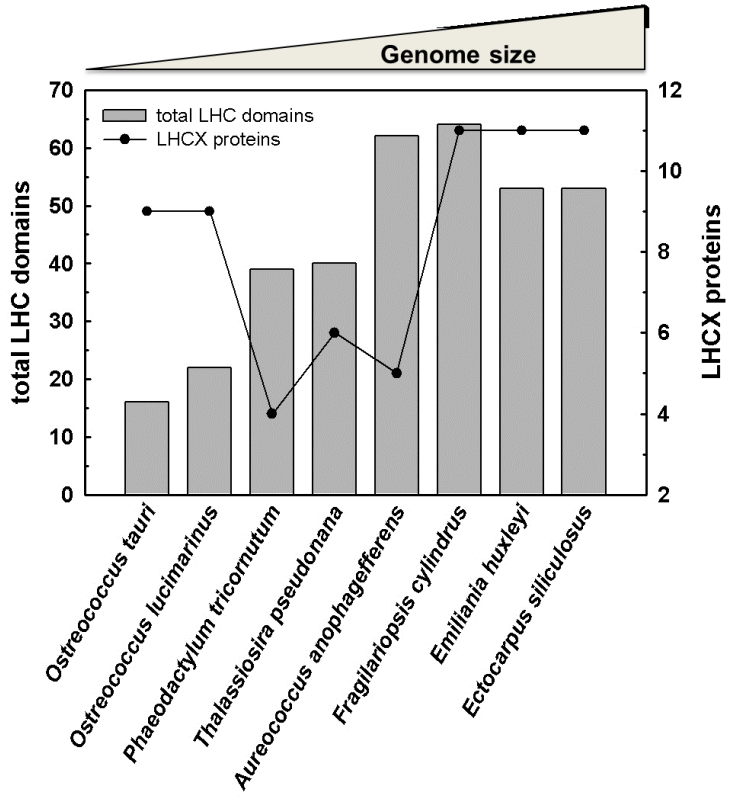



**Figure 17. Total number of light-harvesting complex (LHC) protein domains and number of identified LHC proteins from the LHCX family in selected eukaryotic algae genomes. Genomes are arranged according to genome size (based on data from Beverly E. Green, modified).**

**Table 6. Manually annotated gene models in *F. cylindrus* involved in biosynthesis of photosynthetic pigments.**

| Enzyme Name | Gene Name in *F. cylindrus* | JGI protein identifier |
|---|---|---|
| glutamyl-tRNA synthase | *GTS1* | 291553 |
| glutamyl-tRNA synthase | *GTS2* | 228581 |
| glutamyl-tRNA synthase | *GTS3* | 182520 |
| glutamyl-tRNA synthase | *GTS4* | 249648 |
| glutamyl-tRNA reductase | *GTR1* | 226164 |
| glutamate-1-semialdehyde aminotransferase | *GSAT* | 218589 |

| Enzyme Name | Gene Name in *F. cylindrus* | JGI protein identifier |
|---|---|---|
| 5-aminolevulinic acid dehydratase | *ALAD* | 218256 |
| porphobilinogen deaminase | *PBGD* | 205101 |
| uroporphyrinogen III synthase | *UROS* | 244684 |
| uroporphyrinogen III decarboxylase | *UROD1* | 209393 |
| uroporphyrinogen III decarboxylase | *UROD2* | 268500 |
| uroporphyrinogen III decarboxylase | *UROD3* | 224546 |
| uroporphyrinogen III decarboxylase | *UROD4* | 216482 |
| coproporphyrinogen III oxidase | *CPX1* | 240805 |
| coproporphyrinogen III oxidase | *CPX2* | 267395 |
| coproporphyrinogen III oxidase | *CPX3* | 242067 |
| protoporphyrinogen IX oxidase | *PPX* | 261082 |
| protoporphyrin IX Mg-chelatase subunit D | *CHLD1* | 170289 |
| protoporphyrin IX Mg-chelatase subunit D | *CHLD2* | 247844 |
| protoporphyrin IX Mg-chelatase subunit H | *CHLH1* | 169081 |
| protoporphyrin IX Mg-chelatase subunit H | *CHLH2* | 261055 |
| protoporphyrin IX Mg-chelatase subunit I | *CHLI* | Plastid genome |
| Mg-protoporphyrin IX methyltransferase | *CHLM* | 268444 |
| 3,8-divinyl protochlorophyllide a 8-vinyl reductase | *DVR1* | 268496 |
| 3,8-divinyl protochlorophyllide a 8-vinyl reductase | *DVR2* | 259321 |
| NADPH:protochlorophyllide oxidoreductase | *POR1* | 267731 |
| NADPH:protochlorophyllide oxidoreductase | *POR2* | 188173 |
| chlorophyll synthase | *CHLG* | 223502 |
| geranylgeranyl reductase | *GGR* | 267781 |
| isopentenyl diphosphate synthase | *IDS* | 263072 |
| phytoene dehydrogenase | *PDH2* | 260963 |
| Isopentenyl diphosphate:dimethylallyl diphosphate isomerase | *IDI1* | 226527 |
| Isopentenyl diphosphate:dimethylallyl diphosphate isomerase | *IDI2* | 239201 |
| Phytoene synthase | *PSY2* | 264173 |
| Phytoene synthase | *PSY3* | 233859 |
| Phytoene synthase | *PSY4* | 209449 |
| 15-cis-zeta-carotene isomerase | *Z-ISO* | 291550 |
| zeta-carotene desaturase | *ZDS* | 291551 |
| carotenoid isomerase | *CRTISO1* | 274697 |
| carotenoid isomerase | *CRTISO2* | 206370 |
| carotenoid isomerase | *CRTISO_3* | 186494 |

| Enzyme Name | Gene Name in *F. cylindrus* | JGI protein identifier |
|---|---|---|
| carotenoid isomerase | *CRTISO_4* | 232063 |
| carotenoid isomerase | *CRTISO_5* | 226929 |
| carotenoid isomerase | *CRTISO_6* | 225509 |
| similar to cholin dehydrogenase | *CHDH1* | 246826 |
| lycopene beta cyclase | *LCYB2* | 183412 |
| cytochrome P450, carotenoid hydroxylase | *LTL1* | 261383 |
| cytochrome P450, carotenoid hydroxylase | *LTL2* | 209190 |
| Violaxanthin de-epoxidase | *VDE_1* | 267113 |
| Violaxanthin de-epoxidase | *VDE_2* | 212709 |
| Violaxanthin de-epoxidase | *VDE_3* | 291552 |
| Zeaxanthin epoxidase | *ZEP_1* | 232148 |
| Zeaxanthin epoxidase | *ZEP_2* | 260743 |
| Zeaxanthin epoxidase | *ZEP_3* | 208380 |

#### 3.2.2.6 *F. cylindrus*-specific proteins

A total of 11,511 specific proteins without significant matches to red and green algae genomes (BLAST P e-value cutoff $1e^{-9}$) were identified in the *F. cylindrus* genome including all gene copy variant pairs from heterozygous regions of the genome (Figure 9).

***Ice-binding proteins.*** The genome of *F. cylindrus* encoded for 12 ice-binding proteins (IBPs) and 11 proteins containing C-terminal ice-binding domains, which were not found in the mesophilic diatoms *T. pseudonana* and *P. tricornutum*. Protein sequence analysis predicted N-terminal signal peptides for eight of the 12 IBPs and transmembrane domains for five of the 11 domain fusion proteins with C-terminal IBP domain.

A phylogenetic analysis showed that IBP sequences from *F. cylindrus* clustered with IBPs from other sea ice organisms including diatoms, bacteria, fungi (Raymond and Janech, 2009) as well as a crustacean and formed three major groups (Figure 18). The majority of *F. cylindrus* IBP sequences grouped with sequences from *Fragilariopsis* spp except for two sequences (Fcyl AFP-g1, protein ID 161548; Fcyl AFP-g14, protein ID 219400), which grouped with sequences from other eukaryotic sea ice organisms (Figure 18).

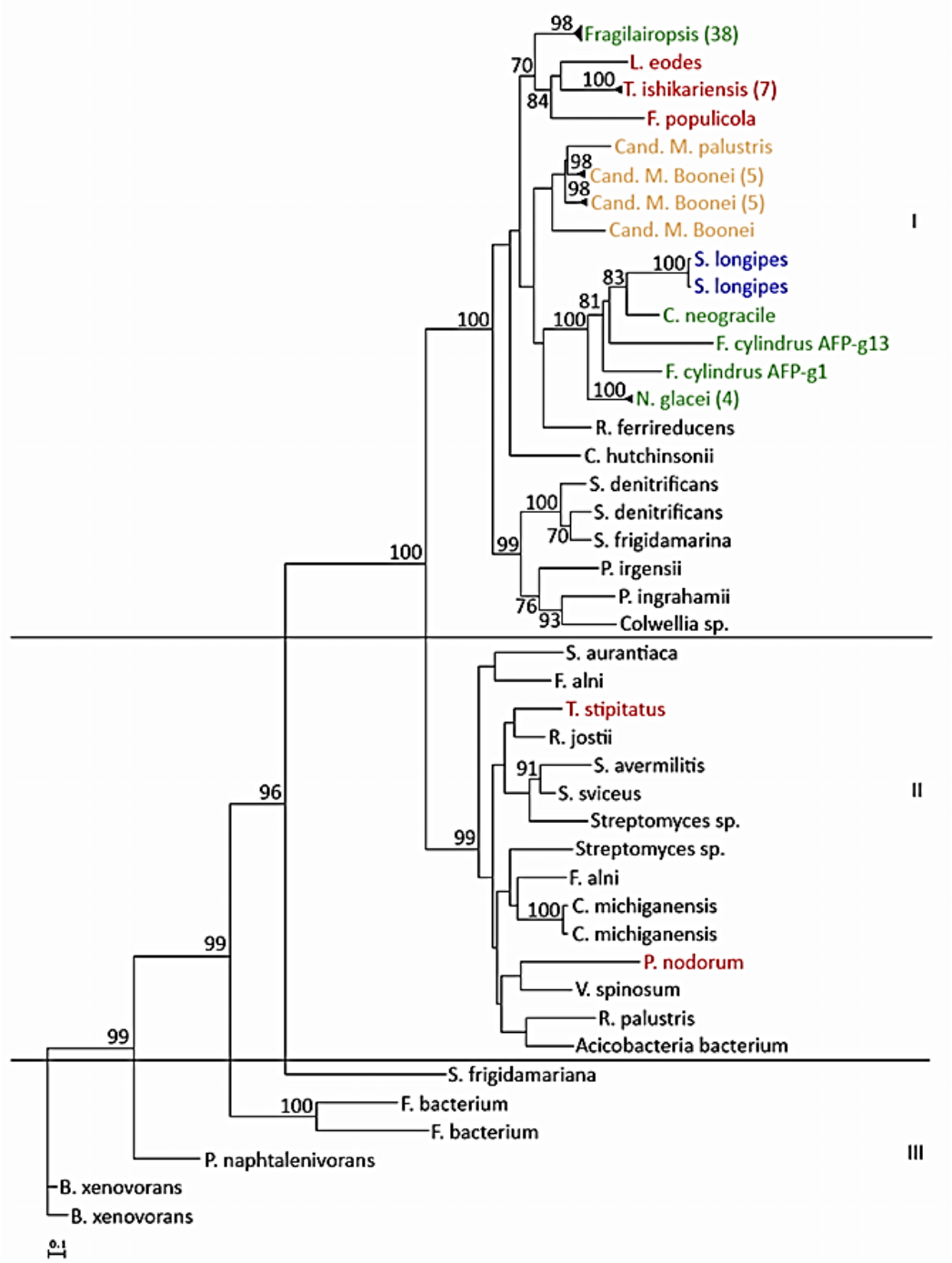



**Figure 18. Phylogeny of ice-binding proteins estimated with PhyML. PyML algorithm v3.0 was applied with the following settings: mode for amino acid substitution WAG, initial tree: BioNJ, 1000 bootstraps. Nodal supports greater than 60 are shown. In cases of multiple isoforms of one species are collapsed into groups, brackets show number of sequences. Archaea sequences are shown in orange, Bacteria in black, diatoms in green, fungi in red and crustaceans in blue (Christiane Uhlig, unpublished data).**

***Novel protein domain combinations.*** A total of 38 protein domain combinations could be identified in *F. cylindrus* which were not found in any other eukaryote. This included a putative aluminium activated malate transporter protein with homology to *Arabidopsis thaliana* (protein ID 242447), two proteins involved in cobalamin

biosynthesis (protein ID 241335 and 234422), four peptidase domain proteins (protein IDs 157190, 240859, 239622 and 239623) as well as the conserved chloroplast protein Ycf34 with unknown function (protein ID 291600). Additionally, manual annotation of the genome revealed a novel protein containing an N-terminal thioredoxin domain fused to a C-terminal hemerythrin metal binding domain (protein ID 240772). Furthermore, two carbonic anhydrases (CA), involved in carbon acquisition were identified in the genome of *F. cylindrus*, which contained novel protein domain combinations and were absent in *P. tricornutum* and *T. pseudonana*. One protein was found encoding for a putative alpha CA and contained an N-terminal frustulin domain (FcACA3, 264424) and a second protein encoding for a putative delta CA contained a fasciclin domain (FcDCA1, 264409).

A neighbour-joining phylogenetic tree was built with CA sequences from all three sequenced diatoms and frustulin sequences from *Cylindrotheca fusiformis* (CfFRU) as well as fasciclin domain-containing protein sequences from *F. cylindrus* (FcFAS1) showing that the putative frustulin domain-containing alpha CA from *F. cylindrus* (FcACA3) clustered intermediate between frustulins from *C. fusiformis* and *F. cylindrus* and other alpha CAs from *F. cylindrus* (Figure 19). In comparison, the putative fasciclin domain-containing delta CA from *F. cylindrus* (FcDCA1) clustered with other delta CAs from *F. cylindrus* and *T. pseudonana* and other fasciclin domain-containing protein sequences from *F. cylindrus* (FcFAS1) formed a separate cluster (Figure 19).

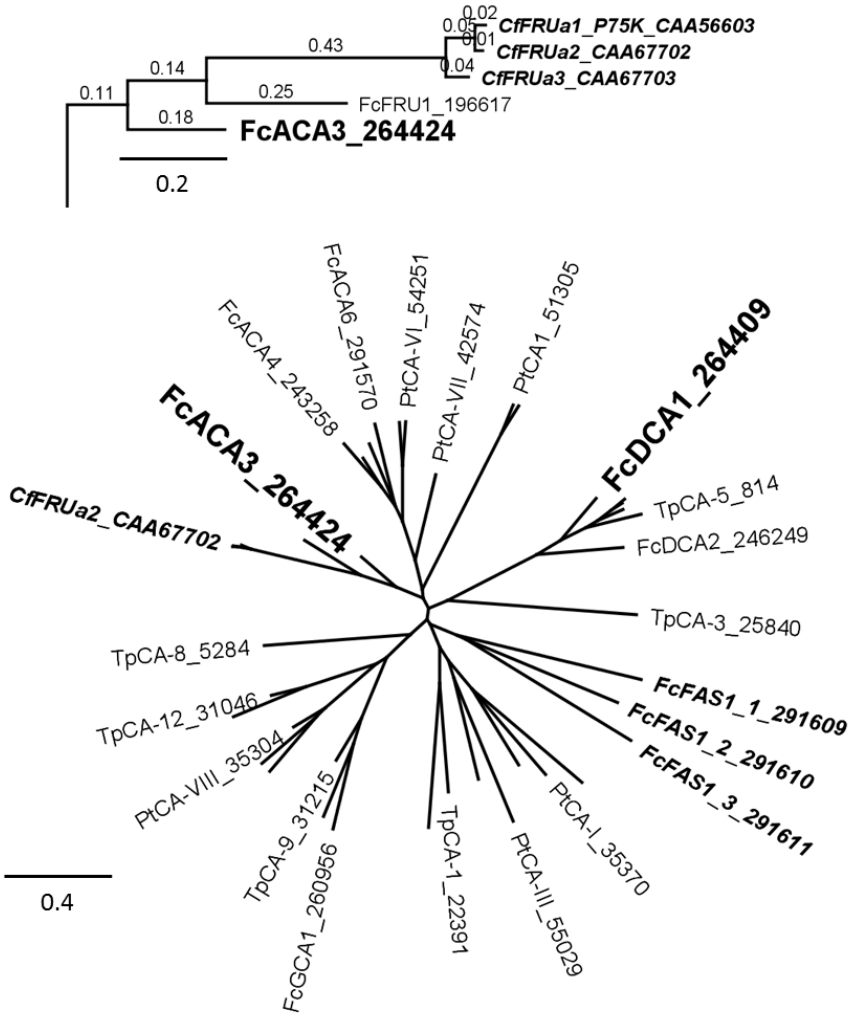



**Figure 19. Neighbour-joining phylogeny of carbonic anhydrases from all three sequenced diatoms (*F. cylindrus*, *P. tricornutum*, *T. pseudonana*), frustulins from *Cylindrotheca fusiformis* (CfFRU) and *F. cylindrus* (FcFRU) and fasciclin domain-containing proteins in *F. cylindrus* (FcFAS1). Labels are composed of species abbreviation, gene name and JGI protein identifier.**

A bestrophin domain (Pfam ID PF01062) involved in chloride membrane transport was found in 21 predicted proteins in *F. cylindrus* but could not be detected by computational analysis of protein family combinations in *P. tricornutum* and *T. pseudonana*. However, five bestrophin domain annotations could be detected by reciprocal best BLAST analysis and were assigned to the diatom core genome (Figure 10). In the same analysis 28 *F. cylindrus*-specific bestrophin Pfam annotations were detected and represented 0.23% in comparison to 0.06% of the total PFAM annotations in the diatom core genome resulting in more than five-fold amplification of bestrophin domains in *F. cylindrus* (Figure 10).

***Two-component and one-component systems including photoreceptors.*** Two-component and one-component signaling systems including photoreceptors could be

identified in *F. cylindrus* (Figure 20). Generally, two-component systems consist of a signal sensing histidine kinase (HisK) and its cognate response regulator (RR), which translates the input signal into a desired output subsequent to phosphorylation. In comparison, one-component systems combine input domain to an output domain in a single protein molecule (Ulrich et al., 2005). Due to the modular architecture of two-component systems additional sensory components have evolved and added to these systems (Cheung and Hendrickson, 2010; Schaller et al., 2011). Such sensory components include GAF (named after domain-containing proteins cGMP-specific phosphodiesterases, Adenylyl cyclases and Formate hydrogen lyase transcriptional activator), light-oxygen voltage (LOV), PAS (Per_ARNT-Sim; named after domain containing proteins period clock protein, aryl hydrocarbon receptor and single-minded protein), periplasmic binding protein (PBP) and phytochrome (PHY) domains. Except for a phytochrome-specific PHY domain, these sensory components could also be identified in *F. cylindrus* (Figure 20). In contrast to that, phytochrome two-component signalling systems, which are red/far-red light receptors related to histidine kinases (Möglich et al., 2010), have been identified in *T. pseudonana* (Tp*DPh*, protein ID 22848) and *P. tricornutum* (Pt*DPh*, protein ID 54330) (Depauw et al., 2012). They consist of two N-terminal PAS and GAF domains in addition to the photoreceptor-specific PHY domain constituting the phytochrome photosensory PAS-GAF-PHY tridomain (Möglich et al., 2010). However, other two-component systems containing GAF and PAS domains could be identified in *F. cylindrus* (Figure 20). Additionally, multiple copies of other photoreceptors including light-oxygen voltage (LOV) sensors and cryptochromes including novel variants of sensory domains could be identified (Figure 20). LOV sensors are blue light sensing photoreceptors, which utilise flavin nucleotide cofactors. The aureochrome family of LOV sensors contains transcription factors that comprise an N-terminal basic region/leucing zipper (bZ) DNA-binding domain and a C-terminal LOV domain that constitutes a subclass of the PAS family (Möglich et al., 2010). Thus, they resemble the modular composition of bacterial one-component systems (Ulrich et al., 2005). A total of seven aureochrome-like proteins could be identified in *F. cylindrus* in comparison to four aureochromes encoded in the genomes of *P. tricornutum* and *T. pseudonana*. Interestingly, in addition to aureochrome one-component systems, a protein with novel combinations of Helix-loop-Helix (HLH)-PAS domain was detected in *F. cylindrus* (protein ID 291565) (Figure 20), which could also be detected in *P. tricornutum* and *T. pseudonana*. Like aureochrome LOV sensors, cryptochromes are blue light photoreceptors, which could be detected in *F. cylindrus*

(Figure 20). They are flavoproteins whose photosensory domains are closely related to DNA photolyases and are composed of an N-terminal photolyase homology region (PHR) and a FAD-binding (FAD) domain (Möglich et al., 2010). A total of 8 genes encoding for proteins of the chryptochrome/photolyase family were identified in *F. cylindrus*, which is comparable to *T. pseudonana* and *P. tricornutum* (courtesy of Antonio E. Fortunato & Angela Falciatore, unpublished data). Finally, a rhodopsin with putative function as a light-driven proton pump was encoded in the genome of *F. cylindrus* (Figure 20), which is lacking in *P. tricornutum* and *T. pseudonana,* and in contrast to the previously described photoreceptors resembled an integral membrane protein.

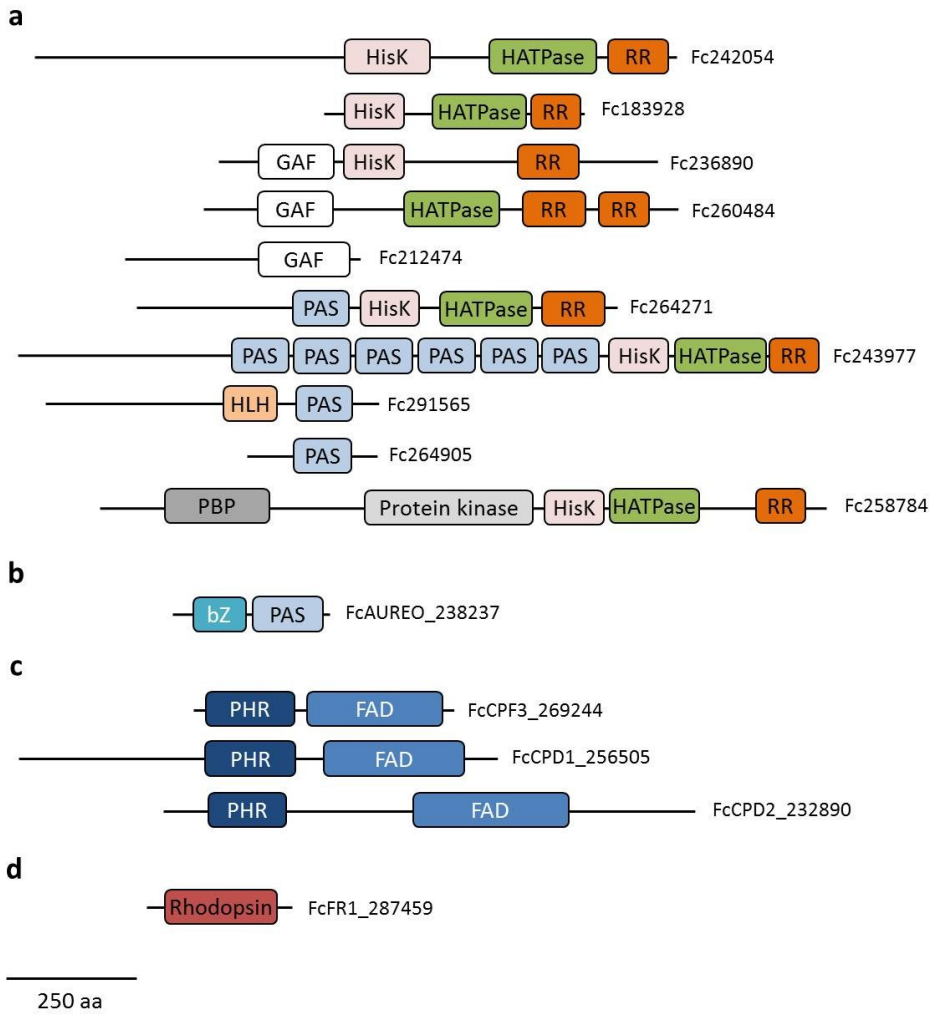



**Figure 20. Domain structures of selected two-component and one-component signaling systems including photoreceptors identified in *F. cylindrus* according to the NCBI conserved domain database (Marchler-Bauer A. *et al.* (2011), CDD: a Conserved Domain Database for the functional annotation of proteins, *Nucleic Acids Res* 39(D) 225-9). (a) Two-component and one-component signaling systems and sensory domain-containing proteins; (b) putative aureochrome photoreceptor belonging to light-oxygen-voltage (LOV) sensor family; (c) putative cryptochromes photoreceptor; (d) rhodopsin. Proteins are drawn approximately to scale and labelled with *F. cylindrus* protein identifiers. Domain abbreviations are PHR (photolyase homology region), FAD (flavin adenine dinucleotide-binding), bZ (basic region/leucine zipper), PAS (Per-ARNT-Sim domain), HisK (histidine kinase), HATPase ($H^+$-ATPase domain), RR (Response Regulator domain), PBP (periplasmic binding protein domain) and HLH (helix-loop-helix domain).**

***F. cylindrus* lacks core meiotic genes.** To explore the evidence whether *F. cylindrus* is a sexual organism, its gene inventory of core meiotic genes was analysed. Although some core meiotic genes homologous to those identified in other organisms could be identified in *F. cylindrus* (Table 7), at least three meiosis-specific core genes including homologs for homologous pairing proteins (e.g. HOP1, HOP2) and the meiotic recombination protein DMC1 could not be detected in *F. cylindrus* and neither in the other sequenced diatoms *T. pseudonana* and *P. tricornutum*.

**Table 7. Manually annotated core meiotic genes in *F. cylindrus*.**

| Enzyme Name | Gene Name in *F. cylindrus* | JGI protein identifier |
|---|---|---|
| meiosis-specific sporulation protein | *SPO11* | 239125 |
| meiotic nuclear division protein | *MND1* | 273989 |
| DNA mismatch repair protein mutS homolog4 | *MSH4* | 253011 |
| DNA mismatch repair protein mutS homolog5 | *MSH5* | 291615 |
| meiosis-specific DEAD-box helicase protein | *MER3* | 187564 |
| double-strand break repair protein RAD21 | *RAD21* | 263327 |
| DNA repair protein RAD51 | *RAD51* | 241710 |
| DNA repair and recombination protein RAD54 | *RAD54* | 259049 |

## 3.3 Discussion

The draft genome sequence of the psychrophilic diatom *F. cylindrus* provided the first genomic insights into the adaptation of diatoms to extreme polar conditions. Its nuclear genome was found to be 80.5 Mb and approximately 18,077 genes were predicted for its single-haplotype (Table 4). In comparison, nuclear genomes of the diatoms *P. tricornutum* (27.4 Mb; 10,402 predicted genes) (Bowler et al., 2008) and *T. pseudonana* (32.4 Mb; 11,776 predicted genes) (Armbrust et al., 2004) were significantly smaller and contained fewer genes (Table 4).

Strikingly, the *F. cylindrus* genome showed a high level of nucleotide sequence polymorphism, which affected the assembly of heterozygous haplotypes into a single haplotype and caused a diffuse haplotype structure resulting in the prediction of 27,137 genes including all gene copy variant pairs from heterozygous regions of the genome. The nucleotide sequence polymorphism was 6% between selected syntenic scaffolds. In comparison, nucleotide sequence polymorphism was found to be 0.75% in the genome of the centric diatom *T. pseudonana* (Armbrust et al., 2004), 4-5% in the sea urchin *Strongylocentrotus purpuratus* (Britten et al., 1978; Sodergren et al., 2006), 5% in the

genome of the sea squirt *Ciona intestinalis* (Dehal et al. (2002); Jeremy Schmutz, JGI, *personal communication* 23/06/2012) and 11.2% in the highly heterozygous genome of grapevine *Vitis vinifera* Pinot Noir (Velasco et al., 2007). While high levels of heterozygosity in marine invertebrates have been ascribed to large effective population sizes (Britten et al., 1978; Dehal et al., 2002), high heterozygosity in plant genomes has been related to genome duplication (The *Arabidopsis* Genome Initiative, 2000; International Rice Genome Sequencing Project, 2005; Tuskan et al., 2006; Jaillon et al., 2007; Velasco et al., 2007) and activity of transposable elements (Morgante et al., 2005; Velasco et al., 2007). For the *F. cylindrus* genome the widespread distribution of single nucleotide polymorphisms (SNPs) and insertion/deletions (in/dels) across the entire genome as well as the absence of large blocks in the alignments of scaffolds to each other in both orientation (data not shown) suggested no large scale genome duplication events consistent with the finding that large-scale genome duplication events are not a major driver in the generation of diatom diversity (Bowler et al., 2008).

To explore the hypothesis whether partial genome duplications or allelic variation contributed to high sequence polymorphism and heterozygosity of gene copy pairs in *F. cylindrus*, their sequence similarity and sequencing coverage was analysed. Two lines of evidence supported the dominance of allelic variation in *F. cylindrus* using assumptions on their sequence similarity and sequence read coverage. First, assuming paralogs to be more divergent than alleles, it was shown that a majority of pairs had a high nucleotide sequence similarity of > 98% (Figure 8). In comparison, allelic nucleotide variation was 1.2% in *C. intestinalis* (Dehal et al., 2002). Secondly, assuming that paralogs show a twofold higher sequencing coverage compared to alleles because they contain sequence reads from two assembled haplotypes, it was found that heterozygous gene copy pairs showed an approximately twofold difference in sequencing coverage (Supplementary Figure S7). These assumptions are, however, limited to the facts that recent gene duplication events would also show high sequence similarity between paralogs and our assumptions on sequencing coverage do not take biases in the assembly method into account, such as the assembly of nearly-identical reads into larger transcripts on larger scaffolds which may cause higher differences in coverage to allelic variants on smaller scaffolds (Robert P. Otillar, JGI, *personal communication* 28/06/2012).

The high degree of allelic polymorphism in *F. cylindrus* may be a result of the absence of sexual reproduction so that divergent alleles remain in a heterozygous state.

Indeed, sexual cycles have never been reported for *F. cylindrus* and neither for *P. tricornutum* and *T. pseudonana*. Consistently, it was shown in *P. tricornutum* that copia-like transposable elements were inserted in only one of the haplotypes at least a century ago and remained in a heterozygous state (Maumus et al., 2009). Furthermore, although some meiosis-related genes could be identified in *F. cylindrus* (Table 7) at least three core meiosis-specific genes including the homologous pairing proteins HOP1 and HOP2 (Leu et al., 1998) and the meiosis-specific DNA strand exchange protein DMC1 (Neale and Keeney, 2006) appeared to be absent and could neither be detected in the genomes of *P. tricornutum* nor in *T. pseudonana*. Yeast deletion mutants of meiosis-specific HOP2 showed recombination failure (Leu et al., 1998; Tsubouchi and Roeder, 2003). In addition, *HOP2* gene knockout mouse spermatocytes showed meiotic defects consistent with a failure in recombination (Petukhova et al., 2003). Moreover, DMC1 yeast mutants were shown deficient in meiotic recombination (Bishop et al., 1992) and DMC1 requirement in meiosis has recently been ascribed to the high resistance of DMC1 D-loops to dissociation by branch-migration proteins (e.g. RAD54, Table 7) during homologous recombination (Bugreev et al., 2011). Detection of some meiotic proteins in *F. cylindrus* (Table 7) may be explained by their function in mitosis and DNA repair (Schurko et al., 2009) and meiosis-specific genes may also be maintained in asexual species owing to neo-functionalization of their gene products (Meselson and Welch, 2007; Pouchkina-Stantcheva et al., 2007; Forche et al., 2008). However, evidence for asexual reproduction only from absence of meiosis-specific genes may be limited due to reports on absence of meiosis genes in species that undergo meiosis (e.g. absence of HOP2, MND1 and DMC1 in *Caenorhabditis* and *Drosophila* (Ramesh et al., 2005; Schurko and Logsdon, 2008)). Moreover, the high proportion of transposable elements of 7.3% in *F. cylindrus* as well as *P. tricornutum* (Table 4)is ambiguous for the absence of sexual reproduction, because LINE-like and gypsy-like retrotransposons were found absent in asexual bdelloid rotifer genomes (Arkhipova and Meselson, 2000). Nevertheless, TEs may also be acquired via horizontal gene transfer which was found pervasive in diatoms (Bowler et al., 2008) and play a key role in long term adaptation of natural diatom populations exposed to environmental stress through generation of genetic diversity (Maumus et al., 2009).

Thus, overall two lines of evidence, high allelic heterozygosity (“Meselson effect”) (Birky, 1996; Mark Welch and Meselson, 2000; Meselson and Welch, 2007) and putative lack of core meiosis-specific genes (Villeneuve and Hillers, 2001; Ramesh et

al., 2005; Malik et al., 2008; Schurko and Logsdon, 2008; Schurko et al., 2009) suggest the absence of sexual reproduction in *F. cylindrus* so that allelic heterozygosity can increase in every generation (Birky, 1996). The high level of allelic variation in *F. cylindrus* may benefit from multilocus heterozygosity-fitness correlations as a result from direct selection on scored genetic loci (Hansson and Westerberg, 2002) providing higher fitness through heterozygote advantage (Sellis et al., 2011; Hedrick, 2012).

The high proportion of TEs found in the *F. cylindrus* genome (Table 4) may have contributed to genome architecture and its high degree of heterozygosity as suggested for plants (Morgante et al., 2005; Velasco et al., 2007). As discussed above the generation of genetic diversity by TEs may also play a role in long term adaptation of diatom exposed to environmental stress (Maumus et al., 2009) and their expansion in the pennate diatom lineage may have been driving pennate diatom diversity through transpositional duplications and subsequent genome fragmentation (Bowler et al., 2008). Overall, the repeat content in the *F. cylindrus* genome was more than twofold higher than in the genomes of *P. tricornutum* and *T. pseudonana* with significant contributions from simple sequence repeats and unclassified repeats (Table 4) and was likely to have contributed to the increased genome size.

Furthermore, the high number of 11,511 species-specific genes in *F. cylindrus* associated with its diffuse haplotype structure as well as the high number of 6913 species-specific genes and 2859 species-specific paralogs estimated for its single-haplotype (Table 4) may have contributed to the increase in genome size consistent with the correlation of genome size with genomic landscape (number of genes, introns, mobile elements) (Lynch and Conery, 2003; van Nimwegen, 2003). Moreover, the high number of species-specific paralogs indicates a role of gene duplication in gene diversification and genome expansion in comparison to *P. tricornutum* and *T. pseudonana*.

As genome size is a biological trait at the intersection of genotype and phenotype it may have evolutionary significance (Oliver et al., 2007) and across all kingdoms of life genome size has been shown to statistically correlate with various phenotypic traits (Bennett, 1987; Gregory, 2001; Kozłowski et al., 2003; Knight et al., 2005; Vinogradov and Anatskaya, 2006; Francis et al., 2008; Veselý et al., 2012). Thus it may be concluded that high proportion of TEs, acquisition of species-specific sequences and gene duplications in *F. cylindrus* and their effect on increasing genome

size may relate to adaptation to environmental stress. Consistently, gene copy number variation has been suggested to provide a mechanism of phenotypic differentiation and evolutionary adaptation to the environment (Hastings et al., 2009; Sudmant et al., 2010; Dassanayake et al., 2011a) and indeed, adaptation to environmental stress by acquisition of lineage-specific sequences and gene duplication has been observed in extremophile plants (Dassanayake et al., 2011b; Oh et al., 2012).

In addition to genome size, G+C content is a basic parameter of genomes and can vary widely for prokaryotes (Hallin and Ussery, 2004) and eukaryotes (e.g. (Yu et al., 2002)). The coding G+C content of 39.8% of the *F. cylindrus* genome was found to be significantly lower compared to G+C contents of 50.6% in *P. tricorntum* and 47.8% in *T. pseudonana* (Table 4) and is, to our knowledge, the lowest G+C in coding sequences observed for autotrophic eukaryotes (The *Arabidopsis* Genome Initiative, 2000; Yu et al., 2002; Matsuzaki et al., 2004; Derelle et al., 2006; Merchant et al., 2007; Worden et al., 2009; Cock et al., 2010; Prochnik et al., 2010; Dassanayake et al., 2011b; D'Hont et al., 2012). It has been suggested that there is a universal mutational bias towards AT in both prokaryotes (Lind and Andersson, 2008; Hershberg and Petrov, 2010; Hildebrand et al., 2010) and eukaryotes (Petrov and Hartl, 1999; Lynch et al., 2008; Denver et al., 2009; Keightley et al., 2009; Lynch, 2010; Ossowski et al., 2010). AT mutational bias for *A. thaliana*, *Drosophila* and mammalian genomes has been at least partly attributed to spontaneous deamination of methylated cytosines (Petrov and Hartl, 1999; Ossowski et al., 2010), which leads to thymine substitutions (Lindahl and Nyberg, 1974; Coulondre et al., 1978; Duncan and Miller, 1980) (for review see (Lindahl, 1993)). However, as a high proportion of G:C sites in *A. thaliana* not reported to be methylated also showed higher rates of mutational bias than A:T sites (Ossowski et al., 2010), the authors suggest an additional mutational effect of ultraviolet (UV) light, which causes a mutational bias towards A:T at dipyrimidine sites (C adjacent to another C or to a T) in pro- and eukaryotes (Friedberg et al., 2006). Additionally, DNA replication and DNA repair mechanisms in fast growing cells including unicellular eukaryotes have been suggested to effect mutations caused by different specificities and fidelity of specific DNA polymerases (Friedberg et al., 2002). Consistently, it has been suggested that GC content variation in bacteria is governed by genome replication and DNA repair mechanisms (Lind and Andersson, 2008) and influenced by variations in the structure of the catalytic subunits of DNA polymerase (Zhao et al., 2007; Wu et al., 2012). In this context it is noteworthy, that DNA-directed DNA polymerase (EC 2.7.7.7)

was the most highly represented European Commission (EC) enzyme number annotation in *F. cylindrus* resulting in three-fold amplification in comparison to the diatom core genome (data not shown) and that the mismatch repair (MMR) DNA repair pathway was expanded in *F. cylindrus* in comparison to *P. tricornutum* and *T. pseudonana* (Antonio E. Fortunato and Angela Falciatore, Génomique Fonctionnelle des Diatomées, Paris, France, *personal communication* 08/07/2011). Last, not least G+C content of microbial communities seems to be globally and actively affected by their environment (Foerstner et al., 2005). Interestingly, it was found that diatom sequences selected from natural phytoplankton metatranscriptomes sampled from polar, temperate and tropical microbial communities showed variable G+C contents with G+C content of ~38% in polar communities (T. Mock *et al*, manuscript in prep.). In conclusion, however, the existence of a mutational bias towards AT and its potential causes and effects in diatoms, such as *F. cylindrus*, remains unknown.

The low G+C content in *F. cylindrus* had a significant impact on codon usage pattern causing a bias towards adenine and thymidine bases (Figure 6; Figure 7), consistent with previous studies (Kanaya et al., 2001; Knight et al., 2001; Chen et al., 2004; Hershberg and Petrov, 2009). We showed a tendency of *F. cylindrus* to favour AT rich codons (Figure 6) as shown for bacteria, archea and fungi (Hershberg and Petrov, 2009). Moreover, we found that the low G+C in *F. cylindrus* also affected the anticodon usage of the 330 identified tRNA genes showing a prevalence of A/T at anticodon position 1 (Figure 7), which is in disagreement with the low variation in tRNA anticodon composition in bacteria (Rocha, 2004). However, it is in good agreement with the thermal adaptation hypothesis proposed by Bernardi (Bernardi and Bernardi, 1986; Bernardi, 2000) based on higher thermal stability of G:C pairs in comparison to A:T pairs due to presence of three hydrogen bonds between G:C pairs and two between A:T pairs (Wada and Suyama, 1986). In higher eukaryotes regional nucleotide compositional changes to high GC accompanied a transition from cold to warm-blooded vertebrates and codon third positions showed a linear dependence of the regional GC level (Bernardi and Bernardi, 1986; Bernardi, 2000). Although, the existence of correlations between genomic G+C content and optimal growth temperature remains controversial for the prokaryotic domain (Marashi and Ghalanbor, 2004; Musto et al., 2004; Basak et al., 2005), enrichment in G+C has been observed for tRNA and rRNA sequences in thermophilic bacteria (Galtier and Lobry, 1997) and A:U base pairing was prevalent in 16S rRNA in psychrophilic prokaryotes suggesting a strong thermo-adaptive

mechanism (Khachane et al., 2005). Moreover, although recently challenged (Lobry and Necşulea, 2006), synonymous codon usage in prokaryotes has been associated with stability of codon-anticodon interaction in thermophilic bacteria (Lynn et al., 2002; Lobry and Chessel, 2003; Singer and Hickey, 2003; Basak and Ghosh, 2005) and for eukaryotes sticky codon-anticodon interactions were avoided whenever possible to maintain uniform interaction energy and smooth progression of translation in yeast (Bennetzen and Hall, 1982). In analogy to that, our finding of codon biases towards adenine and thymidine bases in *F. cylindrus* (Figure 6; Figure 7) may suggest a co-adaptation of relative frequencies of codons and their respective anticodons to optimise protein translation at low temperature, which is the most energetically expensive process in exponentially growing cells (Rocha, 2004; Wilson and Nierhaus, 2007).

In summary, although the cause of the low G+C in *F. cylindrus* remains unknown, it has a significant impact on protein translation through codon-anticodon usage bias towards AT, which may relate to a selective process. It may further have a direct impact on the amino-acid composition of proteins (Sueoka, 1961; Bharanidharan et al., 2004; Foerstner et al., 2005) and act complementary to a drift in amino acid usage over evolutionary timescales (Jordan et al., 2005). In addition to G+C content, purine content (A+G) has been suggested to contribute to nucleotide composition of protein-coding sequences (Zhang and Yu, 2010). Notably, purine metabolism was the most highly represented KEGG pathway annotation in *F. cylindrus* and significantly enriched in comparison to the diatom core genome (Figure 11). As it has been reported that purines play a role in the determination of amino acid physicochemical properties and purines at the second codon position may control the charge and hydrophobicity of amino acids (Taylor and Coates, 1989; Chiusano et al., 2000; Biro et al., 2003; Copley et al., 2005; Yu, 2007), a possible link to cold adaptability of protein in relation to structural rigidity may exist. Indeed, widespread amino acid modifications have been observed in psychrophilic bacteria through metagenomic (Grzymski et al., 2006) and genomic analysis (Saunders et al., 2003; Methe et al., 2005; Ayala-del-Río et al., 2010; Zhao et al., 2010), most notably resulting in a generally reduced hydrophobic amino acid content (see Casanueva *et al.* (2010) for review). However, a global analysis of the *F. cylindrus* proteome was beyond the scope of this work.

The genomic data has also been evaluated in the context of gene repertoire to the evolution and adaptation of *F. cylindrus* to environmental constraints of the Southern Ocean including trace metal availability, low temperatures and low light conditions.

Trace metals, including iron, play a key role in photosynthetic electron transport (Merchant and Dreyfuss, 1998). Many studies have highlighted the impact of iron limitation on the physiology and evolutionary adaptation of diatoms in the high nutrient low chlorophyll (HNLC) oceanic regions (Martin and Fitzwater, 1988; Sunda et al., 1991; Hutchins and Bruland, 1998; Takeda, 1998; Hutchins et al., 1999; Timmermans et al., 2001; Quigg et al., 2003; Strzepek and Harrison, 2004; Boyd et al., 2007) and an adaptive strategy appears to be the substitution of iron with copper, which has similar electrochemical properties and is more abundant in these regions (Bruland, 1980) and iron sparing. The *F. cylindrus* genome encodes for the Cu-based photosynthetic electron carrier plastocyanin, which was shown to replace the Fe-containing cytochrome $c_6$ in the photosynthetic electron transport chain of the oceanic diatom *Thalassiosira oceanica* in comparison to its coastal counterpart *T. weissflogii* (Peers and Price, 2006) and is also lacking in the genomes of the sequenced coastal diatoms *P. tricornutum* and *T. pseudonana* (Armbrust et al., 2004; Bowler et al., 2008). Furthermore, in contrast to *T. pseudonana,* no homolog for beta-carotene hydroxylase (CHYB) involved in carotenoid biosynthesis was detected in *F. cylindrus* and neither in *P. tricornutum* suggesting that pennate diatoms may have replaced the nonheme di-iron hydroxylase CHYB by a cytochrome P450 monooxygenase with only one iron atom in the catalytic centre, which might be advantageous in Fe-limited marine environments (Bertrand, 2010). However, the *F. cylindrus* genome contained ~3% iron-binding proteins, which is comparable to many other sequenced marine algae from iron-replete oceanic regions (Figure 13) suggesting an essential iron requirement even under chronic iron limitation and Fe-binding domains including a cytochrome P450 domain was enriched in comparison to the diatom core genome (Figure 10). Nevertheless, noteworthy in the context of iron usage and acquisition was the presence of putative Fe-binding hemopexin domain-containing proteins in *F. cylindrus*, which were absent in *T. pseudonana* and *P. tricornutum* and may play a role in the recycling of heme-bound iron to maintain iron homeostasis (Tolosano et al., 2010). Moreover, the presence of protein-coding genes involved in high affinity iron uptake systems including several isoenzymes for ferric-chelate reductase as well as an iron permease, a ferroportin and ferritin (Marchetti et al., 2009) may play a role in a low iron environment. Indeed, analysis of functional gene annotations in *F. cylindrus* confirmed the potential importance of iron acquisition pathways by finding significant enrichments for the COG group annotation “Ferric reductase, NADH/NADPH oxidase and related proteins” and Pfam annotation “Ferric reductase-like transmembrane component” in comparison to

the diatom core genome (not shown). Together with the presence of putative proteins serving as iron siderophores, including ferritin (Marchetti et al., 2009) and putative genes involved in enterobactin biosynthesis, may be indicators of adaptation to a low iron environment.

In contrast to iron-binding proteins, *F. cylindrus* has significantly expanded its copper-binding proteins (0.6%; Figure 13; Figure 14) in comparison to other sequenced diatoms. However, it has been shown that the total number of metal-binding domains including copper binding domains scale to the nuclear genome size as a power law, with different slopes for different metals and kingdoms of life (Dupont et al., 2006). The analysis of copper-binding domains in phytoplankton genomes including *F. cylindrus* showed a power law slope of 1.8 and indicated a nearly quadrupling of copper-binding domains with doubling in genome size of phytoplankton (Figure 14). While this scaling is empirical and gives no information about the mode of domain accumulation (e.g. duplications, gene transfer), it shows a selective retention of copper-binding domains. Both copper-binding proteome of 0.6% (Figure 13) in *F. cylindrus* and power law of scaling of phytoplankton (Figure 14) far exceeds that observed for prokaryotes showing ~0.3% copper-binding domains on average (Dupont et al., 2010). This preferential recruitment and retention of copper-binding domains of phytoplankton is consistent with their evolution after hypothesized $O_2$-driven changes in global trace metal geochemistry (Anbar and Knoll, 2002; Anbar, 2008). Even in consideration of the shared scaling of Cu-binding domains in phytoplankton, the genomes of *F. cylindrus*, *Chlamydomonas reinhardtii*, and *Aureococcus anophagefferens* deviated from the global trend (Figure 13).

In addition to Fe and Cu, Zinc (Zn) is involved in many metabolic processes (Vallee and Auld, 1990) and an important metal cofactor in phytoplankton. Zn metalloenzymes include transcription factors (Montsant et al., 2007), alkaline phosphatase (Shaked et al., 2006) and carbonic anhydrase (Morel et al., 1994; Hu et al., 2003), enzymes which are also encoded in *F. cylindrus* (see discussion below). While the total number of zinc-binding protein domains in *F. cylindrus* was comparable to other phytoplankton genomes, the conserved Zn-binding myeloid-Nervy-DEAF-1 (MYND) domain (named after myeloid translocation protein 8, Nervy and DEAF-1) was greatly expanded (Figure 15). Two MYND domains were found associated with the Fe-containing Hypoxia Induction Factor prolyl hydroxylase (HIF) domain (not shown) involved in the cellular response to changing oxygen in Eukarya (Benizri et al., 2008)

and may be involved in the response to extreme changes in oxygen content in sea ice cause by the photosynthetic activity of sea ice diatoms (Thomas and Dieckmann, 2002), which may ultimately lead to oxidative stress in sea ice algae like *F. cylindrus* (McMinn et al., 2005). Interestingly, in the context of oxidative stress was the presence of five Fe-binding hemoproteins containing globin-like domains in *F. cylindrus* including isoenzymes for neuroglobin, which were not detected in *P. tricornutum* and *T. pseudonana* and may be involved in oxidative stress defence through their capacity to bind oxygen and to detoxify reactive oxygen species (Herold et al., 2004; Verde et al., 2009; Giordano et al., 2012). In addition to MYND domains associated with HIF, MYND domains also exist in a large number of proteins including those involved in mediation of protein-protein interactions and transcriptional regulation (Gross and McGinnis, 1996; Liu et al., 2007), and MYND domains in *F. cylindrus* were always in combinations with DNA-binding or protein-protein binding domains suggesting they play a role in signal perception and transductions systems and regulate transcriptional responses to environmental stresses in concert with expanded protein kinase families in *F. cylindrus* (not shown). Interestingly, transcription factors were next to Zn-binding domains the most striking expansion of proteins in *F. cylindrus* (Figure 10; Figure 12; Figure 16) and the enrichment of heat and cold shock factors in *F. cylindrus* in comparison to the diatom core genome may indicate the central importance of transcriptional regulation in response to environmental stress in diatoms (Montsant et al., 2007). Last, but not least, phylogenetic analysis of Zn-binding MYND domains in *F. cylindrus* showed a high nucleotide divergence (data not shown), which is likely to have occurred within the last 30 million years (Myr), suggesting that it may be an evolutionary event coupled with the geological history of the Southern Ocean and its rise through tectonic opening of ocean gateways between Antarctica and Australia (Tasmanian Passage), and Antarctica and South America (Drake Passage) ~34 Myr ago (DeConto and Pollard, 2003). The resulting isolation of Antarctica through the organisation of the Antarctic Circumpolar Current causing extreme polar conditions (Kennett, 1977; Exon et al., 2002) together with relatively high zinc concentrations of the Southern Ocean (Croot et al., 2011) may have been maintained this great expansion of Zn-binding MYND domains and contributed to the evolution and adaptation of *F. cylindrus*.

The capacity to cope with cold stress and survival below the freezing point of seawater within sea ice is a requirement for many phytoplankton species in polar

oceans. *F. cylindrus* is able to thrive in sea ice with temperatures down to −20 °C. Sequencing of expressed sequence tag (EST) libraries from *F. cylindrus* pioneered the discovery of a new class of ice-binding proteins (Janech et al., 2006). The genome of *F. cylindrus* encodes for 12 ice-binding proteins (IBPs) and 11 proteins with C-terminal IBP domains, which resemble adaptations to life in sea ice and are not found in the mesophilic diatoms *T. pseudonana* and *P. tricornutum*. Similar proteins were found in bacteria and fungi (Raymond and Janech, 2009), as well as a sea ice crustacean (Kiko, 2010) suggesting that they were acquired via horizontal gene transfer (Raymond and Kim, 2012) (Figure 19). IBPs have been proposed to serve different functions including antifreeze activity, inhibition of ice crystallisation, attachment to ice and retention of a liquid environment (Raymond, 2011) and gene expression analysis showed that they were expressed under cold and salt stress in *F. cylindrus* (Krell et al., 2008; Bayer-Giraldi et al., 2010). Interestingly, in the context of salt stress, was the significant enrichment of the bestrophin protein family in the genome of *F. cylindrus* in comparison to the diatom core genome (Figure 10), as bestrophins are known to act as chloride channels (Tsunenari et al., 2003) and thus may be involved in the transport of chloride anions to maintain cellular osmolyte homeostasis in salt brine of sea ice (Krell, 2006; Boetius and Joye, 2009).

Moreover, it was striking that the *F. cylindrus* genome showed a significant expansion of helicase associated protein domains in comparison to the diatom core genome (Figure 10). DNA/RNA helicases, catalysing the unwinding of duplexes and secondary structures of nucleic acids, were found up-regulated under cold stress in *F. cylindrus* and a function in the control of secondary structures of DNA/RNA under low temperature stress to keep transcription and translation active was suggested (Mock and Valentin, 2004; Mock et al., 2005).

Furthermore, the significant enrichment of N-glycan metabolic processes (N-Glycan biosynthesis and degradation) in *F. cylindrus* compared to the diatom core genome (Figure 11) was notable and may contribute to the formation of a glycoprotein-rich extracellular matrix (Krembs et al., 2011), functioning in concert with IBPs to shape the microstructure of sea ice (Bayer-Giraldi et al., 2011). Additionally, the enrichment of lipid metabolism in *F. cylindrus* in comparison to the diatom core genome (Figure 11; Figure 12; Figure 16) may represent an adaptation to life in sea ice as an increase in the synthesis of unsaturated fatty acids to maintain membrane fluidity at low temperatures is suggested to be an molecular adaptation to psychrophilic lifestyle

(Suutari and Laakso, 1994; Chattopadhyay, 2006; Morgan-Kiss et al., 2006; Casanueva et al., 2010). Protein families involved in lipid metabolism were also found enriched in the polar terrestrial green alga *Coccomyxa subellipsoidea* (Blanc et al., 2012).

Furthermore, *F. cylindrus* showed a significant enrichment of porphyrin and chorophyll metabolism (Figure 11) and light-harvesting LHCX proteins compared to other sequenced diatoms (Figure 17), which may reflect adaptation to extreme fluctuations in light intensities in sea ice. As expected, the genome of *F. cylindrus* encodes for homologs of most enzymes known to be directly involved in biosynthesis of chlorophyll and carotenoids (Table 6). However, like in other diatoms no homolog for CHL27, a subunit of the Mg-protoporphyrin IX monomethyl ester (MPE) cyclase (Tottey et al., 2003), which is nucleus-encoded in vascular plants and green algae but encoded on the plastid genome in red algae (Lohr et al., 2005), could be found in neither the nuclear nor chloroplast genome of *F. cylindrus*. This supports the suggestion that diatoms might use an unrelated enzyme to form the isocyclic ring of chlorophylls, because no CHL27 homologs could be found in the nuclear and plastid genomes of *P. tricornutum* and *T. pseudonana* (Wilhelm et al., 2006) as well as the plastid genome of the centric diatom *Odontella sinensis* (Kowallik et al., 1995) either. Moreover, like in other diatoms no homolog for subunits of the light-independent protochlorophyllide oxidoreductase (DPOR), which are plastid-encoded in green algae, mosses and gymnosperms, could be detected in the nuclear and plastid genome of *F. cylindrus*. Although some red algal plastids contain genes encoding for DPOR subunits (Reith and Munholland, 1993; Gloeckner et al., 2000), others have lost those genes (Ohta et al., 2003; Hagopian et al., 2004) and no homologs for DPOR subunits could be identified on the plastid genome of *O. sinensis* as well as on the nuclear or the plastid genome of *T. pseudonana* (Wilhelm et al., 2006) and *P. tricornutum*. Additionally, like in other diatoms no homolog with lycopene epsilon-cyclase (LCYE), which catalyses the conversion of lycopene to alpha-carotene in higher plants, could be identified in *F. cylindrus* and was also not detected in found in *T. pseudonana* and *P. tricornutum* (Coesel et al., 2008) explaining why diatoms only contain carotenoids derived from beta-carotene. Nevertheless, the identification of at least six putative isoenzymes involved in carotenoid biosynthesis (Table 6) may have contributed to the enrichment of chlorophyll metabolism in the *F. cylindrus* genome in comparison to the diatom core genome (Figure 11). Interestingly, carotenoids have also been reported to play a role in the regulation of membrane fluidity (Subczynski et al., 1992) and by this means

postulated to play in cold adaptation of psychrophilic bacteria (Chattopadhyay and Jagannadham, 2001; Chattopadhyay, 2006).

The identification of 64 light harvesting complex (LHC) gene models in *F. cylindrus* (Figure 17), out of which 55 models were supported by ESTs indicated their expression and physiological importance. Although number of LHC genes seemed to scale with genome size it appeared that the LHC family of LHCX proteins, which is involved in photoprotection (Peers et al., 2009; Zhu and Green, 2010) did not (Figure 17). Strikingly, the *F. cylindrus* genome contained 11 gene models encoding LHCX proteins (Figure 17) including a LHCX1 homolog, which was demonstrated to be a regulator of photoprotection via non-photochemical quenching (NPQ) in *P. tricornutum* (Bailleul et al., 2010). A comparable high number of LHCX proteins could be detected in the eukaryotic algae genomes of *Ostreococcus* spp (Palenik et al., 2007), *Emiliania huxleyi* (Betsy A. Read *et al.*, unpublished) and *Ectocarpus siliculosus* (Cock et al., 2010) suggesting a relation to the natural light environment of predominantly high light stress in these algae.

Finally, the acclimation to changes in light conditions in *F. cylindrus* requires quantitative regulation of photosynthetic pigments and changes in pigment composition. Particularly physiological adaptations to low light conditions are a prerequisite for life in Antarctic sea ice, because sea ice is an effective barrier to light transmission, especially when covered with snow (Thomas and Dieckmann, 2002; Thomas and Dieckmann, 2010). As shown by the *F. cylindrus* genome this seems to be achieved by expansion of genes involved in chlorophyll and carotenoid biosynthesis as well as LHC proteins supporting the hypothesis that these diatom-specific photoprotective mechanisms play an important role in the ecological adaptation and success of diatoms to fluctuating marine environments (Strzepek and Harrison, 2004).

*F. cylindrus* has to endure up to six months of darkness during polar winter. However, its mode of overwintering in polar sea ice has never been reported and remains unclear. Interestingly, the *F. cylindrus* genome showed a significant enrichment of genes associated with carbohydrate metabolism including “starch and sucrose metabolism” and “glyoxylate and dicarboxylate metabolism” in comparison to the diatom core genome (Figure 11). The enrichment of carbohydrate metabolic processes together with the identification of two complete mitochondrial and peroxisomal pathways for the beta-oxidation of fatty acids in *F. cylindrus* (not shown), which

generate ATP and feed into gluconeogenesis for carbohydrate production, may provide an adaptive mechanism to endure long periods of darkness. In general, sea ice diatoms may survive prolonged darkness by switching from autotrophy to heterotrophic uptake of organic matter (Palmisano and Garrison, 1993), reduction of their metabolism to a lower level of activity (Jochem, 1999), formation of winter growth stages/resting spores (Fryxell and Prasad, 1990), utilisation of intracellular reserves of energy-rich substances (e.g. lipids and carbohydrates) (Weger et al., 1989; Fogg, 1991; Stehfest et al., 2005) and it is likely that a succession of those mechanisms occurs in the environment. In a laboratory study with a mesophilic diatom it could be shown that the degradation of storage products during darkness followed a well-defined sequence starting with an initial decline of lipids, followed by carbohydrates and proteins (Stehfest et al., 2005).

Last, not least physiological adaptations of carbohydrate metabolic processes and organic carbon uptake in *F. cylindrus* may also be related to peculiarities in sea ice carbon chemistry, which arise as a consequence of photosynthetic activity of sea ice algae and lead to depletion of inorganic carbon and increases of pH up to pH 11 in sea ice brine (normal seawater has a pH of ~8) (Gleitz et al., 1995). This causes knock-on effects on carbon acquisition of sea ice algae due to a low substrate affinity and slow turnover rate of the carbon-fixing enzyme Ribulose bisphosphate carboxylase/oxygenase (RubisCO) (Giordano et al., 2005). Thus, diatoms including *F. cylindrus* have evolved mechanisms to concentrate dissolved inorganic carbon via carbon concentrating mechanisms (CCM) (Giordano et al., 2005) using the Zn enzyme carbonic anhydrase (CA), which reversibly catalyses the interconversion of bicarbonate ($HCO_3^-$) and $CO_2$ (Reinfelder et al., 2000). Suprisingly, *F. cylindrus* encodes for two novel CA proteins containing a frustulin domain and a fasciclin domain. As frustulins are calcium-binding proteins involved in diatom cell wall biogenesis (Kroeger and Poulsen, 2008) and fasciclins are extracellular cell adhesion proteins (Huber and Sumper, 1994), both proteins might function as extracellular CAs to facilitate the rate of $CO_2$ formation in the laminar layer surrounding the cells. Their functioning may profit from a postulated proton buffering role of the diatom silicate cell wall (Milligan and Morel, 2002) providing protons involved in the interconversion between $HCO_3^-$ and $CO_2$ more efficiently than water (Tripp and Ferry, 2000).

Last not least, the identification of a light-driven bacteria-like rhodopsin proton pump in *F. cylindrus* was surprising against the background of a chlorophyll-based proton gradient-generating photosynthetic apparatus. In several heterotrophic and

mixotrophic dinoflagellates rhodopsin has been proposed to fuel light-driven ATP synthesis (Lin et al., 2010; Slamovits et al., 2011) and it has been suggested to provide an trace metal-independent mechanism to enhance ATP production in photosynthetic diatoms when photosynthesis is iron-limited (Raven, 2009; Marchetti et al., 2012).

## 3.4 Summary and conclusion

The draft genome of the psychrophilic diatom *F. cylindrus* provides unprecedented insights into how polar environmental conditions can shape the genome of a eukaryotic extremophile organism. Genomic changes relative to mesophilic organisms include a high degree of heterozygosity between homologous chromosomes combined with the presence of highly diverged alleles, an increased genome size with higher number of species-specific genes, a low G+C content and the expansion of genes and protein families related to life in the cold. Against the background that adaptation to an extreme environment is likely to be conferred not only by a specific set of genes but rather a collection of synergistic changes in genome content, structure and amino acid composition of enzymes (Methe et al., 2005), this new genomic data provides new insights into novel physiology and is a starting point to specifically probe adaptive strategies to the polar environment.

# Chapter 4

# Transcriptome analysis of the psychropilic diatom *Fragilariopsis cylindrus* using RNA-Sequencing

## 4.1 Introduction

The obligate psychrophilic polar diatom *F. cylindrus* is a key species found in seawater and sea ice of the Arctic and Southern Ocean (Lundholm and Hasle, 2008) and provides a model to study polar adaptation of phytoplankton (Mock and Valentin, 2004; Mock and Hoch, 2005; Mock et al., 2005; Krell et al., 2007; Bayer-Giraldi et al., 2010; Lyon et al., 2011). The draft genome sequence of the obligate psychrophilic polar diatom *Fragilariopsis cylindrus* (Mock et al.) reveals its genetic blueprint and gives insights into evolution and adaptation of phytoplankton to polar oceans.

However, we cannot predict from a genome sequence when and in which quantities genes are expressed, but these gene expression patterns are crucial to understanding how an organism coordinates its cellular functions in response to environmental changes (Van de Peer, 2011). Furthermore, gene expression patterns provide insights into functions of genes with unknown annotation in diatoms (Allen et al., 2008; Mock et al., 2008) and assist in the prediction of functional elements of genomes in model organisms (Ross-Macdonald et al., 1999; Boone et al., 2007; modENCODE et al., 2011). Applying a transcriptomics approach, we can catalogue different species of transcripts including mRNAs and non-coding RNAs, quantify changing expression levels of each transcript under different conditions and determine the transcriptional structure of genes.

The high-throughput sequencing of complementary DNAs (cDNA) called RNA sequencing (RNA-Seq) provides a powerful tool for transcriptomics (Nagalakshmi et al., 2008; Wang et al., 2009). RNA-Seq relies on the principle that a population of RNA is converted to a library of cDNA, which then is ligated to sequencing adaptors and subjected to high-throughput sequencing. The final mapping of cDNA sequencing reads to a reference genome (Lister et al., 2008; Mortazavi et al., 2008; Nagalakshmi et al., 2008; Wilhelm et al., 2008) or their *de novo* assembly (Birol et al., 2009; Li et al., 2009; Birzele et al., 2010; Robertson et al., 2010; Grabherr et al., 2011) provides a digital measure for the abundance of transcripts. In model organisms RNA-Seq has been used

to improve existing genome (Mortazavi et al., 2008; Nagalakshmi et al., 2008; Wilhelm et al., 2008) annotations as well as to provide condition-specific information on novel, coding and non-coding transcripts, untranslated regions and gene structures (Wilhelm et al., 2008). Accordingly, exploring the condition-specific transcriptome of *F. cylindrus* using RNA-Seq, we aimed to obtain novel insights into the functional elements of the *F. cylindrus* genome including the improvement of the existing genome annotation, the provision of information on genes with unknown annotation, the identification of novel transcripts as well as the identification of genes and pathways involved in acclimation to polar conditions.

In polar waters *F. cylindrus* is exposed to extreme environmental conditions including low temperatures, changing salinities (when sea ice forms and melts) and strong seasonality in solar irradiance (Thomas and Dieckmann, 2002). During polar winter *F. cylindrus* has to endure up to six months of darkness when light is insufficient for photosynthesis. Previous gene expression studies with *F. cylindrus* have been performed using macroarrays (Mock and Valentin, 2004), sequencing of expressed sequence tag (EST) libraries (Mock et al., 2005; Krell et al., 2008) and RT-qPCR assays (Krell et al., 2007; Bayer-Giraldi et al., 2010; Lyon et al., 2011). However, these approaches are limited by their throughput. Furthermore, while the hybridisation-based macroarray approaches is limited by its dynamic range owing to saturation of signals, sequencing-based EST approaches have qualitative and quantitative limitations imposed by bacterial cloning constraints that affect the representation and completeness of cloned sequences. Contrary to EST sequencing, the RNA-Seq approach avoids the need for bacterial cloning of the cDNA input providing a simple and comprehensive way to provide digital gene expression levels (Wang et al., 2009).

In this chapter, I report the first genome-wide expression study using RNA-Seq to investigate the transcriptome of *F. cylindrus* under polar summer growth conditions (nutrient replete, +4 °C, 35 µmol photons $m^{-2}$ $s^{-1}$), freezing temperatures (−2 °C), elevated temperatures (+10 °C), elevated carbon dioxide (1000 ppm $CO_2$), iron starvation (−Fe) and prolonged darkness (one week darkness). I present an overview of the *F. cylindrus* transcriptome providing further understanding of molecular mechanisms of acclimatisation to environmental stresses and highlight selected metabolic pathways and genes involved in acclimation to prolonged darkness.

## 4.2 Results

### 4.2.1 *F. cylindrus* growing under environmental stress conditions

*F. cylindrus* was grown under polar summer growth conditions (nutrient replete, +4 °C, 35 µmol photons $m^{-2}$ $s^{-1}$ continuous light; also referred to as optimal/control/reference growth condition in the following), freezing temperatures (−2 °C), elevated temperatures (+11 °C), elevated carbon dioxide (1000 ppm $CO_2$), iron starvation and prolonged darkness (7 d darkness) to provide total RNA for RNA-Seq library construction and sequencing (Figure 21). Analysis of general growth statistics of *F. cylindrus* grown under the six different growth conditions are shown in Table 8. *F. cylindrus* showed positive growth under all experimental treatments, except for experimental treatments with elevated temperatures (Heat shock, +11 °C) and prolonged darkness (0 µmol photons $m^{-2}$ $s^{-1}$ for 7 days). While cell cultures of *F. cylindrus* treated with polar summer conditions showed optimal growth rates of $\mu = 0.56 \pm 0.01$, cells treated with prolonged darkness showed minimal growth rates of $\mu = 0.01 \pm 4.82e^{-3}$ and cell treated with elevated temperatures showed no growth but formation of cell aggregates leading to a reduction in cell numbers (Figure 21). Additionally, high optimal growth rates of $0.54 \pm 3.62e^{-3}$ were observed in cultures bubbled with elevated $CO_2$ (1000 ppm $CO_2$/Air mixture; Table 8) and significant increase from pH 7.8 to 8.2 could be observed when bubbling was stopped and switched to ambient air (Supplementary Figure S5). Finally, the nutrient, light and temperature limited growth of *F. cylindrus* was verified by increasing cell numbers and photosynthetic quantum yield for PS II ($F_v/F_m$) after add-back of the limiting growth factor to restore optimal growth conditions as found under polar summer conditions (Figure 21).

**Table 8. General growth statistics of *F. cylindrus* under environmental stress conditions. Growth rate μ [cell divisions $d^{-1}$] and photosynthetic quantum yield for PS II $F_v/F_m$ are given at time point of sampling for total RNA extraction.**

| Experimental treatment | Growth conditions | Growth | Growth rate μ [$d^{-1}$] at time of harvest | $F_v/F_m$ at time of harvest |
|---|---|---|---|---|
| Polar summer/ control/reference (Ctrl) | nutrient replete, +4 °C, 35 μmol photons $m^{-2}$ $s^{-1}$ continuous light | + | 0.56 ± 0.01 | 0.58 ± 0.01 |
| Cold shock (Cold) | nutrient replete, −2 °C, 35 μmol photons $m^{-2}$ $s^{-1}$ continuous light | + | 0.23 ± 3.10e$^{-3}$ | 0.42 ± 0.02 |
| Heat shock (Heat) | nutrient replete, +10 °C, 35 μmol photons $m^{-2}$ $s^{-1}$ continuous light | - | n.a. ("negative" μ -0.46 ± 0.06) | 0.41 ± 0.01 |
| Iron (Fe) starvation | −Fe, +4 °C, 35 μmol photons $m^{-2}$ $s^{-1}$ continuous light | + | 0.21 ± 0.07 | 0.29 ± 0.04 |
| High $CO_2$ (CO2) | 1000 ppm $CO_2$/Air, nutrient replete, +4 °C, 35 μmol photons $m^{-2}$ $s^{-1}$ continuous light | + | 0.54 ± 3.62e$^{-3}$ | 0.55 ± 0.01 |
| Prolonged darkness (Dark) | nutrient replete, +4 °C, 0 μmol photons $m^{-2}$ $s^{-1}$ for 7 days | - | 0.01 ± 4.82e$^{-3}$ | 0.53 ± 0.01 |

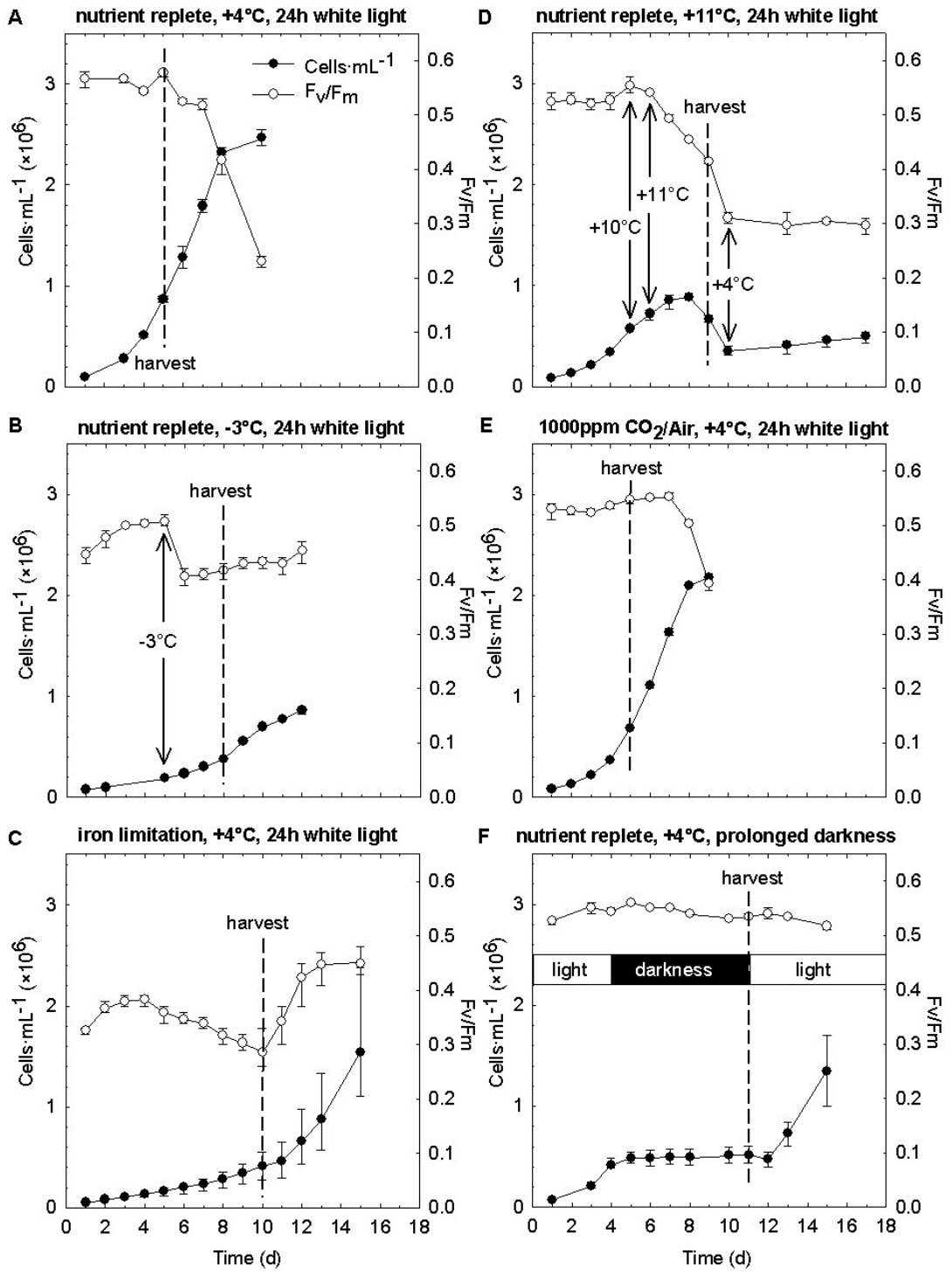



**Figure 21. Cell density and maximum PSII photochemical efficiency ($F_v/F_m$) of *F. cylindrus* grown under (A) optimal conditions (+4 °C, nutrient replete, 35 µmol photons $m^{-2}$ $s^{-1}$), (B) freezing temperatures (−2 °C, nutrient replete, 35 µmol photons $m^{-2}$ $s^{-1}$), (C) iron limitation (−Fe, +4 °C, 35 µmol photons $m^{-2}$ $s^{-1}$), (D) elevated temperatures (+11 °C, nutrient replete, 35 µmol photons $m^{-2}$ $s^{-1}$), (E) elevated carbon dioxide (1000 ppm $CO_2$/Air, +4 °C, nutrient replete, 35 µmol photons $m^{-2}$ $s^{-1}$), and (F) prolonged darkness (1 week darkness, +4 °C, nutrient replete). Shown are median values (n = 3) with maximum and minium values for each time point. Dashed line indicates timpe point of harvest and subsequent add-back of limiting growth factor.**

### 4.2.2 Mapping of sequence reads

The sequencing by synthesis approach using the Illumina HiSeq 2000 platform produced 2.4 to 4.9 million paired-end sequence reads of 101 bases length per sample totalling to 68.8 million reads (Table 9). More than 94% of total reads could be aligned to the *F. cylindrus* draft genome, of which approximately 70% could be uniquely mapped and used for digital gene expression analysis (Table 9; Figure 22). Genome-wide RNA-Seq coverage of the *F. cylindrus* genome was visualised using the Integrative Genomics Viewer (IGV) and an image of a 2.5-kb genomic region is shown in Figure 23. The number of unique fragments mapping to genomic features (i.e. exonic, intronic and intergenic regions of the genome) were counted, giving 1.5 – 3.0 million digital read counts per sample, of which approximately 15 – 30% (21.1% overall mean) did not map to predicted protein coding gene models, indicating novel transcriptionally active genomic regions (TARs) (Table 9; Figure 24). Notably, the percentage of reads that mapped to non-coding features of the genome (i.e., intergenic and intronic regions) was lowest (15%) under optimal growth conditions but twice as high under prolonged darkness (30%) and indicated a high percentage of novel transcriptional activity (e.g., unpredicted genes, alternative splice variants, non-coding RNAs; Table 9; Figure 24).

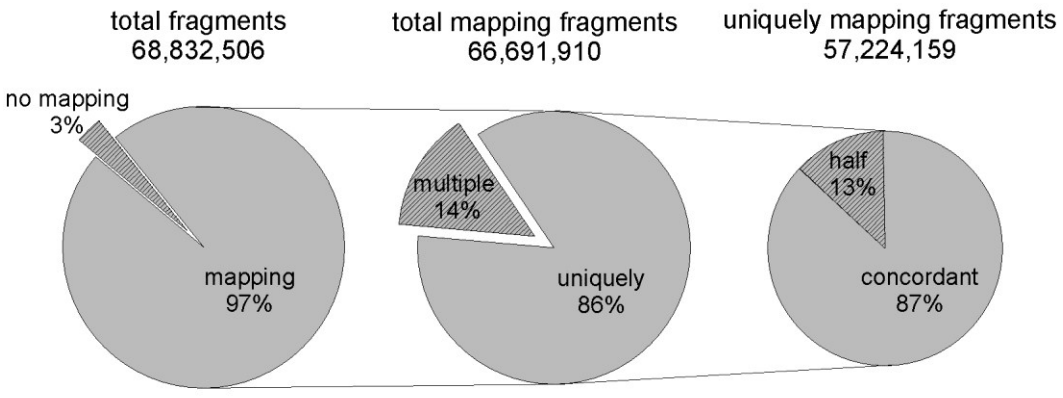



**Figure 22. Overview of the alignment statistics of the fragments mapping onto the *F. cylindrus* reference genome. The reads were mapped to the *F. cylindrus* draft genome assembly (Fracy1). Shading shows the type of mapping; uniquely, reads mapping to only one location in the reference; multiple, reads mapping to more than one location in the reference; concordant, uniquely mapping fragments with both reads mapping in pair; half, uniquely mapping fragments with only one read from the fragment mapping uniquely while other remains unmapped. Only fragments with unique mapping and mapping of both reads properly in pair (concordant) were used for digital transcriptome analysis.**

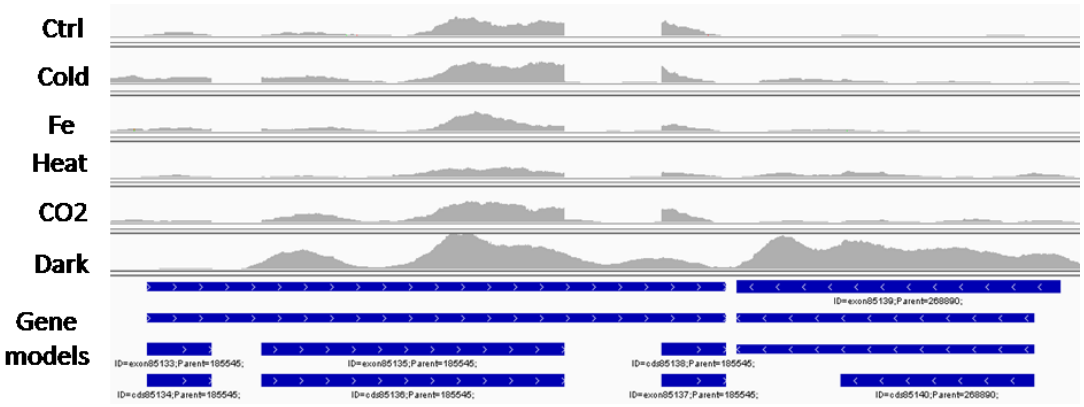



**Figure 23. RNA-Seq coverage for a 2.5-kb region of the *F. cylindrus* genome as displayed by the Integrative Genomics Viewer (IGV v2.1.24). Individual RNA-Seq coverage tracks are shown for experimental treatments on top of the gene models track predicted by the Joint Genome Institute (JGI).**

**Table 9. General statistics for *F. cylindrus* RNA-Seq data. Each experimental treatment was carried out in triplicates (Ctrl = +4 °C; Cold = −2 °C; Iron = Fe; Heat = +10 °C; CO2 = 1000 ppm $CO_2$; Dark = 1 week darkness).**

| Experimental treatment | Total number of reads | % reads mapping | % reads unique mapping | Total digital counts | % counts not mapping to gene model |
|---|---|---|---|---|---|
| Ctrl_1 | 2,481,346 | 97.21 | 71.47 | 1,535,032 | 15.53 |
| Ctrl_2 | 3,413,265 | 97.51 | 72.03 | 2,117,660 | 16.10 |
| Ctrl_3 | 2,969,725 | 97.74 | 72.79 | 1,856,690 | 16.42 |
| Cold_1 | 4,639,725 | 96.07 | 71.41 | 2,804,833 | 18.12 |
| Cold_2 | 3,674,305 | 97.06 | 72.29 | 2,255,951 | 17.74 |
| Cold_3 | 3,932,382 | 96.92 | 71.99 | 2,375,114 | 19.18 |
| Iron_1 | 3,224,916 | 97.95 | 74.68 | 2,011,622 | 19.73 |
| Iron_2 | 3,070,321 | 97.94 | 75.15 | 1,920,484 | 20.14 |
| Iron_3 | 2,961,818 | 97.63 | 74.48 | 1,864,752 | 18.29 |
| Heat_1 | 4,429,184 | 97.09 | 73.66 | 2,644,424 | 23.37 |
| Heat_2 | 4,126,604 | 95.90 | 71.49 | 2,393,030 | 23.27 |
| Heat_3 | 4,497,930 | 94.94 | 69.35 | 2,519,974 | 23.78 |
| CO2_1 | 4,604,350 | 97.50 | 71.80 | 2,823,435 | 17.09 |
| CO2_2 | 3,266,562 | 96.57 | 69.97 | 1,937,864 | 17.94 |
| CO2_3 | 4,820,992 | 97.52 | 72.04 | 2,942,313 | 18.03 |
| Dark_1 | 3,471,467 | 95.89 | 73.45 | 1,961,324 | 30.01 |
| Dark_2 | 4,472,910 | 97.05 | 74.27 | 2,532,937 | 31.15 |
| Dark_3 | 4,774,704 | 96.61 | 73.81 | 2,717,632 | 29.67 |
| **Total** | **68,832,506** | **96.89** | **72.51** | **41,215,071** | **21.09** |

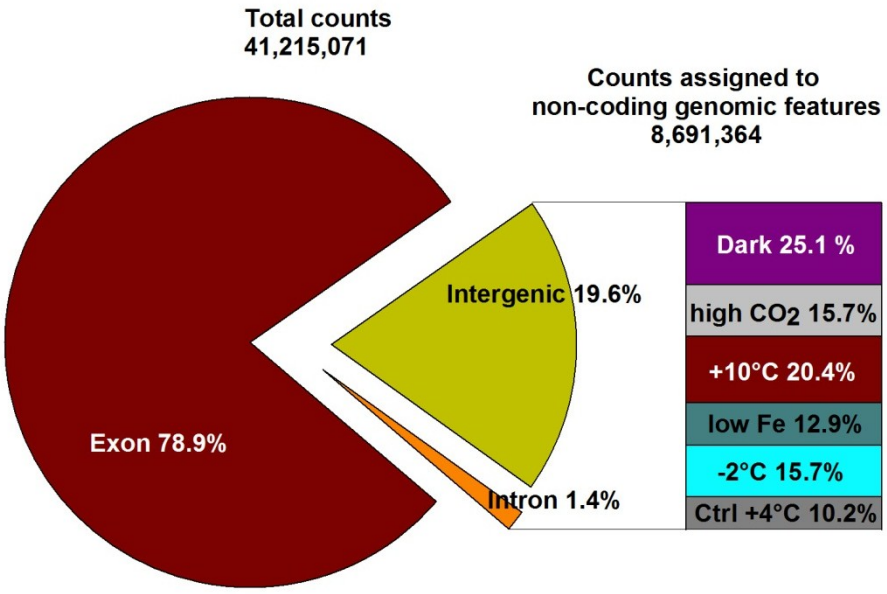



**Figure 24. Pie chart showing the distribution of fragment counts onto the *F. cylindrus* reference genome (Fracy1) in percentage of fragments that map to genomic features including predicted gene models (Exon), intronic regions (Intron) and unannotated regions (Intergenic). The stacked bar shows the percentage of counts from each experimental condition that map to non-coding genomic features (i.e., intergenic and intronic regions).**

The *F. cylindrus* draft genome contained 27,137 predicted gene models including gene copy pairs from highly heterozygous regions of the genome (FilteredModels1 gene model set; see Chapter 3). A total of 25,700 (95%) from the predicted gene models was found transcriptionally active by the definition of having a sum of greater than 0 digital read counts in one or more samples across all experimental treatments (not shown). Additionally, enabled by the unique mapping of fragments to heterozygous regions of the *F. cylindrus* genome, highly heterozygous gene copy pairs, which could not be collapsed into a single haplotype, could be analysed individually. The transcriptome analysis of highly heterozygous gene copy pairs, representing 7966 putative heterozygous alleles (3.2.1), showed that 7832 (> 98%) were transcriptionally active showing unequal expression between putative alleles (4.2.6). In comparison, the digital gene expression analysis based on 18,073 *F. cylindrus* gene models (FilteredModels2 gene model set), which were predicted for its single-haplotype after filtering highly heterozygous gene copy pairs (3.2.1) showed that 17,054 (94%) of the single-haplotype filtered models were transcriptionally active by the above definition (data not shown).

Furthermore, RNA-seq analysis identified a high number of 22,871 genomic regions that were transcriptionally active but located outside of predicted gene models and aligned to unannotated genomic regions and contributed to 19.6% of digital read

counts mapping to intergenic regions (Figure 24). A histogram of the length of transcriptionally active regions (TARs) mapping to unannotated regions using raw unfiltered data showed a length range from 200 bp up to 10 kb but highest frequencies of TAR lengths were found in the 0.6 – 1 kb size range (Figure 25).

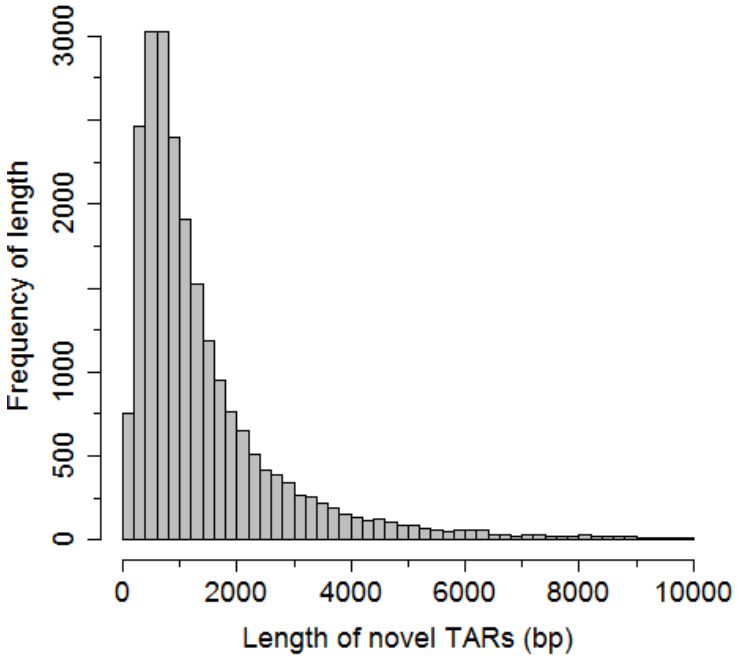



**Figure 25. Histogram showing the length frequencies of novel transcriptional active regions (TARs) with reads mapping to unannotated genomic regions in *F. cylindrus*.**

TARs in proximity of < 250 nucleotides were filtered as they were likely to be 5' or 3' regions of incompletely predicted existing gene models based on experience gained from manual genome annotations. The 57 identified novel TARs were distributed over 36 genomic scaffolds (Figure 26). The length distribution showed highest frequencies of TARs with up to 500 bp length and most TARs were distributed over longer genomic scaffolds (with lower scaffold numbers; Figure 26).

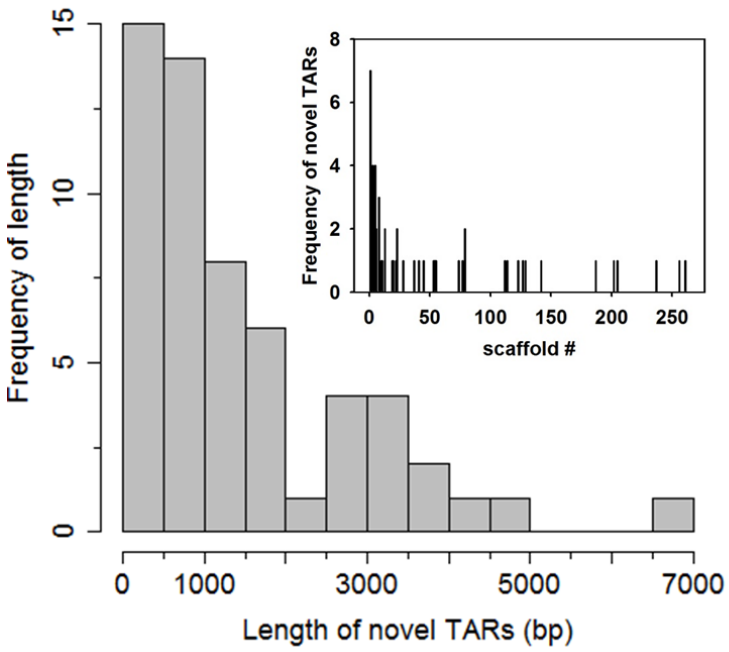



**Figure 26. Histogram showing the length frequencies of filtered novel transcriptional active regions (TARs) with reads mapping to unannotated genomic regions in *F. cylindrus*. TARs were filtered by proximity to annotated gene models (< 250 nt cut off). Insert shows the distribution of novel TARs over all 271 genomic scaffolds.**

To identify undetected protein-coding genes with known sequence homologies, we extracted DNA nucleotide sequences for all 57 novel TARs from the genome and searched for open reading frames (ORFs ≥ 100 nt) on forward and reverse strands. A total of 96 hits were obtained from a sequence similarity search of all 1263 identified ORFs (average protein length of 61) against the Swiss-Prot database (BLASTP, *E*-value ≤ $1e^{-3}$). By using the best BLAST hit 10 ORFs (0.8%) were identified with homologies to deposited reference proteins including two putative transposable elements scoring with low e-values (Table 10).

**Table 10. Novel transcriptionally active genomic regions (TARs) in *F. cylindrus* with hits to Swiss-Prot database (Bairoch et al., 2005). Table is sorted according to e-values. Asterisks (*) mark open reading frames (ORFs), which produced hits to different reference proteins.**

| *F. cylindrus* (Fracy1) genomic coordinates | Best hit | | | | | |
|---|---|---|---|---|---|---|
| | protein | organism | accession# | e-value | query coverage | % identity |
| scaffold_20:1042614-1049163 | Retrovirus-related Pol polyprotein from transposon 17.6 | *Drosophila melanogaster* | P04323 | $2.00E^{-66}$ | 36.03 | 32.41 |
| scaffold_261:1-3337 | Retrovirus-related Pol polyprotein from transposon TNT 1-94 | *Nicotiana tabacum* | P10978 | $1.00E^{-57}$ | 78.04 | 28.94 |
| scaffold_237:1-1970* | Protease Do-like 4, mitochondrial | *Arabidopsis thaliana* | Q9SHZ0 | $4.00E^{-38}$ | 72.67 | 34.55 |
| scaffold_23:481834-482773 | Nucleoporin SEH1 | *Bos taurus* | A7YY75 | $1.00E^{-28}$ | 85.23 | 45.04 |
| scaffold_187:5095-6655 | 5-hydroxyisourate hydrolase 2 | *Rhizobium meliloti* (strain 1021) | Q92UG5 | $2.00E^{-18}$ | 95.89 | 58.57 |
| scaffold_2:3647386-3650927 | 15,16-dihydrobiliverdin:ferredoxin oxidoreductase | *Nostoc punctiforme* (ATCC29133/PCC73102) | Q93TL6 | $4.00E^{-15}$ | 47.54 | 26.46 |
| scaffold_28:1-2671 | Ubiquitin-like protein | *Autographa californica* nuclear polyhedrosis virus | P16709 | $1.00E^{-09}$ | 18.00 | 49.33 |
| scaffold_237:1-1970* | Putative protease Do-like 3, mitochondrial | *Arabidopsis thaliana* | Q9SHZ1 | $1.00E^{-07}$ | 82.54 | 49.06 |
| scaffold_114:1-2693 | Glucosidase 2 subunit beta | *Mus musculus* | O08795 | $2.00E^{-07}$ | 12.48 | 38.67 |
| scaffold_261:1-3337 | Protein LSM14 homolog A | *Mus musculus* | Q8K2F8 | $4.00E^{-05}$ | 90.91 | 70.00 |
| scaffold_5:2016299-2020850 | Autolysin | *Chlamydomonas reinhardtii* | P31178 | $1.00E^{-04}$ | 25.26 | 26.19 |
| scaffold_79:217869-220117 | La protein homolog | *Drosophila melanogaster* | P40796 | $4.00E^{-04}$ | 29.79 | 28.47 |

### 4.2.3 Global analysis of gene expression profiles

A global analysis of transcriptional profiles of all RNA-seq libraries was performed on raw digital read counts. A multi-dimensional scaling (MDS) plot was used to visually explore relationships between transcriptional profiles of samples (Figure 27). Distances on the plot can be interpreted as leading $\log_2$-fold-changes between samples for a common set of genes which was likely to distinguish RNA-Seq libraries. Based on the identification of 4952 genes with substantial high $\log_2$-fold changes of $\geq 5$ in at least one growth condition, a common set of 5000 genes was used to distinguish samples. The MDS plot shows that replicated samples were highly similar and that replicated condition-specific samples clearly separated from each other in both dimensions reflecting the experimental design (Figure 27). Notably, dimension 1 clearly separated samples from the prolonged darkness treatment from all others and indicated the presence of a high number of differentially expressed genes during this experimental treatment in comparison to others (Figure 27). The greatest distances were observed between the experimental treatment with prolonged darkness and polar summer growth conditions with continuous light (Figure 27). The results from the MDS analysis could be confirmed by an independent global hierarchical clustering analysis of transcriptional profiles (not shown).

During explorative data analysis, it was noticed that the sample relationships obtained for RNA-Seq libraries filtered for genes predicted for the *F. cylindrus* single haplotype were inconsistent with the above described relationships obtained for all predicted genes in *F. cylindrus* including putative allelic gene copy pairs from heterozygous regions of the genome. To explore whether specific expression of putative heterozygous allelic copies played a role in the *F. cylindrus* transcriptome, the digital read count data was filtered for the 18,073 gene models as predicted for its single-haplotype and compared to the unfiltered gene set based on 27,137 gene models including gene copy pairs from heterozygous regions of the genome (Figure 27). Assuming that the transcriptome of *F. cylindrus* is affected by differential allele-specific expression, we hypothesized that the MDS relationship between samples would change after filtering digital read count data for single-haplotype transcripts. In comparison, assuming that both heterozygous copies are uniformly expressed under each experimental treatment, we expected no change in the relationship between samples.

Results showed that, although the distinct clustering of samples from the experimental treatment with prolonged darkness remained, a filtering for single-haplotype transcripts strongly affected the relationships of all other samples (Figure 28), concealing the clear separation of replicated samples as observed for non-filtered read count data (Figure 27). Especially the experimental treatments with elevated $CO_2$ and polar summer growth conditions were affected by the filtering process (Figure 28) and indicated a role of allele-specific expression in *F. cylindrus*.

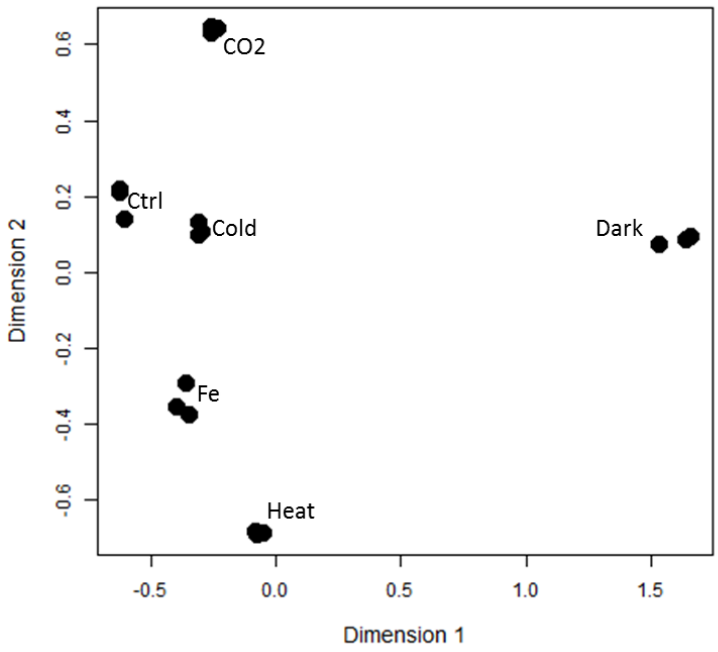



**Figure 27. Multidimensional scaling (MDS) plot of digital gene expression profiles for the *F. cylindrus* RNA-Seq libraries showing the relations between the samples in two dimensions. Distances on the plot represent the biological coefficient of variation of expression between samples using a top set of 5000 genes with highest biological variation and can be interpreted as leading log2-fold changes between the samples. Data was normalised according to TMM scaling normalisation method (Robinson & Oshlack 2010).**

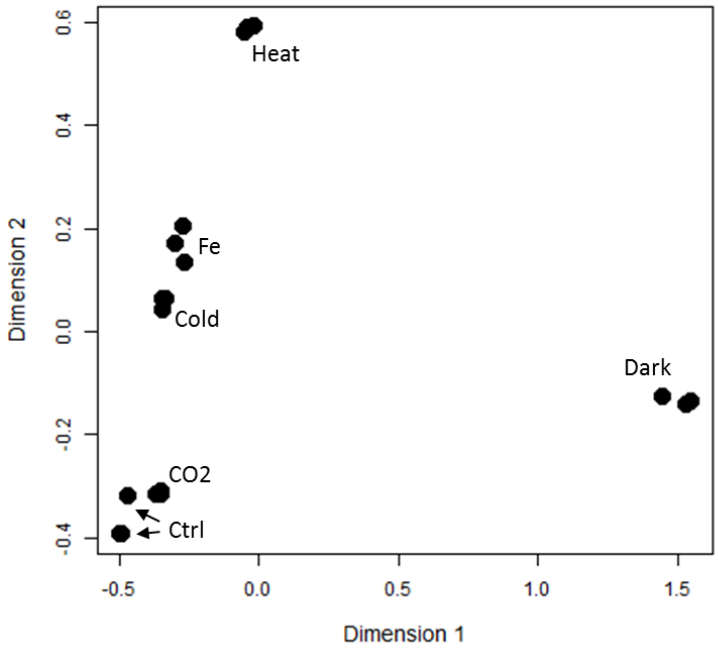



**Figure 28. Multidimensional scaling (MDS) plot of digital gene expression profiles for *F. cylindrus* RNA-Seq libraries filtered for single-haplotype model transcript.**

### 4.2.4 Condition-specific analysis of the *F. cylindrus* transcriptome

A condition-specific analysis of the *F. cylindrus* transcriptome was performed making pair-wise multiple comparisons between experimental conditions testing for differentially expressed genes.

A total of 12,812 genes were differentially expressed in *F. cylindrus* (likelihood ratio test, Benjamini-Hochberg adjusted $P < 0.001$, relative $\log_2$ fold change $\leq -2$ and $\geq +2$) under at least one experimental condition relative to the reference of polar summer growth conditions (nutrient-replete, +4 °C, continuous light). A hierarchical clustering of all 12,812 differentially expressed genes identified condition-specific gene clusters and similarities in genome-wide relative expression between the five experimental stress conditions (Figure 29). Noteworthy was the identification of two main gene clusters, which showed opposite gene expression patterns relative to continuous light (Ctrl) in comparison to other conditions (Figure 29).

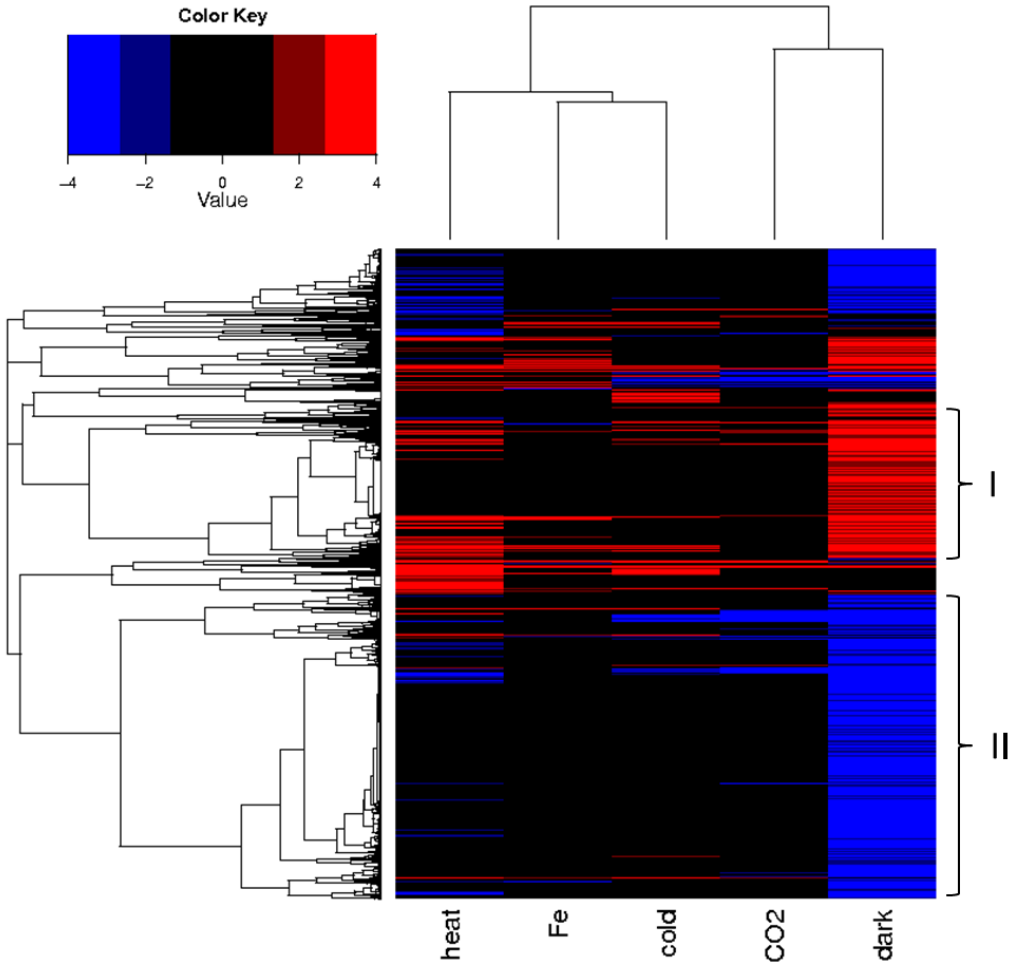



**Figure 29. Hierarchical clustering of 12,812 differentially expressed genes in *F. cylindrus* (likelihood ratio test $P < 0.001$, $\log_2$ fold change $\leq -2$ or $\geq +2$) under iron (Fe)-starvation, freezing temperature (cold, −2 °C), elevated temperature (heat, +10 °C), elevated carbon dioxide ($CO_2$, 1000 ppm $CO_2$/Air) and prolonged darkness (Dark, 1 week darkness) relative to optimal polar summer growth conditions (+4 °C, 35 µmol photons $m^{-2}$ $s^{-1}$ continuous light, nutrient-replete). Each experimental treatment corresponds to one separate column and each single-haplotype gene to a single row. The colour scale ranges from saturated red for up-regulated genes to saturated blue for down regulated genes; black indicates no significant regulation. A one minus Pearson correlation distance metric was applied to cluster rows and columns using complete linking method (created using R software (R Development Core Team, 2012), *heatmap.2* function of gplots package, http://www.r-project.org/).**

As indicated by the global expression analysis (Figure 27), the experimental treatment of *F. cylindrus* with prolonged darkness showed a high number of 7852 up- and 11,004 down regulated genes (likelihood ratio test, Benjamini-Hochberg adjusted $P < 0.05$) relative to treatment with polar summer condition (Ctrl; Table 11). A majority of the down regulated genes in prolonged darkness was upregulated or not differentially expressed in other experimental treatments relative to polar summer growth conditions (Figure 29). The significantly up- and down regulated genes under prolonged darkness relative to polar summer growth with continuous light were functionally analysed using gene ontologies (4.2.5). Moreover, to explore not only the relative differences in the

transcriptional responses of *F. cylindrus* during different treatments compared to polar summer reference growth condition (Ctrl) but also the differences in relative transcriptional responses between all experimental treatments, a pair-wise treatment-by-treatment comparison was performed comparing gene expression profiles every experimental treatment with every other treatment (Table 11). Table 11 shows the number of significantly regulated genes between any two experimental treatments (likelihood ratio test, Benjamini-Hochberg adjusted $P < 0.05$). The numbers represent genes with significant increase in gene expression between an experimental condition on the horizontal axis (row) and an experimental condition on the vertical axis (column). As indicated by MDS (Figure 27) and hierarchical clustering (Figure 29), the treatment-by-treatment comparisons showed that the highest numbers of differentially expressed genes were observed for *F. cylindrus* treated with prolonged darkness (Table 11). Numbers in a small range of ca. 7000 - 8000 genes were upregulated in *F. cylindrus* under prolonged darkness relative to all other treatments, whereas approximately 11,000 genes were down regulated relative to all other conditions (Table 11). Additionally, it is shown that the experimental treatments of *F. cylindrus* with the highest number of differentially regulated genes were detected for the treatment with prolonged darkness (Dark) relative to iron starvation (Fe) showing 7478 upregulated genes and 11,884 down regulated genes, making a total of 19,362 differentially regulated genes between both conditions (Table 11). However, all experimental treatments showed high total numbers (~18,000 – 19,000) of differentially expressed genes (Table 11) and indicated a strong metabolic change in *F. cylindrus* in prolonged darkness.

In comparison the lowest number of differentially expressed genes was observed for the experimental treatment of *F. cylindrus* with elevated $CO_2$ ($CO_2$) relative to treatment with freezing temperatures (Cold) showing 3174 upregulated genes and 4240 down regulated genes making a total of 7414 differentially expressed genes, followed by the treatment comparisons of freezing temperature (CO2) and freezing temperatures (Cold) to polar summer growth conditions (Ctrl) with a total of 8882 differentially expressed genes and 9593, respectively (Table 11).

**Table 11. Treatment-by-treatment comparison of differentially expressed genes in *F. cylindrus*. Table shows the total number of differentially expressed genes (likelihood ratio test, Benjamini-Hochberg adjusted $P < 0.05$) between the row treatment (left) and column treatment (bottom). For genes reported in each cell, there is significant upregulation in the row treatment (left) compared to the column treatment (bottom) and vice versa there is significant down regulation in the column treatment (bottom) compared to the row treatment (left).**

| | | | | | | |
|---|---|---|---|---|---|---|
| **Ctrl** | 0 | 4111 | 4573 | 6912 | 4480 | 11004 |
| **Fe** | 4489 | 0 | 5447 | 6440 | 6116 | 11884 |
| **Cold** | 5020 | 5455 | 0 | 6833 | 4240 | 11458 |
| **Heat** | 6871 | 5876 | 7116 | 0 | 8037 | 11942 |
| **$CO_2$** | 4402 | 5326 | 3174 | 6970 | 0 | 10434 |
| **Dark** | 7852 | 7478 | 7627 | 7115 | 7932 | 0 |
| | **Ctrl** | **Fe** | **Cold** | **Heat** | **$CO_2$** | **Dark** |

Greyscale key for number of differentially expressed genes

| | | | | | |
|---|---|---|---|---|---|
| 0 – 2000 | 2001 – 4000 | 4001 – 6000 | 6001 – 8000 | 8001 – 10000 | 10001 – 12000 |

### 4.2.5 Functional analysis using gene ontologies and metabolic pathway maps

A functional analysis using gene ontologies was carried out on differentially expressed gene sets described in Table 11 to get insights into the metabolic pathways and processes involved in the transcriptional response of *F. cylindrus* to different environmental conditions with main focus on the transcriptional response of *F. cylindrus* to prolonged darkness. Furthermore, metabolic pathways involved in the acclimatory response of *F. cylindrus* to prolonged darkness were identified using a metabolic map based on known metabolic reaction in various organisms (Letunic et al., 2008; Yamada et al., 2011).

The sets of differentially up- and down regulated genes in *F. cylindrus* during prolonged darkness were separately analysed for significantly enriched gene ontology (GO) term annotations (Wallenius approximation, Benjamini-Hochberg adjusted $P < 0.05$) and long lists of GO terms were summarised and visualised using semantic similarity scatter plots (ReViGO scatter plot (Supek et al., 2011)), in which a multidimensional scaling procedure was applied to assign coordinates to each term so that more semantically similar GO terms were closer in the plot (Figure 30; Figure 31).

Figure 30 shows a ReViGO scatter plot showing enriched molecular function GO terms associated with upregulated genes in *F. cylindrus* exposed to prolonged darkness. GO terms are represented as bubbles with colours indicating significance levels of the GO term enrichment test (Wallenius approximation, $P < 0.05$) and sizes indicating the frequency of the GO term in the underlying Gene Ontology Annotation database (UniProt-GOA), which implies that smaller bubbles represent more specific GO terms as they are less frequent than general GO terms. It is shown that many genes involved in regulation of gene expression and DNA replication (GO:0003700 sequence-specific DNA binding transcription factor activity, GO:0043565 sequence-specific DNA binding, GO:0004402 histone acetyltransferase activity, GO:0034061 DNA polymerase activity) were enriched among upregulated genes in *F. cylindrus* during prolonged darkness (Figure 30). Moreover, genes involved in signal transduction as represented by various kinases (GO:0016301 kinase activity, GO:0004713 protein tyrosine kinase,GO:0000285 1-phosphatidylinositol-3-phosphate (PIP) 5-kinase activity) as well as transporter genes (GO:0005215 transporter activity) were enriched among upregulated genes during darkness (Figure 30). Genes with transport activity contributed to carbohydrate transport (GO:0008643) and protein transport (GO:0015031) as shown by analysis of enriched biological process GO terms and which also included proteolysis (GO:0006508; Table 12).

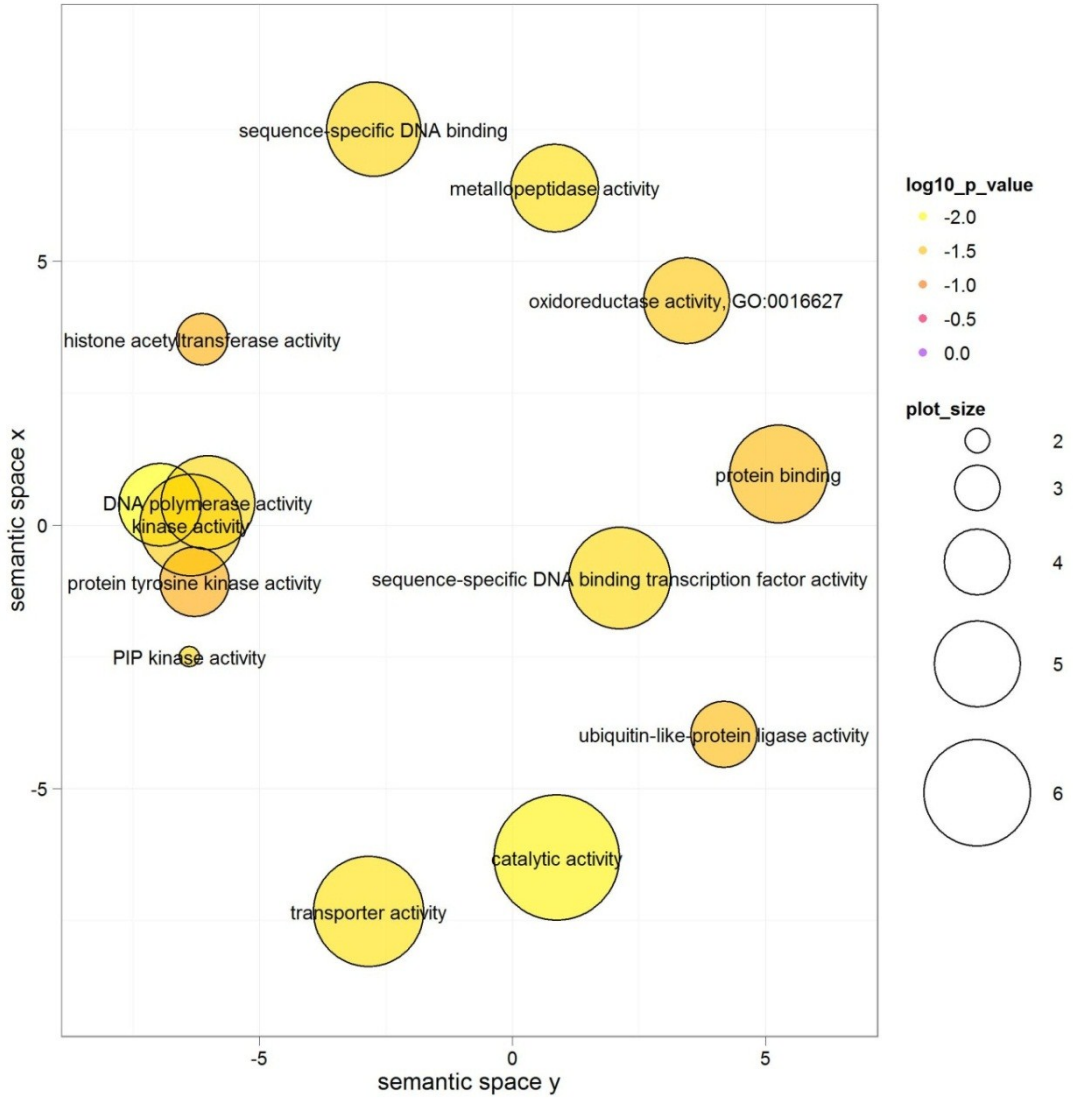



**Figure 30. ReViGO scatterplot (Supek et al., 2011) showing enriched molecular function GO terms of upregulated genes in *F. cylindrus* during prolonged darkness relative to continuous light during polar summer growth condition. Overrepresented GO terms (Wallenius approximation, Benjamini-Hochberg adjusted $P < 0.05$) among upregulated genes (GLM likelihood ratio test, $P < 0.05$) were determined using the goseq Bioconductor R package (Young et al., 2010). A non-redundant GO term set was plotted in a two dimensional space by applying a multidimensional scaling procedure so that more semantically similar GO terms are closer in the plot using the ReViGO Web server (http://revigo.irb.hr/). The bubble colour indicates significance levels and size indicates the frequency of the GO term in the underlying Gene Ontology Annotation (UniProt-GOA) Database.**

**Table 12. Enriched biological process gene ontologies of upregulated genes in *F. cylindrus* during prolonged darkness relative to continuous light under polar summer growth conditions. Given are gene ontology term identifier (GO term ID), term description, frequency of the GO term in the underlying GO Annotation database and *P*-values.**

| GO term ID | Description | Frequency | log10 p-value |
|---|---|---|---|
| GO:0006281 | DNA repair | 1.92% | -1.9309 |
| GO:0008152 | metabolic process | 77.99% | -1.9055 |
| GO:0015031 | protein transport | 1.78% | -1.4753 |
| GO:0006511 | ubiquitin-dependent protein catabolic process | 0.20% | -1.5962 |
| GO:0043687 | post-translational protein modification | 0.01% | -1.4859 |
| GO:0051246 | regulation of protein metabolic process | 0.42% | -1.4859 |
| GO:0008643 | carbohydrate transport | 1.15% | -1.4714 |
| GO:0006396 | RNA processing | 2.59% | -1.362 |
| GO:0006508 | proteolysis | 4.53% | -1.3624 |
| GO:0007264 | small GTPase mediated signal transduction | 0.50% | -1.7327 |
| GO:0007165 | signal transduction | 5.49% | -1.5671 |

In comparison, corresponding to the high number of down regulated genes (Table 11), a higher number of enriched GO terms were enriched among down regulated genes in *F. cylindrus* during prolonged darkness relative to continuous light during polar summer growth condition (Figure 31). A high proportion of genes involved in translation (GO: 0006412; Table 13) was found down regulated as indicated by four enriched molecular function GO terms (GO:0003723 RNA binding, GO:0003735 structural constituent of ribosome, GO:0003743 translation initiation factor activity and GO:0004812 aminoacyl-tRNA ligase activity) (Figure 31). Additionally, enriched GO terms included three proton-pumping ATPase (GO:0003936 F1-ATPase, GO:0046961/GO:0046933 ATP-Synthase activity), cyclophilin (GO:0004600) and sugar transporter (GO:0005351 sugar:hydrogen symporter activity) terms. Last, not least biological process GO terms related to carotenoid biosynthesis and photosynthesis were enriched in the set of down regulated genes in *F. cylindrus* treated with prolonged darkness and included "isoprenoid biosynthetic process" (GO:0008299) and "photosynthesis, light harvesting" (GO:0009765) (Table 13).

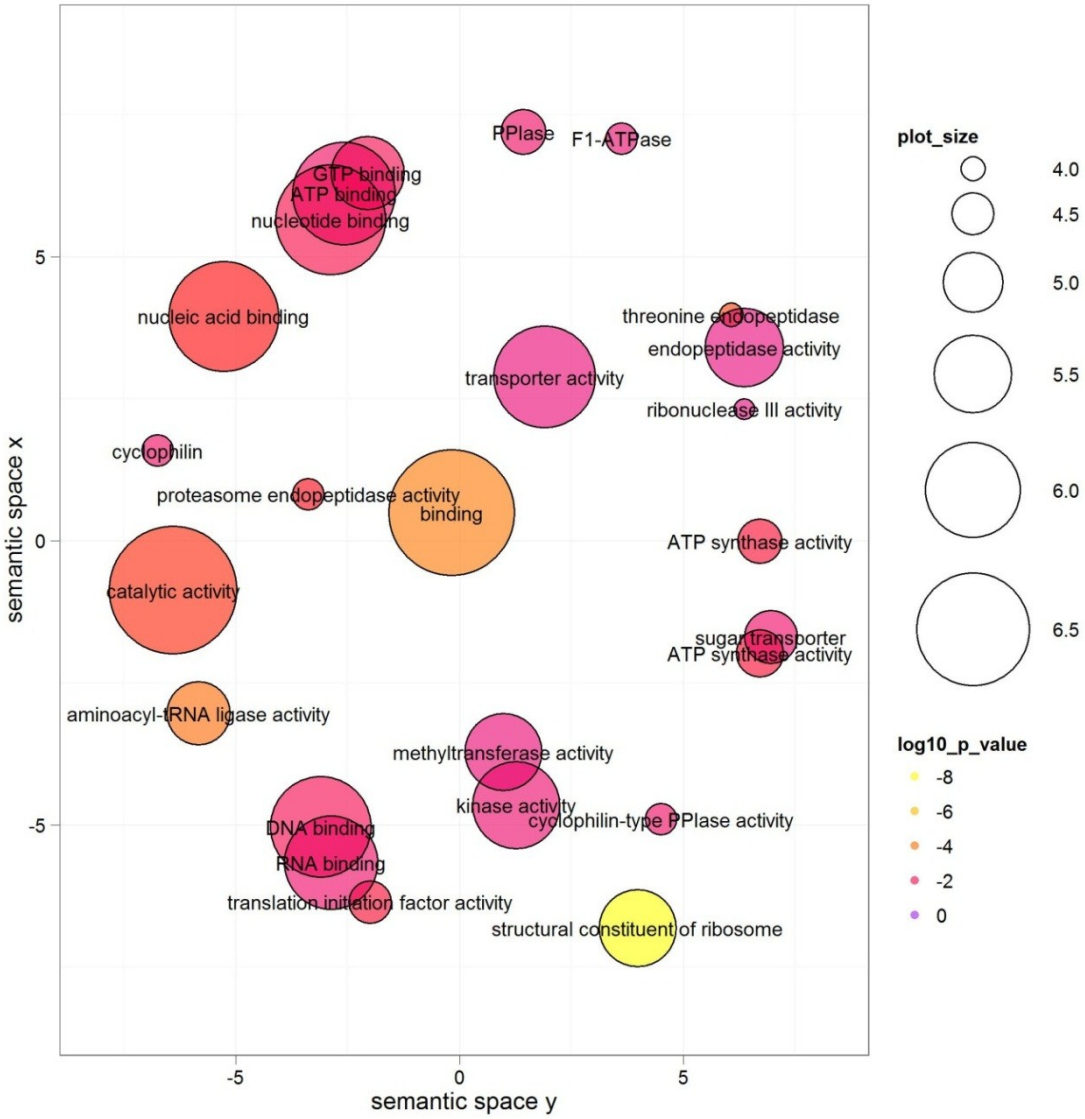



**Figure 31. ReViGO scatterplot (Supek et al., 2011) showing enriched molecular function GO terms of down regulated genes in *F. cylindrus* during prolonged darkness relative to continuous light during polar summer growth condition. Overrepresented GO terms (Wallenius approximation, Benjamini-Hochberg adjusted $P < 0.05$) among upregulated genes (GLM likelihood ratio test, $P < 0.05$) were determined using the goseq Bioconductor R package (Young et al., 2010). A non-redundant GO term set was plotted in a two dimensional space by applying a multidimensional scaling procedure so that more semantically similar GO terms are closer in the plot using the ReViGO Web server (http://revigo.irb.hr/). The bubble colour indicates significance levels and size indicates the frequency of the GO term in the underlying Gene Ontology Annotation (UniProt-GOA) Database. Abbreviation as follows: peptidyl-prolyl cis-trans isomerase (PPIase).**

**Table 13. Enriched biological process gene ontologies of down regulated genes in *F. cylindrus* during prolonged darkness relative to continuous light under polar summer growth conditions. Given are gene ontology term identifier (GO term ID), term description, frequency of the GO term in the underlying GO Annotation database and *P*-values.**

| GO term ID | Description | Frequency | log10 p-value |
|---|---|---|---|
| GO:0006412 | translation | 4.97% | -11.0126 |
| GO:0006885 | regulation of pH | 0.03% | -1.3142 |
| GO:0008152 | metabolic process | 77.99% | -3.773 |
| GO:0015992 | proton transport | 1.03% | -1.6677 |
| GO:0009765 | photosynthesis, light harvesting | 0.01% | -5.5546 |
| GO:0006334 | nucleosome assembly | 0.19% | -2.3075 |
| GO:0008299 | isoprenoid biosynthetic process | 0.39% | -1.6734 |
| GO:0006096 | glycolysis | 0.54% | -1.429 |
| GO:0006810 | transport | 18.62% | -1.3996 |
| GO:0006511 | ubiquitin-dependent protein catabolic process | 0.20% | -1.3985 |
| GO:0006470 | protein dephosphorylation | 0.24% | -1.332 |
| GO:0006260 | DNA replication | 2.88% | -1.9099 |
| GO:0006418 | tRNA aminoacylation for protein translation | 0.97% | -3.5445 |
| GO:0008033 | tRNA processing | 1.17% | -1.4877 |
| GO:0006364 | rRNA processing | 0.62% | -1.6126 |
| GO:0006457 | protein folding | 0.97% | -4.139 |
| GO:0006508 | proteolysis | 4.53% | -2.6896 |
| GO:0006413 | translational initiation | 0.34% | -2.1525 |
| GO:0006414 | translational elongation | 0.67% | -1.5375 |

In addition to GO term analysis, a pathway analysis was performed on differentially regulated genes using metabolic maps to further pinpoint expression levels of different metabolic pathways (Figure 32). Figure 32 shows a metabolic map based on currently known metabolic reactions compiled from various organisms (Letunic et al., 2008; Yamada et al., 2011) and highlights differentially expressed metabolic pathways in *F. cylindrus* under prolonged darkness relative to continuous light during polar summer growth conditions. Nodes on the map correspond to chemical compounds and lines represent series of enzymatic reactions. The width of lines and red colour shading are scaled according to mean absolute expression values (fragments per kilobase of exon per million fragments mapped, FPKM).

It is shown that selected genes encoding for enzymes involved in lipid metabolism (left centre of map, highlighted with dark grey box), nucleotide metabolism (top right of map, highlighted with dark grey box), carbohydrate metabolism (lower

centre of map) including starch and sucrose metabolism (top centre of map, highlighted with dark grey box) showed high expression values for selected pathways (Figure 32). The high expression values of genes involved in starch and sucrose and lipid metabolism in *F. cylindrus* under prolonged darkness indicated utilisation of chrysolaminarin and fatty acid storage products.

In comparison to polar summer growth conditions with continuous light (Ctrl), gene expression analysis of the chrysolaminarin synthesis pathway showed significant down regulation of all identified genes involved in chrysolaminarin synthesis including UDP-sugar pyrophosphorylases (USP1-3, protein IDs: 183667, 211962, 213392) and beta-glucan synthases (BGS1-4, protein IDs: 269043, 186340, 146754, 149475) during prolonged darkness (Dark; Figure 33) with negative relative $\log_2$ fold changes (logFC < −1, $P < 0.001$) for all genes, except BGS3 (logFC = −0.01, $P < 0.001$) and USP2 (logFC = 0.7, $P < 0.001$). Conversely, selected genes involved in the breakdown of chrysolaminarin to free glucose were significantly upregulated under prolonged darkness relative to continuous light under polar summer growth conditions, which included genes for exo-1,3-beta glucanases (EXG2/207213, EXG5/258194; logFC > 1.5 , $P < 0.001$), endo-1,3-beta glucanases (ENG1/206115, ENG3/260039, ENG8/188235, ENG9/241200; logFC > 1.6, $P < 0.001$) and beta-glucosidases (BGL3/182486, BGL5/181839; logFC > 3.4, $P < 0.001$) (Figure 33). Free glucose is a substrate for a significantly upregulated glucokinase (GLK2/216851, logFC = 6.3, $P < 0.001$) and catalyse the first step of glycolysis (Figure 33). Although the upper phase of glycolysis did not appear to be strongly regulated in *F. cylindrus* under prolonged darkness on a global metabolic map (Figure 32), the analysis of the lower ATP and reducing equivalent-producing phase of glycolysis (glyreraldehyde-3-phosphate dehydrogenase to pyruvate kinase) showed significant upregulation of selected genes (Figure 34). A glyceraldehyde-3-phosphate dehydrogenase (GAPDH6/269867) catalysing the reaction glyceraldehyde 3-phosphate (GAP) to 1,3-bisphosphoglycerate (1,3BPG) showed a relative $\log_2$ fold change of 4.1 ($P < 0.001$). Additionally, a phosphoglycerate kinase (PGK3/208673; logFC = 5.5, $P < 0.001$), two phosphoglycerate mutases (PGAM4/185681, PGAM9/161031; logFC > 1.1, $P < 0.001$), an enolase (ENO3/184892, logFC = 1.1, $P < 0.001$) and a pyruvate kinase (PK1/206568; logFC = 1.1, $P < 0.001$) showed upregulation in *F. cylindrus* during prolonged darkness (Dark) relative to control (Figure 34).

The pyruvate decarboxylation to acetyl-CoA is catalysed by pyruvate dehydrogenase enzymes. Although an identified pyruvate dehydrogenase alpha subunit (PDHE1-A/187383) was significantly down regulated during prolonged darkness relative to continuous light (Ctrl) under polar summer growth conditions (logFC = −2.5, $P < 0.001$), a similar E1 component dehydrogenase (DH_E1)-domain containing alpha subunit (BCKDE1-A/186364) with homologies to a 2-oxoisovalerate dehydrogenase (EC 1.2.4.4) showed significant upregulation with a relative $\log_2$ fold change of 9.3 ($P < 0.001$). The final metabolite acetyl-CoA can feed into the tricarboxylic acid (TCA) cycle (Figure 34).

The expression analysis of genes involved in mitochondrial and peroxisomal beta-oxidation of fatty acids in *F. cylindrus* showed upregulation of most genes during prolonged darkness (Dark) relative to continuous light during polar summer growth conditions (Ctrl) (Figure 35). A long chain acyl-CoA synthetase (ACSL1/262994) involved in the initial activation of fatty acid beta-oxidation was significantly upregulated by a relative log2 fold chance of 0.9 ($P < 0.001$). The subsequent peroxisomal oxidation of acyl-CoA is catalysed by an acyl-CoA oxidase (ACOX) and directly uses molecular oxygen generating hydrogen peroxide. A putative peroxisomal ACOX1 (210789) in *F. cylindrus* was significantly upregulated during prolonged darkness by a relative $\log_2$ fold change of 1.5 ($P < 0.001$). In comparison, the FAD-dependent oxidation of acyl-CoA catalysed by acyl-CoA dehydrogenases (ACD1/209571, ACD2/226606, ACD3/268657, ACD4/271673) showed stronger relative $\log_2$ fold changes ranging from 1.9 (for ACD1) to 10.1 (for ACD2). The following hydration reaction is catalysed by enoyl-CoA hydratase (ECH) and selected ECH isoenzymes in *F. cylindrus* showed significant upregulation during prolonged darkness relative to continuous light during polar summer growth conditions. ECH1 (180456) was significantly upregulation by a relative $\log_2$ fold change of 1.1 ($P < 0.001$), whereas ECH3 (273959) and ECH5 (202663) showed higher $\log_2$ fold changes of 4.5 ($P < 0.001$) and 2.7 ($P < 0.001$). All identified isoenzymes for 3-hydroxyacyl-CoA dehydrogenase (HADH) in *F. cylindrus* catalysing the $NAD^+$-dependent oxidation of 3-hydroxyacyl-CoA to 3-ketoacyl-CoA (Figure 35) showed significant upregulation during prolonged darkness relative to continuous light during polar summer growth conditions with $\log_2$ fold changes of 2.0 (for both HADH1/207194 and HADH2/183437) and 4.7 (for HADH3/270026). Notably, the single identified acetyl-CoA acetyltransferase in *F. cylindrus* (ACAT1/274265), which catalyses the ultimate

thiolysis step during beta oxidation of fatty acids was significantly upregulated relative to continuous light during polar summer growth conditions by a $\log_2$ fold change of 5.9 ($P < 0.001$), whereas it showed low FPKM expression values in all other growth conditions (Figure 35). Last, not least oxidative phosphorylation showed high expression values during prolonged darkness (Figure 32; lower centre of map).

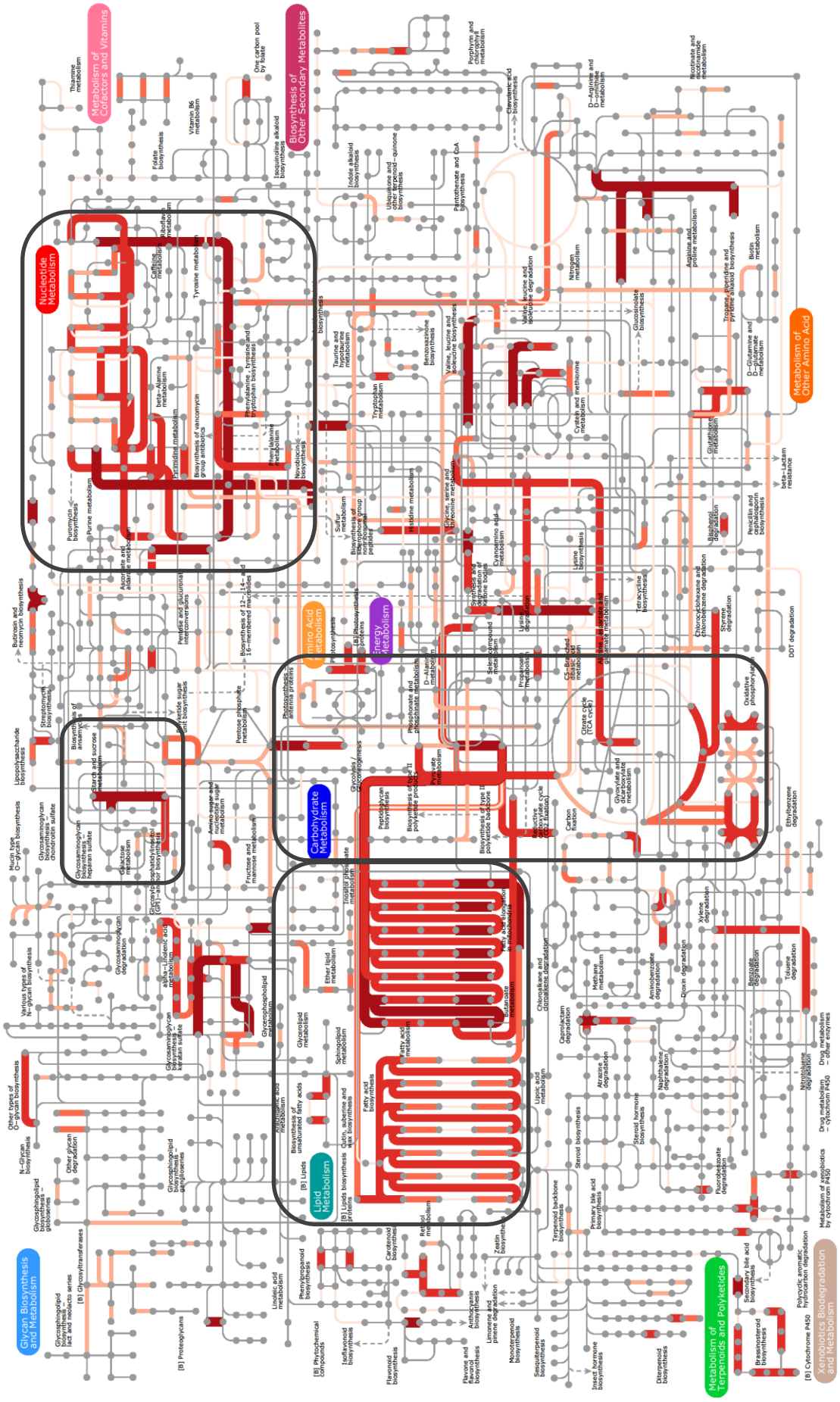



**Figure 32. Metabolic map of differentially expressed genes in *F. cylindrus* grown under prolonged darkness. Nodes on the map correspond to chemical compounds and lines represent series of enzymatic reactions. The width of lines and red colour shading are scaled according to mean absolute expression values (FPKM).**

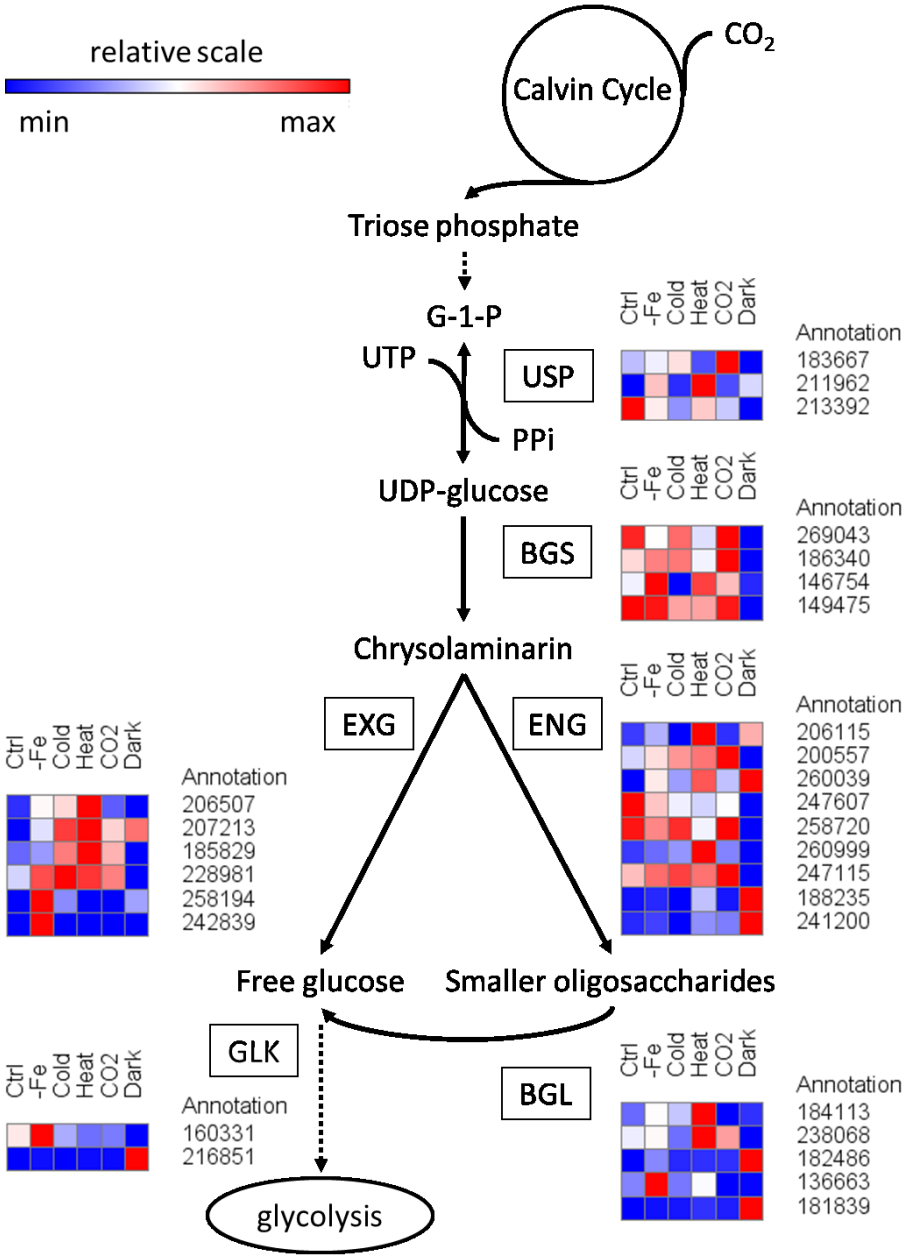



**Figure 33. Expression of genes involved in chrysolaminarin biosynthesis and degradation in *F. cylindrus* under six experimental growth conditions. Colour scale represents absolute FPKM expression values on a relative scale per row (gene). Chemical compound abbreviations: glucose-1-phosphate (G-1-P), Uridine-5'-triphosphate (UTP), pyrophosphate (PPi). Identified isoenzymes are shown in boxes and *F. cylindrus* protein identifiers are reported with annotation labels: UDP-sugar pyrophosphorylase (USP), beta-glucan synthase (BGS), exo-1,3-beta-glucanase (EXG), endo-1,3-beta-glucanase (ENG), beta-glucosidase (BGL), glucokinase (GLK).**

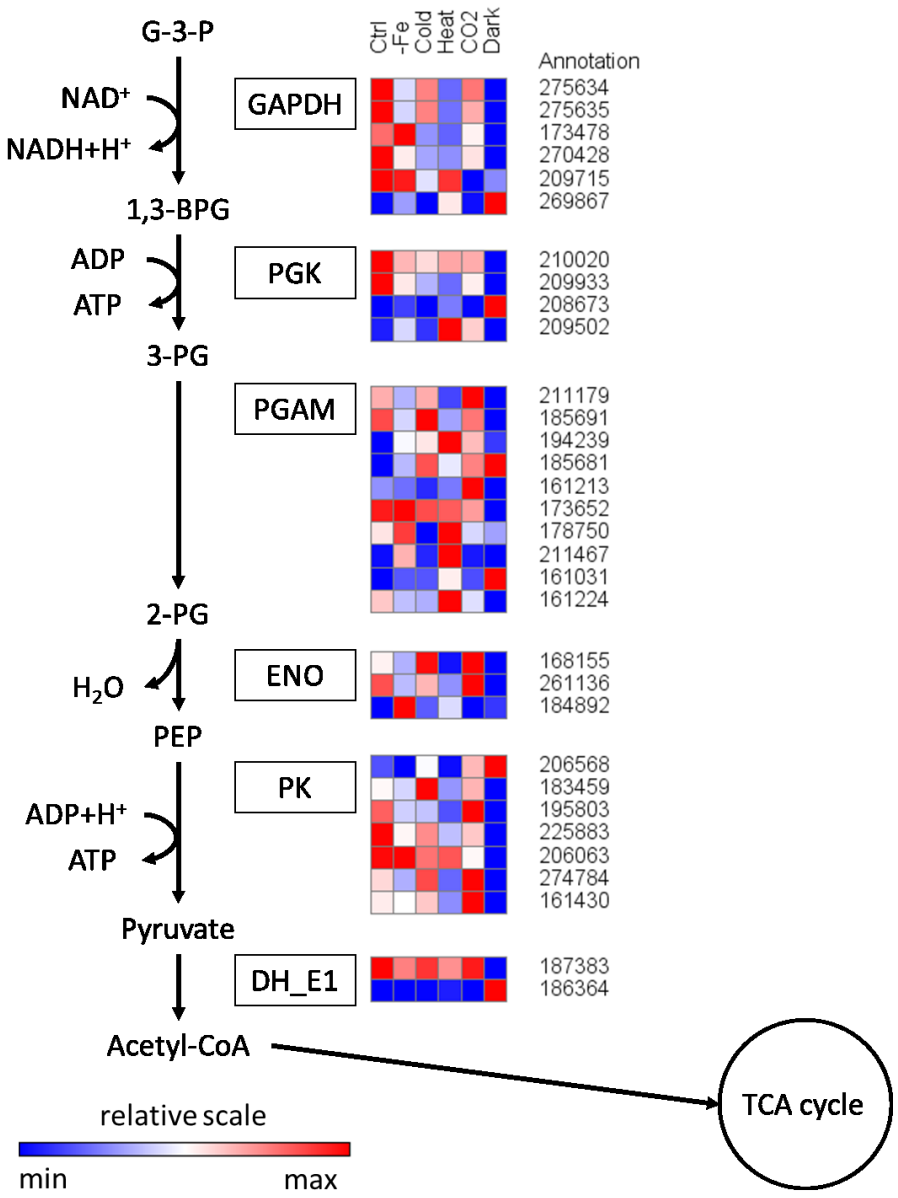



**Figure 34. Expression of genes involved in lower phase of glycolysis in *F. cylindrus* under six experimental growth conditions. Colour scale represents absolute FPKM expression values on a relative scale per row (gene). Chemical compound abbreviations: glyceraldehyde 3-phosphate (G-3-P), 1,3-bisphosphoglycerate (1,3-BPG), 3-phosphoglycerate (3-PG), 2-phosphoglycerate (2-PG), phosphoenolpyruvate (PEP), tricarboxylic acid (TCA), adenosine-5'-triphosphate (ATP), adenosine diphosphate (ADP), coenzyme A (CoA), nicotinamide adenine dinucleotide ($NAD^+$), reduced $NAD^+$ ($NADH+H^+$), tricarboxylic acid cycle (TCA cycle). Identified isoenzymes are shown in boxes and *F. cylindrus* protein identifiers are reported: Glyceraldehyde 3-phosphate dehydrogenase (GAPDH), phosphoglycerate kinase (PGK), phosphoglycerate mutase (PGAM), enolase (ENO), pyruvate kinase (PK), E1 component dehydrogenase (DH_E1).**

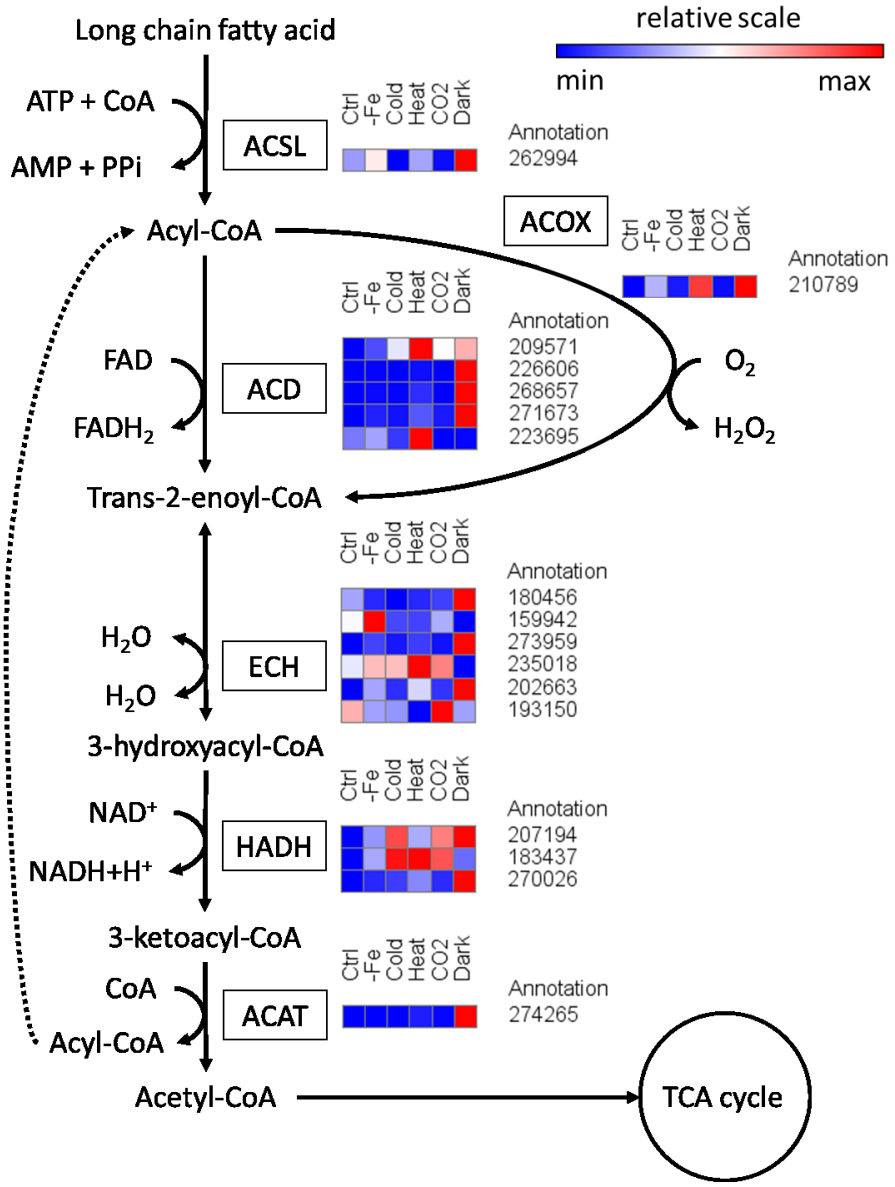



**Figure 35. Expression of genes involved in mitochondrial and peroxisomal fatty acid beta-oxidation in *F. cylindrus* under six experimental growth conditions. Colour scale represents absolute FPKM expression values on a relative scale per row (gene). Chemical compound abbreviations: Adenosine-5'-triphosphate (ATP), Adenosine monophosphate (AMP), pyrophosphate (PPi), coenzyme A (CoA), flavin adenine dinucleotide (FAD), reduced FAD ($FADH_2$), nicotinamide adenine dinucleotide ($NAD^+$), reduced $NAD^+$ ($NADH+H^+$), tricarboxylic acid (TCA). Identified isoenzymes are shown in boxes and *F. cylindrus* protein identifiers are reported with annotation labels: long-chain acyl-CoA synthetase (ACSL), acyl-CoA oxidase (ACOX), acyl-CoA dehydrogenase (ACD), enoyl-CoA hydratase (ECH), 3-hydroxyacyl-CoA dehydrogenase (HADH), acetyl-CoA acetyltransferase (ACAT). The direct oxidation of acyl-CoA using oxygen takes place in peroxisomes and is catalysed by ACOX producing hydrogen peroxide. The FAD-dependent oxidation of acyl-CoA takes place in mitochondria.**

As indicated by the functional analysis of significantly down regulated genes in *F. cylindrus* during prolonged darkness relative to continuous light conditions (polar summer) using biological process gene ontologies, the GO terms "isoprenoid biosynthetic process" (related to carotenoid biosynthesis) and "photosynthesis, light harvesting" showed significant enrichment (Table 13).

Further expression analysis of genes involved in carotenoid biosynthesis and the xanthophyll cycle showed low FPKM expression values relative to continuous light (Ctrl) during polar summer growth conditions (Figure 36) with mostly negative $\log_2$ fold changes (logFC). The single identified putative deoxyxylulose 5-poshpate synthetase in *F. cylindrus* (DXS1/206899), which catalyses the synthesis of deoxyxylulose 5-phosphate (DX-5-P) showed relative $\log_2$ fold change of −5.2 ($P < 0.001$). Genes encoding for putative enzymes involved in the subsequent chain of reactions leading to the generation of isopentenyl and dimethylallyl pyrophosphate (IPP/DMAPP) showed negative relative $\log_2$ fold changes within a similar range from −1.7 for isopentenyl diphosphate:dimethylallyl diphosphate isomerase (IDI2/239201) to −8.0 for 2-C-methylerythritol 2,4-cyclodiphosphate synthase (MCS/205676). Although, a single identified geranylgeranyl pyrophosphate synthase (GGPPS/287526) generating geranylgeranyl pyrophosphate (GGPP) was upregulated in *F. cylindrus* during prolonged darkness by a $\log_2$ fold change of 1.3 ($P < 0.001$), a geranylgeranyl reductase (GGR/179770), ultimately leading to the biosynthesis of phytol and utilised in chlorophyll synthesis (Figure 36) via chlorophyll synthase (see below), was significantly down regulated by a $\log_2$ fold change of −2.5. Similarly to GGR, genes encoding for isoenzymes of phytoene synthase (PSY1-4) were significantly down regulated during prolonged darkness to the same degree ($-4.2 \geq \text{logFC} \leq -2.6$, $P < 0.001$). Additionally, genes encoding enzymes involved in the enzymatic conversion of phytoene to beta-carotene generally showed low FPKM values during prolonged darkness in comparison to continuous light (Figure 36), except for zeta-carotene desaturase (ZDS/291551) and carotenoid isoenzyme 2 (CIS2/206370), which were significantly upregulated during prolonged darkness by a relative $\log_2$ fold change of 0.7 ($P < 0.001$) and 2.1 ($P < 0.001$), respectively. All other genes encoding for enzymes involved in the conversion of phytoene to beta-carotene showed either no significant regulation or negative relative $\log_2$ fold changes. Last, but not least genes encoding for isoenzymes of the zeaxanthin epoxidase (ZEP) and violaxanthin de-epoxidase (VDE) involved in a photoprotective xanthophyll cycle were strongly down regulated in *F. cylindrus* during prolonged darkness relative to continuous light showing negative $\log_2$ fold changes ranging from −3.8 (for VDE2/212709) to −9.1 (for VDE3/207471), with the exception of ZEP1 with no significant regulation. Finally, the single identified beta-carotene monooxygenase (BCMO/228160) in *F. cylindrus*, catalysing the cleavage of beta-carotene into two molecules of retinal (Figure 36) showed significant upregulation during prolonged darkness relative to continuous light by a $\log_2$ fold change of 2.1 ($P <$

0.001). Notably, in the context of retinal synthesis during darkness, a bacteria-like rhodopsin in *F. cylindrus* (FR2/267528), which binds retinal was significantly upregulated during prolonged darkness by a $\log_2$ fold change of 5.4 ($P < 0.001$).

In comparison to gene expression analysis of the carotenoid pathway (Figure 36), the expression analysis of genes involved in chlorophyll synthesis in *F. cylindrus* during prolonged darkness relative to continuous light could not identify a single significantly upregulated gene (Figure 37) and showed negative relative $\log_2$ fold changes ranging from −8.6 for porphobilinogen deaminase (PBGD/267185) to −3.0 for NADPH:protochlorophyllide oxidoreductase (POR2/188173).

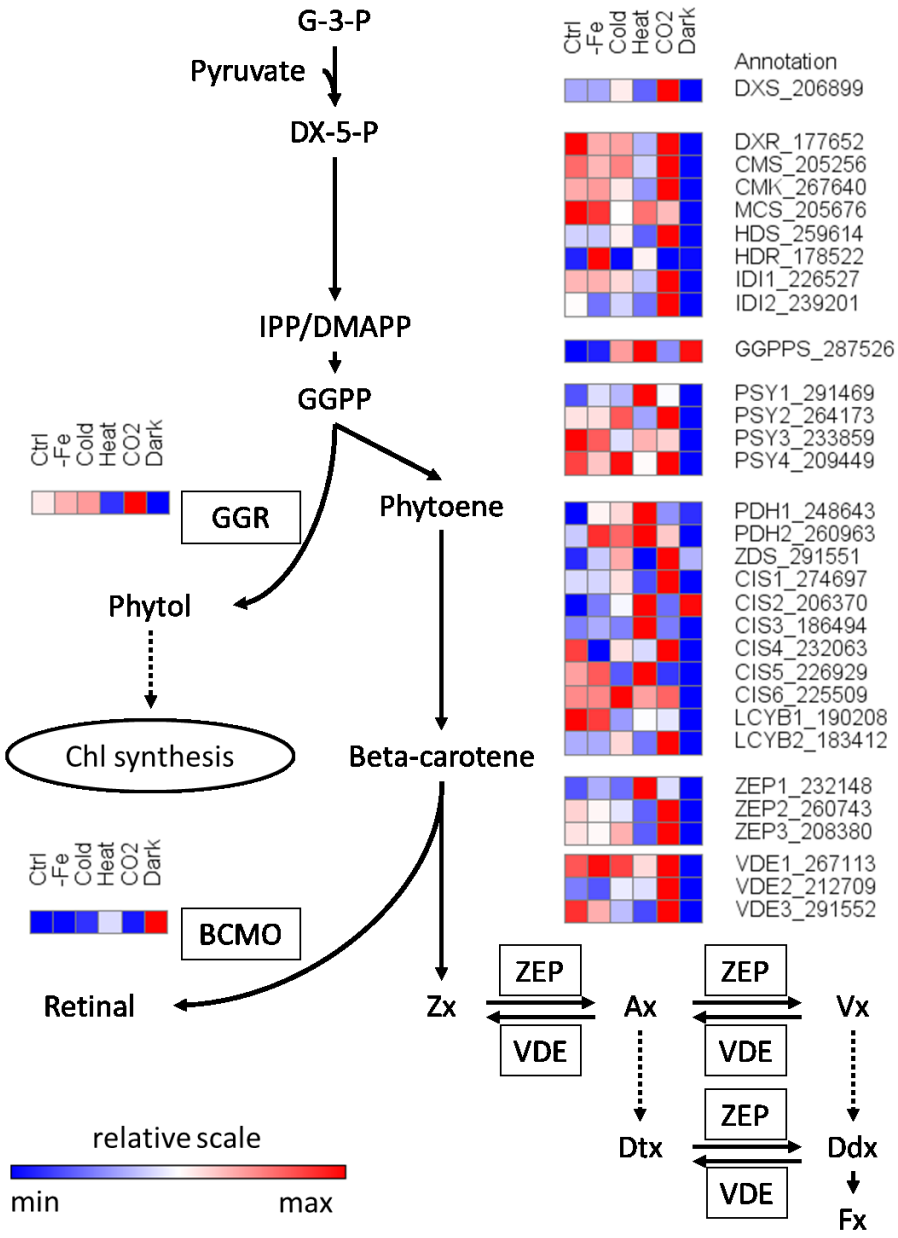



**Figure 36. Expression of genes involved in carotenoid biosynthesis and xanthophyll cycle in *F. cylindrus* under six experimental growth conditions. Colour scale represents absolute FPKM expression values on a relative scale per row (gene). Chemical compound abbreviations: glycerinealdehyde 3-phosphate (G-3-P), deoxyxylulose 5-phosphate (DX-5-P), Isopentenyl pyrophosphate/dimethylallyl pyrophosphate (IPP/DMAPP), geranylgeranyl pyrophosphate (GGPP), zeaxanthin (Zx), antheraxanthin (Ax), violaxanthin (Vx), diatoxanthin (Dtx), diadinoxanthin (Ddx), fucoxanthin (Fx), chlorophyll (Chl). Identified isoenzymes are shown in boxes or annotation labels with *F. cylindrus* protein identifiers: deoxyxylulose 5-phosphate synthase (DXS), deoxyxylulose 5-phosphate reductoisomerase (DXR), diphosphocytidyl-2-C-methylerythritol synthase (CMS), 2-C-methylerythritol kinase (CMK), 2-C-methylerythritol 2,4-cyclodiphosphate synthase (MCS), hydroxy-3-methylbutenyl diphosphate synthase (HDS), hydroxy-3-methylbutenyl diphosphate reductase (HDR), isopentenyl diphosphate:dimethylallyl diphosphate isomerase (IDI), geranylgeranyl pyrophosphate synthase (GGPPS), phytoene synthase (PSY), phytoene dehydrogenase (PDH), zeta-carotene desaturase (ZDS), carotenoid isomerase (CIS), lycopene beta cyclase (LCYB), zeaxanthin epoxidase (ZEP), violaxanthin de-epoxidase (VDE), geranylgeranyl reductase (GGR), beta-carotene monooxygenase (BCMO).**

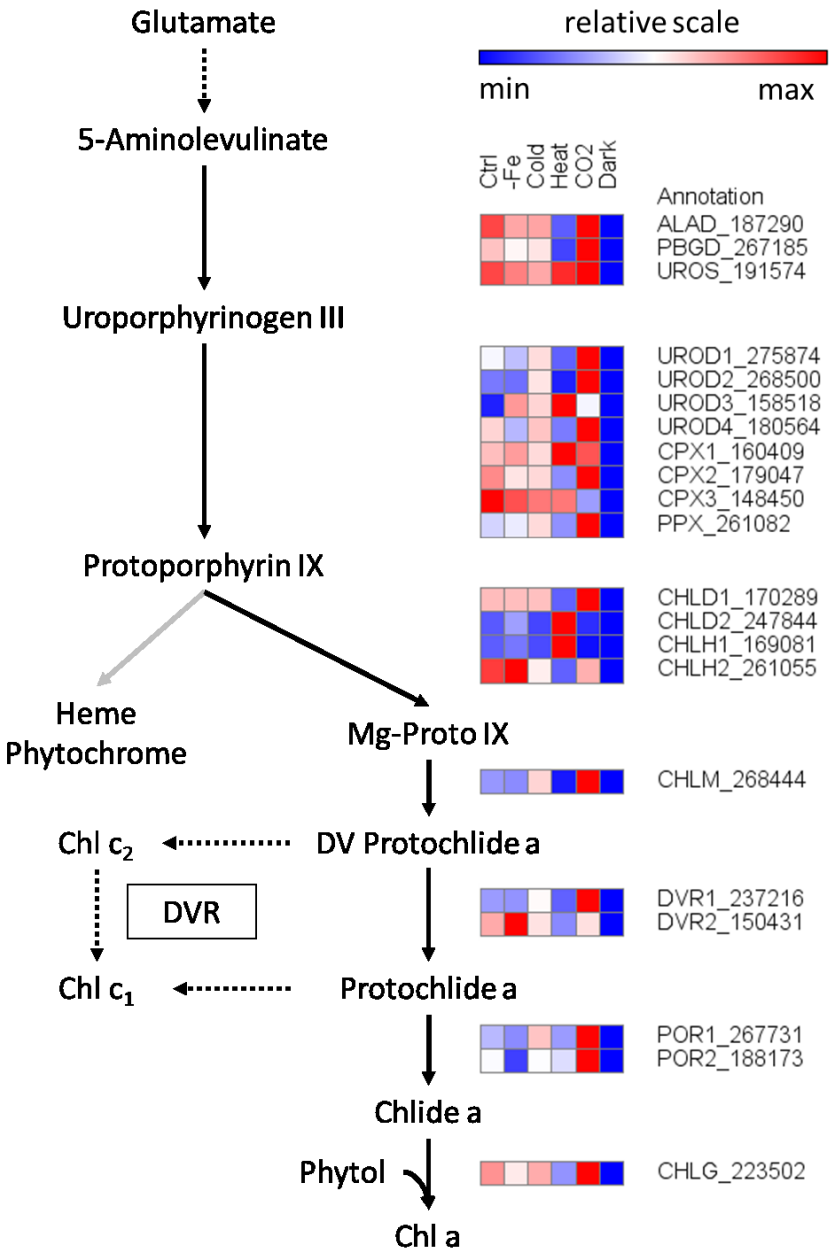



**Figure 37. Expression of genes involved in chlorophyll biosynthesis in *F. cylindrus* under six experimental growth conditions. Colour scale represents absolute FPKM expression values on a relative scale per row (gene). Chemical compound abbreviations: Magnesium-protoporphyrin IX (Mg-Proto IX), divinyl protochlorophyllide a (DV Protochlide a), protochlorophyllide a (Protochlide a), chlorophyllide (Chlide a), chlorophyll (Chl). Identified isoenzymes are shown in boxes or annotation labels with *F. cylindrus* protein identifiers: 5-aminolevulinic acid dehydratase (ALAD), porphobilinogen deaminase (PBGD), uroporphyrinogen III synthase (UROS), uroporphyrinogen III decarboxylase (UROD), coproporphyrinogen III oxidase (CPX), protoporphyrinogen IX oxidase (PPX), protoporphyrin IX Mg-chelatase subunit D (CHLD), protoporphyrin IX Mg-chelatase subunit H (CHLH), Mg-protoporphyrin IX methyltransferase (CHLM), divinyl protochlorophyllide a 8-vinyl reductase (DVR), NADPH:protochlorophyllide oxidoreductase (POR), chlorophyll synthase (CHLG).**

### 4.2.6 Allele-specific analysis of the *F. cylindrus* transcriptome

The draft genome of *F. cylindrus* showed a high degree of sequence polymorphism and prevented haplotypes from heterozygous regions of the genome to be collapsed into a single haplotype causing a diffuse haplotype structure (Chapter 3). These heterozygous regions affected 7966 (30%) of the total 27,137 predicted genes (3.2.1). The analysis of nucleotide similarities of the resulting heterozygous gene copy pairs showed high nucleotide sequence similarities of > 98% for most pairs suggesting the presence of putative heterozygous allelic gene copies (Figure 8).

To explore if putative allelic gene copies from heterozygous parts of the *F. cylindrus* genome showed allele-specific gene expression, a condition-specific analysis of the *F. cylindrus* transcriptome was carried out on the putative 7966 allelic gene copies. It was shown that more than 98% (7832) of the allelic variant genes were transcriptionally active (digital read count > 0 in at least 1 sample) and 5790 allelic genes showed differential expression relative to polar summer growth condition (likelihood ratio test $P < 0.001$, $\log_2$ fold change $\leq -2$ or $\geq +2$) (Figure 38). Notably is a variable bi-allelic expression (unequal expression of alleles) between both allelic copies under different experimental conditions, each represented by two adjacent columns (Figure 38) indicating a high degree of differential bi-allelic expression in *F. cylindrus*. Indeed, if comparing allelic copy 2 against its sister allelic copy 1, 4434 allelic copy pairs (55% of all copy pairs) could be identified showing strong differential bi-allelic expression (likelihood ratio test $P < 0.001$, $\log_2$ fold change $\leq -2$ or $\geq +2$) (Figure 39).

As a result, the individual functional gene ontology analysis of both significantly upregulated allelic gene copy sets in *F. cylindrus* during prolonged darkness relative to continuous light (Ctrl) shows deviant sets of overrepresented molecular function gene ontologies for both allelic copies (Wallenius approximation, Benjamini-Hochberg adjusted $P < 0.05$; Figure 40, Figure 41). Figure 40 shows that upregulated allelic genes from allele set 1 were involved in signal transduction (e.g. GO:0004698 calcium-dependent protein kinase C activity, GO:0004702 receptor signalling protein serine/threonine kinase activity, GO:0019199 transmembrane receptor protein kinase activity), regulation of transcription and translation (e.g. GO:0043565 sequence-specific DNA binding, GO:0004694 eukaryotic translation initiation factor 2 alpha kinase activity, GO:0004711 ribosomal protein S6 kinase activity) and replication of DNA (e.g. GO:0003887 DNA-directed DNA polymerase activity, GO:0034061 DNA polymerase

activity) (Figure 40). In comparison, upregulated allelic genes from allele set 2 were involved in transport (GO: GO:0005215 transporter activity) and protein-protein interaction (GO: GO:0005515 protein binding).

Additionally, allele-specific gene expression also appeared to affect global functional gene ontology analyses (4.2.5; data not shown), because results for significantly overrepresented GO terms were influenced by the gene set used for the analysis. Based on a functional analysis using the complete set of 27,137 predicted genes for *F. cylindrus* including gene copies from heterozygous regions of the genome, 14 enriched GO terms were identified showing that genes involved in signal transduction, transport and regulation of gene expression and DNA replication were enriched in upregulated genes in *F. cylindrus* during prolonged darkness relative to continuous light (Figure 30). In comparison, a functional analysis based on 18,073 genes predicted for the single-haplotype of *F. cylindrus* identified 7 enriched GO terms from the upregulated in *F. cylindrus* during prolonged darkness relative to continuous light and showed a lower overall resolution of enriched GO terms involved in signal transduction (GO:0004672 protein kinase activity), transport (GO:0005215 transporter activity) and regulation of gene expression (GO:0043565 sequence-specific DNA binding, GO:0003700 sequence-specific DNA binding transcription factor activity) and protein metabolism (GO:0005515 protein binding, GO:0004842 ubiquitin-protein ligase activity) (data not shown). Additionally, the functional analysis based on 18,073 single-haplotype genes in *F. cylindrus* identified the molecular function GO term “ion channel activity” (GO:0005216) as enriched in upregulated genes during prolonged darkness relative to continuous light, but which could not be identified based on the functional analysis of 27,137 predicted genes including all genes from heterozygous regions of the *F. cylindrus* genome (Figure 30).

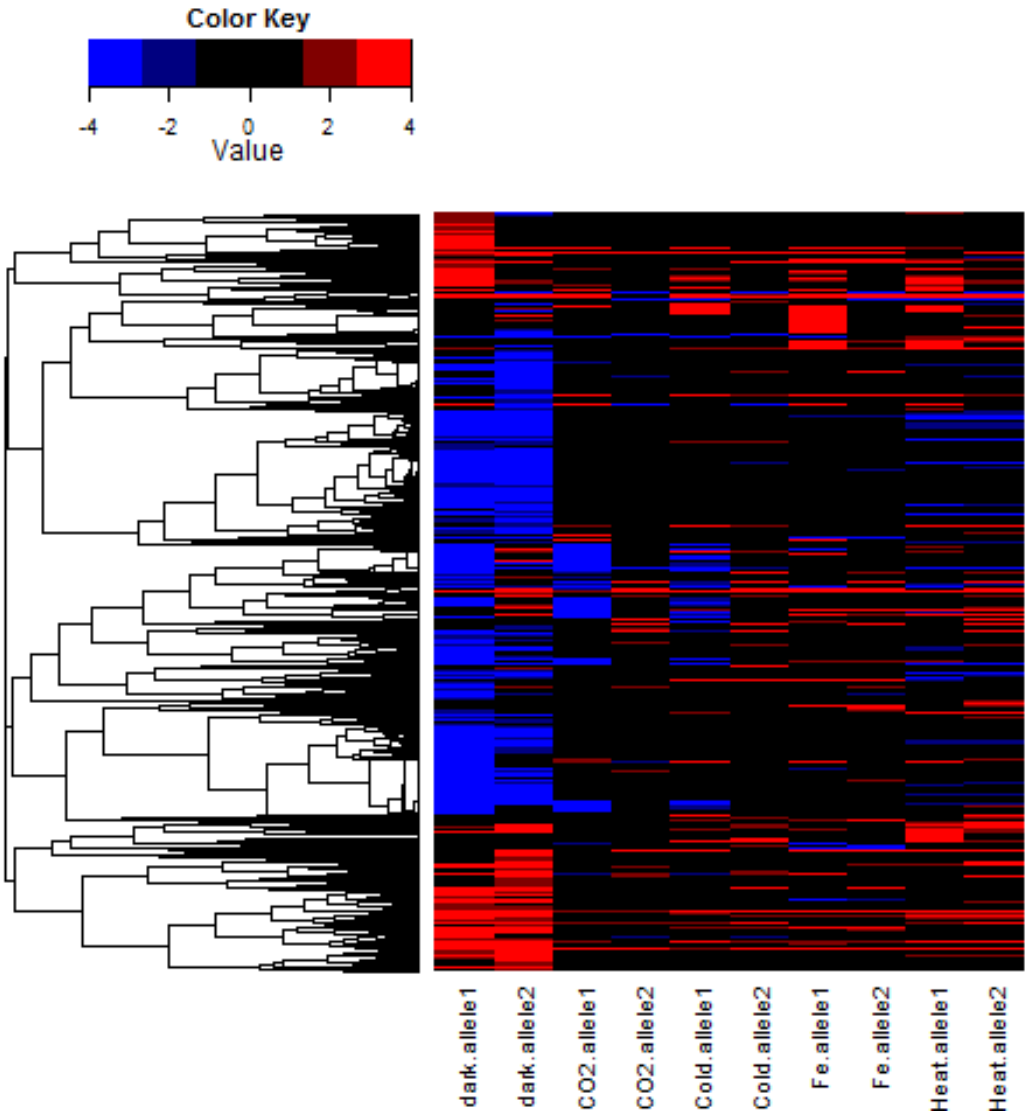



**Figure 38. Hierarchical clustering of 5790 differentially expressed allelic gene pairs in *F. cylindrus* (likelihood ratio test P < 0.001, $\log_2$ fold change ≤ −2 or ≥ +2) under iron (Fe)-starvation, freezing temperature (cold, −2 °C), elevated temperature (heat, +10 °C), elevated carbon dioxide (CO2, 1000 ppm $CO_2$/Air) and prolonged darkness (Dark, 1 week darkness) relative to optimal polar summer growth conditions (+4 °C, 35 µmol photons $m^{-2}$ $s^{-1}$ continuous light, nutrient-replete). Each experimental treatment corresponds to two separate columns for both allelic variants and each single-haplotype gene to a single row. The colour scale ranges from saturated red for up-regulated genes to saturated blue for down regulated genes; black indicates no significant regulation. A one minus Pearson correlation distance metric was applied to cluster rows using complete linking method (created using R software (R Development Core Team 2012), heatmap.2 function of gplots package, http://www.r-project.org/).**

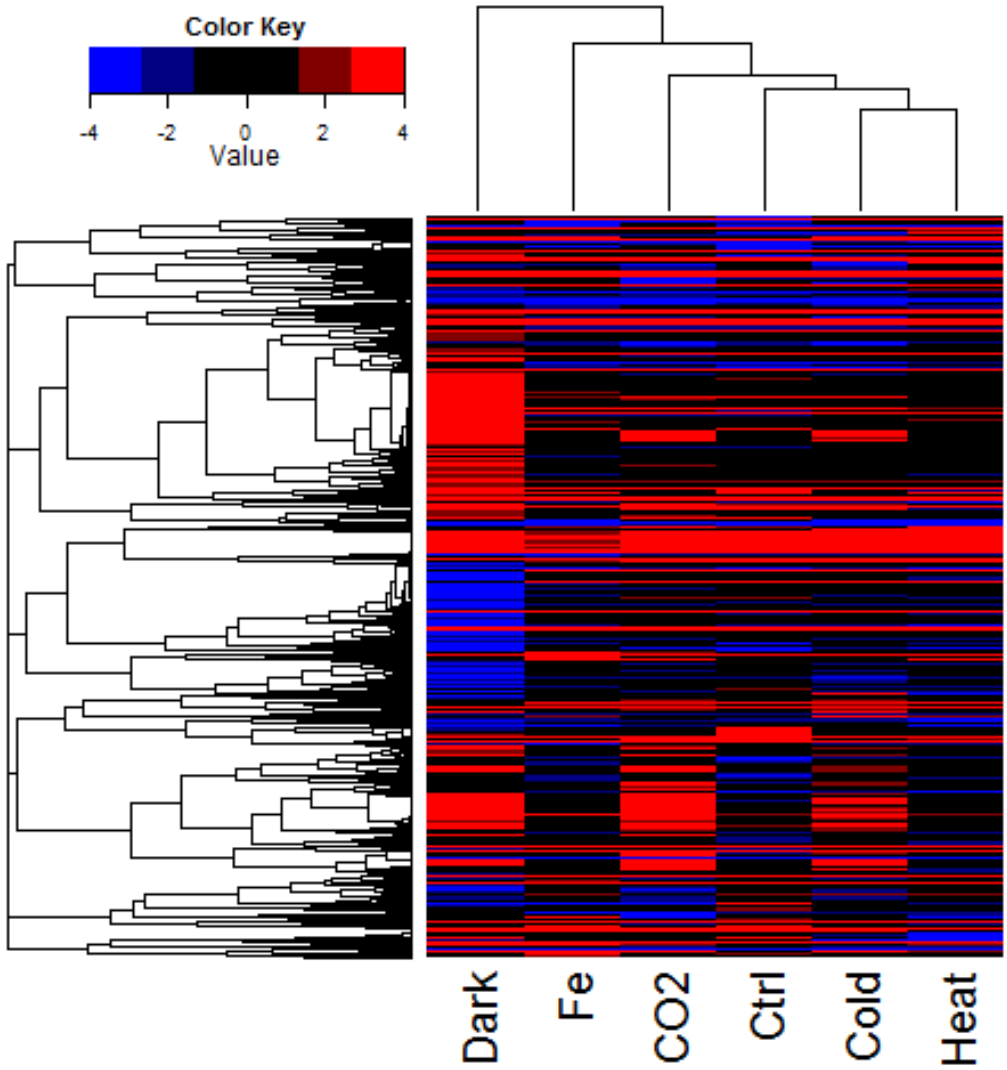



**Figure 39. Differential bi-allelic expression in *F. cylindrus*. Hierarchical clustering of 4434 differentially expressed allelic genes relative to their sister allele (likelihood ratio test P < 0.001, log2 fold change ≤ −2 or ≥ +2) under iron (Fe)-limitation, Cold (−2 °C), Heat (+10 °C), increased carbon dioxide (1000 ppm CO2), prolonged darkness (Dark, 1 week darkness), optimal growth conditions (Ctrl, +4 °C, 35 µmol photons $m^{-2}$ $s^{-1}$ continuous light, nutrient-replete). Each experimental treatment corresponds to one separate column and each allelic gene pair to a single row. The colour scale ranges from saturated red for up-regulated genes to saturated blue for down regulated genes; black indicates no significant regulation. A one minus Pearson correlation distance metric was applied to cluster rows and columns using complete linking method (created using R software (R Development Core Team 2012), heatmap.2 function of gplots package, http://www.r-project.org/).**

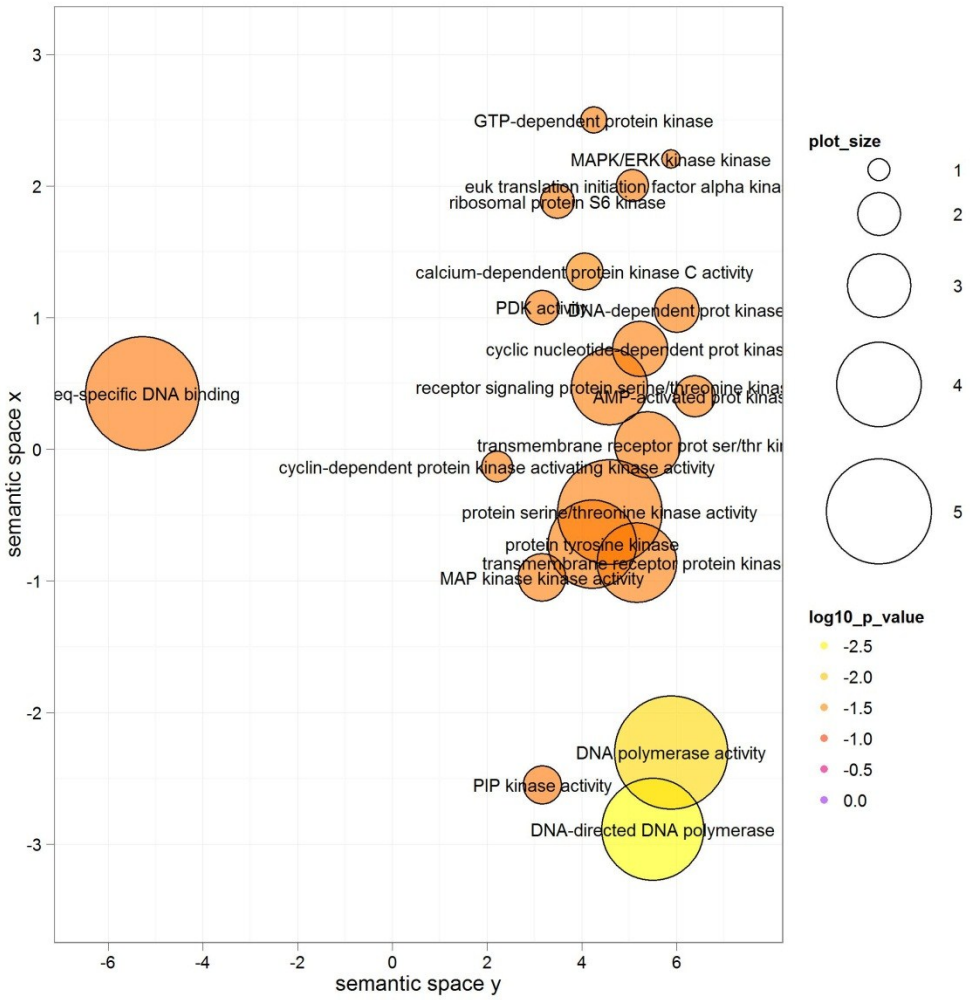



**Figure 40. Allele set 1: ReViGO scatterplot (Supek et al., 2011) showing enriched molecular function GO terms of upregulated allelic genes in *F. cylindrus* during prolonged darkness relative to continuous light during polar summer growth condition. Overrepresented GO terms (Wallenius approximation, Benjamini-Hochberg adjusted P < 0.05) among upregulated genes (GLM likelihood ratio test, P < 0.05) were determined using the goseq Bioconductor R package (Young et al., 2010). A non-redundant GO term set was plotted in a two dimensional space by applying a multidimensional scaling procedure so that more semantically similar GO terms are closer in the plot using the ReViGO Web server (http://revigo.irb.hr/). The bubble colour indicates significance levels and size indicates the frequency of the GO term in the underlying Gene Ontology Annotation (UniProt-GOA) Database. Abbreviation as follows: 3-phosphoinositide-dependent protein kinase (PDK) activity.**

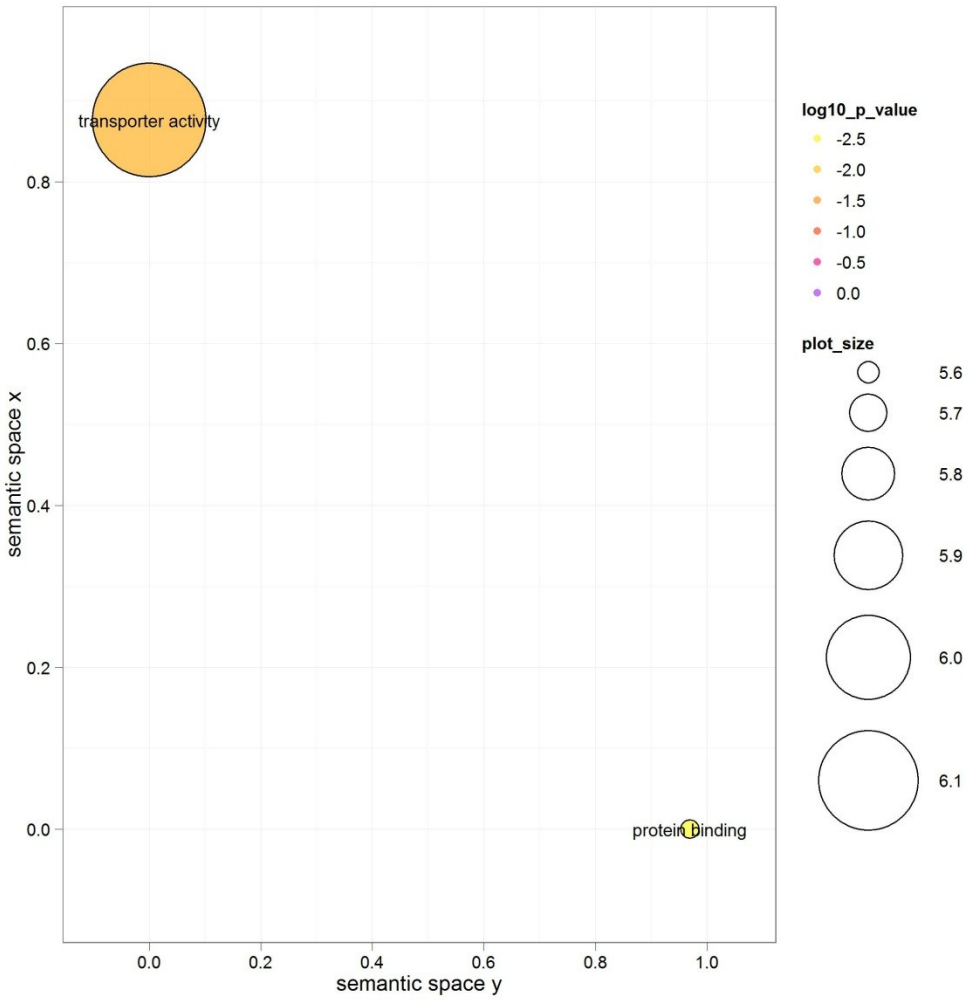



**Figure 41. Allele set 2: ReViGO scatterplot (Supek et al., 2011) showing enriched molecular function GO terms of upregulated allelic genes in *F. cylindrus* during prolonged darkness relative to continuous light during polar summer growth condition. Overrepresented GO terms (Wallenius approximation, Benjamini-Hochberg adjusted $P < 0.05$) among upregulated genes (GLM likelihood ratio test, $P < 0.05$) were determined using the goseq Bioconductor R package (Young et al., 2010). A non-redundant GO term set was plotted in a two dimensional space by applying a multidimensional scaling procedure so that more semantically similar GO terms are closer in the plot using the ReViGO Web server (http://revigo.irb.hr/). The bubble colour indicates significance levels and size indicates the frequency of the GO term in the underlying Gene Ontology Annotation (UniProt-GOA) Database.**

To corroborate RNA-Seq results for allele-specific gene expression, an allele-specific RT-qPCR analysis was performed on the large ribosomal protein subunit L27 in *F. cylindrus*. Allelic gene copies for L27 were encoded by L27/269038 (denoted allele 1) and L27/273430 (denoted allele 2). Both allelic gene copies were strongly expressed showing high absolute FPKM expression values determined by RNA-Seq during all experimental conditions, except for prolonged darkness which showed expression values of 1.5 and 0.6, respectively (Table 14). It is shown that L27/269038 (allele 1) was generally more highly expressed than L27/273430 (allele 2) under all tested experimental conditions (Table 14). This result could be confirmed by the determination

of allele-specific expression of L27 in *F. cylindrus* using RT-qPCR (Figure 42). Although the relative percentages determined for allelic expression of L27 by RT-qPCR deviated from percentages calculated based on RNA-Seq FPKM values, they were in good agreement for all other conditions (Table 14; Figure 42). The variation of percentages in allelic expression of L27 determined by RT-qPCR was caused by its low expression preventing its detection during qPCR analysis.

Finally, a correlation analysis of $\log_2$ fold changes determined by RT-qPCR and RNA-Seq with a coefficient of determination $R^2 = 0.92$ showed the good agreement of both methods over a wide range of transcript abundance (Figure 43).

**Table 14. Absolute RNA-Seq FPKM expression values for the L27 gene copy pair in *F. cylindrus* and percentages of the total FPKM for each allelic gene copy.**

| | **Ctrl** | **Cold** | **Fe** | **Dark** | **Heat** | **CO2** |
|---|---|---|---|---|---|---|
| FPKM (L27/269038) | 341.94 | 414.36 | 188.11 | 1.51 | 99.99 | 348.29 |
| FPKM (L27/273430) | 166.62 | 140.53 | 88.725 | 0.61 | 50.94 | 113.56 |
| FPKM total | 508.57 | 554.9 | 276.84 | 2.12 | 150.9 | 461.85 |
| % Allele 1 (L27/269038) | 67.2 | 74.7 | 68.0 | 71.1 | 66.2 | 75.4 |
| % Allele 2 (L27/273430) | 32.8 | 25.3 | 32.0 | 28.9 | 33.8 | 24.6 |

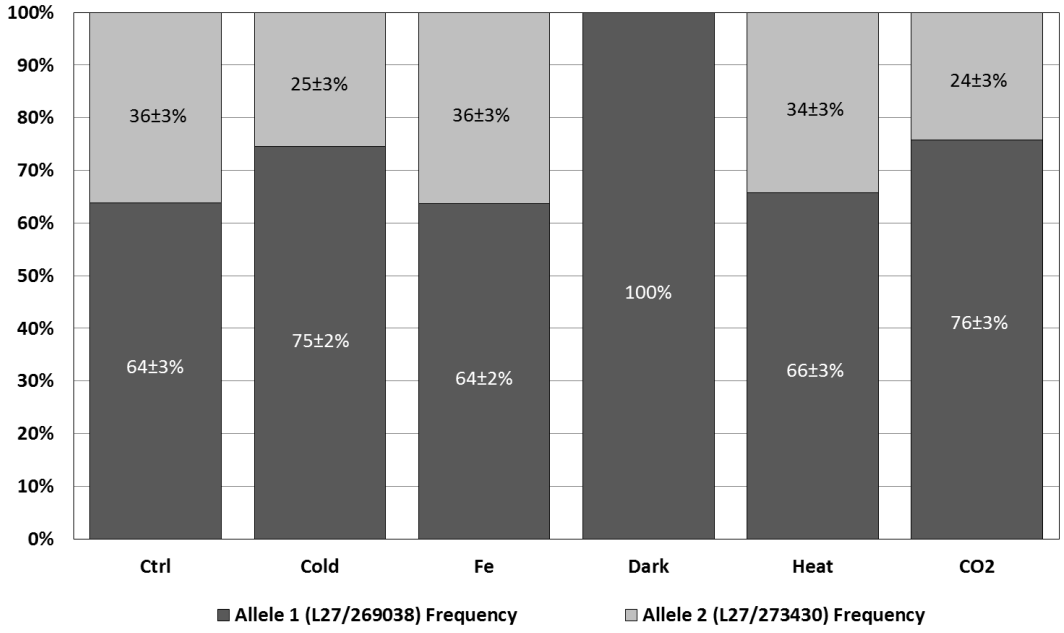



**Figure 42. Relative allelic expression of large ribosomal subunit L27 in *F. cylindrus* under different experimental conditions as determined by RT-qPCR. The allele frequency was calculated according to frequency of allele$_1$ = 1/($2^{\Delta Ct}$ + 1) [Germer et al. (2000), *Genome Research* 10: 258-66].**

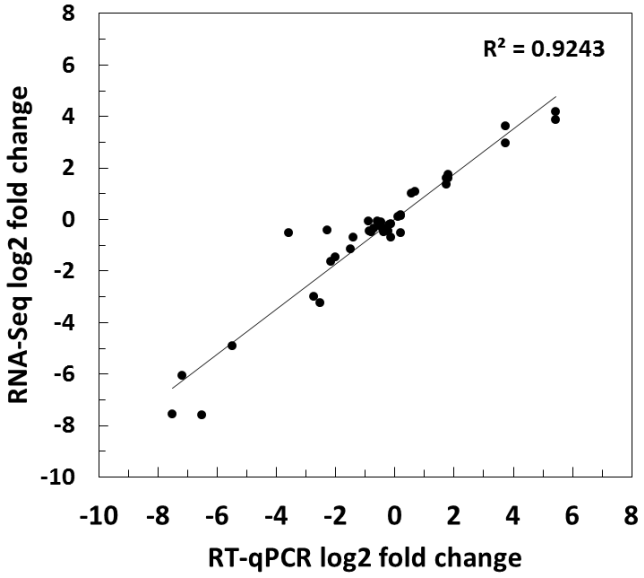



**Figure 43. Comparison of log2 fold expression values determined by RNA-Seq and RT-qPCR in *F. cylindrus*. Fold changes are given for every experimental condition relative to optimal growth at continuous light (Ctrl) for individual transcripts which were quantified by RT-qPCR.**

## 4.3 Discussion

The transcriptome of the polar diatom *F. cylindrus* was studied under six experimental conditions including polar summer growth conditions, freezing temperatures, elevated temperatures, elevated carbon dioxide, iron starvation and prolonged darkness using RNA-Seq. A total of 68.8 million reads was generated, out of which 70% were uniquely mapped to the draft genome of *F. cylindrus* and used for digital expression analysis.

The RNA-Seq approach was sensitive enough to detect widespread transcription of the *F. cylindrus* genome and detected traces of RNA, which could not be detected by RT-qPCR (compare Figure 42 and Table 14). Transcriptional activity was detected for 95% of all predicted genes and interestingly, > 98% of the putative allelic gene copy pairs were identified as transcriptionally active indicating their importance in the *F. cylindrus* genome. The high percentage of transcriptional active genes shows that the reported genome-wide expression profiling of *F. cylindrus* grown under six different growth conditions covered most of its genetic repertoire supporting > 90% of currently predicted gene models. Nevertheless, up to 30% of uniquely mapped reads were mapped to genomic regions in *F. cylindrus* without a predicted genomic feature including intron and intergenic regions and suggesting a high degree of unknown transcriptional activity. Interestingly, the highest percentages (~30%) of reads mapping to no genomic features were detected for *F. cylindrus* cultured under prolonged darkness. As a result, a total of 22,871 transcriptionally active regions (TARs) without annotated genomic feature were identified, which may include novel protein-coding genes, untranslated regions of existing gene models and long non-coding regulatory RNAs. The highest abundances of TARs was found to be in the 0.6 – 1kb size range, which is about half of the average gene length of ~1.6 kb predicted for diatom genes (Armbrust et al., 2004; Bowler et al., 2008) and may suggest that most are non-coding and may have regulatory functions in response to experimental stresses. However, TARs could also be identified in the size range from 1.5 kb up to a maximum of 10kb, which are likely to contain novel protein-coding genes. A similar pattern for length distribution of TARs was reported for the fungus Ascocoryne sarcoides using RNA-Seq analysis (Gianoulis et al., 2012). Furthermore, in the same study, a number of long and highly expressed TARs devoid of open reading frames is reported and the authors suggest a regulatory role (Gianoulis et al., 2012). In comparison, the 57 novel TARs identified for *F. cylindrus* (obtained after filtering for TARs with distance to nearest predicted gene of

> 250 nt) contained a high number of 1263 open reading frames on both strands and 10 ORFs produced hits to protein reference databases suggesting they may be unpredicted genes rather than regulatory non-coding RNAs. Interestingly, two of the identified novel TARs showed high homologies to transposable elements (TEs). TEs have been reported to respond to various stresses in plants (Wessler, 1996) and in the pennate diatom *P. tricornutum* a transposable element was found hypomethylated in response to nitrate starvation providing a direct link between environmental stress and chromatin modelling in diatoms (Maumus et al., 2009). Thus, our finding of two putative unidentified TEs, which were expressed to experimental stresses strengthens their potential importance in genome evolution of pennate diatoms, as suggested by (Maumus et al., 2009).

In summary, identification of high transcriptional activity of the *F. cylindrus* genome including intergenic regions and novel TARs is in agreement with recent transcriptome data from eukaryotic organisms indicating that the transcribed portions of genomes are large and complex, and that many functional properties of transcripts are based not on coding sequences but on regulatory sequences in untranslated regions or non-coding RNAs (Yamada et al., 2003; David et al., 2006; Amaral et al., 2008).

A global analysis of transcriptional profiles from all RNA-Seq libraries showed a separate clustering of replicated sample groups to each other, reflecting the experimental design (Figure 27). Notably was the distinct separate clustering of samples from *F. cylindrus* grown under prolonged darkness to all other conditions in dimension 1 of a MDS analysis, which was likely caused by differential expression. In comparison, the separation of samples from *F. cylindrus* grown under prolonged darkness was less clear in dimension 2 to cells grown under freezing temperatures and optimal growth conditions with continuous light (Figure 27), which may relate to unknown technical noise. Noteworthy, distances in dimension 1 on a MDS plot, separating samples by differential expression, were greatest between samples from *F. cylindrus* grown under prolonged darkness compared to cells grown under continuous light with optimal growth conditions and suggested high numbers of differentially expressed genes between both treatments. Indeed, a condition-specific analysis of the *F. cylindrus* transcriptome showed a high number of 18,475 (68% of predicted gene models) differentially expressed genes ($p < 0.05$) between both experimental conditions.

Strikingly, the sample relationships in a MDS analysis changed if the data set was filtered for genes predicted for the single haplotype of *F. cylindrus* (Figure 28) and suggested specific expression of putative allelic gene copies under different experimental conditions, assuming that MDS relationships would not change significantly if allelic copies were expressed uniformly under each condition. Further allele-specific analysis of the *F. cylindrus* transcriptome indeed showed that 98% of the putative allelic copy pairs from heterozygous regions of the genome were transcriptionally active and 70% of all putative allelic copy were differentially expressed (relative $\log_2$ fold change $\leq -2$ or $\geq +2$ in comparison to optimal growth; Figure 38). Additionally, a strong differential bi-allelic expression (non-uniform expression, relative $\log_2$ fold change $\leq -2$ or $\geq +2$ between allelic copy pairs) could be shown for 55% of all allelic copy pairs suggesting an important role of bi-allelic expression in *F. cylindrus* under different experimental conditions (Figure 39). Furthermore, the functional significance of differential bi-allelic expression on metabolism in *F. cylindrus* was shown by a separate analysis of upregulated allelic genes in cells grown under prolonged darkness relative to continuous light using molecular function GO terms (Figure 40; Figure 41). An overrepresentation of molecular function GO terms referring to signal transduction, regulation of transcription and DNA replication was found for the allelic gene copy set 1 (Figure 40) in comparison to overrepresentation of GO terms related to transport and protein-protein interactions as found for the corresponding allelic gene copy set 2 (Figure 41), suggesting a separation of metabolism between allelic gene copies. Furthermore, the comparison of both analyses suggests that prolonged darkness activated signalling cascades which lead to the induction of DNA-binding proteins that initiated transcription and DNA replication, which was exclusively determined by allelic gene copies from allele set 1 (Figure 40), whereas under the same growth conditions allelic gene copies from allele set 2 dominate transport and protein-protein interaction metabolism (Figure 41). Consequently, it is likely that, under the same growth condition of prolonged darkness, individual representatives from each allelic gene copy contribute differently to a genome-wide functional GO analysis of gene expression in *F. cylindrus* and that GO terms associated with signal transduction and DNA-binding are mostly determined by allelic copies from allele set 1, whereas transport metabolism is determined mainly by allelic copies from allele set 2 under prolonged darkness. Indeed, the functional analysis based on all 27,137 predicted gene models including all allelic gene copies from heterozygous regions of the genome (Figure 30) showed a deviant set

of overrepresented GO terms in comparison to the same analysis based on 18,073 genes predicted for the single haplotype of *F. cylindrus* (not shown). This functional difference of gene expression in the single haplotype gene set is not only likely to be caused by strong differential bi-allelic expression, but also a bias towards functional categories not represented by allelic gene copy pairs, which was the major reason to base our main analysis on all 27,137 predicted genes in *F. cylindrus* including all gene copies from heterozygous regions of the genome.

Although, to my knowledge, this is the first report on allele-specific expression in eukaryotic phytoplankton, the phenomenon of allele-specific gene expression has been widely reported in mammals (Cowles et al., 2002; Enard et al., 2002; Yan et al., 2002; Cheung et al., 2003; Lo et al., 2003), fish (Oleksiak et al., 2002), plants (Guo et al., 2004; Schaart et al., 2005) and yeast (Brem et al., 2002). It has been reviewed that the variations in allelic expression are likely to be context-specific with regard to cell type and stimulus (e.g. environmental condition) and may have physiological implications (Knight, 2004). Moreover a recent study identified allele-specific gene expression in maize in response to environmental stresses and functional diversity of allelic copies was suggested (Guo et al., 2004). Accordingly, the high percentage of transcriptionally active heterozygous allelic gene copies in *F. cylindrus* suggests their importance in adaptation of a polar eukaryote. Moreover, a strong differential expression of allelic gene copies in *F. cylindrus* suggests that individual copies are under different regulatory controls enabling cells to coordinate gene expression in different ways, ultimately leading to a high metabolic flexibility and capacity to adapt to a rapidly changing environment.

A condition-specific differential expression analysis of the *F. cylindrus* transcriptomes from all six tested experimental conditions showed that 12,812 genes were differentially expressed by > 4 fold (likelihood ratio test, $P < 0.001$) in at least one experimental condition relative to optimal growth conditions with continuous light (Figure 29). Strikingly, a hierarchical clustering of all 12,812 differentially expressed genes showed strong differential expression in *F. cylindrus* during prolonged darkness revealing two main gene clusters with an opposite expression pattern (Figure 29). The first cluster showed upregulation under prolonged darkness but mostly down regulation in other conditions relative to optimal growth with continuous light conditions, whereas a second larger cluster showed down regulation under prolonged darkness but little or no expression in other conditions relative to optimal growth under continuous light

conditions (Figure 29; Table 11). In addition to relative comparisons of experimental treatments to optimal growth conditions with continuous light, we also analysed the relative transcriptional differences between all experimental conditions applying pair-wise treatment-by-treatment comparisons (Table 11). As indicated by a hierarchical clustering and multidimensional scaling analysis, the highest numbers of differentially expressed genes in *F. cylindrus* were identified for all experimental conditions compared to prolonged darkness (Table 11) and the numbers of differentially expressed genes are consistent with distances of samples on a MDS plot in dimension 1 confirming the separation of samples in dimension 1 by differential expression found by global analysis of transcriptional profiles using MDS (see discussion above; Figure 27). Although the biological interpretation of numbers for differentially expressed genes in *F. cylindrus* between two different experimental conditions (Table 11) remains purely speculative, because different genes may contribute to the numbers of differentially expressed genes, one main application of this table is to use the underlying gene sets for functional analysis to address specific biological questions. Against the background of temperature adaptation it might, for example, be of interest to compare enriched functional gene categories in the 6833 upregulated genes in *F. cylindrus* grown under freezing temperature (Cold) relative to elevated temperatures (Heat) (Table 11). Interestingly, this comparison revealed a significant enrichment of metabolism relation to translation (data not shown), which could also be found identified by a comparative analysis between polar and tropical metatranscriptome data sets (Toseland *et al.*, unpublished result). However, the focus of the current chapter is on a functional analysis of differentially expressed genes in prolonged darkness relative to continuous light, because prolonged darkness appeared to have the most significant effect on the transcriptome of *F. cylindrus* (Figure 27; Figure 29; Table 11).

The individual functional analysis of up- and down regulated genes in *F. cylindrus* grown under prolonged darkness relative to continuous light using gene ontologies showed that genes involved in regulation of gene expression and cellular transport were significantly enriched ($P < 0.05$) in the upregulated set of genes (Figure 30). Notably, genes with transport activity appeared to contribute to the significant enrichment of carbohydrate transport metabolism, as identified by biological process gene ontologies (Table 12). In comparison, down regulated genes in *F. cylindrus* grown under prolonged darkness relative to continuous light were mainly involved in translation, protein degradation as well as ATP synthesis (Figure 31). Thus, it seems that

*F. cylindrus* reduces its energetic demand by reducing the energetically expensive process of translation (Wilson and Nierhaus, 2007) to compensate for a loss of ATP production. Additionally, the overrepresentation of genes with cyclophilin activity during prolonged darkness was interesting (Figure 31), because cyclophilins are suggested to protect against oxidative stress (Doyle et al., 1999). Thus, the down regulation of cyclophilins under prolonged darkness relative to continuous light may relate to reduced photosynthetic production of reactive oxygen species and oxidative stress. As expected, down regulated genes in *F. cylindrus* under prolonged darkness relative to continuous light were related to photosynthesis and photosynthetic pigment synthesis (Table 13).

This result could be confirmed by the individual analysis of metabolic pathways differentially expressed in *F. cylindrus* (Figure 32). Most of the identified genes involved in the biosynthesis of carotenoids and the photoprotective xanthophyll cycle were significantly down regulated during prolonged darkness (Figure 36). Surprisingly, an opposite expression pattern and strong upregulation by a relative $\log_2$ fold change of 2.1 ($P < 0.001$) was found for a beta-carotene monooxygenase, which catalyses the cleavage of beta-carotene into retinal (Figure 36) and may provide the retinal chromophore for a light-dependent bacteria-like rhodopsin proton pump identified in the genome of *F. cylindrus* (3.2.2.6). Similarly to upregulation of a beta-carotene monooxygenase, both putative retinal-binding *Fragilaripsis* rhodopsin (FR) alleles were significantly upregulated during prolonged darkness by a $\log_2$ fold change of 2.0 for FR1/271123 and 5.4 for FR2/267528 ($P < 0.001$). The upregulation of both *FR* copies in *F. cylindrus* during prolonged darkness may be explained by a strong feedback activation caused by the lack of final protein product but it remains speculative at this stage as protein concentrations were not determined in this work.

In addition to the down regulation of most genes involved in carotenoid biosynthesis, all identified genes involved in chlorophyll biosynthesis were strongly down regulated by relative $\log_2$ fold changes $< -3$ (Figure 37). Although photosynthetic genes were down regulated, it appeared that cells maintained their photosynthetic apparatus without degradation of photosystems, because the photosynthetic quantum yield for photosystem II ($F_v/F_m$) remained on a constant high level of 0.54 during darkness (Table 8) and upon return to light, cells continued rapid growth (data not shown). In comparison to that Reeves *et al.* (2011) reported a drop in $F_v/F_m$ in *F. cylindrus* cultures during seven days of darkness. However, the authors also reported a

much lower $F_v/F_m$ value of 0.20 at the beginning of their darkness treatment, which might relate to a bad photosynthetic health of cells and explain a different response of $F_v/F_m$ values. Additionally, Reeves *et al.* (2011) could show constant levels of chlorophyll a for *F. cylindrus* over a darkness period of up to one month supporting the conclusion that the photosynthetic apparatus was maintained in *F. cylindrus* during 1 week of darkness. In addition to that several other studies on polar microalgae and a macroalgae showed an increase in pigment content during initial exposure to darkness (Peters and Thomas, 1996; Lüder et al., 2002), which might relate to an acclimatory response similar to low light. An early acclimatory response to low light can cause a rearrangement of photosynthesis antenna pigments (Eberhard et al., 2008), which is likely for *F. cylindrus*, too, and explain results showing opposite differential gene expression patterns for genes involved in carotenoid biosynthesis inconsistent with the general down regulation of that pathway (Figure 36).

In contrast, to the down regulation of genes related to photosynthesis and biosynthesis of pigments, starch and sucrose-related pathways as well as fatty acid metabolism showed high expression values during prolonged darkness (Figure 32), suggesting that *F. cylindrus* uses the oxidation of glucan and lipids to provide ATP and reduction equivalents for basal cell maintenance. This finding was supported by the expression pattern of genes involved in the diatom carbon storage product beta-1,3-glucan chrysolaminarin. It was found that most genes predicted to be involved in chrysolaminarin biosynthesis were significantly down regulated in *F. cylindrus* grown under prolonged darkness, whereas the majority of genes involved in the breakdown of chrysolaminarin were significantly upregulated (Figure 33). Ultimately, the complete degradation of chrysolaminarin may lead to free glucose, which can be phosphorylated by glucokinase (Figure 33), the initial step in glycolysis. Interestingly, one of the GLK isoenzymes (GLK2) showed strong upregulation by a $\log_2$ fold change of 6.3 ($P < 0.001$). The subsequent upper phase of glycolysis appeared to have no differential expression based on a metabolic pathway map (Figure 32). However, upregulation was shown for selected genes involved in the lower (payoff) phase of glycolysis (Figure 32), which was confirmed by individual expression analysis of this phase of glycolysis (Figure 34). Interestingly, although a putative pyruvate dehydrogenase was down regulated in *F. cylindrus* grown under prolonged darkness, a mitochondrial E1-component dehydrogenase, which may also catalyse the pyruvate decarboxylation to acetyl-CoA, showed high absolute expression values solely during prolonged darkness

(Figure 34) and thus seems to be a specific acclimatory response of *F. cylindrus* to prolonged darkness. In comparison to upregulation of selected genes involved in glycolysis, genes involved in the mitochondrial and peroxisomal beta-oxidation of fatty acids were more strongly upregulated throughout (Figure 35). Interestingly, the single identified acetyl-CoA acetyltransferase in *F. cylindrus*, catalysing the ultimate step during beta-oxidation of fatty acids showed high absolute expression values relative to all other experimental conditions and seemed to be strongly expressed during prolonged darkness only (Figure 35). Taken together our findings suggest that acetyl-CoA generated by glycolysis and beta-oxidation of fatty acids may feed into the tricarbonic acid cycle contributing to the production of NADH and $FADH_2$ used in oxidative phosphorylation to produce ATP for basal cell maintenance of *F. cylindrus* during prolonged darkness. Notably, parts of the TCA cycle as well as oxidative phorphorylation showed high expression values (Figure 32).

To my knowledge, this is the first report describing a detailed metabolic response of a polar autotrophic organism to prolonged darkness at the gene expression level. However, the reported findings are in general agreement with previous studies showing that sudden darkness did not induce resting spore formation in polar diatoms and cells survive in their vegetative stage maintaining their photosynthetic capacity (Peters and Thomas, 1996). Additionally, it appears a feature of polar phytoplankton species to be able to begin rapid growth upon return to light (Peters and Thomas, 1996; Tang et al., 2009), which was also found for *F. cylindrus* in this study (Figure 21). The high expression values for genes related to fatty acid metabolic pathways in *F. cylindrus* during prolonged darkness may indicate the utilisation of stored lipids for metabolic intermediates and generation of adenosine 5'-triphosphate (ATP), and, as suggested by Armbrust et al. (Armbrust et al., 2004), may explain how diatoms survive long periods of darkness. Furthermore, it has been shown for the mesophilic diatom *Cyclotella meneghiniana* that the degradation of storage products during darkness caused a decrease of all cellular macromolecules over time and degradation followed a well-defined sequence from degradation of the lipid fraction, followed by carbohydrates to proteins (Stehfest et al., 2005). Consequently, the finding that beta-oxidation of fatty acids was more strongly upregulated throughout than chysolaminarin metabolism and glycolysis in *F. cylindrus* under prolonged darkness (Figure 33; Figure 35) may reflect the well-defined sequence of degradation reported for *C. meneghiniana* (Stehfest et al., 2005), suggesting that mesophilic and psychrophilic diatoms share a similar adaptation

strategy to darkness. Additionally, the oxidation of lipids and carbohydrates in the dark was also suggested by studies on mitochondrial respiration in the diatom *Thalassiosira weissflogii*, showing an immediate stop of oxygen production when photosynthesising cells were transferred to darkness (Weger et al., 1989). Ultimately, a decline in protein in cells kept in darkness for a long period, like in overwintering *F. cylindrus*, could be expected due to the degradation of proteins like RubisCO (Geider et al., 1993). The degradation of light-harvesting antennae and the reaction centre proteins PSII and PSI began after 4 month of darkness in the Antarctic macroalgae *Palmaria decipiens* (Lüder et al., 2002).

## 4.4 Summary and conclusions

A genome-wide RNA-Seq analysis of transcriptomes of the psychrophilic diatom *F. cylindrus* grown under six experimental conditions including polar summer growth conditions, freezing temperatures, elevated temperatures, elevated carbon dioxide, iron starvation and prolonged darkness provides unprecedented insights into the transcriptional complexity of a polar eukaryote. Transcriptional activity was detected for 95% of all predicted genes in *F. cylindrus* including putative allelic copies from heterozygous regions of the genome. Furthermore, 98% of heterozygous allelic copies were transcriptionally active and 55% showed ≥4 fold non-uniform bi-allelic expression suggesting that individual copies are under different regulatory controls and enable *F. cylindrus* to coordinate gene expression in different ways, ultimately leading to a high metabolic flexibility and capacity to adapt to a rapidly changing environment. Moreover, up to 30% of RNA sequencing reads mapped to the *F. cylindrus* genome were not associated with a predicted genomic feature and may include novel protein-coding genes and non-coding regulatory RNAs. Additionally, it was found that prolonged darkness caused significant transcriptional changes in *F. cylindrus* providing unprecedented details of the molecular responses of a polar autotrophic organism to initial darkness at the beginning of the polar winter.

# Chapter 5

# A bacteria-like rhodopsin proton pump from the psychrophilic diatom *Fragilariopsis cylindrus*

## 5.1 Introduction

A bacteria-like rhodopsin was discovered in the draft genome of the psychrophilic diatom *F. cylindrus,* but gene products resembling rhodopsins are absent in the sequences of the mesophilic diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* (Chapter 3). Rhodopsins are photoreceptors consisting of seven transmembrane domain proteins, called opsins, and the light-absorbing chromophore retinal. Rhodopsin genes are classified based on their different primary sequences into microbial (type I) and animal (type II) rhodopsins (Spudich et al., 2000). Microbial (type I) rhodopsins are found in prokaryotes, fungi and eukaryotic algae and type II rhodopsins are present in higher eukaryotes, including humans (e.g. visual photoreceptors). In contrast to type II rhodopsins, which use G protein-coupled signal transduction pathways, most microbial type I rhodopsins directly regulate membrane ion conductance. Furthermore, rhodopsins can be classified according to electrical properties into (1) electrically neutral photoreceptors in animal eyes or sensors for phototaxis in prokaryotes (Spudich, 2006), (2) channel rhodopsins causing light-induced passive conductance of $H^+$ and other cations in phototactic algae (Nagel et al., 2002; Sineshchekov et al., 2002) and (3) light-driven ion pumps for $H^+$ and $Cl^-$ providing a mechanism of phototrophy in prokaryotes (Danon and Stoeckenius, 1974).

Light-driven ion pumps have been extensively studied in prokaryotes (Oesterhelt and Stoeckenius, 1971; Matsuno-Yagi and Mukohata, 1977; Grigorieff et al., 1996; Beja et al., 2000; Balashov et al., 2005) and phototrophy has been shown in marine gammaprotobacteria conferred by $H^+$-puming rhodopsins, called proteorhodopsins (Beja et al., 2001). Proteorhodopsins exhibit high genetic mobility (Frigaard et al., 2006; Sharma et al., 2006) and are widespread in the marine environment (Man et al., 2003b; Sabehi et al., 2003; Sabehi et al., 2004; Venter et al., 2004; Atamna-Ismaeel et al., 2008) including the Arctic Ocean (Jung et al., 2008), Southern Ocean (de la Torre et al., 2003) and sea ice (Koh et al., 2010; Qin et al., 2012). Moreover, proteorhodopsin transcripts were abundant in the North Atlantic (Campbell et al., 2008) and in environmental

metatranscriptomes of the Pacific (Frias-Lopez et al., 2008) and Southern Ocean (Andrew Toseland, unpublished data). The ecological role of proteorhodopsins, however, remains unclear. On the one hand light promoted growth and survival of some proteorhodopsins-containing bacterial cultures (Gomez-Consarnau et al., 2007; Gomez-Consarnau et al., 2010) but on the other hand it did not affect growth of others (Giovannoni et al., 2005; Stingl et al., 2007; Giovannoni et al., 2008), suggesting additional functions (Spudich, 2006; Fuhrman et al., 2008). Moreover, bacteria-like $H^+$-pumping rhodopsins have also been discovered in several eukaryotes including fungi (Bieszke et al., 1999b; Idnurm and Howlett, 2001; Waschuk et al., 2005), dinoflagellates (Okamoto and Hastings, 2003; Ruiz-González and Marín, 2004; Lin et al., 2010; Slamovits et al., 2011), the cryptomonad alga *Guillardia theta* (Ruiz-González and Marín, 2004), the unicellular green alga *Acetabularia* (Tsunoda et al., 2006), the haptophyte *Phaeocystis globosa* as well as the pennate diatoms *Pseudo-nitzschia granii* and *F. cylindrus* (Marchetti et al., 2012). While some of these eukaryotic rhodopsins, such as the rhodopsin from the fungus *Neurospora crassa* (Bieszke et al., 1999a) and the marine cryptomonad *G. theta*, appear to be sensory rhodopsins as found in green alga (Nagel et al., 2002), the biochemical properties of most recently discovered eukaryotic rhodopsins have not been investigated. Nevertheless, fast photocycle turnover (< 50 milliseconds) associated with light-driven $H^+$-pumping activity has been shown for the fungal pathogen *Leptosphaeria maculans* (Waschuk et al., 2005) and the giant marine unicellular green alga *Acetabularia acetabulum* (Tsunoda et al., 2006). The *Acetabularia* rhodopsin (Tsunoda et al., 2006) provides the first evidence for an ion-pumping rhodopsin in a photosynthetic eukaryote but no information on its *in vivo* function is available and the physiological role of a light-driven proton pump in photosynthetic algae in the presence of a proton gradient-generating chlorophyll-based photosynthetic apparatus remains puzzling.

Here, light-dependent proton-pumping is also shown for eukaryotic rhodopsins from the Antarctic dinoflagellate *Polarella glacialis* and the polar marine diatom *F. cylindrus*. Additionally, I report on the functional characterisation of the *Fragilariopsis* rhodopsin using reverse genetics, biochemical and biophysical approaches to elucidate its physiologyical role in marine photosynthetic organisms and adaptation to conditions of the Southern Ocean including low iron.

## 5.2 Results

### 5.2.1 *In silico* analysis of *Fragilariopsis* Rhodopsin

A bacterial rhodopsin was predicted in the *F. cylindrus* genome (FR2, 274098) by FGENESH (Salamov and Solovyev, 2000) and Genewise (Birney et al., 2004) algorithms. Its coding sequence length was 777 bp consisting of four exons interspersed with three introns and was strongly supported by EST sequences. Subsequently, a gene copy variant with 100% amino acid sequence identity but a 30 amino acid N-terminal extension was identified (FR1/FRext, 287459) caused by N-terminal non-synonymous base pair exchanges.

No signal peptides were predicted for both FR1 and FR2 gene copies by SignalP v4.0. Contrary, signal peptides were predicted by TargetP v1.1 including prediction of a mitochondrial targeting peptide for FR1 (mTP, score = 0.379) and signal peptide prediction for FR2 (Table 15). By manual inspection of putative N-terminal signal peptide splice sites around phenylalanine (F) residues in both FR protein sequences, similarities to the conserved chloroplast "ASAFAP" motif (Kilian and Kroth, 2005; Gruber et al., 2007) were identified and signal peptide splice sites were found at F-14 (FR1) and F-18 (FR2, F-48 in FR1). Additionally, a lysine rich motif (KNKKKKKAVK) was observed within the extended N-terminus of FR1 which was also predicted as a partial membrane loop segment (Figure 44).

**Table 15. Computational prediction of subcellular targeting of *Fragilariopsis* rhodopsin. Targeting predictions abbreviated as follows: mTP: mitochondrial targeting peptide score, SP: signal peptide score, other: probability for other localisation, Loc: prediction of localisation based on the scores of TargetP, RC: reliability class, 1 = strong, 5 = poor prediction. Targeting predictions were performed by TargetP 1.1 (http://www.cbs.dtu.dk/services/TargetP/) (Emanuelsson et al., 2000).**

| TargetP non-plant networks prediction | | | | | | |
|---|---|---|---|---|---|---|
| | cTP | mTP | SP | other | Loc | RC |
| FR1 | n/a | 0.379 | 0.246 | 0.318 | M | 5 |
| FR2 | n/a | 0.104 | 0.929 | 0.042 | S | 1 |

| TargetP plant networks prediction | | | | | | |
|---|---|---|---|---|---|---|
| | cTP | mTP | SP | other | Loc | RC |
| FR1 | 0.041 | 0.351 | 0.033 | 0.872 | | 3 |
| FR2 | 0.014 | 0.013 | 0.964 | 0.291 | S | 2 |

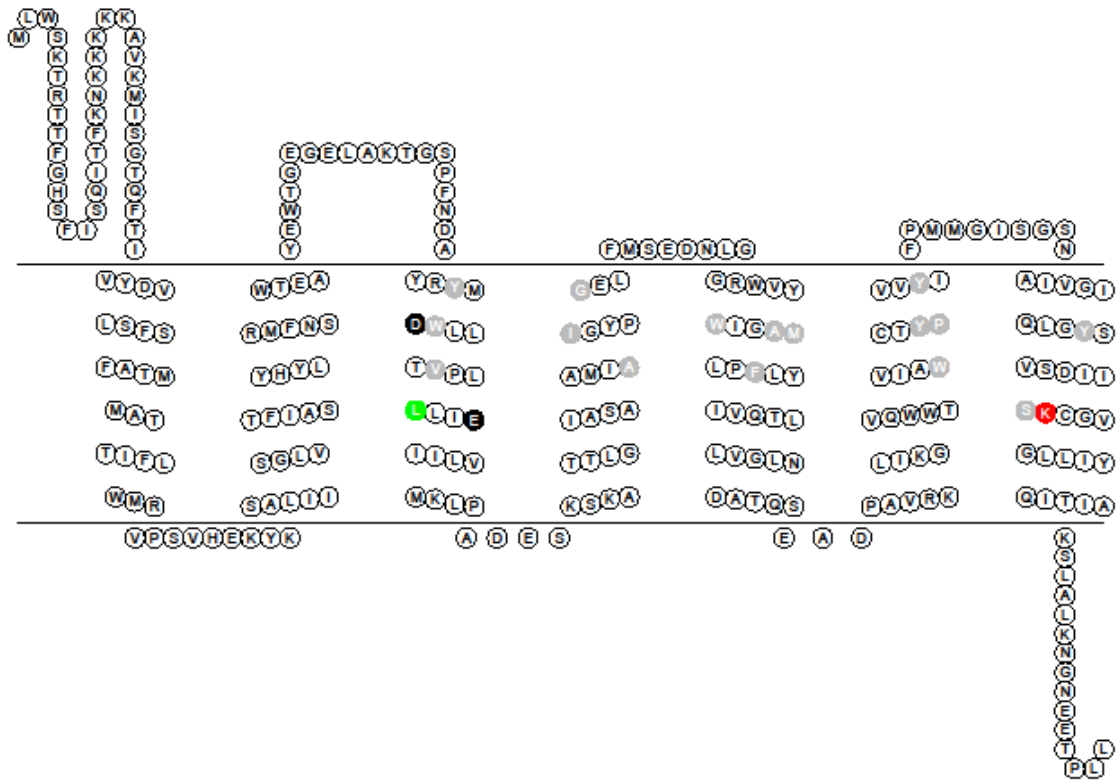


**Figure 44. Secondary protein structure of *F. cylindrus* rhodopsin FR1/FRext. Structure was predicted by I-TASSER (http://zhanglab.ccmb.med.umich.edu/I-TASSER/) and displayed using TOPO2 (http://www.sacs.ucsf.edu/TOPO2/). Single-letter amino acid codes are shown and numbers correspond to the residue positions in *F. cylindrus* and bacteriorhodopsin (1QM8), respectively. Key residues are highlighted as follows: Black: proton acceptor (D-121) and proton donor (E-132); green: spectral tuning (L-129); red: retinal Schiff base linkage (K-261); light grey: retinal binding pocket.**

The FR1 protein sequence showed high sequence similarity (> 45% pairwise identity) to homologs from the haptophyte *Phaeocystis globosa* (AEP68178, 57.0%), an Arctic Oxalobacteraceae bacterium IMCC9480 (ZP_08273891, 52.3%), the alpine glacier bacterium *Janthinobacterium* sp. Strain PAMC 25724 (ZP_10444755, 49.0%) and the Antarctic alphaproteobacterium *Octadecabacter antarcticus* (ZP_05063020, 47.2%). In addition to that, equally high sequence similarities were found with the dinoflagellates *Pyrocystis lunula* (AAO14677, 50.2%), *Oxyrrhis marina* (ADY17806, 49.2%) and *Polarella glacialis* (AEF32712, 46.3%). Next most similar were carotenoid-binding xanthorhodopsins from the cyanobacterium *Gloeobacter violaceus* (NP_923144, 40.6%) and the extreme halophilic *Salinibacter ruber* (YP_445623, 39.0%) followed by the blue-light absorbing rhodopsin from the Antarctic sea-ice bacterium *Glaciecola punicea* ACAM 611T (ZP_09921023, 33.9%). Noteworthy with regard to spectral light tuning, sequence similarity (< 35%) was found with green light-absorbing proteorhodopsin from proteobacterium clone BAC_31A08 (Q9F7P4, 31.4%) and the blue light-absorbing PRs from proteobacterium clones HOT_75m4

(AAK30179, 27.7%) and palE6 (AAK30200, 26.7%). As shown in Figure 44, the FR contained a non-polar leucine residue at postion 129 (L-129), which serves as spectral tuning switch in green and blue light-absorbing PR (Man et al., 2003a) and in bacteriorhodospin (BR; L-93 in BR) (Subramaniam et al., 1991).

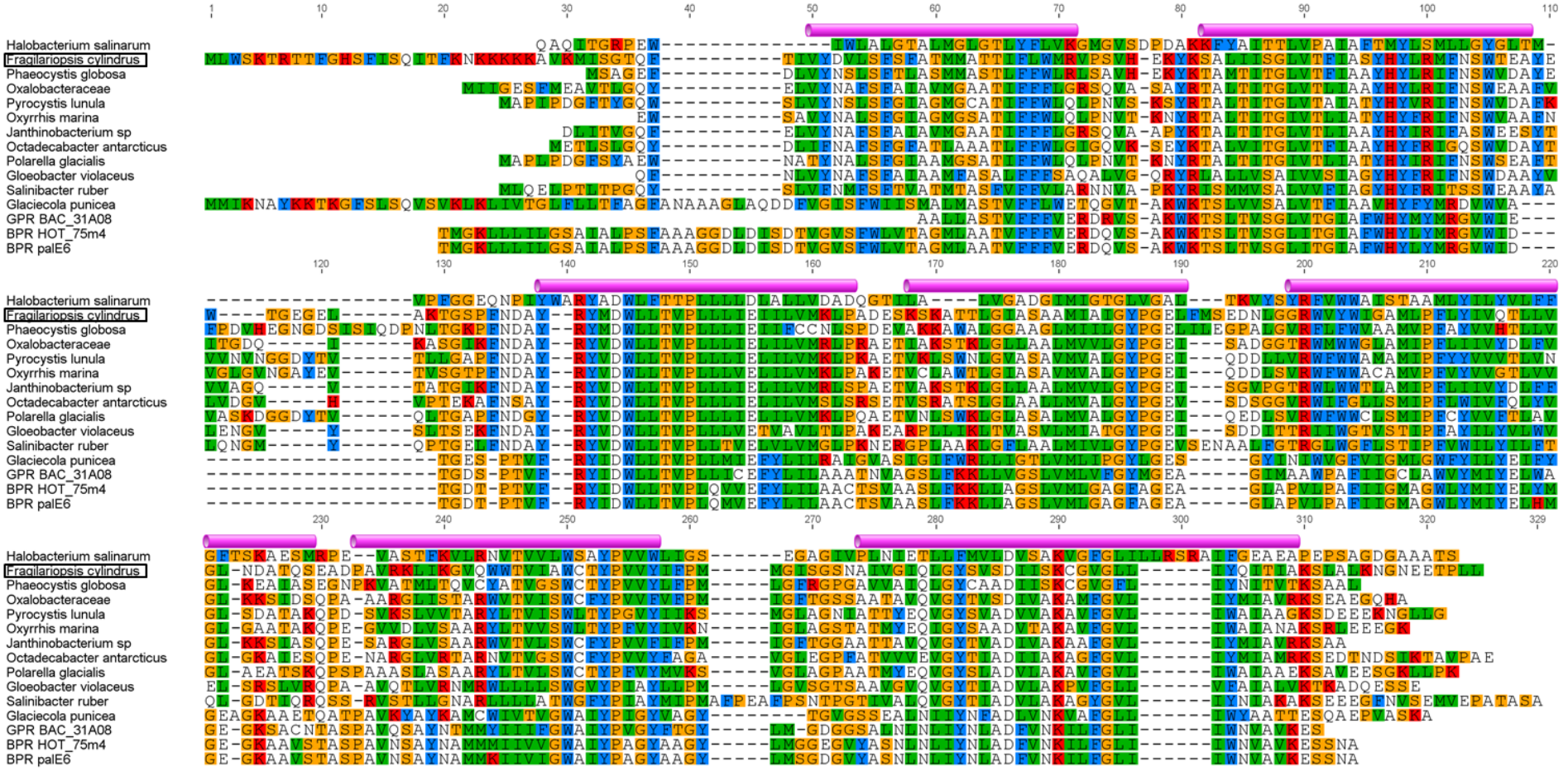



**Figure 45. Protein alignment of bacterial and eukaryotic rhodopsins. Alignment of bacteriorhodopsin from *Halobacterium salinarum* (Accession# 1QM8_A) with rhodopsins from *Fragilariopsis cylindrus* (FR1/FRext, Protein ID 287459), *Phaeocystis globosa* (Accession# AEP68178), Oxalobacteraceae bacterium IMCC9480 (Accession# ZP_08273891), *Pyrocystis lunula* (Accession# AAO14677), *Oxyrrhis marina* (Accession# ADY17806), *Janthinobacterium* sp. Strain PAMC 25724 (Accession# ZP_10444755), *Octadecabacter antarcticus* (Accession# ZP_05063020), *Polarella glacialis* (Accession# AEF32712), *Gloeobacter violaceus* (Accession# NP_923144, xanthorhodopsin), *Salinibacter ruber* (Accession# YP_445623, xanthorhodopsin), *Glaciecola punicea* ACAM 611T (Accession# ZP_09921023), proteobacterium clone BAC_31A08 (Accession# Q9F7P4, green light-absorbing proteorhodopsin), proteobacterium clones HOT_75m4 (Accession# AAK30179, blue light-absorbing proteorhodopsin) and palE6 (Accession# AAK30200, blue light-absorbing proteorhodopsin). Alignment of was performed using MUSCLE alignment in Geneious v5.6 (Drummond et al., 2012) (http://www.geneious.com/). Seven transmembrane (A – G) helices predicted for FR1 by I-TASSER (http://zhanglab.ccmb.med.umich.edu/I-TASSER/) are indicated with magenta bars and asterisks mark functional key residues (see text).**

A whole protein alignment with bacterial and eukaryotic rhodopsins showed conserved key residues responsible for ion transport as established for BR (Figure 45). Ion transport in BR is mediated by alternating proton exchange of the retinal Schiff base (with a lysine residue at position 216, K-216) between a cytoplasmic proton donor (aspartic acid as position 96, D-96) and an extracellular proton acceptor (aspartic acid at position 85, D-85) during its photocycle (Braiman et al., 1988; Butt et al., 1989; Gerwert et al., 1990). The conserved key residues in the *Fragilariopsis* rhodopsin included acidic residues at positions of the proton donor and acceptor comprising an aspartic acid at position 121 (D-121; D-85 in BR), glutamic acid at position 132 (E-132; replaces D-96 in BR) and the retinal Schiff base at position 261 (K-261; K-216 in BR). The *in silico* modelling of the *Fragilariopsis* rhodopsin protein structure based on known protein structures (Zhang, 2008; Roy et al., 2010; Roy et al., 2012) predicted seven transmembrane helices (Figure 44). In Figure 44, key residues are highlighted including proton acceptor D-121 and donor E-132 (both black) and the K-261 (red) forming a Schiff base link with retinal.

In addition to the retinal Schiff base-forming K-261, a retinal binding pocket could be identified by mapping 18 conserved position from BR (Adamian et al., 2006) on the FR protein sequence. The identified FR retinal binding pocket consisted of Y-119, W-122, V-126, L-129, A-160, I-161, G-165, W-181, A-184, M-185, F-188, W-226, Y-229, P-230, Y-233, Y-253 and S-260 (highlighted in light grey in Figure 44). A multiple sequence alignment with structurally known pocket sequences showed that 10 residues were conserved in *F. cylindrus* in comparison to BR (Protein Data Bank ID 1C3W; Figure 46). In a phylogenetic tree, which included all different types of known rhodopsin, the *Fragilariopsis* rhodopsin clustered within the proton-pumping proteorhodopsins (Figure 47).

```
1GU8_SRII        RYW-TIVMAG-FGAFLWAYPFWPDIDTK
1H68_SRII        RYWITIVMAG-FGAF-W-YP-WPDIDTK
1GUE_SRII        RYWITIVMAG-FGAF-W-YPIWPDIDTK
1H2S_SRII        RYW-TIVMAG-FGAFLW-YPIWPDIDTK
FR               -YW-VLAIG--WAMF-W-YP-Y--Y-SK
1XIO_ASR         -YWTLQVIGAWYGVF-W-YP-W--F-SK
1E12_HR          -YW-TIMCGA-YSCF-W-YP-W--Y-AK
1VGO_AR-2        -YWTTLMIG--WSTF-W-YP-W--FDAK
1UAZ_AR-1        -YWTTLMIG--WSTM-W-YP-W--FDAK
1C3W_BR          -YW-TLMIG--WSTM-W-YP-W--F-AK
1F50_BR          -YW-TLMIG--WSTM-W-YP-W--F-AK
1M0L_BR          -YW-TLMIGGWWSTM-W-YP-W--F-AK
1KGB_BR          -YW-TLMIGGWWSTM-W-YP-W--F-AK
1M0K_BR          -YW-TLMIGGWWSTM-W-YP-W--F-AK
1QKO_BR          -YW-TLMIGGWWSTM-W-YP-W--F-AK
1M0M_BR          -YW-TLMIGG-WSTM-W-YP-W--F-AK
1P8H_BR          -YW-TLMIGG-WSTM-W-YP-W--F-AK
1O0A_BR          -YW-TLMIGG-WSTM-W-YP-W--F-AK
1P8U_BR          -YW-TLMIGG-WSTM-W-YP-W--F-AK
                  **      .  :. : * ** :    :*
```


**Figure 46. Retinal-binding pocket residues of *Fragilariopsis* rhodopsin. Multiple sequence alignment of putative FR retinal-binding pocket with sequences of known structures (Adamian et al., 2006) shown on top. A weblogo plot (http://weblogo.berkeley.edu/) highlighting conserved residues is shown at the bottom. The alignment was performed using MUSCLE (3.8) (http://www.ebi.ac.uk/Tools/msa/muscle/). Sequence names are composed of Protein Data Bank ID and protein abbreviations as follows: SRII, Sensory Rhodopsin II; ASR, *Anabaena* Sensory rhodopsin; HR, Halorhodopsin; AR-1, Archaerhodopsin-1; AR-2, Archaerhodopsin-2; BR, Bacteriorhodopsin.**

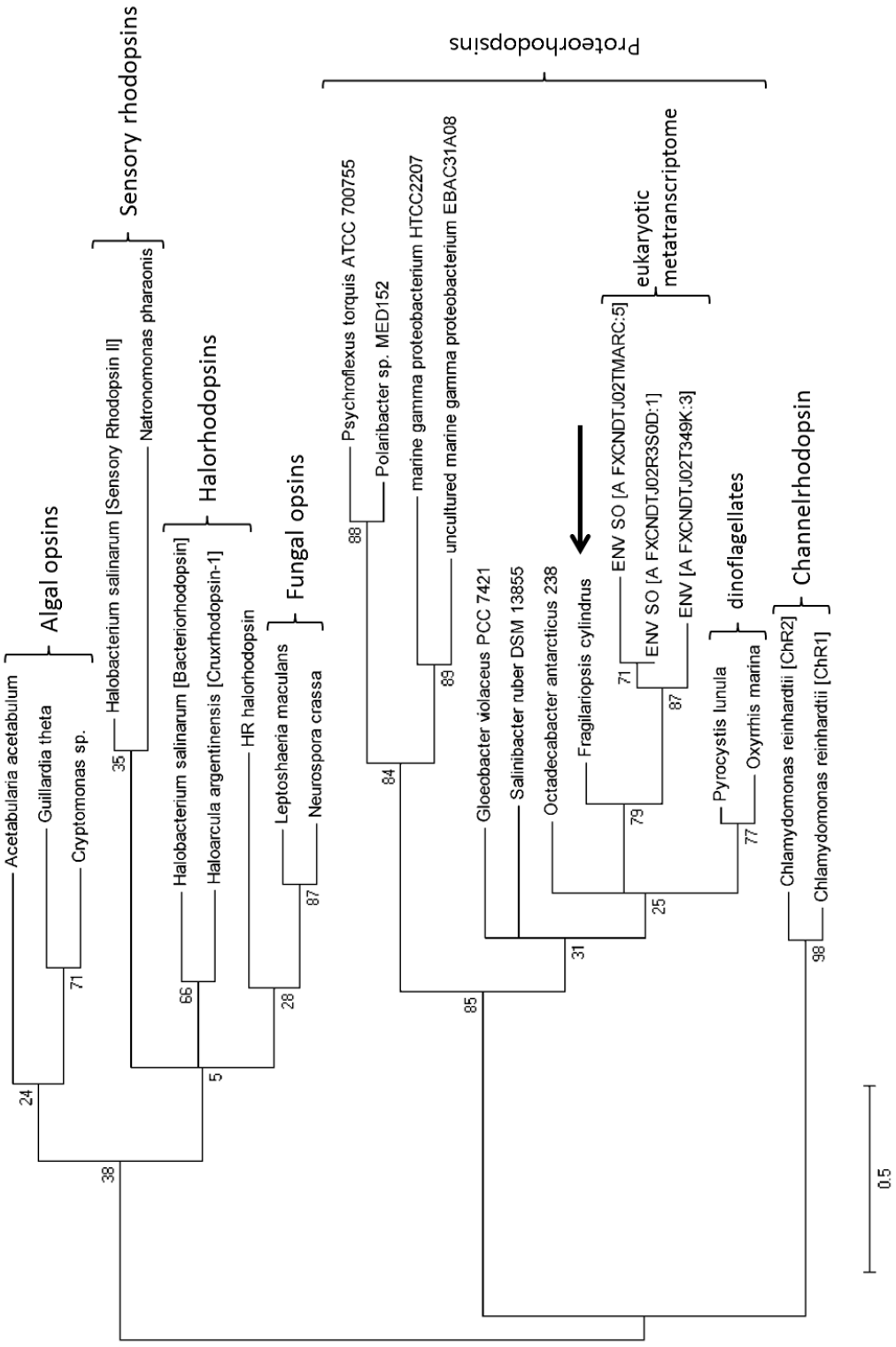



**Figure 47. Maximum likelihood phylogenetic tree of microbial type I rhodopsins with branches showing bootstrap support. Arrow marks the *Fragilariopsis* rhodopsisn.**

### 5.2.2 Gene expression analysis of *Fragilariopsis* rhodopsins gene copies

A RT-qPCR gene expression analysis was performed for both *Fragilariopsis* rhodopsin (FR) gene copies to provide insights into their physiological role by analysis of their specific responses to different experimental conditions. Therefore, *F. cylindrus* was grown at polar summer growth conditions (nutrient replete, +4 °C, 35 µmol photons $m^{-2}$ $s^{-1}$), freezing temperatures (−2 °C), elevated temperatures (+10 °C), elevated carbon dioxide (1000 ppm $CO_2$), iron starvation (−Fe), prolonged darkness (1 week darkness) (4.2.1; p. 111, Figure 21), half-saturation with silicate (0.32 µM Si) as well as red (550 – 700nm) and blue light conditions (480 – 540nm) (Figure 48) to extract RNA for cDNA synthesis.

A preliminary RT-qPCR experiment showed that genes encoding RNA Polymerase II (RNAP) and a TATA-box binding protein (TBP) were most stably expressed under all experimental conditions from a set of commonly used reference genes and used for normalisation of qPCR data. The integrated relative gene expression analysis of both *FR* gene copies showed that the *FR* gene was significantly upregulated (fixed reallocation randomisation test, $P < 0.05$) under most experimental conditions, except for elevated carbon dioxide (high $CO_2$) and red light conditions (red light; Figure 49). Conversely to all other treatments *FR* was significantly down regulated (fixed reallocation randomisation test, $P < 0.05$) and showed no significant relative expression under elevated $CO_2$ (Figure 49). In addition to relative gene expression analysis, absolute cDNA amounts were determined and a gene copy-specific RT-qPCR analysis was performed to analyse the individual contribution of both *FR* gene copies to the total absolute gene expression (Figure 50). Gene copy percentages were calculated according to percentage of gene $copy_1 = 1/(2^{\Delta Ct} + 1)$ (Germer et al., 2000). The accuracy of the approach was shown with the help of calibration mixtures with linearized plasmid DNA containing containing either cloned FR1 or FR2 (Figure 51).

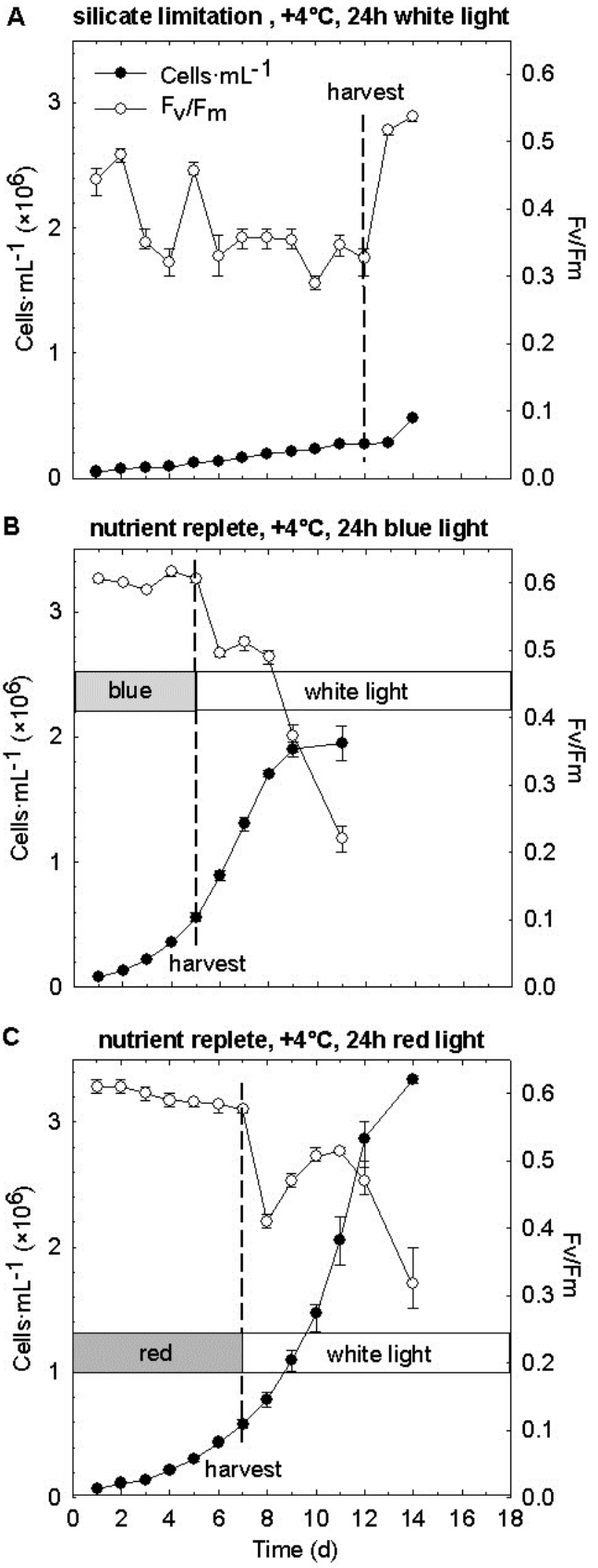



**Figure 48. Cell density and maximum PSII photochemical efficiency ($F_v/F_m$) of *F. cylindrus* grown under (A) half-saturation with silicate (0.3 µM Si, +4 °C, 35 µmol photons $m^{-2}$ $s^{-1}$), (B) blue light illumination (480 – 540 nm, +4 °C, nutrient replete), and (C) red light illumination (550 – 700 nm, +4 °C, nutrient replete).**

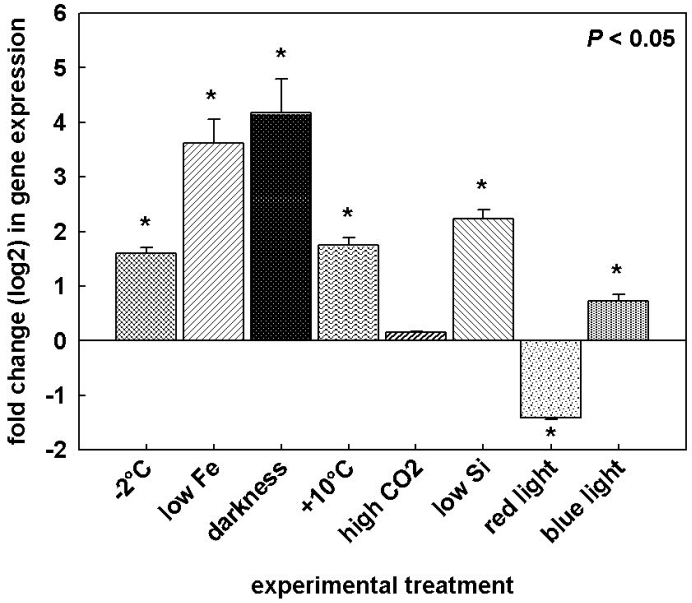



**Figure 49. RT-qPCR analysis of rhodopsin gene determined in the polar diatom *Fragilariopsis cylindrus* in different experimental treatments. Changes in expression are shown as $log_2$ of fold changes relative to *F. cylindrus* grown at reference conditions (+4 °C and white light at 35 µmol photons $m^{-2}$ $s^{-1}$, nutrient-replete). Data was normalised to the geometric mean of 2 reference genes (TBP, RNAP II) using the Relative Expression Software Tool (REST) and represents mean values and standard error from biological replicates (n = 3) and technical replicates (n = 2). Significances ($P < 0.05$) were tested using pair wise fixed reallocation randomisation test using 2000 iterations and are marked with asterisks.**

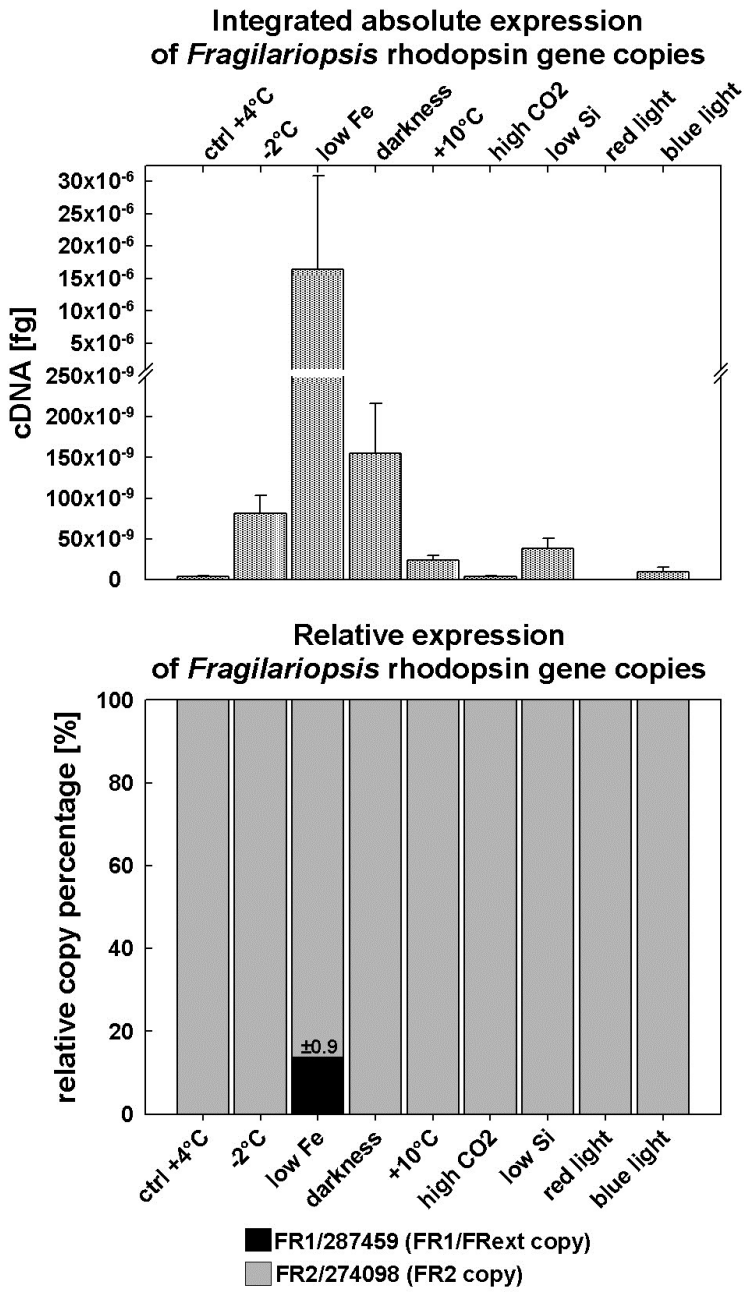



**Figure 50. Expression of of *Fragilariopsis* rhodopsin (*FR*) gene under different experimental conditions as determined by RT-qPCR. Top panel shows the integrated absolute expression of both *FR* gene copies and the bottom panel shows relative gene copy-specific expression. Mean values and standard error were calculated from biological replicates (n = 3) and technical replicates (n = 2).**

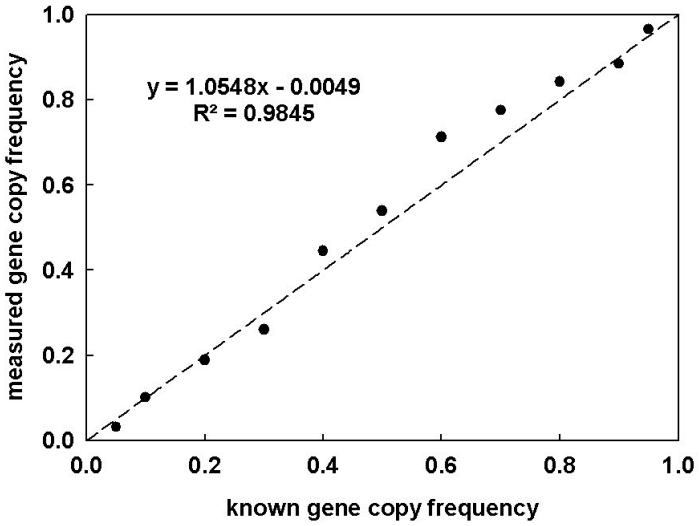



**Figure 51. The accuracy of gene copy frequency measurements by RT-qPCR. Shown is a scatterplot of gene copy frequency measurements comparing known frequencies (determined by calibration mixture with linearised plasmid DNA) with measured frequencies ($\Delta C_t$ was offset by −2 to account for inaccuracy in plasmid DNA concentration determination). The regression equation is shown and the diagonal line is the 1:1 line as expected for complete concordance between known and measured values.**

### 5.2.3 Heterologous expression of rhodopsins from *F. cylindrus* and the Antarctic dinoflagellate *Polarella glacialis*

Common key residues were identified in polar rhodopsins from *F. cylindrus* and the Antarctic dinoflagellate *P. glacialis,* suggesting that both genes encode for light-driven $H^+$-pumping proteins (5.2.1). However, the presence of key residues at positions for primary proton donor and acceptor in rhodopsins is not a general criterion for recognising their functioning as transport rhodopsins as shown in several cases of microbial rhodopsins (see discussion below; 5.3). Therefore, heterologous expression of rhodopsins from *F. cylindrus* and *P. glacialis* was performed in different host systems to provide direct experimental evidence for findings from *in silico* sequence analyses (5.2.1). Moreover, both polar rhodopsins were used for heterologous expression analysis, not only to provide insights into the functioning of rhodopsins in marine photosynthetic organisms, but also to provide insights into their role in adaptation to conditions of the Southern Ocean including low iron. Generally, obtaining milligram quantities of these proteins would allow for biochemical, biophysical and structural analyses. In a first step the full length rhodopsin genes were amplified from from cDNA and cloned. RNA for cDNA synthesis was isolated from exponentially growing cells from *F. cylindrus* and *P. glacialis* cultures. DNA sequencing confirmed the sequence of the cloned products. Interestingly, the FRext (FR1, 287459) gene copy could only be

amplified from cDNA synthesised from RNA from iron-depleted *F. cylindrus* cultures, which was in agreement with results from gene expression analysis (Figure 50). In contrast, it was possible to amplify full-length cDNA sequences of FR2 (274098) and the *Polarella* rhodopsin from nutrient-replete cell cultures. In a second step, the full length rhodopsin sequences were subcloned into the vector pGEMHE for expression in *Xenopus laevis* oocytes to perform electrophysiological measurements and to characterise their proton pumping activity (5.2.3.1). Additionally, both *FR* gene copy variants were subcloned into the diatom expression vector pPha-T1 for overexpression in *P. tricornutum* (5.2.3.2) to analyse subcellular targeting in diatoms using GFP tagging and obtain insights into their physiological role based on their specific subcellular localisation. Furthermore, *P. tricornutum*, which lacks a rhodopsin, was complemented with *Fragilariospsis* rhodopsin to perform phenotype experiments to provide insights into the physiological role of rhodopsins in marine eukaryotic phytoplankton. The heterologous expression of *Fragilariospsis* rhodopsin in the pennate diatom *P. tricornutum* was choosen, because (1) a genetic transformation method for *F. cylindrus* is not available, (2) genetic transformation technologies are most advanced in *P. tricornutum* and (3) *P. tricornutum* is more closely related to *F. cylindrus* than the centric model diatom *T. pseudonana*.

#### 5.2.3.1 Overexpression of rhodopsins from *F. cylindrus* and *P. glacialis* in *Xenopus laevis* oocytes

The work of heterologous expression of rhodopsins from *F. cylindrus* and *P. glacialis* was performed in collaboration with the group of Georg Nagel at the University of Würzburg, Germany, and in particular Shiqiang Gao and Sabrina Förster, who performed all of the expression experiments in oocytes of *Xenopus laevis* and analyses according to published procedures (Nagel et al., 1995; Nagel et al., 1998; Nagel et al., 2002). In addition, S. Gao made a significant contribution and effort to the cloning of FR1. Generally, heterologous expression in *Xenopus* oocytes was performed to carry out electrophysiological measurements *in vitro* using two-electron voltage clamp and to provide first direct experimental evidence for light-dependent proton pumping of rhodopsins from eukaryotic marine phytoplankton. Therefore, the full-length rhodopsin sequences from *F. cylindrus* and *P. glacialis* were amplified from cDNA to be subcloned into the vector pGEMHE (Supplementary Figure S2) for expression in *Xenopus laevis* oocytes. The pGEMHE is a high expression oocyte vector for *in vitro* transcription and expression in *Xenopus* oocytes and contains 3' and 5'

untranslated regions (UTRs) of a *Xenopus* β-globin gene (Liman et al., 1992). In a first step, the *Fragilariopsis* rhodopsin FR2 was expressed in *Xenopus* oocytes and plasma membrane localisation was indicated by a yellow fluorescence protein (YFP)-tag (Sabrina Förster & Georg Nagel, unpublished data; Figure 52). However, no light-driven currents could be detected (Sabrina Förster, *personal communication* 29/03/2011). Subsequently, FR1 (FRext) was cloned and expressed in oocytes. Although, YFP-tagging indicated poor plasma membrane targeting within oocytes (Figure 52), large currents (~200 nA) could be measured when FR1 protein was targeted to the plasma membrane of oocytes likely under processed cell death (Shiqiang Gao & Georg Nagel, unpuplished data; Figure 53). To provide additional evidence for light-dependent proton pumping for rhodopsins from Antarctic phytoplankton species, a rhodopsin from *P. glacialis* was used and similar results were obtained (Georg Nagel, *personal communication* 05/03/2012). As shown in Figure 53, green light seemed to induce higher photocurrents than blue light in *Fragilariopsis* rhodopsin.

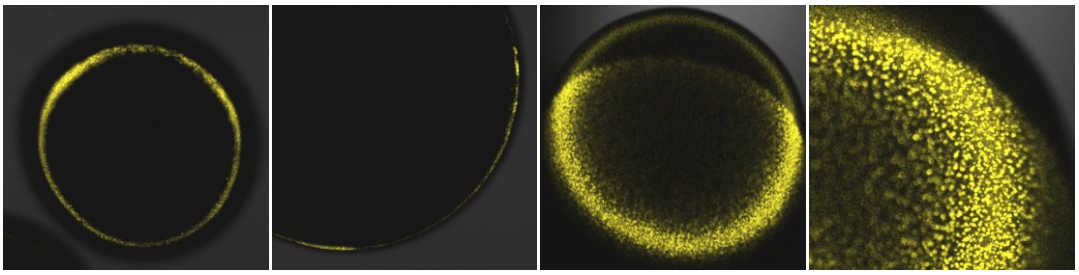


**Figure 52. Expression of *Fragilariopsis* rhodopsin in *Xenopus* oocytes. Left two panels show expression of FR2::YFP and right two panels show FR1::YFP expression. Micrographs show different level of focus (courtesy of S. Förster, S. Gao & G. Nagel).**

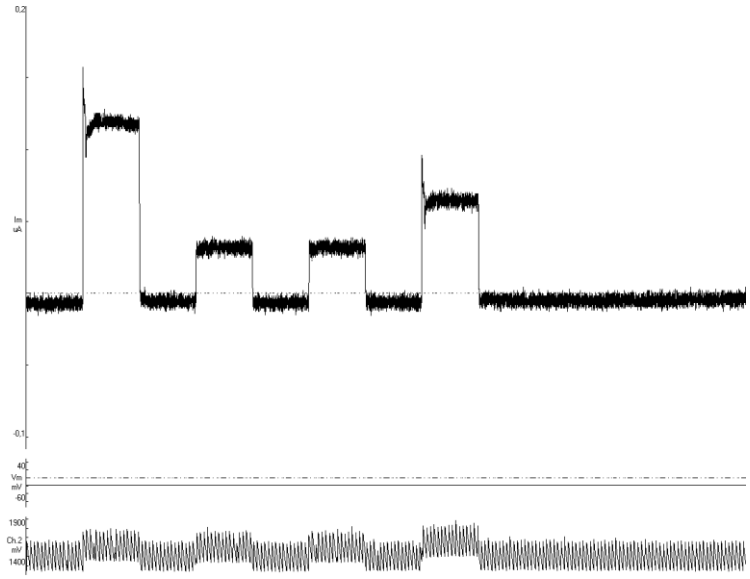


**Figure 53. *Fragilariopsis* rhodopsin FR1 (FRext) photocurrents. Example photocurrent records during exposure to a green light pulse (50% intensity), followed by two blue light pulses (100% intensity) and a final green light pulse (25% intensity). Measurements were performed 6 days post RNA injection at pH 7.6 (courtesy of S. Gao & G. Nagel).**

#### 5.2.3.2 Overexpression of *Fragilariopsis* rhodopsin in the diatom *Phaeodactylum tricornutum*

The heterologous expression of *Fragilariospsis* rhodopsin in *P. tricornutum* was performed to analyse subcellular targeting in diatoms using GFP tagging and obtain insights into its physiological role based on specific subcellular localisation. Additionally, *Fragilariopsis* rhodopsin with C-terminal hexa-histidine tag was expressed in *P. tricornutum* to perform Nickel nitrilo triacetate (Ni-NTA) immobilised metal affinity chromatography (IMAC) protein purification according to Joshi-Deo *et al.* (2010) and spectroscopic analysis of purified recombinant *Fragilariopsis* rhodopsin. Furthermore, rhodopsin-lacking *P. tricornutum* was complemented with non-tagged *Fragilariospsis* rhodopsin to perform phenotype experiments. The identification of phenotypes of *P. tricornutum* complemented with *Fragilariopsis* rhodopsin would allow direct testing of the hypothesis that phytoplankton employing rhodopsin-enabled phototrophy do have a competitive advantage in iron-limited oceans (Raven, 2009; Marchetti et al., 2012) through enhanced growth rates and reduced stress-levels under iron-limitation. Notably, as deduced from from manual gene annotations of the retinal biosynthesis pathway of *F. cylindrus* (4.2.5) and other sequenced diatoms including *P. tricornutum* and *T. pseudonana* (data not shown), diatoms appeared to contain the genetic repertoire to synthesise the chromophore retinal. Thus, the functional expression

of *Fragilariospsis* rhodopsin in *P. tricornutum* was likely without external addition of retinal. In a first step, full length sequences of both FR gene copy variants were cloned from cDNA and different expression vector constructs were generated (Figure 54) by subcloning into the *P. tricornutum* transformation vector pPha-T1 (Zaslavskaia et al., 2000) and its derivative StuI-GFP-pPhaT1 (Supplementary information), which already contains an eGFP gene for GFP-tagging (Gruber et al., 2007).

Protein targeting of both gene copy variants was studied using green fluorescent protein (GFP) tagging. The FR1:GFP fusion protein (PtFR1) was associated with chloroplasts in *P. tricornutum* (Figure 55) and similar results were obtained using a fusion construct consisting of the 49 amino acid-long N-terminal FR1 sequence fused to GFP (Figure 54). Similarly, both FR2 constructs, the PtFR2 fusion protein as well as the 59 amino acid-long N-terminal PtFR2pre, were associated with *P. tricornutum* chloroplasts (Figure 55).

Moreover, *FR*-overexpressing *P. tricornutum* mutant cell lines with high expression of C-terminal Histidin (H)-tagged FR1 and FR2 (Figure 54) for His-affinity protein purification and non-tagged FR1 and FR2 were generated for phenotype analysis of *P. tricornutm*.

```
                1        10        20        30        40        50        60
                |        |         |         |         |         |         |
   PtFR1:GFP    MLWSKTRTTFGHSFISQITFKNKKKKKAVKMISGTQFTIVYDVLSFSFA…240aa….:GFP
   PtFR1:His    MLWSKTRTTFGHSFISQITFKNKKKKKAVKMISGTQFTIVYDVLSFSFA…240aa….:HHHHHH
PtFR1pre:GFP    MLWSKTRTTFGHSFISQITFKNKKKKKAVKMISGTQFTIVYDVLSFSFA:GFP
   PtFR2:GFP    MISGTQFTIVYDVLSFSFATMMATTIFLWMRVPSVHEKYKSALIISGLVTFIASYHYLR…200aa….:GFP
   PtFR2:His    MISGTQFTIVYDVLSFSFATMMATTIFLWMRVPSVHEKYKSALIISGLVTFIASYHYLR…200aa….:HHHHHH
PtFR2pre:GFP    MISGTQFTIVYDVLSFSFATMMATTIFLWMRVPSVHEKYKSALIISGLVTFIASYHYLR:GFP
```

**Figure 54. *Fragilariopsis* rhodopsin (FR) protein sequence constructs fused to enhanced green fluorescent protein (GFP) and hexa-histidin tag (HHHHHH) for expression in *Phaeodactylum tricornutum* (Pt). FR1 denotes protein sequence ID 287459 with 289 amino acid (aa) sequence length and FR2 denotes protein sequence ID 274098 with 259 aa sequence length. N- terminal protein presequence constructs (pre) were generated for 49 aa (ProtID 287459) and 59 aa (ProtID 274098).**

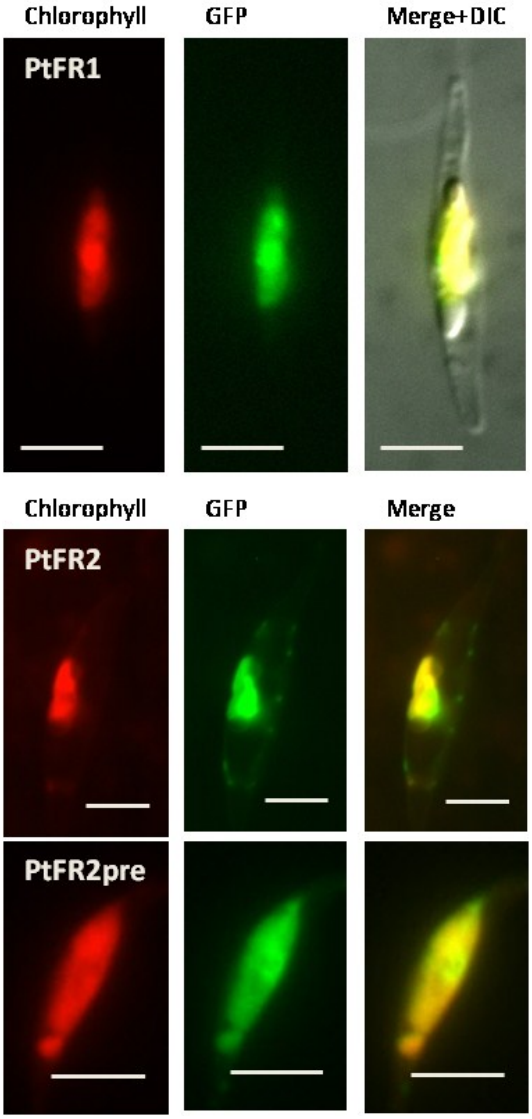



**Figure 55. Localisation of *Fragilariopsis* rhododpsin:GFP fusion proteins after expression in *Phaeodactylum tricornutum*. The top row shows expression of FR1 (FRext), the middle row shows expression of FR2 and the bottom row shows expression of N-terminal FR2. Red chlorophyll autofluorescence (Alexa 568), green GFP fluorescence and a merge of Chlorophyll and GFP fluorescence with (top row) and without (middle and bottom row) Normarski differential interference contrast (DIC) images are shown from left to right, scale bars represent 5 µm.**

## 5.3 Discussion

Two lines of evidence, deduced from sequence analysis, have suggested a role for the *F. cylindrus* rhodopsin as light-driven $H^+$-transporting protein not a sensory rhodopsin. First, the *Fragilariopsis* rhodopsin protein sequence contained key residues of transport rhodopsins. Second, neither homologous sequences of known putative transducer genes (e.g. soluble *Anabaena* sensory rhodopsin transducer, Accession# Q8YSC3; membrane-bound haloarchaeal transducers, Haloarchaeal transducer for SRI, Accession# AAG19913, Haloarchaeal transducer for SRII, Accession# AAG19989) of

sensory rhodopsins could be found in the genome of *F. cylindrus,* nor was an C-terminal extension of more than 15 residues present after the predicted end of helix G of the seven transmembrane domains in FR, which is in contrast to bulky C-terminal domains of ~400 amino acids in *Chlamydomonas* sensory rhodopsins (Spudich, 2006). In a rhodopsin from the polar dinoflagellate *Polarella glacialis* (Lin et al., 2010), similar key residues were found in common with the *Fragilariopsis* rhodopsin. Thus it was likely that both cloned rhodopsin cDNAs from *F. cylindrus* and *P. glacialis* encoded for light-driven $H^+$ proteins. However, the presence of acidic residues at positions for primary proton donor and acceptor, which are absent in most anion transporters, channelrhodopsins and sensors is not a general criteria for recognizing functioning as transport rhodopsins, because acidic residues at positions for proton donor and acceptor could also be found in sensory rhodopsins of the cryptophyte flagellate *Guillardia theta* (Sineshchekov et al., 2005), the filamentous fungus *Neurospora crassa* (Bieszke et al., 1999a) and the freshwater cyanobacterium *Anabaena* (*Nostoc*) sp. PCC7120 (Jung et al., 2003) and signal transduction remained unclear in sensory rhodopsins of cryptophytes (Sineshchekov et al., 2005).

Besides the hypothetical retinal binding amino acid residue lysine, K-261 in FR1, most amino acid protein residues forming the $H^+$ transporting hydrogen bonded network were conserved, namely, Y-87, R-118, Y-119, D-121, E-132, (E-204 of BR not conserved) and D-257, corresponding to Y-57, R-82, Y-83, D-85, D-96, E-204 and D-212 in BR (Luecke et al., 1999). Noteworthy, was the presence of M-238 and I-249 in FR1 because respective residues E-194 and E-204 are involved in extracellular $H^+$ release in BR. However, a respective E-194 residue was not conserved in the $H^+$ pumping eukaryotic rhodopsins from the green algae *Acetabularia* (Tsunoda et al., 2006) and the fungus *Leptoshaeria* (Waschuk et al., 2005), nor in the prokaryotic green and blue-light absorbing proteorhodopsins (Brown and Jung, 2006) either. Similar to FR, both prokaryotic green and blue-light absorbing proteorhodopsins also do not contain a conserved E-204, as found in BR (Brown and Jung, 2006). Furthermore, the proton-collecting antenna of BR (Checover et al., 1997; Checover et al., 2001) was also not conserved but exposed free acidic carboxylates are present at turns in the interior side of FR and together with acidic residues in the interior C-terminus could contribute to $H^+$ recruitment (Turner et al., 2009). Overall, these results support the finding that the proton pumping machinery is sturdy and simple in its nature (Brown and Jung, 2006) and evolutionary pressure to maintain proton pumping functionality has conserved key

residues corresponding to retinal binding in prokaryotic proton pumps (Ihara et al., 1999; Bielawski et al., 2004) but low conservation of the site of $H^+$ collection and release suggests some degree of plasticity.

The sequence of the retinal-binding pocket (constituting residues that interact directly or indirectly with the chromophore) may give clues about the spectral tuning of the *Fragilariopsis* rhodopsin. The *Fragilariopsis* rhodopsin contained 10 conserved residues on the 18 positions forming the retinal binding pocket (Adamian et al., 2006), some of which also contribute to the $H^+$ transporting network. Interestingly, the retinal-binding pocket contained a non-polar leucine residue, L-129 in FR1, because the respective residues L-105 in green light-absorbing proteorhodopsins (GPR) and polar glutamine Q-105 in blue light-absorbing PR (BPR) serve as spectral tuning switch (Man et al., 2003a). Similarly, a mutation at the equivalent position in BR (L-93) was also shown to alter its absorption spectrum (Subramaniam et al., 1991). However, additional residues that directly interact with the chromophore (i.e. in the retinal binding pocket) and those that cause indirect effects by localised changes in the conformation of the retinal binding pocket contribute to spectral fine-tuning in proteorhodopsin (Bielawski et al., 2004; Man-Aharonovich et al., 2004) and the microenvironment of the protonated Schiff base has been found to be a site of wavelength regulation in human rhodopsin (Kochendoerfer et al., 1999). Noteworthy, with regard to spectral fine tuning, was a conserved S-76 residue in FR1 since its corresponding S-65 in PR was shown to produce a small red shift (Man-Aharonovich et al., 2004). However, the non-conserved F-81 residue, corresponding to the red shift causing G-70 in GPR (Man-Aharonovich et al., 2004), may reflect adaptation and spectral fine tuning of the *Fragilariopsis* rhodopsin to different light conditions found in the polar environment. Notably, a tryptophan residue at position 181 (W-181) in the retinal-binding pocket of the *Fragilariopsis* rhodopsin probably impairs the binding of a carotenoid antenna as found in xanthorhodopsins from *Salinibacter ruber* (Balashov et al., 2005) and *Gloeobacter violaceus* (Imasheva et al., 2009), which allows for light-harvesting in a wider spectral range than with retinal alone. Therefore it is expected that the *Fragilariopsis* rhodopsin is likely to have a comparatively narrow spectral range.

Overall, pairwise sequence similarity of FR1 with GPR (Accession# Q9F7P4, 31.4%) was higher than for BPRs (Accessions AAK30179 and AAK30200, ~27%) and together with the L-129 switch suggest an absorption maximum in the green light spectrum as well as a fast photocycle for FR since the photocycle of GPR was found to

be an order of magnitude faster than that of BPR (Wang et al., 2003). However, the correlation between absorption spectrum and speed of the photocycle has been questioned by studies on green light-absorbing proteorhodopsins from the Arctic Ocean (Jung et al., 2008). The same authors suggest that slow photokinetics of Arctic proteorhodopsins may relate to other functions than proton pumping, such regulatory or sensory function (Wang et al., 2003; Jung, 2007) or correspond to low energy requirements due to low metabolic rates and in cold environments. Interestingly, all studied proteorhodopsins from the Arctic Ocean absorbed in the green-light spectrum (Jung et al., 2008), which is in contrast to the blue light-absorbing Antarctic proteorhodopsin clone palE6 (Beja et al., 2001) sharing the same environment with *F. cylindrus*.

Direct experimental evidence for light-driven $H^+$-pumping was provided by heterologous expression in *Xenopus* oocytes for polar phytoplankton rhodopsins from *F. cylindrus* and *Polarella glacialis*, which, to our knowledge, provides the first direct evidence for transport rhodopsins in marine eukaryotic phytoplankton. For FR1, strong positive (outward) photocurrents up to 200 nA could be measured at negative membrane potentials and were indicative of a transport rhodopsin although the targeting of FR1 to the plasma membrane of the oocytes was poor and photocurrents were only measured occasionally. In contrast, sensory channelrhodopsins from *Chlamydomonas reinhardtii* show inward currents at similar membrane potentials indicating a passive light-induced $H^+$ conductance (Nagel et al., 2002; Nagel et al., 2003). Notably, photocurrents were greater using light in the green light spectrum than in light of the blue light spectrum suggesting an absorption maximum of FR1 in the green light spectrum and light saturation probably occurs in unnatural high light intensities (Georg Nagel, *personal communication* 06/03/2012). The putative absorption maximum in green light is counterintuitive with what could be expected from natural low light conditions in the polar marine environment and the dominance of blue wavelengths in increasing water depth and under sea ice. Interestingly, pumping currents could not only be measured at room temperature, but also at reduced temperatures (+11 °C), suggesting that both polar rhodopsins function over a wide temperature range at low temperatures of their native environment. Moreover, as expected, $H^+$ pumping in the *Xenopus* oocyte system of both *Fragilariopsis* and *Polarella* polar microbial rhodopsins in the presence of all-*trans* retinal suggests it is binding to both opsins. However, the binding of other retinal analogues maybe possible as found in *Chlamydomonas* (Foster et al., 1984) and recently

the first microbial type 1 rhodopsin was discovered binding 11-*cis*-retinal similar to type 2 rhodopsins in animals (Sudo et al., 2011) and it may be possible that the *Fragilariopsis* rhodopsin (as well as the *Polarella* rhodopsin) can bind different retinal analogs *in vivo*.

Surprisingly, no $H^+$ pumping activity could be detected for FR2, although it was properly targeted to the plasma membrane of oocytes and may relate to its 30 amino acid amino-terminal truncation negatively affecting protein stability as found for the C-terminal end of bacteriorhodopsin (Turner et al., 2009). Interestingly, the different N-terminus seemed to have little effect on subcellular targeting to the chloroplast in *Phaeodactylum tricornutum* and N-terminal sequences from both FR gene copies showed similar chloroplast targeting suggesting that the N-terminus alone is sufficient for FR targeting. Although inconclusive results were obtained by *in silico* prediction of signal peptides for both FR gene copies, putative signal peptides with homology to the conserved "ASAFAP"-motif (Kilian and Kroth, 2005; Gruber et al., 2007) could be identified by manual inspection of the N-terminal protein sequence suggesting a 13 amino acid-long signal peptide for FR1 and a 17 amino acid-long signal peptide for FR2. Latter contains a lysine rich motif (KNKKKKKAVK) possibly related to lipid binding and which was also predicted to be a partial looping segment by protein structure modelling and may affecting proper membrane targeting in oocytes (Shiqiang Gao, *personal communication* 17/07/2012). Notably, for FR1 a putative N-terminal signal peptide reaches into the first transmembrane helix A as suggested by protein structure prediction and cleavage of signal- and transit-peptide may prevent $H^+$ pumping as in FR1.

Interestingly, FR2 was found predominantly expressed in most tested experimental conditions, whereas FR1 only appeared to be expressed under iron starvation (Figure 50). Thus, it may be speculated that the FR2 gene copy serves as a trace metal-independent mechanism to enhance ATP production by generating a proton gradient, when photosynthesis is iron-limited (Raven, 2009; Marchetti et al., 2012). In contrast, the functioning of the FR1 gene copy is less clear. If proven to function similar to FR2 as a light-driven proton pump, a role in adaptation to low temperature may be hypothesised, because it showed high relative and absolute expression during freezing temperatures (Figure 49, Figure 50). As it has been reported for mitochondrial membranes in plants that high content of polyunsaturated fatty acid (PUFA) can increase proton leaks (Hourton-Cabassa et al., 2009), a link to cold adaptation through

increasing PUFA content to ensure membrane fluidity (Suutari and Laakso, 1994; Chattopadhyay, 2006; Morgan-Kiss et al., 2006; Casanueva et al., 2010) may exist. In this case, FR2 in *F. cylindrus* may counteract increased proton leaks across biological membranes due to high PUFA content. In summary, both hypotheses for the physiological roles of a light-dependent rhodopsin proton pump in *F. cylindrus* under iron-limitation and low temperatures can explain an adaptive strategy to the cold and iron-limited Southern Ocean.

# Chapter 6

# General discussion

## 6.1 Summary of main results

**The *F. cylindrus* genome.** The genome size of the draft genome of the psychrophilic diatom *F. cylindrus* was found to be 80.5 Mb. General features of the genome included a high sequence polymorphism, a low G+C content of 39.8% and a high number 6913 of species-specific genes. The high sequence polymorphism prevented heterozygous haplotypes to be collapsed into a single haplotype resulting in the prediction of 27,137 genes (compared to ~10,000 in other diatoms) including allelic copy pairs from heterozygous regions of the genome and species-specific genes. The low G+C content significantly affected gene codon usage. Additionally, comparative analysis of the *F. cylindrus* genome with other sequenced diatoms revealed the expansion of gene and protein families.

**The *F. cylindrus* transcriptome.** Analysis of transcriptomes of *F. cylindrus* from six different conditions detected transcriptional activity for 95% of predicted genes. 98% of heterozygous allelic copies showed transcriptional activity and 55% of allelic copies showed > 4 fold unequal expression between copies, suggesting allele-based adaptation to different environmental conditions. Additionally, up to 30% of RNA-Seq reads mapped to unannotated regions of the genome. The most significant transcriptional changes were detected in *F. cylindrus* during prolonged darkness significantly affecting ~70% (18,856) genes.

**A bacteria-like rhodopsin in *F. cylindrus*.** A bacteria-like rhodopsin could be identified in the genome of *F. cylindrus*. Two allelic gene copies were identified showing different length of N-termini. Both allelic copies showed non-uniform bi-allelic expression under different conditions in *F. cylindrus*. As determined by RT-qPCR, one allelic copy was only expressed during iron starvation. Both allelic rhodopsin copies were cloned and the subcellular localisation using green fluorescence protein tagging in the diatom *Phaodactylum tricornutum* suggested tight associations of both allelic copies with the diatom plastid. Heterologous expression in *Xenopus* oocytes confirmed the functioning of the iron-induced allelic rhodopsin copy as fast-cycling

rhodopsin capable of light-driven proton transport. However, the physiological role of a light-dependent rhodopsin proton pump in *F. cylindrus* remains unclear.

## 6.2 Discussion

Diatoms are the most successful group of eukaryotic phytoplankton and dominate the permanently cold environment sea ice (Thomas and Dieckmann, 2002). The obligate psychrophilic pennate diatom *Fragilariopsis cylindrus* is a key stone species in the Arctic and Antarctic Ocean (Lundholm and Hasle, 2008) and can form large populations in sea ice brine and the open water column (Kang and Fryxell, 1992) serving at the basis of the polar food chain. However, little is known for adaptation of *F. cylindrus* and other polar eukaryotes to polar conditions and molecular studies to discover the molecular bases of adaptation and gene composition of sea ice algae are sparse. Furthermore, a genome sequence is lacking for an obligate psychrophilic polar eukaryote. Thus, we sequenced the genome of *F. cylindrus* to gain insights into the molecular basis for eukaryotic life below the freezing point of water. Moreover, we intended to identify genes and structural changes of DNA that are necessary to live under polar conditions by comparative analysis of the *F. cylindrus* genome with sequenced genomes from mesophilic diatoms (Armbrust et al., 2004; Bowler et al., 2008). Additionally, we used high-throughput sequencing technology to sequence the transcriptomes of *F. cylindrus* under six experimental conditions, to analyse how genomic information is used to acclimate to important environmental conditions, identify novel genes involved in polar adaptation and improve genome annotation. From our analysis we identified a bacteria-like rhodopsin proton pump highly expressed in the genome under specific environmental conditions and used it for functional analysis to get insights into its physiological role *F. cylindrus*.

The draft genome sequence of *F. cylindrus* provided novel insights into how polar environmental conditions can shape the genome of a eukaryotic extremophile organism. Strikingly, the *F. cylindrus* genome showed a high sequence polymorphism preventing heterozygous haplotypes to be collapsed into a single haplotype resulting in a diffuse haplotype structure and the prediction of 27,137 genes (compared to ~10,000 in other diatoms) including gene copy pairs from heterozygous regions of the genome. High nucleotide sequence similarities between the majorities of gene copies suggested that allelic variation contributed to the high degree of heterozygosity in *F. cylindrus*. The high degree of allelic variation may be a result of the absence of sexual

reproduction and the homologous recombination of chromosomes so that divergent alleles can remain in a heterozygous state. This hypothesis is further strengthened by the absence of meiotic core genes in the genome of *F. cylindrus*. As a result of asexual reproduction, transposable element-mediated genomic rearrangements, rare mitotic recombination and gene conversion may be the principle mechanisms that allow for the shuffling of genes and genetic variation in *F. cylindrus*. In the absence of sex, such recombinational processes are important to avoid or slowdown the accumulation of deleterious mutations (Muller, 1932; Felsenstein, 1974). Interestingly, transposable elements constituted a high proportion (7.3%) of the *F. cylindrus* genome, similar to transposable elements in the pennate diatom *P. tricornutum* (Bowler et al., 2008; Maumus et al., 2009) for which sexual reproduction has never been reported, either (Maumus et al., 2009). Furthermore, two putative transposable elements could be identified in *F. cylindrus* by analysis of novel transcribed regions using transcriptome sequencing. Transposable elements were shown to generate intraspecies diversity in plants (Morgante et al., 2005) and could have an adaptive evolutionary role. On the other hand, the high proportion of transposable elements may also be a consequence of lack of sexual reproduction in *F. cylindrus*, because in the absence of meiotic recombination, purifying selection is less able to purge the load of transposable elements (van Oosterhout, 2009). In summary, both high allelic heterozygosity and lack of core meiosis genes suggests the absence of sexual reproduction and chromosomal recombination in *F. cylindrus*.

Interestingly, we found that 55% of heterozygous alleles in *F. cylindrus* were differentially expressed by > 4 fold under the tested experimental conditions suggesting an important role of unequal bi-allelic expression. Noteworthy, the nucleotide sequence similarity between allelic gene copies did not correlate significantly with the degree of unequal bi-allelic expression as suggested for expression of gene duplications (paralogs) in the water flea *Daphnia pulex* (Colbourne et al., 2011). Additionally, the functional significance of unequal bi-allelic expression on metabolism could be shown by a separate analysis of individual allelic copy sets, which suggested a separation of metabolism between allelic gene copies. Overall, the finding of heterozygous alleles in the genome and their differential expression to different environmental stresses suggests that individual alleles are under different regulatory controls. In general, allelic expression variations are attributed to differences in noncoding DNA sequences and epigenetic regulation (Knight, 2004). In this context, it is noteworthy that nucleotide

sequence analysis of gene promoter regions of selected allelic gene copies identified a lower average sequence identity of 93.5% in comparison to an average of 97.3% sequence identity in regions downstream of the transcription start site, which was, however, not significant (two-tailed t-test, $P = 0.053$; data not shown). Nevertheless, small differences in regulatory regions may be sufficient to influence gene expression via changes in binding affinity for transcription factors, or to altered methylation patterns with influences on epigenetic regulation. Although the phenomenon of allele-specific gene expression has been widely reported in higher (Cowles et al., 2002; Enard et al., 2002; Oleksiak et al., 2002; Yan et al., 2002; Cheung et al., 2003; Lo et al., 2003; Guo et al., 2004; Schaart et al., 2005) and lower (Brem et al., 2002) eukaryotes, virtually nothing is known for eukaryotic phytoplankton. A high degree of heterozygosity paired with allele-specific differences in expression could be particularly useful for environmental adaptation if they are heritable, as reported for humans (Yan et al., 2002). Overall, the finding that heterozygous allelic pairs show condition specific expression in *F. cylindrus* is in agreement with an expression study in plants showing unequal bi-allelic expression in response to different abiotic stresses and suggesting functional diversity of allelic copies (Guo et al., 2004). In the context of functional diversification, it is interesting to note, that two allelic gene copies of a bacteria-like rhodopsin proton pump were identified in the genome of *F. cylindrus*, which showed unequal bi-allelic expression under specific environmental conditions and one allelic copy (FR1/287459) was only expressed under iron limitation. Moreover, as light-dependent proton pumping could only be detected for the iron-induced *Fragilariopsis* rhodopsin (FR) and differences in protein N-termini caused by a point mutation appeared to cause variation in subcellular targeting of the FR allelic copies, the FR allelic gene copies may serve as example of sub- or neofunctionalised allozymes. Sub- or neofunctionalised allelic copies have been reported for other asexual species (Meselson and Welch, 2007; Pouchkina-Stantcheva et al., 2007; Forche et al., 2008). Taken together, the evidence for asexual reproduction of *F. cylindrus* and the resulting increase in heterozygosity of alleles in every generation (Birky, 1996) as well as the pronounced unequal bi-allelic expression may provide an adaptive strategy to polar environments by conferring a high metabolic flexibility and capacity to adapt to a rapidly changing environment. Consequently, *F. cylindrus* may serve as a suitable model to study heterozygosity advantage (Hedrick, 2012) in the polar environment. Even more fundamental, *F. cylindrus* is a species for which the loss of sexual reproduction can have a true adaptive evolutionary advantage. When reproduction is strictly clonal, this avoids

the costs associated with the segregation load (Crow and Kimura, 1970), and this may have facilitated adaptations to extreme conditions in the polar environment.

In addition to a high degree of heterozygosity, the low G+C content of 39.8% in the genome of *F. cylindrus* may represent a structural change of DNA that is necessary to live under polar conditions. The low G+C content in *F. cylindrus* had a significant impact on codon usage causing a bias towards adenine and thymidine bases. We showed that *F. cylindrus* favours AT rich synonymous codons in comparison to genomes of the mesophilic diatoms *T. pseudonana* and *P. tricornutum*. Moreover, we found that the G+C content also affected the anticodon usage of tRNA genes. Both findings are in good agreement with the thermal adaptation hypothesis for vertebrates (Bernardi and Bernardi, 1986; Bernardi, 2000) based on higher thermal stability of G:C pairs in comparison to A:T pairs. In vertebrates, a transition from cold to warm-blooded species was accompanied by regional nucleotide changes to high G+C (Bernardi and Bernardi, 1986; Bernardi, 2000). Moreover, enrichment of G+C has been shown for tRNA and rRNA in thermophilic bacteria and A:U base pairing was prevalent in 16S rRNA in psychrophilic bacteria suggesting a thermo-adaptive mechanism (Khachane et al., 2005). Taking together, our findings may suggest a co-adaptation of relative codon-frequencies and their respective anticodons in *F. cylindrus* to optimise protein translation at low temperature, which is the energetically most expensive process in exponentially growing cells (Rocha, 2004; Wilson and Nierhaus, 2007). Consistent with the high cost of protein translation, genes involved in translation were down regulated in *F. cylindrus* during prolonged darkness relative to continuous light, likely to compensate for a loss of ATP production indicated by down regulation of ATP synthases. Strikingly, the transcriptome of *F. cylindrus* during prolonged darkness was significantly different from other environmental conditions and transcriptional changes affecting ~70% (18,856) genes. As expected, genes involved in photosynthesis, light harvesting and photoprotection were down regulated relative to continuous light. Nonetheless, observations that polar diatoms, such as *F. cylindrus*, withstand long periods of darkness (Peters and Thomas, 1996; Reeves et al., 2011) and begin rapid growth upon a return to light might suggests that they are able to synthesise chlorophyll a not only in light but also during darkness. This would imply that *F. cylindrus* (and other diatoms) might contain an alternative light-independent protochlorophyllide oxidoreductase in addition to the identified light-dependent isoenzymes (POR1/267731; POR2/188173) as hypothesised for angiosperms (Adamson et al., 1997). However, no

candidate genes were suggested by the analysis of the genome and the transcriptome of *F. cylindrus* during darkness. Interestingly, in contrast to the general down regulation of the carotenoid and harvesting pigment synthesis pathway during prolonged darkness, the synthesis of the chromophore retinal appeared to be upregulated in *F. cylindrus* in the dark, as is indicated by the > 4 fold expression change of a beta-carotene monooxygenase catalysing the cleavage of beta-carotene into retinal. Correspondingly, we detected upregulation of an allelic copy of a bacteria-like rhodopsin (FR2/274098; see discussion above). However, the physiological relevance of upregulation of a putative light-dependent rhodopsin proton pump during prolonged darkness as well as its functional role in in the presence of the proton gradient-generating chlorophyll-based photosynthetic apparatus remains speculative.

In contrast to down regulation of genes involved in light harvesting and photosynthesis, fatty acid metabolism was found highly expressed in *F. cylindrus* during prolonged darkness suggesting the oxidation of lipids to provide ATP and reduction equivalents for basal cell maintenance. The utilisation of stored lipids for metabolic intermediates and generation of ATP may explain how diatoms survive long periods of darkness (Armbrust et al., 2004) and the enrichment of genes involved in lipid metabolism in *F. cylindrus* may not only represent an adaption to maintain membrane fluidity by increased synthesis of unsaturated fatty acids at freezing temperatures but also to the long and dark winters in polar oceans, which can last up to 150 days (Lüder et al., 2002; Tang et al., 2009).

In addition to survival of polar winters, the capacity to cope with cold stress is a requirement for phytoplankton species in polar oceans. *F. cylindrus* is able to survive in sea ice brine with temperatures down to −20 °C and it has been established that it uses ice-binding proteins for this purpose (Janech et al., 2006; Krell et al., 2008; Bayer-Giraldi et al., 2010; Bayer-Giraldi et al., 2011; Raymond, 2011; Uhlig et al., 2011). However, novel was the finding that zinc-binding binding MYND protein domains are greatly expanded in the *F. cylindrus* genome and might be involved in acclimation to freezing temperatures, as suggested by the significant enrichment of genes with zinc-binding activity ($P < 0.05$) in upregulated genes at freezing temperatures (−2 °C) relative to optimal growth at +4 °C. Against the background of relatively high zinc concentrations of the Southern Ocean (Croot et al., 2011), the expansion of a Zinc protein domain may have contributed to the evolution and adaptation of *F. cylindrus* to conditions of the Southern Ocean.

## 6.3 Conclusion and future perspectives

The draft genome of the obligate psychrophilic diatom *F. cylindrus* showed at unprecedented detail how the polar environment can shape the genome of eukaryotic phytoplankton. High sequence polymorphism in the *F. cylindrus* genome affected 30% of the 27,137 predicted genes and 55% of putative heterozygous allelic copies showed pronounced unequal bi-allelic expression in response to different environmental stimuli with functional implications for cellular metabolism. This is suggested to be an adaptive strategy to the polar environment by conferring a high metabolic flexibility and capacity to adapt to a rapidly changing environment. Additionally, a low G+C content in the *F. cylindrus* genome significantly affected codon and anticodon usage and is suggested to enable efficient progression of translation at low temperatures. Moreover, high-throughput sequencing of *F. cylindrus* transcriptomes from six environmentally relevant growth conditions revealed the complexity and dynamics of polar eukaryotic transcriptomes as indicated by detection of transcriptional activity for 95% of predicted genes in *F. cylindrus* and the identification of novel transcriptionally active regions. Experimental treatment with prolonged darkness caused the most significant transcriptional changes in *F. cylindrus* and expression patterns suggest the utilisation of lipid storage products to endure darkness periods. Furthermore, the identification of expanded protein families like specific zinc-binding domains as well as specific proteins like a bacteria-like rhodopsin in the *F. cylindrus* genome, combined with their expression patterns in response to important environmental conditions (e.g., low temperatures and iron starvation), provide novel insights into polar adaptation. In summary, these results provide some answers to the initially asked questions: e.g., what genes and genomic features in the genome of a photosynthetic eukaryote are necessary to live under polar conditions? What are the molecular responses of an obligate psychrophilic eukaryote to important environmental conditions and how flexible are these responses to environmental changes? What are the physiological roles of specific key genes like a bacteria-like rhodopsin under polar conditions?

However, the information presented within the preceding chapters is still fragmentary and incomplete. It remains uncertain whether the observed genomic features of *F. cylindrus* represent common eukaryotic adaptations to polar conditions or are species-specific. Moreover, although the comparison of the *F. cylindrus* genome with other available diatom genomes from *P. tricornutum* and *T. pseudonana* provided a glimpse of the range of diatom genome sizes, it remains unclear what the corresponding

range of chromosome numbers is, because the number of chromosomes in *F. cylindrus* remains unknown due to the current draft character of its genome sequence and the presence of 5.4% sequence gaps. Does the larger genome size in *F. cylindrus* relate to a higher number of chromosomes or larger chromosomes? Moreover, although it could be shown in this work that there are significant differences in G+C content and codon usage in *F. cylindrus* compared to other diatom genomes, it remains an open question if there are significant differences in genome methylation, too. Additionally, it remains an open question what the role of allele-specific expression in eukaryotic phytoplankton is and how the observed unequal bi-allelic expression in *F. cylindrus* is regulated. Do different methylation patterns of histones control allele-specific expression as known for other eukaryotic model systems (Fournier et al., 2002)? Furthermore, although gene expression patterns in *F. cylindrus* suggest the utilisation of lipid storage products to endure prolonged darkness periods, the exact mode of dark survival during the polar winter remains uncertain. It also remains unclear how fast *F. cylindrus* can respond to long-term environmental changes in the light of possible climate change scenarios. The hypotheses on the physiological and ecological role of a light-driven rhodopsin proton pump in *F. cylindrus* remains to be tested, too. Nonetheless, taken together, the novel genomic information provided in this thesis suggests a number of starting points for experimental investigations to specifically probe adaptive strategies in *F. cylindrus*. Furthermore, it clearly shows that the use of genomic approaches provides a powerful tool to exploring basic diatom biology. Thus, more genomic studies should be performed to better understand novel aspects of diatom biology.

Future research efforts should include the sequencing of additional diatom genomes to further probe the range of genome sizes between species and investigate corresponding chromosome numbers and genome structure in diatoms. Further genome analysis of the *F. cylindrus* genome including genome mapping (i.e., the assignment of a gene to a particular region of a chromosome and determining the location of and relative distances between genes on the chromosome) to assemble a physical map (i.e., chromosomal or cytogenetic maps representing the chromosomes and providing physical distances between landmarks on individual chromosomes) will facilitate the integration of phenotypic and genetic data and provide definite evidence that heterozygous allelic pairs represent the same divergent genomic loci. The complete genome sequence of *F. cylindrus* would provide the ultimate physical map and should be realised through sequence gap closure or BAC end sequencing of its genome,

allowing for the long-range contiguity of its current draft genome sequence and the identification of chromosomes. The assembly of a physical map should be assisted by optical mapping and/or cytogenetic techniques, which will also provide a rich resource for the analysis of chromosome structures, comparative genomics and functional genomics in diatoms.

Furthermore, to allow researchers to fully exploit the novel genomic information revealed by sequencing of *F. cylindrus*, a genetic transformation system should be developed, which will open the door to advanced functional genomic investigations including the inactivation of specific genes using, e. g., antisense and sense suppression and RNA interference. Moreover, the sequencing of additional transcriptomes of *F. cylindrus* from environmentally relevant growth conditions, including hyperoxic conditions, which may occur in sea ice brine channels, as well as several months of darkness, which annually occurs during the polar winter, will provide a comprehensive resource to study metabolic processes in polar diatoms. This approach will allow determination of the exact mode of darkness survival during the polar winter and provide a more complete picture of photosynthetic life in sea ice. A better understanding of life in sea ice, including high pH and low $CO_2$ conditions, and carbon-concentrating mechanisms would also be facilitated by the study of novel protein domain combinations in *F. cylindrus* consisting of carbonic anhydrases and frustulin and fasciclin membrane domains. Additionally, a global genomic analysis of the *F. cylindrus* proteome would allow analysis for widespread amino acid modifications linked to cold adaptability of proteins and structural rigidity (e.g., reduced hydrophobic amino acid content), which has been observed in psychrophilic microorganisms (Casanueva et al., 2010). In addition to that, the combinatorial application of transcriptomics, proteomics (i.e., large-scale quantitative analysis of proteins in a cell), metabolomics (i.e., the study and quantitative analysis of all small molecules in a cell) and targeted biochemical assays will enable researchers to discover novel gene functions and provide a better understanding of the function of individual genes and metabolic processes in diatoms in response to environmental conditions.

Finally, the elucidation of the physiological role of a bacteria-like rhodopsin in *F. cylindrus* promises novel insights into the life of phytoplankton in contemporary oceans. To confirm the putative role of the *Fragilariopsis* rhodopsin in coping with iron stress, the development of a custom *Fragilariopsis* rhodopsin-specific antibody will allow for analysis of protein expression using western blotting and establish whether

increased expression levels under iron limitation are linked with higher protein levels. Moreover, the heterologous expression of *Fragilariospsis* rhodopsin in other non-rhodopsin containing diatoms using reverse genetics allows for phenotyping experiments and will allow identifying whether proton-pumping rhodopsins in eukaryotic phytoplankton lead to increased generation of ATP supporting enhanced growth under iron stress. Additionally, the further study of individual allelic gene copies of *Fragilariopsis* rhodopsin promises insights into the evolution of sub-or neofunctionalised allozymes. Furthermore, PCR-based screening of environmental metagenome samples for additional *Fragilariopsis* rhodopsin alleles (and other *F. cylindrus* alleles), combined with deep amplicon sequencing of PCR products, will allow estimation of the effective population size of *F. cylindrus* providing a rare opportunity to test whether the two allelic clusters of the same genetic locus have the same number of allelic copies, which could be hypothesised under neutral selection but would be refused if selection acts differently on both allelic subgroups because they are expressed under different conditions and have different functions. In addition to that, the identification and purification of yet unknown eukaryotic rhodopsins from environmental samples as well as isolated phytoplankton cultures will allow for preparation of a gene library for eukaryotic phytoplankton rhodopsins to probe light-driven proton pumping, characterise the influence of genetic variation on wavelength specificity, establish their significance for coping with iron stress and analyse their evolutionary history. Finally, the determination of absorption spectra of rhodopsin protein purifications from phytoplankton including *F. cylindrus* will allow probing for spectral tuning to environmentally predominant light conditions. Additionally, the direct measurement of the rhodopsin chromophore retinal in phytoplankton will allow further assessment of the functionality of eukaryotic phytoplankton rhodopsins.

In a nutshell, the genomic study of *F. cylindrus* presented in this work provides several of new avenues for exploring novel aspects of diatom biology, including high allelic heterozygosity, which may confer high metabolic flexibility and capacity of diatoms to adapt to rapidly changing environments, and promises the discovery of many more molecular secrets of diatoms.

*"We are just scratching the iceberg."*

*Thomas Mock*

# References


Abele, D., and Puntarulo, S. (2004) Formation of reactive species and induction of antioxidant defence systems in polar and temperate marine invertebrates and fish. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology* **138**: 405-415.

Adamian, L., Ouyang, Z., Tseng, Y.Y., and Liang, J. (2006) Evolutionary patterns of retinal-binding pockets of type I rhodopsins and their functions. *Photochemistry and Photobiology* **82**: 1426-1435.

Adamson, H.Y., Hiller, R.G., and Walmsley, J. (1997) Protochlorophyllide reduction and greening in angiosperms: an evolutionary perspective. *Journal of Photochemistry and Photobiology B: Biology* **41**: 201-221.

Alderkamp, A.-C., de Baar, H.J., Visser, R.J., and Arrigo, K.R. (2010) Can photoinhibition control phytoplankton abundance in deeply mixed water columns of the Southern Ocean? *Limnology and Oceanography* **55**: 1248.

Alderkamp, A.-C., Garcon, V., de Baar, H.J.W., and Arrigo, K.R. (2011) Short-term photoacclimation effects on photoinhibition of phytoplankton in the Drake Passage (Southern Ocean). *Deep Sea Research Part I: Oceanographic Research Papers* **58**: 943-955.

Alderkamp, A.-C., Kulk, G., Buma, A.G.J., Visser, R.J.W., Van Dijken, G.L., Mills, M.M., and Arrigo, K.R. (2012) The effect of iron limitation on the photophysiology of *Phaeocystis antarctica* (Prymnesiophyceae) and *Fragilariopsis cylindrus* (Bacillariophyceae) under dynamic irradiance. *Journal of Phycology* **48**: 45-59.

Allen, A.E., Vardi, A., and Bowler, C. (2006) An ecological and evolutionary context for integrated nitrogen metabolism and related signaling pathways in marine diatoms. *Current Opinion in Plant Biology* **9**: 264-273.

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

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

Amaral, P.P., Dinger, M.E., Mercer, T.R., and Mattick, J.S. (2008) The eukaryotic genome as an RNA machine. *Science* **319**: 1787-1789.

Anbar, A.D. (2008) Oceans: elements and evolution. *Science* **322**: 1481-1483.

Anbar, A.D., and Knoll, A.H. (2002) Proterozoic ocean chemistry and evolution: a bioinorganic bridge? *Science* **297**: 1137-1142.

Anderson, M.R., and Rivkin, R.B. (2001) Seasonal patterns in grazing mortality of bacterioplankton in polar oceans: a bipolar comparison. *Aquatic Microbial Ecology* **25**: 195-206.

Arkhipova, I., and Meselson, M. (2000) Transposable elements in sexual and ancient asexual taxa. *Proceedings of the National Academy of Sciences* **97**: 14473-14477.

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

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

Arrigo, K.R., and Thomas, D.N. (2004) Large scale importance of sea ice biology in the Southern Ocean. *Antarctic Science* **16**: 471-486.

Arrigo, K.R., van Dijken, G.L., and Bushinsky, S. (2008) Primary production in the Southern Ocean, 1997-2006. *Journal of Geophysical Research* **113**: C08004.

Arrigo, K.R., Mock, T., and Lizzotte, M.P. (2010a) Primary producers and sea ice. In *Sea ice*. Thomas, D.N., and Dieckmann, G.S. (eds). Chichester: Wiley-Blackwell, pp. 283-325.

Arrigo, K.R., Mills, M.M., Kropuenske, L.R., van Dijken, G.L., Alderkamp, A.-C., and Robinson, D.H. (2010b) Photophysiology in two major Southern Ocean phytoplankton taxa: photosynthesis and growth of *Phaeocystis antarctica* and *Fragilariopsis cylindrus* under different irradiance levels. *Integrative and Comparative Biology* **50**: 950-966.

Arrigo, K.R., Perovich, D.K., Pickart, R.S., Brown, Z.W., van Dijken, G.L., Lowry, K.E. et al. (2012) Massive phytoplankton blooms under Arctic sea ice. *Science* **336**: 1408.

Aslam, S.N., Cresswell-Maynard, T., Thomas, D.N., and 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. *Journal of Phycology* **48**: 1494-1509.

Atamna-Ismaeel, N., Sabehi, G., Sharon, I., Witzel, K.-P., Labrenz, M., Jurgens, K. et al. (2008) Widespread distribution of proteorhodopsins in freshwater and brackish ecosystems. *ISME Journal* **2**: 656-662.

Austin, R.W., and Petzold, T.J. (1986) Spectral dependence of the diffuse attenuation coefficient of light in ocean waters. *Optical Engineering* **25**: 253471-253471.

Ayala-del-Río, H.L., Chain, P.S., Grzymski, J.J., Ponder, M.A., Ivanova, N., Bergholz, P.W. et al. (2010) The genome sequence of *Psychrobacter arcticus* 273-4, a psychroactive Siberian permafrost bacterium, reveals mechanisms for adaptation to low-temperature growth. *Applied and Environmental Microbiology* **76**: 2304-2312.

Bailleul, B., Rogato, A., de Martino, A., Coesel, S., Cardol, P., Bowler, C. et al. (2010) An atypical member of the light-harvesting complex stress-related protein family modulates diatom responses to light. *Proceedings of the National Academy of Sciences* **107**: 18214-18219.

Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S. et al. (2005) The universal protein resource (UniProt). *Nucleic Acids Research* **33**: D154-D159.

Balashov, S.P., Imasheva, E.S., Boichenko, V.A., Anton, J., Wang, J.M., and Lanyi, J.K. (2005) Xanthorhodopsin: a proton pump with a light-harvesting carotenoid antenna. *Science* **309**: 2061-2064.

Bartsch, A. (1989) Sea ice algae of the Weddell Sea (Antarctica): Species composition, biomass, and ecophysiology of selected species. *Reports on Polar Research* **63**: 1-110.

Basak, S., and Ghosh, T.C. (2005) On the origin of genomic adaptation at high temperature for prokaryotic organisms. *Biochemical and Biophysical Research Communications* **330**: 629-632.

Basak, S., Mandal, S., and Ghosh, T.C. (2005) Correlations between genomic GC levels and optimal growth temperatures: some comments. *Biochemical and Biophysical Research Communications* **327**: 969-970.

Bathmann, U.V., Noji, T.T., and Bodungen, B.v. (1990) Copepod grazing potential in late winter in the Norwegian Sea - a factor in the control of spring phytoplankton growth? *Marine Ecology Progress Series* **60**: 225-233.

Baurain, D., Brinkmann, H., Petersen, J., Rodríguez-Ezpeleta, N., Stechmann, A., Demoulin, V. et al. (2010) Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles. *Molecular Biology and Evolution* **27**: 1698-1709.

Bayer-Giraldi, M., Weikusat, I., Besir, H., and Dieckmann, G. (2011) Characterization of an antifreeze protein from the polar diatom *Fragilariopsis cylindrus* and its relevance in sea ice. *Cryobiology* **63**: 210-219.

Bayer-Giraldi, M., Uhlig, C., John, U., Mock, T., and Valentin, K. (2010) Antifreeze proteins in polar sea ice diatoms: diversity and gene expression in the genus *Fragilariopsis*. *Environmental Microbiology* **12**: 1041-1052.

Beja, O., Spudich, E.N., Spudich, J.L., Leclerc, M., and DeLong, E.F. (2001) Proteorhodopsin phototrophy in the ocean. *Nature* **411**: 786-789.

Beja, O., Aravind, L., Koonin, E.V., Suzuki, M.T., Hadd, A., Nguyen, L.P. et al. (2000) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289**: 1902-1906.

Benizri, E., Ginouvès, A., and Berra, E. (2008) The magic of the hypoxia-signaling cascade. *Cellular and Molecular Life Sciences* **65**: 1133-1149.

Benjamini, Y., and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)* **57**: 289-300.

Bennett, M.D. (1987) Variation in genomic form in plants and its ecological implications. *New Phytologist* **106**: 177-200.

Bennetzen, J.L., and Hall, B.D. (1982) Codon selection in yeast. *Journal of Biological Chemistry* **257**: 3026-3031.

Bernardi, G. (2000) Isochores and the evolutionary genomics of vertebrates. *Gene* **241**: 3-17.

Bernardi, G., and Bernardi, G. (1986) Compositional constraints and genome evolution. *Journal of Molecular Evolution* **24**: 1-11.

Bertler, N.A.N., and Barrett, P.J. (2010) Vanishing polar ice sheets. In *Changing Climates, Earth Systems and Society*. Dodson, J. (ed): Springer Netherlands, pp. 49-83.

Bertrand, E.M., Allen, A.E., Dupont, C.L., Norden-Krichmar, T.M., Bai, J., Valas, R.E., and Saito, M.A. (2012) Influence of cobalamin scarcity on diatom molecular physiology and identification of a cobalamin acquisition protein. *Proceedings of the National Academy of Sciences* **109**: E1762–E1771.

Bertrand, E.M., Saito, M.A., Rose, J.M., Riesselman, C.R., Lohan, M.C., Noble, A.E. et al. (2007) Vitamin $B_{12}$ and iron colimitation of phytoplankton growth in the Ross Sea. *Limnology and Oceanography* **52**: 1079-1093.

Bertrand, M. (2010) Carotenoid biosynthesis in diatoms. *Photosynthesis Research* **106**: 89-102.

Bharanidharan, D., Ramya Bhargavi, G., Uthanumallian, K., and Gautham, N. (2004) Correlations between nucleotide frequencies and amino acid composition in 115 bacterial species. *Biochemical and Biophysical Research Communications* **315**: 1097-1103.

Bidle, K.D., and Falkowski, P.G. (2004) Cell death in planktonic, photosynthetic microorganisms. *Nature Reviews Microbiology* **2**: 643-655.

Bidle, K.D., and Bender, S.J. (2008) Iron starvation and culture age activate metacaspases and programmed cell death in the marine diatom *Thalassiosira pseudonana*. *Eukaryotic Cell* **7**: 223-236.

Bielawski, J.P., Dunn, K.A., Sabehi, G., and Beja, O. (2004) Darwinian adaptation of proteorhodopsin to different light intensities in the marine environment. *Proceedings of the National Academy of Sciences* **101**: 14824-14829.

Bieszke, J.A., Spudich, E.N., Scott, K.L., Borkovich, K.A., and Spudich, J.L. (1999a) A eukaryotic protein, NOP-1, binds retinal to form an archaeal rhodopsin-like photochemically reactive pigment. *Biochemistry* **38**: 14138-14145.

Bieszke, J.A., Braun, E.L., Bean, L.E., Kang, S., Natvig, D.O., and Borkovich, K.A. (1999b) The nop-1 gene of *Neurospora crassa* encodes a seven transmembrane helix retinal-binding protein homologous to archaeal rhodopsins. *Proceedings of the National Academy of Sciences* **96**: 8034-8039.

Bimboim, H.C., and Doly, J. (1979) A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Research* **7**: 1513-1523.

Birky, C.W. (1996) Heterozygosity, heteromorphy, and phylogenetic trees in asexual eukaryotes. *Genetics* **144**: 427-437.

Birney, E., Clamp, M., and Durbin, R. (2004) GeneWise and Genomewise. *Genome Research* **14**: 988-995.

Biro, J.C., Benyó, B., Sansom, C., Szlávecz, Á., Fördös, G., Micsik, T., and Benyó, Z. (2003) A common periodic table of codons and amino acids. *Biochemical and Biophysical Research Communications* **306**: 408-415.

Birol, I., Jackman, S.D., Nielsen, C.B., Qian, J.Q., Varhol, R., Stazyk, G. et al. (2009) De novo transcriptome assembly with ABySS. *Bioinformatics* **25**: 2872-2877.

Birzele, F., Schaub, J., Rust, W., Clemens, C., Baum, P., Kaufmann, H. et al. (2010) Into the unknown: expression profiling without genome sequence information in CHO by next generation sequencing. *Nucleic Acids Research* **38**: 3999-4010.

Bishop, D.K., Park, D., Xu, L., and Kleckner, N. (1992) DMC1: A meiosis-specific yeast homolog of *E. coli* recA required for recombination, synaptonemal complex formation, and cell cycle progression. *Cell* **69**: 439-456.

Blanc, G., Agarkova, I., Grimwood, J., Kuo, A., Brueggeman, A., Dunigan, D. et al. (2012) The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. *Genome Biology*.

Boetius, A., and Joye, S. (2009) Thriving in salt. *Science* **324**: 1523-1525.

Boone, C., Bussey, H., and Andrews, B.J. (2007) Exploring genetic interactions and networks with yeast. *Nature Reviews Genetics* **8**: 437-449.

Boroujerdi, A.B., Lee, P., DiTullio, G., Janech, M., Vied, S., and Bearden, D. (2012) Identification of isethionic acid and other small molecule metabolites of *Fragilariopsis cylindrus* with nuclear magnetic resonance. *Analytical and Bioanalytical Chemistry* **404**: 777-784.

Bowler, C., Vardi, A., and Allen, A.E. (2010) Oceanographic and biogeochemical insights from diatom genomes. *Annual Review of Marine Science* **2**: 333-365.

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

Bown, J., Boye, M., and Nelson, D.M. (2012) New insights on the role of organic speciation in the biogeochemical cycle of dissolved cobalt in the southeastern Atlantic and the Southern Ocean. *Biogeosciences* **9**: 2719-2736.

Bown, J., Boye, M., Baker, A., Duvieilbourg, E., Lacan, F., Le Moigne, F. et al. (2011) The biogeochemical cycle of dissolved cobalt in the Atlantic and the Southern Ocean south off the coast of South Africa. *Marine Chemistry* **126**: 193-206.

Boyd, P.W., Crossley, A.C., DiTullio, G.R., Griffiths, F.B., Hutchins, D.A., Queguiner, B. et al. (2001) Control of phytoplankton growth by iron supply and irradiance in the subantarctic Southern Ocean: experimental results from the SAZ Project. *Journal of Geophysical Research: Oceans* **106**: 31573-31583.

Boyd, P.W., Jickells, T., Law, C.S., Blain, S., Boyle, E.A., Buesseler, K.O. et al. (2007) Mesoscale iron enrichment experiments 1993-2005: synthesis and future directions. *Science* **315**: 612-617.

Braiman, M.S., Mogi, T., Marti, T., Stern, L.J., Khorana, H.G., and Rothschild, K.J. (1988) Vibrational spectroscopy of bacteriorhodopsin mutants: light-driven proton transport involves protonation changes of aspartic acid residues 85, 96, and 212. *Biochemistry* **27**: 8516-8520.

Brem, R.B., Yvert, G., Clinton, R., and Kruglyak, L. (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**: 752-755.

Britten, R.J., Cetta, A., and Davidson, E.H. (1978) The single-copy DNA sequence polymorphism of the sea urchin *Strongylocentrotus purpuratus*. *Cell* **15**: 1175-1186.

Brown, L.S., and Jung, K.-H. (2006) Bacteriorhodopsin-like proteins of eubacteria and fungi: the extent of conservation of the haloarchaeal proton-pumping mechanism. *Photochemical & Photobiological Sciences* **5**: 538-546.

Bruland, K.W. (1980) Oceanographic distributions of cadmium, zinc, nickel, and copper in the North Pacific. *Earth and Planetary Science Letters* **47**: 176-198.

Brune, W.H., Anderson, J.G., Toohey, D.W., Fahey, D.W., Kawa, S.R., Jones, R.L. et al. (1991) The potential for ozone depletion in the Arctic polar stratosphere. *Science* **252**: 1260-1266.

Brussaard, C.P.D. (2004) Viral control of phytoplankton populations - a review. *Journal of Eukaryotic Microbiology* **51**: 125-138.

Brussaard, C.P.D., Timmermans, K.R., Uitz, J., and Veldhuis, M.J.W. (2008) Virioplankton dynamics and virally induced phytoplankton lysis versus microzooplankton grazing southeast of the Kerguelen (Southern Ocean). *Deep Sea Research Part II: Topical Studies in Oceanography* **55**: 752-765.

Bugreev, D.V., Pezza, R.J., Mazina, O.M., Voloshin, O.N., Camerini-Otero, R.D., and Mazin, A.V. (2011) The resistance of DMC1 D-loops to dissociation may account for the DMC1 requirement in meiosis. *Nature Structural and Molecular Biology* **18**: 56-61.

Bustin, S.A. (2000) Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *Journal of Molecular Endocrinology* **25**: 169-193.

Butt, H.J., Fendler, K., Bamberg, E., Tittor, J., and Oesterhelt, D. (1989) Aspartic acids 96 and 85 play a central role in the function of bacteriorhodopsin as a proton pump. *EMBO Journal* **8**: 1657-1663.

Calbet, A., and Landry, M.R. (2004) Phytoplankton growth, microzooplankton grazing, and carbon cycling in marine systems. *Limnology and Oceanography* **49**: 51-57.

Campbell, B.J., Waidner, L.A., Cottrell, M.T., and Kirchman, D.L. (2008) Abundant proteorhodopsin genes in the North Atlantic Ocean. *Environmental Microbiology* **10**: 99-109.

Casanueva, A., Tuffin, M., Cary, C., and Cowan, D.A. (2010) Molecular adaptations to psychrophily: the impact of 'omic' technologies. *Trends in Microbiology* **18**: 374-381.

Cefarelli, A., Ferrario, M., Almandoz, G., Atencio, A., Akselman, R., and Vernet, M. (2010) Diversity of the diatom genus *Fragilariopsis* in the Argentine Sea and Antarctic waters: morphology, distribution and abundance. *Polar Biology* **33**: 1463-1484.

Chattopadhyay, M. (2006) Mechanism of bacterial adaptation to low temperature. *Journal of Biosciences* **31**: 157-165.

Chattopadhyay, M.C., and Jagannadham, M.J. (2001) Maintenance of membrane fluidity in Antarctic bacteria. *Polar Biology* **24**: 386-388.

Checover, S., Nachliel, E., Dencher, N.A., and Gutman, M. (1997) Mechanism of proton entry into the cytoplasmic section of the proton-conducting channel of bacteriorhodopsin. *Biochemistry* **36**: 13919-13928.

Checover, S., Marantz, Y., Nachliel, E., Gutman, M., Pfeiffer, M., Tittor, J. et al. (2001) Dynamics of the proton transfer reaction on the cytoplasmic surface of bacteriorhodopsin. *Biochemistry* **40**: 4281-4292.

Chen, S.L., Lee, W., Hottes, A.K., Shapiro, L., and McAdams, H.H. (2004) Codon usage between genomes is constrained by genome-wide mutational processes. *Proceedings of the National Academy of Sciences* **101**: 3480-3485.

Chen, Z., Cheng, C.-H.C., Zhang, J., Cao, L., Chen, L., Zhou, L. et al. (2008) Transcriptomic and genomic evolution under constant cold in Antarctic notothenioid fish. *Proceedings of the National Academy of Sciences* **105**: 12944-12949.

Cheung, J., and Hendrickson, W.A. (2010) Sensor domains of two-component regulatory systems. *Current Opinion in Microbiology* **13**: 116-123.

Cheung, V.G., Conlin, L.K., Weber, T.M., Arcaro, M., Jen, K.-Y., Morley, M., and Spielman, R.S. (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. *Nature Genetics* **33**: 422-425.

Chiusano, M.L., Alvarez-Valin, F., Di Giulio, M., D'Onofrio, G., Ammirato, G., Colonna, G., and Bernardi, G. (2000) Second codon positions of genes and the secondary structures of proteins. Relationships and implications for the origin of the genetic code. *Gene* **261**: 63-69.

Chomczynski, P., and Sacchi, N. (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Analytical Biochemistry* **162**: 156-159.

Clark, M.S., Thorne, M.A.S., Toullec, J.-Y., Meng, Y., Guan, L.L., Peck, L.S., and Moore, S. (2011) Antarctic krill 454 pyrosequencing reveals chaperone and stress transcriptome. *PLoS ONE* **6**: e15919.

Cock, J.M., Sterck, L., Rouze, P., Scornet, D., Allen, A.E., Amoutzias, G. et al. (2010) The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. *Nature* **465**: 617-621.

Coesel, S., Oborník, M., Varela, J., Falciatore, A., and Bowler, C. (2008) Evolutionary origins and functions of the carotenoid biosynthetic pathway in marine diatoms. *PLoS ONE* **3**: e2896.

Colbourne, J.K., Pfrender, M.E., Gilbert, D., Thomas, W.K., Tucker, A., Oakley, T.H. et al. (2011) The ecoresponsive genome of *Daphnia pulex*. *Science* **331**: 555-561.

Collins, R.E., and Deming, J.W. (2011) Abundant dissolved genetic material in Arctic sea ice Part II: viral dynamics during autumn freeze-up. *Polar Biology* **34**: 1831-1841.

Conover, R.J., and Huntley, M. (1991) Copepods in ice-covered seas—Distribution, adaptations to seasonally limited food, metabolism, growth patterns and life cycle strategies in polar seas. *Journal of Marine Systems* **2**: 1-41.

Copley, S.D., Smith, E., and Morowitz, H.J. (2005) A mechanism for the association of amino acids with their codons and the origin of the genetic code. *Proceedings of the National Academy of Sciences* **102**: 4442-4447.

Cota, G.F. (1985) Photoadaptation of high Arctic ice algae. *Nature* **315**: 219-222.

Coulondre, C., Miller, J.H., Farabaugh, P.J., and Gilbert, W. (1978) Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature* **274**: 775-780.

Cowles, C.R., Hirschhorn, J.N., Altshuler, D., and Lander, E.S. (2002) Detection of regulatory variation in mouse genes. *Nature Genetics* **32**: 432-437.

Crame, J.A. (1997) An evolutionary framework for the polar regions. *Journal of Biogeography* **24**: 1-9.

Croft, M.T., Warren, M.J., and Smith, A.G. (2006) Algae need their vitamins. *Eukaryotic Cell* **5**: 1175-1183.

Croft, M.T., Lawrence, A.D., Raux-Deery, E., Warren, M.J., and Smith, A.G. (2005) Algae acquire vitamin $B_{12}$ through a symbiotic relationship with bacteria. *Nature* **438**: 90-93.

Croot, P.L., Baars, O., and Streu, P. (2011) The distribution of dissolved zinc in the Atlantic sector of the Southern Ocean. *Deep Sea Research Part II: Topical Studies in Oceanography* **58**: 2707-2719.

Crow, J.F., and Kimura, M. (1970) *An introduction to population genetics theory*. New York: Harper & Row.

Cullen, J.J. (1991) Hypotheses to explain high-nutrient conditions in the open sea. *Limnology and Oceanography* **36**: 1578-1599.

Cullen, J.J., Neale, P.J., and Lesser, M.P. (1992) Biological weighting function for the inhibition of phytoplankton photosynthesis by ultraviolet radiation. *Science* **258**: 646-650.

D'Amico, S., Collins, T., Marx, J.-C., Feller, G., and Gerday, C. (2006) Psychrophilic microorganisms: challenges for life. *EMBO Reports* **7**: 385-389.

D'Hont, A., Denoeud, F., Aury, J.-M., Baurens, F.-C., Carreel, F., Garsmeur, O. et al. (2012) The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* **488**: 213-217.

Danon, A., and Stoeckenius, W. (1974) Photophosphorylation in *Halobacterium halobium*. *Proceedings of the National Academy of Sciences* **71**: 1234-1238.

Dassanayake, M., Oh, D.-h., Hong, H., Bohnert, H.J., and Cheeseman, J.M. (2011a) Transcription strength and halophytic lifestyle. *Trends in Plant Science* **16**: 1-3.

Dassanayake, M., Oh, D.-H., Haas, J.S., Hernandez, A., Hong, H., Ali, S. et al. (2011b) The genome of the extremophile crucifer *Thellungiella parvula*. *Nature Genetics* **43**: 913-918.

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L. et al. (2006) A high-resolution map of transcription in the yeast genome. *Proceedings of the National Academy of Sciences* **103**: 5320-5325.

de Baar, H.J.W., Boyd, P.W., Coale, K.H., Landry, M.R., Tsuda, A., Assmy, P. et al. (2005) Synthesis of iron fertilization experiments: from the iron age in the age of enlightenment. *Journal of Geophysical Research* **110**.

De La Rocha, C.L., Hutchins, D.A., Brzezinski, M.A., and Zhang, Y. (2000) Effects of iron and zinc deficiency on elemental composition and silica production by diatoms. *Marine Ecology Progress Series* **195**: 71-79.

de la Torre, J.R., Christianson, L.M., Beja , O., Suzuki, M.T., Karl, D.M., Heidelberg, J., and DeLong, E.F. (2003) Proteorhodopsin genes are distributed among divergent marine bacterial taxa. *Proceedings of the National Academy of Sciences* **100**: 12830-12835.

de Santana, C., Rozenfeld, A., Marquet, P., and Duarte, C. (2013) Topological properties of polar food webs. *Marine Ecology Progress Series* **474**: 15-26.

Deacon, G.E.R. (1982) Physical and biological zonation in the Southern Ocean. *Deep Sea Research Part A Oceanographic Research Papers* **29**: 1-15.

DeConto, R., Pollard, D., and Harwood, D. (2007) Sea ice feedback and Cenozoic evolution of Antarctic climate and ice sheets. *Paleoceanography* **22**: PA3214.

DeConto, R.M., and Pollard, D. (2003) Rapid Cenozoic glaciation of Antarctica induced by declining atmospheric $CO_2$. *Nature* **421**: 245-249.

DeConto, R.M., Pollard, D., Wilson, P.A., Palike, H., Lear, C.H., and Pagani, M. (2008) Thresholds for Cenozoic bipolar glaciation. *Nature* **455**: 652-656.

Dehal, P., Satou, Y., Campbell, R.K., Chapman, J., Degnan, B., De Tomaso, A. et al. (2002) The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* **298**: 2157-2167.

Denver, D.R., Dolan, P.C., Wilhelm, L.J., Sung, W., Lucas-Lledó, J.I., Howe, D.K. et al. (2009) A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. *Proceedings of the National Academy of Sciences* **106**: 16310-16314.

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

Derelle, E., Ferraz, C., Rombauts, S., RouzÃ©, P., Worden, A.Z., Robbens, S. et al. (2006) Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proceedings of the National Academy of Sciences* **103**: 11647-11652.

Deschamps, P., and Moreira, D. (2012) Reevaluating the green contribution to diatom genomes. *Genome Biology and Evolution* **4**: 683-688.

Doyle, V., Virji, S., and Crompton, M. (1999) Evidence that cyclophilin-A protects cells against oxidative stress. *Biochemical Journal* **341 ( Pt 1)**: 127-132.

Drummond, A., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Duran, C. et al. (2012). Geneious v5.6. URL http://www.geneious.com

Dubischar, C.D., and Bathmann, U.V. (1997) Grazing impact of copepods and salps on phytoplankton in the Atlantic sector of the Southern ocean. *Deep Sea Research Part II: Topical Studies in Oceanography* **44**: 415-433.

Duncan, B.K., and Miller, J.H. (1980) Mutagenic deamination of cytosine residues in DNA. *Nature* **287**: 560-561.

Dupont, C.L., Yang, S., Palenik, B., and Bourne, P.E. (2006) Modern proteomes contain putative imprints of ancient shifts in trace metal geochemistry. *Proceedings of the National Academy of Sciences* **103**: 17822-17827.

Dupont, C.L., Butcher, A., Valas, R.E., Bourne, P.E., and Caetano-Anollés, G. (2010) History of biological metal utilization inferred through phylogenomic analysis of protein structures. *Proceedings of the National Academy of Sciences* **107**: 10567-10572.

Eberhard, S., Finazzi, G., and Wollman, F.-A. (2008) The dynamics of photosynthesis. *Annual Review of Genetics* **42**: 463-515.

Edwards, R., and Sedwick, P. (2001) Iron in East Antarctic snow: implications for atmospheric iron deposition and algal production in Antarctic waters. *Geophysical Research Letters* **28**: 3907-3910.

Eicken, H. (1992) The role of sea ice in structuring Antarctic ecosystems. *Polar Biology* **12**: 3-13.

Emanuelsson, O., Nielsen, H., and Heijne, G.V. (1999) ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Science* **8**: 978-984.

Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of Molecular Biology* **300**: 1005-1016.

Emanuelsson, O., Brunak, S., von Heijne, G., and Nielsen, H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protocols* **2**: 953-971.

Enard, W., Khaitovich, P., Klose, J., Zöllner, S., Heissig, F., Giavalisco, P. et al. (2002) Intra- and interspecific variation in primate gene expression patterns. *Science* **296**: 340-343.

Engen, Ø., Faleide, J.I., and Dyreng, T.K. (2008) Opening of the Fram Strait gateway: a review of plate tectonic constraints. *Tectonophysics* **450**: 51-69.

Evans, C., and Brussaard, C.P.D. (2012) Viral lysis and microzooplankton grazing of phytoplankton throughout the Southern Ocean. *Limnology and Oceanography* **57**: 1826-1837.

Exon, N., Kennett, J., Malone, M., Leg 189 Scientific Party, and Nürnberg, D. (2002) The opening of the Tasmanian Gateway drove global Cenozoic paleoclimatic and paleoceanographic changes: Results of Leg 189. *Joides Journal* **26**: 11-18.

Fahrbach, E., Beszcynska-Moeller, A., and Rohardt, G. (2009) Polar Oceans - an oceanographic overview. In *Biological studies in polar oceans Exploration of life in icy waters*. Hempel, G., and Hempel, I. (eds). Bremerhaven: Wissenschaftsverlag NW, pp. 17-36.

Falkowski, P.G., Barber, R.T., and Smetacek, V. (1998) Biogeochemical controls and feedbacks on ocean primary production. *Science* **281**: 200-206.

Falkowski, P.G., Katz, M.E., Knoll, A.H., Quigg, A., Raven, J.A., Schofield, O., and Taylor, F.J.R. (2004) The evolution of modern eukaryotic phytoplankton. *Science* **305**: 354-360.

Felsenstein, J. (1974) The evolutionary advantage of recombination. *Genetics* **78**: 737-756.

Ferroni, L., Baldisserotto, C., Zennaro, V., Soldani, C., Fasulo, M.P., and Pancaldi, S. (2007) Acclimation to darkness in the marine chlorophyte *Koliella antarctica* cultured under low salinity: hypotheses on its origin in the polar environment. *European Journal of Phycology* **42**: 91-104.

Fiala, M., and Oriol, L. (1990) Light-temperature interactions on the growth of Antarctic diatoms. *Polar Biology* **10**: 629-636.

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

Fitzwater, S.E., Knauer, G.A., and Martin, J.H. (1982) Metal contamination and its effect on primary production measurements. *Limnology and Oceanography* **27**: 544-551.
Fitzwater, S.E., Johnson, K.S., Gordon, R.M., Coale, K.H., and Smith Jr, W.O. (2000) Trace metal concentrations in the Ross Sea and their relationship with nutrients and phytoplankton growth. *Deep Sea Research Part II: Topical Studies in Oceanography* **47**: 3159-3179.
Foerstner, K.U., von Mering, C., Hooper, S.D., and Bork, P. (2005) Environments shape the nucleotide composition of genomes. *EMBO Reports* **6**: 1208-1213.
Fogg, G.E. (1991) The phytoplanktonic ways of life. *New Phytologist* **118**: 191-232.
Forche, A., Alby, K., Schaefer, D., Johnson, A.D., Berman, J., and Bennett, R.J. (2008) The parasexual cycle in *Candida albicans* provides an alternative pathway to meiosis for the formation of recombinant strains. *PLoS Biology* **6**: e110.
Foster, K.W., Saranak, J., Patel, N., Zarilli, G., Okabe, M., Kline, T., and Nakanishi, K. (1984) A rhodopsin is the functional photoreceptor for phototaxis in the unicellular eukaryote *Chlamydomonas*. *Nature* **311**: 756-759.
Fournier, C., Goto, Y., Ballestar, E., Delaval, K., Hever, A.M., Esteller, M., and Feil, R. (2002) Allele-specific histone lysine methylation marks regulatory regions at imprinted mouse genes. *EMBO Journal* **21**: 6560-6570.
Francis, D., Davies, M.S., and Barlow, P.W. (2008) A strong nucleotypic effect on the cell cycle regardless of ploidy level. *Annals of Botany* **101**: 747-757.
Frank, K.T., Petrie, B., Choi, J.S., and Leggett, W.C. (2005) Trophic cascades in a formerly cod-dominated ecosystem. *Science* **308**: 1621-1623.
Franklin, D.J., Brussaard, C.P.D., and Berges, J.A. (2006) What is the role and nature of programmed cell death in phytoplankton ecology? *European Journal of Phycology* **41**: 1-14.
Franklin, D.J., Airs, R.L., Fernandes, M., Bell, T.G., Bongaerts, R.J., Berges, J.A., and Malin, G. (2012) Identification of senescence and death in *Emiliania huxleyi* and *Thalassiosira pseudonana*: cell staining, chlorophyll alterations, and dimethylsulfoniopropionate (DMSP) metabolism. *Limnology and Oceanography* **57**: 305-317.
Frias-Lopez, J., Shi, Y., Tyson, G.W., Coleman, M.L., Schuster, S.C., Chisholm, S.W., and DeLong, E.F. (2008) Microbial community gene expression in ocean surface waters. *Proceedings of the National Academy of Sciences* **105**: 3805-3810.
Friedberg, E.C., Wagner, R., and Radman, M. (2002) Specialized DNA polymerases, cellular survival, and the genesis of mutations. *Science* **296**: 1627-1630.
Friedberg, E.C., Walker, G.C., Siede, W., Wood, R.D., Schultz, R.A., and Ellenberger, T. (2006) *DNA repair and mutagenesis*. Washington, DC: American Society for Microbiology (ASM) Press.
Friedl, T. (1995) Inferring taxonomic positions and testing genus level assignments in coccoid green lichen algae: a phylogenetic analysis of 18S ribosomal RNA sequences from *Dictyochloropsis reticulata* and from members of the genus *Myrmecia* (Chlorophyta, Trebouxiophyceae cl. nov.). *Journal of Phycology* **31**: 632-639.
Frigaard, N.-U., Martinez, A., Mincer, T.J., and DeLong, E.F. (2006) Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. *Nature* **439**: 847-850.
Frommolt, R., Werner, S., Paulsen, H., Goss, R., Wilhelm, C., Zauner, S. et al. (2008) Ancient recruitment by chromists of green algal genes encoding enzymes for carotenoid biosynthesis. *Molecular Biology and Evolution* **25**: 2653-2667.
Frost, B.W. (1991) The role of grazing in nutrient-rich areas of the open sea. *Limnology and Oceanography* **36**: 1616-1630.
Fryxell, G.A., and Prasad, A.K.S.K. (1990) *Eucampia antarctica* var. *recta* (Mangin) stat. nov. (Biddulphiaceae, Bacillariophyceae): life stages at the Weddell Sea ice edge. *Phycologia* **29**: 27-38.
Fuhrman, J.A., Schwalbach, M.S., and Stingl, U. (2008) Proteorhodopsins: an array of physiological roles? *Nature Reviews Microbiology* **6**: 488-494.
Galtier, N., and Lobry, J.R. (1997) Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *Journal of Molecular Evolution* **44**: 632-636.
Garcia, H.E., Locarnini, R.A., Boyer, T.P., Antonov, J.I., Zweng, M.M., Baranova, O.K., and Johnson, D.R. (2010) World Ocean Atlas 2009, Volume 4: Nutrients (phosphate, nitrate,

silicate). In *NOAA Atlas NESDIS 71*. Levitus, S. (ed). Washington, D.C.: U.S. Government Printing Office, p. 398.
Garrison, D.L., and Buck, K.R. (1989) The biota of Antarctic pack ice in the Weddell sea and Antarctic Peninsula regions. *Polar Biology* **10**: 211-219.
Geider, R.J., Laroche, J., Greene, R.M., and Olaizola, M. (1993) Response of the photosynthetic apparatus of *Phaeodactylum tricornutum* (Bacillariophyceae) to nitrate, phosphate, or iron starvation. *Journal of Phycology* **29**: 755-766.
Germer, S., Holland, M.J., and Higuchi, R. (2000) High-throughput SNP allele-frequency determination in pooled DNA samples by kinetic PCR. *Genome Research* **10**: 258-266.
Gersonde, R., and Zielinski, U. (2000) The reconstruction of late quaternary Antarctic sea-ice distribution—the use of diatoms as a proxy for sea-ice. *Palaeogeography, Palaeoclimatology, Palaeoecology* **162**: 263-286.
Gerwert, K., Souvignier, G., and Hess, B. (1990) Simultaneous monitoring of light-induced changes in protein side-group protonation, chromophore isomerization, and backbone motion of bacteriorhodopsin by time-resolved Fourier-transform infrared spectroscopy. *Proceedings of the National Academy of Sciences* **87**: 9774-9778.
Gianoulis, T.A., Griffin, M.A., Spakowicz, D.J., Dunican, B.F., Alpha, C.J., Sboner, A. et al. (2012) Genomic analysis of the hydrocarbon-producing, cellulolytic, endophytic fungus *Ascocoryne sarcoides*. *PLoS Genetics* **8**: e1002558.
Gibson, G., and Muse, S.V. (2004) *A primer of genome science*. Sunderland, Massachusetts: Sinauer.
Gilstad, M., and Sakshaug, E. (1990) Growth rates of ten diatom species from the Barents Sea at different irradiances and day lengths. *Marine Ecology Progress Series* **64**: 169-173.
Giordano, D., Russo, R., di Prisco, G., and Verde, C. (2012) Molecular adaptations in Antarctic fish and marine microorganisms. *Marine Genomics* **6**: 1-6.
Giordano, M., Beardall, J., and Raven, J.A. (2005) $CO_2$ concentrating mechanisms in algae: mechanisms, environmental modulation, and evolution. *Annual Review of Plant Biology* **56**: 99-131.
Giovannoni, S.J., Hayakawa, D.H., Tripp, H.J., Stingl, U., Givan, S.A., Cho, J.-C. et al. (2008) The small genome of an abundant coastal ocean methylotroph. *Environmental Microbiology* **10**: 1771-1782.
Giovannoni, S.J., Bibbs, L., Cho, J.-C., Stapels, M.D., Desiderio, R., Vergin, K.L. et al. (2005) Proteorhodopsin in the ubiquitous marine bacterium SAR11. *Nature* **438**: 82-85.
Gleitz, M., v.d. Loeff, M.R., Thomas, D.N., Dieckmann, G.S., and Millero, F.J. (1995) Comparison of summer and winter inorganic carbon, oxygen and nutrient concentrations in Antarctic sea ice brine. *Marine Chemistry* **51**: 81-91.
Gloeckner, G., Rosenthal, A., and Valentin, K. (2000) The structure and gene repertoire of an ancient red algal plastid genome. *Journal of Molecular Evolution* **51**: 382-390.
Gomez-Consarnau, L., Gonzalez, J.M., Coll-Llado, M., Gourdon, P., Pascher, T., Neutze, R. et al. (2007) Light stimulates growth of proteorhodopsin-containing marine Flavobacteria. *Nature* **445**: 210-213.
Gomez-Consarnau, L., Akram, N., Lindell, K., Pedersen, A., Neutze, R., Milton, D.L. et al. (2010) Proteorhodopsin phototrophy promotes survival of marine bacteria during starvation. *PLoS Biology* **8**: e1000358.
Gowing, M.M. (2003) Large viruses and infected microeukaryotes in Ross Sea summer pack ice habitats. *Marine Biology* **142**: 1029-1040.
Gowing, M.M., Riggs, B.E., Garrison, D.L., Gibson, A.H., and Jeffries, M.O. (2002) Large viruses in Ross Sea late autumn pack ice habitats. *Marine Ecology Progress Series* **241**: 1-11.
Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I. et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**: 644-652.
Gradinger, R. (1996) Occurrence of an algal bloom under Arctic pack ice. *Marine Ecology Progress Series* **131**: 301-305.
Gregory, T.R. (2001) Coincidence, coevolution, or causation? DNA content, cellsize, and the C-value enigma. *Biological Reviews* **76**: 65-101.

Grigorieff, N., Ceska, T.A., Downing, K.H., Baldwin, J.M., and Henderson, R. (1996) Electron-crystallographic refinement of the structure of bacteriorhodopsin. *Journal of Molecular Biology* **259**: 393-421.

Gross, C.T., and McGinnis, W. (1996) DEAF-1, a novel protein that binds an essential region in a Deformed response element. *EMBO Journal* **15**: 1961-1970.

Gruber, A., Vugrinec, S., Hempel, F., Gould, S., Maier, U.-G., and Kroth, P. (2007) Protein targeting into complex diatom plastids: functional characterisation of a specific targeting motif. *Plant Molecular Biology* **64**: 519-530.

Grunow, A. (1884) Die Diatomeen von Franz Josefs-Land. In *Denkschriften d Kaiserl Akad d W math naturw CIasse*. Wien, pp. 53-112.

Grzymski, J.J., Carter, B.J., DeLong, E.F., Feldman, R.A., Ghadiri, A., and Murray, A.E. (2006) Comparative genomics of DNA fragments from six Antarctic marine planktonic bacteria. *Applied and Environmental Microbiology* **72**: 1532-1541.

Guillard, R.R. (1975) Culture of phytoplankton for feeding marine invertebrates. In *Culture of marine invertebrate animals*. W.L., S., and M.H, C. (eds). New York: Plenum Press, pp. 26-60.

Guillard, R.R.L., and Ryther, J.H. (1962) Studies of marine planktonic diatoms: I. *Cyclotella nana* Hustedt, and *Detonula confervacea* (Cleve) Gran. *Canadian Journal of Microbiology* **8**: 229-239.

Guillard, R.R.L., and Hargraves, P.E. (1993) *Stichochrysis immobilis* is a diatom, not a chrysophyte. *Phycologia* **32**: 234-236.

Guo, M., Rupe, M.A., Zinselmeier, C., Habben, J., Bowen, B.A., and Smith, O.S. (2004) Allelic variation of gene expression in maize hybrids. *The Plant Cell Online* **16**: 1707-1716.

Gupta, M., Yates, C.R., and Meibohm, B. (2005) SYBR Green-based real-time PCR allelic discrimination assay for $beta_2$-adrenergic receptor polymorphisms. *Analytical Biochemistry* **344**: 292-294.

Ha, T., Kang, S., Kwon, T., Ahn, J., Kim, S., and Kim, D. (2006) Antioxidant activity and contents of bioactive components in polar microalgae. *Ocean and Polar Research* **28**: 37-43.

Hagopian, J.C., Reis, M., Kitajima, J.P., Bhattacharya, D., and de Oliveira, M.C. (2004) Comparative analysis of the complete plastid genome sequence of the red alga *Gracilaria tenuistipitata* var. *liui* provides insights into the evolution of rhodoplasts and their relationship to other plastids. *Journal of Molecular Evolution* **59**: 464-477.

Hallin, P.F., and Ussery, D.W. (2004) CBS Genome Atlas Database: a dynamic storage for bioinformatic results and sequence data. *Bioinformatics* **20**: 3682-3686.

Hansson, B., and Westerberg, L. (2002) On the correlation between heterozygosity and fitness in natural populations. *Molecular Ecology* **11**: 2467-2474.

Hart, T.J. (1934) On the phytoplankton of the south-west Atlantic and the Bellingshausen Sea, 1929-31. *Discovery Reports* **8**: 1-268.

Hasle, G.R. (1965) *Nitzschia* and *Fragilariopsis* species in the light and electron microscopes. III. The genus *Fragilariopsis*. Skrifter utgitt av det Norske videnskaps-akademi i Oslo, I Mat-Naturv Klasse, Ny Serie **21**: 5-49.

Hasle, G.R. (1976) The biogeography of some marine planktonic diatoms. *Deep Sea Research and Oceanographic Abstracts* **23**: 319-322.

Hasle, G.R., and Syvertsen, E.E. (1997) Marine diatoms. In *Identifying marine phytoplankton*. Tomas, C.R. (ed). San Diego: Elsevier Academic Press, pp. 5-385.

Hastings, P.J., Lupski, J.R., Rosenberg, S.M., and Ira, G. (2009) Mechanisms of change in gene copy number. *Nature Reviews Genetics* **10**: 551-564.

Hawkins, J., Mahony, D., Maetschke, S., Wakabayashi, M., Teasdale, R.D., and Bodén, M. (2007) Identifying novel peroxisomal proteins. *Proteins: Structure, Function, and Bioinformatics* **69**: 606-616.

Hedrick, P.W. (2012) What is the evidence for heterozygote advantage selection? *Trends in Ecology & Evolution* **27**: 698-704.

Hegseth, E. (1989) Photoadaptation in marine Arctic diatoms. *Polar Biology* **9**: 479-486.

Heimdal, B.R. (1989) Arctic Ocean phytoplankton. In *The Arctic seas: climatology, oceanography, geology, and biology*. Herman, Y. (ed). New York: Van Nostrand Reinhold, pp. 193-222.

Helliwell, K.E., Wheeler, G.L., Leptos, K.C., Goldstein, R.E., and Smith, A.G. (2011) Insights into the evolution of vitamin $B_{12}$ auxotrophy from sequenced algal genomes. *Molecular Biology and Evolution* **28**: 2921-2933.

Hempel, G. (1987) Die Polarmeere - ein biologischer Vergleich. *Polarforschung* **57**: 173-189.

Hempel, G., and Piepenburg, D. (2010) Nord- und Südpolarmeer im Klimawandel. Ein biologischer Vergleich. *Biologie in unserer Zeit* **40**: 386-395.

Herold, S., Fago, A., Weber, R.E., Dewilde, S., and Moens, L. (2004) Reactivity studies of the Fe(III) and Fe(II)NO forms of human neuroglobin reveal a potential role against oxidative stress. *Journal of Biological Chemistry* **279**: 22841-22847.

Hershberg, R., and Petrov, D.A. (2009) General rules for optimal codon choice. *PLoS Genetics* **5**: e1000556.

Hershberg, R., and Petrov, D.A. (2010) Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genetics* **6**: e1001115.

Hildebrand, F., Meyer, A., and Eyre-Walker, A. (2010) Evidence of selection upon genomic GC-content in bacteria. *PLoS Genetics* **6**: e1001107.

Hoogstraten, A., Timmermans, K.R., and de Baar, H.J.W. (2012) Morphological and physiological effects in *Proboscia alata* (Bacillariophyceae) grown under different light and $CO_2$ conditions of the modern Southern Ocean. *Journal of Phycology* **48**: 559-568.

Horner, R. (1990) Ice-associated ecosystems. In *Polar marine diatoms*. Medlin, L.K., and Priddle, J. (eds). Cambridge: British Antarctic Survey, pp. 9-14.

Hou, Y., Zhang, H., Miranda, L., and Lin, S. (2010) Serious overestimation in quantitative PCR by circular (supercoiled) plasmid standard: microalgal *pcna* as the model gene. *PLoS ONE* **5**: e9545.

Hourton-Cabassa, C., Matos, A.R., Arrabaça, J., Demandre, C., Zachowski, A., and Moreau, F. (2009) Genetically modified *Arabidopsis thaliana* cells reveal the involvement of the mitochondrial fatty acid composition in membrane basal and uncoupling protein-mediated proton leaks. *Plant and Cell Physiology* **50**: 2084-2091.

Hu, H., Shi, Y., Cong, W., and Cai, Z. (2003) Growth and photosynthesis limitation of marine red tide alga *Skeletonema costatum* by low concentrations of $Zn^{2+}$. *Biotechnology Letters* **25**: 1881-1885.

Huber, O., and Sumper, M. (1994) Algal-CAMs: isoforms of a cell adhesion molecule in embryos of the alga *Volvox* with homology to *Drosophila* fasciclin I. *EMBO Journal* **13**: 4212-4222.

Hutchins, D., Sedwick, P., DiTullio, G., Boyd, P., Queguiner, B., Griffiths, F., and Crossley, C. (2001) Control of phytoplankton growth by iron and silicic acid availability in the subantarctic Southern Ocean: experimental results from the SAZ Project. *Journal of Geophysical Research* **106**: 559-531.

Hutchins, D.A., and Bruland, K.W. (1998) Iron-limited diatom growth and Si:N uptake ratios in a coastal upwelling regime. *Nature* **393**: 561-564.

Hutchins, D.A., Witter, A.E., Butler, A., and Luther, G.W. (1999) Competition among marine phytoplankton for different chelated iron species. *Nature* **400**: 858-861.

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

Idnurm, A., and Howlett, B.J. (2001) Characterization of an opsin gene from the ascomycete *Leptoshaeria maculans*. *Genome* **44**: 167-171.

Ihara, K., Umemura, T., Katagiri, I., Kitajima-Ihara, T., Sugiyama, Y., Kimura, Y., and Mukohata, Y. (1999) Evolution of the archaeal rhodopsins: evolution rate changes by gene duplication and functional differentiation. *Journal of Molecular Biology* **285**: 163-174.

Imasheva, E.S., Balashov, S.P., Choi, A.R., Jung, K.-H., and Lanyi, J.K. (2009) Reconstitution of *Gloeobacter violaceus* rhodopsin with a light-harvesting carotenoid antenna. *Biochemistry* **48**: 10948-10955.

Imlay, J.A. (2008) Cellular defenses against superoxide and hydrogen peroxide. *Annual Review of Biochemistry* **77**: 755-776.

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* **436**: 793-800.

Jacques, G. (1983) Some ecophysiological aspects of the Antarctic phytoplankton. *Polar Biology* **2**: 27-33.

Jaeckisch, N., Yang, I., Wohlrab, S., Glöckner, G., Kroymann, J., Vogel, H. et al. (2011) Comparative genomic and transcriptomic characterization of the toxigenic marine dinoflagellate *Alexandrium ostenfeldii*. *PLoS ONE* **6**: e28012.

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A. et al. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**: 463-467.

Jakobsson, M., Backman, J., Rudels, B., Nycander, J., Frank, M., Mayer, L. et al. (2007) The early Miocene onset of a ventilated circulation regime in the Arctic Ocean. *Nature* **447**: 986-990.

Janech, M.G., Krell, A., Mock, T., Kang, J.-S., and Raymond, J.A. (2006) Ice-binding proteins from sea ice diatoms (Bacillariophyceae). *Journal of Phycology* **42**: 410-416.

Jickells, T.D., An, Z.S., Andersen, K.K., Baker, A.R., Bergametti, G., Brooks, N. et al. (2005) Global iron connections between desert dust, ocean biogeochemistry, and climate. *Science* **308**: 67-71.

Jochem, F.J. (1999) Dark survival strategies in marine phytoplankton assessed by cytometric measurement of metabolic activity with fluorescein diacetate. *Marine Biology* **135**: 721-728.

Johnson, K.S., Gordon, R.M., and Coale, K.H. (1997) What controls dissolved iron concentrations in the world ocean? *Marine Chemistry* **57**: 137-161.

Jordan, I.K., Kondrashov, F.A., Adzhubei, I.A., Wolf, Y.I., Koonin, E.V., Kondrashov, A.S., and Sunyaev, S. (2005) A universal trend of amino acid gain and loss in protein evolution. *Nature* **433**: 633-638.

Joshi-Deo, J., Schmidt, M., Gruber, A., Weisheit, W., Mittag, M., Kroth, P.G., and Buechel, C. (2010) Characterization of a trimeric light-harvesting complex in the diatom *Phaeodactylum tricornutum* built of FcpA and FcpE proteins. *Journal of Experimental Botany* **61**: 3079-3087.

Jung, G., Lee, C.G., Kang, S.H., and Jin, E. (2007) Annotation and expression profile analysis of cDNAs from the Antarctic diatom *Chaetoceros neogracile*. *Journal of Microbiology and Biotechnology* **17**: 1330-1337.

Jung, J.Y., Choi, A.R., Lee, Y.K., Lee, H.K., and Jung, K.-H. (2008) Spectroscopic and photochemical analysis of proteorhodopsin variants from the surface of the Arctic Ocean. *FEBS Letters* **582**: 1679-1684.

Jung, K.-H. (2007) The distinct signaling mechanisms of microbial sensory rhodopsins in Archaea, Eubacteria and Eukarya. *Photochemistry and Photobiology* **83**: 63-69.

Jung, K.-H., Trivedi, V.D., and Spudich, J.L. (2003) Demonstration of a sensory rhodopsin in eubacteria. *Molecular Microbiology* **47**: 1513-1522.

Kanaya, S., Kinouchi, M., Abe, T., Kudo, Y., Yamada, Y., Nishi, T. et al. (2001) Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): characterization of horizontally transferred genes with emphasis on the *E. coli* O157 genome. *Gene* **276**: 89-99.

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

Kang, S.-H., Kang, J.-S., Lee, S., Chung, K.H., Kim, D., and Park, M.G. (2001) Antarctic phytoplankton assemblages in the marginal ice zone of the northwestern Weddell Sea. *Journal of Plankton Research* **23**: 333-352.

Keightley, P.D., Trivedi, U., Thomson, M., Oliver, F., Kumar, S., and Blaxter, M.L. (2009) Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Research* **19**: 1195-1201.

Kennett, J.P. (1977) Cenozoic evolution of Antarctic glaciation, the circum-Antarctic Ocean, and their impact on global paleoceanography. *Journal of Geophysical Research* **82**: 3843-3860.

Kettle, A.J., Andreae, M.O., Amouroux, D., Andreae, T.W., Bates, T.S., Berresheim, H. et al. (1999) A global database of sea surface dimethylsulfide (DMS) measurements and a procedure to predict sea surface DMS as a function of latitude, longitude, and month. *Global Biogeochemical Cycles* **13**: 399-444.

Khachane, A.N., Timmis, K.N., and dos Santos, V.A.P.M. (2005) Uracil content of 16S rRNA of thermophilic and psychrophilic prokaryotes correlates inversely with their optimal growth temperatures. *Nucleic Acids Research* **33**: 4016-4022.

Kiko, R. (2010) Acquisition of freeze protection in a sea-ice crustacean through horizontal gene transfer? *Polar Biology* **33**: 543-556.

Kilian, O., and Kroth, P.G. (2005) Identification and characterization of a new conserved motif within the presequence of proteins targeted into complex diatom plastids. *The Plant Journal* **41**: 175-183.

Kirchman, D.L. (1999) Oceanography: phytoplankton death in the sea. *Nature* **398**: 293-294.

Kirk, J.T.O. (2011) *Light and photosynthesis in aquatic ecosystems*. Cambridge: Cambridge University Press.

Kirst, G.O., and Wiencke, C. (1995) Ecophysiology of polar algae. *Journal of Phycology* **31**: 181-199.

Knight, C.A., Molinari, N.A., and Petrov, D.A. (2005) The large genome constraint hypothesis: evolution, ecology and phenotype. *Annals of Botany* **95**: 177-190.

Knight, J.C. (2004) Allele-specific gene expression uncovered. *Trends in Genetics* **20**: 113-116.

Knight, R., Freeland, S., and Landweber, L. (2001) A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biology* **2**: research0010.0011 - research0010.0013.

Knox, G., A (2007) *Biology of the Southern Ocean*. Boca Raton: CRC Press.

Kobayashi, M., and Shimizu, S. (1999) Cobalt proteins. *European Journal of Biochemistry* **261**: 1.

Kochendoerfer, G.G., Lin, S.W., Sakmar, T.P., and Mathies, R.A. (1999) How color visual pigments are tuned. *Trends in Biochemical Sciences* **24**: 300-305.

Koh, E.Y., Atamna-Ismaeel, N., Martin, A., Cowie, R.O.M., Beja, O., Davy, S.K. et al. (2010) Proteorhodopsin-bearing bacteria in Antarctic sea ice. *Applied and Environmental Microbiology* **76**: 5918-5925.

Kooistra, W.H.C.F., Gersonde, R., Medlin, L.K., and Mann, D.G. (2007) The origin and evolution of the diatoms: their adaptation to a planktonic existence. In *Evolution of primary producers in the sea*. Falkowski, P.G., and Knoll, A.H. (eds). Amsterdam: Elsevier Academic Press, pp. 207-249.

Kopczynska, E.E. (2008) Phytoplankton variability in Admiralty Bay, King George Island, South Shetland Islands: six years of monitoring. *Polish Polar Research* **29**: 117-139.

Kowallik, K., Stoebe, B., Schaffran, I., Kroth-Pancic, P., and Freier, U. (1995) The chloroplast genome of a chlorophyll a+c -containing alga, *Odontella sinensis*. *Plant Molecular Biology Reporter* **13**: 336-342.

Kozłowski, J., Konarzewski, M., and Gawelczyk, A.T. (2003) Cell size as a link between noncoding DNA and metabolic rate scaling. *Proceedings of the National Academy of Sciences* **100**: 14080-14085.

Krell, A. (2006) Salt stress tolerance in the psychrophilic diatom *Fragilariopsis cylindrus*. In. Bremen: University of Bremen, p. 124.

Krell, A., Funck, D., Plettner, I., John, U., and Dieckmann, G. (2007) Regulation of proline metabolism under salt stress in the psychrophilic diatom *Fragilariopsis cylindrus* (Bacillariophyceae). *Journal of Phycology* **43**: 753-762.

Krell, A., Beszteri, B., Dieckmann, G., Gloeckner, G., Valentin, K., and 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). *European Journal of Phycology* **43**: 423 - 433.

Krembs, C., Eicken, H., and Deming, J.W. (2011) Exopolymer alteration of physical properties of sea ice and implications for ice habitability and biogeochemistry in a warmer Arctic. *Proceedings of the National Academy of Sciences* **108**: 3653-3658.

Krembs, C., Eicken, H., Junge, K., and Deming, J.W. (2002) High concentrations of exopolymeric substances in Arctic winter sea ice: implications for the polar ocean carbon cycle and cryoprotection of diatoms. *Deep Sea Research Part I: Oceanographic Research Papers* **49**: 2163-2181.

Kroeger, N., and Poulsen, N. (2008) Diatoms - from cell wall biogenesis to nanotechnology. *Annual Review of Genetics* **42**: 83-107.

Kropuenske, L.R., Mills, M.M., van Dijken, G.L., Bailey, S., Robinson, D.H., Welschmeyer, N.A., and Arrigo, K.R. (2009) Photophysiology in two major Southern Ocean phytoplankton taxa: photoprotection in *Phaeocystis antarctica* and *Fragilariopsis cylindrus*. *Limnology and Oceanography* **54**: 1176.

Kroth, P.G. (2007) Genetic transformation: a tool to study protein targeting in diatoms. In *Methods in Molecular Biology*. van der Giezen, M. (ed). Totowa, NJ: Humana Press, pp. 257-267.

La Roche, J., Geider, R.J., Graziano, L.M., Murray, H., and Lewis, K. (1993) Induction of specific proteins in eukaryotic algae grown under iron-, phosphorus-, or nitrogen-deficient conditions. *Journal of Phycology* **29**: 767-777.

Lannuzel, D., Schoemann, V., de Jong, J., Tison, J.-L., and Chou, L. (2007) Distribution and biogeochemical behaviour of iron in the East Antarctic sea ice. *Marine Chemistry* **106**: 18-32.

Lanyon-Hogg, T., Warriner, S.L., and Baker, A. (2010) Getting a camel through the eye of a needle: the import of folded proteins by peroxisomes. *Biology of the Cell* **102**: 245-263.

Lehman, J.T. (1991) Interacting growth and loss rates: the balance of top-down and bottom-up controls in plankton communities. *Limnology and Oceanography* **36**: 1546-1554.

Letunic, I., Yamada, T., Kanehisa, M., and Bork, P. (2008) iPath: interactive exploration of biochemical pathways and networks. *Trends in Biochemical Sciences* **33**: 101-103.

Leu, J.Y., Chua, P.R., and Roeder, G.S. (1998) The meiosis-specific Hop2 protein of *S. cerevisiae* ensures synapsis between homologous chromosomes. *Cell* **94**: 375-386.

Levitus, S., Conkright, M.E., Reid, J.L., Najjar, R.G., and Mantyla, A. (1993) Distribution of nitrate, phosphate and silicate in the world oceans. *Progress In Oceanography* **31**: 245-273.

Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K., and Wang, J. (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**: 1966-1967.

Li, W.K.W. (1985) Photosynthetic response to temperature of marine phytoplankton along a latitudinal gradient (16°N to 74°N). *Deep Sea Research Part A Oceanographic Research Papers* **32**: 1381-1391.

Li, W.K.W., Smith, J.C., and Piatt, T. (1984) Temperature response of photosynthetic capacity and carboxylase activity in Arctic marine phytoplankton. *Marine Ecology Progress Series*: 237-243.

Ligowski, R., Godlewski, M., and Lukowski, A. (1992) Sea ice diatoms and ice edge planktonic diatoms at the northern limit of the Weddell Sea pack ice. *Proceedings of the Polar Biology Symposium of the National Institute of Polar Research, Tokyo, Japan* **5**: 9-20.

Liman, E.R., Tytgat, J., and Hess, P. (1992) Subunit stoichiometry of a mammalian $K^+$ channel determined by construction of multimeric cDNAs. *Neuron* **9**: 861-871.

Lin, S., Zhang, H., Zhuang, Y., Tran, B., and Gill, J. (2010) Spliced leader-based metatranscriptomic analyses lead to recognition of hidden genomic features in dinoflagellates. *Proceedings of the National Academy of Sciences* **107**: 20033-20038.

Lind, P.A., and Andersson, D.I. (2008) Whole-genome mutational biases in bacteria. *Proceedings of the National Academy of Sciences* **105**: 17878-17883.

Lindahl, T. (1993) Instability and decay of the primary structure of DNA. *Nature* **362**: 709-715.

Lindahl, T., and Nyberg, B. (1974) Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry* **13**: 3405-3410.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**: 523-536.

Littlepage, J.L. (1965) Oceanographic investigations in McMurdo Sound, Antarctica. In *Biology of the Antarctic Seas II*. Washington, DC: AGU, pp. 1-37.

Liu, Y., Chen, W., Gaudet, J., Cheney, M.D., Roudaia, L., Cierpicki, T. et al. (2007) Structural basis for recognition of SMRT/N-CoR by the MYND domain and its contribution to AML1/ETO's activity. *Cancer Cell* **11**: 483-497.

Lizotte, M.P. (2001) The contributions of sea ice algae to Antarctic marine primary production. *American Zoologist* **41**: 57-73.

Lo, H.S., Wang, Z., Hu, Y., Yang, H.H., Gere, S., Buetow, K.H., and Lee, M.P. (2003) Allelic variation in gene expression is common in the human genome. *Genome Research* **13**: 1855-1862.

Lobry, J.R., and Chessel, D. (2003) Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria. *Journal of Applied Genetics* **44**: 235-261.
Lobry, J.R., and Necşulea, A. (2006) Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes. *Gene* **385**: 128-136.
Locarnini, R.A., Mishonov, A.V., Antonov, J.I., Boyer, T.P., Garcia, H.E., Baranova, O.K. et al. (2010) World Ocean Atlas 2009, Volume 1: Temperature. In *NOAA Atlas NESDIS 68*. Levitus, S. (ed). Washington, D.C.: U.S. Government Printing Office, p. 184.
Lohr, M., Im, C.-S., and Grossman, A.R. (2005) Genome-based examination of chlorophyll and carotenoid biosynthesis in *Chlamydomonas reinhardtii*. *Plant Physiology* **138**: 490-515.
Lommer, M., Specht, M., Roy, A.-S., Kraemer, L., Andreson, R., Gutowska, M. et al. (2012) Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. *Genome Biology* **13**: R66.
Longhurst, A. (1995) Seasonal cycles of pelagic production and consumption. *Progress in Oceanography* **36**: 77-167.
Lovejoy, C., Galand, P., and Kirchman, D. (2011) Picoplankton diversity in the Arctic Ocean and surrounding seas. *Marine Biodiversity* **41**: 5-12.
Lüder, U.H., Wiencke, C., and Knoetzel, J. (2002) Acclimation of photosynthesis and pigments during and after six month of darkness in *Palmaria decipiens* (Rhodophyta): a study to simulate antarctic winter sea ice cover. *Journal of Phycology* **38**: 904-913.
Luecke, H., Schobert, B., Richter, H.-T., Cartailler, J.-P., and Lanyi, J.K. (1999) Structural changes in bacteriorhodopsin during ion transport at 2 Angstrom resolution. *Science* **286**: 255-260.
Lundholm, N., and Hasle, G.R. (2008) Are *Fragilariopsis cylindrus* and *Fragilariopsis nana* bipolar diatoms? - morphological and molecular analyses of two sympatric species. *Nova Hedwigia Beihefte* **133**: 231-250.
Lundholm, N., Daugbjerg, N., and Moestrup, Ø. (2002) Phylogeny of the Bacillariaceae with emphasis on the genus *Pseudo-nitzschia* (Bacillariophyceae) based on partial LSU rDNA. *European Journal of Phycology* **37**: 115-134.
Lynch, M. (2010) Rate, molecular spectrum, and consequences of human mutation. *Proceedings of the National Academy of Sciences* **107**: 961-968.
Lynch, M., and Conery, J.S. (2003) The origins of genome complexity. *Science* **302**: 1401-1404.
Lynch, M., Sung, W., Morris, K., Coffey, N., Landry, C.R., Dopman, E.B. et al. (2008) A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proceedings of the National Academy of Sciences* **105**: 9272-9277.
Lynn, D.J., Singer, G.A.C., and Hickey, D.A. (2002) Synonymous codon usage is subject to selection in thermophilic bacteria. *Nucleic Acids Research* **30**: 4272-4277.
Lyon, B.R., Lee, P.A., Bennett, J.M., DiTullio, G.R., and Janech, M.G. (2011) Proteomic analysis of a sea-ice diatom: Salinity acclimation provides new insight into the dimethylsulfoniopropionate production pathway. *Plant Physiology* **157**: 1926-1941.
M. Franck, V., Brzezinski, M.A., Coale, K.H., and Nelson, D.M. (2000) Iron and silicic acid concentrations regulate Si uptake north and south of the Polar Frontal Zone in the Pacific Sector of the Southern Ocean. *Deep Sea Research Part II: Topical Studies in Oceanography* **47**: 3315-3338.
Maheswari, U., Mock, T., Armbrust, E.V., and Bowler, C. (2009) Update of the diatom EST database: a new tool for digital transcriptomics. *Nucleic Acids Research* **37**: D1001-D1005.
Malik, S.-B., Pightling, A.W., Stefaniak, L.M., Schurko, A.M., and Logsdon, J.M., Jr. (2008) An expanded inventory of conserved meiotic genes provides evidence for sex in *Trichomonas vaginalis*. *PLoS ONE* **3**: e2879.
Man-Aharonovich, D., Sabehi, G., Sineshchekov, O.A., Spudich, E.N., Spudich, J.L., and Beja, O. (2004) Characterization of RS29, a blue-green proteorhodopsin variant from the Red Sea. *Photochemical & Photobiological Sciences* **3**: 459-462.
Man, D., Wang, W., Sabehi, G., Aravind, L., Post, A.F., Massana, R. et al. (2003a) Diversification and spectral tuning in marine proteorhodopsins. *The EMBO Journal* **22**: 1725-1731.
Man, D., Wang, W., Sabehi, G., Aravind, L., Post, A.F., Massana, R. et al. (2003b) Diversification and spectral tuning in marine proteorhodopsins. *EMBO Journal* **22**: 1725-1731.

Mann, D.G., and Droop, S.J.M. (1996) 3. Biodiversity, biogeography and conservation of diatoms. *Hydrobiologia* **336**: 19-32.
Maranger, R., Bird, D.R., and Juniper, S.K. (1994) Viral and bacterial dynamics in Arctic sea ice during the spring algal bloom near Resolute, N.W.T., Canada. *Marine Ecology Progress Series* **111**: 121-127.
Marashi, S.-A., and Ghalanbor, Z. (2004) Correlations between genomic GC levels and optimal growth temperatures are not 'robust'. *Biochemical and Biophysical Research Communications* **325**: 381-383.
Marchetti, A., Parker, M.S., Moccia, L.P., Lin, E.O., Arrieta, A.L., Ribalet, F. et al. (2009) Ferritin is used for iron storage in bloom-forming marine pennate diatoms. *Nature* **457**: 467-470.
Marchetti, A., Schruth, D.M., Durkin, C.A., Parker, M.S., Kodner, R.B., Berthiaume, C.T. et al. (2012) Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability. *Proceedings of the National Academy of Sciences*.
Mark Welch, D.B., and Meselson, M. (2000) Evidence for the evolution of bdelloid rotifers without sexual reproduction or genetic exchange. *Science* **288**: 1211-1215.
Martin, J.H. (1990) Glacial-Interglacial $CO_2$ change: the iron hypothesis. *Paleoceanography* **5**: 1-13.
Martin, J.H., and Fitzwater, S.E. (1988) Iron deficiency limits phytoplankton growth in the north-east Pacific subarctic. *Nature* **331**: 341-343.
Matsuno-Yagi, A., and Mukohata, Y. (1977) Two possible roles of bacteriorhodopsin; a comparative study of strains of *Halobacterium halobium* differing in pigmentation. *Biochemical and Biophysical Research Communications* **78**: 237-243.
Matsuzaki, M., Misumi, O., Shin-i, T., Maruyama, S., Takahara, M., Miyagishima, S.-y. et al. (2004) Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. *Nature* **428**: 653-657.
Maumus, F., Allen, A., Mhiri, C., Hu, H., Jabbari, K., Vardi, A. et al. (2009) Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. *BMC Genomics* **10**: 624.
Maxwell, K., and Johnson, G.N. (2000) Chlorophyll fluorescence - a practical guide. *Journal of Experimental Botany* **51**: 659-668.
Maykut, G.A., and Grenfell, T.C. (1975) The spectral distribution of light beneath first-year sea ice in the Arctic Ocean. *Limnology and Oceanography* **20**: 554-563.
McMinn, A., Pankowski, A., and Delfatti, T. (2005) Effect of hyperoxia on the growth and photosynthesis of polar sea ice microalgae. *Journal of Phycology* **41**: 732-741.
Medigue, C., Krin, E., Pascal, G., Barbe, V., Bernsel, A., Bertin, P.N. et al. (2005) Coping with cold: the genome of the versatile marine Antarctica bacterium *Pseudoalteromonas haloplanktis* TAC125. *Genome Research* **15**: 1325-1335.
Medlin, L.K., Kooistra, W.C.H.F., and Schmid, A.-M.M. (2000) A review of the evolution of the diatoms - a total approach using melecules, morphology and geology. In *The origin and early evolution of the diatoms, fossil, molecular and biogeographical approaches*. Witkowski, A., and Sieminska, J. (eds). Cracow, Poland: Szafer Institute of Botany, Polish Academy of Sciences, pp. 13-35.
Merchant, S., and Dreyfuss, B.W. (1998) Posttranslational assembly of photosynthetic metalloproteins. *Annual Review of Plant Physiology and Plant Molecular Biology* **49**: 25-51.
Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B. et al. (2007) The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* **318**: 245-250.
Meselson, M., and Welch, D.M. (2007) Stable heterozygosity? *Science* **318**: 202-203.
Methe, B.A., Nelson, K.E., Deming, J.W., Momen, B., Melamud, E., Zhang, X. et al. (2005) The psychrophilic lifestyle as revealed by the genome sequence of *Colwellia psychrerythraea* 34H through genomic and proteomic analyses. *Proceedings of the National Academy of Sciences* **102**: 10913-10918.
Milligan, A.J., and Morel, F.M.M. (2002) A proton buffering role for silica in diatoms. *Science* **297**: 1848-1850.

Mills, M.M., Kropuenske, L.R., van Dijken, G.L., Alderkamp, A.-C., Berg, G.M., Robinson, D.H. et al. (2010) Photophysiology in two Southern Ocean phytoplankton taxa: photosynthesis of *Phaeocystis antarctica* (Prymnesiophyceae) and *Fragilariopsis cylindrus* (Bacillariophyceae) under simulated mixed-layer irradiance. *Journal of Phycology* **46**: 1114-1127.

Mitchell, B.G. (1992) Predictive bio-optical relationships for polar oceans and marginal ice zones. *Journal of Marine Systems* **3**: 91-105.

Mitchell, B.G., Brody, E.A., Holm-Hansen, O., McClain, C., and Bishop, J. (1991) Light limitation of phytoplankton biomass and macronutrient utilization in the Southern Ocean. *Limnology and Oceanography* **36**: 1662-1677.

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

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**: 307-317.

Mock, T., and Thomas, D.N. (2005) Recent advances in sea-ice microbiology. *Environmental Microbiology* **7**: 605-619.

Mock, T., Krell, A., Glöckner, G., Kolukisaoglu, Ü., and Valentin, K. (2005) Analysis of expressed sequence tags (ESTs) from the polar diatom *Fragilariopsis cylindrus*. *Journal of Phycology* **42**: 78-85.

Mock, T., Dieckmann, G.S., Haas, C., Krell, A., Tison, J.-L., Belem, A.L. et al. (2002) Micro-optodes in sea ice: a new approach to investigate oxygen dynamics during sea ice formation. *Aquatic Microbial Ecology* **29**: 297-306.

Mock, T., Samanta, M.P., Iverson, V., Berthiaume, C., Robison, M., Holtermann, K. et al. (2008) Whole-genome expression profiling of the marine diatom *Thalassiosira pseudonana* identifies genes involved in silicon bioprocesses. *Proceedings of the National Academy of Sciences* **105**: 1579-1584.

Mock, T., Strauss, J., Allen, A.E., Green, B.R., Gruber, A., Kroth, P.G. et al. (2012) The *Fragilariopsis cylindrus* genome and its environmental transcriptome reveal insights into evolution and adaptation of phytoplankton to the Southern Ocean. In.

modENCODE, Roy, S., Ernst, J., Kharchenko, P.V., Kheradpour, P., Negre, N. et al. (2011) Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330**: 1787-1797.

Möglich, A., Yang, X., Ayers, R.A., and Moffat, K. (2010) Structure and function of plant photoreceptors. *Annual Review of Plant Biology* **61**: 21-47.

Montsant, A., Allen, A.E., Coesel, S., Martino, A.D., Falciatore, A., Mangogna, M. et al. (2007) Identification and comparative genomic analysis of signaling and regulatory components in the diatom *Thalassiosira pseudonana*. *Journal of Phycology* **43**: 585-604.

Moran, K., Backman, J., Brinkhuis, H., Clemens, S.C., Cronin, T., Dickens, G.R. et al. (2006) The Cenozoic palaeoenvironment of the Arctic Ocean. *Nature* **441**: 601-605.

Moran, M.A., and Armbrust, E.V. (2007) Genomes of sea microbes. *Oceanography* **20**: 47-55.

Morel, F.M.M., Milligan, A.J., and Saito, M.A. (2006) Marine bioinorganic chemistry: the role of trace metals in the oceanic cycles of major nutrients. In *The oceans and marine geochemistry*. Elderfield, H. (ed). Amsterdam: Elsevier, pp. 113-143.

Morel, F.M.M., Rueter, J.G., Anderson, D.M., and Guillard, R.R.L. (1979) Aquil: a chemically defined phytoplankton culture medium for trace metal studies. *Journal of Phycology* **15**: 135-141.

Morel, F.M.M., Reinfelder, J.R., Roberts, S.B., Chamberlain, C.P., Lee, J.G., and Yee, D. (1994) Zinc and carbon co-limitation of marine phytoplankton. *Nature* **369**: 740-742.

Morgan-Kiss, R.M., Priscu, J.C., Pocock, T., Gudynaite-Savitch, L., and Huner, N.P.A. (2006) Adaptation and acclimation of photosynthetic microorganisms to permanently cold environments. *Microbiology and Molecular Biology Reviews* **70**: 222-252.

Morgante, M., Brunner, S., Pea, G., Fengler, K., Zuccolo, A., and Rafalski, A. (2005) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nature Genetics* **37**: 997-1002.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**: 621-628.

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

Muller, H.J. (1932) Some genetic aspects of sex. *American Naturalist* **66**: 118-138.

Musto, H., Naya, H., Zavala, A., Romero, H., Alvarez-Valín, F., and Bernardi, G. (2004) Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. *FEBS Letters* **573**: 73-77.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**: 1344-1349.

Nagel, G., Möckel, B., Büldt, G., and Bamberg, E. (1995) Functional expression of bacteriorhodopsin in oocytes allows direct measurement of voltage dependence of light induced $H^+$ pumping. *FEBS Letters* **377**: 263-266.

Nagel, G., Kelety, B., Möckel, B., Büldt, G., and Bamberg, E. (1998) Voltage dependence of proton pumping by bacteriorhodopsin is regulated by the voltage-sensitive ratio of $M_1$ to $M_2$. *Biophysical Journal* **74**: 403-412.

Nagel, G., Ollig, D., Fuhrmann, M., Kateriya, S., Musti, A.M., Bamberg, E., and Hegemann, P. (2002) Channelrhodopsin-1: a light-gated proton channel in green algae. *Science* **296**: 2395-2398.

Nagel, G., Szellas, T., Huhn, W., Kateriya, S., Adeishvili, N., Berthold, P. et al. (2003) Channelrhodopsin-2, a directly light-gated cation-selective membrane channel. *Proceedings of the National Academy of Sciences* **100**: 13940-13945.

Neale, M.J., and Keeney, S. (2006) Clarifying the mechanics of DNA strand exchange in meiotic recombination. *Nature* **442**: 153-158.

Neale, P.J., Davis, R.F., and Cullen, J.J. (1998) Interactive effects of ozone depletion and vertical mixing on photosynthesis of Antarctic phytoplankton. *Nature* **392**: 585-589.

Nelson, D.M., Brzezinski, M.A., Sigmon, D.E., and Franck, V.M. (2001) A seasonal progression of Si limitation in the Pacific sector of the Southern Ocean. *Deep Sea Research Part II: Topical Studies in Oceanography* **48**: 3973-3995.

Nelson, D.M., Treguer, P., Brzezinski, M.A., Leynaert, A., and Queguiner, B. (1995) Production and dissolution of biogenic silica in the ocean: revised global estimates, comparison with regional data and relationship to biogenic sedimentation. *Global Biogeochemical Cycles* **9**: 359-372.

Neori, A., and Holm-Hansen, O. (1982) Effect of temperature on rate of photosynthesis in Antarctic phytoplankton. *Polar Biology* **1**: 33-38.

Newman, P.A., Gleason, J.F., McPeters, R.D., and Stolarski, R.S. (1997) Anomalously low ozone over the Arctic. *Geophysical Research Letters* **24**: 2689-2692.

Newton, C.R., Graham, A., Heptinstall, L.E., Powell, S.J., Summers, C., Kalsheker, N. et al. (1989) Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). *Nucleic Acids Research* **17**: 2503-2516.

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering* **10**: 1-6.

Noethig, E.-M., Assmy, P., Klaas, C., and Renate, S. (2009) Phyto- and protozooplankton in polar waters. In *Biological studies in polar oceans Exploration of life in icy waters*. Hempel, G., and Hempel, I. (eds). Bremerhaven: Wissenschaftsverlag NW, pp. 65-74.

Nolan, T., Hands, R.E., and Bustin, S.A. (2006) Quantification of mRNA using real-time RT-PCR. *Nature Protocols* **1**: 1559-1582.

Norton, T.A., Melkonian, M., and Andersen, R.A. (1996) Algal biodiversity. *Phycologia* **35**: 308-326.

O'Brien, K.M., and Crockett, E.L. (2013) The promise and perils of Antarctic fishes. *EMBO Reports* **14**: 17-24.

Oesterhelt, D., and Stoeckenius, W. (1971) Rhodopsin-like protein from the purple membrane of *Halobacterium halobium*. *Nature New Biology* **233**: 149-152.

Oh, D.-H., Dassanayake, M., Bohnert, H., and Cheeseman, J. (2012) Life at the extreme: lessons from the genome. *Genome Biology* **13**: 241.

Ohta, N., Matsuzaki, M., Misumi, O., Miyagishima, S.-y., Nozaki, H., Tanaka, K. et al. (2003) Complete sequence and analysis of the plastid genome of the unicellular red alga *Cyanidioschyzon merolae*. *DNA Research* **10**: 67-77.

Okamoto, O.K., and Hastings, J.W. (2003) Novel dinoflagellate clock-related genes identified through microarray analysis. *Journal of Phycology* **39**: 519-526.

Okimoto, R., and Dodgson, J.B. (1996) Improved PCR amplification of multiple specific alleles (PAMSA) using internally mismatched primers. *BioTechniques* **21**: 20-26.

Oleksiak, M.F., Churchill, G.A., and Crawford, D.L. (2002) Variation in gene expression within and among natural populations. *Nature Genetics* **32**: 261-266.

Oliver, M.J., Petrov, D., Ackerly, D., Falkowski, P., and Schofield, O.M. (2007) The mode and tempo of genome size evolution in eukaryotes. *Genome Research* **17**: 594-601.

Olli, K., Wassmann, P., Reigstad, M., Ratkova, T.N., Arashkevich, E., Pasternak, A. et al. (2007) The fate of production in the central Arctic Ocean – top–down regulation by zooplankton expatriates? *Progress In Oceanography* **72**: 84-113.

Ossowski, S., Schneeberger, K., Lucas-Lledó, J.I., Warthmann, N., Clark, R.M., Shaw, R.G. et al. (2010) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**: 92-94.

Oudot-Le Secq, M.-P., and Green, B.R. (2011) Complex repeat structures and novel features in the mitochondrial genomes of the diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*. *Gene* **476**: 20-26.

Oudot-Le Secq, M.-P., Grimwood, J., Shapiro, H., Armbrust, E., Bowler, C., and Green, B. (2007) Chloroplast genomes of the diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*: comparison with other plastid genomes of the red lineage. *Molecular Genetics and Genomics* **277**: 427-439.

Pabi, S., van Dijken, G.L., and Arrigo, K.R. (2008) Primary production in the Arctic Ocean, 1998–2006. *Journal of Geophysical Research: Oceans* **113**.

Pagny, S., Lerouge, P., Faye, L., and Gomord, V. (1999) Signals and mechanisms for protein retention in the endoplasmic reticulum. *Journal of Experimental Botany* **50**: 157-164.

Palenik, B., Grimwood, J., Aerts, A., Rouze, P., Salamov, A., Putnam, N. et al. (2007) The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. *Proceedings of the National Academy of Sciences* **104**: 7705-7710.

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

Palmisano, A.C., SooHoo, J.B., Moe, R.L., and Sullivan, C.W. (1987) Sea ice microbial communities. VII. Changes in under-ice spectral irradiance during the development of Antarctic sea ice microalgal communities. *Marine Ecology Progress Series* **35**: 165-173.

Pankowski, A., and McMinn, A. (2009) Iron availability regulates growth, photosynthesis, and production of ferredoxin and flavodoxin in Antarctic sea ice diatoms. *Aquatic Biology* **4**: 273-288.

Panzeca, C., Beck, A.J., Leblanc, K., Taylor, G.T., Hutchins, D.A., and Sañudo-Wilhelmy, S.A. (2008) Potential cobalt limitation of vitamin $B_{12}$ synthesis in the North Atlantic Ocean. *Global Biogeochemical Cycles* **22**.

Park, S., Jung, G., Hwang, Y.-s., and Jin, E. (2010) Dynamic response of the transcriptome of a psychrophilic diatom, *Chaetoceros neogracile*, to high irradiance. *Planta* **231**: 349-360.

Parkhill, J.-P., Maillet, G., and Cullen, J.J. (2001) Fluorescence-based maximal quantum yield for PSII as a diagnostic of nutrient stress. *Journal of Phycology* **37**: 517-529.

Parkinson, C.L., and Gloersen, P. (1993) Global sea ice coverage. In *Atlas of satellite observations related to global change*. Gurney, R.J., Foster, J.L., and Parkinson, C.L. (eds). Cambridge: Cambridge University Press, pp. 371-383.

Paterson, H., and Laybourn-Parry, J. (2012) Antarctic sea ice viral dynamics over an annual cycle. *Polar Biology* **35**: 491-497.

Payet, J.P. (2012) Ecology and diversity of marine viruses on the Canadian Arctic Shelf, Arctic Ocean. In *Faculty of Graduate Studies (Oceanography)*. Vancouver: University of British Columbia, p. 182.

Payet, J.P., and Suttle, C.A. (2008) Physical and biological correlates of virus dynamics in the southern Beaufort Sea and Amundsen Gulf. *Journal of Marine Systems* **74**: 933-945.

Peers, G., and Price, N.M. (2006) Copper-containing plastocyanin used for electron transport by an oceanic diatom. *Nature* **441**: 341-344.

Peers, G., Truong, T.B., Ostendorf, E., Busch, A., Elrad, D., Grossman, A.R. et al. (2009) An ancient light-harvesting protein is critical for the regulation of algal photosynthesis. *Nature* **462**: 518-521.

Peters, E., and Thomas, D.N. (1996) Prolonged darkness and diatom mortality I: marine Antarctic species. *Journal of Experimental Marine Biology and Ecology* **207**: 25-41.

Petersen, J., Teich, R., Brinkmann, H., and Cerff, R. (2006) A "green" phosphoribulokinase in complex algae with red plastids: evidence for a single secondary endosymbiosis leading to haptophytes, cryptophytes, heterokonts, and dinoflagellates. *Journal of Molecular Evolution* **62**: 143-157.

Petersen, T.N., Brunak, S., Heijne, G.v., and Nielsen, H. (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* **8**: 785–786.

Petrou, K., and Ralph, P.J. (2011) Photosynthesis and net primary productivity in three Antarctic diatoms: possible significance for their distribution in the Antarctic marine ecosystem. *Marine Ecology Progress Series* **437**: 27-40.

Petrov, D.A., and Hartl, D.L. (1999) Patterns of nucleotide substitution in *Drosophila* and mammalian genomes. *Proceedings of the National Academy of Sciences* **96**: 1475-1479.

Petukhova, G.V., Romanienko, P.J., and Camerini-Otero, R.D. (2003) The Hop2 protein has a direct role in promoting interhomolog interactions during mouse meiosis. *Developmental Cell* **5**: 927-936.

Pfaffl, M.W., Horgan, G.W., and Dempfle, L. (2002) Relative expression software tool (REST©) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. *Nucleic Acids Research* **30**: e36.

Pfaffl, M.W., Tichopad, A., Prgomet, C., and Neuvians, T.P. (2004) Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper – Excel-based tool using pair-wise correlations. *Biotechnology Letters* **26**: 509-515.

Plettner, I. (2002) Streßphysiologie bei antarktischen Diatomeen: Ökophysiologische Untersuchungen zur Bedeutung von Prolin bei der Anpassung an hohe Salinitäten und tiefe Temperaturen. *Fachbereich Biologie*.

Pouchkina-Stantcheva, N.N., McGee, B.M., Boschetti, C., Tolleter, D., Chakrabortee, S., Popova, A.V. et al. (2007) Functional divergence of former alleles in an ancient asexual invertebrate. *Science* **318**: 268-271.

Poulin, M. (1990) Ice diatoms: the Arctic. In *Polar marine diatoms*. Medlin, L.K., and Priddle, J. (eds). Cambridge: British Antarctic Survey, pp. 15-18.

Powell, W.M., and Clarke, G.L. (1936) The reflection and absorption of daylight at the surface of the ocean. *Journal of the Optical Society of America* **26**: 111-119.

Price, N.M., Harrison, G.I., Hering, J.G., Hudson, R.J., Nirel, P.M.V., Palenik, B., and Morel, F.M.M. (1988/89) Preparation and chemistry of the artificial algal culture medium Aquil. *Biological Oceanography* **6**: 443-461.

Priddle, J., and Fryxell, G.A. (1985) *Handbook of the common plankton diatoms of the Southern Ocean: centrales except the genus Thalassiosira*: British Antarctic Survey, Natural Environment Research Council.

Prochnik, S.E., Umen, J., Nedelcu, A.M., Hallmann, A., Miller, S.M., Nishii, I. et al. (2010) Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. *Science* **329**: 223-226.

Qin, Q.-L., Xie, B.-B., Shu, Y.-L., Rong, J.-C., Zhao, D.-L., Zhang, X.-Y. et al. (2012) Genome sequence of proteorhodopsin-containing sea Ice bacterium *Glaciecola punicea* ACAM 611T. *Journal of Bacteriology* **194**: 3267.

Quesneville, H., Bergman, C.M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M., and Anxolabehere, D. (2005) Combined evidence annotation of transposable elements in genome sequences. *PLoS Computational Biology* **1**: e22.

Quigg, A., Finkel, Z.V., Irwin, A.J., Rosenthal, Y., Ho, T.-Y., Reinfelder, J.R. et al. (2003) The evolutionary inheritance of elemental stoichiometry in marine phytoplankton. *Nature* **425**: 291-294.

Quillfeldt, C.H.v. (2004) The diatom *Fragilariopsis cylindrus* and its potential as an indicator species for cold water rather than for sea ice. *Vie et Milieu* **54**: 137-143.

R Development Core Team (2012). R: A language and environment for statistical computing. URL http://www.R-project.org

Ramesh, M.A., Malik, S.-B., and Logsdon Jr, J.M. (2005) A phylogenomic inventory of meiotic genes: evidence for sex in *Giardia* and an early eukaryotic origin of meiosis. *Current Biology* **15**: 185-191.

Rasmussen, R. (2001) Quantification on the LightCycler. In *Rapid Cycle Real-time PCR, Methods and Applications*. Meuer, S., Wittwer, C., and Nakagawara, K. (eds). Heidelberg: Springer, pp. 21-34.

Raven, J.A. (2009) Functional evolution of photochemical energy transformations in oxygen-producing organisms. *Functional Plant Biology* **36**: 505-515.

Raymond, J.A. (2011) Algal ice-binding proteins change the structure of sea ice. *Proceedings of the National Academy of Sciences* **108**: E198.

Raymond, J.A., and Janech, M.G. (2009) Ice-binding proteins from enoki and shiitake mushrooms. *Cryobiology* **58**: 151-156.

Raymond, J.A., and Kim, H.J. (2012) Possible role of horizontal gene transfer in the colonization of sea ice by algae. *PLoS ONE* **7**: e35968.

Reeves, S., McMinn, A., and Martin, A. (2011) The effect of prolonged darkness on the growth, recovery and survival of Antarctic sea ice diatoms. *Polar Biology* **34**: 1019-1032.

Regoli, F., Benedetti, M., Krell, A., and Abele, D. (2011) Oxidative challenges in polar seas. In *Oxidative stress in aquatic ecosystems*: John Wiley & Sons, Ltd, pp. 20-40.

Reinfelder, J.R., Kraepiel, A.M.L., and Morel, F.M.M. (2000) Unicellular $C_4$ photosynthesis in a marine diatom. *Nature* **407**: 996-999.

Reith, M., and Munholland, J. (1993) A high-resolution gene map of the chloroplast genome of the red alga *Porphyra purpurea*. *The Plant Cell Online* **5**: 465-475.

Rintoul, S.R., and Bullister, J.L. (1999) A late winter hydrographic section from Tasmania to Antarctica. *Deep Sea Research Part I: Oceanographic Research Papers* **46**: 1417-1454.

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S.D. et al. (2010) De novo assembly and analysis of RNA-seq data. *Nature Methods* **7**: 909-912.

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

Rocha, E.P.C. (2004) Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. *Genome Research* **14**: 2279-2286.

Roche, J.L., Boyd, P.W., McKay, R.M.L., and Geider, R.J. (1996) Flavodoxin as an *in situ* marker for iron stress in phytoplankton. *Nature* **382**: 802-805.

Rodhouse, P.G., and White, M.G. (1995) Cephalopods occupy the ecological niche of epipelagic fish in the Antarctic polar frontal zone. *Biological Bulletin* **189**: 77-80.

Rogers, A.D. (2007) Evolution and biodiversity of Antarctic organisms: a molecular perspective. *Philosophical Transactions of the Royal Society B: Biological Sciences* **362**: 2191-2214.

Ross-Macdonald, P., Coelho, P.S.R., Roemer, T., Agarwal, S., Kumar, A., Jansen, R. et al. (1999) Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402**: 413-418.

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

Roy, A., Kucukural, A., and Zhang, Y. (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols* **5**: 725-738.

Roy, A., Yang, J., and Zhang, Y. (2012) COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Research* **40**: W471-W477.

Rue, E.L., and Bruland, K.W. (1995) Complexation of iron (III) by natural organic ligands in the Central North Pacific as determined by a new competitive ligand equilibration/adsorptive cathodic stripping voltammetric method. *Marine Chemistry* **50**: 117-138.

Ruiz-González, M.X., and Marín, I. (2004) New insights into the evolutionary history of type 1 rhodopsins. *Journal of Molecular Evolution* **58**: 348-358.

Rysgaard, S., Nielsen, T.G., and Hansen, B.W. (1999) Seasonal variation in nutrients, pelagic primary production and grazing in a high-Arctic coastal marine ecosystem, Young Sound, Northeast Greenland. *Marine Ecology Progress Series* **179**: 13-25.

Sabehi, G., Béjà, O., Suzuki, M.T., Preston, C.M., and DeLong, E.F. (2004) Different SAR86 subgroups harbour divergent proteorhodopsins. *Environmental Microbiology* **6**: 903-910.

Sabehi, G., Massana, R., Bielawski, J.P., Rosenberg, M., Delong, E.F., and Béjà, O. (2003) Novel proteorhodopsin variants from the Mediterranean and Red Seas. *Environmental Microbiology* **5**: 842-849.

Saito, M.A., Goepfert, T.J., and Jason, T.R. (2008) Some thoughts on the concept of colimitation: three definitions and the importance of bioavailability. *Limnology and Oceanography* **53**: 276-290.

Saito, M.A., Goepfert, T.J., Noble, A.E., Bertrand, E.M., Sedwick, P.N., and DiTullio, G.R. (2010) A seasonal study of dissolved cobalt in the Ross Sea, Antarctica: micronutrient behavior, absence of scavenging, and relationships with Zn, Cd, and P. *Biogeosciences* **7**: 4059-4082.

Sakshaug, E., and Holm-Hansen, O. (1984) Factors governing pelagic production in polar oceans. In *Marine Phytoplankton and Productivity*. Washington, DC: AGU, pp. 1-18.

Sakshaug, E., Slagstad, D., and Holm-Hansen, O. (1991) Factors controlling the development of phytoplankton blooms in the Antarctic Ocean — a mathematical model. *Marine Chemistry* **35**: 259-271.

Salamov, A.A., and Solovyev, V.V. (2000) Ab initio gene finding in *Drosophila* genomic DNA. *Genome Research* **10**: 516-522.

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

Sañudo-Wilhelmy, S.A., Gobler, C.J., Okbamichael, M., and Taylor, G.T. (2006) Regulation of phytoplankton dynamics by vitamin $B_{12}$. *Geophysical Research Letters* **33**: n/a-n/a.

Sañudo-Wilhelmy, S.A., Cutter, L.S., Durazo, R., Smail, E.A., Gómez-Consarnau, L., Webb, E.A. et al. (2012) Multiple B-vitamin depletion in large areas of the coastal ocean. *Proceedings of the National Academy of Sciences* **109**: 14041-14045.

Saunders, N.F.W., Thomas, T., Curmi, P.M.G., Mattick, J.S., Kuczek, E., Slade, R. et al. (2003) Mechanisms of thermal adaptation revealed from the genomes of the Antarctic archaea *Methanogenium frigidum* and *Methanococcoides burtonii*. *Genome Research* **13**: 1580-1588.

Schaart, J.G., Mehli, L., and Schouten, H.J. (2005) Quantification of allele-specific expression of a gene encoding strawberry polygalacturonase-inhibiting protein (PGIP) using pyrosequencing. *The Plant Journal* **41**: 493-500.

Schaller, G.E., Shiu, S.-H., and Armitage, Judith P. (2011) Two-component systems and their co-option for eukaryotic signal transduction. *Current Biology* **21**: R320-R330.

Schriek, R. (2000) Effect of light and temperature on the enzymatic antioxidative defense systems in the Antarctic ice diatom *Entomoneis kufferathii* Manguin. In *Fachbereich Biologie/Chemie*. Bremen: University of Bremen, p. 130.

Schurko, A.M., and Logsdon, J.M. (2008) Using a meiosis detection toolkit to investigate ancient asexual "scandals" and the evolution of sex. *BioEssays* **30**: 579-589.

Schurko, A.M., Neiman, M., and Logsdon Jr, J.M. (2009) Signs of sex: what we know and how we know it. *Trends in Ecology & Evolution* **24**: 208-217.

Schuster, S.C. (2008) Next-generation sequencing transforms today's biology. *Nature Methods* **5**: 16-18.

Sedwick, P.N., and DiTullio, G.R. (1997) Regulation of algal blooms in Antarctic shelf waters by the release of iron from melting sea ice. *Geophysical Research Letters* **24**: 2515-2518.

Sellis, D., Callahan, B.J., Petrov, D.A., and Messer, P.W. (2011) Heterozygote advantage as a natural consequence of adaptation in diploids. *Proceedings of the National Academy of Sciences* **108**: 20666-20671.

Shackleton, N., and Kennett, J. (1975) Paleotemperature history of the Cenozoic and the initiation of Antarctic glaciation: oxygen and carbon isotope analyses in DSDP Sites 277, 279, and 281. *Initial reports of the deep sea drilling project* **29**: 743-755.

Shaked, Y., Xu, Y., Leblanc, K., and Morel, F.M.M. (2006) Zinc availability and alkaline phosphatase activity in *Emiliania huxleyi*: implications for Zn-P co-limitation in the ocean. *Limnology and Oceanography* **51**: 299-309.

Sharma, A.K., Spudich, J.L., and Doolittle, W.F. (2006) Microbial rhodopsins: functional versatility and genetic mobility. *Trends in Microbiology* **14**: 463-469.

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

Sineshchekov, O.A., Jung, K.-H., and Spudich, J.L. (2002) Two rhodopsins mediate phototaxis to low- and high-intensity light in *Chlamydomonas reinhardtii*. *Proceedings of the National Academy of Sciences* **99**: 8689-8694.

Sineshchekov, O.A., Govorunova, E.G., Jung, K.-H., Zauner, S., Maier, U.-G., and Spudich, J.L. (2005) Rhodopsin-mediated photoreception in cryptophyte flagellates. *Biophysical Journal* **89**: 4310-4319.

Singer, G.A.C., and Hickey, D.A. (2003) Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene* **317**: 39-47.

Slamovits, C.H., Okamoto, N., Burri, L., James, E.R., and Keeling, P.J. (2011) A bacterial proteorhodopsin proton pump in marine eukaryotes. *Nature Communications* **2**: 183.

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

Smetacek, V., and Nicol, S. (2005) Polar ocean ecosystems in a changing world. *Nature* **437**: 362-368.

Smetacek, V., Assmy, P., and Henjes, J. (2004) The role of grazing in structuring Southern Ocean pelagic ecosystems and biogeochemical cycles. *Antarctic Science* **16**: 541-558.

Smetacek, V., Klaas, C., Menden-Deuer, S., and Rynearson, T.A. (2002) Mesoscale distribution of dominant diatom species relative to the hydrographical field along the Antarctic Polar Front. *Deep Sea Research Part II: Topical Studies in Oceanography* **49**: 3835-3848.

Smith, D.C., and Steward, G.F. (1992) Virus and bacteria abundances in the Drake Passage during January and August 1991. *Antarctic Journal of the United States* **27**: 125.

Smith, R., Prezelin, B., Baker, K., Bidigare, R., Boucher, N., Coley, T. et al. (1992) Ozone depletion: ultraviolet radiation and phytoplankton biology in Antarctic waters. *Science* **255**: 952-959.

Smith, W.O., and Nelson, D.M. (1985) Phytoplankton bloom produced by a receding ice edge in the Ross Sea: spatial coherence with the density field. *Science* **227**: 163-166.

Smith, W.O., and Sakshaug, E. (1990) Polar phytoplankton. In *Polar oceanography, Part B: Chemistry, biology, and geology*. Smith, W.O. (ed). San Diego: Academic Press, p. 760.

Smith, W.O., and Harrison, W.G. (1991) New production in polar regions: the role of environmental controls. *Deep Sea Research Part A Oceanographic Research Papers* **38**: 1463-1479.

Sodergren, E., Weinstock, G.M., Davidson, E.H., Cameron, R.A., Gibbs, R.A., Angerer, R.C. et al. (2006) The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science* **314**: 941-952.

Spudich, J.L. (2006) The multitalented microbial sensory rhodopsins. *Trends in Microbiology* **14**: 480-487.

Spudich, J.L., Yang, C.-S., Jung, K.-H., and Spudich, E.N. (2000) Retinylidene proteins: structures and functions from archaea to humans. *Annual Review of Cell and Developmental Biology* **16**: 365-392.

Stehfest, K., Toepel, J., and Wilhelm, C. (2005) The application of micro-FTIR spectroscopy to analyze nutrient stress-related changes in biomass composition of phytoplankton algae. *Plant Physiology and Biochemistry* **43**: 717-726.

Stickley, C.E., St John, K., Koc, N., Jordan, R.W., Passchier, S., Pearce, R.B., and Kearns, L.E. (2009) Evidence for middle Eocene Arctic sea ice from diatoms and ice-rafted debris. *Nature* **460**: 376-379.

Stingl, U., Desiderio, R.A., Cho, J.-C., Vergin, K.L., and Giovannoni, S.J. (2007) The SAR92 clade: an abundant coastal clade of culturable marine bacteria possessing proteorhodopsin. *Applied and Environmental Microbiology* **73**: 2290-2296.

Strelnikova, N.I., and Simola, H. (1990) Evolution of diatoms during the Cretaceous and Paleogene periods. In *Proceedings of the 10th International Diatom Symposium*. Simola, H. (ed). Koenigstein: Koeltz Scientific Books pp. 195-204.

Strzepek, R.F., and Harrison, P.J. (2004) Photosynthetic architecture differs in coastal and oceanic diatoms. *Nature* **431**: 689-692.

Subczynski, W.K., Markowska, E., Gruszecki, W.I., and Sielewiesiuk, J. (1992) Effects of polar carotenoids on dimyristoylphosphatidylcholine membranes: a spin-label study. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1105**: 97-108.

Subramaniam, S., Greenhalgh, D.A., Rath, P., Rothschild, K.J., and Khorana, H.G. (1991) Replacement of leucine-93 by alanine or threonine slows down the decay of the N and O intermediates in the photocycle of bacteriorhodopsin: implications for proton uptake and 13-*cis*-retinal -> all-*trans*-retinal reisomerization. *Proceedings of the National Academy of Sciences* **88**: 6873-6877.

Sudmant, P.H., Kitzman, J.O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A. et al. (2010) Diversity of human copy number variation and multicopy genes. *Science* **330**: 641-646.

Sudo, Y., Ihara, K., Kobayashi, S., Suzuki, D., Irieda, H., Kikukawa, T. et al. (2011) A microbial rhodopsin with a unique retinal composition shows both sensory rhodopsin II and bacteriorhodopsin-like properties. *Journal of Biological Chemistry* **286**: 5967-5976.

Sueoka, N. (1961) Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. *Proceedings of the National Academy of Sciences* **47**: 1141-1149.

Sullivan, C.W., Palmisano, A.C., and SooHoo, J.B. (1984) Influence of sea ice and sea ice biota on downwelling irradiance and spectral composition of light in McMurdo Sound *SPIE Proceedings - The International Society for Optics and Photonics* **489**: 159-577.

Sunda, W., and Huntsman, S. (2003) Effect of pH, light, and temperature on Fe–EDTA chelation and Fe hydrolysis in seawater. *Marine Chemistry* **84**: 35-47.

Sunda, W.G., Swift, D.G., and Huntsman, S.A. (1991) Low iron requirement for growth in oceanic phytoplankton. *Nature* **351**: 55-57.

Sunda, W.G., Price, N.M., and Morel, F.M.M. (2005) Trace metal ion buffers and their use in culture studies. In *Algal Culturing Techniques*. Anderson, R. (ed). Burlington, MA: Academic Press, pp. 35-63.

Supek, F., Bosnjak, M., Skunca, N., and Smuc, T. (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **6**: e21800.

Suttle, C.A. (1994) The significance of viruses to mortality in aquatic microbial communities. *Microbial Ecology* **28**: 237-243.

Suttle, C.A. (2005) Viruses in the sea. *Nature* **437**: 356-361.

Suutari, M., and Laakso, S. (1994) Microbial fatty acids and thermal adaptation. *Critical Reviews in Microbiology* **20**: 285-328.

Takeda, S. (1998) Influence of iron availability on nutrient consumption ratio of diatoms in oceanic waters. *Nature* **393**: 774-777.

Tang, K.W., Smith, W.O., Shields, A.R., and Elliott, D.T. (2009) Survival and recovery of *Phaeocystis antarctica* (Prymnesiophyceae) from prolonged darkness and freezing. *Proceedings of the Royal Society B: Biological Sciences* **276**: 81-90.

Taylor, F.J., and Coates, D. (1989) The code within the codons. *Biosystems* **22**: 177-187.

Taylor, G.T., and Sullivan, C.W. (2008) Vitamin $B_{12}$ and cobalt cycling among diatoms and bacteria in Antarctic sea ice microbial communities. *Limnology and Oceanography* **53**: 1862-1877.

The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796-815.

Thomas, D.N. (2003) Iron limitation in the Southern Ocean. *Science* **302**: 565c-566.

Thomas, D.N., and Dieckmann, G.S. (2002) Antarctic sea ice - a habitat for extremophiles. *Science* **295**: 641-644.

Thomas, D.N., and Dieckmann, G.S. (eds) (2010) *Sea Ice*. Oxford, UK: Wiley-Blackwell.

Thomas, M.K., Kremer, C.T., Klausmeier, C.A., and Litchman, E. (2012) A global pattern of thermal adaptation in marine phytoplankton. *Science* **338**: 1085-1088.

Thuróczy, C.E., Boye, M., and Losno, R. (2010) Dissolution of cobalt and zinc from natural and anthropogenic dusts in seawater. *Biogeosciences* **7**: 1927-1936.

Tilzer, M.M., and Dubinsky, Z. (1987) Effects of temperature and day length on the mass balance of Antarctic phytoplankton. *Polar Biology* **7**: 35-42.

Tilzer, M.M., Elbrächter, M., Gieskes, W.W., and Beese, B. (1986) Light-temperature interactions in the control of photosynthesis in Antarctic phytoplankton. *Polar Biology* **5**: 105-111.

Timmermans, K.R., Veldhuis, M.J.W., and Brussaard, C.P.D. (2007) Cell death in three marine diatom species in response to different irradiance levels, silicate, or iron concentrations. *Aquatic Microbial Ecology* **46**: 253-261.

Timmermans, K.R., Gerringa, L.J.A., Baar, H.J.W.d., Wagt, B.v.d., Veldhuis, M.J.W., Jong, J.T.M.d. et al. (2001) Growth rates of large and small Southern Ocean diatoms in relation to availability of iron in natural seawater. *Limnology and Oceanography* **46**: 260-266.

Tolosano, E., Fagoonee, S., Morello, N., Vinchi, F., and Fiorito, V. (2010) Heme scavenging and the other facets of hemopexin. *Antioxidant & Redox Signaling* **12**: 305-320.

Toseland, A., Daines, S., Clark, J., Kirkham, A., Strauss, J., Uhlig, C. et al. (2013) Temperature controls marine phytoplankton metabolism with implications for resource allocation and the Redfield N:P ratio. In.

Tottey, S., Block, M.A., Allen, M., Westergren, T., Albrieux, C., Scheller, H.V. et al. (2003) Arabidopsis CHL27, located in both envelope and thylakoid membranes, is required for the synthesis of protochlorophyllide. *Proceedings of the National Academy of Sciences* **100**: 16119-16124.

Treguer, P., Nelson, D.M., Van Bennekom, A.J., DeMaster, D.J., Leynaert, A., and Queguiner, B. (1995) The silica balance in the world ocean: a reestimate. *Science* **268**: 375-379.

Trenerry, L., McMinn, A., and Ryan, K. (2002) In situ oxygen microelectrode measurements of bottom-ice algal production in McMurdo Sound, Antarctica. *Polar Biology* **25**: 72-80.

Tripati, A., Backman, J., Elderfield, H., and Ferretti, P. (2005) Eocene bipolar glaciation associated with global carbon cycle changes. *Nature* **436**: 341-346.

Tripp, B.C., and Ferry, J.G. (2000) A structure-function study of a proton transport pathway in the γ-class carbonic anhydrase from *Methanosarcina thermophila*. *Biochemistry* **39**: 9232-9240.

Tsubouchi, H., and Roeder, G.S. (2003) The importance of genetic recombination for fidelity of chromosome pairing in meiosis. *Developmental Cell* **5**: 915-925.

Tsuda, A., and Kawaguchi, S. (1997) Microzooplankton grazing in the surface water of the Southern Ocean during an austral summer. *Polar Biology* **18**: 240-245.

Tsuda, A., Takeda, S., Saito, H., Nishioka, J., Nojiri, Y., Kudo, I. et al. (2003) A mesoscale iron enrichment in the western subarctic Pacific induces a large centric diatom bloom. *Science* **300**: 958-961.

Tsunenari, T., Sun, H., Williams, J., Cahill, H., Smallwood, P., Yau, K.-W., and Nathans, J. (2003) Structure-function analysis of the bestrophin family of anion channels. *Journal of Biological Chemistry* **278**: 41114-41125.

Tsunoda, S.P., Ewers, D., Gazzarrini, S., Moroni, A., Gradmann, D., and Hegemann, P. (2006) $H^+$-pumping rhodopsin from the marine alga *Acetabularia*. *Biophysical Journal* **91**: 1471-1479.

Turner, G.J., Chittiboyina, S., Pohren, L., Hines, K.G., Correia, J.J., and Mitchell, D.C. (2009) The bacteriorhodopsin carboxyl-terminus contributes to proton recruitment and protein stability. *Biochemistry* **48**: 1112-1122.

Turner, J., and Marshall, G.J. (2011) *Climate change in the polar regions*. Cambridge: Cambridge University Press.

Tuskan, G.A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U. et al. (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**: 1596-1604.

Uhlig, C., Kabisch, J., Palm, G.J., Valentin, K., Schweder, T., and Krell, A. (2011) Heterologous expression, refolding and functional characterization of two antifreeze proteins from *Fragilariopsis cylindrus* (Bacillariophyceae). *Cryobiology* **63**: 220-228.

Ulrich, L.E., Koonin, E.V., and Zhulin, I.B. (2005) One-component systems dominate signal transduction in prokaryotes. *Trends in Microbiology* **13**: 52-56.

Um, M.Y., Kang, S.H., Ahn, J.Y., and Ha, T.Y. (2012) Antioxidant activity of *Fragilariopsis pseudonana* and protective effect against hydrogen peroxide-induced inhibition of gap junctional intercellular communication. *Food Science and Biotechnology* **21**: 435-441.

Vallee, B.L., and Auld, D.S. (1990) Zinc coordination, function, and structure of zinc enzymes and other proteins. *Biochemistry* **29**: 5647-5659.

Van de Peer, Y. (2011) Genomes: the truth is in there. *EMBO Reports* **12**: 93-93.

van Nimwegen, E. (2003) Scaling laws in the functional content of genomes. *Trends in Genetics* **19**: 479-484.

van Oosterhout, C. (2009) Transposons in the MHC: the Yin and Yang of the vertebrate immune system. *Heredity* **103**: 190-191.

Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D.A., Cestaro, A., Pruss, D. et al. (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* **2**: e1326.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A. et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66-74.

Verde, C., Giordano, D., Russo, R., Riccio, A., Vergara, A., Mazzarella, L., and di Prisco, G. (2009) Hemoproteins in the cold. *Marine Genomics* **2**: 67-73.

Veselý, P., Bureš, P., Šmarda, P., and Pavlíček, T. (2012) Genome size and DNA base composition of geophytes: the mirror of phenology and ecology? *Annals of Botany* **109**: 65-75.

Villeneuve, A.M., and Hillers, K.J. (2001) Whence meiosis? *Cell* **106**: 647-650.

Vincent, W. (2002) Cyanobacterial dominance in the polar regions. In *The Ecology of Cyanobacteria*. Whitton, B., and Potts, M. (eds): Springer Netherlands, pp. 321-340.

Vinogradov, A.E., and Anatskaya, O.V. (2006) Genome size and metabolic intensity in tetrapods: a tale of two lines. *Proceedings of the Royal Society B: Biological Sciences* **273**: 27-32.

von Dassow, P., and Montresor, M. (2011) Unveiling the mysteries of phytoplankton life cycles: patterns and opportunities behind complexity. *Journal of Plankton Research* **33**: 3-12.

von der Gathen, P., Rex, M., Harris, N.R.P., Lucic, D., Knudsen, B.M., Braathen, G.O. et al. (1995) Observational evidence for chemical ozone depletion over the Arctic in winter 1991-92. *Nature* **375**: 131-134.

Voronina, N.M. (1998) Comparative abundance and distribution of major filter-feeders in the Antarctic pelagic zone. *Journal of Marine Systems* **17**: 375-390.

Wada, A., and Suyama, A. (1986) Local stability of DNA and RNA secondary structure and its relation to biological functions. *Progress in Biophysics and Molecular Biology* **47**: 113-157.

Walsh, J.J. (1981) A carbon budget for overfishing off Peru. *Nature* **290**: 300-304.

Wang, W.-W., Sineshchekov, O.A., Spudich, E.N., and Spudich, J.L. (2003) Spectroscopic and photochemical characterization of a deep ocean proteorhodopsin. *Journal of Biological Chemistry* **278**: 33985-33991.

Wang, Z., Gerstein, M., and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**: 57-63.

Waschuk, S.A., Bezerra, A.G., Shi, L., and Brown, L.S. (2005) *Leptosphaeria* rhodopsin: Bacteriorhodopsin-like proton pump from a eukaryote. *Proceedings of the National Academy of Sciences* **102**: 6879-6883.

Weger, H.G., Herzig, R., Falkowski, P.G., and Turpin, D.H. (1989) Respiratory losses in the light in a marine diatom: measurements by short-term mass spectrometry. *Limnol Oceanogr* **34**: 1153–1161.

Wessler, S.R. (1996) Plant retrotransposons: turned on by stress. *Current Biology* **6**: 959-961.

Wheeler, P.A., Gosselin, M., Sherr, E., Thibaultc, D., Kirchman, D.L., Benner, R., and Whitledge, T.E. (1996) Active cycling of organic carbon in the central Arctic Ocean. *Nature* **380**: 697-699.

Wilhelm, B.T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I. et al. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**: 1239-1243.

Wilhelm, C., Buchel, C., Fisahn, J., Goss, R., Jakob, T., LaRoche, J. et al. (2006) The regulation of carbon and nutrient assimilation in diatoms is significantly different from green algae. *Protist* **157**: 91-124.

Wilkening, S., Hemminki, K., Thirumaran, R.K., Bermejo, J.L., Bonn, S., Forsti, A., and Kumar, R. (2005) Determination of allele frequency in pooled DNA: comparison of three PCR-based methods. *Biotechniques* **39**: 853-858.

Wilson, D.L., Smith Jr, W.O., and Nelson, D.M. (1986) Phytoplankton bloom dynamics of the western Ross Sea ice edge - I. Primary productivity and species-specific production. *Deep Sea Research Part A Oceanographic Research Papers* **33**: 1375-1387.

Wilson, D.N., and Nierhaus, K.H. (2007) The weird and wonderful world of bacterial ribosome regulation. *Critical Reviews in Biochemistry and Molecular Biology* **42**: 187-219.

Worden, A.Z., Lee, J.-H., Mock, T., Rouze, P., Simmons, M.P., Aerts, A.L. et al. (2009) Green evolution and dynamic adaptations revealed by genomes of the marine micoeukaryotes *Micromonas*. *Science* **324**: 268-272.

Wu, H., Zhang, Z., Hu, S., and Yu, J. (2012) On the molecular mechanism of GC content variation among eubacterial genomes. *Biology Direct* **7**: 2.

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

Yamada, K., Lim, J., Dale, J.M., Chen, H., Shinn, P., Palm, C.J. et al. (2003) Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* **302**: 842-846.

Yamada, T., Letunic, I., Okuda, S., Kanehisa, M., and Bork, P. (2011) iPath2.0: interactive pathway explorer. *Nucleic Acids Research* **39**: W412-W415.

Yamamoto, Y., Fujisawa, A., Hara, A., and Dunlap, W.C. (2001) An unusual vitamin E constituent (alpha-tocomonoenol) provides enhanced antioxidant protection in marine organisms adapted to cold-water environments. *Proceedings of the National Academy of Sciences* **98**: 13144-13148.

Yan, H., Yuan, W., Velculescu, V.E., Vogelstein, B., and Kinzler, K.W. (2002) Allelic variation in human gene expression. *Science* **297**: 1143.

Yandell, M., and Ence, D. (2012) A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics* **13**: 329-342.

Young, M., Wakefield, M., Smyth, G., and Oshlack, A. (2010) Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology* **11**: R14.

Yu, J. (2007) A content-centric organization of the genetic code. *Genomics, Proteomics & Bioinformatics* **5**: 1-6.

Yu, J., Hu, S., Wang, J., Wong, G.K.-S., Li, S., Liu, B. et al. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**: 79-92.

Zachos, J., Pagani, M., Sloan, L., Thomas, E., and Billups, K. (2001) Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* **292**: 686-693.

Zachos, J.C., Stott, L.D., and Lohmann, K.C. (1994) Evolution of early Cenozoic marine temperatures. *Paleoceanography* **9**: 353-387.

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

Zentara, S.J., and Kamykowski, D. (1981) Geographic variations in the relationship between silicic acid and nitrate in the South Pacific Ocean. *Deep Sea Research Part A Oceanographic Research Papers* **28**: 455-465.

Zhang, Y. (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **9**: 40.

Zhang, Z., and Yu, J. (2010) Modeling compositional dynamics based on GC and purine contents of protein-coding sequences. *Biology Direct* **5**: 63.

Zhao, J.-S., Deng, Y., Manno, D., and Hawari, J. (2010) *Shewanella* spp. genomic evolution for a cold marine lifestyle and *in-situ* explosive biodegradation. *PLoS ONE* **5**: e9109.

Zhao, X., Zhang, Z., Yan, J., and Yu, J. (2007) GC content variability of eubacteria is governed by the pol III α subunit. *Biochemical and Biophysical Research Communications* **356**: 20-25.

Zhu, S.-H., and Green, B.R. (2010) Photoprotection in the diatom *Thalassiosira pseudonana*: role of LI818-like proteins in response to high light stress. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1797**: 1449-1457.

Zielinski, U., and Gersonde, R. (1997) Diatom distribution in Southern Ocean surface sediments (Atlantic sector): implications for paleoenvironmental reconstructions. *Palaeogeography, Palaeoclimatology, Palaeoecology* **129**: 213-250.

# Supplementary information

## Supplementary protocol S1

### Extraction of genomic DNA from *Fragilariopsis cylindrus* using CTAB

1. Preparation of 3% CTAB solution and heating to 65 °C.
2. Cells were spun down in a centrifuge tube. Supernatant was discarded.
3. Application of 10× volume of pre-heated CTAB solution onto cell pellet.
4. Addition of Proteinase K (e.g., 20 µL Proteinase K to 6 mL).
5. Addition of 5 µL RNase A (100 µg/mL) per mL of reaction mix.
6. Incubation for 3 h at 60 °C (± 5 °C) with shaking.
7. Addition of 1× volume Phenol/Chloroform/Isoamylalcohol (25:24:1) (pH 8).
8. Centrifugation for ≥ 60 min, ≥ 10,000 rpm (full speed) at RT.
9. Careful transfer of upper aqeous phase into fresh clean centrifuge tube to avoid carry over of the interphase.
10. Addition of ⅔ volume Isopropanol and incubation at RT ≥15 min to precipitate DNA.
11. Centrifugation for 30 min, ≥ 10,000 rpm (full speed) at 4 °C. Supernatant was discarded.
12. Addition of ice-cold 80% ethanol to fully cover DNA pellet. Centrifugation for 30 min, full speed, 4 °C. Supernatant was discarded. The washing step was performed twice.
13. Resuspension of DNA pellet in 50 – 100 µL molecular grade water.

## Supplementary protocol S2

### Extraction of total RNA from *Fragilariopsis cylindrus* using Trizol

1. Trizol was heated to 60 °C and 1 mL of pre-heated Trizol was directly applied onto frozen sample filters.

2. Addition of glass beads (425-600 µm, Sigma-Aldrich, MO, USA) to samples, followed by cell disruption using a Mini-Beadbeater (BioSpec Products, Bartlesville, OK, USA) and two disruption cycles of 60 s and 30 s.
3. Transfer of samples into 50 mL centrifuge tubes.
4. Addition of 1× volume chloroform to glass bead-trizol-cell mix and mixing by vortexing for 15 s.
5. Incubation for ≥ 5 min at RT.
6. Centrifugation for 30 min at 12,000 g, 4 °C.
7. Careful transfer of transparent upper aqueous phase into fresh 1.5 mL centrifuge tube to avoid carry-over of contaminanting debris.
8. To precipitate RNA 1× volume ice-cold Isopropanol was added and samples were mixed by vortexing for 15 s, followed by incubation for ≥ 20 min at −20 °C.
9. Samples were centrifuged for 30 min at 12,000 g, 4 °C to pellet RNA. Supernatant was discarded.
10. RNA pellet was washed twice by adding 1 mL ice-cold 75% ethanol (molecular grade), vortexing for 15 s and centrifugation for 2 min at 12,000 g, 4 °C. Supernatant was discarded.
11. Samples were incubated under sterile laminar flow hood until RNA pellets were dry and turned transparent.
12. Depending on the size of RNA pellets, 20 – 100 µL RNase/DNase-free water was added to resuspend RNA.
13. Samples were flash-frozen in liquid nitrogen and stored at −80 °C until downstream processing.

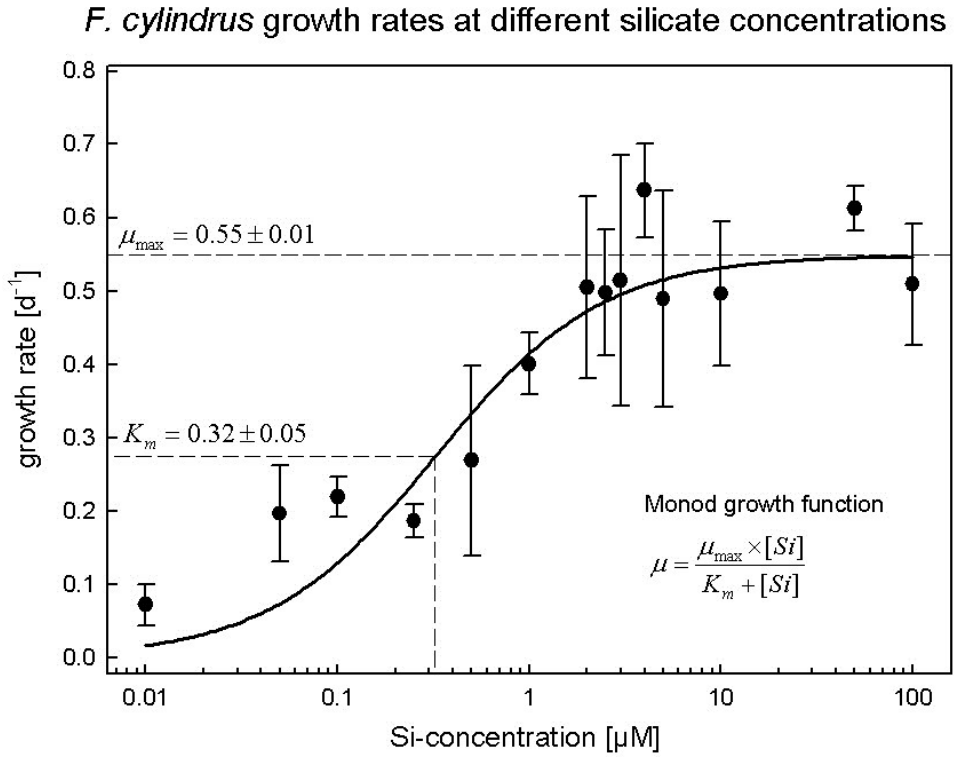
**Supplementary note S1**

The lowest iron concentrations which are achievable with artificial seawater are variable, and depend on a number of factors and include cleaning of bottles, quality of seawater salts and purity of trace metal and vitamin stocks (e.g., both trace metal and vitamin stock solutions cannot be chelexed and need to be of purist form). Iron contaminations in the range of 0.05 – 2 nmol $L^{-1}$ total iron [$Fe_T$] may be observed (A. Marchetti, *personal communication* 05/05/2009). Accordingly, computational predictions of dissolved inorganic iron concentration [Fe'] and free ferric iron concentration [$Fe^{3+}$], resulting from $Fe_T$ contaminations, may range between 0.06 – 2.3

pmol $L^{-1}$ for [Fe'] and $2.24 \times 10^{-12} - 89.6 \times 10^{-12}$ pmol $L^{-1}$ for [$Fe^{3+}$] based on chemical mass balance equations (Sunda et al., 2005). Parameters for computations of [Fe'] and [$Fe^{3+}$], including conditional stability constants $K'_{Fe'EDTA}$ and $K^*_{Fe3+EDTA}$, as well as the side reaction coefficient for inorganic complexation $\alpha_{Fe}$, were obtained from Sunda et al. (Sunda et al., 2005). They were determined for a 100 µmol $L^{-1}$ EDTA metal ion buffer system in seawater at 20 °C at pH 8.2, salinity of 36, and light intensities of 175 µmol photons $m^{-2\ -1}$ ($K'_{Fe'EDTA} = 10^{6.94}$, $K^*_{Fe3+EDTA} = 10^{17.35}$ and $\alpha_{Fe} = 2.6 \times 10^{10}$). However, for the culture experiments performed in this study, actual conditional stability constants may differ due to variations in light and growth temperatures. Although different temperatures have a limited effect on conditional stability constants, the combinatorial effect of cold temperatures and different light intensities may increase the importance of photo-dissociation of Fe-EDTA chelates (Sunda and Huntsman, 2003). Finally, if ~99% of the maximum potential iron contamination (i.e., $Fe_T$ = 2 nmol $L^{-1}$) is bound to strong organic complexes (Rue and Bruland, 1995), the resulting [Fe'] is ~23 pmol $L^{-1}$ and $\sim 8.96 \times 10^{-10}$ pmol $L^{-1}$ for [$Fe^{3+}$], rendering iron contaminations insignificant.

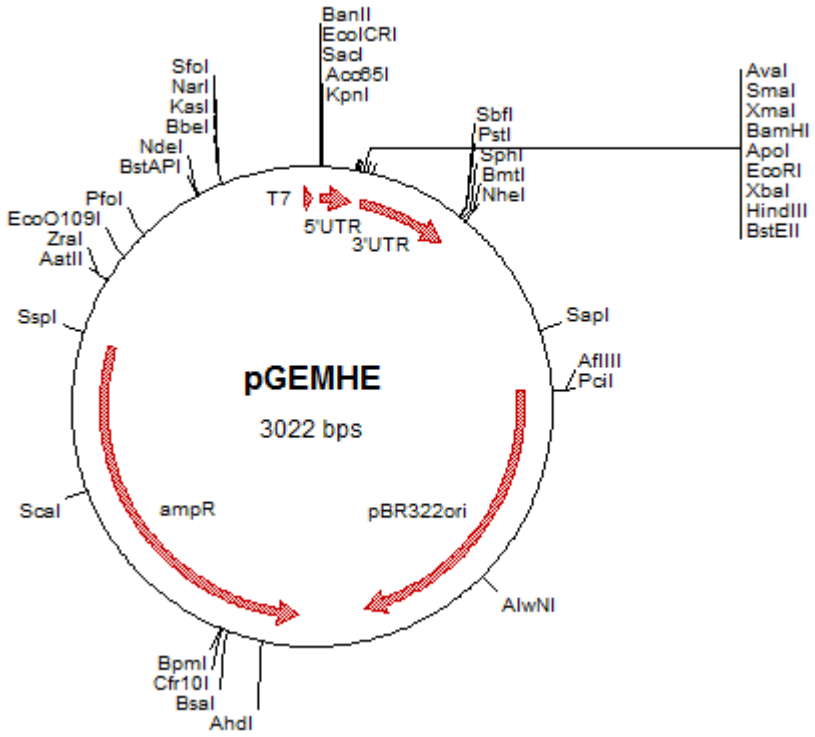
## Supplementary figures

### Supplementary Figure S1: *F. cylindrus* half-saturation constant for silicate

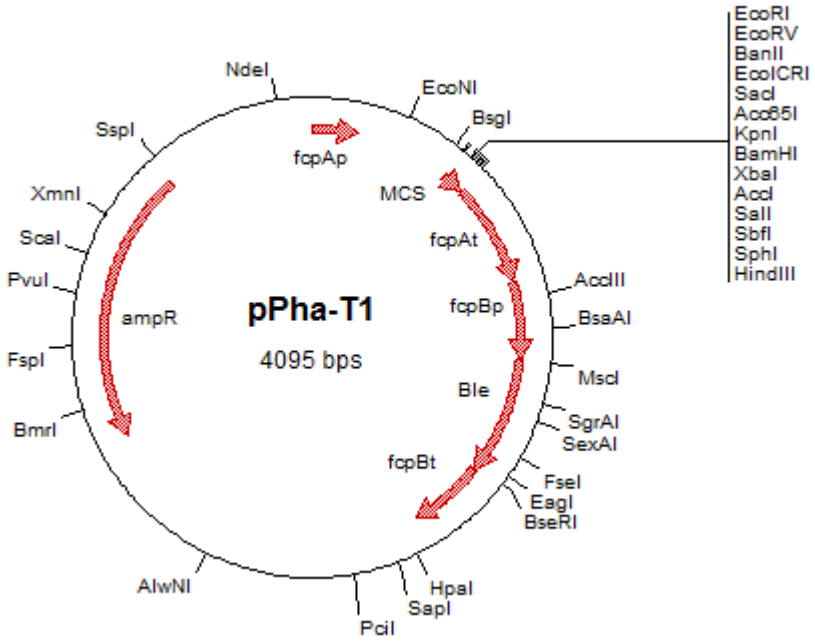



**Supplementary Figure S1.** ***Fragilariopsis cylindrus*** **mean growth rates ($d^{-1}$) in relation to different silicate [Si] concentrations (log scale). Error bars indicate standard deviation (n ≥ 3). Line represents fitting to nonlinear Monod growth function.**

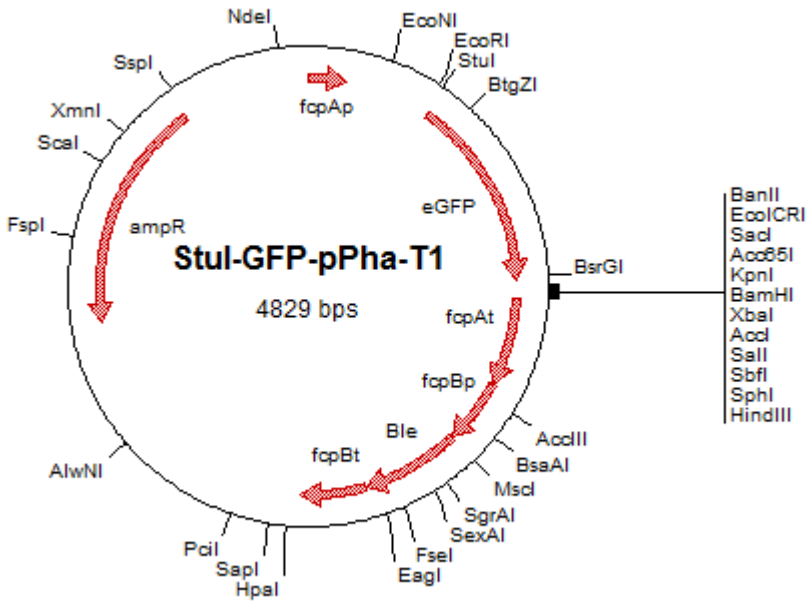
## Supplementary Figures S2-S4: Vectors used in this study




**Supplementary Figure S2.** ***Xenopus laevis*** **expression vector pGEMHE containing 5' and 3'UTRs from a** ***Xenopus*** **β-globin gene (Kreig, P. A., and Melton, D.A. (1984),** ***Nucl. Acids Res.*** **72, 7057-70), which flank a polylinker with restriction enzyme sites (Liman, E. R. et al. (1992),** ***Neuron*** **9(5): 861-71).**

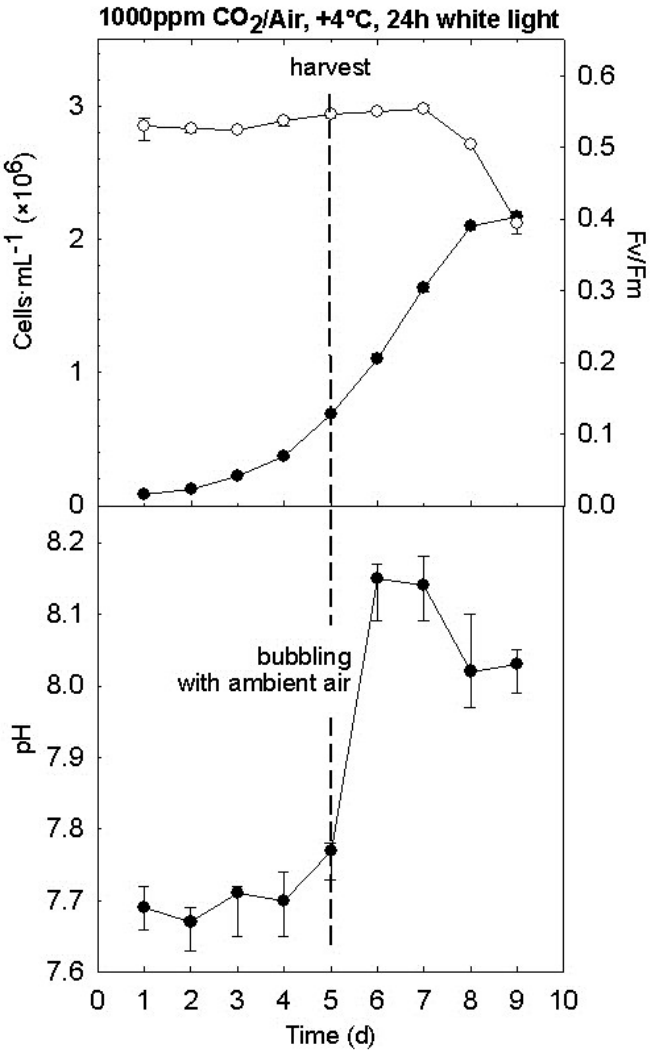



**Supplementary Figure S3.** ***Phaeodactylum tricornutum*** **transformation vector pPha-T1 (GenBank AF219942; Zaslavskaia** ***et al.*** **(2000),** ***J. Phycol.*** **36: 379) containing fucoxanthin chlorophyll-binding protein (fcp) regulatory sequences (p: promoter; t: terminator) to drive constitutive expression of bleomycin resistance protein (Ble) conferring zeocin resistance and gene of interest, which is to be cloned into multiple cloning site (MCS).**

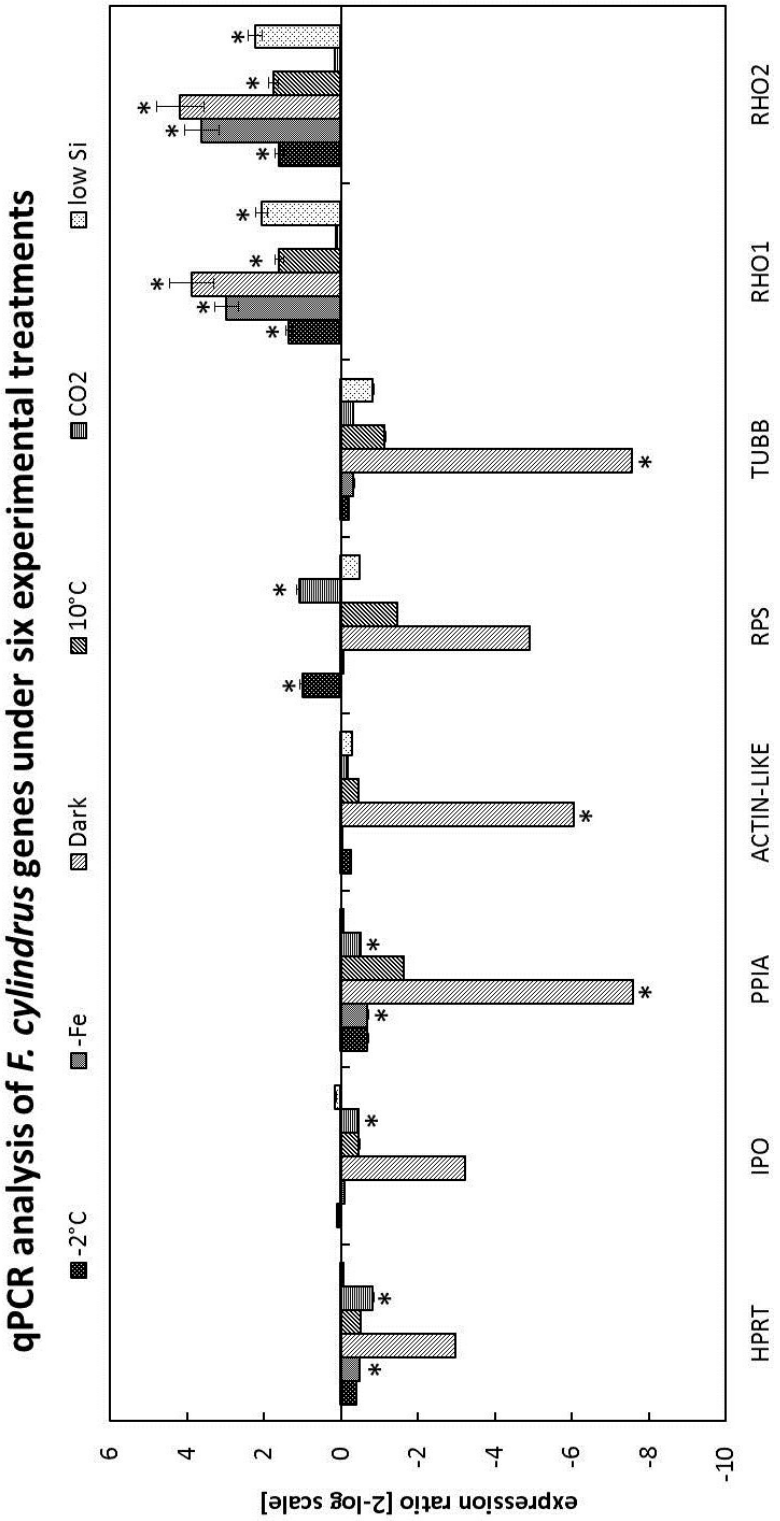



**Supplementary Figure S4.** ***Phaeodactylum tricornutum*** **transformation vector StuI-GFP-pPha-T1 (Gruber** ***et al.*** **(2007),** ***Plant Mol. Biol.*** **64: 519) for generation of eGFP fusion proteins. To generate GFP fusion constructs, the sequence of interest is to be cloned into the StuI restriction site at 5' end of eGFP. Vector represents a derivated of the pPha-T1 vector (Zaslavskaia et al. (2000),** ***J. Phycol.*** **36: 379).**

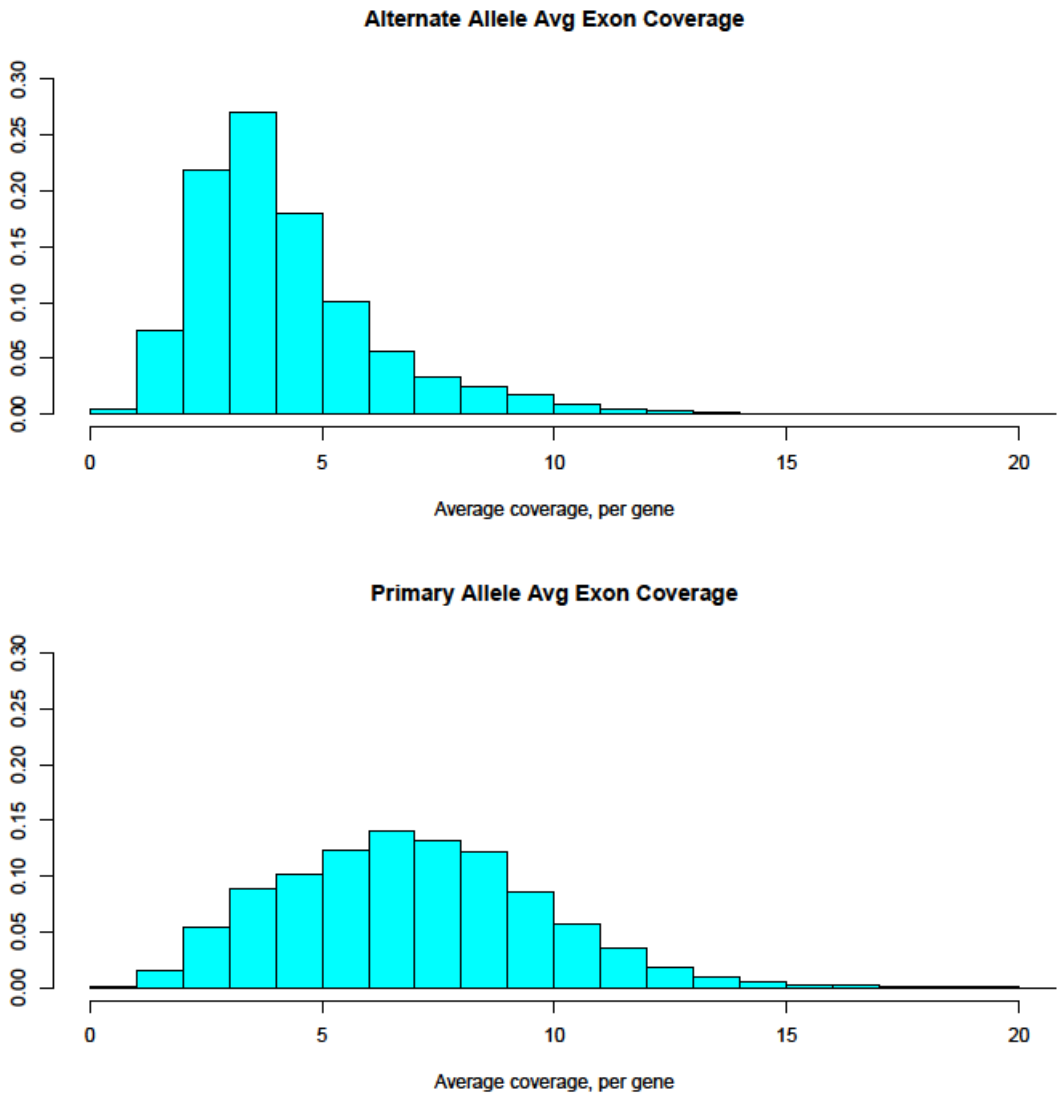
**Supplementary Figure S5: Changes in pH during growth of *F. cylindrus* under elevated (1000 ppm) $CO_2$**




**Supplementary Figure S5. Growth of *Fragilariospsis cylindrus* under elevated carbon dioxide (1000 ppm $CO_2$/Air, +4 °C, nutrient replete, 35 µmol photons $m^{-2}$ $s^{-1}$). Dashed line indicates time point of harvest for RNA extraction and shift to bubbling with ambient air. Top panel shows cell counts and photosynthetic performance ($F_v/F_m$) and bottom panel shows changes in pH.**




**Supplementary Figure S6. RT-qPCR analysis of selected *Fragilariopsis cylindrus* genes under six experimental treatments. Genes for TATA-box binding protein and RNA polymerase II were the most stable reference genes under all six experimental treatments as determined by Repeated Pair-wise Correlation Analysis using the Excel-based tool BestKeeper. A geometric mean was calculated using both genes (reference gene index). REST-MCS © – version 2 (Relative Expression Software Tool - Multiple Condition Solver) was used to test the expression of target genes under six experimental conditions, normalised by a by a reference gene index containing TBP and RNAP as reference genes. The expression ratio results of the investigated transcripts were tested for significance by a Pair Wise Fixed Reallocation Randomisation Test and plotted using standard error estimation via a complex Taylor algorithm using REST. Asterisks (*) indicate significant gene regulation compared to the control condition.**




**Supplementary Figure S7. Sequencing depth coverage of the *F. cylindrus* genome investigated for heterozygous gene copy pairs (courtesy of Robert P. Otillar, JGI, unpublished data).**