


5-2018

Phylogeny and Evolutionary Genomics of Non-Photosynthetic Diatoms

Anastasiia Onyshchenko
University of Arkansas, Fayetteville

Follow this and additional works at: <https://scholarworks.uark.edu/etd>

 Part of the [Bioinformatics Commons](#), [Evolution Commons](#), [Genomics Commons](#), and the [Molecular Genetics Commons](#)

Recommended Citation

Onyshchenko, Anastasiia, "Phylogeny and Evolutionary Genomics of Non-Photosynthetic Diatoms" (2018). *Theses and Dissertations*. 2695.
<https://scholarworks.uark.edu/etd/2695>

This Thesis is brought to you for free and open access by ScholarWorks@UARK. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of ScholarWorks@UARK. For more information, please contact scholar@uark.edu, ccmiddle@uark.edu.

Phylogeny and Evolutionary Genomics of Non-Photosynthetic Diatoms

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Cell and Molecular Biology

by

Anastasiia Onyshchenko
Taras Shevchenko National University of Kyiv
Bachelor of Science in Biology, 2016

May 2018
University of Arkansas

This thesis is approved for recommendation to the Graduate Council

Dr. Andrew Alverson
Thesis Director

Dr. Andy Pereira
Committee member

Dr. Jeffrey Lewis
Committee member

Abstract

Diatoms are prolific photosynthesizers responsible for some 20% of global primary production. In real terms, the oxygen in one of every five breaths traces back to photosynthesis by marine diatoms. Among the tens of thousands of diatom species, a small handful of colorless diatom species in the genus *Nitzschia* have lost photosynthesis altogether and rely exclusively on extracellular organic carbon for growth. I used DNA sequence data to reconstruct the phylogeny of this group, and found that nonphotosynthetic diatoms are monophyletic, indicating that photosynthesis was lost just one time over the course of some 200 million years of diatom evolution. Carbon metabolism in nonphotosynthetic diatoms, including the exact source of carbon used by these species, has not been fully characterized. We sequenced the nuclear genome of one species and used it to develop a comprehensive model of central carbon metabolism. Preliminary analysis of *Nitzschia* metabolism showed that it generally matches to the pattern of previously reported diatom metabolic networks. As well we found some hints regarding *Nitzschia* external carbon acquisition which possibly can help to explain its heterotrophic mode of life. Overall, this study has provided novel insights into the evolutionary origin and metabolism of non-photosynthetic diatoms, which are unique among diatoms in their ability to sustain their growth solely from extracellular carbon.

Acknowledgments

I would like to thank Fulbright Program for providing me opportunity to complete Master's Degree in Cell and Molecular Biology at University of Arkansas. I am grateful to my adviser, Andrew Alverson, for a chance to work on such a great research project and support and opportunities he provided to me during two years of my studies. I would like to thank to all the lab members, Teofil Nakov, Elizabeth Ruck, Kala Downey for excellent work environment, moral and professional support and for the research they are doing without which none of my work could happen. I'd like to thank my committee members Jeffrey Lewis and Andy Pereira for agreeing to advise and help me during completion of my degree. I am grateful to Douglas Rhoads (Head of CEMB Program) and David McNabb (Head of Biology Department) for letting me be part of their programs.

I would also like to thank to my mother Tetyana and my boyfriend Lance for unconditional emotional support whenever I needed it.

Table of Contents

| | |
|--|-----------|
| Introduction | 1 |
| The Diatoms | 1 |
| Loss of photosynthesis in diatoms and across the tree of life | 2 |
| Nuclear genomic insights into carbon metabolism in nonphotosynthetic diatoms | 4 |
| Rationale and significance | 5 |
| References | 7 |
| Chapter 1. A Single Loss of Photosynthesis In Diatoms | 10 |
| Abstract | 10 |
| Introduction | 11 |
| Materials and Methods | 14 |
| Collection and culturing of apochloritic <i>Nitzschia</i> | 14 |
| DNA extraction, PCR, and DNA sequencing | 14 |
| Phylogenetic analyses | 15 |
| Plastid genome sequencing and analysis | 17 |
| Results | 17 |
| Phylogeny of apochloritic <i>Nitzschia</i> | 17 |
| Plastid genome sequencing and analysis | 19 |
| Discussion | 21 |
| Monophyly of apochloritic diatoms | 21 |
| The ecology and biogeography of apochloritic diatoms | 23 |
| Plastid genome reduction in apochloritic diatoms | 24 |
| Conclusions | 25 |
| References | 27 |
| Chapter 2. Core carbon metabolism and characterization of a β-ketoacid pathway inferred from the genome of a non-photosynthetic diatom (Bacillariophyta) | 48 |
| Abstract | 48 |
| Introduction | 49 |
| Materials and Methods | 50 |
| Collection and culturing of <i>Nitzschia</i> sp. | 50 |
| DNA and RNA extraction and sequencing | 51 |
| Genome and transcriptome assembly and annotation | 51 |
| Genome annotation | 53 |
| Construction of orthologous clusters | 54 |
| Characterization of carbon metabolism genes | 55 |
| Prediction of protein localization | 56 |

| | |
|--|-----------|
| Results and Discussion | 56 |
| Genome characteristics | 56 |
| Central carbon metabolism | 57 |
| A β -ketoacid pathway in diatoms | 61 |
| Conclusions | 65 |
| References | 67 |
| Conclusions | 84 |

List of tables

| | |
|--|-----------|
| Chapter 1. A Single Loss of Photosynthesis In Diatoms | 10 |
| Table 1-1. Taxa and sources of DNA sequences analyzed in this study. | 32 |
| Table S1-1. Primers used to amplify and sequence cob and nad1 fragments for study taxa. | 37 |
| Table S1-2. Overlapping and adjacent genes in plastid genomes of <i>Nitzschia</i> sp. nitz4 and <i>Nitzschia</i> sp. NIES-3581. Groups of overlapping or adjacent genes are separated by empty rows with the coordinates showing the degree of overlap. All but two groups are shared between the two genomes. | 38 |
| Chapter 2. Core carbon metabolism and characterization of a β-ketoacid pathway inferred from the genome of a non-photosynthetic diatom (Bacillariophyta) | 48 |
| Table 2-1. <i>Nitzschia</i> sp. Nitz4 genome annotation progress statistics | 72 |
| Table 2-2. General genome annotation statistics for analyzed diatom species | 72 |
| Table 2-3. Diatom β -ketoacid pathway annotation for analyzed species | 73 |

List of figures

Chapter 1. A Single Loss of Photosynthesis In Diatoms 10

Figure 1-1. Light micrographs of newly sequenced nonphotosynthetic *Nitzschia* species. Clockwise from top: *Nitzschia* sp. nitz2, *Nitzschia* sp. nitz7, *Nitzschia* sp. nitz8, *Nitzschia* sp. nitz4. 33

Figure 1-2. Phylogenetic trees inferred from plastid 16S (A), plastid 16S transformed into purine/pyrimidine (R/Y) coding (B), nuclear 28S d1–d2 genes for all newly sequenced taxa and data from GenBank [”lsu (new + ncbi)”, C], mitochondrial cob (D), mitochondrial nad1 (E), concatenated cob and nad1 genes [”mito (cob + nad1)”, F], concatenated cob, nad1, and nuclear 28S d1–d2 genes for all newly sequenced taxa [”mito (cob + nad1) + lsu (new)”, G], and a concatenated alignment of cob, nad1, and nuclear 28S d1–d2 genes for taxa from this study and GenBank [”mito (cob + nad1) + lsu (new + ncbi)”, H]. For the phylogenies in panels C and H, we removed branches shorter than 0.00001 units for clarity. The full phylogenies are available in Supplementary Figure 1. Thicker black branches correspond to apochlorotic taxa. White points identify nodes with bootstrap support >70%. 34

Figure 1-3. Conserved synteny, gene content, and sequence in the plastid genomes of two nonphotosynthetic diatoms in the genus *Nitzschia*, Nitz4 (this study) and NIES-3581 (Kamikawa et al. 2015a). Unlabeled genes are shared between the two species. 36

Figure S1-1. Phylogenetic trees for plastid 16S (A), plastid 16S transformed into purine/pyrimidine (R/Y) coding (B), mitochondrial cob (C), nad1 (D), a densely sampled 28S d1–d2 matrix with sequences from this study and GenBank (E), cob and nad1 combined (F), and a large combined nuclear and mitochondrial gene matrix with sequences from this study and GenBank (G). In the 16S tree, *Nitzschia* sp. NIES-3581 is identified by its synonym, iriis04. 40

Chapter 2. Core carbon metabolism and characterization of a β -keto adipate pathway inferred from the genome of a non-photosynthetic diatom (Bacillariophyta) 48

Figure 2-1. Protocatechuate dioxygenase genes localization on *Nitzschia* sp. Nitz4 scaffolds 78

Figure 2-2. Schematic annotation of protocatechuate branch of β -keto adipate pathway in analyzed diatom species 79

Figure S2-1. Statistics for original genome scaffolding for range of K-mers 80

Figure S2-2. Statistics for organellar reads-free genome scaffolding for range of K-mers 81

Figure S2-3. Comprehensive mitochondrial metabolic pathways of *Nitzschia* sp. Nitz4 derived from genome annotation 82

Figure S2-4. Comprehensive plastid metabolic pathways of *Nitzschia* sp. Nitz4 derived from genome annotation 83

List of published papers

Chapter 1: Onyshchenko, A., Ruck, E.C., Nakov, T. & Alverson, A.J. 2018. A single loss of photosynthesis in diatoms. *bioRxiv*. doi: <http://dx.doi.org/10.1101/298810>

Introduction

The Diatoms

Diatoms (Bacillariophyta) are a group of widely distributed unicellular algae belonging to the chromalveolate clade, which includes photosynthetic lineages such as brown algae, dinoflagellates, haptophytes, and non-photosynthetic groups such ciliates and apicomplexans (Cavalier-Smith 1999). Diatoms are ancestrally photosynthetic, and they are responsible for roughly 20% of global primary production. They are also highly diverse, with as many as 200,000 different species (Mann and Droop 1996). Diatoms are also known for their elaborate and highly intricate silica cell walls (Round, Crawford, and Mann 1990) and are classified historically based on cell wall features and, more recently, their phylogenetic relationships (Sims, Mann, and Medlin 2006). For example, “centric” diatoms (Coscinodiscophyceae) are radially symmetrical, and pennate diatoms have bipolar symmetry. Raphe-bearing pennate diatoms (Bacillariophyceae) possess a slit for their cell wall that allows them to glide along a surface. Diatoms are ancestrally oogamous, and this mode of reproduction is found throughout the centric diatom lineages. The pennate diatoms are isogamous or anisogamous (Williams and Kociolek 2011). Raphid pennates constitute the vast majority of diatom diversity, which likely reflects a combination of life history and active motility (Nakov, Beaulieu, and Alverson 2018).

Diatom plastids surrounded by four membranes (in contrast to green algae and land plants, which have only two membranes), which reflects the origin of their plastids in which a photosynthetic red alga cell was engulfed by a heterotrophic eukaryote. Diatom cells are highly compartmentalized, including a periplastid compartment derived from the cytoplasm of the ancient red algal endosymbiont (Gruber et al. 2007). As a result, diatom genomes often

possess multiple isozymes related to central carbon pathways that have distinct cellular localizations. For example, several diatoms have complete or partial glycolytic pathways localized to all three cellular compartments (Smith, Abbriano, and Hildebrand 2012; Kroth et al. 2008). It is clear from the very small sample of sequenced diatom genomes that diatoms are genomically and physiologically diverse, so we expect that newly sequenced genomes are likely to reveal many new features and unsuspected peculiarities.

Loss of photosynthesis in diatoms and across the tree of life

Loss of photosynthesis is common in nearly all major groups of photosynthetic eukaryotes (Hadariová et al. 2018). For example, at least five different lineages within green the algal orders Chlamydomonadales and Chlorellales lost photosynthesis and are now obligate parasites (Těšitel 2016). Multiple lineages of flowering plants either completely or incompletely dispensed with photosynthesis in favor of parasitism as well (Barkman et al. 2007; Těšitel 2016). Photosynthesis has been lost dozens or more times in florideophycean red algae as well. In a phenomenon known as adelphoparasitism, these species go on to parasitize their sister species (Goff et al. 1996; Blouin and Lane 2012). In addition to lineage with primary green and red plastids, species with secondary plastids have repeatedly lost photosynthesis as well. These include euglenoids (Marin 2004), apicomplexans ((McFadden et al. 1996),(Waller and McFadden 2005)), ciliates and dinoflagellates (Taylor, Hoppenrath, and Saldarriaga 2008). In most cases, it is thought that secondarily nonphotosynthetic species evolved from mixotrophic ancestors, which in addition to photosynthesis, can also use extracellular carbon for growth.

Many pennate diatoms are mixotrophic (Hellebust and Lewin 1977), and although it is likely very costly to maintain these dual modes of nutrition (Raven 1997), it may help these

diatoms survive through periods of low sunlight when photosynthesis alone cannot sustain cell viability (Hellebust and Lewin 1977; Tuchman et al. 2006). Despite recurrent loss of photosynthesis in all major clades of photosynthetic organisms, loss of photosynthesis in diatoms has been rare. Only one subset of 20 or so mostly undescribed free-living species in the genus *Nitzschia* that have given up photosynthesis and are now obligate heterotrophs (Lewin and Lewin 1967; Kamikawa et al. 2015). These “apochloritic” diatoms are often found in environments that are rich in organic carbon such as mangrove forests and decaying seaweeds (Pringsheim 1956; Lewin and Lewin 1967; Kamikawa et al. 2015; Blackburn, Hannah, and Rogerson 2009a). Although few in number and limited in their taxonomic scope, these species nevertheless encompass a broad range of morphological diversity and are worldwide in distribution (Blackburn, Hannah, and Rogerson 2009b; Kamikawa et al. 2015), which together raises the possibility that loss of photosynthesis occurred more than once in this group—a hypothesis supported by phylogenetic analyses of one plastid and one nuclear gene (Kamikawa et al. 2015). These data could not reject monophyly of apochloritic species either (Kamikawa et al. 2015).

The first part of my thesis used combined DNA datasets from the nuclear, plastid, mitochondrial genomes, as well as expanded species sampling, to test the number of losses of photosynthesis in this group. I found that photosynthesis was lost just one time in the common ancestor of this one small subclade within the genus *Nitzschia* (Onyshchenko et al. 2018). Thus, this radical trophic shift from autotrophy to obligate heterotrophy represents a singular rare event in diatoms. For this study, we also sequenced and analyzed the plastid genome of one nonphotosynthetic diatom, *Nitzschia* sp. Nitz4, and compared it with a previously sequenced

nonphotosynthetic diatom plastid genome (Kamikawa et al. 2016). The genomes were highly similar in gene content, synteny and sequence—consistent with rapid genomic streamlining following the loss of photosynthesis in their common ancestor.

Nuclear genomic insights into carbon metabolism in nonphotosynthetic diatoms

Genomic and experimental studies alike have shown that central carbon metabolism in diatoms is highly complex and compartmentalized (Kroth et al. 2008; Smith, Abbriano, and Hildebrand 2012; Armbrust et al. 2004; Traller et al. 2016; Kamikawa et al. 2017). These studies have shown that relatively few generalizations about carbon metabolism can be made across diatom species, thereby cautioning against extrapolations from model to non-model species. This would include species that have lost photosynthesis and are likely to have unique carbon metabolic pathways. For the second part of my thesis, I analyzed the nuclear genome sequence of *Nitzschia* sp. Nitz4 to better understand carbon metabolism in nonphotosynthetic diatoms. I was specifically interested in identifying the mechanism of carbon carbon acquisition and forms of organic carbon used by these species.

I provide the first characterization of a β -keto adipate pathway in diatoms. This pathway is involved in degradation of aromatic components derived from decaying plant material (e.g., lignin) or from toxic environment pollutants (Harwood and Parales 1996). The β -keto adipate pathway was previously thought to be restricted exclusively to fungi and bacteria, which have slightly different versions of the pathway. I expanded my analysis to other sequenced diatom genomes and found that genes in the β -keto adipate pathway were present in all sequenced diatoms, though the pathway was incomplete in most cases. I used the genome to predict that diatoms transport and subsequently degrade protocatechuic acid, an aromatic compound derived

from lignin, and that the terminal products of the pathway are direct intermediates of the Krebs cycle. I also developed a model of central carbon metabolism in this species and compared it photosynthetic diatoms. It appears that *Nitzschia* sp. Nitz4 has retained the conserved compartmentalization pattern for carbon metabolism as photosynthetic species. Modifications include the ability to initiate gluconeogenesis exclusively within the mitochondrion and the absence of pyruvate metabolising enzymes found in other diatoms.

Rationale and significance

A clear understanding of the loss of photosynthesis in diatoms requires a robust phylogenetic hypothesis, and I showed that photosynthesis was lost once in this group. This allows us to better understand the evolutionary and ecological context of this event, and in addition, it guides strategies for future research in this group. For example, comparative approaches offer limited insights into events that occurred just one time.

Trophic shifts from autotrophy to heterotrophy can have profound and cascading genomic consequences, and these have been most intensively studied in the plastid genome. I focused most of my attention on the nuclear genome, using it to understand and model carbon metabolism in apochlorotic diatoms. I found and characterized a novel carbon pathway, a finding that I will test further with experimental techniques. In addition, this finding raised several questions about the origin and phylogenetic distribution of this pathway across diatoms as a whole.

Diatoms show great promise for industrial biofuel production (Traller et al. 2016; Tanaka et al. 2015), and the discovery of a β -ketoacid pathway in diatoms raises new possibilities for biotechnology applications. Finally, nonphotosynthetic diatoms are highly amenable to

laboratory experimentation, making it possible to test the genome-based predictions made in this study and establishing them as a useful model for understanding carbon metabolism in diatoms more generally.

References

- Armbrust, E. Virginia, John A. Berges, Chris Bowler, Beverley R. Green, Diego Martinez, Nicholas H. Putnam, Shiguo Zhou, et al. 2004. "The Genome of the Diatom *Thalassiosira Pseudonana*: Ecology, Evolution, and Metabolism." *Science* 306 (5693): 79–86.
- Barkman, Todd J., Joel R. McNeal, Seok Hong Lim, Gwen Coat, Henrietta B. Croom, Nelson D. Young, and Claude W. DePamphilis. 2007. "Mitochondrial DNA Suggests at Least 11 Origins of Parasitism in Angiosperms and Reveals Genomic Chimerism in Parasitic Plants." *BMC Evolutionary Biology* 7 (1). BioMed Central: 248.
- Blackburn, Michele V., Fiona Hannah, and Andrew Rogerson. 2009a. "First Account of Apochlorotic Diatoms from Mangrove Waters in Florida." *The Journal of Eukaryotic Microbiology* 56 (2): 194–200.
- . 2009b. "First Account of Apochlorotic Diatoms from Intertidal Sand of a South Florida Beach." *Estuarine, Coastal and Shelf Science* 84 (4): 519–26.
- Blouin, Nicolas A., and Christopher E. Lane. 2012. "Red Algal Parasites: Models for a Life History Evolution That Leaves Photosynthesis behind Again and Again." *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 34 (3): 226–35.
- Cavalier-Smith, T. 1999. "Principles of Protein and Lipid Targeting in Secondary Symbiogenesis: Euglenoid, Dinoflagellate, and Sporozoan Plastid Origins and the Eukaryote Family Tree." *The Journal of Eukaryotic Microbiology* 46 (4). Blackwell Publishing Ltd: 347–66.
- Goff, Lynda J., Debra A. Moon, Pi Nyvall, Birgit Stache, Katrina Mangin, and Giuseppe Zuccarello. 1996. "The Evolution of Parasitism in the Red Algae: Molecular Comparisons of Adelphoparasites and Their Hosts." *Journal of Phycology* 32 (2). Wiley Online Library: 297–312.
- Gruber, Ansgar, Sascha Vugrinec, Franziska Hempel, Sven B. Gould, Uwe-G Maier, and Peter G. Kroth. 2007. "Protein Targeting into Complex Diatom Plastids: Functional Characterisation of a Specific Targeting Motif." *Plant Molecular Biology* 64 (5): 519–30.
- Hadariová, Lucia, Matej Vesteg, Vladimír Hampl, and Juraj Krajčovič. 2018. "Reductive Evolution of Chloroplasts in Non-Photosynthetic Plants, Algae and Protists." *Current Genetics* 64 (2): 365–87.
- Harwood, C. S., and R. E. Parales. 1996. "The Beta-Ketoadipate Pathway and the Biology of Self-Identity." *Annual Review of Microbiology* 50: 553–90.
- Hellebust, Johan A., and Joyce Lewin. 1977. "Heterotrophic Nutrition." *The Biology of Diatoms*.

- Kamikawa, Ryoma, Daniel Moog, Stefan Zauner, Goro Tanifuji, Ken-Ichiro Ishida, Hideaki Miyashita, Shigeki Mayama, et al. 2017. “A Non-Photosynthetic Diatom Reveals Early Steps of Reductive Evolution in Plastids.” *Molecular Biology and Evolution* 34 (9). academic.oup.com: 2355–66.
- Kamikawa, Ryoma, Goro Tanifuji, Sohta A. Ishikawa, Ken-Ichiro Ishii, Yusei Matsuno, Naoko T. Onodera, Ken-Ichiro Ishida, et al. 2016. “Proposal of a Twin Arginine Translocator System-Mediated Constraint against Loss of ATP Synthase Genes from Nonphotosynthetic Plastid Genomes.” *Molecular Biology and Evolution* 33 (1). academic.oup.com: 303.
- Kamikawa, Ryoma, Naoji Yubuki, Masaki Yoshida, Misaka Taira, Noriaki Nakamura, Ken-Ichiro Ishida, Brian S. Leander, et al. 2015. “Multiple Losses of Photosynthesis in *Nitzschia* (Bacillariophyceae).” *Phycological Research* 63 (1): 19–28.
- Kroth, Peter G., Anthony Chiovitti, Ansgar Gruber, Veronique Martin-Jezequel, Thomas Mock, Micaela Schnitzler Parker, Michele S. Stanley, et al. 2008. “A Model for Carbohydrate Metabolism in the Diatom *Phaeodactylum Tricornutum* Deduced from Comparative Whole Genome Analysis.” *PLoS One* 3 (1): e1426.
- Lewin, Joyce, and R. A. Lewin. 1967. “Culture and Nutrition of Some Apochlorotic Diatoms of the Genus *Nitzschia*.” *Microbiology* 46 (3). Microbiology Society: 361–67.
- Mann, D. G., and S. J. M. Droop. 1996. “Biodiversity, Biogeography and Conservation of Diatoms.” In *Biogeography of Freshwater Algae*, 19–32. Developments in Hydrobiology. Springer, Dordrecht.
- Marin, Birger. 2004. “Origin and Fate of Chloroplasts in the Euglenoida.” *Protist* 155 (1). Urban & Fischer: 13–14.
- McFadden, G. I., M. E. Reith, J. Munholland, and N. Lang-Unnasch. 1996. “Plastid in Human Parasites.” *Nature* 381 (6582). Nature Publishing Group: 482.
- Nakov, Teofil, Jeremy M. Beaulieu, and Andrew J. Alverson. 2018. “Accelerated Diversification Is Related to Life History and Locomotion in a Hyperdiverse Lineage of Microbial Eukaryotes (Diatoms, Bacillariophyta).” *The New Phytologist*, April. <https://doi.org/10.1111/nph.15137>.
- Onyshchenko, Anastasiia, Elizabeth C. Ruck, Teofil Nakov, and Andrew J. Alverson. 2018. “A Single Loss of Photosynthesis in Diatoms.” *bioRxiv*. <https://doi.org/10.1101/298810>.
- Pringsheim, E. G. 1956. “Micro-Organisms from Decaying Seaweed.” *Nature* 178 (4531): 480–81.
- Raven, J. A. 1997. “Phagotrophy in Phototrophs.” *Limnology and Oceanography* 42 (1): 198–205.

- Round, F. E., R. M. Crawford, and D. G. Mann. 1990. *Diatoms: Biology and Morphology of the Genera*. Cambridge University Press.
- Sims, Patricia A., David G. Mann, and Linda K. Medlin. 2006. "Evolution of the Diatoms: Insights from Fossil, Biological and Molecular Data." *Phycologia* 45 (4). The International Phycological Society: 361–402.
- Smith, Sarah R., Raffaella M. Abbriano, and Mark Hildebrand. 2012. "Comparative Analysis of Diatom Genomes Reveals Substantial Differences in the Organization of Carbon Partitioning Pathways." *Algal Research* 1 (1): 2–16.
- Tanaka, Tsuyoshi, Yoshiaki Maeda, Alaguraj Veluchamy, Michihiro Tanaka, Heni Abida, Eric Maréchal, Chris Bowler, et al. 2015. "Oil Accumulation by the Oleaginous Diatom *Fistulifera Solaris* as Revealed by the Genome and Transcriptome." *The Plant Cell* 27 (1): 162–76.
- Taylor, F. J. R., Mona Hoppenrath, and Juan F. Saldarriaga. 2008. "Dinoflagellate Diversity and Distribution." *Biodiversity and Conservation* 17 (2): 407–18.
- Těšitel, Jakub. 2016. "Functional Biology of Parasitic Plants: A Review." *Plant Ecology and Evolution* 149 (1): 5–20.
- Traller, Jesse C., Shawn J. Cokus, David A. Lopez, Olga Gaidarenko, Sarah R. Smith, John P. McCrow, Sean D. Gallaher, et al. 2016. "Genome and Methylome of the Oleaginous Diatom *Cyclotella Cryptica* Reveal Genetic Flexibility toward a High Lipid Phenotype." *Biotechnology for Biofuels* 9 (November): 258.
- Tuchman, Nancy C., Marc A. Schollett, Steven T. Rier, and Pamela Geddes. 2006. "Differential Heterotrophic Utilization of Organic Compounds by Diatoms and Bacteria under Light and Dark Conditions." *Hydrobiologia* 561 (1). Kluwer Academic Publishers: 167–77.
- Waller, Ross F., and Geoffrey I. McFadden. 2005. "The Apicoplast: A Review of the Derived Plastid of Apicomplexan Parasites." *Current Issues in Molecular Biology* 7 (1): 57–79.
- Williams, David M., and J. Patrick Kociolek. 2011. "An Overview of Diatom Classification with Some Prospects for the Future." In *The Diatom World*, edited by Joseph Seckbach and Patrick Kociolek, 47–91. Dordrecht: Springer Netherlands.

Chapter 1. A Single Loss of Photosynthesis In Diatoms

Abstract

Loss of photosynthesis is a common and often repeated trajectory in nearly all major groups of photosynthetic eukaryotes. One small subset of ‘apochloritic’ diatoms in the genus *Nitzschia* have lost their ability to photosynthesize and require extracellular carbon for growth. Similar to other secondarily nonphotosynthetic taxa, apochloritic diatoms maintain colorless plastids with highly reduced plastid genomes. Although the narrow taxonomic breadth of apochloritic diatoms suggests a single loss of photosynthesis in their common ancestor, previous phylogenetic analyses suggested that photosynthesis was lost multiple times. We sequenced phylogenetic markers from the nuclear and mitochondrial genomes for a broad set of taxa and found that the best trees for data sets representing all three genetic compartments supported monophyly of apochloritic *Nitzschia*, consistent with a single loss of photosynthesis in diatoms. We sequenced the plastid genome of one apochloritic species and found that it was highly similar in all respects to the plastid genome of another apochloritic *Nitzschia* species, indicating that streamlining of the plastid genome had completed prior to the split of these two species. Finally, it is increasingly clear that some locales host relatively large numbers apochloritic *Nitzschia* species that span the phylogenetic diversity of the group, indicating that these species co-exist because of resource abundance or resource partitioning in ecologically favorable habitats. A better understanding of the phylogeny and ecology of this group, together with emerging genomic resources, will help identify the factors that have driven and maintained the loss of photosynthesis in this group, a rare event in diatoms.

Introduction

Photosynthetic eukaryotes (Archaeplastida) trace back to a single common ancestor, in which an eukaryotic host paired with a cyanobacterial endosymbiont that would eventually become the plastid, a fully integrated cellular organelle that is the site of photosynthesis (Archibald 2009, Keeling 2010). Although most archaeplastids remain photoautotrophic, loss of photosynthesis has occurred—often repeatedly—in nearly all major archaeplastid lineages (Hadariová et al. 2017). As many as five different lineages within the green algal orders Chlamydomonadales and Chlorellales have traded off photosynthesis for trophic strategies that include heterotrophy and obligate parasitism (Rumpf et al. 1996, Tartar and Boucias 2004, Yan et al. 2015). Partial or complete loss of photosynthesis has occurred in at least 11 different lineages of flowering plants, representing hundreds of species (Barkman et al. 2007); parasitic angiosperms obtain extracellular carbon from the vascular tissues of host plants for all (holoparasites) or part (hemiparasites) of their life history (Těšitel 2016). Perhaps most strikingly, photosynthesis has been lost dozens or more times across the florideophycean red algae (e.g., Goff et al. 1996, Kurihara et al. 2010), which go on to parasitize a closely related photosynthetic species—a phenomenon known as adelphoparasitism (Blouin and Lane 2012). Photosynthesis has been lost in a broad range of taxa with secondary plastids as well, including euglenoids (Marin 2004), apicomplexans (McFadden et al. 1996, Waller and McFadden 2005), ciliates (Reyes-Prieto et al. 2008), dinoflagellates (Saldarriaga et al. 2001), cryptophytes (Donaher et al. 2010, Martin-Cereceda et al. 2010), and stramenopiles (Tyler et al. 2006). Most nonphotosynthetic algae evolved from mixotrophic ancestors (Figuroa-Martinez et al. 2015),

likely because they already had the means to secure extracellular carbon and because the energetic costs of mixotrophy are thought to be high (Raven 1997).

Diatoms are a lineage of stramenopile algae responsible for roughly 20% of global primary production (Field et al. 1998). They are ancestrally photosynthetic, and although the overwhelming majority of the estimated 100,000 or so diatom species remain photosynthetic, many species are mixotrophic, which allows them to use external sources of carbon for growth in fluctuating light conditions (Hellebust and Lewin 1977, Tuchman et al. 2006). A much smaller set of 20 or so mostly undescribed, colorless, free-living species in the genus *Nitzschia*, and one species in the closely related and morphologically similar genus *Hantzschia*, have abandoned photosynthesis altogether and rely exclusively on extracellular carbon for growth (Lewin and Lewin 1967). These “apochloritic” diatoms—the only known nonphotosynthetic diatom species—are often found in association with mangroves, and decaying seaweeds and sea grasses (Pringsheim 1956, Blackburn et al. 2009a, Kamikawa et al. 2015b).

The small number of species and narrow taxonomic range of apochloritic diatoms leads to the prediction, based on parsimony, that nonphotosynthetic diatoms are monophyletic, tracing back to a single loss of photosynthesis in their common ancestor. Despite the small number of species, however, they encompass a relatively broad range of morphological diversity (Blackburn et al. 2009b, Kamikawa et al. 2015b) and use a variety of carbon sources (Lewin and Lewin 1967, Hellebust and Lewin 1977), raising the possibility that obligate heterotrophy evolved multiple times—a hypothesis supported by a phylogenetic analysis of nuclear 28S d1–d2 sequences that separated apochloritic taxa into three separate clades (Kamikawa et al. 2015b). Monophyly of apochloritic species could not be rejected, however (Kamikawa et al. 2015b).

These two competing hypotheses (one vs. multiple origins) have important implications for our understanding of the underlying phylogenetic, ecological, and genomic contexts of this radical trophic shift, which has occurred far less frequently in diatoms than it has in other groups.

Like most other lineages that have lost photosynthesis, apochloritic diatoms maintain highly reduced plastids and plastid genomes (Kamikawa et al. 2015a). Their plastids lack chlorophyll and thylakoids (Kamikawa et al. 2015b), and their plastid genomes have lost most photosynthesis-related genes, including all photosystem genes (Kamikawa et al. 2015a). The nearly complete set of ATP synthase genes in the plastid genome might function in ATP hydrolysis, creating a proton gradient that fuels protein import into the plastid (Kamikawa et al. 2015a). Carbon metabolism is highly compartmentalized in diatoms (Smith et al. 2012), and the large number of nuclear-encoded proteins targeted to the plastid point to a highly metabolically active and, as a result, indispensable (Kamikawa et al. 2017) organelle. Although comparative genomics is greatly improving our understanding of carbon metabolism in both photosynthetic (Smith et al. 2012) and nonphotosynthetic diatoms (Kamikawa et al. 2015a, 2017), the power of comparative genomics can only be fully leveraged within the framework of an accurate, densely sampled phylogenetic hypothesis.

We collected, isolated, and cultured several apochloritic *Nitzschia* species and sequenced common phylogenetic markers to test whether photosynthesis was lost one or multiple times. A combined dataset of nuclear, mitochondrial, and plastid genes support monophyly of apochloritic *Nitzschia* species, consistent with a single loss of photosynthesis in diatoms. Species were split between two major subclades that co-occur in habitats with apparently favorable, albeit poorly defined, conditions. We also sequenced and characterized the plastid genome of *Nitzschia* sp. and

found it to be highly similar to that of another species in the same subclade, indicating rapid genomic streamlining following loss of photosynthesis in their common ancestor. The results presented here help frame a number of questions about the evolution, ecology, and genomic consequences of this rare, radical trophic shift in diatoms.

Materials and Methods

Collection and culturing of apochloritic Nitzschia

All cultures originated from a single composite sample, collected on 10 November 2011 from Whiskey Creek, which is located in Dr. Von D. Mizell-Eula Johnson State Park (formerly John U. Lloyd State Park), Dania Beach, Florida, USA (26.081330 lat, -80.110783 long). This site was previously shown to host a diverse assemblage of apochloritic *Nitzschia* species (Blackburn et al. 2009a). Our sample consisted of near-surface plankton collected with a 10 μ M mesh net, submerged sand (1m and 2m depth), and nearshore wet (but unsubmerged) sand. We selected for nonphotosynthetic species by storing the sample in the dark at room temperature (21°C) for several days before isolating colorless diatom cells with a Pasteur pipette. Clonal cultures were grown in the dark at 21°C on agar plates made with L1+NPM medium (Guillard 1960, Guillard and Hargraves 1993) and 1% Penicillin–Streptomycin–Neomycin solution (Sigma-Aldrich P4083) to retard bacterial growth.

DNA extraction, PCR, and DNA sequencing

Cells were rinsed with L1 medium and removed from agar plates by pipetting, briefly centrifuged, then broken with MiniBeadbeater-24 (BioSpec Products). We then extracted DNA with a Qiagen DNeasy Plant Mini Kit. Nuclear *SSU* and partial *LSU* rDNA genes were

PCR-amplified and sequenced using published PCR conditions and primer sequences (Alverson et al. 2007). PCR and sequencing primers for two mitochondrial markers, cytochrome b (*cob*) and NADH dehydrogenase subunit 1 (*nad1*), are listed in Table S1. PCRs for mitochondrial genes used: 1.0–5.0 µL of DNA, 6.5 µL of Failsafe Buffer E (Epicentre Technologies), 0.5 µL of each primer (20 µM stocks), 0.5 units Taq polymerase, and ddH₂O to a final volume of 25 µL. In a few cases, we used a nested PCR strategy to amplify the *nad1* gene. PCR conditions for *cob* and *nad1* genes were as follows: 95 °C for 5 minutes, 36 cycles of (95 °C for 60 s, 45 °C for 60 s, 72 °C for 60 s), and a final extension at 72 °C for 5 minutes. PCR products were sequenced on an ABI 3100 capillary sequencer. Raw sequences were assembled and edited with Geneious ver. 7.1.4 (Biomatters Ltd.) and deposited in GenBank (Table 1).

Phylogenetic analyses

In addition to the data generated here, we compiled previously published data for the plastid 16S and nuclear 28S rDNA genes for Bacillariales to better enable comparisons with previous studies. We downloaded all Bacillariales sequences from Genbank, checked percent identity and coverage of each sequence to a local BLAST database comprised of apochloritic *Nitzschia* sequences, and reverse-complemented sequences if necessary. We kept Bacillariales sequences that were at least 20% identical and covered at least 40% of at least one target in the database (BLAST options: identity_cutoff=20, hsp_coverage_cutoff=40). We kept only the longest sequence for cases in which multiple downloaded sequences had the same NCBI Taxid identifier. The total number of sequences meeting these criteria were 46 for the 16S and 116 for the 28S rDNA genes.

We used SSU-ALIGN ver. 0.1 (Nawrocki et al. 2009) to align rDNA sequences (plastid 16S and the d1–d3 region of the nuclear 28S rDNA), using SSU-ALIGN’s built-in covariance models of secondary structure for bacteria for the 16S alignment and a heterokont-specific covariance model for the 28S alignment (Nakov et al. 2014). We used SSU-MASK to remove poorly aligned regions and used these more conservative alignments for downstream analyses. The protein-coding *cob* and *nad1* genes from the mitochondrial genome were aligned by hand with Mesquite ver. 3.31 (Maddison, W. P. and Maddison, D. R. 2008) after color-coding nucleotide triplets by their conceptual amino acid translations. We built phylogenetic trees from each individual gene alignment and three concatenated alignments: (1) a concatenation of the two mitochondrial genes into a *cob+nad1* (heretofore *mito* dataset), (2) a concatenation of the two mitochondrial genes and the masked 28S alignment of newly generated 28S sequences (heretofore *mito-lsu* dataset), and (3) a concatenation of the two mitochondrial genes and the masked alignment of all 28S sequences (newly generated and downloaded from GenBank, heretofore *mito-ncbi-lsu* dataset). The alignments have been archived in a ZENODO online repository (<https://doi.org/10.5281/zenodo.1211571>).

We used RAxML ver. 8.2.4 (Stamatakis 2014) to infer phylogenetic trees from each of the three concatenated alignments. For each dataset, we performed 10 maximum likelihood searches to find the best-scoring tree and a rapid bootstrap analysis of 250 pseudoreplicates per search (Stamatakis et al. 2008), for a total of 2500 bootstrap samples per alignment. We used the general time-reversible model (GTR) of nucleotide substitution, and we used a Gamma distribution to accommodate rate variation across sites within each alignment (GTR+GAMMA, the default model in RAxML).

Plastid genome sequencing and analysis

We sequenced the plastid genome of *Nitzschia* sp. (Nitz4) using the Illumina HiSeq2000 platform, with a 300-bp library and 90-bp paired-end reads. We removed adapter sequences and trimmed raw reads with Trimmomatic ver. 0.32 and settings ‘ILLUMINACLIP=<TruSeq_adapters.fasta>:2:30:10, TRAILING=5, SLIDINGWINDOW=6:18, HEADCROP=9, MINLEN=50’ (Bolger et al. 2014). We assembled trimmed reads using Ray ver. 2.3.1 with default settings and k-mer length of 45 (Boisvert et al. 2012). We assessed the assembly quality with QUASt ver. 2.3 (Gurevich et al. 2013), confirmed high genome-wide read coverage by mapping the trimmed reads to the assembly with Bowtie ver. 0.12.8 (Langmead et al. 2009), and evaluated these results with SAMtools. We used DOGMA (Wyman et al. 2004) and ARAGORN (Laslett and Canback 2004) to annotate the genome. The annotated genome has been archived in GenBank under accession MG273660. We used Easyfig (Sullivan et al. 2011) to perform a BLASTN-based synteny comparison of the plastid genomes of two apochloritic *Nitzschia* taxa, strains Nitz4 (this study) and NIES-3581 (Kamikawa et al. 2015a).

Results

Phylogeny of apochloritic Nitzschia

We isolated and cultured four strains (Nitz2, Nitz4, Nitz7, and Nitz8) of apochloritic *Nitzschia* from Whiskey Creek (Florida, USA), including both linear and undulate forms (Fig. 1-1). Although cells grew well in culture, we never observed sexual reproduction and, unlike

some other apochloritic *Nitzschia* strains that have been maintained in culture for decades, all of our strains experienced substantial size reduction and eventually died off.

We sequenced the nuclear 28S rDNA and mitochondrial *cob* and *nad1* genes for 26 Bacillariales taxa, including seven apochloritic strains (Table 1-1). We combined these data with GenBank sequences to reconstruct phylogenetic trees for individual and concatenated alignments. The plastid 16S rDNA gene supported monophyly of apochloritic *Nitzschia* species (Fig. 1-2A, Bootstrap proportion (BS)=100), though the exceptionally long branch lengths—driven by decreased GC content—raises the possibility that this grouping simply reflects shared nucleotide composition (Steel et al. 1993, Galtier and Gouy 1995, Kamikawa et al. 2015b). One way to reduce the influence of GC bias in tree inference is to transform the alignment into purine/pyrimidine (R/Y) coding. By treating all A and G nucleotides as R and all C and T nucleotides as Y, the GC bias present in some sequences is masked, the base frequencies are normalized, and the majority of phylogenetic signal originates from transversions (i.e., R ↔ Y). The phylogeny resulting from the R/Y-transformed alignment should better reflect the history of the lineage rather than the nucleotide composition bias. Performing this transformation for the plastid 16S alignment, we again found strong support for monophyly of the apochloritic *Nitzschia* (bootstrap support = 95%) (Fig. 1-2B). This result suggests that monophyly of apochloritic *Nitzschia* species reconstructed with the plastid-encoded 16S gene might not necessarily be an artifact of the decreased GC content in the plastid genomes of nonphotosynthetic taxa.

A previous analysis of nuclear 28S d1–d2 rDNA sequences weakly supported non-monophyly of apochloritic taxa, though monophyly could not be rejected (Kamikawa et al.

2015b). Expanding the analysis to include the larger number of Bacillariales 28S d1–d2 sequences from GenBank and newly sequenced taxa from this study (Table 1-1) returned a monophyletic grouping of apochloritic *Nitzschia* (Fig. 1-2C). We also sequenced two mitochondrial genes, *cob* and *nad1*, and although neither of them individually supported monophyly of apochloritic taxa (Fig. 1-2 D, E), analysis of a concatenated mitochondrial *cob+nad1* alignment did return a clade of apochloritic *Nitzschia* (Fig. 1-2F). Concatenated alignments of nuclear 28S d1–d2 and the two mitochondrial genes supported monophyly of apochloritic *Nitzschia* as well, both for an analysis restricted to just newly sequenced taxa (Fig. 1-2G) and one that included both newly sequenced and GenBank taxa (Fig. 2H). In all cases, branch support for monophyly of apochloritic *Nitzschia* with nuclear and mitochondrial markers, analyzed individually or in combination, was less than 70% (Fig. 1-2). The fully labeled trees are available in Fig. S2.

Plastid genome sequencing and analysis

The plastid genome of *Nitzschia* sp. Nitz4 maps as a circular chromosome and has the conserved quadripartite architecture typical of most plastid genomes, with two inverted repeat (IR) regions and small and large single copy regions (SSC and LSC, respectively). At 67,764 bp in length, the genome is roughly half the size of plastid genomes from photosynthetic diatoms (Ruck et al. 2014, Yu et al. 2018). The genome contains minimal intergenic DNA, with 12 genes that overlap by anywhere from 1–61 bp in length and another two genes that are immediately adjacent to one another. (Table S1-2). The genome consists mainly of genes with conserved housekeeping functions, including 32 tRNA and 6 rRNA genes. All three rRNA and six of the tRNA genes are present twice in the genome because of their location in the IR region. More

than half (45) of the 73 protein-coding genes encode ribosomal proteins or subunits of RNA polymerase, with ATP synthase genes and ORFs constituting most of the remaining genes. Some of the ORFs include ones (*ycf41*, *ycf89* and *ycf90*) that are highly conserved among diatoms (Ruck et al. 2014, Yu et al. 2018). The genome also contains subunit of Sec-independent protein translocator TatC and SecA subunit of Sec-mediated transport system, iron-sulfur clusters forming proteins SufB and SufC, molecular chaperone *dnaK* and protease (*ClpC*). The genome is highly AT-rich (77.6% A+T).

Similar to the plastid genome of another apochloritic diatom, *Nitzschia* sp. NIES-3581 (aka IriIs04) (Kamikawa et al. 2015a), the genome of *Nitzschia* sp. Nitz4 is missing all genes encoding subunits of photosystem I and II, proteins of cytochrome *b6f* complex, carbon fixation system, porphyrin and thiamine metabolism. As well Nitz4 lacks several conserved ORFs and membrane translocators subunits, *dnaB* helicase, chaperonine *groEL* and *ftsH* cell division protein. The genomes of Nitz4 and NIES-3581 are highly similar in size, gene content, gene order, nucleotide composition, and sequence (Fig. 3). The loss of *rps20* in Nitz4, which appears to have been replaced by a unique ORF (*orf122*), is among the few minor differences between the two genomes (Fig. 1-3). Two gene fusions, *orf96-atpB* and *orf122-rpoB*, present in Nitz4 are separated in NIES-3581. Likewise, the overlapping *dnaK-tRNA-Arg* genes in NIES-3581 are adjacent but non-overlapping in Nitz4.

Discussion

Monophyly of apochloritic diatoms

The small number of species (roughly 20) and narrow taxonomic range (all Bacillariales, mostly all *Nitzschia*) of apochloritic diatoms suggests that these species—the only known nonphotosynthetic diatoms—are monophyletic, reflecting a single loss of photosynthesis and transition to obligate heterotrophy in their common ancestor. Previous phylogenetic tests of this hypothesis were equivocal, however, with the plastid 16S rDNA sequences supporting monophyly of apochloritic species and the nuclear 28S d1–d2 gene splitting them, albeit with low bootstrap support, into three separate clades (Kamikawa et al. 2015b). Further underscoring the uncertainty in their relationships, the non-normalized, highly AT-rich plastid 16S sequences may have resulted in long-branch artifacts (Steel et al. 1993, Galtier and Gouy 1995), and the nuclear 28S dataset could not reject monophyly of apochloritic species (Kamikawa et al. 2015b), pointing to insufficient signal in this relatively short sequence fragment to address this question. As efforts to understand the causes and genomic consequences of loss of photosynthesis in diatoms accelerate (Kamikawa et al. 2015a, 2015b, 2017), it is important to determine if the shift occurred one or multiple times. The number, pattern, and timing of shifts has important implications for understanding of the ease of transition(s) to obligate heterotrophy, the ecological setting of the transition(s), and, in practical terms, the power of comparative approaches to resolve these questions—as evolutionary replication, represented in this case by multiple losses of photosynthesis—maximizes the power of comparative methods to reveal possible mechanisms underlying these kinds of evolutionary transitions (e.g., Maddison and FitzJohn 2015). A single transition would, by contrast, greatly limit the power of comparative methods to uncover the

underlying genomic and ecological drivers of the switch away from auto- or mixotrophy to obligate heterotrophy.

To address this problem, we increased both the number of taxa and genes available for phylogenetic analyses. We also normalized the plastid 16S rDNA sequences to guard against artefactual grouping of apochloritic species with highly AT-rich plastid genomes. Although support was generally low, the best trees inferred from genes representing all three genetic compartments uniformly supported monophyly of apochloritic *Nitzschia* (Fig.1- 2). In short, the best available data support a single loss of photosynthesis in *Nitzschia* and, by extension, diatoms as a whole. In light of this, future research efforts can focus attention away from questions that naturally arise for characters that evolve multiple times (e.g., the role of convergent evolution) and focus more on firmly placing the apochloritic clade within the broader Bacillariales phylogeny to help identify, for example, the ecological and mixotrophic characteristics of their closest relatives. Considerable divergence and phylogenetic structure exists within the apochloritic clade, with the largest combined-data tree (mito-lsu-new-ncbi) splitting taxa into two major subclades (Figs. 1-2 and S1-1). The modest level of species diversity, in turn, might make it feasible to sample each subclade more-or-less exhaustively and apply comparative approaches that will be useful for understanding the genomic and metabolic consequences of the switch to heterotrophy, including, for example, the rate of decay of the photosynthetic apparatus. All of this requires a robust, time-calibrated phylogeny of Bacillariales, which is one of the largest and most species-rich taxonomic orders of diatoms (Kociolek et al. 2018). This and other phylogenetic studies have begun resolving some parts of the Bacillariales tree using small numbers of commonly used phylogenetic markers (e.g., Lundholm et al. 2002, Rimet et al. 2011,

Smida et al. 2014), but many relationships remain unresolved. With most of the traditional markers now more-or-less exhausted, effort should turn to much deeper sampling of the nuclear genome, which appears to hold great promise for resolving relationships across many phylogenetic scales within diatoms (Parks et al. 2018).

The ecology and biogeography of apochloritic diatoms

Although relatively few apochloritic *Nitzschia* species have been formally described and named, the amount of phylogenetic diversity described in this (Fig. 1-2) and other studies (Kamikawa et al. 2015b), suggests that this clade could easily comprise on the order of 20 species. This seems even more probable when considering the relatively modest historical efforts to collect and characterize apochloritic diatoms. One emerging, and quite striking, theme among the small number of studies that focused on characterizing species diversity in this group is the large number and diversity of apochloritic *Nitzschia* that co-occur over very small spatial scales (e.g., within a sample from a single site). The diversity is apparent from both morphological and molecular phylogenetic evidence alike (Figs. 1-1 and 1-2; Blackburn et al. 2009a, Kamikawa et al. 2015b). Although some of the apochloritic taxa in our trees came from culture collections (Table 1-1 and Fig. S1-1), most of the taxa derive from a small number of mangrove-dominated habitats in Japan (Kamikawa et al. 2015b) and the United States (Figs. 1-1 and 1-2; Blackburn et al. 2009a). Both of these sites host multiple (apparently undescribed) apochloritic *Nitzschia* species that span the full phylogenetic breadth of the clade (Fig. S1-1). Focused efforts to collect and culture apochloritic *Nitzschia* from amenable habitats (e.g., seagrasses and mangroves) worldwide will show whether this anecdotal trend holds. If it is upheld, the next natural step will be to determine whether species co-occurrence is made possible by the sheer abundance of local

resources in these habitats or rather by resource partitioning among species, either in fine-scale microhabitats or through specialization on different carbon sources. Of course, both of these alternatives require knowledge of the organic carbon sources used by these species.

Nevertheless, the phylogeny clearly shows species are not clustered geographically (Fig. S1-1), which suggests that many or most apochloritic diatom species have broad, presumably worldwide, geographic distributions and that the local species pool at any one site is the product of dispersal, not *in situ* speciation.

Plastid genome reduction in apochloritic diatoms

Although probably different species, the two *Nitzschia* strains with sequenced plastid genomes belong to the same subclade (Fig. S1A, bottom) and so, in the context of the entire apochloritic clade, are very close relatives. Their close relationship was also evident in their plastid genomes, which were highly similar in nearly all respects. Additional sampling across the entire apochloritic clade will show whether all species share the same fundamental, highly reduced plastid genome—indicative of rapid genomic streamlining following the loss of photosynthesis in their common ancestor—or whether decay of the plastid genome is ongoing in some taxa. Similar to some green algae (Knauf and Hachtel 2002), diatoms have retained a nearly full set of ATP synthase genes whose products, instead of functioning in ATP hydrolysis, presumably generate a proton gradient for tat-dependent protein translocation across the thylakoid membrane (Kamikawa et al. 2015a). This model underscores both the compartmentalized nature of carbon metabolism in diatoms (Smith et al. 2012, Kamikawa et al. 2017) and, consequently, the indispensable nonphotosynthetic plastids in diatoms.

In addition to understanding the specific consequences for the plastid genome to following the loss of photosynthesis, a fuller understanding of these plastid genomes can shed light on photosynthetic diatom plastids as well. For example, the retention several conserved ORFs in these nonphotosynthetic genomes—in the context of near wholesale loss of the photosynthetic apparatus—strongly suggests these conserved, diatom-specific ORFs have functions that are unrelated, or only peripherally related, to photosynthesis. These ORFs include *ycf41*, *ycf89*, and *ycf90*. Finally, the loss of photosynthesis can have cascading effects on rates of sequence evolution all three genomes (Nickrent et al. 1998). Branch lengths based on plastid, mitochondrial, and nuclear genes showed that the AT-driven rate acceleration in the plastid genomes of apochloritic species appears to be restricted to the plastid genome alone (Fig.1- 2). This does not rule out that there have not been other effects on the mitochondrial and nuclear genomes including, for example, gene content and whether some of the missing plastid genes have been transferred to the nucleus.

Conclusions

The discovery that apochloritic *Nitzschia* are monophyletic represents an important step forward in understanding the loss of photosynthesis and switch to obligate heterotrophy in diatoms—a transition that has occurred just once in a lineage of some 100,000 species of photoautotrophs. A clearer view of this relationship highlights new research avenues and priorities, including more intensive taxon sampling within Bacillariales to identify the closest relatives of the apochloritic clade, which almost certainly evolved from a mixotrophic ancestor (Hellebust and Lewin 1977, Figueroa-Martinez et al. 2015). The exact carbon sources, modes of

carbon uptake and utilization, and the degree of carbon specialization within and across species remains unclear for photosynthetic and nonphotosynthetic *Nitzschia* alike. Bacillariales features a diverse set of taxa with fascinating biology, motivating the development of excellent genomic resources for this group (Basu et al. 2017, Kamikawa et al. 2017, Mock et al. 2017). A nuclear genome sequence for a nonphotosynthetic *Nitzschia* could help address several of these outstanding questions, leading to hypotheses that can be directly tested in a group of diatoms that has proven highly amenable to experimental manipulation (e.g., Lewin and Lewin 1967, Azam and Volcani 1974, McGinnis and Sommerfeld 2000).

References

- Alverson, A.J., Jansen, R.K. & Theriot, E.C. 2007. Bridging the Rubicon: Phylogenetic analysis reveals repeated colonizations of marine and fresh waters by thalassiosiroid diatoms. *Mol. Phylogenet. Evol.* 45:193–210.
- Archibald, J.M. 2009. The puzzle of plastid evolution. *Curr. Biol.* 19:R81–8.
- Azam, F. & Volcani, B.E. 1974. Role of silicon in diatom metabolism. *Arch. Microbiol.* 101:1–8.
- Barkman, T.J., McNeal, J.R., Lim, S.H., Coat, G., Croom, H.B., Young, N.D. & DePamphilis, C.W. 2007. Mitochondrial DNA suggests at least 11 origins of parasitism in angiosperms and reveals genomic chimerism in parasitic plants. *BMC Evol. Biol.* 7:248.
- Basu, S., Patil, S., Mapleson, D., Russo, M.T., Vitale, L., Fevola, C., Maumus, F. et al. 2017. Finding a partner in the ocean: Molecular and evolutionary bases of the response to sexual cues in a planktonic diatom. *New Phytol.* 215:140–56.
- Blackburn, M.V., Hannah, F. & Rogerson, A. 2009a. First account of apochlorotic diatoms from mangrove waters in Florida. *J. Eukaryot. Microbiol.* 56:194–200.
- Blackburn, M.V., Hannah, F. & Rogerson, A. 2009b. First account of apochlorotic diatoms from intertidal sand of a south Florida beach. *Estuar. Coast. Shelf Sci.* 84:519–26.
- Blouin, N.A. & Lane, C.E. 2012. Red algal parasites: Models for a life history evolution that leaves photosynthesis behind again and again. *Bioessays.* 34:226–35.
- Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F. & Corbeil, J. 2012. Ray Meta: Scalable *de novo* metagenome assembly and profiling. *Genome Biol.* 13:R122.
- Bolger, A.M., Lohse, M. & Usadel, B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics.* 30:2114–20.
- Donaher, N., Tanifuji, G., Onodera, N.T., Malfatti, S.A., Chain, P.S.G., Hara, Y. & Archibald, J.M. 2010. The complete plastid genome sequence of the secondarily nonphotosynthetic alga *Cryptomonas paramecium*: Reduction, compaction, and accelerated evolutionary rate. *Genome Biol. Evol.* 1:439–48.
- Field, C.B., Behrenfeld, M.J., Randerson, J.T. & Falkowski, P. 1998. Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science.* 281:237–40.
- Figuerola-Martinez, F., Nedelcu, A.M., Smith, D.R. & Reyes-Prieto, A. 2015. When the lights go out: The evolutionary fate of free-living colorless green algae. *New Phytol.* 206:972–82.
- Galtier, N. & Gouy, M. 1995. Inferring phylogenies from DNA sequences of unequal base compositions. *Proc. Natl. Acad. Sci. U. S. A.* 92:11317–21.

- Goff, L.J., Moon, D.A., Nyvall, P., Stache, B., Mangin, K. & Zuccarello, G. 1996. The evolution of parasitism in the red algae: Molecular comparisons of adelphoparasites and their hosts. *J. Phycol.* 32:297–312.
- Guillard, R.R.L. 1960. A mutant of *Chlamydomonas moewusii* lacking contractile vacuoles. *J. Eukaryot. Microbiol.* 7:262–8.
- Guillard, R.R.L. & Hargraves, P.E. 1993. *Stichochrysis immobilis* is a diatom, not a chrysophyte. *Phycologia.* 32:234–6.
- Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. 2013. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics.* 29:1072–5.
- Hadariová, L., Vesteg, M., Hampl, V. & Krajčovič, J. 2017. Reductive evolution of chloroplasts in non-photosynthetic plants, algae and protists. *Curr. Genet.*
- Hellebust, J.A. & Lewin, J. 1977. Heterotrophic nutrition. In Werner, D. [Ed.] *The Biology of Diatoms*. University of California Press, pp. 169–97.
- Kamikawa, R., Moog, D., Zauner, S., Tanifuji, G., Ishida, K.-I., Miyashita, H., Mayama, S. et al. 2017. A non-photosynthetic diatom reveals early steps of reductive evolution in plastids. *Mol. Biol. Evol.* 34:2355–66.
- Kamikawa, R., Tanifuji, G., Ishikawa, S.A., Ishii, K.I., Matsuno, Y., Onodera, N.T., Ishida, K.I. et al. 2015a. Proposal of a twin arginine translocator system-mediated constraint against loss of ATP synthase genes from nonphotosynthetic plastid genomes. *Mol. Biol. Evol.* 32:2598–604.
- Kamikawa, R., Yubuki, N., Yoshida, M., Taira, M., Nakamura, N., Ishida, K.I., Leander, B.S. et al. 2015b. Multiple losses of photosynthesis in *Nitzschia* (Bacillariophyceae). *Phycological Res.* 63:19–28.
- Keeling, P.J. 2010. The endosymbiotic origin, diversification and fate of plastids. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 365:729–48.
- Knauf, U. & Hachtel, W. 2002. The genes encoding subunits of ATP synthase are conserved in the reduced plastid genome of the heterotrophic alga *Prototheca wickerhamii*. *Mol. Genet. Genomics.* 267:492–7.
- Kocielek, J.P., Balasubramanian, K., Blanco, S., Coste, M., Ector, L., Liu, Y., Kulikovskiy, M. et al. 2018. DiatomBase. Available at: <http://www.diatombase.org/> (last accessed April 5, 2018).
- Kurihara, A., Abe, T., Tani, M. & Sherwood, A.R. 2010. Molecular phylogeny and evolution of red algal parasites: a case study of *Benzaitenia*, *Janczewskia*, and *Ululania* (Ceramiales). *J. Phycol.* 46:580–90.

- Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10:R25.
- Laslett, D. & Canback, B. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* 32:11–6.
- Lewin, J. & Lewin, R.A. 1967. Culture and nutrition of some apochlorotic diatoms of the genus *Nitzschia*. *Microbiology.* 46:361–7.
- Lundholm, N., Daugbjerg, N. & Moestrup, Ø. 2002. Phylogeny of the Bacillariaceae with emphasis on the genus *Pseudo-nitzschia* (Bacillariophyceae) based on partial LSU rDNA. *Eur. J. Phycol.* 37:115–34.
- Maddison, W. P. and Maddison, D. R. 2008. Mesquite: A modular system for evolutionary analysis. *Evolution.* 62:1103–18.
- Maddison, W.P. & FitzJohn, R.G. 2015. The unsolved challenge to phylogenetic correlation tests for categorical characters. *Syst. Biol.* 64:127–36.
- Marin, B. 2004. Origin and fate of chloroplasts in the euglenoida. *Protist.* 155:13–4.
- Martin-Cereceda, M., Roberts, E.C., Wootton, E.C., Bonaccorso, E., Dyal, P., Guinea, A., Rogers, D. et al. 2010. Morphology, ultrastructure, and small subunit rDNA phylogeny of the marine heterotrophic flagellate *Goniomonas* aff. *amphinema*. *J. Eukaryot. Microbiol.* 57:159–70.
- McFadden, G.I., Reith, M.E., Munholland, J. & Lang-Unnasch, N. 1996. Plastid in human parasites. *Nature.* 381:482.
- McGinnis, K.M. & Sommerfeld, M.R. 2000. PLASTID FATTY ACID BIOSYNTHESIS IN THE DIATOMS *NITZSCHIA ALBA* AND *NITZSCHIA LAEVIS*. *J. Phycol.* 36:46–46.
- Mock, T., Otilar, R.P., Strauss, J., McMullan, M., Paajanen, P., Schmutz, J., Salamov, A. et al. 2017. Evolutionary genomics of the cold-adapted diatom *Fragilariopsis cylindrus*. *Nature.* 541:536–40.
- Nakov, T., Ruck, E.C., Galachyants, Y., Spaulding, S.A. & Theriot, E.C. 2014. Molecular phylogeny of the Cymbellales (Bacillariophyceae, Heterokontophyta) with a comparison of models for accommodating rate variation across sites. *Phycologia.* 53:359–73.
- Nawrocki, E.P., Kolbe, D.L. & Eddy, S.R. 2009. Infernal 1.0: Inference of RNA alignments. *Bioinformatics.* 25:1335–7.

- Nickrent, D.L., Duff, R.J., Colwell, a. E., Wolfe, a. D., Young, N.D., Steiner, K.E. & DePamphilis, C.W. 1998. Molecular phylogenetic and evolutionary studies of parasitic plants. In *Molecular Systematics of Plants II. DNA Sequencing*. Springer US, Boston, MA, p. 211–41.
- Parks, M.B., Wickett, N.J. & Alverson, A.J. 2018. Signal, Uncertainty, and Conflict in Phylogenomic Data for a Diverse Lineage of Microbial Eukaryotes (Diatoms, Bacillariophyta). *Mol. Biol. Evol.* 35:80–93.
- Pringsheim, E.G. 1956. Micro-organisms from decaying seaweed. *Nature*. 178:480–1.
- Raven, J.A. 1997. Phagotrophy in phototrophs. *Limnol. Oceanogr.* 42:198–205.
- Reyes-Prieto, A., Moustafa, A. & Bhattacharya, D. 2008. Multiple genes of apparent algal origin suggest ciliates may once have been photosynthetic. *Curr. Biol.* 18:956–62.
- Rimet, F., Kermarrec, L., Bouchez, A., Hoffmann, L., Ector, L. & Medlin, L.K. 2011. Molecular phylogeny of the family Bacillariaceae based on 18S rDNA sequences: focus on freshwater *Nitzschia* of the section Lanceolatae. *Diatom Res.* 26:273–91.
- Ruck, E.C., Nakov, T., Jansen, R.K., Theriot, E.C. & Alverson, A.J. 2014. Serial gene losses and foreign DNA underlie size and sequence variation in the plastid genomes of diatoms. *Genome Biol. Evol.* 6:644–54.
- Rumpf, R., Vernon, D., Schreiber, D. & Birky, C.W. 1996. Evolutionary consequences of the loss of photosynthesis in Chlamydomonadaceae: Phylogenetic analysis of *Rrn18* (18S rDNA) in 13 *Polytoma* strains (Chlorophyta). *J. Phycol.* 32:119–26.
- Saldarriaga, J.F., Taylor, F.J.R., Keeling, P.J. & Cavalier-Smith, T. 2001. Dinoflagellate nuclear SSU rRNA phylogeny suggests multiple plastid losses and replacements. *J. Mol. Evol.* 53:204–13.
- Smida, D.B., Lundholm, N., Kooistra, W.H.C.F., Sahraoui, I., Ruggiero, M.V., Kotaki, Y., Ellegaard, M. et al. 2014. Morphology and molecular phylogeny of *Nitzschia bizertensis* sp. nov.—A new domoic acid-producer. *Harmful Algae*. 32:49–63.
- Smith, S.R., Abbriano, R.M. & Hildebrand, M. 2012. Comparative analysis of diatom genomes reveals substantial differences in the organization of carbon partitioning pathways. *Algal Research*. 1:2–16.
- Stamatakis, A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 30:1312–3.
- Stamatakis, A., Hoover, P. & Rougemont, J. 2008. A rapid bootstrap algorithm for the RAxML Web servers. *Syst. Biol.* 57:758–71.

- Steel, M.A., Lockhart, P.J. & Penny, D. 1993. Confidence in evolutionary trees from biological sequence data. *Nature*. 364:440–2.
- Sullivan, M.J., Petty, N.K. & Beatson, S.A. 2011. Easyfig: A genome comparison visualizer. *Bioinformatics*. 27:1009.
- Tartar, A. & Boucias, D.G. 2004. The non-photosynthetic, pathogenic green alga *Helicosporidium* sp. has retained a modified, functional plastid genome. *FEMS Microbiol. Lett.* 233:153–7.
- Těšitel, J. 2016. Functional biology of parasitic plants: A review. *Plant Ecol. Evol.* 149:5–20.
- Tuchman, N.C., Schollett, M.A., Rier, S.T. & Geddes, P. 2006. Differential heterotrophic utilization of organic compounds by diatoms and bacteria under light and dark conditions. *Hydrobiologia*. 561:167–77.
- Tyler, B.M., Tripathy, S., Zhang, X., Dehal, P., Jiang, R.H.Y., Aerts, A., Arredondo, F.D. et al. 2006. *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science*. 313:1261–6.
- Waller, R.F. & McFadden, G.I. 2005. The apicoplast: A review of the derived plastid of apicomplexan parasites. *Curr. Issues Mol. Biol.* 7:57–79.
- Wyman, S.K., Jansen, R.K. & Boore, J.L. 2004. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics*. 20:3252–5.
- Yan, D., Wang, Y., Murakami, T., Shen, Y., Gong, J., Jiang, H., Smith, D.R. et al. 2015. *Auxenochlorella protothecoides* and *Prototheca wickerhamii* plastid genome sequences give insight into the origins of non-photosynthetic algae. *Sci. Rep.* 5:14465.
- Yu, M., Ashworth, M.P., Hajrah, N.H., Khiyami, M.A., Sabir, M.J., Alhebshi, A.M., Al-Malki, A.L. et al. 2018. Evolution of the plastid genomes in diatoms. *Adv. Bot. Res.*

Table 1-1. Taxa and sources of DNA sequences analyzed in this study.

| Genus | Species | Culture ID | Apoch- lorotic | cob | nad1 | lsu | 16s |
|----------------------|-----------------------|---------------|-------------------|-----------------------|-----------------------|----------|----------|
| Nitzschia | sp. | ccmp2144 | | MH017326 | MH017300 | MH017352 | |
| Nitzschia | dippelii | AJA014-6 | | MH017327 | MH017301 | MH017353 | |
| Hantzschia | spectabilis | UTEX FD269 | | MH017328 | MH017302 | MH017354 | |
| Hantzschia | elongata | UTEX FD421 | | MH017329 | MH017303 | MH017355 | |
| Hantzschia | amphioxys | AJA013-11 | | MH017330 | MH017304 | MH017356 | |
| Fragilariopsis | cylindrus | ccmp1022 | | JGI proj ID: 16035 | JGI proj ID: 16036 | | |
| Pseudo- nitzschia | sp. | ccmp1447 | | MH017331 | MH017305 | MH017357 | |
| Nitzschia | frustulum | ccmp1532 | | MH017332 | MH017306 | MH017358 | |
| Nitzschia | cf. ovalis | ccmp1118 | | MH017333 | MH017307 | MH017359 | |
| Cylindrotheca | closterium | ccmp1855 | | MG271845 | MG271845 | | |
| Psammodictyon | constrictum | ccmp576 | | MH017334 | MH017308 | MH017360 | |
| Nitzschia | laevis | ccmp1092 | | MH017335 | MH017309 | MH017361 | |
| Nitzschia | sp. | OHI12.10 | | MH017336 | MH017310 | MH017362 | |
| Nitzschia | cf. frequens | ccmp1500 | | MH017337 | MH017311 | MH017363 | |
| Nitzschia | pusilla | ccmp2526 | | MH017338 | MH017312 | MH017364 | |
| Nitzschia | cf. pusilla | ccmp581 | | MH017339 | MH017313 | MH017365 | |
| Nitzschia | sp. | ccmp2533 | | MH017340 | MH017314 | MH017366 | |
| Nitzschia | draveiliensis | AJA010-17 | | MH017341 | MH017315 | MH017367 | |
| Nitzschia | palea var. debilis | AJA013-2 | | MH017342 | MH017316 | MH017368 | |
| Nitzschia | gracilis | AJA010-53 | | MH017343 | MH017317 | MH017369 | |
| Nitzschia | capitellata | AJA012-22 | | MH017344 | MH017318 | MH017370 | |
| Nitzschia | alba | ccmp2426 | yes | MH017345 | MH017319 | MH017371 | |
| Nitzschia | leucosigma | ccmp2197 | yes | MH017346 | MH017320 | MH017372 | MH017379 |
| Nitzschia | sp. | ccmp579 | yes | MH017347 | MH017321 | MH017373 | |
| Nitzschia | sp. | Nitz2 | yes | MH017348 | MH017322 | MH017377 | |
| Nitzschia | sp. | Nitz4 | yes | MH017349 | MH017323 | MH017375 | MH017378 |
| Nitzschia | sp. | Nitz7 | yes | MH017350 | MH017324 | MH017376 | |
| Nitzschia | sp. | Nitz8 | yes | MH017351 | MH017325 | MH017374 | MH017380 |

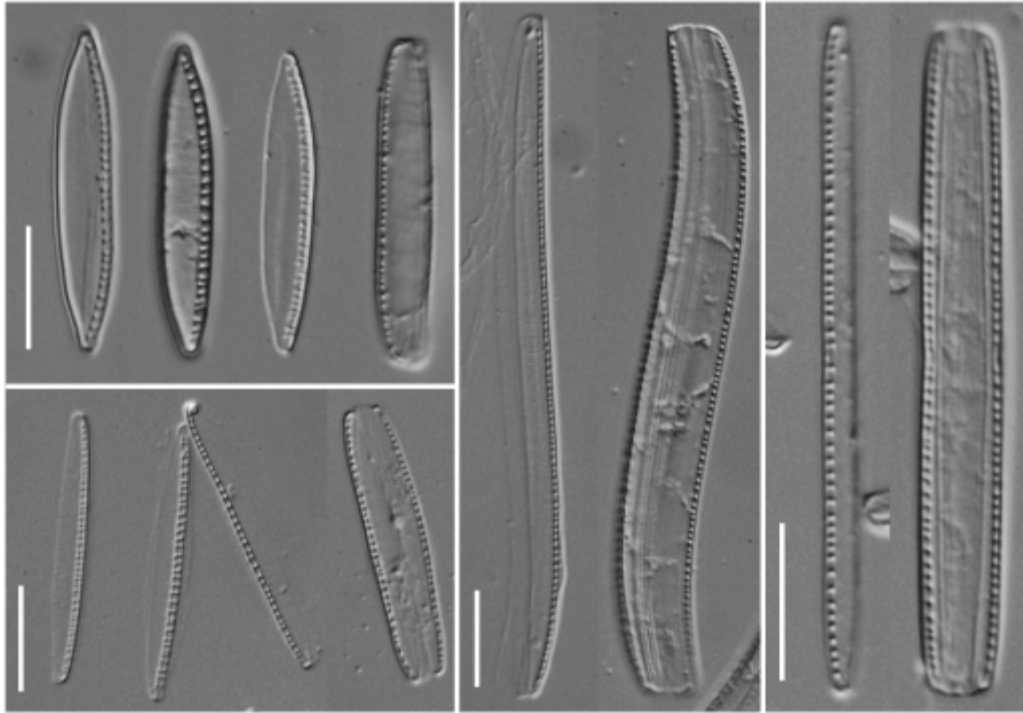


Figure 1-1. Light micrographs of newly sequenced nonphotosynthetic *Nitzschia* species. Clockwise from top: *Nitzschia* sp. nitz2, *Nitzschia* sp. nitz7, *Nitzschia* sp. nitz8, *Nitzschia* sp. nitz4.

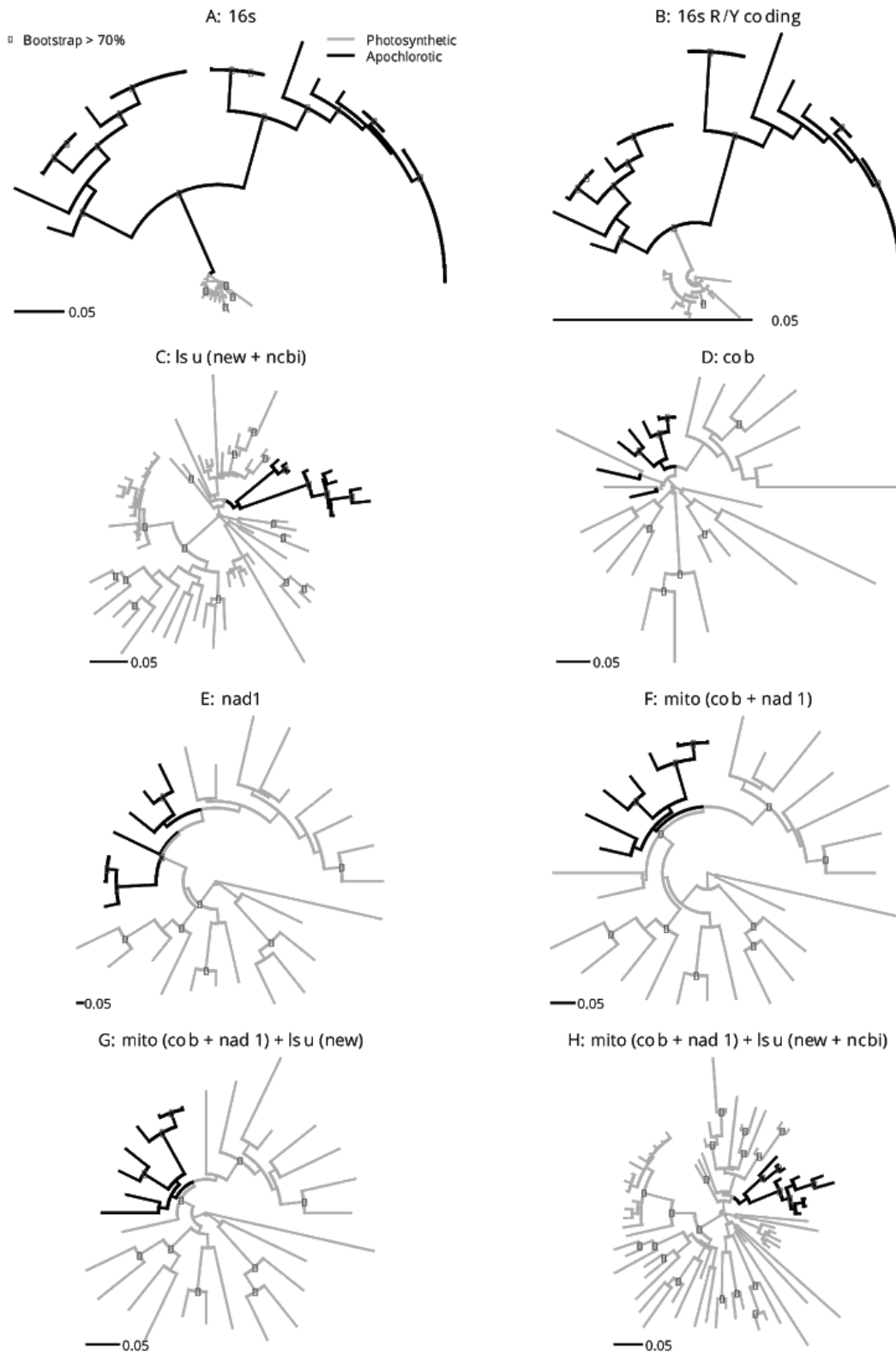


Figure 1-2. Phylogenetic trees inferred from plastid 16S (A), plastid 16S transformed into purine/pyrimidine (R/Y) coding (B), nuclear 28S d1–d2 genes for all newly sequenced taxa and

data from GenBank [”lsu (new + ncbi)”, C], mitochondrial cob (D), mitochondrial nad1 (E), concatenated cob and nad1 genes [”mito (cob + nad1)”, F], concatenated cob, nad1, and nuclear 28S d1–d2 genes for all newly sequenced taxa [”mito (cob + nad1) + lsu (new)”, G], and a concatenated alignment of cob, nad1, and nuclear 28S d1–d2 genes for taxa from this study and GenBank [”mito (cob + nad1) + lsu (new + ncbi)”, H]. For the phylogenies in panels C and H, we removed branches shorter than 0.00001 units for clarity. The full phylogenies are available in Supplementary Figure 1. Thicker black branches correspond to apochlorotic taxa. White points identify nodes with bootstrap support >70%.

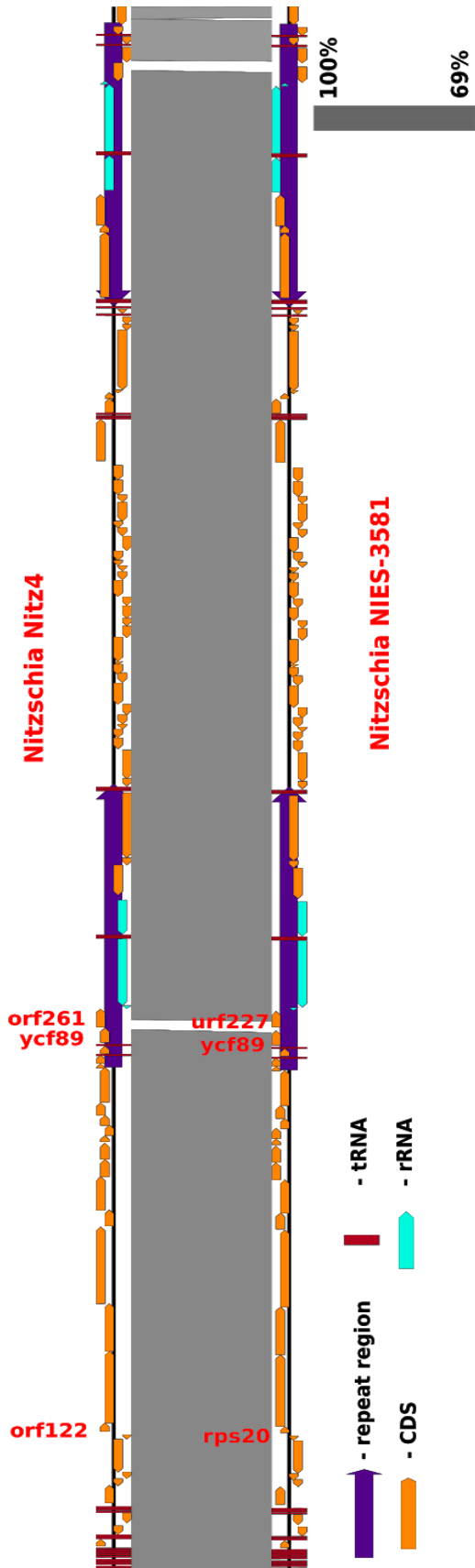


Figure 1-3. Conserved synteny, gene content, and sequence in the plastid genomes of two nonphotosynthetic diatoms in the genus *Nitzschia*, *Nitz4* (this study) and *NIES-3581* (Kamikawa et al. 2015a). Unlabeled genes are shared between the two species.

Table S1-1. Primers used to amplify and sequence *cob* and *nad1* fragments for study taxa.

| Primer Name | Primer Sequence (5'-3') |
|------------------------|--------------------------------|
| CobnF | GGAGTTTYGGYTCTTTAGCWGG |
| CobnR | GGHARAAAATACCACTCWGGVAC |
| Nad1nF ^a | ATGGGAGCAATYCAAAGRCG |
| Nad1nR ^a | CATTTCCAACCTAAATACATTAATTG |
| Nad1F-mod ^b | AAAGACGACGAGGWCCWAATGTKATAGGTT |
| Nad1R-mod ^b | CATTAATTGGTCATAYCGGTAWCGWGG |

^a Forward and reverse primers for initial amplification reaction.

^b Forward and reverse primers for second amplification reaction when nested PCR was performed.

Table S1-2. Overlapping and adjacent genes in the plastid genomes of *Nitzschia* sp. nitz4 and *Nitzschia* sp. NIES-3581. Groups of overlapping or adjacent genes are separated by empty rows with the coordinates showing the degree of overlap. All but two groups are shared between the two genomes.

| <i>Nitzschia</i> sp. Nitz4 | | | | <i>Nitzschia</i> sp. NIES-3581 | | | |
|-----------------------------------|---------------------------|---------------------|-----------------|---------------------------------------|---------------------------|---------------------|-----------------|
| Gen name | Gene coordinates | Overlap ping | Adjacent | Gene name | Gene coordinates | Overlap ping | Adjacent |
| rpl19 | Complement (3679..3993) | | | rpl19 | Complement (3681..3995) | | |
| orf96 | Complement (3986..4276) | + | | urf97 | Complement (3970..4260) | + | |
| atpB | Complement (4267..5694) | + | | | | | |
| | | | | | | | |
| orf122 | 6041..6409 | | | | | | |
| rpoB | 6406..9513 | + | | | | | |
| | | | | | | | |
| rpoC1 | 9517..11580 | | | rpoC1 | 9410..11479 | | |
| rpoC2 | 11577..14900 | + | | rpoC2 | 11476..14820 | + | |
| | | | | | | | |
| rps2 | 14938..15627 | | | rps2 | 14858..15547 | | |
| sufB | 15624..17072 | + | | sufB | 15544..16992 | + | |
| sufC | 17072..17815 | | + | sufC | 16992..17735 | | + |
| | | | | | | | |
| atpG | 18859..19311 | | | atpG | 18777..19229 | | |
| atpF | 19298..19732 | + | | atpF | 19216..19650 | + | |
| | | | | | | | |
| atpA | 20300..21805 | | | atpA | 20206..21720 | | |
| rpl35 | 21783..21977 | + | | rpl35 | 21714..21908 | + | |
| | | | | | | | |
| rps12 | Complement (36097..36471) | | | rps12 | Complement (36024..36398) | | |
| rpl31 | Complement (36471..36680) | | + | rpl31 | Complement (36398..36604) | | + |

Table S1-2. (Cont.)

| Nitzschia sp. Nitz4 | | | | Nitzschia sp. NIES-3581 | | | | |
|----------------------------|---------------------------|---------------------|-----------------|--------------------------------|------------------|---------------------------|---------------------|-----------------|
| Gen name | Gene coordinates | Overlap ping | Adjacent | | Gene name | Gene coordinates | Overlap ping | Adjacent |
| rps9 | Complement (36687..37091) | | | | rps9 | Complement (36612..37016) | | |
| rpl13 | Complement (37088..37537) | + | | | rpl13 | Complement (37013..37462) | + | |
| rpoA | Complement (37534..38466) | + | | | rpoA | Complement (37459..38391) | + | |
| | | | | | | | | |
| rps11 | Complement (38487..38891) | | | | rps11 | Complement (38412..38822) | | |
| rps13 | Complement (38830..39273) | + | | | rps13 | Complement (38755..39198) | + | |
| | | | | | | | | |
| secY | Complement (39409..40455) | | | | secY | Complement (39334..40398) | | |
| rps5 | Complement (40427..40930) | + | | | rps5 | Complement (40370..40873) | + | |
| | | | | | | | | |
| rps17 | Complement (43317..43556) | | | | rps17 | Complement (43288..43527) | | |
| rpl29 | Complement (43546..43782) | + | | | rpl29 | Complement (43517..43756) | + | |
| | | | | | | | | |
| | | | | | dnaK | 48083..49930 | | |
| | | | | | tRNA-Arg | 49930..50003 | | + |
| | TOTAL: | 12 | 2 | | | TOTAL: | 10 | 3 |

A: 16s

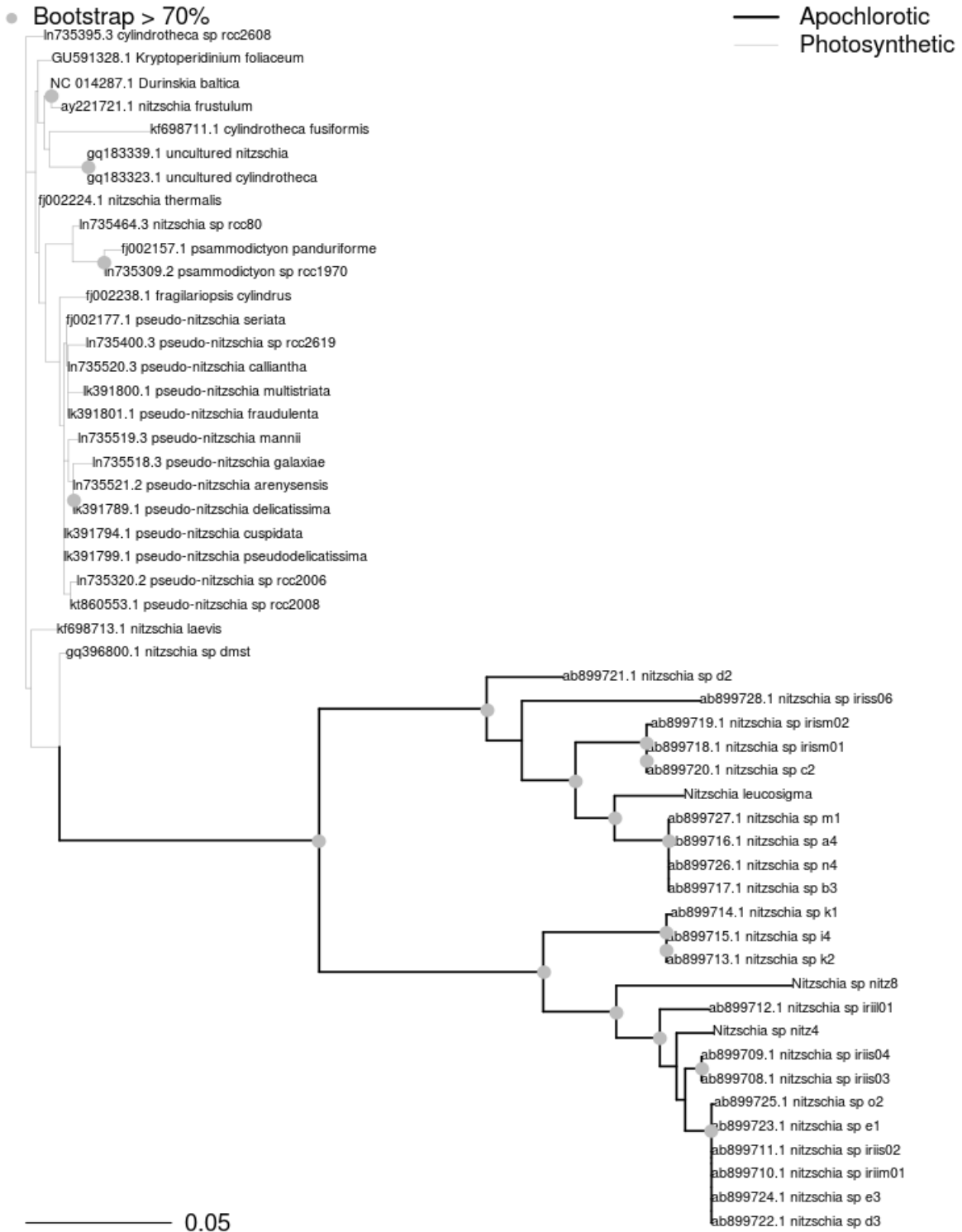


Fig S1-1. (Cont.)

B: 16s R/Y coding

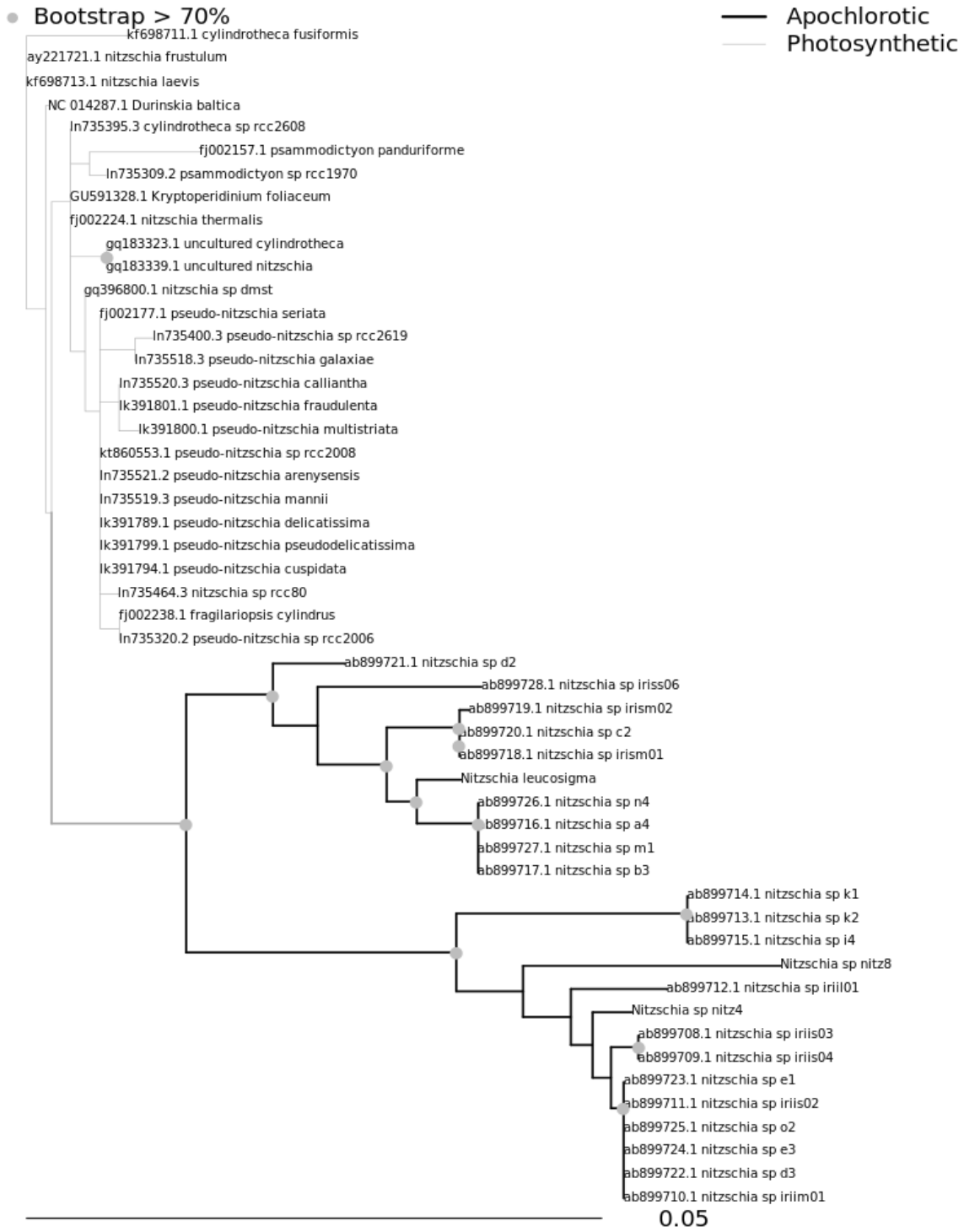


Fig S1-1. (Cont.)

C: cob

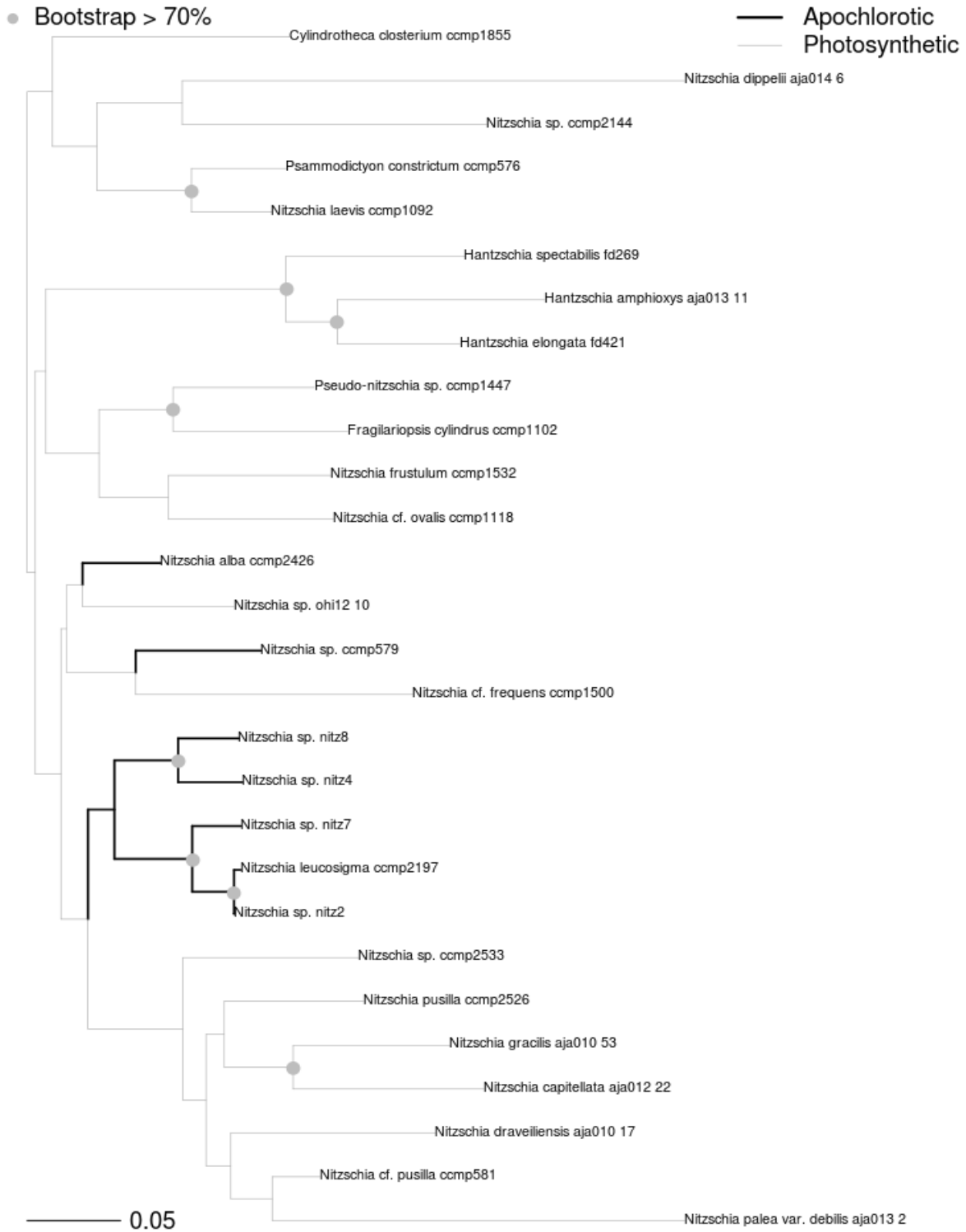


Fig S1-1. (Cont.)

D: nad1

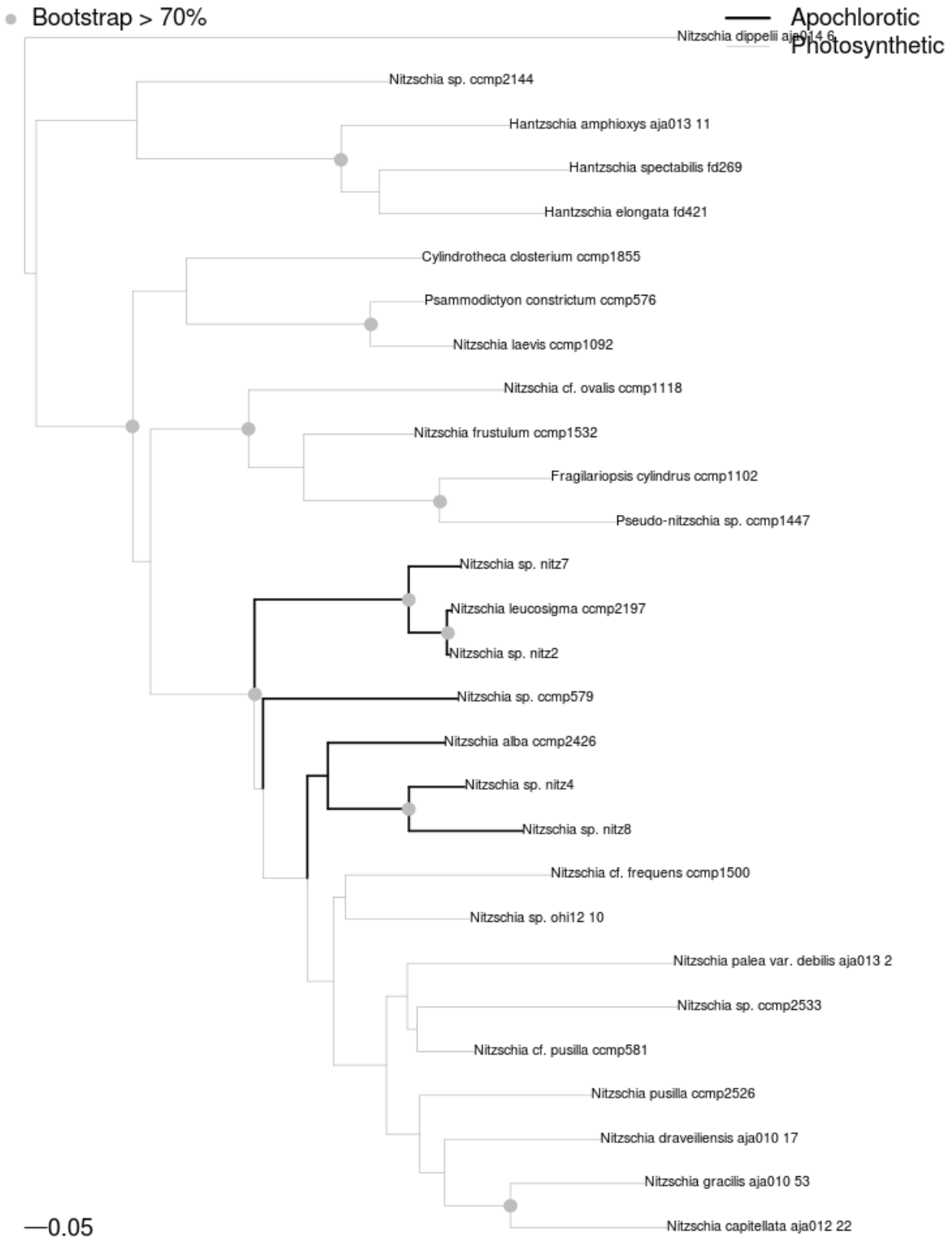


Fig S1-1. (Cont.)

E: Isu (new + ncbi)

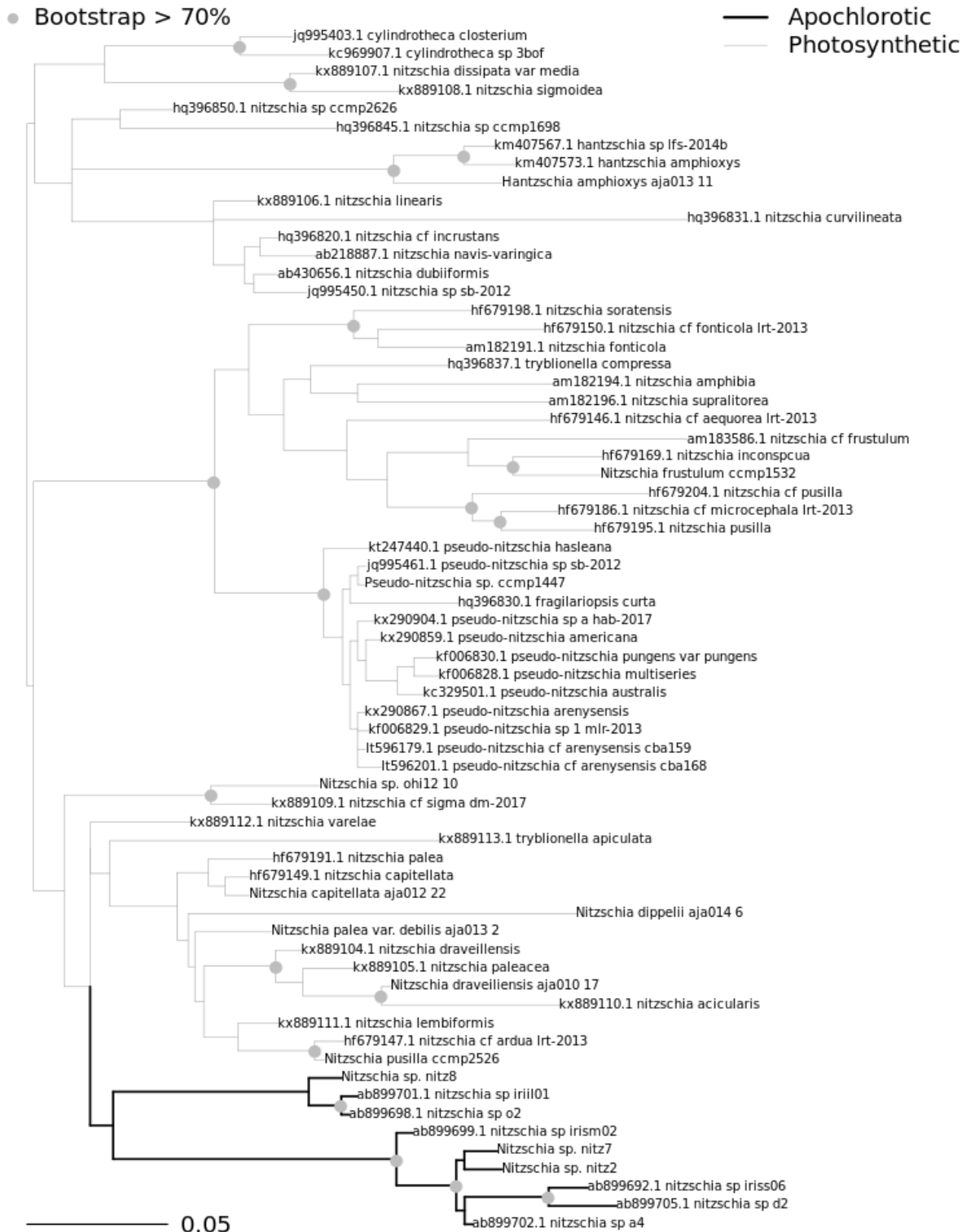


Fig S1-1. (Cont.)

F: mito (cob + nad1)

● Bootstrap > 70%

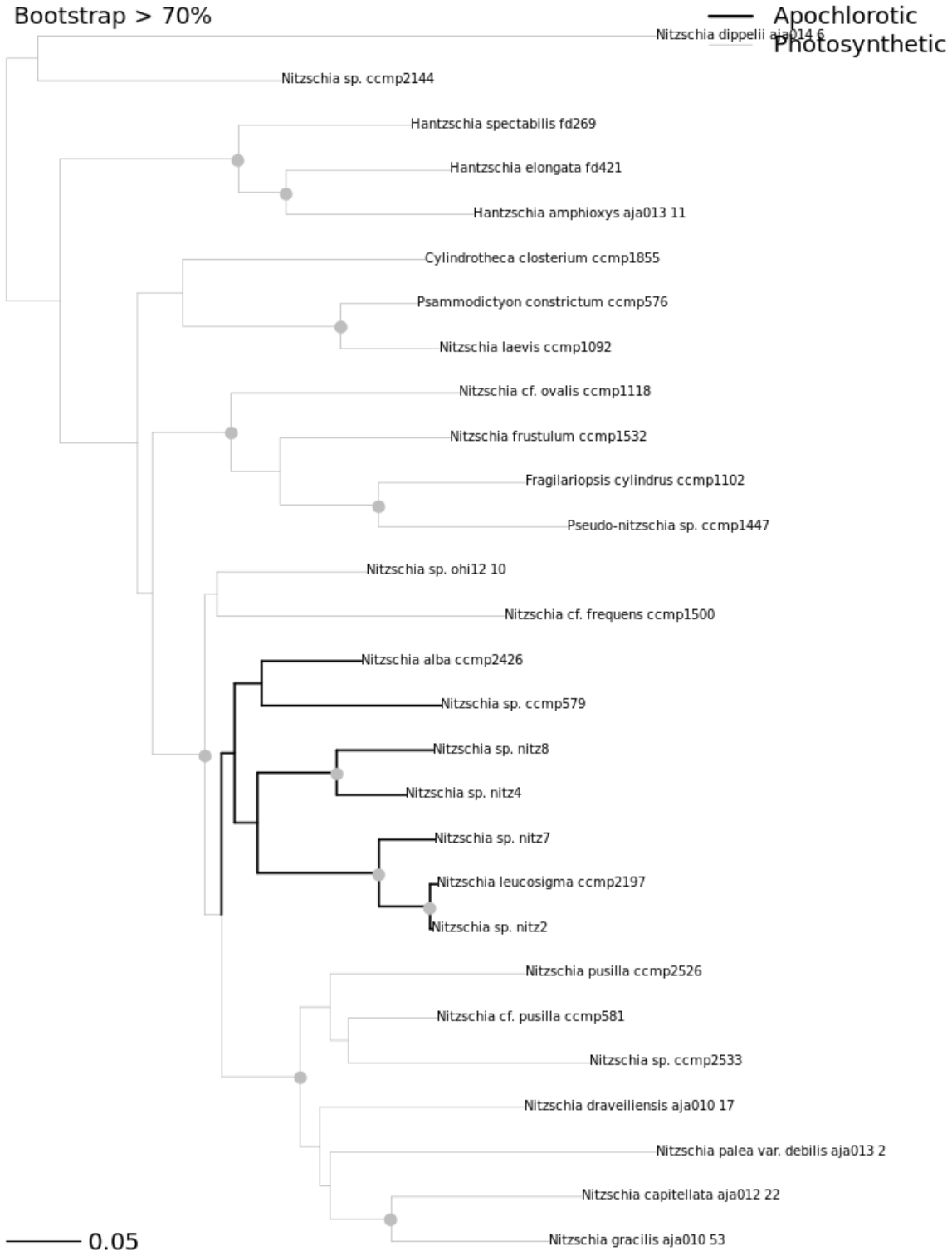


Fig S1-1. (Cont.)

G: mito (cob + nad1) + Isu (new + ncbi)

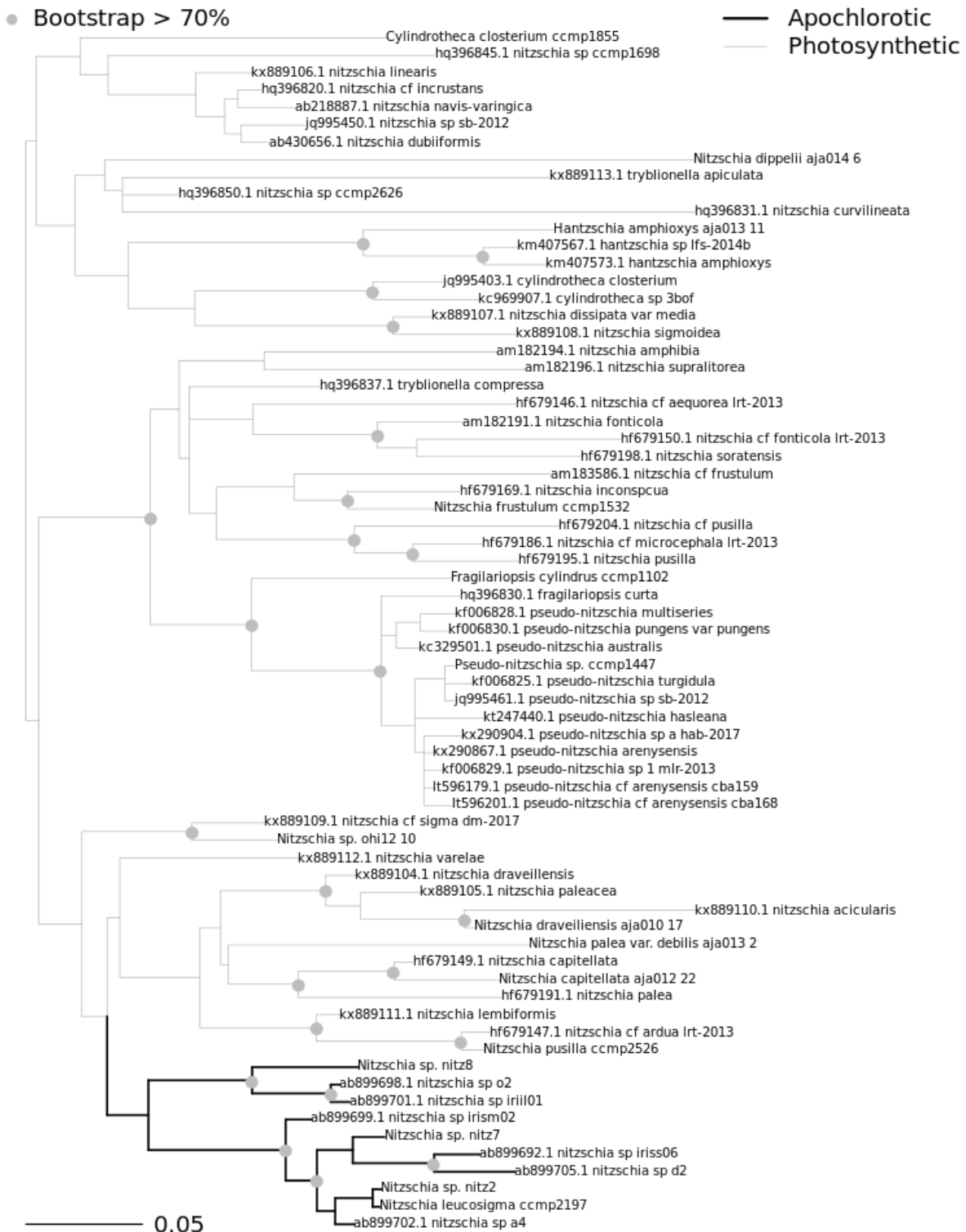


Fig S1-1. (Cont.)

Figure S1-1. Phylogenetic trees for plastid 16S (A), plastid 16S transformed into purine/pyrimidine (R/Y) coding (B), mitochondrial cob (C), nad1 (D), a densely sampled 28S d1–d2 matrix with sequences from this study and GenBank (E), cob and nad1 combined (F), and a large combined nuclear and mitochondrial gene matrix with sequences from this study and GenBank (G). In the 16S tree, *Nitzschia* sp. NIES-3581 is identified by its synonym, *iriis04*.

Chapter 2. Core carbon metabolism and characterization of a β -ketoacid pathway inferred from the genome of a non-photosynthetic diatom (Bacillariophyta)

Abstract

Although most of the tens of thousands of diatom species are photoautotrophs, many mixotrophic species can also use extracellular sources of organic carbon, and one small lineage of obligate heterotrophs has lost the ability to photosynthesize and requires extracellular organic carbon for growth. We lack an understanding of the exact sources of carbon used by these species, however. We sequenced the genome of a non-photosynthetic diatom, *Nitzschia* sp., and used it to develop a comprehensive model of carbon metabolism in these species. The genome is relatively small (31 Mbp), gene dense, and contains a β -ketoacid pathway, a pathway known mostly from bacteria and fungi. The β -ketoacid pathway potentially allows diatoms to metabolize lignin-derived aromatic compounds, the products of which are predicted to be delivered to mitochondria and fed directly into the Krebs cycle. Genes in this pathway are also present in genomes of some photosynthetic diatoms, suggesting that this mode of carbon utilization is an ancestral feature of diatoms, allowing them to exploit derivatives of one of the most abundant sources of organic carbon on the planet.

Introduction

Diatoms are microscopic unicellular algae that are widely distributed throughout marine and fresh waters. They are prolific photosynthesizers responsible for some 20% of global primary production (Field et al. 1998, Smetacek 1999). Diatom plastids trace back to a secondary endosymbiosis between a bi-flagellated host—the common ancestor of the chromalveolate clade—and a red algal endosymbiont (Archibald 2009). As a result, diatom plastids are surrounded by four membranes, including two relic primary plastid membranes, which are remnants of red algal plasmalleme, and the diatom endoplasmic membrane (Stoebe and Maier 2002). The first sequenced diatom genome captured this history, revealing many genes transferred from the red algal endosymbiont into the host nuclear genome (Armbrust et al. 2004). Most functions in the diatom plastids are, in fact, carried out by proteins that are encoded in the nuclear genome and synthesized in the cytoplasm. To be properly targeted and transported into the plastid, plastid-targeted proteins must contain N-terminal bipartite presequences with signal and transit peptide domains separated by a cleavage site that contains highly conserved amino acid motif (Lang et al. 1998, Gruber et al. 2007).

Carbon metabolism is highly compartmentalized in diatoms. Diatom plastids possess a full Calvin cycle, but lack complete pentose phosphate pathway (PPP), whereas, and the lower part of the glycolytic pathway takes place in the mitochondrion. Many genes associated with primary metabolism enzymes exist in multiple copies in the genome and differ in their targeting sequences (Kroth et al. 2008, Smith et al. 2012). Models of carbon metabolism in diatoms are based on photosynthetic species, but a broad diversity of diatoms, mostly in the pennate lineage, are able to use external sources of carbon as well, probably as a means of maintaining their

growth in low light conditions (Tuchman et al. 2006). One small subclade of raphid pennate diatoms in the genus *Nitzschia* have lost the ability to photosynthesize altogether and exist now as free-living heterotrophs. The 20 or so apochloritic diatoms trace back to a single loss of photosynthesis in their common ancestor (Onyshchenko et al. 2018). These species are often found in habitats rich in organic matter, including mangrove forests (Blackburn et al. 2009, Kamikawa et al. 2015, Onyshchenko et al. 2018) and decaying seaweeds (Pringsheim 1956, Lewin and Lewin 1967), potentially providing some clues about the sources of organic carbon available to them.

We sequenced, assembled, and annotated the genome of one non-photosynthetic diatom, *Nitzschia* sp. (Nitz4) to understand how these species acquire and metabolize organic carbon for energy. Apochloritic diatoms also maintain colorless plastids with highly reduced plastid genomes (Kamikawa et al. 2016, Onyshchenko et al. 2018). A model of carbon metabolism in these plastids suggests that this organelle maintains a high degree of carbon metabolic activity and biochemical complexity even in the absence of a photosynthetic apparatus (Kamikawa et al. 2017). The goal of our study was to use the *Nitzschia* sp. Nitz4 genome to identify possible sources of exogenous carbon and to develop a comprehensive, whole-cell model of carbon metabolism in nonphotosynthetic diatoms.

Materials and Methods

Collection and culturing of Nitzschia sp.

We collected a composite sample on 10 November 2011 from Whiskey Creek, which is located in Dr. Von D. Mizell-Eula Johnson State Park (formerly John U. Lloyd State Park), Dania

Beach, Florida, USA (26.081330 latitude, -80.110783 longitude). The sample consisted of near-surface plankton collected with a 10 μ M mesh net, submerged sand (1 m and 2 m depth), and nearshore wet (but unsubmerged) sand. We selected for non-photosynthetic species by storing the sample in the dark at room temperature (21 °C) for several days before isolating colorless diatom cells with a Pasteur pipette. Clonal cultures were grown in the dark at 21 °C on agar plates made with L1+NPM medium (Guillard 1960, Guillard and Hargraves 1993) and 1% Penicillin–Streptomycin–Neomycin solution (Sigma-Aldrich P4083) to retard bacterial growth.

DNA and RNA extraction and sequencing

We rinsed cells with L1 medium and removed them from agar plates by pipetting, lightly centrifuged them, and then disrupted them with a MiniBeadbeater-24 (BioSpec Products). We extracted DNA with a Qiagen DNeasy Plant Mini Kit. We sequenced total extracted DNA from a single culture strain, *Nitzschia* sp. Nitz4, with the Illumina HiSeq2000 platform housed at the Beijing Genomics Institute, with a 500-bp library and 90-bp paired-end reads. We separately extracted total RNA with a Qiagen RNeasy kit, and sequenced a 300-bp Illumina TruSeq RNA library using the the Illumina HiSeq2000 platform.

Genome and transcriptome assembly and annotation

A total of 15.4 GB of pair-end DNA reads were recovered and used to assemble the Nitz4 nuclear, plastid, and mitochondrial genomes. We used FastQC (ver. 0.11.5) (Andrews 2010) to check read quality then used ACE (Sheikhzadeh and de Ridder 2015) to correct predicted sequencing errors. We subsequently trimmed and filtered the reads using Trimmomatic (ver. 0.32) (Bolger et al. 2014) with settings “ILLUMINACLIP=<TruSeq_adapters.fa>:2:40:15, LEADING=2 TRAILING=2, SLIDINGWINDOW=4:2, MINLEN=30, TOPHRED64.” We

assembled the genome using RAY (ver. 2.3.1) (Boisvert et al. 2012) with a range of k-mer lengths (15, 21, 27, 33, 39, 45, 51, 59, 65) to determine which k-mer length maximized N50, maximum scaffold length, and total assembly length. Based on these criteria, the assembly based on a k-mer length of 45 was judged to be best (Fig. S2-1). We then used bowtie2 (ver. 2.2.8) (Langmead and Salzberg 2012) to remove reads that mapped to the previously assembled plastid genome (Onyshchenko et al. 2018). The remaining nuclear and mitochondrial reads were reassembled following the strategy outlined above with a k-mer length of 45. To identify mitochondrial contigs, we used NCBI-BLASTN to search all of the assembled scaffolds against a database of diatom mitochondrial genes. This search returned a single mitochondrial scaffold. We then mapped and removed all mitochondrial reads as described above for the plastid genome. The remaining set of organelle-filtered reads were assembled again with RAY and range of k-mer lengths, and the assembly that used a k-mer length of 45 was again the best one (Fig. S2-2). We used Blobtools (ver. 1) (Laetsch and Blaxter 2017) to visualize the assembly and verify that the final nuclear assembly was free of organelle and contaminating bacterial contigs. We then used AGOUTI (ver 0.3.3) (Zhang et al. 2016) to scaffold together contigs that could be joined based on paired-end information in the RNA-seq reads. The AGOUTI step used the fourth Maker-based genome annotation (see “Annotation” methods below).

We used REAPR (Hunt et al. 2013) with the smalt map aligner and default parameter settings to check the quality of our *de novo* nuclear genome assembly. REAPR uses read-mapping data to assess read and read-pair depth, as well as read-pair conflicts, across the genome, breaking potentially misassembled contigs that have, for example, local concentrations of read-pair conflicts or coverage values that fall outside of an expected distribution. We found

that original, unbroken assembly did not differ critically from the REAPR-produced assembly, so we proceeded with original assembly.

RNA extraction, sequencing, read processing, and assembly of RNA-seq data followed Parks and Nakov et al. (2017).

Genome annotation

We used the Maker pipeline (ver. 2.31.8) (Cantarel et al. 2008) to annotate the Nitz4 nuclear genome. In order to successively improve the annotation quality, we ran Maker a total of six times with default settings unless stated otherwise. We used the assembled Nitz4 transcriptome (Maker's expressed sequence tag [EST] evidence) and the predicted proteome of *Fragilariopsis cylindrus* (GCA_001750085.1) (Maker's protein homology evidence) to inform the annotation. We used RepeatMasker (ver. open-4.0.7) (Chen 2004) to generate a repeat library for Nitz4, which was also used by Maker for the annotation. We used Augustus (ver. 3.2.2) for *de novo* gene prediction with settings "max_dna_len = 200,000" and "min_contig = 300." We trained Augustus using the set of annotated genes from *Phaeodactylum tricornutum* (ver. 2), which we filtered as follows: (1) remove "hypothetical" and "predicted" proteins, (2) remove all but one splice variant of a gene, (3) remove genes with no introns, and (4) remove genes that overlap with neighboring gene models or the 1000 bp flanking regions of adjacent genes, as these regions are included in the training set along with the enclosed gene. If possible, we generated UTR annotations for the selected gene models by subtracting 5' and 3' CDS coordinates from the corresponding 5' and 3' end coordinates of the associated mRNA sequence; among these, we only retained those with UTRs that extended ≥ 25 bp beyond both the 5' and 3' ends of the CDS. The final filtered training set included a total of 726 genes.

We generated a separate set of gene models for UTR training by using all of the filtering criteria outlined above except that we retained intronless genes, as we assumed intron presence or absence was less relevant for training the UTR annotation parameters. In addition, we retained only those genes with UTRs that were ≥ 40 bp in length. In total, our UTR training set included 531 genes. We trained and optimized the Augustus gene prediction parameters (with no UTRs) on the first set of 726 genes and optimized UTR prediction parameters with the second set of 531 genes. We then used these parameters to perform a series of six successive gene predictions within Maker. We assessed the first Maker annotation for completeness using BUSCO (ver. 2.0) (Simão et al. 2015) with the protist database (Table 2-1). Following recommendations of the developers, for the five subsequent Maker runs, we used a different *ab initio* gene predictor, SNAP (Korf 2004), trained with Maker-generated gene models from the previous run. Although the fifth SNAP-based Maker run discovered more genes, the number of complete BUSCOs decreased by one, so we discarded this assembly and used the previous one (the fifth Maker run, the fourth SNAP-based run) as the final assembly.

Scripts used for genome assembly, annotation and MAKER gene training set search are available at https://github.com/Nastassia/Nitz4_annot_methods_clean.git.

Construction of orthologous clusters

We used Orthofinder (ver. 1.1.4) (Emms and Kelly 2015) with default parameters to build orthologous clusters from the complete set of predicted proteins from the genomes of Nitz4, *Fragilariopsis cylindrus* (GCA_001750085.1), *Phaeodactylum tricornutum* (GCA_000150955.2), *Cyclotella nana* (GCA_000149405.2), and the transcriptome of another *Nitzschia* species (Nitz2144) (accession forthcoming). We tried to characterize species-specific

Nitz4 genes by using NCBI-BLASTP to search orthogroups that consisted exclusively of Nitz4 genes against NCBI's nonredundant (nr) database (GenBank release 223.0 or 224.0).

Characterization of carbon metabolism genes

We performed a thorough manual annotation of primary carbon metabolism genes in the Nitz4 genome. We started with a set of core carbon metabolic genes in diatoms (e.g., Smith et al. 2012, Kamikawa et al. 2017), downloaded all genes from GenBank's nr database (release 223.0 or 224.0) using the annotation as the search term, and filtered them to include just RefSeq accessions. As we characterized carbon metabolic pathways, we expanded our searches as necessary to include genes that were predicted to be involved in those pathways. We constructed local BLAST databases of Nitz4 genome scaffolds, predicted proteins, and assembled transcripts and used the NCBI-BLAST (ver. 2.6.0) package to separately search the set of sequences for a given carbon metabolism gene against the three Nitz4 databases. In most cases, each gene had thousands of annotated sequences on GenBank, so if the GenBank sequences for a given gene were, in fact, homologous and accurately annotated, then we expected Nitz4 homologs to likewise return thousands of strong hits, and this was often the case. We did not consider further any putative carbon genes with relatively few query hits (e.g., tens of weak hits vs. the thousands of strong hits in verified, correctly annotated sequences). We extracted the Nitz4 subject matches and used NCBI-BLASTX or BLASTP to search the transcript or predicted protein sequence against NCBI's nr protein database to verify the annotation. Nitz4 subject matches in scaffold regions were checked for overlap with annotated genes, and if the subject match did overlap with an annotated gene, the largest annotated CDS in this region was extracted and BLASTed to the nr

protein database to verify the annotation. As necessary, we repeated this procedure for other diatom species used to construct clusters of orthologous genes.

Prediction of protein localization

We used the software programs SignalP (Petersen et al. 2011), ASAFind (Gruber et al. 2015), ChloroP (Emanuelsson et al. 1999), TargetP (Emanuelsson et al. 2007), MitoProt (Claros 1995), and HECTAR (Gschloessl et al. 2008) to predict whether proteins were targeted to the plastid, mitochondrion, cytoplasm, or endoplasmic reticulum (ER) for the secretory pathway. Our approach was slightly modified from that of Traller et al. (2016). Proteins with a SignalP-, ASAFind, and HECTAR-predicted plastid signal peptide were classified as plastid targeted. We placed less weight on ChloroP target predictions, which are optimized for targeting to plastids bound by two membranes and not the four-membrane plastids found in diatoms. Proteins predicted as mitochondrial-targeted by any two of the TargetP, MitoProt, and HECTAR programs were classified as localized to the mitochondrion. Those proteins in the remaining set were classified as ER targeted if they had ER signal peptides predicted by SignalP and HECTAR. Most of the remaining proteins were cytoplasmic.

Results and Discussion

Genome characteristics

A total of 74,996,078 100-bp paired-end DNA reads assembled into 4,447 nuclear scaffolds totaling 30.4 Mbp in length and with an average read depth of 180x. AGOUTI joined 350 contigs that were broken within transcribed regions, reducing the number of scaffolds to 4,097 and increasing the assembly size to 30.7 Mbp in length. Half of the Nitz4 genome is

contained within 195 scaffolds, each one longer than 43,578 bp (Table 2-2, N50). The average scaffold length is 7,507 bp, and the single largest scaffold is 340,060 bp in length. The Nitz4 genome is similar in size to other small diatom genomes (Table 1), but it is the smallest so far sequenced from Bacillariales, the lineage that also includes *F. cylindrus* (61 Mbp) and *Pseudo-nitzschia multistriata* (55 Mbp) (Table 2-2). The genomic GC content is 47.8%, which is similar to most other diatoms (Table 2-2).

The fifth iteration of Maker resulted in 9,235 gene models, the vast majority (9,017) of which were supported by protein or EST evidence. The gene models included a total of 209 of 234 protist BUSCOs (Table 2-1). Although Nitz4 has fewer genes than all other diatoms sequenced to date, it also has the largest BUSCO count, suggesting that the smaller overall number of genes is not due to incomplete sequencing of the genome (Table 2-1). Genes models from *C. nana*, *F. cylindrus*, *Nitzschia* sp. 2144, and Nitz4 clustered into 10,834 multi-sequence (≥ 2 gene models) orthogroups. A total of 8,304 of the Nitz4 genes were assigned to orthogroups, and 21 genes fell into six Nitz4-specific orthogroups. The remaining genes did not cluster with any other sequences, either from Nitz4 or another species.

Central carbon metabolism

Glycolysis, gluconeogenesis, and pyruvate metabolism—Glycolysis is a core cytosolic pathway for degradation of glucose to pyruvate, which can then be targeted to mitochondria for ATP production and conversion into a range of other compounds that can, in turn, be used for further energy production. Through gluconeogenesis, several pyruvate metabolizing enzymes convert pyruvate back into hexoses when the energy demands of the cell have been met, or when necessary, into glycolysis carbon intermediates for anabolism. In diatoms, including Nitz4, the

latter half of the glycolytic pathway can occur in both the cytosol and mitochondria, and part of the pathway can be carried out in the plastid (Kroth et al. 2008, Smith et al. 2012). Overall, Nitz4 contains fewer copies of genes encoding glycolytic enzymes compared to other diatoms. These include the plastid-targeted FBP, FBI, TPI genes and cytosolic GADPH and PGK genes (Fig. S2-4). The first enzyme involved in glycolysis, glucokinase (GK), exists as a single copy in the Nitz4 genome and has no targeting signal, indicating that the first step of glycolysis likely occurs in the cytosol. Similar to other diatoms, we did not find a hexokinase gene in Nitz4. The second glycolytic enzyme, glucose-6-phosphate isomerase (GPI), was predicted to be localized to both the plastid and the cytosol, though none of the three phosphofructokinase-1 (PFK) genes possessed plastid targeting signals. Enzymes for subsequent steps in glycolysis, fructose-bisphosphate aldolase (FBA) and pyruvate kinase (PK), are localized to both the cytosolic and plastid compartments, as predicted in other diatoms as well (Kroth et al. 2008, Smith et al. 2012, Traller et al. 2016).

As in other diatoms (Smith et al. 2012), enzymes catalyzing reactions in the latter half of glycolysis (from triosephosphate isomerase [TPI] to pyruvate kinase [PK]) were also predicted to be localized to mitochondria (Fig. S2-3), which appears to be a conserved feature across diatoms (Kroth et al. 2008, Smith et al. 2012). The Nitz4 genome contains two copies of phosphoglycerate kinase (PGK), which catalyzes the seventh reversible step of glycolysis and has a possible role in gluconeogenesis. Surprisingly, however, neither copy contains a mitochondrial targeting signal—instead, one copy is predicted to be plastid-targeted and the other is cytosolic. We verified the absence of a mitochondrial PGK through careful manual searching of both the scaffolds and the transcriptome. Lack of a mitochondrial PGK suggests that

Nitz4 either lacks half of the mitochondrial glycolysis/gluconeogenesis pathway or that one or both of the plastid and cytosolic PGK genes are, in fact, dual-targeted to the mitochondrion. In the true, and seemingly unlikely absence, of a mitochondrial-targeted PGK, the mitochondrion would be missing a single critical enzyme for glycolysis and gluconeogenesis. We did not find mitochondrial phosphoenolpyruvate (PEP) exporters, which would suggest that this intermediate compound is shuttled into a cytosolic gluconeogenic pathway after conversion from pyruvate.

We then characterized the presence and localization of genes in the so-called “pyruvate hub” (Smith et al. 2012), which describes the set of genes involved in conversion of pyruvate for reuse in gluconeogenesis or other metabolic pathways (Fig. S2-3). Overall, the “pyruvate hub” in Nitz4 is reduced by comparison to photosynthetic diatoms both in the number of different enzymes present in the genome and in protein localization, which is restricted mostly to the mitochondrial compartment in Nitz4. Among the possible enzymes involved in pyruvate conversion to phosphoenolpyruvate (PEP), Nitz4 possesses only the pyruvate carboxylase (PC) and PEPCK enzymes, which convert pyruvate through an intermediate compound, oxaloacetate (OAA). Both enzymes, which are encoded by single copy genes, were predicted to be localized to the mitochondrion. Another unidirectional gluconeogenesis enzyme, fructose biphosphatase (FBP), is targeted to the cytosol and plastids, as predicted for other diatoms (Kroth et al. 2008, Smith et al. 2012). Nitz4 is missing glucose-6-phosphatase (G6Pase), which is responsible for the third committed gluconeogenic reaction and, as a result, Nitz4 cannot produce free glucose in this pathway.

Like other diatoms (Kroth et al. 2008, Smith et al. 2012), Nitz4 targets enzymes involved in pyruvate conversion to the mitochondrion. As previously stated, the PC and PEPCK enzymes

that convert pyruvate to PEP through OAA are restricted to mitochondrial compartment, but the reverse reaction—conversion of PEP back to OAA by phosphoenolpyruvate carboxylase (PEPC)—is carried out in the cytosol. This differs from other diatoms, which localize pyruvate carboxylase (PC) to the plastid and PEPC to the plastid/periplastidial space (Smith et al. 2012). Phosphoenolpyruvate synthetase (PEPS) and phosphate dikinase (PPDK), which can catalyze the first reaction of gluconeogenesis, are not present in Nitz4 (Fig. S2-3). Malic enzyme (ME), which is involved in pyruvate metabolism and C₄-photosynthesis in photosynthetic diatoms, is missing from Nitz4. The role of ME in diatom metabolism is unclear (Smith et al. 2012), the the apparent loss of ME in Nitz4 could be a direct consequence of the loss of photosynthesis or a secondary effect related to other metabolic changes. It is also hard to predict how the inability to convert malate directly to pyruvate might influence carbonic flux through the TCA cycle. Malate dehydrogenase enzyme (MDH) is targeted to the mitochondrion, which is expected considering its indispensable role in the TCA cycle. MDH might also function in gluconeogenesis by converting malate to OAA and, subsequently, to pyruvate in conjunction with PEPCK. MDH is targeted to the mitochondrion in Nitz4, ruling out any involvement in the Asp–Malate shuttle and, consequently, the transfer of reducing equivalents (Mikulášová 1998).

Apart from PFK, a complete set of glycolytic enzymes are targeted to the plastid in Nitz4, a feature that is conserved in other diatoms (e.g., *C. cryptica* and *F. cylindrus*) as well. In the absence of a photosynthesis-derived carbon source in the plastid, Nitz4 may import glycolysis intermediates into the plastid for further metabolism.

Detailed annotations of genes constituting *Nitzschia* sp. Nitz4 central metabolic pathways are available at Appendix 1.

A β -ketoadipate pathway in diatoms

Experimental data have shown that mixotrophic diatoms apparently can use a range of exogenous carbon sources (Hellebust and Lewin 1977, Tuchman et al. 2006). The exact source of carbon in strictly heterotrophic, nonphotosynthetic diatoms is unclear, however, and identifying the carbon species used by these diatoms was a principal goal of this study. The genome sequences allows us to compile the gene models and make such predictions based on the presence of specific carbon-related genes and pathways or, alternatively, based on the expansions of gene families related to specific modes of carbon metabolism relative to photosynthetic species. Species in the pennate diatom clade (to which *Nitz4* belongs) are mixotrophic, and mixotrophy is naturally thought to be the intermediate trophic strategy between autotrophy and heterotrophy (Figueroa-Martinez et al. 2015), so genome comparisons to other raphid pennate diatoms may conceal obvious features related to obligate vs. facultative heterotrophy. For example, major differences between mixotrophs and strict heterotrophs may involve other kinds of changes in, for example, transcription levels of or post-translational modifications.

Orthogroup clustering revealed six *Nitz4*-specific orthogroups with multiple gene models, and we could not identify or ascribe functions to most of these groups. A systematic survey of orthogroups enriched for *Nitz4* homologs (≥ 3 *Nitz4* sequences) returned one species-specific cluster (OG0006126) with five *Nitz4* genes that had strong matches to genes encoding protocatechuate dioxygenase (PCA), an enzyme involved in the degradation of aromatic compounds, specifically through intradiol ring cleavage of a hydroxylated derivative of benzoate, protocatechuic acid. Four of these genes are located on relatively long scaffolds with identifiable diatoms genes (Figure 2-1). A further search of the genomic scaffolds revealed

several additional copies of unannotated PCA enzymes, and BLASTP searches to GenBank's nr protein database returned to hits to protocatechuate 3,4-dioxygenase (P3,4O) enzymes. With no prior knowledge of the presence or role of this enzyme in diatoms, we focused a great deal of attention characterizing this gene, its role in the larger protocatechuate pathway, and the overall contribution of this pathway to central carbon metabolism in *Nitzschia*.

The major role of PCA is degradation of aromatic substrates through the β -ketoacid pathway (Stanier and Ornston 1973, Harwood and Parales 1996, Fuchs et al. 2011), which is known mostly from soil fungi and eubacteria. The β -ketoacid does not function in protocatechuate degradation *per se*, rather it serves as a funneling pathway for further metabolism of a variety of aromatic compounds commonly found in nature (Fuchs et al. 2011). Aromatic rings are a common feature of many organic compounds, and organisms degrade these compounds into a range of aromatic intermediates (e.g., protocatechuate, catechol, gentisate, and benzoyl-CoA), which are then subject to ring-cleavage reactions and further degradation through a limited number of metabolic pathways. Thus, derivatives of aromatic environmental pollutants and natural substrates, such as lignin, released by plants may be subject to catabolism through protocatechuate intermediates (Harwood and Parales 1996, Masai et al. 2007, Fuchs et al. 2011, Wells and Ragauskas 2012). Lignin is of particular interest here because it comprises a large fraction of plant biomass and is, as a result, one of the most abundant biopolymers in nature. Lignin is an unordered heterogeneous polymer consisting of aromatic rings and is highly resistant to biodegradation, requiring a suite of microbial enzymes and non-enzymatic steps for its decomposition (Janusz et al. 2017). Several lignin derived monomers, which are the products of decaying plants, are converted to protocatechuate and metabolized through one branch of the

β -keto adipate pathway (Stanier and Ornston 1973, Harwood and Parales 1996). In summary, lignin from degrading plant material represents a major source of naturally occurring aromatic compounds that could feed directly into aromatic degradation pathways, including the β -keto adipate pathway.

In addition to PCA, we discovered that Nitz4 has a nearly complete set of enzymes for the β -keto adipate pathway, including single copies of 3-carboxymuconate lactonizing enzyme (CMLE), 4-carboxymuconolactone decarboxylase (CMD), and beta-keto adipate enol-lactone hydrolase (ELH) (Table 2-3, Fig. 2-2). PCA is the first enzyme in the protocatechuic branch of the β -keto adipate pathway (Fig. 2-2) and is responsible for ortho-cleavage of the activated benzene ring between two hydroxyl groups, which results in the production of β -carboxymuconic acid (Fig. 2-2). CMLE, CMD, and ELH catalyze subsequent steps of the pathway after fission of the benzene ring (Fig. 2-2). The last two enzymes in the pathway, 3-keto adipate:succinyl-CoA transferase (TR) and 3-keto adipyl-CoA thiolase (TH), convert β -keto adipate to TCA intermediates, succinyl-CoA and acetyl-CoA, are missing from the Nitz4 genome. However, these two enzymes are highly similar to ones involved in central carbon metabolism—Succinyl-CoA:3-ketoacid CoA transferase (SCOT) and 3-ketoacyl-CoA thiolase (3-KAT), respectively (Parales and Harwood n.d., Kaschabek et al. 2002). Thus, we predict that these two core carbon metabolism enzymes substitute in for the last two steps of the β -keto adipate pathway. This hypothesis will require experimental validation, but the presence of intact genes in the upper portion of the pathway strongly suggests that an entire pathway exists, and these enzymes are the best candidates for fulfilling these roles. Moreover, both TH and TR are strongly predicted to be targeted to the mitochondrion.

Products of the two last steps of β -ketoacid pathway, succinyl-CoA and acetyl-CoA, are TCA cycle intermediates. The predicted mitochondrial localization of these steps therefore strongly suggests that these products are funneled directly into the TCA cycle, providing carbon skeletons for generation of ATP. This hypothesis requires, however, transport of β -ketoacid into the mitochondrion, a process that may involve the mitochondrial 2-oxoglutarate/malate antiporter (Zeman et al. 2016), which is usually a component of the malate/aspartate shuttling system. Both cytoplasmic and mitochondrial MDH enzymes are necessary for this shuttling system, however, and most diatoms, including *Nitzschia*, only have a mitochondrial MDH. This suggests that the 2-oxoglutarate/malate transporter in *Nitzschia* may function β -ketoacid transport into the mitochondrion. In addition, the diatom also needs to import protocatechuic acid into the cell. All diatom genomes, including *Nitzschia*, contain a putative 4-hydroxybenzoate transporter, which can reportedly transport protocatechuic acid (3,4-dihydroxybenzoate) (Nichols and Harwood n.d.). Alternatively, protocatechuic acid might be synthesized from other metabolites transported into the cell, but we found no evidence for enzymes involved in degradation of protocatechuic acid precursors. In summary, it is difficult to predict which diatom transport system is involved in protocatechuic acid acquisition or whether protocatechuic acid is, in fact, the specific aromatic compound that is transported into the diatom cell. Experimental evidence will be necessary to test these hypotheses.

We found evidence for all or part of the β -ketoacid pathway in all sequenced diatom genomes as well as the transcriptome of *Nitzschia* 2144. *Fragilariopsis cylindrus* lacks both TR and 3-ketoacid:succinyl-CoA transferase, whereas *P. tricornutum* and *C. nana* lack PCA, the first enzyme in the pathway. *Phaeodactylum tricornutum* is also missing ELH (Fig. 2-2). Overall,

only the two *Nitzschia* species, Nitz4 and Nitz2144, possessed a complete set of enzymes for the β -ketoadipate pathway. Given the presence of β -ketoadipate genes across diatoms, we wanted to determine whether Nitz4 was enriched for one or more of the genes in this pathway. The first gene in this pathway that we discovered, PCA, catalyzes aromatic ring fission, which is the first and potentially most difficult step in the β -ketoadipate pathway because of the high resonance energy and, therefore, enhanced chemical stability of the carbon ring structure in protocatechuic acid (Fuchs et al. 2011). Consequently, this first critical reaction is one potential “bottleneck” in the β -ketoadipate pathway. Interestingly, Nitz4 possess fully 10 copies of the PCA gene, whereas other diatom genomes contain four or fewer copies. Major expansion of the PCA gene family may suggest that Nitz4 is more efficient than photosynthetic diatom at overcoming the bottleneck reaction of the β -ketoadipate pathway.

Conclusions

The goal of our study was to comprehensively assess core carbon metabolism in a non-photosynthetic diatom, one of a few such species in a lineage of some 100,000 photoautotrophic species (Mann and Vanormelingen 2013, Onyshchenko et al. 2018). The genome sequence revealed the presence of a β -ketoadipate pathway that appears to have been completed through co-option of core carbon metabolism proteins. The genomic hypotheses raised here require follow-up validation with metabolomic experiments. The β -ketoadipate is mostly known from fungi and bacteria (Harwood and Parales 1996), so its presence in Nitz4—and possibly across diatoms as a whole—raises important questions about the origin and ancestral function of this pathway in diatoms. CMD is absent in fungi, which instead possess

3-carboxymuconolactone hydrolase (CMH) converting carboxymuconolactone (Harwood and Parales 1996). The presence of CMD, rather than CMH, in Nitz4, together with relatively strong sequence similarity of Nitz4 CMD to bacterial CMDs suggests that the β -keto adipate pathway in diatoms may be of bacterial rather than fungal origin. A full assessment of the origins of this pathway in diatoms will require detailed phylogenetic analyses of each of the enzymes in the pathway.

The presence of this pathway raises the possibility that mixotrophic diatoms, and obligately heterotrophic diatoms like Nitz4, are able to use lignin-derived aromatic compounds for growth, allowing them to capitalize on one of the most abundant sources of organic carbon worldwide. Experimental data will be necessary to test and validate this hypothesis, including, for example determination of whether diatoms can import protocatechuic acid and whether its derivatives are shuttled directly into the TCA cycle, allowing us to draw a direct line between plant-derived environmental carbon and ATP production in nonphotosynthetic diatoms.

References

- Andrews, Simon. 2010. "FastQC: A Quality Control Tool for High Throughput Sequence Data."
- Archibald, J. M. 2009. "The Puzzle of Plastid Evolution." *Current Biology: CB* 19 (2): R81–88.
- Armbrust, E. Virginia, John A. Berges, Chris Bowler, Beverley R. Green, Diego Martinez, Nicholas H. Putnam, Shiguo Zhou, et al. 2004. "The Genome of the Diatom *Thalassiosira Pseudonana*: Ecology, Evolution, and Metabolism." *Science* 306 (5693): 79–86.
- Blackburn, Michele V., Fiona Hannah, and Andrew Rogerson. 2009. "First Account of Apochlorotic Diatoms from Mangrove Waters in Florida." *The Journal of Eukaryotic Microbiology* 56 (2): 194–200.
- Boisvert, Sébastien, Frédéric Raymond, Elénie Godzaridis, François Laviolette, and Jacques Corbeil. 2012. "Ray Meta: Scalable *de Novo* Metagenome Assembly and Profiling." *Genome Biology* 13 (12): R122.
- Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics* 30 (15): 2114–20.
- Cantarel, Brandi L., Ian Korf, Sofia M. C. Robb, Genis Parra, Eric Ross, Barry Moore, Carson Holt, Alejandro Sánchez Alvarado, and Mark Yandell. 2008. "MAKER: An Easy-to-Use Annotation Pipeline Designed for Emerging Model Organism Genomes." *Genome Research* 18 (1): 188–96.
- Chen, Nansheng. 2004. "Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences." *Current Protocols in Bioinformatics / Editorial Board, Andreas D. Baxevanis ... [et Al.]* Chapter 4 (May): Unit 4.10.
- Claros, M. G. 1995. "MitoProt, a Macintosh Application for Studying Mitochondrial Proteins." *Computer Applications in the Biosciences: CABIOS* 11 (4): 441–47.
- Emanuelsson, Olof, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. 2007. "Locating Proteins in the Cell Using TargetP, SignalP and Related Tools." *Nature Protocols* 2 (4): 953–71.
- Emanuelsson, O., H. Nielsen, and G. von Heijne. 1999. "ChloroP, a Neural Network-Based Method for Predicting Chloroplast Transit Peptides and Their Cleavage Sites." *Protein Science: A Publication of the Protein Society* 8 (5): 978–84.
- Emms, David M., and Steven Kelly. 2015. "OrthoFinder: Solving Fundamental Biases in Whole Genome Comparisons Dramatically Improves Orthogroup Inference Accuracy." *Genome Biology* 16 (August): 157.

- Field, C. B., M. J. Behrenfeld, J. T. Randerson, and P. Falkowski. 1998. "Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components." *Science* 281 (5374): 237–40.
- Figuroa-Martinez, Francisco, Aurora M. Nedelcu, David R. Smith, and Reyes-Prieto Adrian. 2015. "When the Lights Go out: The Evolutionary Fate of Free-Living Colorless Green Algae." *The New Phytologist* 206 (3): 972–82.
- Fuchs, Georg, Matthias Boll, and Johann Heider. 2011. "Microbial Degradation of Aromatic Compounds – from One Strategy to Four." *Nature Reviews. Microbiology* 9 (11): 803–16.
- Gruber, Ansgar, Gabrielle Roca, Peter G. Kroth, E. Virginia Armbrust, and Thomas Mock. 2015. "Plastid Proteome Prediction for Diatoms and Other Algae with Secondary Plastids of the Red Lineage." *The Plant Journal: For Cell and Molecular Biology* 81 (3): 519–28.
- Gruber, Ansgar, Sascha Vugrinec, Franziska Hempel, Sven B. Gould, Uwe-G Maier, and Peter G. Kroth. 2007. "Protein Targeting into Complex Diatom Plastids: Functional Characterisation of a Specific Targeting Motif." *Plant Molecular Biology* 64 (5): 519–30.
- Gschloessl, Bernhard, Yann Guermeur, and J. Mark Cock. 2008. "HECTAR: A Method to Predict Subcellular Targeting in Heterokonts." *BMC Bioinformatics* 9 (September): 393.
- Guillard, Robert R. L. 1960. "A Mutant of *Chlamydomonas Moewusii* Lacking Contractile Vacuoles." *The Journal of Eukaryotic Microbiology* 7 (3). Wiley Online Library: 262–68.
- Guillard, R. R. L., and P. E. Hargraves. 1993. "*Stichochrysis Immobilis* Is a Diatom, Not a Chrysophyte." *Phycologia* 32 (3). The International Phycological Society: 234–36.
- Harwood, C. S., and R. E. Parales. 1996. "The Beta-Ketoadipate Pathway and the Biology of Self-Identity." *Annual Review of Microbiology* 50: 553–90.
- Hellebust, Johan A., and Joyce Lewin. 1977. "Heterotrophic Nutrition." In *The Biology of Diatoms*, edited by Dietrich Werner, 13:169–97. University of California Press.
- Hunt, Martin, Taisei Kikuchi, Mandy Sanders, Chris Newbold, Matthew Berriman, and Thomas D. Otto. 2013. "REAPR: A Universal Tool for Genome Assembly Evaluation." *Genome Biology* 14 (5): R47.
- Janusz, Grzegorz, Anna Pawlik, Justyna Sulej, Urszula Swiderska-Burek, Anna Jarosz-Wilkolazka, and Andrzej Paszczyński. 2017. "Lignin Degradation: Microorganisms, Enzymes Involved, Genomes Analysis and Evolution." *FEMS Microbiology Reviews* 41 (6): 941–62.

- Kamikawa, Ryoma, Daniel Moog, Stefan Zauner, Goro Tanifuji, Ken-Ichiro Ishida, Hideaki Miyashita, Shigeki Mayama, et al. 2017. “A Non-Photosynthetic Diatom Reveals Early Steps of Reductive Evolution in Plastids.” *Molecular Biology and Evolution* 34 (9). academic.oup.com: 2355–66.
- Kamikawa, Ryoma, Goro Tanifuji, Sohta A. Ishikawa, Ken-Ichiro Ishii, Yusei Matsuno, Naoko T. Onodera, Ken-Ichiro Ishida, et al. 2016. “Proposal of a Twin Aarginine Translocator System-Mediated Constraint against Loss of ATP Synthase Genes from Nonphotosynthetic Plastid Genomes.” *Molecular Biology and Evolution* 33 (1). academic.oup.com: 303.
- Kamikawa, Ryoma, Naoji Yubuki, Masaki Yoshida, Misaka Taira, Noriaki Nakamura, Ken-Ichiro Ishida, Brian S. Leander, et al. 2015. “Multiple Losses of Photosynthesis in *Nitzschia* (Bacillariophyceae).” *Phycological Research* 63 (1): 19–28.
- Kaschabek, Stefan R., Bernd Kuhn, Dagmar Müller, Eberhard Schmidt, and Walter Reineke. 2002. “Degradation of Aromatics and Chloroaromatics by *Pseudomonas* Sp. Strain B13: Purification and Characterization of 3-Oxoadipate:succinyl-Coenzyme A (CoA) Transferase and 3-Oxoadipyl-CoA Thiolase.” *Journal of Bacteriology* 184 (1): 207–15.
- Korf, Ian. 2004. “Gene Finding in Novel Genomes.” *BMC Bioinformatics* 5 (May): 59.
- Kroth, Peter G., Anthony Chiovitti, Ansgar Gruber, Veronique Martin-Jezequel, Thomas Mock, Micaela Schnitzler Parker, Michele S. Stanley, et al. 2008a. “A Model for Carbohydrate Metabolism in the Diatom *Phaeodactylum Tricornutum* Deduced from Comparative Whole Genome Analysis.” *PloS One* 3 (1): e1426.
- Laetsch, Dominik R., and Mark L. Blaxter. 2017. “BlobTools: Interrogation of Genome Assemblies.” *F1000Research* 6 (July). <https://doi.org/10.12688/f1000research.12232.1>.
- Lang, Markus, Kirk E. Apt, and Peter G. Kroth. 1998. “Protein Transport into ‘Complex’ Diatom Plastids Utilizes Two Different Targeting Signals.” *The Journal of Biological Chemistry* 273 (47): 30973–78.
- Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59.
- Lewin, Joyce, and R. A. Lewin. 1967. “Culture and Nutrition of Some Apochlorotic Diatoms of the Genus *Nitzschia*.” *Microbiology* 46 (3). Microbiology Society: 361–67.
- Mann, David G., and Pieter Vanormelingen. 2013. “An Inordinate Fondness? The Number, Distributions, and Origins of Diatom Species.” *The Journal of Eukaryotic Microbiology* 60 (4): 414–20.
- Masai, Eiji, Yoshihiro Katayama, and Masao Fukuda. 2007. “Genetic and Biochemical Investigations on Bacterial Catabolic Pathways for Lignin-Derived Aromatic Compounds.” *Bioscience, Biotechnology, and Biochemistry* 71 (1): 1–15.

- Mikulášová, D. 1998. “Malate Dehydrogenase: Distribution, Function and Properties.” *Gen . Physiol . Biophys* 17: 193–121.
- Nichols, Nancy N., and Caroline S. Harwood. n.d. “PcaK, a High-Affinity Permease for the Aromatic Compounds 4-Hydroxybenzoate and Protocatechuate from *Pseudomonas Putida*.”
- Onyshchenko, Anastasiia, Elizabeth C. Ruck, Teofil Nakov, and Andrew J. Alverson. 2018. “A Single Loss of of Photosynthesis in Diatoms.” *bioRxiv*. <https://doi.org/10.1101/298810>. *In Review*.
- Parales, Rebecca E., and Caroline S. Harwood. n.d. “Succinyl-Coenzyme Transferase in *Pseudomonas Putida*.”
- Parks, Matthew B., Teofil Nakov, Elizabeth C. Ruck, Norman J. Wickett, and Andrew J. Alverson. 2017. “Phylogenomics Reveals an Extensive History of Genome Duplication in Diatoms (Bacillariophyta).” *bioRxiv*. <https://doi.org/10.1101/181115>.
- Petersen, Thomas Nordahl, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. 2011. “SignalP 4.0: Discriminating Signal Peptides from Transmembrane Regions.” *Nature Methods* 8 (10): 785–86.
- Pringsheim, E. G. 1956. “Micro-Organisms from Decaying Seaweed.” *Nature* 178 (4531): 480–81.
- Sheikhzadeh, Siavash, and Dick de Ridder. 2015. “ACE: Accurate Correction of Errors Using K-Mer Tries.” *Bioinformatics* 31 (19): 3216–18.
- Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. “BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs.” *Bioinformatics* 31 (19): 3210–12.
- Smetacek, V. 1999. “Diatoms and the Ocean Carbon Cycle.” *Protist* 150 (1): 25–32.
- Smith, Sarah R., Raffaella M. Abbriano, and Mark Hildebrand. 2012. “Comparative Analysis of Diatom Genomes Reveals Substantial Differences in the Organization of Carbon Partitioning Pathways.” *Algal Research* 1 (1): 2–16.
- Stanier, R. Y., and L. N. Ornston. 1973. “The Beta-Ketoadipate Pathway.” *Advances in Microbial Physiology* 9 (0): 89–151.
- Stoebe, Bettina, and Uwe-G Maier. 2002. “One, Two, Three: Nature’s Tool Box for Building Plastids.” *Protoplasma* 219 (3-4): 123–30.

- Traller, Jesse C., Shawn J. Cokus, David A. Lopez, Olga Gaidarenko, Sarah R. Smith, John P. McCrow, Sean D. Gallaher, et al. 2016a. "Genome and Methylome of the Oleaginous Diatom *Cyclotella Cryptica* Reveal Genetic Flexibility toward a High Lipid Phenotype." *Biotechnology for Biofuels* 9 (November): 258.
- Tuchman, Nancy C., Marc A. Schollett, Steven T. Rier, and Pamela Geddes. 2006. "Differential Heterotrophic Utilization of Organic Compounds by Diatoms and Bacteria under Light and Dark Conditions." *Hydrobiologia* 561 (1). Kluwer Academic Publishers: 167–77.
- Wells, Tyrone, Jr, and Arthur J. Ragauskas. 2012. "Biotechnological Opportunities with the β -Ketoacid Pathway." *Trends in Biotechnology* 30 (12): 627–37.
- Zeman, Igor, Martina Neboháčová, Gabriela Gérecová, Kornélia Katonová, Eva Jánošíková, Michaela Jakúbková, Ivana Centárová, et al. 2016. "Mitochondrial Carriers Link the Catabolism of Hydroxyaromatic Compounds to the Central Metabolism in *Candida Parapsilosis*." *G3* 6 (12): 4047–58.
- Zhang, Simo V., Luting Zhuo, and Matthew W. Hahn. 2016. "AGOUTI: Improving Genome Assembly and Annotation Using Transcriptome Data." *GigaScience* 5 (1): 31.

Table 2-1. *Nitzschia* sp. Nitz4 genome annotation progress statistics

| Annotation № | Gene models number | Complete BUSCOs | Complete and Duplicated BUSCOs | |
|---|--------------------|-----------------|--------------------------------|-------------------------|
| 1 | 8676 | 202 | 5 | |
| 2 (SNAP HMM1) | 9291 | 208 | 5 | |
| 3 (SNAP HMM2) | 9324 | 208 | 5 | |
| 4 (SNAP HMM3) | 9327 | 208 | 5 | |
| 5 (after agouti rescaffolding; SNAP HMM4) | 9235 | 209 | 5 | FINAL annotation |
| 6 (after agouti rescaffolding; SNAP HMM5) | 9295 | 208 | 5 | |

Table 2-2. General genome annotation statistics for analyzed diatom species

| Genome feature | <i>Cyclotella nana</i> | <i>Cyclotella cryptica</i> | <i>Phaeodactylum tricorutum</i> | <i>Fragilariopsis cylindrus</i> | <i>Nitzschia</i> sp. Nitz4 | <i>Pseudonitzschia multistriata</i> |
|------------------------------|------------------------|----------------------------|---------------------------------|---------------------------------|----------------------------|-------------------------------------|
| Genome size (Mbp) | 32.4 | 161.7 | 27.4 | 61.1 | 30.7 | 55.1 |
| GC content (%) | 47 | 43 | 48.8 | 38.8 | 47.8 | 46.4 |
| Protein-coding genes | 11,673 | 21,121 | 10,408 | 18,111 | 9,235 | not annotated |
| Complete/Duplicated BUSCOs | 202/196 | 198/37 | 201/192 | 193/7 | 209/5 | not annotated |
| Average gene size (bp) | 992 | 1471 | 1621 | 1575 | 1953 | not annotated |
| Gene density (genes per Mbp) | 360 | 131 | 380 | 296 | 301 | not annotated |
| Reference | Armbrust et al. 2004 | Traller et al. 2016 | Bowler et al. 2008 | Mock et al. 2017 | This study | GCA_900005105.1 |

Table 2-3. Diatom β -ketoacid pathway annotation for analyzed species

| Species name | Gene ID/ Scaffold name | Approximate gene location coordinates | Orthogroup ID |
|--|---|---------------------------------------|---------------|
| PCA (protocatechuate 3,4-dioxygenase) | | | |
| Nitzschia sp. Nitz4 | | | |
| | augustus_masked-agouti_scaf_108-processed-gene-0.1-mRNA-1 | | OG0006126 |
| | augustus_masked-scaffold-1189-processed-gene-0.1-mRNA-1 | | OG0006126 |
| | augustus_masked-scaffold-393-processed-gene-0.2-mRNA-1 | | OG0006126 |
| | augustus_masked-scaffold-393-processed-gene-0.9-mRNA-1 | | OG0006126 |
| | maker-scaffold-393-augustus-gene-0.15-mRNA-1 | | OG0006126 |
| | agouti_scaf_71 | 65156-66154 | - |
| | agouti_scaf_82 | 25240-26280 | - |
| | scaffold-320 | 103557-104515 | - |
| | scaffold-357 | 23928-24971 | - |
| | scaffold-420 | 11602-12576 | - |
| F. cylindrus | | | |
| | OEU10566.1 | | OG0003380 |
| | OEU05863.1 | | OG0003380 |
| | OEU16075.1 | | OG0003380 |
| | KV784367.1_FRACYscaffold_15 | 963661-965124 | - |
| T. pseudonana | | | |
| | XP_002290083.1 | | - |
| P. tricornutum | | | |
| | | 0 | 0 - |
| Nitzschia sp. Nitz2144 | | | |
| | TRINITY_DN15091_c0_g1_i1 m.24867_nitz2144 | | OG0003380 |
| | TRINITY_DN28518_c0_g1_i1 m.64186_nitz2144 | | OG0003380 |
| P. multiseriis | | | |
| | LN865384.1_PsnmuV1.4_scaffold_220 | 7444-8199 | - |
| | LN865619.1_PsnmuV1.4_scaffold_455 | 25524-26258 | - |

Table 2-3. (Cont.)

| Species name | Gene ID/ Scaffold name | Approximate gene location coordinates | Orthogroup ID |
|---|--|---------------------------------------|---------------|
| CMLE (3-carboxy-cis,cis-muconate lactonizing enzyme) | | | |
| Nitzschia sp. Nitz4 | | | |
| | maker-scaffold-212-augustus-gene-0.118 | | OG0005692 |
| F. cylindrus | | | |
| | OEU21789.1 | | OG0005692 |
| T. pseudonana | | | |
| | XP_02294624.1 | | OG0005692 |
| P. tricornutum | | | |
| | XP_002186108.1 | | OG0005692 |
| Nitzschia sp. Nitz2144 | | | |
| | TRINITY_DN17337_c0_g1_i1 m.31016 | | OG0005692 |
| P. multiseriis | | | |
| | LN865398.1_PsnmuV1.4_scaffold_234-size_84702-1 | 10312 10944 | - |
| CMD (4-carboxy-muconolactone decarboxylase) | | | |
| Nitzschia sp. Nitz4 | | | |
| | TRINITY_DN5181_c0_g1_i2 m.19823 | | - |
| | TRINITY_DN5181_c0_g1_i3 m.19827 | | - |
| | TRINITY_DN5181_c0_g1_i5 m.19835 | | - |
| | TRINITY_DN5181_c0_g1_i9 m.19850 | | - |
| F. cylindrus | | | |
| | OEU17373.1 | | OG0007118 |
| | KV784448.1_FRACYscaffold_97 | 126002-126649 | - |
| P. tricornutum | | | |
| | NC_011679.1_chromosome_11_complete_sequence | 672175-672819 | - |
| T. pseudonana | | | |
| | XP_002296700.1 | | - |
| | XP_02286031.1 | | OG0007118 |

Table 2-3. (Cont.)

| Species name | Gene ID/ Scaffold name | Approximate gene location coordinates | Orthogroup ID |
|--|--|---------------------------------------|---------------|
| Nitzschia sp. Nitz2144 | | | |
| | TRINITY_DN14112_c0_g1_i1 m.22486 | | OG0007118 |
| | TRINITY_DN14112_c0_g1_i2 m.22488 | | OG0007118 |
| P. multiseriis | | | |
| | LN865211.1_PsnmuV1.4_scaffold_47 | 182326-183117 | - |
| ELH (enol-lactone hydrolase) | | | |
| Nitzschia sp. Nitz4 | | | |
| | snap_masked-scaffold-381-processed-gene-0.9-mRNA-1 | | OG0003891 |
| F. cylindrus | | | |
| | OEU08712.1 | | OG0009503 |
| T. pseudonana | | | |
| | XP_002294595.1 | | OG0010576 |
| P. tricornutum | | | |
| | | 0 | 0 - |
| Nitzschia sp. Nitz2144 | | | |
| | TRINITY_DN18903_c0_g1_i1 m.37015 | | OG0007049 |
| P. multiseriis | | | |
| | | 0 | 0 - |
| TR (3-ketoacid:succinyl-CoA transferase) | | | |
| Nitzschia sp. Nitz4 | | | |
| | augustus_masked-scaffold-630-processed-gene-0.27 | | OG0007884 |
| F. cylindrus | | | |
| | | 0 | 0 - |

Table 2-3. (Cont.)

| Species name | Gene ID/ Scaffold name | Approximate gene location coordinates | Orthogroup ID |
|--|--|---------------------------------------|---------------|
| T. pseudonana | | | |
| | XP_002288066.1 | | OG0007884 |
| P. tricornutum | | | |
| | XP_002184327.1 | | OG0007884 |
| Nitzschia sp. Nitz2144 | | | |
| | TRINITY_DN13259_c0_g1_i1 m.20451 | | OG0007884 |
| P. multiseriis | | | |
| | | 0 | 0 - |
| TH (3-ketoacyl-CoA thiolase) | | | |
| Nitzschia sp. Nitz4 | | | |
| | maker-agouti_scaf_70-snap-gene-0.94-mRNA-1 | | OG0005358 |
| F. cylindrus | | | |
| | OEU20070.1 | | OG0005358 |
| T. pseudonana | | | |
| | XP_002291557.1 | | OG0005358 |
| P. tricornutum | | | |
| | XP_002180306.1 | | OG0005358 |
| Nitzschia sp. Nitz2144 | | | |
| | TRINITY_DN11969_c0_g1_i1 m.17323 | | OG0005358 |
| P. multiseriis | | | |
| | LN865426.1_PsnmuV1.4_scaffold_262 | 60224 61021 | - |
| PcaK (4-hydroxybenzoate and protocatechuate transporter) | | | |
| Nitzschia sp. Nitz4 | | | |
| | augustus_masked-scaffold-375-processed-gene-0.7-mRNA-1 | | OG0003582 |

Table 2-3. (Cont.)

| Species name | Gene ID/ Scaffold name | Approximate gene location coordinates | Orthogroup ID |
|-------------------------------|-----------------------------------|--|----------------------|
| F. cylindrus | | | |
| | OEU08398.1 | | OG0003582 |
| | OEU07405.1 | | OG0003476 |
| T. pseudonana | | | |
| | XP_002288818.1 | | OG0003582 |
| | XP_002287244.1 | | no_ortho |
| | XP_002294330.1 | | OG0003476 |
| P. tricornutum | | | |
| | XP_002178175.1 | | OG0003582 |
| | XP_002177299.1 | | OG0003476 |
| Nitzschia sp. Nitz2144 | | | |
| | TRINITY_DN22222_c0_g2_i1 m.53129 | | OG0003582 |
| P. multiseriis | | | |
| | LN865266.1_PsnmuV1.4_scaffold_102 | 96577-97404 | - |
| | LN865522.1_PsnmuV1.4_scaffold_358 | 9025-9774 | - |

Legend:

- diatom gene
- PCA
(bacterial in all cases)
- unknown origin gene
- bacterial gene

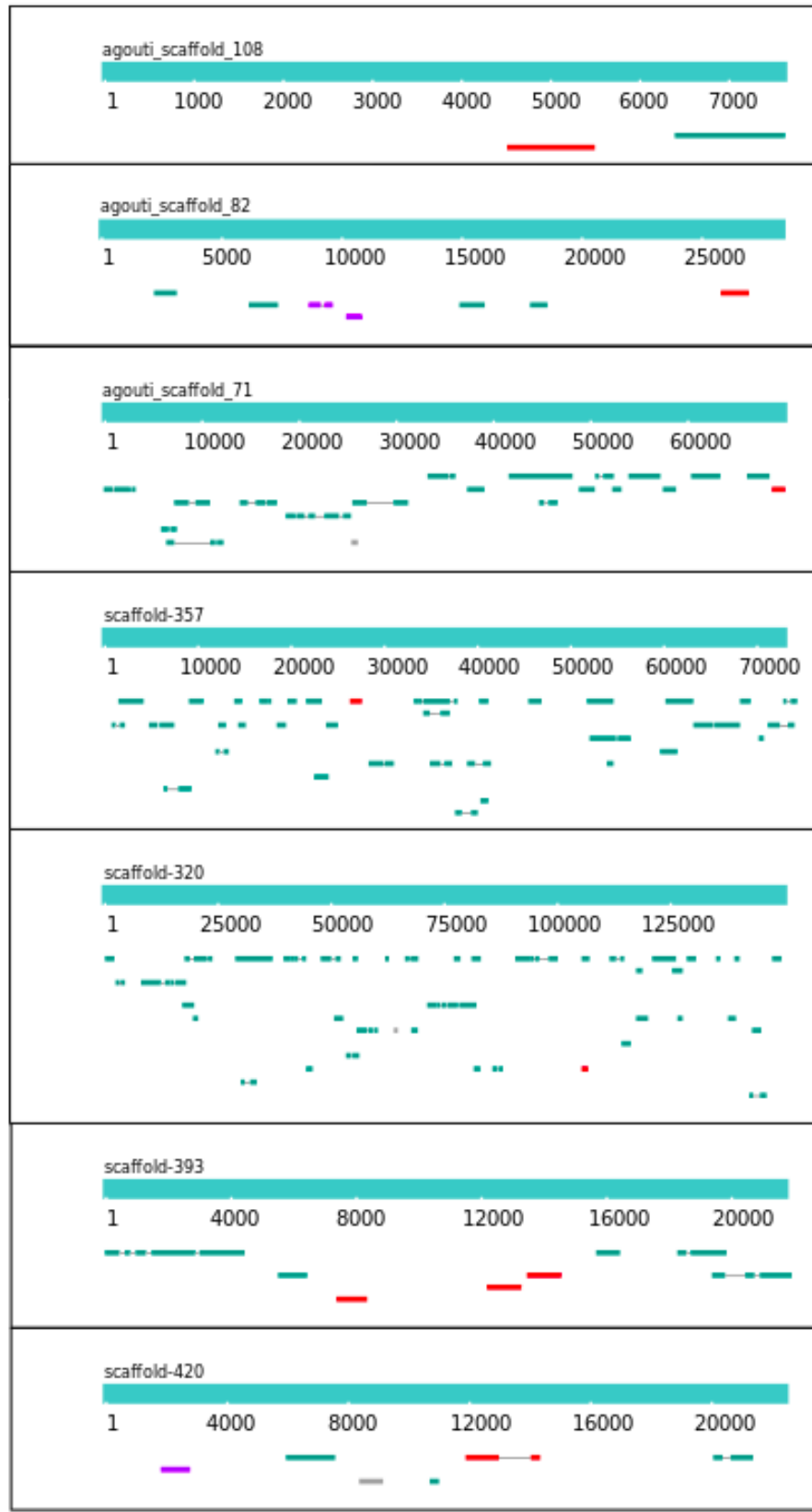
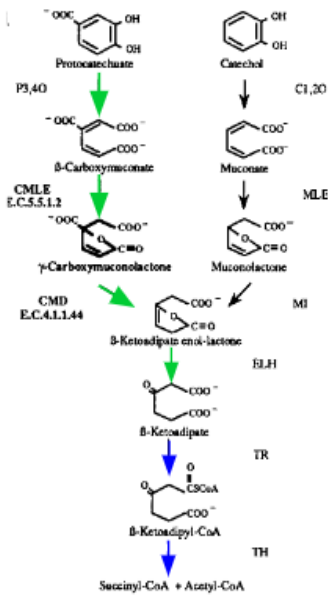
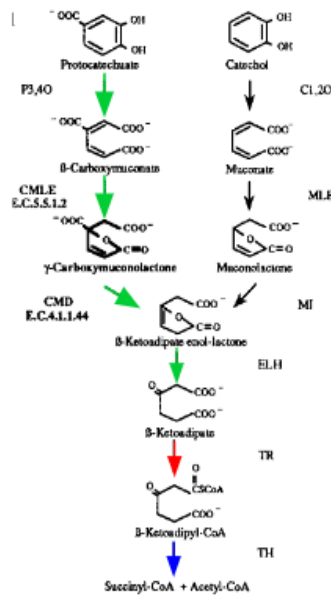


Figure 2-1. Protocatechuate dioxygenase genes localization on *Nitzschia* sp. Nitz4 scaffolds

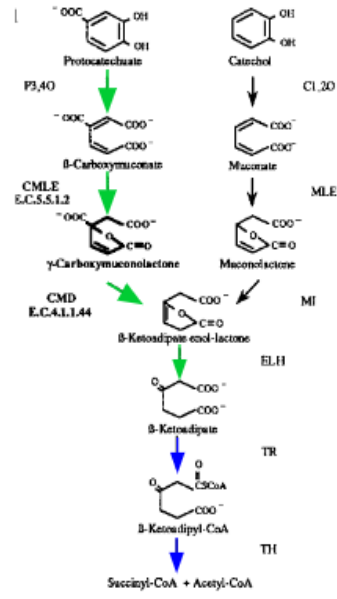
Nitzschia sp. Nitz4



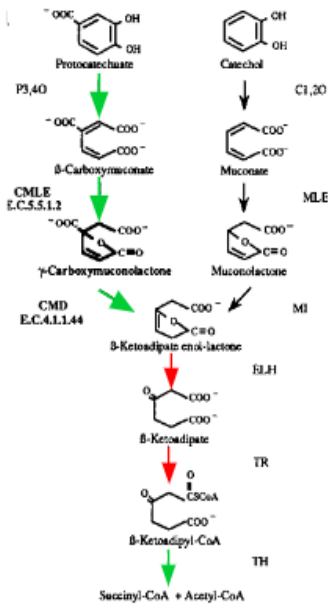
F. cylindrus



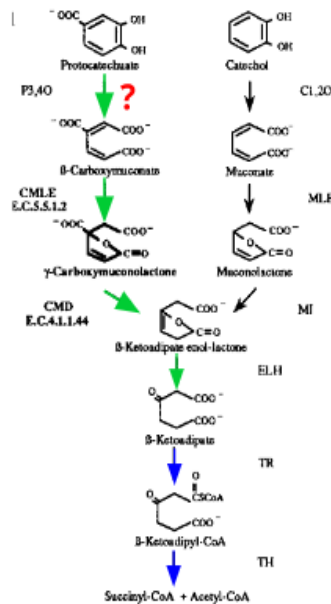
Nitzschia sp. Nitz2144



P. multiseriis



T. pseudonana



P. tricornutum

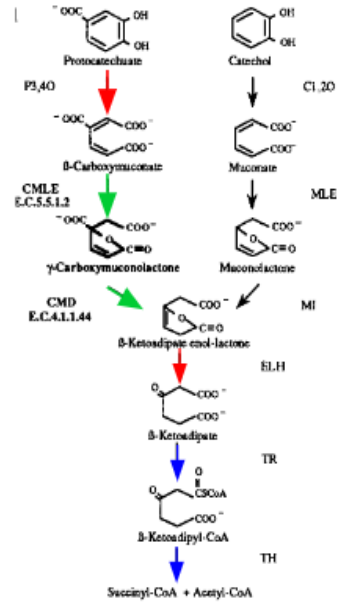


Figure 2-2. Schematic annotation of protocatechuate branch of β -ketoadipate pathway in analyzed diatom species. Green arrow -gene present, red -gene was not found, blue -putative gene was found.

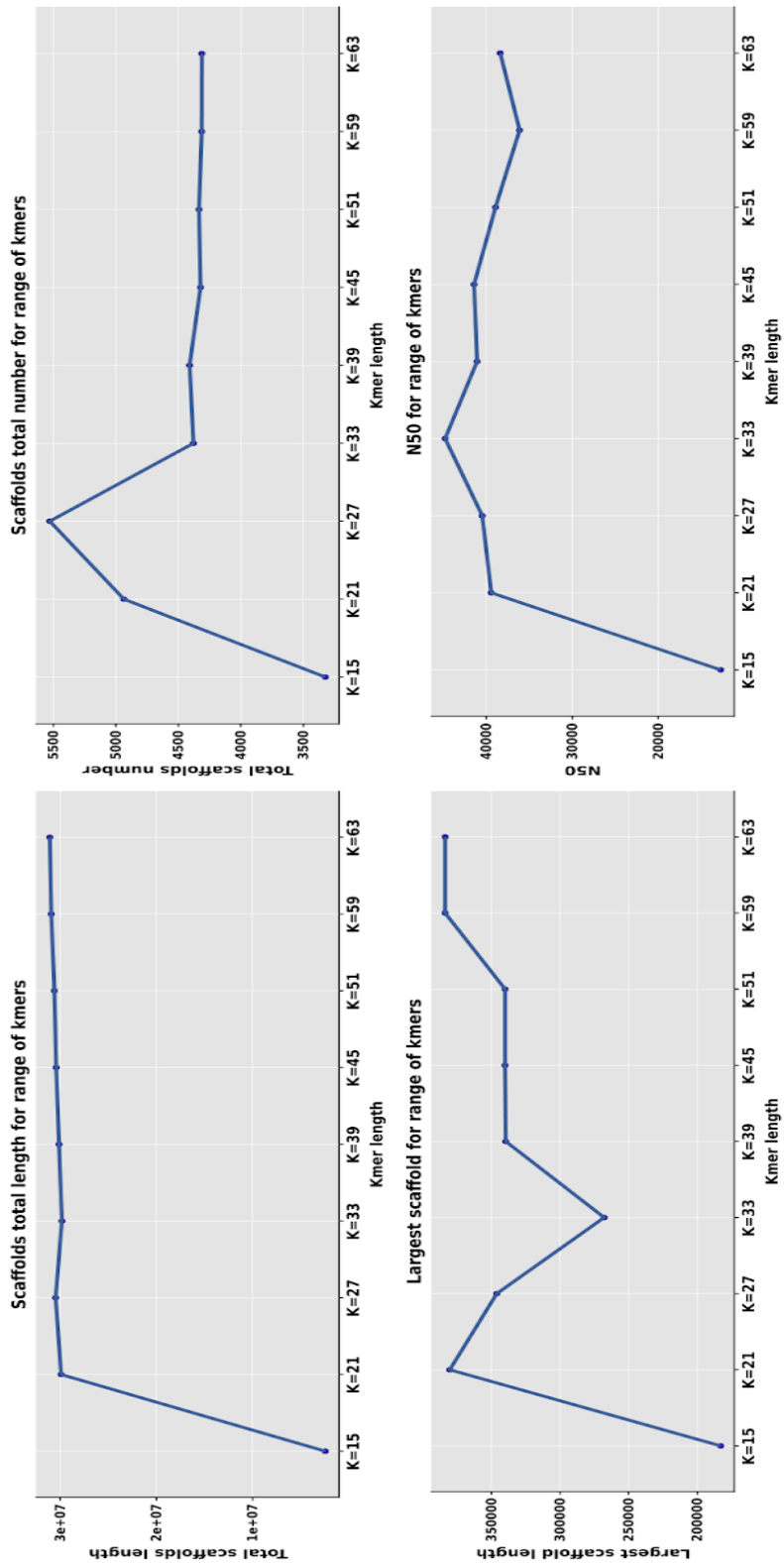


Figure S2-1. Statistics plots for original genome scaffolding for range of K-mers

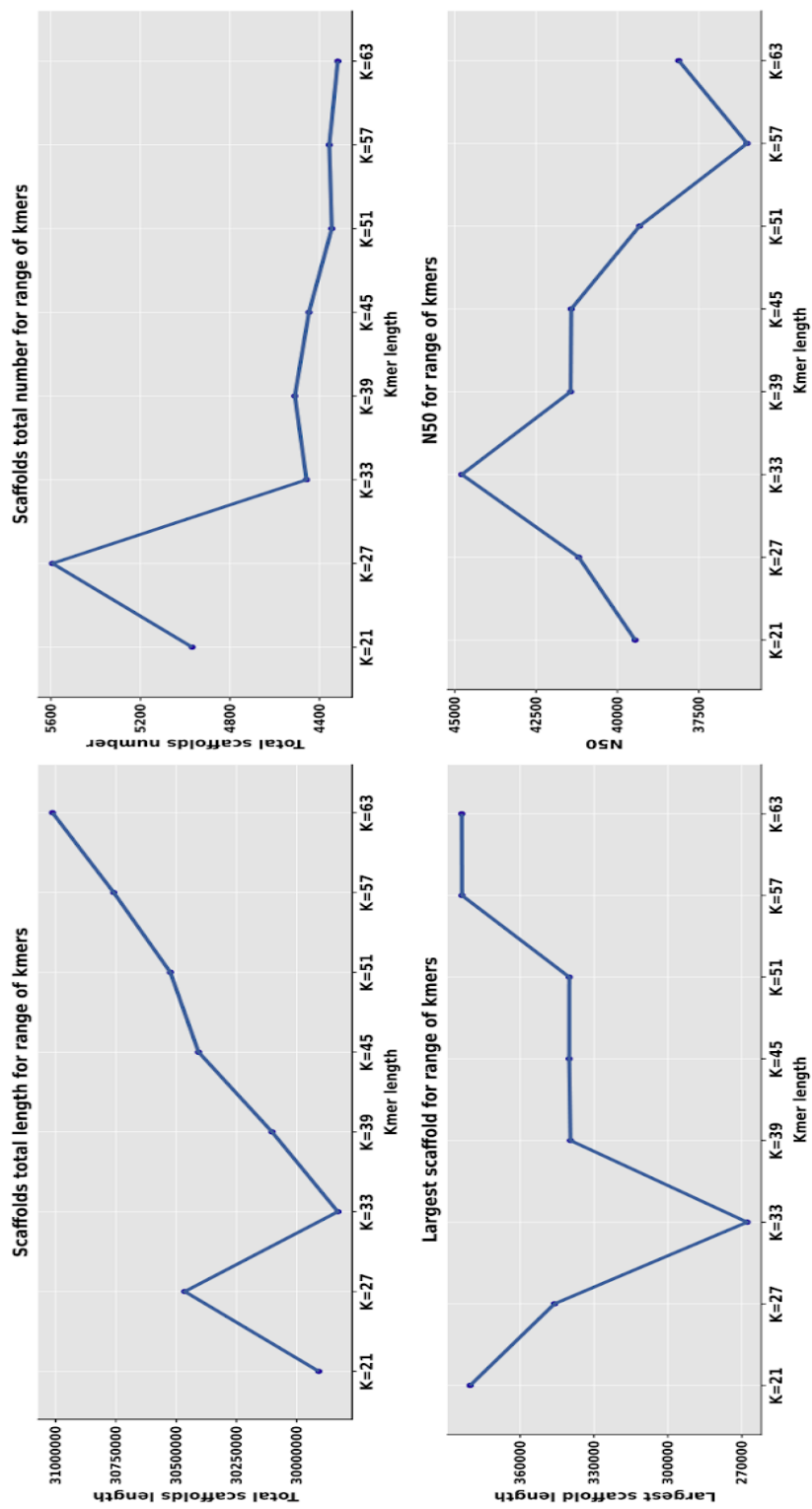


Figure S2-2. Statistics plots for organellar reads-free genome scaffolding for range of K-mers

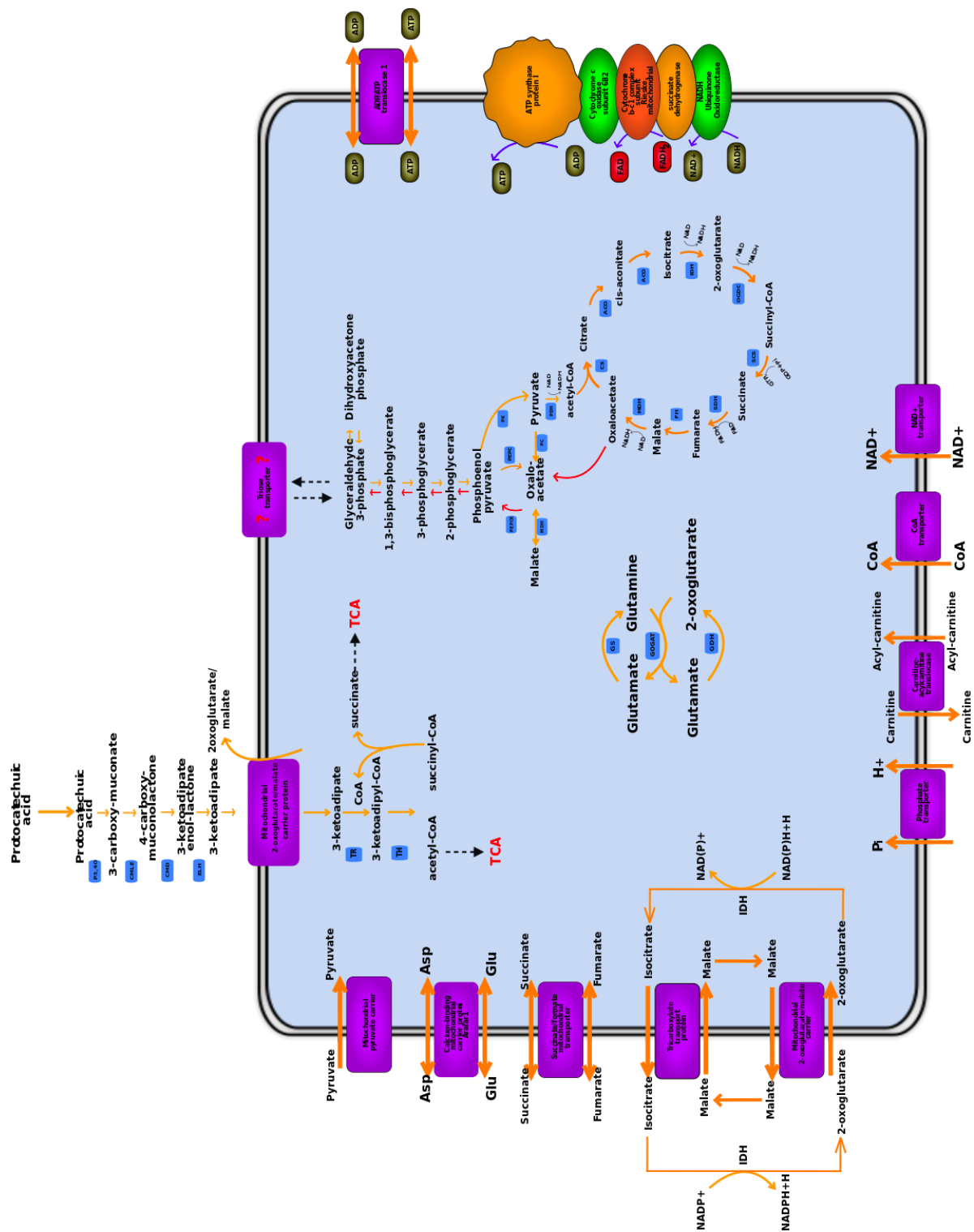


Figure S2-3. Comprehensive mitochondrial metabolic pathways of *Nitzschia* sp. Nitz4 derived from genome annotation

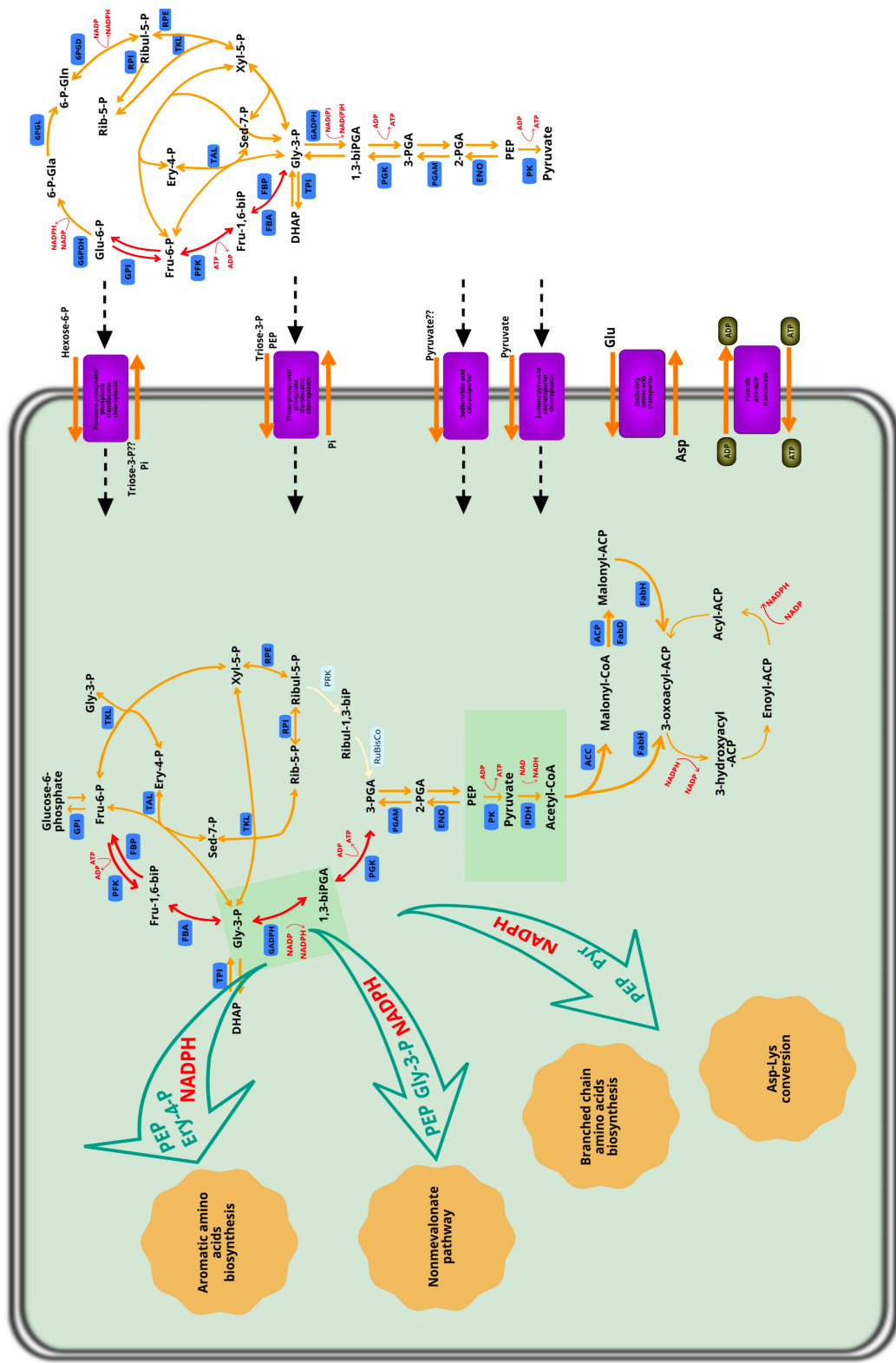


Figure S2-4. Comprehensive plastid metabolic pathways of *Nitzschia* sp. Nitz4 derived from genome annotation

Conclusions

The goal of this study was to determine the number of losses of photosynthesis in diatoms, a diverse photosynthetic lineage of eukaryotic algae. In the first chapter of my thesis, I showed that photosynthesis was lost just one time. This finding reframes both our understanding of this radical trophic shift but also helps shape future research priorities aimed at understanding the evolutionary, ecological, and genomic contexts of the switch from autotrophy to heterotrophy in diatoms. The second overall goal of my project was to use the nuclear genome sequence of a nonphotosynthetic diatom species to characterize core carbon metabolism and identify the sources of extracellular carbon used by these species. This work, described in chapter two of my thesis, led to the discovery of a novel carbon metabolic pathway involved in the degradation of lignin-derived aromatic compounds. If verified experimentally, this would show that nonphotosynthetic diatoms are able to capitalize on one of the most abundant sources of carbon on the planet, further cementing their place among the most intriguing and important contributors to the global carbon cycle.