UC San Diego
UC San Diego Electronic Theses and Dissertations

Title
Coordinate regulation of carbon and energy metabolism at the transcript level in the diatom *Thalassiosira pseudonana*

Permalink
https://escholarship.org/uc/item/7282b0w1

Author
Smith, Sarah R

Publication Date
2014

Supplemental Material
https://escholarship.org/uc/item/7282b0w1#supplemental

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Coordinate regulation of carbon and energy metabolism at the transcript level in the diatom *Thalassiosira pseudonana*

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Marine Biology

by

Sarah R. Smith

Committee in charge:

Mark Hildebrand, Chair
Andrew E. Allen
Eric E. Allen
Stephen Mayfield
Brian Palenik
Immo Scheffler

2014

Copyright

Sarah R. Smith, 2014

All rights reserved.

The Dissertation of Sarah R. Smith is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2014

DEDICATION

This dissertation is dedicated to my parents, Dr. Steven C. Smith and Mrs. Kristin S. Smith, for their steadfast support and encouragement. I admire them as both parents and people, and am continually inspired by and grateful for their life-long commitment to musical, artistic, spiritual, culinary, and intellectual pursuits.

TABLE OF CONTENTS

LIST OF SUPPLEMENTAL FILES

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Chapter 1, in full, is a reprint of the material as it appears in Algal Research 2012. Smith, Sarah R.; Abbriano, Raffaela. M., and Hildebrand, Mark. “Comparative analysis of diatom genomes reveals substantial differences in the organization of carbon partitioning pathways”, Algal Research, 1: 2-16, 2012. The dissertation author was the primary researcher and author of this paper.

A very special thanks to my advisor, Dr. Mark Hildebrand, who has been patient, supportive, and encouraging during this journey. I am particularly grateful for his flexibility, which permitted me to be creative and thorough while balanced by a strong sense of security and gravity. His guidance has greatly improved the quality of the work in this dissertation and has shown me by example how to be a good scientist. I would like to thank my committee members for their support and guidance throughout my academic career. Specifically, Dr. Andrew Allen contributed to this dissertation substantially through providing many resources and stimulating conversations. Thank you to Dr. Eric Allen for always keeping it real and for invaluable advice on multiple fronts. I am grateful for the guidance from Dr. Brian Palenik whom I have long admired. Thank you Dr. Scheffler for teaching me metabolic biochemistry and for a carefully considered and refreshing perspective in committee meetings. Thank you to Dr. Steve Mayfield for straightforward advice and great questions.

Thank you to mentors at Scripps Institution of Oceanography. Especially to Dr. Vic Vacquier who has taught me so much about how science works and provided advice for a future career. Thank you for opening the gate for me on this academic path. A special thanks also to Dr. Ron Burton for your support and encouragement. Thank you to Dr. Bill Gerwick, Dr. Lena Gerwick, Dr. Greg Mitchell, Dr. Dominick Mendola, Dr. Terry Gaasterland, and others in the algae biotechnology and bioinformatics community at SIO and at UCSD who have provided a community that fosters intellectual development. I am grateful to the Hubbs Hall community,

especially the third floor group including Dr. Martin Tresguerres, Dr. Amro Hamdoun, and Dr. Bianca Brahamsha for sharing your space, your guidance, and your friendship.

I would also like to thank the Hildebrand lab group, specifically Jesse Traller and Raffaela Abbriano for friendship, laughs, and for all your help in and out of the lab. Special thanks also to Emily Trentacoste, Orna Cook, Aubrey Davis, Eva Sanchez, Trina Norden-Krichmar, Roshan and Kalpana Shrestha, and Hildebrand lab members past, and present for being great lab mates.

Thank you to my SIO friends, especially the 2008 cohort for weathering all the ups and downs together. A special thanks to my mom, dad, and sister Stefanie for their patience, advice, and encouragement.

VITA

2003	Bachelor of Science, Santa Clara University, Santa Clara
2009	Master of Science, Moss Landing Marine Laboratories, San José State University
2014	Doctor of Philosophy, University of California, San Diego

ABSTRACT OF THE DISSERTATION

Coordinate Regulation of Carbon and Energy Metabolism at the Transcript Level in the Diatom *Thalassiosira pseudonana*

by

Sarah R. Smith

Doctor of Philosophy in Marine Biology

University of California, San Diego, 2014

Mark Hildebrand, Chair

In the last decade, advances in high throughput sequencing and computational analysis of DNA and RNA have revolutionized biological research from environmental science to biomedicine. Genome and transcriptome sequences from diverse eukaryotes are now relatively easy to obtain providing unprecedented opportunities for comparative and functional genomics studies. Despite the availability of several genomes of ecologically relevant microbial eukaryotes, including the marine diatom *Thalassiosira pseudonana*, the application of sequence data to obtain a more complete understanding of the physiology, ecology, and evolution of eukaryotic marine microorganisms remains in its infancy. Practically, the relationship between sequence data and carbon and energy metabolism is of interest for biotechnological applications such as the engineering of algae for the development of renewable biofuels. The primary objective of this dissertation was to characterize aspects of the regulation of carbon and energy metabolism in *T. pseudonana* using both comparative and functional genomics approaches. In the first chapter, an analysis of the conservation of carbon partitioning enzymes was conducted by comparing three genomes of distantly related diatoms. In the second chapter, the effect of silicon starvation on the physiology and transcriptome of *T. pseudonana* was characterized to provide insight into the mechanisms by which metabolism is regulated at the transcript level. In the third chapter, the existence of a specific mode of transcriptional regulation (organization of genes into inverted gene pairs) was documented. Taken together, these chapters represent a significant advance in our understanding of mechanisms that regulate metabolism and cellular energetics in *T. pseudonana* with implications for both environmental studies and biotechnological applications.

Introduction

Phytoplankton are responsible for nearly half of the inorganic carbon fixation on earth thus playing an integral role in fueling oceanic food webs and biogeochemical cycles (Field et al. 1998). Diatoms alone are thought to be responsible for up to 40% of oceanic primary productivity making them one of the most productive classes of oceanic phytoplankton (Nelson et al. 1995). Despite their ecological importance, relatively little is known about how diatoms acquire, utilize, and store carbon at the biochemical and molecular levels. This is unfortunate since understanding carbon metabolism in diatoms is essential to both comprehend their ecological and biogeochemical significance and for applications in biotechnology.

Genomics and advances in high throughput sequencing have revolutionized biological research in a variety of fields from environmental microbiology to biomedicine and biotechnology. The sequencing of the genome of *Thalassiosira pseudonana*, a small photosynthetic diatom, was a major milestone in molecular biological oceanography (Armbrust et al. 2004). Since then, genomes of several other diatoms have been sequenced including *Phaeodactylum tricornutum, Fragilariopsis cylindrus, Pseudo-nitzschia multiseries* and *Thalassiosira oceanica* with more organisms in the pipeline (Bowler et al. 2008, Lommer et al. 2012). The analysis of these genomes both singly and in a comparative way provides information on the physiology, ecology, and evolution of phytoplankton and other marine microorganisms that is useful in understanding their role in marine ecosystems. Furthermore, genome data facilitate functional genomics studies in which more dynamic aspects of genomes (i.e. gene expression) can be investigated to better understand how genes are used to modify physiology and metabolism to adapt to changing environmental conditions. This understanding is important not only in the application of sequence data to investigating how organisms behave in complex ocean ecosystems but is useful for biotechnological applications.

Diatoms have emerged as a useful class of organisms for many applied purposes ranging from nanotechnology to the development of renewable biofuels (Bozarth et al. 2009, Hildebrand et al. 2012). Diatoms, and many algae in general, are excellent candidate organisms for renewable liquid fuels production for many reasons including their high productivity relative to terrestrial crops, lack of competition for agricultural land, and the potential to utilize wastewater as a nutrient source (Chisti 2007). Algal biofuels are currently not economically competitive with fossil fuels (Bozarth et al. 2009, Lundquist et al. 2010). While improvements to production systems may reduce the cost of algal biofuels, the greatest impact on the economics of microalgal fuels would be improvement of the productivity (biomass and oil or lipid yield) of existing strains through genetic modifications (Chisti 2007, Dunahay et al. 1996, Leon-Banares et al. 2004, Roessler 1990, Roessler 1994). Work conducted in the DOE Aquatic Species Program pioneered genetic transformation in diatoms as a means to optimize the lipid production of microalgae (Sheehan et al. 1998). A gene thought to be important in regulating lipid synthesis (acetyl-CoA carboxylase; ACCase) was introduced and overexpressed in a diatom, and while increased protein activity was demonstrated, there was no enhanced lipid accumulation phenotype suggesting that there are additional factors that regulate the flux of carbon into fatty acid biosynthesis (Dunahay et al. 1996). Labeling experiments in the diatom *Cyclotella cryptica* during a silicon starvation and lipid induction time course demonstrated that carbon is repartitioned from storage carbohydrate to lipid which highlights the importance of carbon flux regulation in the lipid accumulation process. Ultimately, this work highlights that in spite of the ability to genetically engineer diatoms, the approach is not necessarily fruitful unless it is applied within a framework of understanding native regulatory control of flux of carbon into lipid to be successful.

In autotrophic microorganisms, growth and division are major outputs for cellular carbon fixed during photosynthesis. Nutrient-starvation induced growth arrest is often concomitant with enhanced lipid accumulation suggesting that intracellular carbon flux is being redirected from growth-related processes towards lipid accumulation. How this is regulated, and whether there is a transcriptional component to this redirection is not well understood. Broadly, this thesis seeks to better elucidate the molecular mechanisms that underlie the regulation of carbon and energy metabolism in diatoms. This was done not only to better characterize fundamental aspects of diatom biology that have implications for how we study their ecology, physiology, and evolution, but to inform effective metabolic engineering approaches to enhance lipid productivity for the develoment of renewable fuels from algae, including diatoms. This was accomplished by both comparative and functional genomics approaches.

In the first chapter, the organization and conservation of enzymes in carbon partitioning pathways (i.e. glycolysis and gluconeogenesis) was characterized. Carbon partitioning pathways regulate the flux of carbon intracellularly towards several fates, including storage carbohydrates (gluconeogenesis) or towards respiration or fatty acid biosynthesis (glycolysis). Despite the fundamental role of these pathways in partitioning carbon, relatively little was known about the localization and evolutionary conservation of these enzymes in diatoms. Evolutionarily, diatoms arose through a secondary endosymbiosis in which a free-living eukaryotic red alga was engulfed and enslaved as a chloroplast, followed by large-scale transfer of the endosymbiont genetic material to the host nuclear genome (Armbrust 2009). As a result of this evolutionary history, diatom intracellular compartmentation is distinct from that of terrestrial plants and primary endosymbionts (Archibald 2009, Keeling 2010). Furthermore, their nuclear genomes are a mixture of genetic material from a variety of evolutionarily distinct sources (Armbrust et al. 2004, Lopez et al. 2005). This understanding, combined with the knowledge that many protists are known to re-organize metabolism by modifying the targeting of enzymes to different compartments (Ginger et al. 2010) was the motivation to characterize the distribution of these pathways in diatoms.

There are elements of carbon partitioning pathways that have been absolutely conserved throughout diatom diversification (cytosolic preparatory phase of glycolysis and mitochondrial payoff phase of glycolysis, Smith et al. 2012), there are aspects of these pathways that have been modified in lineage-specific ways. For example, *Phaeodactylum tricornutum* has additional copies of an ATP-phosphofructokinase, one of which is found in the chloroplast indicating that only this diatom (of the three investigated) has the capacity to completely catabolize glucose in the chloroplast (Smith et al. 2012). In the future, the expansion of this analysis to include additional diatom genomes as they are sequenced will illuminate how alterations to carbon partitioning pathways may have accompanied or even driven diatom diversification.

For the second chapter of this thesis, the transcriptomic response of *Thalassiosira pseudonana* on a time course of silicon starvation and lipid induction was investigated. Experimentally manipulated environmental stressors such as nitrogen or phosphorus limitation, low light, high light, or low temperature have been shown to increase lipid content in a variety of microalgal strains (Cohen et al. 1988, Harrison et al. 1990, Sicko-Goad 1988, Shifrin and Chisholm 1981, Tadros and Johansen 1988). Specifically, silicon starvation induces lipid accumulation in the diatoms *T. pseudonana* and *Cyclotella cryptica* (Roessler 1988, Traller et al. 2013). Transcriptomic studies have examined the response of the *T. pseudonana* transcriptome to phosphorus stress (Dhyrman et al. 2012), silicon re-addition (Shrestha et al. 2012), iron starvation (Thamatrakoln et al. 2012), and diel cycles (Ashworth et al. 2013). Mock et al. (2008) showed that silicon limitation elicits a transcriptomic response in *T. pseudonana* though the conditions tested were likely not severe enough to induce lipid accumulation. For the first time, the transcriptomic response of *T. pseudonana* on a time course of silicon starvation was characterized and resulting transcript level changes were interpreted within the context of other documented physiological changes.

During silicon starvation there are significant changes in the transcriptome of *T. pseudonana*. Generally, the up regulation of enzymes within metabolic pathways is consistent with observed physiological changes. Overall, genes involved in carbon and energy acquisition and storage, such as light harvesting, carbon fixation, and fatty acid biosynthesis are up regulated during silicon starvation whereas genes associated with growth, such as ribosomal proteins, glycolysis enzymes, and complex lipid biosynthesis are down regulated. Furthermore, a highly coordinated response of large suites of genes was documented illustrating a high degree of connectivity within regulatory networks that act at the transcript level. Finally, many of these coordinated changes were correlated with cell cycle progression, supporting the idea that in wild-type diatoms the regulation of cellular physiology and metabolism, and cellular carbon and energy inputs have been mechanistically integrated with growth. With this understanding, metabolic engineering approaches that seek to de-couple growth and aspects of metabolism such as lipid accumulation for the development of renewable biofuels can be more successful (Trentacoste et al. 2013).

The third chapter of this thesis sought to examine the functional significance of gene order in diatoms with respect to transcriptional regulation. In diatoms, it has been observed that genes that are found to be near one another are often functionally linked or co-expressed (Allen et al. 2008, Sapriel et al. 2009). Though gene order in eukaryotic genomes is generally assumed to be random due to shuffling throughout evolution, this is not always the case and genes that are found near to one another are often co-expressed through a variety of mechanisms (Hurst et al. 2004). To date, there has not been any global analysis of the functional significance of gene order in unicellular photosynthetic eukaryotes. This is only recently possible as there are an increasing number of genomes and time course transcriptomes available. A major finding of the third chapter of this thesis was that in *T. pseudonana* the organization of genes into inverted gene pairs appears to be a specific mechanism by which the co-expression of genes at the transcript level can be regulated. From a survey of diverse eukaryotic genomes, it also appears that organization of genes into inverted gene pairs is more common in small genomes and in the genomes of eukaryotes with a red-algal derived plastid. Interestingly, in *T. pseudonana,* several genes that are organized into inverted gene pairs appear to have key roles in cell cycle progression, carbon metabolism, and photosynthesis, suggesting that this mode of regulation may be important in cellular carbon and energy metabolism. In summary, the findings presented in chapter 3 ultimately raise more questions than answers, and future work should expand on this analysis.

In summary, the findings of this thesis provide insight into how diatoms regulate the flux of intracellular carbon both by compartmentalizing metabolism and by coordinating cellular carbon and energy sources and sinks at the transcript level. Furthermore, a mechanism by which specific genes are co-expressed (i.e. the organization of genes into inverted gene pairs) was documented for the first time. Overall, the research presented in this thesis represents a significant advance in our understanding of how carbon and energy metabolism are regulated in an ecologically important group of organisms (diatoms) that is evolutionarily distinct from most model eukaryotes. This understanding is not only fundamental to better appreciate the ecology, physiology, and evolution of diatoms but also to inform metabolic engineering strategies in a group of organisms that is increasingly being recognized for its value in biotechnology.

REFERENCES FOR THE INTRODUCTION

Allen, A.E., Laroche, J., Maheswari, U., Lommer, M., Schauer, N., Lopez, P.L., Finazzi, G., Fernie, A.R., and Bowler, C. (2008) Whole-cell response of the pennate diatom *Phaeodactylum tricornutum* to iron starvation. Proc. Natl. Acad. Sci. U. S. A. **105:** 10438-10443.

Archibald, J.M. (2009) The puzzle of plastid evolution. Curr. Biol. **19:** R81-R88.

Armbrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H., Zhou, S., Allen, A.E., Apt, K.E., Bechner, M., Brzezinski, M.A., Chaal, B.K., Chiovitti, A., Davis, A.K., Demarest, M.S., Detter, J.C., Glavina, T., Goodstein, D., Hadi, M.Z., Hellsten, U., Hildebrand, M., Jenkins, B.D., Jurka, J., Kapitonov, V.V., Kröger, N., Lau, W.W.Y., Lane, T.W., Larimer, F.W, Lippmeier, J.C., Lucas, S., Medina, M., Monsant, A., Obornik, M., Schnitzler Parker, M., Palenik, B., Pazour, G.J., Richardson, P.M., Rynearson, T.A., Saito, M.A., Schwartz, D.C, Thamatrakoln, K., Valentin, K., Vardi, A., Wilkerson, F.P., Rokhsar, and Rokhsar, D.S. (2004) The genome of the diatom *Thalassiosira pseudonana*: Ecology, evolution, and metabolism. Science. **306:** 79-86.

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

Ashworth, J., Coesel, S., Lee, A., Armbrust, E.V., Orellana, M.V., and Baliga, N.S. (2013) Genome-wide diel growth state transitions in the diatom *Thalassiosira pseudonana*. Proc. Natl. Acad. Sci. U. S. A. **110:** 7518-7523.

Bowler, C., Allen, A.E., Badger, J.H., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., Martens, C., Maumus, F., Otillar, R.P., Rayko, E., Salamov, A., Vandepoele, K., Beszteri, B., Gruber, A., Heijde, M., Katinka, M., Mock, T., Valentin, K., Verret, F., Berges, J.A., Brownlee, C., Cadoret, J., Chiovitti, A., Choi, C.J., Coesel, S., De Martino, A., Detter, J.C., Durkin, C., Falciatore, A., Fournet, J., Haruta, M., Huysman, M.J.J., Jenkins, B.D., Jiroutova, K., Jorgensen, R.E., Joubert, Y., Kaplan, A., Kröger, N., Kroth, P.G., La Roche, J., Lindquist, E., Lommer, M., Martin-Jézéquel, V., Lopez, P.J., Lucas, S., Mangogna, M., McGinnis, K., Medlin, L.K., Montsant, A., Oudot-Le Secq, M., Napoli, C., Obornik, M., Parker, M.S., Petit, J., Porcel, B.M., Poulsen, N., Robison, M., Rychlewski, L., Rynearson, T.A., Schmutz, J., Shapiro, H., Siaut, M., Stanley, M., Sussman, M.R., Taylor, A.R., Vardi, A., von Dassow, P., Vyverman, W., Willis, A., Wyrwicz, L.S., Rokhsar, D.S., Weissenbach, J., Armbrust, E.V., Green, B.R., Van de Peer, Y., and I.V. Grigoriev. (2008) The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. Nature. **456:** 239-244.

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

Chisti, Y. (2007) Biodiesel from microalgae. Biotechnol. Adv. **25:** 294-306.

Cohen, Z., Vonshak, A., and Richmond, A. (1988) Effect of environmental conditions on fatty acid composition of the red alga *Porphyridium cruentum:* Correlation to growth rate. J. Phycol. **24:** 328-332.

Dunahay, T.G., Jarvis, E.E., Dais, S.S., and P.G. Roessler. (1996) Manipulation of microalgal lipid production using genetic engineering. Appl. Biochem. Biotechnol. **57-58:** 223-231.

Dyhrman, S.T., Jenkins, B.D., Rynearson, T.A., Saito, M.A., Mercier, M.L., Alexander, H., Whitney, L.P., Drzewianowski, A., Bulygin, V., Bertrand, E.M., Wu, Z., Benitez-Nelson, C., and Heithoff, A. (2012) The transcriptome and proteome of the diatom *Thalassiosira pseudonana* reveal a diverse phosphorus stress response. PLOS ONE. **7:** ee33768

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

Ginger, M.L., McFadden, G.I., and Michels, P.A. (2010) Rewiring and regulation of cross-compartmentalized metabolism in protists. Phil. Trans. R. Soc. B. **365:** 831-845.

Harrison, P.J., Thompson, P.A., and Calderwood, G.S. (1990) Effects of nutrient and light limitation on the biochemical composition of phytoplankton. J. Appl. Phycol. **2:** 45-56.

Hildebrand, M., Davis, A.K., Smith, S.R., Traller, J.C., and Abbriano, R.M. (2012) The place of diatoms in the biofuels industry. Biofuels. **3:** 221-240.

Hurst, L.D., Pál, C., and Lercher, M.J. (2004) The evolutionary dynamics of eukaryotic gene order. Nat. Rev. Genet. **5:** 299-310.

Keeling, P.J. (2010) The endosymbiotic origin, diversification and fate of plastids. Phil. Trans. R. Soc. B. **365:** 729-748.

León-Bañares, R., Gonzáles-Ballester, D., Galván, A., and Fernández, E. (2004) Transgenic microalgae as green cell-factories. Trends Biotechnol. **22:** 45-52.

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

Lopez, P.J., Desclés, J., Allen, A.E., and Bowler, C. (2005) Prospects in diatom research. Curr. Opin. Biotechnol. **16:** 180-186.

Lundquist, T.J., Woertz, I.C., Quinn, N.W.T., and Benemann, J.R. (2010) A realistic technology and engineering assessment of algae biofuel production. Energy Biosciences Institute, Berkeley, California

Mock, T., Samanta, M.P., Iverson, V., Berthiaume, C., Robison, M., Holtermann, K., Durkin, C., Bondurant, S.S., Richmond, K., Rodesch, M., Kallas, T., Huttlin, E.L., Cerrina, F., Sussman, M.R., and Armbrust, E.V. (2008) Whole-genome expression profiling of the marine diatom *Thalassiosira pseudonana* identifies genes involved in silicon bioprocesses. Proc. Natl. Acad. Sci. U. S. A. **105:** 1579-1584.

Nelson, D.M., Tréguer, P., Brzezinski, M.A., Leynaert, A., and Quéguiner, B. (1995) Production and dissolution of biogenic silica in the ocean: Revised global estimates, comparison with regional data and relationship to biogenic sedimentation. Global Biogeochem. Cy. **9:** 359-372.

Roessler, P.G. (1988) Effects of silicon deficiency on lipid composition and metabolism in the diatom *Cyclotella cryptica.* J. Phycol. **24:** 394-400.

Roessler, P.G. (2004) Enviromental control of glycerolipid metabolism in microalgae: Commercial implications and future research directions. J. Phycol. **26:** 393-399.

Roessler, P.G., Brown, L.M., Dunahay, T.G., Heacox, D.A., Jarvis, E.E., Schneider, J.C., Talbot, S.G., and Zeiler, K.G. (1994) Genetic-engineering approaches for enhanced production of biodiesel fuel from microalgae. ACS Symp. Ser. **566:** 255-270.

Sapriel, G., Quinet, M., Heijde, M., Jourdren, L., Tanty, V., Luo, G., Le Crom, S., and Lopez, P.J. (2009) Genome-wide transcriptome analyses of silicon metabolism in *Phaeodactylum tricornutum* reveal the multilevel reglulation of silicic acid transporters. PLOS ONE. **4:** e7458.

Sheehan, J., Dunahay, T., Benemann, J., and Roessler, P. (1998) A look back at the U.S. Department of Energy's Aquatic Species Program – biodiesel from microalgae. National Renewable Energy Laboratory, Golden, CO. Report NREL/TP-580-24190

Shifrin, N.S., and Chisholm, S.W. (2008) Phytoplankton lipids: Interspecific differences and effects of nitrate, silicate, and light-dark cycles. J. Phycol. **17:** 374-384.

Shrestha, R.P., Tesson, B., Norden-Krichmar, T., Federowicz, S., Hildebrand, M., and Allen, A.E. (2012) Whole transcriptome analysis of the silicon response of the diatom *Thalassiosira pseudonana*. BMC Genomics. **13:** 499.

Sicko-Goad, L., Simmons, M.S., Lazinsky, D., and Hall, J. (1988) Effect of light cycle on diatom fatty acid composition and quantitative morphology. J. Phycol. **24:** 1-7.

Tadros, M.G., and Johansen, J.R. (1988) Physiological characterization of six lipid-producing diatoms from the southeastern United States. J. Phycol. **24:** 445-452.

Thamatrakoln, K., Korenovska, O., Niheu, A.K., and Bidle, K.D. (2012) Whole-genome expression analysis reveals a role for death-related genes in stress acclimation of the diatom *Thalassiosira pseudonana*. Environ. Microbiol. **14:** 67-81.

Traller, J.C., and Hildebrand, M. (2013) High throughput imaging to the diatom *Cyclotella cryptica* demonstrates substantial cell-to-cell variability in the rate and extent of triacylglycerol accumulation. Algal Res. **2:** 244-252.

Trentacoste, E.M., Shrestha, R.P., Smith, S.R., Glé, C., Hartmann, A.C., Hildebrand, M., and Gerwick, W.H. (2013) Metabolic engineering of lipid catabolism increases microalgal lipid accmulation without compromosing growth. Proc. Natl. Acad. Sci. U. S. A. **110:** 19748-19753.

Chapter 1

Comparative analysis of diatom genomes reveals substantial differences in the organization of carbon partitioning pathways

Algal Research 1 (2012) 2–16

Contents lists available at SciVerse ScienceDirect

Algal Research

journal homepage: www.elsevier.com/locate/algal

Comparative analysis of diatom genomes reveals substantial differences in the organization of carbon partitioning pathways

Sarah R. Smith, Raffaela M. Abbriano, Mark Hildebrand *

Marine Biology Research Division, Scripps Institution of Oceanography, UCSD, La Jolla, CA, United States

ARTICLE INFO

Article history:
Received 24 February 2012
Accepted 16 April 2012

Keywords:
Diatom
Glycolysis
Gluconeogenesis
Carbon metabolism
Evolution
Compartmentation

ABSTRACT

A major challenge in the development of microalgal strains for large-scale production is the optimization of biomass accumulation and production of fuel-relevant molecules such as triacylglycerol. Selecting targets for genetic manipulation approaches will require a fundamental understanding of the organization and regulation of carbon metabolic pathways in these organisms. Functional genomic and metabolomics data is becoming easier to obtain and process, however interpreting the significance of these data in a physiological context is challenging since the metabolic framework of all microalgae remains poorly understood. Owing to a complex evolutionary history, diatoms differ substantially from many other photosynthetic organisms in their intracellular compartmentation and the organization of their carbon partitioning pathways. A comparative analysis of the genes involved in carbon partitioning metabolism from *Thalassiosira pseudonana*, *Phaeodactylum tricornutum*, and *Fragilariopsis cylindrus* revealed that diatoms have conserved the lower half of glycolysis in the mitochondria, the upper half of glycolysis (including key regulatory enzymes) in the cytosol, and several mitochondrial carbon partitioning enzymes. However, some substantial differences exist between the three diatoms investigated, including the translocation of metabolic pathways to different compartments, selective maintenance and horizontal acquisition of genes, and differential gene family expansions. A key finding is that metabolite transport between intracellular compartments is likely to play a substantial role in the regulation of carbon flux. Analysis of the carbon partitioning components in the mitochondria suggests an important role of this organelle as a carbon flux regulator in diatoms. Differences between the analyzed species are specific examples of how diatoms may have modified their carbon partitioning pathways to adapt to environmental niches during the diversification of the group. This comparative analysis highlights how even core central pathways can be modified considerably within a single algal group, and enables the identification of suitable targets for genetic engineering to enhance biofuel precursor production.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Algal biofuels research is gaining momentum, fueled in part by the ease of obtaining genomic and transcriptomic information. A goal of many researchers is to mine this data to identify gene targets for manipulation that will improve growth and lipid, or more specifically, triacylglycerol (TAG), accumulation characteristics that will drive down the cost of production. There are several approaches that can be taken to enhance the TAG content in microalgae including over-expressing fatty acid or TAG biosynthesis genes, inhibiting lipid catabolism, and inhibiting metabolic pathways that compete with lipid biosynthesis for carbon intermediates such as the synthesis of storage carbohydrates [1]. Many carbon metabolic pathways are extensively studied and consequently it may be assumed that these pathways and their genes are well understood in microalgae. However, algae have an evolutionary history that is quite divergent from many model organisms, and the biochemistry of these pathways in algae is in general poorly characterized. Characterization of these pathways and their regulation in microalgae is

Abbreviations: ACCase, acetyl-CoA carboxylase; ATP-PFK, ATP-dependent phosphofructokinase; CCM, carbon concentrating mechanism; DNRA, dissimilatory nitrate reduction to ammonia; ED, Entner–Dourdoff glycolysis; EGT, endosymbiotic gene transfer; EMP, Embden–Meyerhof–Parnas glycolysis; ENO, enolase; FA, fatty acid; FBA, fructose-bisphosphate aldolase; FBP, fructose 1,6 bisphosphatase; Fru 2,6 bisP, fructose 2,6 bisphosphate; GAPDH, glyceraldehyde 3-phosphate dehydrogenase; GLK, glucokinase; GPI, glucose-6-phosphate isomerase; HGT, horizontal gene transfer; MDH, malate dehydrogenase; ME, malic enzyme; OAA, oxaloacetate; PC, pyruvate carboxylase; PDRP, pyruvate phosphate dikinase regulatory protein; PEP, phosphoenolpyruvate; PEPC, phosphoenolpyruvate carboxylase; PEPCK, phosphoenolpyruvate carboxykinase; PEPS, phosphoenolpyruvate synthase; PFK, phosphofructokinase; PF2K/F26BP, bifunctional 6-phosphofructo-2-kinase/fructose 2,6 bisphosphatase; PGAM, phosphoglycerate mutase; PGK, phosphoglycerate kinase; PK, pyruvate kinase; PPDK, pyruvate phosphate dikinase; PP_i-PFK, pyrophosphate-dependent phophofructokinase; TAG, triacylglycerol; TCA, tricarboxylic acid; TPI, triose phosphate isomerase.

* Corresponding author at: 9500 Gilman Dr., La Jolla, CA 92093-0202, United States. Tel.: +1 858 822 0167.

E-mail address: mhildebrand@ucsd.edu (M. Hildebrand).

2211-9264/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.algal.2012.04.003

essential to identify the appropriate gene targets to modify for increased productivity.

In early work to modify algal strains for improved fuel precursor molecule production, acetyl-CoA carboxylase (ACCase), the first committed step of fatty acid biosynthesis, was successfully overexpressed in the diatom *Cyclotella cryptica* [2]. Despite an increased enzyme activity, there was no resulting increase in TAG content, suggesting that there are other factors that regulate the flux of carbon into fatty acid biosynthesis. Generally, microalgae do not accumulate TAG during growth, and only develop lipid bodies during the stationary growth phase or when nutrient-limited [3,4]. Under silicon-limited TAG accumulation conditions in *C. cryptica*, it was also demonstrated that the flux of carbon was repartitioned from storage carbohydrates into lipid over the course of TAG induction [5,6]. Blocking storage carbohydrate synthesis, as demonstrated in starchless mutants of the green alga *Chlamydomonas reinhardtii*, also enhances TAG accumulation [7]. From these data, it is reasonable to expect that enhancing carbon flux towards TAG by reducing flux to competing pathways is a viable approach to improve de novo TAG biosynthesis and accumulation in algae.

Central carbon metabolic pathways are reasonable targets for the modification of intracellular carbon flux in algae. In all known photosynthetic organisms, the primary pathways involved in the partitioning of fixed carbon into either storage carbohydrates or TAG are glycolysis, gluconeogenesis, and pyruvate metabolism (Fig. 1). Glycolysis, or the catabolism of hexoses to produce pyruvate and ATP, provides the cell with energy and metabolic intermediates required to supply either the TCA cycle or fatty acid biosynthesis. Gluconeogenesis is essentially the reverse of the glycolysis pathway in that pyruvate is converted to hexoses that supply storage carbohydrate biosynthetic pathways. In this way, these pathways and their subsequent branch points act as partitioning regulators of intracellular carbon flux.

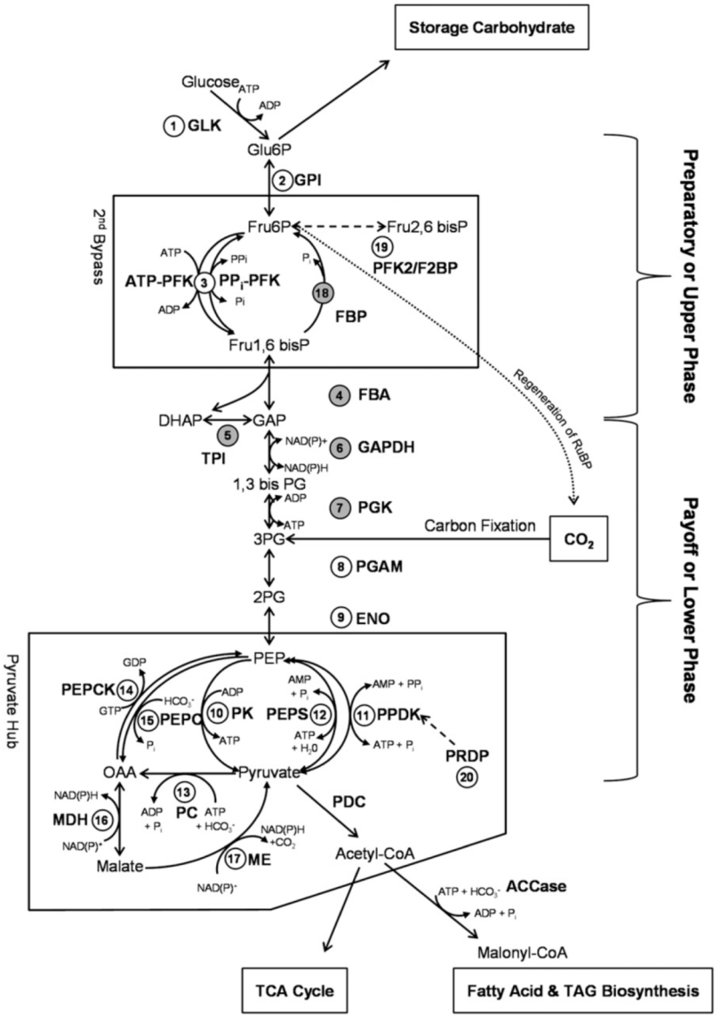


Fig. 1. Schematic of carbon partitioning enzymes involved in glycolysis, gluconeogenesis and pyruvate metabolism. Circled numbers adjacent to enzyme abbreviations correspond to those in Table 1. Shaded circles denote enzymes that are also involved in the Calvin–Benson cycle. Arrows indicate the directionality of the reaction catalyzed by each enzyme. Brackets indicate the two phases of glycolysis; the upper phase (GLK to TPI) requires an initial energy investment, while the lower phase (GAPDH to PFK) produces ATP and reducing equivalents.

From decades of research on terrestrial plants and other model organisms, it is well known that the direction of carbon flux through these core and conserved pathways is regulated by mass action (the concentration of reactants relative to products) and by dedicated glycolysis or gluconeogenesis enzymes at a few key regulatory points [8,9]. The regulatory points are known as glycolysis bypass points (glycolysis reactions must be "bypassed" by dedicated gluconeogenic enzymes; Fig. 1). The directionality and magnitude of metabolite flux through these bypass points is tightly regulated by a combination of fine (enzyme phosphorylation, allosteric effectors) and coarse (transcription, translation, and protein turnover) controls [9]. Glycolysis bypass points are appropriate targets for genetic engineering since they are natural points of regulation, and control the flux of carbon in one favorable direction. Another important mechanism of carbon flux regulation in eukaryotes is their ability to adjust the concentration of metabolic intermediates in sub-cellular compartments, thereby affecting the supply of these intermediates into various pathways. Compartmentation makes it possible for eukaryotic cells to run competing pathways simultaneously, regulate cellular and organellar energetics, and control the rate of intermediate supply to other compartments via specialized transporters [10–12]. Considering the latter point, the transport of metabolites between compartments can be important regulatory steps.

Compartmentation in diatoms is significantly different from green algae and terrestrial plants due to differences in their respective evolutionary histories [13,14]. Photosynthetic eukaryotes arose via a primary endosymbiosis event where chloroplasts were derived from a cyanobacterial endosymbiont (Fig. 2). The progenitor plant cell gave rise to the glaucophytes, green algae (and eventually plants derived from this lineage) and red algae (Fig. 2). Diatoms and other chromalveolates arose through a secondary endosymbiotic event, in which a red algal endosymbiont was enslaved as a plastid, followed by the eventual loss of the red algal nucleus and mitochondria (Fig. 2). One consequence of the complex evolutionary history of secondary endosymbionts is the distinct chloroplast membrane organization [15,16]. In addition to the chloroplast inner and outer membranes typical of all photosynthetic eukaryotes, diatom (and other chromalveolate) plastids are surrounded by a periplastid membrane (the relic plasma membrane of the red algal symbiont), which defines the periplastid compartment (PPC), and the endoplasmic reticulum, commonly called the "chloroplast ER" (Fig. 2, [17]). This additional complexity adds to the challenge of reconstructing algal metabolic networks.

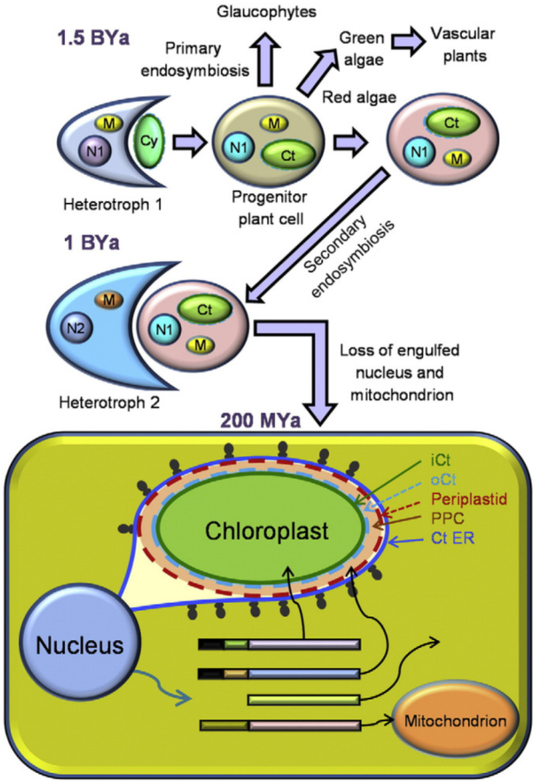


Fig. 2. The primary and secondary endosymbiotic events that gave rise to modern diatoms. The upper panel shows the evolutionary progression leading to the heterokonts. Organelles are colored to denote different origins and labeled as such (N = nucleus, M = mitochondria, Ct = chloroplast, Cy = cyanobacterium). The lower panel is a diagram of diatom intracellular compartmentation, with an emphasis on the chloroplast and associated extra membranes relative to the progenitor plant cell. Colored bars represent proteins with different leader sequences targeting them to different intracellular locations. The inner chloroplast membrane (iCt), outer chloroplast membrane (oCt), periplastid membrane, periplastid compartment (PPC), and chloroplast ER (ct ER) are labeled.

As secondary endosymbionts, the nuclear genomes of diatoms are a combination of genes from several evolutionarily distinct organisms, and analysis of diatom genomes has revealed both plant-like and animal-like features [18,19]. The genomes of both non-photosynthetic and photosynthetic protists, including diatoms, are known to have been shaped by the acquisition of genes from endosymbiotic/horizontal gene transfer events and duplication events, as well as by the selective deletion of certain genes or gene families [20,21]. Furthermore, there is evidence that some organisms have re-targeted nuclear-encoded genes to distinct organelles, and in some cases have single isozymes that can be dually targeted to more than one subcellular location, adding to the flexibility of their pathway compartmentation significantly [22]. Additionally, diatoms seem to possess several distinct genes encoding isozymes of conserved metabolic pathways, indicating that they have acquired genes from several different sources and that these isozymes may be functionally differentiated [23]. Analysis of the genome of the first diatom sequence available, *Thalassiosira pseudonana*, first identified enzymes for the complete cytosolic glycolysis and gluconeogenesis pathways [18]. Later work investigating the genome of *Phaeodactylum tricornutum* suggested that all the reactions of glycolysis might also occur in diatom plastids [23]. Additionally, Kroth et al. showed that there are isozymes for the complete lower half of glycolysis (TPI to PK) predicted to be mitochondrially located, a finding that supported earlier work demonstrating the presence of a triose phosphate isomerase (TPI)/glyceraldehyde 3P dehydrogenase (GAPDH) fusion protein in diatom mitochondria [23,24]. The evolutionary divergence of the centric and pennate classes of diatoms occurred at least 90 mya (based on the fossil record), yet the genome of *T. pseudonana* (centric) and *P. tricornutum* (pennate) differ to the same extent as those of fish and mammals, which diverged approximately 550 mya [25]. Diatoms evolved over a time of substantial environmental change on the planet, particularly with regards to CO_2 and O_2 levels [26,27], and it is reasonable to assume that some of the differences distinguishing different classes of diatoms are adaptations for optimal productivity under the particular conditions during which they arose. There have been substantial differences described for carbon concentrating mechanisms in different classes of diatoms [28,29], and recently, a phosphoketolase pathway was identified in *P. tricornutum*, but not in *T. pseudonana* [30] consistent with evolutionarily based alterations in fundamental aspects of carbon metabolism. Despite previous valuable insights, a comprehensive comparative analysis of the homology and targeting of the carbon partitioning enzymes and the extent to which diatoms differ from one another has not been conducted. The extent to which the core carbon partitioning proteome has been modified by the selective maintenance or deletion of genes, duplications, re-targeting, or acquisition via horizontal gene transfer is not well-understood but should provide insight into unique adaptive metabolic capabilities characteristic of a group or species.

The primary aim of this study was to utilize the available genome sequences for the centric *T. pseudonana*, the raphid pennate *P. tricornutum*, and the assembly scaffolds for the psychrophilic raphid pennate diatom *Fragilariopsis cylindrus*, to perform a comprehensive in silico evaluation of the carbon partitioning proteome (Fig. 3). The analysis resulted in the identification of a rich variety of features, including annotation of absolutely core metabolic genes (likely to have an indispensable metabolic function), characterization of gene duplication events, description of unique genes arising from selective deletion, evidence of genes acquired by horizontal gene transfer, and differential intracellular targeting, and enabled the reconstruction of metabolic networks, with an emphasis on intracellular compartmentation. The analysis highlights distinctions between representatives of the centric and pennate diatom lineages, and enables generalizations to be made about diatom carbon metabolism that distinguishes them from known model organisms and other algae. The information gained from this type of analysis is essential to inform metabolic engineering approaches to improve precursor molecule production for biofuel applications.

2. Materials and methods

2.1. Sequence screening and functional annotation

KEGG Pathway Database and Gene Ontology annotations were used to identify protein sequences for all genes of interest from the genomes of *T. pseudonana* v3.0 (http://genome.jgi-psf.org/Thaps3/Thaps3.home.html), *P. tricornutum* v2.0 (http://genome.jgi-psf.org/Phatr2/Phatr2.home.html), and *F. cylindrus* v1.0 (http://genome.jgi-psf.org/Fracy1/Fracy1.home.html). It is common that gene models in these genomes are not full-length, therefore in all cases, protein models were checked to ensure they were extended to the first in-frame methionine, and occasionally EST data was used to manually obtain the full ORF. These manual modifications are noted in the supplemental table (Table S1). To verify that no sequences were missed, any proteins identified from the annotations were queried against the other diatom genomes using BLAST [31]. In many cases, predicted gene models with significant homology but lacking annotations were identified (Table S1). When appropriate, functional domains that were used as criteria to identify predicted proteins were included in the supplemental table (Table S1). Some enzymes with weak annotated functions did not meet our criteria as the functional protein of interest and are not included in Table 1 but are included in the supplemental table (Table S1).

2.2. Bioinformatics-based targeting analysis

Several bioinformatics software programs were used for intracellular targeting prediction. For general localization predictions, the programs HECTAR, Predotar, and TargetP were used [32–34]. The presence of mitochondrial target peptides was determined using Mitoprot [35]. Diatom plastid-targeted peptides must cross four membranes, so their plastid pre-sequences are distinct from green algae and plants [36,37]. Identification of plastid-targeted proteins based on sequence data can

Table 1
Number of genes encoding enzymes involved in carbon partitioning metabolism (and their putative metabolic pathways) in the genomes of *T. pseudonana*, *P. tricornutum*, *F. cylindrus*, and *Chlamydomonas reinhardtii*. Gray shadowing indicates the occurrence of more than three isoenzymes in any of the diatom genomes. Gly = glycolysis, GNG = gluconeogenesis, Pyr = pyruvate metabolism, CB = Calvin–Benson cycle, TCA = tricarboxylic acid cycle, C4 = C4-type photosynthesis.

#	Enzyme name	EC number	Putative roles in metabolism						Total #	# of Isoenzymes/genome			
			Gly	GNG	Pyr	CB	TCA	C4	genes	T. pseudo	P. tricorn	F. cylin	C. reinh.
1	HXK	2.7.1.1	x	––	––	––	––	––	0	0	0	0	1
1	GLK	2.7.1.2	x	––	––	––	––	––	1	1	1	1	1
2	GPI	5.3.1.9	x	x	––	––	––	––	4	4	4	4	1
3	ATP -PFK	2.7.1.11	x	––	––	––	––	––	3	1	3	1	1
3	PP$_i$-PFK	2.7.1.90	x	x	––	––	––	––	1	1	1	1	2
4	FBA	4.1.2.13	x	x	––	x	––	––	6	4	5	6	4
5	TPI	5.3.1.1	x	x	––	x	––	––	5	3	3	5	1
5/6	TPI/GAPDH	5.3.1.1/1.2.1.12	x	x	––	––	––	––	1	1	1	1	0
6	GAPDH	1.2.1.12/1.2.1.13	x	x	––	x	––	––	5	4	4	4	4
7	PGK	2.7.2.3	x	x	––	x	––	––	6	5	3	3	2
8	PGAM	5.4.2.1	x	x	––	––	––	––	10	6	7	7	1
9	ENO	4.2.1.11	x	x	––	––	––	––	3	2	2	3	1
10	PK	2.7.1.40	x		x	––	––	––	8	5	7	5	5
18	FBP	3.1.3.11	––	x	––	x	––	––	6	3	5	7	1
11	PPDK	2.7.9.1	x	x	x	––	––	x	1	1	1	1	2
12	PEPS	2.7.9.2	x	x	x	––	––	––	3	1	0	2	0
13	PC	6.4.1.1	––	x	x	––	x	––	4	3	2	2	1
14	PEPCK	4.1.1.32	––	x	x	––	––	x	1	1	1	1	1
15	PEPC	4.1.1.31	––	––	x	––	x	x	2	2	2	2	1
16	NAD(P)-MDH	1.1.1.37/1.1.1.82	––	––	x	––	x	x	2	2	1	1	4
17	NAD-ME	1.1.1.39	––	––	x	––	––	x	1	1	1	1	1
17	NADP-ME	1.1.1.40	––	––	x	––	––	x	1	0	1	0	4
19	P2FK/F2BP	3.1.3.46/2.7.1.105	x	x	––	––	––	––	2	2	2	2	2
20	PRDP	N/A	x	x	x	––	––	––	2	0	1	1	0
								Total:	74	51	55	58	N/A
									Conserved in all diatoms:	40, 78%	40, 73%	40, 69%	N/A
									Unique to genome:	7, 14%	8, 15%	10, 17%	N/A
									Found in 1 other genomes:	5, 8%	7, 13%	8, 14%	N/A

be challenging in green algae [38], however, the requirement for an ER signal peptide facilitates these predictions in diatoms. Additionally, a motif has been detected for targeting proteins to the periplastid compartment in diatoms [39]. To detect plastid and periplastid compartment targeting, the programs SignalP v4.0 and SignalP v3.0 were used along with manual inspection of the predicted cleavage site [39–41]. The program ChloroP was used to detect the presence of predicted chloroplast transit domains [42]. In cases where individual targeting predictions did not agree with one another, the localization was predicted based on weighing the relative predictions. The results of all the predictions are summarized in Table S1. In diatoms, several predictions using the bioinformatics approaches we outline have been confirmed experimentally [24,37,39,43,44]. In some cases there were high predictions for mitochondria and also for the presence of a signal peptide and/or chloroplast targeting. A signal peptide would be a default for chloroplast targeting, however, it is possible that different transcription starts or splicing could eliminate the signal peptide and/or chloroplast targeting sequences [45,46]. Because the mitochondrial predictions were so high, these enzymes were stated to be potentially dually targeted.

2.3. Phylogenetic analysis and criteria for homology and monophyly

Homologs (enzymes that catalyze the same reaction but are not necessarily evolutionarily related) which were identified in all three diatom genomes were aligned to determine sequence similarities and to assess phylogenetic relationships. The web-based ClustalW2 at the European Bioinformatics Institute (www.ebi.ac.uk/Tools/msa/clustalw2/) was used to generate pairwise output scores (Table S1). The slow pairwise alignment option and otherwise default multiple sequence alignment options were selected. Sequences homologous to the diatom enzymes were identified by searching the nonredundant GenBank CDS database (nr). Sequences were aligned with ClustalX 2.1, and RAxML-HPC BlackBox (7.2.8, Cipres Science Gateway, www.phylo.org) was used to generate protein maximum likelihood trees. Trees were used to identify monophyletic clusters of diatom genes. Monophyletic clusters of diatom genes found in all three diatom genomes were considered orthologs (genes that are derived from an enzyme found in the last common ancestor of diatoms). Genes that arose through an apparent duplication event are considered paralogs. When the duplication occurred prior to the evolutionary divergence of the different diatoms, these duplications were considered out-paralogs. Duplications that occurred after the divergence (and are only found in one diatom for example) are considered in-paralogs.

2.4. Nomenclature

Monophyletic clusters of diatom enzymes were arbitrarily numbered because at the outset it was not possible to assign functional differences or their relative importance in metabolism. In Table S1, protein ID numbers (PIDs) are given for each gene cluster. The PIDs are prefixed with Tp, Pt, or Fc to indicate *T. pseudonana*, *P. tricornutum*, and *F. cylindrus* respectively. In the case of FBA, the nomenclature from previous studies was maintained [23,44].

3. Results and discussion

3.1. Overview of glycolysis, gluconeogenesis, and pyruvate metabolism

Although glycolysis, gluconeogenesis, and pyruvate metabolism may be familiar core metabolic pathways, the details of their regulation are not as familiar; therefore we present an overview of the pathways and regulatory steps here to set the stage for subsequent analyses (Fig. 1). Most steps of glycolysis are carried out by enzymes that function bi-directionally in the pathway. These steps tend to be regulated by mass action, allostery, or post-translational modification. Steps that play key regulatory roles generally involve enzymes catalyzing unidirectional reactions.

Overall, glycolysis is an energy-generating pathway, but the pathway is divided into two main phases (Fig. 1). The preparatory, or upper phase of glycolysis (GLK to TPI) involves an initial investment of ATP in order to ultimately generate ATP and NADH in the payoff, or lower phase (GAPDH to PK). Gluconeogenesis is not divided into preparatory or payoff phases since all of the reactions in the pathway are either at equilibrium (freely reversible) or require energy, but can be divided into upper and lower phases. The production of metabolites (such as pyruvate for fatty acid biosynthesis) not only involves carbon flux, but also is intimately connected with cellular energetics. In order to prevent the wasteful consumption of ATP, cells must achieve a balance between the supply of biosynthetic precursors and their energy budget through careful regulation of the competing pathways of glycolysis and gluconeogenesis. The key regulatory steps in glycolysis (Fig. 1) are catalyzed by glucokinase (GLK), phosphofructokinase (ATP-PFK, PP$_i$-PFK), and pyruvate kinase (PK), which are reversed by the committed enzymes of gluconeogenesis. The first committed reaction of gluconeogenesis is the conversion of pyruvate to phosphoenolpyruvate (PEP, Fig. 1). This bypass can be accomplished in several ways in different organisms. In mammals, gluconeogenesis is initiated from pyruvate by the combined activities of pyruvate carboxylase (PC) and phosphoenolpyruvate carboxykinase (PEPCK). In plants, gluconeogenesis at the first bypass is initiated with the intermediate oxaloacetate (during seed germination) and therefore requires only PEPCK [47]. Alternatively, plants use a pyruvate phosphate dikinase (PPDK) to directly initiate gluconeogenesis [47]. Finally, bacteria and archaea are known to use both PPDK and phosphoenolpyruvate synthase (PEPS) in gluconeogenesis [48]. Clearly, there is considerable flexibility in the organization of the first bypass across the domains of life. In addition to some regulatory capacity over the direction of the glycolysis/gluconeogenesis pathway, these enzymes may govern flux into other pathways since they exist at the intersection of

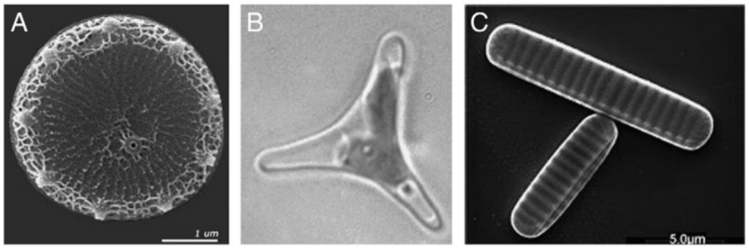


Fig. 3. Diatom species with sequenced genomes. A) The multipolar centric diatom *Thalassiosira pseudonana* has a cosmopolitan distribution and a relatively small genome (34Mb). B) The raphid pennate diatom *Phaeodactylum tricornutum* (30Mb) is not naturally abundant, but is a good model organism for laboratory studies in diatom physiology, morphogenesis, and silicification. C) The raphid pennate diatom *Fragilariopsis cylindrus* is abundant in polar regions and is a model organism for studying tolerance to extreme environmental conditions. Images A&B are provided courtesy of the DOE Joint Genome Institute and image C is provided courtesy of Henrik lange and Gerhard Dieckmann, Alfred-Wegener Institute for Polar and Marine Research, Germany.

glycolysis/gluconeogenesis with the TCA cycle, fatty acid biosynthesis, and amino acid metabolism. Since they may have a variety of different and variable roles in metabolism, defining them as glycolysis bypass enzymes could be misleading. Therefore, the direct conversion enzymes, along with PEPC, PEPCK, PC, MDH, and ME, which also involve the interconversion of 3-carbon and 4-carbon intermediates, will here be referred to as the pyruvate hub instead of the first bypass (Fig. 1).

The second bypass step (Fig. 1) in the glycolysis direction (fructose 6 phosphate to fructose 1, 6 bisphosphate) can be catalyzed by either an ATP-dependent phosphofructokinase (ATP-PFK) or a pyrophosphate-dependent phosphofructokinase (PP_i-PFK). The gluconeogenic reaction (Fig. 1) is catalyzed by fructose 1,6 bisphosphatase (FBP). In eukaryotes, the second bypass is reciprocally regulated by a potent allosteric effector molecule, fructose 2,6 bisphosphate (Fru 2,6 bisP). Typically, Fru 2,6 bisP activates glycolytic ATP-PFK while acting as a competitive inhibitor of gluconeogenic FBP [49]. Interestingly, plant ATP-PFK is insensitive to Fru 2,6 bisP, and PP_i-PFK is responsive instead [50]. In diatoms, it is unknown whether Fru 2,6 bisP activates ATP-PFK or PP_i-PFK. Fru 2,6 bisP is produced and degraded from the glycolytic intermediate fructose 6-phosphate by the activity of a bi-functional 6-phosphofructo-2-kinase/fructose 2,6 bisphosphatase (PF2K/F2BP). The only known function of Fru 2,6 bisP is to act as an allosteric effector of the second bypass, and therefore the concentration of Fru 2,6 bisP is an important determinant of the direction of carbon flux into either glucose oxidation or carbohydrate synthesis.

The final unidirectional glycolysis step is catalyzed by glucokinase (GLK). However, a glucose-6-phosphatase, the gluconeogenic enzyme which would reverse the activity of GLK, has not been identified in diatoms [23]. Thus, there is no third bypass in diatoms, presumably since glucose in its non-phosphorylated form is not an important metabolite in these organisms [23].

3.2. The conservation of diatom carbon metabolism genes

Examination of the three diatom genomes identified a total of 164 genes encoding enzymes with putative roles in carbon metabolism pathways, with 51 of these genes found in *T. pseudonana*, 55 in *P. tricornutum*, and 58 in *F. cylindrus* (Table 1). Genes were clustered according to sequence similarity to identify homologs (Table S1), and phylogenies were constructed to validate monophyletic clusters of diatom genes. Monophyletic clusters were given a gene name, and were considered to be orthologs (inherited from the most recent common ancestor) if present in at least two diatom genomes. There were 40 orthologs identified in all three diatom genomes which comprised the core carbon partitioning proteome (Fig. 4). The majority of the carbon partitioning enzymes in a given diatom genome belong to this core proteome with 78%, 73%, and 69% in *T. pseudonana*, *P. tricornutum*, and *F. cylindrus* respectively. Most of the enzymes that belong to the core proteome (33/40, or 83%) were predicted to be targeted to the same sub-cellular location, a finding that lends some support to the accuracy of the localization predictions.

In the core proteome, 7 of the 40 orthologs (17%) were found to not share targeting predictions (Fig. 4). In most of these cases, it is not clear whether this was because the enzymes are truly differently targeted, or whether there were inaccuracies in the prediction software or incorrect gene models. For example, the glycolysis enzyme phosphoglycerate mutase 5 (PGAM5) is predicted to be targeted to the chloroplast in both *T. pseudonana* and *P. tricornutum*, whereas it does not have a predicted chloroplast localization in *F. cylindrus* (Table S1). Both N-termini of the *T. pseudonana* and *P. tricornutum* models are supported by EST data, however the *F. cylindrus* model is likely erroneously predicted owing to the presence of a 666 nucleotide intron directly following the predicted start codon. Despite the possibility that the correct ORFs were not used to make the targeting predictions, there are two cases where gene models that are well-supported by EST data in *T. pseudonana* and *P. tricornutum* give different targeting predictions. These include one of the enolase orthologs (ENO2) and pyruvate phosphate dikinase (PPDK), which are both predicted to be chloroplast-localized in the pennate diatoms but not in *T. pseudonana*. Since these enzymes catalyze two sequential reactions of the glycolysis/gluconeogenesis pathway, they may represent a translocation of a portion of this metabolic pathway following the divergence of centric and pennate diatoms, and are an example of how core metabolism has been rearranged in diatoms. This metabolic rearrangement will be discussed more in later sections.

Carbon partitioning genes that are not common to all diatom genomes are specific examples of enzymatic steps that have been modified following the diversification of diatoms and may confer some metabolic adaptation specialized for a given species or class. These modifications include species or class-specific duplications (in-paralogs), the selective maintenance or deletion of orthologs, and variable horizontal acquisition of foreign genes. A significant portion (23–31%) of genes in any given genome is either unique to that species or found only in one other diatom genome, indicating that carbon metabolic pathways are not static in diatoms and that there have been several adjustments throughout the evolution of modern species (Table 1, Fig. 4). The next several sections will explore both the common features of the organization and regulation of diatom carbon partitioning metabolic pathways as well as the features of these pathways that appear to be more flexible within the realm of diatom diversity represented by the sequenced genomes.

3.3. Enzymes of carbon metabolism exist as several isozymes

Many of the enzymes involved in carbon metabolism exist as a number of isozymes in the three diatom genomes (Table 1). Diatoms have on average twice as many enzymes involved in glycolysis/gluconeogenesis than were found in the *C. reinhardtii* genome, and additional analysis suggests that this feature is conserved in other green algal genomes (Table 1, and unpublished observations). A possible explanation for this is that unlike green algae, diatoms are secondary endosymbionts, and could have acquired many of these additional isozymes from endosymbiotic gene transfer (EGT). However it is also possible that the additional isozymes found in both the core and accessory proteomes of diatoms have arisen through duplications or through horizontal gene transfer events (HGT). Regardless of the origin of these additional enzymes, many seem to have been maintained in diatoms at least since the divergence of centrics and pennates, indicating that they are useful and may be functionally differentiated (for example GPI). In multicellular organisms, variations in isozyme form and function may be useful at different times during development or in differentiated tissues [51]. In unicellular organisms, different isozyme forms may be useful under certain environmental conditions or may be targeted to different sub-cellular location or organelles. The high number of isozymes found in these pathways in diatoms suggests flexibility or optimization in regulating carbon metabolism.

3.4. The organization of carbon partitioning in diatoms

Isozymes of the glycolysis/gluconeogenesis pathways and the pyruvate hub are predicted to be distributed across several sub-cellular compartments, including the plastid, periplastid compartment, cytosol, and mitochondria. There are several conserved features of this organization that enable generalizations regarding the organization of carbon metabolism in diatoms (Fig. 5). First, all three diatoms have conserved orthologs of the lower half of the glycolysis pathway (TPI to PK) that are predicted to be targeted to the mitochondria [22,23]. Second, only a partial cytosolic glycolysis pathway is conserved in all three diatoms (Fig. 5). Specifically, orthologs could be identified for the upper half and final step of the cytosolic glycolysis pathway, but not for the mid-payoff phase (reactions catalyzed by PGK, PGAM, and ENO). Several

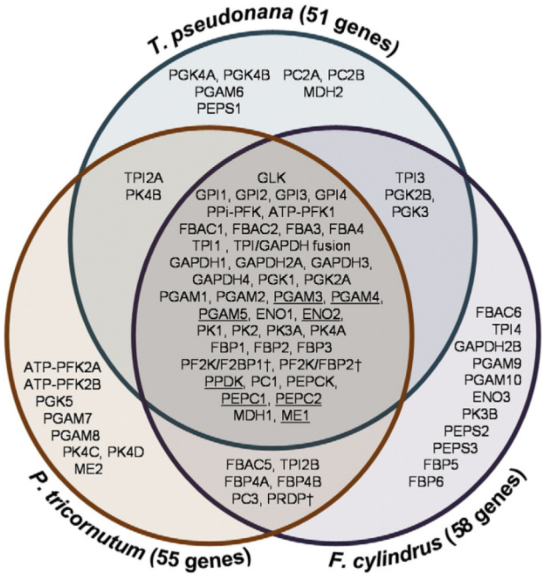


Fig. 4. Venn diagram of glycolysis, gluconeogenesis, or pyruvate hub enzymes showing the proportion of shared genes (the core proteome), genes found in two genomes, and unique genes in the genomes of *T. pseudonana*, *P. tricornutum*, and *F. cylindrus*. See Table 1 for enzyme identifications. † Regulatory enzyme, not included in the tallies for carbon partitioning enzymes. Underlined enzymes are orthologs for which the targeting predictions are not shared. See text and Table S1 for details.

gluconeogenesis and regulatory enzymes are conserved in diatoms, particularly at the second bypass. Finally, there are several unifying features in the organization of pyruvate metabolism that illustrate the role of the mitochondria as a regulatory hub for the distribution of intracellular organic carbon, but there are many features that suggest that the pyruvate hub is a metabolic network that has been subject to modification as diatoms diversified.

3.4.1. Metabolism in the mitochondria

Mitochondrial glycolysis is currently believed to occur only in diatoms and non-photosynthetic oomycetes, which remain relatively understudied with respect to metabolism as compared to mitochondria in conventional model organisms [22]. Consequently, there is little known about the origin or function of this pathway [23,24]. The first enzyme of the diatom Embden–Meyerhoff–Parnas (EMP) mitochondrial glycolysis pathway is a TPI-GAPDH fusion protein, which has been experimentally localized to mitochondria in *P. tricornutum* with immunogold labeling [24]. Orthologs of this TPI-GAPDH fusion, along with enzymes that catalyze the complete lower phase of glycolysis (PGK to PK) predicted to be mitochondrially targeted were identified in all three diatoms with a remarkable degree of conservation (Fig. 5). There have been very few duplication events, substitutions, or variable acquisitions of enzymes involved in mitochondrial glycolysis.

Interestingly, slightly upstream of the TPI-GAPDH fusion, on the opposite strand in the reverse orientation, a duplicated mitochondrially targeted GAPDH (GAPDH1) is found in an arrangement that is strikingly well conserved in all three diatom genomes. The conservation of this gene order is consistent with their coordinated regulation using a bidirectional promoter in the intergenic region [52]. This arrangement is apparently found only in diatoms, since a TPI-GAPDH fusion protein was not identified in the genome of the related photosynthetic heterokont *Aureococcus anophagefferens*, and though the TPI-GAPDH fusion protein is also found in oomycetes like *Phytophthora sojae* (PID: 285408, JGI v3.0), there was no adjacent GAPDH found in an inverted orientation. Taken together, it seems likely diatoms have a unique way to coordinately regulate mitochondrial TPI-GAPDH and GAPDH, and that both of the triose phosphate metabolic intermediates DHAP and GAP can be a starting point for mitochondrial glycolysis in diatoms.

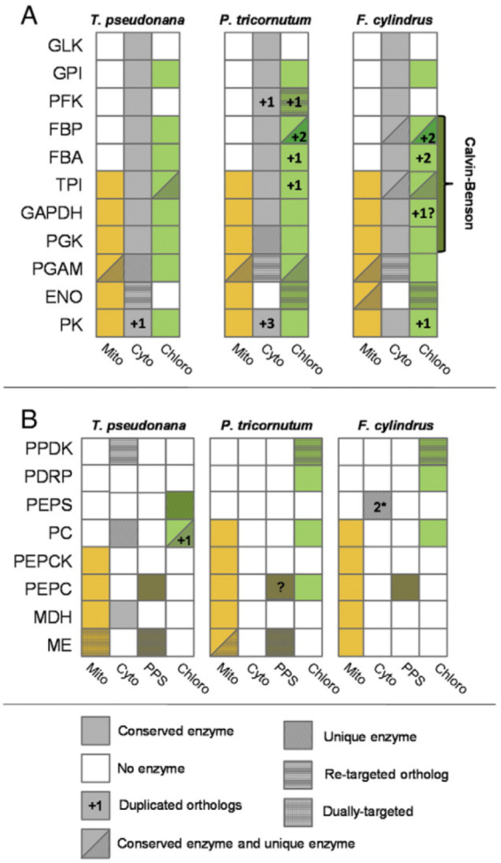


Fig. 5. The compartmentation of A) glycolysis/gluconeogenesis and B) pyruvate hub enzymes in diatoms. Filled boxes indicate the presence of an enzyme targeted to the mitochondria (orange), cytosol (grey), chloroplast (green), or periplastid compartment (brown). White boxes designate the absence of an enzyme. All enzymes are referenced in Table S1. Orthologs are indicated by solid colors while unique enzymes (found only in a single diatom genome) are indicated by diagonal stripes. Horizontal stripes show enzymes with orthologs that have been re-targeted in all three diatom genomes. Cross-hatching denotes possible targeting of an enzyme to two different intracellular locations (dual targeting). The numbers indicate the number of additional enzymes identified per compartment, when the number is on a solid box it indicates a duplication event. When the number is on a horizontal striped box it indicates an enzyme that was either horizontally acquired or selectively maintained. The 2* in *F. cylindrus* PEPS indicates two unique enzymes.

The next steps of glycolysis are catalyzed by phosphoglycerate kinase and phosphoglycerate mutase. A single mitochondrially targeted isoform of phosphoglycerate kinase (PGK1) was identified in all three diatom genomes and is phylogenetically related to PGKs from other heterokonts (Fig. 6). In contrast, there are several phosphoglycerate mutases (PGAM) that are putatively targeted to the mitochondria, however, only one (PGAM1) is conserved in all three diatoms and is of the expected length (approx. 288aa or 32 kDa). PGAM1 sequences are most closely related to the chloroplast-localized PGAM2. PGAM1 and PGAM2 are apparently distantly related to other diatom PGAM

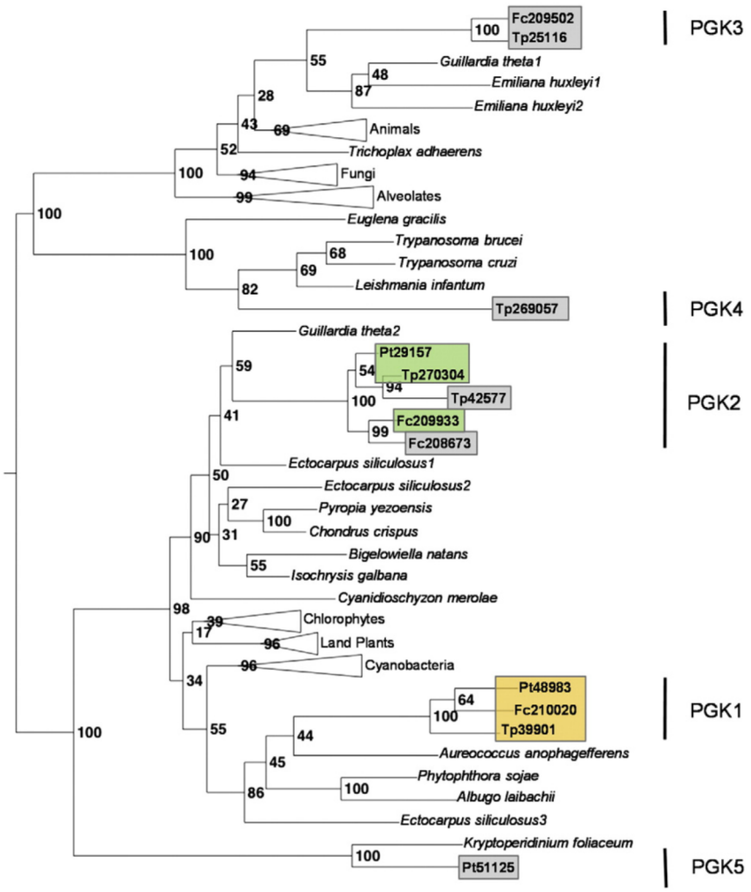


Fig. 6. Maximum likelihood tree of PGK isozymes. The phylogenetic tree was generated using RAxML-HPC BlackBox (CIPRES Science Gateway) from 64 PGK sequences. The tree has been mid-point rooted. Diatom sequences discussed in this study are shown boxed and in bold font. Filled boxes indicate the presence of an enzyme targeted to the mitochondria (orange), cytosol (grey), chloroplast (green). Bootstrap values are labeled at each node and are presented in %. Sequence accession numbers are listed in Table S2.

orthologs (PGAM3–PGAM10), some of which have predicted mitochondrial-targeting (Table S1). Since several of the gene models for the remaining PGAM predicted proteins appear to be partial or incorrect, and they are larger than the typical PGAM subunit, these ORFs may just be part of proteins of unknown function that include the catalytic activity of PGAM but may not be directly involved in the glycolysis pathway (Table S1).

The final two steps of the glycolysis pathway are catalyzed by enolase (ENO) and pyruvate kinase (PK). A mitochondrial ortholog of enolase (ENO1) was identified in all three diatom genomes. ENO1 is similar to sequences both from other heterokonts and the chromalveolate *Emiliania huxleyi* and is similar to the other orthologous enolase (ENO2), which is predicted to be cytosolic in *T. pseudonana* but chloroplast-targeted in the pennates. A second mitochondrially targeted enolase (ENO3) was identified only in the genome of *F. cylindrus*, and this enzyme is more highly divergent from ENO1 and ENO2 (Table S1). The final step of glycolysis, catalyzed by pyruvate kinase (PK) is the only unidirectional reaction of the mitochondrial glycolysis pathway. Pyruvate kinase is a key enzyme in the regulation of carbon metabolism and is known to possess a wide range of kinetic and regulatory properties [53]. Two mitochondrial PKs (PK1, PK2) were found to be conserved and orthologous in all three diatom genomes. Both diatom mitochondrial PKs belong to a family of PKs that are conserved within heterokonts. This heterokont group of PKs is most closely related to enzymes from metazoans than the other PKs found in diatom genomes suggesting that they may be ancient eukaryotic enzymes.

Major questions remain concerning both the origin and function of mitochondrial glycolysis. It is currently unknown whether mitochondrial glycolysis is an ancient trait or a trait derived from gene transfer after the acquisition of a plastid [22]. The finding that non-photosynthetic oomycetes also have mitochondrial glycolysis seems to support the idea that mitochondrial glycolysis is an ancient eukaryotic trait. However, since oomycetes have several genes that suggest they may have harbored a transient plastid, it's possible that enzymes acquired via EGT were targeted to the mitochondria prior to plastid loss and the diversification of diatoms [14]. Therefore, without more rigorous phylogenetic analysis, the origin of mitochondrial glycolysis remains unclear.

There are several possible reasons for and consequences of a complete mitochondrial glycolysis payoff phase in diatoms. In plants, cytoplasmic glycolysis enzymes have been reported to be physically associated with mitochondria under conditions of high respiration to facilitate the channeling of pyruvate to the TCA cycle [54,55]. By internalizing the lower glycolysis pathway into the mitochondria, diatoms would not need to transport pyruvate to supply the TCA cycle since pyruvate would be produced within the target organelle. Intermediates for the mitochondrial glycolysis pathway would need to be supplied from the cytosol by a triose phosphate transporter that has yet to be identified. Alternatively, the supply of GAP for mitochondrial glycolysis may be predominantly from a recently discovered mitochondrial Entner–Doudoroff (ED) pathway [30]. Neither intracellular transporters nor diatom ED glycolysis are well-characterized, and it is presently unclear what the relative importance of each supply route would be. Another implication of an internalized mitochondrial glycolysis pathway is that NADH is produced within the organelle rather than in the cytosol. Typically, NADH must be imported into the mitochondria via the malate–aspartate shuttle so that the reducing power can be used to drive oxidative phosphorylation [56]. By translocating the production of NADH into the mitochondria, diatoms would no longer require these enzymes for this purpose. Finally, glycolate produced from photorespiration can be enzymatically converted to glycerate in diatom mitochondria where it is believed to then be imported to the chloroplast, phosphorylated, and can re-enter the Calvin–Benson cycle [23]. Recently, a chloroplast-localized glycerate kinase was identified [30]. However, if a mitochondrial glycerate kinase could be identified, it's possible that this photorespiration by-product could then directly enter mitochondrial glycolysis and the TCA cycle and be used to either drive energy generation or to replenish TCA cycle intermediates.

The malate–aspartate shuttle involves specialized enzymes and transporters, but the key enzyme is malate dehydrogenase (MDH), which is part of the pyruvate hub (Fig. 1). Mitochondrial MDH is an essential TCA cycle enzyme, and not surprisingly, orthologs for mitochondrial MDH were conserved in diatom genomes (MDH1). However, the malate–aspartate shuttle requires a cytosolic MDH, which could be identified in the *T. pseudonana* genome (MDH2) but was not found in the genome of either pennate diatom. Therefore, it appears that pennate diatoms are unable to import reducing equivalents into the mitochondria from the cytosol in this manner while the centric *T. pseudonana* may possess this capacity. This finding supports the idea that pennate diatoms do not need to import reducing equivalents (NADH) since they are produced in the mitochondrial matrix.

Like most pyruvate hub enzymes, there are several possible metabolic roles for MDH in addition to its essential role as a TCA cycle intermediate, and in the malate–aspartate shuttle. The OAA produced by MDH1 in mitochondria could be used to supply mitochondrial gluconeogenesis, the next step of which would be catalyzed by PEPCK. In mammals, PEPCK is known to be the rate-controlling step of gluconeogenesis and is under transcriptional control [57]. Alternatively, in C4 plants, the decarboxylating activity of PEPCK is also part of a biochemical carbon concentrating mechanism [56]. The exact role of PEPCK in diatoms is not clear; however it exists as a single-copy ortholog in all three diatom genomes and is strongly predicted to be mitochondrial (Table S1). Based on its mitochondrial localization, PEPCK is most likely a gluconeogenesis enzyme (rather than a C4-type enzyme) in diatoms. If PEPCK is a gluconeogenesis enzyme, it could use the OAA produced either by MDH1, or by mitochondrial pyruvate carboxylase (PC). Though there are several isozymes of PC in diatom genomes, only one mitochondrial isoform was found (PC3), but was absent in the genome of *T. pseudonana*. Therefore, pennate diatoms have the capacity to initiate gluconeogenesis from pyruvate through the combined activities of PC and PEPCK in the mitochondria, while centric diatoms may be restricted to initiating gluconeogenesis from OAA in this organelle.

Under normal growth conditions, the metabolic intermediates used to initiate gluconeogenesis in the mitochondria are typically drawn away from the TCA cycle for biosynthetic reactions, such as the synthesis of amino acids. These intermediates can be replenished by activity of anapleurotic enzymes, including mitochondrial PC and PEPC. All three diatoms have two PEPC, and one isozyme (PEPC2) is strongly predicted to be mitochondrially localized, indicating its most likely role is anapleurotic (Table S1). The other isozyme, PEPC1, has high scores predicting mitochondrial localization, but is also predicted to be localized to the periplastid compartment (Table S1). In both PEPC1 and PEPC2, there is some ambiguity about whether the correct gene models are predicted, and it may be possible that there are several transcript isoforms originating from the same locus (Table S1). Although it is tempting to speculate on the role of PEPC1 and PEPC2 in diatom metabolism, any significance cannot be ascribed without more confidence in the sub-cellular localization of these enzymes.

The final enzyme conserved in the mitochondria of diatoms is malic enzyme (ME1). The role of diatom ME is currently unknown, but it could participate in both pyruvate metabolism and C4-type photosynthesis. In all three diatoms, ME1 has high predicted mitochondrial localization; however, a signal peptide that targets ME1 to the periplastid compartment was detected in *T. pseudonana* (Table S1). *P. tricornutum* has an additional ME (ME2) for which there is no apparent ortholog in the other diatoms though there are closely related enzymes found in oomycetes and *Ectocarpus*, suggesting it is at least partially conserved in heterokonts. The additional *P. tricornutum* ME (ME2), like *T. pseudonana* ME1 has predicted dual targeting to the mitochondria and the periplastid compartment.

Although there is compelling bioinformatics evidence as well as experimental evidence to support the existence of a mitochondrial glycolysis pathway in diatoms, these predictions should be validated experimentally [23,24]. However, if it is assumed that the mitochondrial glycolysis pathway will be demonstrated in diatoms, more questions will remain to be answered. For example, what is the magnitude of carbon flux through this pathway and what is the nature of the carbon intermediates and transporters that supply this pathway? Also, what proportion of the pyruvate produced in this pathway is catabolized completely to make ATP or to make 4C skeletons for other biosynthetic processes? Since the enzymes of the mitochondrial glycolysis pathway catalyze reversible reactions, it is possible that this pathway is used to run gluconeogenesis, allowing the mitochondria to export triose phosphates. The complement of pyruvate hub enzymes found in diatom mitochondria includes those capable of both anapleurotic and key glycolysis bypass (gluconeogenic) reactions, so this is theoretically possible. Running this pathway in the gluconeogenic direction would require carbon intermediates, which could theoretically be supplied from mitochondrial fatty acid oxidation or photorespiration [18]. In sum, if the manipulation of carbon flux in diatoms is proposed, it must be done with a better idea of the importance of this compartmentalized pathway relative to other metabolically compartmentalized pathways in contributing to the production, destruction, or recycling of fatty acids.

3.4.2. Distribution of carbon partitioning enzymes between the cytosol and chloroplast

Cytosolic orthologs were identified for all steps in the upper phase of the glycolysis/gluconeogenesis pathways (GLK to TPI), including those involved in the second bypass (PFK and FBP). All three diatoms possess an orthologous copy of a cytosol-localized ATP-PFK. Interestingly, the *P. tricornutum* genome contains two additional copies of ATP-PFK (ATP-PFK2A, ATP-PFK2B), which are arranged adjacent to one another on chromosome 29 and share a high degree of sequence similarity (89%), suggesting that they arose through a relatively recent in-paralogous duplication. ATP-PFK2B from *P. tricornutum* is putatively targeted to the chloroplast (Table S1), whereas the other two diatoms apparently lack a chloroplast-localized PFK despite possessing near-complete plastidic glycolysis (Fig. 4). This suggests that *P. tricornutum* has an increased capacity for hexose metabolism in the chloroplast relative to *T. pseudonana* and *F. cylindrus*.

In a deviation from classical EMP glycolysis, all diatom genomes also encode an alternative enzyme to ATP-PFK that utilizes pyrophosphate as the phosphoryl donor (PP_i-PFK; Table 1, Table S1). This glycolysis variant is thought to be energetically advantageous, since ATP is not consumed in the preparatory phase [58]. This strategy may be particularly important in anaerobic conditions, when the generation of ATP from the combined actions of the TCA cycle and oxidative phosphorylation ceases [58–60]; for example, amitochondriate protists and anaerobic bacteria both lack ATP-PFK and instead rely the PP_i-dependent enzyme [9]. The presence of PP_i-PFK is widespread in photosynthetic organisms [61]; in terrestrial plants PP_i-PFK activity is more responsive to regulatory controls than ATP-PFK and therefore is thought to be a method of adjusting plant carbon metabolism under changing environmental conditions [62]. However, diatom PP_i-PFK enzymes contain the bacterial PRK06555 domain (ProtClustDB ID 417822) and share similarity to several bacterial sequences (Table S1), suggesting that PP_i-PFK may have been acquired horizontally from bacteria at some point prior to the divergence of diatoms. The high conservation of PP_i-PFK amino acid sequence among the diatoms is consistent with this enzyme playing an important role as a regulatory step in central carbon metabolism. As one of the key regulatory steps in glycolysis, PFK is a potentially promising target for genetic manipulation. However, the relative importance of ATP-PFK and PP_i-PFK in regulating diatom intracellular carbon flux is currently unknown.

FBP, at the second bypass step, is an important site of regulation to supply carbon for carbohydrate synthesis. Previous FBP phylogenies in diatoms suggest a bacterial origin for cytosolic FBP [63,64], while chloroplast-localized FBPs are thought to have arisen from an ancient duplication of a cytosolic FBP that was redirected to the plastid [43,64]. In the chloroplast, these enzymes have been co-opted to function in carbon fixation and are involved in the regeneration of ribulose 1,5 bisphosphate in the Calvin–Benson cycle [65]. Additional chloroplast-localized FBP genes were acquired by diatoms via the red algal endosymbiont [66].

The cytosolic form of FBP is well conserved among diatoms; each diatom genome contains a single orthologous copy (FBP3; Fig. 4). The *F. cylindrus* genome contains an additional cytosolic FBP (FBP6) that is highly divergent from other diatom FBPs (<30% similar), suggesting that it is evolving rapidly. Multiple chloroplast-localized FBPs were identified in each diatom genome; two orthologous groups were conserved among the three diatoms (FBP1, 2). Localization studies using GFP fusions to FBP N-terminal bipartite pre-sequences in *P. tricornutum* confirmed the plastid localization of these enzymes [43]. *P. tricornutum* and *F. cylindrus* have two additional orthologous FBPs (FBP4) that are absent in *T. pseudonana*, and *F. cylindrus* has another unique FBP (FBP5). Interestingly, in both the *P. tricornutum* and *F. cylindrus* genome, the FBP4 genes have high sequence similarity (>60%) and neighbor one other on the same chromosome/scaffold, suggesting that perhaps they are a result of a duplication event early in pennate evolution.

The widespread duplication and diversification of chloroplast targeted FBPs observed in pennates (especially in *F. cylindrus*) may confer an advantage to this lineage by allowing the cell to quickly adjust its rate of carbon fixation to accommodate its metabolic demands. Alternatively, the multiple copies may have different activities or substrate affinities, and therefore be optimal only under certain conditions. For example, FBP catalyzes a similar reaction to the enzyme sedoheptulose-1,7-bisphosphatase (SBP). A plastid-targeted SBP is apparently absent in diatom chloroplasts, and it is possible that a plastidic FBP could serve as a substitute for SBP in the reductive pentose phosphate pathway [43]. An additional possibility is that certain FBPs function as dedicated Calvin–Benson enzymes associated with a sub-chloroplast location (e.g. the pyrenoid [67]), while other FBPs are located elsewhere in the chloroplast where they may participate in plastidal gluconeogenesis.

In addition to FBP, the enzymes FBA, TPI, GAPDH, and PGK could potentially play dual roles in the Calvin–Benson cycle if targeted to the plastid. Localization studies have demonstrated pyrenoid targeting for FBA isozymes in *P. tricornutum* [44]. Although all diatom genomes have multiple copies of TPI targeted to the chloroplast, they only have a single chloroplast-targeted ortholog of GAPDH. Therefore, this GAPDH likely has an essential role in the Calvin–Benson cycle and is probably not involved in plastidial glycolysis under photosynthetic conditions. Although diatoms have a nearly complete complement of genes that would permit the upper phase of glycolysis/gluconeogenesis in the plastid, several of those genes may be partially or exclusively involved in carbon fixation.

In contrast to the upper phase glycolysis/gluconeogenesis pathway, in which orthologs are remarkably well-conserved, the enzymes that catalyze the 7th–9th steps of cytosolic glycolysis or the "mid-payoff phase" seem to have been subjected to deletion and re-targeting throughout the diversification of diatoms. First, though all diatoms seem to have homologs of PGAM and PGK predicted to be localized to the cytosol, there are no conserved orthologs found in all three diatom genomes in this space. Furthermore, of the three diatoms, only *T. pseudonana* appears to have a full cytosolic glycolysis pathway while both *P. tricornutum* and *F. cylindrus* apparently lack cytosolic isoforms of enolase (ENO).

A more careful analysis reveals that there are several versions of cytosolic PGK and PGAM found in diatom genomes. Only mitochondrial and chloroplast versions of PGAM were identified with confidence and no obvious candidates for cytosolic PGAMs were found (See Section 3.4.1). While it is possible that diatoms have cytosolic PGAM activity, they have not maintained conserved orthologs of this enzyme in the cytoplasm.

There is a surprising amount of diversity within diatom PGK, and the acquisition of several PGK genes from a variety of sources was revealed through phylogenetic analysis (Fig. 6). The distribution of PGK among the diatoms genomes highlights the ability of these organisms to selectively duplicate, re-target and/or delete acquired genes, presumably in order to optimize metabolic processes. For example, although each diatom genome encodes for a single plastid-localized PGK (PGK2A), *T. pseudonana* and *F. cylindrus* both contain an additional PGK with high sequence similarity but without chloroplast targeting information (PGK2B). No PGK2B was found in *P. tricornutum*. The PGK2 group clusters with sequences from red algae and secondary endosymbionts from the red algal lineage, indicating that theses enzymes likely originated via EGT from the acquisition of a red algal plastid. Curiously, a sequence from the chlorachniophyte *Bigelowiella natans* (a green algae-derived secondary endosymbiont) also clusters with this group, the significance of which is difficult to interpret without a more comprehensive understanding of the evolutionary origin of the *B. natans* PGK. Additional cytosolic PGKs were found in *T. pseudonana* and *F. cylindrus* (PGK3) that belong to clade comprised of eukaryotic organisms, including chromalveolates and opisthokonts (Fig. 6). These two groups are distantly related, and it is possible that this enzyme is an ancient isoform, perhaps a remnant from the secondary exosymbiont, which has been lost in *P. tricornutum*. Interestingly, *P. tricornutum* does have a unique version of PGK (PGK5), which is most similar to a sequence from the dinoflagellate *Kryptoperidinium foliaceum*. This dinoflagellate is known to maintain a tertiary plastid derived from a diatom endosymbiont, and since neither *T. pseudonana* nor *F. cylindrus* has a PGK5, the diatom endosymbiont found in *K. foliaceum* is likely related to a group of diatoms including *P. tricornutum* [68]. Finally, *T. pseudonana* has a unique isoform of PGK (PGK4) that is not found in the other diatom genomes. PGK4 clades with a well-supported cluster with sequences from excavates, an unrelated group to diatoms or any of their endosymbionts. It's possible that *T. pseudonana* acquired this gene horizontally or that it is a metabolic relic from deep in eukaryotic evolutionary history. Despite the inability to determine the absolute origins of these distinct enzyme types, it is clear that cytosolic PGKs, unlike their highly conserved mitochondrial homologs, have not been conserved throughout the diversification of diatoms (Fig. 6). Whether these isozymes are functional equivalents, or have kinetic or regulatory differences that confer an adaptive advantage to one diatom or another is not clear without functional characterization.

Finally, the only diatom with a predicted cytosolic enolase is *T. pseudonana* (ENO2). In pennates, a signal peptide and chloroplast transit peptide were detected for this protein. The predicted ENO2 model in *T. pseudonana* was validated with 5′ RACE (unpublished results) to eliminate the possibility that the model was missing an upstream exon with a targeting motif. The localization of ENO2 in the chloroplast should be validated experimentally in the pennate diatoms. If confirmed, the presence of the penultimate glycolysis step either in the cytoplasm of centrics or in the chloroplast of pennates represents a significant difference in the organization of metabolism between the two major diatom lineages.

The lack of conservation in the mid-payoff phase (Fig. 5A) suggests that cytosolic glycolysis is not an essential energy-producing pathway in diatoms. Though the pennates are apparently missing a full cytosolic glycolysis pathway, they have a more complete plastid-localized pathway that *T. pseudonana* seems to lack and this variability in pathway organization has major implications for how the regulation of the direction of carbon flux may be different in these major groups.

Diatoms have several isozymes of PK, which catalyzes the final and unidirectional step of the payoff phase of glycolysis (Table 1, Fig. 1). The mitochondrial isoforms have already been discussed (PK1, PK2, Section 3.4.1), however there are additional PKs found in both the chloroplast (PK3) and the cytosol (PK4). Both *T. pseudonana* and *P. tricornutum* have a single chloroplast PK3 ortholog, while *F. cylindrus* has two (Table S1). Each diatom has a different number of cytosolic PK4s, *T. pseudonana* has two, *P. tricornutum* has 4, and *F. cylindrus* has a single copy. It is unclear why the number of cytosolic PKs varies among species, but one explanation is that PK4 may have been duplicated and deleted independently several times during diatom evolution. Supporting this idea is evidence for a recent duplication event in *P. tricornutum* that gave rise to two in-paralogs (homologs arising from a within-species duplication) on chromosome 8 (data not shown).

The pyruvate produced by PK3 in the chloroplast or PK4 in the cytosol enters the pyruvate metabolic hub. Compared with the mitochondria, there are relatively few enzymes of the pyruvate hub found in the chloroplast and cytosol of diatoms. Only a single chloroplast enzyme (PC1) was found in all three diatom genomes. Though *T. pseudonana* is apparently missing a mitochondrial PC3, there are two apparently recently duplicated copies of a PC2 found only in the centric diatom. One of the PC2 copies has predicted chloroplast localization, while the other copy is apparently cytosolic (Table S1). Typically, PC is a mitochondrial or cytosolic enzyme involved in gluconeogenesis or an anapleurotic role, and little is known about its role in chloroplasts. Since PEPCK is only predicted to be mitochondrial (Section 3.4.1) it is unlikely that plastid PC is functioning to initiate gluconeogenesis, since there are no other known enzymes capable of converting OAA to PEP found in this compartment. It is more likely that the OAA produced in chloroplasts is used in other biosynthetic pathways. In *E. huxleyi*, the transcription of chloroplast PC has been shown to be light-regulated and has been proposed to be involved in the production of OAA as a precursor for the biosynthesis of amino acids [69]. OAA can also be produced by PEPC, and as discussed previously (Section 3.4.1), all three diatoms have a PEPC2 with some predicted periplastid compartment/chloroplast localization. In C4 plants, PEPC has a role in a biochemical carbon concentrating mechanism to facilitate the delivery of CO_2 to RuBisCO in photosynthesis [70]. If PEPC1 is in fact periplastid localized in diatoms, it could only function in this capacity provided there is an enzyme near RuBisCO that can decarboxylate OAA. Though there is some experimental evidence to suggest that diatoms may have a biochemical carbon concentrating mechanism, the localization of a decarboxylating enzyme near RuBisCO has not been demonstrated conclusively [23,70].

The only non-mitochondrial MDH that could be identified in diatoms was the cytosolic MDH2 from *T. pseudonana*, indicating that diatoms apparently lack the ability to convert chloroplast OAA to malate in either the plastid or the periplastid compartment. However, as mentioned previously (Section 3.4.1), both *T. pseudonana* and *P. tricornutum* have isozymes of ME (ME1, and ME2 respectively) that are possibly targeted to the periplastid compartment, suggesting that malate may be either imported or produced by still unknown enzymes in this space. The decarboxylation of malate by ME in this compartment would produce CO_2, pyruvate, and NAD(P)H, and if the malate is imported from the mitochondria this reaction could effectively serve both as a biochemical CCM and a conduit for the precursors of fatty acid biosynthesis. However, ME1 in *T. pseudonana* and ME2 in *P. tricornutum* also have high predicted targeting to the mitochondria and they may be incorrectly predicted to the periplastid compartment. Because of the potentially important role in intracellular carbon flux, validating the sub-cellular localization of these enzymes should be a priority for the field.

Another enzyme of the pyruvate hub, PPDK, is found as a single copy in all three diatom genomes. Though diatom PPDK was determined to be monophyletic, the sequence from *F. cylindrus* appears to be more highly divergent (lower sequence similarity) to the sequences from *T. pseudonana* and *P. tricornutum*, suggesting it may be experiencing a higher rate of mutation. PPDK is phylogenetically limited to prokaryotes, protists (including diatoms), some fungi, and green plants, but its function can be quite variable [71]. PPDK is capable of acting bidirectionally, by replacing the activity of PK in the final step of glycolysis or by initiating gluconeogenesis [72]. Anaerobic bacteria utilize a pyrophosphate-dependent variant of glycolysis in which PPDK acts as an alternative to PK in the final glycolysis step. In C4 plants the role of PPDK is specialized, as it resupplies PEP in the stroma of leaf-mesophyll cell chloroplasts, however it is believed that PPDK first became functionally seated in C3 plants, where it has a role in balancing the flux of carbon through glycolysis or gluconeogenesis, and was only slightly modified to achieve a new function in C4 plants [71]. Interestingly, just like the ENO2, PPDK is predicted to be localized to the plastid in both pennate diatoms, while the PPDK from *T. pseudonana* is cytosolic. Whether the differential localization of PPDK between centrics and pennates indicates some functional distinction, or is connected to the translocation of ENO2, remains unclear

The final enzyme of the pyruvate hub is PEPS. PEPS is found mostly in prokaryotes where it is believed to catalyze the conversion of pyruvate to PEP in the gluconeogenic direction [48,73,74]. PEPS was identified in the *T. pseudonana* and *F. cylindrus* genomes but appeared to be absent from the *P. tricornutum* genome. The *T. pseudonana* PEPS is chloroplast-localized and is most similar to a sequence from *Cyanothece* sp. PCC 7424 (ZP_01910935.1). Both the *F. cylindrus* PEPS genes are apparently cytosolic and are most similar (but with low similarity scores) to PEPS from the heterotrophic bacteria *Plesiocystis pacifica* (ZP_01910935.1,) and *Bacillus pumilis* (ZP_03053952.1). Since none of these enzymes are found in any other heterokonts or related eukaryotes, the most parsimonious explanation for the occurrence of PEPS in these diatom genomes is that they have been acquired horizontally following the evolutionary divergence of all three diatoms investigated.

The specific functions of PPDK and PEPS in diatoms remain unclear, but they are undoubtedly important enzymes in the distribution of carbon intermediates in the pyruvate hub. Very little is known about the role of PEPS in eukaryotes though its role in bacteria as a regulatory enzyme is more well characterized [48]. It has been proposed that plastidic PPDK in C3 plants may have an important role in supplying PEP for the biosynthesis of aromatic amino acids [75]. If it has this role in diatoms, it may be an important target for down-regulation since it represents a sink for pyruvate (a fatty acid biosynthesis precursor) though there may be detrimental effects on growth by inhibiting aromatic amino acid production.

Overall, there are some generalizations about diatom carbon partitioning in the chloroplast and cytosol that emerge from this analysis that have implications for the way intracellular carbon flux is regulated. First, diatoms do not share a conserved upper half or preparatory phase of glycolysis in chloroplasts, which means that diatom plastids (with the exception of *P. tricornutum*) are not metabolically equipped

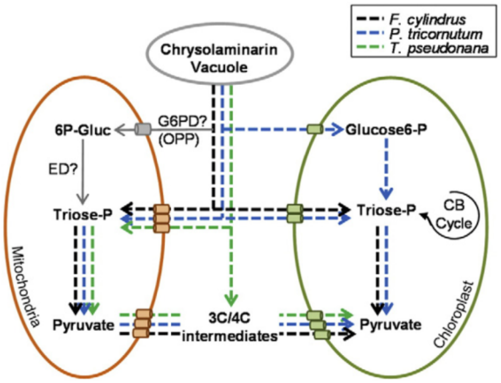


Fig. 7. Schematic diagram of the compartmentation of glycolytic flux in *T. pseudonana*, *P. tricornutum*, and *F. cylindrus*. Arrows follow the direction of glycolytic flux from cytosolic Glucose 6-P (produced by degradation of chrysolaminarin) to oxidized 3C or 4C metabolic intermediates. The pathway organization for each diatom is indicated by a different color (see key). All diatoms have a cytosolic preparatory phase and a mitochondrial payoff phase of classical Emden–Meyerhoff–Parnas glycolysis. Only *T. pseudonana* has a complete cytosolic pathway and only the pennate diatoms have a chloroplast-localized payoff phase. All compartments are shown with unknown transporters. Gray arrows indicate a possible pathway from cytosolic Glucose 6-P into mitochondrial glycolysis via the activity of the first enzyme of the oxidative pentose phosphate pathway (OPP), which is G6PD (glucose 6-P dehydrogenase, EC 1.1.1.49). To enter the mitochondrial Entner–Doudoroff glycolysis pathway, the 6-phosphoglucono-δ-lactone (6P-Gluc) produced in the cytosol would have to be imported by an unknown transporter.

for the catabolism of hexoses (Fig. 7). This is not surprising, since unlike terrestrial plants and green algae, which store and break down starch inside the plastid, diatom chrysolaminarin vacuoles (where diatoms store carbohydrate) are extra-plastidial. This is supported by the conservation of the upper half of glycolysis in the cytosol of diatoms. In diatoms, fatty acids are synthesized de novo in the chloroplast by a type II fatty acyl synthase, meaning that pyruvate must be either produced in or imported to this organelle [18]. Since pennate diatoms have isozymes for the lower half of glycolysis targeted to the plastid, they are theoretically capable of producing pyruvate in the chloroplast from triose phosphates that are either produced during photosynthesis or imported from the cytosol. In contrast, *T. pseudonana* apparently does not have a chloroplast-localized lower half of glycolysis meaning that in order for organic carbon fixed during photosynthesis to be incorporated into fatty acids it must first be exported from the chloroplast. Therefore, in *T. pseudonana* (and possibly other centric diatoms), the transporters that export triose phosphates and import organic carbon precursors like oxaloacetate may be an important site of carbon flux regulation into other fatty acid supply pathways (Fig. 7). In pennates the regulation may also include a fine level of regulatory control within the plastid, at the import of glucose 6-P (in *P. tricornutum*) or triose phosphates. As a result of these analyses, it is now clear that there are some significant differences between the centric and pennate diatom lineages with respect to the organization of the carbon partitioning pathways in the cytosol and the chloroplast, and these differences have implications for the way intracellular carbon flux is regulated (Fig. 7).

The most notable difference with respect to the organization of the pyruvate metabolic hub between centrics and pennates is that the capacity to initiate gluconeogenesis (by bypassing PK and catalyzing the pyruvate → PEP reaction) has not been conserved in the plastid or cytosol in diatoms. This may indicate that 1) gluconeogenesis is mostly initiated from pyruvate (or OAA) in the mitochondria, 2) there has been significant adaptability of this carbon flux regulatory node during the diversification of diatoms, or 3) a combination of both. The adaptability of this carbon flux distribution node is supported by finding that the orthologous PPDK has been re-targeted in centrics and pennates and that PEPS seems to have been acquired horizontally independently in *T. pseudonana*, and *P. tricornutum*. The extent to which these enzymes affect carbon flux distribution in a primary or accessory role cannot be predicted bioinformatically.

3.4.3. Pyrophosphate-dependent glycolysis and other EMP variants

Both PP_i-PFK and PPDK are found in the genomes of all three diatoms investigated suggesting that these organisms may utilize the PP_i-dependent variant of EMP glycolysis. As was mentioned previously, PP_i-dependent glycolysis functions in anaerobic bacteria and amitochondriate protists to confer an energetic advantage when oxidative phosphorylation cannot be used to generate cellular ATP. Though this glycolysis variant generates less ATP than the combined activities of EMP glycolysis and oxidative phosphorylation, it may be sufficient to sustain cellular energetic demands temporarily or for long periods of time when cells are in a resting stage. Diatoms are known to survive long periods (months to decades) in dark and anoxic sediment layers [76]. Recently, the respiration of intracellular nitrate stores under anoxic conditions in diatoms (dissimilatory nitrate reduction to ammonia, DNRA) has been implicated as a potential survival mechanism under these conditions as they transition to a resting state [77]. While DNRA could be occurring in diatoms, it's also possible that PP_i-dependent glycolysis may also serve to sustain cellular energetic demands during dark anoxic conditions and may be a longer-term strategy for sustaining metabolic demands throughout the resting stage.

The frequency with which EMP glycolysis variants occur in diatoms is just beginning to be appreciated. Recently, a phosphoketolase pathway was identified in *P. tricornutum*, but not in *T. pseudonana* [30]. Additionally, a mitochondrial Entner–Dourdoff (ED) glycolysis pathway was discovered in both *P. tricornutum* and *T. pseudonana* [30]. The relative importance of these glycolysis variants in diatoms is currently unknown, as is the effect that running each pathway would have on cellular energetics. What is clear is that diatoms are complex organisms and much remains to be learned about the organization and regulation of their most conserved, core metabolic pathways.

3.5. Regulation of carbon partitioning pathways

Carbon flux through glycolysis or gluconeogenesis in the cytosol must be precisely regulated to prevent futile cycling between these opposing pathways. Reciprocal regulation coordinates the simultaneous activation of one pathway and suppression of the other. The reciprocal regulation of the second bypass is accomplished by the synthesis and breakdown of the allosteric effector Fru 2,6 bisP. Fru 2,6 bisP is synthesized from the glycolytic intermediate fructose 6-phosphate by the activity of 6-phosphofructo-2-kinase (PF2K) and converted back to fructose 6-phosphate by fructose 2,6 bisphosphatase (F2BP). In animals, these reactions are catalyzed by a single polypeptide, with a kinase domain on the C-terminal end and a phosphatase domain on the N-terminal end; this bifunctional enzyme arose early in eukaryotic evolution by the fusion of these two functional units [78]. Each of the diatom genomes examined encode for two PF2K/F2BP isozymes, all of which lack signal peptide sequences and are localized to the cytosol. The PF2K/F2BP enzymes form two orthologous clusters (PF2K/F2BP 1,2); the PF2K/F2BP 2 group was not previously annotated and was identified using BLAST.

The function of these two PF2K/F2BP isozymes in respect to the control of carbon metabolism in diatoms is currently unknown, although in other eukaryotes it has been shown that different PF2K/F2BP isozymes are expressed during a change in environmental condition or developmental stage [49]. In mammals, four different isozymes are known and are differentially expressed in various tissues to maintain glucose homeostasis. In contrast, higher plants typically encode for only one PFK-2/F2BP [50]. There is evidence for the presence of monofunctional enzymes in yeast (and in some plants), where the phosphatase domain has been inactivated due to a critical amino acid substitution in a universally conserved "RHG motif" [78]. The polypeptide still contains a vestigial phosphatase domain, but lacks the histidine amino acid required to accept a phosphate group during catalysis. Alignment of PF2K/F2BP 1 and 2 sequences with those of other eukaryotes reveals that the PF2K/F2BP 1 group contains the critical histidine amino acid, while the PF2K/F2BP 2 group has replaced it either with the non-polar amino acid alanine (*T. pseudonana*) or proline (*P. tricornutum* and *F. cylindrus*, Fig. 8). This suggests that the phosphatase domain in the PF2K/F2BP 2 homologues may not be active. Furthermore, this inactivation may have arisen independently in centric and pennate diatoms, given the difference in the interfering amino acid.

Another bifunctional kinase/phosphatase regulatory protein (PDRP) is known to govern the activity of the PPDK enzyme via light-mediated phosphorylation of the PPDK active site [79]. PDRP was discovered and initially described in C4 plants, where it works to coordinate PPDK activity with photosynthesis [75]. Chloroplast-localized PDRPs were putatively identified in the genomes of both *P. tricornutum* and *F. cylindrus* (Table S1) based on the presence of a PRK05339 domain that is also found in the PRDP from *Zea mays*. Therefore, in pennates, both PPDK and its regulatory protein appear to occur exclusively in the chloroplast. In contrast, no PDRP could be identified for *T. pseudonana*, which is consistent with the absence of a plastidal PPDK (Fig. 5). This arrangement suggests that pennate and centric diatoms may utilize these enzymes in functionally distinct capacities. As discussed in the previous section, the metabolic role of PPDK in diatoms is not entirely clear. In other photosynthetic organisms, plastid-localized PPDK plays a specialized role in C4 photosynthesis, although PPDK is also found in C3 organisms where it may work to supplement the PEP requirement for aromatic amino acid biosynthesis [75]. If PPDK performs the latter role in diatoms, it may be an important target for down-regulation since it represents a sink for pyruvate (a fatty acid biosynthesis precursor).

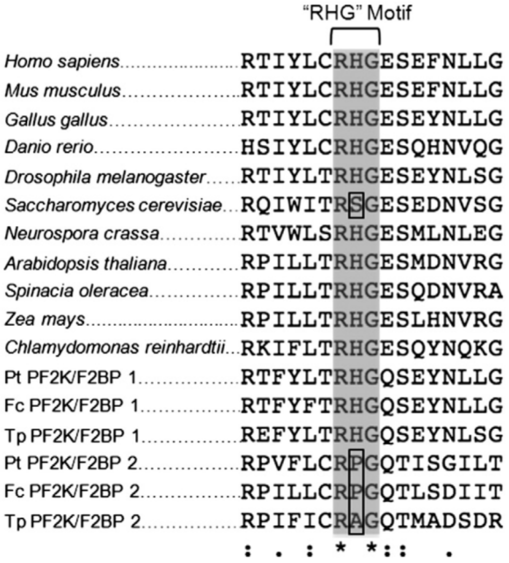

"RHG" Motif
Homo sapiens........................RTIYLCRHGESEFNLLG
Mus musculus....................RTIYLCRHGESEFNLLG
Gallus gallus......................RTIYLCRHGESEYNLLG
Danio rerio........................HSIYLCRHGESQHNVQG
Drosophila melanogaster.....RTIYLTRHGESEYNLSG
Saccharomyces cerevisiae...RQIWITRSGESEDNVSG
Neurospora crassa.............RTVWLSRHGESMLNLEG
Arabidopsis thaliana............RPILLTRHGESMDNVRG
Spinacia oleracea...............RPILLTRHGESQDNVRA
Zea mays............................RPILLTRHGESLHNVRG
Chlamydomonas reinhardtii...RKIFLTRHGESQYNQKG
Pt PF2K/F2BP 1.................RTFYLTRHGQSEYNLLG
Fc PF2K/F2BP 1.................RTFYFTRHGQSEYNLLG
Tp PF2K/F2BP 1.................REFYLTRHGQSEYNLSG
Pt PF2K/F2BP 2.................RPVFLCRPGQTISGILT
Fc PF2K/F2BP 2.................RPILLCRPGQTLSDIIT
Tp PF2K/F2BP 2.................RPIFICRAGQTMADSDR
: . : * *:: .

Fig. 8. Alignment of PFK2/F2BP bifunctional regulatory proteins. Shows the conserved "RHG" motif in PFK2/F2BP bifunctional proteins among diverse taxa including animals, fungi, higher plants, green algae and diatoms. Both diatom PFK2/F2BP group 2 sequences and a 860 known monofunctional PFK2/F2BP sequence from *Saccharomyces cerevisiae* (with an inactivated phosphatase domain) have substituted a different amino acid at the place of a critical histidine residue, as indicated by boxes. The accession numbers for sequences used in the alignment are as follows: *Homo sapiens* (GenBank ID: NP_006203.2), *Mus musculus* (GenBank ID: NP_032851.2), *Gallus gallus* (GenBank ID: XP_417979.2), *Danio rerio* (GenBank ID: NP_957302.1), *Drosophila melanogaster* (UniProt ID: Q9Y1W3_DROME), *Saccharomyces cerevisiae* (GenBank ID: AAA34858.1), *Neurospora crassa* (GenBank ID: XP_958926.1), *Arabidopsis thaliana* (GenBank ID: AEE28077.1), *Spinacia oleracea* (UniProt ID: O64983_SPIOL), *Zea mays* (GenBank ID: AAL09471.1), *Chlamydomonas reinhardtii* (UniProt ID: A8JAE8_CHLRE). Diatom PIDs are listed in Table S1.

With an ultimate goal of identifying gene targets for modification to improve fuel precursor molecule production, it is important to consider what sorts of regulatory processes are most amenable to manipulation. Eukaryotic cells employ several mechanisms to regulate the synthesis and breakdown of carbon intermediates. Transcriptional regulation enables a short-term response and large-scale control; for example, the activity of transcription factors allows for up- or down-regulation of entire pathways of related function in response to environmental cues. The bypass points and secondary regulatory molecules previously discussed are unidirectional control points, which would be especially attractive to alter the direction of carbon flux. Most of the steps in glycolysis are bidirectional, and are regulated by mass action, allostery, or post-translational modification. Modification of these steps to improve carbon flux would be more difficult than choosing a unidirectional control point, or genes that are largely regulated by transcription. Since changes in transcription are commonly major indicators of changes in cellular metabolism, determining transcript levels for each of the steps involved in carbon metabolism would greatly contribute to our understanding of carbon flux in diatoms and other algae and would aide in the identification of interesting targets for genetic manipulation.

4. Concluding remarks

The decreasing cost and increased output of high-throughput sequencing have meant that algal genome and transcriptome sequences can now be obtained easily, and the characterization of metabolomes and proteomes is becoming commonplace. This systems biology approach is incredibly powerful, but the challenge remains to distill these data into useful information that gives insight into the functional organization of a cell or organism [80]. With the aim of developing algal strains for biofuel production, such information is essential to characterize pathways that govern the flux of carbon from the fixation of CO_2 to the de novo biosynthesis of fatty acids and TAG. For the first time, the availability of genome sequence data from three diatoms representing both major diatom lineages has allowed for a comparative analysis of the organizing principles of one of the most fundamental and conserved metabolic pathways. Through this comparative analysis, several core features of diatom metabolism emerged that distinguish diatoms from other model organisms. First, it is clear that metabolic pathways are organized into various sub-cellular compartments in diatoms, and the conservation of the cytosolic preparatory phase and mitochondrial payoff phase of glycolysis is a unifying feature. The mitochondrial glycolysis pathway is likely to be the primary supply pathway for energy generation in the TCA cycle since a cytosolic pathway is not conserved. Consequently, an important point of carbon flux regulation should be the selective transport of metabolites between the cytosol and organelles (such as the export of organic carbon in the plastid, import of organic carbon in the mitochondria). Second, several isozymes of the pyruvate hub (PEPCK, PEPC, MDH) are conserved and targeted to diatom mitochondria. The presence or absence of isozymes in each compartment determines the possible fates for fixed carbon, which has associated metabolic and energetic consequences, so it can be inferred that mitochondrial PEPCK, PEPC, and MDH have indispensable roles in diatom metabolism.

These examples are in contrast to the significant proportion of the enzymes of the core carbon partitioning pathways that is not strictly conserved in diatom genomes. The differences between lineages and species are examples of how diatoms may have modified their carbon partitioning pathways to adapt to specific environmental niches during the diversification of the group. These modifications include the translocation of metabolic pathways (such as the payoff phase of glycolysis, which is plastid-localized in both pennates and cytosolic in *T. pseudonana*), selective maintenance and horizontal acquisition of gene families for enzymes like PGK and PEPS, and novel gene family expansions (as in the case of pennate FBPs). Further experimental work is required to determine what the physiological implications of these metabolic variations are, and additional genome sequences that span a greater diversity within the diatom family will be important to determine how significant some of these species-specific or lineage-specific modifications are in an evolutionary context. However, documenting these differences is an important first step in identifying which isozymes may confer adaptive flexibility within this algal group and should help in the interpretation and annotation of functional-omics datasets.

In summary, this analysis has characterized both conserved and variable features of central carbon metabolism pathways within a single class of algae. Characterizing the conserved features establishes a blueprint for the metabolic organization of diatoms over which functional data can be overlaid and interpreted in an ecological, evolutionary, or physiological context. Additionally, it facilitates the identification of enzymes with important roles in regulating carbon flux that are suitable targets for metabolic engineering. The variability in the organization of metabolic pathways in diatoms that was documented here illustrates how even core central pathways can be modified considerably within a single algal group. Overall, we have aimed to generate a framework using available genomic information to highlight gaps in our understanding, identify important areas for clarification, and to facilitate the interpretation of future functional-omics studies.

Supplementary data related to this article can be found online at http://dx.doi.org/10.1016/j.algal.2012.04.003.

Acknowledgments

These sequence data were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/ in collaboration with the user community. Diatom biofuels research in the Hildebrand lab is supported by AFOSR grant FA9550-08-1-0178, DOE grant DE-EE0001222, and NSF grant CBET-0903712. SRS was supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program. RA was supported by the San Diego Fellowship awarded through the UCSD Temporal Dynamics of Learning Center. The authors would like to thank Thomas Mock (University of East Anglia, UK), Ansgar Gruber, and Peter Kroth (University of Konstanz, Germany) for permission to use data acquired from the *F. cylindrus* genome sequencing project. The authors would also like to thank Juan Ugalde and Eric Allen for assistance with the construction of phylogenetic trees.

References

[1] R. Radakovits, R.E. Jinkerson, A. Darzins, M.C. Posewitz, Genetic engineering of algae for enhanced biofuel production, Eukaryotic Cell 9 (2010) 486–501.

[2] T.G. Dunahay, E.E. Jarvis, S.S. Dais, P.G. Roessler, Manipulation of microalgal lipid production using genetic engineering, Applied Biochemistry and Biotechnology 57–58 (1996) 223–231.

[3] E.T. Yu, F.J. Zendejas, P.D. Lane, S. Gaucher, B.A. Simmons, T.W. Lane, Triacylglycerol accumulation and profiling in the model diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* (Baccilariophyceae) during starvation, Journal of Applied Phycology 21 (2009) 669–681.

[4] Q. Hu, M. Sommerfeld, E. Jarvis, M. Ghirardi, M. Posewitz, M. Seibert, et al., Microalgal triacylglycerols as feedstocks for biofuel production: perspectives and advances, The Plant Journal 54 (2008) 621–639.

[5] P. Roessler, Effects of silicon deficiency on lipid composition and metabolism in the diatom *Cyclotella cryptica*, Journal of Phycology 24 (1988) 394–400.

[6] P. Roessler, Changes in the activities of various lipid and carbohydrate biosynthetic enzymes in the diatom *Cyclotella cryptica* in response to silicon deficiency, Archives of Biochemistry and Biophysics 267 (1988) 521–528.

[7] Y. Li, D. Han, G. Hu, M. Sommerfeld, Q. Hu, Inhibition of starch synthesis results in overproduction of lipids in *Chlamydomonas reinhardtii*, Biotechnology and Bioengineering 107 (2010) 258–268.

[8] C. Givan, Evolving concepts in plant glycolysis: two centuries of progress, Biological Reviews 74 (1999) 277–309.

[9] W. Plaxton, The organization and regulation of plant glycolysis, Annual Review of Plant Physiology 47 (1996) 185–214.

[10] N. Linka, A.P.M. Weber, Intracellular metabolite transporters in plants, Molecular Plant 3 (2010) 21–53.

[11] U.I. Flügge, R.E. Häusler, F. Ludewig, M. Gierth, The role of transporters in supplying energy to plant plastids, Journal of Experimental Botany 62 (2011) 2381–2392.

[12] J.E. Lunn, Compartmentation in plant metabolism, Journal of Experimental Botany 58 (2007) 35–47.

[13] J.M. Archibald, The puzzle of plastid evolution, Current Biology 19 (2009) R81–R88.

[14] P.J. Keeling, The endosymbiotic origin, diversification and fate of plastids, Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 365 (2010) 729–748.

[15] G. McFadden, Primary and secondary endosymbiosis and the origin of plastids, Journal of Phycology 37 (2001) 951–959.

[16] B. Stoebe, U.G. Maier, One, two, three: nature's tool box for building plastids, Protoplasma 219 (2002) 123–130.

[17] K. Bolte, L. Bullmann, F. Hempel, A. Bozarth, S. Zauner, U.G. Maier, Protein targeting into secondary plastids, Journal of Eukaryotic Microbiology 56 (2009) 9–15.

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

[19] P.J. Lopez, J. Desclés, A.E. Allen, C. Bowler, Prospects in diatom research, Current Opinion in Biotechnology 16 (2005) 180–186.

[20] J.N. Timmis, M.A. Ayliffe, C.Y. Huang, W. Martin, Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes, Nature Reviews Genetics 5 (2004) 123–135.

[21] P.J. Keeling, J.D. Palmer, Horizontal gene transfer in eukaryotic evolution, Nature Reviews Genetics 9 (2008) 605–618.

[22] M.L. Ginger, G.I. McFadden, P.A.M. Michels, Rewiring and regulation of cross-compartmentalized metabolism in protists, Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 365 (2010) 831–845.

[23] P.G. Kroth, A. Chiovitti, A. Gruber, V. Martin-Jezequel, T. Mock, M.S. Parker, et al., A model for carbohydrate metabolism in the diatom *Phaeodactylum tricornutum* deduced from comparative whole genome analysis, PloS One 3 (2008) e1426.

[24] M. Liaud, C. Lichtl, K. Apt, W. Martin, R. Cerff, Compartment-specific isoforms of TPI and GAPDH are imported into diatom mitochondria as a fusion protein: evidence in favor of a mitochondrial origin of the eukaryotic glycolytic pathway, Molecular and Biological Evolution 17 (2000) 213–223.

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

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

[27] P.G. Falkowski, M.E. Katz, A.H. Knoll, A. Quigg, J.A. Raven, O. Schofield, et al., The evolution of modern eukaryotic phytoplankton, Science 305 (2004) 354–360.

[28] K. Roberts, E. Granum, R.C. Leegood, J.A. Raven, Carbon acquisition by diatoms, Photosynthesis Research 93 (2007) 79–88.

[29] M. Tachibana, A.E. Allen, S. Kikutani, Y. Endo, C. Bowler, Y. Matsuda, Localization of putative carbonic anhydrases in two marine diatoms, *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*, Photosynthesis Research 109 (2011) 205–221.

[30] M. Fabris, M. Matthijs, S. Rombauts, W. Vyverman, A. Goossens, G.J.E. Baart, The metabolic blueprint of *Phaeodactylum tricornutum* reveals a eukaryotic Entner–Doudoroff glycolytic pathway, The Plant Journal (2012).

[31] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Research 25 (1997) 3389–3402.

[32] B. Gschloessl, Y. Guermeur, J.M. Cock, HECTAR: a method to predict subcellular targeting in heterokonts, BMC Bioinformatics 9 (2008) 393.

[33] I. Small, N. Peeters, F. Legeai, C. Lurin, Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences, Proteomics 4 (2004) 1581–1590.

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

[35] M.G. Claros, P. Vincens, Computational method to predict mitochondrially imported proteins and their targeting sequences, European Journal of Biochemistry 241 (1996) 779–786.

[36] K.E. Apt, L. Zaslavkaia, J.C. Lippmeier, M. Lang, O. Kilian, R. Wetherbee, et al., In vivo characterization of diatom multipartite plastid targeting signals, Journal of Cell Science 115 (2002) 4061–4069.

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

[38] L. Franzen, J. Rochaix, G. von Heijne, Chloroplast transit peptides from the green alga *Chlamydomonas reinhardtii* share features with both mitochondrial and higher plant chloroplast presequences, FEBS Letters 260 (1990) 165–168.

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

[40] T.N. Petersen, S. Brunak, G. von Heijne, H. Nielsen, SignalP 4.0: discriminating signal peptides from transmembrane regions, Nature Methods 8 (2011) 785–786.

[41] J. Bendtsen, H. Nielsen, G. von Heijne, S. Brunak, Improved prediction of signal peptides: SignalP 3.0, Journal of Molecular Biology 340 (2004) 783–795.

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

[43] A. Gruber, T. Weber, C.R. Bártulos, S. Vugrinec, P.G. Kroth, Intracellular distribution of the reductive and oxidative pentose phosphate pathways in two diatoms, Journal of Basic Microbiology 49 (2009) 58–72.

[44] A.E. Allen, A. Moustafa, A. Montstat, A. Eckert, P. Kroth, C. Bowler, Evolution and Functional Diversification of Fructose Bisphosphate Aldolase Genes in Photosynthetic Marine Diatoms, , 2011, pp. 1–48.

[45] N. Peeters, I. Small, Dual targeting to mitochondria and chloroplasts, Biochimica et Biophysica Acta (BBA) — Molecular Cell Research 1541 (2001) 54–63.

[46] C. Carrie, E. Giraud, J. Whelan, Protein transport in organelles: dual targeting of proteins to mitochondria and chloroplasts, FEBS Journal 276 (2009) 1187–1195.

[47] S.S. Sung, D. Xu, C.M. Galloway, C.C. Black, A reassessment of glycolysis and gluconeogenesis in higher plants, Physiologia Plantarum 72 (1988) 650–654.

[48] U. Sauer, B.J. Eikmanns, The PEP–pyruvate–oxaloacetate node as the switch point for carbon flux distribution in bacteria, FEMS Microbiology Reviews 29 (2005) 765–794.

[49] D.A. Okar, A.J. Lange, Fructose-2,6-bisphosphate and control of carbohydrate metabolism in eukaryotes, Biofactors 10 (1999) 1–14.

[50] T.H. Nielsen, J.H. Rung, D. Villadsen, Fructose-2,6-bisphosphate: a traffic signal in plant metabolism, Trends in Plant Science 9 (2004) 556–563.

[51] L.D. Gottlieb, Conservation and duplication of isozymes in plants, Science 216 (1982) 373–380.

[52] L.D. Hurst, C. Pál, M.J. Lercher, The evolutionary dynamics of eukaryotic gene order, Nature Reviews Genetics 5 (2004) 299–310.

[53] M. Munoz, E. Ponce, Pyruvate kinase: current status of regulatory and functional properties, Comparative Biochemistry and Physiology. B 135 (2003) 197–218.

[54] P. Giege, J.L. Heazlewood, U. Roessner-Tunali, A. Millar, A.R. Fernie, C.J. Leaver, et al., Enzymes of glycolysis are functionally associated with the mitochondrion in *Arabidopsis* cells, The Plant Cell 15 (2003) 2140–2151.

[55] J.W.A. Graham, Mitochondrial Glycolysis in Plants, University of Oxford, Oxford, 2007.

[56] D. Nelson, M.M. Cox, Lehninger Principles of Biochemistry, fifth ed. W.H. Freeman, 2008.

[57] K. Chakravarty, H. Cassuto, L. Reshef, R.W. Hanson, Factors that control the tissue-specific transcription of the gene for phosphoenolpyruvate carboxykinase-C, Critical Reviews in Biochemistry and Molecular Biology 40 (2005) 129–154.

[58] E. Mertens, Pyrophosphate-dependent phosphofructokinase, an anaerobic glycolytic enzyme? FEBS Letters 285 (1991) 1–5.

[59] E. Mertens, ATP versus pyrophosphate — glycolysis revisited in parasitic protists, Parasitology Today 9 (1993) 122–126.

[60] C. Slamovits, P. Keeling, Pyruvate-phosphate dikinase of oxymonads and parabasalia and the evolution of pyrophosphate-dependent glycolysis in anaerobic eukaryotes, Eukaryotic Cell 5 (2006) 148–154.

[61] N. Carnal, C.W. Black, Phosphofructokinase activities in photosynthetic organisms — the occurrence of pyrophosphate-dependent 6-phosphofructokinase in plants and algae, Plant Physiology 71 (1983) 150–155.

[62] C. Black, L. Mustardy, S. Sung, P. Kormanik, D. Xu, N. Paz, Regulation and roles for alternative pathways of hexose metabolism in plants, Physiologia Plantarum 69 (1987) 387–394.

[63] M. Rogers, P. Keeling, Lateral transfer and recompartmentalization of Calvin cycle enzymes of plants and algae, Journal of Molecular Evolution 58 (2004) 367–375.

[64] W. Martin, A. Mustafa, K. Henze, C. Schnarrenberger, Higher-plant chloroplast and cytosolic fructose-1,6-bisphophosphatase isoenzymes: origins via duplication rather than prokaryote–eukaryote divergence, Plant Molecular Biology 32 (1996) 485–491.

[65] C.A. Raines, The Calvin cycle revisited, Photosynthesis Research 75 (2003) 1–10.

[66] R. Teich, S. Zauner, D. Baurain, H. Brinkmann, J. Petersen, Origin and distribution of Calvin cycle fructose and sedoheptulose bisphosphatases in plantae and complex algae: a single secondary origin of complex red plastids and subsequent propagation via tertiary endosymbioses, Protist 158 (2007) 263–276.

[67] K. Suss, I. Prokhorenko, K. Adler, In-situ association of Calvin cycle enzymes, ribulose-1,5-bisphosphate carboxylase oxygenase activase, ferredoxin-NADP (+) reductase, and nitrite reductase with thylakoid and pyrenoid membranes of *Chlamydomonas reinhardtii* chloroplasts as revealed by immunoelectron microscopy, Plant Physiology 107 (1995) 1387–1397.

[68] B. Imanian, P.J. Keeling, The dinoflagellates *Durinskia baltica* and *Kryptoperidinium foliaceum* retain functionally overlapping, BMC Evolutionary Biology 7 (2007).

[69] Y. Tsuji, I. Suzuki, Y. Shiraiwa, Photosynthetic carbon assimilation in the coccolithophorid *Emiliania huxleyi* (Haptophyta): evidence for the predominant operation of the C3 cycle and the contribution of B-carboxylases to the active anaplerotic reaction, Plant & Cell Physiology 50 (2009) 318–329.

[70] P.J. McGinn, F.M.M. Morel, Expression and inhibition of the carboxylating and decarboxylating enzymes in the photosynthetic C4 pathway of marine diatoms, Plant Physiology 146 (2008) 300–309.

[71] C. Chastain, C. Failing, L. Manandhar, M.A. Zimmerman, M.M. Lakner, T.H.T. Nguyen, Functional evolution of C4 pyruvate, orthophosphate dikinase, Journal of Experimental Botany 62 (2011) 3083–3091.

[72] B. Tjaden, A. Plagens, C. Dorr, B. Siebers, R. Hensel, Phosphoenolpyruvate synthetase and pyruvate, phosphate dikinase of *Thermoproteus tenax*: key pieces in the puzzle of archaeal carbohydrate metabolism, Molecular Microbiology 60 (2006) 287–298.

[73] A. Hutchins, J. Holden, M. Adams, Phosphoenolpyruvate synthetase from the hyperthermophilic Archaeon *Pyrococcus furiosus*, Journal of Bacteriology 183 (2001) 709–715.

[74] H. Imanaka, A. Yamatsu, T. Fukui, H. Atomi, T. Imanaka, Phosphoenolpyruvate synthase plays an essential role for glycolysis in the modified Embden–Meyerhof pathway in *Thermococcus kodakarensis*, Molecular Microbiology 61 (2006) 898–909.

[75] C.J. Chastain, R. Chollet, Regulation of pyruvate, orthophosphate dikinase by ADP-/Pi-dependent reversible phosphorylation in C3 and C4 plants, Plant Physiology and Biochemistry 41 (2003) 523–532.

[76] J. Lewis, A. Harris, K. Jones, R. Edmonds, Long-term survival of marine planktonic diatoms and dinoflagellates in stored sediment samples, Journal of Plankton Research 21 (1999) 343–354.

[77] A. Kamp, D. de Beer, J.L. Nitsch, G. Lavik, P. Stief, Diatoms respire nitrate to survive dark and anoxic conditions, Proceedings of the National Academy of Sciences of the United States of America 108 (2011) 5649–5654.

[78] M. Rider, L. Bertrand, D. Vertommen, P. Michels, G. Rousseau, L. Hue, 6-Phosphofructo-2-kinase/fructose-2,6-bisphosphatase: head-to-head with a bifunctional enzyme that controls glycolysis, Biochemistry Journal 381 (2004) 561–579.

[79] C.J. Chastain, J.P. Fries, J.A. Vogel, C.L. Randklev, A.P. Vossen, S.K. Dittmer, et al., Pyruvate, orthophosphate dikinase in leaves and chloroplasts of C3 plants undergoes light-/dark-induced reversible phosphorylation, Plant Physiology 128 (2002) 1368–1378.

[80] F.J. Bruggeman, H.V. Westerhoff, The nature of systems biology, Trends in Microbiology 15 (2007) 45–50.

ACKNOWLEDGEMENTS

Chapter 1, in full, is a reprint of the material as it appears in Algal Research 2012. Smith, Sarah R.,; Abbriano, Raffaela M.; and Hildebrand, Mark. “Comparative analysis of diatom genomes reveals substantial differences in the organization of carbon partitioning pathways”, Algal Research, 1:2, 2012. The dissertation author was the primary researcher and author of this paper.

Chapter 2

Integration of carbon and energy metabolism at the transcript level in *Thalassiosira pseudonana*

2.1 ABSTRACT

Analysis of the transcriptional and physiological response of *T. pseudonana* on a silicon starvation time-course revealed a complex response of several carbon metabolic pathways. Generally, lipid metabolism genes are up regulated along with genes for carbon fixation, and carbohydrate storage. Genes involved in nitrogen metabolism and growth, such as ribosomal proteins and pathways that supply cellular respiration, are down regulated. A highly coordinated transcriptional response of large suites of genes was documented, indicating that the regulation of many different metabolic pathways and cellular processes are synchronized. This observation revealed the existence of a global mechanism that integrates multiple cellular functions. Transcriptional changes are correlated with documented changes in growth (i.e. chloroplast replication) indicating that cell cycle progression globally affects gene expression in *T. pseudonana.* These findings provide valuable insight into the significance of transcript level changes in marine microalgae, and have informed successful metabolic engineering of *T. pseudonana* to enhance lipid production.

2.2 INTRODUCTION

Diatoms are photosynthetic marine microbial eukaryotes that form the base of oceanic food webs and are important in biogeochemical cycling of many elements (C, N, Si, Fe). Additionally, many diatoms are highly lipid productive organisms, making them candidates for the development of renewable biofuels from microalgae (Hildebrand et al. 2012, Levitan et al. 2014). They were the first group of marine phytoplankton with a representative to have a full genome sequence available (Armbrust et al. 2004) and subsequently, many additional diatom genomes have been sequenced (Bowler et al. 2008, Lommer et al. 2012, www.jgi.doe.gov). The availability of genome data and focused efforts to develop tools for genetic manipulation of diatoms, facilitate in-depth investigation of many aspects of the biology of this group of organisms (Poulsen et al. 2005, Poulsen et al. 2006, Bozarth et al. 2009). However, while genomes provide valuable insight into the evolutionary history and metabolic potential of these important organisms, there is still little known about dynamic aspects of genomes such as gene expression and its regulation (transcription, translation, etc.). Functional genomics can be used to characterize how organisms use the information encoded in genomes to replicate and adapt to fluctuating environmental conditions.

Transcriptomics generally aims to characterize how mRNA-level changes underlie a variety of physiological, metabolic, and developmental processes. Advances in high-throughput sequencing have facilitated culture-based transcriptomic studies in diatoms, which have provided insight into the adaptive response of these organisms to environmental change such as high light (Park et al. 2010), phosphorus stress (Dyhrman et al. 2012), nitrogen and silicon starvation (Hockin et al. 2012, Mock et al. 2008), and iron starvation (Allen et al. 2008). A more complete understanding of the significance of transcript data as related to cellular response is important to understand many aspects of the basic biology of diatoms and other microalgae and has implications for studies of environmental populations, and for biotechnology. For example, metabolic engineering to improve lipid productivity is widely considered to be an essential element of economic feasibility and commercialization of algae as a feedstock for renewable biofuels (Radakovits 2010, Davis et al. 2011). The appropriate selection of targets for effective genetic engineering is required to accomplish this. Targets not only include the genes to be manipulated, but the steps in gene expression that will have the largest overall impact on metabolic activity.

Regulation of cellular function is multi-leveled and complex and as a result, transcriptomes alone are not necessarily predictive of the true physiological or metabolic response. At one level, metabolic flux is regulated by the activities of enzymes in an individual pathway, and there are many factors that affect enzyme activity such as mass action, allosteric interactions, and post-translational modifications (Plaxton 1996). Transcript levels affect overall enzyme activity through regulating cellular potential to synthesize new proteins. In plants this level of regulation is considered coarse and generally occurs during development or under long-term adaptation (Plaxton 1996). In other unicellular eukaryotes (i.e. yeast), a correlation between transcript level changes and metabolic responses has been demonstrated (Tu et al. 2005). The relationship between transcript changes and metabolic shifts has been examined only in a limited context in algae. Since most genetic engineering techniques involve artificially regulating mRNA levels through over-expression or knock-down techniques, a large contribution by transcript level over metabolic response would be advantageous.

In diatoms, many genes are transcriptionally responsive to specific environmental conditions suggesting that changes in transcript levels correlate with metabolic shifts. For example, the expression of silicon transporters is up regulated during silicon starvation in diatoms (Hildebrand et al. 1998), and genes in the photorespiratory pathway are up regulated during increased glycolate production (Parker et al. 2004). Additionally, genes such as LHCX1 and AUREOCHROME 1-a in *Phaeodactylum tricornutum* are known to be light responsive (Bailleul et al. 2010, Costa et al. 2013). In the case of AUREOCHROME 1-a the specific mechanism by which light regulates cell division through the activity of dsCYC2 (a diatom-specific cyclin) has been elucidated (Huysman et al. 2013). Alternatively, recent functional genomics studies in diatoms have shown that transcription of large suites of genes is coordinated with growth-related processes such as chloroplast division, release from silicon starvation and cell wall synthesis, and circadian shifts, suggesting that many genes may be under the control of master regulators (i.e. redox state, transcription factors) which choreograph genome-wide transcription (Gillard et al. 2008, Ashworth et al. 2013, and Shrestha et al. 2012). Consequently, transcript level changes can be interpreted as either a response to adapt to a specific environmental condition or as a part of a coordinated regulatory program associated with growth. Since eukaryotic gene expression is complex, it is possible that both factors are involved.

In many algae, including diatoms, starvation for essential nutrients (i.e. nitrogen, phosphorus) induces the formation of triacylglycerol-rich lipid droplets coincident with an arrest in growth (Yu et al. 2009, Hu et al. 2008). Silicon is required for cell wall synthesis and growth in most diatoms, and silicon-starvation also induces growth arrest and the formation of lipid droplets (Roessler 1988). In contrast to nitrogen starvation (Hockin et al. 2012, Gasch et al. 2000), silicon starvation causes little decrease in overall metabolic activities making it a unique system to distinguish between cell cycle arrest and secondary effects that arise during nutrient limitation (Darley and Volcani, 1969).

We present a transcriptomic analysis of the response of *T. pseudonana* during silicon limitation. To relate transcript-level changes to cellular processes, we also evaluated growth data (cell concentration, cell cycle progression), cellular composition (lipid levels, pigment levels), and photophysiology (carbon fixation, physiological fluorescence). The data provide insight into a variety of cellular processes, and indicate that transcripts of many carbon and energy metabolism genes are de-coupled from shifts in metabolism and physiology, and that transcript levels of certain genes are more likely regulated by growth and division rather than in response to certain environmental stimuli. A high degree of coordinate regulation of transcript levels for genes involved in distinct photosynthetic and metabolic processes in different cellular compartments was documented. This indicates that many distinct carbon and energy sources and sinks are integrated at the transcript level, perhaps through the existence of a master regulator or regulators.

2.3 METHODS

2.3.1 Culture conditions

Axenic 8L cultures of *Thalassiosira pseudonana* (CCMP1335) were grown in artificial seawater medium (NEPC, http://www3.botany.ubc.ca/cccm/NEPCC/esaw.html) at 18°C under continuous light (150 µmol m^{-2} s^{-1}) to a concentration of approximately $1x10^6$ cells ml^{-1} and then harvested by centrifugation for 12 min at 3100 x *g*. and placed in 8L of silicic acid free medium in a polycarbonate bottle at a concentration of approximately $5x10^5$ cells ml^{-1}. Cultures were stirred and bubbled with air under constant continuous light. Cultures were sampled at 0, 4, 8, 12, 18, and 24 hour time points following inoculation into silicic acid free medium to evaluate the parameters in Table 2. and described in the following. Experiments were conducted over the course of several years from 2009 – 2013.

2.3.2 Cell counts and cell cycle analysis

Cell concentrations were determined by Neubauer hemocytometer with a minimum of 200 cells counted per sample. For cell cycle stage determination, 12.5 ml of experimental culture was harvested by centrifugation (6 min at 4000 x *g*), extracted in 12ml of 100% ice-cold methanol, and kept at 4°C until analysis. Cells were pelleted, and washed 3X with TE (pH 8.0), and then re-suspended in 1ml TE to treat with RNase A (0.3 mg ml^{-1}) at 37°C for 40 minutes. Cells were stained with SYBR® Green I (1X final concentration from a 100X SYBR® Green I stock made in DMSO, Life Technologies^{TM}) for >10 minutes and kept on ice and in the dark until analysis within 4 hours. Cells were analyzed using the Becton Dickinson In-flux sorting cytometer (BD Biosciences, San Jose, CA). Data were analyzed using FlowJo cytometry software (Tree Star Inc. Ashland, OR).

2.3.3 Monitoring lipid accumulation by Nile Red and BODIPY

At each time point, triplicate samples (10 ml each) were sampled from experimental culture and pelleted by centrifugation (6 min at 3,200 x *g*). Pellets were kept frozen at -20°C until

Table 2. Summary of silicon starvation experiments conducted and parameters quantified. Experiments were conducted over the course of several years 2009-2013.

Experiment	Cell Counts	Lipid Dye	Instrument	FAME Analysis	HPLC Pigments	Cell Cycle	RNA	Pvs. E	FRRf	Chloroplast Analysis
Si- #1	--	Nile Red	Mirotiter Plate Reader	--	--	--	--	--	--	--
Si- #2	--	Nile Red	Mirotiter Plate Reader	--	--	--	--	--	--	--
Si- #3	--	Nile Red	Mirotiter Plate Reader	--	--	--	Microarray	--	--	--
Si- #4	--	Nile Red	Mirotiter Plate Reader	--	--	--	--	--	--	--
Si- #5	x	Bodipy	Imaging Flow Cytometer	--	x	x	--	--	--	--
Si- #6	x	Bodipy	Imaging Flow Cytometer	--	x	x	--	--	--	--
Si- #7	x	Bodipy	Imaging Flow Cytometer	x	x	--	--	x	x	--
Si- #8	x	--	Imaging Flow Cytometer	--	--	--	--	--	--	x
Si- #9	x	Bodipy	Imaging Flow Cytometer	x	--	x	RNA-Seq	--	--	--
Si- #10	x	Bodipy	Imaging Flow Cytometer	x	--	x	--	--	--	--

analysis. For Nile Red analysis, each pellet was re-suspended in 1ml 2.3% NaCl solution and stained with 6.25 µl Nile Red working stock (12.5mg L^{-1} in acetone). Samples were incubated in the dark for 20-30 min before being read in triplicate on a microtiter plate reader (200 µl) at exc/em: 485/576nm. For BODIPY analysis, each pellet was re-suspended in 500 µl 2.3% NaCl and stained with a final concentration of 2.6 µg ml-1 of a 1 mg ml-1 BODIPY stock dissolved in DMSO (BODIPY 493/503, Life Technologies^{TM}). Samples were stained on ice for a minimum of 10 min and a maximum of 4 hr and kept on ice. BODIPY fluorescence was quantified using imaging flow cytometry (ImageStreamX, Amnis Corp., Seattle, WA) with a 488 nm laser with two neutral density filters (0.6 and 1.0). Post acquisition spectral compensation and data analysis were performed using IDEAS software (Amnis Corp.). Fluorescence intensity was determined for a minimum of 5,000 cells.

2.3.4 FAME analysis

For lipid analysis, approximately 800ml of culture was harvested by centrifugation and rinsed with 0.4M ammonium formate. Cells were placed in a glass tube under N2 gas and stored at -80°C until analysis. Before analysis, cells were lyophilized overnight and the amount of matter submitted to the extraction was accurately weighed (between 5-13 mg). Lipids were extracted following the Folch et al. (1957) method. First, lipid extracts were homogenized with a 2:1 chloroform:methanol (v:v) with 0.01% of BHT (butylated hydroxytoluene) solvent mixture and ultrasonicated during 20 min in an ice bath. The mixture was then washed with a saline solution of 0.88% KCL.

Fatty acids in the extracts were determined by transmethylation with 14% boron trifluoride/methanol at 70°C. The concentrations of the fatty acid methyl esters (FAME) were determined with a Hewlett Packard 5890 Series II gas chromatograph equipped with a flame ionization detector (GC/MS-FID; Dodds et al. 2005). Separation was achieved in a DB-Wax column (30m, 0.25mm, and 0.25um) with a helium carrier gas at a flow of 1.4 ml min^{-1}. The injector, FID, and MS sources temperatures were 220°, 300°, and 230°C, respectively. Samples (1 µl) were injected with a split ratio of 50:1. FAMEs were identified by comparison of retention times to authentic lipid standards (68A and 68D, Nu-Check Prep, USA). Quantification of C14 to C18 and C20 to C24 fatty acids was performed relative to known concentrations and Flame Ionization Detection (FID) peak areas of C13 and C19 FAME internal standards, respectively. The response factor for each standard compound (in GLC 68A and 68D) of interest was applied to the respective sample peak areas to calculate the final concentration of each peak of interest. The fatty acid methylation reaction (i.e., transesterification) efficiency was determined from the C15 FA internal standard added to each lyophilized algal pellet. Each sample was analyzed in duplicate.

2.3.5 Pigment analysis

Samples (10mL) were collected and immediately filtered through GFF Whatman filters (25mm) and stored at -80°C until their analysis by HPLC. Filters were extracted in 90% HPCL-grade acetone and ultrasonicated in an ice bath for 10 min, stored at -20°C for 24 hr and pre-filtered through a 0.45 um Whatman nylon Puradisk filter before injection. Afterwards, pigments were separated using a reverse phase C-8 column following the method of Zapata et al. (2000). External pigment standards were used for system calibration (DHI, Denmark). Samples were analyzed on a Waters 600 HPLC system, equipped with a Thermo Separation Products AS3000 sampler, a TSP Spectra 100 variable wavelength (fixed at 440 nm), a Waters 996 diode array (scanning 330-800 nm) and a Waters 470 scanning fluorescence detector. Data were collected using the Waters Millennium 32 software package.

2.3.6 Transcriptomics

2.3.6.1 RNA isolation

For experiments Si- #3 and Si- #9, 750ml of culture was removed, treated with cycloheximide (20 µg ml-1) and harvested by filtration. Cells were pelleted and stored at -80°C prior to total RNA isolation (Hildebrand and Dahlin 2000). RNA from experiment Si- #3 was processed for hybridization to an Affymetrix GeneChip whole genome tiling array while RNA from experiment Si- #9 was processed for Illumina-based RNA-Seq.

2.3.6.2 Microarray hybridization and processing

The Affymetrix microarray was designed and analyzed at the gene and exon level with a total of 524,909 sense strand probes (average of 16 probes per gene), based on gene model predicted transcripts for *T. pseudonana,* version 3.0 (http://genome.jgi-psf.org/Thaps3/Thaps3.home.html). Included on the array were 33,886 antigenomic probes to account for nonspecific hybridization. For each hybridization, double stranded cDNA was synthesized from 7 µg of total RNA with no amplification using the GeneChip® WT Amplified Double-Stranded cDNA Synthesis Kit (Affymetrix). Cleanup of double-stranded cDNA was done with the GeneChip® Sample Cleanup Module (Affymetrix). Fragmentation and end-labeling was performed using the GeneChip® WT Double-stranded DNA Terminal Labeling Kit (Affymetrix). Hybridization of labeled targets on the arrays was carried out using the GeneChip® Hybridization, Wash, and Stain Kit (Affymetrix). The arrays were then scanned with the GeneChip® Scanner, to generate the probe cell intensity data files. Initial data analysis was performed as described in (Shrestha et al. 2012).

2.3.6.3 RNA-Seq sequencing and processing

Prior to library preparation, each RNA sample was subjected to quality control evaluation as follows. The concentration and purity of the RNA samples were assayed by a NanoDrop spectrophotometer (Thermo Scientific). Each sample was required to have an A_{260}/A_{280} ratio between 2.0 and 2.2, and an A_{260}/A_{230} ratio above 2.0. The RNA samples failed to pass the A_{260}/A_{230} ratio test, and were subjected to an additional precipitation to remove contaminants. Using standard laboratory procedures, the samples were precipitated by the addition of 0.1 volumes of sodium acetate (pH = 7.5) and 2.5 volumes of 100% ethanol. After a 30 min incubation at -20°C, the RNA was pelleted by centrifugation for 30 min at 4°C at maximum speed in a refrigerated microfuge, and the supernatant was decanted. The RNA pellet was washed with 70% ethanol, decanted, air dried for 15 min at room temperature, and finally dissolved in 50 µL dH_2O. Following this precipitation, all samples had A_{260}/A_{230} ratios above 2.0. RNA quality was evaluated by BioAnalyzer (Agilent Technologies) on an Agilent RNA 6000 Nano chip following the manufacturer's instructions. RNA integrity was quantified by Agilent 2100 Expert software.

Next, cDNA libraries were prepared from 4 µg of total RNA for each sample using the Illumina TruSeq Stranded mRNA Sample Prep kit (Illumina) according to manufacturer's protocols (Rev. D). The resulting libraries were evaluated for size on an Agilent DNA High Sensitivity chip following the manufacturer's instructions, and quantified by Q-PCR according to Illumina's protocols. Each library was diluted to 2 nM with 10 mM Tris-Cl, 0.1% Tween 20. Bar coded libraries were then pooled (10 or 11 per pool), and sequenced on a Illumina HiSeq 2000 with 50 nt single end reads.

Raw reads were demultiplexed and quality trimmed and all reads mapped to the Tp3 exons (Thaps3 genome downloaded from JGI) using the CLC Assembly Cell (CLCbio, http://www.clc-bio.com). Quality trimming was done to a minimum score of 33 and a minimum length of 20bp. Genes were further annotated using PhyloDB, a comprehensive in-house database of proteins at JCVI. RPKMs were calculated for each library using the sum of lengths of exons for each gene and the coverage of reads mapped to each exon, using a Perl script.

2.3.6.4 Comparison of microarray and RNA-Seq data

To quantify the replication of the response in both experiments, the Pearson's correlation coefficient (PCC) between the response of a given gene from the microarray and RNA-Seq data was determined. The distribution of PCCs shows a generally good agreement, indicating an overall correlation between experiments (Supplemental Figure 1). Microarray data are plotted throughout as fold-change (log base 2) relative to the 0 time point, and cases where microarray and RNA-Seq expression patterns are not replicated (PCC <0.5) are noted.

2.3.6.5 Expression clustering and functional enrichment analysis

K-means clustering was repeated using Genesis (1.7.6) using 4, 9, 16, 25, 36, 49, 64, and 81 clusters with 100 iterations. Accurate clustering is reflected by a low mean within cluster variance, which decreases with an increased number of clusters. The minimum number of clusters with an acceptable mean within cluster variance was chosen to (n = 36, 0.18). The numbers of genes annotated with KOG Classes were tabulated for each cluster. Since both clusters and KOG Classes have different numbers of genes attributed to them a functional enrichment index (FEI) was calculated to determine if clusters were disproportionately enriched in any functional categories:

$$FEI = (G_{clusterclass}/(G_{cluster} * G_{class}))*1000$$

Where $G_{clustercategory}$ is the number of genes assigned to both a given expression cluster and KOG class, $G_{cluster}$ is the total number of genes represented in that cluster, and G_{class} is the number of genes in a given KOG Class.

2.3.7 Photosynthesis vs. irradiance experiments

Photosynthetic characteristics were obtained from photosynthesis–irradiance (P–I) measurements using a modified ^{14}C-bicarbonate incorporation technique (Lewis and Smith 1983; Arrigo et al. 1999). P-I incubations were carried out in a photosynthetron incubator at 24 irradiances ranging from 0 to 830 μmol photons m^{-2} s^{-1}. Illumination was provided by two 150W tungsten-halogen lamps and adjusted within each vial chamber by neutral density filters. Total photosynthetically available radiation (PAR, 400–700 nm) within each illumination chamber was measured using a Biospherical Instrument QSL 101 sensor. P-I experiments were carried out at 18°C.

For each P–I curve, a 60 ml sample of culture was spiked with 0.06 mCi $NaH^{14}CO_3$ to obtain a final activity of 0.001 mCi mL^{-1}. The spiked sample was distributed in 2 ml aliquots to 7 mL glass scintillation vials. 24 of these vials were placed in the individually illuminated chambers within the incubator and incubated for 1 hr. Two vials (time-zero samples) were acidified immediately with 200 μL of 20% HCl and placed under the hood during the 24 hr experiment. Total activity in the samples was determined by adding 100 μL of sample at t=0 into two scintillation vials containing 100 μL of 1M NaOH. 5 mL of liquid scintillation cocktail (ECOLUME^{TM}) was then added.

After incubation, the samples were acidified with 200 μL of 20% HCl and placed under the hood for 24 hr to drive off unincorporated inorganic radioisotope. After 24 hr ventilation, all samples received 5 mL of ECOLUME^{TM} liquid scintillation cocktail and were vigorously shaken.

Carbon uptake, normalized by Chl *a* and/or cell concentration, was calculated from radioisotope incorporation. The photosynthetic response was modeled by curve fitting as suggested by Platt et al (1975):

$$P^B = P_m^B \tanh(\alpha * E / P_m^B)$$

Where P^B is Photosynthesis per unit biomass (Chl *a* or cell concentration) in units mgC mgChla^{-1} h^{-1}, P_m^B is maximum rate of photosynthesis per unit of biomass, α is the initial slope in units of [mgC mgChla^{-1} h^{-1} (µmol photons m^{-2} s^{-1})$^{-1}$] and E is irradiance in units of µmol photons m^{-2} s^{-1} .

2.3.8 Imaging flow cytometry for chloroplast replication analysis

At each time point, samples (10 ml each) were sampled from experimental culture and pelleted by centrifugation (6 min at 3,200 x *g*). Pellets were kept frozen at -20°C until analysis. Pellets were re-suspended in 500 µl 2.3% NaCl solution and analyzed using imaging flow cytometry as described above. Post acquisition data analysis was performed using IDEAS software using a minimum of 2000 cells (Amnis Corp.). Chloroplasts were quantified by creating a custom spot count feature in IDEAS.

2.4 RESULTS

2.4.1 Silicon-starvation induced lipid accumulation

Following transfer of *Thalassiosira pseudonana* to silicon-free medium, cell concentration (cells ml^{-1}) did not increase significantly, however there was a detectible shift in the proportion of the population in the G1 and G2+M phases of cell cycle between 0-4 hr indicating that while cell division in the majority of the population arrests immediately, progression through the cell cycle does not (Figure 9A, 9B). RNA isolated from 6 time points (0, 4, 8, 12, 18, and 24h) from experiments Si- #3 and Si- #9 was processed for hybridization to Affymetrix whole genome tiling arrays designed for the *T. pseudonana* genome, and for Illumina-based RNA-Seq respectively. Both types of data were analyzed as described in Methods. All data plotted throughout are microarray results supported by RNA-Seq replication. Instances in which the transcript changes did not replicate the pattern seen in microarray data are noted. Transcripts encoding silicon transporters (SIT1 and SIT2) were up regulated significantly at 4hr and remained high throughout the time course confirming at the transcript level that the culture was experiencing silicon limitation (Figure 9C, Hildebrand et al. 1998).

Cellular lipid levels were reproducibly induced between 8-12 hr of silicon starvation in all experiments (Figure10). This induction corresponded to a nearly 3-fold increase in cellular fatty acid methyl ester (FAME) levels by 24 hr accompanied by a shift in the FAME composition (Figure 10B). The dominant fatty acid methyl ester (FAME) produced initially was palmitic acid (C16:0), but after 12 hr of silicon starvation the FAME composition was dominated by palmitoleic acid (C16:1) (Figure 10B). Palmitoleic acid is a major constituent of diatom triacylglycerol (Yu et al. 2009), and this shift in FAME coincided with the onset of lipid droplet formation as seen by BODIPY staining (Figure 10B).

Several genes involved in lipid production, including genes for fatty acid biosynthesis, modification, and glycerolipid biosynthesis, were differentially expressed during the 24-hour silicon starvation period (Figure 11). De novo biosynthesis of fatty acids occurs in the chloroplast of diatoms via a type II fatty acid synthase (FAS-II, Armbrust et al. 2004). The first committed step of fatty acid synthesis is catalyzed by acetyl-coA carboxylase (ACCase). There are two copies of ACCase in the *T. pseudonana* genome, one of which is predicted to be localized to the chloroplast (Thaps3_6770) and the other to the cytosol (Thaps3_12234), which are both up-regulated during the 24-hour silicon starvation period (Figure 11). Transcripts for nearly all of the

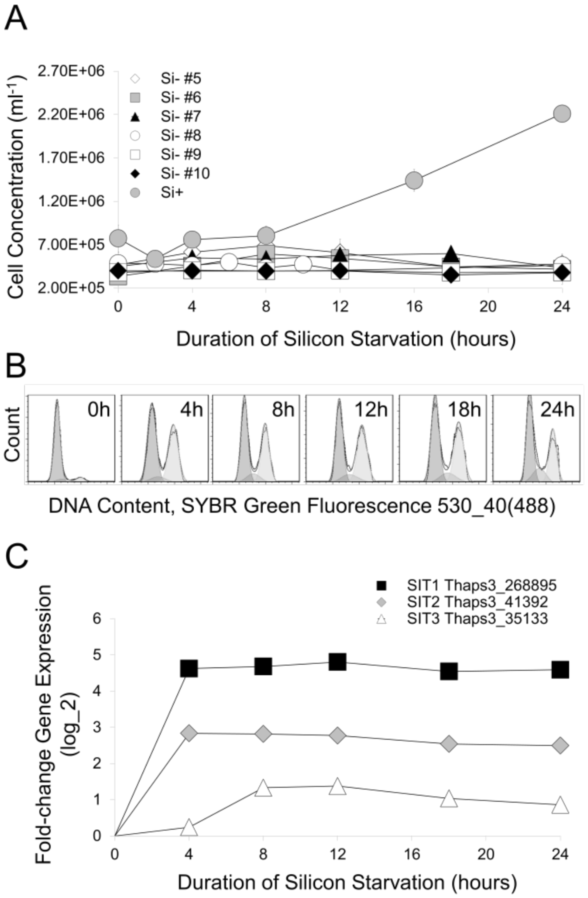


Figure 9. Effect of silicon starvation on A) growth, B) cell cycle stage, and C) transcript levels of silicon transporters in *T. pseudonana* incubated in continuous light. A) Data from silicon starvation experiments are shown with a silicon replete control (Si+). B) Representative cell cycle histograms (from experiment Si- #5) are shown. In each histogram, the left peak are cells in the G1 phase of the cell cycle, and the right peak G2+M. C) Microarray fold change gene expression (relative to t=0) is plotted. Pearson's Correlation Coefficients with RNA-Seq data are Thaps3_268895 (0.62), Thaps3_41392 (0.76), Thaps3_35133 (0.73).

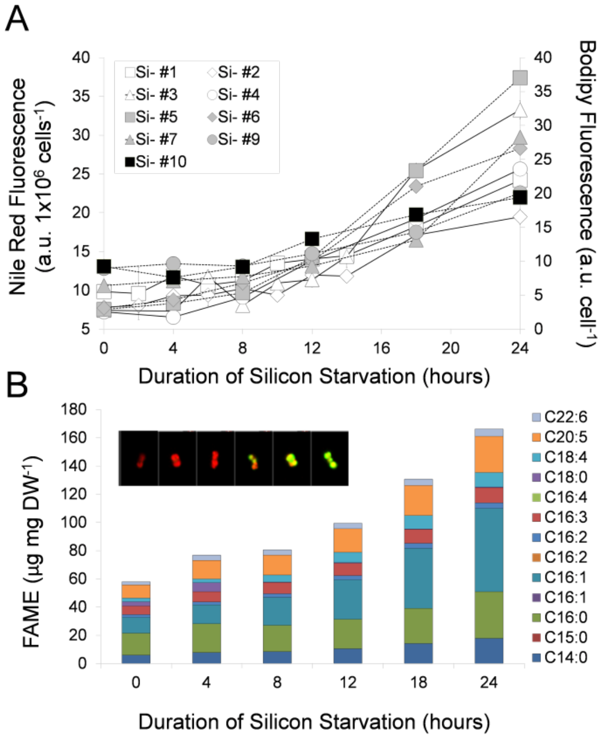


Figure 10. Lipid accumulation profiles of silicon-starved *T. pseudonana*. (A) Lipid accumulation determined fluorometrically using lipophilic fluorescent dyes Nile Red (solid lines) and BODIPY (dashed lines). (B) Representative fatty acid methyl ester (FAME) profile (from experiment Si- #7). Inset micrographs show chlorophyll (red) and BODIPY-stained lipid droplets (green) in a cell representing the average of the population at each experimental time point, showing the induction in lipid droplet formation at 12 hr.

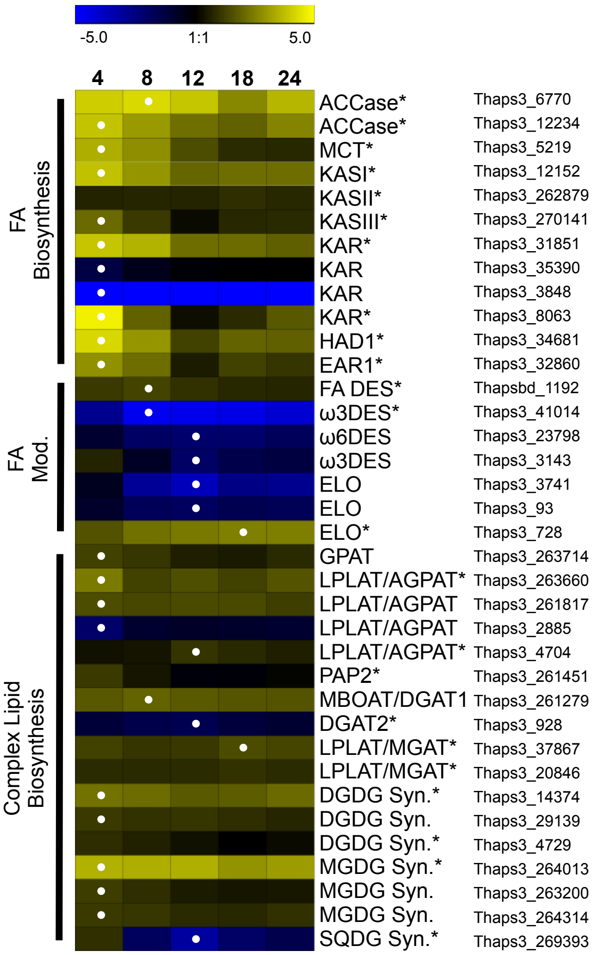


Figure 11. Transcript level changes of genes involved in *de novo* fatty acid synthesis, modification of fatty acids, and complex lipid biosynthesis during silicon starvation in *T. pseudonana*. Heatmap shows microarray data as log_2 fold-change in transcript level relative to t=0. White dots indicate the time of maximum or minimum transcript-level change for each gene. Asterisks following the gene name indicate the pattern was replicated in RNA-Seq data. (Pearson's Correlation Coefficient >0.5).

enzymes catalyzing subsequent steps of the fatty acyl synthase pathway were also significantly up regulated (Figure 11). Several desaturases responsible for the synthesis of unsaturated fatty acids through the introduction of double bonds were identified in the *T. pseudonana* genome. Most notably, a delta-9 fatty acid desaturase (FAD, Thaps3_1192_bd) was the most up-regulated and highly expressed (>300 RPKM) desaturase during silicon-starvation. This FAD (Thaps3_1192_bd) is predicted to be located in the chloroplast and is co-expressed with the chloroplast acetyl-CoA carboxylase (ACCase, Thaps3_6770) implicating it as the enzyme responsible for the detected shift from palmitic acid (C16:0) to palmitoleic acid (C16:1).

Cellular polyunsaturated fatty acid (PUFA) levels increased during silicon starvation though most genes predicted to be involved in the synthesis of PUFA were down regulated (Figure 10B, Figure 11). Two omega-3 fatty acid desaturases (Thaps3_41014, Thaps3_3143), and one omega-6 (Thaps3_23798) fatty acid desaturase were down regulated (Figure 11). Of the three fatty acid elongases from the GNS1/SUR4 family (IPR002076) in the *T. pseudonana* genome, two were down regulated (Thaps3_3741, Thaps3_93). Comparison with PUFA levels (Figure 10B) suggests that if the transcriptional regulation of PUFA synthesis enzymes acts to regulate the synthesis of PUFA, it occurs on longer time-scales than those observed in this short-term silicon starvation and that PUFA synthesis is more likely to be regulated by other factors regulating the flux of metabolite precursors into the pathway.

Finally, several components of complex lipid biosynthesis machinery were up regulated including enzymes putatively involved in TAG biosynthesis and in the synthesis of polar thylakoid lipids (Figure 11). Changes in the transcript levels of TAG biosynthesis enzymes were generally not well replicated between experiments with the exception of two putative LPLAT/AGPATs (Thaps3_263660, Thaps3_4704), a LPLAT/MGAT (Thaps3_37867). Therefore, these enzymes are the most likely to be transcriptionally regulated in the TAG biosynthesis pathway under silicon starvation.

Monogalactosyldiacylglycerol (MGDG) and digalactosyldiacylglycerol (DGDG) are neutral galactolipids found in thylakoid membranes. In algae, thylakoid membranes can comprise a majority of total cellular lipid (Janero and Barrnett 1982). Several enzymes responsible for the synthesis of these compounds were identified and their transcripts were up regulated under silicon-starvation, suggesting that the synthesis of thylakoid membranes is also up regulated (Figure 11). Sulfoquinovosyl diacylglycerol (SQDG) acts as a strong repressor of the diadinoxanthin de-epoxidation reaction, thereby inhibiting the dissipation of excess light energy through non-photochemical quench (NPQ, Lepetit et al. 2012). A single gene codes for SQDG synthase in *T. pseudonana*, and is strongly down regulated within 12 hr of silicon starvation (Figure 11). Though thylakoid lipid composition was not quantified, the simultaneous down-regulation of SQDG synthase with the up-regulation of MGDG and DGDG synthases suggests that silicon-starved *T. pseudonana* are adapting to the experimental conditions by modifying the lipid composition of the thylakoids.

2.4.2 Effect of silicon-starvation on light-harvesting complex, photosystem function, and carbon fixation

During autotrophic growth, the energy used to drive biosynthesis of lipids and other macromolecules is acquired from photons absorbed by the light-harvesting complex. Therefore, changes to the size or efficiency of the light-harvesting complex should directly affect carbon fixation rates and the flux of this fixed carbon into lipids. In response to silicon starvation, transcripts for nearly all components of the peripheral light-harvesting complex were strongly and coordinately up regulated for the entire experiment, and many components of the photosynthetic reaction centers were also up regulated (Figure 12A). Pigment biosynthesis genes (chlorophyll and carotenoids) were also significantly up regulated with a peak in transcript levels detected for most enzymes at 4 hr (Figure 12A). Total cellular pigment levels increased by 2.4-2.7 fold after

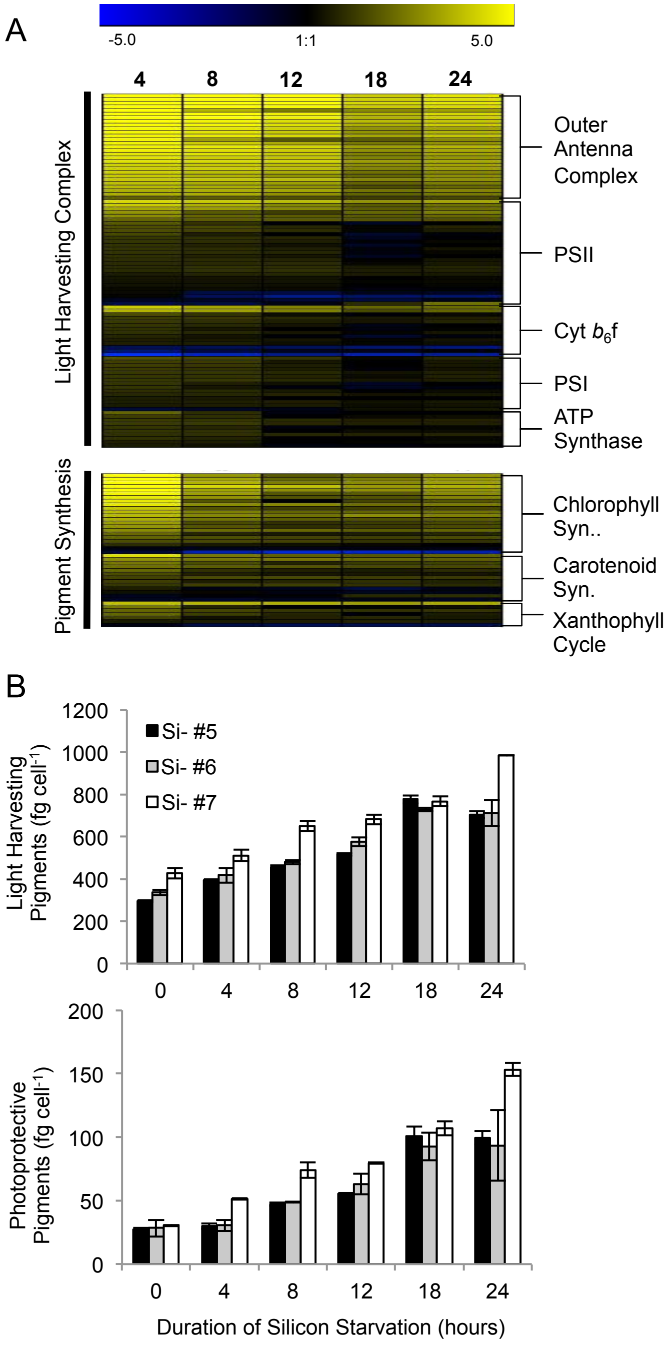


Figure 12. Effect of silicon starvation on A) transcript levels of light harvesting complex and pigment biosynthesis genes, and B) cellular pigment levels in *T. pseudonana*. Heatmap shows microarray data as log_2 fold-change in transcript level relative to t=0. Light harvesting pigments are defined as chlorophyll *a,* chlorophyll *c,* and fucoxanthin. Photoprotective pigments are diatoxanthin and diadinoxanthin. Error bars on pigment data show standard deviation of two technical replicates.

24 hr of silicon starvation (Figure 12B). Typically, an increase in cellular pigment levels is a photoadaptive response to low light as a means to increase light absorption (Falkowski and Raven 2007). Initially, the data seem to suggest that under these conditions, *T. pseudonana* is increasing capacity for light absorption through an increase in the light-harvesting antenna. However, a disproportionate increase in photoprotective xanthophyll pigments (diatoxanthin and diadinoxanthin, 3.3-5.1-fold per cell) relative to light-harvesting pigments (chlorophyll *a*, chlorophyll *c*, and fucoxanthin, 2.1-2.4-fold per cell), along with a large increase in the de-epoxidation state (DES) of the xanthophyll pool (Figure 13) is consistent with absorbed light energy being dissipated through NPQ. Somewhat paradoxically, silicon-starved *T. pseudonana* increases the size of the light-harvesting complex (increasing light absorption) while simultaneously dissipating at least some of this newly acquired light energy through increased NPQ.

Light energy collected with the light-harvesting complex is transferred to photosystems I and II (PSI and PSII). The quantum yield (measured as Fv/Fm) is an important indicator of what proportion of the light absorbed by PSII is then transferred to the photosynthetic electron transport chain where it is used to generate the NADPH and ATP that drives carbon fixation. There was a significant decrease in Fv/Fm at 8 hr (Figure 14A), which signifies either damage to the reaction center core or saturation of linear electron transport. This damage could be caused or worsened by the observed increase of the light-harvesting complex. After 8 hr, Fv/Fm recovers to approximately 80% of initial levels for the remainder of the experiment. This recovery is likely facilitated by the increase in NPQ as evidenced by the DES of the xanthophyll pool, which reduces the magnitude of the flow of damaging photons to PSII. Additionally, several proteins found at the reaction center core that are encoded on the plastid genome (D1, D2, CP43, CP47, cytochrome b559alpha, cytochrome b559beta) were up-regulated between 4-12 hr, perhaps to repair any damage sustained at the reaction center core (Figure 14B).

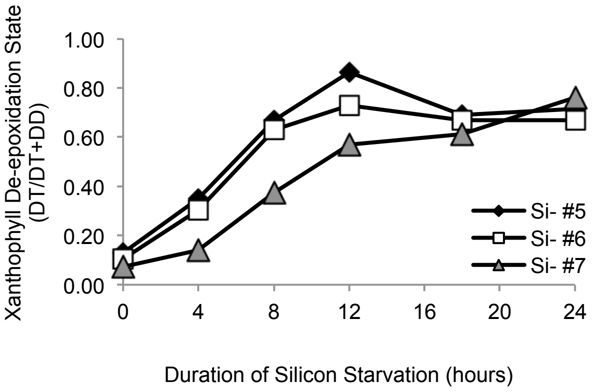


Figure 13. Xanthophyll de-epoxidation state (DES) during silicon starvation in *T. pseudonana* under continuous light. Data show the average DES from two technical replicate measurements from three experiments.

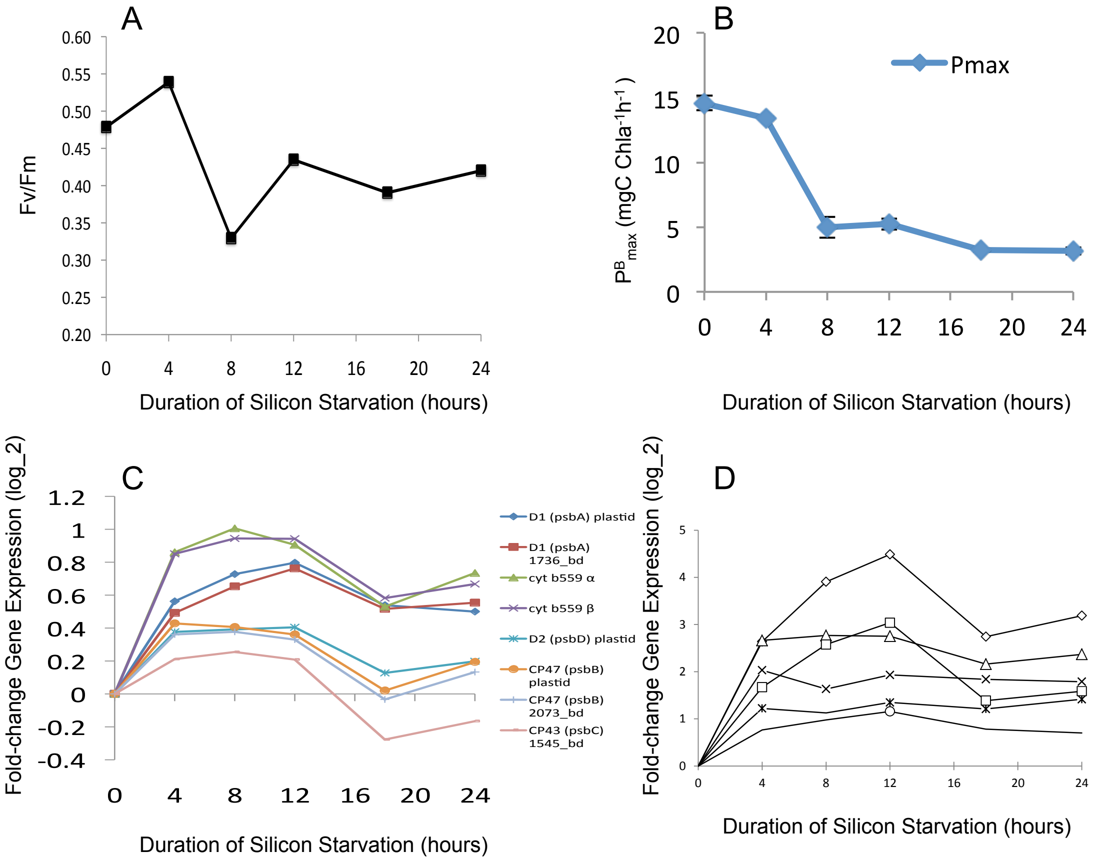


Figure 14. Shifts in photosynthetic parameters and transcripts of reaction center core proteins of *T. pseudonana* during silicon starvation. A) Plot of Fv/Fm showing a decrease in the photochemical quantum yield of photosystem II at 8 hr. B) Plot of Pmax, showing a decrease at 8 hr while the rate of carbon fixation at the experimental light intensity remains high. For A and B, data represent the average of two technical replicates from Si- #7. C) Microarray data showing that transcripts of genes encoding proteins in the reaction center core of PSII are up regulated maximally between 8-12 hr. D) Microarray data showing that transcripts of photorespiratory genes are up regulated maximally at 12 hr.

It is clear that there are several changes to the light-harvesting capacity of *T. pseudonana* during silicon-starvation induced lipid accumulation. To quantify any associated changes to carbon fixation during the time-course of silicon starvation, P-I curves were generated. Overall, carbon fixation rates were high at the experimental light intensity, however there was a dramatic and persistent decrease in Pmax at 8 hr indicating an impaired ability to fix carbon at higher light intensities (Figure 14B). There are several factors that can affect Pmax such as the magnitude of the supply of NADPH and ATP through photosynthetic electron transport. The observed decrease in Pmax coincided with a decrease in Fv/Fm, indicating that any damage sustained at the PSII reaction center core may be at least partially affecting carbon fixation at high light intensities. Another factor that can affect Pmax is photorespiration, or the fixation of O_2 instead of CO_2 by RuBisCO under high oxygen conditions. It was found that several key photorespiration genes are up-regulated maximally between 8-12 hr consistent with an increase in O_2 fixation is at least partially being responsible for the decrease in Pmax (Figure 14D, Parker et al. 2004).

2.4.3 Cell cycle progression coordinates genome-wide transcript-level changes

To put the expression patterns for photosynthetic and lipid metabolism genes in context of the genome-wide response of *T. pseudonana,* a global transcriptomic analysis was conducted. In the microarray experiment, nearly half (45%, 5347/11856) of the genes in the *T. pseudonana* genome were differentially expressed at some point during the 24 hr time course, and at any given time point at least 20% of the transcripts in the genome were up or down regulated relative to exponential conditions. The magnitude of the expression response was maximal at the 4hr time point (Figure 15), when 64% of the genes that were significantly expressed were at their maximum or minimum level. This dramatic, rapid, and coordinated transcriptional response of large suites of genes to silicon-starvation conditions occurred after cell growth arrested but preceded the onset of complete cell cycle arrest and lipid accumulation (Figure 9, Figure 10).

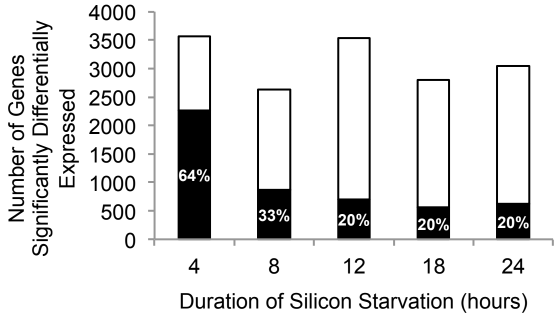


Figure 15. The number of differentially expressed genes in the *T. pseudonana* genome relative to t=0 during silicon starvation. Data labels show number of genes maximally expressed as a percentage of the total genes that are differentially expressed at any given experimental time point.

The function of genes with coordinated expression was investigated by grouping their expression patterns of genes (using *k*-means clustering, see Methods) into 36 clusters (Supplemental Figure 2) and by investigating which KOG classes were enriched in certain expression categories (Figure 15). Several clusters with a transcript minimum or maximum response at 4hr were identified. Clusters that were transiently up regulated at 4hr were disproportionately enriched in cell division related processes such as cell signaling, cell cycle, chromatin structure and dynamics (histones), cytoskeleton, nuclear structure, and transcription (Figure 16Bi). In contrast, down regulated clusters were enriched in genes associated with nucleotide transport and metabolism, RNA processing and metabolism, and translation, ribosomal structure, and biogenesis (Figure 16Bii). Overall, this shows evidence for transcript-level control of sustaining cell cycle progression in the short-term, while other growth processes such as the synthesis of nucleotides and new proteins are being shut down.

Several expression clusters contain genes that had a maximum at 4 hr but stayed up regulated in response to silicon starvation indicative of a prolonged response in addition to cell cycle related phenomena (Figure 16Biii). Generally, these clusters are functionally enriched in genes associated with energy production and conversion, as well as in metabolic pathways including coenzyme transport and metabolism and carbohydrate transport and metabolism. However, genes associated with certain metabolic pathways (i.e. glycolysis, lipid metabolism) were found in both up and down regulated clusters indicating that within given pathways the transcript-level responses were diverse (Figure 16iv).

The degree to which the transcript response of several genes is coordinated points to the existence of a global regulatory mechanism that integrates several cellular processes, many of which are related to coordinating the progression of cell cycle and growth. In addition to cell cycle progression related genes, many genes associated with chloroplast function showed a distinct expression response (generally up regulated) at 4 hr following silicon starvation. This

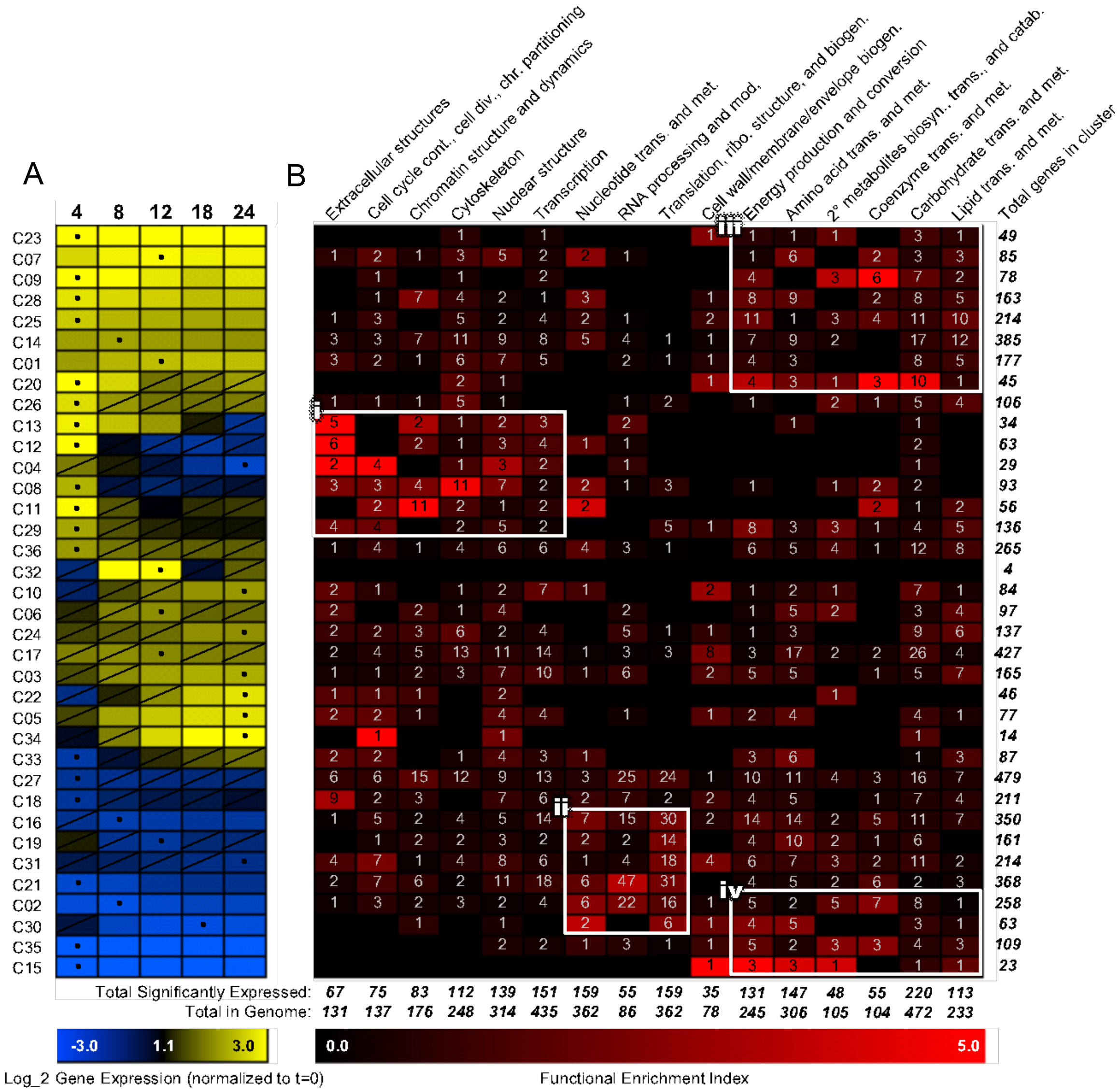


Figure 16. Function (KOG class) of genes clustered by expression pattern in *T. pseudonana* during silicon starvation. A) Heatmap shows microarray data for expression clusters (C1 – C36) and represents the average log_2 fold-change in transcript level relative to t=0 at 5 experimental time points (hours at top). The number of genes in each cluster is shown in the column to the right of part B. Dots indicate the time of maximum or minimum transcript-level change for each cluster. The functional roles of genes found within expression clusters are shown in B) where the number of genes with a given KOG class assignment is labeled for each category and expression cluster. The intensity of red in B) shows the functional enrichment index (see methods for equation) which demonstrates how enriched a given expression cluster is for a given KOG class by normalizing the genes found in that category and cluster by the total number of genes in that category (bottom row) and cluster (right hand column). Bi-iv) show regions of the heatmap referred to in the text.

includes genes from the plastid type II fatty acid synthase (FAS-II) as well as several genes from the Calvin-Benson cycle (Figure 17A, 17B). Imaging flow cytometric analysis of chloroplasts showed that during the 0-8 hr time period of silicon starvation, a significant proportion of cells in the *T. pseudonana* population replicate their chloroplasts (Figure 17C). Since both the FAS-II and Calvin-Benson cycle enzymes are chloroplast-localized, it seems likely that the peak in expression observed at 4 hr can be explained by the requirement to produce new chloroplasts rather than strictly by the up regulation of fatty acid biosynthesis or carbon fixation pathways. Several other genes encoding chloroplast proteins also peak in expression at 4 hr (Figure 11A). It is clear from this genome-wide analysis that many genes are up-regulated at 4 hr, and these genes are often involved in cell cycle and growth-related processes including the replication of chloroplasts in at least part of the *T. pseudonana* population. Furthermore, transcript-level changes in genes associated with carbon metabolism (lipid metabolism and carbon fixation) appear to be regulated by cell cycle progression in this experiment.

2.4.4 The coordinate regulation of carbon and energy metabolism

In addition to coordinate regulation of genes involved in cell cycle progression it was observed that the expression patterns of genes involved in distinct cellular processes and metabolic pathways clustered together, indicative of coordinate regulation. The up regulated, chloroplast localized ACCase (Thaps3_6770) is strikingly correlated with transcript-level changes of several other genes (Figure 18A). ACCase is co-expressed with both of the plastid-encoded subunits of RuBisCO (rbcS, rbcL). RuBisCO and ACCase are carboxylases requiring inorganic carbon (CO_2 or bicarbonate) as a substrate, and it is possible that they are co-expressed in this experiment in response to a shift in intracellular inorganic carbon conditions. The expression pattern of these enzymes is correlated with an un-annotated cadmium β-carbonic anhydrase (Thaps3_233), and with two genes (Thaps3_4819, Thaps3_4820) that possess a

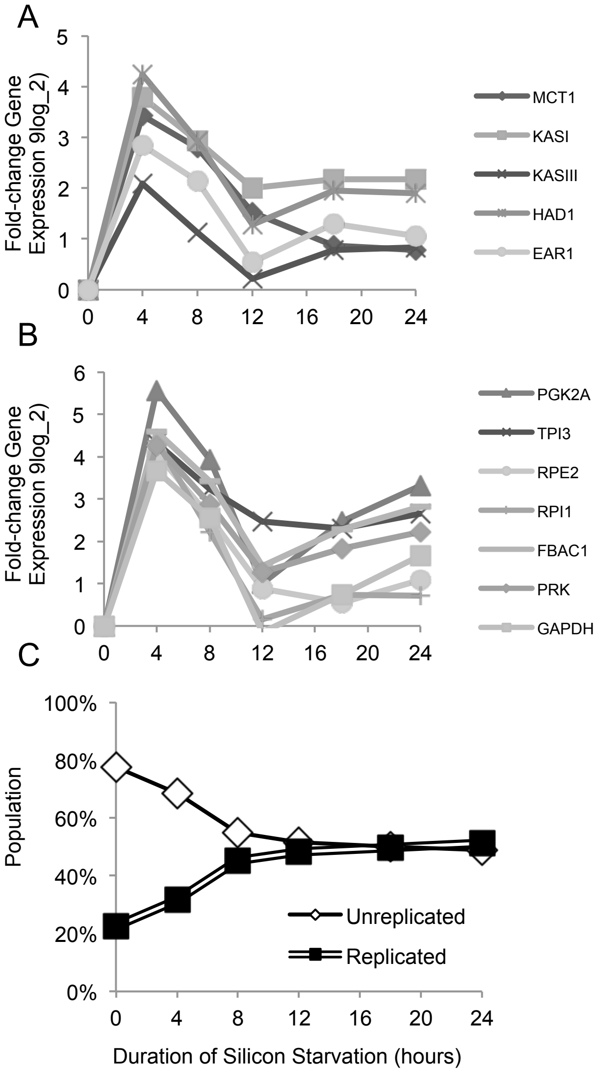


Figure 17. Correlation of transcripts of A) fatty acid biosynthesis enzymes and B) Calvin-Benson cycle enzymes with C) chloroplast replication in *T. pseudonana* during silicon starvation. Asterisks following the gene name indicate the pattern was replicated in RNA-Seq data (Pearson Correlation Coefficient >0.5).

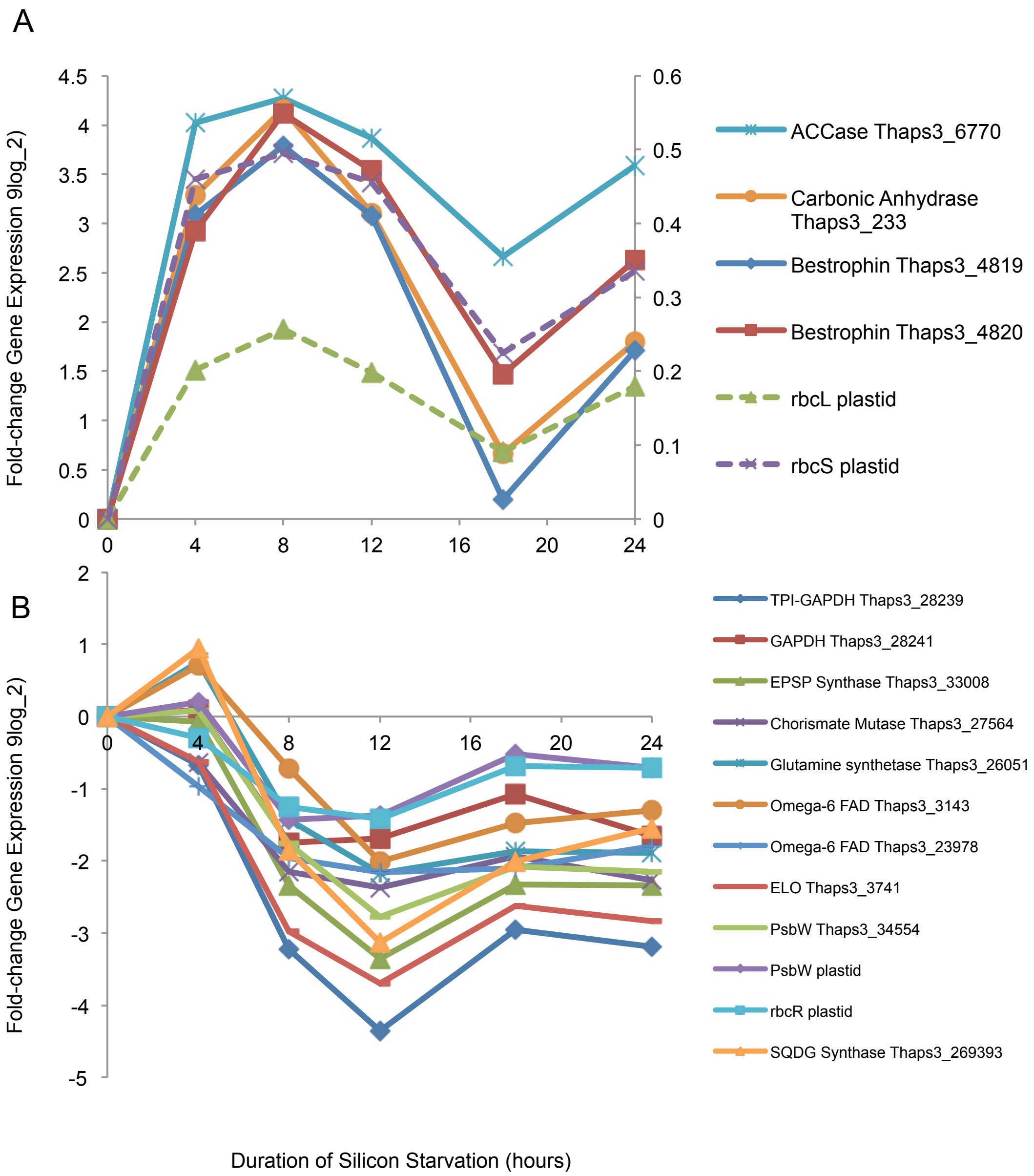


Figure 18. Correlation of transcript level changes of genes previously not considered linked by pathway or function. Transcript level changes of A) ACCase, carbonic anhydrase, bestrophin, and RuBisCO subunits, and B) several different proteins involved in processes ranging from photosynthesis and thylakoid lipid biosynthesis to cellular respiration and aromatic amino acid biosynthesis.

bestrophin (pfam01062) domain. In humans, bestrophins are believed to be calcium-activated ion channels with a role in intracellular calcium signaling and have been shown to be permeable to a variety of substrates including bicarbonate, glutamate, and GABA. In diatoms, these proteins are clearly targeted to the chloroplast, though their specific role remains unclear. If diatom bestrophins function as either bicarbonate transporters to supply RuBisCO or ACCase, or as glutamate or amino acid transporters to supply carbon skeletons or amino groups to or from the chloroplast, they may be important mediators of intracellular carbon flux.

Two clusters that are similarly down regulated (C19, C30) include several key enzymes involved in a variety of cellular processes including growth, carbon metabolism, and photosynthesis (Figure 16). Both clusters contain a high proportion of genes in the translation, ribosomal structure, and biogenesis KOG class. Similarly, many enzymes involved in nitrogen and amino acid metabolism are down regulated such as key enzymes in the shikimate pathway for aromatic amino acid biosynthesis, EPSP synthase and chorismate mutase (Thaps3_33008, Thaps3_27564 respectively) as well as glutamine synthetase (Thaps3_26051, Figure 16).

The observed down regulation of transcripts for enzymes involved in protein synthesis and nitrogen metabolism likely reflects the arrest in growth that results from silicon starvation. Interestingly, transcript levels of several key enzymes from carbon metabolism pathways are correlated with these changes (Figure 18B). For example, down regulated expression clusters C19 and C30 contain the mitochondrial isoforms of glyceraldehyde 3-P dehydrogenase (GAPDH, Thaps3_28241) and a triosephosphate isomerase-GAPDH fusion protein (TPI-GAPDH, Thaps3_28239). Due to their mitochondrial localization, these isoforms of key glycolysis enzymes are thought to be disproportionately important in the supply of carbon to the TCA cycle for cellular respiration thereby providing both energy and molecular precursors necessary for growth and division (Smith et al. 2012). Several genes with key roles in the synthesis of complex lipids were also found in C19 and C30 including an omega-3 fatty acid desaturase

(Thaps3_3143), an omega-6 fatty acid desaturase (Thaps3_23798), and a fatty acid elongase (Thaps3_3741) previously presented (See 2.4.1, Figure 11). Proportionately, complex lipids are less abundant during growth arrest (Figure 10B). Considering the biological roles of all of these key enzymes, it seems likely that regulation of these genes occurs at the transcript level to coordinately integrate several processes known to be involved in cellular proliferation.

Finally, while most enzymes involved in photosynthesis and related processes (light harvesting complex, Calvin-Benson cycle, thylakoid lipid biosynthesis) are up-regulated with a characteristic spike at 4 hr correlating with chloroplast replication (Figures 12, 17), there are a few enzymes that are in contrast, down regulated (Figure 18B). The plastid-encoded transcriptional regulator of the RuBisCO operon (rbcR) is also down regulated. Additionally, SQDG synthase (Thaps3_269393) along with the nuclear and plastid copies of psbW (Thaps3_34554, plastid_psbW), a protein thought to be involved in the contact and stability between the outer antenna complex and photosystem II (García-Cerdán et al. 2011) are down regulated. It is tempting to speculate that the coordinate regulation of both isoforms of psbW with SQDG synthase, which manufactures a lipid known to repress NPQ, alludes to a role of this gene in photosynthesis under balanced light and growth conditions.

2.5 DISCUSSION

2.5.1 Environmental control vs. cell cycle progression

Transcript-level changes drive a variety of cellular processes including growth, division, and adaptation to environmental change. Consequently, it is often challenging to interpret the biological significance of transcript level changes. For example, during the experimental conditions, *T. pseudonana* was starved for silicon, growth and cell cycle progression was arrested, and cells began to experience high light stress despite no manipulation to light intensities. During this time-course, the transcriptome could represent a dynamic response to any or all of these conditions.

The strong and transient expression maxima/minima observed at the 4 hr time point during silicon starvation was a dramatic feature in the global response of *T. pseudonana* to the experimental conditions. At the 4 hr time point, a portion of the *T. pseudonana* population is still progressing through the cell cycle, and chloroplast replication is occurring (Figure 17). This, along with the finding that genes with this expression pattern were functionally enriched for DNA replication, cytoskeletal processes, and cell cycle control, is strong evidence that other genes with a similar pattern are being regulated in a cell cycle-dependent manner. At the pathway level, transcripts of genes for Calvin-Benson cycle, fatty acid biosynthesis, and chlorophyll biosynthesis enzymes exhibit a 4 hr peak followed by a return to the levels of transcript seen at t=0. These changes are de-coupled from carbon fixation rates, fatty acid levels, and pigment levels, which is further support for transcript level regulation by cell cycle control (specifically plastid replication). Though the increased transcript levels may influence rates of carbon fixation, fatty acid biosynthesis, and pigment biosynthesis ultimately, this is likely an artifact of arrested cell division rather than a specific adaptive response to silicon starvation conditions.

In contrast to genes in which the transcript level changes are under cell cycle control, there are genes that appear to be responsive to other environmental signals, such as high light stress. During the silicon starvation time course, *T. pseudonana* experiences high light stress, as indicated by the disproportionate increase in photoprotective pigments relative to total pigments and the increase in the xanthophyll DES. When photosynthetic cells experience high light conditions, they photo adapt by decreasing the size of the antenna complex so as to minimize the flux of damaging photons to PSII (Falkowski and Raven 2007). Surprisingly, *T. pseudonana* increases the size of the outer antenna complex while simultaneously increasing the dissipation of light energy through NPQ. We propose that this represents a novel photoprotective strategy by which thylakoid membranes are remodeled such that cells become self-shaded (perhaps to protect DNA from photo damage) and that light energy captured by this larger antenna is funneled to regions of enhanced xanthophyll pigment pools (Nymark et al. 2009). The up regulated outer light harvesting antenna proteins (fucoxanthin chlorophyll binding proteins) therefore represent a class of genes in which the transcript levels are responsive to environmental signals in this experiment rather than being explicitly under cell cycle control. As cells are known to adapt to light conditions within minutes, it would make sense for these genes to be de-coupled from cell cycle progression which is likely to occur on longer time scales.

2.5.2 The significance of highly coordinated gene expression clusters

Several clusters of genes with correlated changes in transcript level were identified in the experimental transcriptomes. With the increasing availability of time course resolved transcriptomes, it is becoming more apparent that many diverse cell types follow highly choreographed transcriptional programs by which large suites of genes are coordinated, often on a circadian cycle (i.e. Zinser et al. 2009, Singh et al. 2010, Dhyrman et al. 2012, Kanesaki et al. 2012, Ashworth et al. 2012). This is advantageous for the cell since whole suites of genes participating in similar or complementary metabolic pathways can be coordinated, presumably by transcription factors to maintain a balance between energetic and carbon inputs as well as outputs in order to sustain growth and survival.

During balanced growth conditions (i.e. t=0 in these experiments), the energy and carbon inputs from light harvesting and carbon fixation are partitioned to a variety of outputs including division, respiration, and to carbohydrates and lipids. Experimentally removing silicon from the growth medium eliminates division as a carbon or energy output, thereby forcing *T. pseudonana* to reorganize the distribution of carbon (and energy in the form of reduced carbon molecules). Physiologically, this is evident as an increase in cellular lipid levels. Along with growth arrest and lipid induction, there are coordinated changes in transcript abundance of several enzymes that are key steps in a variety of different cellular energy and carbon acquisition and partitioning processes. This indicates that there is a transcriptional component to this reorganization.

2.5.3 Metabolic engineering to improve algal lipid productivity for biofuels

Metabolic engineering to improve lipid productivity is considered to be an important part of reducing the cost and improving feasibility of production of renewable biofuels from algae. Early attempts to engineer lipid metabolism in algae such as the overexpression of ACCase were unsuccessful in improving lipid yields, likely since metabolite flux into the pathway was the bottleneck and not enzyme activity (Dunahay et al. 1996). Selecting single targets to engineer metabolism is difficult without *a priori* knowledge of which steps may limit metabolic flux through that pathway (Broun 2004). However, there has been success in engineering algae to produce enhanced lipid without compromising growth rates at the single gene level (Trentacoste et al. 2013)

The findings presented here illustrate how transcript levels of several enzymes that integrate photosynthesis, fatty acid biosynthesis, respiration, and growth in silicon-starved *T. pseudonana* are correlated, suggesting that these genes may be under the control of the same or similar transcription factors. In terrestrial plants, manipulating transcription factors can be more successful as an approach to engineer metabolism as transcription factors can target multiple suites of genes in a given pathway (Broun 2004). Clustering of expression patterns of genes in different pathways and cellular processes in *T. pseudonana* suggests that coordinated global metabolic changes could result from manipulation of the appropriate regulatory factors.

2.6 ACKNOWLEDGEMENTS

Special thanks to several collaborators without whom the research described in this chapter would not be possible. First, I would like to thank Raffaela M. Abbriano and Jesse C. Traller (Scripps Institution of Oceanography, SIO) for invaluable assistance conducting experiments, processing samples, analyzing data, and for many helpful conversations. Dr. Corine Glé (SIO) assisted with experiments and did the pigment, lipid, and photosynthetic measurements. These analyses were done in the laboratory of Dr. Maria Vernet (SIO). A special thanks to both Dr. Glé and Dr. Vernet for their tangible contributions to this project, in addition to valuable discussions and assistance with data interpretation. Thank you to SIO colleagues including Dr. Orna Cook, Dr. Aubrey Davis, Dr. Emily Trentacoste, Dr. Roshan Shrestha, and Dr. Kalpana Shrestha for help with experiments, data analysis, and discussions, which greatly improved the quality of this study. Thank you to several collaborators who provided technical assistance including Steve Federowicz (UCSD) and Trina Norden-Krichmar (SIO and J. Craig Venter Institute, JCVI) for processing the microarray. Thank you to Dr. Eric Allen (SIO) for downstream bioinformatics assistance for the microarray data. Thank you to Sean Gallaher (Merchant Lab, UCLA), who prepared the libraries and sequenced mRNA for RNA-Seq, and to John McCrow (JCVI) for processing the raw reads. I am particularly grateful to Dr. Andrew Allen (JCVI) for providing the microarray and helpful guidance throughout the analysis. Thank you to Dr. Mark Hildebrand (SIO) for developing the experimental framework and establishing the collaboration with Dr. Allen (JCVI). Dr. Hildebrand's patience, enthusiasm, and deep knowledge of gene expression and diatom cell biology guided this project and greatly enhanced the quality of the analysis. Diatom biofuels research in the Hildebrand lab is supported by AFOSR grant FA9550-08-1-0178, DOE grant DE-EE0001222, and NSF grant CBET-0903712. Funding for Sarah R. Smith was provided by a National Defense Science and Engineering Graduate Research Fellowship.

2.7 REFERENCES

Allen, A.E., Laroche, J., Maheswari, U., Lommer, M., Schauer, N., Lopez, P.L., Finazzi, G., Fernie, A.R., and Bowler, C. (2008) Whole-cell response of the pennate diatom *Phaeodactylum tricornutum* to iron starvation. Proc. Natl. Acad. Sci. U. S. A. **105:** 10438-10443.

Armbrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H., Zhou, S., Allen, A.E., Apt, K.E., Bechner, M., Brzezinski, M.A., Chaal, B.K., Chiovitti, A., Davis, A.K., Demarest, M.S., Detter, J.C., Glavina, T., Goodstein, D., Hadi, M.Z., Hellsten, U., Hildebrand, M., Jenkins, B.D., Jurka, J., Kapitonov, V.V., Kröger, N., Lau, W.W.Y., Lane, T.W., Larimer, F.W, Lippmeier, J.C., Lucas, S., Medina, M., Monsant, A., Obornik, M., Schnitzler Parker, M., Palenik, B., Pazour, G.J., Richardson, P.M., Rynearson, T.A., Saito, M.A., Schwartz, D.C, Thamatrakoln, K., Valentin, K., Vardi, A., Wilkerson, F.P., Rokhsar, and Rokhsar, D.S. (2004) The genome of the diatom *Thalassiosira pseudonana*: Ecology, evolution, and metabolism. Science. **306:** 79-86.

Arrigo, K.R., Robinson, D.H., Worthen, D.L., Dunbar, R.B., DiTullio, G.R., VanWoert, M., and Lizotte, M.P. (1999) Phytoplankton community structure and the drawdown of nutrients and CO_2 in the Southern Ocean. Science. **283:** 365-367.

Ashworth, J., Coesel, S., Lee, A., Armbrust, E.V., Orellana, M.V., and Baliga, N.S. (2013) Genome-wide diel growth state transitions in the diatom *Thalassiosira pseudonana*. Proc. Natl. Acad. Sci. U. S. A. **110:** 7518-7523.

Bowler, C., Allen, A.E., Badger, J.H., Grimwood, J., Jabbari, K., Kuo, A., Maheswari, U., Martens, C., Maumus, F., Otillar, R.P., Rayko, E., Salamov, A., Vandepoele, K., Beszteri, B., Gruber, A., Heijde, M., Katinka, M., Mock, T., Valentin, K., Verret, F., Berges, J.A., Brownlee, C., Cadoret, J., Chiovitti, A., Choi, C.J., Coesel, S., De Martino, A., Detter, J.C., Durkin, C., Falciatore, A., Fournet, J., Haruta, M., Huysman, M.J.J., Jenkins, B.D., Jiroutova, K., Jorgensen, R.E., Joubert, Y., Kaplan, A., Kröger, N., Kroth, P.G., La Roche, J., Lindquist, E., Lommer, M., Martin-Jézéquel, V., Lopez, P.J., Lucas, S., Mangogna, M., McGinnis, K., Medlin, L.K., Montsant, A., Oudot-Le Secq, M., Napoli, C., Obornik, M., Parker, M.S., Petit, J., Porcel, B.M., Poulsen, N., Robison, M., Rychlewski, L., Rynearson, T.A., Schmutz, J., Shapiro, H., Siaut, M., Stanley, M., Sussman, M.R., Taylor, A.R., Vardi, A., von Dassow, P., Vyverman, W., Willis, A., Wyrwicz, L.S., Rokhsar, D.S., Weissenbach, J., Armbrust, E.V., Green, B.R., Van de Peer, Y., and I.V. Grigoriev. (2008) The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. Nature. **456:** 239-244.

Bailleul, B., Rogato, A., de Martino, A., Coesel, S., Cardol, P., Bowler, C., Falciatore, A, and Finazzi, G. (2010) An atypical member of the light-harvesting complex stress-related protein family modulates diatom responses to light. Proc. Nat. Acad. Sci. U. S. A. **107:** 18214-18214.

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

Broun, P. (2004) Transcription factors as tools for metabolic engineering in plants. Curr. Opin. Plant Biol. **7:** 202-209.

Costa, B.S., Sachse, M., Jungandreas, A., Bartulos, C.R., Gruber, A., Jakob, T., Kroth, P.G., and Wilhelm, C. (2013) Aureochrom 1a is involved in the photoacclimation of the diatom *Phaeodactylum tricornutum.* PLOS ONE. **8:** e74451.

Darley, W.M., and Volcani, B.E. (1969) Role of silicon in diatom metabolism. A silicon requirement for deoxyribonucleic acid synthesis in *Cylindrotheca fusiformis.* Exp. Cell Res. **58:** 334-342.

Davis, R., Aden, A., and Pienkos, P.T. (2011) Techno-economic analysis of autotrophic microalgae for fuel production. Appl. Energy. **88:** 3524-3531.

Dodds, E.D., McCoy, M.R., Rea, L.D., and Kennish, J.M. (2005) Gas chromatographic quantification of fatty acid methyl esters: Flame ionization detection vs. electron impact mass spectrometry. Lipids. **4:** 419-428.

Dunahay, T.G., Jarvis, E.E., Dais, S.S., and P.G. Roessler. (1996) Manipulation of microalgal lipid production using genetic engineering. Appl. Biochem. Biotechnol. **57-58:** 223-231.

Dyhrman, S.T., Jenkins, B.D., Rynearson, T.A., Saito, M.A., Mercier, M.L., Alexander, H., Whitney, L.P., Drzewianowski, A., Bulygin, V., Bertrand, E.M., Wu, Z., Benitez-Nelson, C., and Heithoff, A. (2012) The transcriptome and proteome of the diatom *Thalassiosira pseudonana* reveal a diverse phosphorus stress response. PLOS ONE. **7:** ee33768.

Falkowski, P.G., and Raven, J.A. (2007) Photosynthesis in continuous light. In *Aquatic Photosynthesis*. (Falkowski, P.G., and Raven, J.A., eds). New Jersey: Princeton University Press, pp. 237-277.

Folch, J., Lees, M., and Sloane Stanley, G.H. (1957) A simple method for the isolation and purification of total lipides from animal tissues. J. Biol. Chem. **226:** 497-509.

García-Cerdán, J.G., Kovács, L., Tóth, T., Kereïche, S., Aseeva,E., Boekema, E.J., Mamedov, F., Funk, C. and Schröder, W.P. (2011) The PsbW protein stabilizes the supramolecular organization of photosystem II in higher plants. Plant J. **65:** 368-381.

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., and Brown, P.P. (2000) Genomic ezpression programs in the response of yeast cells to environmental changes. Mol. Biol. Cell. **11:** 4241-4257,

Gillard, J., Devos, V., Huysman, M.J.J., De Veylder, L., D'Hondt, S., Martens, C., Vanormelingen, P., Vannerum, K., Sabbe, K., Chepurnov, V.A., Inzé, D., Vuylsteke, M., and Vyverman, W. (2008) Physiological and transcriptomic evidence for a close coupling between chloroplast ontogeny and cell cycle progression in the pennate diatom *Seminavis robusta*. Plant Physiol. **148:** 1394-1411.

Hildebrand, M., and Dahlin, K. (2000) Nitrate transporter genes from the diatom *Cylindrotheca fusiformis* (Bacillariophyceae): mRNA levels controlled by nitrogen source and by the cell cycle. J. Phycol. **36:** 702-713.

Hildebrand, M., Dahlin, K., and Volcani, B.E. (1998) Characterization of a silicon transporter gene family in *Cylindrotheca fusiformis:* Sequences, expression analysis, and identification of homologs in other diatoms. Mol. Gen. Genet. **260:** 480-486.

Hildebrand, M., Davis, A.K., Smith, S.R., Traller, J.C., and Abbriano, R.M. (2012) The place of diatoms in the biofuels industry. Biofuels. **3:** 221-240.

Hockin, N.L., Mock, T., Mulholland, F., Kopriva, S., and Malin, G. (2012) The response of diatom central carbon metabolism to nitrogen starvation is different from that of green algae and higher plants. Plant Physiol. **158:** 299-312.

Hu, Q., Sommerfeld, M., Jarvis, E., Ghirardi, M, Posewitz, M., Seibert, M., and Darzins, Al. (2008) Microalgal triacylglycerols as feedstocks for biofuel production: Perspectives and advances. Plant J. **54:** 621-639.

Huysman, M.J., Fortunato, A.E., Matthijs, M., Costa, B.S., Vanderhaeghen, R., Van den Daele, H., Sachse, M., Inzé, D., Bowler, C., Kroth, P.G., Wilhelm, C., Falciatore, A., Vyverman, W., and De Veylder, L. (2013) AUREOCHROME1a-mediated induction of the diatom-specific cyclin dsCYC2 controls the onset of cell division in diatoms (*Phaeodactylum tricornutum*). Plant Cell. **25:** 215-228.

Janero, D.R., and R. Barrnett. Isolation and characterization of an ether-linked homoserine lipid from the thylakoid membrane of *Chlamydomonas reinhardtii* 137+. J. Lipid Res. **23:** 307-316.

Kanesaki, Y., Imamura, S., Minoda, A., and Tanaka, K. (2012) External light conditions and internal cell cycle phases coordinate accumulation of chloroplast and mitochondrial transcripts in the red alga *Cyanidioscyzon merolae.* DNA Res. **19:** 289-303.

Lepetit, B., Goss, R., Jakob, T., and Wilhelm, C. (2012) Molecular dynamics of the diatom thylakoid membrane under different light conditions. Photosynth. Res. **1-2:** 245-257.

Levitan, O., Dinamarca, J., Hochman, G., and Falkowski, P.G. (2014) Diatoms: A fossil fuel of the future. Trends Biotechnol. **32:** 117-124.

Lewis, M.R., and J.C. Smith. (1983) A small volume, short-incubation-time method for measurement of photosynthesis as a function of incident irradiance. Mar. Ecol. Prog. Ser. **13:** 99-102.

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

Mock, T., Samanta, M.P., Iverson, V., Berthiaume, C., Robison, M., Holtermann, K., Durkin, C., Bondurant, S.S., Richmond, K., Rodesch, M., Kallas, T., Huttlin, E.L., Cerrina, F., Sussman, M.R., and Armbrust, E.V. (2008) Whole-genome expression profiling of the marine diatom *Thalassiosira pseudonana* identifies genes involved in silicon bioprocesses. Proc. Natl. Acad. Sci. U. S. A. **105:** 1579-1584.

Nymark, M., Valle, K.C., Brembu, T., Hancke, K., Winge, P., Andresen, K., Johnsen, G., and Bones, A.M. (2009) An integrated analysis of molecular acclimation to high light in the marine diatom *Phaeodactylum tricornutum.* PLOS ONE. **4:** e7743.

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.

Parker, M.S., Armbrust, E.V., Piovia-Scott, J., and Keil, R.G. (2004) Induction of photorespiration by light in the centric diatom *Thalassiosira weissflogii* (Bacillariophyceae): Molecular characterization and physiological consequences. J. Phycol. **40:** 557-567.

Platt, T., Denman, K.L., and Jassby, A.D. (1975) The mathematical representation and prediction of phytoplankton productivity. Fish. Mar. Serv. Tech. Rep. **523** 110p.

Plaxton, W.C. (1996) The organization and regulation of plant glycolysis. Annu. Rev. Plant Physiol. Plant Mol. Biol. **47:** 185-214.

Poulsen, N., and Kröger, N. (2005) A new molecular tool for transgenic diatoms. FEBS J. **272:** 3413-3423.

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

Radakovits, R., Jinkerson, R.E., Darzins, A., and Posewitz, M.C. (2010) Genetic engineering of algae for enhanced biofuel production. EC. **9:** 486-501.

Roessler, P.G. (1988) Effects of silicon deficiency on lipid composition and metabolism in the diatom *Cyclotella cryptica.* J. Phycol. **24:** 394-400.

Shrestha, R.P., Tesson, B., Norden-Krichmar, T., Federowicz, S., Hildebrand, M., and Allen, A.E. (2012) Whole transcriptome analysis of the silicon response of the diatom *Thalassiosira pseudonana*. BMC Genomics. **13:** 499.

Singh, A.K., Elvitigala, T., Cameron, J.C., Ghosh, J.K., Bhattacharyya-Pakrasi, M, and Pakrasi, H.B. (2010). Integrative analysis of large scale expression profiles reveals core transcriptional response and coordination between multiple cellular processes in a cyanobacterium. BMC Syst. Biol. **4:** 105.

Smith, S.R., Abbriano, R.M., and Hildebrand, M. (2012) Comparative analysis of diatom genomes reveals substantial differences in the organization of carbon partitioning pathways. Algal Res. **1:** 2-16.

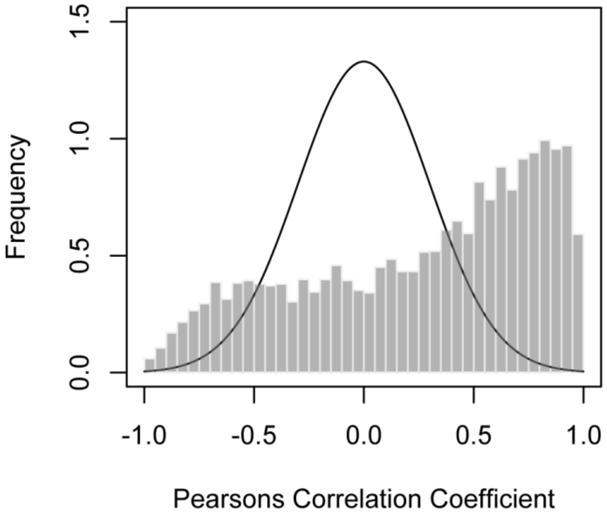
Trentacoste, E.M., Shrestha, R.P., Smith, S.R., Glé, C., Hartmann, A.C., Hildebrand, M., and Gerwick, W.H. (2013) Metabolic engineering of lipid catabolism increases microalgal lipid accmulation without compromosing growth. Proc. Natl. Acad. Sci. U. S. A. **110:** 19748-19753.

Tu, B.P., Kudlicki, A., Rowicka, M., and McKnight, S.L. (2005) Logic of the yeast metabolic cycle: Temporal compartmentalization of cellular processes. Science. **310:** 1152-1158.

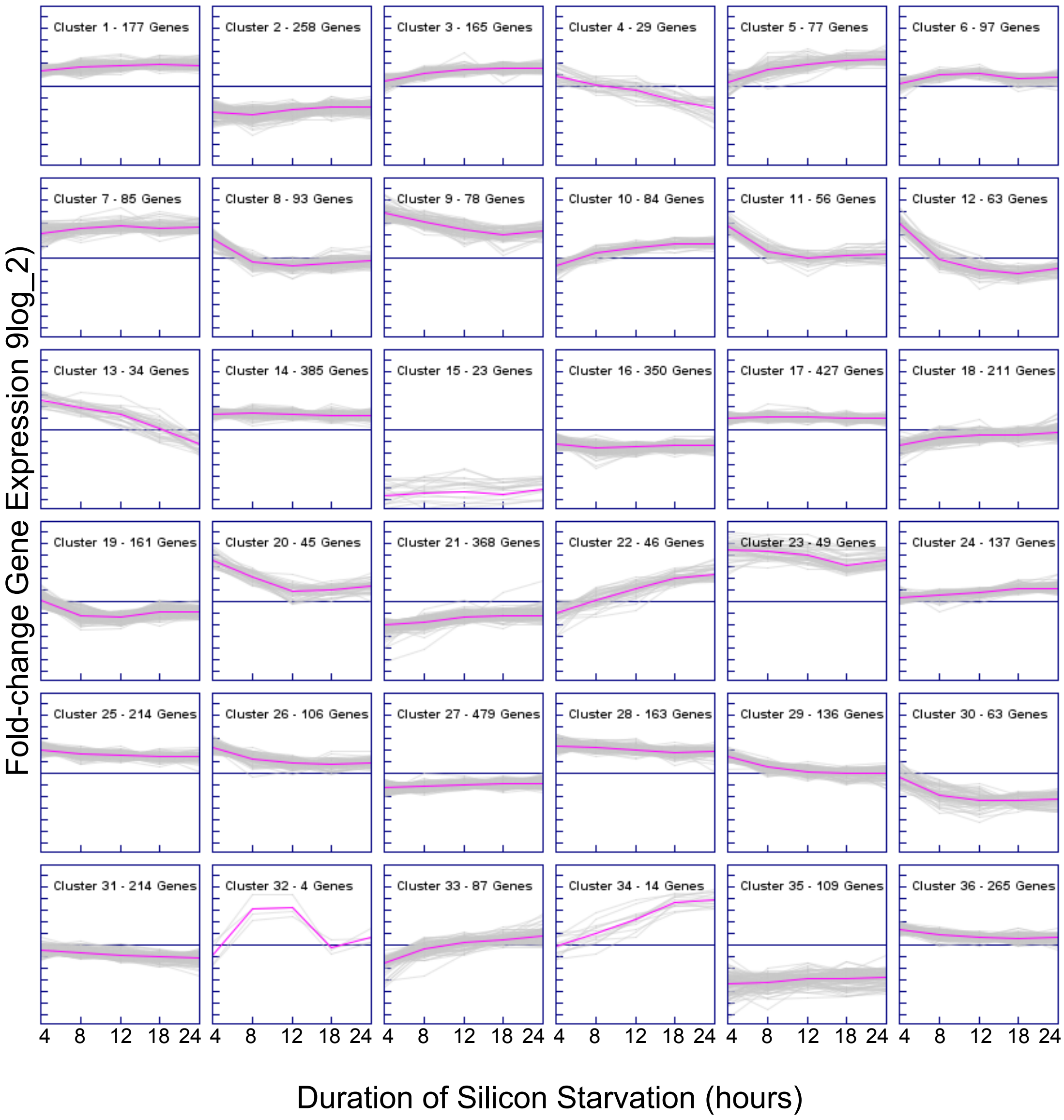
Yu, E.T., Zendejas, F.J., Lane, P.D., Gaucher, S., Simmons, B.A., and Lane, T.W. (2009) Triacylglycerol accumulation and profiling in the model diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* (Bacillariophyceae) during starvation. J. Appl. Phycol. **21:** 669-681.

Zapata, M., Rodríguez, F., and Garrido, J.L. (2000) Separation of chlorophylls and carotenoids from marine phytoplankton: A new HPLC method using a reversed phase C_8 column and pyridine-containing mobile phases. Mar. Ecol. Prog. Ser. **195:** 29-45.

Zinser, E.R., Lindell, D., Johnson, Z.I., Futschik, M.E., Steglich, C., Coleman, M.L., Wright, M.A., Rector, T., Steen, R., McNulty, N., Thompson, L.R., and Chisholm, S.W. (2009) Choreography of the transcriptome, photophysiology, and cell cycle of a minimal photoautotroph, *Prochlorococcus*. PLOS ONE. **4:** e5135.



Supplemental Figure 1. Agreement of the transcriptomic response in Si- #3 and Si- #9. Data represent the distribution of Pearson Correlation Coefficients between the microarray data and RNA-Seq data and show the number of genes that were represented in both data sets and were significantly differentially expressed in the microarray experiment**..** For PCC generation, the core signal fluorescence intensities from the microarray were correlated with RPKMs from the RNA-Seq analysis.



Supplemental Figure 2. Line plots of 36 expression clusters. Data represent microarray fold-change relative to t=0. Y-axis range is from (-7 to 7, fold change log_2).

Chapter 3

Towards characterizing the functional and evolutionary significance of inverted gene pairs in marine microbial eukaryotes

3.1 ABSTRACT

The distribution of genes in genomes of eukaryotic organisms can have profound implications for transcriptional regulation. Relatively recently, a new class of genes that are arranged in a “head to head” fashion on chromosomes were found to be co-expressed through the control of a bidirectional promoter in several model eukaryotes. This bidirectional arrangement is a unique mechanism by which a significant number of genes can be co-expressed, however the extent to which this mechanism is found across diverse eukaryotic phyla is unknown. With advances in high-throughput sequencing, analysis of the functional significance of gene order, and specifically bidirectional arrangements, can be done. For the first time, this work shows that bidirectional gene pairs (inverted gene pairs) are prevalent and are co-expressed in a diatom *Thalassiosira pseudonana*. Additionally, the relationship between genome size and prevalence of inverted gene pairs in diverse eukaryotes was characterized. Findings suggest that organisms with a small genome, and photosynthetic microbial eukaryotes with a red algal derived plastid may utilize this unique regulatory mechanism more than non photosynthetic microbial eukaryotes.

3.2 INTRODUCTION

Full genome sequencing has made studies of gene order or gene organization (the distribution of genes within chromosomes) possible. Though gene order in eukaryotes has generally been assumed to be random due to genome shuffling during evolution, whole-genome wide studies of gene order and gene expression have shifted this paradigm by showing that genes with coordinated patterns of expression are often clustered at a variety of scales (reviewed in Hurst et al. 2004, Michalak 2008). In the human genome, a major group of genes is organized in pairs with an inverted orientation (Figure 19) with transcription start sites separated by less than 1kb, compared with the average intergenic distance of 100kb (Adachi & Lieber 2002, Trinklein et

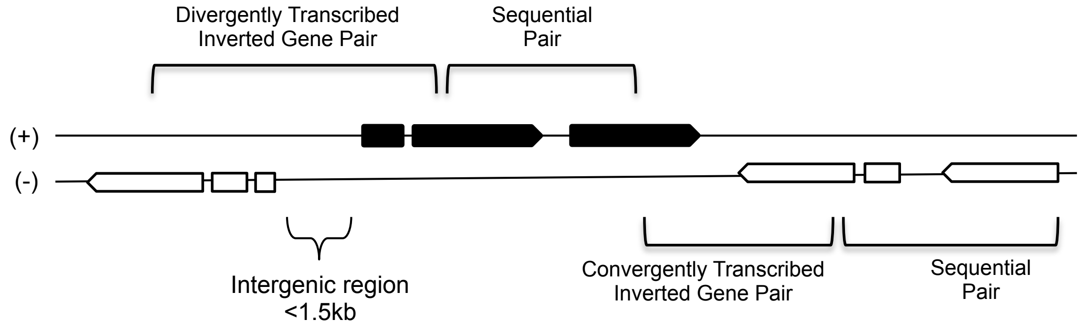


Figure 19. Schematic of gene arrangements on a stretch of genomic DNA. Genes are either coded on the sense (+) or antisense (-) strand. Divergently transcribed inverted gene pairs are putatively co-regulated through a bidirectional regulatory element in the intergenic region.

al. 2004). These divergently transcribed inverted gene pairs (referred to throughout as inverted gene pairs) comprise more than 10% of the human genome, and in many cases are coordinately regulated through the existence of bidirectional regulatory elements (promoter) located in the intergenic region (Trinklein et al. 2004). Similar inverted gene pairs have been documented in the *Arabidopsis thaliana* genome indicating that this bidirectional arrangement may be a common mechanism for coordinating regulation of a significant number of genes in eukaryotes (Wang et al. 2009).

The frequency and coordinate regulation of inverted gene pairs in genomes from a broad range of eukaryotic diversity have not been well investigated. Eukaryotic diversity is vast, and genomes in different taxa can vary greatly in gene content (number of genes) and size. For example, genome sizes are known to range over several orders of magnitude from 3 Mbp (parasitic *Encephalitozoon cuniculi*) to over 3000 Mbp in the case of dinoflagellates (Hou & Lin 2009). It is likely that these organisms employ a variety of different strategies to regulate expression of their extremely different genomes. One strategy for regulating expression from genomes (genome regulation) is to arrange genes in a manner that affects transcription. Arranging two genes under the control of a shared regulatory element may enable organisms with small genomes to streamline genome regulation. Understanding common features of genome organization that are important to regulate gene expression in eukaryotes versus those that are uniquely present in specific taxa will provide insight into factors that contribute to the specialization and evolution of genomes.

Initial observations of inverted gene pair frequency in diverse eukaryotes have shown that photosynthetic heterokonts such as the diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum,* and the oleaginous microalga *Nannochloropsis gaditana* (Eustigmatophyceae) are disproportionately enriched in inverted gene pairs relative to non-photosynthetic eukaryotes and photosynthetic eukaryotes from the green lineage (Jinkerson et al. 2013). This suggests that this arrangement could be particularly important in the genome regulation of heterokonts that possess red-algal derived plastids. However, these organisms all have relatively small genomes with high gene densities, and to determine if there is any kind of genome size or taxonomic bias in the frequency of inverted gene pairs, diverse genomes with a broader size range should be considered. If certain organisms rely more heavily on genome regulation through inverted gene pairs it would be of substantial evolutionary significance.

Currently, there is essentially nothing known regarding the functional or evolutionary significance of the abundance of inverted gene pairs in the genomes of marine microbial eukaryotes. There has been no work to demonstrate if inverted gene pairs in photosynthetic eukaryotes are co-expressed as they are in other eukaryotes. Furthermore, it is not known if the initial observation that the genomes of certain taxa (photosynthetic heterokonts) are disproportionately enriched with inverted gene pairs will be supported by incorporating additional genomes from a broader range of diversity and genome size. A fundamental understanding of specific mechanisms that underlie how diverse eukaryotes streamline transcriptional regulation would be useful both in understanding factors that have contributed to eukaryotic microbial diversification and practically for the interpretation of transcriptomic and metatranscriptomic data sets.

In this analysis, the functional and evolutionary significance of inverted gene pairs in microbial eukaryote genomes was investigated using a variety of approaches. First, the extent of co-expression of inverted gene pairs in the diatom *T. pseudonana* was determined with three different time course resolved transcriptomic data sets. Second, the functional roles of co-expressed inverted gene pairs in *T. pseudonana* were analyzed to gain insight into the biological significance of this mode of regulation. Finally, the frequency of inverted gene pairs from a variety of diverse eukaryotic genomes was analyzed to determine if this gene arrangement is generally enriched in photosynthetic heterokonts or photosynthetic eukaryotes.

3.3 METHODS

***3.3.1 Co-expression and functional roles of inverted gene pairs in* T. pseudonana**

All neighboring gene pairs (inverted gene pairs, convergently transcribed inverted gene pairs, and sequential gene pairs on both sense and antisense strands, Figure 19.) were identified by screening the *T. pseudonana* gene models v3.0 (genome.jgi-psf.org/Thaps3/Thaps3.home.html, Armbrust et al. 2004). Experimental transcriptomes, comprised of two different time course resolved silicon starvations (Si- #3, Si- #9, Smith, Ch. 2 this dissertation) along with a silicon re-addition experiment (Shrestha et al. 2012), were used to determine co-expression. The strength of correlation between each pair was determined by calculating the Pearson correlation coefficient (PCC) within each experiment and then the averaging these values across the experiments. Co-expressed genes are those that are defined by a PCC >0.5. The functional roles of genes that are co-expressed were analyzed and described. The density of inverted gene pairs in *T. pseudonana* was calculated by determining the percentage of gene pairs that are inverted out of 50 pairs in frames across a chromosome.

3.3.2 Frequency of inverted gene pairs in diverse eukaryotic genomes

Genomes of diverse eukaryotic microorganisms were computationally screened for frequency of inverted gene pairs, with a restricted intergenic distance of <1500 nucleotides. Most genomes were downloaded from the DOE JGI Genome Portal (genome.jgi.doe.gov). The genome of *Cyclotella cryptica,* a centric diatom, was obtained collaboratively (J. C. Traller, Scripps

Institution of Oceanography). The frequency of inverted gene pairs was plotted as a function of genome size.

3.4 RESULTS

3.4.1 Frequency and identity of co-expressed inverted gene pairs in T. pseudonana

Without restricting intergenic distance, the frequency of all neighboring gene pair types (inverted gene pairs, convergently transcribed inverted gene pairs, and sequential gene pairs, Figure 19) was quantified in *T. pseudonana.* Inverted gene pairs and convergently transcribed inverted gene pairs each comprise 33% of the genome whereas sequential gene pairs combined comprise 32%. This represents enrichment in inverted gene pairs in the *T. pseudonana* genome since the null expectation for neighboring gene pairs is 25% for both types of inverted pairs and 50% sequential gene pairs. When the intergenic distance of inverted gene pairs is restricted to 1.5kb, the frequency of inverted gene pairs decreases slightly to 26%. Since gene density is high in the small and compact *T. pseudonana* genome (Armbrust et al. 2004), the high proportion of inverted gene pairs alone does not indicate that inverted gene pairs are an important mode of regulation.

The Pearson Correlation Coefficient distributions show that there is a slight increase in the likelihood of co-expression for any neighboring gene pair (Figure 20). However, there is clearly an increase in the likelihood of co-expression for divergently transcribed inverted gene pairs relative to any other gene pair class (Figure 20A). There were 1560 genes (780 gene pairs) with a PCC > 0.5, which is 13.7% of the *T. pseudonana* genome. Most of the co-expressed genes found in this arrangement are hypothetical proteins, but several (n = 10) are annotated as core histones, which are co-expressed in humans via bidirectional promoters (Trinklein et al. 2004). Among the histones and hypothetical proteins, several co-expressed inverted genes in *T.*

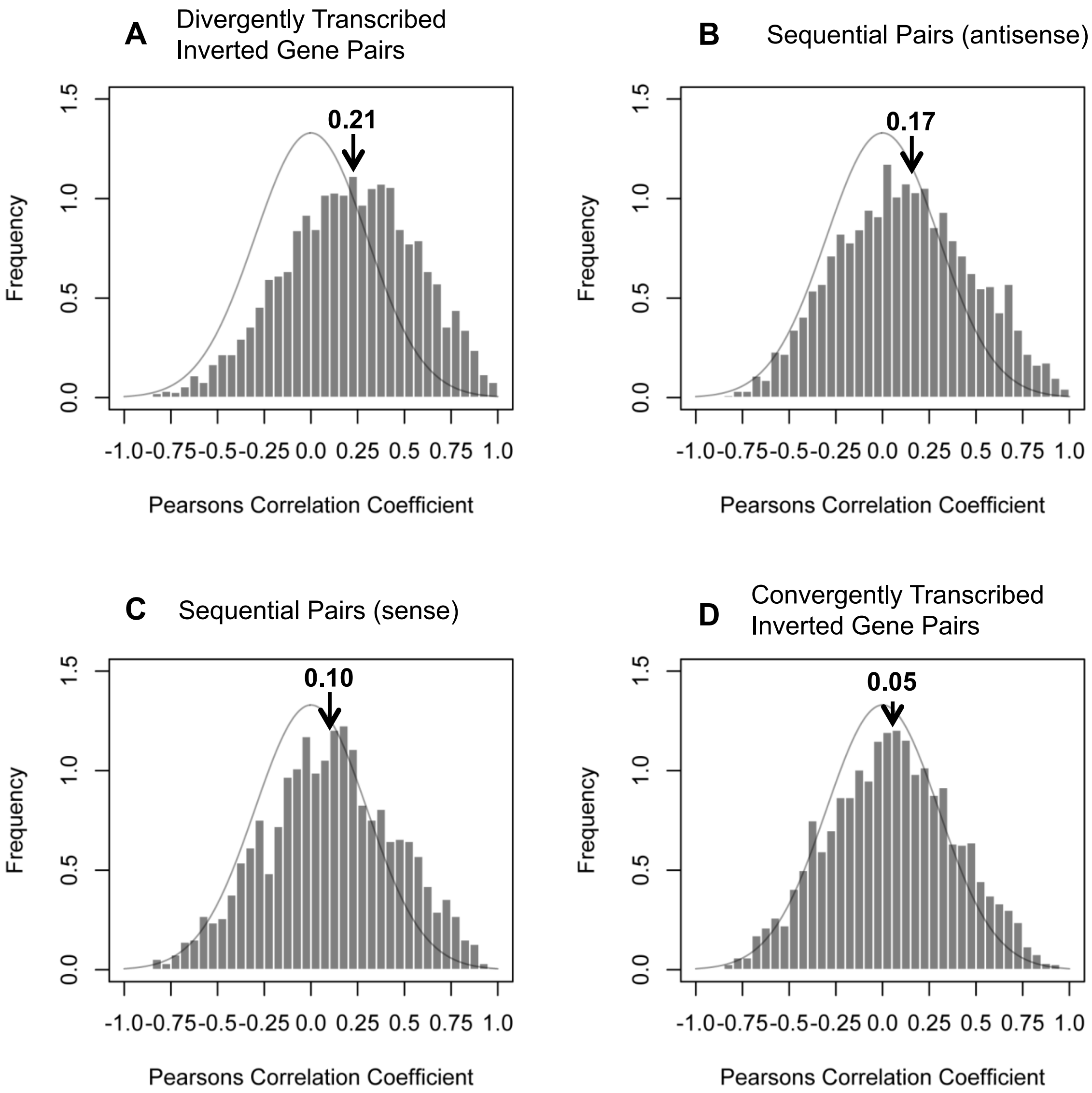


Figure 20. Distribution of average Pearson Correlation Coefficients for each type of neighboring gene pair. The curve on each plot show a normal distribution, arrows indicate the mean PCC for each distribution, and histogram data show distributions for A) divergently transcribed inverted gene pairs, B) sequential gene pairs on the antisense strand, C) sequential gene pairs on the sense strand, and D) convergently transcribed inverted gene pairs. Data represent the average PCC for each gene pair across three expression datasets.

pseudonana are important genes involved in light harvesting (fucoxanthin-chlorophyll a-c binding proteins), and carbon metabolism (glycolysis, fatty acid biosynthesis). In many cases the co-expression of inverted gene pairs is remarkably robust (Figure 21). Whether genes with these functional roles are also found as inverted gene pairs in other diatoms or photosynthetic eukaryotes remains unclear. One specific example in which microsynteny has been conserved for an inverted pair within diatoms is the arrangement of a mitochondrial glyceraldehyde phosphate dehydrogenase (GAPDH) and a triosephosphate isomerase-GAPDH fusion protein (TPI-GAPDH). This arrangement is conserved in all diatom genomes investigated to date, but not in the genome of the closely related photosynthetic heterokont *Aureococcus anophagefferens*, indicating this arrangement is diatom-specific (Smith et al. 2012).

3.4.2 The frequency of inverted gene pairs in diverse microbial eukaryotes

Eukaryotic genomes spanning a range of sizes were screened for the frequency of divergently transcribed inverted gene pairs (Table 3). Overall, there is a clear relationship between genome size and the frequency of inverted gene pairs in that small genomes are much more likely to have inverted gene pairs than small genomes (Figure 22). In addition, genomes of photosynthetic eukaryotes do seem to have more inverted gene pairs overall when compared to non-photosynthetic eukaryotes, however a limited number of non-photosynthetic heterokont genomes were included for comparison.

3.5 DISCUSSION

Arranging genes as divergently transcribed inverted gene pairs under the control of a bidirectional promoter is a well-recognized strategy to regulate co-expression in model organisms

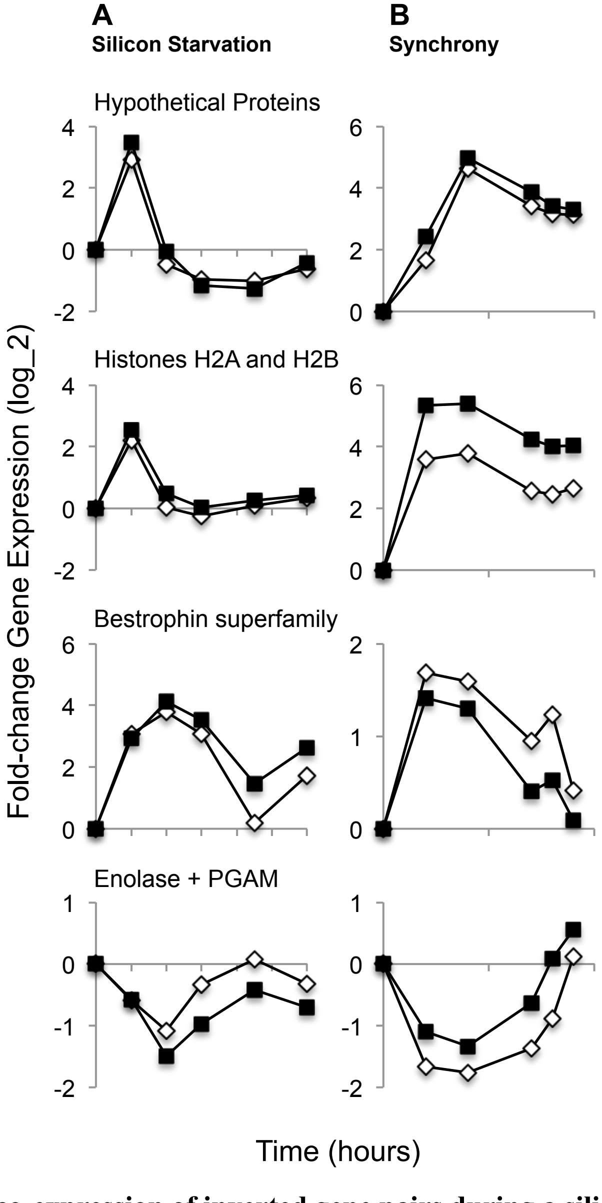


Figure 21. Robust co-expression of inverted gene pairs during a silicon starvation time course (24 hr) and a synchronized growth time course (9 hr). Data show transcript level changes (log_2) normalized to the initial time point (t=0).

Table 3. Eukaryotic genomes screened for the frequency of inverted gene pairs.

			Number of Models					Expressed as a Percentage of Total Protein Models				
	Size (Mbp)	**Total Protein Models**	**(-/+)**	**(-/-)**	**(+/+)**	**(+/-)**	**(-/+) <1500**	**(-/+)**	**(-/-)**	**(+/+)**	**(+/-)**	**(-/+) <1500**
Guillardia theta	87.16	24840	8148	3987	4053	8041	2178	33%	16%	16%	32%	9%
Emiliania huxleyi	167.7	33330	8462	6710	7004	8364	1347	25%	20%	21%	25%	4%
Fragilariopsis cylindrus	80.5	27137	8624	4837	4787	8623	2435	32%	18%	18%	32%	9%
Thalassiosira pseudonana	31	11390	3724	1927	1988	3724	2953	33%	17%	17%	33%	26%
Phaeodactylum tricornutum	26.1	10025	3202	1814	1779	3197	2550	32%	18%	18%	32%	25%
Pseudo-nitschia multiseries	218.73	19703	4982	3157	3553	5266	1103	25%	16%	18%	27%	6%
Aureococcus anophagefferens	56.7	11501	3111	2151	2174	3035	1820	27%	19%	19%	26%	16%
Cyclotella cryptica	180	33678	4630	-	-	-	1931	14%	-	-	-	6%
Schizochytrium aggregatum	40.85	10612	3080	2199	2074	3088	1540	29%	21%	20%	29%	15%
Aplanochytrium kerguelense	35.77	11892	3294	2552	2657	3277	1701	28%	21%	22%	28%	14%
Aurantiochytrium limacinum	60.93	14859	4338	3079	2950	4351	1191	29%	21%	20%	29%	8%
Phytophthora sojae	82.6	26584	6695	6555	6553	6700	1956	25%	25%	25%	25%	7%
Naegleria gruberi	41	15753	3842	3631	3824	3859	2552	24%	23%	24%	24%	16%
Bigelowiella natans	94.7	21708	7767	2955	3005	7721	3489	36%	14%	14%	36%	16%
Micromonas pusilla	21.95	10672	3568	1776	1746	3561	3279	33%	17%	16%	33%	31%
Ostreococcus tauri	12.56	7725	2391	1356	1488	2386	2283	31%	18%	19%	31%	30%
Ostreococcus lucimarinus	13.2	53273	10000	16138	16971	9932	3095	19%	30%	32%	19%	6%

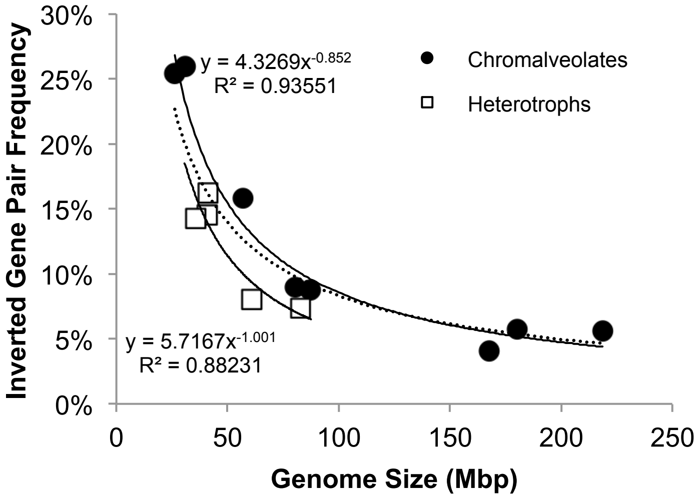


Figure 22. The relationship between genome size (Mbp) and frequency of inverted gene pairs for several chromalveolates (n = 8) and heterotrophic eukaryotes (n = 5). Overall large genomes have fewer inverted gene pairs than small genomes (dashed line shows power fit). Chromalveolates have a higher proportion of inverted gene pairs than heterotrophic eukaryotes .

like humans, mice, and *Arabidopsis,* however little work has been done to demonstrate if this arrangement is common in diverse eukaryotes. Using a comparative genomics approach in 64 species representing a wide range of eukaryotic phyla, Dávila López et al. (2010) identified several deeply evolutionarily conserved inverted gene pairs (microsynteny), which strongly suggests there is a functional significance of gene order for these pairs that arose early in eukaryotic evolution. However, no work has been done to demonstrate co-expression of inverted gene pairs in diverse eukaryotes.

For the first time, we have characterized the significance of gene order for co-expression of neighboring gene pairs at the genome scale, and have demonstrated that there is a significant increase in the likelihood of co-expression when neighboring genes are arranged as inverted gene pairs relative to any other neighboring pair type in the diatom *T. pseudonana.* Generally, genes that are nearby one another are more likely to be co-expressed due to a variety of factors that can be independent of sharing a promoter. For example, gene pairs may be co-localized in a region of a chromosome subject to similar DNA methylation, or histone modifications (Hurst et al. 2004). This is seen in co-expression data in *T. pseudonana* for all neighboring gene pairs. Convergently transcribed gene pairs are the least likely pair of genes to be co-expressed. However, both sequential gene pairs on the sense and antisense strand are slightly more likely to be co-expressed. It is possible that these genes are polycistronic (two genes are contained within the same mRNA) which is generally assumed to be important only in prokaryotes, yet is known to occur in some “lower” eukaryotes such as *Caenorhabditis elegans* and may be more important in microbial eukaryotes than previously assumed (Huang et al. 2001). Alternatively, there are several examples of genes in the *T. pseudonana* genome in which a single gene model is erroneously predicted as two sequential open reading frames (personal observation). This could also explain an increase in the likelihood of co-expression of sequential gene pairs. This should be investigated by more careful examination of the accuracy of gene models for co-expressed sequential pairs.

Functionally, several different types of genes were found as inverted gene pairs and co-expressed in *T. pseudonana.* In addition to the types of genes conserved in non-photosynthetic eukaryotes (histones, ribosomal proteins, etc., Dávila López et al. 2010), several genes involved in photosynthetic processes and carbon partitioning (i.e. glycolysis) were arranged as inverted gene pairs. The extent to which this microsynteny is conserved in other diatoms or other chromalveolates is currently unknown. The diatom-specific conservation of mitochondrial GAPDH and TPI-GAPDH fusion proteins suggests that certain pairs of inverted genes confer adaptive advantage to certain groups but not others. Valuable insight into factors that affect microbial eukaryotic evolution would be gained by characterizing conserved vs. novel inverted gene pairs across eukaryotic diversity.

The relationship between genome size and frequency of inverted gene pairs is quite striking (Figure 22). There is not a linear relationship between predicted protein content and genome size; larger genomes contain proportionately more non-coding DNA (Hou & Lin 2009). It is possible that the decrease in inverted gene pairs seen in larger genomes may simply be a factor of distributing fewer genes across larger stretches of the chromosome. However, it is known that there can be variability in average gene content, average gene size, and gene distribution in different organisms and these factors should be accounted for when determining the relationship between genome size and frequency of inverted gene pairs. It is possible that regulation through inverted gene pairs may be a particular strategy for the efficient regulation of small genomes. Genome size varies over several orders of magnitude in diverse eukaryotes. Increasingly, it is becoming appreciated that non-coding regions of large genomes are not "junk DNA", rather that they serve a variety of biological functions, many with regulatory consequences (Gregory 2005). Though inverted gene pairs are important in large genomes, they may be particularly important in small genomes as a way to minimize the need for other regulatory strategies that require more non-coding DNA. Evolutionarily this could allow for genome size reduction at some potential cost of reducing regulatory network complexity. Additional genomes that cover a broader range in size and eukaryotic diversity should be incorporated to better address the evolutionary questions that these preliminary findings raise.

In humans, the bidirectional promoters that regulate co-expression of inverted gene pairs lack TATA boxes and are rather enriched in CpG islands (Trinklein et al. 2004). Though there is now convincing evidence that many inverted gene pairs are co-expressed in *T. pseudonana*, there has been no work to characterize the putative bidirectional regulatory elements. This will be important to conclusively show that inverted gene pair organization is functionally significant in *T. pseudonana.* Molecular characterization of the intergenic regions of co-expressed inverted gene pairs should be done to isolate the sequences thought to regulate transcription bidirectionally. These intergenic regions can also be characterized bioinformatically by searching for conserved motifs or palindromic sequences (i.e. through MEME, Bailey et al. 2009).

Our understanding of the frequency, function, biological role, and evolutionary significance of inverted gene pairs in diverse eukaryotes remains limited. However, there are many exciting research opportunities as high-throughput sequencing of genomes and transcriptomes combined with bioinformatics based analysis continue to improve. These questions are of fundamental interest in the studies of eukaryotic cellular function and evolution.

3.6 ACKNOWLEDGEMENTS

Funding for Sarah R. Smith was provided by a National Defense Science and Engineering Graduate Fellowship. Special thanks to Dr. Mark Hildebrand (Scripps Institution of Oceanography, SIO) for providing steadfast, valuable guidance, and constructive comments that greatly enhanced the quality of this work. Thanks to Dr. Andrew Allen (J. Craig Venter Institute) for encouragement, guidance, and enthusiasm for continuing this work. I would also like to acknowledge Jesse C. Traller (SIO) for providing unpublished data from the genome of *Cyclotella cryptica.*

3.7 REFERENCES

Adachi, N., and Lieber, M.R. (2002) Bidirectional gene organization: A common architectural feature of the human genome. Cell. **109:** 807-809.

Armbrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H., Zhou, S., Allen, A.E., Apt, K.E., Bechner, M., Brzezinski, M.A., Chaal, B.K., Chiovitti, A., Davis, A.K., Demarest, M.S., Detter, J.C., Glavina, T., Goodstein, D., Hadi, M.Z., Hellsten, U., Hildebrand, M., Jenkins, B.D., Jurka, J., Kapitonov, V.V., Kröger, N., Lau, W.W.Y., Lane, T.W., Larimer, F.W, Lippmeier, J.C., Lucas, S., Medina, M., Monsant, A., Obornik, M., Schnitzler Parker, M., Palenik, B., Pazour, G.J., Richardson, P.M., Rynearson, T.A., Saito, M.A., Schwartz, D.C, Thamatrakoln, K., Valentin, K., Vardi, A., Wilkerson, F.P., Rokhsar, and Rokhsar, D.S. (2004) The genome of the diatom *Thalassiosira pseudonana*: Ecology, evolution, and metabolism. Science. **306:** 79-86.

Bailey, T.L., Bodén, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009) MEME SUITE: Tools for motif discovery and searching. Nucleic Acids Res. **37:** W202-W208.

Dávila López, M, Martínez-Guerra, J.J., and Samuelsson, T. (2010) Analysis of gene order conservation in eukaryotes identifies transcriptionally and functionally linked genes. PLOS ONE. **5:** e10654.

Gregory, T.R. (2005) Synergy between sequence and size in large-scale genomics. Nat. Rev. Genet. **6:** 699-708.

Hou, Y., and Lin, S. (2009) Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: Gene content estimation for dinoflagellates. PLOS ONE. **9:** e6978.

Huang, T., Kuersten, S., Deshpande, A.M., Spieth, J., MacMorris, M., and Blumenthal, T. (2001) Intercistronic region required for polycistronic pre-mRNA processing in *Caenorhabditis elegans*. Mol. Cell. Biol. **21:** 1111-1120.

Hurst, L.D., Pál, and Lercher, M.J. (2004) The evolutionary dynamics of eukaryotic gene order. Nat. Rev. Genet. **5:** 299-310.

Jinkerson, R.E, Radakovits, R., and Posewitz, M.C. (2013) Genomic insights from the oleaginous model alga *Nannochloropsis gaditana.* Bioengineered. **4:** 37-43.

Michalak, P. (2008) Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomics. **91:** 243-248.

Shrestha, R.P., B. Tesson, T. Norden-Krishmar, S. Federowics, M. Hildebrand, and A.E. Allen. (2012) Whole transcriptome analysis of the silicon response of the diatom *Thalassiosira pseudonana.* BMC Genomics. **13:** 499

Smith, S.R., Abbriano, R.M., and Hildebrand, M. (2012) Comparative analysis of diatom genomes reveals substantial differenes in the organization of carbon partitioning pathways. Algal Res. **1:** 2-16.

Trinklein, N.D., Aldred, S.F., Hartman, S.J., Schroeder, D.I., Otillar, R.P., and Myers, R.M. (2004). An abundance of bidirectional promoters in the human genome. Genome Res. **14:** 62-66.

Wang, Q., Wan, L., Li, D., Zhu, L., Qian, M., and Deng, M. (2009) Searching for bidirectional promoters in *Arabidopsis thaliana.* BMC Bioinformatics **10 (Suppl 1):** S29