A GENOMIC SURVEY OF TWO DINOTOMS

by

Behzad Imanian

MSc., The University of British Columbia, 2006

BSc., The University of British Columbia, 2002

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Botany)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2013

Abstract

Endosymbiosis has played a major role in shaping eukaryotic cells, their success and diversity. At the base of the eukaryotic tree, an α-proteobacterium endosymbiont in a protoeukaryotic cell was converted into the mitochondrion through its reductive evolution, endosymbiotic gene transfer (EGT) and the development of a protein targeting system to direct the products of the transferred genes to this organelle. Similar events mark the plastid evolution from a cyanobacterium. However, the primary endosymbiosis of plastid, unlike the mitochondrion, was followed by the secondary and tertiary movement of this organelle between eukaryotes through analogous endosymbiotic reduction, EGT and evolution of a protein targeting system and many subsequent independent losses from different eukaryotic lineages.

The obligate tertiary diatom endosymbiont in a small group of dinoflagellates called 'dinotoms' is exceptional in that it retains most of its ancestral characters including a large nucleus, its own mitochondria, plastids and many other eukaryotic organelles and structures in a large cytoplasm all enclosed in and separated from its dinoflagellate host by a single membrane. This level of conservation of ancestral features in the endosymbiont suggests an early stage of integration. In order to investigate the impacts of endosymbiosis on the organelle genomes and to determine the extent of EGT and the contribution of the host nuclear genome to the proteomes of the organelles, I conducted mass pyrosequencing of the A+T-rich portion of the DNA extracted from two dinotoms, *Durinskia baltica* and *Kryptoperidinium foliaceum*, and the SL cDNA library constructed for *D. baltica*.

The plastid and mitochondrial genomes of these two dinotoms were sequenced, and the results indicated that, despite the permanent symbiosis between the host and its endosymbiont in dinotoms and in spite of small variations, the dinotom organelle genomes have changed very

little from those of free-living diatoms and dinoflagellates. There was also no sign of EGT to the host in *D. baltica*, suggesting a strict compartmentalization in which the host mitochondria remain reliant on the host nucleus while the endosymbiont organelles, mitochondria and plastids, stay entirely dependent on the endosymbiont nucleus with no genetic exchange between the host and endosymbiont.

Preface

A version of chapter 2 has been published. Imanian B, Pombert J-F, Keeling PJ. 2010. The complete plastid genomes of the two "dinotoms" *Durinskia baltica* and *Kryptoperidinium foliaceum*. PloS ONE. 5:e10711. doi: 10.1371/journal.pone.0010711. The project was conceived by PJK, J-FP and I. J-FP worked on the genome assembly, annotation and Sanger sequencing of the plastid genome of *K. foliaceum*. I conducted the culturing, DNA and RNA extractions, CsCl gradient density centrifugations, DNA amplifications, PCRs, RT-PCRs, Sanger sequencing, final chromosome walking steps, genome assembly, base calling, annotation and finishing of both plastid genomes. J-FP and I analyzed the data. I wrote the first draft. J-FP, PJK and I contributed in editing and writing the final draft.

A version of chapter 3 has been published as well. Imanian B, Pombert J-F, Dorrell RG, Burki F, Keeling PJ. 2012. Tertiary endosymbiosis in two dinotoms has generated little change in the mitochondrial genomes of their dinoflagellate hosts and diatom endosymbionts. PLoS ONE. 7:e43763. doi: 10.1371/journal.pone.0043763. PJK and I conceived and designed the experiments. J-FP and RGD helped with the PCRs, RT-PCRs and Sanger sequencing for the endosymbiont mitochondrial genomes. FB prepared the PolyA cDNA library of *K. foliaceum*. I conducted culturing, DNA and RNA extractions, CsCl gradient density centrifugations, PCRs, RT-PCRs, cDNA amplifications, genome assemblies, base callings, chromosome walking, Sanger sequencing and annotation of all the four mitochondrial genomes. I analyzed the data and wrote the first draft. J-FP, PJK and I contributed in editing and writing the final draft.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my supervisor, Patrick Keeling, who gave me the chance to work in his excellent lab where I met some of the brightest scientists and some of my best friends. I would like to extend my gratitude to my committee members, Naomi Fast, Brian Leander, Keith Adams as well as my graduate advisor Gary Bradfield, who have supported and encouraged me over the years in UBC. I want to thank all my co-authors, Patrick Keeling again, Jean-François Pombert, Fabien Burki and Richard Dorrell for all their hard work and contributions to my research. I would like to give special thanks to Fabien Burki and Elisabeth Hehenberger for their valuable advice and significant help on phylogenetic analyses and 5′ RACE experiments, respectively, whose results are documented and discussed in chapter 4.

This research was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) grant to Patrick Keeling. I cordially thank the NSERC for granting me a Doctoral Scholarship.

I am thankful to my lab mates, past and present, for their great passion for contributing to our collective knowledge, for their help, for their friendship: Bryony Williams, Matthew Rogers, Michelle McEwan, Audrey de Koning, Nicola Patron, Ross Waller, Claudio Slamovits, Ales Horak, Lex Howe, Rowena Stern, Gillian Gile, Kevin Carpenter, Lena Burri, James Todd Harper, Ruijuan Kang, Noriko Okamoto, Vera Tai, Floyd Bardell, David Smith, Eric James, Yoshihisa Hirakawa, Juan Saldarriaga, Jean-François Pombert, Jan Janouškovec and Fabien Burki.

I also thank Beverley Green and Yunkun Dang for their assistance with my experiments on CsCl gradient density centrifugation; Beverley Green, Marie-Pierre Oudot-Le Secq and Chris Bowler for providing me access to the data for mitochondrial genome of *Phaeodactylum*

tricornutum prior to its publication; the US Department of Energy Joint Institute (http://www.jgi.doe.gov) for producing the sequence data for *Fragilariopsis cylindrus* and Thomas Mock for his permission to use the data; Beverley Green again for her constant interest on my project and her guidance; Curtis Suttle and James Berger for their time, efforts and thought-provoking questions during my candidacy exam; Steven Hallam and Young Song for their help with bioinformatics work; Julie Brame for her great and determined efforts to send me *Peridinium quinquecorne* samples, which unfortunately I could not completely isolate and culture.

My thanks also go to all the scientists, technicians and other workers in the Génome Québec Innovation Centre for performing the pyrosequencing and in the CCMP and CSIRO culture collections for providing us with the cultures, as well as the hard working secretaries and office workers in the botany office and the cleaning personnel in both the biodiversity and biology buildings.

I am also sincerely grateful to my dear friends Michael and Linda Lipsen; David and Nancy Crawford; to my fond mother Shokofeh Jahanbakhsh, my tireless brother Hashem Imanian and my caring sister Banafsheh Imanian; and to my loving wife Netsanet Tsegay.

Dedication

To all those who saw further and became a shoulder for us to see better.

Chapter 1: Introduction

The mitochondrion and plastid endosymbioses

Endosymbiotic events are at the core of evolution of eukaryotic cells. Through endosymbiosis, unrelated cells were forged together, and new chimeras were born. These chimeric cells took giant leaps forward together to generate new level of complexity and diversity. The engulfment, reduction, modification, and integration of the ancestors of an α-proteobacterium and a cyanobacterium by another cell in separate occasions gave rise to mitochondria and plastids, respectively (Archibald and Keeling 2002; Gray et al. 1999; Keeling 2010; Palmer 2003). These new additions to the host cells through endosymbiosis resulted in far more complex cells at both structural and physiological levels, and expanded the ability of new resulting cells to explore, adapt to and colonize new environments. Among other functions oxidative phosphorylation and photosynthesis were added to the repertoire of what these new cells collectively could do.

The α-proteobacterial-like endosymbiont transformed to an organelle very early on in the evolution of eukaryotic cells prior to their radiation, and as a consequence nearly all contemporary eukaryotes have at least a mitochondrion or one of its derivatives (i.e. hydrogenosomes, mitosome) or some of its derived genes (Bui et al. 1996; Roger et al. 1996; Roger and Silberman 2002; Tovar et al. 2003; Williams et al. 2002; Williams and Keeling 2003). Long after the establishment of the mitochondrion in the eukaryotic cell, the cyanobacterial endosymbiosis led to the evolution of another organelle that later on diversified into the plastids found in the glaucophytes, red algae, green algae and plants. The rich intracellular environment and availability of nutrients and metabolites remove the necessity of producing these essentials by the engulfed autonomous cells (in both endosymbiosis and parasitism). In such an

environment, many biochemical pathways in the endosymbiont (and the parasite) with an equivalent in the host cell become redundant, and the corresponding proteins and genes become dispensable, and they can be eliminated over time from the proteome and genome of the engulfed cell. The transformation of a free living prokaryotic cell into an organelle in both mitochondrial and plastid endosymbioses has been accompanied by miniaturization of the symbiont, characterized in part by massive gene losses from the bacterial endosymbiont genome, reducing its size and coding capacity to a small fraction of its estimated original size and capability. More than 95% of the genes found in the closest free-living relatives of mitochondria and plastids are missing from the genomes of these two organelles. Modern free-living α-proteobacteria (closest relatives of mitochondria) and cyanobacteria (closest relatives of plastids) possess genomes encoding three to several thousand protein-coding genes (Fogel et al. 1999, Kaneko et al. 1996, Timmis et al. 2004). The known mitochondrial genomes, however, retain only up to 97 genes (Adams and Palmer 2003; Gray et al. 1999) while most plastids maintain only about 1% of the coding capacity of the genomes in their closest free-living prokaryotes (Dagan and Martin 2009). The missing genes from the genomes of these two organelles have been either lost completely or transferred to the nucleus of the host cells by hundreds or even thousands (Martin 2009; Martin et al. 2002; Timmis et al. 2004).

Complementary and subsequent to the endosymbiotic gene transfer (EGT), a protein targeting system has also evolved independently but with analogous features in both endosymbiotic events (Pfanner and Geissler 2001; Vesteg et al. 2009). The transferred genes are encoded and transcribed in the host nucleus, translated in its cytosol, and then some but not all are targeted to whence they originated. The analogous features and components of the protein targeting systems for the two organelles include the translocons of outer and inner membranes of

the mitochondrion and plastid (TOM, TIM and TOC, TIC), their associated receptor proteins, carrier proteins and others that recognize, receive and translocate the organelle proteins through the double-membrane of these organelles to their proper destinations (Cline and Dabney-Smith 2008; Dolezal et al. 2006; Gutensohn et al. 2006; Kovács-Bogdán et al. 2010). Another important analogous feature of these transport systems are the addition of a targeting signal or transit peptide (TP) to the organelle proteins, in many cases to their N-terminal site. The transferred genes are tagged and the products carry this added signal that specifies the correct destination in each case. There is no consensus for primary sequences of these sorting signals. However, both mitochondrial and plastid transit peptides (mTPs and cTPs) do share certain characteristics in their own amino acid compositions and secondary structures (Danne and Waller 2011; Duby et al. 2001; Emanuelsson et al. 2000; Franzén et al. 1990; Hammen and Weiner 1998; von Heijne et al. 1989; von Heijne 1986). The successful integration of the host and endosymbiont would not have been possible without the large-scale enrichment of the host nucleus through EGT and the subsequent development of the protein targeting systems that keep the organelles viable, functional and beneficial.

The striking difference between the evolutionary histories of the mitochondrion and plastid lies in their relative complexity. The endosymbiosis that gave rise to the mitochondrion seems to have occurred only once and very early on at or near the base of the eukaryotic tree. The discoveries of mitochondrion-derived organelles such as hydrogenosomes and mitosomes in highly reduced anaerobic parasites (Bui et al. 1996; Roger 1999; Williams et al. 2002; Williams and Keeling 2003; Tovar et al. 2003) shook the foundations of Archezoa Hypothesis (Cavalier-Smith 1983) and convincingly argued against the hypothetical group of primitively amitochondriate eukaryotes. These discoveries also implied that complete disposal of

mitochondria is a rare event. With one extraordinary exception (dinotoms), there is also no evidence of the secondary acquisition of a mitochondrion by a eukaryote from another eukaryote. The plastid evolution, on the other hand, appears much more eventful. The rise of glaucophytes, red and green algal lineages after the primary endosymbiosis was just the beginning of the plastid succeeding movement between eukaryotes, its secondary acquisitions, replacements and losses.

Secondary endosymbioses

The successful procurement of these two organelles set the conditions for further experimentations in endosymbiosis by eukaryotes. In the following endosymbiotic events known as secondary endosymbioses, a eukaryotic cell with a primary plastid was engulfed by and integrated within another eukaryotic cell. The secondary endosymbioses with red and green algae played a significant role in restructuring and diversifying many eukaryotic lineages (Keeling 2009, 2010). While the number of red algal secondary endosymbioses is still contentious, the fact that it has occurred at least once is not. The red algal endosymbiont or its derived plastids have been discovered in a large group of eukaryotic taxa such as haptophytes, cryptophytes, dinoflagellates, apicomplexans and stramenopiles (heterokonts) (Archibald and Keeling 2002; Cavalier-Smith 1999; Gould et al. 2008; Keeling 2010; Palmer 2003). The secondary green algal derived chloroplasts have been found in two distantly related eukaryotic lineages, euglenids (excavates) and chlorarachniophytes (rhizarians), as well as in the unrelated dinoflagellate genus *Lepidodinium* (*L. viride* and *L. chlorophorum*) (Archibald 2009; Gibbs 1978; Gould et al. 2008; Hansen et al. 2007; Keeling 2010; Kim and Archibald 2009; Matsumoto et al. 2011; Minge et al. 2010; Van de Peer et al. 1996).

The integration of these secondary endosymbionts (both red and green) with their respective hosts has resulted in their extensive phenotypic and genetic reduction, comparable in its nature and extent to the reduction of bacterial ancestors of the mitochondrion and the primary plastid. In most cases, the eukaryotic endosymbiont has lost its nucleus, its mitochondria and nearly all other organelles except the plastids that are maintained and wrapped in one or two extra membranes. In two unrelated lineages, cryptophytes and chlorarachniophytes, which have plastids derived from a red and a green alga, respectively, the miniaturized nucleus (nucleomorph) of the secondary endosymbiont is still maintained in a tiny remnant of its own cytosol (Archibald 2007; Gilson et al. 2006; Lane et al. 2005). The discovery of the nucleomorphs and later the complete sequencing of their genomes demonstrated compellingly that the plastid acquisition could occur indirectly or secondarily through another eukaryote (Archibald 2007; Greenwood 1974; Hibberd and Norris 1984; Lane et al. 2007a; Lane and Archibald 2006; Lane et al. 2006; Lane et al. 2007b). These endosymbionts, with the retention of their nucleomorph, represent a transitional state from a eukaryotic endosymbiont to an organelle (Gilson and McFadden 2002). The genomes of these nucleomorphs are highly compacted and severely reduced with a very limited coding capacity (Archibald 2007; Gilson et al. 2006; Lane et al. 2005). These genomes encode only up to 30 genes with plastid functions while the majority of the genes for plastid-targeted proteins have already been transferred to and are now encoded in the host nuclear genome (Archibald 2007; Gilson et al. 2006; Gilson and McFadden 2002). Most of the proteins encoded in the nucleomorph contribute in the maintenance of its genome, but their functions have to be complemented by the functions of many other proteins whose genes are now, after their transfer, encoded in the nucleus of the host (Douglas et al. 2001; Gilson et al. 2006; Keeling 2010; Lane et al. 2007).

Although the EGT to the host is one of the hallmarks of both primary and secondary endosymbioses, two distinctions between the two events should be noted. First, in the secondary endosymbioses most of the primary plastid genes (derived from the cyanobacterial ancestor) had already been transferred to and assimilated by the nuclear genome of primary host which became the secondary endosymbiont (primary EGT). Thus, the secondary EGT should have occurred mainly as a result of the successful migration of many genes encoded in the nucleus of the endosymbiont to the nuclear genome of the host (Archibald 2007; Gould et al. 2008; Keeling 2009, 2010; Kim and Archibald 2009). Recent studies have started to track and assess the genetic footprints and the extent of the secondary EGT as well as other sources of the horizontally transferred genes in the nucleus of the hosts in these complex systems (Bachvaroff et al. 2004; Burki et al. 2012; Deschamps and Moreira 2012; Minge et al. 2010; Moustafa et al. 2009; Patron et al. 2006).

Second, the extra membrane or membranes that envelope the secondarily derived plastids have added one or more barriers in the way of the protein products of the transferred genes to their destination, the plastid. The plastids of haptophytes, cryptophytes, stramenopiles, apicomplexans and chlorarachniophytes are enveloped in four membranes, the first two (from inside out) derived from the original or primary endosymbiont (cyanobacterium-like), the third from the engulfed red or green algal cell membrane, and the fourth from the phagosomal or food vacuole membrane of the host (Archibald 2009; Archibald and Keeling 2002; Keeling 2010). In dinoflagellates and euglenids that share many convergent features (Lukes et al. 2009), the plastids are surrounded by three instead of the expected four membranes as a consequence of the loss of one of the two outermost membranes either the cell membrane of the secondary endosymbiont or the phagosomal membrane of the host (Archibald 2009; Archibald and Keeling

2002; Keeling 2010). The extra membrane barriers in the secondary endosymbioses have been dealt with, in most cases, by the addition of another targeting signal, called signal peptide (SP), to the N-terminus of the proteins targeted to the plastid. Since many of the nuclear-encoded plastid-targeted proteins in the red or green algae that became the eukaryotic endosymbiont already had a targeting signal, cTP, the addition of SP to cTP has resulted in a bi-partite targeting signal (Deane et al. 2000; Hirakawa et al. 2009; Lang et al. 1998; Van Dooren et al. 2001; Wastl and Maier 2000). Some of the nuclear-encoded plastid-targeted pre-proteins in euglenids and dinoflagellates have modified targeting signals with three functional domains and include, in addition to the SP and cTP, a hydrophobic signal called stop transfer membrane anchor (STMA) (Agrawal and Striepen 2010; Minge et al. 2010; Nassoury and Morse 2005; Patron and Waller 2007; Patron et al. 2005; Sheiner and Striepen 2012). These bi- or tri-partite signals direct many of the secondary plastid proteins first to the protein secretory pathway through the host endomembrane system. From there, they are directed to the TOC and TIC homologues and their associated proteins found in the two innermost membranes of secondary plastids (DeRocher et al. 2000; Durnford and Gray 2006; Felsner et al. 2011; Lang et al. 1998; Sheiner and Striepen 2012; Tonkin et al. 2006; Waller et al. 2000).

Tertiary endosymbiosis in dinoflagellates with a cryptophyte or a haptophyte endosymbiont

In yet another round of endosymbiotic events, dinoflagellates have experimented with new partners, this time with secondary plastid-containing eukaryotes, generating new and extremely complex chimeras. Roughly half of dinoflagellate species are autotrophic, and there is a growing consensus that they along with their parasitic sister group apicomplexans have descended from an ancestor that already had a red algal-derived plastid (Archibald 2009;

Cavalier-Smith 1982, 1999; Janouskovec et al. 2010; Keeling 2010; Moore et al. 2008). Independent plastid losses have occurred many times in dinoflagellates. In several dinoflagellate genera and species, however, the old red algal-derived plastid has been replaced through the uptake of other eukaryotes with secondary plastids such as cryptophytes, haptophytes and diatoms (stramenopiles). Interestingly, the plastids in these three eukaryotic taxa are also derived from red algae.

The cryptophyte-derived plastids are found in several dinoflagellate species from *Amphidinium*, *Gymnodinium* and *Dinophysis* genera (Garcia-Cuetos et al. 2010), but in most cases they are not permanently retained within the dinoflagellate host. In order to keep the plastid functional, the dinoflagellate host needs to feed on a cryptophyte prey directly or indirectly through another eukaryote that feeds on the cryptophyte such as the ciliate *Myrionecta rubra* (synonym, *Mesodinium rubrum*) that maintains the cryptophyte plastid, mitochondria and nucleus for days in isolation and starvation. In a recent transcriptome analysis of the dinoflagellate host in *D. acuminata*, only 5 plastid-targeted proteins were discovered, and phylogenetic analyses indicated that they were derived from various algal groups (1 from haptophytes, 3 from dinoflagellates and only 1 from cryptophytes) (Wisecaver and Hackett 2010). The transient, sequestered, cryptophyte plastids in two phagotrophic dinoflagellates, *A. poecilochroum* and *G. acidotum*, experience little or no modification, whereas in *Dinophysis* species they undergo visible ultrastructural alterations (Garcia-Cuetos et al. 2010). These modifications have been interpreted as evidence for the permanent nature of the relationship between the host/predator and its endosymbiont/prey while the lack of evidence for massive EGT in *Dinophysis* is used to argue for its transient or transitional nature (Garcia-Cuetos et al. 2010; Wisecaver and Hackett, 2010). Whether the cryptophyte plastid in *Dinophysis* is an

established organelle, an organelle-in-the-making, or just a monthly ration of food has been the subject of many studies and heated debates and needs further investigations (Garcia-Cuetos et al. 2010; Hackett et al. 2003; Hallegraeff and Lucas 1988; Lucas and Maret 1990; Park et al. 2010; Qiu et al. 2011; Schnepf and Elbraechter 1988; Wisecaver and Hackett 2010).

Although a transient relationship between a dinoflagellate from Antarctica and a haptophyte is also reported (Gast et al. 2007), the permanent nature of the haptophyte-derived plastids in the two dinoflagellate genera *Karenia* and *Karlodinium* is less controversial (Tengs et al. 2000; Yoon et al. 2002). From the haptophyte endosymbionts in *Karenia* and *Karlodinium* only their plastids remain, and there is no sign of a nucleus, mitochondria or any other organelles. It is, unfortunately, not clear whether these haptophyte-derived plastids are surrounded by 2, 3 or 4 membranes (Dodge 1989; Hackett, et al. 2004; Tengs et al., 2000). It is known, however, that in *Karlodinium micrum* (synonym *Karlodinium veneficum*) the plastid genome has suffered gene losses and shows signs of gene degeneration, massive genome rearrangements and intergenic space expansion (Gabrielsen et al. 2011). There is evidence of EGT in these tertiary plastid-containing dinoflagellates (Ishida and Green 2002; Nosenko et al. 2006; Patron et al. 2006; Yokoyama et al. 2011). The expressed sequence tag (EST) surveys and phylogenetic analyses of the putative plastid-targeted proteins in *K. micrum* and *K. brevis* have revealed that the plastid is maintained by the proteins mostly derived from the haptophyte endosymbiont along with several proteins derived from the dinoflagellate host as well as other sources (Nosenko et al. 2006; Patron et al. 2006). These results suggested that the haptophyte-derived plastid might have coexisted for some time side by side the original dinoflagellate peridinin plastid (Patron et al. 2006) or that the host might have acquired some of the genes for the plastid chimeric proteome through HGT by enduring mixotrophy (Nosenko et al. 2006).

Interestingly, the bipartite targeting signals of these proteins included a typical SP followed by a cTP that differed from cTPs in both haptophytes and dinoflagellates in its lack of net positive charge, the phenylalanine at position +1 or nearby and the FVAP-domain (Patron et al. 2006). While the EGT from the haptophyte endosymbiont to the dinoflagellate host in *K. micrum* has played a significant role in restructuring the plastid proteome, it has not affected at all the mitochondrial proteome of the dinoflagellate host (Danne et al. 2011).

Tertiary endosymbiosis in dinotoms

One of the most extraordinary instances of tertiary endosymbioses is found in the so called dinotoms, a small group of dinoflagellates that harbor a diatom endosymbiont. Dinotoms have a wide distribution around the world. With only 10 or so described members, dinotoms are amazingly diverse: some live in fresh water, but most are marine species; some are benthic, some planktonic; some are thecate, some naked; some are dominantly motile, some are mainly sessile. Their hosts are classified under several different dinoflagellate genera (Carty and Cox 1986; Dodge 1971; Horiguchi and Pienaar, 1991, 1994; Pienaar et al. 2007; Tamura et al. 2005; Tomas et al. 1973; Tomas and Cox, 1973; Zhang et al. 2011) while their endosymbionts seem to belong to a few different diatom taxa (Chesnick et al. 1997; Horiguchi and Takano 2006; Horiguchi 2004; Imanian and Keeling 2007; McEwan and Keeling 2004; Pienaar et al. 2007; Takano et al. 2008). The union of dinoflagellates and diatoms in dinotoms is in itself bewildering. Diatoms constitute one of the most diverse and influential microscopic phytoplankton groups, with about 200,000 species (Armbrust et al. 2004; Falciatore and Bowler 2002a; Mann and Droop 1996) and with an annual organic carbon output rivaled only by the combined efforts of all terrestrial rainforests (Field et al. 1998; Mann 1999). Dinoflagellates make up another diverse and cosmopolitan group of algae with about 2,000 classified autotrophic, mixotrophic or

heterotrophic species, living free or as the symbiont or parasite of others (Taylor 2004; Taylor et al. 2007). Both of these remarkably intricate, impressively diverse and ecologically important groups of algae have acquired their own plastids secondarily from a red alga (Moore et al. 2008; Keeling 2008; Janouskovec et al. 2010). These two very complicated eukaryotic cells have come together in dinotoms, generating a rare, confounding and intriguing complexity.

The obligate and permanent relationship between the diatom endosymbiont and its dinoflagellate host in dinotoms has been well studied and documented in at least two species, *Kryptoperidinium foliaceum* and *Durinskia baltica* (Chesnick and Cox 1987, 1989; Figueroa et al. 2009; Tippit and Pickett-Heaps 1976; Tomas and Cox, 1973). The endosymbiont is ever-present in all different stages of the dinoflagellate host's life cycle, sexual and asexual, in the vegetative cell, the gametes, the zygotes and the cysts (Chesnick and Cox 1987, 1989; Cox and Rizzo 1976; Dodge 1971; Jeffrey and Vesk 1976; Kite and Dodge 1985; Tippit and Pickett-Heaps 1976; Tomas and Cox 1973; Figueroa et al. 2009). Like other endosymbionts, the diatom endosymbiont in dinotoms has experienced reduction. In two extreme cases, a strain of *K. foliaceum* isolated from South Carolina and *Peridiniopsis niei* from China, the diatom endosymbiont of the dinotom seems to have completely lost its nucleus, but in these cases no information about the retained plastid has been provided (Kempton et al. 2002; Zhang et al. 2011). The characteristic diatom cell wall and motility are lost in all dinotom endosymbionts. Also, in most dinotoms, while the host nucleus undergoes normal dinoflagellate mitosis, the endosymbiont nucleus does not: the chromosomes do not condense, and neither a spindle apparatus nor any microtubules are observed. The mitotic division and perhaps meiosis do not occur. The amitotic division of this nucleus during and as a result of the cytokinesis of the host

cell produces unequal daughter nuclei (Chesnick and Cox 1987, 1989; Figueroa et al. 2009; Tippit and Pickett-Heaps 1976).

What differentiates this tertiary diatom endosymbiont from other known endosymbionts is the retention of many of its original features and characters including a large nucleus, all the plastids with the expected four surrounding membranes, the outermost of which is continuous with the nuclear envelope, the endoplasmic reticulum (ER), many mitochondria with tubular cristae, ribosomes, dictyosomes, a large cytoplasm and a single membrane that separates it from its host (Tomas et al. 1973; Tomas and Cox 1973; Schnepf and Elbrachter 1999; Jeffrey and Vesk 1976; Dodge 1971; Cox and Rizzo 1976). Each one of these features is unique and is found only in the diatom endosymbiont of dinotoms. The nucleus of this endosymbiont is much larger than the inconspicuous nucleomorphs of either chlorarachniophytes or cryptophytes, and it contains huge amounts of DNA (Kite et al. 1988), roughly 700 × more than that in one of its closest free-living relatives, the pennate diatom *Phaeodactylum tricornutum*. The stable maintenance of its own mitochondria is not seen in any other endosymbiont. This has generated an exceptional mitochondrial redundancy in dinotoms not found in any other cell. The dinotom endosymbiont has also retained more membranes than any other secondary or tertiary endosymbionts, the extra membrane most likely being its own cell membrane (Eschbach et al. 1990). Interestingly, the host in dinotoms retains most of the ultrastructural features found in other autotrophic, mixotrophic and heterotrophic dinoflagellates, including a dinokaryon with its permanently condensed chromosomes, an intricate endomembrane system, conspicuous pusules, trichocysts, accumulation bodies, and in most cases also a triple-membraned eyespot, thought to be the relic of the original dinoflagellate plastid (Cavalier-Smith 1993; Cox and Rizzo 1976; Dodge 1971; Horiguchi 2004; Jeffrey and Vesk 1976; Schnepf and Elbrachter 1999; Tomas et al.

1973; Tomas and Cox 1973). In sum, dinotoms are among the most complicated cells, with at least five DNA-containing compartments: a plastid, two mitochondrial and two nuclear genomes (Figure 1.1).

The dinotom host species for which the data are available appear as closest relatives of each other in small subunit ribosomal DNA (SSU rDNA) and cytochrome c oxidase subunit 1 (Cox1) phylogenetic trees (Inagaki et al. 2000; Tamura et al. 2005; Imanian and Keeling 2007), sometimes to the exclusion of other dinoflagellates from the same genus (Takano et al. 2008). In constructed phylogenetic trees (i.e. SSU rDNA, rbcL, α-Tubulin, Actin, Cox1, Cox2, Cox3, Cob, and mitochondrial LSU rDNA) most dinotom endosymbionts group with pennate diatoms (Chesnick et al. 1997; Imanian and Keeling 2007; McEwan and Keeling 2004; Pienaar et al. 2007). However, the endosymbionts of *Peridinium quinquecorne* and three *Peridiniopsis* species in SSU rDNA, rbcL and internal transcribed spacer region (ITS rDNA) trees group with centric diatoms (genus *Chaetoceros*, *Thalassiosira* or *Discostella*) (Horiguchi and Takano 2006; Takano et al. 2008; Zhang et al. 2011).

Research objectives

Although dinotoms, especially *K. foliaceum* and *D. baltica*, had attracted a great deal of attention, and a wealth of ultrastructural information was available for most of them, many important questions had remained unexplored and unanswered especially at genetic or genomic level. Prior to this study, only a handful of nuclear genes (i.e. *ssu rDNA*, *lsu rDNA*, *actin*, *α-tubulin* and *hsp90*), plastid genes (i.e. *rbcL*), and mitochondrial genes (i.e. *cox1-3*, *cob*, *lsu rDNA*, *ssu rDNA*) were sequenced from a few of these organisms. There were also no complete organelle genomes available for any of the other dinoflagellates with tertiary endosymbionts. The need for having more insight into the complex genome of dinotoms is better understood in

the context of endosymbiosis, the process that has given rise to indispensable eukaryotic organelles (mitochondria and plastids) and to certain extent to protist diversity. With their well-preserved endosymbiont, dinotoms epitomize an earlier transitional stage in the complicated process of transformation of a free-living eukaryote to an organelle. They, therefore, present a rare, if not unique, opportunity to study endosymbiosis in its initial stages. In order to examine the impact of endosymbiosis on the genome content and structure of mitochondria and plastids in these extraordinarily complex cells and the contribution of the host nuclear genome to the proteomes of the two organelles, I conducted the following three projects:

1. Complete sequencing of the plastid genomes of *D. baltica* and *K. foliaceum*

From the three examples of tertiary plastids mentioned earlier, none had been completely sequenced. Except a few genes from the haptophyte plastids of *Karenia* and *Karlodinium* and the cryptophyte plastids of *Dinophysis* species no genetic data were available for these rare plastids. Complete sequencing of the plastid genomes of *K. foliaceum* and *D. baltica* had the potential to provide the first insights into these genomes. Comparing the genome content and structure of these plastids with that of the free-living diatoms such as *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*, for which the complete plastid genome was available, could indicate how endosymbiosis might have impacted the evolution of genomes in tertiary plastids, and more specifically whether they had experienced any reduction, gene loss or degradation, and genome rearrangements. Additionally, since it had been proposed that *D. baltica* and *K. foliaceum* acquired their pennate diatom endosymbiont prior to their divergence (Imanian and Keeling 2007; Inagaki et al. 2000), comparing these two plastid genomes could reveal how similarly or differently they had evolved in parallel after speciation.

2. Complete and/or mass sequencing of the mitochondrial genomes of the endosymbiont and the host in *D. baltica* and *K. foliaceum*

The tertiary plastids are rare, but more unusual are the mitochondria of the tertiary endosymbionts of dinotoms since none of the other known secondary or tertiary endosymbionts have retained their own mitochondria. The genomic data from these rare mitochondria, just like the plastids, could shed light not only on the organizational properties of these uncommon organelles, their content, and their possible reduction, expansion or degeneration, but also on their parallel evolution and their conformity to or deviation from those of the free-living diatoms.

Dinoflagellates have unusual mitochondrial genomes. In terms of gene content, the dinoflagellate mitochondrial genomes are among the smallest with only three protein-encoding genes: *cox1*, *cox3*, and cytochrome b (*cob*) (Nash et al. 2007, 2008; Norman and Gray 2001). The mitochondrial ribosomal RNA genes in dinoflagellates, like those of their sister group apicomplexans, are highly fragmented, and only several of these fragments had been identified prior to this study (Kamikawa et al. 2007; Waller and Jackson 2009). The transcripts of these genes are also extensively edited, with some editions occurring uniquely in dinoflagellates (Lin et al. 2002; Zhang and Lin 2005). A large-scale survey of the mitochondrial genome of the host in *D. baltica* and *K. foliaceum* could disclose whether the coexistence of this dinoflagellate organelle with its diatom counterpart over evolutionary time had affected its genome and its organization, and more specifically whether there was any sign of reduction, gene loss or degeneration and genome remodeling.

3. A survey of the host nuclear transcriptome in *D. baltica*

The permanent endosymbiosis has generally been associated with the drastic reduction of the endosymbiont and EGT to the host nuclear genome, and this has been shown also in the

dinoflagellates with the tertiary haptophyte endosymbionts, where many host nuclear-encoded plastid targeted proteins have been identified through large-scale transcriptome surveys (Nosenko et al. 2006; Patron et al. 2006). Dinoflagellates have very large nuclear genomes, and no dinoflagellate genome has been sequenced to date. In recent years and as an alternative to the whole genome sequencing, several large-scale dinoflagellate EST projects were completed (Bachvaroff et al. 2004; Hackett et al. 2005; Hackett et al. 2004; Leggat et al. 2007; Nosenko and Bhattacharya 2007; Patron et al. 2005; Patron et al. 2006; Sanchez-Puerta et al. 2007). The transcriptome survey of *D. baltica* could provide an additional set of data of this sort to help identify the expressed gene content of dinoflagellate genomes as a whole. More importantly, a survey of the host nuclear transcriptome could reveal whether the host nucleus in *D. baltica* was the recipient of any transferred genes from the diatom endosymbiont. There is little doubt about the permanence of the relationship between the diatom endosymbiont and the dinoflagellate host in dinotoms including *D. baltica*, yet the dinotom endosymbiont uniquely retains most of its ancestral features. The transcriptome survey of the host in *D. baltica* could show whether its nuclear genome contributed to the proteomes of its own and its endosymbiont mitochondria and the plastid, and if so, to what extent.

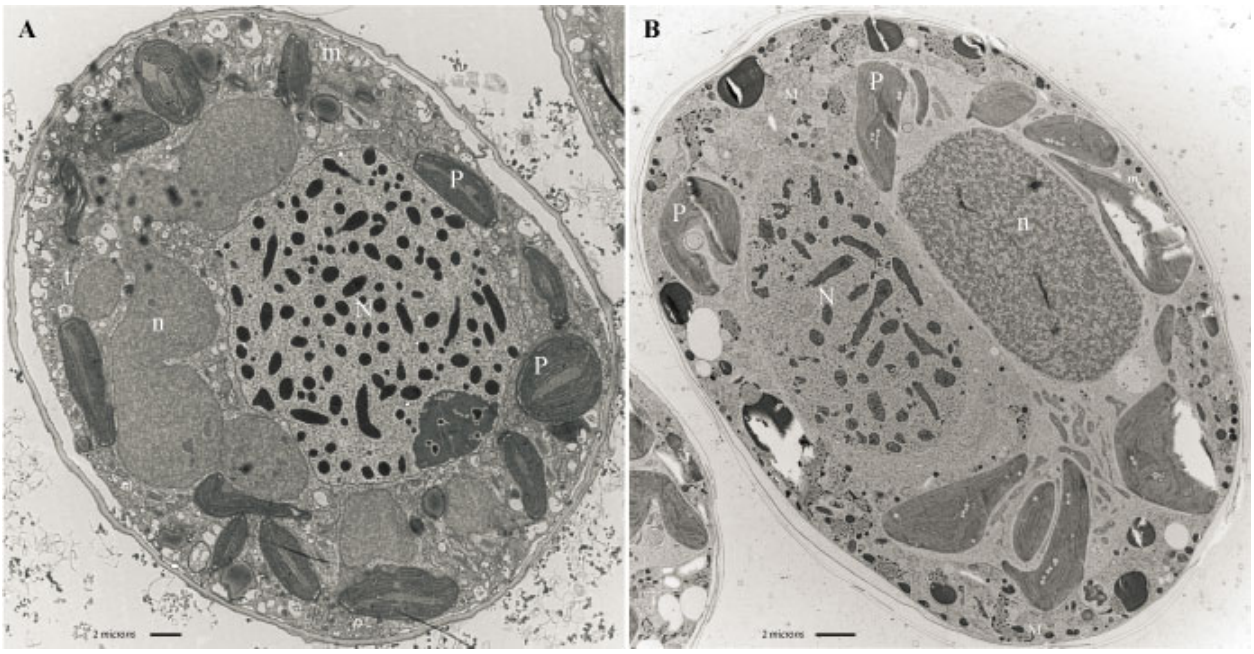

Figure 1.1: Transmission electron micrographs of *Kryptoperidinium foliaceum* (A) and *Durinskia baltica* (B).

The nucleus of the host (N) with its permanently condensed chromosomes, the nucleus of the endosymbiont (n), host mitochondria (M), endosymbiont mitochondria (m), host trichocysts (t) and plastids (P), which are all within the endosymbiont cytoplasm, are visible. Courtesy of Kevin Carpenter, Patrick Keeling and BI.

Chapter 2: The plastid genomes of two dinotoms

Introduction

The path of plastid evolution has been neither simple nor linear, but rather full of twists and turns. After the divergence of glaucophytes, red and green algae following primary endosymbiosis, plastids spread by the secondary and tertiary uptake of these eukaryotic algae by new eukaryotic hosts (Archibald and Keeling 2002; Bhattacharya et al. 2004; McFadden 2001; Palmer 2003). Each of these endosymbiotic events involved a massive loss of genes from the symbiont as well as a large scale transfer of other genes to its new host. In the primary endosymbiosis this meant gene transfers from the ancient cyanobacterium, whereas in secondary and tertiary endosymbioses most gene transfer would be from the nucleus of the endosymbiont alga to the nucleus of its new host (Archibald et al. 2003; Deane et al. 2000; Patron et al. 2006). The products of many of these genes would be targeted to the plastid, which necessitated the development of a new protein targeting system to direct the protein products back to their correct location (Bruce 2001; Jarvis and Soll 2002).

These processes have been most thoroughly studied in primary and secondary plastids, but tertiary endosymbioses add another layer of complexity to the process. In tertiary endosymbiosis an alga with a secondary plastid is taken up by another eukaryote, and to date the only lineage known to take up tertiary plastids is dinoflagellates, where tertiary plastids derived from three different lineages are known: *Karenia* and *Karlodinium* species with plastids derived from a haptophyte (Patron et al. 2006; Tengs et al. 2000); *Dinophysis* species with cryptophyte derived-plastids (Hewes et al. 1998; Schnepf and Elbraechter 1988; Hackett et al. 2003); and a small but growing group of dinoflagellates harboring a diatom endosymbiont (Dodge 1971; Horiguchi and Pienaar 1991, 1994; Tamura et al. 2005; Tomas and Cox 1973), which we refer to

as dinotoms. By dinotoms, we will refer to the whole biological system that includes both the dinoflagellate host and the diatom endosymbiont.

Dinotoms are widely distributed in both freshwater and marine environments and some, most notably *Kryptoperidinium foliaceum* and *Peridinium quinquecorne*, form blooms with occasional harmful effects (Demadariaga et al. 1989; Kempton et al. 2002; Garate-Lizarraga and Muneton-Gomez 2008). The Dinoflagellate host components are currently divided into at least five distinct genera, *Kryptoperidinium*, *Durinskia*, *Peridinium*, *Gymnodinium*, and *Galeidiniium* (Carty and Cox 1986; Dodge 1971; Horiguchi and Pienaar 1991, 1994; Tamura et al. 2005; Tomas and Cox 1973), while the endosymbiont components have been shown to originate from three different diatom lineages, one pennate (Chesnick et al. 1997; Imanian and Keeling 2007; McEwan and Keeling 2004; Pienaar et al. 2007) and two centric (Horiguchi and Takano 2006; Takano et al. 2008).

In haptophyte and cryptophyte endosymbiont-containing dinoflagellates, the endosymbiont has reduced to the point that only the plastid itself remains. In contrast, the diatom endosymbionts in dinotoms have preserved more of their genetic and cellular identity than any other secondary or tertiary plastid. The endosymbiont has lost some characters such as its cell wall, motility, and the ability to condense its chromosomes normally or divide mitotically (Chesnick and Cox 1989; Tomas and Cox 1973; Tippit and Pickett-Heaps 1976), but it retains a large nucleus and the nuclear genome, mitochondria and the mitochondrial genome (Imanian and Keeling 2007; Imanian et al. 2007), as well as cytosolic ribosomes, endoplasmic reticulum (ER), and dictyosomes in an extensive cytoplasm that is separated from the host by a single membrane (Eschbach et al. 1990; Tomas and Cox 1973). Despite such unusual degree of character retention, the endosymbiont is permanently integrated within its host, and it is present at all

different stages of the life cycle including cell division, sexual reproduction, and cyst formation (Chesnick and Cox 1989, 1987; Figueroa et al. 2009).

The number of plastids in dinotoms varies from one or two (in gametes) to as many as 30 to 40 (in zygotes). Chlorophyll *a*, *c1*, and *c2* are among the plastid pigments found in the best-studied dinotoms, *K. foliaceum* and *Durinskia baltica* (Jeffrey et al. 1975; Withers et al. 1977). The main carotenoid in the plastids of these two dinotoms is identified as fucoxanthin (Jeffrey et al. 1975; Kite and Dodge 1985; Mandelli 1968; Withers et al. 1977) as expected of a diatom and opposed to peridinin, which is the typical plastid carotenoid in dinoflagellate plastids (Schnepf and Elbrachter 1999). The peripherally distributed plastids are enclosed in the endosymbiont ER (which is continuous with the nuclear envelope), and retain thylakoids in stacks of three, girdle lamellae, and an internal pyrenoid (Horiguchi and Pienaar 1991, 1994; Tamura et al. 2005; Jeffrey and Vesk 1976; Tomas and Cox 1973).

Although tertiary endosymbiosis has been subject to a good deal of investigation in recent years, the actual genomes of tertiary plastids have received little attention, and to date no tertiary plastid genome has been sequenced from any lineage. Here, we describe the complete plastid genomes from two dinotom endosymbionts, *K. foliaceum* and *D. baltica*, in order to investigate the impact of tertiary endosymbiosis on the content and organization of these genomes. By comparing these genomes with each other and with available plastid genomes from free-living diatoms we find that the tertiary endosymbiosis has led to little change in either form or content of the plastid genome. However, the plastid genome of the endosymbiont of *K. foliaceum* is much larger than that of either free-living pennate diatoms or *D. baltica*, apparently due to the acquisition, incorporation, and maintenance of integrase/recombinase-encoding plasmid-like elements that are sporadically distributed in other heterokonts.

Results

Genome structure, gene repertoire, and GC content of the *D. baltica* and *K. foliaceum* genomes

The *D. baltica* CSIRO CS-38 plastid genome (GenBank: GU591327) assembly contained 18,704 Titanium pyrosequencing 454 reads (363 bp average), amounting to 6.8 Mbp, or 58–fold coverage of the genome. The *K. foliaceum* CCMP 1326 plastid genome (GenBank: GU591328) assembly included 7,274 reads (383 bp average) amounting to 2.8 Mbp, or 20-fold coverage. Over 20 kb of the *D. baltica* and 75 kb of *K. foliaceum*'s plastid genome sequences were also ascertained by PCR and Sanger sequencing (see Methods).

The *D. baltica* and *K. foliaceum* plastid genomes (Figure 2.1) map as circular molecules divided into large single-copy (LSC) and small single-copy (SSC) regions by the two inverted repeats (IRs), a quadripartite structure that is common to many other algal plastid genomes including the pennate and centric diatoms *P. tricornutum* and *Thalassiosira pseudonana*, respectively (Oudot-Le Secq et al. 2007). The general characteristics of all diatom and diatom-derived plastid genomes are juxtaposed in Table 2.1.

Like other related plastids, both dinotom plastid genomes use standard plastid/bacterial genetic code, with GTG as alternative start codon to ATG. This alternative start codon is found in the same four plastid genes (*rbcS*, *rpl23*, *rps8*, and *rpl3*) in all four diatom and diatom-derived plastid genomes.

The IRs in *D. baltica* are very similar to those of the free-living diatom *P. tricornutum* and feature almost the same gene composition (*trnP*, *ycf89*, *rrs*, *trnI*, *trnA*, *rrl*, and *rrn5*) and size. The slight difference in the size and composition of the IRs in these two plastids is due to the presence of *psbY* in the IRs of *P. tricornutum* instead of partial *ccsA* in the IRs of *D. baltica*.

The plastid genome size and gene content of *D. baltica* are remarkably similar to those of *P. tricornutum*. The *D. baltica* plastid genome is only about 900 bp shorter than that of *P. tricornutum*, and the two genomes share 159 genes in common. The *D. baltica* plastid genome encodes 127 protein-coding genes, three rRNAs, 27 tRNAs, a sufficient set for their plastid protein synthesis machinery, one transfer-messenger RNA (tmRNA), *ssra*, and one plastid signal recognition particle RNA, *ffs*. Interestingly, like *P. tricornutum* it has retained *syfB*, encoding a trnF synthetase, which is missing from the plastid genome of *T. pseudonana* but is present in red algal plastid genomes (Oudot-Le Secq et al. 2007). Only three genes present in the plastid genome of *P. tricornutum* are absent from the *D. baltica* genome: *tsf* (not found in other diatom plastid genomes), *acpP*, and *ycf42*.

In contrast, the *K. foliaceum* plastid genome is considerably larger than the plastid genomes of *D. baltica* and *P. tricornutum*, by about 24 and 23 kb, respectively. The IRs in *K. foliaceum* are shorter than those of *D. baltica* and *P. tricornutum* by almost 1 kb because of the absence of *trnP* and *ycf89* in the *K. foliaceum* IRs, so its larger size is not due to the increased size of the IRs as seen in *T. pseudonana* (Oudot-Le Secq et al. 2007). Instead, both SSC and LSC in *K. foliaceum* are sizably larger than those observed in other diatoms, owing to the presence of more apparently non-coding DNA (see below) and protein-coding genes. In addition to the same 159 genes found in both *D. baltica* and *P. tricornutum*, the plastid genome of *K. foliaceum* encodes a putative tyrosine recombinase gene, *tyrC*, two putative serine recombinase genes, *serC1* and *serC2*, two smaller ORFs, ORF93 and ORF92 both related to *serC1*, and seven putative open reading frames (ORFs) larger than 150 amino acids (aa), or 15 ORFs if the threshold for annotation is lowered to 100 aa.

Compactness of dinotom plastid genomes

Like other chromist plastid genomes, the plastid genomes of the two dinotoms possess some of the features of a compact genome. They lack introns, and the same four overlapping pairs of genes found in diatoms (Oudot-Le Secq et al. 2007) are also found in both dinotoms with the identical length of overlap: *psbD-psbC*, *atpD-atpF*, *sufC-sufB*, and *rpl4-rpl23* with 53, 4, 1, and 8 nucleotides (nt) overlap, respectively. In addition, *dnaB* and *trnF* have no intergenic spacer in *D. baltica* and *P. tricornutum*, whereas this gene pair is separated by 1 nt in *K. foliaceum*. Similarly, *rpl14* and *rpl24* are separated by a single nt in *D. baltica*, *K. foliaceum*, and *P. tricornutum*.

The plastid genomes of *D. baltica*, *K. foliaceum*, and *P. tricornutum* demonstrate no considerable change in the length of their genes (Figure 2.S1). Out of the common 159 genes, 108 are invariant in length and the sum of all differences between *P. tricornutum* genes and those of *D. baltica* and *K. foliaceum* amount to a mere 199 and 142 bp, respectively (and only 57 bp between *K. foliaceum* and *D. baltica*; Figure 2.S1).

Average intergenic spaces in *D. baltica* (94.3 bp) are only slightly longer than those of *P. tricornutum* (88.4 bp), but in *K. foliaceum* the spacing is more than twice as long (246.7 bp on average) (Table 2.1). Even when putative ORFs in *K. folicaeum* are brought into account, the average spacing is 180 bp, but more importantly when the average is calculated based only on the 159 shared genes, the average is only 94.1 bp, about equivalent to *D. baltica* and *P. tricornutum*.

Conserved ordered gene blocks

To investigate the conservation of genome structure, MAUVE (Darling et al. 2004) was used to detect gene clusters. Overall, 23 conserved clusters were found in *T. pseudonana*, *P.*

tricornutum, *D. baltica*, and *K. foliaceum*. If *T. pseudonana* (a more distantly related centric diatom) is removed from analysis, 14 larger blocks are found. In pairwise comparisons, nine large conserved blocks are shared between *P. tricornutum* and *D. baltica*, 13 between *P. tricornutum* and *K. foliaceum*, and nine between *D. baltica* and *K. foliaceum*. However, taking into account the presence or absence of a single gene between large blocks extends these blocks (to 16 conserved blocks among the three species amounting to more than 108 kb, 10 blocks between *P. tricornutum* and *D. baltica*, 14 blocks between *P. tricornutum* and *K. foliaceum*, and 13 blocks between *D. baltica* and *K. foliaceum*) (Figure 2.2). The largest block conserved among the three species spans more than 31 kb and includes 46 genes appearing in the same order, encoded on the same strands (*ycf33*, *trnI*, *trnS* … *rpoC1*, *rpoC2*, *rps2*). The largest conserved gene block between *P. tricornutum* and *D. baltica* is about 33 kb and contains 51 genes (*rpl32*, *trnL*, *rbcR* … *rps7*, *tufA*, *rps10*). This conserved gene block is broken into four smaller, dispersed blocks of genes in *K. foliaceum* (*rpl32-psbA* and *ycf35-psb28*, which are also inverted; *trnQ-groEL*; and *dnaK-rps10*). There are two small blocks of tRNAs (*trnR*, *trnV*, *trnY*, and *trnL*, *trnC*) that are conserved in three species, but they are inverted in *D. baltica* and *K. foliaceum* with respect to *P. tricornutum*. Similarly, two small conserved blocks of genes (*rpl20*, *rpl35*, *ycf45* and *psbC*, *psbY*) appear in inverted orientation in *D. baltica* with respect to the other two species.

To see how the organization of blocks of shared genes might have evolved, GRIMM (Tesler 2002) was used to identify 14 inversions in the transition of the three plastid genomes of *P. tricornutum*, *D. baltica*, and *K. foliaceum*. If *T. pseudonana* is added, 23 inversions are required. In pairwise analyses, GRIMM also proposes 6 inversions for *P. tricornutum* and *D. baltica*, 9 for *P. tricornutum*, and *K. foliaceum*, and 8 for *D. baltica* and *K. foliaceum*.

Closer manual inspections reveal that compared to the plastid genome of *P. tricornutum* fewer rearrangements of the conserved gene blocks distinguish *D. baltica* from *K. foliaceum*: only three inversions (blocks 2, 8, and 9) and two translocations (block 10 and *clpC* gene) are detected in *D. baltica* versus two inversions (blocks 1, 4), six inversions/translocations (blocks 10, 11, 7, 6, 8, and 12) and three translocations (blocks 9, 13, and *clpC* gene) in *K. foliaceum* (Figure 2.2). Compared to the plastid genome of *D. baltica, K. foliaceum* shows two inversions (blocks 1, 4), five inversions/translocations (blocks 6, 10, 9, 8, and 7) and two translocations (blocks 12, and 11).

All the three missing genes from the plastid genomes of *D. baltica* and *K. foliaceum*, present in *P. tricornutum,* are located in its LSC region. Curiously, however, most of the rearrangements seem to have occurred in the SSC regions of the plastid genomes of *D. baltica* and *K. foliaceum* (Figure 2.2).

Low gene density regions of the *Kryptoperidinium foliaceum* plastid genome

There are nine distinct regions (labeled with Roman numerals in Figures 2.1 and 2.2) within the *K. foliaceum* plastid genome that have a low gene density and do not show any similarity to *D. baltica, P. tricornutum*, or *T. pseudonana*. Six of the nine regions are dispersed within the SSC (regions III-VIII, totaling to more than 17 kb) and three within the LSC (regions I, II, and IX, amounting to about 7.5 kb). All four junctions of the IRs with the SSC and LSC include such regions: II and IX at the boundary of IRa and LSC, and III and VIII at the junction of IRb and SSC. These nine distinct regions collectively amount to more than 24 kb ranging in size from 905 bp (region IX at the boundary of IRa and LSC) to 4852 bp (region III at the junction of IRb and SSC) with an overall GC content of 30.4%, which is 2% lower than the GC content of the genome as a whole (Table 2.1), and 2.4% lower than the rest of the genome.

Interestingly, regions I and III are each bounded by two imperfect palindromes. A 35 bp palindrome is located near the *rps2* gene and a 44 bp palindrome is located at its other end, near *rbcS*. Region III is similarly bounded by two palindromic sequences: a 25 bp sequence near the *rrn5* gene and a 42 bp sequence near *dnaK*. Another 32 bp palindrome is close to one end of region V (near *psbA*).

The *tyrC* gene in *K. foliaceum* and *Heterosigma akashiwo*

The *tyrC* gene located in region III shows strong similarity to a putative site-specific tyrosine recombinase protein (TyrC) encoded within the plastid genome of the raphidophyte heterokont *H. akashiwo* (Cattolico et al. 2008). The conceptual translation of *tyrC* also shows similarity, albeit much weaker, to putative integrase/recombinase proteins encoded in the plastid genome of the chlorophycean alga *Oedogonium cardiacum* and in the mitochondrion of the charophyte *Chaetosphaeridium globosum*. As revealed by NCBI Conserved Domain Database (CDD) searches (Marchler-Bauer et al. 2009), the *K. foliaceum* TyrC has conserved all the major catalytic, active, and DNA-binding sites required by this protein for integrase/recombinase activity, including His 250 and the four invariably conserved sites Arg 145, Arg 253, Lys 172, and Tyr 285 (Figure 2.S2) (Esposito and Scocca 1997; Friesen and Sadowski 1992; Han et al. 1993). RT-PCR was performed on *tyrC* and the amplicon sequenced (data not shown), confirming that this gene is transcribed and most likely expressed in the *K. foliaceum* plastid genome.

Similarity between the *K. foliaceum* plastid genome and pCf1 and pCf2 plasmids in *Cylindrotheca fusiformis*

A total of five ORFs (*orf141*, *serC1* (*orf205*), *serC2* (*orf212*), *orf93*, and *orf92*) in the *K. foliaceum* plastid genome show strong similarity to ORFs found in the pCf1 and pCf2 plasmids

of the pennate diatom *C. fusiformis*. Each of these two plasmids includes several ORFs, two pairs of which share considerable similarity (ORF217 of pCf2 and ORF218 of pCf1 with almost 80% aa identity and ORF484 of pCf2 and ORF482 of pCf1 with 54%) (Hildebrand et al. 1992). *K. foliaceum* ORF141 (region VI) shares 57% and 47% aa identity with ORF484 (aa 186 to aa 324) from pCf2 and ORF482 from pCf1 plasmid, respectively. The *K. foliaceum* SerC1 shares 76% and 66% aa identity with ORF218 from pCf1 and ORF217 from pCf2, respectively, while SerC2 displays 60% aa identity with *C. fusiformis* ORF218 and 61% with ORF217. Interestingly, SerC1 and SerC2 share less similarity to each other (57% of aa identity) than they do with *C. fusiformis* ORF218 and ORF217, and *serC2* also shares a single codon insertion specifically with ORF217. *K. foliaceum orf93* and *orf92* (region I) appear to be truncated versions of the *C. fusiformis* ORF218, corresponding to amino acids 1 to 93 and 117 to 206, respectively. The two *K. foliaceum* fragments are separated from each other by 69 bp, the conceptional translation of which shares 87% identity with *C. fusiformis* ORF218 amino acids 95 to 116, however this region contains two stop codons suggesting it is a pseudogene.

CDD searches (Marchler-Bauer et al. 2009) reveal that SerC1 and SerC2 in *K. foliaceum* have retained almost all the catalytic, DNA-binding, presynaptic, and synaptic residues found in other site-specific serine recombinases (Figure 2.S3). Once again, RT-PCR (data not shown) showed that both *serC1* and *serC2* are transcribed and most likely expressed in *K. folicaeum*.

In addition to the five abovementioned ORFs, a number of dispersed non-coding stretches of DNA in several distinct regions of the *K. foliaceum* plastid genome show strong similarity to the *C. fusiformis* pCf1/pCf2 plasmid sequences. A 350-bp sequence in region III of *K. foliaceum*'s plastid genome (Figure 2.1) shows 92% nt identity to a portion of ORF484 (corresponding to aa 322 to 431) from pCf2. This 350-bp sequence does not include any ORF

but contains two stop codons in the same frame that shows similarity to ORF484. Similarly, in region VI about 600 bp immediately downstream of ORF141 shows 71% nt identity to *C. fusiformis* ORF484 and region IV contains 240 bp, 110 bp, and 120 bp sequences with strong similarity to non-coding regions of pCf2 (72%, 72% and 74% nt identity). There is also a 156 bp sequence in region VIII with strong similarity (66% nt identity) to the non-coding region at the end of pCf1.

We also searched the non-plastid 454 sequence data (both assembled and singleton reads) for potential plasmids similar to those of *C. fusiformis*, however, none were found. Since ORF218 and ORF482 are close to each other in pCf1 and similarly ORF217 and ORF484 are close in pCf2, we also designed outward primers for the two corresponding *K. foliaceum* ORFs (ORF205 or *serC1* and ORF141). All attempts to PCR amplify a small product were unsuccessful.

Discussion

The divergent evolution of two tertiary plastid genomes of diatom origin

The plastid genomes of the tertiary endosymbionts of *K. foliaceum* and *D. baltica* share numerous common features with those of free-living diatoms, including gene content, ordered gene blocks, and overall genome structure. Especially striking is the similarity to the pennate diatom *P. tricornutum*, with which they share more than 108 kb of syntenic gene clusters, reconfirming the pennate diatom ancestry for these endosymbionts also suggested by molecular phylogeny (Chesnick et al. 1997; Horiguchi and Takano 2006b; Imanian and Keeling 2007; McEwan and Keeling 2004; Pienaar et al. 2007). Recent phylogenetic analyses suggest a particularly close relationship with the genus *Nitzschia* (Imanian and Keeling 2007; Pienaar et al. 2007; Takano et al. 2008). Unfortunately, at present the only pennate diatom plastid genome

known is from the more distantly related *P. tricornutum*. Considering the high degree of conservation between its plastid genome content, composition, and organization and those of *K. foliaceum* and *D. baltica*, we suggest the plastid genome of closer free-living pennate diatom relatives will reveal even fewer structural changes have taken place since the tertiary endosymbiosis.

Notwithstanding the high degree of conservation between the tertiary plastid genomes and their free-living relative *P. tricornutum*, the plastid genome of *K. foliaceum* is different in one interesting respect. Its genome is more than 23 kb larger than those of its close relatives and the majority of the additional sequence falls into a handful of specific regions. Most of this sequence shows no strong similarity to known sequences, but a few regions share a strong similarity to the plasmids pCf1 and pCf2 in the pennate diatom *C. fusiformis*. The genome also encodes two site-specific serine recombinase genes also shared with those plasmids, as well as a site-specific tyrosine recombinase gene present in the plastid genome of another heterokont, the raphidophyte *H. akashiwo*.

Earlier hybridization experiments suggest that either pCf1/ pCf2 or plasmids with considerable sequence similarity existed in three strains of *C. fusiformis*, in one of three strains of *C. closterium*, in *Nitzschia angularis*, and in *N. curvilineata* (Hildebrand et al. 1992; Jacobs et al. 1992), but no sequence data were available to indicate the possible sites of hybridization or integration of such plasmids with the plastids. In *K. foliaceum* we find no evidence for the presence of intact plasmids, only the putative *serC1* and *serC2* genes and some degenerated fragments integrated into the plastid genome. Overall we can conclude that both plasmids were present in an ancestor of *K. foliaceum*, and fragments of both have persisted by integration into the plastid genome in *K. foliaceum*, but not *D. baltica*.

While the two serine recombinase genes in the *K. foliaceum* plastid genome clearly originated from plasmids and probably functioned in spreading those plasmids, the origin of the tyrosine recombinase/integrase is less clear. TyrC in *H. akashiwo* has been speculated to be involved in converting multimeric plastid molecules to monomeric forms (Cattolico et al. 2008), similar to what other recombinases do in certain bacteria with circular chromosomes. In *Escherichia coli*, homologous recombination of the two sister chromatids results in formation of a chromosome dimer, and reversion of the dimer to monomers before cell division is accomplished through the functions of two related recombinases, XerC and XerD (Barre et al. 2001; Blakely and Sherratt 1994; Lesterlin et al. 2004). These two proteins break and re-ligate DNA strands at conserved specific binding sites (dif), found in the chromosomal segregation region. The dif sites are usually 28 bp long with two arms, 11 bp each, separated by 6 bp in the center (Barre et al. 2001; Blakely and Sherratt 1994; Lesterlin et al. 2004). We did not find any sequences similar to the proposed dif sites of *H. akashiwo* or other known bacteria (Cattolico et al. 2008), but whether TyrC in *H. akashiwo* and *K. foliaceum* bind dif sites similar to those of bacterial or viral recombinases, and whether this protein is active in conversion of multimeric forms of plastid genome or even those of the plasmids to monomers are unknown. However, the conservation of all the active sites in the conceptual translation of *K. foliaceum tyrC* (Figure 2.S2) and its transcription both imply that the protein is functional. The presence of palindromic sequences at the boundaries of at least two of the distinct regions (I and III, the latter of which also contains a recombinase gene) in the plastid genome of *K. foliaceum* further suggests these elements may remain mobile by some means that generates such ends during movement/replication.

The ancestral state of the tertiary endosymbiont genome

Despite the more recent common ancestry of the *D. baltica* and *K. foliaceum* plastids, the *D. baltica* and *P. tricornutum* plastid genomes share a much greater overall similarity in structure, in large part due to the presence of plasmid-associated sequences in *K. foliaceum*. Which of these two tertiary endosymbionts better represents the state of their common ancestor is not entirely clear. On one hand, the close similarity between the genomes of *D. baltica* and *P. tricornutum* might suggest this represents the ancestral state and that the genome of *K. foliaceum* subsequently acquired plasmids (which are selfish and frequently mobile elements) leading to its expansion and reorganization. On the other hand the plasmids are known to exist in some form in other pennate diatoms that are more closely related to the tertiary plastids, most notably some *Nitzschia* species (Hildebrand et al. 1991). If they were in the ancestor of the tertiary endosymbionts then *D. baltica* would have to have ridded itself of all evidence of both plasmids to revert to a highly similar form as *P. tricornutum*. Both explanations, the multiple movements of plasmids between close relatives or the complete loss of plasmids in certain lineages, are consistent with the seemingly sporadic interspecies and intraspecies distribution of these plasmids in diatoms: only 5 out of 18 examined diatom species and only 1 out of 3 strains of the pennate diatom *C. closterium* are suggested to have similar plasmids (Hildebrand et al. 1991). Perhaps the most likely explanation is that the ancestor possessed unintegrated plasmids and a plastid genome with a structure highly similar to that of *P. tricornutum* and *D. baltica*. In *K. foliaceum* the plasmids would have integrated into the main plastid genome, degenerated, and promoted the reorganization of many gene blocks, whereas in *D. baltica* all traces of the plasmids were lost, which is not unlikely if they never integrated into the plastid genome.

Conclusions

Here we describe the first completely sequenced plastid genomes from tertiary endosymbionts, specifically the diatom-derived plastids of two dinoflagellates, *D. baltica* and *K. foliaceum*. Both genomes have retained many characteristics of the ancestral, free-living diatom, including elements of genome structure, gene content, and ordered gene clusters. The plastid genome of *K. foliaceum* is much larger than that of *D. baltica*, and contains a site-specific tyrosine recombinase gene also found in the heterokont *H. akashiwo*, and the incorporation, maintenance, and degradation of genetic material from two similar plasmids found in other pennate diatoms, which have resulted in the addition of two site-specific serine recombinases.

Materials and methods

Strains and culture conditions

Cultures of *Durinskia baltica* (*Peridinium balticum*) CSIRO CS-38 and *Kryptoperidinium foliaceum* CCMP 1326 were obtained respectively from the CSIRO Microalgae Supply Service (CSIRO Marine and Atmospheric Research Laboratories, Tasmania, Australia) and from the Provasoli-Guillard National Center for Culture of Marine Phytoplankton (West Boothbay Harbor, ME, USA). *D. baltica* cultures were maintained in GSe medium at 22°C (12:12 light:dark cycle) whereas *K. foliaceum* cultures were maintained in F/2-Si medium under the same conditions.

DNA and RNA extractions, PCR, RT-PCR, and DNA fractionation and precipitation

Cells were collected and ground as described previously (Imanian et al. 2007). Ground cells were lysed in 50 mM Tris-Hcl, 100 mM EDTA, 100 mM NaCl, pH 8.0 in the presence of β-mercaptoethanol (2%), SDS (2%) and proteinase K (300 μg/ml) at 50°C for 1 hour. In case of *D. baltica*, 6 phenol and 1 phenol/chloroform extractions were performed, whereas for *K.*

foliaceum, 3 phenol, 1 phenol/chloroform, and 2 chloroform extractions were conducted. Organellar A+T-rich DNA was separated from nuclear DNA using CsCl gradient density centrifugation. The initial CsCl reflective index was adjusted to 1.3995 and 1.4000 for *D. baltica* and *K. foliaceum* respectively and Hoechst 33258 (Invitrogen, Carlsbad, CA, USA) was added to the solution (100 μg/ml for *D. baltica* and 120 μg/ml for *K. foliaceum*). Ultracentrifugation was conducted in a Beckman L8 80M ultracentrifuge, using a VTi 80 (Beckman) rotor at 55000 rpm and 20°C for 22 and 20 hours for *D. baltica* and *K. foliaceum*, respectively. The extracted A+T-rich satellite bands were washed 4 times with CsCl/TE buffer-saturated isopropanol to remove the Hoechst dye. The DNA was precipitated from CsCl as described previously (Kite et al. 1988) and eluted in Tris HCl pH 8.0. The purified DNA was amplified using the REPLI-g mini kit (Qiagen, Missisauga, ON, Canada). The total genomic DNA from both species used for the PCR reactions was obtained after 2 phenol, 1 Phenol:Chloroform:Isoamyl Alcohol (25:24:1), and 2 chloroform extractions and ethanol precipitation. Total RNA extraction and RT-PCR were carried out as described previously (Imanian et al. 2007) using the following primers: tyrC_F, CCATAACTGCGTAATATAGCCG, tyrC_R, TCTGAAGGAATTAAATCTAATCAAGG, serC1_F, CCAGTTAACTTGCTACTGTCGG, serC1_R TTGGCTCTGCTGCTAACG, serC2_F TGTGTCTTCAAAGTCACAAGAGG, and serC2_R AACTAATCGGTTATATGGTATGTAATTCA. PCR was performed using the EconoTaq PLUS GREEN kit (Lucigen, Middleton, WI, USA).

Genome sequencing

The *D. baltica* and *K. foliaceum* plastid genomes were sequenced using massively parallel GS-FLX DNA pyrosquencing (Roche 454 Life Sciences, Branford, CT, USA). The GS-FLX shotgun libraries and pyrosequencing using the GS-FLX Titanium reagents were carried out

at the Génome Québec Innovation Centre. The Newbler *de novo* assemblies were edited and re-assembled with CONSED 19 (Gordon 2004). Plastid sequences in assembled and unassembled sequence pools were identified by BLAST searches (Altschul et al. 1990). Ambiguous pyrosequencing homopolymer stretches in the assemblies were verified by PCR/Sanger sequencing, which invariably yielded sequence that preserved the open reading frame. The only exceptions were fragments of plasmid-derived genes in the *K. foliaceum* plastid genomes that are concluded to be pseudogenes.

Genome annotation and analysis

Genes were identified by DOGMA searches (Wyman et al. 2004) and by BLAST homology searches (Altschul et al. 1990) against the NCBI nonredundant database (http://www.ncbi.nlm.nih/BLAST), and annotated using Artemis 11 (Rutherford et al. 2000). Protein-coding genes were identified using GETORF from EMBOSS 6.0.1 (Rice et al. 2000) and ORFFINDER at NCBI, with start codons ascertained by comparison with known homologues. Positions of tRNA-encoding genes were determined with tRNAscan-SE (Schattner et al. 2005). Ribosomal and miscellaneous RNA-encoding genes were annotated by comparison with *P. tricornutum* and *T. pseudonana* homologues. Repeated elements were searched for using PipMaker (Schwartz et al. 2000), REPuter (Kurtz et al. 2001), and FUZZNUC from the EMBOSS package (Rice et al. 2000). Physical maps were generated using GenomeVx (Conant and Wolfe 2008) and further edited manually. Conserved gene clusters between the *D. baltica*, *K. foliaceum* and *P. tricornutum* plastid genomes were identified using MAUVE (Darling et al. 2004) and by visual inspection of the physical maps. Hypothetical gene inversions between the 159 genes that are shared between the three genomes were examined using GRIMM (Tesler 2002). Translocations were identified by manual inspections and defined as homologous portions

of genomes (*i.e.* a gene or a conserved blocks of genes) appearing at different loci in the same orientation.

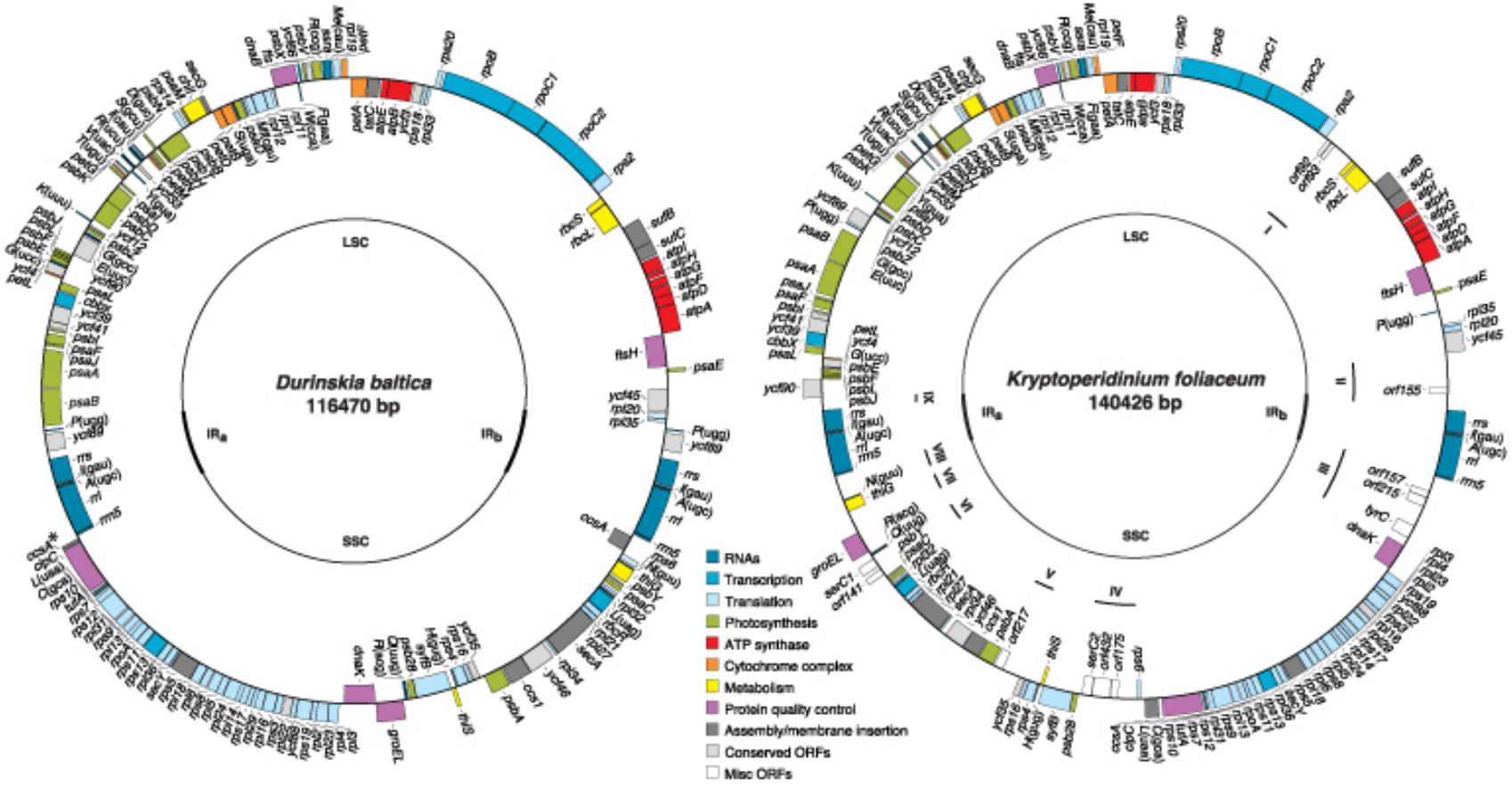

Figure 2.1: The plastid genome maps of *Durinskia baltica* and *Kryptoperidinium foliaceum*.

Functionally related genes are indicated by color and transcriptional direction is indicated by boxes outside the ring (clockwise) or inside the ring (counterclockwise). Genes for tRNAs are indicated by their single letter code. The large single copy (LSC), small single copy (SSC), and inverted repeats (IRa and IRb) are shown on the inner circle. Roman numerals (I-IX) mark the locations of 9 distinct regions in the plastid genome of *K. foliaceum*.

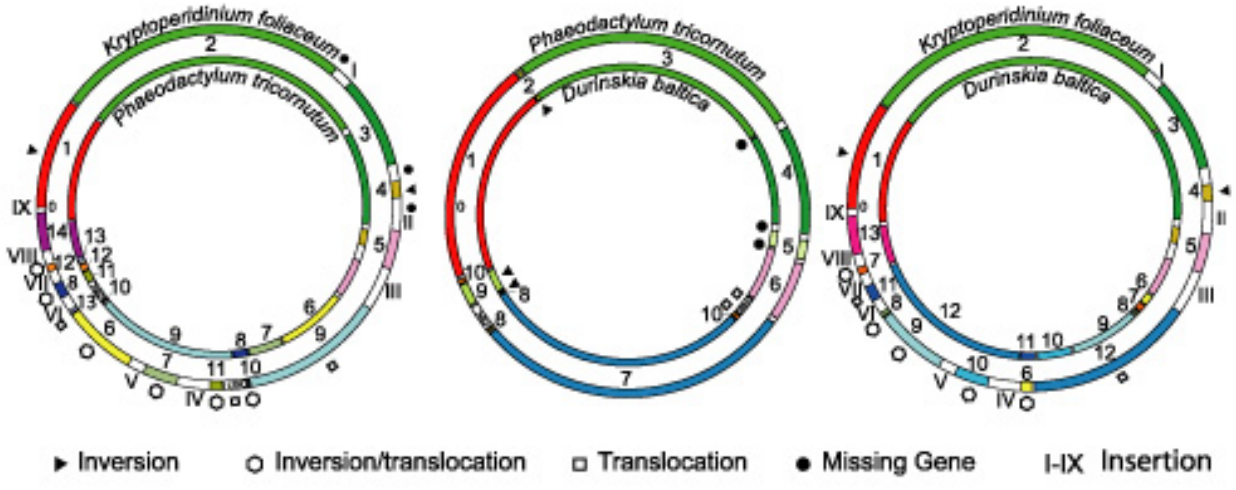

Figure 2.2: Conserved ordered gene blocks among three plastid genomes.

All possible two-way comparisons between plastid genomes of *K. foliaceum*, *D. baltica*, and *P. tricornutum*. Conserved blocks of genes are indicated by color, inversions are marked by a black triangle, inversions/translocations by a hexagon, translocations by a rectangle, missing genes by a black circle and insertions by Roman numerals I-IX.

Table 2.1: General characteristics of plastid genomes in dinotoms compared to diatoms

	Durinskia baltica	*Kryptoperidinium foliaceum*	*Phaeodactylum tricornutum* [a]	*Thalassiosira pseudonana* [a]
Size (bp)				
Total	116470	140426	117369	128814
IR	7067	6017	6912	18337
SSC	39813	56521	39871	26889
LSC	62523	71871	63674	65250
GC content (%)				
Total	32.55	32.4	32.56	30.66
rRNA genes	46.9	47.0	47.2	47.0
tRNA genes	53.5	53.7	53.0	52.6
Other RNAs	27.3	28.3	26.0	25.6
Protein-coding genes	32.4	33.0	32.9	31.5
Intergenic spacer [b]	22.1	26.5	18.8	16.3
Coding sequence (%) [c]	86.7	71.9	87.5	85.2
Gene content [d]				
Total	159	160	162	159
Protein-coding genes	127	128	130	127
rRNA genes	3	3	3	3
tRNA genes	27	27	27	27
Other RNAs	2	2	2	2
Introns	0	0	0	0
Overlapping genes	4	4	4	4
Average intergenic spacer (bp)	94.3	246.7	88.4	108.2
Start codons				
ATG	123	123	124	121
GTG	4	5	5	5
Other	0	0	1 ATT	1 ATA

[a] Data taken from Oudot-Le Secq *et al* (2007).
[b] Duplicated genes were taken into account (size/number of genes).
[c] Conserved genes (unique and duplicated) and ORFs were considered as coding sequences.
[d] Duplicated genes and unique ORFs were not taken into account.

Chapter 3: The mitochondrial genomes of the endosymbiont and host in two dinotoms

Introduction

Reduction is a universal theme in the symbiotic events that gave rise to mitochondrial and plastid diversity. In primary endosymbiosis, the α–proteobacterial and cyanobacterial ancestors of mitochondria and plastids were drastically reduced to organelles that encode only a small fraction of their original genes (Gray et al. 1999; Kaneko et al. 1996; Nierman et al. 2001; Palmer 2003). In plastid evolution, this was followed by further rounds of primary and secondary endosymbiosis. Secondary endosymbionts, derived from red or green algae, have also lost nearly everything except their plastids (Archibald and Keeling 2002; McFadden 2001), and even in those exceptions where secondary endosymbionts retained a miniature nucleus (nucleomorph), it is highly reduced and nearly all its cytoplasmic features are gone (Archibald 2007; Gilson and McFadden 2002; Gilson et al. 2006; Greenwood 1974; Lane et al. 2005). In tertiary endosymbionts generally only the plastids remains (Tengs et al. 2000), with one interesting exception, the so-called 'dinotoms'.

With 10 known species, dinotoms are a small group of closely related dinoflagellates whose endosymbionts are thought to belong to at least three different diatom clades (Horiguchi and Pienaar 1994, 1991; Pienaar et al. 2007; Takano et al. 2008; Tamura et al. 2005). Considering the small size of this group, dinotoms are very diverse in their morphologies (for example, with or without thecal plates with different plate configurations among the thecate species), their habitats (fresh water or marine environments), and their life styles (planktonic or

benthic, dominantly motile or prevailingly sessile), and have consequently been classified into five distinct genera.

The tertiary diatom endosymbiont of dinotoms has, like other tertiary endosymbionts' reduced to some degree: it has lost its distinctive cell wall, motility, and the ability to divide mitotically (Dodge 1971; Tomas and Cox 1973). Despite these losses and integration within its host, however, the endosymbiont has also retained many of its original characters, including a large nucleus with vast amounts of DNA, a large volume of cytoplasm separated from the host by a single membrane, and perhaps most surprisingly its own mitochondria (Chesnick and Cox 1987, 1989; Cox and Rizzo 1976; Imanian and Keeling 2007; Jeffrey and Vesk 1976; Tippit and Pickett-Heaps 1976).

In two dinotom species, *Durinskia baltica* and *Kryptoperidinium foliaceum*, it has been shown that the mitochondria of the endosymbionts still express genes for cytochrome c oxidase subunit 1 (*cox1*) and cytochrome b (*cob*) (Figueroa et al. 2009; Imanian and Keeling 2007). The host mitochondria in *D. baltica* also expresses *cox1* and *cob*, so this species at least is thought to possess uniquely redundant mitochondria (Imanian and Keeling 2007; Imanian et al. 2007). While diatom and dinoflagellate mitochondria are similar morphologically, they could not be more dissimilar in terms of genomic content and organization. Sequenced diatom mitochondrial genomes range from 43 to 77 kbp, have a circular map, and encode about 60 genes. While generally compact, they usually feature one large intergenic spacer composed of repetitive sequences (from nearly 5 kbp in the centric diatom *Thalassiosira pseudonana* and the araphid pennate diatom *Synedra acus*, to about 35 kbp in the raphid pennate diatom *Phaeodactylum tricornutum*) (Oudot-Le Secq and Green 2011; Ravin et al. 2010). In contrast, dinoflagellate mitochondria encode only three protein-coding genes (*cox1*, *cox3* and *cob*) and many fragments

of ribosomal RNA (rRNA), and these appear to be organised on multiple chromosomes that may be linear, and which are greatly expanded in number and include numerous incomplete copies or pseudogenes along with highly dispersed short or long stretches of non-coding and repetitive sequences (Jackson et al. 2007; Slamovits et al. 2007; Waller and Jackson 2009). The disposal of the canonical start and stop codons of the 3 protein-coding genes, trans-splicing of *cox3* in at least a few species, polyadenylation and editing of the mitochondrial transcripts are among other oddities observed in the dinoflagellate mitochondrial genomes (Gray et al. 2004; Jackson et al. 2007; Slamovits et al. 2007; Waller and Jackson 2009).

The co-occurrence of these two distinct mitochondria within dinotoms raises questions about whether or not either or both genomes have been reduced in any way due to this unique mitochondrial redundancy; or more specifically, do host and symbiont mitochondrial genomes encode a similar suite of genes found in mitochondria of free-living diatoms and dinoflagellates that lack a symbiont? In endosymbiotic partnerships, the symbiont is generally the more reduced, so it is of interest to know whether the dinotom symbiont has retained a full suite of diatom mitochondrial genes or not. However, in this case the host genome is also of interest because dinoflagellate mitochondrial genomes are already highly reduced so that all the genes they originally encoded are also found in the symbiont. To address these questions and investigate the outcome of the permanent and obligate tertiary endosymbiosis on the content and organization of the two distinct mitochondrial genomes in dinotoms, we sequenced the endosymbiont mitochondrial genomes of *D. baltica* and *K. foliaceum*. We also extensively sequenced the *D. baltica* host mitochondrial genome (but not completely since the nature of dinoflagellate mitochondrial genomes is not compatible with 'complete' sequencing), and produced the first sequencing data from the host mitochondrial genome in *K. foliaceum* in addition to extra

sequencing data pertaining to the transcription in both genomes. Then, we compared these data from endosymbiont and host in dinotoms with available diatom and dinoflagellate mitochondrial genomes and sequences, respectively, to see if they are in any way reduced in relation to their free-living counterparts. We find both endosymbiont genomes are almost identical in gene content to other diatoms and even genome organization is almost identical to that of the raphid pennate diatom *Fragilariopsis cylindrus*. We also find that the host mitochondrion in *D. baltica* encodes complete copies of *cox1*and *cob* genes and a bipartite *cox3* gene, many pseudogenes of all three genes, along with several fragments of the large subunit of ribosomal RNA gene (*LSU rRNA*), exactly as described in other dinoflagellates (Gray et al. 2004; Jackson et al. 2007; Slamovits et al. 2007; Waller and Jackson 2009;). From the host mitochondrion in *K. foliaceum*, we also characterized the first identified fragments of the three protein-coding genes, their corresponding transcripts along with the transcripts of several *LSU rRNA* fragments, all of which show a high degree of homology with their counterparts in other dinoflagellates. Overall, it appears that the endosymbiotic integration of the diatom with its dinoflagellate host has had no detectable effect on the evolution of its two distinct mitochondrial genomes, which contrasts with all other secondary and tertiary endosymbionts, where the organelle is lost altogether.

Results

The endosymbiont mitochondrial genomes of *D. baltica* and *K. foliaceum*

From the A+T-rich fraction of DNA of *D. baltica* and *K. foliaceum*, 299 and 635 pyrosequencing reads with an average length of 366 bp and 386 bp were respectively identified as endosymbiont mitochondrial sequences. A total of 169 and 123 Sanger reads were also used in the assemblies, resulting in single contigs of 35,505 bp (*D. baltica*) and 39,686 bp (*K. foliaceum*) with an overall coverage of 5.46 × and 7.73 ×, respectively. We were unable to bridge the final

gap in both genomes, despite numerous attempts using different long-range PCR protocols under different conditions, buffer systems, and primers. This is most likely due to the presence of a large intervening sequence, as is common to other diatom mitochondrial genomes (for example the 35 kb insertion in *P. tricornutum* (Oudot-Le Secq and Green 2011)), and/or to the presence of repetitive elements that may form complex secondary structures that inhibit PCR. Since all the other sequenced diatom mitochondrial genomes map as circular molecules (Oudot-Le Secq and Green 2011; Ravin et al. 2010), it is likely that the *D. baltica* and *K. foliaceum* genomes share the same configuration.

General features of the endosymbiont mitochondrial genomes of *D. baltica* and *K. foliaceum*

The coding regions of the endosymbiont mitochondrial genomes of *D. baltica* (34,242 bp) (GenBank: JN378735) and *K. foliaceum* (34,742 bp) (GenBank: JN378734) are very similar in size, form and content to those of other diatoms (Table 3.1). They are compact, featuring small intergenic spacers and a number of overlapping genes, and encode 58 and 59 genes, respectively (figure 3.1, Table 3.1). In addition to two rRNA genes, *D. baltica* and *K. foliaceum* mitochondria respectively encode 33 and 35 protein-coding, and 23 and 22 tRNA genes. Both code for the initiator and elongator methionine tRNAs but seem to lack tRNAs for threonine, like all other known diatoms and heterokonts (Gray et al. 2004). The apparent absence of a tRNA for glutamic acid (*trnE*) is shared with *S. acus* but not with their closer relative *P. tricornutum*, and the histidine tRNA is missing from *K. foliaceum* but not *D. baltica*. In the latter case, it is possible that the missing tRNA genes are encoded in the unsequenced portion of the genomes, as they are encoded in other diatom mitochondria. The two dinotom mitochondrial genomes also share two potentially spurious open reading frames (ORFs) larger than 100 amino acids (aa), *orf138* and *orf105* in *K. foliaceum* and *orf124* and *orf102* in *D. baltica*, respectively displaying 67% and

55% aa identity to each other. These ORFs are not found in other diatoms and show no significant homology in BLAST searches (Altschul et al. 1990). Interestingly, the endosymbiont mitochondrial gene complement is well-conserved across the larger group of stramenopiles or heterokonts that include diatoms (Ehara et al. 2000). Gene length comparisons between the mitochondrial genes in the two endosymbionts and those of diatoms indicate that their protein-coding and rRNA genes are also very similar in size (Figure 3.S1). Only the *rpl2* gene in *D. baltica* seems shorter at the 5′-end, however, it still retains both the conserved RNA-binding and the C-terminal domains.

The overall G+C content is very similar in the two endosymbiont mitochondrial genomes, albeit slightly less so in their intergenic regions (Table 3.1). Their G+C content is also consistent with that of the other diatom mitochondrial genomes, with the higher total G+C content observed in that of *P. tricornutum* due at least in part to the presence of a large 35 kb-long insertion (nearly half of its genome) with repetitive elements having 36.7% G+C content (33.6% GC content without). Like their pennate diatom counterparts in *S. acus* and *P. tricornutum*, the endosymbiont mitochondrial genomes of *D. baltica* and *K. foliaceum* use the universal genetic code. In contrast, the centric diatom *T. pseudonana* (Oudot-Le Secq and Green 2011) and possibly two other *Thalassiosirales*, *T. nordenskioldii* and *Skeletonema costatum* (Ehara et al. 2000) use TGA for tryptophan rather than as a signal for translational termination. In addition to the canonical ATG, the two dinotoms use ATA (*rps2*, *rpl2*, *nad3* in *D. baltica* and *atp8* in *K. foliaceum*) and ATT (*rps2* in *K. foliaceum*) as alternative start codons. The alternative start codons are utilized by other organisms including diatoms. *S. acus*, for example, uses GTG (*tatC*, *nad5* and *cox2*), *P. tricornutum* uses TTG (*cox3*, *cob* and *tatC*) and GTG (*nad7*), and *T. pseudonana* uses ATT (*atp8*) as alternatives for ATG. The two endosymbiont mitochondrial

genomes use all the codons for their proteins just like their diatom and brown algal counterparts (Oudot-Le Secq et al. 2006), hence the missing tRNAs must be imported from cytosol. As with most A+T rich genomes, *D. baltica* and *K. foliaceum* endosymbiont mitochondrial genomes display a bias towards A or T in the third codon position of their protein-encoding genes (79% and 76%, respectively), as do their diatom counterparts (*T. pseudonana* 79%, *S. acus* 76%, and *P. tricornutum* 72%).

Gene fission

One of the protein-coding genes, *nad11*, in the endosymbiont mitochondrial genomes of *D. baltica* and *K. foliaceum* is broken into two parts corresponding to its two functional domains: the iron-sulfur (FeS) binding (*nad11a*) and the molybdopterin-binding (*nad11b*) domains. These two new segments have acquired a new stop codon (*nad11a*) and a new start codon (*nad11b*) and now reside on opposite strands, distantly separated in the genome. In *T. pseudonana* and *S. acus*, *nad11* remains intact. However, in the pennate diatom *P. tricornutum* it is divided into two segments at about the same position but on the same strand and only 13 bp apart, while in *F. cylindrus nad11a* and *nad11b* are configured exactly as in dinotoms (Oudot-Le Secq and Green 2011). It is noteworthy that the molybdopterin-binding domain of *nad11* in brown algae is highly divergent, and has been relocated to the nucleus of at least one species, *Ectocarpus siliculosus* (Oudot-Le Secq and Green 2011).

An in-frame insertion

Another distinguishing feature of both endosymbiont mitochondrial genomes is the presence of a long insertion in *nad2*. This nearly 500 bp-long in-frame insertion (from amino acid 213 in both to aa 377 in *D. baltica* and aa 381 in *K. foliaceum*) is not found in *P. tricornutum*, *S. acus* or *T. pseudonana*, and falls within the NDH/q1-type oxidoreductase

domain of the Nad2 protein, between two conserved α-helices (Figure 3.S2). The insertion sequence shares no similarity to any known sequence, and is highly divergent between the two dinotoms: endosymbiont *nad2* genes share 93% and 88% amino acid identity before and after the insertion site, respectively, whereas the inserts share only 40% identity. This insertion is not spliced at the mRNA level, as indicated by RT-PCR and sequencing.

Gene fusions in *D. baltica*

The mitochondrial genome of the endosymbiont in *D. baltica* also contains two pairs of genes that have fused: *rps3-rpl16* and *rps13-nad9* (red arrows in figure 3.1). In both pairs, the first gene has lost its stop codon while the second has kept its first methionine. In *K. foliaceum, P. tricornutum* and *T. pseudonana,* the *rps3* and *rpl16* genes are adjacent but not fused, whereas in *S. acus, rps3* is degenerated and remains in the genome as a pseudogene near the *rpl16* gene (Ravin et al. 2010). The other two genes, *rps13* and *nad9*, are adjacent and in close proximity in *K. foliaceum* but not in the other diatoms.

Introns in *K. foliaceum*

The *K. foliaceum* endosymbiont mitochondrion contains three ORF-encoding introns, whereas *D. baltica* has none. One *K. foliaceum* intron is found in *rnl* (group I) and two (group I and group II) in *cox1* (figure 3.1 and figure 3.2). The *orf168* located in the *rnl* intron codes for a putative single LAGLIDADG endonuclease while *orf339* from the *cox1* group I intron encodes a putative heterodimeric endonuclease carrying two LAGLIDADG motifs. The *orf715* from the *cox1* group II intron encodes a reverse-transcriptase maturase (RTM). Of the three *K. foliaceum* introns, only one is inserted at a site in common with other diatoms (Table 3.1): the *cox1* group II intron being found in *T. pseudonana* and *P. tricornutum*, and sharing 91% and 81% nucleotide identity with the conserved cores (510 and 496 aligned residues), respectively. The *K.*

foliaceum's *orf715* is also highly similar to *orf718* in the *T. pseudonana* intron and slightly less so with *orf728*, a pseudo-RTM, present in two adjacent pieces in the *P. tricornutum* intron (85% and 67% amino acid identity over 718 and 730 aligned residues, respectively). The close phylogenetic relationship between *K. foliaceum*'s *ORF715* and *T. pseudonana*'s *ORF718* has been corroborated independently through phylogenetic analysis (Kamikawa et al. 2009).

Synteny

The endosymbiont mitochondrial genomes of *D. baltica* and *K. foliaceum* are perfectly syntenic, and demonstrate striking similarity to that of the raphid pennate diatom *F. cylindrus*. Two large gene blocks (*rps8-rpl6-rps2-rps4-trnN* and *rpl2-rps19-rps3-rpl16-atp9-trnK-nad4L-trnD-nad11a*) are also conserved with *P. tricornutum* and *T. pseudonana* (the green arcs in figure 3.1), whereas a third (*rps12-rps7-trnR-rpl14-rpl5-trnG-trnS-trnC-nad1-tatC-trnW-trnI*) is shared with *P. tricornutum* (the orange arc in figure 3.1). With the exception of *trnC*, this third block is also conserved in *T. pseudonana*. Compared to other diatom mitochondrial genomes, there is a small inversion unique to the dinotoms (*trnA-atp8*).

Table 3.2 summarizes the estimated minimum number of inversions required for the interconversions of the diatom mitochondrial genomes. Transition from either dinotom mitochondrial genome to that of *P. tricornutum*, and vice versa, requires only 5 inversions while their transition to that of *T. pseudonana* requires 6 inversions. A minimum of 8 inversions are required to interconvert *T. pseudonana* with either *P. tricornutum* or *S. acus*.

Transcription of the endosymbiont mitochondrial genes

We had previously shown that the endosymbiont *cox1*, *cob*, *cox2*, *cox3* and *rnl* genes in *D. baltica* and *K. foliaceum* are transcribed with no signs of editing, that the *cox1* introns in *K. foliaceum* are removed from its mRNA, and that *cox3* and *cob* are transcribed as an operon in

both *D. baltica* and *K. foliaceum* (Imanian and Keeling 2007; Imanian et al. 2007). In this study we further expanded our sampling of the transcripts of mitochondrial genes in the endosymbionts of dinotoms. Using RT-PCRs with DNase-treated total RNA and specific primers, we obtained partial *nad5* and *nad2* products from both genomes. We also investigated and confirmed the polycistronic transcription of the conserved gene block *rps19-rps3-rpl16*, which includes the *rps3-rpl16* fused gene in *D. baltica*. All cDNA sequences were identical to their corresponding genes, consistent with the lack of editing in diatom mitochondrial transcripts as opposed to those of dinoflagellates which are heavily edited by substitutions (Lin et al. 2002).

The mitochondrial genome of the dinoflagellate host in *D. baltica*

From the 454 sequencing data of the A+T-rich fraction of DNA in *D. baltica*, we identified more than 29,000 reads (average length of 349 bp amounting to more than 10 million bp) corresponding to putative dinoflagellate host mitochondrial sequences. These reads were subsequently assembled into hundreds of unique contigs. Of these, we further analyzed 123 high quality contigs that included 4,569 reads covering 89,634 bp of unique consensus sequences from the host's mitochondrial DNA in *D. baltica*, providing the most comprehensive assemblage of any dinoflagellate mitochondrial genome to date. The contigs vary in size from 210 to 2,740 bp, with an average length of 711 bp. We identified full-length copies of the *cox1* and *cob* genes, the *cox3* gene that is split into two parts (GenBank: JX001475-JX001478) along with several fragments of the large subunit ribosomal RNA (*LSU rRNA*) gene (GenBank: JX001584-JX001600). We have also recovered 102 contigs containing pseudogenes of *cox1* (GenBank: JX001520- JX001583), *cob* (GenBank: JX001497- JX001519) and *cox3* (GenBank: JX001482-JX001496).

Host mitochondrial protein-coding genes, transcription and editing

The contig containing *cox1* is 2,740 bp long with 99 reads (12.6 × coverage), while the contig that includes *cob* is 2,020 bp long with 82 reads (14.2 × coverage). As is the case in several other dinoflagellates (Jackson et al. 2007; Jackson et al. 2012), the *D. baltica cox3* gene is broken in two separate parts: *cox3* part 1 (*cox3-1*) is 733 bp long with 48 reads (22.9 × coverage), while the second contig, *cox3* part 2 (*cox3-2*), is 595 bp long, with 12 reads (7.0 × coverage). The 5′ end of *cox1* gene is preceded by non-coding sequence with no significant homology to any known sequences. The 3′ end of the gene is followed by 81 bp, non-coding, and then, by a *cob* pseudogene (339 bp) and a short *cox1* pseudogene (110 bp). The *cob* gene is also flanked by 115 bp and 259 bp non-coding sequences at its 5′ and 3′ ends, respectively, and it is followed by 2 separate *cox3* pseudogenes.

In the dinoflagellate *Crypthecodinium cohnii*, the *cox1* gene appears in multiple copies bounded by distinct flanking sequences (Norman and Gray 2001). It is also reported, though not definitively shown, that there is more than one copy of *cox1* and *cob* genes in *K. micrum* mitochondrial genome (Jackson et al. 2007). In our extensive sequencing survey and careful assembly of the host mitochondrial genome of *D. baltica*, we were unable to find any evidence of multiple copies of the full-length *cox1* and *cob* genes and *cox3-1*, each of which appears only in one genomic context. However, the *cox3-2* that encodes the short 3′ end of the gene appears in multiple contexts (see GenBank: JX001478, JX001487, JX001488, JX001494) flanked by distinct sequences like the 3′ segment of this gene in *K. micrum* (Jackson et al. 2007).

The host mitochondrial protein-coding genes of *D. baltica* have very similar GC content to their homologues in other dinoflagellates: 33.3% , 29.8% and 28.5% GC content for *cox1*, *cob* and *cox3*, respectively, compared to an average of 33.2%, 29.6% and 28.4% for the same genes,

respectively, in other dinoflagellates (Excel file S1 available online: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0043763). These genes also show high degree of nucleotide and amino acid identities to their counterparts in other dinoflagellates: *cox1*, *cob* and *cox3* have an average of 95%, 95% and 89% nucleotide identities and 90%, 88% and 72% amino acid identities to their homologues in other dinoflagellates (Excel file S1 available online: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0043763).

One of the distinguishing characteristics of the mitochondrial protein-coding genes in dinoflagellates is the genes themselves do not encode canonical start and stop codons to direct the initiation and termination of translation (Jackson et al. 2007; Jackson et al. 2012; Slamovits et al. 2007). The only exception to date is the *cox3* gene of the basal dinoflagellate *Hematodinium* which encodes a canonical stop codon (Jackson et al. 2012), and the *cox1* gene of *C. cohnii* which seems to encode a canonical start codon (Norman and Gray 2001). In some dinoflagellates the *cox3* transcript apparently obtains a stop codon through polyadenylation, while others simply lack a stop codon (Jackson et al. 2007; Jackson et al. 2012). The *cox1*, *cob* and *cox3* genes in *D. baltica* resemble homologues in other dinoflagellates, in lacking canonical start and stop codons as well. There is one in-frame TGA codon in the middle of *cox3*, but in all likelihood this is edited at the mRNA level as has been shown in the *cox1* transcript of *Amphidinium carterae* (Nash et al. 2007), the *cox3* transcript of *K. micrum* (Jackson et al. 2007), and others (Lin et al. 2002; Zhang and Lin 2005). Indeed, TGA, which typically codes for stop and sometimes for tryptophan, is unassigned in dinoflagellates (Jackson et al. 2007; Jackson et al. 2012).

The comparison between the complete *cox1* gene and its nearly complete transcript (GenBank: JX001479) obtained through RT-PCR, reveals extensive substitutional editing occurring at either the first or second codon positions, resulting without exception in an amino acid change (see Table 3.S1). Most of the edits substitute a G for an A, while some replace a T with a C or a C with a U or more infrequently a G with a C. Most of these replacements result in a conservative substitution of an amino acid (for example, an isoleucine with a valine). The number of editing sites, their codon positions and the types of edits all are consistent with those reported for other dinoflagellates (Jackson et al. 2007; Jackson et al. 2012; Lin et al. 2002; Nash et al. 2007; Zhang & Lin 2005).

A novel feature of the *cob* gene is the presence of a 150-nucleotide-long in-frame insert starting at amino acid 121 to 170. The insert sequences show no homology to any other sequences in the public databases except to a 69-nucleotide-long portion of another insert within a *cox1* pseudogene in *D. baltica* (GenBank: EF434626.1). The insert is located between the two predicted transmembrane helices, conserved also in *Alexandrium catenella* and *Pfiesteria piscicida*, without disrupting them (figure 3.S3). The RT-PCR results show that this insert is transcribed along with the flanking conserved regions of this gene and remains unedited (GenBank: JX001480) unlike other parts of the transcript that is edited in the dinoflagellate fashion (Imanian and Keeling 2007).

The *cox3* gene in the basal dinoflagellates *Oxyrrhis marina* and *Hematodinium* sp. is unbroken (Jackson et al. 2012; Slamovits et al. 2007), whereas in at least five other dinoflagellates it is broken into two parts, transcribed and polyadenylated separately and then trans-spliced together to produce the full-length transcript (Jackson et al. 2007; Jackson et al. 2012). In *D. baltica*, *cox3* is similarly encoded as two separate sections. The *cox3-1* segment

encodes the first 705 nucleotides (corresponding to the first 235 amino acids), the 5′ end of the gene, and it is followed by 27 nucleotides of non-coding sequences. The *cox3-2* encodes the 153 nucleotides corresponding to the 3′ end of the gene, and it is flanked by stretches of 297 and 145 nucleotides unrelated to *cox3* sequences. In *K. micrum*, the trans-splicing site is predicted to occur between the codons for the amino acid 235 and 236 (Jackson et al. 2007), which is the same position where the two parts are patched together in *D. baltica* (amino acid 235-236). The evidence for the conserved site of trans-splicing comes from the RT-PCR results. The *cox3* transcript in *D. baltica* (GenBank: JX001481) covers the nucleotides 306 to 768 (corresponding to amino acids 102 to 258) traversing the two separate parts of the gene including their junction while there is not even a single 454 sequence (out of more than 29,000 host mitochondrial sequences we identified from the A+T-rich fraction of the DNA) that spans the two parts of the gene. The comparison between the *cox3* gene and its transcript reveals extensive editing especially upstream the trans-splicing site (about 36 substitutions), which also includes five A residues at the junction site. This penta-A is also found at the junction of the two parts of the *cox3* gene in *K. micrum* and is thought to have been derived from the poly A tail of the part one of the gene (Jackson et al. 2007).

Host mitochondrial ribosomal RNA gene fragments

The ribosomal RNA genes in both apicomplexans and dinoflagellates are highly fragmented, and 20 or more fragments have been identified in a few species from both taxa (Feagin et al. 1997; Jackson et al. 2007; Jackson et al. 2012). We have identified 8 unique fragments of the *LSU rRNA* in *D. baltica*: *LSUA*, *LSUD*, *LSUE*, *LSUF*, *LSUG*, *RNA2*, *RNA7* and *RNA10*-like fragments. The *LSUA*, *LSUE* and *RNA10*-like fragments appear in two copies, each of which within a different genomic context. Compared to their homologous sequences in other

dinoflagellates (for example, in *K. micrum*, *A. catenella* and *P. piscicida*) the *D. baltica LSU rRNA* fragments are highly conserved (on average between 88% to 96% nucleotide identities).

The host mitochondrial genome is dominated by pseudogenes

The mitochondrial genomes of apicomplexans are among the smallest mitochondrial genomes, encoding only 3 protein-coding genes and highly fragmented rRNA genes in a short linear chromosome (about 6 kbp). Although the dinoflagellate mitochondrial genomes seem to be as gene-poor, their genome is expanded enormously through amplification of the few genes and gene fragments they encode, generating in some species multiple copies of these genes and more often myriads of their gene fragments or pseudogenes (Feagin et al. 1997; Jackson et al. 2007; Jackson et al. 2012; Imanian and Keeling 2007; Nash et al. 2007; Norman and Gray 2001; Slamovits et al. 2007). In this regard the mitochondrial genome of the host in *D. baltica* is a typical dinoflagellate mitochondrial genome with hundreds if not thousands of pseudogenes of both the protein-coding and *LSU rRNA* gene fragments. These pseudogenes appear in a wide variety of sizes, orientations and genomic contexts. They generally include a highly conserved portion of the true genes (usually with 99% to 100% nucleotide identity to their corresponding sequences found in the full-length genes), flanked by different non-coding and/or repetitive sequences (figure 3.3A). The conserved regions of these pseudogenes appear in various lengths, and we present the sequence data, for the first time, demonstrating that they are derived from all different regions of the full-length genes without any apparent preference or hot spots for any specific region (figure 3.3B).

Although the majority of the pseudogenes show a high degree of sequence identity to different regions of the true genes, we identified a number of pseudogenes with different degrees of degeneration. For example, a *cox1* pseudogene (GenBank: JX001555) is highly conserved

along the first 327 nucleotides (99% identity), but it is followed by a *cob* pseudogene that is highly degenerated (only 44% identity to other dinoflagellates' *cob*). In another example (GenBank: JX001543) a degenerated *cox3* pseudogene (46% identity) is located between two conserved *cob* and *cox1* pseudogenes. These degenerate sequences in the presence of many well-conserved gene fragments may indicate that rampant amplification and recombination not only play a role in sequence conservation of many pseudogenes (Jackson et al. 2012) but also simultaneously generate many mutations elsewhere.

The mitochondrial genome of the dinoflagellate host in *K. foliaceum*

While we recovered thousands of sequences with significant homology to dinoflagellate mitochondrial sequences from the A+T-rich fraction of DNA in *D. baltica*, we were unable to find any such sequences from the A+T-rich fraction of DNA in *K. foliaceum*. Our initial attempts to amplify and sequence the protein-coding genes and their transcripts using degenerate or dinoflagellate specific primers through PCR and RT-PCR, respectively, were unsuccessful. However, the 454 sequencing data from the *K. foliaceum* cDNA library (see Materials and Methods) generated hundreds of short sequences (average length of 76 bp) that show significant homology to mitochondrial sequences of other dinoflagellates. The assembly of these reads generated larger contigs and after subsequent PCR and RT-PCR based on these new data, we were able to recover larger fragments of all the three protein-coding genes but not their full-length sequences. These results are summarized in Table 3.3. We also recovered several fragments of the *LSU rRNA* transcripts (some in 2 copies within distinct flanking sequences) including *LSUA*, *LSUE*, *LSUG* and *RNA7*-like fragments (GenBank: JX001601-JX001608) with 358, 65, 67 and 409 pyrosequencing reads, respectively. Our attempts to recover the full-length genes and their transcripts through further PCR and RT-PCR failed. Nested primers were also

tested without any results. We also tested the possibility that gene fragments were encoded on separate circular chromosomes using outward primers in PCR and long range PCR, but they did not produce any product.

The host's mitochondrial protein-coding gene fragments in *K. foliaceum* have very similar GC content to their corresponding homologous sequences in other dinoflagellates: 34.3%, 29.6% and 28.9% GC content for *cox1*, *cob* and *cox3* fragments, respectively (Excel file S1 available online: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0043763). These gene fragments also show high degree of nucleotide and amino acid identities to their counterparts in *D. baltica*: *cox1*, *cob* and *cox3* fragments have an average of 99%, 98% and 88% nucleotide identities and 96%, 93% and 84% amino acid identities to their homologous sequences in *D. baltica* (Excel file S1 available online: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0043763).

A comparison between the *cox1* gene fragments and their corresponding cDNAs reveals similar substitutional mRNA editing to those occurring in *D. baltica* and other dinoflagellates (see Table 3.S1). Most of the edits affect either the first or second codon positions, resulting in an amino acid change. Just like in *D. baltica*, most of the edits in *K. foliaceum* are from A to G, but changes from T to C, C to U and G to C are also observed. Out of 11 editing sites in the *cox1* mRNA of *K. foliaceum* 8 are conserved in *D. baltica* as well (Table 3.S1).

Discussion

The mitochondrial genomes of the endosymbionts in *D. baltica* and *K. foliaceum* have not been reduced

The mitochondrial genomes of the tertiary endosymbionts in *D. baltica* and *K. foliaceum* share nearly all the characteristics found in mitochondrial genomes of free-living diatoms, including gene repertoire, gene length, GC content, and gene order. Their diatom gene set is also packaged in the diatom style: they are densely packed, with short intergenic sequences, a few overlapping genes, and no scattered stretches of repeated elements. The only repetitive elements in diatom mitochondrial genomes are sequestered into one or two long contiguous regions (Oudot-Le Secq and Green 2011; Ravin et al. 2010), and it is likely that the unsequenced region of the two endosymbionts corresponds to a similar repetitive element-rich region. In short, the tertiary endosymbiosis event has had little if any effect on the endosymbiont mitochondrial genome, which is of interest since in all other comparable cases, the organelle is totally lost.

Recently, Gabrielsen et al. (2011) sequenced the plastid genome of the tertiary haptophyte in the dinoflagellate *Karlodinium veneficum*, providing the only available haptophyte-derived plastid genome for comparison in this study. They showed that it maintains a genome, but with extensive gene losses, enlarged intergenic regions and substantial rearrangements compared to that of free-living haptophytes. Some of the existing genes in this genome have diverged so markedly that they might have become pseudogenes or reliant on RNA editing to produce functional proteins (Gabrielsen et al. 2011). In contrast to this, we have shown that the plastid genomes of *D. baltica* and *K. foliaceum* are not reduced, and encode well-conserved genes that are organized similarly to those in the plastid genomes of free-living diatoms (Imanian et al. 2010). Moreover, the *K. foliaceum* plastid genome is much larger and

more re-arranged, mainly because of the integration and partial maintenance of at least two relict plasmids also found in other diatoms (Imanian et al. 2010).

The endosymbiont mitochondrial genomes of the two dinotoms appear equally unaffected by their integration with the dinoflagellate. Indeed, we were only able to identify a handful of features that distinguish dinotom mitochondria, or link them to a subset of free-living diatom lineages (Figure 3.S4). First, the homologous (but divergent) long in-frame insert within *nad2* is found in dinotoms but not in *P. tricornutum*, *S. acus* or *T. pseudonana*. Second, the dinotoms share a small unique inversion (*trnA-atp8*). Third, the fragmented *nad11* gene and translocated *nad11b* is found in both dinotoms, but also in *F. cylindrus* (Oudot-Le Secq and Green 2011), suggesting the dinotom endosymbionts are more closely related to this raphid pennate diatom than any other diatom for which mitochondrial genome data exist.

The mitochondrial genomes of the host in *D. baltica* and *K. foliaceum* retain nearly all their dinoflagellate characteristics

The dinoflagellate host in *D. baltica* retains a typical dinoflagellate mitochondrion with tubular cristae (Imanian and Keeling 2007), and we have shown here that this organelle maintains a genome with all the typically unusual traits of this genome in other dinoflagellates, including the gene content, the GC composition, gene and amino acid identities, abandonment of canonical start or stop codons, and genome organization (Jackson et al. 2007; Jackson et al. 2012; Slamovits et al. 2007; Waller and Jackson 2009; Nash et al. 2007). The *cox3* gene in *D. baltica* is encoded as two separate sections, and the transcripts are trans-spliced at the same general region of the gene in at least five other dinoflagellates (and the same nucleotide position as in *K. micrum cox3*) to produce the full-length mRNA (Jackson et al. 2007; Jackson et al. 2012; Waller and Jackson 2009). Despite being gene poor, the host's mitochondrial genome in *D.*

baltica has expanded enormously through amplification and recombination, harboring numerous pseudogenes. We have also shown here that extensive substitutional mRNA editing occurs in *D. baltica* (Jackson et al. 2007; Jackson et al. 2012; Lin et al. 2002). Indeed, the only novel trait we have found in the *D. baltica* host mitochondrial genome is the 150-nucleotide in-frame insert within its *cob* gene.

The mitochondrial genome of the host in *K. foliaceum* has been more elusive, but we have characterized several fragments of all three protein-coding genes and their transcripts along with several nearly full-length *LSU rRNA* fragments. These data indicate that the host in *K. foliaceum* has a mitochondrial genome that encodes at least the same three protein-coding genes, with very similar GC content, nucleotide and amino acid identities to those in other dinoflagellates (Excel file S1 available online: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0043763). We have also demonstrated that the *K. foliaceum cox1* mRNA editing is substitutional, and its types, codon positions, and sites show consistency with those seen in other dinoflagellates (Table 3.S1). Overall, the data seem to be consistent with a conventional dinoflagellate mitochondrial genome in the host of *K. foliaceum*, though it is curiously hard to characterise.

These genomes raise the important question of why the endosymbiont mitochondrial genomes have not been completely eliminated or significantly reduced, and why the host mitochondrial genomes remain almost completely unaffected by the endosymbiosis. We have previously suggested that the mitochondrial genome redundancy (with two sets of *cox1, cob* and *cox3* genes, one from dinoflagellate host and one from the diatom endosymbiont) found in dinotoms might be due to spatial differentiation rather than functional specialization (Imanian and Keeling 2007). The nearly complete endosymbiont genomes are consistent with this, but

additional data from the host mitochondrial genome in *K. foliaceum* and from mitochondrion-targeted proteins in both nuclear genomes will be required to really determine whether the function of either organelle has been affected by the presence of the other.

Conclusions

Despite the full integration of the diatom tertiary endosymbiont within the dinoflagellate host and the consequent unique mitochondrial genome redundancy within dinotoms, we have found no evidence of significant changes in the mitochondrial genome of the host in *D. baltica* or *K. foliaceum* compared to those in free-living dinoflagellates. Our results also indicate that the endosymbiont mitochondrial genomes in the two dinotoms closely resemble those of their counterparts in free-living diatoms, following nearly the same evolutionary path to those in other diatoms but starkly distinct from those in other secondary and tertiary endosymbionts where mitochondria are lost altogether.

Materials and methods

Strains and culture conditions

Cultures of *Kryptoperidinium foliaceum* CCMP 1326 and *Durinskia baltica* (*Peridinium balticum*) CSIRO CS-38 were respectively obtained from the Provasoli-Guillard National Center for Culture of Marine Phytoplankton (West Boothbay Harbor, ME, USA) and from the CSIRO Microalgae Supply Service (CSIRO Marine and Atmospheric Research Laboratories, Tasmania, Australia). *K. foliaceum* cultures were maintained in F/2-Si medium at 22 ºC (12:12 light:dark cycle) whereas *D. baltica* cultures were maintained under the same conditions in GSe medium.

Nucleic acids extraction, preparation and amplification

Exponentially growing cells were collected and ground as described previously (Imanian et al. 2007). Cells lysis, DNA extractions, precipitations, fractionations, adenine+thymine-rich

(A+T-rich) DNA isolations, purifications and amplifications were performed for both species as described earlier (Imanian et al. 2010). Total genomic DNA was extracted for polymerase chain reactions (PCR) either as described previously (Imanian et al. 2010), or using Master Pure Complete DNA and RNA Purification Kit (EPICENTRE Biotechnologies, Madison, WI, USA) following the manufacturer's instructions. Total RNA for RT-PCR was obtained as described earlier (Imanian et al. 2007). RNeasy MinElute Cleanup kit (Qiagen, Mississauga, ON) was used to clean up the total RNA after DNase treatment according to the manufacturer's instructions. PCR and RT-PCR reactions were performed using specific primers designed based on the obtained genomic data as described elsewhere (Imanian et al. 2007, 2010). Long range PCRs were conducted either as described earlier (Imanian et al. 2007, 2010), or using Expand Long Template PCR System kit (Roche Applied Science, Indianapolis, IN, USA) following the manufacturer's instructions.

The cDNA construction for *K. foliaceum*

Approximately 5 µg of total RNA was used as template for producing cDNA with SMARTer Pico PCR cDNA Synthesis kit (Clontech, CA) according to manufacturer's protocol. In order to optimize the number of PCR cycles for our sample, we performed between 15 and 30 cycles, and, based on agarose gel, determined that the optimal amplification was reached after 18 cycles.

Genome sequencing

The mt genomes of the endosymbionts and hosts in *K. foliaceum* and *D. baltica* and the cDNA library in *K. foliaceum* were sequenced using massively parallel GS-FLX DNA pyrosequencing (Roche 454 Life Sciences, Branford, CT, USA) using GS-FLX shotgun libraries prepared and sequenced at the Génome Québec Innovation Centre. Sequences were assembled

de novo using gsAssembler 2.5p1 (formerly known as Newbler), edited and re-assembled with CONSED 20 (Gordon et al. 1998, 2001). Gaps between contigs and ambiguous pyrosequencing homopolymer stretches were linked/ascertained by PCR and Sanger sequencing of the resulting products.

Genome annotation and analyses

Genes were identified through BLAST homology searches (Altschul et al. 1990) against the NCBI non-redundant databases [http://www.ncbi.nlm.nih/BLAST] and annotated in Artemis 12 (Rutherford et al. 2000). Protein-coding genes of endosymbionts were positioned with ORFFINDER at NCBI and GETORF from EMBOSS 6.0.1 (Rice et al. 2000) and their start codons determined by orthologous comparisons with close relatives while transfer-RNA (tRNA) genes were identified with tRNAscan-SE 1.21 (Schattner et al. 2005). The 5′ and 3′ ends of the mitochondrial protein-coding genes of the dinoflagellate hosts were determined after alignments were made with those in other dinoflagellates. Ribosomal RNA (rRNA) genes of the endosymbionts were annotated after comparison with their homologues in *P. tricornutum* and *T. pseudonana*, while those of the hosts' were annotated after comparison with their homologues in other dinoflagellates especially *K. micrum*, *A. catenella* and *P. piscicida*. Physical circular maps were prepared using GenomeVx (Conant and Wolfe 2008) and refined manually. Group I and group II intron secondary structures were predicted manually according to the conventions described in Burke *et al.* (1987) and Michel *et al*. (1989).

Transmembrane helices domains and the insertion site in the nad2 genes and the *D. baltica*'s cob were predicted using Domain homology searches (Marchler-Bauer et al. 2009), SeaView 4.0 (Gouy et al. 2010) and the TMHMM Server 2.0 [http://www.cbs.dtu.dk/services/TMHMM-2.0/] (Krogh et al. 2001). Conserved gene blocks

between the mitochondrial genomes of dinotoms and diatoms were identified through MAUVE 2.3.1 (Darling et al. 2004) and by manual examination of the physical maps. The hypothetical numbers of inversions between the dinotom and diatom mitochondrial genomes were estimated with GRIMM 1.04 (Tesler 2002).

The sequence data for *F. cylindrus* mitochondrial genome were downloaded through jgi website [http://genome.jgi-psf.org/Fracy1/Fracy1.download.html] and annotated as described above.

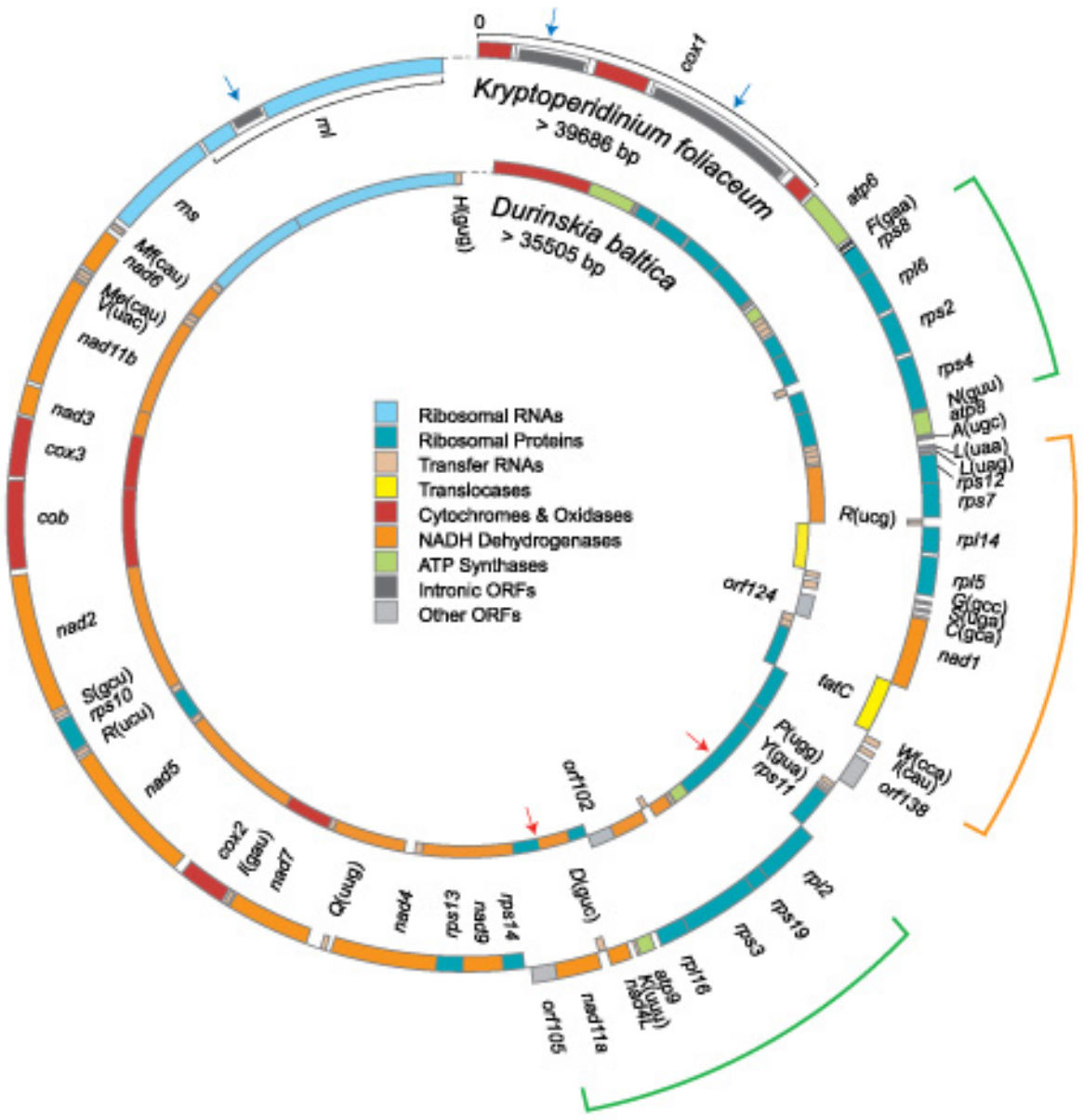

Figure 3.1: The mitochondrial genome maps of the endosymbionts in *Durinskia baltica* and *Kryptoperidinium foliaceum*.

Functionally related genes are colour-coded and transcriptional direction is clockwise (boxes outside the ring) or counterclockwise (inside). Genes for tRNAs are indicated by their single letter code. The dashed lines represent the gap in the genomes. The blue arrows specify the locations of the introns in the map for *K. foliaceum*, and the red arrows point at the locations of gene fusions in the map of *D. baltica*. The arcs show the conserved gene blocks in the two dinotoms and *P. tricornutum* (green and orange arcs) and *T. pseudonana* (the green arcs). The two genomes are not represented in scale with respect to one another.

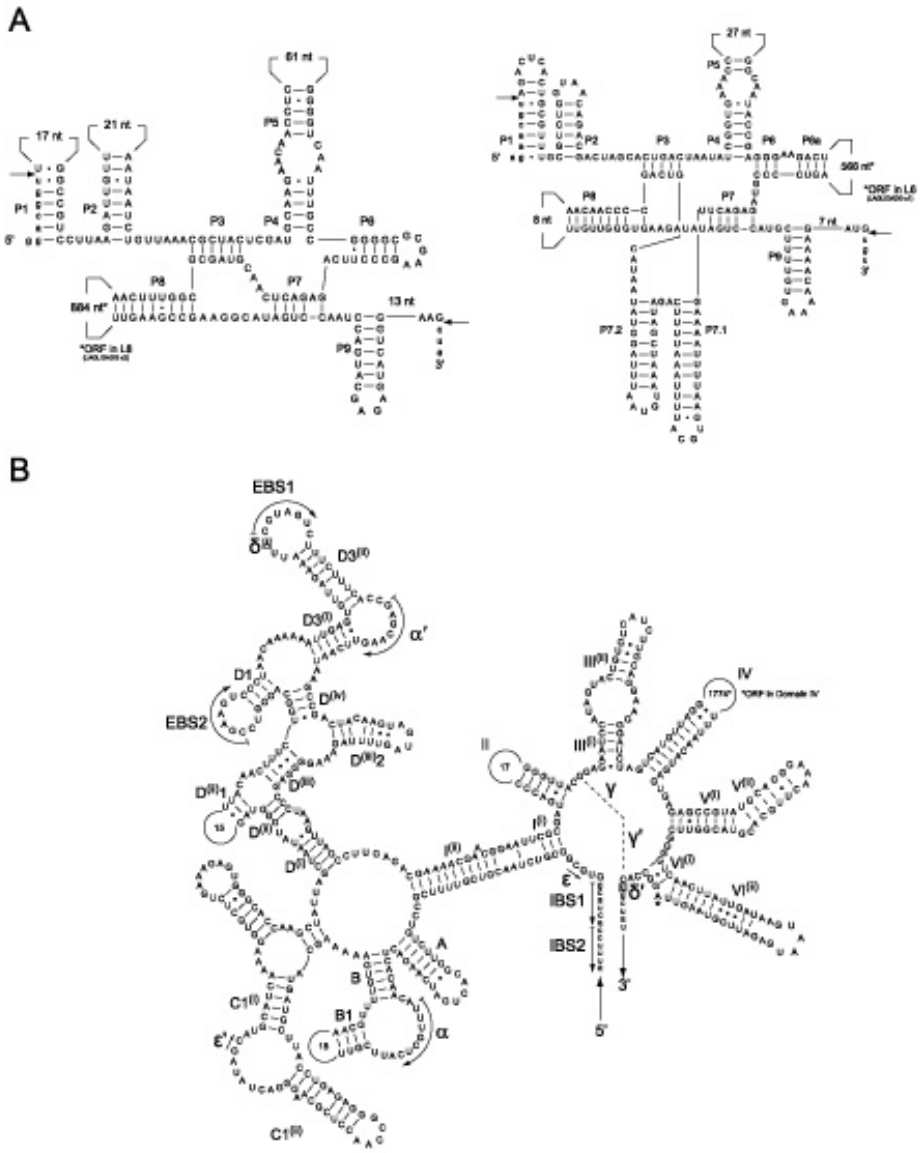

Figure 3.2: Predicted secondary structure of the three *Kryptoperidinium foliaceum* endosymbiont mitochondrial introns modeled according to the conventions described in Burke et al. (1987) and Michel et al. (1989).

(A) Group I introns. Left, the first *cox1* intron; Right, the *rnl* intron. The *K. foliaceum cox1* group I intron (left) had been previously mistakenly referred to as a group II intron (Imanian et al. 2007). (B) Group II intron. The second *cox1* intron. Panels A and B: canonical Watson-Crick base pairings are denoted by dashes whereas guanine-uracyl pairings are marked by dots. Numbers inside variable loops indicate the sizes of these loops. Exon sequences are shown in lowercase letters. Panel A: splice sites between exon and intron residues are denoted by arrows; Panel B: the major structural domains are indicated by roman numerals and capital letters A to D, whereas tertiary interactions are represented by dashed lines, curved arrows, and/or Greek letters. Nucleotides potentially involved in the δ-δ' interaction are boxed. Intron-binding and exon-binding sites are indicated by IBS and EBS, respectively. The putative site of lariat formation is denoted by an asterisk.

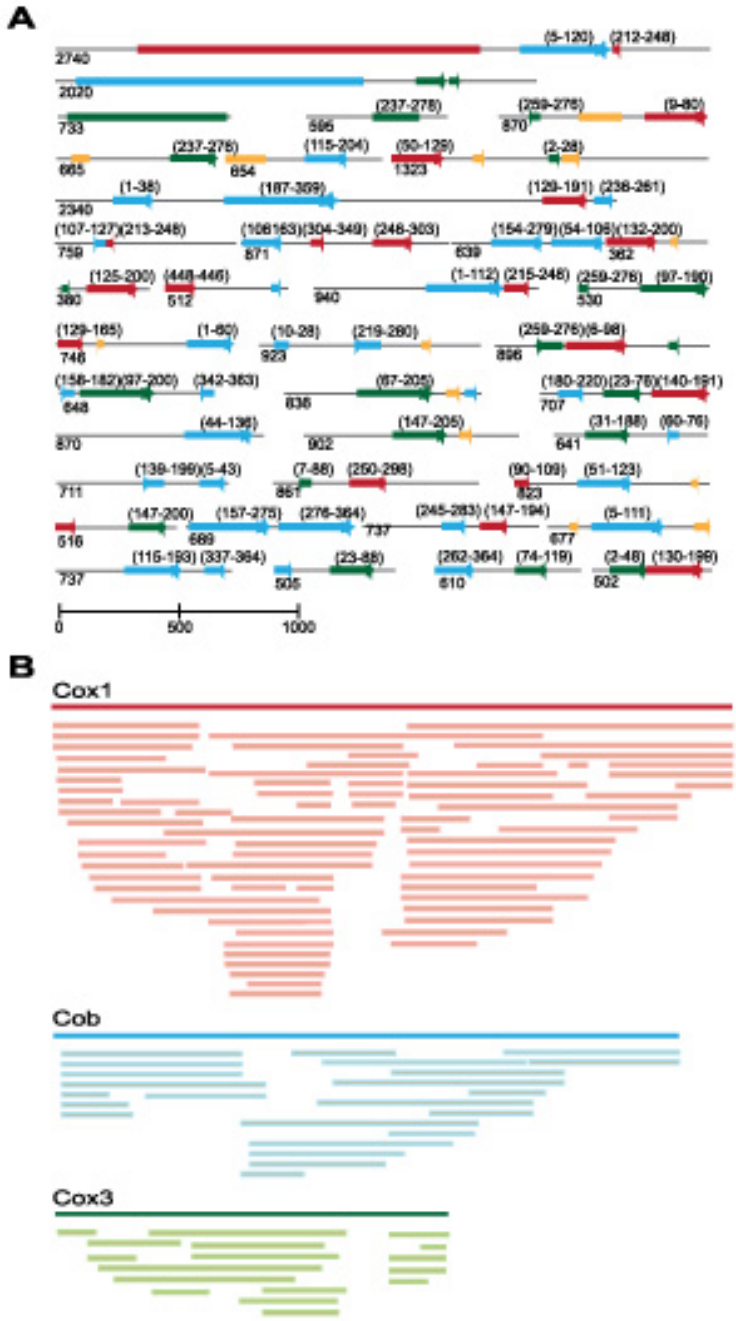

Figure 3.3: Genes and their pseudogenes in the mitochondrial genome of *Durinskia baltica*.

(A) The full-length genes and their derived pseudogenes. The full-length protein-coding genes and the *LSU rRNA* gene fragments are represented by colored blocks, while the pseudogenes are shown by colored blocks with a broken tip. The lines represent non-coding sequences. The genes and their matching sequences within the pseudogenes are color-coded: *cox1* in red; *cob* in blue; *cox3* in green; *LSU rRNA* fragments in yellow. The sequences are drawn in scale. The numbers at the bottom of the contigs show their sizes in nucleotides, while the numbers on the top within parentheses specify the number of the first and last amino acids on the full-length gene corresponding to the conserved sequences of the pseudogenes. (B) The Alignment of the conserved regions of many pseudogenes with their corresponding full-length gene.

Table 3.1: General characteristics of mitochondrial genomes in dinotoms compared to diatoms

	Durinskia baltica	*Kryptoperidinium foliaceum*	*Phaeodactylum tricornutum*	*Synedra acus*	*Thalassiosira pseudonana*
Size (bp)					
Total	> 35505	> 39686	77356[a]	46657[b]	43827[a]
Coding and intergenic	34242	34742	35177[a]	35944[b]	36519[a]
GC content (%)					
Total	31.02	32.41	35.08	31.78	30.11
rRNA genes	36.27	36.57	36.66	34.03	33.03
tRNA genes	44.03	43.72	43.01	38.52	40.55
Protein-coding genes	30.25	31.64	32.84	30.73	28.96
Intergenic spacer	22.14	26.15	26.17[c]	26.74	23.53[d]
Gene content					
Total	58	59	60	61[b]	61
Protein-coding genes	33	35	34	33[b]	34
rRNA genes	2	2	2	2[b]	2
tRNA genes	23	22	24	24[b]	25
Intronic ORFs	0	3	2	2[b]	1
Other ORFs	2	2	0	3[b]	0
Coding sequence (%)	90.45	83.03	77.01[e]	88.87	82.88[f]
Introns	0	3	4	3[b]	1
Gene overlaps (pairs)[g]	4	2	6	1	1
Fused genes (pairs)[h]	2	0	1	0	0
Intergenic spacer (bp)	58	109	841[a]	73	157[a]
Gene length[i]	793 (554)	709 (540)	770 (538)	758(531)	741 (519)

[a] Data from Oudot-Le Secq and Green (2011).
[b] Data from Ravin et. al. 2010.
[c] Calculated without repeat region (with repeat region it is 36.28%).
[d] Calculated without repeat region (with repeat region it is 30.10%).
[e] Calculated without repeat region (with repeat region it is 41.72%).
[f] Calculated without repeat region (with repeat region it is 73.48%).
[g] In *D. baltica*: *rps12-rps7*, *nad1-tatC*, *rps19-rps3-rpl16* fusion, *orf124-trnP*. In *K. foliaceum*: *rps12-rps7*, *nad1-tatC*. In *P. tricornutum nad4-rps13*, *rps2-rps4*, *nad1-tatC*, *rpl2-rps19*, *rps19-rpl16*, *rpl5-trnG*. In *S. acus* and *T. pseudonana nad1-tatC*.
[h] In *D. baltica*: *rps3-rpl16*, *rps13-nad9*. In *P. tricornutum*: *nad9-rps14*.
[i] First number is the average length of protein-coding genes, the number in parentheses is the average length of all genes.

Table 3.2: Number of inversions for the inter-conversions of the mitochondrial genomes of the two dinotoms and those of diatoms (predicted by GRIMM)

	D. baltica	*K. foliaceum*	*P. tricornutum*	*S. acus*	*T. pseudonana*
D. baltica	0	0	5	7	6
K. foliaceum	0	0	5	7	6
P. tricornutum	5	5	0	7	8
S. acus	7	7	7	0	8
T. pseudonana	6	6	8	8	0

Table 3.3: Partial protein-coding genes and their transcripts found from the host mitochondrial genome of *Kryptoperidinium foliaceum*

	GenBank Accession	Number of Contigs	Total Length (bp)	454 Reads	Sanger Reads
cox1	JX001614	2	968		37
cox1 transcript	JX001613	3	1173	69	12
cob	JX001611	4	579		13
cob transcript	JX001612	3	927	105	9
cox3	JX001609	1	88		4
cox3 transcript	JX001610	3	398	25	3

Chapter 4: A survey of the host nuclear transcriptome in *D. baltica*

Introduction

The transformation of an autonomous free-living bacterium into an essential organelle such as the mitochondrion or plastid through endosymbiosis has been accomplished, at least in part, by successful endosymbiotic gene transfers (EGTs) to the host nucleus. The contemporary mitochondrial and plastid genomes encode only a fraction of the genes whose protein products keep these organelles viable and functional. The majority of the organelle proteins are encoded in the nuclear genome. The estimates of the scope of the EGT from the bacterial ancestors of the mitochondrion and plastid to their respective host nucleus hovers around hundreds to over 1,000 genes (Archibald 2006; Gray et al. 2001; Martin 2009; Martin et al. 2002; Moustafa and Bhattacharya 2008; Reyes-Prieto et al. 2006; Timmis et al. 2004).

The parallel development of a protein targeting system in these two endosymbiotic events has complemented the EGT so that the protein products of the transferred genes can be sent to whence they originated. The independently evolved components of the protein targeting systems for these two organelles have analogous features found in their protein machinery (i.e. the organelle carrier proteins, the receptor proteins, TOM, TIM and TOC, TIC) (Cline and Dabney-Smith 2008; Dolezal et al. 2006; Gutensohn et al. 2006; Kovács-Bogdán et al. 2010) and in their targeting signals (mTPs and cTPs). While the primary sequences of these targeting signals are not conserved, the amino acid compositions and secondary structures of both mitochondrial and plastid transit peptides share certain features that are, indeed, conserved (Danne and Waller 2011; Duby et al. 2001; Emanuelsson et al. 2000; Franzén et al. 1990; Hammen and Weiner 1998; von Heijne et al. 1989; von Heijne 1986).

In the secondary endosymbioses, the red and green algae were engulfed by and integrated within other eukaryotes, and, in most cases, they were reduced extensively to just the plastid with one or two extra membranes (Archibald 2009; Archibald and Keeling 2002; Cavalier-Smith 1999; Gould et al. 2008; Keeling 2010; Kim and Archibald 2009; Matsumoto et al. 2011; Minge et al. 2010; Palmer 2003). The secondary plastid genomes, like those in primary plastids, encode only about 200 genes and rely heavily on their host nuclear genomes, which is enriched by EGT from both the plastid genome and more prominently the nuclear genome of the red or green algal endosymbiont (Archibald 2007; Gould et al. 2008; Keeling 2009; Kim and Archibald 2009). Even in cryptophytes and chlorarachniophytes, whose endosymbionts retain their highly reduced nucleomorphs, the endosymbionts remain vitally dependent on their host nuclear genomes, where the majority of the plastid-targeted proteins and the proteins for maintenance of the nucleomorph are now encoded (Archibald 2007; Gilson et al. 2006; Gilson and McFadden 2002; Lane et al. 2005). The EGT in the secondary endosymbioses has been complemented by the amendments of the protein targeting system, partly, with the addition of a signal peptide (SP) to the plastid transit peptide (cTP), which enables the plastid proteins to overcome the extra membrane barriers (Deane et al. 2000; Hirakawa et al. 2009; Lang et al. 1998; van Dooren et al. 2001; Wastl and Maier 2000). The extent of EGT from the secondary endosymbionts to the host has been evaluated, sometimes with drastically different results, in diatoms (Bowler et al. 2008; Deschamps and Moreira 2012; Dorrell and Smith 2011; Moustafa et al. 2009), in chromerids (Burki et al. 2012; Woehle et al. 2011) and in dinoflagellates (Minge et al. 2010).

There is also evidence of EGT in *Karenia* and *Karlodinium*, whose plastid is derived from a tertiary haptophyte endosymbiont (Ishida and Green 2002; Patron et al. 2006; Yokoyama et al. 2011). An expressed sequence tag (EST) survey and phylogenetic analyses in *Karlodinium*

micrum has revealed that the plastid is maintained by a chimeric proteome derived mostly from the haptophyte endosymbiont in addition to some plastid-targeted proteins derived from the dinoflagellate host (none of which are involved in photosynthesis) and other sources (Patron et al. 2006). Interestingly, the bipartite targeting signals of these proteins included a typical SP followed by a cTP that differed from those of both haptophytes and dinoflagellates (Patron et al. 2006). These results suggested that the haptophyte-derived plastid might have coexisted for some time side by side the original dinoflagellate peridinin plastid (Patron et al. 2006).

Durinskia baltica and *Kryptoperidinium foliaceum* are the best-studied dinotoms, the dinoflagellates with a tertiary diatom endosymbiont (Chesnick and Cox 1987, 1989; Cox and Rizzo 1976; Dodge 1971; Figueroa et al. 2009; Imanian et al. 2007; Imanian and Keeling 2007; Imanian et al. 2010, 2012; Jeffrey and Vesk 1976; Kite et al. 1988; Kite and Dodge 1985; Tippit and Pickett-Heaps 1976; Tomas and Cox 1973; Tomas et al. 1973). Despite experiencing certain character losses (Chesnick and Cox 1987, 1989; Figueroa et al. 2009; Tippit and Pickett-Heaps 1976), the dinotom endosymbiont is unique in retaining many of its original features including an extra surrounding membrane, most likely derived from its original cell membrane (Eschbach et al. 1990), its own mitochondria and its prominent nucleus (Cox and Rizzo 1976; Dodge 1971; Jeffrey and Vesk 1976; Schnepf and Elbrachter 1999; Tomas and Cox 1973; Tomas et al. 1973). The nucleus of this endosymbiont is much larger and contains much more DNA (Kite et al. 1988) than the nucleomorphs of either chlorarachniophytes or cryptophytes, or even the nucleus of its close free-living relatives, the diatoms *Thalassiosira pseudonana* or *Phaeodactylum tricornutum*.

The obligate and permanent symbiosis of the tertiary diatom endosymbiont with its dinoflagellate host in dinotoms raises important and interesting questions regarding the extent of

genetic and genomic integration of the endosymbiont and its host. These questions become more intriguing in the light of the recent studies that indicate neither the plastid genome nor the mitochondrial genomes have been substantially reduced or affected in any significant way compared to their free-living counterparts (Imanian et al. 2010, 2012). In this study, one main question is asked: Does the host nuclear genome encode any gene acquired through EGT? Despite the established nature of endosymbiosis in dinotoms, very little change in their organelle genomes has been detected (Imanian et al. 2010, 2012), and this promotes the expectation of few or no EGTs to the host nuclear genome with respect and in response to its 'permanent guest' and its organelles. In the case of plastid, there is an extra layer of complexity to reflect on since the original dinoflagellate plastid has been replaced by that of the diatom endosymbiont. This implies that the host nucleus, at least once, encoded many genes (mostly of a red algal origin) for its original peridinin plastid. With its loss, the dinoflagellate old plastid-targeted genes might be expected to be lost or gone awry as well, or alternatively, mutated, modified and targeted to the new endosymbiont plastid.

In order to address and answer these questions and evaluate the above-mentioned expectations, a dinoflagellate splice leader cDNA library was prepared for the dinotom *D. baltica* and subsequently subjected to 454 sequencing. The sequences were extensively examined especially through BLAST searches and phylogenetic analyses. Our results indicate that the host nucleus encodes and expresses many mitochondrial genes and just a few plastid genes mainly with a dinoflagellate affinity. More interestingly, our results corroborate with our expectations arising from the small degree of endosymbiotic reduction since out of thousands of sequences only a handful of diatom genes were found in the host cDNA sequences, which most likely represent a small contamination by the endosymbiont nuclear transcripts.

Results

The assembly of SL cDNA sequences of *D. baltica*

The pyrosequencing of SL cDNA of *D. baltica* produced a total of 553,695 reads with an average length of 351 bp and 59.7% GC content. The *de novo* assembly was carefully examined, using consed 23 (Gordon et al. 1998), and the misaligned reads were removed. The final assembly contained 65% of all the reads, assembled into 5,625 large sequences with an average of 63.0% GC content. This Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GAAT00000000. The version described here is the first version, GAAT01000000.

The host putative nuclear-encoded mitochondrial proteins of *D. baltica*

Through BLASTP homology searches 42 protein-encoding sequences with putative mitochondrial functions were identified from the SL cDNA library of *D. baltica* (Table 4.1). The coding sequences of these proteins have an average of 62.5% GC content, ranging from 57.0% to 70.0% while the GC content of the protein-coding genes encoded in the mitochondria of *D. baltica* and other dinoflagellates and diatoms have an average closer to 30%. The GC content of these 42 proteins is also noticeably higher than that in the nuclear genomes of *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* (48.9% and 46.9%, respectively) and their coding sequences (50.0% and 48.0%, respectively) (Armbrust et al. 2004; Bowler et al. 2008). Since the mitochondrial genome of dinoflagellates and that of the host in *D. baltica* encode only three protein-encoding genes (*cox1*, *cob* and *cox3*), the possibility of transfer of the genes commonly found in the mitochondrial genomes (for example, *nad* genes and *cox2*) from the mitochondrion to its host nucleus was explored. However, no such genes were found. One of the hallmarks of the nuclear genome of many dinoflagellates is the presence of multi-copy genes sometimes

appearing in 1,000 or even 5,000 copies (Bachvaroff et al. 2004; Bachvaroff and Place 2008; Lin 2006). In the set of the *D. baltica* host mitochondrion-targeted sequences 10 proteins have multiple copies, ranging from 2 to 13 paralogues, whereas 32 out of 42 proteins appear only in a single copy (Table 4.1). The amino acid identity of the paralogous sequences ranges from 77% to 99%. These 42 putative nuclear-encoded mitochondrial proteins belong to a variety of functional categories such as amino acid, lipid, and fatty acid metabolism, electron transport, protein processing, transcription and translation, and they include the enzymes of tricarboxylic acid (TCA) cycle, electron transport chain, and subunits of ATP synthase (Table 4.1).

The targeting signals of the host putative mitochondrion-targeted proteins

Multiple sequence alignments of the *D. baltica* 42 putative nuclear-encoded mitochondrial proteins with their homologues in other eukaryotic and/or prokaryotic taxa indicated that 35 had a putative N-terminus, and 23 had an extended N-terminus (marked with a star in Table 4.2) ranging from 16 to 130 amino acids with an average length of 58. In order to amplify and sequence the 5′ end of the truncated sequences, RT-PCR was tried with both the total RNA and cDNA of *D. baltica* as template and a specific primer paired with the splice leader (SL) primer for each sequence, but in most cases they resulted in amplification of many products. The 5′ ends of four truncated sequences were eventually recovered using 5′ RACE. Despite several trials, the 5′ ends of the remaining three truncated sequences could not be recovered (Table 4.1). All the cDNAs with a confirmed complete 5′ end had the conserved dinoflagellate SL (marked with a caret in Table 4.1).

Four out of the 39 proteins with complete N-terminus were mitochondrial carrier proteins and lacked mitochondrial targeting signals or transit peptides (mTP) (Emanuelsson et al. 2000). Mitochondrial carrier proteins usually lack an mTP and instead carry an internal targeting signal

(Habib et al. 2007). The lack of mTP in some mitochondrial carrier proteins has been experimentally confirmed in the dinoflagellate *Karlodinium micrum* (synonym, *K. veneficum*) (Danne and Waller 2011). The same four carrier proteins (Table 4.1) also lacked N-terminal extensions when aligned with their respective prokaryotic homologues.

The N-terminus of each of the remaining 35 proteins was examined in search for mitochondrial targeting signals, and 19 proteins were predicted (TargetP algorithm) to have putative mTPs (Table 4.1). The alignment of the N-terminal peptides of the 35 proteins and analyses of their amino acid compositions and secondary structure (Table 4.2) revealed a similar pattern observed also in the mTPs of mitochondrion-targeted proteins of other eukaryotes (Bedwell et al. 1989; Habib et al. 2007; Hammen and Weiner 1998; Roise et al. 1988; von Heijne 1986; von Heijne et al. 1989) especially the dinoflagellate *K. micrum* (Danne and Waller 2011). An excess of positively charged basic residues (the red boxes in Table 4.2), much fewer negatively charged acidic residues (the blue boxes), in the background of hydrophobic amino acids (the yellow boxes) make the majority of the N-terminal peptides carry a net positive charge, a feature of mTPs of nuclear-encoded mitochondrial proteins in plants, animals, fungi and also dinoflagellates (Bedwell et al. 1989; Danne and Waller 2011; Habib et al. 2007; Hammen and Weiner 1998; von Heijne 1986; von Heijne et al. 1989). More specifically, the positively charged amino acid Arg and the hydrophobic amino acid Ala are used more frequently in these targeting peptides compared to their respective mature proteins while the negatively charged amino acids, Asp and Glu, have been used less frequently (Figure 4.1). A similar pattern is reported for the average percentage of amino acid compositions of most mTPs compared to their mature nuclear-encoded mitochondrial proteins in the dinoflagellate *K. micrum* (Danne and Waller 2011). The other common feature found in many mTPs is the presence of the amphipathic

α-helical secondary structure (Danne and Waller 2011; Roise et al. 1988), which is also detected in the N-terminal peptides of many of *D. baltica* putative mitochondrion-targeted proteins (marked with Xs in Table 4.2).

The host putative nuclear-encoded mitochondrial proteins of *D. baltica* with a likely dinoflagellate ancestry

In order to elucidate the phylogenetic origins of the 42 putative mitochondrion-targeted proteins in *D. baltica*, RAxML 7.2.8 (Stamatakis 2006) was used to reconstruct the Maximum Likelihood phylogenetic trees for each of these proteins. Despite using the strict e-value of 1e-25, the blast output file for many of the *D. baltica* proteins contained large numbers of hits, resulting in large phylogenetic trees. In these cases only the partial tree is shown. The BLASTP output for 8 proteins contained fewer than 5 hits, and/or their length was shorter than 50% of the total length of the alignment. For these proteins no tree reconstruction was attempted, and only their best blast hit against the NCBI non-redundant (NR) database is reported (Table 4.1). Also, the position of *D. baltica* in 7 phylogenetic trees remained unresolved (marked by a question mark in Table 4.1) (Figures 4.S1-4.S4). In 4 of these unresolved trees, *D. baltica* is separated from the well-supported diatom clade (Figures 4.S1, 4.S3B and 4.S4B).

Of the 27 resolved phylogenies, *D. baltica* groups with dinoflagellates in 21 trees with varying degrees of bootstrap support: 68%-89% in 6 trees and more than 90% in the other 15 trees. In 6 protein trees, *D. baltica* shows a strong dinoflagellate affinity, but the trees are comprised of a limited number of taxa, only dinoflagellates and apicomplexans (Figures 4.S5A-C) or dinoflagellates plus two or a few more taxa (Figures 4.S5D-F). The remaining resolved phylogenetic trees contain a large number of taxa, and they are more informative but complex. For example, the position of *D. baltica* in the cysteine desulfurase 1 tree is within dinoflagellates

with strong bootstrap support, and it is separated from well-supported diatom clade (Figure 4.2). A similar pattern is also found in the prohibitin tree, where *D. baltica* branches with dinoflagellates (98% bootstrap support), and the dinoflagellate clade is the sister clade to apicomplexans, and they are separated from the diatom clade (Figure 4.S6A). In 10 other resolved phylogenies the dinoflagellate clade that includes *D. baltica* is strongly supported, and they are separated from the well-supported diatom clades in 9 of these trees (Figures 4.S6B, 4.S7, 4.S8, 4.3-4.5). In two of these trees, diatoms have two separate clades in each tree, with 100% bootstrap support, suggesting that the two proteins have each at least two isoforms in one or more taxa (Figure 4.5).

The phylogeny of the multi-copy putative mitochondrion-targeted proteins in *D. baltica* is more complex due to their different copy numbers, varying evolutionary rates of different isoforms in different taxa and perhaps limited sampling. For instance, in some phylogenies nearly all the *D. baltica* isoforms branch strongly with other dinoflagellates some of which have also more than one copy of the protein (Figure 4.3B and 4.S9). More complex phylogenetic trees are shown in Figure 4.S10. Despite the complicated evolutionary histories of these multi-copy proteins reflected in these trees, the putative dinoflagellate ancestry of *D. baltica* proteins are strongly supported at least for some of the isoforms.

The host putative nuclear-encoded mitochondrial proteins of *D. baltica* with a non-dinoflagellate affinity

Dinoflagellates are not represented in 5 out of the remaining 6 resolved phylogenetic trees where *D. baltica* branches with a non-dinoflagellate taxon. In the cytochrome P450 704C1 isoform 1 phylogeny, *D. baltica* is the sister to the only bacterium present in the tree, and its position is strongly supported (98%) to the exclusion of all other taxa that include green algae

and plants, fungi, oomycetes and a stramenopile (Figures 4.6A). In the small phylogenetic tree of hydroxymethylglutaryl-CoA lyase, *D. baltica* groups with bacteria again with 78% bootstrap support (Figures 4.6B). In the other two small phylogenetic trees, tricarboxylate transport protein and saccharopine dehydrogenase domain-containing protein, *D. baltica* branches with the stramenopile *Aureococcus anophagefferens* (to the exclusion of both a ciliate and diatoms) and the rhizarian *Bigelowiella natans* (to the exclusion of apicomplexans), respectively, and their positions are weakly supported (Figures 4.6C and 4.6D). In two trees, elongation factor Tu and acetyl-CoA carboxylase, *D. baltica* groups with apicomplexans with weak bootstrap support (Figure 4.7). The grouping of *D. baltica* with the apicomplexans, the dinoflagellate sister group, with weak support, especially in elongation factor Tu tree where the dinoflagellate *P. marinus* and *D. baltica* belong to the same strongly supported clade, may still imply a dinoflagellate ancestry for the *D. baltica* protein rather than an HGT from apicomplexans. The dinoflagellate host in *K. micrum* is reported to have acquired three horizontally transferred genes for mitochondrion-targeted proteins from different sources (Danne et al. 2011), and it is possible that *D. baltica* has also acquired the above-mentioned proteins through recent HGT events. However, the limited number of taxa and only weak or moderate support for the position of *D. baltica* in most of these cases do not make a strong case for HGTs in *D. baltica* host nuclear-encoded mitochondrial proteins.

The putative nuclear-encoded plastid proteins in the SL cDNA library of *D. baltica*

Through BLAST homology searches 8 putative nuclear-encoded plastid proteins were identified from the SL cDNA library of *D. baltica* (Table 4.3). These cDNAs have an average of 61.4% GC content, ranging from 48.2% to 68.1% (Table 4.S1) while the GC content of the protein-encoding genes in the plastid genomes of *D. baltica* and other diatoms have an average

closer to 30% (Imanian et al. 2010; Oudot-Le Secq et al. 2007). Four of these cDNAs appear in only one copy while the other 4 have multiple paralogues, from 2 to 6 copies (Table 4.3), with the amino acid identity of the isoforms ranging from 73% to 96%. There was also an RT-PCR-confirmed fused bi-partite cDNA, encoding a plastid adenylate kinase fused to a U-box domain containing protein (ADK-UBOX fusion), which is unique to *D. baltica*. The *D. baltica* putative nuclear-encoded plastid proteins fall under different functional categories such as photosynthesis, carbon utilization, ion transport, cell maintenance and growth and amino acid biosynthesis.

The 5′ ends of most of the nuclear-encoded plastid sequences were recovered and/or confirmed through 5′ RACE. Despite numerous attempts through 5′ RACE and also RT-PCR using the SL and specific primers, the presence or absence of the dinoflagellate SL in 2 transcripts could not be determined (Table 4.3). Based on multiple sequence alignments of these 8 proteins with their respective homologues in other eukaryotes and/or prokaryotes, all the proteins seem to have a complete N-terminus, and 4 are predicted to have an extended N-terminal sequence, with an average length of 76 amino acids, ranging from 25 to 160 amino acids (marked with a star in Table 4.3). Of the cDNAs with the confirmed 5′ end (5′ RACE results) , only 2 have the conserved dinoflagellate SL. Also, only 3 proteins are predicted to have a SP, 2 of which are confirmed to lack the SL and have low GC content, fucoxanthin chlorophyll a/c binding protein (FCP) and thylakoid bound ascorbate peroxidase (APXT), and the third, FeS assembly ATPase SufC (SufC), which in addition to a SP is also predicted to have a cTP (Table 4.3).

The putative nuclear-encoded plastid proteins of *D. baltica* with a dinoflagellate affinity or origin

In order to examine the ancestry of the putative nuclear-encoded plastid proteins, they were subjected to similar phylogenetic analyses. Since the fusion protein was unique to *D. baltica*, separate trees were reconstructed for each component of the fused protein. The position of *D. baltica* in 2 Maximum Likelihood trees, CASTOR and ADK, remained unresolved (Figures 4.8A and 4.8C). The *D. baltica* CASTOR protein lacks the SL and it may be encoded in the endosymbiont nucleus. In the ADK (N-terminal component of the fusion protein) phylogeny, *D. baltica* branches with the dinoflagellate *Heterocapsa triquetra* (49% bootstrap support), but it does so to the exclusion of the strong diatom clade (Figure 4.8A). In the UBOX tree, despite the strong support for the *Durinskia*/*Roombia* clade, the limited number of taxa makes drawing any strong conclusion about its origin in *D. baltica* difficult (Figure 4.8B). The *D. baltica* cDNA for ADK-UBOX fusion protein has a confirmed SL, and it is very likely encoded in the host nucleus, but its origin or origins remain unclear.

In 4 resolved phylogenies, *D. baltica* branches with other dinoflagellates with ≥ 50% bootstrap support. In the chloroplast ascorbate peroxidase (APX) tree, the *D. baltica* protein is grouped with those in the dinoflagellates *Oxyrrhis marina* and *H. triquetra* with 50% bootstrap support (Figure 4.9A). The lack of SL in the *D. baltica* APX protein may suggest that it is encoded in the nucleus of the diatom endosymbiont rather than that of the dinoflagellate host, but grouping of *D. baltica* with other dinoflagellates, though weakly supported, excludes moderately supported diatom and stramenopile clades (Figure 4.9A). In the carbonic anhydrase tree, the position of *D. baltica* at the base of the dinoflagellate clade gains 95% bootstrap support (Figure 4.9B) while in the SufC phylogeny, the *Durinskia*/*Perkinsus* clade is only weakly supported but

separated from the strongly supported clade that includes the plastid copy of SufC (Figure 4.9C). A copy of SufC in *D. baltica* as well as in *K. foliaceum*, in diatoms and phaeophyceans is encoded in the plastid genome, but it seems to have been transferred to the nucleus in other lineages with plastids. Also, in the chloroplast o-acetyl serine lyase (OASL) tree, the *D. baltica* protein (which has a confirmed SL) and its isoforms cluster with the moderately supported dinoflagellate clade that is separated from the two strongly supported diatom clades (Figure 4.9D). Despite the weak bootstrap support in two cases, these 4 proteins seem to be encoded in the nuclear genome of the host in *D. baltica* and have a dinoflagellate ancestry.

The *D. baltica* SufC is the only putative plastid protein with a dinoflagellate affinity that has targeting signals: it is predicted to have both a SP and a cTP, which is rich in the amino acid proline (21.6%) and does not include any phenylalanine (Phe). Presence of Phe, elevated level of hydroxylated amino acids (Ser and Thr), positively charged Arg along with that of hydrophobic Ala and lower usage of negatively charged basic residues characterize the amino acid composition of the cTPs in dinoflagellates (Patron and Waller 2007). None of the remaining three putative nuclear-encoded plastid proteins have a SP.

The putative nuclear-encoded plastid proteins of *D. baltica* with a diatom origin

In the remaining two resolved phylogenetic trees, *D. baltica* is branched with diatoms with high bootstrap support or within the strongly supported diatom clade. In the phylogenetic tree for APXT, the *D. baltica* protein and its paralogue are grouped with two diatoms with 93% bootstrap support to the exclusion of dinoflagellates (Figure 4.11A), and in the FCP tree *D. baltica* is nested within the well-supported diatom clade (Figure 4.11B). The cDNAs for these two proteins have lower GC content than that of the cDNAs in other nuclear-encoded plastid proteins of *D. baltica* (Table 4.S1, see also Table 4.S2), which is closer to the GC content of

their homologues as well as that of the genomes and coding sequences in the diatoms *P. tricornutum* and *T. pseudonana* (~ 50%) (Bowler et al. 2008; Armbrust et al. 2004). Both of these proteins are also predicted to have a SP, but no cTP (Table 4.3). For these two proteins, the leading 30 amino acids after the SP cleavage site were further analyzed, and the average percentage of their amino acid composition was compared to that of the mature proteins (Figure 4.11C). All the hallmarks of the diatom cTPs are found in the leader sequences of these two proteins: the amino acid Phe appears at position +1 and a Pro residue at position +3 of both leader sequences after the SP predicted cleavage site, a feature found in the diatom cTPs (Armbrust et al. 2004; Patron and Waller 2007); the majority of leader residues are hydrophobic (53.5% in both leader peptides); both peptides are enriched in the hydroxylated amino acids and depleted of the polar acidic residues (Figure 4.11C). Based on their phylogeny and the shared features of their targeting signals with those of diatoms, these two *D. baltica* proteins are most likely of a diatom origin and targeted to the plastid. The two proteins also lack the dinoflagellate SL, and in all likelihood they are still encoded in the nucleus of the diatom endosymbiont of *D. baltica*.

Horizontally acquired genes for the tryptophan biosynthesis in *D. baltica*

From the SL cDNA library of *D. baltica* a cDNA encoding anthranilate synthase containing both component I and II (ASase) was recovered along with that of a PCR-confirmed tripartite fused protein, composed of a phosphoribosylanthranilate isomerase, a phosphoribosyltransferase plus a GTP cyclohydrolase domain containing protein (PRAI-PRT-GTPCH fusion). The *D. baltica* PRAI-PRT-GTPCH fusion protein is found in neither dinoflagellates nor diatoms but only in the stramenopile *A. anophagefferens*. ASase (component I and II), PRAI and PRT proteins comprise four of the seven enzymes involved in tryptophan

biosynthesis. The *D. baltica* SL cDNA library was searched for the missing three enzymes (indole-3-glycerol-phosphate synthase and tryptophan synthase subunits α and β), but none was found. Since tryptophan biosynthesis is suggested to be localized in the plastid of diatoms as well as plants, green and red algae (Jiroutová et al. 2007), these proteins were subjected to similar analyses conducted for the nuclear-encoded plastid proteins. Separate trees were reconstructed for the PRAI and PRT components of the fusion protein since the fusion was unique to only *D. baltica* and *A. anophagefferens*. The BLASTP output file for the GTPCH component of PRAI-PRT-GTPCH fusion protein contained fewer than five hits, and thus no phylogenetic tree reconstruction was attempted for this portion of the fused protein. In ASase phylogenetic tree, *D. baltica* is the sister to the haptophyte *Emiliania huxleyi* with no significant support, and in both trees reconstructed for the fusion protein PRAI-PRT-GTPCH, it branches with the stramenopile *A. anophagefferens* with 100% bootstrap support to the exclusion of the strong diatom clade in all three trees (Figure 4.10).

The BLAST results indicate that the phototrophic stramenopiles such as *A. anophagefferens* and *E. siliculosus*, in addition to diatoms and *D. baltica* encode the fused gene for ASase, while dinoflagellates either do not have the genes encoding the two components of ASase or the genes are divergent beyond detection. In the ASase phylogeny, *Durinskia*/*Emiliania* clade is separated from the strong diatom clade, but it remains within the weakly supported stramenopile clade that is in turn nested within a mixed bacterial clade with 70% bootstrap support (Figure 4.10A). This along with the shared fusion marker in *D. baltica,* the photosynthetic stramenopiles and the members of the bacterial clade they belong to imply an early bacterial HGT to this eukaryotic clade, which is consistent with the results of phylogenetic analyses conducted elsewhere (Jiroutová et al. 2007). The presence of the SL in the ASase

cDNA is confirmed through 5′ RACE, and its presence implies that the ASase gene now resides in the nucleus of the host in *D. baltica*. Since ASase is not found in any other dinoflagellate, it is possible that the dinoflagellate host in *D. baltica* acquired it through an HGT.

In the two trees reconstructed for the two components of the PRAI-PRT-GTPCH fusion protein, *D. baltica* and stramenopile *A. anophagefferens* make a strong clade to the exclusion of the strong diatom clade (Figures 4.10B and 4.10C). This and the shared unique character (fusion protein) and the absence of PRAI and PRT from the dinoflagellate sequence data make a strong case for a possible HGT event directly and recently from *A. anophagefferens* to *D. baltica*. Given the functional relatedness of the ASase, PRAI and PRT proteins, it is also possible that their source in *D. baltica* is one and the same, *A. anophageferrens* or its close stramenopile relative. The 5′ RACE result for the transcript of the fusion protein in *D. baltica* indicates that it, unlike the ASase, does not have the dinoflagellate SL, suggesting that the gene might be encoded in the nucleus of the diatom endosymbiont rather than that of the host.

Interestingly, no targeting signals (SP or cTP) are predicted for ASase in *D. baltica*, *Ectocarpus siliculosus* and *A. anophagefferens*. The *D. baltica* fusion protein like its homologue in *A. anophagefferens* is not predicted to have any targeting signal either, implying a cytosolic localization for these enzymes in the two organisms. If this is true, diatoms are the only group of stramenopiles in which tryptophan biosynthesis is localized in the plastid.

Various genetic signals in the entire dinoflagellate host SL cDNA library of *D. baltica*

In order to assess the extent of possible EGT/HGT to the dinoflagellate host in *D. baltica*, the Maximum Likelihood phylogenetic trees were reconstructed for 1,856 proteins from the SL cDNA library of *D. baltica*. Since the sequences in this library were expected to originate almost entirely from the dinoflagellate host, the dominant signal was expected to be a dinoflagellate

signal. Thus, as control, we looked for trees with topologies in which *D. baltica* sequences branched exclusively with a dinoflagellate or within a dinoflagellate clade with ≥ 80% bootstrap support. The strong dominant dinoflagellate signal is indeed what was found: the automatic search identified 886 trees in which *D. baltica* grouped exclusively with dinoflagellates with ≥ 80% bootstrap support. Lowering the bootstrap support to ≥ 50% resulted in retrieving 90 extra trees where *D. baltica* queries branched exclusively with other dinoflagellates. There were also 207 trees exclusively comprised of only *D. baltica* proteins (the query and its paralogues only). Then, we automatically sorted the trees with topologies where the *D. baltica* proteins grouped exclusively with those of a taxon of interest or its clade with ≥ 80% bootstrap support.

PhyloSort (Moustafa and Bhattacharya 2008) was used to estimate the number of unique gene families and to cluster the repetitive trees for the queries with paralogues in both the entire set of trees and the subset of trees with a dinoflagellate signal. The total 1,856 trees were clustered into 590 families with the minimum number of gene overlap set to one. The trees where *D. baltica* branched with dinoflagellates (886) were grouped into 291 unique clusters, and the 207 trees comprised only of the *D. baltica* proteins and their paralogues were clustered into 33 gene families. The repetitive trees in the rest of the sorted trees were manually identified and clustered.

All the trees with the non-dinoflagellate signal were then manually inspected. In all these trees *D. baltica* showed a definite phylogenetic affinity with a non-dinoflagellate taxon or its clade (with ≥ 80% bootstrap support), but the limited number of taxa (< 8) or the absence of any other dinoflagellates or diatoms did not allow us assigning a putative origin for many of these non-dinoflagellate gene/proteins. Thus, the number of non-dinoflagellate proteins decreased in nearly all different classes after the manual inspection. For instance, the automatic search

identified 17 proteins that branched strongly with an apicomplexan or within an apicomplexan clade, but after further inspection this number was reduced to only 3 mainly due to the absence of any other dinoflagellate in these protein trees. Presence of at least another dinoflagellate or a dinoflagellate clade is necessary in distinguishing an HGT to *D. baltica* from apicomplexans which are the sister group of dinoflagellates and in their absence *D. baltica* is expected to branch with them.

Figure 4.12 summarizes the results of automatic search for different phylogenetic affinities of *D. baltica* proteins, and Table 4.S3 provides the list of ids and their possible source for all the non-dinoflagellate, non-diatom sequences found through automatic search (with black font) and after manual inspection (in red). Several examples of these trees where the gene seems to have been acquired through putative HGTs are shown in Figure 4.13. The *D. baltica* putative peptidase, for example, branches within the oomycete clade to the exclusion of the dinoflagellate clade, with high bootstrap support for both clades (Figure 4.13A). In the ubiquitin-activating enzyme E1 tree, *D. baltica* is the sister taxon to the stramenopile *A. anophagefferens* to the exclusion of well-supported alveolate group that also includes the dinoflagellate *P. marinus* (Figure 4.13B). In the putative nexus protein phylogeny, *D. baltica* is nested within the prokaryotes excluded from the alveolates, both clades backed by ≥ 80% bootstrap support (Figure 4.13C). In the acid phosphatase tree, *D. baltica* and the haptophyte *Emiliania huxleyi* are sister taxa, and they branch within the strongly supported clade of green algae and plants to the exclusion of alveolate group (Figure 4.13D). In two other trees, *D. baltica* is the sister to an excavate (Figure 4.13E) and a glaucophyte (Figure 4.13F), respectively, separated strongly from other dinoflagellates and/or alveolates. *D. baltica* is also well-separated from other diatoms or diatom clades in the five out of these six protein trees (Figures 4.13A, 4.13B, 4.13D-F).

The diatom genetic footprint in the SL cDNA library of *D. baltica*

The automatic search through 1,856 reconstructed phylogenetic trees resulted in recovering only 14 trees in which *D. baltica* branched strongly (≥ 80% bootstrap support) with a diatom or a diatom clade. Lowering the bootstrap support to 50% led to retrieving only an additional tree. These 15 trees comprised the initial candidates for EGT in *D. baltica* (Table 4.4). The average GC content of the 15 putative diatom-derived candidate sequences is 58.7%, higher than that of their respective diatom homologues (Table 4.S2). A closer look at these numbers reveals that some of the *D. baltica* cDNAs have a closer GC content to their homologues than others, which might simply reflect the wide range of the GC content within diatoms or limited sampling. Two of the recovered candidate proteins were APXT and FCP proteins, which were also detected during our search for nuclear-encoded plastid proteins and discussed earlier. The strong diatom affinity of the *D. baltica* proteins is apparent in all the 15 trees. However, in some trees the number of taxa is simply inadequate, and the tree does not provide any more information other than a strong diatom kinship (Figure 4.S11). The absence of any dinoflagellate or alveolate taxon or clade with a supported position in the tree is also another shortcoming of some of these phylogenies (Figure 4.S12).

In two large and complex multi-copy protein trees, P-type ATPase and trypsin-like serine protease, *D. baltica* groups strongly with one or more diatoms to the exclusion of apicomplexans or the dinoflagellate/apicomplexan clade (Figure 4.S13). In the phylogenetic tree for DNA topoisomerase 3-beta-1, diatoms appear in three separate branches: two clades with 100% bootstrap support and the third clade that includes only the diatom *Pseudonitzschia multiseries*, the *D. baltica* protein and its paralogue (94% bootstrap support) (Figure 4.14). In this tree, the dinoflagellates are well separated from the diatom/*D. baltica* clade. In a slightly simpler tree, *D.*

baltica branches within one of the two strongly supported diatom clades to the exclusion of both ciliates and apicomplexans (Figure 4.15A). In the last phylogenetic tree in this set, replication protein a large 70 kD subunit, *D. baltica* is nested within the multi-membered diatom clade with 100% bootstrap support, and the dinoflagellate clade is strongly supported and well separated from the diatom clade (Figure 4.15B).

Since the *D. baltica* endosymbiont still retains its own large nucleus, the sequences emanating the diatom signal might have still been encoded in the endosymbiont nucleus and contaminated the host SL cDNA library. In order to address this issue, the 5′ RACE was carried out to recover the 5′ end of the truncated sequences and also to confirm the presence or absence of the dinoflagellate SL, which is absent from the diatom transcripts. The 5′ end of all the truncated sequences was successfully recovered. However, despite numerous trials, we failed to confirm the 5′ end and presence or absence of the SL sequences in four of these cDNAs (denoted by a dash under the SL column in Table 4.4) that appeared to have the complete 5′ end in their alignment with their eukaryotic and/or prokaryotic homologues. Of the 11 sequences for which we were able to amplify and/or confirm the completeness of their 5′ end, one cDNA (isotig02507, coding for a conserved predicted protein) was found to contain several frame shifts resulting in four stop codons in its coding sequence, and in all likelihood it is originated from a non-functional pseudogene (Table 4.4). More importantly, none of these 11 cDNAs with confirmed 5′ end had the dinoflagellate SL. This finding implies that none of the diatom-derived candidate transcripts with the confirmed 5′ end is encoded in the nuclear genome of the host, but derived most likely from that of the diatom endosymbiont.

Discussion

The host nucleus in *D. baltica* encodes putative mitochondrion-targeted proteins predominantly of a dinoflagellate ancestry, none with a diatom origin

From the SL cDNA library of *D. baltica*, 42 protein-encoding sequences were identified with putative mitochondrial functions, many of which were predicted to have an N-terminal targeting signal with conserved mTP features. The phylogenetic analyses with strict criteria suggest that while a few of these proteins might have been acquired through recent HGT events from various sources excluding diatoms (Figures 4.6 and 4.7), the majority have a putative dinoflagellate origin as expected, signifying their vertical inheritance. This, in turn, implies that the nuclear-encoded genes for the diatom-derived endosymbiont mitochondrion (Imanian et al. 2012) are not in the host nucleus but in the nucleus of the endosymbiont.

In a similar study on the dinoflagellate *K. micrum*, whose endosymbiont has lost its mitochondrion and nucleus, it is shown that EGT from the tertiary haptophyte endosymbiont has not contributed at all to the mitochondrial proteome of the dinoflagellate host (Danne et al. 2011). *Karlodinium micrum* and *D. baltica* have independently acquired their tertiary endosymbionts from different lineages, and the obligate and permanent symbiosis in these two organisms is at different stages. Nevertheless, the hosts in these two dinoflagellates have converged in not recruiting any mitochondrial gene through EGT. There is also no report of any significant contribution by other secondary or tertiary endosymbionts through EGT to the mitochondrial proteome of their hosts. Even in *Arabidopsis thaliana*, which has a primary plastid, only a handful of mitochondrion-targeted genes with a putative cyanobacterial origin have been identified (Martin et al. 2002). The small or lack of contribution of the EGT to the mitochondrial proteome of the host in primary, secondary and tertiary endosymbioses may be

due to the selective pressure or lack thereof for replacing or remodeling an already functioning mitochondrial proteome in the already mitochondriate hosts including the dinoflagellate host in *D. baltica*. This selective pressure may also explain the only large scale EGT of mitochondrial genes that of the α-proteobacterial ancestor of the mitochondrion to its most likely amitochondriate host (Cavalier-Smith 2009).

The plastid in *D. baltica* remains almost entirely independent of its host nucleus

All plastids rely heavily on the nucleus of their hosts where hundreds of their essential proteins are encoded. Through BLAST homology searches in the *D. baltica* SL cDNA sequences, only several nuclear-encoded plastid proteins were identified, a few showing a putative dinoflagellate ancestry (Figure 4.9). Three of these proteins lack canonical bipartite targeting signals (SP-cTP), and they may be no longer targeted to the plastid. Instead, they may have found a niche in the dinoflagellate cytosol, especially in the case of carbonic anhydrase and ascorbate peroxidase which have both cytosolic and plastid isozymes. The third protein, o-acetylserine lyase (OASL), is one of the enzymes in the cysteine biosynthesis that occurs in the plastid via acetylserine. The cytosolic pathway proceeds via cystathionine, using different and distantly related enzymes including cystathionine-beta-synthase (CBS), also found in *D. baltica* (isogig03677). Despite the possible presence of both pathways in dinoflagellates (Patron et al. 2006), as the enzymes are only distantly related, the *D. baltica* OASL may not be engaged in the cytosolic cysteine biosynthesis, and it may have found a new function.

The *D. baltica* SufC is the only protein with putative dinoflagellate ancestry that is predicted to have the canonical bipartite plastid targeting signal. Thus, the evidence suggests that it might be targeted to the plastid. However, for two reasons it remains uncertain whether it is actually targeted to the plastid within the endosymbiont: first, a copy of the *sufC* gene is encoded

in the diatom-derived plastid genome of *D. baltica*, so making it biochemically unnecessary and redundant for the host nuclear-encoded copy to be targeted to the plastid; second, since the endosymbiont of *D. baltica*, like other dinotoms, is unique in retaining one extra membrane (the fifth membrane counting from inside the plastid stroma, derived either from the original diatom cell membrane or the dinoflagellate host phagocytic membrane), it is unknown whether the canonical targeting signal on the nuclear copy of SufC could actually take the protein through this unique membrane barrier. If the dinoflagellate SufC protein is not targeted to the plastid of the diatom endosymbiont, it might have found a function in other compartments within the dinoflagellate host. The conservation of the targeting signal of the SufC might even suggest that it might be targeted to the relic plastid of the dinoflagellate, the triple-membraned eyespot.

Also two of the recovered putative nuclear-encoded plastid proteins, FCP and APXT, are most likely encoded in the nuclear genome of the endosymbiont and not the host, and they both have bi-partite plastid targeting signals with the conserved features especially in their diatom homologues (Figure 4.11). This is corroborated with the results of a recent study suggesting that the endosymbiont nucleus in the dinotom *K. foliaceum* encodes the gene for the plastid-targeted oxygen evolving enhancer protein (PsbO) (Yokoyama et al. 2011). In *D. baltica* as well as in *K. foliaceum* the original dinoflagellate peridinin plastid is gone and replaced by the plastids of the endosymbiont, which retains nearly all its organelles including its own nucleus (Cox and Rizzo 1976; Dodge 1971; Jeffrey and Vesk 1976; Schnepf and Elbrachter 1999; Tomas and Cox 1973; Tomas et al. 1973). Not surprisingly, the loss of original peridinin plastid in *D. baltica* seems to have been followed by the loss of the cDNAs for dinoflagellate plastid-targeted proteins and perhaps their genes. On the other hand, retention of the endosymbiont nucleus appears to have made unnecessary the EGT of nuclear-encoded plastid genes to the host nucleus. In *K. micrum*,

the nuclear genome of the host should encode all the genes for the plastid-targeted proteins as it is the only present nuclear genome, and it is shown that most of the recovered nuclear-encoded plastid-targeted proteins have a haptophyte origin (EGT from the tertiary endosymbiont) while some are derived from the red alga that gave rise to the peridinin plastid (EGT from the secondary endosymbiont) (Patron et al. 2006). In *D. baltica*, however, the contribution of the dinoflagellate host nucleus in encoding plastid-targeted proteins seems to be minimal if not null, and the plastids seem to rely entirely on the endosymbiont nucleus.

D. baltica host nuclear genome has acquired many genes from a variety of sources but none from its diatom endosymbiont

Through phylogenetic analyses, automatic sorting algorithms and manual inspections 28 protein trees congruent with HGT to *D. baltica* were identified. Identifying an HGT event and determining its source in an organism is generally challenging (Keeling and Palmer 2008) and more so in an extremely complex organism such as *D. baltica*. Based on our phylogenetic analyses it seems possible that *D. baltica* has gained many genes through multiple putative HGTs not from a single source but from a variety of sources including apicomplexans, stramenopiles, haptophytes, plantae, fungi, excavates and more prominently bacteria (see Table 4.S3 and Figure 4.13). It should be emphasized that some of these lineages (i.e. apicomplexans) are very unlikely sources of an HGT to *D. baltica*, and their grouping with *D. baltica* in the phylogenetic trees is the result of sampling errors or other tree making artifacts. Interestingly, two more plausible cases of HGT in *D. baltica* (Figure 4.10) were not detected through our phylogenetic analyses and automatic sorting pipeline but through BLAST homology searches because they involved two fused proteins, ASase (component I and II) and PRAI-PRT-GTPCH. The two fused proteins are not found in any dinoflagellate. The immediate source of HGT for the

fusion protein ASase cannot be determined confidently. However, in case of PRAI-PRT-GTPCH fusion protein, it seems that *D. baltica* has acquired it from a stramenopile like *A. anophagefferens*, the only organism which has this rare fusion protein. The 5′ RACE results introduce a complicated twist in the story of these two HGTs: while the ASase cDNA has a SL and perhaps is encoded in the dinoflagellate host nucleus, the cDNA for PRAI-PRT-GTPCH fusion protein does not have it and is probably encoded in the endosymbiont nucleus. Where exactly the two fused genes are located, whether they were acquired through one or two HGT events and when they were transferred all remain unknown.

Out of 1,856 reconstructed phylogenetic trees for the *D. baltica* SL cDNA sequences and using automatic sorting algorithms, only 15 proteins with a diatom affinity were identified, 11 of which were found to lack the dinoflagellate SL. This implied that most of these diatom-derived genes were originated from the endosymbiont nucleus. These results in conjunction with the results of the BLAST homology searches for the nuclear-encoded mitochondrion- and plastid-targeted proteins all indicate that there has been no EGT to the host nucleus in *D. baltica*. Considering the permanent and obligate nature of symbiosis in dinotoms (Chesnick and Cox 1987, 1989; Cox and Rizzo 1976; Figueroa et al. 2009; Tippit and Pickett-Heaps 1976; Tomas et al. 1973; Tomas and Cox 1973) and the close association of the two partners over evolutionary time, one would expect a large scale EGT to the nuclear genome of the host. This is not the case. The small degree of reduction in the endosymbiont, the little loss and change in its organelle genomes (Imanian et al. 2010, 2012), and the lack of almost any EGT to the host nuclear genome reveal a strict compartmentalization and division of labor between the two partners in *D. baltica* not seen in any other endosymbiont. Most other secondary and tertiary endosymbionts/plastids reside within the endomembrane system of their host and surrounded by 3 or 4 membranes, the

outermost of which is thought to have been derived from the host phagocytic membrane (Archibald and Keeling 2002; Archibald 2009). The diatom endosymbiont in dinotoms including *D. baltica* has a fifth membrane that separates it from the host cytosol, and it is thought to have been originated from its ancestral diatom cell membrane (Eschbach et al. 1990). This membrane might be the actual physical barrier between the diatom and dinoflagellate in dinotoms, and its endurance over evolutionary time might be the simple reason behind not only the lack of any EGT to the host but also the little reduction seen in the endosymbiont.

Conclusions

It is generally assumed and shown in many instances that permanent symbiosis and EGT go hand in hand. While the permanent nature of symbiosis between the host and endosymbiont in dinotoms such as *D. baltica* is well documented, in this study no evidence for any EGT was found in this dinotom. One of the implications of the lack of EGT to the host in *D. baltica* is that the host mitochondria remain almost entirely dependent on the host nucleus while the endosymbiont mitochondria and plastids seem to rely exclusively on the endosymbiont for their nuclear-encoded proteins. This strict compartmentalization in *D. baltica* is unique, suggesting that the permanent symbiosis is not always accompanied by EGT as seen in other organisms with permanent endosymbionts or endosymbiont-derived organelles.

Materials and methods

Strains and culture conditions

The culture of *Durinskia baltica* (*Peridinium balticum*) CSIRO CS-38 and was obtained from the CSIRO Microalgae Supply Service (CSIRO Marine and Atmospheric Research Laboratories, Tasmania, Australia). The culture was maintained in GSe medium at 22 ºC (12:12 light:dark cycle).

Nucleic acids extraction, preparation, amplification and 5´ RACE

Exponentially growing cells were collected and ground as described previously (Imanian et al. 2007). Cells lysis, DNA extractions, precipitations and purifications were performed for both species as described earlier (Imanian et al. 2010). Total genomic DNA was extracted for polymerase chain reactions (PCR) either as described previously (Imanian et al. 2010), or using Master Pure Complete DNA and RNA Purification Kit (EPICENTRE Biotechnologies, Madison, WI, USA) following the manufacturer's instructions. Total RNA for RT-PCR was obtained as described earlier (Imanian et al. 2007). RNeasy MinElute Cleanup kit (Qiagen, Mississauga, ON) was utilized to clean up the total RNA after DNase treatment according to the manufacturer's instructions. Oligotex mRNA Mini Kit (Qiagen, Mississauga, ON) was used to purify poly A RNA from approximately 25 µg of cleaned-up total RNA based on the manufacturer's instructions. PCR and RT-PCR reactions were performed using specific primers designed based on the genomic and/or the obtained cDNA data as described elsewhere (Imanian et al. 2007, 2010). Long range PCRs were conducted either as described earlier (Imanian et al. 2007, 2010), or using Expand Long Template PCR System kit (Roche Applied Science, Indianapolis, IN, USA) following the manufacturer's instructions. The 5´ ends of truncated transcripts were recovered/ascertained using FirstChoice RLM-RACE kit (Life Technologies, Burlington, ON) and sequenced on both strands using BigDye terminator chemistry.

Splice leader (SL) cDNA construction and amplification for *D. baltica*

Approximately 500 ng of poly A RNA from *D. baltica* was used as template for constructing first and second strand cDNA with Just cDNA Double Stranded cDNA Synthesis kit (Agilant Technologies Canada, Mississauga, ON) according to manufacturer's protocol with one modification: instead of oligo (dT) and random 9mer primers, a dinoflagellate splice leader

(SL) primer (5´- CCGTAGCCATTTTGGCTCAAG-3´) was used. The resulting double-stranded cDNA sample was amplified through PCR and/or long-range PCR with the SL primer in conjunction with the random 9mer primer. The amplified cDNA sample was purified using QIAquick PCR Purification kit (Qiagen, Mississauga, ON), and re-amplified once more through PCR and/or long-range PCR. The optimized PCR conditions were determined to be: 94 ºC for 2 min, 39 cycles of 94 ºC for 15 s, 42 ºC for 30 s, 72 ºC for 5 min, followed by 72 ºC for 6 min, while the long-range PCR conditions were optimized at 92 ºC for 2 min, 34 cycles of 94 ºC for 10 s, 45 ºC for 15 s, 68 ºC for 20 min, followed by 68 ºC for 7 min using buffer 3 from Expand Long Template PCR System kit (Roche Applied Science, Indianapolis, IN, USA).

The cDNA sequencing and assembly

The amplified SL cDNA of *D. baltica* was sequenced using massively parallel GS-FLX DNA pyrosquencing (Roche 454 Life Sciences, Branford, CT, USA). The GS-FLX shotgun libraries and pyrosequencing using the GS-FLX Titanium reagents were carried out at the Génome Québec Innovation Centre. Sequences were assembled *de novo* using gsAssembler 2.5p1 (formerly known as Newbler), edited and re-assembled with CONSED 23 (Gordon et al. 1998, 2001), which was also used for designing various primers including outer and inner primers to amplify the 5´ ends of transcripts paired with 5´ RACE outer and inner primers, respectively.

Assessing the phylogenetic footprints of diatoms and other taxa in the SL cDNA sequences of *D. baltica*

ORFPredictor (Min et al. 2005) was used to translate the *D. baltica* SL cDNA sequences, which were subsequently used as queries in a BLASTP (Altschul et al. 1990) homology search with an e-value < 1e-5 against the protein collections from complete genomes and EST databases

(the complete list of the taxa is found in Table 4.S4). In order to retrieve all the aligned sequences (hits) for each query, the default value for blastp parameter –max_target_seqs was changed from 100 to 100,000. The sequence retrieval, alignment and tree reconstruction were conducted as described elsewhere (Burki et al. 2012) with the following modifications. CDHIT (Li and Godzik 2006) was utilized to remove redundant sequences and close paralogues from each protein database to simplify interpretations of the resulting phylogenetic trees (with 85% identity threshold for clustering). The blast output file was parsed with a strict e-value threshold of 1e-25 to reduce the number of distantly related paralogues and to generate multiple fasta files including each protein query and the corresponding hits. The sequences in each file were aligned using MAFFT (Katoh and Toh 2008) with the fftnsi option, and alignment positions were selected and sites containing more than 10% of were removed using TRIMAL (Capella-Gutiérrez et al. 2009). The alignment files with fewer than 5 species or when the query sequences were shorter than 50% of the total length of the alignments were discarded at this stage. FastTree (Price et al. 2009) with the WAG model of evolution (Whelan and Goldman 2001) was used to reconstruct initial trees. A Ruby script was used to reduce the complexity of these trees by keeping only a subset of representative operational taxonomic units (OUT) in well-supported clades (> 0.9 Shimodaria-Hasegawa or SH (Shimodaira and Hasegawa 1989) support); dinoflagellate and diatom taxa were flagged and left out of this procedure. From other taxa, 10 prokaryotes, 10 green algae, 10 red algae, 10 glaucophytes, 5 streptophytes, and 2 from all the rest of the taxa retained. The sequences for the retained taxa were retrieved anew into multiple fasta files, and MAFFT (Katoh and Toh 2008) with the fftnsi option was used to align them and TRIMAL (Capella-Gutiérrez et al. 2009) was used as described above to choose the aligned positions and remove the gaps. RAxML 7.2.8 (Stamatakis 2006) was run to reconstruct the

phylogenetic trees, with LG substitution matrix + Γ4 + F evolutionary model with 100 bootstrap replicates. A Perl script (Chan, Reyes-Prieto, et al. 2011; Chan, Yang, et al. 2011) was used in the initial sorting of these trees with a variety of preconditions (i.e. the query sequences from *D. baltica* should be monophyletic with the members of diatoms, or dinoflagellates, or others such as green algae, stramenopiles, prokaryotes, etc. with a specified percentage of support, in most cases at least 80%). PhyloSort was also used to estimate the number of gene families and to cluster the repetitive phylogenetic trees for the queries with multiple paralogues (Moustafa and Bhattacharya 2008). Then, the trees under all the preconditions, with the exception of monophyletic grouping of the *D. baltica* queries with dinoflagellates, which constituted most of the trees, were manually examined. The trees with a non-dinoflagellate signal that contained fewer than 8 taxa were deemed non-informative and discarded. Presence of at least a dinoflagellate or a diatom taxon or their respective clades was considered as a necessary criterion in order to assign a non-dinoflagellate non-diatom, but a eukaryotic taxon's signal to a *D. baltica* protein query. Also, presence of both at least a dinoflagellate and a diatom taxon in the trees with diatom signals was considered a necessary precondition for assigning a diatom signal to a *D. baltica* protein query.

Identification and annotation of organelle-targeted genes

The protein sequences for the SL cDNA sequences (described above) were used as queries in BLASTP homology searches against the protein collections from the following subsets of the NCBI non-redundant (NR) databases and/or the Joint Genome Institute (JGI) downloaded on 2012/03/31; a) the available mitochondrial and/or plastid genomes from red, green and glaucophyte algae, diatoms including the endosymbiont of *D. baltica* and *K. foliaceum* and their host's mitochondrial genes, other stramenopiles, haptophytes, cryptophytes, apicomplexans

(mitochondrial and/or apicoplast genomes), ciliates, chromerids (*Chromera velia*), amoebozoans, several representatives from opistokonts including human, rat and yeast, and *Malawimonas jakobiformis* and *Reclinomonas americana*, and the organelle-encoded genes of other dinoflagellates; b) mitochondrial genes including nuclear-encoded, organelle-targeted genes; c) plastid genes including nuclear-encoded, organelle-targeted genes. The initial candidates for putative mitochondrial or plastid targeted genes in the SL cDNA sequences of *D. baltica* were selected based on their BLAST score (threshold e-value < 1e-5) against the sequences in the first two databases.

The candidate sequences were, then, used as queries against the entire NR databases. The BLAST results were examined manually at this stage. In order to be selected as a putative organelle-targeted protein, the candidate sequence had to meet at least one of the following criteria: its best BLAST homologues against the entire NR database should be encoded within a mitochondrial or plastid genome; it should be known to be targeted to the mitochondrion or plastid; its putative function should be part of a biochemical pathway or process known to occur in one of the two organelles. Presence of targeting signal was also considered in the final selection of the nuclear-encoded organelle proteins.

The putative organelle-targeted proteins were annotated based on their best and most informative homologues found in BLASTP searches against the entire NR database and/or the domain homology searches against the Conserved Domain database (Marchler-Bauer et al. 2009).

Targeting signal predictions

The presence/absence of the 5´ end of transcripts was determined after aligning them with their best eukaryotic (mitochondrial, plastid and cytosolic copies included whenever

available) and/or prokaryotic homologues. TargetP (Emanuelsson et al. 2000) was used to check for mitochondrial transit peptide (mTP), while ChloroP (Emanuelsson et al. 1999) and SignalP 3.0 (Bendtsen et al. 2004) with NN option were used to search for a plastid transit peptide (cTP) and a signal peptide (SP), respectively. If a signal peptide was predicted, its predicted sequences were removed prior to search for cTP. Amino Acid Calculator (http://proteome.gs.washington.edu/cgi-bin/aa_calc.pl) was used to calculate the amino acid composition of the mTP, cTP, SP and the mature proteins. Webserver SCRATCH (http://scratch.proteomics.ics.uci.edu/) was used to predict the putative α-helices, and their potential amphipathic properties were examined by helical wheel projection (http://cti.itc.virginia.edu/~cmg/Demo/wheel/wheelApp.html). The peptides (≥ 5 amino acids) that contributed to both the α-helical secondary structure and amphipathic properties were predicted to make up an amphipathic α-helix (Danne and Waller 2011).

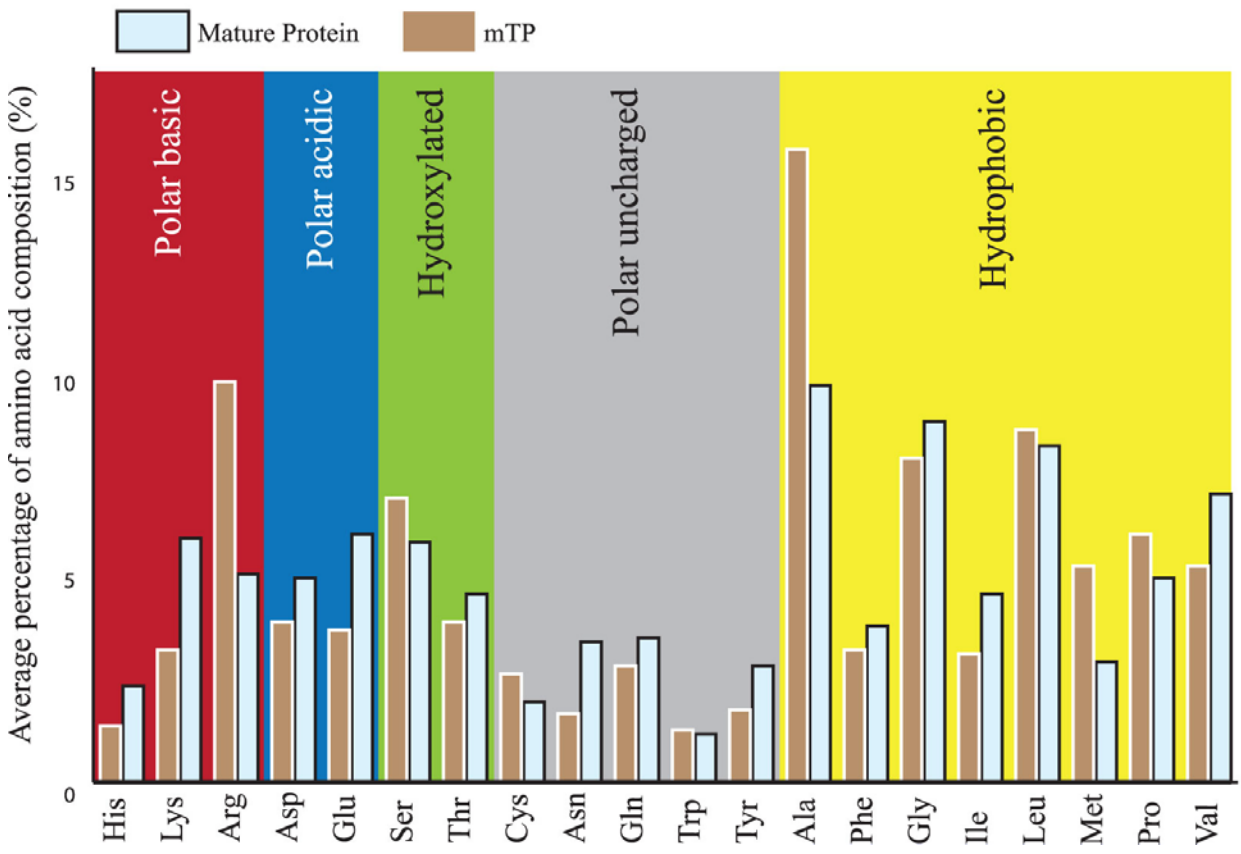

Figure 4.1: Average percentage of amino acid composition in the *Durinskia baltica* mitochondrial transit peptides (mTPs) compared to that of the mature proteins.

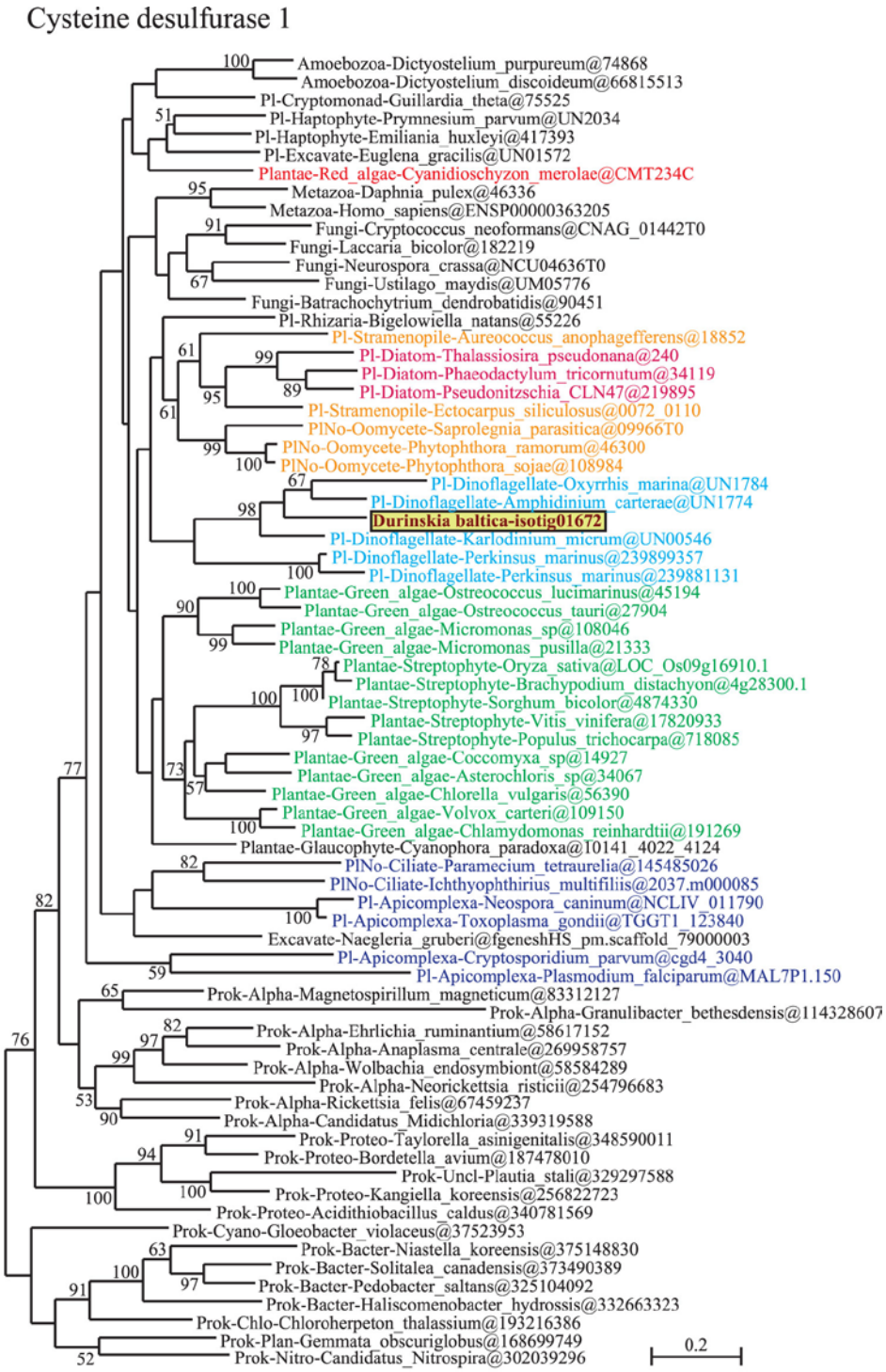

Figure 4.2: The maximum likelihood trees for cysteine desulfurase 1, partial tree.

Numbers at the nodes indicate bootstrap support ≥ 50%. A yellow box highlights the position of *D. baltica*. Several taxa of interest are color-coded: Dinoflagellates in light blue; Apicomplexans and Ciliates in dark blue; Diatoms in scarlet; Stramenopiles in orange; Green algae and Plants in green; Red algae in red; All other taxa in black. Pl, taxon with a plastid; PlNo, taxon without plastids.

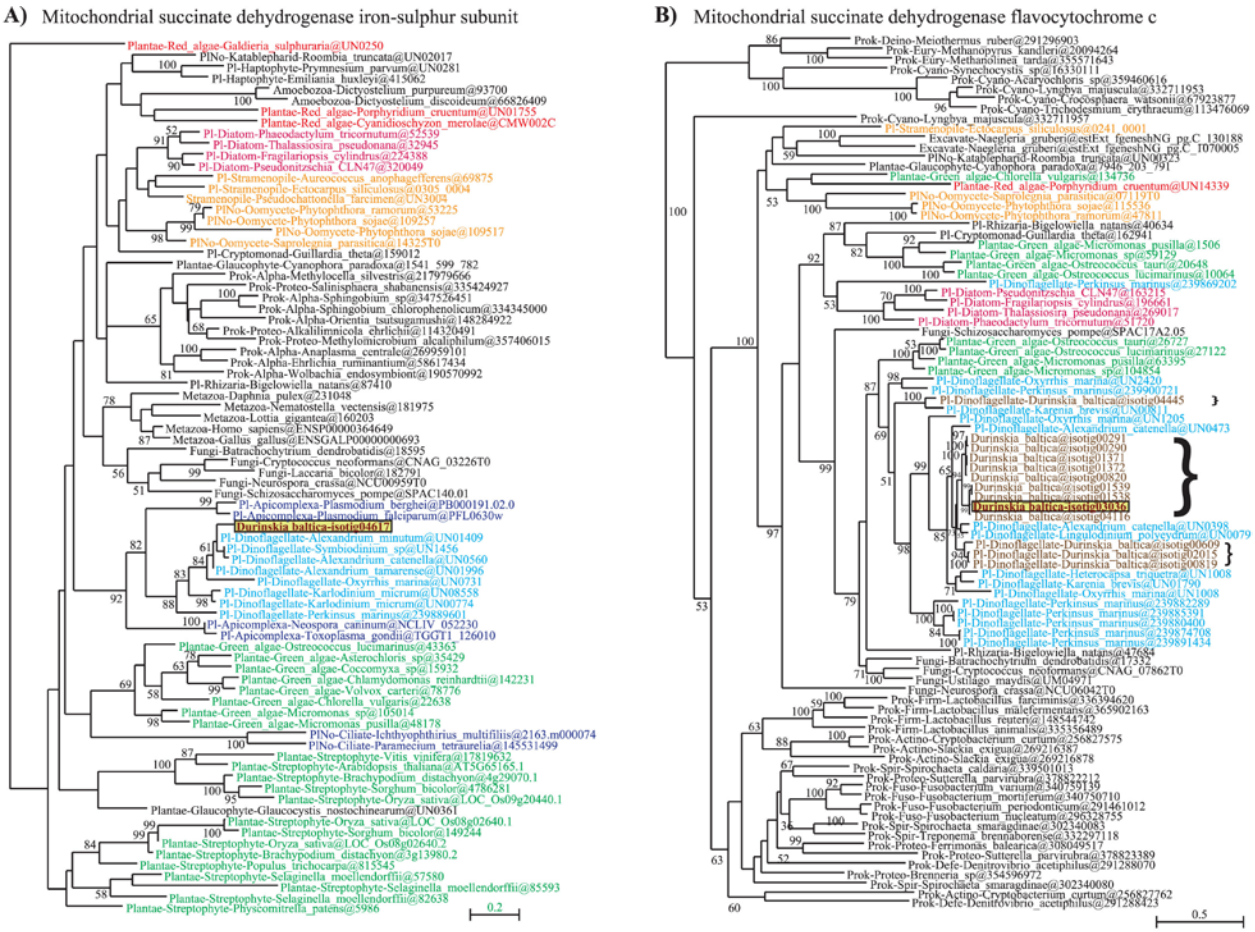

Figure 4.3: The maximum likelihood trees for the host putative nuclear-encoded mitochondrial proteins in *Durinskia baltica* (SdH FeS subunit and SdH FCytC).

A) Mitochondrial succinate dehydrogenase iron-sulphur subunit, partial tree, B) Mitochondrial succinate dehydrogenase flavocytochrome c. Numbers at the nodes indicate bootstrap support ≥ 50%. A yellow box highlights the position of *D. baltica*. The braces indicate the *D. baltica* protein's isoforms. Several taxa of interest are color-coded: Dinoflagellates in light blue; Apicomplexans and Ciliates in dark blue; Diatoms in scarlet; Stramenopiles in orange; Green algae and Plants in green; Red algae in red; All other taxa in black. Pl, taxon with a plastid; PlNo, taxon without plastids.

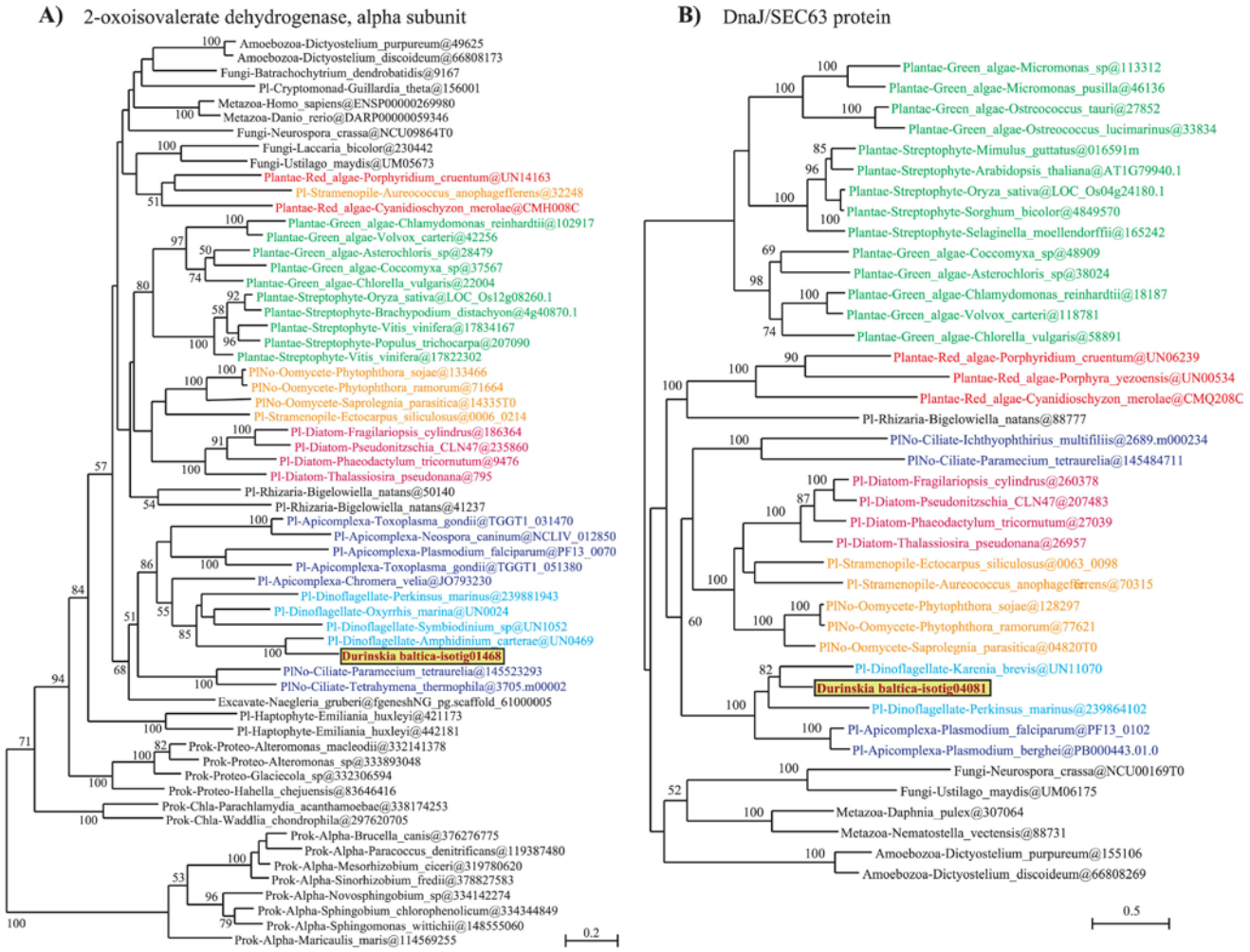

Figure 4.4: The maximum likelihood trees for the host putative nuclear-encoded mitochondrial proteins in *Durinskia baltica* (OIVDH Alpha subunit and DnaJ/SEC63).

A) 2-oxoisovalerate dehydrogenase, alpha subunit, partial tree, B) DnaJ/SEC63 protein. Numbers at the nodes indicate bootstrap support ≥ 50%. A yellow box highlights the position of *D. baltica*. Several taxa of interest are color-coded: Dinoflagellates in light blue; Apicomplexans and Ciliates in dark blue; Diatoms in scarlet; Stramenopiles in orange; Green algae and Plants in green; Red algae in red; All other taxa in black. Pl, taxon with a plastid; PlNo, taxon without plastids.

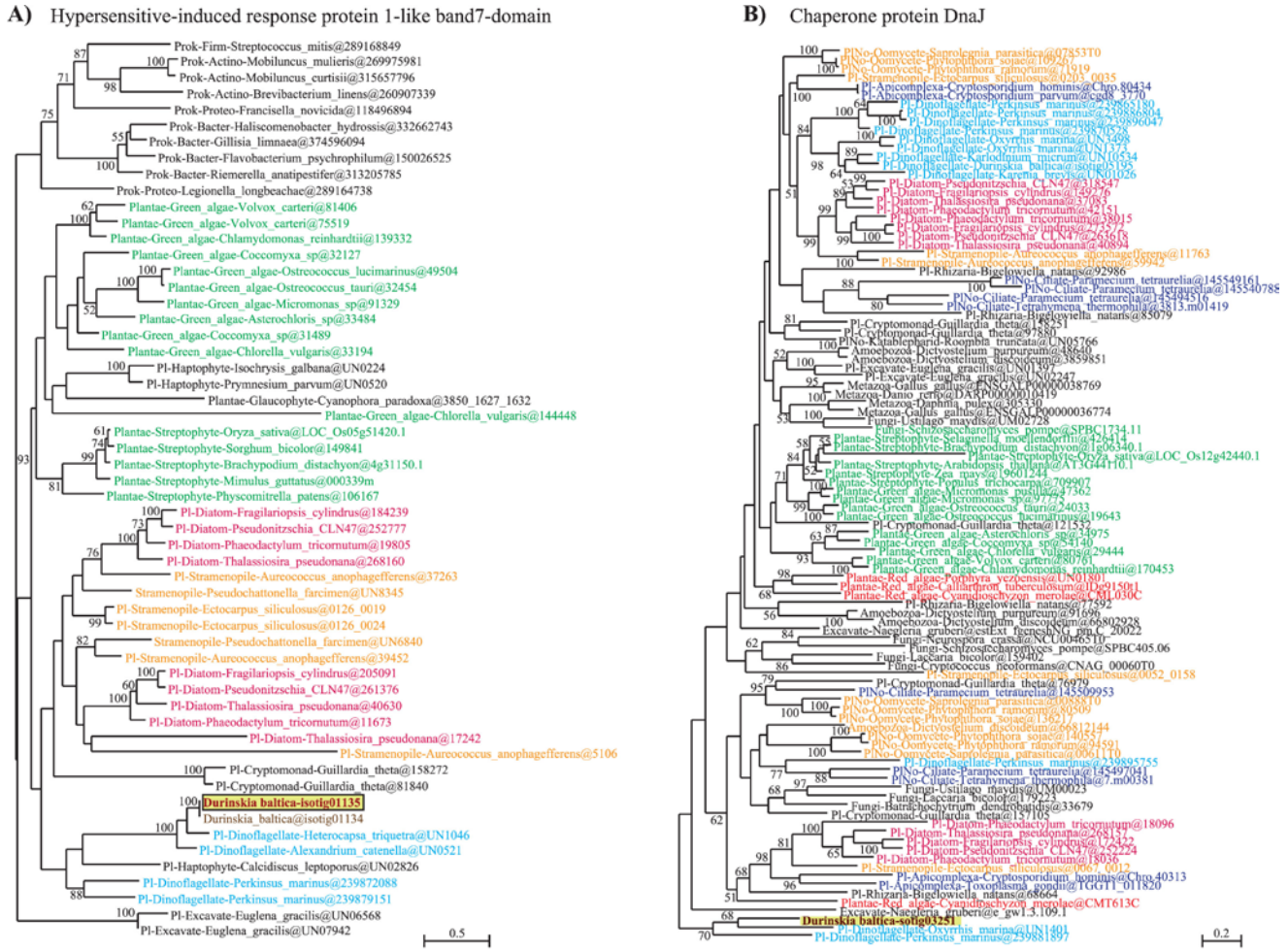

Figure 4.5: The maximum likelihood trees for the host putative nuclear-encoded mitochondrial proteins in *Durinskia baltica* (HIRP and DnaJ).

A) Hypersensitive-induced response protein 1-like, band7-domain, B) Chaperone protein DnaJ, partial tree. Numbers at the nodes indicate bootstrap support ≥ 50%. A yellow box highlights the position of *D. baltica*. Several taxa of interest are color-coded: Dinoflagellates in light blue; Apicomplexans and Ciliates in dark blue; Diatoms in scarlet; Stramenopiles in orange; Green algae and Plants in green; Red algae in red; All other taxa in black. Pl, taxon with a plastid; PlNo, taxon without plastids.

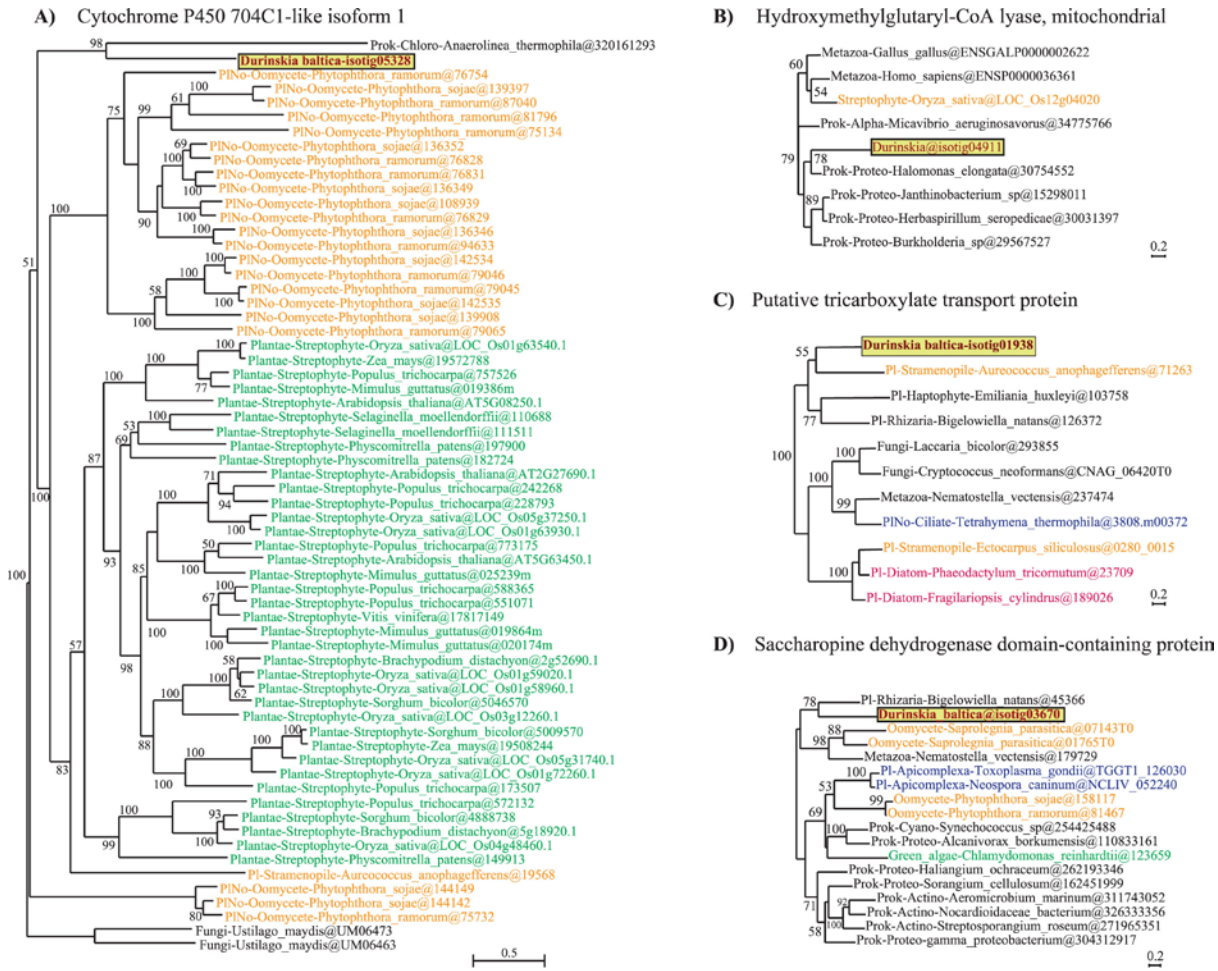

Figure 4.6: The maximum likelihood trees for the host putative nuclear-encoded mitochondrial proteins in *Durinskia baltica* (CytP450 and HMG CoAL).

A) Cytochrome P450 704C1-like isoform 1, B) Hydroxymethylglutaryl-CoA lyase, mitochondrial, C) Putative tricarboxylate transport protein, D) Saccharopine dehydrogenase domain-containing protein. Numbers at the nodes indicate bootstrap support ≥ 50%. A yellow box highlights the position of *D. baltica*. Several taxa of interest are color-coded: Dinoflagellates in light blue; Apicomplexans and Ciliates in dark blue; Diatoms in scarlet; Stramenopiles in orange; Green algae and Plants in green; Red algae in red; All other taxa in black. Pl, taxon with a plastid; PlNo, taxon without plastids.

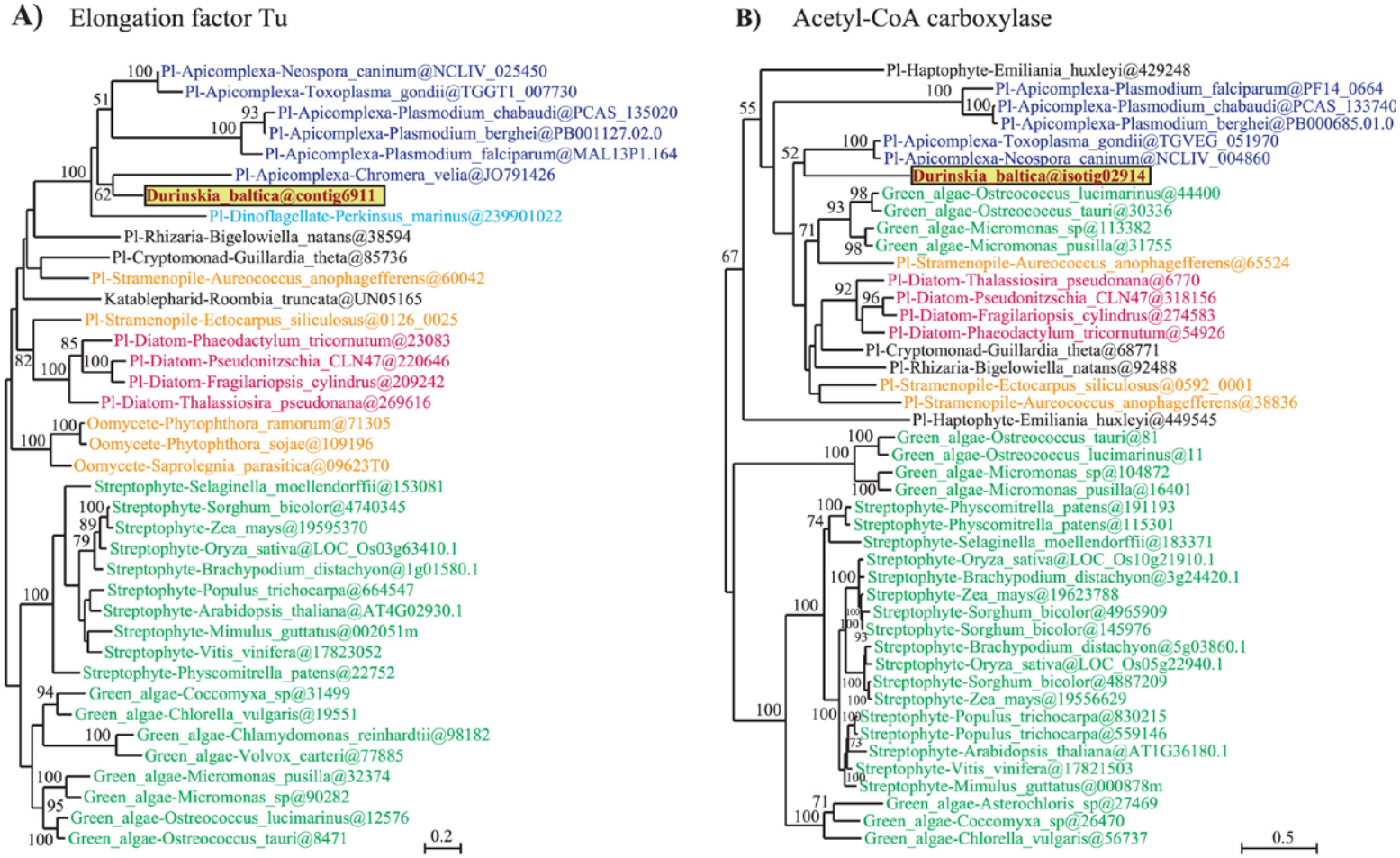

Figure 4.7: The maximum likelihood trees for the host putative nuclear-encoded mitochondrial proteins in *Durinskia baltica* (EFTu and AcCoAC).

A) Elongation factor Tu, partial tree, B) Acetyl-CoA carboxylase, partial tree. Numbers at the nodes indicate bootstrap support ≥ 50%. A yellow box highlights the position of *D. baltica*. Several taxa of interest are color-coded: Dinoflagellates in light blue; Apicomplexans and Ciliates in dark blue; Diatoms in scarlet; Stramenopiles in orange; Green algae and Plants in green; Red algae in red; All other taxa in black. Pl, taxon with a plastid; PlNo, taxon without plastids.

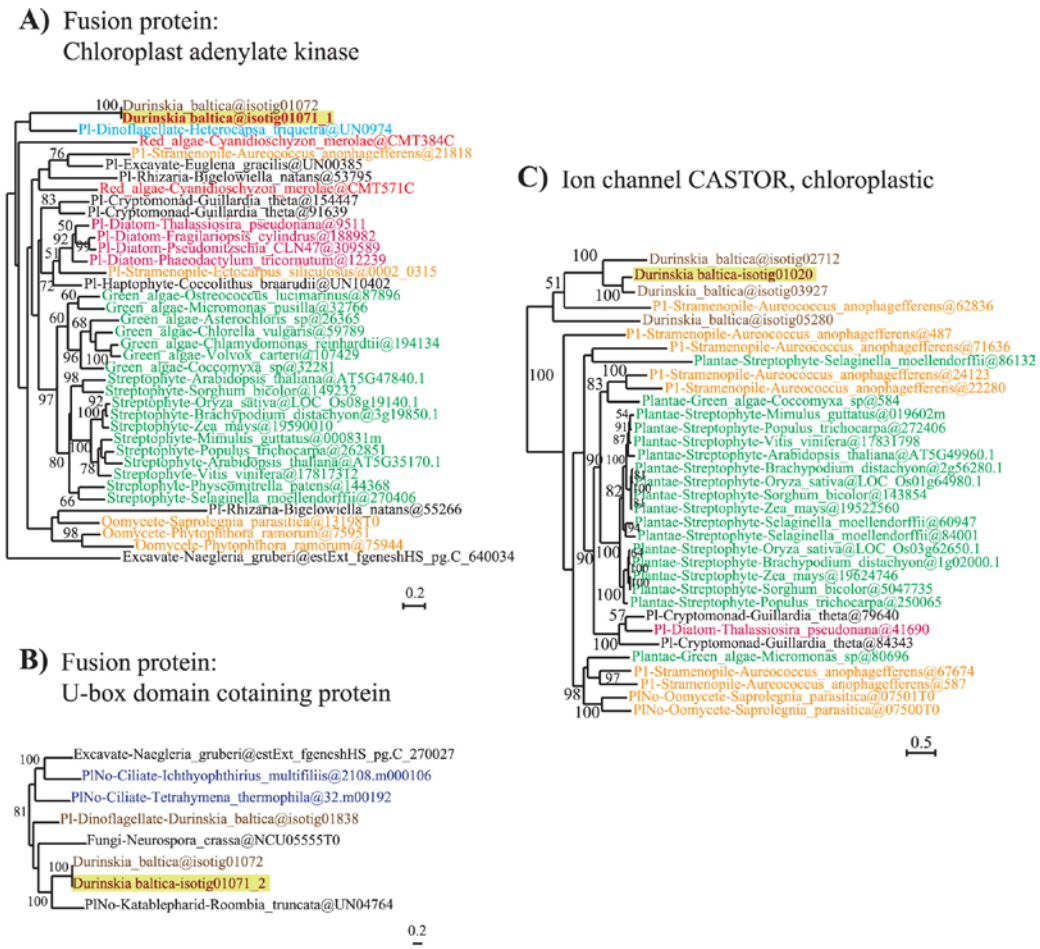

Figure 4.8: The maximum likelihood trees for the host putative nuclear encoded plastid proteins in *Durinskia baltica* (Fusion Protein AK-UBox and CASTOR).

A) Soluble starch synthase 1, chloroplastic/amyloplastic, B) Ion channel CASTOR, chloroplastic, C) Fusion protein, chloroplast adenylate kinase, partial tree, D) Fusion protein, U-box domain containing protein. Numbers at the nodes indicate bootstrap support ≥ 50%. A yellow box highlights the position of *D. baltica*. Several taxa of interest are color-coded: Dinoflagellates in light blue; Apicomplexans and Ciliates in dark blue; Diatoms in scarlet; Stramenopiles in orange; Green algae and Plants in green; Red algae in red; All other taxa in black. The *D. baltica* protein's isoforms are in dark brown. Pl, taxon with a plastid; PlNo, taxon without plastids.

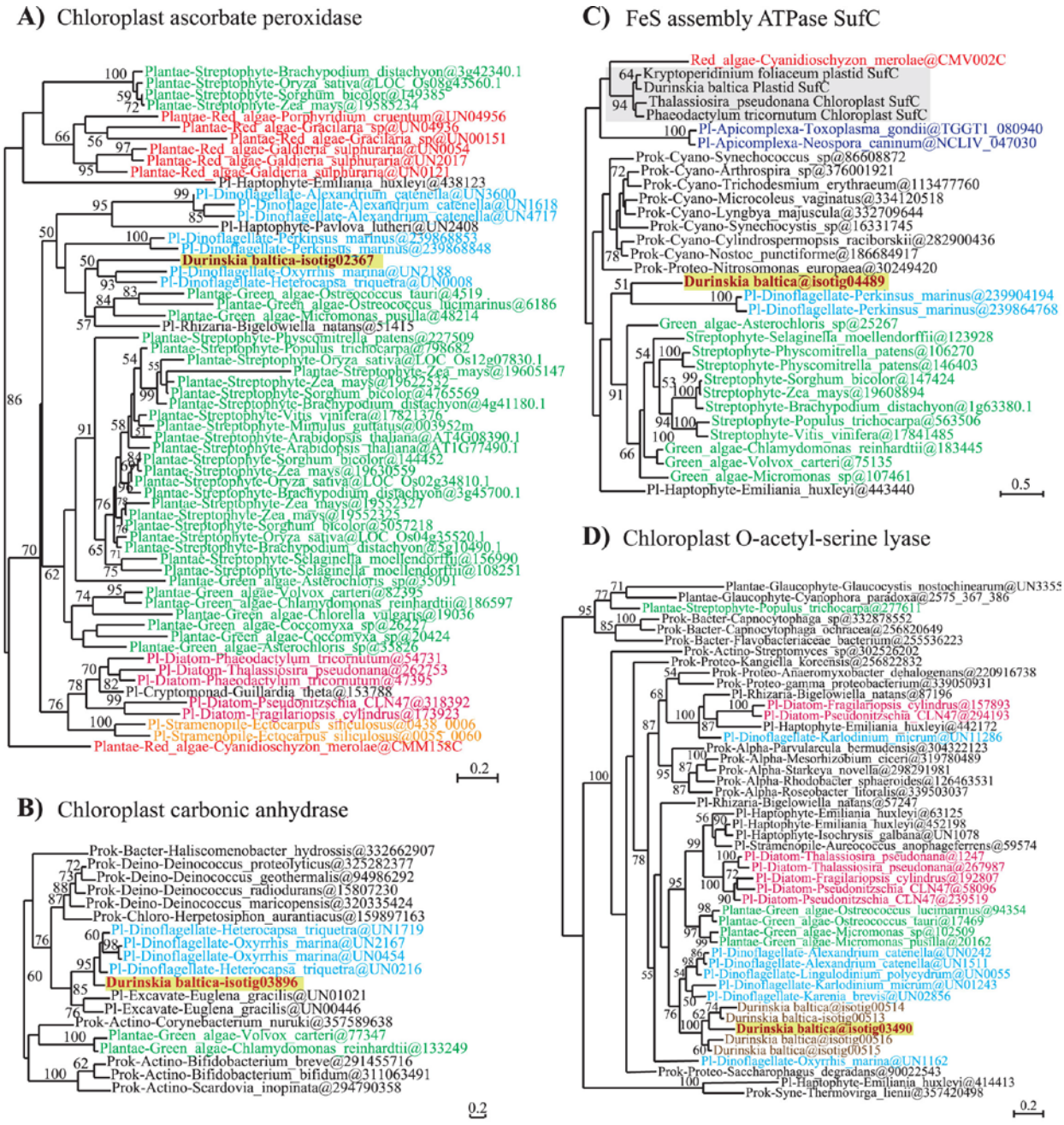

Figure 4.9: The maximum likelihood trees for the host putative nuclear encoded plastid proteins in *Durinskia baltica* (APX, CA, SufC and OASL).

A) Chloroplast ascorbate peroxidase, B) FeS assembly ATPase SufC, C) Chloroplast carbonic anhydrase, D) Chloroplast o-acetyl serine lyase, partial tree. Numbers at the nodes indicate bootstrap support ≥ 50%. A grey box indicates the plastid-encoded SufC in B.A yellow box highlights the position of *D. baltica*. Several taxa of interest are color-coded: Dinoflagellates in light blue; Apicomplexans and Ciliates in dark blue; Diatoms in scarlet; Stramenopiles in orange; Green algae and Plants in green; Red algae in red; All other taxa in black. The *D. baltica* protein's isoforms are in dark brown. Pl, taxon with a plastid; PlNo, taxon without plastids.

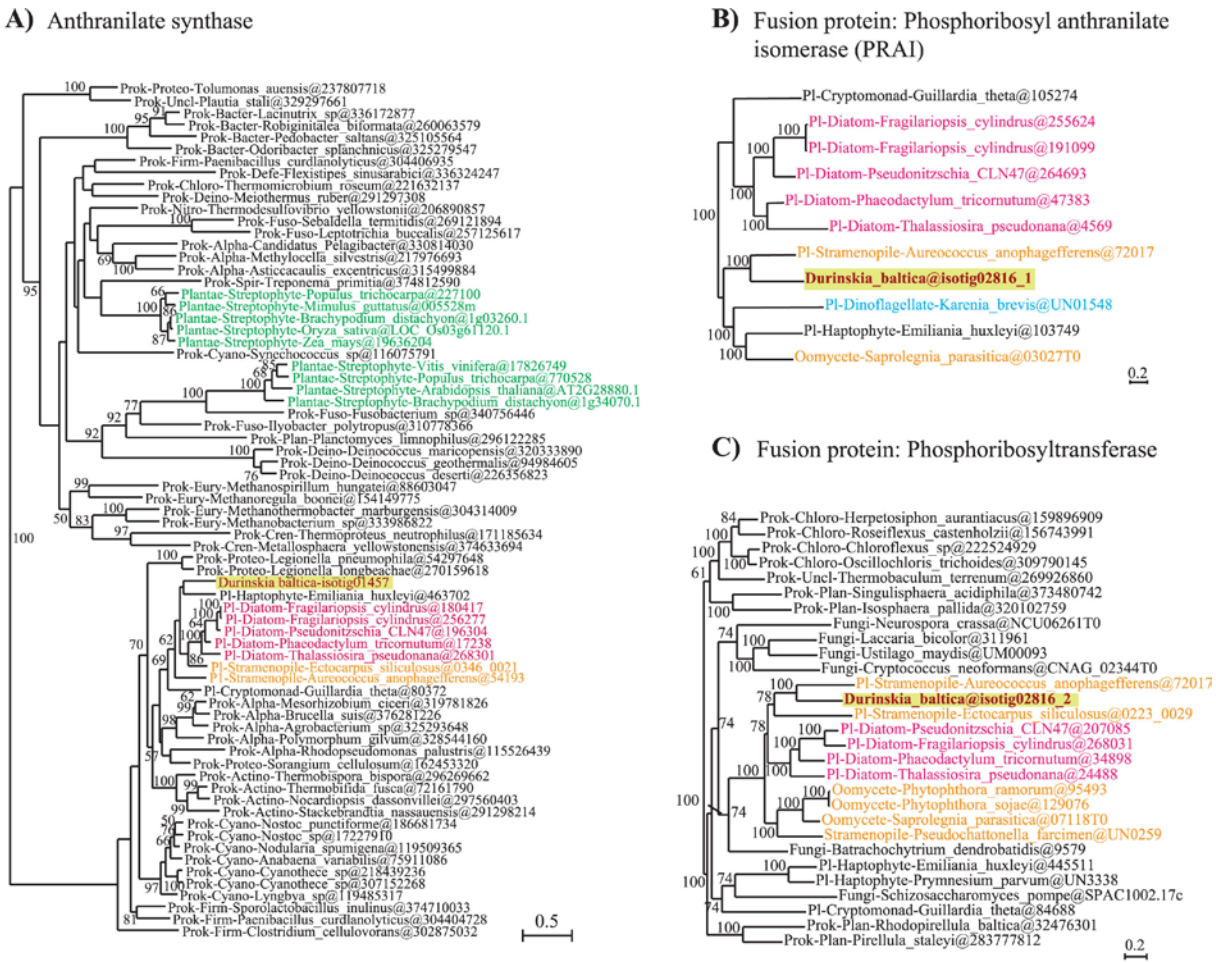

Figure 4.10: The maximum likelihood trees for the host proteins in *Durinskia baltica* inferring horizontal gene transfer events (ASase and the fusion protein PRAI-PRT) .

A) Anthranilate synthase, B) Fusion protein, phosphoribosyl anthranilate isomerase, C) Fusion protein, phosphoribosyltransferase, partial tree. Numbers at the nodes indicate bootstrap support ≥ 50%. A yellow box highlights the position of *D. baltica*. Several taxa of interest are color-coded: Dinoflagellates in light blue; Apicomplexans and Ciliates in dark blue; Diatoms in scarlet; Stramenopiles in orange; Green algae and Plants in green; Red algae in red; All other taxa in black. Pl, taxon with a plastid; PlNo, taxon without plastids.

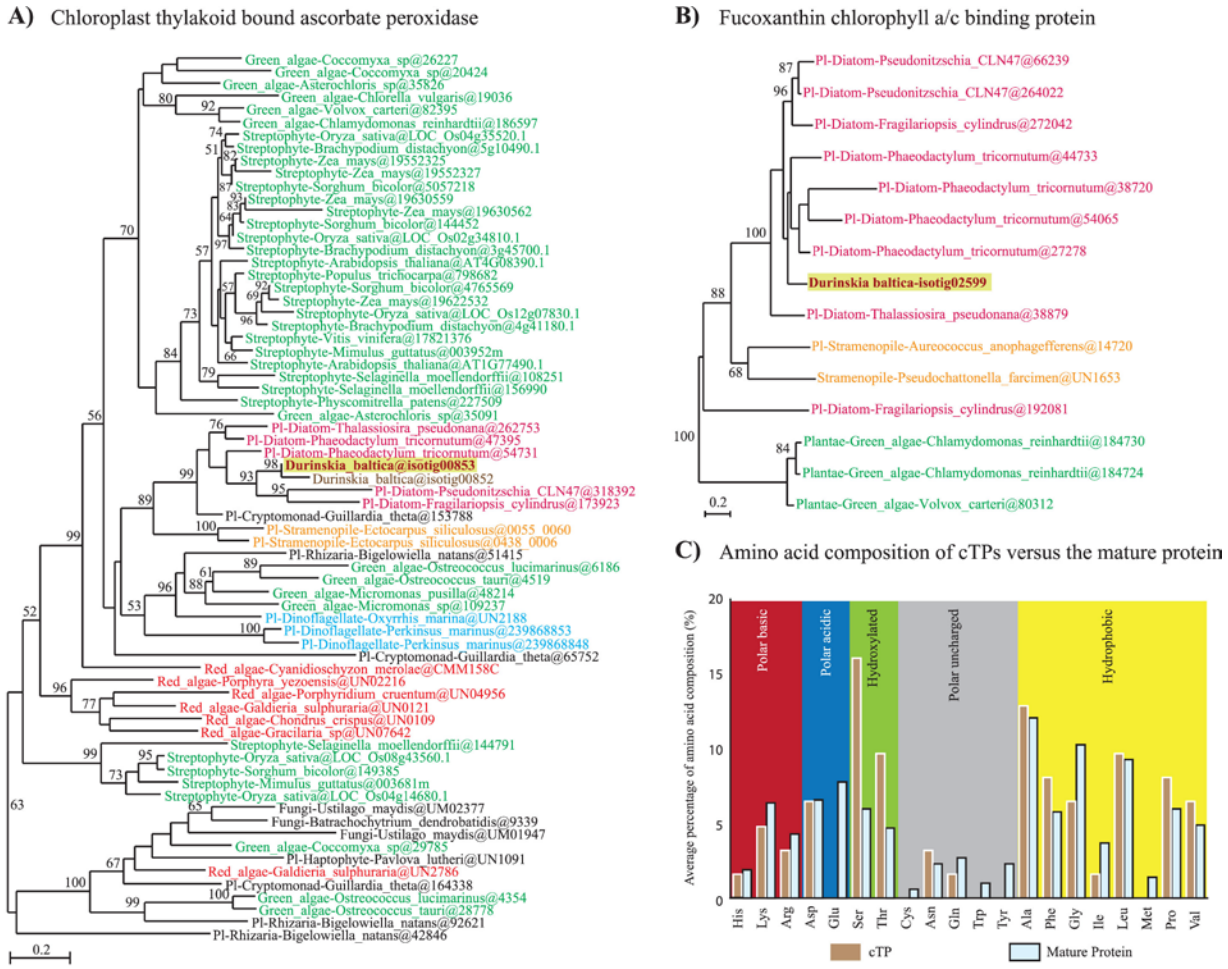

Figure 4.11: The maximum likelihood trees for the host putative nuclear encoded plastid proteins in *Durinskia baltica* (APXT and FCP).

A) Chloroplast thylakoid bound ascorbate peroxidase, B) Fucoxanthin chlorophyll a/c binding protein. Numbers at the nodes indicate bootstrap support ≥ 50%. A yellow box highlights the position of *D. baltica*. Several taxa of interest are color-coded: Dinoflagellates in light blue; Apicomplexans and ciliates in dark blue; Diatoms in scarlet; Stramenopiles in orange; Green algae and plants in green; Red algae in red; All other taxa in black. C) Amino acid composition of plastid transit peptides versus that of the mature protein in these two proteins. Pl, taxon with a plastid; PlNo, taxon without plastids.

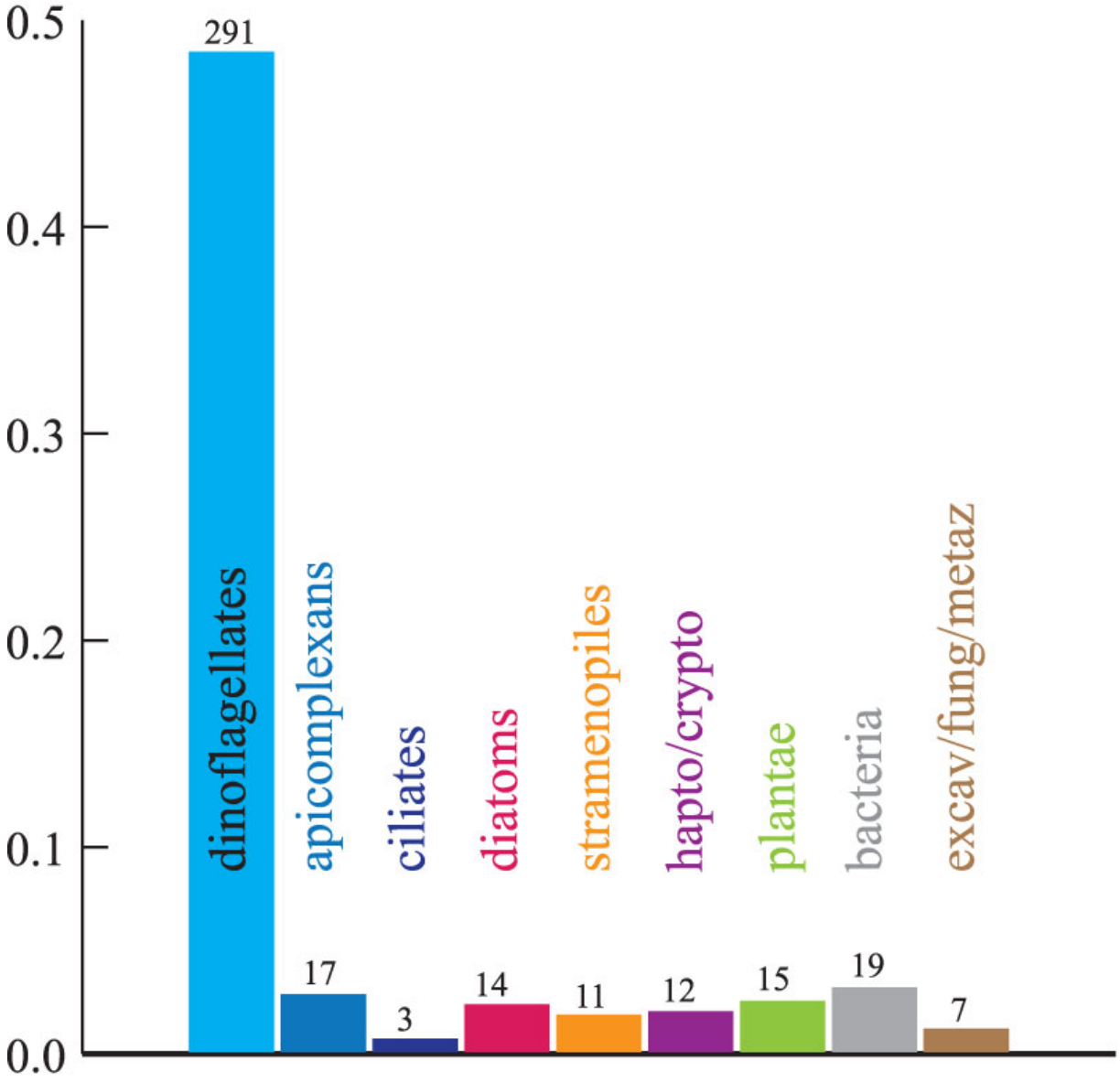

Figure 4.12: Sequences with various phylogenetic signals identified through automatic phylogenetic analyses of the SL cDNA library of *D. baltica*.

The 1856 reconstructed protein trees for the *D. baltica* sequences were automatically sorted into various groups based on the phylogenetic affinity and the bootstrap support for the *D. baltica* query protein (≥ 80%). The Y-axis shows the percentage of the phylogenies calculated based on the number of trees after clustering. The numbers on top of the bars indicate the number of phylogenetic trees (also after clustering for dinoflagellates and before clustering for all other taxa) in which *D. baltica* is grouped with that taxon.

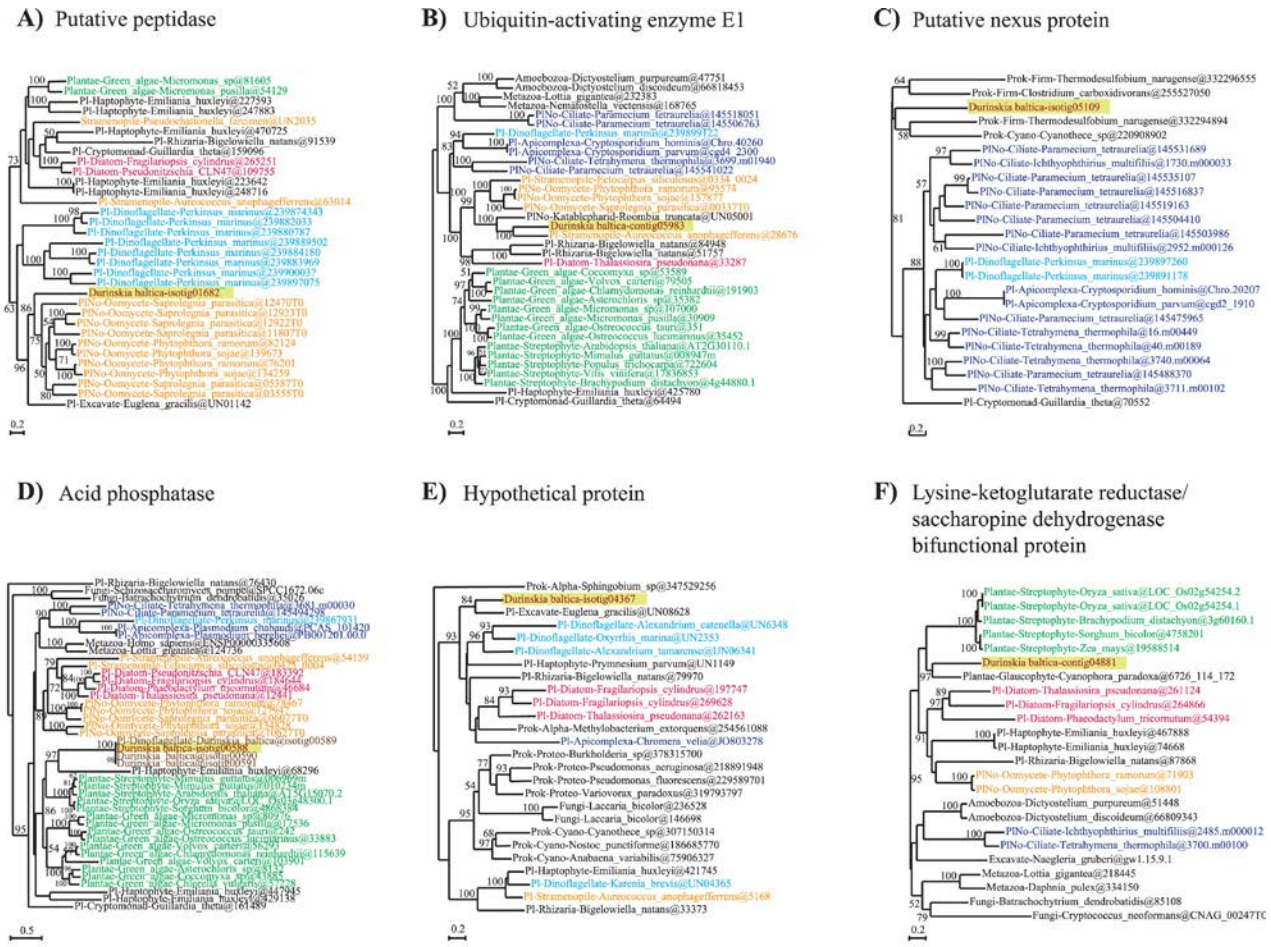

Figure 4.13: Examples of maximum likelihood trees congruent with HGT from various sources found in the SL cDNA library of *D. baltica*.

A) Putative peptidase, partial tree, B) Ubiquitin-activating enzyme E1, partial tree, C) Putative nexus protein, D) Acid phosphatase, E) Hypothetical protein, F) Lysine-ketoglutarate reductase/saccharopine dehydrogenase bifunctional protein. Numbers at the nodes indicate bootstrap support ≥ 50%. A yellow box highlights the position of *D. baltica*. Several taxa of interest are color-coded: Dinoflagellates in light blue; Apicomplexans and ciliates in dark blue; Diatoms in scarlet; Stramenopiles in orange; Green algae and plants in green; Red algae in red; All other taxa in black. . Pl, taxon with a plastid; PlNo, taxon without plastids.

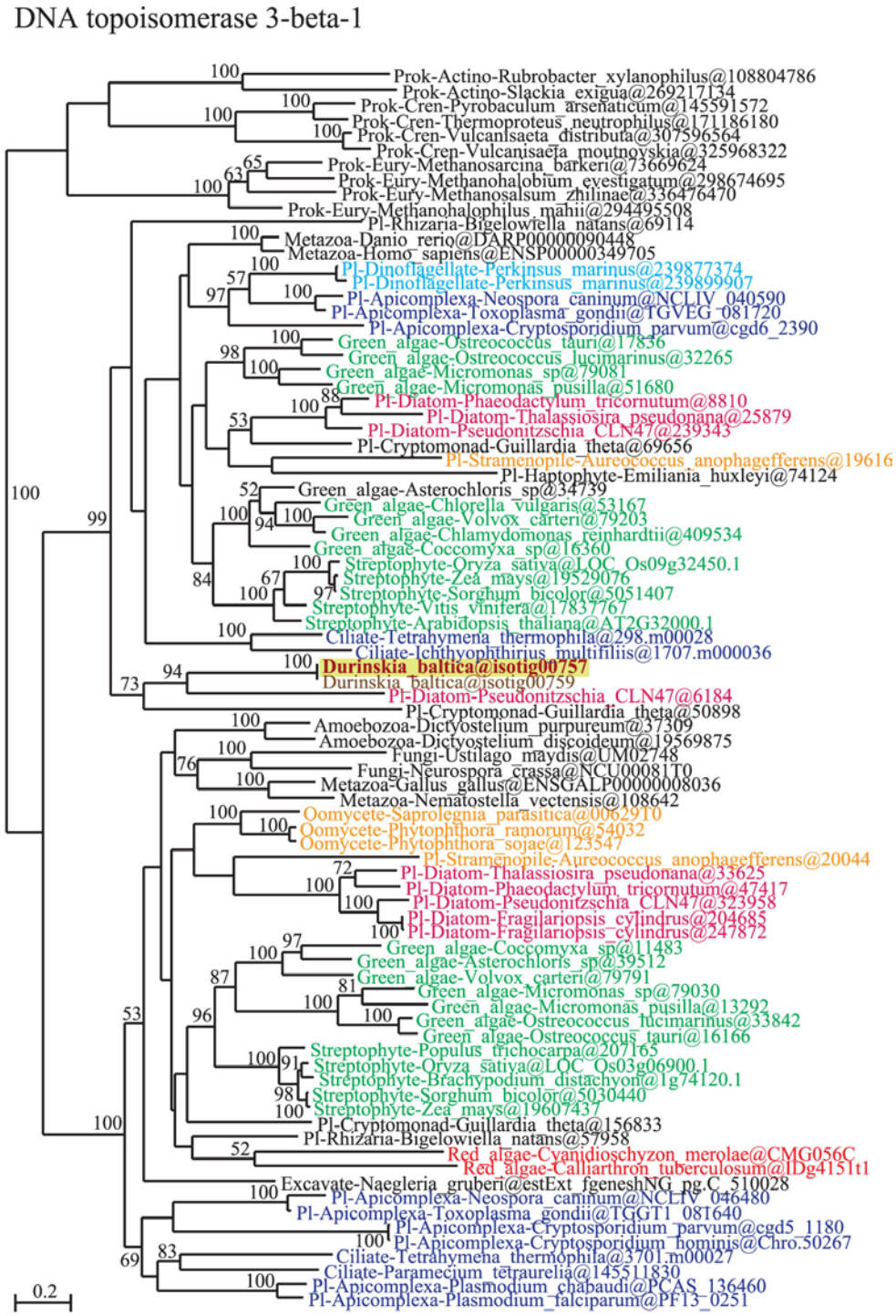

Figure 4.14: The maximum likelihood tree for DNA topoisomerase 3-beta-1 showing a diatom affinity for the *D. baltica* protein to the exclusion of alveolates.

Numbers at the nodes indicate bootstrap support ≥ 50%. A yellow box highlights the position of *D. baltica*. Several taxa of interest are color-coded: Dinoflagellates in light blue; Apicomplexans and ciliates in dark blue; Diatoms in scarlet; Stramenopiles in orange; Green algae and plants in green; Red algae in red; All other taxa in black. Pl, taxon with a plastid; PlNo, taxon without plastids.

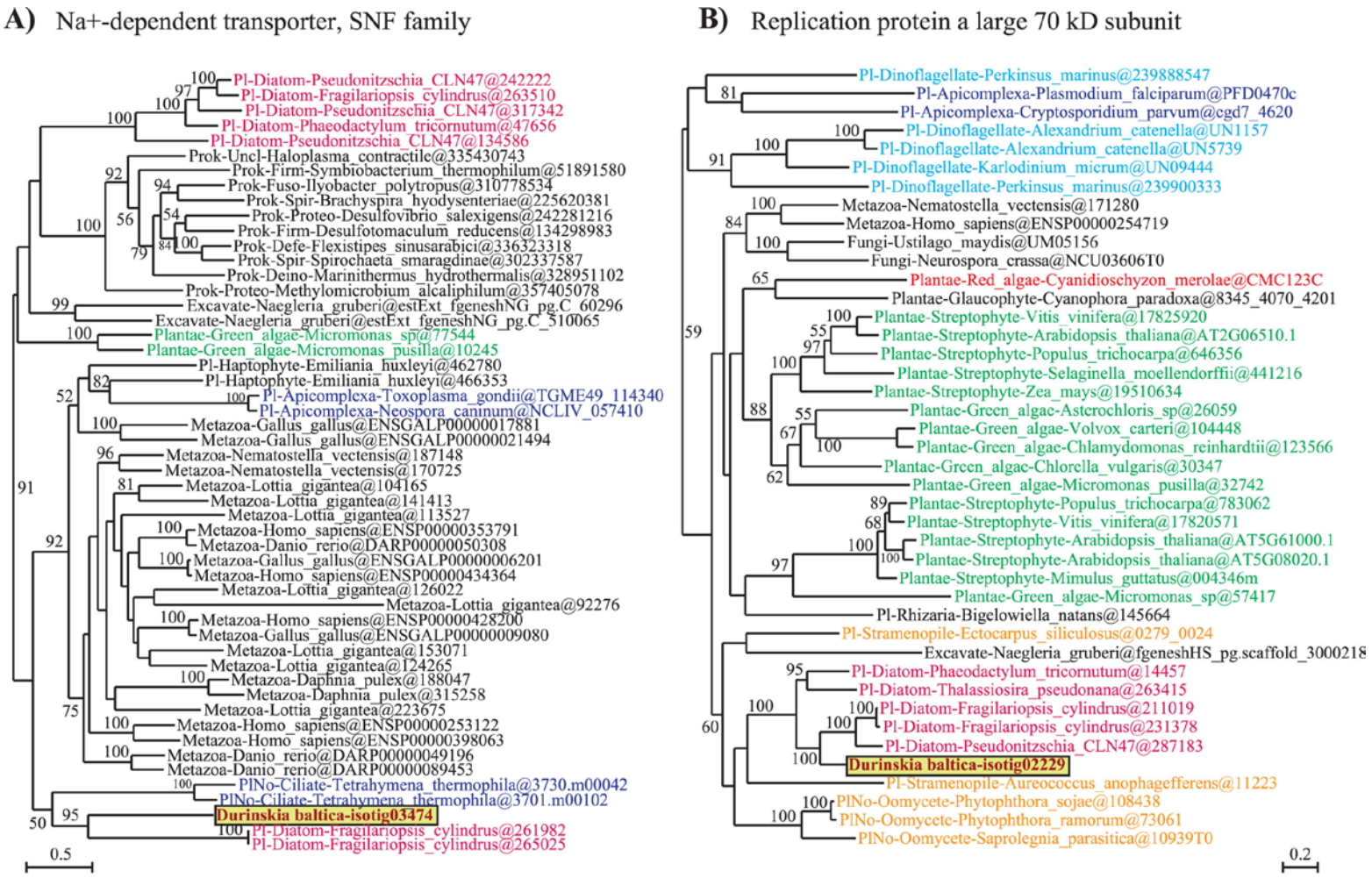

Figure 4.15: The maximum likelihood trees for the putative nuclear encoded proteins in *Durinskia baltica* congruent with a diatom affinity or origin to the exclusion of alveolates (SDTSNF and RPA1).

A) Na+-dependent transporter, SNF family, B) Replication protein a large 70 kD subunit. Numbers at the nodes indicate bootstrap support ≥ 50%. A yellow box highlights the position of *D. baltica*. Several taxa of interest are color-coded: Dinoflagellates in light blue; Apicomplexans and ciliates in dark blue; Diatoms in scarlet; Stramenopiles in orange; Green algae and plants in green; Red algae in red; All other taxa in black. Pl, taxon with a plastid; PlNo, taxon without plastids.

Table 4.1: Putative mitochondrion-targeted proteins in *Durinskia baltica*

Function	Protein (# of paralogues)	BLAST e-value	5′ end	3′ end	mTP	Phylogeny	Contig/Isotig
alternative energy metabolism	Pyruvate:Ferredoxin (flavodoxin) Oxidoreductase (PFO) (13)	0.00E+00	yes	no	yes	dino ++++	isotig00473
amino acid break down	3-hydroxyisobutyrate dehydrogenase, mitochondrial precursor (HIBADH)	1.00E-58	yes	no	yes	?	isotig03674
	Hydroxymethylglutaryl-CoA lyase, mitochondrial	2.00E-21	yes	no	yes	prok ++++	isotig04911
amino acid metabolism	cysteine desulfurase 1	1.00E-135	yes	no	yes	dino ++++	isotig01672
	2-oxoisovalerate dehydrogenase, alpha subunit	1.00E-86	yes	yes	yes	dino ++++	isotig01468
	saccharopine dehydrogenase domain-containing protein	1.00E-33	yes	no	no	rhiz ++	isotig03670
ATP synthase complex	mitochondrial ATP synthase F0 lipid binding subunit-like protein 3	1.00E-36	yes	yes	yes	dino ++++	isotig05154
	mitochondrial ATP synthase F1 delta subunit	1.00E-78	yes	yes	yes	dino ++++	isotig03014
	mitochondrial ATP synthase oligomycin sensitivity-conferring protein	2.00E-34	yes	no	yes	*Karlodinium micrum*	isotig03806
carrier protein	mitochondrial tricarboxylate transporter-like protein 2	7.00E-50	yes	no	no	dino ++++	isotig05539
	ATP-binding cassette protein 3	4.00E-107	yes	no	no	?	isotig02716
	mitochondrial carnitine/acylcarnitine carrier protein	3.00E-34	yes	yes	no	dino ++++	isotig05364
	putative tricarboxylate transport protein	6.00E-26	yes	no	no	stram +	isotig01938
cell cycle, cristae morphogenesis, functional integrity	Prohibitin	2.00E-95	yes	no	no	dino ++++	isotig01743
cytochrome-independent oxygen consumption	AOX alternative oxidase isoform A	2.00E-52	yes^	no	yes	?	isotig05153
detoxification	manganese superoxide	7.00E-36	yes	no	no	?	isotig04461

Function	Protein (# of paralogues)	BLAST e-value	5´ end	3´ end	mTP	Phylogeny	Contig/Isotig
	dismutase						
electron transport chain	electron transfer flavoprotein subunit beta	3.00E-56	no	yes	n/a	dino ++++	isotig03016
	cytochrome P450 704C1-like isoform 1	6.00E-44	yes	no	yes	prok ++++	isotig05328
	flavoprotein subunit of succinate dehydrogenase	9.00E-86	yes	no	no	dino ++++	isotig02801
	mitochondrial cytochrome c-like protein 2	6.00E-63	yes	yes	no	dino ++	isotig02327
faty acid synthesis	acetyl-CoA carboxylase	2.00E-52	yes^	no	yes	api +	isotig02914
short chain fatty acid oxidation	3-hydroxyacyl-CoA dehydrogenase (2)	0.00E+00	yes	yes	no	dino ++++	isotig00972
lipid metabolism, fatty acid beta oxidation	Medium-chain specific acyl-CoA dehydrogenase (8)	0.00E+00	yes	no	no	dino ++	isotig03561
metabolic homeostasis	protein ETHE1, mitochondrial-like	7.00E-41	yes	no	no	?	isotig03564
mitochondrial fusion?	Dynamin-like protein	2.00E-24	yes	no	no	*Salpingoeca* sp	isotig02685
nucleotide metabolic process	oligoribonuclease, mitochondrial	4.00E-10	yes	no	no	*Dictyostelium discoideum*	isotig03838
other	hypersensitive-induced response protein 1-like band7-domain (2)	5.00E-54	yes	yes	no	dino ++++	isotig01135
	hypothetical protein CAEBREN_09431	1.00E-12	yes	no	no	*Caenorhabditis brenneri*	isotig03091
protein processing, modification, transport/folding	mitochondrial processing peptidase alpha subunit (3)	7.00E-20	yes^	no	yes	dino ++++	isotig03833
	peptidase M16 domain protein (2)	2.00E-78	yes	no	no	dino ++++	isotig04357
	chaperone protein DnaJ (2)	5.00E-67	yes	no	yes	dino ++	isotig03251
	DnaJ/SEC63 protein	7.00E-127	yes	no	no	dino +++	isotig04081
	DnaJ heat shock protein HSP40 homolog	8.00E-26	yes	yes	yes	*Perkinsus marinus*	contig00705
TCA cycle	2-oxoglutarate dehydrogenase E1 component	1.00E-13	yes	no	yes	*Phytophthora infestans*	isotig01713
	mitochondrial malate dehydrogenase (NAD)-like protein 1 (3)	6.00E-112	no	no	n/a	?	isotig04548
	mitochondrial succinate	5.00E-163	yes	yes	yes	dino +++	isotig04617

Function	Protein (# of paralogues)	BLAST e-value	5´ end	3´ end	mTP	Phylogeny	Contig/Isotig
	dehydrogenase iron-sulphur subunit-like protein 2						
	mitochondrial succinyl-CoA synthetase alpha subunit	2.00E-83	yes	no	yes	?	isotig01842
	succinate dehydrogenase flavocytochrome c (13)	0.00E+00	yes	yes	no	dino ++++	isotig03036
TCA cycle/electron transport	dihydrolipoamide dehydrogenase Dld1	6.00E-29	yes	no	yes	*Schizosaccharomyces pombe*	isotig04842
transcription	Mitochondrial transcription termination factor family protein (3)	9.00E-95	yes	no	no	dino +++	isotig01563
mRNA editing	pentatricopeptide repeat-containing protein	2.00E-51	no	no	n/a	*Perkinsus marinus*	isotig04561
translation	elongation factor Tu	7.00E-91	yes^	no	yes	api ++	contig6911

The number in the parenthesis indicates the number of paralogues for the protein including the protein itself. The BLAST e-values are those of the best BLAST hits against the NCBI non-redundant (nr) protein sequence database. The 5´ end and presence (^) or absence (#) of the splice leader (SL) has been also confirmed with 5´ RACE. The presence of a mitochondrial transit peptide (mTP) is predicted using the algorithm TargetP. The phylogeny indicates whether the position of *D. baltica* in the protein maximum likelihood phylogenies is resolved or not (?), the phylogenetic affinity of the *D. baltica* protein (api, Apicomplexans; dino, Dinoflagellates; prok, Prokaryotes; rhiz, Rhizarians; stram, Stramenopiles), the level of bootstrap support (+, 50-59%; ++, 60-79%; +++, 80-89%; ++++, 90-100%). For the sequences with fewer than 5 hits no phylogenetic tree reconstruction was attempted. In those cases the best BLAST hit is reported.

Table 4.2: The putative mTPs of the host mitochondrion-targeted proteins in *Durinskia baltica*

Protein	Amino acid composition and secondary structure	mTP length	TargetP	Seq Id
PFO*		33	mTP	isotig00473
HIBADH*		22	mTP	isotig03674
HMGCL*		88	mTP	isotig04911
NFS1*		43	mTP	isotig01672
BCKDHA*		55	mTP	isotig01468
SACDH		-	cyto	isotig03670
ATPC*		34	mTP	isotig05154
ATPδ*		16	mTP	isotig03014
ATPOSCP*		30	mTP	isotig03806
PHB		-	cyto	isotig01743
AOXA		81	mTP	isotig05153
MnSOD*		-	sec	isotig04461
CYP704I1*		86	mTP	isotig05328
SDHA		-	cyto	isotig02801
CYT-C2		-	cyto	isotig02327
MPPA*		21	mTP	isotig03833
ACADM*		-	cyto	isotig03561
ETHE1*		-	cyto	isotig03564
DRP*		-	cyto	isotig02685
ORN*		-	cyto	isotig03838
HIR1		-	cyto	isotig01135
CAEBREN*		-	cyto	isotig03091
PM16		-	cyto	isotig04357
DnaJ*		35	mTP	isotig03251
DnaJ/SEC63*		-	cyto	isotig04081
DnaJ/HSP40H*		27	mTP	contig00705
HCD		-	cyto	isotig00972
OGDHE1*		82	mTP	isotig01713
SDHB2		20	mTP	isotig04617
SUCD*		20	mTP	isotig01842
SDH		-	cyto	isotig03036
DLD1*		25	mTP	isotig04842
MTERF		-	cyto	isotig01563
EFTU*		117	mTP	contig6911
ACoAC		87	mTP	isotig02914

The N-terminal sequences of 35 putative mitochondrion-targeted proteins in *Durinskia baltica*. Only the first 30 amino acids are shown if the length of transit peptide (mTP) is predicted to be larger than 30. The amino acids are color-coded based on their chemical properties in a similar manner to Danne and Waller (2011): red, polar basic (H, K, R); blue, polar acidic (D, E); green, hydroxylated (S, T); grey, polar uncharged (C, N, Q, W, Y); yellow, hydrophobic (A, F, G, I, L, M, P, V). The crosses indicate the amino acids predicted to form amphipathic α-helix

secondary structure. TargetP predictions for mTP length and localization are also shown (mTP, mitochondrion; cyto, cytoplasm; sec, secretory). A star (*) marks the proteins with 5´ extended sequences found in its alignment with its orthologs from other eukaryotes and/or prokaryotes. Abbreviation for protein names: ACADM, Medium-chain specific acyl-CoA dehydrogenase; ACoAC, acetyl-CoA carboxylase; AOXA, alternative oxidase isoform A; ATPC, ATP synthase F0 lipid binding subunit-like protein 3; ATPδ, ATP synthase F1 delta subunit; ATPOSCP, ATP synthase oligomycin sensitivity-conferring protein; BCKDHA, 2-oxoisovalerate dehydrogenase alpha subunit; CAEBREN, hypothetical protein CAEBREN_09431; CYP704I1, cytochrome P450 704C1-like isoform 1; CYT-C2, cytochrome c-like protein 2; DLD1, dihydrolipoamide dehydrogenase; DnaJ, chaperone protein DnaJ; DnaJ/HSP40H, DnaJ heat shock protein HSP40 homolog; DnaJ/SEC63, DnaJ/SEC63 protein; DRP, Dynamin-like protein; EFTU, elongation factor Tu; ETHE1, protein ETHE1; HCD, 3-hydroxyacyl-CoA dehydrogenase; HIBADH, 3-hydroxyisobutyrate dehydrogenase, mitochondrial precursor; HIR1, hypersensitive-induced response protein 1-like band7-domain; HMGCL, Hydroxymethylglutaryl-CoA lyase; MnSOD, manganese superoxide dismutase; MPPA, mitochondrial processing peptidase alpha subunit; MTERF, mitochondrial transcription termination factor family protein; NFS1, cysteine desulfurase 1; OGDHE1, 2-oxoglutarate dehydrogenase E1 component; ORN, oligoribonuclease; PFO, Pyruvate:Ferredoxin (flavodoxin) Oxidoreductase; PHB, prohibitin; PM16, Peptidase M16 domain protein; SACDH, saccharopine dehydrogenase domain-containing protein; SDH, succinate dehydrogenase; SDHA, flavoprotein subunit of succinate dehydrogenase; SDHB2, succinate dehydrogenase iron-sulphur subunit-like protein 2; SUCD, succinyl-CoA synthetase alpha subunit.

Table 4.3: Putative plastid-targeted proteins in *Durinskia baltica*

Function	Protein (paralogue)	e-value	5′ end	3′ end	SL	SP	cTP	Phylogeny	Seq ID
photosynthesis	FCP	7.00E-48	yes^	yes	no	yes	no	diatom ++++	isotig02599
response to oxidative stress	APXT (2)	1.00E-104	yes^	yes	no	yes	no	diatom ++++	isotig00853
	APX	2.00E-50	yes^	no	no	no	no	dino +	isotig02367
atp catabolism and transport	SufC*	3.00E-95	yes	yes	-	yes	yes	dino +	isotig04489
carbon utilization	CA	4.00E-59	yes	no	-	no	no	dino ++++	isotig03896
amino acid biosynthesis	OASL* (5)	9.00E-164	yes^	yes	yes	no	no	dino ++	isotig03490
maintenance and cell growth	ADK-UBOX Fusion* (3)	3.00E-51	yes^	yes	yes	no	yes	1: ? 2: kata ++++	isotig01071
ion transport	CASTOR* (6)	2.00E-58	yes^	no	no	no	no	?	isotig01020

The number in the parenthesis indicates the number of paralogues for the protein including the protein itself. A star (*) means that the protein is predicted to have an N-terminus extension based on protein alignments. The BLAST e-values are those of the best BLAST hits against the NCBI non-redundant (nr) protein sequence database. A caret (^) means that the 5′ end and presence or absence of the splice leader (SL) has been confirmed with 5′ RACE. The presence of a plastid transit peptide (cTP) is predicted using the algorithm ChloroP. The phylogeny indicates whether the position of *D. baltica* in the protein maximum likelihood phylogenies is resolved or not (?), the phylogenetic affinity of the *D. baltica* protein (crypto, Cryptophytes; diatom, Diatoms; dino, Dinoflagellates; kata, Katablepharidophytes; stram, Stramenopiles), the level of bootstrap support (+, 50-59%; ++, 60-79%; ++++, 90-100%). A dash (-) indicates that the presence or absence of the dinoflagellate splice leader (SL) was not successfully verified through 5′ RACE. Abbreviations: SP, signal peptide; stram, Stramenopiles; ADK-UBOX Fusion, chloroplast adenylate kinase and U-box domain containing protein; APX, chloroplast ascorbate peroxidase; APXT, chloroplast thylakoid bound ascorbate peroxidase; CA, chloroplast carbonic anhydrase ; CASTOR, chloroplastic ion channel CASTOR; FCP, fucoxanthin chlorophyll a/c binding protein; OASL, chloroplast O-acetyl-serine lyase.

Table 4.4: Putative diatom-derived proteins in *Durinskia baltica*

Biological process or function	Protein (paralogues)	e-value	5′ end	3′ end	SL	SP	mTP/cTP	Phylogeny	Seq ID
photosynthesis	FCP	7.00E-48	yes^	yes	no	yes	no/no	++++	isotig02599
response to oxidative stress	APXT (2)	1.00E-104	yes^	yes	no	yes	no/no	++++	isotig00853
energy production/conversion	NADHD-FAD	9.00E-10	yes^	no	no	no	no/no	++++	isotig05486
carbohydrate metabolism	PFL	0	yes^	no	no	yes	no/no	+++	isotig04324
pyruvate metabolism	DLD	100E-50	yes^	no	no	yes	no/no	+++	isotig01223
proteolysis	TLP	5.00E-69	yes^	no	no	no	no/no	+++	isotig01523
ATP catabolic process	ATPaseP	8.00E-126	yes	no	-	no	no/no	++++	isotig04442
transport: amino acid; ions	LCNL	5.00E-28	yes^	no	no	yes	no/no	++++	isotig04201
	SDTSNF	1.00E-83	yes	no	-	no	no/no	++++	isotig03474
	CHX2	3.00E-33	yes	no	-	yes	no/no	++	isotig03741
	CorAMIT (2)	4.00E-31	yes^	yes	no	no	no/no	++++	isotig05045
acyltransferase	AcylT (3)	0	yes^	yes	no	no	no/no	++++	isotig00328
DNA topological change	TOP3B (2)	1.00E-114	yes	no	-	no	no/no	+++	isotig00757
nucleotide excision repair	RPA1	5.00E-53	yes^	no	no	no	no/yes	++++	isotig02229
unknown	PrPr $^{\Psi}$	2.00E-58	yes^	no	no	yes	no/no	++++	isotig02507

The number in the parenthesis indicates the number of paralogues for the protein including the protein itself. The BLAST e-values are those of the best BLAST hits against the NCBI non-redundant (nr) protein sequence database. A caret (^) means that the 5′ end and presence or absence of the splice leader (SL) has been confirmed with 5′ RACE. The presence of a mitochondrial transit peptide (mTP) and a plastid transit peptide (cTP) are predicted using the algorithm TargetP and ChloroP, respectively. The phylogeny indicates the level of bootstrap support for the *D. baltica* grouping with a diatom or within a diatom clade (++, 60-79%; +++, 80-89%; ++++, 90-100%). A dash (-) indicates that the presence or absence of the dinoflagellate splice leader (SL) was not successfully verified through 5′ RACE. $^{\Psi}$ means a pseudogene. Abbreviations: AcylT, acyltransferase family protein; APXT, chloroplast thylakoid bound ascorbate peroxidase; ATPaseP, P-type ATPase; CHX2, monovalent Cation:Proton antiporter-2 family; CorAMIT, CorA metal ion transporter family; DLD, D-lactate dehydrogenase; FCP, fucoxanthin chlorophyll a/c binding protein; LCNL, lipocalin-like protein; NADHD-FAD, NADH dehydrogenase, FAD-containing subunit; PFL, pyruvate-formate lyase; PrPr, predicted protein; RPA1, replication protein a large 70 kD subunit; SDTSNF, Na+-dependent transporter, SNF family; TLP, trypsin-like serine protease; TOP3B, DNA topoisomerase 3-beta-1.

Chapter 5: Conclusions

Summary

Prior to the work presented here, many things about several dinotoms, especially *D. baltica* and *K. foliaceum*, were already known. Their pigments (Withers et al. 1977), their ultrastructure (Eschbach et al. 1990; Jeffrey and Vesk 1976; Tomas and Cox 1973; Tomas, Cox, et al. 1973, Tomas et al. 1973) their cell division (Tippit and Pickett-Heaps 1976), sexual life cycle (Chesnick and Cox 1987, 1989), and their endosymbiont (Chesnick et al. 1996, 1997) had been studied relatively well. However, their sequence data was scarce. In the course of this work, the plastid genomes of *D. baltica* and *K. foliaceum* were completely sequenced, the first complete tertiary plastid genomes. Shortly thereafter, their endosymbiont mitochondrial genomes were nearly completely sequenced, another first, and their host mitochondrial genomes were surveyed, producing one of the best sampled mitochondrial genomes of any dinoflagellates (for *D. baltica*) and the first sequence data from the genome of *K. foliaceum*. The pyrosequencing of the SL cDNA library in *D. baltica* added thousands of new sequences almost entirely from the host dinoflagellate to the public databases.

The results of this study indicated that the diatom endosymbiont organelle genomes in these two dinotoms have changed very little from those in their free-living cousins, showing no sign of reduction or degeneration, and this is in accordance with the small degree of morphological reduction observed in the endosymbionts. The plastid genome of *K. foliaceum* has, in fact, expanded conspicuously and undergone more reorganization compared to their counterparts in free-living diatoms mostly due to the integration, maintenance, degradation and rearrangements of the two plasmids also found in other diatoms. The host mitochondrial genome in *D. baltica* was found to have the same gene content and a very similar organization to that in

other dinoflagellates. The host mitochondrial genome in *K. foliaceum* was much more elusive and hard to sequence, but the well conserved fragments of all the mitochondrial genes found in other dinoflagellates were recovered from this dinotom, altogether implying that the host mitochondrial genomes of the two dinotoms remain mainly unchanged in spite of their co-existence with their endosymbiont counterparts. These endosymbiont counterparts showed even fewer signs of change, being nearly identical in gene content, gene order and organization to that in other diatoms especially the pennate diatom *Fragilariopsis cylindrus*.

The results of the transcriptome survey of the *D. baltica* host revealed that no EGT to the host has occurred, and despite the permanent and obligate nature of symbiosis in dinotoms, the *D. baltica* endosymbiont retains its genetic integrity and self-reliance with respect to its own organelles. The finding of the diatom-derived plastid genes with conventional bi-partite targeting signals, which are most likely encoded in the nuclear genome of the endosymbiont, (i.e. FCP and APXT) hints at a functional targeting system within the diatom endosymbiont of *D. baltica*. However, the lack of EGT to the host in *D. baltica* implies that a protein targeting system that targets the products of the transferred genes from the host to endosymbiont in dinotoms is unnecessary and most likely non-existent.

If the genetic integration and the complementary targeting system are the criteria to distinguish an endosymbiont-derived organelle from an endosymbiont (Cavalier-Smith and Lee 1985), then, strictly speaking, the dinotom endosymbionts are not or have not yet become organelles. Although strict definitions might offer convenience and clarity, they usually do not reflect the true complexity of the subject matter in real life. The transformation of a free-living cell to an organelle through endosymbiosis is not a linear progression, and it has happened independently many times. A wide variety of intermediary stages and a wide range of symbiotic

interactions have been discovered, and the binomial terminology, endosymbiont and organelle, does not describe well the nearly continuous spectrum of the endosymbionts in their transition to an organelle (Keeling and Archibald 2008). The dinotom endosymbiont is not genetically integrated with its host, but at the cellular level it is. Does this not make it an organelle as well as an endosymbiont?

Future directions

The genetic survey of the two dinotoms in this study has produced a wealth of information about their complex genome, but it is far from complete. Despite many trials, the mitochondrial genomes of the endosymbiont could not be completely sequenced. More importantly, this survey did not include the nucleus of the endosymbiont. Its survey could shed light on the extent of genetic reduction in this rare eukaryotic nucleus that divides amitotically (Tippit and Pickett-Heaps 1976). The recent advances in sequencing technology has made whole genome sequencing much quicker and cheaper than before, and the progress in the bioinformatics fronts is promising more accurate and faster assembly algorithms. Soon, it will be possible perhaps to sequence and assemble the whole genome of at least one of the dinotoms. Until such a time, a polyA EST or a direct RNA sequencing (Ozsolak et al. 2009) project from dinotoms that includes the whole transcriptome of the cell could be very informative about the gene content of the endosymbiont nuclear genome and perhaps even its extent of reduction. Alternatively, the CsCl gradient density could be applied to the total DNA extracted from a dinotom, the band enriched in the endosymbiont nuclear DNA (the middle band in the three band profile of the dinotom DNA in the gradient column) could be isolated, amplified and massively sequenced through pyrosequencing (as done in this study for the organelle-enriched DNA, the top or satellite band) or illumina sequencing.

In this study, two putative diatom-derived (FCP and APXT) and at least one putative dinoflagellate-derived (SufC) plastid-targeted proteins were identified. It would be curious to examine whether they are actually targeted to the plastid. No dinoflagellate has been successfully transformed permanently or transiently with any reporter gene, and a model system for such experimentations in dinoflagellates is lacking. Until a simpler dinoflagellate model becomes available, other organisms can be used to test for the targeting destination of these three proteins. Perhaps the best suited organism is *P. tricornutum*, a pennate diatom, which has been successfully transformed (Apt et al. 1996; Niu et al. 2012). A reporter gene such as a green fluorescent protein (GFP) gene can be used to test whether the targeting signals of the three *D. baltica* proteins are able to transport the GFP to its plastid. Transmission electron microscopy in conjunction with primary antibodies against these proteins and gold-conjugated secondary antibodies against the first ones could reveal where the proteins destined to, a technique used in investigating the targeting system for the nuclear-encoded plastid proteins in dinoflagellates (Nassoury et al. 2003).

There is a report of a strain or variety of *K. foliaceum* that lacks the endosymbiont nucleus (Kempton et al. 2002) and other dinotoms that seem to bear different centric diatom endosymbionts (Horiguchi and Pienaar 1991; Takano et al. 2008) rather than the usual pennate one. A survey of the SL cDNA library of the first and a survey of the organelle genomes from the second ones could provide invaluable data for comparison with those gathered for *D. baltica* and *K. foliaceum* in this study. However, most of the known dinotoms, including the mono-nucleate *K. foliaceum* and the ones with a centric diatom endosymbiont, have not been successfully cultured, and some of the most basic aspects of dinotom cell biology, even in *D. baltica* and *K. foliaceum*, have not been explored. Successful culturing of the dinotoms that are

not available in the culture collections could encourage further studies in the dinotom basic cell physiology, metabolism and ecology. Such studies in combination with genomic surveys such as the one presented here could enrich one another and produce invaluable insight into these truly complex and beautiful organisms.

References

Agrawal S, Striepen B. 2010. More membranes, more proteins: complex protein import mechanisms into secondary plastids. Protist. 161:672–87. doi: 10.1016/j.protis.2010.09.002.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

Apt KE, Kroth-Pancic PG, Grossman AR. 1996. Stable nuclear transformation of the diatom *Phaeodactylum tricornutum*. Molecular & general genetics : MGG. 252:572–9.

Archibald JM. 2006. Algal genomics: exploring the imprint of endosymbiosis. Curr. Biol. 16:R1033–5. doi: 10.1016/j.cub.2006.11.008.

Archibald JM. 2007. Nucleomorph genomes: structure, function, origin and evolution. BioEssays. 29:392–402.

Archibald JM. 2009. The puzzle of plastid evolution. Curr. Biol. 19:R81–8. doi: 10.1016/j.cub.2008.11.067.

Archibald JM, Keeling PJ. 2002. Recycled plastids: A “green movement” in eukaryotic evolution. Trends Genet. 18:577–584.

Archibald JM, Rogers MB, Toop M, Ishida K, Keeling PJ. 2003. Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga *Bigelowiella natans*. Proc. Natl. Acad. Sci. U. S. A. 100:7678–7683.

Armbrust EV et al. 2004. The genome of the diatom *Thalassiosira pseudonana*: Ecology, evolution, and metabolism. Science. 306:79–86.://000224304000039.

Bachvaroff TR, Concepcion GT, Rogers CR, Herman EM, Delwiche CF. 2004. Dinoflagellate expressed sequence tag data indicate massive transfer of chloroplast genes to the nuclear genome. Protist. 155:65–78.

Bachvaroff TR, Place AR. 2008. From stop to start: tandem gene arrangement, copy number and trans-splicing sites in the dinoflagellate *Amphidinium carterae*. PloS one. 3:e2929. doi: 10.1371/journal.pone.0002929.

Barre FX et al. 2001. Circles: The replication-recombination-chromosome segregation connection. Proc. Natl. Acad. Sci. U. S. A. 98:8189–8195.

Bedwell DM, Strobel S a, Yun K, Jongeward GD, Emr SD. 1989. Sequence and structural requirements of a mitochondrial protein import signal defined by saturation cassette mutagenesis. Mol. Cell. Biol. 9:1014–25.

Bendtsen JD, Nielsen H, Von Heijne G, Brunak Søren. 2004. Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol. 340:783–95. doi: 10.1016/j.jmb.2004.05.028.

Bhattacharya D, Yoon HS, Hackett JD. 2004. Photosynthetic eukaryotes unite: Endosymbiosis connects the dots. BioEssays. 26:50–60.

Blakely GW, Sherratt DJ. 1994. Interactions of the Site-Specific Recombinases Xerc and Xerd with the Recombination Site Dif. Nucleic Acids Res. 22:5613–5620.

Bowler C et al. 2008. The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. Nature. 456:239–44. doi: 10.1038/nature07410.

Bruce BD. 2001. The paradox of plastid transit peptides: Conservation of function despite divergence in primary structure. Biochim. Biophys. Acta. 1541:2–21.

Bui ETN, Bradley PJ, Johnson PJ. 1996. A common evolutionary origin for mitochondria and hydrogenosomes. Proc. Natl. Acad. Sci. U. S. A. 93:9651–9656.

Burke JM et al. 1987. Structural convention for group I introns. Nucleic Acids Res. 15:7217–7221. doi: 10.1093/nar/gkn942.

Burki F et al. 2012. Re-evaluating the green versus red signal in eukaryotes with secondary plastid of red algal origin. Genome Biol. Evol. 4:evs049. doi: 10.1093/gbe/evs049.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics (Oxford, England). 25:1972–3. doi: 10.1093/bioinformatics/btp348.

Carty S, Cox ER. 1986. *Kansodinium* Gen-Nov and *Durinskia* Gen-Nov - 2 Genera of Fresh-Water Dinoflagellates (Pyrrhophyta). Phycologia. 25:197–204.

Cattolico RA et al. 2008. Chloroplast genome sequencing analysis of *Heterosigma akashiwo* CCMP452 (West Atlantic) and NIES293 (West Pacific) strains. BMC genomics. 9:211. doi: 10.1186/1471-2164-9-211.

Cavalier-Smith T. 1983. A 6-kingdom classification and a unified phylogeny. In: Endocytobiology II: Intracellular Space as Oligogenetic. Schenk, HEA et al., editor. Walter de Gruyter & Co: Berlin pp. 1027–1034.

Cavalier-Smith T. 1993. Kingdom protozoa and its 18 phyla. Microbiol. Rev. 57:953–94.

Cavalier-Smith T. 2009. Predation and eukaryote cell origins: a coevolutionary perspective. Int. J. Biochem. Cell Biol. 41:307–22. doi: 10.1016/j.biocel.2008.10.002.

Cavalier-Smith T. 1999. Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. J. Euk. Microbiol. 46:347–66.

Cavalier-Smith T, Lee JJ. 1985. Protozoa as hosts for endosymbioses and the conversion of symbionts into organelles. J. Euk. Microbiol. 32:376–379. doi: 10.1111/j.1550-7408.1985.tb04031.x.

Cavalier-Smith T. 1982. The origins of plastids. Biol. J. Linn. Soc. 17:289–306.

Chan CX, Yang EC, et al. 2011. Red and green algal monophyly and extensive gene sharing found in a rich repertoire of red algal genes. Curr. Biol. 21:328–33. doi: 10.1016/j.cub.2011.01.037.

Chan CX, Reyes-Prieto A, Bhattacharya D. 2011. Red and green algal origin of diatom membrane transporters: insights into environmental adaptation and cell evolution. PloS one. 6:e29138. doi: 10.1371/journal.pone.0029138.

Chesnick JM, Cox E. 1989. Fertilization and zygote development in the binucleate dinoflagellate *Peridinium balticum* (Pyrrhophyta). Am. J. Bot.. 76:1060–1072.

Chesnick JM, Cox E. 1987. Synchronized sexuality of an algal symbiont and its dinoflagellate host, *Peridinium balticum* (Levander) Lemmermann. Biosystems. 21:69–78.

Chesnick JM, Kooistra WHC, Wellbrock U, Medlin LK. 1997. Ribosomal RNA analysis indicates a benthic pennate diatom ancestry for the endosymbionts of the dinoflagellates *Peridinium foliaceum* and *Peridinium balticum* (Pyrrhophyta). J. Euk. Microbiol. 44:314–20.

Chesnick JM, Morden CW, Schmieg AM. 1996. Identity of the endosymbiont of *Peridinium foliaceum* (Pyrrophyta): Analysis of the rbcLS operon. J. Phycol. 32:850–857.

Cline K, Dabney-Smith C. 2008. Plastid protein import and sorting: different paths to the same compartments. Curr. Opin. Plant Biol. 11:585–92. doi: 10.1016/j.pbi.2008.10.008.

Conant G, Wolfe K. 2008. GenomeVx: simple web-based creation of editable circular chromosome maps. Bioinformatics. 24:861–862.

Cox E, Rizzo PJ. 1976. Observations on cell division in a bi nucleate dinoflagellate. J. Phycol. 12:21.

Danne JC, Gornik SG, Waller RF. 2011. An Assessment of vertical inheritance versus endosymbiont transfer of nucleus-encoded genes for mitochondrial proteins following tertiary endosymbiosis in *Karlodinium micrum*. Protist. 163:76–90. doi: 10.1016/j.protis.2011.03.002.

Danne JC, Waller RF. 2011. Analysis of dinoflagellate mitochondrial protein sorting signals indicates a highly stable protein targeting system across eukaryotic diversity. J. Mol. Biol. 408:643–53. doi: 10.1016/j.jmb.2011.02.057.

Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve : Multiple Alignment of Conserved Genomic Sequence With Rearrangements. Genome Res. 1394–1403. doi: 10.1101/gr.2289704.

Deane JA et al. 2000. Evidence for nucleomorph to host nucleus gene transfer: Light-harvesting complex proteins from cryptomonads and chlorarachniophytes. Protist. 151:239–252.

Demadariaga I, Orive E, Boalch GT. 1989. Primary Production in the Gernika Estuary During a Summer Bloom of the Dinoflagellate *Peridinium quinquecorne* Abe. Botanica Marina. 32:159–165.

DeRocher a, Hagen CB, Froehlich JE, Feagin JE, Parsons M. 2000. Analysis of targeting sequences demonstrates that trafficking to the *Toxoplasma gondii* plastid branches off the secretory system. J. Cell Sci. 113 (Pt) 2:3969–77.

Deschamps P, Moreira D. 2012. Re-evaluating the green contribution to diatom genomes. Genome Biol. Evol. 4:683–688. doi: 10.1093/gbe/evs053.

Dodge JD. 1971. A dinoflagellate with both a mesokaryotic and a eukaryotic nucleus: Part 1 fine structure of the nuclei. Protoplasma. 73:145–157.

Dodge JD. 1989. Phylogenetic relationships of dinoflagellates and their plastids. In: The chromophyte algae: problems and perspectives. Green, JC, Leadbeater, BSC & Diver, WL, editors. Clarendon Press: Oxford, England pp. 207–227.

Dolezal P, Likic V, Tachezy J, Lithgow T. 2006. Evolution of the molecular machines for protein import into mitochondria. Science (New York, N.Y.). 313:314–8. doi: 10.1126/science.1127895.

Douglas S et al. 2001. The highly reduced genome of an enslaved algal nucleus. Nature. 410:1091–1096.

Duby G, Oufattole M, Boutry M. 2001. Hydrophobic residues within the predicted N-terminal amphiphilic alpha-helix of a plant mitochondrial targeting presequence play a major role in in vivo import. Plant J.. 27:539–49.

Durnford Dion G, Gray MW. 2006. Analysis of *Euglena gracilis* plastid-targeted proteins reveals different classes of transit sequences. Eukaryotic cell. 5:2079–91. doi: 10.1128/EC.00222-06.

Ehara M, Inagaki Y, Watanabe KI, Ohama T. 2000. Phylogenetic analysis of diatom coxI genes and implications of a fluctuating GC content on mitochondrial genetic code evolution. Curr. Genet. 37:29–33.

Emanuelsson O, Nielsen H, Brunak S, Von Heijne G. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J. Mol. Biol. 300:1005–16. doi: 10.1006/jmbi.2000.3903.

Emanuelsson O, Nielsen H, Von Heijne G. 1999. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein science : a publication of the Protein Society. 8:978–84. doi: 10.1110/ps.8.5.978.

Eschbach S, Speth V, Hansmann P, Sitte P. 1990. Freeze-fracture study of the single membrane between host cell and endocytobiont in the dinoflagellates *Glenodinium foliaceum* and *Peridinium balticum*. J. Phycol. 26:324–328.

Esposito D, Scocca JJ. 1997. The integrase family of tyrosine recombinases: evolution of a conserved active site domain. Nucleic Acids Res. 25:3605–3614.

Falciatore A, Bowler C. 2002. Revealing the molecular secrets of marine diatoms. Ann. Rev. Plant Biol. 53:109–30. doi: 10.1146/annurev.arplant.53.091701.153921.

Feagin JE, Mericle BL, Werner E, Morris M. 1997. Identification of additional rRNA fragments encoded by the *Plasmodium falciparum* 6 kb element. Nucleic Acids Res. 25:438–46.

Felsner G et al. 2011. ERAD components in organisms with complex red plastids suggest recruitment of a preexisting protein transport pathway for the periplastid membrane. Genome Biol. Evol. 3:140–50. doi: 10.1093/gbe/evq074.

Field CB, Behrenfeld MJ, Randerson JT, Falkowski P. 1998. Primary production of the biosphere: Integrating terrestrial and oceanic components. Science. 281:237–240. doi: 10.1126/science.281.5374.237.

Figueroa RI et al. 2009. The life history and cell cycle of *Kryptoperidinium foliaceum*, a dinoflagellate with two eukaryotic nuclei. Protist. 160:285–300. doi: 10.1016/j.protis.2008.12.003.

Franzén LG, Rochaix JD, Von Heijne G. 1990. Chloroplast transit peptides from the green alga Chlamydomonas reinhardtii share features with both mitochondrial and higher plant chloroplast presequences. FEBS letters. 260:165–8.

Friesen H, Sadowski PD. 1992. Mutagenesis of a conserved region of the gene encoding the Flp recombinase of *Saccharomyces cerevisiae* - a role for arginine-191 in binding and ligation. J. Mol. Biol. 225:313–326.

Gabrielsen TM et al. 2011. Genome evolution of a tertiary dinoflagellate plastid. PloS one. 6:e19132. doi: 10.1371/journal.pone.0019132.

Garate-Lizarraga I, Muneton-Gomez MD. 2008. Bloom of *Peridinium quinquecorne* Abe, in La Ensenada de La Paz, Gulf of California (July 2003). Acta Botanica Mexicana. 83:33–47.

Garcia-Cuetos L, Moestrup Ø, Hansen PJ, Daugbjerg N. 2010. The toxic dinoflagellate *Dinophysis acuminata* harbors permanent chloroplasts of cryptomonad origin, not kleptochloroplasts. Harmful Algae. 9:25–38. doi: 10.1016/j.hal.2009.07.002.

Gast RJ, Moran DM, Dennett MR, Caron D a. 2007. Kleptoplasty in an Antarctic dinoflagellate: caught in evolutionary transition? Environ. Microbiol. 9:39–45. doi: 10.1111/j.1462-2920.2006.01109.x.

Gilson P et al. 2006. Complete nucleotide sequence of the chlorarachniophyte nucleomorph: Nature's smallest nucleus. Proc. Natl. Acad. Sci. 103:9566–9571.

Gilson P, McFadden GI. 2002. Jam packed genomes: A preliminary, comparative analysis of nucleomorphs. Genetica (Dordrecht). 115:13–28.

Gordon D. 2004. Viewing and editing assembled sequences using Consed. In: Current Protocols in Bioinformatics. Baxevanis, A & Davidson, D, editors. John Wiley & Co: New York pp. 11.12.11–11.12.43.

Gordon D, Abajian C, Green P. 1998. Consed: a graphical tool for sequence finishing. Genome Res. 8:195–202.

Gordon D, Desmarais C, Green P. 2001. Automated finishing with autofinish. Genome Res. 11:614–25. doi: 10.1101/gr.171401.

Gould SB, Waller RF, Mcfadden GI. 2008. Plastid evolution. Ann. Rev. Plant Biol. 59:491–517. doi: 10.1146/annurev.arplant.59.032607.092915.

Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. 27:221–4. doi: 10.1093/molbev/msp259.

Gray MW, Burger G, Lang B. 2001. The origin and early evolution of mitochondria. Genome Biol. 2: Reviews 1018.

Gray MW, Burger G, Lang BF. 1999. Mitochondrial evolution. Science (Washington D C). 283:1476–1481.

Gray MW, Lang BFF, Burger G. 2004. Mitochondria of protists. Annu. Rev. Genet. 38:477–524. doi: 10.1146/annurev.genet.37.110801.142526.

Greenwood AD. 1974. The Cryptophyta in relation to phylogeny and photosynthesis. In: 8th International Congress of Electron Microscopy. Sanders, J & Goodchild, D, editors. Australian Academy of Sciences: Canberra pp. 566–567.

Gutensohn M et al. 2006. Toc, Tic, Tat et al. Structure and function of protein transport machineries in chloroplasts. J. plant Physiol. 163:333–47. doi: 10.1016/j.jplph.2005.11.009.

Habib SJ, Neupert W, Rapaport D. 2007. Analysis and prediction of mitochondrial targeting signals. Methods Cell Biol. 80:761–81. doi: 10.1016/S0091-679X(06)80035-X.

Hackett J, Anderson D, Erdner D, Bhattacharya DB. 2004. Dinoflagellates: a remarkable evolutionary experiment. Am. J. Bot.. 91:1523–1534.

Hackett JD, Maranda L, Yoon Hwan Su, Bhattacharya D. 2003. Phylogenetic evidence for the cryptophyte origin of the plastid of *Dinophysis* (Dinophysiales, Dinophyceae). J. Phycol. 39:440–448.

Hallegraeff GM, Lucas I A N. 1988. The marine dinoflagellate genus *Dinophysis* (Dinophyceae)--Photosynthetic, neritic and non-photosynthetic, oceanic species. Phycologia. 27:25–42.

Hammen PK, Weiner H. 1998. Mitochondrial leader sequences: structural similarities and sequence differences. J. Exp. Zool. 282:280–3.

Han YPW, Gumport RI, Gardner JF. 1993. Complementation of Bacteriophage-Lambda Integrase Mutants - Evidence for an intersubunit active-site. Embo Journal. 12:4577–4584.

Hewes CD, Mitchell BG, Moisan T a., Vernet M, Reid FMH. 1998. the Phycobilin signatures of chloroplasts from three dinoflagellate species: a microanalytical study of *Dinophysis caudata*, D. Fortii, and *D. Acuminata* (Dinophysiales, Dinophyceae). J. Phycol. 34:945–951. doi: 10.1046/j.1529-8817.1998.340945.x.

Hildebrand M et al. 1992. Nucleotide-sequence of diatom plasmids - Identification of open reading frames with similarity to site-specific recombinases. Plant Mol. Biol. 19:759–770.

Hildebrand M et al. 1991. Plasmids in diatom species. Journal of Bacteriology. 173:5924–5927.

Hirakawa Y, Nagamune K, Ishida K. 2009. Protein targeting into secondary plastids of chlorarachniophytes. Proc. Natl. Acad. Sci. U. S. A. 106:12820–5. doi: 10.1073/pnas.0902578106.

Horiguchi T, Pienaar R. 1991. Ultrastructure of a marine dinoflagellate, *Peridinium quinquecorne* Abe (Peridiniales) from South Africa with special reference to its chrysophyte endosymbiont. Botanica Marina. 34:123–131.

Horiguchi T, Pienaar R. 1994. Ultrastructure of a new marine sand-dwelling dinoflagellate, *Gymnodinium quadrilobatum* sp. nov. (Dinophyceae) with special reference to its endosymbiotic alga. Eur. J. Phycol. 29:237–245. doi: 10.1080/09670269400650691.

Horiguchi T. 2004. Origin and evolution of dinoflagellates with a diatom endosymbiont. Neo-Science of Natural History: Integration of Geoscience and Biodiversity StudiesNatural History. 5:53–59.

Horiguchi T, Takano Y. 2006a. Serial replacement of a diatom endosymbiont in the marine dinoflagellate *Peridinium quinquecorne* (Peridiniales, Dinophyceae). Phycological Res. 54:193–200. doi: 10.1111/j.1440-1835.2006.00426.x.

Imanian B, Carpenter KJ, Keeling PJ. 2007. The mitochondrial genome of a tertiary endosymbiont retains genes for electron transport proteins. J. Euk. Microbiol. 54:146–153.

Imanian B, Keeling PJ. 2007. The dinoflagellates *Durinskia baltica* and *Kryptoperidinium foliaceum* retain functionally overlapping mitochondria from two evolutionarily distinct lineages. BMC Evol. Biol. 7:172. doi: 10.1186/1471-2148-7-172.

Imanian B, Pombert J-F, Dorrell RG, Burki F, Keeling PJ. 2012. Tertiary endosymbiosis in two dinotoms has generated little change in the mitochondrial genomes of their dinoflagellate hosts and diatom endosymbionts. PLoS ONE. 7:e43763. doi: 10.1371/journal.pone.0043763.

Imanian B, Pombert J-F, Keeling PJ. 2010. The complete plastid genomes of the two “dinotoms” *Durinskia baltica* and *Kryptoperidinium foliaceum*. PloS one. 5:e10711. doi: 10.1371/journal.pone.0010711.

Ishida K, Green BR. 2002. Second- and third-hand chloroplasts in dinoflagellates: Phylogeny of oxygen-evolving enhancer 1 (PsbO) protein reveals replacement of a nuclear-encoded plastid gene by that of a haptophyte tertiary endosymbiont. Proc. Natl. Acad. Sci. U. S. A. 99:9294–9299.

Jackson CJ, Gornik S G, Waller RF. 2012. The mitochondrial genome and transcriptome of the basal dinoflagellate *Hematodinium* sp.: Character evolution within the highly derived mitochondrial genomes of dinoflagellates. Genome Biol. Evol. 4:59–72. doi: 10.1093/gbe/evr122.

Jackson CJ et al. 2007. Broad genomic and transcriptional analysis reveals a highly derived genome in dinoflagellate mitochondria. BMC Biol. 5:41. doi: 10.1186/1741-7007-5-41.

Jacobs JD et al. 1992. Characterization of 2 circular plasmids from the marine diatom *Cylindrotheca fusiformis* - Plasmids hybridize to chloroplast and nuclear-DNA. Mol. Gen. Genet. 233:302–310.

Janouskovec J, Horák A, Oborník M, Lukes J, Keeling PJ. 2010. A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. Proc. Natl. Acad. Sci. U. S. A. 107:10949–54. doi: 10.1073/pnas.1003335107.

Jarvis P, Soll J. 2002. Toc, Tic, and chloroplast protein import (vol 1541, pg 64, 2001). Biochim. Biophys. Acta-Molecular Cell Research. 1590:177–189.

Jeffrey SW, Sielicki M, Haxo F T. 1975. Chloroplast pigment patterns in dinoflagellates. J. Phycol. 11:374–384.

Jeffrey SW, Vesk M. 1976. Further evidence for a membrane bound endosymbiont within the dinoflagellate *Peridinium foliaceum*. J. Phycol. 12:450–455.

Jiroutová K, Horák A, Bowler C, Oborník M. 2007. Tryptophan biosynthesis in stramenopiles: eukaryotic winners in the diatom complex chloroplast. J. Mol. Evol. 65:496–511. doi: 10.1007/s00239-007-9022-z.

Kamikawa R et al. 2009. Mitochondrial group II introns in the raphidophycean flagellate *Chattonella* spp. suggest a diatom-to-*Chattonella* lateral group II intron transfer. Protist. 160:364–75. doi: 10.1016/j.protis.2009.02.003.

Kaneko T et al. 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res. 3:109–116,185–209.

Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. Briefings in bioinformatics. 9:286–98. doi: 10.1093/bib/bbn013.

Keeling PJ. 2008. Bridge over troublesome plastids. Nature. 451.

Keeling PJ, Archibald JM. 2008. Organelle evolution: what's in a name? Curr. Biol. 18:R345–7. doi: 10.1016/j.cub.2008.02.059.

Keeling PJ. 2009. Chromalveolates and the evolution of plastids by secondary endosymbiosis. J. Euk. Microbiol. 56:1–8. doi: 10.1111/j.1550-7408.2008.00371.x.

Keeling PJ, Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. Nat. Rev. Genet. 9:605–618. doi: 10.1038/nrg2386.

Keeling PJ. 2010. The endosymbiotic origin, diversification and fate of plastids. Philosophical transactions of the Royal Society of London. Series B, Biological sciences. 365:729–48. doi: 10.1098/rstb.2009.0103.

Kempton JW et al. 2002. *Kryptoperidinium foliaceum* blooms in South Carolina: A multi-analytical approach to identification. Harmful Algae. 1:383–392.

Kim E, Archibald JM. 2009. Diversity and evolution of plastids and their genomes. Mol. Biol. doi: 10.1007/7089.

Kite GC, Rothschild LJ, Dodge JD. 1988. Nuclear and plastid DNAs from the binucleate dinoflagellates *Glenodinium* (*Peridinium*) *foliaceum* and *Peridinium balticum*. Biosystems. 21:151–163.

Kite GC, Dodge J. 1985. Structural organization of plastid DNA in two anomalously pigmented dinoflagellates. J. Phycol. 21:50–56.

Kovács-Bogdán E, Soll Jürgen, Bölter B. 2010. Protein import into chloroplasts: the Tic complex and its regulation. Biochim. Biophys. acta. 1803:740–7. doi: 10.1016/j.bbamcr.2010.01.015.

Krogh A, Larsson B, Von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. 305:567–80. doi: 10.1006/jmbi.2000.4315.

Kurtz S et al. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 29:4633–4642.

Lane CE et al. 2007. Nucleomorph genome of *Hemiselmis andersenii* reveals complete intron loss and compaction as a driver of protein structure and function. Proc. Natl. Acad. Sci. U. S. A. 104:19908–19913.

Lane CE et al. 2005. Insight into the Diversity and evolution of the cryptomonad nucleomorph Genome. Mol. Biol. doi: 10.1093/molbev/msj066.

Lang M, Apt KE, Kroth PG. 1998. Protein transport into “complex” diatom plastids utilizes two different targeting signals. J. Biol. Chem. 273:30973–8.

Lesterlin C, Barre FX, Cornet F. 2004. Genetic recombination and the cell cycle: what we have learned from chromosome dimers. Mol. Microbiol. 54:1151–1160.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics (Oxford, England). 22:1658–9. doi: 10.1093/bioinformatics/btl158.

Lin S. 2006. The smallest dinoflagellate genome is yet to be found: a Comment on Lajeunesse et al. “*Symbiodinium* (Pyrrhophyta) genome sizes (DNA content) are smallest among dinoflagellates”1. J. Phycol. 42:746–748. doi: 10.1111/j.1529-8817.2006.00213.x.

Lin S, Zhang H, Spencer DF, Norman JE, Gray MW. 2002. Widespread and extensive editing of mitochondrial mRNAS in Dinoflagellates. J. Mol. Biol. 320:727–739.

Lucas IA. N, Maret V. 1990. The fine structure of two photosynthetic species of *Dinophysis* (Dinophysiales, Dinophyceae). J. Phycol. 26:345–357.

Lukes J, Leander BS, Keeling PJ. 2009. Cascades of convergent evolution: the corresponding evolutionary histories of euglenozoans and dinoflagellates. Proc. Natl. Acad. Sci. U. S. A. 106 Suppl :9963–70. doi: 10.1073/pnas.0901004106.

Mandelli EF. 1968. Carotenoid pigments of dinoflagellate *Glenodinium foliaceum* Stein. J. Phycol. 4:347–348.

Mann D G, Droop SJM. 1996. 3 . Biodiversity , biogeography and conservation of diatoms. Hydrobiologia. 19–32.

Mann David G. 1999. The species concept in diatoms. Phycologia. 38:437–495.

Marchler-Bauer A et al. 2009. CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res. 37:D205–D210.

Martin W et al. 2002. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc. Natl. Acad. Sci. U. S. A. 99:12246–12251.

Martin W. 2009. Gene transfer from organelles to nucleus: Frequent and in big chunks. Sciences-New York. 100:8612–8614. doi: 10.1073/pnas.

Matsumoto T et al. 2011. Green-colored plastids in the dinoflagellate genus *Lepidodinium* are of core chlorophyte origin. Protist. 162:268–76. doi: 10.1016/j.protis.2010.07.001.

McEwan ML, Keeling PJ. 2004. HSP90, tubulin and actin are retained in the tertiary endosymbiont genome of *Kryptoperidinium foliaceum*. J. Euk. Microbiol. 51:651–659.

McFadden GI. 2001. Primary and secondary endosymbiosis and the origin of plastids. J. Phycol. 37:951–959.

Michel F, Umesono K, Ozeki H. 1989. Comparative and functional anatomy of group II catalytic introns: a review. Gene. 82:5–30.

Min XJ, Butler G, Storms R, Tsang A. 2005. OrfPredictor: predicting protein-coding regions in EST-derived sequences. Nucleic Acids Res. 33:W677–80. doi: 10.1093/nar/gki394.

Minge MA et al. 2010. A phylogenetic mosaic plastid proteome and unusual plastid-targeting signals in the green-colored dinoflagellate *Lepidodinium chlorophorum*. BMC Evol. Biol. 10:191. doi: 10.1186/1471-2148-10-191.

Moore RB et al. 2008. A photosynthetic alveolate closely related to apicomplexan parasites. Nature. 451:959–63. doi: 10.1038/nature06635.

Moustafa A et al. 2009. Genomic footprints of a cryptic plastid endosymbiosis in diatoms. Science (New York, N.Y.). 324:1724–6. doi: 10.1126/science.1172983.

Moustafa A, Bhattacharya D. 2008. PhyloSort: a user-friendly phylogenetic sorting tool and its application to estimating the cyanobacterial contribution to the nuclear genome of *Chlamydomonas*. BMC Evol. Biol. 8:6. doi: 10.1186/1471-2148-8-6.

Nash EA, Nisbet RER, Barbrook AC, Howe CJ. 2008. Dinoflagellates: a mitochondrial genome all at sea. Trends Genet : TIG. 24:328–35. doi: 10.1016/j.tig.2008.04.001.

Nash EA et al. 2007. Organization of the mitochondrial genome in the dinoflagellate *Amphidinium carterae*. Biosystems. 24:1528–1536. doi: 10.1093/molbev/msm074.

Nassoury N, Cappadocia M, Morse D. 2003. Plastid ultrastructure defines the protein import pathway in dinoflagellates. Journal of cell science. 116:2867–74. doi: 10.1242/jcs.00517.

Nassoury N, Morse D. 2005. Protein targeting to the chloroplasts of photosynthetic eukaryotes: getting there is half the fun. Biochim. Biophys. acta. 1743:5–19. doi: 10.1016/j.bbamcr.2004.09.017.

Nierman W et al. 2001. Complete genome sequence of *Caulobacter crescentus*. Proc. Natl. Acad. Sci. USA. 98:4136–4141.

Niu Y-F et al. 2012. Transformation of diatom *Phaeodactylum tricornutum* by electroporation and establishment of inducible selection marker. BioTechniques. 1–3. doi: 10.2144/000113881.

Norman JE, Gray MW. 2001. A complex organization of the gene encoding cytochrome oxidase subunit 1 in the mitochondrial genome of the dinoflagellate *Crypthecodinium cohnii*: humologous recombination generates two different cox1 open reading frames. J. Mol. Evol. 53:351–363. doi: 10.1007/s002390010225.

Nosenko T et al. 2006. Chimeric plastid proteome in the Florida “red tide” dinoflagellate *Karenia brevis*. Mol. Biol. Evol. 23:2026–38. doi: 10.1093/molbev/msl074.

Oudot-Le Secq M-P, GR. 2011. Complex repeat structures and novel features in the mitochondrial genomes of the diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*. Gene. 476:20–6. doi: 10.1016/j.gene.2011.02.001.

Oudot-Le Secq M-P, Loiseaux-de Goër S, Stam WT, Olsen JL. 2006. Complete mitochondrial genomes of the three brown algae (Heterokonta: Phaeophyceae) *Dictyota dichotoma*, *Fucus vesiculosus* and *Desmarestia viridis*. Curr. Genet. 49:47–58. doi: 10.1007/s00294-005-0031-4.

Oudot-Le Secq MP et al. 2007. Chloroplast genomes of the diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*: comparison with other plastid genomes of the red lineage. Mol. Genet. Genomics. 277:427–439.

Ozsolak F et al. 2009. Direct RNA sequencing. Nature. 461:814–8. doi: 10.1038/nature08390.

Palmer JD. 2003. The symbiotic birth and spread of plastids: How many times and whodunit? J. Phycol. 39:4–11. doi: 10.1046/j.1529-8817.2003.02185.x.

Park MG, Kim M, Kim S, Yih W. 2010. Does *Dinophysis caudata* (Dinophyceae) have permanent plastids? J. Phycol. 46:236–242. doi: 10.1111/j.1529-8817.2009.00777.x.

Patron NJ, Waller RF, Keeling PJ. 2006. A tertiary plastid uses genes from two endosymbionts. J. Mol. Biol. 357:1373–1382.

Patron NJ, Waller RF. 2007. Transit peptide diversity and divergence: A global analysis of plastid targeting signals. BioEssays : news and reviews in molecular, cellular and developmental biology. 29:1048–58. doi: 10.1002/bies.20638.

Patron NJ, Waller RF, Archibald JM, Keeling PJ. 2005. Complex protein targeting to dinoflagellate plastids. Gene. 1015–1024. doi: 10.1016/j.jmb.2005.03.030.

Pienaar RN, Sakai H, Horiguchi Takeo. 2007. Description of a new dinoflagellate with a diatom endosymbiont, *Durinskia capensis* sp nov (Peridiniales, Dinophyceae) from South Africa. J. Plant Res. 120:247–258.

Price MN, Dehal PS, Arkin AP. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. 26:1641–50. doi: 10.1093/molbev/msp077.

Qiu D, Huang L, Liu S, Lin S. 2011. Nuclear, Mitochondrial and plastid gene phylogenies of *Dinophysis miles* (Dinophyceae): Evidence of variable types of chloroplasts. PloS one. 6:e29398. doi: 10.1371/journal.pone.0029398.

Ravin N V et al. 2010. Complete sequence of the mitochondrial genome of a diatom alga *Synedra acus* and comparative analysis of diatom mitochondrial genomes. Curr. Genet. 56:215–23. doi: 10.1007/s00294-010-0293-3.

Reyes-Prieto A, Hackett JD, Soares MB, Bonaldo MF, Bhattacharya D. 2006. Cyanobacterial contribution to algal nuclear genomes is primarily limited to plastid functions. Curr. Biol. 16:2320–5. doi: 10.1016/j.cub.2006.09.063.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: The European molecular biology open software suite. Trends Genet. 16:276–277.

Roger AJ, Clark CG, Doolittle WF. 1996. A possible mitochondrial gene in the early-branching amitochondriate protist *Trichomonas vaginalis*. Proc. Natl. Acad. Sci. U. S. A. 93:14618–14622.

Roger AJ, Silberman JD. 2002. Mitochondria in hiding. Nature. 419: 827:9.

Roise D et al. 1988. Amphiphilicity is essential for mitochondrial presequence function. EMBO J. 7:649–653.

Rutherford K et al. 2000. Artemis: sequence visualization and annotation. Bioinformatics. 16:944–945.

Schattner P, Brooks AN, Lowe TM. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 33:W686–W689.

Schnepf E, Elbraechter M. 1988. cryptophycean-like double membrane-bound chloroplast in the dinoflagellate, *Dinophysis ehrenb* evolutionary, phylogenetic and toxicological implications. Botanica Acta. 101:196–203.

Schnepf E, Elbrachter M. 1999. Dinophyte chloroplasts and phylogeny-A review. Grana. 38:81–97.

Schwartz S et al. 2000. PipMaker - A Web server for aligning two genomic DNA sequences. Genome Res. 10:577–586.

Sheiner L, Striepen B. 2012. Protein sorting in complex plastids. Biochim. Biophys. acta. doi: 10.1016/j.bbamcr.2012.05.030.

Shimodaira H, Hasegawa M. 1989. Letter to the editor multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16:1114–1116.

Slamovits CH, Saldarriaga JF, Larocque A, Keeling PJ. 2007. The highly reduced and fragmented mitochondrial genome of the early-branching dinoflagellate *Oxyrrhis marina* shares characteristics with both apicomplexan and dinoflagellate mitochondrial genomes. J. Mol. Biol. 372:356–68. doi: 10.1016/j.jmb.2007.06.085.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics (Oxford, England). 22:2688–90. doi: 10.1093/bioinformatics/btl446.

Takano Y, Hansen G, Daisuke F, Horiguchi T. 2008. Serial replacement of diatom endosymbionts in two freshwater dinoflagellates , *Peridiniopsis* spp. (Peridiniales, Dinophyceae). Phycologia. 47:41–53. doi: 10.2216/07-36.1.

Tamura M, Shimada S, Horiguchi T. 2005. *Galeidiniium rugatum* gen. et sp nov (Dinophyceae), a new coccoid dinoflagellate with a diatom endosymbiont. J. Phycol. 41:658–671. doi: 10.1111/j.1529-8817.2005.00085.x.

Taylor FJR. 2004. Extraordinary dinoflagellates: past and present. In: Neo-science of natural history: Integration of geoscience and biodiversity studies. Mawatari, SF & Okada, H, editors. Sapporo pp. 61–66.

Taylor FJR., Hoppenrath M, Saldarriaga JF. 2007. Dinoflagellate diversity and distribution. Biodiversity and Conservation. 17:407–418. doi: 10.1007/s10531-007-9258-3.

Tengs T et al. 2000. Phylogenetic analyses indicate that the 19'hexanoyloxy-fucoxanthin-containing dinoflagellates have tertiary plastids of haptophyte origin. Mol. Biol. Evol. 17:718–729.

Tesler G. 2002. GRIMM: genome rearrangements web server. Bioinformatics. 18:492–493.

Timmis JN, Ayliffe MA, Huang CY, Martin W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat. Rev. Genet. 5:123–35. doi: 10.1038/nrg1271.

Tippit DH, Pickett-Heaps JD. 1976. Apparent amitosis in the binucleate dinoflagellate *Peridinium balticum*. J. Cell. Sci. 21:273–289.

Tomas RN, Cox E. 1973. Observations on the symbiosis of *Peridinium balticum* and its intracellular alga .1. Ultrastructure. J. Phycol. 9:304–323.://A1973Q857400013.

Tomas RN, Cox E R, Steiding KA. 1973. *Peridinium-balticum* (Levander) Lemmermann, an unusual dinoflagellate with a mesocaryotic and an eukaryotic nucleus. J. Phycol. 9:91–98.://A1973P572900009.

Tonkin CJ, Roos DS, McFadden GI. 2006. N-terminal positively charged amino acids, but not their exact position, are important for apicoplast transit peptide fidelity in *Toxoplasma gondii*. Molecular and biochemical parasitology. 150:192–200. doi: 10.1016/j.molbiopara.2006.08.001.

Tovar J et al. 2003. Mitochondrial remnant organelles of *Giardia* function in iron-sulphur protein maturation. Nature (London). 426:172–176.

Van Dooren GG, Schwartzbach SD, Osafune T, McFadden GI. 2001. Translocation of proteins across the multiple membranes of complex plastids. Biochim. Biophys. acta. 1541:34–53.

Von Heijne G, Steppuhn J, Herrmann RG. 1989. Domain structure of mitochondrial and chloroplast targeting peptides. Eur. J. Biochem / FEBS. 180:535–45.

Von Heijne G. 1986. Mitochondrial targeting form amphiphilic helices. The EMBO Journal. 5:1335–1342.

Waller RF, Jackson CJ. 2009. Dinoflagellate mitochondrial genomes: stretching the rules of molecular biology. BioEssays. 31:237–45. doi: 10.1002/bies.200800164.

Waller RF, Reed MB, Cowman AF, Mcfadden GI. 2000. Protein traficking to the plastid of *Plasmodium falciparum* is via the secretory pathway. The EMBO Journal. 19.

Wastl J, Maier U G. 2000. Transport of proteins into cryptomonads complex plastids. J. Biol. Chem. 275:23194–8. doi: 10.1074/jbc.M003125200.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18:691–9.

Williams BAP, Hirt RP, Lucocq JM, Embley TM. 2002. A mitochondrial remnant in the microsporidian *Trachipleistophora hominis*. Nature (London). 418:865–869.

Williams BAP., Keeling PJ. 2003. Cryptic organelles in parasitic protists and fungi. Advances in Parasitol. 54:9–68.

Wisecaver JH, Hackett JD. 2010. Transcriptome analysis reveals nuclear-encoded proteins for the maintenance of temporary plastids in the dinoflagellate *Dinophysis acuminata*. BMC genomics. 11:366. doi: 10.1186/1471-2164-11-366.

Withers NW, Cox ER., Tomas R, Haxo FT. 1977. Pigments of the dinoflagellate *Peridinium balticum* and its photosynthetic endosymbiont1. J. Phycol. 13:354–358. doi: 10.1111/j.1529-8817.1977.tb00610.x.

Woehle C, Dagan T, Martin WF, Gould SB. 2011. Red and problematic green phylogenetic signals among thousands of nuclear genes from the photosynthetic and apicomplexa-related *Chromera velia*. Genome Biol. Evol. 3:1220–30. doi: 10.1093/gbe/evr100.

Wyman SK, Jansen RK, Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics. 20:3252–3255.://000225361400041.

Yokoyama A, Takahashi F, Kataoka H, Hara Y, Nozaki H. 2011. Evolutionary analyses of the nuclear-encoded photosynthetic gene *psbO* from tertiary plastid-containingalgae in dinophyta. J. Phycol. 47:407–414. doi: 10.1111/j.1529-8817.2011.00961.x.

Yoon HS, Hackett JD, Bhattacharya D. 2002. A single origin of the peridinin- and fucoxanthin-containing plastids in dinoflagellates through tertiary endosymbiosis. Proc. Natl. Acad. Sci. U. S. A. 99:11724–11729.

Zhang H, Lin S. 2005. Mitochondrial cytochrome b mRNA editing in dinoflagellates: Possible ecological and evolutionary associations? J. Euk. Microbiol. 52:538–545.

Zhang Q, Liu G, Hu Z. 2011. Morphological differences and molecular phylogeny of freshwater blooming species, *Peridiniopsis* spp. (Dinophyceae) from China. Eur. J. Protistology. 47:149–60. doi: 10.1016/j.ejop.2011.03.001.

Appendices

Appendix 1: Supplementary figures and tables of chapter 2

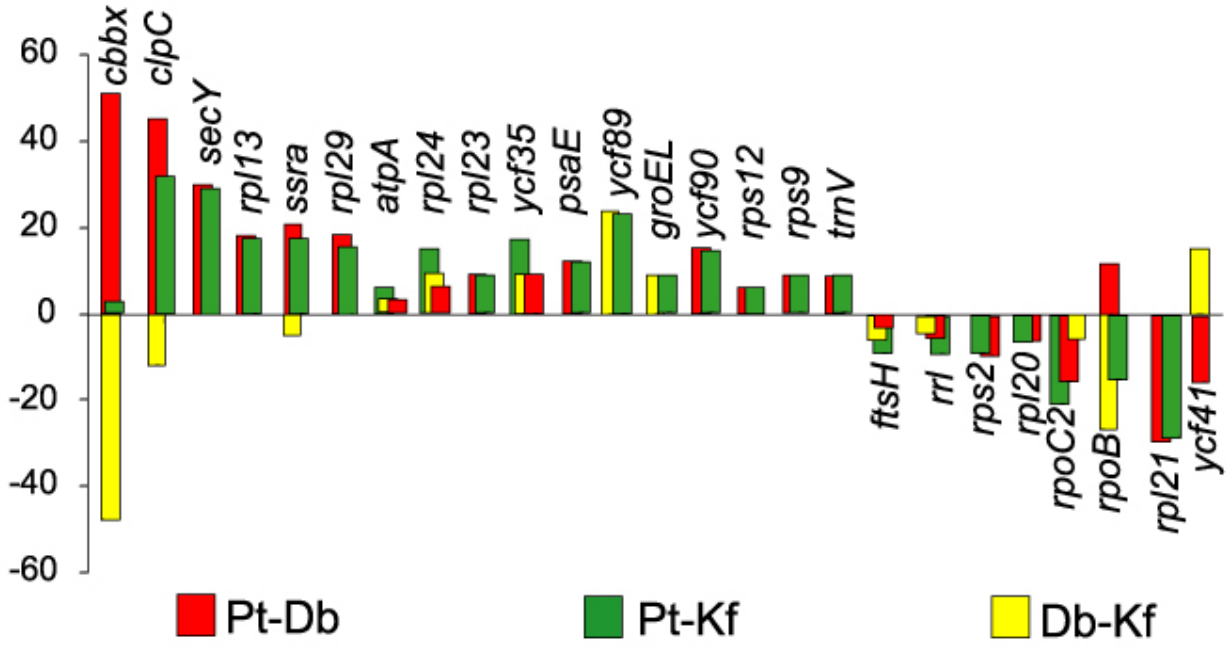

Figure 2.S1: Length comparison of the genes encoded in the plastid genomes of *D. baltica*, *K. foliaceum* and the pennate diatom *Phaeodactylum tricornutum*.

Abbreviations: Db, *D. baltica*; Kf, *K. foliaceum*; Pt, *Phaeodactylum tricornutum*.

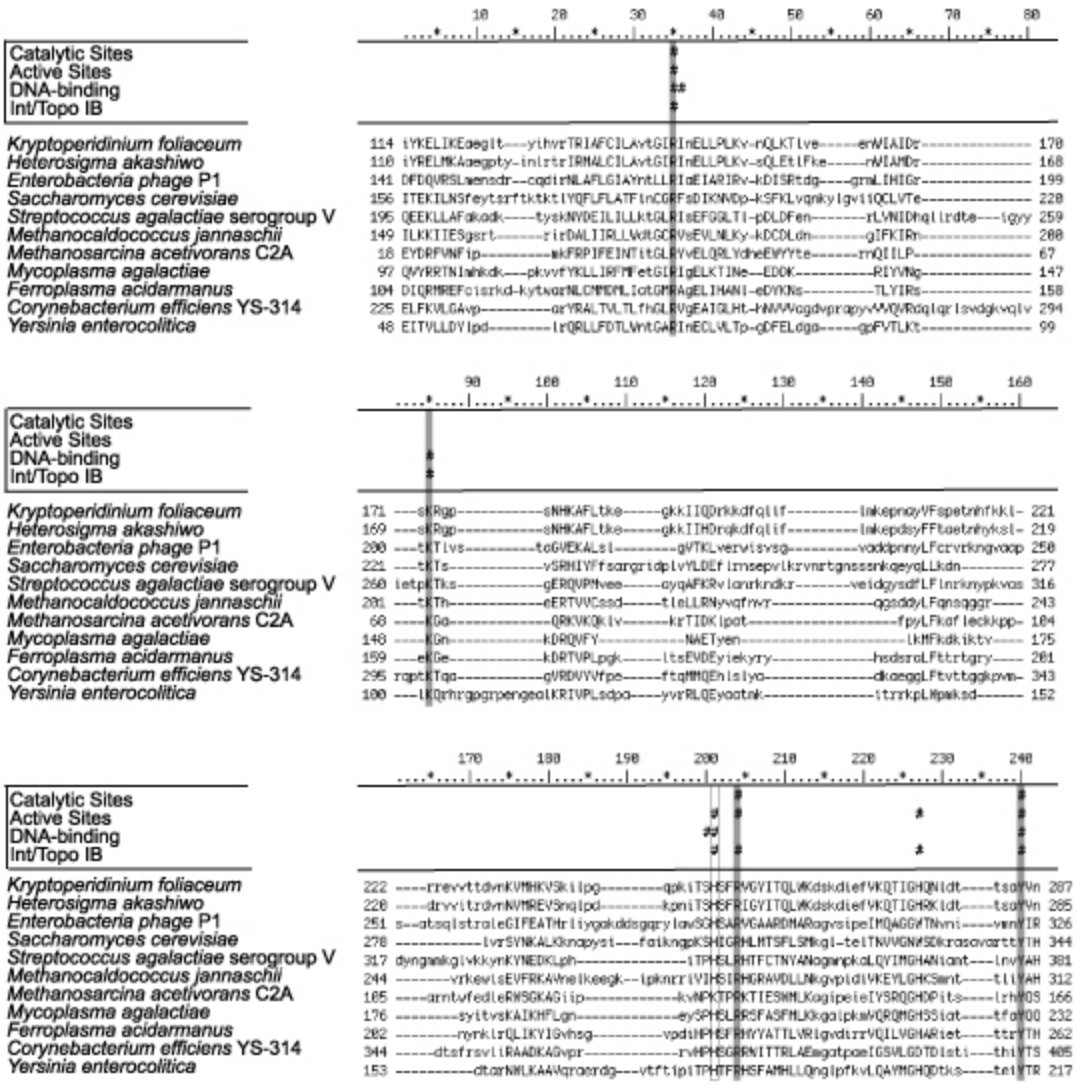

Figure 2.S2: *K. foliaceum* TyrC conserved catalytic, active, and DNA-binding sites.

The sequences of *Heterosigma akashiwo*'s TyrC were manually added to the Conserved Domain Database (CDD) alignment for *K. foliaceum*'s TyrC. The conserved residues with specific functions are marked with a number sign (#) above the alignments. Shaded residues indicate invariable sites among all the recombinases in the alignment, and the long rectangular boxes highlight conserved sites among all the recombinases in the alignment except one.

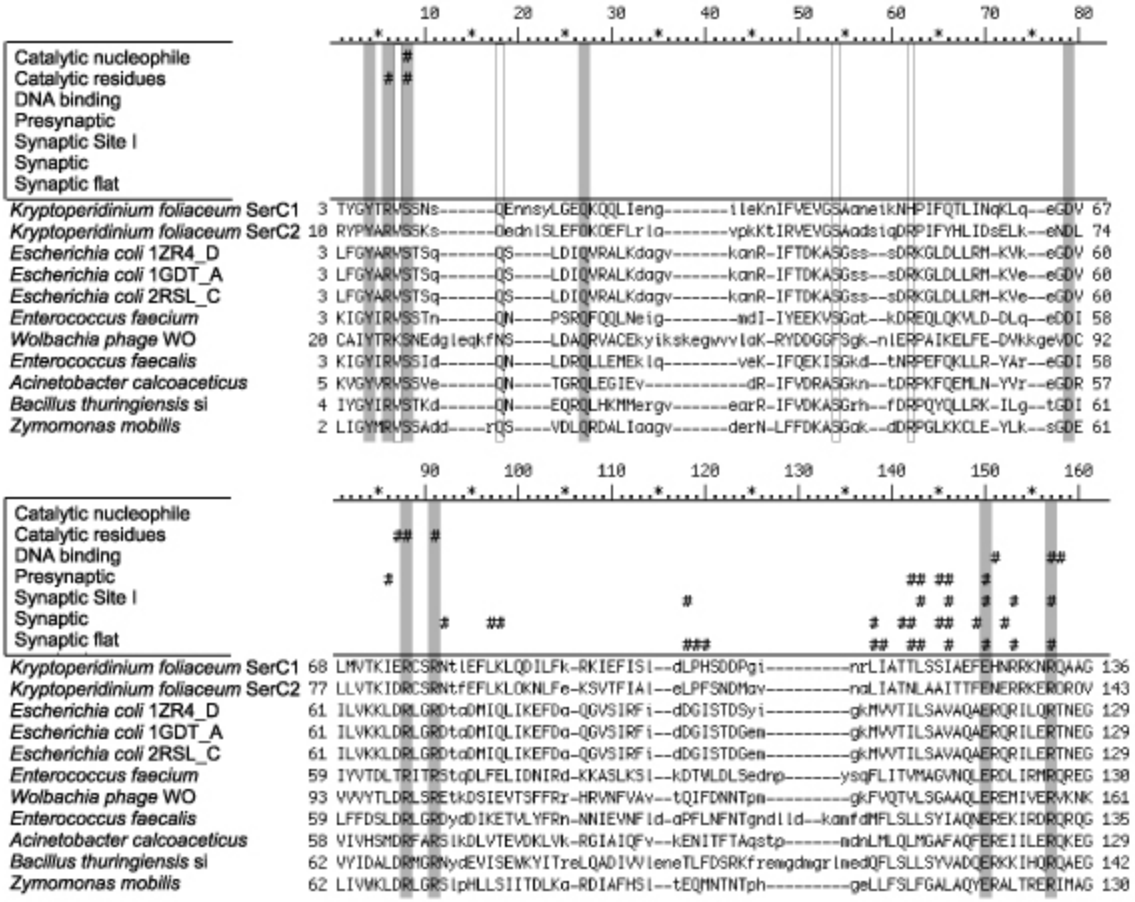

Figure 2.S3: The conserved residues found in the SerC1 and SerC2 recombinases encoded in the plastid genomes of *Kryptoperidinium foliaceum* and other site-specific serine recombinases.

The sequences of SerC2 were manually added to the Conserved Domain Database (CDD) alignment for SerC1. The conserved residues with specific functions are marked with a number sign (#) above the alignments. Shaded residues indicate invariable sites among all the recombinases in the alignment, and the long rectangular boxes highlight conserved sites among all the recombinases in the alignment except one.

Appendix 2: Supplementary figures and tables of chapter 3

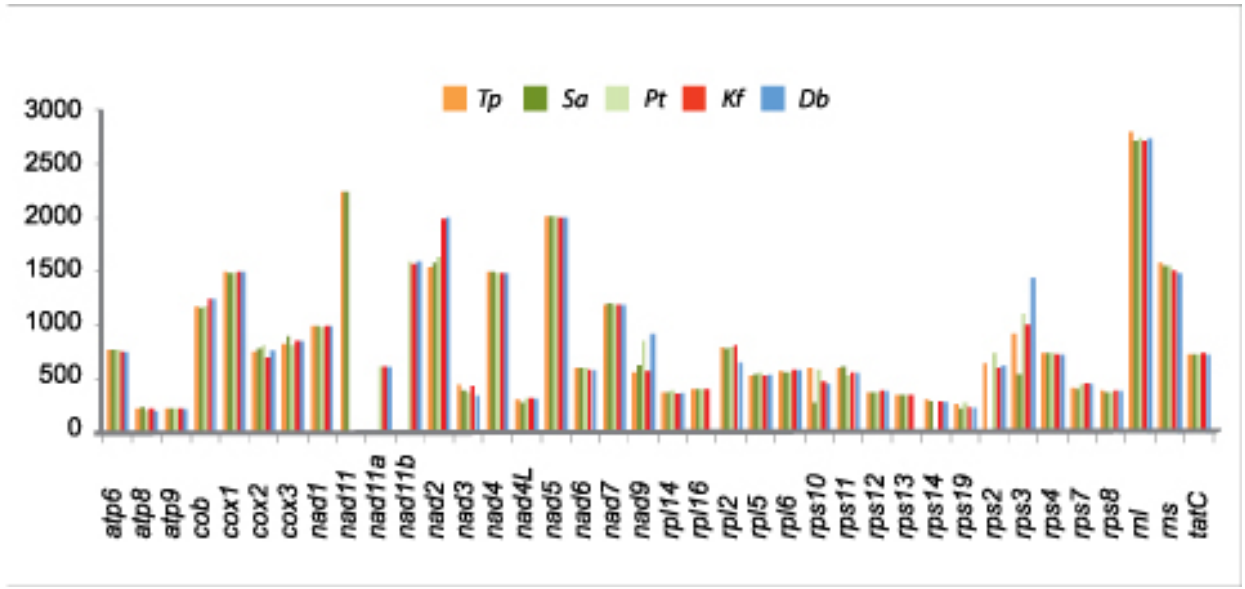

Figure 3.S1: Gene size comparisons between the protein-coding and rRNA genes in the two mitochondrial genomes of the dinotom endosymbionts and those of three diatoms.

Ts, *Thalassiosira pseudonana*; Sa, *Synedra acus*; Pt, *Phaeodactylum tricornutum*; Kf, *Kryptoperidinium foliaceum*; Db, *Durinskia baltica*.

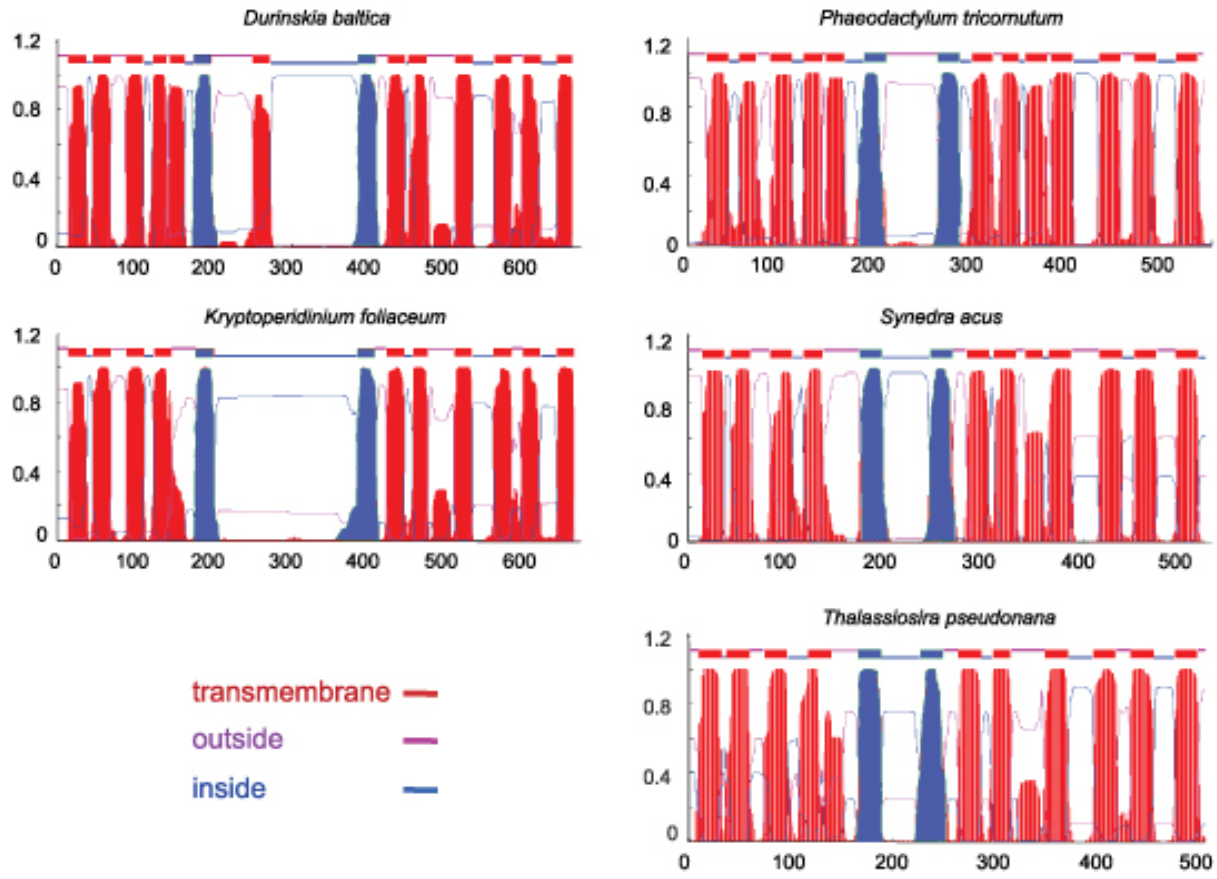

Figure 3.S2: Posterior probabilities for transmembrane helices in *nad2* gene of the two endosymbionts and other diatoms.

The X-axis shows the amino acid number, and the Y-axis the probability. The two conserved transmembrane helices flanking the dinotoms' inserts are painted blue in dinotoms and diatoms.

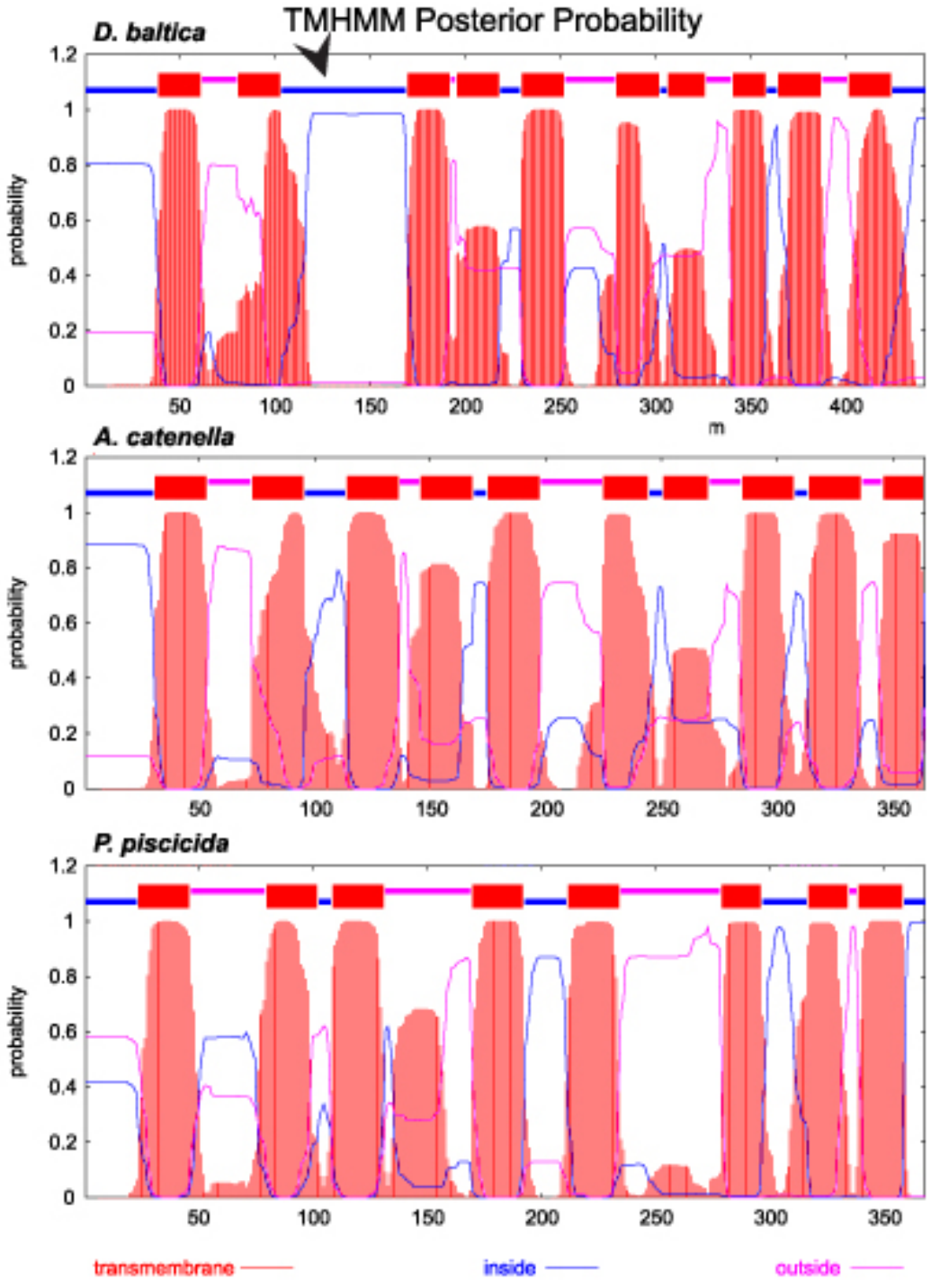

Figure 3.S3: Posterior probabilities for transmembrane helices in *cob* gene of the host in *D. baltica* compared to that in *Pfiesteria piscicida* and *Alexandrium catenella*.

The X-axis shows the amino acid number, and the Y-axis the probability. The black arrow head marks the position of the insert within the *cob* gene in *D. baltica*.

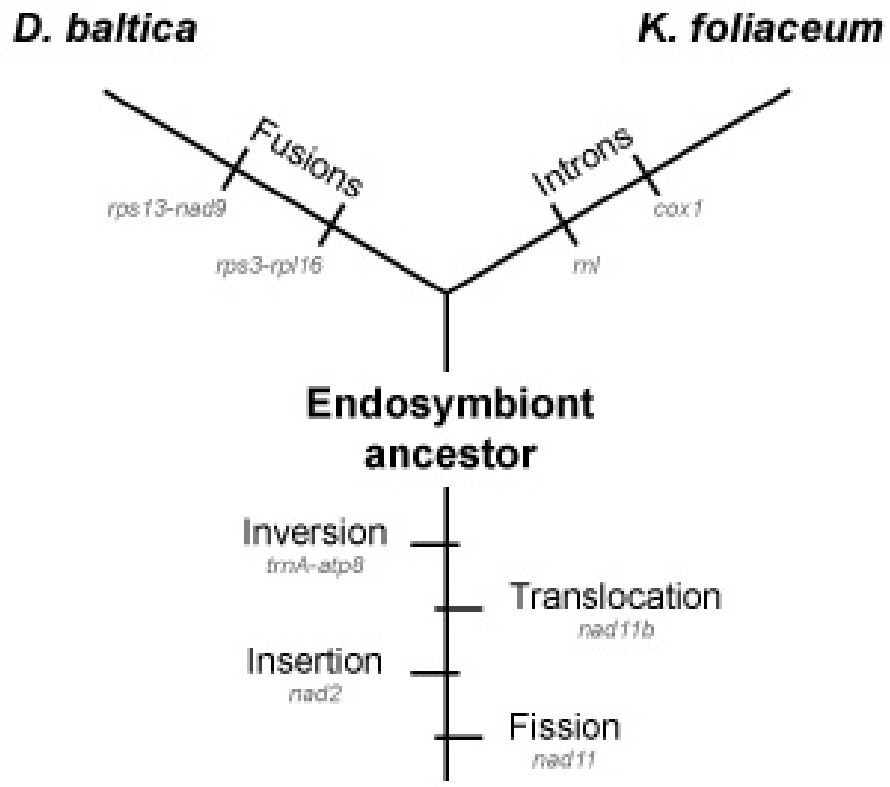

Figure 3.S4: A few ancestral and derived characters in the mitochondrial genomes of the endosymbionts in the two dinotoms inferred based on the most parsimonious scenario.

The sequence of events is arbitrary.

Table 3.S1: Editing sites in the cox1 mRNA of *Durinskia baltica* and *Kryptoperidinium foliaceum*

Durinskia baltica				
DNA Site	DNA	RNA	Codon Site	Change aa
154	**A**	**G**	**1st**	**I → V**
175	T	C	1st	F → L
305	C	U	2nd	S → F
445	A	G	1st	I → V
515	A	G	2nd	Y → C
658	A	G	1st	I → V
736	A	G	1st	I → V
739	T	C	1st	F → L
748	A	G	1st	I → V
776	T	C	2nd	L → S
998	**A**	**G**	**2nd**	**K → R**
1004	**A**	**G**	**2nd**	**N → S**
1009	**C**	**U**	**1st**	**P → S**
1012	**T**	**C**	**1st**	**F → L**
1019	**G**	**C**	**2nd**	**G → A**
1063	**A**	**G**	**1st**	**I → V**
1094	**G**	**C**	**2nd**	**G → A**
1114	A	G	1st	T → A
1198	G	C	1st	V → L
1211	A	G	2nd	N → S
1225	T	C	1st	S → P
1267	A	G	1st	I → V

Kryptoperidinium foliaceum				
DNA Site relative to *D. baltica*	DNA	RNA	Codon Site	Change aa
76	A	G	1st	I → V
90	A	G	3rd	I → M
154	**A**	**G**	**1st**	**I → V**
676	A	G	1st	I → V
998	**A**	**G**	**2nd**	**K → R**
1004	**A**	**G**	**2nd**	**N → S**
1009	**C**	**U**	**1st**	**P → S**
1012	**T**	**C**	**1st**	**F → L**
1019	**G**	**C**	**2nd**	**G → A**
1063	**A**	**G**	**1st**	**I → V**
1094	**G**	**C**	**2nd**	**G → A**

Editing sites on the *cox1* mRNA in the dinoflagellate host of *D. baltica* and *K. foliaceum* and the deduced resulting amino acid change in the Cox1 protein inferred from the differences found in the gene and its corresponding transcript sequences. The bold fonts mark the conserved changes seen in the two species.

Appendix 3: Supplementary figures and tables of chapter 4

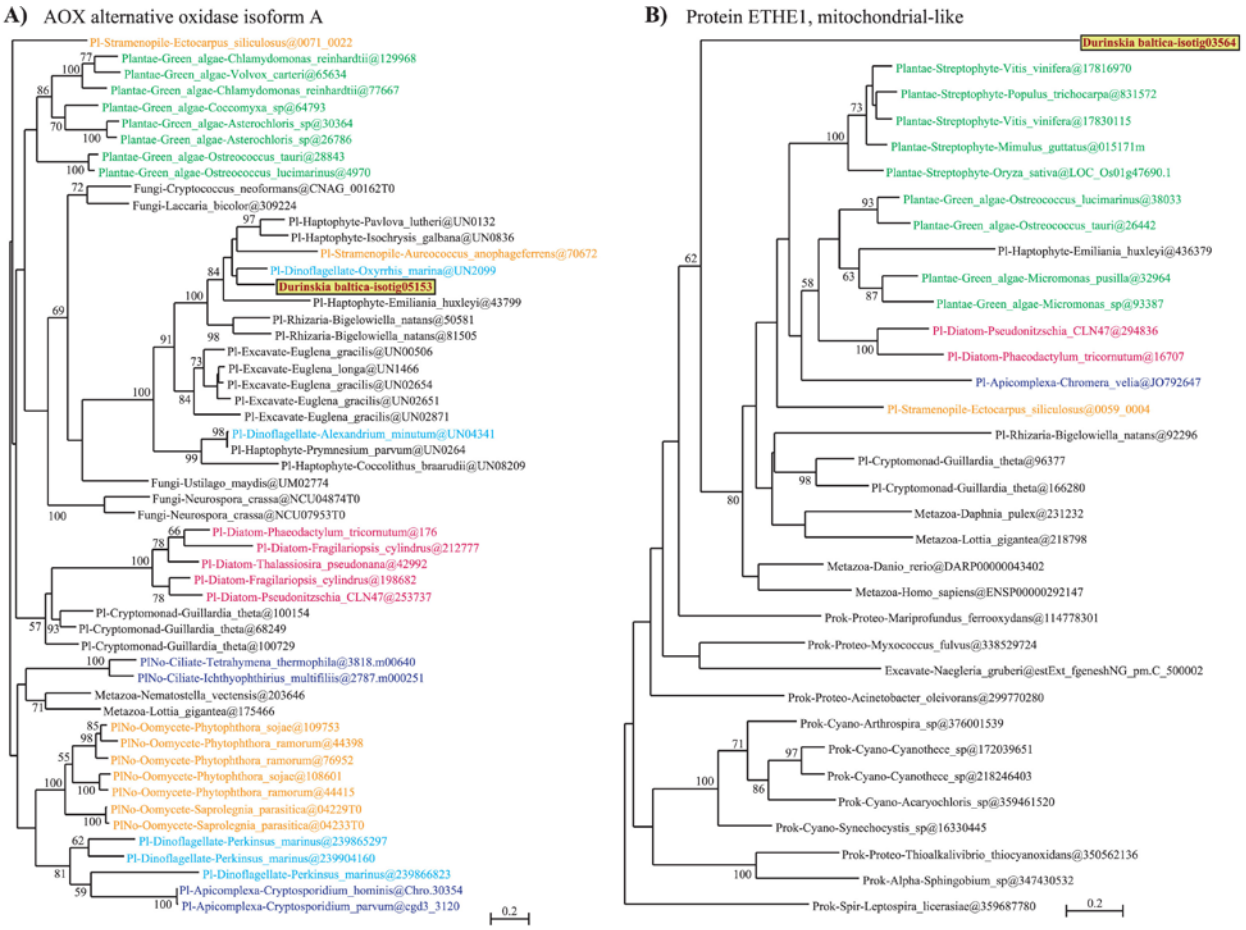

Figure 4.S1: The maximum likelihood trees with an unclear phylogenetic affinity and/or origin for the host putative nuclear-encoded mitochondrial proteins in *Durinskia baltica*.

A) AOX alternative oxidase isoform A, partial tree, B) Protein ETHE1, mitochondrial-like. Numbers at the nodes indicate bootstrap support ≥ 50%. A yellow box highlights the position of *D. baltica*. Several taxa of interest are color-coded: Dinoflagellates in light blue; Apicomplexans and Ciliates in dark blue; Diatoms in scarlet; Stramenopiles in orange; Green algae and Plants in green; Red algae in red; All other taxa in black. Pl, taxon with a plastid; PlNo, taxon without plastids.

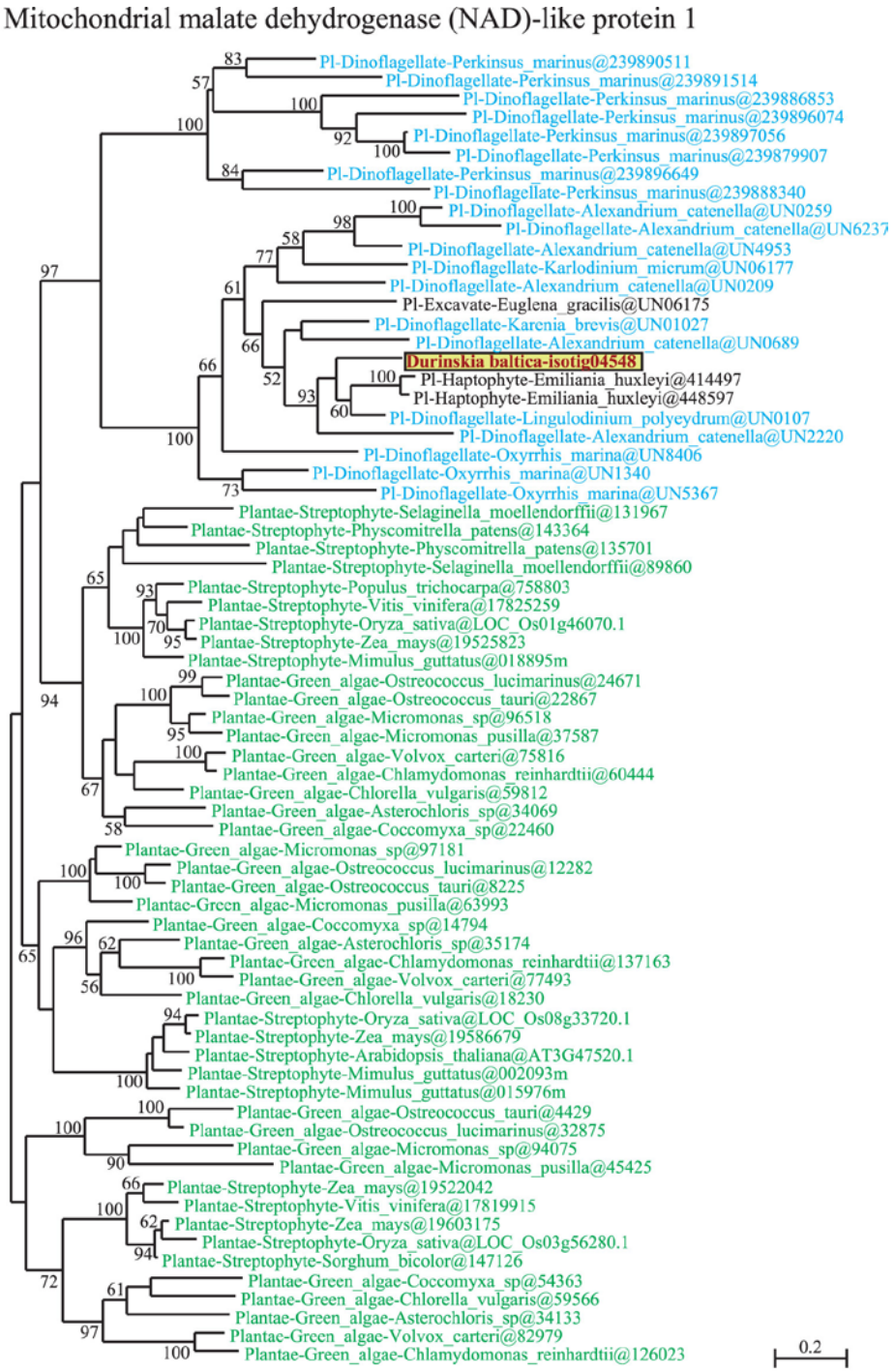

Figure 4.S2: The maximum likelihood tree for mitochondrial malate dehydrogenase (NAD)-like protein 1, partial tree.

Numbers at the nodes indicate bootstrap support ≥ 50%. A yellow box highlights the position of *D. baltica*. Several taxa of interest are color-coded: Dinoflagellates in light blue; Apicomplexans and Ciliates in dark blue; Diatoms in scarlet; Stramenopiles in orange; Green algae and Plants in green; Red algae in red; All other taxa in black. Pl, taxon with a plastid; PlNo, taxon without plastids.

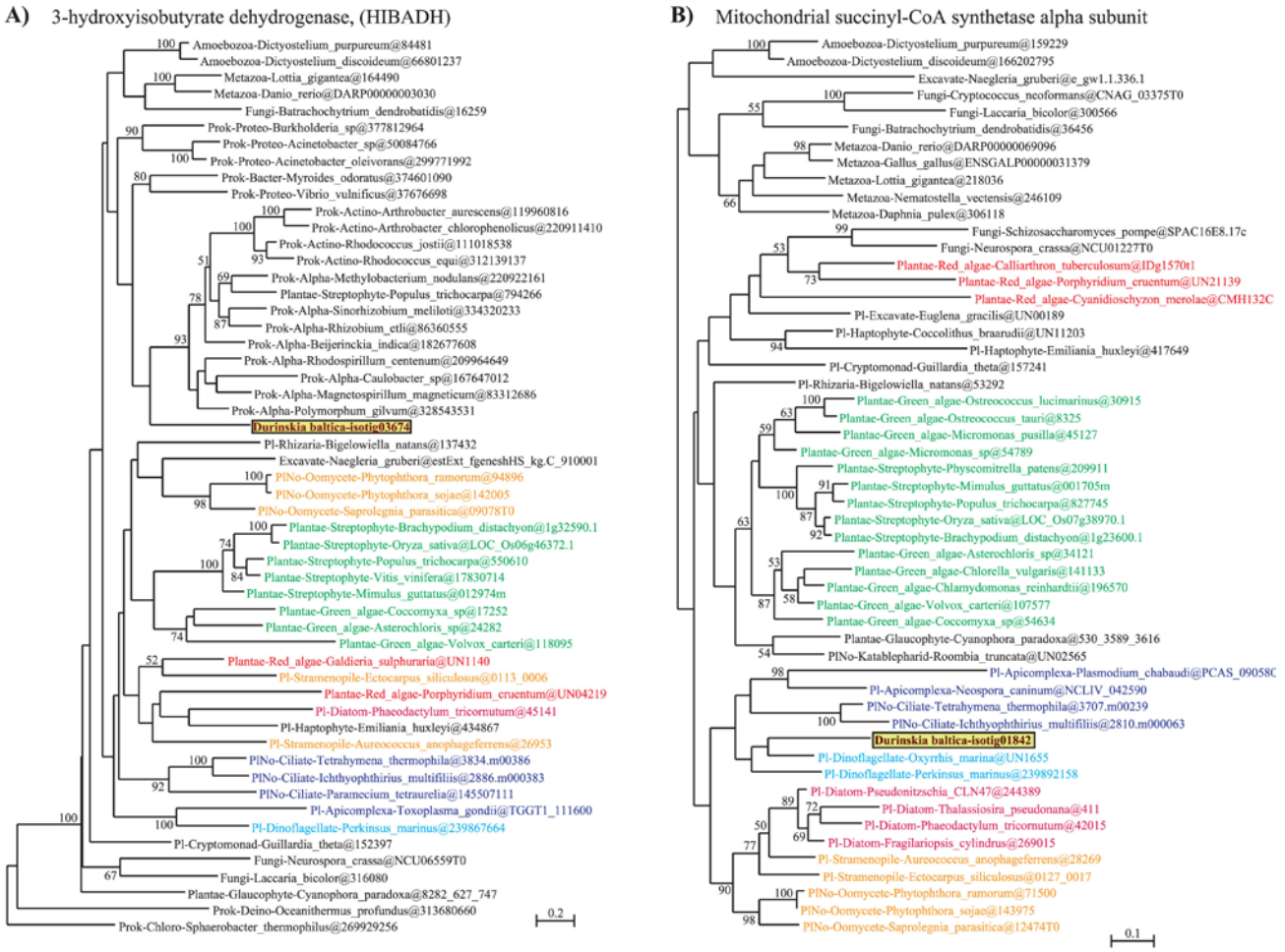

Figure 4.S3: The maximum likelihood trees with an unclear phylogenetic affinity and/or origin for the host putative nuclear-encoded mitochondrial proteins in *Durinskia baltica*.

A) 3-hydroxyisobutarate dehydrogenase (HIBADH), partial tree, B) Mitochondrial succinyl-CoA synthetase alpha subunit, partial tree. Numbers at the nodes indicate bootstrap support ≥ 50%. A yellow box highlights the position of *D. baltica*. Several taxa of interest are color-coded: Dinoflagellates in light blue; Apicomplexans and Ciliates in dark blue; Diatoms in scarlet; Stramenopiles in orange; Green algae and Plants in green; Red algae in red; All other taxa in black. Pl, taxon with a plastid; PlNo, taxon without plastids.

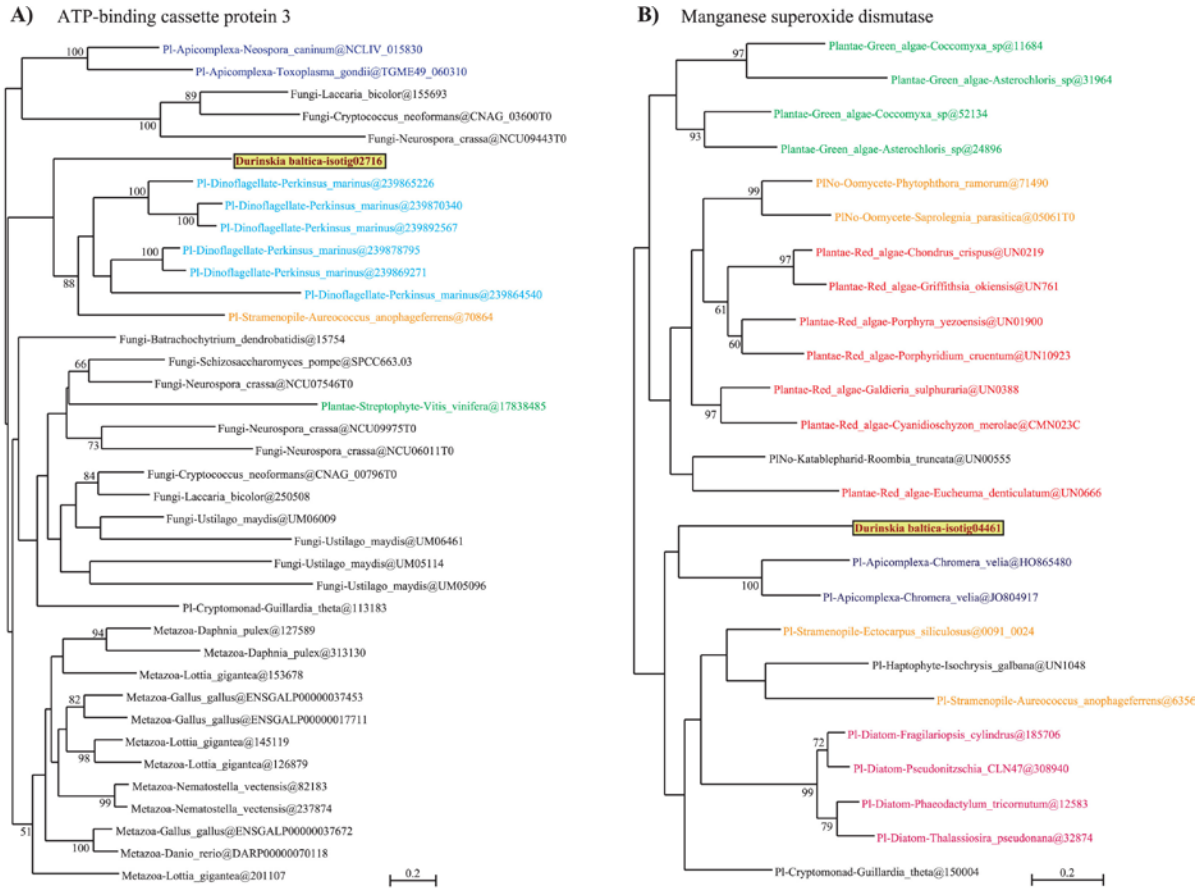

Figure 4.S4: The maximum likelihood trees with an unclear phylogenetic affinity and/or origin for the host putative nuclear-encoded mitochondrial proteins in *Durinskia baltica*.

A) ATP binding cassette protein 3, partial tree, B) Manganese superoxide dismutase, partial tree. Numbers at the nodes indicate bootstrap support ≥ 50%. A yellow box highlights the position of *D. baltica*. Several taxa of interest are color-coded: Dinoflagellates in light blue; Apicomplexans and Ciliates in dark blue; Diatoms in scarlet; Stramenopiles in orange; Green algae and Plants in green; Red algae in red; All other taxa in black. Pl, taxon with a plastid; PlNo, taxon without plastids.

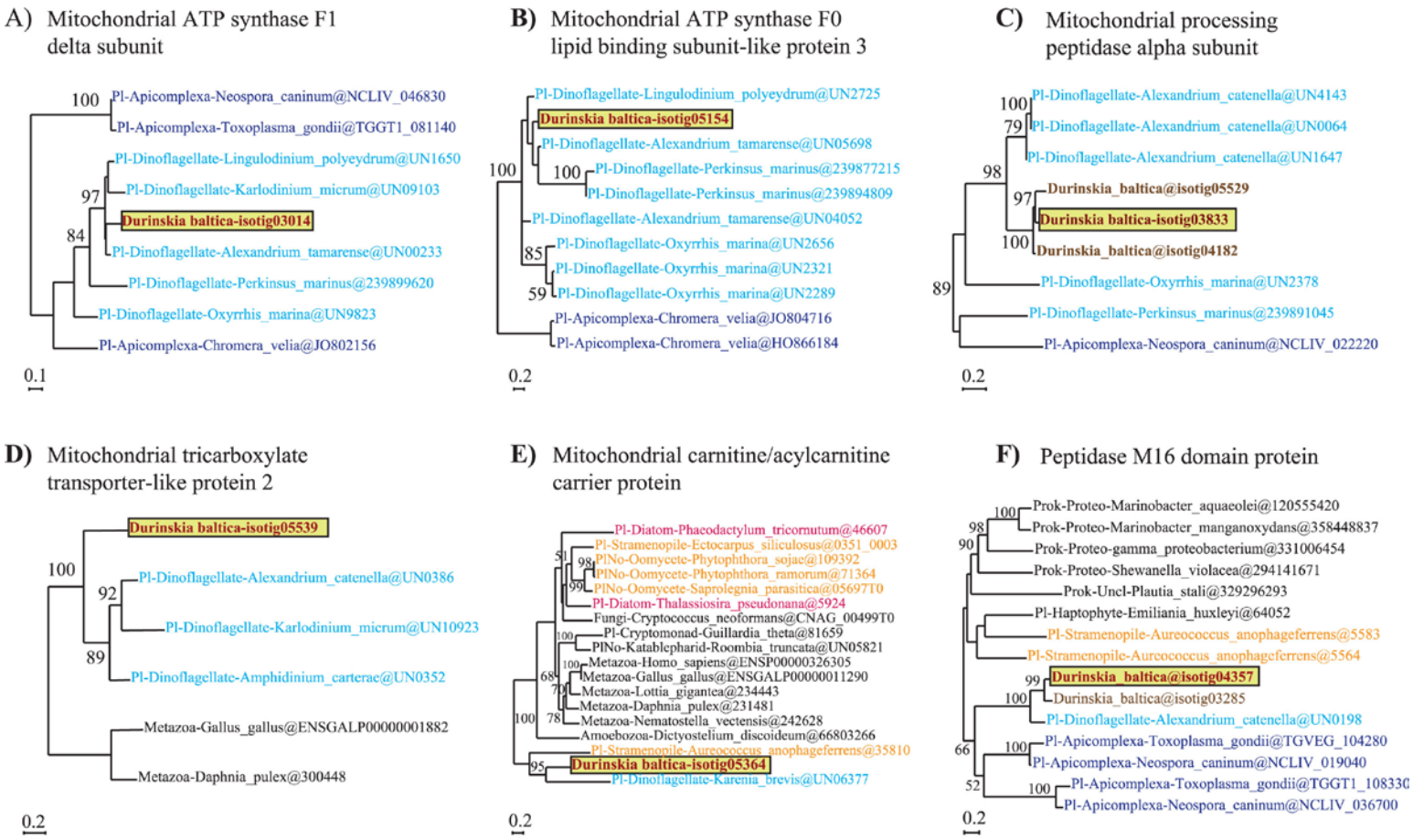

Figure 4.S5: The maximum likelihood trees with a limited number of taxa showing a dinoflagellate affinity for the host putative nuclear-encoded mitochondrial proteins in *Durinskia baltica*.

A) Mitochondrial ATP synthase F1 delta subunit, B) Mitochondrial ATP synthase F0 lipid binding subunit-like protein 3, C) Mitochondrial processing peptidase alpha subunit, D) Mitochondrial tricarboxylate transporter-like protein 2, E) Mitochondrial carnitine/acylcarnitine carrier protein, F) Peptidase M16 domain protein. Numbers at the nodes indicate bootstrap support ≥ 50%. A yellow box highlights the position of *D. baltica*. Several taxa of interest are color-coded: Dinoflagellates in light blue; Apicomplexans and Ciliates in dark blue; Diatoms in scarlet; Stramenopiles in orange; Green algae and Plants in green; Red algae in red; All other taxa in black. Pl, taxon with a plastid; PlNo, taxon without plastids.

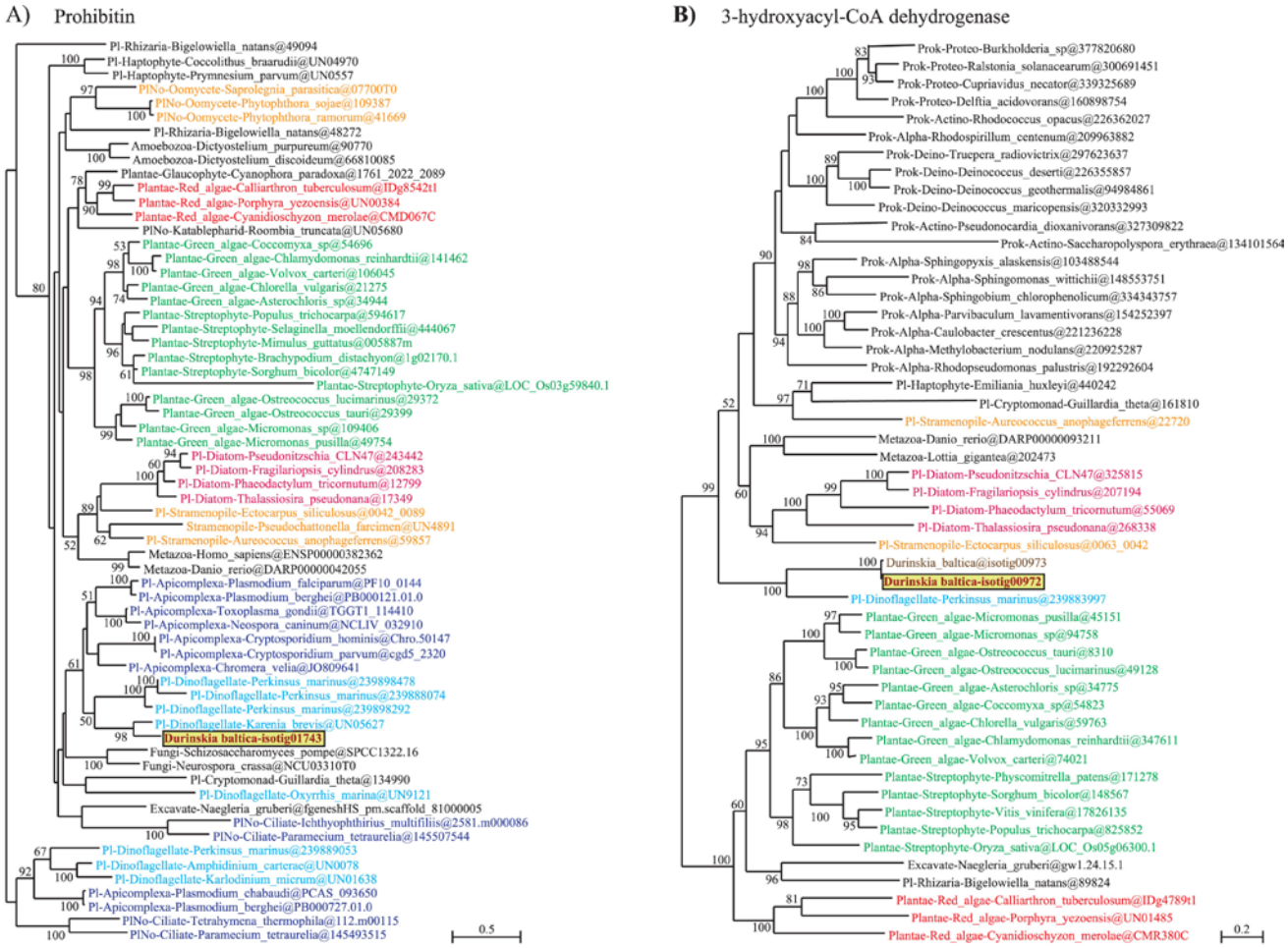

Figure 4.S6: The maximum likelihood trees with a dinoflagellate affinity and/or origin for the host putative nuclear-encoded mitochondrial proteins in *Durinskia baltica*.

A) Prohibitin, partial tree, B) 3-hydroxyacyl-CoA dehydrogenase, partial tree. Numbers at the nodes indicate bootstrap support ≥ 50%. A yellow box highlights the position of *D. baltica*. Several taxa of interest are color-coded: Dinoflagellates in light blue; Apicomplexans and Ciliates in dark blue; Diatoms in scarlet; Stramenopiles in orange; Green algae and Plants in green; Red algae in red; All other taxa in black. Pl, taxon with a plastid; PlNo, taxon without plastids.

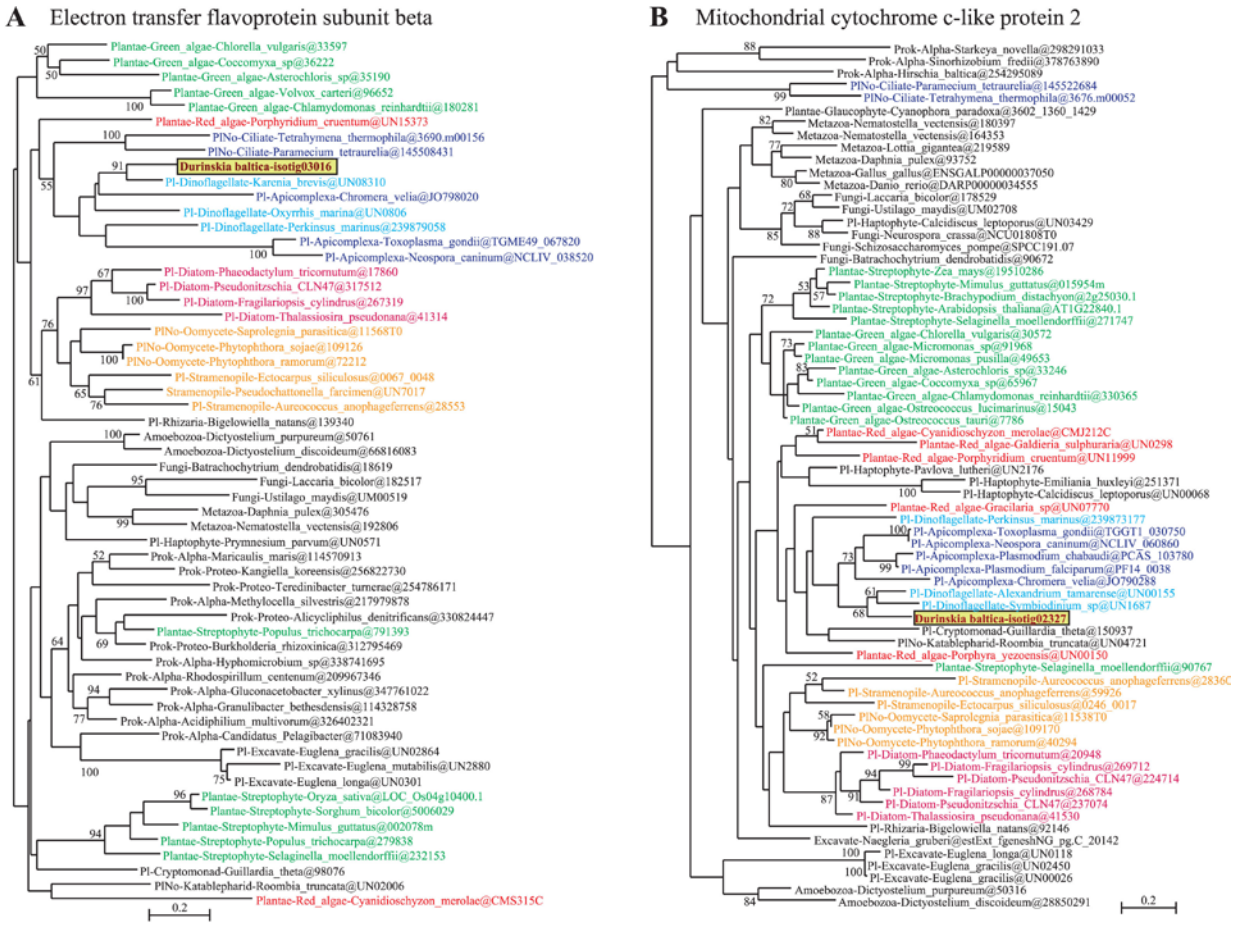

Figure 4.S7: The maximum likelihood trees with a dinoflagellate affinity and/or origin for the host putative nuclear-encoded mitochondrial proteins in *Durinskia baltica*.

A) Electron transfer flavoprotein subunit beta, B) Mitochondrial cytochrome c-like protein 2. Numbers at the nodes indicate bootstrap support ≥ 50%. A yellow box highlights the position of *D. baltica*. Several taxa of interest are color-coded: Dinoflagellates in light blue; Apicomplexans and Ciliates in dark blue; Diatoms in scarlet; Stramenopiles in orange; Green algae and Plants in green; Red algae in red; All other taxa in black. Pl, taxon with a plastid; PlNo, taxon without plastids.

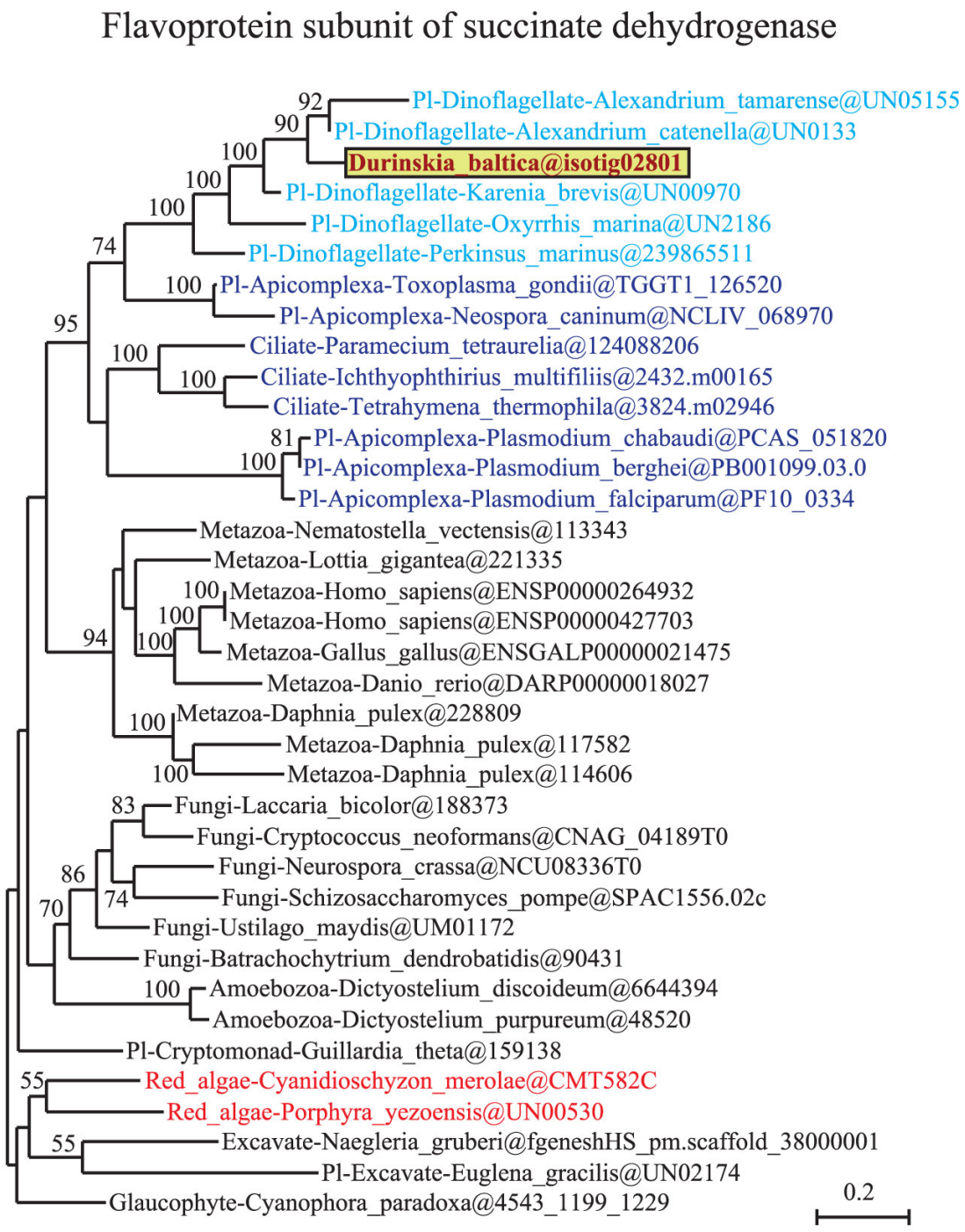

Figure 4.S8: The maximum likelihood tree for flavoprotein subunit of succinate dehydrogenase congruent with a dinoflagellate origin for the host putative nuclear-encoded mitochondrial protein in *Durinskia baltica*, partial tree.

Numbers at the nodes indicate bootstrap support ≥ 50%. A yellow box highlights the position of *D. baltica*. Several taxa of interest are color-coded: Dinoflagellates in light blue; Apicomplexans and Ciliates in dark blue; Diatoms in scarlet; Stramenopiles in orange; Green algae and Plants in green; Red algae in red; All other taxa in black. Pl, taxon with a plastid; PlNo, taxon without plastids.

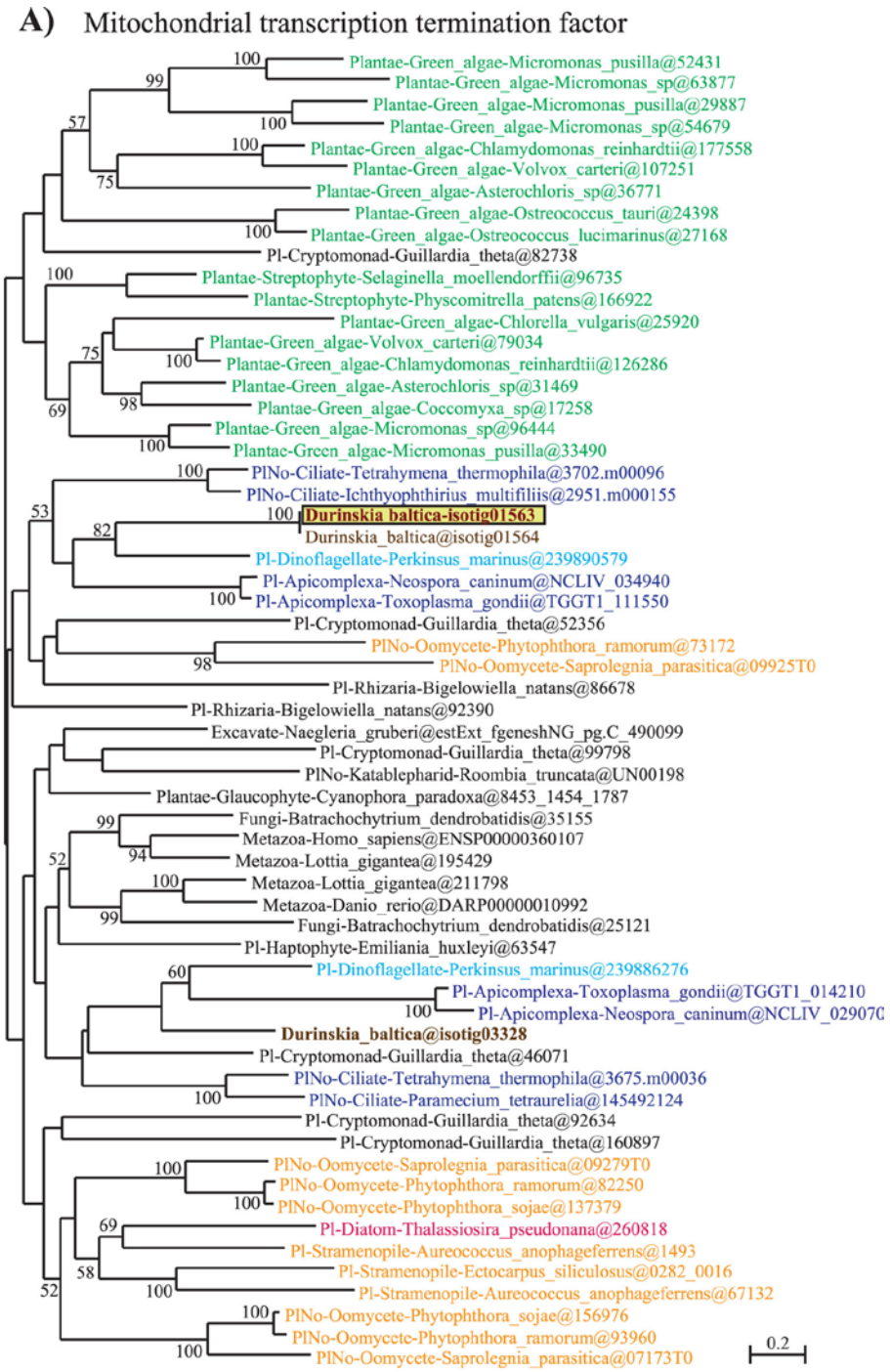

Figure 4.S9: The maximum likelihood tree for mitochondrial transcription termination factor congruent with a dinoflagellate affinity for both copies of the host putative nuclear-encoded mitochondrial protein in *Durinskia baltica*.

Numbers at the nodes indicate bootstrap support ≥ 50%. A yellow box highlights the position of *D. baltica*. Several taxa of interest are color-coded: Dinoflagellates in light blue; Apicomplexans and Ciliates in dark blue; Diatoms in scarlet; Stramenopiles in orange; Green algae and Plants in green; Red algae in red; All other taxa in black. Pl, taxon with a plastid; PlNo, taxon without plastids.

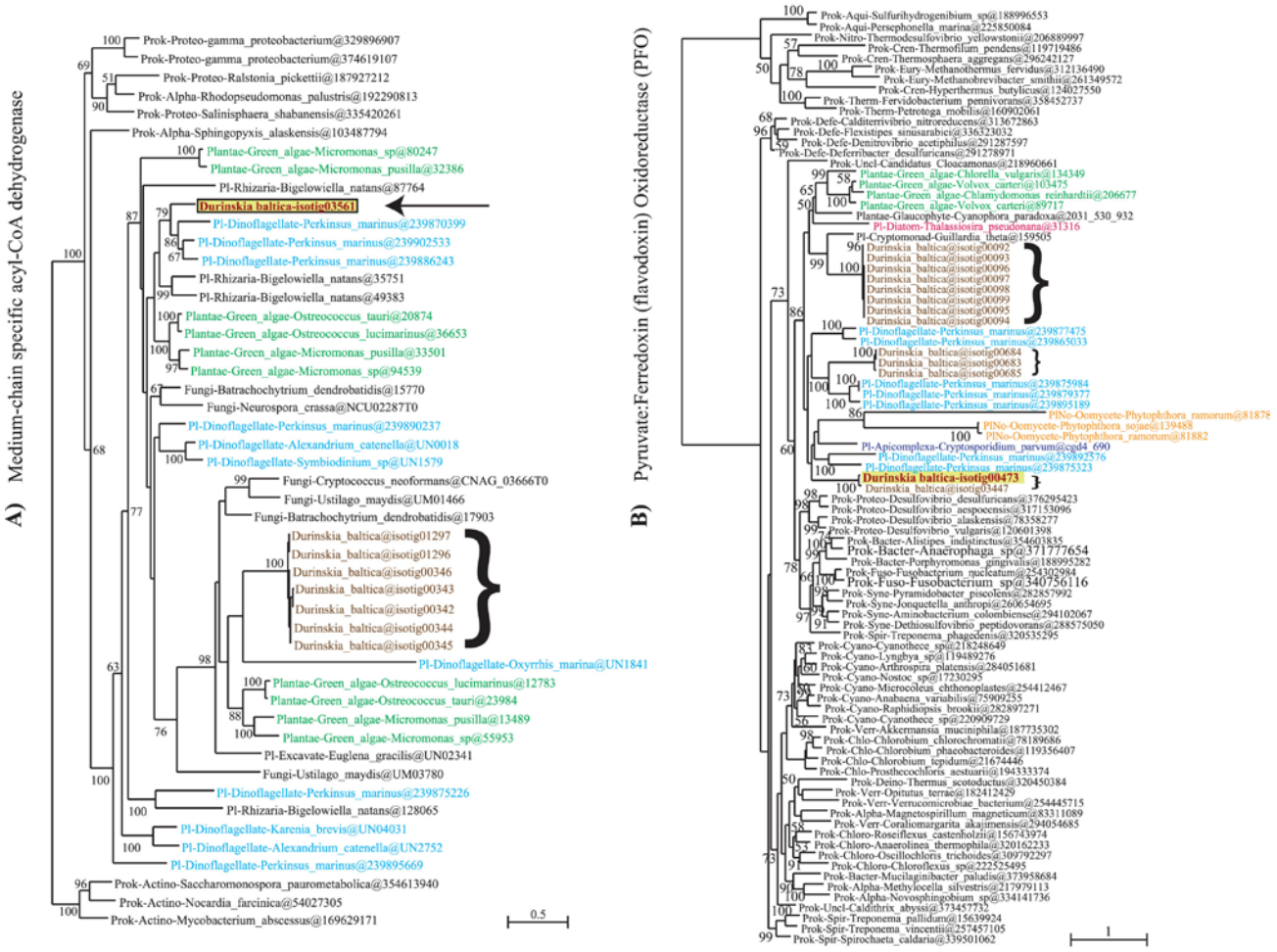

Figure 4.S10: The maximum likelihood trees for the host putative nuclear-encoded mitochondrial multi-copy proteins in *Durinskia baltica*.

A) Medium-chain specific acyl-CoA dehydrogenase, partial tree, B) Pyruvate:Ferrodoxin (flavodoxin) oxidoreductase (PFO). Numbers at the nodes indicate bootstrap support ≥ 50%. A yellow box highlights the position of *D. baltica*. Several taxa of interest are color-coded: Dinoflagellates in light blue; Apicomplexans and Ciliates in dark blue; Diatoms in scarlet; Stramenopiles in orange; Green algae and Plants in green; Red algae in red; All other taxa in black. Pl, taxon with a plastid; PlNo, taxon without plastids.

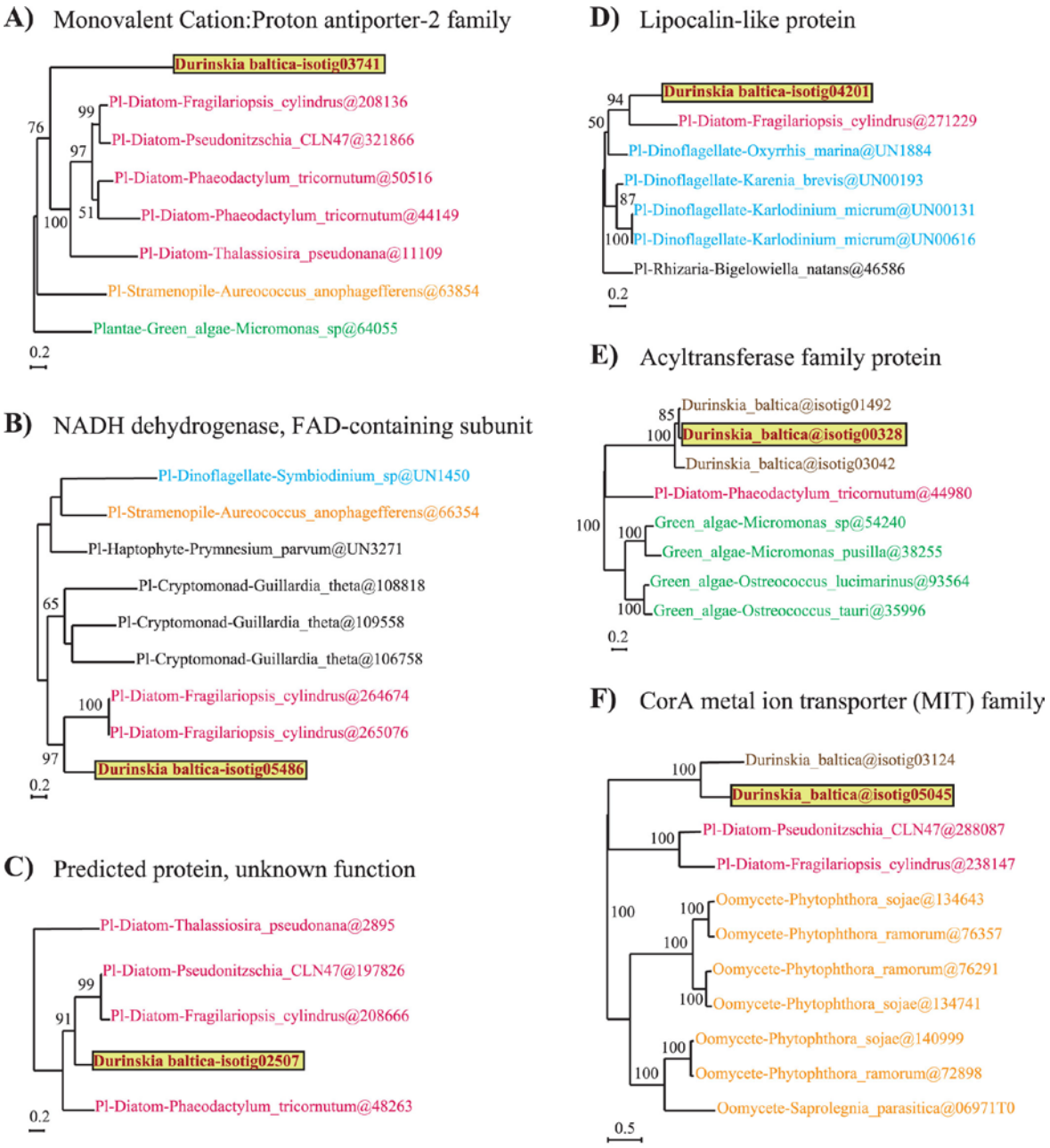

Figure 4.S11: The maximum likelihood trees with a limited number of taxa showing a diatom affinity for the putative nuclear-encoded proteins in *Durinskia baltica*.

A) Monovalent cation:proton antiporter-2 family, B) NADH dehydrogenase, FAD containing subunit, C) Predicted protein, unknown function, D) Lipocalin-like protein, E) Acyltransferase family protein, F) CorA metal ion transporter. Numbers at the nodes indicate bootstrap support ≥ 50%. A yellow box highlights the position of *D. baltica*. Several taxa of interest are color-coded: Dinoflagellates in light blue; Apicomplexans and Ciliates in dark blue; Diatoms in scarlet; Stramenopiles in orange; Green algae and Plants in green; Red algae in red; All other taxa in black. Pl, taxon with a plastid; PlNo, taxon without plastids.

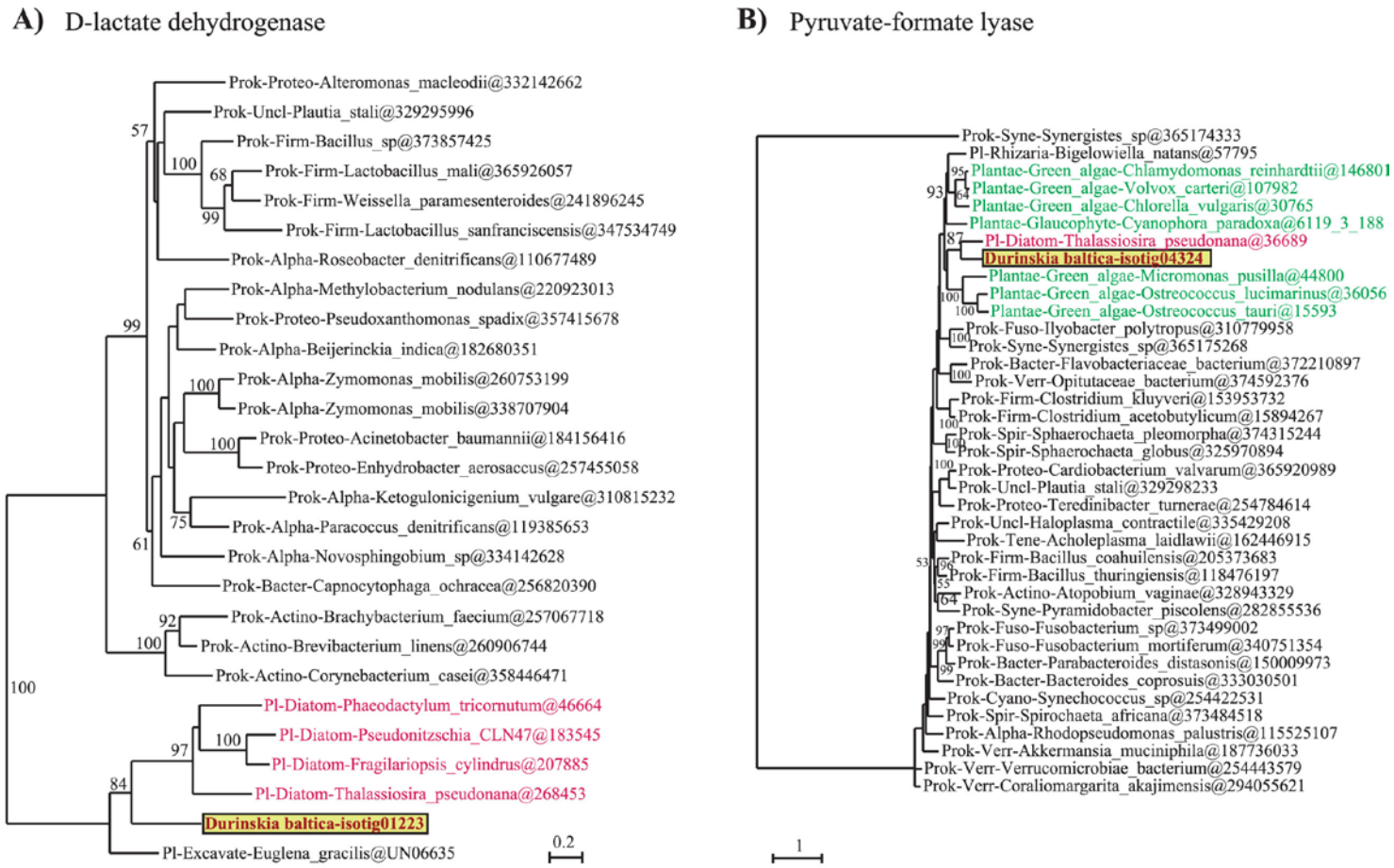

Figure 4.S12: The maximum likelihood trees showing a diatom affinity for the putative nuclear-encoded proteins in *Durinskia baltica*.

A) D-lactate dehydrogenase, B) Pyruvate-formate lyase. Numbers at the nodes indicate bootstrap support ≥ 50%. A yellow box highlights the position of *D. baltica*. Several taxa of interest are color-coded: Dinoflagellates in light blue; Apicomplexans and Ciliates in dark blue; Diatoms in scarlet; Stramenopiles in orange; Green algae and Plants in green; Red algae in red; All other taxa in black. Pl, taxon with a plastid; PlNo, taxon without plastids.

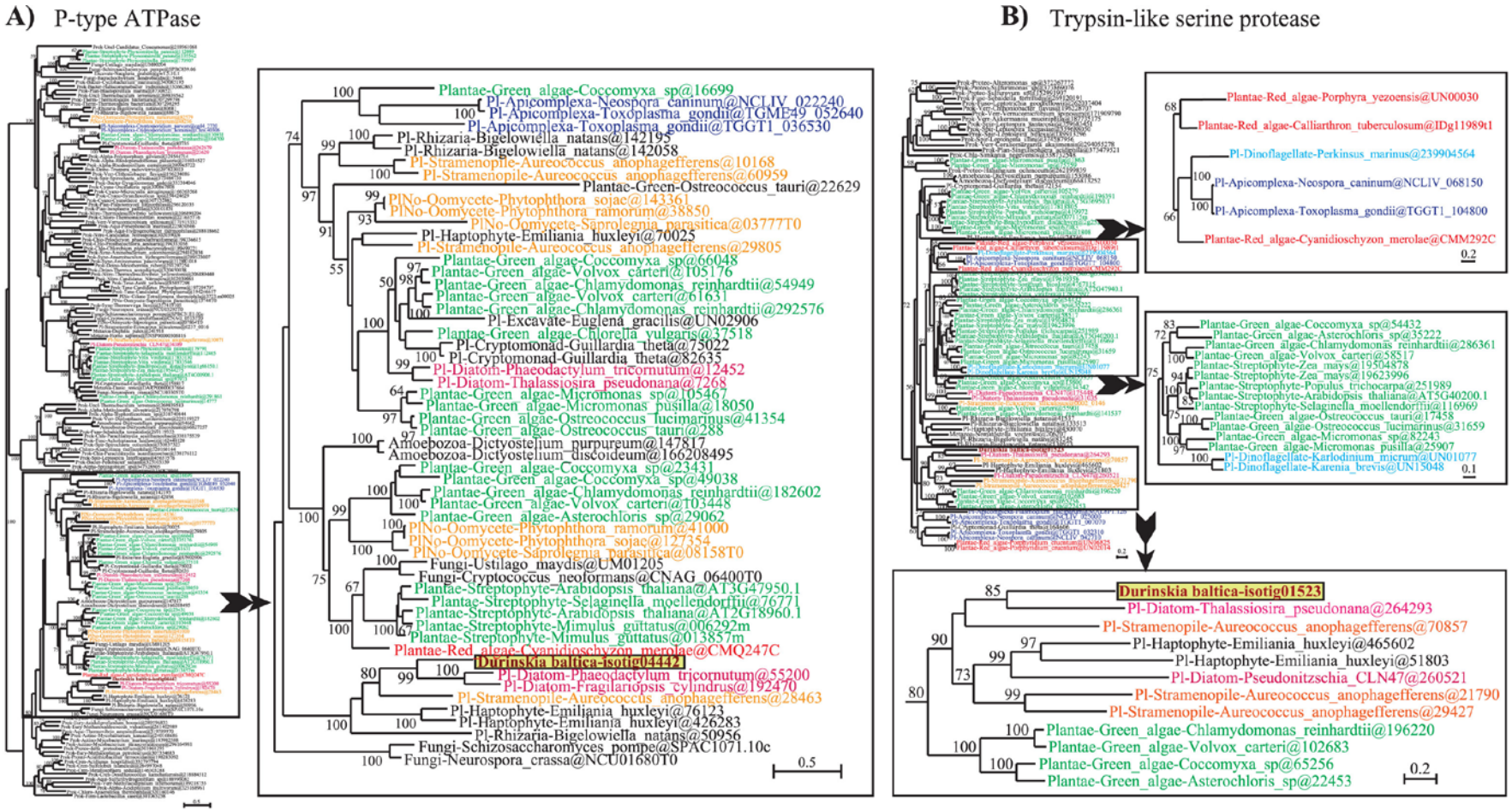

Figure 4.S13: The maximum likelihood trees showing a diatom origin or affinity for the putative nuclear-encoded proteins in *Durinskia baltica*.

A) P-type ATPase, left: complete tree, right: partial tree, B) Trypsin-like serine protease, top left: complete tree, arrows point at partial sections of the tree. Numbers at the nodes indicate bootstrap support ≥ 50%. A yellow box highlights the position of *D. baltica*. Several taxa of interest are color-coded: Dinoflagellates in light blue; Apicomplexans and Ciliates in dark blue; Diatoms in scarlet; Stramenopiles in orange; Green algae and Plants in green; Red algae in red; All other taxa in black. Pl, taxon with a plastid; PlNo, taxon without plastids.

Table 4.S1: The GC content of the *D. baltica* nuclear-encoded plastid cDNAs

Protein	GC %	Seq ID
APXT	48.2	isotig00853
FCP	52.1	isotig02599
ADK-UBOX Fusion	64.2	isotig01071
APX	63.3	isotig02367
CA	68.1	isotig03896
OASL	66.4	isotig03490
SufC	66.8	isotig04489
CASTOR	61.8	isotig01020

Abbreviations: ADK-UBOX Fusion, chloroplast adenylate kinase and U-box domain containing protein; APX, chloroplast ascorbate peroxidase; APXT, chloroplast thylakoid bound ascorbate peroxidase; CA, chloroplast carbonic anhydrase ; CASTOR, chloroplastic ion channel CASTOR; FCP, fucoxanthin chlorophyll a/c binding protein; SufC, FeS assembly ATPase SufC; OASL, chloroplast O-acetyl-serine lyase.

Table 4.S2: The GC content of the *D. baltica* diatom-derived candidate cDNAs compared to that of their orthologues in other diatoms

Db Seq ID	GC%	Pt (gi)	GC%	Tp (gi)	GC%	Fc (jgi)	GC%	To (gi)	GC%
isotig02599	52.1	219117949	53.6	224013063	57.0				
isotig02507	46.0	219124587	50.5	223996744	46.6				
isotig00853	48.2	219122836	51.2	224003374	53.4				
isotig02229	45.1	219123922	56.4	224006214	49.1				
isotig04324	65.7			224012632	48.7				
isotig01523	66.9			224010520	47.5				
isotig03741	60.1	219130627	51.0	224014501	48.5				
isotig04442	62.0	219119116	47.9	224004641	48.6				
isotig05486	58.0					264674	40.6		
isotig03474	64.8							397600967	54.5
isotig04201	62.8					271229	38.8		
isotig05045	66.6					238147	39.6		
isotig00328	60.0	219116143	49.0						
isotig00757	59.1	219111226	48.0	223999502	48.9			397627398	49.7
isotig01223	62.9	219121196	53.9	223998195	46.4				

Abbreviations: Db, *Durinskia baltica*; Pt, *Phaeodactylum tricornutum*; Tp, *Thalassiosira pseudonana*; Fc, *Fragilariopsis cylindrus*; To, *Thalassiosira pseudonana*; gi, NCBI gi accession number; jgi, JGI accession number.

Table 4.S3: The *D. baltica* sequence ids with an automatically assigned non-dinoflagellate non-diatom phylogenetic signal

Stramenopile	Haptophyte	Cryptophyte	Plantae	Ciliate	Apicomplexan	Metazoan	Fungi	Excavate	Bacteria
contig05983	isotig00588	isotig05221	contig01636	isotig03489	isotig01194	isotig04588	isotig01034	isotig03931	contig01594
isotig02918	isotig01463		contig04881	isotig04283	isotig01239	isotig05539	isotig03143	isotig04367	isotig01822
isotig04218	isotig01508		isotig01785	isotig04644	isotig01330		isotig04879		isotig02388
isotig05115	isotig01995		isotig01864		isotig01471				isotig03134
isotig05144	isotig02679		isotig01921		isotig01606				isotig03143
isotig01360	isotig03458		isotig02632		isotig01971				isotig04056
isotig01682	isotig03625		isotig03401		isotig02292				isotig04086
isotig02295	isotig03693		isotig03433		isotig02710				isotig04216
isotig03793	isotig04008		isotig03694		isotig03285				isotig04327
isotig04435	isotig04485		isotig04023		isotig03760				isotig04351
isotig03869	isotig02912		isotig04707		isotig03868				isotig04467
			isotig04984		isotig04593				isotig05109
			isotig05314		isotig04960				isotig05328
			isotig03526		isotig05158				isotig05511
			isotig04214		isotig05248				contig01594
					isotig03614				isotig01822
					isotig04221				isotig02388
									isotig03984
									isotig04005
									isotig04656
									isotig04739
									isotig05042

The red font marks the sequences that were assigned a non-dinoflagellate signal after manual inspection.

Table 4.S4: The list of taxa included in the phylogenetic analyses

	Group	Taxon	Plastid	Data type
	Prokaryote	*Actinobacteria*	no	Genomes
Alveolate	Dinoflagellate	*Alexandrium catenella*	yes	ESTs
Alveolate	Dinoflagellate	*Alexandrium minutum*	yes	ESTs
Alveolate	Dinoflagellate	*Alexandrium ostenfeldii*	yes	ESTs
Alveolate	Dinoflagellate	*Alexandrium tamarense*	yes	ESTs
	Prokaryote	*Alphaproteobacteria*	no	Genomes
Alveolate	Dinoflagellate	*Amphidinium carterae*	yes	ESTs
	Prokaryote	*Aquificae*	no	Genomes
	Streptophyte	*Arabidopsis thaliana*	yes	Genomes
	Green Alga	*Asterochloris sp*	yes	Genomes
Stramenopile	Pelagophyte	*Aureococcus anophageferrens*	yes	Genomes
	Prokaryote	*Bacteroides fragilis*	no	Genomes
	Prokaryote	*Batrachochytrium dendrobatidis*	no	Genomes
	Rhizaria	*Bigelowiella natans*	yes	Genomes
	Streptophyte	*Brachypodium distachyon*	yes	Genomes
	Haptophyte	*Calcidiscus leptoporus*	yes	ESTs
	Red Alga	*Calliarthron tuberculosum*	yes	ESTs
	Prokaryote	*Chlamydiae*	no	Genomes
	Green Alga	*Chlamydomonas reinhardtii*	yes	Genomes
	Green Alga	*Chlorella vulgaris*	yes	Genomes
	Prokaryote	*Chlorobi*	no	Genomes
	Prokaryote	*Chloroflexi*	no	Genomes
	Red Alga	*Chondrus crispus*	yes	ESTs
Alveolate	Apicomplexa	*Chromera velia*	no	ESTs
	Haptophyte	*Coccolithus braarudii*	yes	ESTs
	Green Alga	*Coccomyxa sp*	yes	Genomes
	Prokaryote	*Crenarchaeota*	no	Genomes
	Fungi	*Cryptococcus neoformans*	no	Genomes
Alveolate	Apicomplexa	*Cryptosporidium hominis*	no	Genomes
Alveolate	Apicomplexa	*Cryptosporidium parvum*	no	Genomes
	Red Alga	*Cyanidioschyzon merolae*	yes	Genomes
	Prokaryote	*Cyanobacteria*	no	Genomes
	Glaucophyte	*Cyanophora paradoxa*	yes	ESTs
	Metazoa	*Danio rerio*	no	Genomes
	Metazoa	*Daphnia pulex*	no	Genomes
	Prokaryote	*Deferribacteres*	no	Genomes
	Prokaryote	*Deinococcus*	no	Genomes

	Group	Taxon	Plastid	Data type
	Amoebozoa	*Dictyostelium discoideum*	no	Genomes
	Amoebozoa	*Dictyostelium purpureum*	no	Genomes
Stramenopile	Phaeophyte	*Ectocarpus siliculosus*	yes	Genomes
	Haptophyte	*Emiliania huxleyi*	yes	Genomes
	Red Alga	*Eucheuma denticulatum*	yes	ESTs
	Excavate	*Euglena gracilis*	yes	ESTs
	Excavate	*Euglena longa*	yes	ESTs
	Excavate	*Euglena mutabilis*	yes	ESTs
	Prokaryote	*Euryarchaeota*	no	Genomes
	Prokaryote	*Firmicutes*	no	Genomes
Stramenopile	Diatom	*Fragilariopsis cylindrus*	yes	Genomes
	Prokaryote	*Fusobacteria*	no	Genomes
	Red Alga	*Galdieria sulphuraria*	yes	ESTs
	Red Alga	*Furcellaria lumbricalis*	yes	ESTs
	Metazoa	*Gallus gallus*	no	Genomes
	Glaucophyte	*Glaucocystis nostochinearum*	yes	ESTs
	Red Alga	*Gracilaria sp*	yes	ESTs
	Red Alga	*Griffithsia okiensis*	yes	ESTs
	Cryptomonad	*Guillardia theta*	yes	Genomes
Alveolate	Dinoflagellate	*Heterocapsa triquetra*	yes	ESTs
	Metazoa	*Homo sapiens*	no	Genomes
Alveolate	Ciliate	*Ichthyophthirius multifiliis*	yes	ESTs
Alveolate	Haptophyte	*Isochrysis galbana*	yes	ESTs
Alveolate	Dinoflagellate	*Karenia brevis*	yes	ESTs
Alveolate	Dinoflagellate	*Karlodinium micrum*	yes	ESTs
	Fungi	*Laccaria bicolor*	no	Genomes
Alveolate	Dinoflagellate	*Lingulodinium polyeydrum*	no	ESTs
	Metazoa	*Lottia gigantea*	no	Genomes
	Green Alga	*Micromonas pusilla*	yes	Genomes
	Green Alga	*Micromonas sp*	yes	Genomes
	Streptophyte	*Mimulus guttatus*	yes	Genomes
	Excavate	*Naegleria gruberi*	no	Genomes
	Metazoa	*Nematostella vectensis*	no	Genomes
Alveolate	Apicomplexa	*Neospora caninum*	yes	Genomes
	Fungi	*Neurospora crassa*	no	Genomes
	Prokaryote	*Nitorospirae*	no	Genomes
	Streptophyte	*Oryza sativa*	yes	Genomes
	Green Alga	*Ostreococcus lucimarinus*	yes	Genomes

	Group	Taxon	Plastid	Data type
	Green Alga	*Ostreococcus tauri*	yes	Genomes
Alveolate	Dinoflagellate	*Oxyrrhis marina*	NA	ESTs
Alveolate	Ciliate	*Paramecium tetraurelia*	no	Genomes
	Haptophyte	*Pavlova lutheri*	yes	ESTs
Alveolate	Dinoflagellate	*Perkinsus marinus*	yes	Genomes
Stramenopile	Diatom	*Phaeodactylum tricornutum*	yes	Genomes
	Streptophyte	*Physcomitrella patens*	yes	Genomes
Stramenopile	Oomycete	*Phytophthora ramorum*	no	Genomes
Stramenopile	Oomycete	*Phytophthora sojae*	no	Genomes
	Prokaryote	*Planctomycetes*	no	Genomes
Alveolate	Apicomplexa	*Plasmodium berghei*	yes	Genomes
Alveolate	Apicomplexa	*Plasmodium chabaudi*	yes	Genomes
Alveolate	Apicomplexa	*Plasmodium falciparum*	yes	Genomes
	Streptophyte	*Populus trichocarpa*	yes	Genomes
	Red Alga	*Porphyra haitanensis*	yes	ESTs
	Red Alga	*Porphyra yezoensis*	yes	ESTs
	Red Alga	*Porphyridium cruentum*	yes	ESTs
	Prokaryote	*Proteobacteria-nonalpha*	yes	Genomes
	Haptophyte	*Prymnesium parvum*	yes	ESTs
Stramenopile	Dictyochophyte	*Pseudochattonella farcimen*	yes	ESTs
Stramenopile	Diatom	*Pseudonitzschia multiseries CLN47*	yes	ESTs
	Katablepharid	*Roombia truncata*	no	ESTs
Stramenopile	Oomycete	*Saprolegnia parasitica*	no	Genomes
	Fungi	*Schizosaccharomyces pompe*	no	Genomes
	Streptophyte	*Selaginella moellendorffii*	yes	Genomes
	Streptophyte	*Sorghum bicolor*	yes	Genomes
	Prokaryote	*Spirochaetes*	no	Genomes
Alveolate	Dinoflagellate	*Symbiodinium sp*	yes	ESTs
	Prokaryote	*Synergistetes*	yes	Genomes
	Prokaryote	*Tenericutes*	no	Genomes
Alveolate	Ciliate	*Tetrahymena thermophila*	no	Genomes
Stramenopile	Diatom	*Thalassiosira pseudonana*	yes	Genomes
	Prokaryote	*Thermotogae*	no	Genomes
	Prokaryote	*Thumarchaeota*	no	Genomes
Alveolate	Apicomplexa	*Toxoplasma gondii*	yes	Genomes
	Prokaryote	*Unclassifides*	no	Genomes
	Fungi	*Ustilago maydis*	no	Genomes
	Prokaryote	*Verrucomicrobia*	no	Genomes

Group	Taxon	Plastid	Data type
Streptophyte	*Vitis vinifera*	yes	Genomes
Green Alga	*Volvox carteri*	yes	Genomes
Streptophyte	*Zea mays*	yes	Genomes