Coevolution of plastid genomes and transcript processing pathways in photosynthetic alveolates

Richard George Dorrell; King's College

Initial submission dated 31/05/2014

Corrected submission dated 08/08/2014

This dissertation is submitted for the degree of Doctor of Philosophy

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

None of the material in this dissertation has been previously submitted for any other academic qualification.

Some of the material within this dissertation has previously been published, in the following papers:

Barbrook AC, Dorrell RG, Burrows J, Plenderleith LJ, Nisbet RER, Howe CJ. 2012. Polyuridylylation and processing of transcripts from multiple gene minicircles in chloroplasts of the dinoflagellate *Amphidinium carterae*. *Plant Molecular Biology* **79**, 347-357.

Dorrell RG, Howe CJ. 2012. What makes a chloroplast? Reconstructing the establishment of photosynthetic symbioses. *Journal of Cell Science* **125**, 1865-1875.

Dorrell RG, Howe CJ. 2012. Functional remodeling of RNA processing in replacement chloroplasts by pathways retained from their predecessors. *Proceedings of the National Academy of Sciences USA*. 109: 18879-18884

Dorrell RG, Butterfield ER, Nisbet RER, Howe CJ. 2013. Evolution: unveiling early alveolates. *Current Biology* **23**, 1093-1096.

Dorrell RG, Drew J, Nisbet RE, Howe CJ. 2014. Evolution of chloroplast transcript processing in *Plasmodium* and its chromerid algal relatives. *PLoS Genetics* **10**, e1004008.

Richardson E*, Dorrell RG* (joint first authors), Howe CJ. 2014. Genome-wide transcript profiling reveals the coevolution of chloroplast gene sequences and transcript processing pathways in the fucoxanthin dinoflagellate *Karlodinium veneficum*. *Molecular Biology and Evolution,* **in press.**

All of the material from these publications that is included in this thesis was my own work, and was written by me.

This dissertation is 51903 words length, excluding preface, abstract, figure legends, bibliography and appendices.

Thanks

Thank you to Chris, for being such a generous and patient supervisor, and to members of the Howe Group for providing a warm and intellectually supportive research environment over the last four years. Particular thanks are due to Ellen Nisbet, for providing critical feedback on a great deal of the material in this thesis; to David Lea-Smith, Joanna McKenzie, Adrian Barbrook, and Davy Kurniawan, for training me in experimental techniques used in my PhD; and to Erin Butterfield, for reminding me to view everything with perspective. Thanks also to Martin Embley and to Ross Waller for a most enlightening viva examination, which I feel has made a very positive impact on this thesis.

Thank you to the BBSRC for providing financial support for my work, and the British Phycological Society, British Society for Protist Biology, International Society of Protistologists, and King's College, Cambridge for enabling me to attend some very intellectually stimulating conferences.

Thank you to my students- each and every one- for having provided me with a great deal to think about, challenging me on the assumptions that I make, and reminding me on a weekly basis of why I got into this job in the first place. Particular thanks to Beth Richardson, George Hinksman, and James Drew, who worked so hard and (I think) grew so much as research students under my supervision.

Thank you to my parents, Mary and Peter Dorrell, for teaching me early on and continuing to teach me today to be inquisitive, limit myself only in the capacity of my imagination, and strive to be the kind of person that I want to see in the world. Finally, thank you to Anil: my love; the most beautiful person I know inside and out; for keeping me in one piece over the months spent writing this.

OK, let's go!

Abstract

Following their endosymbiotic uptake, plastids undergo profound changes to genome content and to their associated biochemistry. I have investigated how evolutionary transitions in plastid genomes may impact on biochemical pathways associated with plastid gene expression, focusing on the highly unusual plastids found in one group of eukaryotes, the alveolates. The principal photosynthetic alveolate lineage is the dinoflagellate algae. Most dinoflagellate species harbour unusual plastids derived from red algae. The genome of this plastid has been fragmented into small, plasmid-like elements termed “minicircles”. Transcripts of this genome receive a 3’ poly(U) tail and, in some species, undergo extensive sequence editing. Some dinoflagellates have replaced their original plastids with others, in a process termed “serial endosymbiosis”. The major non-photosynthetic alveolates are the apicomplexans, which include the malaria parasite *Plasmodium.* Apicomplexans are descended from free-living algae and possess a vestigial plastid, which originated through the same endosymbiosis as the ancestral red dinoflagellate plastid. This plastid has lost all genes involved in photosynthesis and does not possess a poly(U) tail addition pathway.

I have investigated the consequences of the fragmentation of the red algal dinoflagellate plastid genome on plastid transcription. I have characterised non-coding transcripts in plastids of the dinoflagellate *Amphidinium carterae*, including the first evidence for antisense transcripts in an algal plastid. Antisense transcripts in dinoflagellate plastids do not receive poly(U) tails, suggesting that poly(U) tail addition may play a role in strand discrimination during transcript processing.

I have additionally characterised transcript processing in dinoflagellate plastids that were acquired through serial endosymbiosis. I have shown that poly(U) tail addition and editing occur in the haptophyte-derived serial endosymbionts of the fucoxanthin-containing dinoflagellates *Karenia mikimotoi* and *Karlodinium veneficum*. This is the first evidence that plastids acquired through serial endosymbiosis may be supported by pathways retained from previous symbioses. Transcript editing constrains the phenotypic consequences of divergent mutations in fucoxanthin plastid genomes, whereas poly(U) tail addition plays a central role in recognising and processing translationally functional fucoxanthin plastid mRNAs. I have additionally shown that certain genes within fucoxanthin plastids are located on minicircles. This demonstrates convergent evolution in the organisation of the fucoxanthin and red algal dinoflagellate plastid genomes since their endosymbiotic acquisition.

Finally, I have investigated transcript processing in the algae *Chromera velia* and *Vitrella brassicaformis*. These species are closely related to apicomplexans but are still

photosynthetic and apply poly(U) tails to plastid transcripts, as with dinoflagellates. I have shown that poly(U) tails in these species are preferentially associated with translationally functional mRNAs of photosynthesis genes. This is the first plastid transcript processing pathway documented to target a specific functional gene category. Poly(U) tail addition may direct transcript cleavage and allow photosynthesis gene transcripts to accumulate to high levels. The loss of this pathway from ancestors of apicomplexans may have contributed to their transition from photosynthesis to parasitism.

Contents

Chapter One- Thesis Introduction

The origins of photosynthesis in the eukaryotes

Eukaryotic life is believed to have been present on earth for nearly two billion years, and over this time it has had fundamental effects on planetary ecosystems and geochemistry (Embley and Martin, 2006; Parfrey *et al.*, 2011). Eukaryotes originated from the symbiotic integration of at least two distantly related prokaryotes, which occurred once, generating the common ancestor of all eukaryotic species (Cox *et al.*, 2008; Walker *et al.*, 2011). The phylogenetic identity of the prokaryotic lineages involved, and the exact evolutionary events that gave rise to the first eukaryotes remain debated (Cox *et al.*, 2008; Embley and Martin, 2006). However, it is widely agreed that the most recent common ancestor of all eukaryotes was complex, possessing many cellular structures, including nuclei, mitochondria, an endomembrane system, and a cytoskeleton, which are found in modern-day descendants (Embley and Martin, 2006; Walker *et al.*, 2011).

Since their origin, specific eukaryotic lineages have undergone profound transitions in lifestyle. Many of these transitions have involved dramatic changes to the genomes, and to the cellular organisation of these lineages. Some eukaryotes, for example, have secondarily lost the capacity for aerobic respiration (Burki *et al.*, 2014; Embley and Martin, 2006). The mitochondria of these lineages have been converted into alternative organelles (e.g. hydrogenosomes, mitosomes) that allow the generation of ATP under anaerobic conditions, or do not synthesise ATP at all (Hjort *et al.*, 2010; Lindmark and Muller, 1973). Multicellularity has evolved in at least seven phylogenetically distinct eukaryotic lineages (Brown *et al.*, 2012). In multicellular eukaryotes, specific nuclear gene families, particularly those associated with cell signalling and differentiation, have undergone dramatic diversifications, which have occurred concurrent to the divergence of these lineages from single-celled relatives (Cock *et al.*, 2010; de Mendoza *et al.*, 2013).

The most fundamental of these evolutionary transitions, in terms of its consequences for planetary ecology and climate, is the transition of some eukaryotes from depending on the phagocytotic consumption of other organisms for the acquisition of organic carbon, to the direct fixation of inorganic carbon through photosynthesis (Dorrell and Smith, 2011; Igamberdiev and Lea, 2006). Photosynthesis was acquired within the eukaryotes through the endosymbiotic internalisation and domestication of free-living cyanobacteria as plastids, also termed “chloroplasts” (Fig. 1.1) (Howe *et al.*, 2008a; Sagan, 1967). Permanent plastids provide many beneficial functions for photosynthetic eukaryotes, including carbon fixation, the biosynthesis of specific amino acids (e.g. aromatic amino acids, and lysine) and phenolic

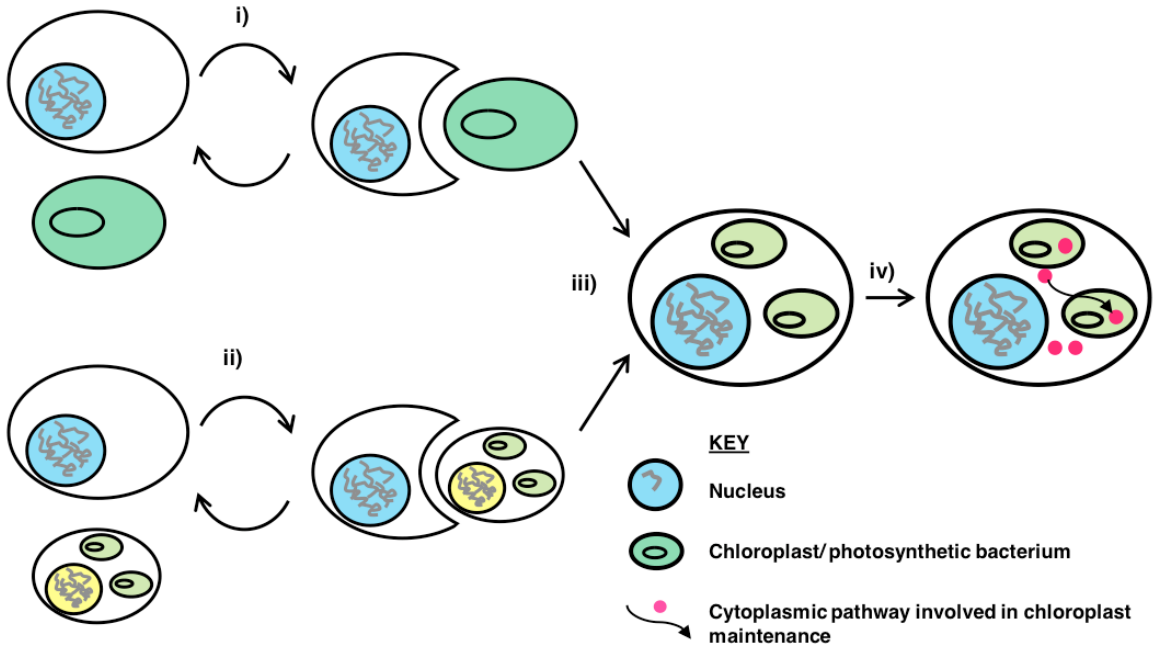


Fig. 1.1: Principles of plastid endosymbiosis.

This diagram shows the fundamental events that occur in conventional (i.e. non-serial) endosymbiosis events. A non-photosynthetic eukaryote consumes a free-living photosynthetic prokaryote in the case of primary endosymbiosis **(i)**, or a eukaryote in the case of secondary or tertiary endosymbiosis **(ii)**, and converts it into an intracellular organelle. This may not occur immediately, and there may be multiple cycles of uptake and loss before a permanent plastid is established **(iii)**. As part of this process, pathways evolve within the host that facilitate the long-term retention of the symbiont **(iv)**.

compounds, and the dissipation of excess mitochondrial reducing potential (Herrmann and Weaver, 1999; Hoefnagel *et al.*, 1998).

In return, extant plastids are supported by other cellular organelles (Fig. 1.1). Many of the proteins essential for plastid function, including the vast majority of proteins involved in expression of the plastid genome, are not expressed from genes located on the plastid genome, but are instead expressed from nuclear genes, and imported into plastids from the cytoplasm (Barkan, 2011; Suzuki and Miyagishima, 2010). The mitochondria may also play important roles in supporting plastids, for example by providing specific intermediates for particular plastid metabolic pathways, and eliminating excess electron potential and reducing intermediates generated through photosynthesis (Prihoda *et al.*, 2012). Thus, the biology of the eukaryotic host has played a fundamental role in shaping the biology of plastid lineages.

Taxonomic distribution of plastid lineages

Plastids have been acquired by multiple eukaryote lineages. Almost all documented plastids originated through the endosymbiosis of a β-cyanobacterium in an ancestor of the archaeplastid supergroup, containing red algae, glaucophytes, and green algae and plants (Figs. 1.1, 1.2). An independent primary plastid endosymbiosis, involving an α-cyanobacterium, is understood to have occurred in the rhizarian amoeba *Paulinella chromatophora* (Marin *et al.*, 2005). A further cyanobacterial endosymbiosis has been identified in the diatom *Rhopalodia gibba*, although as this species also contains plastids of conventional endosymbiotic origin, it is not clear whether the cyanobacterial endosymbionts function as plastids (Kneip *et al.*, 2008; Prechtl *et al.*, 2004).

Other major photosynthetic eukaryotic lineages (e.g. diatoms, haptophytes) have arisen subsequently through similar endosymbiotic events. In these lineages, the host has taken up a free-living alga from within the archaeplastid clade, which itself contained a plastid of cyanobacterial origin, in a process termed secondary endosymbiosis (Figs. 1.1, 1.2) (Dorrell and Smith, 2011; Walker *et al.*, 2011). Some lineages, such as the euglenids and chlorarachniophytes, possess plastids of green algal origin, while many ecologically prominent groups of algae, including diatoms, haptophytes and some dinoflagellates, possess plastids derived from red algae. A few species, within the dinoflagellates, are known to have acquired plastids from diatoms or haptophytes, thus possessing tertiary endosymbionts (Fig. 1.2). In many cases, the exact progression of endosymbiotic events that gave rise to specific plastid lineages remains controversial, for example due to difficulty in assigning precise phylogenetic origins for several major plastid lineages, and due to conflicting results in nuclear and plastid gene phylogenies regarding the taxonomic relationships between photosynthetic eukaryotes (Baurain *et al.*, 2010; Dorrell and Smith, 2011; Shalchian-Tabrizi *et al.*, 2006).

The majority of plastid lineages are believed to have originated through the endosymbiotic acquisition of a photosynthetic symbiont by a non-photosynthetic host lineage, which did not previously possess plastids (Dorrell and Smith, 2011; Sagan, 1967). In some cases, however, a previously photosynthetic eukaryote has replaced its original plastid with one of a new phylogenetic derivation. This process is termed “serial endosymbiosis”. Serial endosymbiosis accounts for the green algal-, haptophyte- and diatom-derived plastids in dinoflagellate algae (Fig. 1.2) (Dorrell and Smith, 2011). In these lineages, the ancestral plastid derived from red algae has been lost, and a plastid of new phylogenetic origin, with distinct differences in genome content and organisation, has been acquired in its place (Gabrielsen *et al.*, 2011; Imanian *et al.*, 2012; Matsumoto *et al.*, 2011a). In addition, genes

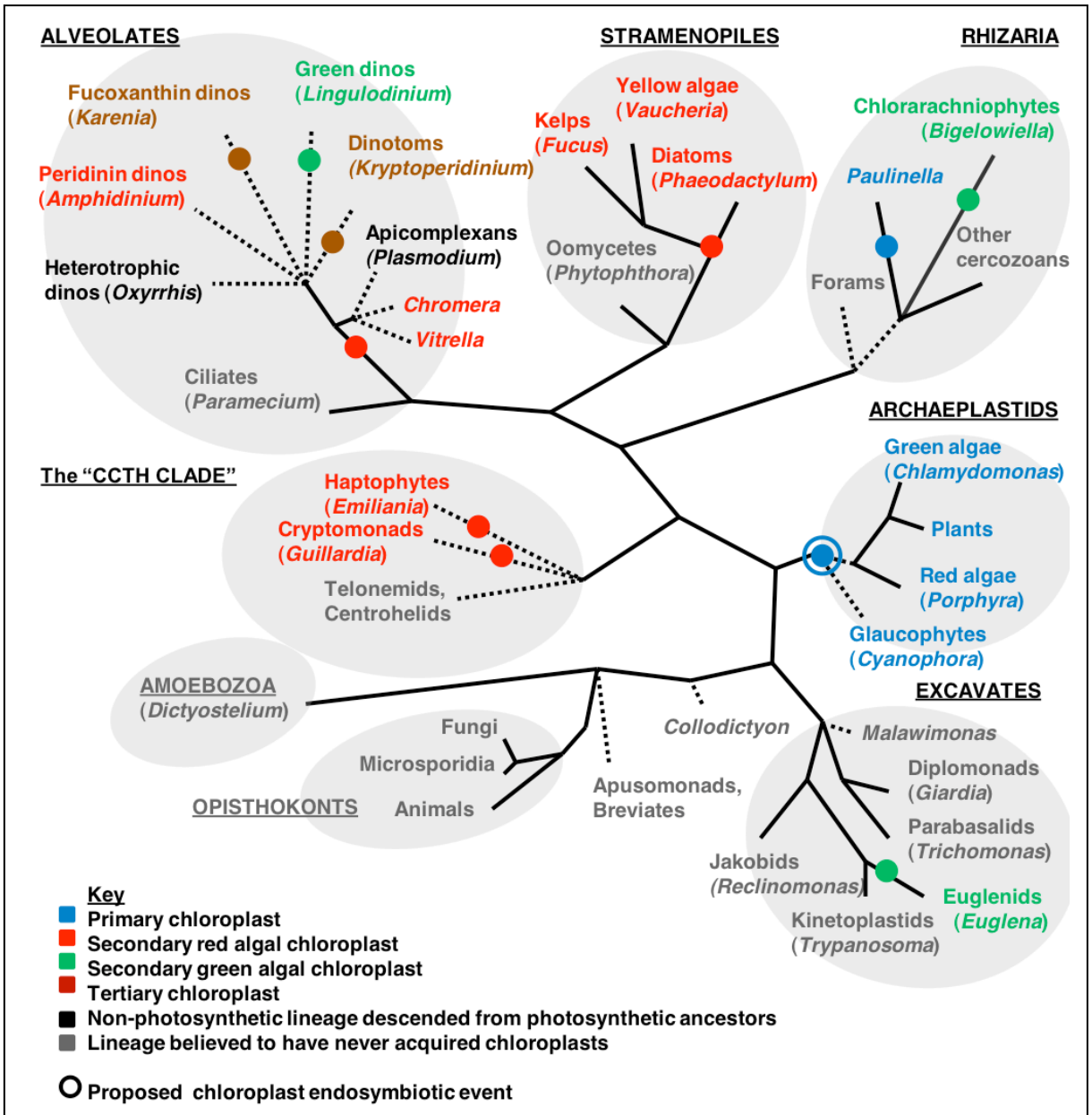


Fig. 1.2: The distribution of plastids across the eukaryotes.

This tree shows the distribution and phylogenetic origins of permanent plastids across the eukaryotes. The phylogenetic relationships shown are adapted from Dorrell and Smith, 2011. For clarity, some phylogenetic relationships have been condensed; branch positions that are currently under debate are shown by dotted lines. Lineages are grouped into higher-order phylogenetic affiliations, as demarcated by a shaded ellipse, which is named in block capitals. Names in parentheses refer to type species for particular lineages. Each lineage is coloured according to the type of plastids present. Proposed endosymbiotic events that have given rise to individual plastid lineages are shown as coloured circles. The primary endosymbiosis that gave rise to the common ancestor of all archaeplastids, and consequently all documented secondary and tertiary plastids, is labelled with a larger, annulated circle.

derived from the nucleus of the incoming symbiont, which may encode proteins involved in plastid maintenance, have also been acquired by the nuclear genome of the host (Burki *et al.*, 2014; Ishida and Green, 2002; Minge *et al.*, 2010).

Identifying model systems for studying plastid evolution

Understanding the processes that underpin plastid evolution may provide important insights into eukaryotic cell biology. For example, investigating what pathways have evolved in specific plastid lineages since their endosymbiotic acquisition may provide valuable insights into why certain photosynthetic eukaryotes have risen to positions of ecological prominence (Dorrell and Smith, 2011; Prihoda *et al.*, 2012). Similarly, identifying pathways encoded within the host that are important for plastid function might enable the creation of synthetic strategies to optimise and engineer plastid metabolism for agricultural and industrial use (Weber and Osteryoung, 2010).

A great deal is known about plastid biology in some photosynthetic eukaryotes, such as plants and diatoms. However, many of the established species for studying plastid biology are not ideal models for studying dynamic events in plastid evolution, as they are evolutionarily ancient. For example, the plastids found in the archaeplastid lineage may have been acquired upwards of one billion years ago, as has been inferred from molecular clock and fossil evidence (Fig. 1.3) (Parfrey *et al.*, 2011; Yoon *et al.*, 2004). Likewise, many of the most well-studied plastid lineages derived from secondary endosymbioses, for example those of diatoms and haptophytes, were acquired at least 500 million years before the present, and potentially much earlier (Fig. 1.3) (Medlin, 2011; Parfrey *et al.*, 2011). Since their acquisition, many of these ancient plastid lineages have progressed to relatively stable evolutionary states, with relatively few recent changes to plastid genome content or function (Janouškovec *et al.*, 2013a; Palmer, 1987). Other plastid lineages, which have been acquired more recently or are more evolutionarily dynamic, may provide better models for investigating plastid evolution.

Evolutionary diversity of the alveolates

My research has focussed on the extremely diverse plastids found within one group of eukaryotes, the alveolates. Alveolates form a sister-group to the stramenopiles, which include diatom algae, kelps, and oomycete pathogens, but are distantly related to plants, animals and fungi (Fig. 1.1) (Dorrell and Smith, 2011; Walker *et al.*, 2011). Alveolates principally consist of three lineages, ciliates, dinoflagellates, and apicomplexans, with a number of additional groups, some of which have only recently been established in culture

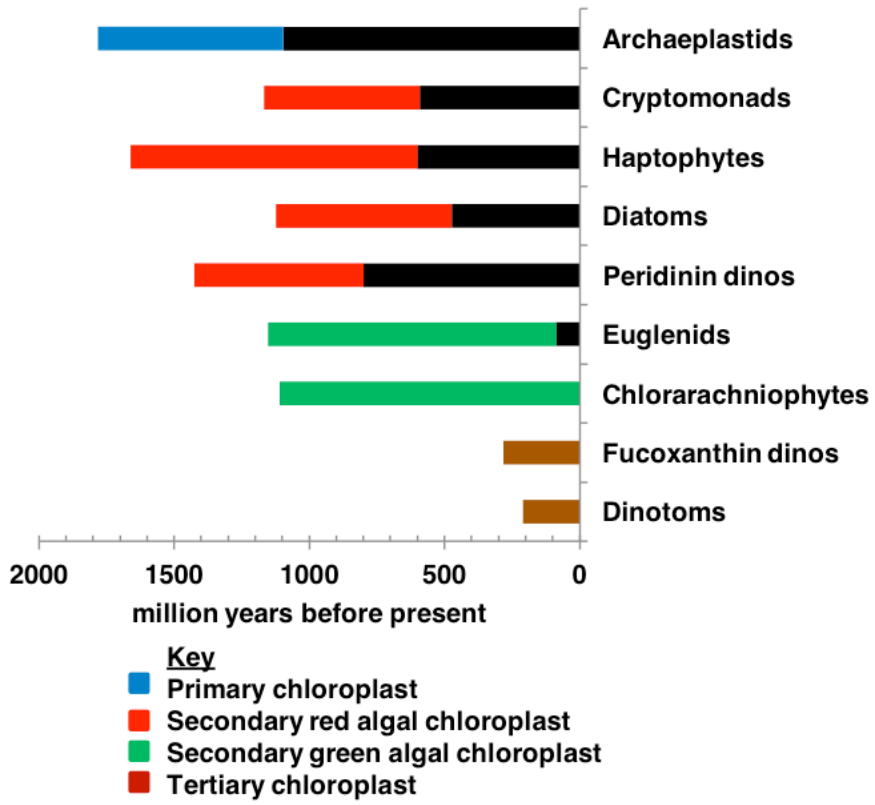


Fig. 1.3: A timeline of the origins of plastids across the eukaryotes.

This timeline shows the range of dates over plastid endosymbiotic events are believed to have occurred for different photosynthetic taxa. Each endosymbiotic event is shown as a coloured bar; the colour of each bar corresponds to the origin of the respective plastid lineage as shown in Fig. 1.2. The range of each bar extends from the earliest date at which the taxon may have diverged from its closest relatives that do not possess the given plastid lineage, to the latest date at which the taxon may have radiated; black regions correspond to dates after which the taxon is believed to have radiated, i.e. definitely after endosymbiosis has occurred. Divergence dates are taken from fossil calibrated molecular estimates as calculated in: Berney and Pawlowski 2006; and Parfrey *et al.*, 2011.

(e.g. *Colponema, Psammosa*), are inferred to be present from environmental sequence data but have not formally been identified (e.g. the apicomplexan-related ARL clades), or for which only limited molecular data are available (e.g. *Colpodella*) (Gile and Slamovits, 2014; Janouškovec *et al.*, 2012; Janouškovec *et al.*, 2013d; Okamoto and Keeling, 2014).

The ciliates are a group of protists that play important roles in microbial food webs. Ciliates diverge at the base of the alveolates (Fig. 1.2) (Walker *et al.*, 2011). Each ciliate cell contains multiple nuclei, in two different forms: a vegetative “micronucleus”, containing large numbers of gene fragments, and a “macronucleus”, containing translationally functional genes, which are generated via the rearrangement of the micronuclear genome (Eisen *et al.*, 2006). This unusual genome has provided insights into processes of broad biological significance: for example, early work on telomere maintenance used the ciliate *Tetrahymena* (Greider and Blackburn, 1985).

The dinoflagellates and apicomplexans group within the alveolates as a single clade (Fig. 1.2) (Janouškovec *et al.*, 2010). Despite the relatively close evolutionary relationships between them, the dinoflagellates and apicomplexans have adopted radically different lifestyle strategies. The dinoflagellates contain both heterotrophic and photosynthetic members, although it is clear from genetic and morphological evidence that the heterotrophic species are descended from photosynthetic ancestors (Matsuzaki *et al.,* 2008; Slamovits and Keeling, 2008). Some photosynthetic dinoflagellates are free-living and form an important contribution to oceanic primary production, while others form symbiotic associations, such as members of the genus *Symbiodinium*, the principal photosynthetic component of coral (Barbrook *et al.*, 2013). Other photosynthetic dinoflagellates additionally have detrimental effects on marine fauna, as the principal component of fish-killing “red tides” and other harmful algal blooms (Walker *et al.*, 2011). The apicomplexans, in contrast, are a largely parasitic lineage, and include *Plasmodium* (the causative agent of malaria) and *Toxoplasma* (toxoplasmosis), although some lineages are speculated to form commensal associations with their hosts (Saffo *et al.*, 2010; Walker *et al.*, 2011).

Many of the features of the cell biology of apicomplexans and dinoflagellates are extremely different from each other. The dinoflagellate nuclear genome contains large tandem arrays of intron-rich genes, and is extremely large in size (Shoguchi *et al.*, 2013). Dinoflagellate chromosomes are in a permanently condensed state, and appear to be predominantly packaged via a histone-independent strategy, using a protein of viral origin that has not been identified in any other eukaryotic lineage (Gornik *et al.*, 2012). In contrast, the apicomplexan nuclear genome, while extremely AT-rich, is conventionally organised, and contains few introns (Walker *et al.*, 2011). In other aspects of their cell biology, however, apicomplexans and dinoflagellates share many unusual and highly derived characteristics, which point to a shared ancestry. The mitochondrial genomes of both dinoflagellates and apicomplexans, for example, are highly reduced in content, only encoding three proteins and ribosomal RNA (Jackson *et al.*, 2007; Nash *et al.*, 2007). A similar degree of reduction is not found in the

mitochondrial genomes of other alveolates, or any other eukaryote lineage studied to date (Janouškovec *et al.*, 2013d).

Alveolates possess highly unusual plastids

Reflecting their extremely divergent life strategies, alveolates possess an extremely diversified range of plastids. A few (<20) genes of red algal origin have been found in some ciliate nuclear genomes, which have been interpreted as evidence of a historical plastid symbiosis, although it is possible that they arose from other lateral gene transfer events, or were misidentified in the phylogenies performed (Reyes-Prieto *et al.*, 2008; Stiller *et al.*, 2009). Other ciliates form transient associations with photosynthetic symbionts (Baker, 1994; Johnson, 2011). However, no ciliate species has yet been identified to contain permanent plastids, and they therefore will not be discussed in further detail here.

Extant plastids within the alveolates are confined to the dinoflagellates, apicomplexans, and their close relatives (Fig. 1.2). The majority of photosynthetic dinoflagellates contain plastids derived from red algae. These plastids contain the accessory light-harvesting carotenoid pigment peridinin (Figs. 2, 4) (Haxo *et al.*, 1976). This plastid is believed, from molecular and fossil evidence, to have originated approximately 500 million years ago, at roughly the same time as other secondary red algal plastids, such as those of stramenopiles (Fig. 1.3) (Parfrey *et al.*, 2011). However, relative to these lineages, peridinin dinoflagellate plastid genes evolve at a dramatically faster rate, forming extremely long branches on plastid phylogenies (Barbrook *et al.*, 2013; Inagaki *et al.*, 2004; Zhang *et al.*, 2000).

Several dinoflagellate species possess plastids acquired through alternative endosymbiotic events. Phylogenies of nuclear genes clearly show that peridinin-containing dinoflagellates are paraphyletic to the species that harbour alternative plastids (Bachvaroff *et al.*, 2014; Shalchian-Tabrizi *et al.*, 2006). Thus, the alternative plastid lineages must have arisen through the serial endosymbiotic replacement of the original peridinin lineage. Dinoflagellates that contain the accessory light harvesting pigment fucoxanthin, typified by the genera *Karenia* and *Karlodinium*, have plastids derived from haptophyte algae (Ishida and Green, 2002; Katoh *et al.*, 1989; Takishita *et al.,* 1999). Similarly, members of the genus *Lepidodinium* possess plastids derived from green algae (Matsumoto *et al.*, 2011a; Minge *et al.*, 2010); and the “dinotoms”, consisting of members of the Peridiniaceae (e.g. *Kryptoperidinium, Durinskia*) have undergone at least three distinct endosymbiosis events involving diatom plastids (Figs. 1.2, 1.4) (Horiguchi and Takano, 2006; Imanian *et al.*, 2012). Further endosymbiosis events have been postulated to occur in other dinoflagellate species (Escalera *et al.*, 2011; Garcia-Cuetos *et al.,* 2008). These serial endosymbiosis events must

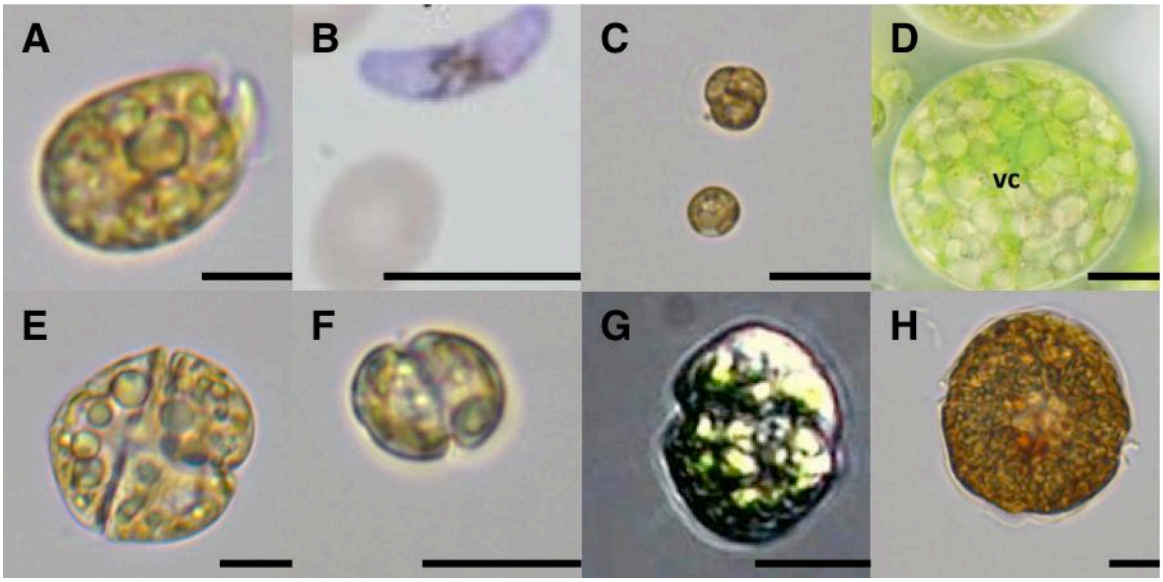


Fig. 1.4. Plastid diversity in photosynthetic alveolates.

This figure shows a representative array of alveolate species, harbouring different types of plastids. **Panels A-D**: lineages with red algal plastids. **A**: *Amphidinium carterae* (photosynthetic dinoflagellate); **B**: *Plasmodium falciparum* (non-photosynthetic apicomplexan); **C**: *Chromera velia* (chromerid); and **D**: *Vitrella brassicaformis* vegetative cell (labelled vc; chromerid). **Panels E-H**: dinoflagellates harbouring plastids of serial endosymbiotic origin. **E**: *Karenia mikimotoi* (fucoxanthin dinoflagellate with haptophyte plastids); **F**: *Karlodinium veneficum* (fucoxanthin dinoflagellate); **G**: *Lepidodinium chlorophorum* (dinoflagellate with green algal plastids); and **H**: *Kryptoperidinium foliaceum* ("dinotom", dinoflagellate with diatom plastids). Scale bars on each image are 10 μm long. Images A, C, E, F and H were taken by the author. Images B is reproduced from Encyclopaedia of Life (www.eol.org) and Image G is reproduced from Planktonnet (planktonnet.awi.de), per the associated Creative Commons licenses. Image D is reproduced, with the permission of the authors, from Oborník *et al.*, 2012.

have occurred following the radiation of extant dinoflagellates, which is believed to have occurred a maximum of 250 million years ago (Fig. 1.3) (Medlin, 2011; Parfrey *et al.*, 2011). The serially acquired dinoflagellate plastids thus represent some of the most recently acquired plastid lineages known (Fig. 1.3).

Although the apicomplexans are no longer photosynthetic, it is clear that they are descended from photosynthetic ancestors, as all extant species, barring members of the genus *Cryptosporidium*, retain a vestigial, non-photosynthetic plastid, termed the "apicoplast"

(Janouškovec *et al.*, 2010; Lim and McFadden, 2010). The apicoplast resolves phylogenetically as a sister group of the peridinin dinoflagellate plastid (Janouškovec *et al.*, 2010). Recently, two fully photosynthetic species, the "chromerid" algae *Chromera velia* and *Vitrella brassicaformis*, which were identified from coral reefs, have been shown to resolve as sister-groups to the parasitic apicomplexan species, and possess similar red algal-derived plastids to peridinin dinoflagellates, confirming that these plastids originate through a common endosymbiotic event (Figs. 1.2, 1.4) (Janouškovec *et al.*, 2010; Moore *et al.*, 2008; Oborník *et al.*, 2012).

Thesis aims

My PhD was conceived to investigate evolutionary transitions in the highly diversified plastids in alveolate lineages. In the following chapters, I will demonstrate how studying alveolate plastids may provide valuable insights into the evolution of the divergent life strategies employed by different alveolate lineages, and into fundamental processes that underpin plastid evolution across the eukaryotes. In particular, I will focus on the extremely unusual transcript processing pathways found in dinoflagellate and chromerid plastids as a model system for which to understand alveolate plastid biology and evolution.

First, I will outline three major conceptual themes in plastid evolution, and demonstrate how alveolate plastids provide ideal systems in which to resolve major outstanding questions for each theme. The first of these concerns what biological factors may lead to genome reduction and gene loss from plastid lineages, and in particular what may have given rise to the extremely different sets of genes retained in the plastids of photosynthetic and parasitic alveolates. The second of these examines how post-endosymbiotic changes to plastid genome organisation, which are particularly noticeable in the extremely divergent plastid genomes found in dinoflagellates, may impact on the evolution of biochemical pathways associated with plastids. The final major theme investigates whether plastids acquired through serial endosymbiosis, such as those identified in dinoflagellates, are supported by a "shopping bag" of pathways derived from different phylogenetic sources (Larkum *et al.*, 2007). I will then outline the specific features of transcript processing in alveolate plastids. Finally, I will provide a brief overview of the subsequent thesis chapters.

Theme 1- Genome reduction in plastid evolution

The post-endosymbiotic evolution of plastid genomes is characterised by extensive gene loss. Whereas the genomes of free-living cyanobacteria, for example, possess more than 3000 genes, even the largest red algal plastid genomes possess only 250 genes (Fig. 1.5)

(Green, 2011; Janouškovec *et al.*, 2013a; Kaneko *et al.*, 1996). Similarly, with few known exceptions, the nuclei and mitochondria of eukaryotic symbionts were lost during the acquisition of secondary and tertiary plastids (Curtis *et al.*, 2012; Imanian *et al.*, 2012). While a few plastid genomes have been proposed to have gained genes since their endosymbiotic acquisition, this process has not been documented to occur in every plastid lineage, and where it does occur appears to contribute only a small number of genes (Janouškovec *et al.*, 2010; Moszczynski *et al.*, 2012; Ruck *et al.*, 2014).

Some of the genes that are no longer retained in plastid genomes have been relocated to the host nucleus. Other genes may be lost completely, leading to streamlining of plastid metabolism. For example, a high-throughput phylogenetic study of the genes of cyanobacterial origin in the nuclear genomes of archaeplastid lineages concluded that the symbiont that gave rise to the primary plastid was likely to be closely related to heterocyst-forming, nitrogen-fixing cyanobacterial lineages (e.g. *Anabaena, Nostoc)* (Deusch *et al.*, 2008). From this, it was suggested that the early ancestor of all extant plastids possessed the ability to fix nitrogen, but that the genes required for this metabolic pathway were lost following the primary endosymbiosis event, as nitrogen fixation is not known to occur in any eukaryotic lineage (Deusch *et al.*, 2008). In other cases, plastid genes may have been lost in individual eukaryotic lineages. Some of the most dramatic examples of this occur in plastid-containing eukaryotes that have secondarily lost the ability to photosynthesise, many of which have adopted parasitic life strategies. This has not only occurred in apicomplexans, but has also occurred in at least fourteen lineages of plants, as well as in many algal lineages (Blouin and Lane, 2012; Gornik *et al.*, 2012; Tillich and Krause, 2010).

Understanding why certain genes have been lost from plastid genomes, while others have been retained, may provide fundamental insights into post-endosymbiotic organelle evolution. The evolutionary factors that underpin which genes have been retained in plastid genomes have been discussed extensively. For example, it has been suggested that genes whose products affect the redox poise of the plastid, such as the reaction core subunits of complexes involved in photosynthetic electron transfer, are preferentially retained in plastid genomes to allow rapid redox regulation of expression (Allen, 1993, 2003). Although a great deal has been written about the processes that bias a gene in favour of being retained in plastid genomes, less has been said about what factors render certain genes especially susceptible to being lost from plastid genomes. It has been suggested that gene loss from plastid genomes is influenced by mutations in the underlying sequence. Localised high levels of sequence mutation have been suggested to have contributed to preferential gene loss and transfer in some plastid lineages (Magee *et al.*, 2010). Furthermore, several

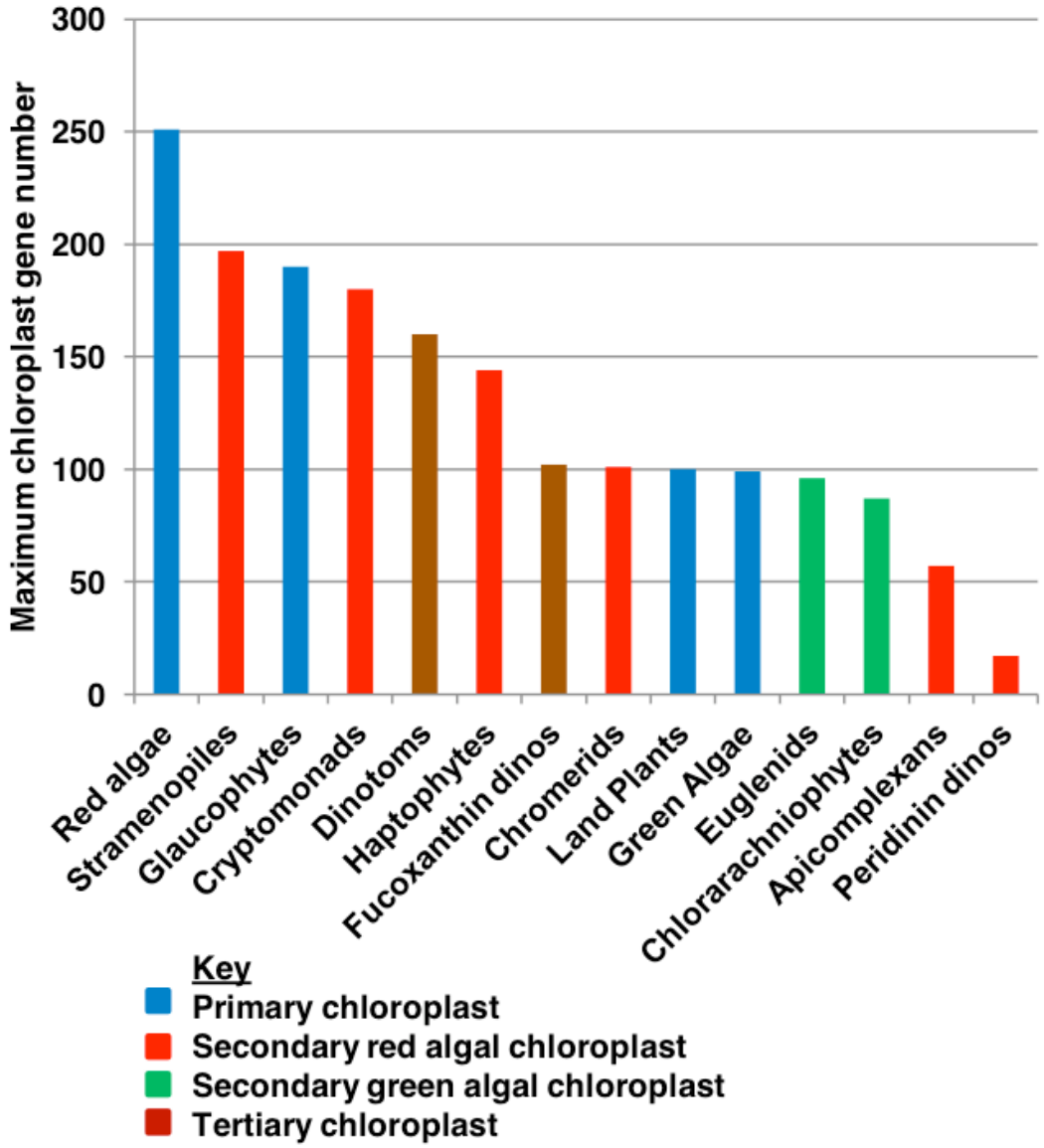


Fig. 1.5. Reductive evolution of plastid genomes.

This diagram plots the maximum number of genes found in plastids from different phylogenetic groups of photosynthetic eukaryotes. This diagram was generated using information from: Gabrielsen *et al.*, 2011; Green, 2011; Imanian *et al.*, 2010; Janouškovec *et al.*, 2010; and Ruck *et al.*, 2014. Plastids are coloured corresponding to their endosymbiotic origin as per Fig. 1.2.

species that have secondarily lost the ability to photosynthesise retain genes in their plastids that contain mutations that render them translationally non-functional (Randle and Wolfe, 2005; Siemeister and Hachtel, 1990). This suggests that gene inactivation precedes gene

loss from plastid genomes. However, it is additionally possible that plastid gene loss events may also be predicated by changes to *trans*-acting factors that interact with plastid genes. For example, the loss of a factor required for the expression of a particular plastid gene might lead to its inactivation, indirectly causing its loss from the plastid genome.

Alveolates provide an exceptional opportunity for testing the factors that influence plastid gene loss. The plastid genomes of peridinin dinoflagellates are the smallest known of any photosynthetic eukaryote, containing as few as 20 genes (Fig. 1.5) (Barbrook *et al.*, 2013; Howe *et al.*, 2008b; Mungpakdee *et al.*, 2014). All of the genes of recognisable function within the peridinin dinoflagellate plastid, with the exception of ribosomal and transfer RNAs, encode proteins that function directly in photosynthesis (Green, 2011; Howe *et al.*, 2008b). Some peridinin dinoflagellate plastid genomes additionally contain large open reading frames that do not encode recognisable photosystem subunits (Barbrook *et al.*, 2006; Nisbet *et al.*, 2004). These open reading frames appear to be transcribed (A.C. Barbrook, pers. comm.) (Barbrook *et al.*, 2012; Nisbet *et al.*, 2008). However, these open reading frames contain no obvious similarity to any previously identified sequence, are generally not conserved between different peridinin dinoflagellate species, and frequently contain large numbers of codons that are otherwise rarely found in the protein-coding genes of recognisable function in peridinin dinoflagellate plastids (Barbrook *et al.*, 2006; Hiller, 2001; Iida *et al.*, 2010). Thus, it is thus debatable as to whether these open reading frames encode functional proteins. Evidence has additionally emerged for the possible acquisition of a small number of genes of non-photosynthesis function in the plastid of the dinoflagellate *Ceratium horridum* through lateral gene transfer, although these genes are not known in any other peridinin dinoflagellate plastid genome (Moszczynski *et al.*, 2012; Howe *et al.*, 2008). All other plastid genes of non-photosynthesis function that have been identified in peridinin dinoflagellates are located in the nucleus (Bachvaroff *et al.*, 2004; Hackett *et al.*, 2004; Howe *et al.*, 2008b).

In contrast to the peridinin dinoflagellate plastid, the plastids in parasitic apicomplexans have lost all genes of photosynthetic function, and only retain genes of non-photosynthesis function (Janouškovec *et al.*, 2010; Lim and McFadden, 2010). Thus, the apicomplexan and peridinin dinoflagellate plastid lineages, while closely related to each other, retain almost entirely non-overlapping sets of genes. This dramatic a divergence in genome content has not been documented in any other plastid lineage (Green, 2011). Recently, plastid genomes have been sequenced for the chromerid algae *Chromera velia* and *Vitrella brassicaformis* (Janouškovec *et al.*, 2010). Notably, these genomes retain more genes than those of either peridinin dinoflagellates or apicomplexans, and include genes of photosynthesis and of non-photosynthesis function (Fig. 1.5) (Janouškovec *et al.*, 2010).

The chromerid algae accordingly provide a valuable model system for inferring possible ancestral features of the peridinin dinoflagellate and apicomplexan plastids. From this, it may be possible to reconstruct some of the evolutionary events that are likely to have preceded the divergent evolution of the peridinin plastid and apicoplast genomes. For example, it would be interesting to determine whether there are biochemical pathways specifically associated with the functional expression of either the photosynthesis genes, or genes of non-photosynthesis function in chromerid plastids, which might similarly have been specifically involved in the function of one set of genes in an early ancestor of dinoflagellates or apicomplexans. If, for example, the most recent photosynthetic ancestors of extant apicomplexans possessed a pathway specifically required for the expression of the photosynthesis genes within their plastids, the loss or corruption of this machinery might have led to the loss of photosynthesis function, and the eventual loss of photosynthesis genes and divergence towards a heterotrophic, and ultimately parasitic lifestyle.

Theme 2- Post-endosymbiotic changes to plastid genome organisation

In addition to undergoing extensive gene loss, plastids undergo extensive post-endosymbiotic changes to genome organisation. These may include the disruption of operons through genome rearrangement, lineage-specific gains of introns, and expansions in intergenic non-coding DNA (Green, 2011; Palmer, 1987; Shimada and Sugiura, 1991). In addition, the global architecture of the plastid genome may change following endosymbiosis. Whereas most characterised plastid genomes consist of a single circular chromosome, some plastid genomes (including that of *Chromera*) have alternative linear and branched forms (Janouškovec *et al.*, 2013b; Oldenburg and Bendich, 2004). It is not known, however, whether changes to genome organisation and gene structure affect plastid biology, and in particular alter plastid gene expression pathways.

Dinoflagellate plastids represent an excellent model in which to study the effects of changes to genome organisation. In addition to having undergone extensive gene loss events, the peridinin dinoflagellate plastid genome is fragmented into a series of small, circular elements termed “minicircles”, most of which contain only one gene (Owari *et al.*, 2014; Zhang *et al.*, 1999). Although some multiple gene minicircles have been identified in specific dinoflagellate species (*Adenoides eludens, Amphidinium carterae, Heterocapsa triquetra*), these do not contain combinations of genes found in other plastid lineages, and appear to have arisen through recent recombination events (Hiller, 2001; Nelson *et al.*, 2007; Nelson and Green, 2005). Thus, the common ancestor of peridinin dinoflagellates is likely to have possessed a plastid genome composed entirely of single gene minicircles (Howe *et al.*, 2008b). The

consequences of this fragmentation event for the gene expression pathways associated with peridinin dinoflagellate plastids are poorly understood.

A plastid genome sequence has recently been published for the fucoxanthin dinoflagellate *Karlodinium veneficum* (Figs. 1.4, 1.5) (Gabrielsen *et al.*, 2011). This genome has lost over a third of the genes present in the plastid genomes of free-living haptophytes (Gabrielsen *et al.*, 2011). In addition, the *K. veneficum* plastid genome has undergone extensive rearrangement, and contains many in-frame insertions and premature in-frame termination codons in coding sequences, which are not found in haptophyte plastids (Gabrielsen *et al.*, 2011). A recent study has suggested that some genes within the *K. veneficum* plastid genome are located on episomal elements that are not contiguous with the primary chromosomal genome, which may constitute a population of minicircles that has evolved in parallel to those of the peridinin lineage, although the complete sequence of an episomal element has not previously been identified (Espelund *et al.*, 2012).

Complete plastid genome sequences have also been published for the dinotom algae *Kryptoperidinium foliaceum* and *Durinskia baltica* (Imanian *et al.*, 2010). Unlike the situation in fucoxanthin dinoflagellates, dinotom plastid genomes have undergone very few independent gene loss events since their divergence from free-living diatoms, and have undergone no major rearrangement events (Imanian *et al.*, 2010). A complete plastid genome has yet to be published for the green algal plastids present in *Lepidodinium.* However, preliminary studies have suggested that the genome of this plastid may be divergently organised from those of other green algae, for example utilising alternative translation initiation codons (Matsumoto *et al.*, 2011a). Understanding why certain dinoflagellate plastid genomes are divergently organised, while others are not, may provide insights into the physiological factors underpinning post-endosymbiotic genome evolution.

Theme 3- Serial endosymbiosis and the "shopping bag" model

The serially acquired plastids found in dinoflagellates provide an important opportunity to investigate an additional process in plastid evolution. The conventional model of plastid evolution involves the endosymbiotic integration of two organisms: a host, and a free-living symbiont. This involves the establishment of pathways within the host to support the plastids (Howe *et al.*, 2008a; Larkum *et al.*, 2007). In serial endosymbioses, however, three organisms may participate in the establishment of a permanent plastid: the host, the incoming symbiont, and the ancestral plastid lineage that is replaced (Fig. 1.6). It is possible that pathways established to support the ancestral plastid might be retained and applied to the incoming replacement plastid, thus supporting its biology (Fig. 1.6).

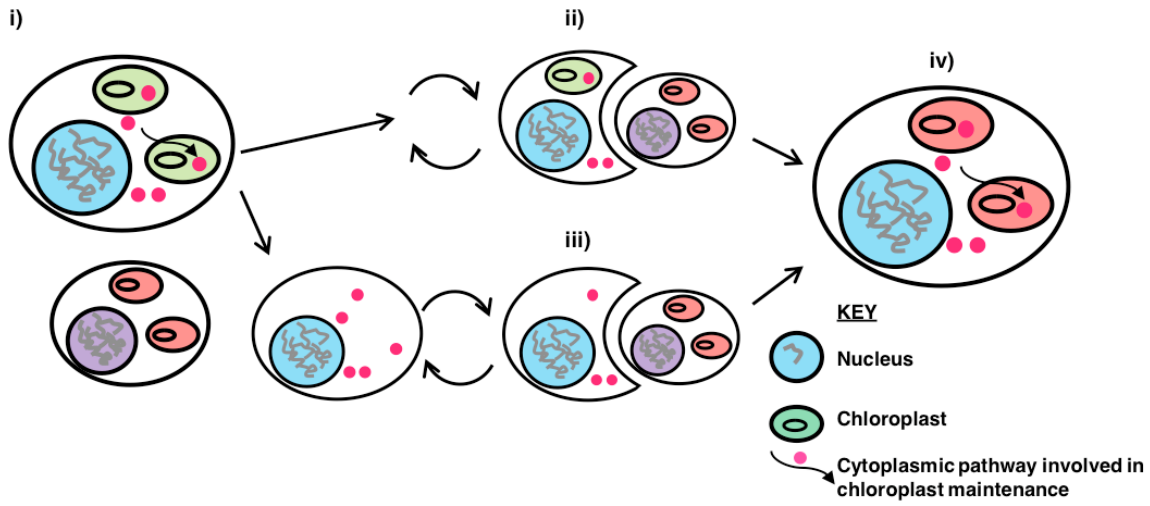


Fig. 1.6: Serial endosymbiosis and the "shopping bag" model.

This diagram shows the events proposed to occur in the "shopping bag" model for the evolution of organelles acquired through serial endosymbiosis. Serial endosymbiosis occurs where a eukaryotic host that is already photosynthetic replaces its original plastids with ones of different phylogenetic origin **(i)**.The replacement plastids may be acquired before the original plastids are lost **(ii)**, or the host may lose its original plastids and pass through a non-photosynthetic intermediate state before consuming the new symbiont lineage **(iii)**, and establishing it as a permanent organelle **(iv).**

The host lineage that acquires the replacement plastid already possesses pathways that may be useful for plastid maintenance, established to support the original plastid lineage **(i)**. These pathways may be retained following the loss of the original plastid **(ii, iii)**, and applied to the incoming replacement, supporting and altering its biology **(iv)**.

The idea that serial acquired plastids may be supported by pathways retained from predecessor symbioses is one of several predictions of the "shopping bag" model for plastid evolution (Larkum *et al.*, 2007). The "shopping bag" model furthermore predicts that many extant plastids, including the primary plastids found in archaeplastids, may be supported by pathways derived from symbiotic associations in which the host lineage engaged, prior to the endosymbiotic acquisition of the plastid (Larkum *et al.*, 2007). Recently, evidence has emerged that many plastid lineages are supported by pathways that are derived from a different phylogenetic source to either the host nucleus, or the extant plastid itself. The genes that are associated with these pathways may constitute the "footprints" of cryptic plastid symbioses that are no longer retained in any extant lineage. For example, some algae that

Table 1.1. Genes retained from the peridinin plastid symbiosis in dinoflagellates that have undergone serial endosymbiosis.

The genes listed have been identified in previously published single-gene phylogenies to encode proteins of plastid function, and resolve with peridinin dinoflagellates to the exclusion of all other lineages. For each gene, the source organism, bootstrap support for the phylogenetic association, and the study in which each gene was identified are given. Information is taken from **A:** Nosenko *et al.*, 2006. **B:** Patron *et al.*, 2006. **C:** Minge *et al.*, 2010. **D:** Takishita *et al.*, 2008. Genes highlighted in bold are ones for which a second gene copy, derived from the replacement plastid or another phylogenetic source, has been identified in the host.

Gene	Species	Support	Study
Glyceraldehyde-3-phosphate dehydrogenase isoform C1	*Karenia brevis*	76	A
Ascorbate peroxidase	*Karlodinium veneficum*	96	B
ATP synthase gamma	*Karlodinium veneficum*	61	B
Cysteine synthase	*Karlodinium veneficum*	100	B
Ferredoxin	*Karlodinium veneficum*	<50	B
Glutamate semialdehyde mutase	*Karlodinium veneficum*	<50	B
Phosphoribulokinase	*Karlodinium veneficum*	100	B
Ribulose bisphosphate isomerase	*Karlodinium veneficum*	<50	B
1-deoxy-D-xylulose 5-phosphate reductoisomerase	*Lepidodinium chlorophorum*	100	C
4-diphosphocytidyl-2C-methyl-D-erythritol kinase	*Lepidodinium chlorophorum*	95	C
Ferredoxin NADP reductase	*Lepidodinium chlorophorum*	95	C
Glyceraldehyde-3-phosphate dehydrogenase isoform C1	*Lepidodinium chlorophorum*	100	D

possess secondary, red algal plastids, such as diatoms, contain genes purported to be of green algal origin (Dorrell and Smith, 2011; Moustafa *et al.*, 2009). Similarly, the genomes of some archaeplastids contain numerous genes of predicted chlamydiobacterial origin (Ball *et al.*, 2013; Huang and Gogarten, 2007). These may be the remnants of a chlamydiobacterial endosymbiosis, which preceded the cyanobacterial plastid endosymbiosis in these lineages. The validity of these hypotheses remains controversial, due to problems in confirming the phylogenetic affinities of the genes identified, and the absence of extant descendants of the cryptic endosymbionts, which would provide clear taxonomic evidence of the serial endosymbiosis event (Deschamps and Moreira, 2012; Dorrell and Smith, 2011; Woehle *et al.*, 2011).

Previous studies of the nuclear transcriptomes of dinoflagellates that harbour serially acquired plastids have identified genes of red algal origin, that may have been acquired by endosymbiotic gene transfer from the preceding peridinin plastid (Table. 1.1) (Minge *et al.*, 2010; Nosenko *et al.*, 2006; Patron *et al.*, 2006). However, the expression products of these genes have not been confirmed to function in the plastids of serial dinoflagellates, and therefore their exact contributions to the physiology of the replacement plastid lineages

remain uncertain. In addition, each of the genes identified is of widespread distribution across photosynthetic eukaryotes (Table 1.1). It is likely that orthologues of these genes were present in the free-living ancestors of the replacement plastid, and might also have been acquired by the dinoflagellate host. For example, *Karlodinium veneficum* has been shown to possess two copies of the cysteine synthase gene, one of which is retained from the peridinin symbiosis, and one of which is derived from the replacement fucoxanthin plastid (Patron *et al.*, 2006). A similar situation has been shown for isoform C1 glyceraldehyde-3-phosphate dehydrogenase in other fucoxanthin dinoflagellates, and in *Lepidodinium* (Nosenko *et al.*, 2006; Takishita *et al.*, 2008). It is possible that the ancestral copies of these genes are no longer functional, and only the serially acquired copies support the replacement plastid. To date, no pathways have been identified in a serially acquired dinoflagellate plastid that originate from the ancestral peridinin plastid symbiosis, and are not found in the free-living relatives of the replacement plastid lineage. The identification of such a pathway would unambiguously confirm that the biology of replacement plastids is affected by pathways retained by the host from previous symbioses.

Transcript processing in plastids

The expression of plastid genes involves a complex suite of transcription and transcript processing events. Plant plastids utilise two RNA polymerases: a plastid-encoded polymerase similar to that employed by free-living cyanobacteria, and a nucleus-encoded polymerase similar to the phage-type polymerase employed by mitochondria (Barkan, 2011; Green, 2011; McBride *et al.*, 1994). The phage-type plastid polymerase is not found in other photosynthetic eukaryotes, where the bacterial-type polymerase is solely responsible for transcription of the plastid genome (Barkan, 2011; Green, 2011). Within the alveolates, the apicoplast genome, and the plastid genomes of chromerids, fucoxanthin dinoflagellates and dinotoms all encode the conventional bacterial-type polymerase (Gabrielsen *et al.*, 2011; Imanian *et al.*, 2010; Janouškovec *et al.*, 2010). There are no genes of obvious homology to an RNA polymerase in peridinin dinoflagellate plastids (Howe *et al.*, 2008b).

Typically, plastid genes are cotranscribed. This has been documented in multiple plastid lineages, including those of plants (Strittmatter and Kössel, 1984), diatoms (Hwang and Tabita, 1991), and haptophytes (Fujiwara *et al.*, 1993). Similar events have been documented in peridinin dinoflagellate plastids. For example, in some dinoflagellates, transcripts equal to or longer than the corresponding minicircle sequence have been identified (Dang and Green, 2010; Nisbet *et al.*, 2008). This has been interpreted as evidence for a “rolling circle” form of transcription, in which long transcripts containing multiple copies of the minicircle sequence are generated by each RNA polymerase (Dang

and Green, 2010). Cotranscription has also been documented to occur in apicoplast transcripts (Wilson *et al.*, 1996). However, before the work in this thesis, the extent of cotranscription in other alveolate plastid lineages, such as those of chromerids and dinoflagellates that possess replacement plastids, had not been directly investigated.

Following transcription, plant plastid transcripts undergo extensive processing events. Introns within transcript sequence are removed as a result of *cis*-splicing, and exons of individual genes that are transcribed from distinct parts of the plastid genome may be ligated together through *trans*-splicing (Asano *et al.*, 2013; Glanz and Kück, 2009; Tillich and Krause, 2010). In addition, polycistronic transcripts, generated through the cotranscription of plastid genes, are cleaved to form mature and often monocistronic mRNAs, via 5' and 3' nuclease activities (Barkan, 2011; Pfalz *et al.*, 2009). Dinoflagellate plastid genomes do not possess recognisable introns, and transcript splicing has not been reported (Gabrielsen *et al.*, 2011; Howe *et al.*, 2008b; Imanian *et al.*, 2010). However, the predominant plastid transcripts in peridinin dinoflagellates, as identified through northern blotting, correspond to monocistronic mRNAs, indicating that plastid transcripts undergo extensive cleavage following transcription (Barbrook *et al.*, 2001; Nisbet *et al.*, 2008).

In plant plastids, the generation of mature mRNAs involves alternative cleavage events, in which the cleavage of an mRNA at a site associated with one gene prevents the generation of mature mRNAs of adjacent genes from the same polycistronic precursor (Barkan *et al.,* 1994; Rock *et al.*, 1987). There is evidence for similar alternative cleavage events, associated with transcripts of multigene minicircles, in dinoflagellate plastids. The *Amphidinium carterae petB/atpA* minicircle, for example, has been shown to give rise to mature, monocistronic *atpA* transcripts that extend at the 5' end up to 45 nt into the upstream *petB* CDS (A.C. Barbrook, pers. comm.) (Barbrook *et al.*, 2012). However, before the work in this thesis, similar cleavage events had not been characterised in other photosynthetic alveolates.

A further function of transcript cleavage appears to be the degradation of non-coding RNA (Barkan, 2011; Hotto *et al.*, 2012). At least some of the degradation events in plant plastids are programmed by the addition of a 3' poly(A) tail onto unwanted transcripts, allowing them to be distinguished from functional transcripts (Kudla *et al.*, 1996). Polyadenylylated plastid transcripts have not formally been identified in any algal plastid lineages, although poly(A) tail addition has recently been inferred to also occur in the secondary, green algal derived plastids of euglenids (Lange *et al.,* 2009; Záhonová *et al.,* 2014). It is not clear what processes enable the degradation of non-coding transcripts in other plastid lineages. It has been shown in peridinin dinoflagellates that transcripts covering non-coding sequence are

much less abundant than the mature mRNAs (Dang and Green, 2010; Nisbet *et al.*, 2008). This suggests that non-coding plastid transcripts are preferentially degraded. However, studies prior to this thesis had not identified any processing events in a dinoflagellate plastid that discriminate non-coding transcripts from functional mRNAs.

Antisense transcripts are a particularly important component of non-coding RNA in plant plastids (Hotto *et al.*, 2012; Sharwood *et al.*, 2011). These are generated either from promoters located on the template strand of plastid genes, or via transcriptional read-through from pairs of genes located in opposing orientation to each other (Georg *et al.*, 2010; Sharwood *et al.*, 2011). The targeted removal of antisense transcripts is important for plant plastid function, as antisense transcripts can anneal to and inhibit the expression of the complementary sense transcripts (Sharwood *et al.*, 2011; Zghidi-Abouzid *et al.*, 2011). The presence of antisense plastid transcripts in the apicomplexan *Toxoplasma gondii* has been inferred from microarray data, and have subsequently also been detected in *Plasmodium falciparum* (Bahl *et al.*, 2010; Kurniawan, 2013). However, before the work in this thesis, antisense transcripts had not been reported in algal plastids, and it was not known whether antisense transcripts play further functional roles in gene expression in any plastid lineage.

Peridinin dinoflagellates utilise two additional very distinctive plastid RNA processing pathways. Plastid transcripts may undergo extensive substitutional editing events, in which up to one tenth of the nucleotides in a given transcript sequence are altered to form other nucleotides, with a wide variety of different forms of editing event found in different species (Green, 2011; Howe *et al.*, 2008b; Zauner *et al.*, 2004). The extent of editing varies between different species, with far greater numbers of editing events observed in the species *Ceratium horridum* and *Alexandrium tamarense* than the basally divergent dinoflagellates *Amphidinium carterae* and *Heterocapsa triquetra* (Bachvaroff *et al.*, 2014; Howe *et al.*, 2008b; Iida *et al.*, 2009; Zauner *et al.*, 2004). Transcript editing has been reported in plant plastids, but is very different from that observed in alveolates, as it is restricted to fewer than 100 sites across the entire genome, and is limited to C to U interconversions (Fujii and Small, 2011; Yoshinaga *et al.*, 1996). Editing has not been reported in published transcript sequences from other plastids, such as those of green algae, haptophytes or diatoms, indicating that the plant and dinoflagellate transcript editing pathways arose independently (Fujii and Small, 2011; Fujiwara *et al.*, 1993; Hwang and Tabita, 1991). Before the work in this thesis, transcript editing had not been documented in any other algal plastid lineage.

Most unusually, transcripts in peridinin dinoflagellate plastids receive a 3' poly(U) tail (Wang and Morse, 2006). This transcript modification has not been reported in the plastids of plants, or any other non-alveolate plastid lineage. The function of the poly(U) tail is poorly

understood, although it has been suggested to enable other transcript processing events, such as 5' terminal cleavage (Dang and Green, 2010; Nisbet *et al.*, 2008) and editing (Dang and Green, 2009). Before the work in this thesis, poly(U) tail addition had been shown to occur on transcripts of three plastid photosynthesis genes (*psaA, psbB, psbC*) in the chromerid alga *Chromera velia*, and had been shown not to occur on plastid transcripts of parasitic apicomplexans (R.E.R. Nisbet, pers. comm..), (Dorrell *et al.*, 2014; Janouškovec *et al.*, 2010). However, it was not known whether poly(U) tails were found in the chromerid *Vitrella brassicaformis*, or were applied to plastid genes of non-photosynthesis function in any photosynthetic alveolate. In addition, the presence of poly(U) tails and transcript editing had not been investigated in any dinoflagellate plastid acquired through serial endosymbiosis.

Outline of thesis chapters

Chapter Two outlines the key experimental techniques employed in each subsequent chapter.

Chapter Three, "Processing of core-containing and antisense transcripts generated from plastid minicircles in the peridinin dinoflagellate *Amphidinium carterae*" describes the diversity and processing of transcripts associated with the *petB/ atpA* and *psbA* minicircles in the model peridinin dinoflagellate species *Amphidinium carterae* (Fig. 1.4). I wished to identify whether multi-copy transcripts were generated from these minicircles, and whether these transcripts undergo similar 5' cleavage and 3' poly(U) tail addition events to mature mRNAs. I additionally wished identify the non-coding transcripts produced from each minicircle, and in particular determine whether antisense transcripts are present in peridinin dinoflagellate plastids. I finally wished to determine whether non-coding transcripts undergo different processing events to mature mRNAs.

I demonstrate that rolling circle transcription is a ubiquitous feature across the peridinin dinoflagellates, and that transcripts generated through this process can undergo similar 5' end cleavage and 3' end poly(U) tail addition events to monocistronic mRNAs. I additionally provide the first evidence for antisense transcripts in an algal plastid lineage. These antisense transcripts do not receive poly(U) tails, indicating that poly(U) tail addition may have a role in discriminating between coding and non-coding transcripts in plastid RNA processing.

Chapter Four, "Transcript processing pathways retained from an ancestral plastid symbiosis function in serially acquired dinoflagellate plastids", investigates whether serially acquired

dinoflagellate plastids are supported by pathways retained from the predecessor peridinin symbiosis. I wished to determine whether transcripts in serially acquired dinoflagellate plastids may receive 3' poly(U) tails or undergo substitutional sequence editing, as in the ancestral peridinin plastid. I additionally wished to determine whether poly(U) tail addition and editing are found in non-alveolate plastid lineages, or are specifically associated with the plastids of dinoflagellates and their closest relatives.

I report that transcripts in the plastids of the fucoxanthin dinoflagellate *Karenia mikimotoi* (Fig. 1.4) receive 3' poly(U) tails, and are edited. I show that these pathways are not found in free-living haptophytes or other lineages containing secondary, red algal plastids, indicating that they have been retained from the ancestral peridinin symbiosis through serial endosymbiosis, and applied to the replacement plastid. This represents a major development to existing theories of plastid evolution, as it demonstrates that the biology of plastids may be actively altered by pathways retained from prior symbioses.

Chapter Five, "Plastid genome sequences and transcript processing pathways have evolved together in the fucoxanthin dinoflagellate *Karlodinium veneficum*" profiles poly(U) tail addition and transcript editing events across the entire published plastid genome of *K. veneficum* (Fig. 1.4). I wished to determine the extent to which these pathways, which have been acquired by the fucoxanthin plastid following its endosymbiotic acquisition by the dinoflagellate host, have been co-opted to enable the expression of the plastid genome. I additionally wished to identify what poly(U) tail addition and editing events were associated with transcripts of highly divergent regions of the *K. veneficum* genome, and from this infer how the transcript processing machinery has responded to the rapid genome evolution of fucoxanthin dinoflagellate plastids.

I demonstrate that poly(U) tail addition and editing are associated with effectively every transcript in the *K. veneficum* plastid, including transcripts of genes of non-photosynthesis function not present in the ancestral peridinin plastid. I additionally provide evidence that transcript processing pathways in fucoxanthin dinoflagellates have evolved alongside the underlying genome sequence. For example, the *K. veneficum* plastid genome has undergone a parallel fragmentation event to that observed in peridinin dinoflagellates, in which the *dnaK* gene is located on episomal minicircles, which give rise to polyuridylylated and edited transcripts.

Chapter Six, "Poly(U) tail addition plays a central role in plastid transcript processing in the fucoxanthin dinoflagellate *Karenia mikimotoi*" characterises the role of poly(U) tail addition in plastid transcript processing in fucoxanthin dinoflagellates. I wished to determine whether

poly(U) tail addition was extensively associated with plastid transcript processing in *K. mikimotoi,* as it is in *Karlodinium veneficum,* and determine whether poly(U) tail addition is associated with other events in transcripts processing, as has been inferred to occur in peridinin dinoflagellate plastids. I additionally wished to confirm whether antisense plastid transcripts are present in fucoxanthin dinoflagellates, as I have previously shown to be present in peridinin dinoflagellate plastids, and determine whether these antisense transcripts receive 3' poly(U) tails and are edited.

I have reconstructed a polyuridylylated plastid transcriptome for *K. mikimotoi* via a novel next-generation sequencing pathway. I find evidence for a wide diversity of polyuridylylated plastid transcripts, and also find evidence for the post-endosymbiotic divergence of fucoxanthin plastid genomes. I additionally find evidence for functional roles of poly(U) tail addition in facilitating editing, and the stoichiometric adjustment of different transcripts through alternative processing. As in peridinin dinoflagellates, non-polyuridylylated antisense transcripts are widespread in fucoxanthin dinoflagellates. I demonstrate that these antisense transcripts are edited in complementary patterns to the corresponding sense transcripts, suggesting that they play a previously unidentified role in directing plastid transcript processing events.

Chapter Seven, "Evolution and function of plastid transcript processing in algal relatives of malaria parasites" documents the distribution and function of poly(U) sites in the plastids of *Chromera velia* and *Vitrella brassicaformis* (Fig. 1.4). I wished to determine whether poly(U) tail addition in chromerid algae is an ubiquitous feature of plastid transcript processing, as in fucoxanthin dinoflagellates, or whether poly(U) tails are specifically associated with transcripts of photosynthesis genes, which have been lost from apicomplexan plastids. I additionally wished to identify potential roles for poly(U) tail addition in chromerid plastid transcript processing. From this, I wished to infer whether the loss of the poly(U) tail addition machinery from apicomplexans might be associated with the loss of photosynthesis genes from the apicoplast genome, and the transition of ancestors of apicomplexans from photosynthesis towards parasitism.

I present evidence that poly(U) tails in chromerids are specifically added to transcripts that encode components of the photosynthetic electron transport chain, and are not associated with transcripts of plastid genes of non-photosynthesis function. This represents the first documented example of a plastid transcript processing pathway that preferentially targets one functional category of genes. I provide evidence that this differential poly(U) tail addition may drive differences in transcript abundance between non-photosynthesis in photosynthesis genes, by directing the maturation of individual genes from polycistronic precursor

transcripts. The loss or inactivation of a poly(U) tail addition pathway essential for high levels of photosynthesis gene expression might accordingly have driven the transition towards a parasitic life strategy in early ancestors of apicomplexans.

Chapter Eight presents a synoptic view of the evolution and function of transcript processing pathways across alveolate plastids, and outlines future potential directions for experimental research.

Chapter Two- Materials and Methods

Cultures

Amphidinium carterae CCMP 1314, *Phaeodactylum tricornutum* CCAP 1052/6, *Emiliania huxleyi* CCMP 1516, *Kryptoperidinium (Glenodinium) foliaceum* PCC 499 and *Chromera velia* CCMP 2878 were cultured in f/2 medium, which was prepared with Ultramarine Synthetica artificial sea water (Waterlife) and buffered with 500 µg/ ml tricine to pH 8. *Vitrella brassicaformis* RRM 111-2 was cultured in f/2 medium supplemented with 100 µg / ml spectinomycin, and 20 µg/ ml each ampicillin and kanamycin. Cultures were maintained at 18 °C, under 30 µE $m^{-2}s^{-1}$ illumination on a 16:8h L:D cycle.

Karenia mikimotoi RCC1513, *Karlodinium veneficum* UIO297 and *Lepidodinium chlorophorum* AC195 were grown in modified k/2 medium as per http://www.sb-roscoff.fr/Phyto/RCC/index.php?option=com_content&task=view&id=8&Itemid=14#K_lan at 15 °C under 50 µE $m^{-2}s^{-1}$ continuous illumination.

The identity of each culture was confirmed by microscopy, and by DNA barcoding, using PCR primers specific to the plastid *psbA* and nuclear 18S ribosomal RNA genes. *A. carterae* CCMP 1314 was found to be genetically identical to the strain CCAP 1102/6, for which extensive plastid genome sequence is available (Barbrook and Howe, 2000; Barbrook *et al.*, 2001; Gachon *et al.* 2013; Nisbet *et al.*, 2004); and *Karlodinium veneficum* UIO 297 and *V. brassicaformis* RRM 111-2 were respectively found to be identical to the strains UIO 083 and CCMP 3155, for which complete plastid genome sequences have been published (Gabrielsen *et al.*, 2011; Janouškovec *et al.*, 2010). *Karenia mikimotoi* RCC 1513 was found to be close in sequence identity to *Gymnodinium mikimotoi* strain G303ax-2, for which some plastid gene sequences have previously been published (Takishita *et al.*, 1999). *Kryptoperidinium foliaceum* PCC 499 was found to be substantially different in sequence from the strain of *Kryptoperidinium foliaceum* (CCMP 1326) for which plastid genome sequences have previously been published (Imanian *et al.*, 2010; Imanian *et al.*, 2012).

RNA isolation

Cultures used for RNA isolation for RT-PCR were harvested in late log phase (21 days post-inoculation for *Amphidinium carterae, Emiliania huxleyi, Phaeodactylum tricornutum, Chromera velia*; 45 days post-inoculation for *Karenia mikimotoi, Karlodinium veneficum, Kryptoperidinium foliaceum, Lepidodinium chlorophorum, Vitrella brassicaformis*). Cultures used for RNA isolation for northern blotting were harvested in early stationary phase (35 days post-inoculation for *A. carterae, C. velia*; 60 days post-inoculation for *Karenia*

mikimotoi). At the time of harvesting, *C. velia* cells were predominantly coccoid, and *V. brassicaformis* were predominantly pigmented (i.e. in the vegetative stage of the life cycle) (Oborník *et al.*, 2012; Oborník *et al.*, 2011). Cells were harvested by centrifugation of the liquid culture at 4380 x g for 10 minutes at 21˚C. Cell pellets were washed three times with sterile artificial sea water prior to isolation of nucleic acids.

Cell pellets were lysed by resuspension in 1ml Trizol reagent (Life Technologies): 50 mg cell pellets, in RNAse-free 2 ml Eppendorf tubes. Trizol-resuspended *A. carterae* cells were immediately used for RNA isolation, as detailed below. Trizol-resuspended *P. tricornutum, E. huxleyi, Karenia mikimotoi, Karlodinium veneficum, Kryptoperidinium foliaceum* and *L. chlorophorum* cells were initially frozen at -80 °C and thawed on ice to facilitate cell lysis, and then immediately used for RNA isolation. Trizol-resuspended *C. velia* and *V. brassicaformis* cells were ground to a powder in liquid nitrogen in a clean pestle and mortar that had been prewashed in 10% hydrogen peroxide. The powdered cells were resuspended in an additional 1 ml Trizol/ 50 mg cells, and immediately used for RNA isolation.

RNA was isolated from the Trizol resuspensions by phase extraction with chloroform, as previously described (Barbrook *et al.*, 2012). 200µl chloroform was added to each 50 mg pellet resuspension, and the samples were centrifuged at 4°C for two minutes, at 8000 x g. The aqueous phase of each centrifugation product was transferred into a clean 2ml Eppendorf tube, a further 500 µl chloroform was added, and the samples were centrifuged at 4 °C for two minutes to remove any residual contamination from the organic phase. The aqueous phase of the chloroform separation was transferred to a clean RNAse-free 1.5 ml Eppendorf tube. 500 µl RNAse free isopropanol was added, and the samples were precipitated at -20^{o} C overnight. The RNA was pelleted by centrifugation at 4 ^{o}C for 15 minutes at 8000 x g. Pellets were washed with ethanol, pelleted again by centrifugation under the same conditions for 5 minutes, and cleaned of all residual ethanol.

RNA to be used for RT-PCR was resuspended in diethylpyrocarbonate-treated water. This was immediately incubated with 10 U RNase-free DNase I, pre-diluted in RNA DNAse Digest buffer (both QIAgen) following the manufacturer's instructions, for 90 minutes at 37 °C. The digestion products were re-purified with RNeasy Kit (QIAgen) and eluted with diethylpyrocarbonate-treated water. RNA to be used for northern blotting was not DNase-treated, and instead was resuspended immediately in formamide. All RNA samples were stored at -80 °C.

The concentration of each RNA sample obtained was quantified using a nanodrop photospectrometer. RNA integrity was confirmed by electrophoresis of 1 µg of each sample

in an RNase-free TBE gel containing 1% agarose, and 0.003% ethidium bromide. Each RNA sample to be used for RT-PCR was tested for DNA contamination by direct PCR amplification with primers specific to the *psbA* gene of the species concerned. The primary products from each PCR were used as the template for a second round of PCR amplification with the same primers. Only RNA samples that produced no products for both the primary and secondary PCR amplifications were used for subsequent experiments.

DNA Isolation

Genomic DNA was harvested by phase extraction with water-saturated pH 8.0 phenol essentially as previously described (Nash *et al.*, 2007). Late log phase cultures of each species were harvested by centrifugation, and each cell pellet was washed with artificial sea water, as before. 50 mg cell pellets were resuspended in 0.01 mol NaCl, 0.001 mol Tris pH 8.0, and 0.0001 mol EDTA, in sterile water to a total volume of 10 ml, and stored at -80 °C. The frozen cells were thawed, and 1 ml 10% SDS was added, along with approximately 20 U Pronase (Roche), which had been pre-activated in 1 ml diethylpyrocarbonate-treated water for one hour at 37 °C. The thawed cells were incubated at 37 °C for four hours to enable cell lysis, with a further 20 U pre-activated pronase added after the first two hours incubation.

The cell lysate was mixed with 10 ml water-saturated phenol pH 8.0 (Roche), and centrifuged at 4380 x g for 15 minutes at room temperature. The aqueous phase was extracted, mixed with 5 ml water-saturated phenol, and 5 ml chloroform, and centrifuged as before. A final phase extraction was performed with the second aqueous phase and 10 ml pure chloroform to remove any residual contamination from the organic phase, and the aqueous phase of this extraction was divided equally into two DNase-free 15 ml falcon tubes. 6 ml pure ethanol, containing 0.01 mol sodium acetate, was added to each tube, and DNA was precipitated overnight at -20 °C.

DNA was pelleted from each precipitation by centrifugation at 12000 x g for 15 minutes at 4 °C. The pellets were washed in ethanol, and precipitated by centrifugation again, under the same conditions as before. Samples were cleaned of all residual ethanol, resuspended in diethylpyrocarbonate-treated water and stored at -20 °C. For all preparations, nucleic acid concentrations were quantified using a nanodrop photospectrometer.

RNA ligation

RNA circularisation was performed using 750 ng freshly harvested total cellular RNA, and T4 RNA ligase (Promega), essentially as previously described (Nash *et al.*, 2007). 750 ng freshly harvested total cellular RNA was incubated with 10 U T4 RNA ligase, 4 µl T4 10x

buffer and 40 U RNAsin (all Promega), 20 µl 40% PEG, and diethylpyrocarbonate-treated water to 40 µl at 37 °C for 1 hour, and then at 16 °C for 16 hours.

RNA ligase-mediated 5' RACE was performed using a variant of previously described conditions (Dang and Green, 2010; Scotto-Lavino *et al.*, 2006). 4 µg freshly harvested total cellular RNA, quantified by a nanodrop spectrophotometer, was ligated to 1 µg of a custom synthesised RNA adapter sequence (GCUGAUGGCGAUGAGCACUGGGUUGCAA) using 10 U T4 RNA ligase, 6 µl T4 10x buffer, 40 U RNAsin, 30 µl 40% PEG and diethylpyrocarbonate-treated water to 60 µl, under the same conditions as used for circularisation. Products of each RNA ligation were cleaned using an RNeasy Mini kit (Qiagen) according to the manufacturer's instructions, eluted in diethylpyrocarbonate-treated water, and stored at -80 °C.

RT-PCR

Reverse transcriptions were performed using a Superscript III First Strand Synthesis kit (Invitrogen), essentially following the manufacturer's instructions. 100 ng RNA template, as quantified by a nanodrop photospectrometer, 10 nmol premixed RNase-free dNTPs, and 2 pmol cDNA synthesis primer, were combined with diethylpyrocarbonate-free water to a final volume of 13 µl, incubated at 65 °C for 5 minutes, and snap cooled on ice. The reactants were collected by centrifugation, and 4 µl 10 x first strand synthesis mixture, 100 nmol DTT, and 200 U Superscript III (all Invitrogen), and 20 U RNAsin (Promega) were added, to a final volume of 20 µl. Reverse transcriptions were performed at 50 °C for 50 minutes (for crude RNA templates) or for 20 minutes (for RNA circularisation and adapter ligation products, and incubated at a further 15 minutes at 75 °C to denature the enzyme. Reverse transcription products were stored at -20 °C.

PCRs were performed using the GoTaq polymerase kit (Promega) essentially as previously described (Barbrook *et al.*, 2012). Approximately 100 ng PCR template was mixed with 10 µl 5 X PCR reaction buffer, 75 nmol $MgCl_2$ and 5U GoTaq Flexi polymerase (all Promega), along with 10 nmol premixed dNTPs, and 10 pmol each of the PCR forward and reverse primers, to a total volume of 50 µl. PCR primers for each experiment are tabulated in the corresponding chapter. The reactants were collected in the tube by centrifugation, then incubated for 10 minutes at 95 °C, followed by 40 cycles of: 45 seconds at 95 °C, 45 seconds at 55 °C, and 3 minutes at 72 °C. A final incubation step at 72 °C was not performed, to minimise the abundance of chimeric PCR products (Lahr and Katz, 2009; Smyth *et al.*, 2010). PCR products were stored at -20 °C.

Thermal asymmetric interlaced PCR (TAiL-PCR)

TAiL-PCRs were performed using reaction mixtures and cycling conditions using a modified version of a previously described protocol (http://dps.plants.ox.ac.uk/langdalelab/protocols/PCR/TAIL_PCR.pdf). The TAiL-PCR protocol consists of three reactions, each of which utilises a PCR primer specific to the template, and an arbitrary degenerate (AD) primer. Eight AD primers, of between 64-fold and 1028-fold degeneracy, were designed based on primers used in previous studies (Liu *et al.*, 1995; Takishita *et al.*, 1999). Each individual combination of PCR template and gene specific primer was tested with each individual AD primer.

For the initial PCR, approximately 1 ng DNA template was mixed with 4 µl 10 X PCR reaction buffer, 30 nmol $MgCl_2$, and 4U GoTaq Flexi polymerase (all Promega), and 2 nmol premixed dNTPs, 3 pmol of the template-specific PCR primer, 80 pmol of the selected AD primer, and 0.4 µl DMSO, in diethylpyrocarbonate-treated water, to a total volume of 20 µl. These reagents were incubated at for 2 minutes at 92 °C, and 1 minute at 95 °C. 5 PCR cycles were then performed of: 30 seconds at 94 °C, 1 minute at 55 °C, and 2 minutes at 72 °C, to amplify sequence using the template-specific PCR primer. The reaction mixture was immediately cooled to 25 °C, and then heated at a rate of 0.4 °C/s to a temperature of 72 °C to allow the annealing of the AD primer to the template. 15 PCR cycles were then performed of: 30 seconds at 94 °C, 1 minute at 55 °C, 2 minutes at 72 °C, 30 seconds at 94 °C, 1 minute at 55 °C, 2 minutes at 72 °C, 30 seconds at 94 °C, 1 minute at 45 °C, and 2 minutes at 72 °C. The reactions were then incubated for a single cycle of 5 minutes at 72 °C, and stored at 4 °C.

For the second PCR, 25 nl of the initial PCR product was mixed with 5 µl 10 X PCR reaction buffer, 37.5 nmol $MgCl_2$, buffer and 5U GoTaq Flexi polymerase, 5 nmol premixed dNTPs, 4 pmol of the second template-specific PCR primer, 100 pmol of the selected AD primer, and 0.5 µl DMSO, in diethylpyrocarbonate-treated water, to a total volume of 25 µl. The second template-specific PCR primer was positioned downstream of the first template-specific PCR primer, to specifically amplify products of the desired template. Reaction conditions were twelve cycles of: 30 seconds at 94 °C, 1 minute at 55 °C, 2 minutes at 72 °C, 30 seconds at 94 °C, 1 minute at 55 °C, 2 minutes at 72 °C, 30 seconds at 94 °C, 1 minute at 45 °C, and 2 minutes at 72 °C. The reactions were then incubated for a single cycle of 5 minutes at 72 °C, and stored at 4 °C.

The final PCR was set up using the same reaction mixture as the second PCR, only using 100 nl of the secondary PCR product as template, and 5 pmol of the third template-specific

PCR, which was positioned downstream of the second template-specific primer. Reaction conditions were 20 cycles of: 30 seconds at 94 °C, 1 minute at 45 °C, and 2 minutes at 72 °C. The reactions were then incubated for a single cycle of 5 minutes at 72 °C, and stored at 4 °C.

Sequencing of PCR products

PCR products were separated by electrophoresis on a TBE gel containing 1% agarose, and .003% ethidium bromide for 30 minutes at 100V, and visualised using a UV transilluminator. PCR products were purified either from the crude PCR products (if only one band were visible on the electrophoresis gel) or from excised gel pieces (if more than one band were visible) using the MinElute gel extraction kit (Qiagen). PCR products were sequenced using an Applied Biosystems 3730xl DNA Analyser, using one of the primers used for the initial PCR amplification.

RT-PCR products generated from circularised or ligated RNA, were purified and ligated into pGEM-T Easy plasmid vector (Promega), following the manufacturer's instructions, and were then introduced into transformation competent *Escherichia coli* DH5α cells. Transformation competent cells were generated via an adapted version of a $MgCl_2$ protocol (D. J. Lea-Smith, pers. comm.). Untransformed cells, taken from a liquid culture grown from a single colony, were grown to mid-log phase in 400 ml LB, supplemented with 0.6 mol $MgCl_2$. Cell growth was arrested on ice, and cells were then collected by centrifugation at 4 °C for 10 minutes, at 4380 x g. Cell pellets were then washed in 100 ml Solution A (containing 0.005 mol $CaCl_2$, 0.001 mol MES, 0.001 mol $MnCl_2$), incubated on ice for a further 20 minutes and collected again by centrifugation at 4 °C for 10 minutes, at 4380 x g. Cell pellets were resuspended in 2 ml Solution A and 300 µl glycerol. 50 µl volumes of the resuspension were aliquotted into sterile 1.5 ml Eppendorf tubes over dry ice. Cell preparations were stored at -80 °C.

For transformation, 5 µl ligation mix was introduced into one aliquot of competent cells, and incubated on ice for 20 minutes. Cells were heat-shocked at 42 °C for 50 seconds, and cooled on ice for two minutes. 600 µl sterile LB was added, and the cells were recovered at 37 °C for 90 minutes, prior to plating on LB-agarose plates containing 100 µg/ ml each ampicillin, X-Gal and IPTG, and incubation overnight at 37 °C. Individual white colonies were picked from each transformant plate, and incubated overnight in LB containing 100 µg/ ml ampicillin. Plasmids were harvested from liquid cultures using a GeneJET miniprep kit (Thermo), per the manufacturer's instructions, and sequenced as before, using primers specific to the pGEM vector sequence.

Generation and assembly of next generation sequencing products.

Double-stranded cDNA was synthesised for next generation sequencing using a Maxima H Minus synthesis kit (Thermo), and a modified cDNA synthesis protocol. 4 µg *Karenia mikimotoi* total cellular RNA, as quantified by a nanodrop photospectrometer, was mixed with 100 pmol of an oligo-d(A) primer previously determined to anneal to polyuridylylated plastid transcripts (Barbrook *et al.*, 2012; Dorrell *et al.*, 2014; Dorrell and Howe, 2012a), and diethylpyrocarbonate-treated water to a total volume of 14 µl. The mixture was incubated at 65 °C for 5 minutes and cooled on ice. The reactants were collected by centrifugation, and 5 µl 4x First Strand synthesis mix and 1 µl First Strand enzyme mix (both Thermo) were added. First strand synthesis was performed at 50 °C for 30 minutes and 85 °C for 5 minutes following the manufacturer's instructions. The reaction products were then cooled on ice, collected by centrifugation, and immediately mixed with 20 µl Second Strand synthesis mix and 5 µl Second Strand enzyme mix (both Thermo), and diethylpyrocarbonate-treated water to a final volume of 100 µl. Second strand synthesis was performed at 16 °C at 60 minutes following the manufacturer's instructions. Reaction products were immediately cleaned with a MinElute spin column (Qiagen) using a guanidine thiocyanate binding buffer, and were eluted in pH 8 Tris-EDTA buffer.

Double stranded cDNA was quantified using a Qubit fluorometer (Invitrogen) following the manufacturer's instructions. A sequencing library was generated from 100 ng purified product using a NexteraXT tagmentation kit (Illumina). The library was sequenced over 500 cycles using a MiSeq sequencer. Reads were trimmed using the Miseq reporter version 2.0.26, and contigs were assembled using ELAND (Illumina) and GeneIOUS version 4.736.

Gene identification in next generation sequencing products

Sequences of potential plastid origin in the *Karenia mikimotoi* next generation sequencing libraries were identified by reciprocal BLAST searches against protein sequences, generated by conceptual translations of plastid genes, from the fucoxanthin dinoflagellate *Karlodinium veneficum* (Gabrielsen *et al.*, 2011; Richardson *et al.*, 2014), the cultured haptophytes *Emiliania huxleyi, Phaeocystis globosa*, and *Pavlova lutheri* (Baurain *et al.*, 2010; Puerta *et al.*, 2005), and the uncultivated haptophyte C19847 (Cuvelier *et al.*, 2010). For *Karlodinium veneficum,* protein sequences were based on the conceptual translation products of published plastid transcript sequences, to account for the effect of transcript editing on protein sequence (Jackson *et al.*, 2013; Richardson *et al.*, 2014).

Initially, a tBLASTn search was performed of the complete read data using protein queries from all five species, using a cut-off threshold of 0.01. These reads were assembled into contigs using GeneIOUS version 4.736, and compared with the entire NCBI database using BLASTx. Only contigs that recovered plastid or cyanobacterial sequences as the first hit were selected for further analysis. Read coverage over each contig was quantified by reciprocal BLASTn alignment of the complete contig sequence against the primary read data.

Additional gene sequences in multigene contigs were identified using BLASTx, and NCBI ORF finder (http://ncbi.nlm.nih.gov/gorf/gorf.html) under the default conditions (Rombel *et al.*, 2002). 5' and 3' UTR sequences were identified using NCBI ORF finder and the Expasy translate servers (http://web.expasy.org/translate/). Transfer RNA genes were identified using the ARAGORN (http://mbio-serv2.mbioekol.lu.se/ARAGORN/) and tRNAscan (http://lowelab.ucsc.edu/tRNAscan-SE/) web servers (Laslett and Canback, 2004; Lowe and Eddy, 1997).

Identification of novel genes in previously published sequences

Previously unannotated genes (*atpE, petG, rps10)* in the *Karlodinium veneficum* plastid, and genes encoding potential plastid-targeted proteins (*psaD, rpl22, rpl23*) in *Karlodinium veneficum* EST libraries, were identified using a similar reciprocal BLAST programme used to inspect the *Karenia mikimotoi* next generation sequencing data. The predicted translation products of every gene annotated in three haptophyte genomes (*Emiliania huxleyi, Phaeocystis globosa, Pavlova lutheri*) (Baurain *et al.*, 2010; Puerta *et al.*, 2005) that had not previously been identified in *Karlodinium veneficum* was searched against published *Karlodinium veneficum* plastid genome and EST sequences using tBLASTn. As before, regions of homology with an expect score below 0.01 were selected, and a reciprocal BLASTx search was performed for each of these sequences against the entire NCBI database. Only regions of sequence that were judged to be homologous to the query genes were selected for further analysis.

Analysis of plastid transcript terminus positions

Poly(U) sites in *Karlodinium veneficum, Chromera velia* and *Vitrella brassicaformis* were identified by aligning the oligo-d(A) RT-PCR products with the most recently published plastid genome sequence of each species, using GENEious (http://www.geneious.com/) (Gabrielsen *et al.*, 2011; Janouškovec *et al.*, 2010; Janouškovec *et al.*, 2013a; Kearse *et al.*, 2012).

To identify putative sequences associated with poly(U) sites in the plastid genomes of *Karlodinium veneficum, C. velia* and *V. brassicaformis*, alignments of every 3' UTR

sequence, and the 100 bp of genomic sequence downstream of each poly(U) site in each plastid genome were constructed. To search for sequences with conserved patterns of purines and pyrimidines, sequences were manually recoded using RY IUPAC nomenclature, as has previously been described (Phillips *et al.*, 2004). Conserved primary sequences were searched by reciprocal BLASTn searches of each sequence against each other sequence within the alignment, and with the Bioprospector (http://robotics.stanford.edu/~xsliu/BioProspector/) (Liu *et al.*, 2001), and Improbizer web servers (http://users.soe.ucsc.edu/~kent/improbizer/improbizer.html) (Siddharthan *et al.*, 2005). GC contents over each transcript sequence were quantified using GeneIOUS. Conserved secondary structures were searched using the WAR web server (http://genome.ku.dk/resources/war/) (Torarinsson and Lindgreen, 2008). The minimum Gibbs free energy of folding of each sequence was calculated using the Mfold server, under the default folding conditions (http://mfold.rit.albany.edu/?q=mfold) (Zuker, 2003).

Analysis of plastid transcript editing

Sequence editing was quantified for *Karenia mikimotoi* and *Karlodinium veneficum* transcripts by GENEious alignments, as before. To determine the effect of transcript editing on protein sequence conservation between *Karlodinium veneficum* transcripts and haptophyte orthologues, the transcript and genomic sequence of each gene in the *Karlodinium veneficum* were aligned to plastid protein sequences from the haptophytes *Emiliania huxleyi* and *Phaeocystis globosa* using BLASTx (Puerta *et al.*, 2005). For each alignment, the number of residues conserved between the *Karlodinium veneficum* and haptophyte protein sequences were recorded.

To determine whether editing events were clustered within certain regions of the *Karlodinium veneficum psaA* and *tufA* transcripts, editing sites across the entire coding sequence of each gene were identified by comparison of transcript and genetic sequences, as detailed above. Editing sites were identified in each alignment, and scored over a 60 bp sliding sequence window, and regions with elevated frequencies of editing relative to the entire CDS were identified by a binomial test. Sequence conservation between the *Karlodinium veneficum* and *E. huxleyi* protein sequences was scored over each window using BLAST alignment, as before.

Analysis of plastid genome sequences

Potential recombination events associated with the *Karlodinium veneficum* plastid were identified by comparison of the complete plastid genome sequence with plastid genomes of the free-living haptophytes *Emiliania huxleyi, Phaeocystis globosa, Pavlova lutheri*, the uncultured prymnesiophyte C19487, and the diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* (Baurain *et al.*, 2010; Cuvelier *et al.*, 2010; Oudot-le-Secq *et al.*, 2007; Puerta *et al.*, 2005). Recombination events were determined to be those where (i) the *K. veneficum* gene order differed from all other plastid lineages, (ii) an identical gene order was found within all the haptophyte genomes (i.e. there was no evidence for recombination events within the haptophytes), and (iii) a similar gene order was found in diatoms and in haptophytes, taking into account differences in coding content between diatom and haptophyte plastid genomes (i.e. the haptophyte gene order is likely to be ancestral) (Oudot-le-Secq *et al.*, 2007; Ruck *et al.*, 2014). Potential recombination events in *Karenia mikimotoi* were identified using similar comparisons.

Potential indels in fucoxanthin plastid genomes were identified by performing alignments of the predicted translation products of *Karenia mikimotoi* and *Karlodinium veneficum* transcript sequences with orthologous plastid protein sequences from seventeen different sequenced algal plastids (haptophytes: *Emiliania huxleyi, Phaeocystis globosa, Pavlova lutheri;* stramenopiles: *Thalassiosira pseudonana, Phaeodactylum tricornutum, Ectocarpus siliculosus, Aureococcus anophageferrens*; cryptomonads: *Guillardia theta; Rhodomonas salina*; red algae: *Cyanidioschyzon merolae, Porphyridium purpureum, Porphyra yezoensis;* green algae and plants: *Chlamydomonas reinhardtii, Ostreococcus tauri, Arabidopsis thaliana, Marchantia polymorpha*; glaucophyte: *Cyanophora paradoxa*). Indels were only recorded if they were not found in any non-dinoflagellate lineage studied. Terminal extensions were only recorded if the complete terminal region of the corresponding CDS had been identified. Indels were recorded as being conserved between *Karenia mikimotoi* and *Karlodinium veneficum* if the insertions or deletions were idiomorphic to each other (i.e. in the same position of the gene sequence, although not necessarily of the same length or sequence).

Putative bacterial promoters in the *C. velia* plastid were identified from the 5' UTR sequences of each plastid gene using the Neural Network Promoter Prediction server (Reese, 2001) (http://www.fruitfly.org/seq_tools/promoter.html). A pilot experiment was performed using the barley plastid genome, for which promoters have been extensively characterised (Berends Sexton *et al.*, 1990; Zhelyazkova *et al.*, 2012), and a cutoff value of 0.8 was selected as identifying the highest number of promoters with a minimal false positive rate.

Targeting prediction of nuclear transcripts

Plastid targeting sequences were identified in the *Karenia mikimotoi rps18* transcript sequence, and in *Karlodinium veneficum* ESTs encoding proteins of predicted plastid function, using the HECTAR, TargetP and ChloroP web servers (Emanuelsson *et al.*, 2007; Emanuelsson *et al.*, 1999; Gschloessl *et al.*, 2008). The predicted translation products of each contig were searched for a bipartite targeting sequence, consisting of a hydrophobic N-terminal signal peptide, followed by a hydrophilic transit peptide that is enriched in positively charged residues relative to the transit peptides of other plastid lineages, as these are characteristic of proteins imported into fucoxanthin plastids (Ishida and Green, 2002; Patron and Waller, 2007).

Phylogenetic analysis

A 32x 1796aa concatenated PsaA/PsbA/PsbC/PsbD phylogeny was assembled using MAFFT version 5.08 (http://mafft.cbrc.jp/alignment/software/) (Katoh *et al.*, 2005), and hand-curated using MacClade (http://macclade.org/macclade.html) . Protein sequences for *Karenia mikimotoi* were defined by the conceptual translation products of polyuridylylated transcript RT-PCR sequences, generated using the Expasy translate server as before. Sites that were absent or gapped in more than two taxa were manually removed from the alignment.

PhyML phylogenies were calculated using the MABL online server (http://www.phylogeny.fr/version2_cgi/one_task.cgi?task_type=phyml). Initially, three different substitution matrices (Dayhoff, JTT, and WAG) were tested with Γ correction (Dereeper *et al.*, 2008). Fast site removal analyses were performed using MacClade and TIGER (Cummins and McInerney, 2011). Bootstrap values for each tree were calculated using 100 replicate phylogenetic analyses with the same model.

RNA separation and transfer for northern blots

For each northern blot, 30 µg total cellular RNA (*Amphidinium carterae*) or 3µg total cellular RNA (*Karenia mikimotoi* and *Chromera velia*) was diluted to 20 µl in formamide, melted at 65 °C for 5 minutes and snap frozen. These samples were separated by electrophoresis on an RNase-free TBE gel containing 1% agarose and 500 mg/ l guanidine thiocyanate, for 90 minutes at 100V. 4 µl DIG-labelled RNA ladder I (Roche), again diluted to 20 µl in formamide, melted and snap frozen, was run alongside as a size marker, and a formamide-only lane was run as a negative control. To confirm RNA integrity during electrophoresis, an additional lane of total cellular RNA was run out, stained post-hoc in ethidium bromide, and visualised with UV.

The northern transfer was set up following the manufacturers' instructions, using a positively charged nitrocellulose membrane (Nytran) overnight. Following the transfer, the membrane was incubated for 30s under a 1200 μE $m^{-2}s^{-1}$ lamp (Stratagene) to crosslink the RNA to the membrane, and washed to remove any residual transfer medium. The compressed gel slice from the transfer was stained with ethidium bromide and visualised with UV as before, to confirm that the RNA had not degraded during the transfer time period.

Generation of northern probes and hybridisation of northern blots

Probes for each northern blot were generated using a DIG Northern Starter kit (Roche), essentially following the manufacturer's instructions. This kit allows the transcription of digoxigenin-labelled RNA probes, complementary to transcripts of interest, from a DNA template consisting of the desired probe sequence that has been fused to a T7 promoter. Probe template sequences were generated by ligating PCR products derived from desired regions of *Amphidinium carterae, Karenia mikimotoi* and *Chromera velia* plastid gene sequences into pGEM-T Easy vector sequence (Promega). Each ligation product was then amplified by PCR, using a primer specific to the insert sequence, and a primer specific to the T7 promoter, to generate products containing the desired insert sequence, and the 69 bp T7 arm of the vector sequence. To visualise sense transcripts, constructs were selected where the insert was fused in an antisense orientation, and for antisense transcripts, constructs were selected where the insert was fused in a sense orientation relative to the T7 promoter, such that they would generate probes complementary to the desired transcripts. Probe sequences are listed in the corresponding chapter.

Crosslinked membranes from each transfer were incubated at 65 °C for one hour in 12 ml DIG Easy hyb solution (Roche). The DIG Easy hyb solution was decanted, and replaced by a further 12 ml Easy hyb solution containing a complementary RNA probe sequence, and hybridised at 65 °C overnight. Probe sequences were generated with a DIG northern starter kit (Roche), essentially following the manufacturer's instructions. 150 ng probe template sequence, prepared as previously described, was mixed with 2.4 μl each of 5 x digoxigenin labelling mix and 5 x transcription buffer, 24 U T7 RNA polymerase (all Roche), and 20 U RNAsin (Promega), with diethylpyrocarbonate-treated water to a final volume of 15 μl. The reaction mixture was incubated at 42 °C for one hour. 15 U RNase-free DNase (Roche) was

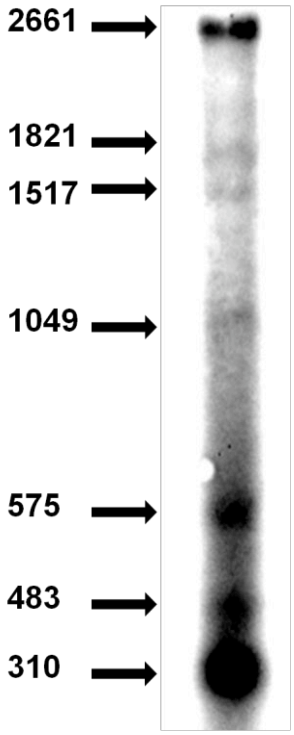


Fig. 2.1. DIG-Labelled RNA ladder I (Roche). This ladder sample was separated by electrophoresis on an RNase-free TBE-agarose gel, and fluorescence was detected using a DIG northern starter kit (Roche) as detailed in this chapter. The arrows correspond to the migration distances of the RNA size markers present.

added, and the reaction mixture was incubated at 37 °C for a further 15 minutes to eliminate any residual DNA, before hybridisation to the blot.

Detection of northern hybridisation.

Hybridised membranes were washed as previously described (Kurniawan, 2013), blocked with BSA medium (Roche) and an HRP-coupled anti-digoxigenin antibody supplied with the Northern Starter kit (Roche) was applied following the manufacturer's instructions. Excess antibody was removed by further washes, and the remaining antibody was activated by incubation in a detection solution, as previously described (Kurniawan, 2013). Antibody labelling patterns were visualised by incubation with a chemiluminescent substrate (CPD-star) in a dark cabinet, following the manufacturer's instructions (Roche).

The cumulative fluorescence signal for each blot was visualised at 30 minute intervals over a twelve hour period, and the image with the clearest hybridisation selected for further analysis. To estimate the sizes of the bands obtained, a logarithmic curve was constructed

using the migration distances of the bands present in the size marker lane. A representative size marker lane under optimal exposure is shown in fig. 2.1.

Sequence deposition

Sequences that had not been identified in any previous study were deposited in GenBank (*Karenia mikimotoi:* JX899682-JX899726, KM065572-KM065732; *Karlodinium veneficum*: KF133369-KF133441, KF135651-KF135653, KF954775-KM954778, KM062161-KM062180 KM065532-KM065533; *C. velia* KC568536-KC568563, KM062122-KM062150; *V. brassicaformis* KC568564-KC618583, KM062113-KM062121). Transcript sequences that were too short to be uploaded to GenBank are listed in Appendix 3.

Chapter Three- Processing of core-containing and antisense transcripts generated from plastid minicircles in the peridinin dinoflagellate *Amphidinium carterae*.

Introduction

Much is known about the content and organisation of plastid genomes (Barbrook *et al.*, 2010; Green, 2011). The plastids of different photosynthetic eukaryotes retain different numbers of genes, with fewer than 100 genes in plant plastids and over 250 in some red algal plastids (Green, 2011). Almost all plastid genomes are organised as a single, circular chromosome, although some may have alternative linear or branched forms (Barbrook *et al.*, 2010; Janouškovec *et al.*, 2013b; Oldenburg and Bendich, 2004). Plastid genes are typically arranged in operons, located downstream of promoters, although recent next generation sequencing surveys in plants have identified additional promoters located at internal positions within predicted operons, and in regions of non-coding plastid DNA (Hotto *et al.*, 2012; Zhelyazkova *et al.*, 2012). Plastid genomes are additionally believed to lack functional terminator elements, with transcription extending far downstream of predicted operons (Rott *et al.*, 1996; Stern and Gruissem, 1987).

The organisation of the plastid genome influences plastid transcript processing events. As plastid genes are arranged in operons, they are cotranscribed before being cleaved into mature mRNAs, which in many cases may be monocistronic, via the activity of 5' and 3' end nucleases (Barkan, 2011; Stern *et al.*, 2010). In the absence of efficient transcription termination, cleavage may delineate the 3' ends of primary transcripts (Rott *et al.*, 1996). A further important role of the plastid transcript processing machinery is the discrimination of sense and antisense transcripts (Georg *et al.*, 2010; Sharwood *et al.*, 2011). In plant plastids, antisense transcripts are generated from promoters located on the reverse, i.e. template, strands of plastid genes (Hotto *et al.*, 2012; Zhelyazkova *et al.*, 2012). At certain loci, antisense transcripts may also be generated as a result of inefficient transcript termination, as the polymerase may extend into and transcribe genes located downstream that are in opposing transcriptional orientation (Rott *et al.*, 1996; Sharwood *et al.*, 2011). Antisense plastid transcripts are deleterious as they anneal to and inhibit the function of the corresponding sense transcripts (Hotto *et al.*, 2010; Zghidi-Abouzid *et al.,* 2011). Thus, they are believed to be preferentially removed from plant plastid transcript pools, for example by the activity of the 5' exonuclease RNAse J (Sharwood *et al.*, 2011).

The plastid genomes of peridinin-containing dinoflagellates are organised in a very different way from those of other plastid lineages. The peridinin plastid genome contains fewer than 20 genes (Barbrook *et al.*, 2013; Howe *et al.*, 2008b) and is fragmented into small circular DNA molecules, termed "minicircles" (Howe *et al.*, 2008b; Zhang *et al.*, 1999). Typically, each minicircle contains a single, complete gene, gene fragments, or no gene whatsoever (Hiller, 2001; Howe *et al.*, 2008b; Iida *et al.*, 2010). Minicircles additionally possess a non-coding "core" region, which is rich in elements predicted to form secondary structures (Moore *et al.*, 2003; Nelson and Green, 2005). The sequence of the core region is broadly conserved across all of the minicircles present in a particular species, and the coding regions of each minicircle are in the same orientation relative to the core sequence (Barbrook *et al.*, 2013; Zhang *et al.*, 2002). Minicircles containing more than one gene have been found in some dinoflagellates (Hiller, 2001; Moszczynski *et al.*, 2012; Nelson and Green, 2005). However, these multigene minicircles are specific to the dinoflagellate species concerned, suggesting that the last common ancestor of the peridinin dinoflagellates possessed a plastid genome in which each gene was located on a separate genetic element (Barbrook *et al.*, 2013; Howe *et al.*, 2008b).

The organisation of the peridinin dinoflagellate plastid genome has influenced the diversity of transcripts produced from each minicircle. In the dinoflagellate *Heterocapsa triquetra*, transcripts that are longer than the underlying minicircle sequence have been identified (Dang and Green, 2010). This is consistent with a "rolling circle" form of transcription, in which a plastid RNA polymerase that has already transcribed a complete minicircle CDS proceeds to transcribe through the minicircle core region, and then transcribes a second, tandem copy of the minicircle CDS (Dang and Green, 2010). However, the predominant bands identified in northern blots of dinoflagellate plastid transcripts correspond to monocistronic mRNAs, suggesting that the majority of transcripts undergo further cleavage (Barbrook *et al.*, 2001; Dang and Green, 2009; Nisbet *et al.*, 2008). This includes 5' end cleavage, as dinoflagellate plastid transcripts can be sequenced through RNA ligase-mediated 5' RACE (5' RLM-RACE) of native RNA, which can identify only those transcripts with processed 5' ends (Dang and Green, 2010; Scotto-Lavino *et al.*, 2006). More unusually, 3' end cleavage of plastid mRNAs involves the addition of a poly(U) tail (Howe *et al.*, 2008b; Wang and Morse, 2006). This processing event has not been reported in any other plastid lineage, except that of the close relative of dinoflagellates, *Chromera velia* (Green, 2011; Janouškovec *et al.*, 2010).

Previous studies of transcript processing in dinoflagellate plastids have predominantly characterised the processing events associated with monocistronic mRNAs in peridinin

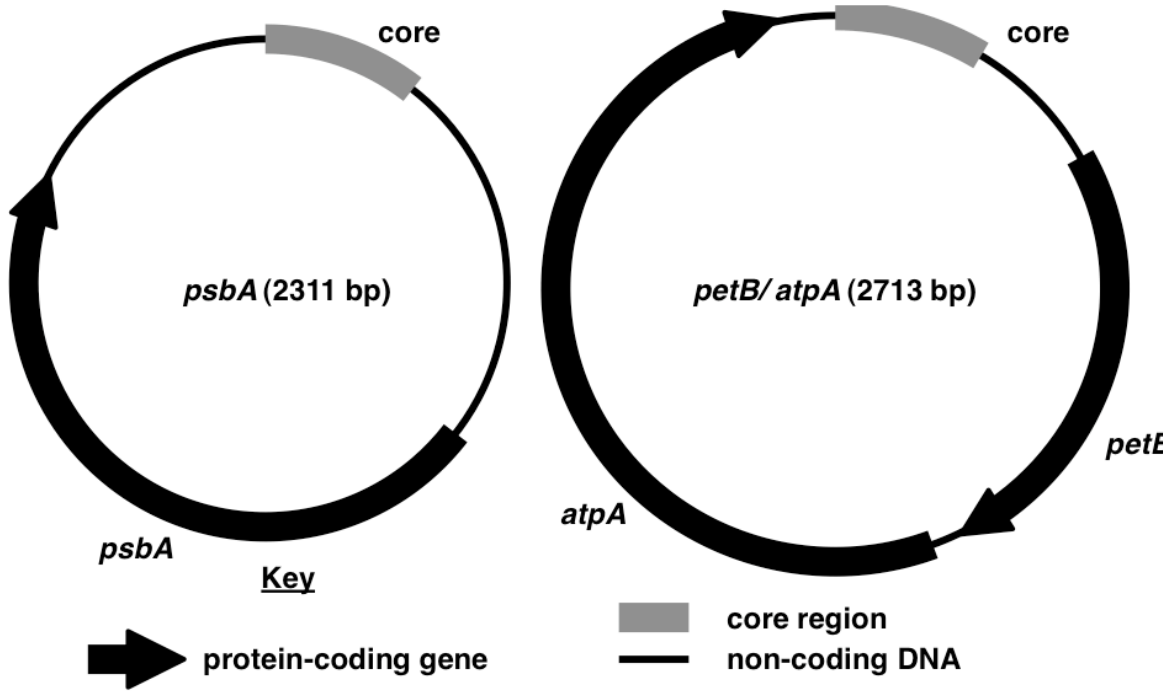


Fig. 3.1: The *Amphidinium carterae psbA* and *petB/atpA* minicircles.

This diagram shows gene maps of the *A. carterae psbA* and *petB/atpA* minicircles, as previously sequenced in: Barbrook and Howe, 2000; Barbrook *et al.*, 2001.

dinoflagellate plastids (Barbrook *et al.*, 2012; Nelson *et al.*, 2007; Wang and Morse, 2006). This project was conceived to characterise the diversity and processing events associated with high molecular weight transcripts, and non-coding transcripts generated from minicircle sequences, about which less was previously known. I have studied non-coding transcripts generated from both strands of the single gene *psbA* minicircle, and the multigene *petB/ atpA* minicircle, in the model peridinin dinoflagellate *Amphidinium carterae* (Fig. 3.1). These minicircles have previously been shown to give rise to highly abundant monocistronic transcripts for each gene, which may possess a poly(U) tail (Barbrook *et al.*, 2012; Barbrook *et al.*, 2001; Nisbet *et al.*, 2008). The 5' termini of the mature mRNAs derived from each minicircle have previously been identified by circular RT-PCR, allowing the direct comparison of processing events associated with non-coding and translationally functional transcripts.

I initially wished to determine whether multi-copy transcripts generated through rolling circle transcription are present in *A. carterae* plastids, as found in *H. triquetra* (Dang and Green, 2010). I identified multi-copy transcripts of both minicircles, suggesting that rolling circle transcription is a widespread feature of peridinin plastid gene expression. I additionally

wished to determine whether multi-copy transcripts undergo similar 5' end cleavage and poly(U) tail addition events to those observed on monocistronic mRNAs in dinoflagellate plastids. I have identified evidence that some multi-copy transcripts are cleaved at the 5' end at the same position associated with monocistronic mRNAs, and that multi-copy transcripts may possess 3' poly(U) tails. This provides clear evidence that multi-copy transcripts undergo similar processing events to mature mRNAs, suggesting that they represent processing precursors in plastid transcript maturation pathways.

I finally wished to determine whether antisense transcripts are present in dinoflagellate plastids. To date, no dinoflagellate minicircle has been identified to contain genes in opposing orientation to each other (Green, 2011; Howe *et al.*, 2008b). Thus, antisense transcripts could not be generated in dinoflagellate plastids via transcriptional run-through between genes, as occurs in plant plastids (Sharwood *et al.*, 2011). I have identified antisense transcripts from both the *psbA* and *petB/ atpA* minicircles. This constitutes the first report of antisense transcripts in an algal plastid, and indicates that both strands of minicircle sequences are transcribed. The antisense transcripts are low in abundance, undergo different terminal cleavage events from the corresponding sense transcripts, and universally lack poly(U) tails. This indicates that poly(U) tails are specifically added to coding transcripts in dinoflagellate plastids. The absence of poly(U) tails from non-coding RNA, such as antisense transcripts, might indirectly enable their degradation during plastid transcript processing.

Results

Rolling circle transcription occurs in *Amphidinium carterae* plastids

I wished to determine whether multi-copy transcripts, and other transcripts generated through rolling circle transcription, are present in *A. carterae,* as have been documented in the dinoflagellate *Heterocapsa triquetra* (Dang and Green, 2010). To do this, cDNA was synthesised using primers positioned within the 5' end of the CDS of the *psbA* and *petB/atpA* minicircles (Table 3.1). These cDNA preparations were then used as templates for PCRs using combinations of primers flanking the core region of each minicircle. This would amplify transcripts that specifically contain core and CDS regions of each minicircle, which must have been generated through rolling circle transcription (Fig. 3.2, panel A, PCR i; Table 3.1). Additional PCRs were performed using each cDNA template, and pairs of PCR primers positioned further upstream of the cDNA synthesis primer annealing site, to amplify sequences associated with transcripts containing the CDS upstream of the core, and

Table 3.1. Primers for RT-PCR to identify multi-copy transcripts

The annealing site of the 5' end of each primer is given relative to the sequence of the corresponding minicircle, where position 1 corresponds to the 5' end of the core region.

Minicircle	Accession	Size	cDNA primer	Annealing site
psbA	AJ250262.1	2311 bp	AGTTAGAGCGAATAAGGCTTG	894 R
PCR	**PCR forward primer**	**Annealing site**	**PCR reverse primer**	**Annealing site**
1	CGAGTCAGAGGCATCAAAC	264 F	AGTTAGAGCGAATAAGGCTTG	894 R
2	TACATTGAGTAGGCATCTTTAATAGC	547 F	AGTTAGAGCGAATAAGGCTTG	894 R
3	CTGGGGTTCTTTCGTTCAAAC	896 F	GATACCAATTACAGGCCAAGC	1706 R
4	TACATTGAGTAGGCATCTTTAATAGC	547 F	GATACCAATTACAGGCCAAGC	1706 R
5	CGAGTCAGAGGCATCAAAC	264 F	TGCAGGAGCAAGGAAGAAAG	1028 R
6	ACGCTCATAACTTCCCTCTTG	1861 F	TGCAGGAGCAAGGAAGAAAG	1028 R
7	ACGCTCATAACTTCCCTCTTG	1861 F	GATACCAATTACAGGCCAAGC	1706 R
Minicircle	**Accession**	**Size**	**cDNA primer**	**Annealing site**
petB/atpA	AY048664.1	2713 bp	GCAACTCAAGACGCTCTTCAC	629 R
PCR	**PCR forward primer**	**Annealing site**	**PCR reverse primer**	**Annealing site**
1	GGTCTTCTTGGGTTATTTCC	2611 F	GCAACTCAAGACGCTCTTCAC	629 R
2	TCAGTCTGTCTGCGAACCAC	1660 F	CCTTTCCGTATCCTTCATTCG	97 R
3	TCAGTCTGTCTGCGAACCAC	1660 F	TCGTTCAACCACACTTTATACAGAAC	25 R
4	GCAGACGATATCCTCTCTAAG	638 F	ACAAGGCCATATACGACATC	1407 R
5	TCCCGATCTCACAAGTCTCC	337 F	ACAAGGCCATATACGACATC	1407 R
6	CGAATGAAGGATACGGAAAGG	77 F	ACAAGGCCATATACGACATC	1407 R
7	GTTCTGTATAAAGTGTGGTTGAACGA	5 F	ACAAGGCCATATACGACATC	1407 R
8	GGTCTTCTTGGGTTATTTCC	2611 F	ACAAGGCCATATACGACATC	1407 R
control 1	ATGGTTGTTCGTCTTCCTTATGTC		CACTCTGAGGTAGACGAGACAGC	769 bp
control 2	CTCTTGAACTACTCATGGAAGC		ATCTCAAGAGGAGTTGCAAAC	1892 bp

transcripts containing two or more complete core regions (Fig. 3.2, panel A ,PCRs ii and iii respectively; Table 3.1). Products were obtained of the expected size for transcripts covering 5' UTR or core sequences (Fig.3.2, panel B: image i, lanes 1-2; image ii, lane 1), and transcripts extending into the upstream CDS for each minicircle (Fig. 3.2, panel B: image i, lanes 3-4; ii, lanes 2-5). To exclude the possibility that these were chimeric, the identity of representative products of expected size from each minicircle (Fig. 3.2, panel B: image i, lane 3; ii, lane 1) was confirmed by sequencing.

Low intensity bands consistent with transcripts containing two core regions were recovered for the *petB/atpA* minicircle, and the identity of these products was confirmed by sequencing (Fig. 3.2, panel B; image ii, lanes 7-8). Although products were recovered for comparable PCRs for the *psbA* minicircle (Fig. 3.2, panel B; image i, lanes 6-7), these were not of the expected size, and were determined from sequencing to be PCR chimeras. Although the PCRs tested with these reactions were generally of larger expected size than those

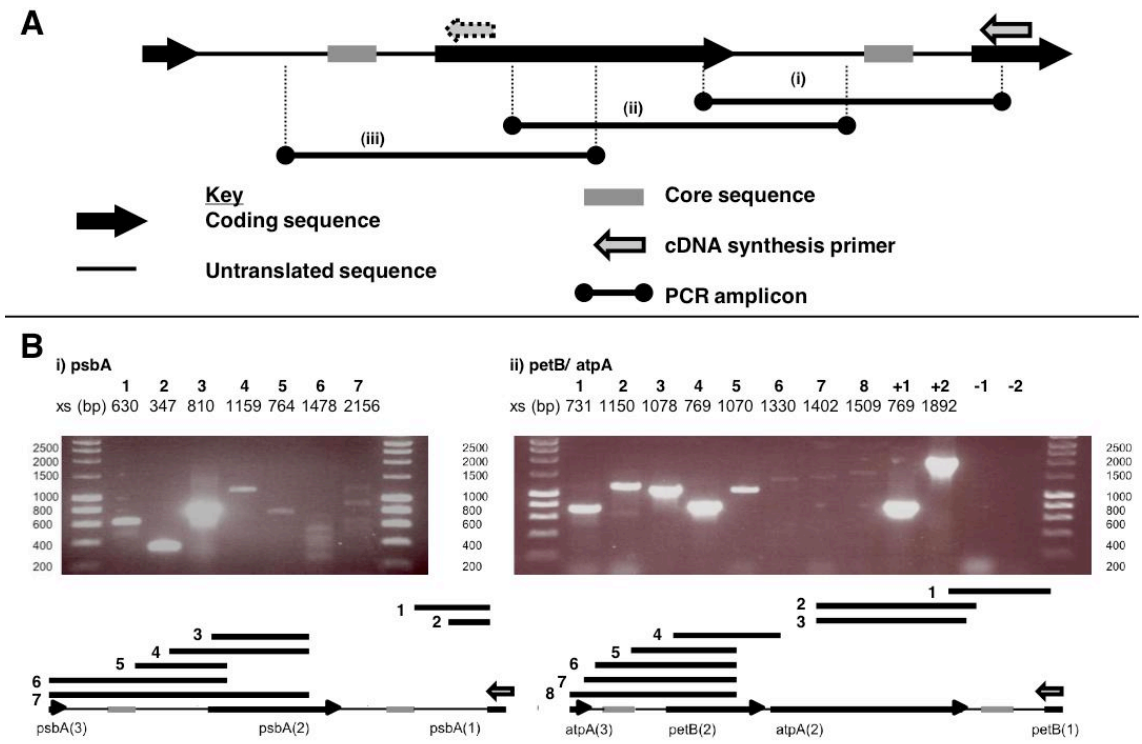


Fig. 3.2: Rolling circle transcription in *Amphidinium carterae* plastids.

Panel A shows a diagram of the PCRs employed to detect transcription over the core regions of the *A. carterae psbA* and *petB/atpA* minicircles. Minicircle sequences are shown as in Fig. 3.1. cDNA was generated using primers specific to the 5' ends of minicircle genes. PCR was performed with various combinations of primers to amplify sequences specific to **(i)** core-containing transcripts; **(ii)** transcripts containing the CDS upstream of the core as well as the CDS downstream; and **(iii)** transcripts containing two or more core regions. The dotted arrow shows sites where the cDNA primer could anneal but which would not generate cDNA that could serve as a template for PCR **(ii)** and **(iii)**.

Panel B shows gel photographs of the RT-PCRs performed for the *psbA* **(i**) and *petB/atpA* **(ii)** minicircles. The size marker displayed is DNA Hyperladder I (Bioline); the sizes of markers up to 2.5 kb are shown to the left of each gel photograph. A transcript diagram is shown beneath each gel photograph, consisting of a hypothetical linearised multi-copy transcript for each minicircle, shaded as in Panel A. The transcript diagrams additionally show the regions of sequence amplified in each PCR. These are numbered with the corresponding lane in the gel photograph. Lanes **1–8** in each gel photograph show the results of RT-PCRs of minicircle transcripts. Lanes **+1** and **+2** represent PCR positive controls, using an *A. carterae* DNA template, and amplifying short and long regions of the *psbA* minicircle. Lane **-1** is a negative control lacking reverse transcriptase, and Lane **-2** a template negative control for the PCR.

designed to identify lower molecular weight transcripts, the reduction in yield is unlikely to be due to lower PCR efficiency, as a positive control to amplify a 1.9 kbp long product from a *psbA* cDNA template yielded abundant product (Fig. 3.2, panel B; image ii, lane +2). Furthermore, the difference in product abundance is unlikely to be due to decreased reverse transcriptase efficiency with longer RNA templates, as a continuous decline in product abundance was not observed. Instead, for both minicircles, a sharp decline in product abundance was observed, associated with transcripts containing a 5' UTR upstream of the core region (Fig. 3.2, panel B; compare image i, lanes 4-5; ii, lanes 5-6). It is thus likely that transcripts containing two or more core regions are present only at extremely low copy numbers in *A. carterae*.

Multi-copy transcripts can receive poly(U) tails

Previously, it has been suggested that poly(U) tails are added in dinoflagellates during the concerted cleavage of transcript precursors containing multiple copies of minicircle sequence into mature mRNAs (Dang and Green, 2010). However, polyuridylylated polycistronic transcripts have been identified from the multigene *petB/atpA* and *psbD/psbE/psbI A. carterae* minicircles (Barbrook *et al*., 2012; Nisbet *et al*., 2008). This indicates that poly(U) tails may be added prior to the processing of polycistronic precursors into monocistronic mRNAs. I wished to determine whether poly(U) tails are ever added to multi-copy transcripts of more than one minicircle length.

To test for the existence of polyuridylylated multi-copy transcripts for the *petB/atpA* and *psbA* minicircles, cDNA was synthesised using primers containing an oligo-d(A) region, which would anneal to poly(U) tails in transcript sequence. This technique has previously been used to identify polyuridylylated mRNA in *A. cartera*e (A. C. Barbrook, pers. comm.) (Barbrook *et al*., 2012). Each cDNA synthesis primer was designed to contain an oligo-d(A) sequence at its 5' end, and a sequence at the 3' end complementary to the 3' UTR immediately upstream of the poly(U) sites previously identified on monocistronic *petB, atpA* and *psbA* transcripts (Table 3.2) (A.C. Barbrook, pers. comm..) (Barbrook *et al*., 2012). Each primer would therefore specifically anneal to the polyuridylylated transcripts of one gene only. The predicted annealing temperatures of the complementary region between each cDNA synthesis primer and the corresponding transcript 3' UTR was less than 20˚C, and therefore the cDNA primers could not have annealed to non-polyuridylylated transcripts during reverse transcription. PCR amplifications of the generated cDNA were performed using PCR primers that flanked the poly(U) site, to identify transcripts that contained a second copy of the minicircle CDS (Fig. 3.3, panel A; Table 3.2).

Table 3.2. RT-PCR to detect polyuridylylated multi-copy transcripts

Primer annealing positions are given relative to the sequence of the corresponding minicircle, where position 1 corresponds to the 5' end of the core region, as per Table 3.1.

Gene	cDNA primer	Annealing site
psbA	AAAAAAARATAAAGGGG	1870/ 1872 R
petB	AAAAAAAWAAGAATAGAAGT	1123/ 1226 R
atpA	AAAAAAAAAAAAAAAAAAATATACAGAAC	2592 R

Gene	PCR forward primer	Annealing site	PCR reverse primer	Annealing site
psbA	CAAGCCTTATTCGCTCTAACT	838 F	ATCGTTAATCAGAAAGCCTAGTC	1918 R
petB	GCAGACGATATCCTCTCTAAG	507 F	ACAAGGCCATATACGACATC	1276 R
atpA	TCAGTCTGTCTGCGAACCAC	1529 F	CCTTTCCGTATCCTTCATTCG	2679 R

For all three genes, products were identified consistent with the presence of polyuridylylated multi-copy transcripts (Fig. 3.3, panel B; lanes 1, 3, 5). Products could not be identified for any PCR under reverse transcriptase negative conditions, confirming that these products were not due to residual gDNA contamination (Fig. 3.3, panel B; lanes 2, 4, 6). To confirm that the cDNA primers employed annealed specifically to polyuridylylated transcripts of one gene, PCRs were performed using template generated with the *psbA* cDNA primer and the *petB* PCR primers, and using template generated with either the *petB* or *atpA* primer and the *psbA* PCR primers (Fig. 3.3, panel B; lanes 9-11). Products could not be identified in any case, confirming that the cDNA primers used were specific to the intended template, and were not annealing to other transcripts in the RNA samples at detectable levels. Thus, poly(U) tails can be added to transcripts of more than one minicircle length.

Multi-copy transcripts can possess mature 5' ends

A previous study inferred from 5' RACE data that multi-copy transcripts in *Heterocapsa triquetra* possess different 5' ends from monocistronic mRNAs (Dang and Green, 2010). Multi-copy transcripts identified in this study were found to have 5' termini upstream of the mature 5' terminus, at a position corresponding to the poly(U) site (Dang and Green, 2010). This study additionally detected transcripts with the same 5' end as mature mRNAs, which might be derived from multi-copy transcripts (Dang and Green, 2010). However, as 5' RACE only identifies transcript 5' ends instead of complete transcript sequences, and the cDNA synthesis primers used in this study were positioned within the CDS of each minicircle, it was not possible to determine whether these transcripts corresponded to multi-copy precursors, or instead were monocistronic (Dang and Green, 2010).

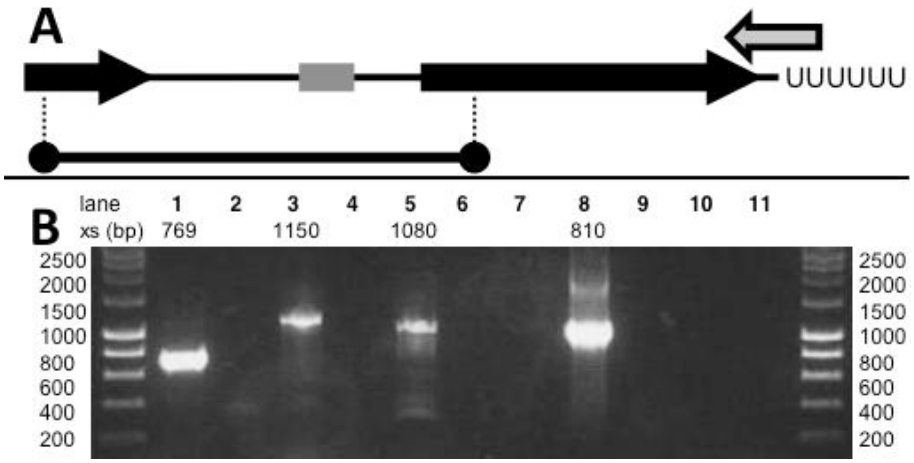


Fig. 3.3: Presence of polyuridylylated multi-copy transcripts.

Panel A shows a diagram of the PCRs employed to detect polyuridylylated multi-copy transcripts from the *A. carterae psbA* and *petB/atpA* minicircles. Transcript sequences and PCRs are drawn as before. cDNA synthesis primers were designed containing a 5' oligo-d(A) region, and a 3' region complementary to the 3' UTR sequence region directly upstream of either the *psbA, petB* or *atpA* poly(U) site. PCRs were performed as before using the cDNA preparations generated using these synthesis primers, and pairs of PCR primers that flank the cDNA primer annealing site. These reactions will specifically amplify polyuridylylated transcripts that additionally contain a second copy of minicircle sequence, in which the 3' UTR sequence is complete, and not interrupted by the poly(U) tail.

Panel B shows a gel photograph demonstrating the presence of polyuridylylated multi-copy transcripts. The gel photograph and size markers are shown as in Fig. 2. Lane **1**: RT-PCR to detect polyuridylylated multi-copy transcripts from the *petB/atpA* minicircle using a cDNA synthesis primer specific to the *petB* poly(U) site. Lane **2**: reverse transcriptase negative control for lane 1. Lane **3**: RT-PCR to detect polyuridylylated multi-copy transcripts using a cDNA synthesis primer specific to the *atpA* poly(U) site. Lane **4**: reverse transcriptase negative control for lane 3. Lane **5**: RT-PCR to detect polyuridylylated multi-copy transcripts from the *psbA* minicircle. Lane **6**: reverse transcriptase negative control for lane 5. Lane **7**: blank lane. Lane **8**: reaction positive control, using a genomic DNA template and PCR primers internal to the *psbA* CDS. Lane **9-10**: RT-PCRs using the *petB* and *atpA* poly(U) site cDNA synthesis primers, and PCR primers internal to the *psbA* CDS, confirming the minicircle specificity of cDNA synthesis.

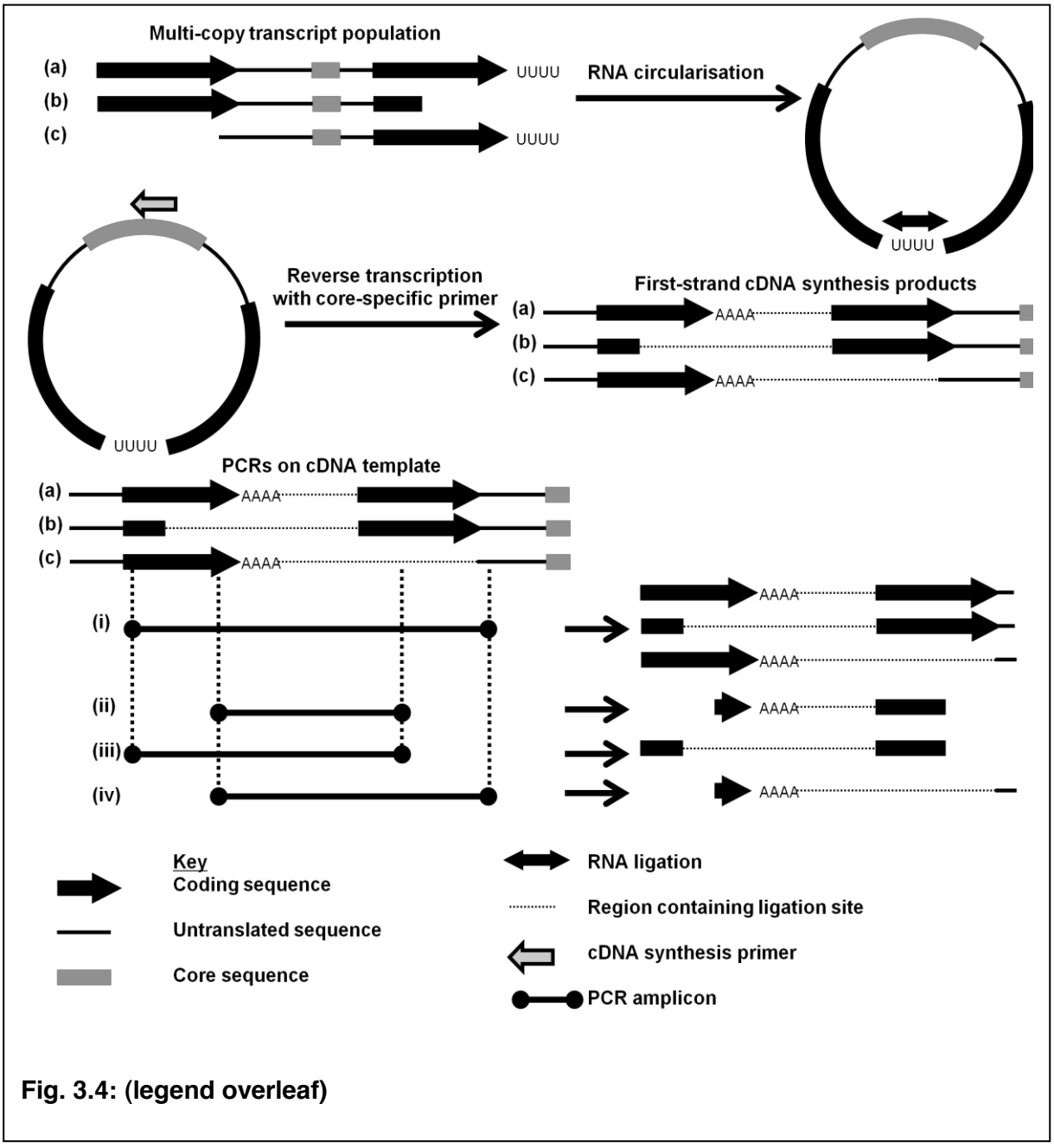


Fig. 3.4: (legend overleaf)

I wished to determine whether any of the multi-copy transcripts in the *Amphidinium carterae* plastid possessed mature 5' ends. I additionally wished to determine whether multi-copy transcripts containing alternative 3' end positions to the poly(U) sites previously identified by oligo-d(A) primed RT-PCR were present. To determine the 5' terminus positions associated with multi-copy transcripts, RT-PCRs were performed using circularised total cellular RNA (Fig. 3.4). This technique allows the simultaneous identification of transcript 5' and 3' ends

Fig. 3.4: Circular RT-PCR of core-containing minicircle transcripts.

This figure (**shown on previous page**) shows a diagram of the circular RT-PCR protocol used to map the termini of transcripts from the *A. carterae psbA* and *petB/atpA* minicircles. A heterogeneous population of transcripts, including multi-copy transcripts with mature 5' ends **(a, b)**, 5' ends that terminate within the upstream UTR **(c)**, and transcripts with poly(U) tails **(a, c)**, or that terminate at the 3' end upstream of the poly(U) site **(b)**, is treated with T4 RNA ligase, generating circularised RNA. RNA is not pretreated to remove the 3' triphosphate caps from primary transcripts, and therefore only transcripts that have undergone prior 5' cleavage are ligated. The circular RNA is reverse transcribed using cDNA synthesis primers specific to minicircle core regions, generating cDNA specifically from the different core-containing transcripts present.

PCRs are then performed using the cDNA preparations using different combinations of primers. Four schematic PCR primer combinations are shown. Primer combination **(i)** consists of a PCR forward primer located at the 5' end of the minicircle CDS, and a PCR reverse primer located in the 3' UTR of the minicircle. This PCR combination will amplify all three transcripts (a, b, and c) shown. Transcripts in which the PCR primers anneal far from the ligation site (e.g. transcript a) may not be amplified by these primers, due to outcompetition by transcripts in which the ligation site is closer to the PCR primer annealing sites. To allow a greater diversity of transcripts to be mapped, additional PCRs are performed with PCR primers in different positions. For example, transcript (a), which possesses a mature 5' end and a poly(U) tail, will be preferentially amplified by PCR reaction **(ii)**, which utilises a PCR reverse primer positioned at the 5' end of the minicircle CDS, and a PCR forward primer positioned within the CDS 3' end. Similarly, transcript (b), which possesses a mature 5' end but terminates upstream of the poly(U) site will be amplified by the PCR reverse and forward primers positioned within the CDS 5' end **(iii)**,

(Barbrook *et al.*, 2012). It is therefore possible to infer the complete sequences of individual transcripts and determine whether individual transcripts are of greater than one minicircle length. Reverse transcriptions were performed using cDNA synthesis primers specific to core regions of the *psbA* and *petB/atpA* minicircles (Fig. 3.4; Table 3.3). To identify the full diversity of 5' and 3' termini associated with multi-copy transcripts, five different PCR forward, and five different PCR reverse primers were designed against different regions of each minicircle sequence (Fig. 3.4; Table 3.3). For example, for the *psbA* minicircle, a PCR reverse primer was designed specific to the 5' end of the *psbA* CDS, which would preferentially amplify multi-copy transcripts with mature 5' termini. Three further reverse

Table 3.3. Primers for circular RT-PCR of multi-copy transcripts

Primer annealing positions are shown relative to the sequence of the corresponding minicircle, where position 1 corresponds to the 5' end of the core region, as per Table 3.1

Minicircle			core-specific cDNA primer	Annealing site
psbA			AGTCTCCCGATTGTCTATTCTC	41 R
	PCR forward primers	**Annealing site**	**PCR reverse primers**	**Annealing site**
1	CGAGTCAGAGGCATCAAAC	228 F (core)	CTTTAGACTGCGGTGTGAAC	563 R (5' UTR)
2	TACATTGAGTAGGCATCTTTAATAGC	512 F (5' UTR)	AGTTAGAGCGAATAAGGCTTG	858 R (CDS 5' end)
3	CTGGGGTTCTTTCGTTCAAAC	860 F (CDS 5' end)	GATACCAATTACAGGCCAAGC	1670 R (CDS 3' end)
4	ACGCTCATAACTTCCCTCTTG	1825 F (CDS 3' end)	ATCGTTAATCAGAAAGCCTAGTC	1918 R (3' UTR)
5	CCTCCTACCGAAAGTCAATTC	2238 F (3' UTR)	ATTGACTTTCGGTAGGAGGC	2256 R (3' UTR)
Minicircle	**positive control cDNA primer**	**Annealing site**	**core-specific cDNA primer**	**Annealing site**
petB/ atpA	GCATTGCTGTGGAATAGAC	2417 R	CCTTTCCGTATCCTTCATTCG	2679 R
	PCR forward primers	**Annealing site**	**PCR reverse primers**	**Annealing site**
1	GAAAATCCAGGTCATATCATAGGAG	133 F (core)	GCAACTCAAGACGCTCTTCAC	498 R (petB 5' end)
2	GCAGACGATATCCTCTCTAAG	507 F (petB 5' end)	CAAACACTGTACCCAACGAAG	963 R (petB 3' end)
3	CCTTCTCCTTACTCATTTCCTAATG	1058 F (petB 3' end)	ACAAGGCCATATACGACATC	1276 R (atpA 5' end)
4	TCAGTCTGTCTGCGAACCAC	1529 F (atpA 5' end)	CTTCTGACCCACAGGGACAT	1715 R (atpA 3' end)
5	GGTCTTCTTGGGTTATTTCC	2480 F (atpA 3' end)	CCTTTCCGTATCCTTCATTCG	2679 R (3' UTR)
control	GGTCTTCTTGGGTTATTTCC	2480 F (atpA 3' end)	ACAAGGCCATATACGACATC	1276 R (atpA 5' end)

primers were designed specific to non-coding regions of the *psbA* minicircle, upstream of the *psbA* mature transcript 5' terminus position that would preferentially amplify multi-copy transcripts containing extensive UTR sequence. A final reverse primer was designed specific to the 3' end of *psbA* that would preferentially amplify transcripts with 5' termini located within the CDS (Table 3.3). PCRs were then performed using each possible combination of forward and reverse primer, to amplify the ligated termini of transcripts covering minicircle core regions (Fig. 3.4; Table 3.3). Each circular RT-PCR was repeated three times using independently isolated RNA samples, to identify the full range of transcripts present. As a positive control, cDNA was synthesised using a primer positioned internal to the *atpA* CDS, and PCRs were performed with outward-directed primers for the *atpA* gene (Table 3.3).

A small number of core-containing transcripts were identified for each minicircle through circular RT-PCR. The terminus positions of these transcripts are listed in Table 3.4, along with the specific combination of PCR primers used to identify each transcript. Five core-containing *petB/ atpA* transcripts, and one core-containing *psbA* transcript, were predicted to

Table 3.4. Core-containing transcripts identified through circular RT-PCR.

This table lists the multi-copy transcripts and core-containing transcripts of less than one minicircle length mapped for each minicircle. The terminus positions and primer combinations used for each circular RT-PCR product are shown relative to the sequence of the corresponding minicircle, where position 1 corresponds to the 5' end of the core region, as per Table 3.1. For reference, the terminus positions of the minicircle core region and CDS are given. In addition, the consensus terminus positions associated with monocistronic polyuridylylated *psbA*, *petB* and *atpA* transcripts, as identified in a previous study are shown (Barbrook et al., 2012). Two multi-copy *atpA* transcripts that terminate at a similar 5' end position to that observed for monocistronic transcripts are shown in bold text.

1. psbA	**Minicircle length**		2311 bp	**Core**		1-281
	CDS		**monocistronic transcript**			
	5' end	**3' end**	**5' end**	**poly(U) site**		
psbA	834	1856	600-829	1870-1872		
Transcripts	**5' end**	**3' end**	**R primer**	**F primer**	**Poly(U)**	**Length (bp)**	**Notes**
multi-copy transcript 1	1407	2228 (2)	1670 R	1825 F	0	3132	
core-containing transcript 1	1310	108 (2)	1670 R	2238 F	0	1109	
core-containing transcript 2	1347	146 (2)	1670 R	1825 F	0	1110	
core-containing transcript 3	1371	67 (2)	1918 R	2238 F	0	1007	
core-containing transcript 4	1416	257 (2)	1670 R	2238 F	0	1152	
core-containing transcript 5	1416	257 (2)	1669 R	1825 F	0	1152	
core-containing transcript 6	1430	202 (2)	1670 R	2238 F	0	1083	
core-containing transcript 7	1542	72 (2)	1670 R	2238 F	0	841	
core-containing transcript 8	1549	76 (2)	1670 R	1825 F	0	838	
core-containing transcript 9	1806	201 (2)	1918 R	2238 F	0	706	
core-containing transcript 10	1843	225 (2)	1918 R	2238 F	0	693	
core-containing transcript 11	2127	328 (2)	2256 R	228 F	0	512	
core-containing transcript 12	2172	262 (2)	2256 R	228 F	0	401	
core-containing transcript 13	2188	262 (2)	2256 R	228 F	0	385	

2. petB/ atpA	**Minicircle length**		2713 bp	**Core**		1-281
	CDS		**monocistronic transcript**			
	5' end	**3' end**	**5' end**	**poly(U) site**		
petB	456	1115	310-424	1122-1126		
atpA	1206	2582	1081-1088	2591		
Transcripts	**5' end**	**3' end**	**R primer**	**F primer**	**Poly(U)**	**Length (bp)**	**Notes**
atpA mRNA 1	1086	2591	1276 R	2480 F	24	1529	
atpA mRNA 2	1086	2591	1276 R	2480 F	32	1537	
atpA mRNA 3	1087	2591	1276 R	2480 F	26	1530	
atpA mRNA 4	1087	2591	1276 R	2480 F	35	1539	
atpA mRNA 5	1089	2591	1276 R	2480 F	26	1528	
multi-copy transcript 1	501	1276 (2)	963 R	1058 F	0	3488	
multi-copy transcript 2	**1080**	**2246 (2)**	**1276 R**	**1529 F**	**0**	**3879**	**Possesses mature 5' end**
multi-copy transcript 3	**1085**	**1861 (2)**	**1276 R**	**1529 F**	**0**	**3489**	**Possesses mature 5' end**
multi-copy transcript 4	1151	1919 (2)	1276 R	1529 F	0	3481	
multi-copy transcript 5	1151	1919 (2)	1276 R	1529 F	0	3481	
core-containing transcript 1	1157	277 (2)	1276 R	2480 F	0	1833	
core-containing transcript 2	1224	76 (2)	1276 R	2480 F	0	1565	
core-containing transcript 3	1224	77 (2)	1276 R	2480 F	0	1566	
core-containing transcript 4	1461	14 (2)	1715 R	2480 F	0	1266	
core-containing transcript 5	1483	603 (2)	1715 R	2480 F	0	1833	
core-containing transcript 6	1526	295 (2)	1715 R	2480 F	0	1482	
core-containing transcript 7	1562	74 (2)	1715 R	2480 F	0	1225	

be greater than one minicircle length (*petB/ atpA*: 2713 bp, *psbA:* 2311 bp). These correspond to multi-copy transcripts (Table 3.4). None of the multi-copy transcripts identified by circular RT-PCR was polyuridylylated. The multi-copy *psbA* transcript terminated downstream of the *psbA* poly(U) site, in the 3' UTR, whereas all of the 3' termini associated with multi-copy *petB/ atpA* transcripts were located within the *atpA* CDS (Table 3.4). In contrast, all of the monocistronic *atpA* transcripts identified in the positive control reaction were polyuridylylated (Table 3.4), consistent with *atpA* transcripts identified in previous circular RT-PCR studies (A.C. Barbrook, pers. comm.) (Barbrook *et al.*, 2012).

Surprisingly, two of the multi-copy transcripts from the *petB/atpA* minicircle had similar 5' ends to those identified for monocistronic, polyuridylylated *atpA* mRNAs, located 340 and 345 bp upstream of the *atpA* CDS (Table 3.4). Similar 5' end positions for monocistronic *atpA* transcripts have also been identified in previous circular RT-PCR studies (A.C. Barbrook, pers. comm.) (Barbrook *et al.*, 2012). Thus, some multi-copy transcripts undergo similar 5' cleavage events as monocistronic mRNAs. Taken together with the presence of poly(U) tails on some multi-copy transcripts, as identified by oligo-d(A) RT-PCR, it appears that multi-copy transcripts may undergo similar terminal processing events to mature transcripts in dinoflagellate plastids.

Short core-containing transcripts are present in dinoflagellate plastid RNA pools

In addition to the multi-copy transcripts identified by circular RT-PCR, thirteen transcripts were identified from the *psbA* minicircle, and seven transcripts were identified from the *petB/atpA* minicircle, that covered part of the core region, but were predicted to be of less than one minicircle length. All of the core-containing transcripts from the *psbA* minicircle terminated at the 5' end within the *psbA* CDS, and all of the core-containing transcripts from the *petB/atpA* minicircle terminated at the 5' end either within the *atpA* CDS, or within the *atpA* 5' UTR, downstream of the 5' ends of the monocistronic, polyuridylylated *atpA* mRNAs identified by circular RT-PCR (Barbrook *et al.*, 2012) (Table 3.4). Only one of the 5' ends associated with these transcripts was conserved between more than one sequence (Table 3.4). All of the core-containing *psbA* transcripts, and all but two of the core-containing *petB/atpA* transcripts identified terminated at the 3' end within the minicircle core region (Fig. 3.5; Table 3.4). PCR amplifications using forward primers specific to the 5' end of the *psbA* and *petB* genes failed to recover any end ligation sequences regardless of PCR reverse primer position, suggesting that the majority of the short core-containing transcripts from each minicircle do not extend into the downstream CDS.

Table 3.5 Northern blot probes to detect sense *A. carterae* plastid transcripts.

This table lists the sequence of the T7 arm of the pGEM-T Easy vector, alongside the first 50 bp of each probe sequence complementary to minicircle sequence. The positions covered by each probe are shown relative to the sequence of the corresponding minicircle, where position 1 corresponds to the 5' end of the core region, as per Table 3.1.

Probe	Start	End	Sequence
T7 arm			TAATACGACTCACTATAGGGCGAATTGGGCCCGACGTCGCATGCTCCCGGCCGCCATGGCCGCGGGATT
psbA			
5' UTR	852	510	AGCGAATAAGGCTTGTCATAATAGATATGGAAAAAGGTTTACGCTGGCCT…
5' CDS	1077	838	AAGAAGGAATAACAGCACCAGAGATGATGTTGTTACCATAGATAAGGGAA…
3' CDS	1845	1616	CAAGAGGGAAGTTATGAGCGTTACGCTCGTGCATTACCTCGATACCAAGA…
3' UTR	2256	1896	ATTGACTTTCGGTAGGAGGCTCAAAGAAGGAAAGGCACTATTACCTAACC…
petB/ atpA			
5' UTR	498	206	GCAACTCAAGACGCTCTTCACACCAATCGTAAATGAAACCCATGTAGAGA…
petB	961	504	AACACTGTACCCAACGAAGGGAACACGTTATTGAGAGCCTCAGGAGTCGC…
atpA 5'	1715	1531	CTTCTGACCCACAGGGACATAGACGGATAGAACCTTCTCGTACTTAAGGT…
atpA 3'	2525	2190	TAGGAAACACCAAGACGGTAGGCTAAGGAAATAACCCAAGAAGACCTTCA…

Core-containing transcripts are present at low abundance in *A. carterae* plastids

It has been shown through quantitative RT-PCR that UTR sequences from the *Amphidinium carterae psbA* minicircle are present at no more than one-fiftieth the abundance of that of the *psbA* CDS in plastid RNA pools (Nisbet *et al.*, 2008). It has also been inferred from northern blotting studies that transcripts of one minicircle length or greater are much less abundant than mature mRNAs (Dang and Green, 2010; Nisbet *et al.*, 2008). This is consistent with data from plastids in other lineages, including *Chromera velia*, a close relative of peridinin dinoflagellates, which indicate that long polycistronic transcripts, and transcripts covering non-coding regions of sequence accumulate to much lower abundance than mature mRNAs (Barkan *et al.* 1994; Janouškovec *et al.*, 2013b; Zhelyazkova *et al.*, 2012).

I wished to determine whether any of the core-containing transcripts identified by circular RT-PCR are abundant in *A. carterae* plastids. To do this, northern blots of *A. carterae* RNA were hybridised to single-stranded RNA probes complementary to sense transcripts from the *psbA* and *petB/atpA* minicircles. Although previous northern blotting studies in dinoflagellates have not detected substantial levels of non-coding transcripts, the probes used in those studies were derived from the minicircle CDS, and may not have detected low molecular weight transcripts that cover regions of non-coding minicircle sequence (Dang and

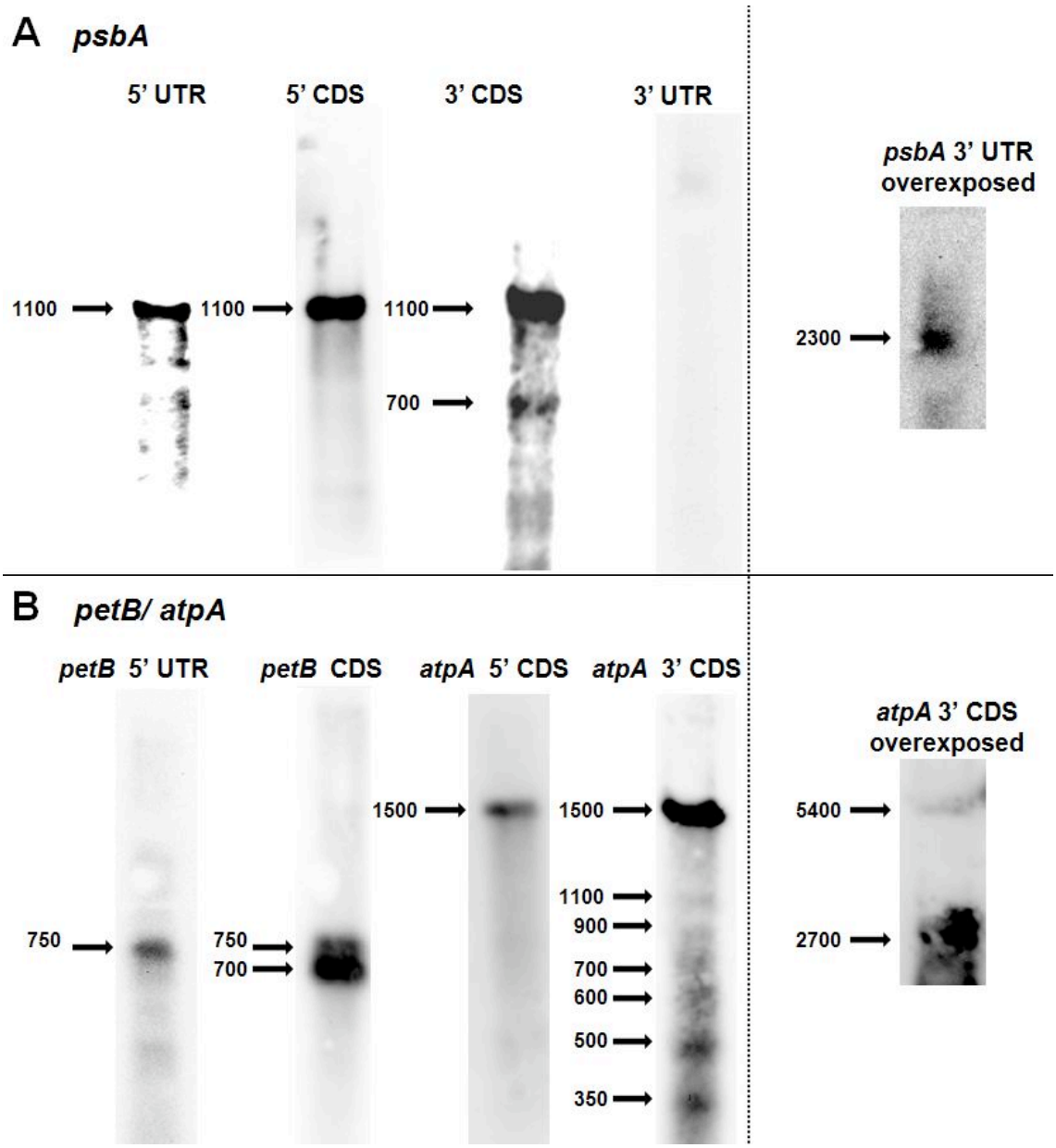


Fig. 3.5: Northern blots of *psbA* and *petB/atpA* minicircle transcripts.

This diagram shows northern blots hybridised with single-stranded RNA probes corresponding to different regions of non-template strand transcript sequence from the *psbA* **(A)** and *petB/atpA* **(B)** minicircles. Key bands are identified with arrows. Sizes of each band were calculated by comparison to a DIG-labelled RNA ladder separated on the same gel, as detailed in Chapter 2. To the right of the dotted line, overexposed regions of the *psbA* 5' UTR and *atpA* CDS 3' end northern blots are shown, demonstrating the presence of transcripts the length of one and two linearised minicircles.

Green, 2010; Nisbet *et al.*, 2008). To determine the full diversity of transcripts produced from the *psbA* and *petB/ atpA* minicircles, RNA probes were synthesised that were specific to short regions of both coding and non-coding sequence from each minicircle (Table 3.5). Probes were designed that were specific to the 5' and 3' UTRs of the *psbA* minicircle, as well as probes that were specific to the 5' and 3' ends of the *psbA* CDS (Table 3.5). For the *petB/atpA* minicircle, probes were designed that were specific to the *petB* CDS*,* and the 5' and 3' ends of the *atpA* CDS (Fig. 3.5, panel B; Table 3.3). The intergenic region between *petB* and *atpA*, and the *atpA* 3' UTR sequence were too short to design appropriate northern probes, but an additional RNA probe was designed specific to the *petB* 5' UTR (Table 3.5). Each blot was performed using 30 µg total cellular RNA, as this has been shown to be adequate to detect very low abundance and multi-copy transcripts in *H. triquetra* (Dang and Green, 2010).

High intensity bands were detected in each of the blots hybridised with CDS probes (*psbA* : 1100 nt, *petB*: 700 nt, *atpA*: 1500 nt). These corresponded in size to the monocistronic mRNAs of each gene identified in previous studies (Fig. 3.5) (Barbrook *et al.*, 2012; Barbrook *et al.*, 2001; Nisbet *et al.*, 2008). Bands consistent with transcripts the length of one minicircle, similar to those identified in previous studies, were identifiable for both *psbA* (2300 nt, corresponding to a 2311 bp minicircle) and *petB/atpA* (2700 nt, corresponding to a 2713 bp minicircle) on overexposure of specific blots (Nisbet *et al.*, 2008). Using a probe complementary to the 3' end of *atpA*, a very low intensity band was detected corresponding to transcripts twice the length (5400 nt) of the *petB/atpA* minicircle sequence (Fig. 3.5, panel B). Overall, the northern blots indicate that multi-copy transcripts are present at much lower abundance than the corresponding monocistronic mRNAs.

Several bands were identified in individual northern blots that correspond to transcripts of less than one minicircle length (Fig. 3.5). These bands were much lower in intensity than the monocistronic mRNAs identified in the CDS blots. A 750 nt band was found in blots probed for the *petB* 5' UTR and CDS (Fig. 3.5). This band was lower in abundance, and corresponded to a higher molecular weight transcript than the 700 nt *petB* mRNA identified in the *petB* CDS blot, and therefore might correspond to a 5' end processing precursor of the mature transcript. Low abundance bands were also identified in northern blots probed for the 3' ends of the *psbA* (700 nt) and *atpA* (350-1100 nt) (Fig. 3.5). In contrast to the 750 nt *petB* band, these bands were shorter than the corresponding mature mRNA of each gene, and are therefore likely to correspond to translationally non-functional transcripts. However, it is unlikely that any of these bands extend into the minicircle core region. The 700 nt additional band identified in the blot probed for the *psbA* CDS 3' end could not be identified in the blot

Table 3.6. Primers for RT-PCR of antisense transcripts

Primer annealing positions are shown relative to the sequence of the corresponding minicircle, where position 1 corresponds to the 5' end of the core region, as per Table 3.1.

		Minicircle	cDNA synthesis primer	Annealing site
		psbA	CAAGCCTTATTCGCTCTAACT	838 F
		petB/ atpA	TCAGTCTGTCTGCGAACCAC	1529 F
Reaction	**PCR forward primer**	**Annealing site**	**PCR reverse primer**	**Annealing site**
psbA	CTTCTAACGCAATCGGTGTCC	1075 F	GCTCGTGCATTACCTCGATAC	1821 R
petB	ATCATCCAAGCGGCAACT	588 F	GACACAATGGACGGTGC	1525 R
atpA	CAGCGTGAACTAATTATTGGTG	1599 F	TCGTTCAACCACACTTTATACAGAAC	2607 R
psbA core	GACTAGGCTTTCTGATTAACGAT	1896 F	AGTTAGAGCGAATAAGGCTTG	858 R
petB/ atpA core	CGAATGAAGGATACGGAAAGG	2679 F	GCAACTCAAGACGCTCTTCAC	498 R
Spliced leader	GTACCCATTTTGGCTCAAG			

probed for the *psbA* 3' UTR probe, indicating that this transcript did not extend downstream of the CDS (Fig. 3.5, panel A). Although the additional bands identified in the blot hybridised with the 3' end of *atpA* may extend into the core sequence, all of these bands were of lower molecular weight (of 350-1100 nt length) than the core-containing *atpA* transcripts identified by circular RT-PCR (of 1200-1800 nt length) (Fig. 3.5, panel B; Table 3.4). It is therefore likely that these bands represent the degradation products of mature mRNAs, rather than core-containing transcripts. Overall, it appears that core-containing transcripts are likely to be present only at extremely low abundance. This may be because the core region is only transcribed infrequently, or because transcripts containing core regions are targeted for immediate processing and degradation.

Presence of antisense transcripts in dinoflagellate plastids

Given the unusual organisation of the dinoflagellate plastid genome, I wished to test whether antisense transcripts of plastid minicircles were present in *Amphidinium carterae*. To do this, cDNA was synthesised using primers with the same sequence as the non-template strands of the *psbA* and *petB/atpA* minicircles (Table 3.6; Fig. 3.6, panel A), which would anneal to antisense transcripts. Each cDNA synthesis primer was confirmed by BLAST not to be similar to the sequence of the template strand of the minicircle in question, and thus should not anneal promiscuously to sense transcripts. PCRs were then performed using combinations of primers internal to the *psbA, petB* and *atpA* genes, and the core regions of each minicircle (Fig. 3.6, panel A).

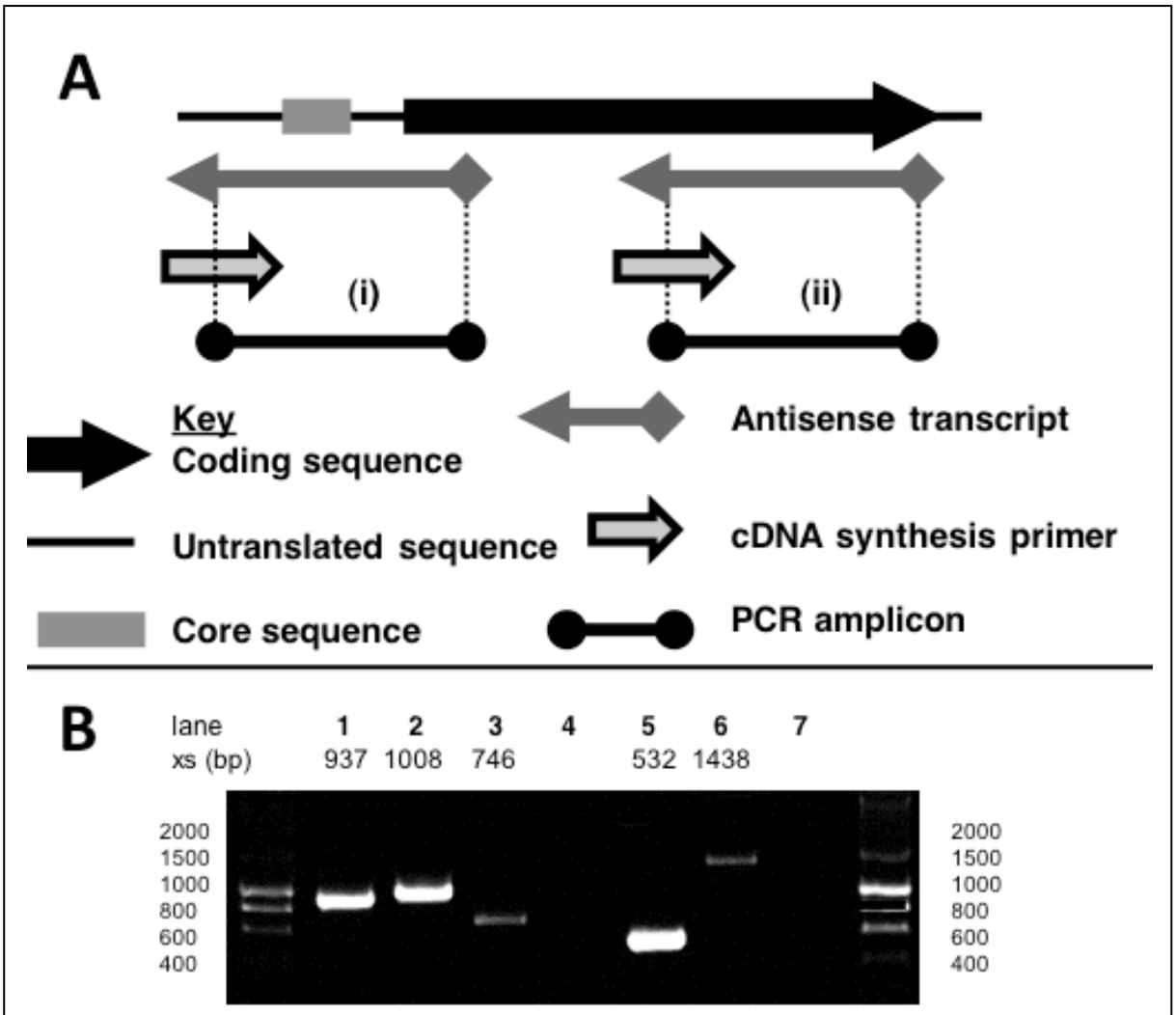


Fig. 3.6: Antisense transcripts in dinoflagellate plastids.

Panel A shows a diagram of the RT-PCRs used to identify antisense transcripts. cDNA was synthesised using primers with the same sequence as the non-template strand of minicircle sequence, which anneal to antisense transcripts. PCRs are then performed as before, using different combinations of primers **(i, ii)** to detect antisense transcripts covering different regions of minicircle sequence. **Panel B** shows a gel photograph confirming the existence of antisense transcripts from both minicircles. Lanes **1-3**: RT-PCRs to detect antisense transcripts covering the *petB, atpA* and *psbA* coding sequences. Lane **4**: template negative control for lane 1. Lanes **5-6**: RT-PCRs to detect antisense transcripts covering the *petB/atpA* and *psbA* minicircle core regions. Lane **7**: template negative control for lane 5.

Products were obtained for PCRs against each gene (Fig. 3.6, panel B; lanes 1-3) and core region tested (Fig. 3.6, panel B; lanes 5-6; Table 3.6). These products were not visible in

Table 3.7. Primers for 5' RACE of antisense transcripts

Primer annealing positions are shown relative to the sequence of the corresponding minicircle, where position 1 corresponds to the 5' end of the core region, as per Table 3.1.

RNA adapter	GCUGAUGGCGAUGAGCACUGGGUUGCAA			
		Transcript	**cDNA primer**	**Annealing site**
Adapter primer 1	GCTGATGGCGATAGC	**psbA antisense**	CAAGCCTTATTCGCTCTAACT	838 F
Adapter primer 2	GATGAGCACTGGGTTGC	**atpA antisense**	TCAGTCTGTCTGCGAACCAC	1529 F
Transcript	**Gene-specific PCR primer 1**	**Annealing site**	**Gene-specific PCR primer 2**	**Annealing site**
psbA antisense	CTGGGGTTCTTTCGTTCAAAC	860 F	CTTCTAACGCAATCGGTGTCC	1075 F
atpA antisense	CAGCGTGAACTAATTATTGGTG	1599 F	ACGAGAAGGTTCTATCCGTCTATG	1675 F

reverse transcriptase negative controls, hence were not the result of gDNA contamination (Fig. 3.6, panel B; lanes 4, 7). Each product was sequenced, and confirmed to be identical to the previously sequenced *A. carterae psbA* and *petB/ atpA* minicircles (Barbrook and Howe, 2000; Barbrook *et al.*, 2001). The core regions sequenced were found to be in the correct orientation relative to the 5' and 3' UTR sequences, indicating that the transcripts were not generated via the transcription of novel minicircles containing reversed fragments of plastid CDS. I additionally performed RT-PCRs using the antisense cDNA preparations, the primer used for cDNA synthesis, and a PCR primer designed to be similar to the spliced-leader (SL) sequence, a short motif associated with the 5' end of most dinoflagellate nuclear transcripts (Table 3.6) (Lin, 2011; Lin *et al.*, 2010). Products could not be detected suggesting that it is unlikely that these transcripts were generated within the dinoflagellate nucleus (data not shown).

Theoretically, the RT-PCR products may have been generated through the promiscuous annealing of the cDNA synthesis primers to sense transcripts from the same minicircle. To confirm that antisense transcripts were present, RNA-ligase mediated 5' RACE (5' RLM-RACE) was performed for antisense *psbA* and *atpA* transcripts. This technique uses the direct ligation of an RNA adapter to clone the 5' end of transcript sequence (Dang and Green, 2010; Scotto-Lavino *et al.*, 2006). cDNA was synthesised from adaptor ligated *A. carterae* RNA, using the same primer previously used to identify antisense *psbA* and *atpA* transcripts (Table 3.7). These cDNA products were used as template for PCRs, using primers with the same sequence as the non-template strand of the *psbA* and *atpA* genes, and a PCR primer with the same sequence as the RNA adapter used (Table 3.7). This primer combination would solely amplify the 5' ends of antisense minicircle transcripts, and would not be able to amplify sequences generated through promiscuous annealing of the cDNA synthesis primer to sense transcripts (Fig. 3.7, panel A). In each case, the sequences

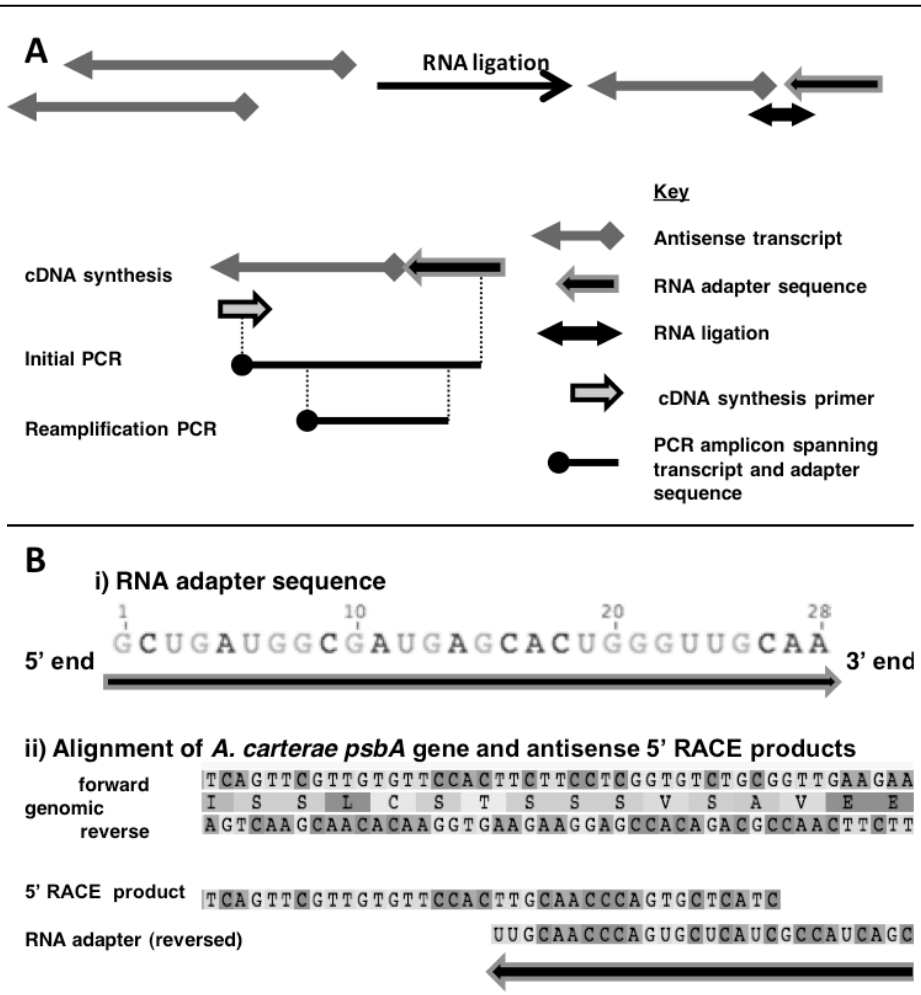


Fig. 3.7: 5' RLM-RACE of antisense plastid transcripts.

Panel A shows a diagram of RNA ligase-mediated 5' RACE. An RNA adapter is ligated to a sample of total cellular RNA. The 3' end of the RNA adapter ligates to the 5' ends of processed transcript sequences. RT-PCR is performed using a cDNA synthesis primer with the same sequence as the non-template strand of minicircle sequence, which will anneal to antisense transcripts. PCR is performed with the cDNA template, a nested PCR primer specific to the antisense transcript sequence, and a forward directed primer specific to RNA adapter sequence. This will amplify antisense transcripts onto which an RNA adapter has been ligated at the 5' end. A second amplification is performed using the initial PCR product, and nested transcript-specific and adapter-specific primers. Products of the second PCR are cloned and sequenced. **Panel B** shows **(i)** the complete RNA adapter sequence employed, and **(ii)** an alignment of the genomic *psbA* sequence with the terminal region of a 5' RACE product. The 5' RACE product contains sequence that aligns with the *psbA* template strand, followed by a sequence that aligns with the 3' end of the RNA adapter. This product is the 5' end of a minicircle antisense transcript.

of ligation products between the RNA adapter and the 5' ends of antisense *psbA* and *atpA* transcripts were identified (Fig. 3.7, panel B; Table 3.6). This confirms that antisense transcripts are present in dinoflagellate plastids.

Antisense transcripts undergo different end cleavage events from sense transcripts

In plants, complementary sense and antisense transcripts have been documented to possess different consensus terminus positions (Georg *et al.*, 2010; Zghidi-Abouzid *et al.*, 2011). The end cleavage events specifically associated with antisense transcripts may be associated with transcript degradation pathways (Sharwood *et al.*, 2011). I wished to determine whether sense and antisense transcripts in dinoflagellate plastids likewise possessed different associated terminus positions.

RT-PCRs were performed using circularised RNA to identify the termini of antisense transcripts from the *A. carterae psbA* and *petB/atpA* minicircles (Table 3.8). cDNA was synthesised using a range of different primers to different regions of the non-template strands of each minicircle sequence, to map the diversity of antisense transcripts present (Table 3.8). To identify antisense transcripts from the *psbA* minicircle, cDNA synthesis primers were designed with the same sequences as the non-template strands of the *psbA* 5' UTR, CDS 5' end, CDS 3' end, and 3' UTR (Table 3.8). In the case of the *petB/ atpA* minicircle, cDNA synthesis primers were designed with the same sequences as the non-template strands of the 5' end of the *petB* CDS, the 5' and 3' ends of the *atpA* CDS, and the *atpA* 3' UTR. For each cDNA template, PCRs were performed using a reverse primer positioned upstream of the cDNA primer annealing site, and a PCR forward primer positioned downstream of the cDNA primer annealing site (Table 3.8).

To determine which of the antisense transcripts identified by circular RT-PCR were predominant, northern blots of *A. carterae* RNA were hybridised to single-stranded RNA probes with the same sequence as the non-template strands of the *psbA* and *petB/atpA* minicircles, which would specifically anneal to antisense transcripts (Table 3.8). To facilitate the direct comparison of sense and antisense transcripts, the RNA probes were complementary in sequence to the probes previously designed for sense transcripts from each minicircle, and identical RNA electrophoresis and detection conditions were used for sense and antisense transcript blots (Table 3.8). Each blot was performed twice using independently isolated RNA samples, and consistent banding patterns were identified in each case.

Table 3.8. Circular RT-PCR primers and northern blot probes for antisense transcripts

Primer annealing positions are shown relative to the sequence of the corresponding minicircle, where position 1 corresponds to the 5' end of the core region, as per Table 3.1. Northern probes are shown as per Table 3.5.

Primers	psbA	Annealing site	petB/ atpA	Annealing site
cDNA synthesis primers				
1	CGAGTCAGAGGCATCAAAC	228 F (5' UTR)	GCAGACGATATCCTCTCTAAG	507 F (petB 5' end)
2	CTGGGGTTCTTTCGTTCAAAC	860 F (CDS 5' end)	ACGAGAAGGTTCTATCCGTCTATG	1675 F (atpA 5' end)
3	CCTCTCTTGGTGTTGCTACTATG	1678 F (CDS 3' end)	GTAGGTATCTCGGTTACACG	2190 F (atpA 3' end)
4	GACTAGGCTTTCTGATTAACGAT	1896 F (3' UTR)	CGAATGAAGGATACGGAAAGG	2659 F (3' UTR)
PCR forward primers				
1	AGTCTCCCGATTGTCTATTCTC	41 R (core)	GCAACTCAAGACGCTCTTCAC	498 R (petB 5' end)
2	AGTTAGAGCGAATAAGGCTTG	858 R (CDS 5' end)	CACCAATAATTAGTTCACGCTG	1620 (atpA 5' end)
3	GATACCAATTACAGGCCAAGC	1670 R (CDS 3' end)	GCATTGCTGTGGAATAGAC	2417 (atpA 3' end)
4	ATCGTTAATCAGAAAGCCTAGTC	1918 R (3' UTR)	TCGTTCAACCACACTTTATACAGAAC	2607 (3' UTR)
PCR reverse primers				
1	TACATTGAGTAGGCATCTTTAATAGC	512 (5' UTR)	ATCATCCAAGCGGCAACT	588 F (petB 5' end)
2	CTTCTAACGCAATCGGTGTCC	1075 (CDS 5' end)	TCCCTGTGGGTCAGAAG	1699 F (atpA 5' end)
3	ACGCTCATAACTTCCCTCTTG	1825 F (CDS 3' end)	GGTCTTCTTGGGTTATTTCC	2480 F (atpA 3' end)
4	CCTCCTACCGAAAGTCAATTC	2238 (3' UTR)	GAAAATCCAGGTCATATCATAGGAG	133 F (core)

Northern blot probes			
T7 arm	TAATACGACTCACTATAGGGCGAATTGGGCCCGACGTCGCATGCTCCCGGCCGCCATGGCCGCGGGATT		
psbA	**Start**	**End**	
5' UTR	510	852	TTACATTGAGTAGGCATCTTTAATAGCTTGTAGGGTTCACACCGCAGTCT…
5' CDS	838	1077	CAAGCCTTATTCGCTCTAACTCCTGGGGTTCTTTCGTTCAAACAATCACT…
3' CDS	1616	1845	CAACAACTCCCGTTCTCTTCACTTCTTCCTTGCAGCTTGGCCTGTAATTG…
3' UTR	1896	2256	GACTAGGCTTTCTGATTAACGATGATGTAATATTATTAAGAACGATGGCT…
petB/ atpA			
5' UTR	206	498	TCCCGATCTCACAAGTCTCCATTAAGGACTAGAGCTTGAAAGACAAATGA…
petB	504	961	ATTGCAGACGATATCCTCTCTAAGTTCGTTCCTTCTCATGTTAACATCTT…
atpA 5'	1531	1715	AGTCTGTCTGCGAACCACTCGCTACTGGTATCGTATCAATCGATGCAATG…

A diverse population of antisense transcripts was detected for each minicircle through circular RT-PCR (Table. 3.9). None of these transcripts contained a region of sequence complementary to the corresponding cDNA primer used (as determined using BLAST), and thus are unlikely to constitute sense transcripts amplified via promiscuous annealing of the cDNA synthesis primer. In addition, bands were identified in several of the antisense transcript northern blots (Fig. 3.8). These were not the same size as the bands identified in northern blots of sense transcripts (Fig. 3.5), indicating that they were not the result of in vitro reverse transcription of the antisense probe sequences by the T7 RNA polymerase to generate probes complementary to sense transcripts (Cazenave and Uhlenbeck, 1994). Thus, these bands are likely to correspond to minicircle antisense transcripts.

Table 3.9. Antisense transcripts identified by circular RT-PCR.

This table lists the antisense transcripts mapped for each minicircle. The terminus positions and primer combinations used for each circular RT-PCR product are given as per in Table 3.4. Antisense transcripts that may correspond to hybridisation observed in antisense transcript northern blots are shown in bold text.

1. psbA **Minicircle length** 2311 bp **Core** 1-281

	CDS		sense transcript	
	5' end	3' end	5' end	poly(U) site
psbA	834	1856	600-829	1870-1872

Transcript	5' end	3' end	cDNA primer	R primer	F primer	Poly(U)	Size (bp)	Northern bands
antisense transcript 1	146 (2)	1347	1896 F	2238 F	1670 R	0	964	
antisense transcript 2	72 (2)	1542	1896 F	2238 F	1670 R	0	769	
antisense transcript 3	1868	1071	1678 F	1825 F	1670 R	0	797	
antisense transcript 4	1867	1520	1678 F	1825 F	1670 R	0	347	
antisense transcript 5	1678	544	860 F	1075 F	858 R	0	1134	
antisense transcript 6	1196	715	860 F	1075 F	858 R	0	481	
antisense transcript 7	1188	765	860 F	1075 F	858 R	0	423	
antisense transcript 8	**1092**	**168**	**860 F**	**860 F**	**858 R**	**0**	**924**	**May correspond to band (i)**
antisense transcript 9	**916**	**168**	**860 F**	**860 F**	**858 R**	**0**	**748**	**May correspond to band (ii)**

2. petB/ atpA **Minicircle length** 2713 bp **Core** 1-281

	CDS		sense transcript	
	5' end	3' end	5' end	poly(U) site
petB CDS	456	1115	310-424	1122-1126
atpA CDS	1206	2582	1081-1088	2591

Transcript	5' end	3' end	cDNA primer	R primer	F primer	Poly(U)	Size (bp)	Northern bands
antisense transcript 1	301 (2)	2086	2659 F	133 F	2417 R	0	928	
antisense transcript 2	294 (2)	2084	2659 F	133 F	2417 R	0	923	
antisense transcript 3	204 (2)	1218	2659 F	133 F	1620 R	0	1699	
antisense transcript 4	**192 (2)**	**1677**	**2659 F**	**133 F**	**2417 R**	**0**	**1228**	**May correspond to band (v)**
antisense transcript 5	**2040**	**1532**	**1675 F**	**1699 F**	**1620 R**	**0**	**508**	**May correspond to band (iii)**
antisense transcript 6	**1770**	**1509**	**1675 F**	**1699 F**	**1620 R**	**0**	**261**	**May correspond to band (iv)**
antisense transcript 7	916	168	860 F	860 F	858 R	0	748	

For the *psbA* minicircle, two antisense transcripts, one 748 nt length, and one 942 nt length, were identified through circular RT-PCR that extended from the 5' end of the *psbA* CDS to the interior of the core region at the 3' end (Table 3.9; transcripts labelled i, ii). Bands corresponding in size to these transcripts were visible in blots probed with the *psbA* 5' UTR and CDS 5' end (Fig. 3.8; transcripts labelled ia, ib, iia, iib). A diverse range of further antisense transcripts covering the 3' end of the minicircle, from 347 to 1110 nt length, were identified through circular RT-PCR (Table 3.9). However, no hybridisation was identified in northern blots of either the *psbA* CDS 3' end or 3' UTR (data not shown), suggesting that they are present at low abundance.

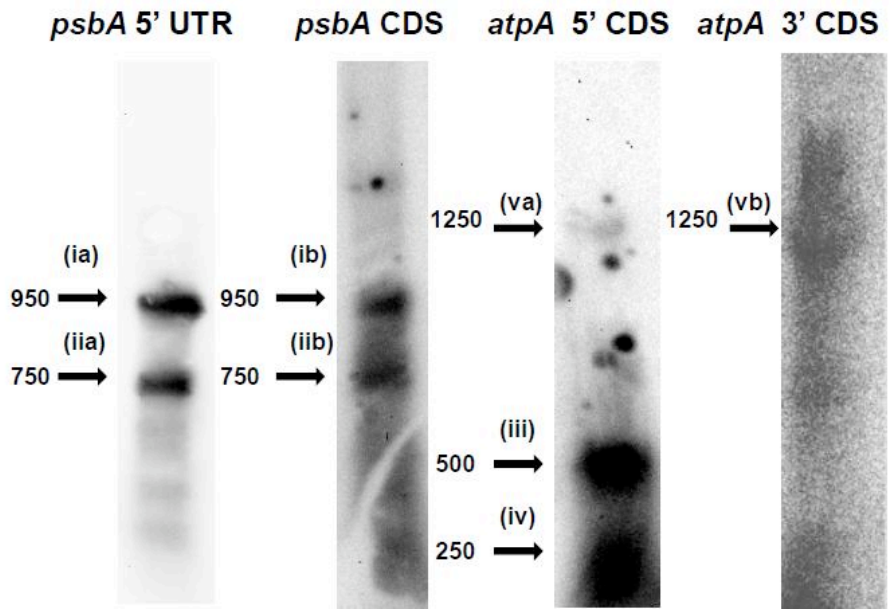


Fig. 3.8: Northern blots of antisense transcripts.

The results of northern blots probed for antisense *psbA* and *atpA* transcripts are shown as per fig. 3.5. Sizes of each band were calculated by comparison to a DIG-labelled RNA ladder separated on the same gel, as detailed in Chapter 2. Bands that may correspond to transcripts identified through circular RT-PCR are labelled with numbered arrows, as per in Table 3.9. As the probes complementary to antisense transcripts covering the *psbA* CDS 3' end and 3' UTR, and the *petB* CDS and 5' UTR, failed to yield any distinct bands, the corresponding blots are not shown. As the *atpA* 3' CDS blot only produced very weak fluorescence, an overexposed blot image is shown.

For the *petB/atpA* minicircle, antisense transcripts were principally identified extending over the *atpA* CDS, and core region. Two high intensity bands, one 500 nt length and the other 250 nt length, were detected in the *atpA* 5' CDS blot (Fig. 3.8; transcripts labelled iii, iv). These transcripts are likely to correspond to 508 nt and 261 nt length antisense transcripts identified by circular RT-PCR (Table 3.9; transcripts labelled iii, iv). A low intensity 1250 nt band was additionally detected in the blot probed for the *atpA* 5' end (Fig. 3.8; transcript labelled va). A 1250 nt band was additionally detected in the blot probed for the *atpA* 3' end, although this band was only visible on overexposure of the blot (Fig. 3.8; transcript labelled vb). These bands may correspond to a 1228 nt transcript identified to extend from the core into the region complementary to the 5' end of *atpA* (Table 3.9). Transcripts were also identified through circular RT-PCR of 923-1699 nt length that terminate at the 5' end within the core or *petB* 5' UTR, and at the 3' end within the *atpA* CDS (Table 3.9). A further 1108 nt transcript was identified through circular RT-PCR that extended from the 5' end of *petB* to

Table 3.10. Primers for semi-quantitative RT-PCR of sense and antisense transcripts

Primer annealing positions are shown relative to the sequence of the corresponding minicircle, where position 1 corresponds to the 5' end of the core region, as per Table 3.1.

Amplicon	Antisense cDNA primer	Annealing site	Sense cDNA primer	Annealing site
psbA 5' end	CGAGTCAGAGGCATCAAAC	228 F	AGTTAGAGCGAATAAGGCTTG	858 R
psbA 3' end	TCAACAACTCCCGTTCTC	1615 F	AAGAGGGAAGTTATGAGCGTTAC	1844 R
atpA	TCAGTCTGTCTGCGAACCAC	1529 F	GCATTGCTGTGGAATAGAC	2417 R
	PCR forward primer	**Annealing site**	**PCR reverse primer**	**Annealing site**
psbA 5' end	TACATTGAGTAGGCATCTTTAATAGC	512 F	TGCAGGAGCAAGGAAGAAAG	992 R
psbA 3' end	TCTCTTCACTTCTTCCTTG	1629 F	GCTCGTGCATTACCTCGATAC	1821 R

the 3' end of *atpA* (Table 3.9). However, hybridisation corresponding to these transcripts could not be detected in any northern blot, suggesting that they are very low in abundance.

The vast majority of the antisense transcripts identified through circular RT-PCR terminated at positions internal to the corresponding CDS. Only one antisense transcript, complementary to *atpA*, extended over an entire CDS (Table 3.2). However, this transcript was not detectable in northern blots, suggesting that it is low in abundance (Fig. 3.8, panel B). None of the antisense transcripts identified through circular RT-PCR terminated either at a position complementary to the consensus 5' end positions or poly(U) sites of mature sense previously identified for monocistronic sense *psbA, petB* or *atpA* transcripts (A.C. Barbrook, pers. comm.) (Barbrook *et al.*, 2012). This indicates that sense and antisense transcripts undergo different end cleavage events in dinoflagellate plastids.

Antisense transcripts are lower in abundance than sense transcripts

In plant plastids, antisense transcripts are less abundant than sense transcripts (Sharwood *et al.*, 2011; Zghidi-Abouzid *et al.*, 2011). The antisense transcripts observed through northern blotting were less abundant than the corresponding sense transcripts, requiring longer exposure times, or even overexposure of the blots, to be visible (Fig. 3.8). I wished to estimate the ratio of abundance of sense and antisense transcripts over the *psbA* and *petB/atpA* minicircles.

Transcript abundance was investigated through semi-quantitative RT-PCR. cDNA was synthesised using 100 ng isolated RNA, and primers were designed for sense or antisense transcripts from the *psbA* and *petB/atpA* minicircles (Table 3.10). PCRs were performed using serial dilutions of each cDNA template generated from the RT-PCR, with primers

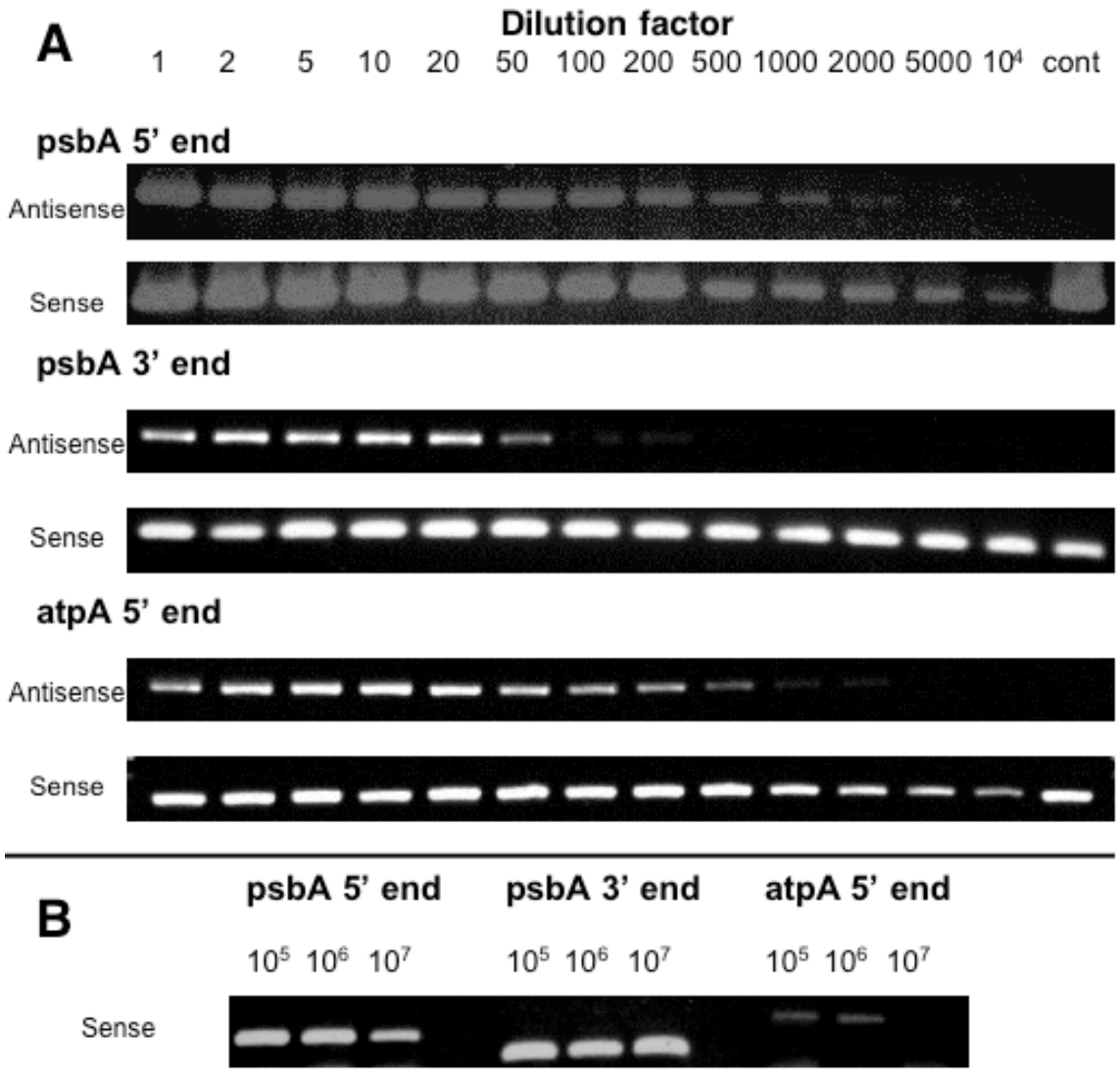


Fig. 3.9: Semi-quantitative RT-PCR of sense and antisense transcripts.

Panel A shows the result of RT-PCRs using up to 10^5-fold dilutions of cDNA template generated from sense and antisense transcripts covering the 5' and 3' ends of the *psbA* CDS, and the 5' end of *atpA* . The fold dilution of each lane is given at the top of the figure. The final lane corresponds to control reactions performed for each reaction using template negative conditions (shown on antisense transcript gel photo) and with gDNA template (shown on sense transcript gel photo). Products were only obtained with antisense transcript cDNA templates for up to 2000 fold dilution for the *psbA* 5' end, 1000 fold dilution for the *atpA* 5' end, and 200 fold dilution for the *psbA* 3' end, while products were identified for sense transcripts at every dilution tested.

Panel B shows the result of RT-PCRs performed using sense transcript cDNA under even greater degrees of dilution. Sense transcript *atpA* transcripts were detected following 10^6-fold dilution, and for *psbA* following 10^7-fold dilution of the cDNA template.

positioned between the sense and antisense transcript cDNA synthesis primers, to determine the relative abundance of each transcript (Table 3.10). PCRs were performed for a region that overlapped the 5' CDS end and 5' UTR of *psbA*, and a region covering the 5' end of *atpA.* These regions gave rise to the most intense hybridisation in northern blots probed for antisense transcripts, and are therefore likely to correspond to the most abundant antisense transcripts from each minicircle (Fig. 3.8). Additional PCRs were performed for a region of the 3' end of the *psbA* CDS (Table 3.10). Antisense transcripts covering this region were detectable through circular RT- PCR, but not through northern blotting, indicating that the transcripts might be lower in abundance than those for the 5' end of *psbA* and *atpA* (Table 3.9; Fig. 3.8). To enable direct comparison of the results, the same PCR primer combination was used to amplify sense and antisense cDNA transcripts for each region, and a minimum of three independent replicates were performed, using RNA samples from different *A. carterae* cultures.

For each region tested, antisense transcripts were lower in abundance than the corresponding sense transcripts (Fig. 3.9). Antisense products were not detected for the 5' end of *psbA* and *atpA* using greater than 2000-fold dilutions, or for the 3' end of *psbA* using greater than 200-fold cDNA dilutions (Fig. 3.9, panel A). In contrast, sense transcripts were detected for the *atpA* region with 10^6-fold dilution of the cDNA template, and for *psbA* following 10^7-fold dilution of the sense transcript cDNA templates (Fig. 3.9, panel B). This indicates that the sense transcripts of the *petB/atpA* and *psbA* minicircles may be at least 500-fold more abundant, and at some loci up to 50000 times more abundant, than the corresponding antisense transcripts.

Antisense transcripts lack poly(U) tails

As antisense transcripts are much less abundant than sense transcripts in dinoflagellate plastids, specific pathways may limit their accumulation. None of the antisense transcripts that identified by circular RT-PCR possessed poly(U) tails (Fig. 3.8, panel A; Table 3.2). I wished to determine whether polyuridylylated antisense transcripts were generated from either minicircle.

To test whether antisense transcripts possess poly(U) tails, cDNA was synthesised from *A. carterae* total cellular RNA using an oligo-d(A) primer. This cDNA synthesis primer consisted of a 3' oligo-d(A) region which would anneal to transcript poly(U) tails, and an additional region at the 5' end that did not correspond to any *A. carterae* minicircle sequence, which would act as a sequence anchor, to enable subsequent amplification of the cDNA template at the annealing temperature (55 °C) used. This primer should accordingly anneal to every

Table 3.11. Primers for RT-PCR to detect polyuridylylated antisense transcripts

oligo-d(A) primer	GGGACTAGTCTCGAGAAAAAAAAAAAAAAAAA

Amplicon	PCR gene-specific primer	Annealing site	Gene-specific cDNA primer	Annealing site
Antisense psbA-1	GCTCGTGCATTACCTCGATAC	1821 R	CAAGCCTTATTCGCTCTAACT	838 F
Antisense psbA-2	CTTTAGACTGCGGTGTGAAC	563 R	GACTAGGCTTTCTGATTAACGAT	1896 F
Antisense petB	AAGGTGTGAGCCTGATAGAAC	1033 R	GCAGACGATATCCTCTCTAAG	507 F
Antisense atpA	CTTCTGACCCACAGGGACAT	1715 R	ACGAGAAGGTTCTATCCGTCTATG	1675 F
Sense psbA	CAAGCCTTATTCGCTCTAACT	838 F	n/a	
Sense petB	GCAGACGATATCCTCTCTAAG	507 F	n/a	
Sense atpA	ACGAGAAGGTTCTATCCGTCTATG	1675 F	n/a	

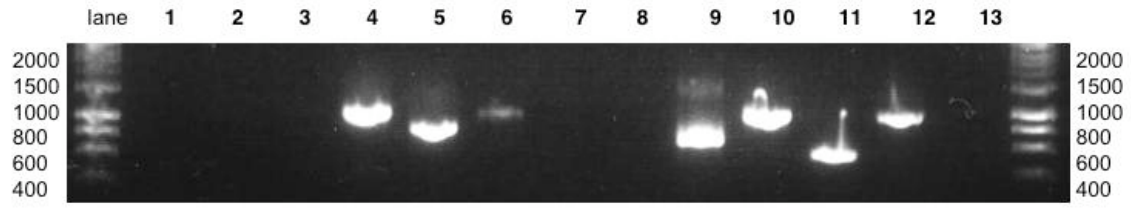


Fig. 3.10: Absence of poly(U) tails from antisense transcripts.

This gel photograph shows the result of a series of RT-PCRs to test for poly(U) tails on antisense transcripts of the *psbA* and *petB/atpA* minicircles. Lanes **1, 13**: blank lane. Lanes **2-3, 7-8**: RT-PCRs performed with an oligo-d(A) primer for cDNA synthesis, and PCR with the same oligo-d(A) primer and a primer with the same sequence as the template strand of the *psbA* CDS **(2)** and UTR **(3)**, and the *petB* **(7)** and *atpA* CDS **(8)**, demonstrating the absence of polyuridylylated antisense transcripts extending over these regions. Lanes **4, 9-10**: RT-PCR performed with oligo-d(A) primed cDNA as before, and PCR with oligo-d(A) and primers with the same sequence as the non-template strands of the *psbA* **(4)**, *petB* **(9)** and *atpA* CDS **(10)**, confirming the presence of polyuridylylated sense transcripts in the RNA sample. Lanes **5, 6, 11, 12**: positive controls for the presence of antisense transcripts over the *psbA* CDS **(5)** and UTR **(6)**, and the *petB* **(11)** and *atpA* CDS **(12)**, using a gene-specific cDNA synthesis and PCR primer with the same sequence as the non-template strands of minicircle sequence, and the same template strand PCR primer as used in the corresponding oligo-d(A) primed PCR for each reaction.

polyuridylylated transcript present, regardless of the minicircle from which it was generated (Table 3.11). PCRs were then performed using the same oligo-d(A) primer, and PCR primers

with the same sequence as the template strands of the *psbA* and *petB/atpA* minicircles, to identify polyuridylylated antisense transcripts (Table 3.11).

Products could not be identified using any of the template strand primers, indicating that polyuridylylated antisense transcripts were not present (Fig. 3.10; lanes 2-3, 7-8). Products could not be detected even following a second round of PCR amplification, using the primary PCR product as a PCR template. This result was confirmed independently through three repeats of the RT-PCR, using different RNA samples for each cDNA synthesis reaction. Polyuridylylated sense transcripts were amplified from each cDNA preparation, by PCR with the oligo-d(A) primer, and PCR primers with the same sequence as the non-template strand of the *psbA, petB* and *atpA* genes, confirming that the oligo-d(A) primed cDNA synthesis reactions were successful (Fig. 3.10; lanes 4, 9, 10). In addition antisense transcripts covering each of the regions of sequence tested were amplified from each RNA samples, using gene-specific cDNA synthesis and PCR primers, as previously described (Fig. 3.6, panel A; Fig. 3.10, lanes 5-6; 11-12). Thus, while a diverse range of antisense *psbA* and *petB/ atpA* transcripts are generated in *A. carterae* plastids, the antisense transcripts present do not possess a 3' poly(U) tail. The different polyuridylylation of sense and antisense transcripts may limit the accumulation of antisense transcripts in dinoflagellate plastids.

Discussion

I have characterised non-coding transcripts from plastid minicircles in the dinoflagellate *Amphidinium carterae*. I have identified core-containing transcripts, and transcripts of greater than one minicircle length, as have previously been found in *Heterocapsa triquetra* (Figs. 3.2-3) (Dang and Green, 2010). These multi-copy transcripts might have been generated from concatemers generated by minicircle fusion. Large minicircles containing multiple copies of core sequence have previously been found in the dinoflagellate *Adenoides eludens*, suggesting that individual minicircles may fuse to form larger polymers (Nelson and Green, 2005). However, Southern blots of *A. carterae* plastid DNA have not identified minicircles of an equivalent size to the multi-copy transcripts detected by RT-PCR and by northern blotting (Figs. 3.2, 3.4) (Barbrook and Howe, 2000; Barbrook *et al.*, 2001). It is therefore likely that these transcripts are generated by rolling circle transcription. This may occur as a result of inefficient termination of plastid transcription, similarly to what occurs in other plastid lineages (Barkan, 2011; Rott *et al.*, 1996).

The multi-copy transcripts identified might be non-functional, generated at low levels through inefficient transcription termination and 3' processing. Alternatively, these multi-copy transcripts may represent processing precursors of mature mRNAs (Barbrook *et al.*, 2012;

Dang and Green, 2010). I identified the complete sequences of multi-copy transcripts that have mature 5' ends, and other multi-copy transcripts that possess 3' poly(U) tails, which suggests that multi-copy transcripts undergo similar processing events to mature mRNAs (Fig. 3.2; Table 3.4). I additionally identified short transcripts that terminate at the 3' end within minicircle core regions (Table 3.4). None of these transcripts appeared to contain a complete CDS, or possess a 5' or 3' end associated with mature mRNAs, indicating that they are unlikely to possess a coding function.

I additionally report the presence of antisense transcripts generated from peridinin dinoflagellate minicircle sequences (Figs. 3.5-10). It is possible that these antisense sequences are not generated from plastid gene sequences, but from copies of plastid sequences located in the dinoflagellate nucleus (NUPTs). It is well understood that fragments of plastid sequence are frequently transferred to the nuclei of plants, and some of these fragments may be transcribed at low levels (Huang *et al.* 2004; Kleine *et al.*, 2009; Wang *et al.,* 2014). It has been suggested similarly that minicircles, or fragments of minicircle-derived sequence, reside in the nuclei of peridinin dinoflagellates, although recent studies have indicated that the overwhelming majority of minicircle sequences are located within the dinoflagellate plastid (Laatsch *et al.*, 2004; Owari *et al.*, 2014). In theory, a fragment of minicircle sequence might insert in antisense orientation within a transcriptionally active region of a dinoflagellate nucleus, and give rise to antisense transcripts, although if this does occur these transcripts seem not to receive 5' spliced leaders (Lin, 2011). Whether the antisense transcripts are of nuclear origin will only be conclusively answered by sequencing and quantification of NUPTs in the *A. carterae* nuclear genome.

The antisense transcripts in *A. carterae* might equally be derived from copies of plastid genes. If so, this would represent the first evidence for antisense transcripts in an algal plastid. The presence of antisense transcripts in the dinoflagellate plastid is surprising, as even in species such as *A. carterae* that possess minicircles containing more than one CDS, there are no minicircles in which more than one gene is present in an opposing transcriptional orientation, in which antisense transcripts might be generated through transcriptional run-through (Howe *et al.*, 2008b; Sharwood *et al.*, 2011). The antisense transcripts might be generated via an RNA-dependent RNA polymerase activity located within the plastid, such as has previously been indicated to be present in plant plastids (Zandueta-Criado and Bock, 2004). Alternatively, the antisense transcripts might be generated via the bidirectional transcription of minicircle sequence. There may be specific promoters located on minicircle template strands that allow the generation of antisense transcripts. Alternatively, antisense transcripts might be generated as a result of transcription

initiation events that are not dependent on specific primary sequence motifs, with the plastid RNA polymerase recruited to random sites, or to features such as stem loops or single-stranded nicks in minicircle sequence (Dang and Green, 2009; Leung and Wong, 2009; Moore *et al.*, 2003; Zhang *et al.*, 2002).

The antisense transcripts observed in the *A. carterae* plastid are substantially less abundant than the corresponding sense transcripts. This may be due to a difference in the associated transcriptional activity of promoters located on the forward and template strands of minicircle sequence, as has been shown at some loci in plant plastids (Zhelyazkova *et al.*, 2012). Alternatively, if antisense transcripts have deleterious effects on plastid gene expression, as occurs in plants (Hotto *et al.*, 2012; Sharwood *et al.*, 2011), dinoflagellates may possess pathways to eliminate them from plastid RNA pools. The addition of a poly(U) tail to sense transcripts might enable the plastid to identify and degrade the non-polyuridylylated antisense transcripts (Fig. 3.10). This would support previous hypotheses that the dinoflagellate plastid poly(U) tail has a role in transcript 3' end stabilisation, similar to the nuclear poly(A) tail, or the poly(U) tail in kinetoplastid mitochondria (Barbrook *et al.*, 2012; Fisk *et al.*, 2008; Norbury, 2010). It remains to be determined whether the accumulation of antisense transcripts has deleterious effects on the expression of sense transcripts in dinoflagellate plastids, similarly to as in plants (Hotto *et al.*, 2010; Sharwood *et al.*, 2011). It likewise remains to be determined whether the poly(U) tail is directly involved in transcript stabilisation. Further experimentation may provide valuable insights into the function of poly(U) tail addition and the evolution of transcript processing in this remarkable plastid genome.

Chapter Four- Transcript processing pathways retained from an ancestral plastid symbiosis function in serially acquired dinoflagellate plastids

Introduction

The endosymbiont hypothesis for the origin of plastids is one of the most well-established tenets of eukaryotic cell biology (Howe *et al.*, 2008a; Sagan, 1967). Each plastid lineage found within the eukaryotes arose through the endosymbiotic integration of two organisms: a free-living photosynthetic prokaryote or eukaryote, of varying phylogenetic origin, which was taken up by a eukaryotic host and converted into a permanent organelle (Dorrell and Howe, 2012b). This process involved the establishment of pathways within the host, which evolved as a consequence of endosymbiosis, to support the plastid (Dorrell and Howe, 2012b).

Genomic and phylogenetic evidence has suggested that several major photosynthetic eukaryote lineages have replaced their original plastids with others of different phylogenetic origin, in a process termed serial endosymbiosis (Burki *et al.*, 2014; Dorrell and Howe, 2012b). The best supported examples of this are within the dinoflagellate algae, in which the ancestral plastid, containing the pigment peridinin and derived from a red alga, has been replaced in at least three lineages. For example, dinoflagellates that contain the pigment fucoxanthin have replaced their ancestral peridinin-containing plastids, with ones derived from haptophytes (Ishida and Green, 2002; Takishita *et al.*, 1999). Similarly, the “dinotom” algae are believed to have acquired replacement plastids derived from diatoms, and dinoflagellates of the genus *Lepidodinium* possess replacement plastids derived from green algae (Burki *et al.*, 2014; Matsumoto *et al.*, 2011b). Two even more dramatic examples of serial endosymbiosis have recently been put forward, based on “footprints” of genes that may have been acquired from historical plastids in the nuclear genomes of major photosynthetic eukaryotic lineages. The first proposes that the ancestors of taxa currently harbouring red algal-derived plastids, such as diatoms and apicomplexan parasites, contained green algal symbionts (Frommolt *et al.*, 2008; Moustafa *et al.*, 2009). In the second example, the cyanobacterial-derived plastids of plants and their closest relatives were acquired following the loss of a previous endosymbiont derived from chlamydiobacteria (Becker *et al.*, 2008; Huang and Gogarten, 2007). Although both these proposed replacement events remain controversial (Burki *et al.*, 2012; Deschamps and Moreira, 2012; Woehle *et al.*, 2011), serial endosymbioses may constitute a widespread feature of plastid evolution.

Regardless of the number of serial endosymbiosis events that have occurred, one outstanding question is whether the ancestral plastid symbiosis might affect the biology of its replacement (Dorrell and Howe, 2012b; Larkum *et al.*, 2007). In theory, pathways established to support the ancestral plastid could be retained, following serial endosymbiosis, and applied to the incoming replacement plastid. If these pathways had not previously existed in the replacement plastid, its biology might be dramatically changed as a result. It has been demonstrated that genes encoding plastid proteins, which were derived from the ancestral peridinin plastid symbiosis, may be retained in dinoflagellates that have undergone serial endosymbiosis (Minge *et al.*, 2010; Nosenko *et al.*, 2006; Patron *et al.*, 2006; Takishita *et al.*, 2008). However, none of these genes has been confirmed to encode a product that functions in the associated serially acquired plastid. In addition, all of the genes of peridinin origin that have been identified in these species encode proteins that are associated with a wide phylogenetic distribution of plastid lineages. It is therefore likely that the free-living ancestors of the replacement plastid possessed homologues of these genes (Dorrell and Howe, 2012b). Consequently, the retention of these genes from the ancestral peridinin symbiosis might not confer a biochemical activity to the replacement plastid that it previously lacked.

This project was conceived to determine whether transcript processing pathways from the ancestral peridinin plastid have been retained, and applied to serially acquired dinoflagellate plastids. Transcripts in the ancestral peridinin dinoflagellate plastid undergo unusual processing events. Plastid transcripts in all studied peridinin dinoflagellates receive a 3' poly(U) tail (Barbrook *et al.*, 2012; Wang and Morse, 2006). This pathway is additionally found in the plastids of the "chromerid" alga *Chromera velia*, a photosynthetic alveolate that is closely related to dinoflagellates, and is thus likely to be an ancestral feature of the peridinin plastid, but has not been documented in any other plastid lineage (Green, 2011; Janouškovec *et al.*, 2010). Plastid transcripts in some peridinin dinoflagellate species also undergo extensive substitutional sequence editing, including transition substitutions between both purines and pyrimidines, and transversion substitutions (Iida *et al.*, 2009; Zauner *et al.*, 2004). Editing has been identified in plant plastids, but is restricted to pyrimidine transition substitutions, and appears to have evolved independently to the editing in dinoflagellates (Fujii and Small, 2011; Knoop, 2011). Editing has not been reported in any other plastid lineage. Poly(U) tail addition and editing are thus unlikely to have occurred in the free-living ancestors that gave rise to serially acquired dinoflagellate plastids. The presence of either pathway in a serially acquired dinoflagellate plastid, alongside unambiguous evidence that poly(U) addition and editing were not present in free-living relatives of the replacement

Table 4.1 Primers for oligo-d(A) RT-PCRs for *Karenia mikimotoi*

Oligo-d(A) primer	GGGACTAGTCTCGAGAAAAAAAAAAAAAAAAAA	
Gene	**PCR forward primer**	**Internal cDNA primer**
psbA	GCTATCAGGCTCACTTTTATATGC	CCATCGTAGAAACTCCCATAG
psbC	CGACGGCTGCTGAAG	
psbD	GCTATTCACGGAGCGAC	
psaA	CACGTAGTTCAGCTCTGATACC	
rbcL	GATGCGTATGGCAGGTG	
PCNA	GCACTCGTCGCCCTC	AGTCGGGACCAAGGC
cox1	GATTGTTTGGAGGATTTGG	TCCACTGCTGCATTTCC
***A. carterae* psbA**	CTTCTAACGCAATCGGTGTCC	

plastid lineage, would provide definitive evidence that these plastids may be supported by pathways retained from previous symbioses.

I wished to determine whether either transcript poly(U) tail addition or editing occurs in serially acquired dinoflagellate plastid lineages. I investigated transcript processing events in *Karenia mikimotoi*, a particularly well-studied fucoxanthin dinoflagellate species that is of major ecological importance, as a component of harmful algal blooms (Brand *et al.*, 2012; Takishita *et al.*, 1999, 2000). I additionally investigated transcript processing events in representative dinotom (*Kryptoperidinium foliaceum*) and green dinoflagellate *(Lepidodinium chlorophorum*) species, which have been studied extensively elsewhere (Imanian *et al.*, 2012; Imanian *et al.*, 2010; Matsumoto *et al.*, 2011a; Matsumoto *et al.*, 2011b). I report that 3' poly(U) tail addition, and extensive substitutional sequence editing occur in the fucoxanthin plastids of *Karenia mikimotoi*, but not in the plastids found in dinotoms or green dinoflagellates. I show that these transcript processing pathways do not occur in the plastids of free-living haptophytes, confirming that they were retained from the ancestral peridinin plastid, and applied to the fucoxanthin plastid following its endosymbiotic acquisition. This demonstrates that the biology of replacement plastids can be dramatically remodelled by host functions remaining from previous symbioses.

Results

Plastid transcripts in *Karenia mikimotoi* receive poly(U) tails.

I wished to test whether transcripts in the serially-acquired plastids of *Karenia mikimotoi* received 3' poly(U) tails, as in the ancestral peridinin plastid lineage (Janouškovec *et al.*, 2010; Wang and Morse, 2006). cDNA was generated from total cellular RNA of *K. mikimotoi* using an oligo-(dA) primer. This primer has been shown to anneal to polyuridylylated

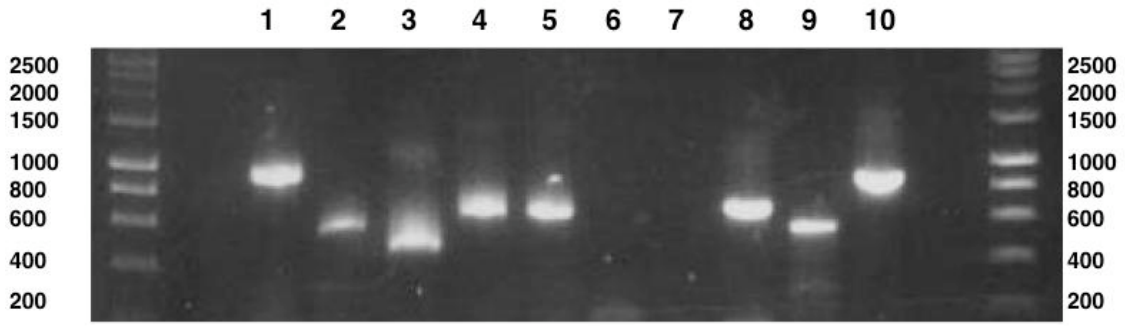


Fig. 4.1: Oligo-d(A) and gene-specific RT-PCRs for transcripts from *Karenia mikimotoi.*

This gel photograph displays the products from a series of RT-PCRs to detect polyuridylylated transcripts from *K. mikimotoi*. The size standard is DNA Hyperladder I (Bioline). **Lanes 1-5:** oligo-d(A) RT-PCR of *K. mikimotoi psbA, psbC, psbD, psaA* and *rbcL*. **6-7:** oligo-d(A) RT-PCR of *K. mikimotoi PCNA* and *cox1*. **8-10:** gene-specific RT-PCR of *K. mikimotoi psbA, PCNA* and *cox1*.

transcripts in the peridinin dinoflagellate *Amphidinium carterae* (Barbrook *et al.*, 2012). PCR reactions were then carried out using the oligo-(dA) primer as a reverse primer, and forward primers that annealed within specific genes (Fig. 4.1; Table 4.1). All five *K. mikimotoi* plastid transcripts tested (*psbA, psbC, psbD, psaA, rbcL*) gave PCR products of between 500 and 1000 bp. These were consistent (based on product size and annealing positions of the PCR forward primers employed) with monocistronic transcripts, possessing a poly(U) sequence in the 3' UTR (Fig. 4.1, lanes 1-5). Representative nuclear (*PCNA*) and mitochondrial sequences (*cox1*) for *K. mikimotoi* were also tested (Fig. 4.1, lanes 6-7). No products were amplified for either gene, whereas RT-PCRs using internal gene-specific cDNA primers that did not depend on a 3' poly(U) sequence generated products of the expected sizes (Fig. 4.1, lanes 8-10), indicating that poly(U) sequences are only found on plastid transcripts.

The products for each reaction were sequenced directly using the gene-specific PCR primer (Fig. 4.2, panel A). The sequences identified were very similar to previously published transcript sequences for *K. mikimotoi* (Takishita *et al.*, 1999; Takishita *et al.*, 2005), and much less similar to orthologous sequences from peridinin plastids. For example, the polyuridylylated *K. mikimotoi rbcL* transcript identified was of a form ID rubisco large subunit gene, as present in haptophytes and most other plastid lineages derived from red algae (Tabita *et al.*, 2008; Takishita *et al.*, 2000). In contrast, peridinin dinoflagellates and *C. velia*

Fig. 4.2: Cloned 3' ends of *K. mikimotoi* plastid transcripts.

This figure shows the 3' ends of individual cloned oligo-d(A) RT-PCR products of the *K. mikimotoi psbA, psbC, psbD, psaA* and *rbcL* genes, showing the 3' end of the CDS and the 3' UTR sequence. Termination codons are labelled with vertical black arrows. Transcript sequences are shown for each unique poly(U) site observed. The first 40 nt of longest poly(U) tail identified for each corresponding poly(U) site is shown. Numbers in parentheses correspond to the full length of the poly(U) sequence as sequenced in different clones. Asterisks correspond to the poly(U) tail position obtained by direct sequencing of crude RT-PCR products, i.e. the predominant poly(U) site utilised for transcripts of each gene.

possess a form II *rbcL* gene acquired through lateral gene transfer, which has replaced the form ID gene (Janouškovec *et al.*, 2010; Morse *et al.*, 1995). Thus, the PCR products amplified correspond to plastid transcripts from a fucoxanthin dinoflagellate, as opposed to contaminants from a peridinin plastid or other phylogenetic source.

Each transcript sequenced terminated at the 3' end in a poly(U) tract, similar to those previously reported in peridinin dinoflagellates and *Chromera velia* (Barbrook *et al.*, 2012; Janouškovec *et al.*, 2010; Wang and Morse, 2006). Each poly(U) tract began in the 3' UTR of the transcript concerned, 8-22 nt downstream of the translation termination codon (Fig.4.2, panel A). The poly(U) sites identified through direct sequencing of the PCR products correspond to the predominant poly(U) site associated with each gene. To determine whether alternative poly(U) sites were utilised by individual transcripts of each gene, each RT-PCR reaction product was cloned, and individual colonies were sequenced.

Table 4.2. Primers to amplify *K. mikimotoi* 3' UTR sequences.

1. Thermal asymmetric interlaced PCR.

	gene-specific primer 1	gene-specific primer 2	gene-specific primer 3
psbC	CGACGGCTGCTGAAG	CTCCTCTTGGTTCTTTAAATTCG	CCTGTTCTTTATATGCGTCCG
psbD	GCTATTCACGGAGCGAC	CAAACGGTGGTTACACTTCTTC	TGGTAATGGTCTCTAACACGTC
psaA	CACGTAGTTCAGCTCTGATACC	CCCCTTCTCAAGCAATCTC	CGACTACTACCCGCTAAAAGG
rbcL	GATGCGTATGGCAGGTG	CTCTCCGTAAATGCGTACC	AGTAAGTACAACTGGCGGGG

Arbitrary degenerate primers		Degeneracy
1	TTNTCGASTWTSGWGTT	64
2	CCTTNTWGAWTWTWGWWTT	256
3	TTWGTGNAGWANCANAGA	256
4	CCTTWGTGNAWWANCANAWA	256
5	GGAACWACNTWTWNGTNTTW	256
6	TTACWACANGWWGNTGNTWT	1024
7	GGAANACTWAWAWCWWAWA	1024
8	TTAANCWAGWCWCWAWWAA	1024

2. Confirmatory PCR.

	3' UTR reverse primer	5' end forward primer
psbA	GAGGTCTAATTTGAATGTCAGTG	CGGTTTCGTGTTGAAAATTG
psbC	TTTTAACGTTACATTAATACTTCTCTGG	TAGGTGCGCATGTGGCCC
psbD	AGTTGAGGAGAAGATTGAACG	TATCAGTGGGAGGTTGGTTAAC
psaA	CTAGCGGAATCAAATAAACGAC	TTCCTTAGATTGGTTTCAAAATG
rbcL	CTAAAAATTTAGAAAGGGATAATTGC	GATGCGTATGGCAGGTG

In the case of *psbA*, a single poly(U) site was observed in every clone, whereas the precise poly(U) site varied by up to 5 nt in all other genes (Fig. 4.2).

To confirm that the poly(U) tracts were not transcribed from the underlying genomic sequence, the 3' UTR of each gene from genomic DNA templates was amplified by thermal asymmetric interlaced PCR (TAiL-PCR) and sequenced. The sequence of each UTR sequence was then confirmed with a PCR, using a forward primer positioned within the CDS, and a reverse primer positioned within the proposed UTR sequence of the gene, as identified by TAiL-PCR (Table 4.2). For each gene investigated, the poly(U) sites identified did not correspond to poly(T) tracts in the underlying genomic 3' UTR sequence (Fig. 4.2).

Finally, to confirm that the poly(U) sequences represented 3' terminal modifications, as opposed to internal sequence insertions, RT-PCR was performed on circularised transcripts

Table 4.3. Primers for circular RT-PCR of *Karenia mikimotoi* RNA

Gene	cDNA primer	PCR reverse primer	PCR forward primer
psbA	CCATCGTAGAAACTCCCATAG	CAATTTTAGATGCTTGTGGATG	TACCCCCATTGTAAAGCC
psbC	CGTCCCTGCTATTTCACC	CAATCTAAGGAAGGAGCCG	CCTGTTCTTTATATGCGTCCG

Fig. 4.3: Circular RT-PCR products *K. mikimotoi psbA, psbC*.

This figure shows aligned 3' ends of *K. mikimotoi psbA* and *psbC* transcripts as obtained by circular RT-PCR, aligned against genomic and oligo-d(A) RT-PCR sequences, shown as per fig. 4.2. This alignment demonstrates that sequence-independent poly(U) sequences are added onto transcript 3' ends as opposed to inserted internally; transcript 5' ends are not displayed. The full length of each poly(U) sequence is shown.

of *psbA* and *psbC* (Fig. 4.3; Table 4.3). This technique has previously been used to identify polyuridylylated transcripts in *Amphidinium carterae* (Barbrook *et al.*, 2012). For both genes, products were identified that contained homopolymeric poly(U) sequences of between 15 and 30 nt, between the transcript termini. As the oligo-d(A) RT-PCR indicates that these sequences are located on the 3' end of the transcript, they must correspond to poly(U) tails. Non-polyuridylylated *psbA* transcripts were also identified through this approach, as have previously been identified in *A. carterae* (Barbrook *et al.*, 2012). However, all of the non-polyuridylylated transcripts identified terminated upstream of the *psbA* poly(U) site (Fig. 4.3), suggesting that they represent the 3' degradation products of polyuridylylated *psbA* transcripts, as opposed to *psbA* transcripts that have undergone alternative maturation events (Fig. 4.3). Thus, transcripts in fucoxanthin plastids are modified post-transcriptionally with 3' terminal poly(U) tails, as occurs in the ancestral peridinin plastid.

Editing of plastid transcripts in *K. mikimotoi.*

I additionally wished to determine whether plastid transcripts in *K. mikimotoi* were edited, as occurs in some peridinin dinoflagellate species (Howe *et al.*, 2008b; Zauner *et al.*, 2004). To do this, the sequences of oligo-d(A) primed RT-PCR products for each gene were compared to the corresponding sequences amplified from genomic DNA. To ensure that the sequences compared were correct, each oligo-d(A) primed RT-PCR was repeated twice, and each gDNA sequence amplified twice, using independently isolated nucleotide template samples.

Table 4.4: Overview of editing of plastid transcripts in *Karenia mikimotoi.*

This table lists all of the editing events observed within 5473bp polyuridylylated *K. mikimotoi* transcript sequence. Editing events are shown in the form (DNA sequence residue - Polyuridylylated transcript sequence residue).

Gene	psbA	psbC	psbD	psaA	rbcL	Total
Sequence length (bp)	1107	1225	892	1737	512	5473
Total editing events	52	37	22	117	32	260
% bases edited	4.70	3.02	2.47	6.74	6.25	4.75
A-C	2	4	2	16	2	26
A-G	7	5	1	31	15	59
C-A	1	0	0	0	0	1
C-U	5	1	4	5	2	17
G-A	2	4	1	5	3	15
G-C	3	5	0	11	5	24
U-C	31	18	14	48	5	116
U-G	1	0	0	1	0	2
of which	**psbA**	**psbC**	**psbD**	**psaA**	**rbcL**	**Total**
Completely edited	41	35	22	106	31	235
Partially edited	11	2	0	11	1	25
Non-synonymous	13	22	9	88	20	152
Synonymous	35	15	13	29	12	104
Codon position 1	7	15	6	47	15	90
Codon position 2	7	9	7	49	10	82
Codon position 3	34	13	9	21	8	85
In UTR	4	0	0	0	0	4
% complete	78.85	94.59	100.00	90.60	96.88	90.38
% partial	21.15	5.41	0.00	9.40	3.13	9.62
% non-synonymous	25.00	59.46	40.91	75.21	62.50	58.46
% position 1	13.46	40.54	27.27	40.17	46.88	34.62
% position 2	13.46	24.32	31.82	41.88	31.25	31.54
% position 3	65.38	35.14	40.91	17.95	25.00	32.69

An overview of the editing events observed across all five plastid transcripts is shown in Table 4.4, with detailed editing data for one exemplar transcript (*psaA*) shown in Table 4.5. Plastid transcripts were found to be extensively edited, with 4.8% of bases differing between corresponding oligo-d(A) RT-PCR and genomic DNA sequences (Table 4.4). Although the oligo-(dA) RT-PCR sequencing products should be representative of the entire population of transcripts and might therefore contain a mixture of edited and unedited sequences, only a small proportion of bases (8.1%) in the oligo-d(A) RT-PCR sequences were ambiguous (Table 4.4). Likewise, individual cloned RT-PCR products showed few differences in sequence (data not shown). Thus, editing at the majority of individual sites had essentially gone to completion.

Table 4.5. Detailed editing data for *K. mikimotoi psaA*.

This table lists all of the editing events observed within the polyuridylylated *psaA* transcript sequence. The predicted effect of each editing event on the transcript translation product is given, in the format (unedited translation product - edited translation product). Where no translation products are given, the editing event is predicted to have a synonymous effect on transcript sequence. Two events, shown in bold text, are predicted to remove in-frame premature termination codons.

Base	Editing	Extent	Position	Translation	Base	Editing	Extent	Position	Translation
2	A-C	Complete	2	H-P	656	A-C	Complete	2	Y-S
6	U-C	Complete	3	-	659	U-C	Complete	2	L-P
7	U-C	Complete	1	-	665	G-C	Complete	2	W-S
11	U-C	Complete	2	I-T	679	U-C	Complete	1	F-L
43	A-G	Complete	1	T-A	685	C-U	Complete	1	L-S
97	A-G	Complete	1	T-V	686	U-C	Complete	2	-
106	U-C	Complete	1	C-R	692	U-C	Complete	2	L-P
109	A-G	Complete	1	I-V	697	A-G	Complete	1	S-G
110	U-G	Partial	2	V-G	716	U-C	Complete	2	I-T
124	A-G	Complete	1	T-A	729	U-C	Complete	3	-
133	G-C	Complete	1	V-L	731	A-G	Complete	2	Q-R
146	A-C	Complete	2	K-T	749	G-C	Complete	2	S-T
175	A-G	Partial	1	T-A	752	A-C	Complete	2	N-T
196	C-U	Complete	1	L-F	760	A-G	Complete	1	T-A
200	A-C	Complete	2	K-T	786	U-C	Complete	3	-
206	G-C	Complete	2	G-A	794	U-C	Complete	2	M-T
212	U-C	Complete	2	V-A	807	U-C	Complete	3	-
229	U-C	Complete	1	F-L	810	C-U	Complete	3	F-L
278	G-C	Complete	2	S-T	**850**	**U-C**	**Complete**	**1**	**UAG Stop-Q**
310	A-G	Partial	1	I-V	855	G-C	Complete	3	E-D
318	U-C	Complete	3	-	857	U-C	Complete	2	I-T
322	C-U	Complete	1	L-F	878	U-C	Complete	2	L-S
340	A-C	Complete	1	-	883	G-A	Complete	1	D-S
342	U-C	Complete	3	S-R	884	A-G	Complete	2	-
347	U-C	Complete	2	V-A	896	U-C	Complete	2	V-A
379	U-C	Partial	1	-	922	A-G	Complete	1	K-E
387	U-C	Complete	3	V-A	949	A-G	Complete	1	S-A
406	A-G	Complete	1	K-G	950	G-C	Complete	2	-
407	A-G	Complete	2	-	**980**	**G-C**	**Complete**	**2**	**UAU Stop-S**
410	U-C	Complete	2	V-A	991	A-G	Complete	1	I-A
417	A-C	Complete	3	Q-H	992	U-C	Complete	2	-
419	A-C	Complete	2	K-T	993	U-C	Complete	3	-
424	A-C	Complete	1	I-L	994	A-G	Complete	1	I-V
427	A-G	Complete	1	N-D	1004	A-G	Complete	2	Q-R
440	G-C	Complete	2	R-T	1081	A-G	Complete	1	T-A
443	U-C	Partial	2	L-S	1093	U-C	Partial	1	-
481	G-A	Complete	1	V-M	1095	A-G	Partial	3	-
571	U-C	Complete	1	-	1096	A-C	Complete	1	I-L
577	A-G	Complete	1	T-V	1126	A-G	Complete	1	N-A
578	C-U	Complete	2	-	1127	A-C	Complete	2	-
603	U-C	Complete	3	-	1148	A-G	Complete	2	K-R
634	G-C	Complete	1	V-L					

Table 4.5 (continued)

Base	Editing	Extent	Position	Translation	Base	DNA	Extent	Position	Translation
1158	U-C	Complete	3	-	1438	U-C	Complete	1	S-P
1164	U-C	Complete	3	-	1441	A-G	Complete	1	T-A
1224	U-C	Partial	3	-	1453	A-G	Complete	1	T-A
1277	A-C	Complete	2	Y-S	1481	U-C	Partial	2	M-T
1307	U-C	Complete	2	I-T	1503	U-C	Complete	3	-
1313	U-C	Complete	2	I-T	1509	U-C	Complete	3	-
1315	U-C	Complete	1	S-P	1544	A-G	Complete	2	T-A
1324	A-G	Complete	1	T-A	1562	A-G	Complete	2	H-R
1331	U-C	Complete	2	L-S	1571	A-G	Complete	2	H-R
1355	A-C	Complete	2	E-A	1579	U-C	Complete	1	S-P
1357	A-C	Complete	1	S-R	1591	U-C	Complete	1	S-P
1359	U-C	Partial	3	-	1674	U-C	Partial	3	-
1360	G-A	Complete	1	D-S	1688	U-C	Complete	2	V-A
1361	A-G	Complete	2	-	1696	G-A	Complete	1	A-T
1405	G-C	Complete	1	E-Q	1697	A-C	Partial	2	D-A
1415	A-C	Complete	2	E-A	1699	U-C	Complete	1	C-R
1437	G-A	Complete	3	-	1700	A-G	Partial	2	Y-C

Eight types of base interconversion were identified, including transition and transversion substitutions (Table 4.4). Although the extent of bias varied, the *psbA*, *psbC*, *psbD* and *psaA* transcripts appeared to contain particularly high frequencies of uracil-to-cytosine and adenosine-to-guanosine conversions (Table 4.4). The diversity of substitutions observed in the *Karenia mikimotoi* plastids are similar to those observed in peridinin dinoflagellate plastids (Dang and Green, 2009; Zauner *et al.*, 2004). In contrast, the editing events found in plant plastids principally consist of cytosine-to-uracil conversions, while uracil-to-cytosine conversions are only identified in a few lineages, and interconversions between purine bases, and transversion substitutions have not been identified in any species (Fujii and Small, 2011; Knoop, 2011 Yoshinaga *et al.*, 1996). 58% of the editing events were predicted to result in non-synonymous substitutions, i.e. alter the translation product of the codon in question (Table 4.4). Notably, within the *psaA* transcripts, there were two instances where predicted premature in-frame termination codons were converted into coding sequence by editing (Table 4.5). It is therefore likely that editing of plastid transcripts plays an important role in enabling the expression of a functional photosystem I A1 subunit in the *K. mikimotoi* plastid (Table 4.5)

Absence of poly(U) tails and editing from haptophyte plastids

Previous studies of plastid transcription in taxa that are closely related to alveolate lineages, such as haptophytes and diatoms, have not reported the presence of either poly(U) tails or

Table 4.6. Primers for RT-PCRs of *Emiliania huxleyi* and *Phaeodactylum tricornutum*

1. Oligo-d(A) RT-PCR		
Oligo-d(A) primer	GGGACTAGTCTCGAGAAAAAAAAAAAAAAAAA	
Emiliania huxleyi	**Internal PCR forward primer**	**Internal cDNA primer**
psbA	AAAGCGCAAGCTTCTGG	AACTACTGGCCATGCACC
psbD	GTGACCGTTTCGTTTTCG	CGCCATCCATGAACG
psbC	CGTGGGCTCCAGGTG	
psaA	TTTGTGGGGCAGCAG	
rbcL	TGCGTTACCGTGAGCG	
Phaeodactylum tricornutum	**Internal PCR forward primer**	**Internal cDNA primer**
psbA	GCGGTTTTTGTGGTTGGATTAC	TAAAGCACGAGAGTTGTTAAATGAAG
psbD	GTGGCATTTTTGATCTAATTGACG	ACGTTTCAAATTCAGGATCTTCAG
psbC	CAGGTGGTGGCGATG	
psaA	ACGACCTGGGCCATC	
rbcL	GCTGCGATTTGGGCG	
2. Circular RT-PCR	***Emiliania huxleyi* psbA**	
cDNA primer	AACTACTGGCCATGCACC	
PCR reverse primer	CCAGAAGCTTGCGCTTT	
PCR forward primer	GCGTAACGCTCACAACTTCC	

editing of plastid transcripts (Fujiwara *et al.*, 1993; Hwang and Tabita, 1991). I wished to determine whether transcript poly(U) tail addition and editing occur in the plastids of haptophytes and other related lineages, or whether they are specifically associated with the plastids of dinoflagellates and other alveolates.

The presence of poly(U) tails and editing was investigated for plastid transcripts in the model haptophyte species *Emiliania huxleyi*, and in the diatom *Phaeodactylum tricornutum*, a representative of the stramenopiles, which are the closest related major lineage of photosynthetic eukaryotes to the alveolates (Janouškovec *et al.*, 2010; Puerta *et al.*, 2005). Oligo-d(A) primed RT-PCRs were performed for the *psbA, psbC, psbD, psaA* and *rbcL* transcripts of each species, using similar reaction conditions as before (Table 4.6; Fig. 4.4, panel A; lanes 1-5, 11-15). None of the transcripts were detected by oligo-d(A) primed RT-PCR, indicating that they do not possess poly(U) tails. The same results were observed when the primary product for each reaction was used as template for an additional 40 cycles of PCR amplification. *psbA* and *psbD* transcripts of each species could, however, be detected by RT-PCR using an internal gene-specific cDNA synthesis primer, as before (Table 4.6; Fig. 4.4, panel A; lanes 6-7, 16-17). As these transcripts could not be amplified with the oligo-d(A) primer, they are likely to be non-polyuridylylated.

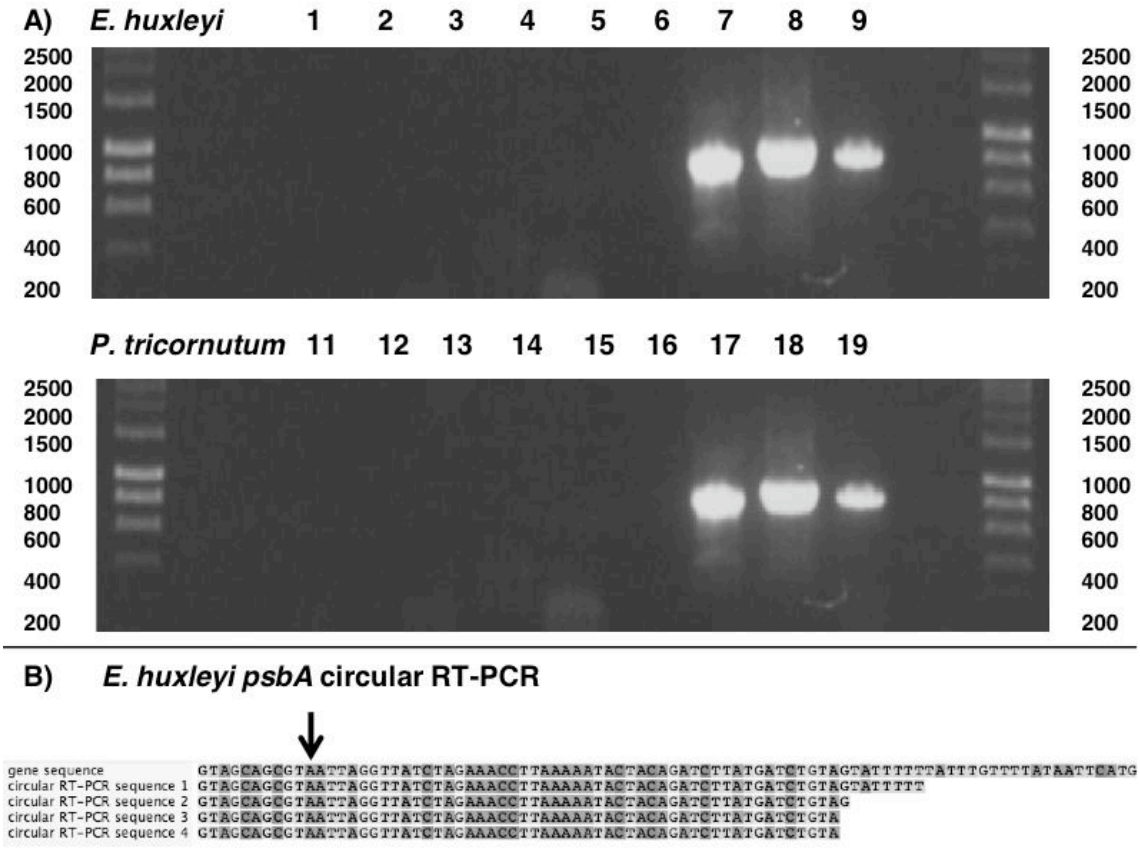


Fig. 4.4: Absence of poly(U) tails from plastid transcripts in *Emiliania huxleyi* and *Phaeodactylum tricornutum.*

Panel A shows gel photographs of a series of RT-PCRs to detect plastid transcripts in the haptophyte *Emiliania huxleyi* (top) and the diatom *Phaeodactylum tricornutum* (bottom). The size standard is DNA Hyperladder I (Bioline). Reactions are ordered identically for each panel. **Lanes 1-5, 11-15:** oligo-d(A) RT-PCR of *psbA, psbC, psbD, psaA* and *rbcL* from *E. huxleyi* (1-5) and *P. tricornutum* (11-15), indicating that polyuridylylated transcripts of each gene are not present. **6, 16:** reverse transcriptase negative control for gene-specific RT-PCR of *E. huxleyi* and *P. tricornutum psbA.* **7-8, 17-18:** gene-specific RT-PCRs of *E. huxleyi* and *P. tricornutum psbA* and *psbD*, indicating that non-polyuridylylated transcripts are present. **9, 19:** oligo-d(A) RT-PCR of *Amphidinium carterae psbA* (oligo-d(A) cDNA synthesis reaction positive control).

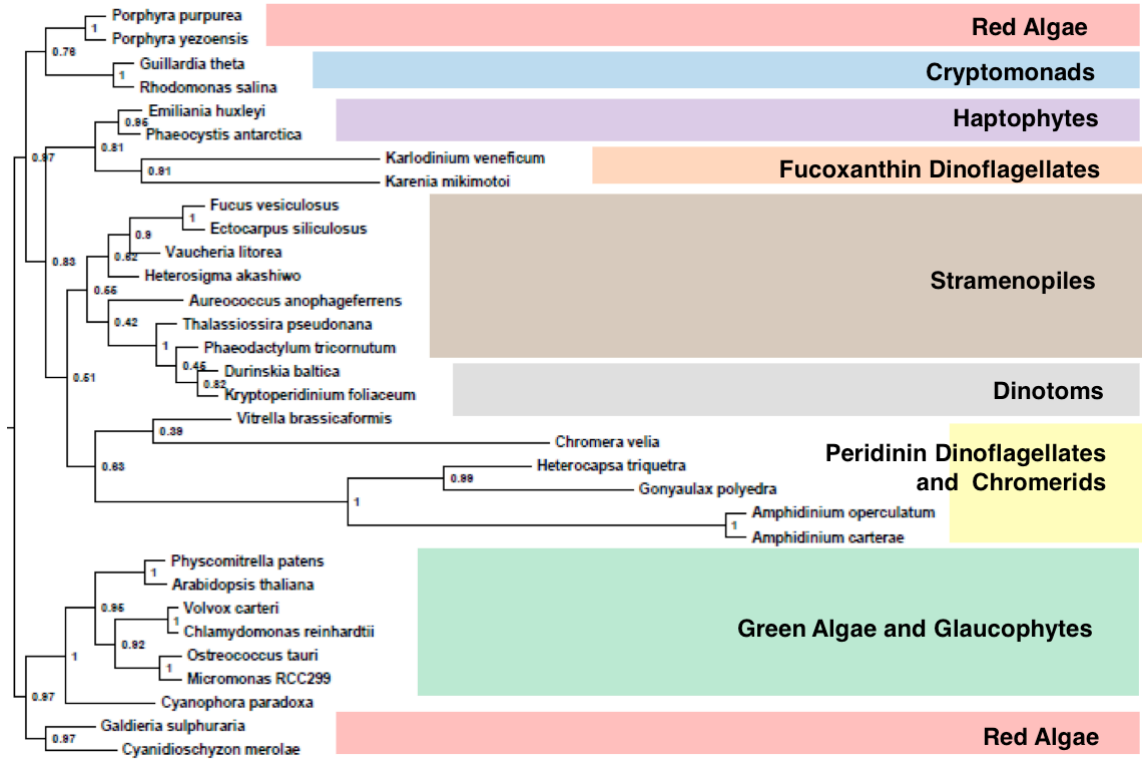
Panel B shows an alignment of the 3' ends of *E. huxleyi psbA* transcripts identified through circular RT-PCR, shown against the corresponding genomic 3' UTR sequence, as per Fig. 4.2. In each case, the transcript identified terminates in the 3' UTR, with no poly(U) tail or other 3' terminal modification.

To confirm that plastid transcripts in free-living haptophytes do not receive poly(U) tails, circular RT-PCR was performed on *E. huxleyi psbA* (Table 4.6; Fig. 4.4, panel B). All of the transcripts identified through this approach terminated in the 3' UTR, and did not possess a poly(U) tail or any further 3' end modification (Fig. 4.4, panel B). None of the transcripts sequenced for either species contained any evidence of editing. Thus, poly(U) tail addition and sequence editing are specifically associated within the plastids of dinoflagellates and their closest relatives within the alveolates (i.e. *C. velia*), and were most likely not present in the plastids of the free-living haptophyte ancestors of the fucoxanthin plastid.

Serial endosymbiotic remodelling of transcript processing in fucoxanthin plastids

The absence of poly(U) tail addition and editing from haptophyte plastids suggests that they originated in fucoxanthin plastids following a serial endosymbiotic event. Alternatively, these pathways may have originated much earlier in a common ancestor of the fucoxanthin and peridinin plastid lineages. The phylogenetic relationship between the peridinin-containing and fucoxanthin-containing plastid lineages has historically proved controversial. Early studies suggested that the fucoxanthin plastid lineage is a sister-group of the peridinin plastid, and that these plastids were acquired through a common endosymbiosis (Takishita *et al.*, 1999; Yoon *et al.*, 2002), although subsequent studies have indicated that the two plastid lineages have arisen through separate endosymbiotic events (Gabrielsen *et al.*, 2011; Inagaki *et al.*, 2004; Takishita *et al.*, 2005). Recent phylogenetic studies that have included plastid sequences from the chromerid algae *Chromera velia* and *Vitrella brassicaformis* have indicated that the peridinin dinoflagellate plastid is closely related to other alveolate plastid lineages, and thus represents the ancestral plastid type in dinoflagellates (Janouškovec *et al.*, 2010; Janouškovec *et al.*, 2012; Moore *et al.*, 2008). However, to date, no plastid phylogenies have been constructed that include sequences from fucoxanthin dinoflagellates as well as other alveolate plastid lineages.

I wished to confirm that the fucoxanthin plastid lineage arose through a serial endosymbiotic replacement of an ancestral peridinin-type plastid, and thus determine whether poly(U) tail addition and editing arose in the fucoxanthin plastids arose as a result of serial endosymbiosis. To test this, a concatenated alignment of four plastid genes (*psbA, psbC, psbD, psaA*) investigated in this study was constructed, including sequences from fucoxanthin and peridinin dinoflagellates, as well as sequences from chromerids, haptophytes and a broad sample of other plastid lineages. The *rbcL* gene was excluded as the form II isoform utilised by peridinin dinoflagellates and *Chromera velia* is understood to have been acquired via a recent lateral gene transfer event from a bacterial donor, and its inclusion might lead to the retrieval of artifactual phylogenetic relationships (Janouškovec *et*



Substitution model	Dayhoff	JTT	WAG
Fucoxanthin dinoflagellates			
Monophyletic	91	95	94
With haptophytes	81	99	100
With peridinin dinoflagellates	x	x	x
Peridinin dinoflagellates			
Monophyletic	100	100	100
With Chromera + Vitrella	63	93	100
(+Chromera + Vitrella) with Stramenopiles	51	86	88
With haptophytes	x	x	x
Control groups			
Cyanidiales monophyletic	97	97	98
Green Algae monophyletic	95	98	99
Diatoms (inc. dinotoms) monophyletic	100	100	100
Stramenopiles monophyletic	66	89	88

Fig. 4.5: PhyML protein phylogeny of concatenated *K. mikimotoi psaA, psbA, psbC,* and *psbD.*

This phylogeny, of a 32 x 1796aa protein alignment shows the phylogenetic derivation of the polyuridylylated transcripts sequenced in this study. The topology obtained with the Dayhoff substitution matrix is shown. The table below lists the bootstrap values obtained using the Dayhoff, JTT and WAG substitution matrices.

al., 2010; Morse *et al.*, 1995). Evolutionary relationships within this alignment were calculated using PhyML, and three different Γ-corrected substitution matrices (Dayhoff, JTT, WAG) (Fig. 4.5).

In each of the phylogenies, the fucoxanthin dinoflagellate and haptophyte plastids grouped together with robust bootstrap support, distinct from the peridinin plastids, confirming that the fucoxanthin plastid arose from a haptophyte endosymbiotic source (Fig. 4.5). In contrast, the peridinin plastids form a well supported group with the plastids of the chromerid algae *C. velia* and *V. brassicaformis* (Fig. 4.5).This confirms previous conclusions that the peridinin plastid shares a common endosymbiotic ancestry to the plastids found in other alveolate lineages, and that this plastid is likely to have been acquired by an ancestor of all extant dinoflagellates (Janouškovec *et al.*, 2010). Thus, the haptophyte derived plastid in the fucoxanthin dinoflagellates serially replaced this original plastid lineage. Consistent with previous studies, the plastids of peridinin dinoflagellates and chromerids together form a sister group to the plastids of stramenopiles (Fig. 4.5) (Janouškovec *et al.*, 2010). As stramenopile plastid transcripts do not possess poly(U) tails or undergo editing, this indicates that this machinery arose independently in the peridinin dinoflagellate lineage, and was not secondarily lost from the free-living haptophytes studied (Fig. 4.4, panel A). Thus, the poly(U) tail addition and editing pathways associated with fucoxanthin dinoflagellate plastids were retained from an ancestral peridinin plastid symbiosis, and applied to the replacement fucoxanthin plastid following its serial endosymbiotic acquisition.

The fucoxanthin and peridinin dinoflagellates formed exceptionally long branches on the phylogenies obtained (Fig. 4.5). This raises the question of whether the phylogenetic associations recovered are genuine, or artifacts caused by fast sequence evolution within the dinoflagellates. To determine whether the phylogenetic relationships obtained in this study were genuine, fast-evolving sites were progressively removed from the alignment (Dacks *et al.*, 2002; Hampl *et al.*, 2009) (Table 4.7). The total conservation of each site within the alignment was calculated, and alignments were constructed that only contained sites with fixed threshold levels of conservation. Phylogenies were constructed for each alignment using the JTT substitution matrix (Table 4.7).

Removal of fast-evolving sites from the total alignment disrupted the phylogenetic affinity of the fucoxanthin lineage for the haptophytes, and of the peridinin lineage for the stramenopiles (Table 4.7). However, no consistent alternative topology was obtained following fast site removal. The phylogenetic affinities obtained within these trees for the fucoxanthin dinoflagellates were generally weakly supported, and several were clearly artifactual, for example identifying separate phylogenetic affinities for the *Karenia mikimotoi*

Table 4.7. Effects of fast site removal on relationships recovered by PhyML phylogenies

This table shows the bootstrap support obtained for a series of clades in trees constructed with and without peridinin dinoflagellate sequences, using the JTT matrix with Γ correction. "x" denotes that the given relationship was not retained.

	Alignment of sites with conservation >						
1. With peridinin dinoflagellates	**0%**	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**
Fucoxanthin dinoflagellates							
Monophyly	95	x	x	x	x	60	63
With haptophytes	99	x	x	x	x	x	11
With peridinin dinoflagellates	x	x	x	x	34	23	x
Karenia only with peridinin dinoflagellates	x	89	18	23	25	x	x
Karlodinium only with haptophytes	x	51	16	20	x	x	x
Peridinin dinoflagellates							
Monophyly	100	100	100	100	100	100	100
With Chromera + Vitrella	93	x	x	x	x	x	59
(+Chromera + Vitrella) with Stramenopiles	86	x	x	x	x	x	x
With haptophytes	x	96	34	45	x	18	x
Control groups							
Cyanidiales monophyletic	97	88	71	74	55	57	49
Green Algae monophyletic	98	100	100	100	99	97	99
Diatoms (inc. dinotoms) monophyletic	100	91	100	98	95	30	x
Stramenopiles monophyletic	89	x	36	46	x	x	x
2. Without peridinin dinoflagellates	**0**	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**
Fucoxanthin dinoflagellates							
Monophyly	89	88	91	88	92	93	95
With haptophytes	57	90	99	88	98	77	87
With Chromera + Vitrella	x	x	x	x	x	x	x
Chromera + Vitrella							
Monophyly	99	93	99	90	99	91	94
With Stramenopiles	81	55	84	46	64	63	x
With haptophytes	x	x	x	x	x	x	x
Control groups							
Cyanidiales monophyletic	68	67	69	65	77	62	79
Green Algae monophyletic	94	81	94	84	93	90	95
Diatoms (inc. dinotoms) monophyletic	94	98	94	98	94	99	20
Stramenopiles monophyletic	x	x	x	x	x	29	x

and *Karlodinium veneficum* plastids (Table 4.7). Only one tree (sites with >10% conservation) produced robust support for a clade of fucoxanthin and peridinin plastids, and this association was limited to *Karenia mikimotoi*, with *Karlodinium veneficum* grouping with moderate support with the haptophytes (Table 4.7). In addition, other phylogenetic groups

Table 4.8. Primers for RT-PCRs of *Kryptoperidinium* and *Lepidodinium*

1. Oligo-d(A) primed RT-PCR

Oligo-d(A) primer	GGGACTAGTCTCGAGAAAAAAAAAAAAAAAAAA

Kryptoperidinium foliaceum	**PCR forward primer**	**Internal cDNA primer**
psbA	GCAACACCAGCCATGTG	CGCAGCTCCTCCAGTTG
psbC	GCTTTCGTTTGGTCAGG	
psbD	GTCCAGAAGCACAAGGTG	
psaA	TGTGATGGTCCAGGTCG	
rbcL	GAAGCAGAGCAGCAGTAG	

Lepidodinium chlorophorum	**PCR forward primer**	**Internal cDNA primer**
psbA	ACATCATTTCGGGAGCC	CCGATAACAGGCCAAGC
psbC	TGGGTGCCATTTCGG	
psbD	CTTTGCGCTATTCACGG	
psaA	ATCGCCCATCACCATC	
rbcL	CAGTTTGGGGGTGGTACTC	

2. Circular RT-PCR	***K. foliaceum* psbA**	***L. chlorophorum* psbA**
cDNA primer	CGCAGCTCCTCCAGTTG	CCGATAACAGGCCAAGC
PCR reverse primer	AACCAAACCAACCGATG	CTACAGGAGGAGCAGCG
PCR forward primer	CACAATGGCGTTCAAC	CTGTAGTAGATTCTCAAGGACGTG

well supported from previous studies (e.g. monophyly of the cyanidiales, and of the diatoms) were disrupted by the fast site removal, so it is unlikely that any novel relationships uncovered within these trees were genuine (Table 4.7).

It is possible that the results obtained within the fast site removal phylogenies were the results of additional experimental artifacts that could not be contained by eliminating fast-evolving sites. In particular, plastid sequences from peridinin dinoflagellates are known to contain several other sources of phylogenetic artifacts, including uneven rate evolution, and unusual patterns of codon usage (Inagaki *et al.*, 2004; Shalchian-Tabrizi *et al.*, 2006). To avoid potential artifacts generated from within the peridinin dinoflagellates, the fast site removal series was repeated, using an alignment from which the peridinin dinoflagellate sequences were removed. Sequences from *Chromera velia* and *Vitrella brassicaformis* were retained as representatives of the peridinin plastid lineage (Janouškovec *et al.*, 2010). In the absence of peridinin dinoflagellate sequences, the fucoxanthin dinoflagellates grouped with moderate support with the haptophytes in the complete phylogeny, and with robust support in each of the fast site removal phylogenies. *Chromera* and *Vitrella* grouped either within or as sister to the stramenopiles in all but one fast site removal phylogeny and never grouped with the haptophytes. Thus, the separate origins identified of the peridinin and fucoxanthin

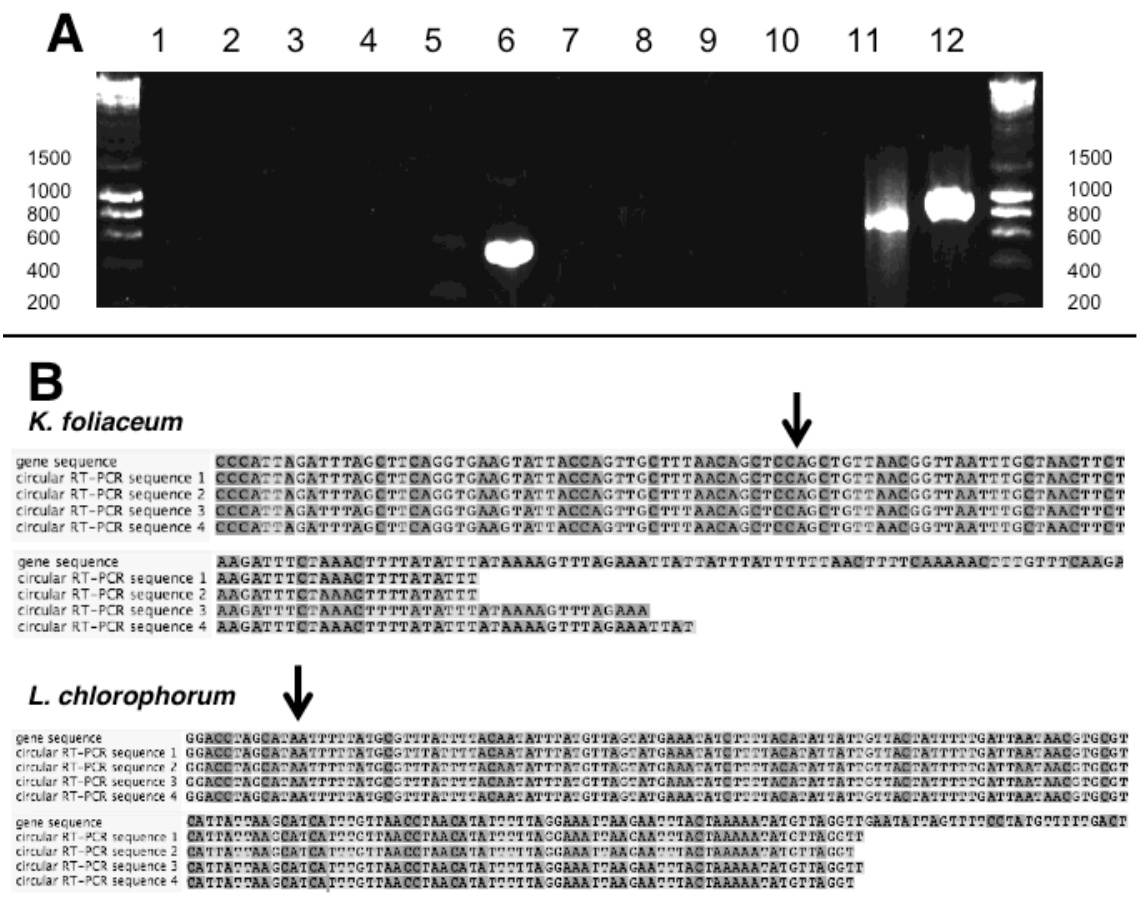


Fig. 4.6: Poly(U) tail addition is found only in fucoxanthin-containing serial dinoflagellate plastids.

Panel A shows a gel photo for a series of oligo-d(A) RT-PCRs against representative dinotom *(Kryptoperidinium foliaceum*) and green dinoflagellate *(Lepidodinium chlorophorum*) plastid transcripts. **Lanes 1-5, 7-11**: oligo-d(A) RT-PCRs *psbA, psbC, psbD, psaA* and *rbcL*, for *K. foliaceum* (**1-5**) and *L. chlorophorum* (**7-11**). The RT-PCR product from lane 5 was found to be a PCR chimera. **6, 12**: RT-PCRs of *K. foliaceum* (**6**) and *L. chlorophorum* (**12**) *psbA*, with internal gene-specific cDNA synthesis primers, and the same PCR forward primer used for the corresponding oligo-d(A) primed RT-PCR. **13:** reaction positive control.

Panel B shows the aligned 3' termini of *Kryptoperidinium foliaceum* and *Lepidodinium chlorophorum psbA* transcript sequences, as identified by circular RT-PCR, with the underlying genomic sequence as per Fig.4.2. In each case, the transcript sequences identified terminate within the 3' UTR of the gene, without a poly(U) tail or any other 3' terminal modification. This confirms that transcripts in dinotom and green dinoflagellate plastids do not receive poly(U) tails.

lineage plastids are not the result of phylogenetic artifact, and poly(U) addition and editing were likely to have been acquired by the fucoxanthin plastid following its serial endosymbiotic acquisition by the dinoflagellate host.

Absence of poly(U) tail addition and editing from diatom and green algal-derived serially acquired dinoflagellate plastids

I wished to determine whether poly(U) tail addition and transcript editing are found in either dinotom or green dinoflagellate plastids, as in the fucoxanthin and peridinin-containing lineages. As before, oligo-d(A) primed cDNA was generated from total cellular RNA of the dinotom *Kryptoperidinium foliaceum* and green dinoflagellate *Lepidodinium chlorophorum* (Table 4.7). PCRs were then performed as before using the oligo-d(A) primed cDNA and PCR reverse primer, and PCR forward primers specific to five genes (*psbA, psbC, psbD, psaA, rbcL*) from each species (Table 4.7; Fig. 4.6, panel A; lanes 1-5, 7-11). In each case, products corresponding to polyuridylylated transcripts could not be obtained. The same results were observed when the primary product for each reaction was used as template for an additional 40 cycles of PCR amplification.

As before, non-polyuridylylated *psbA* transcripts were detected for both species by RT-PCR using gene-specific cDNA synthesis primers (Fig. 4.6, panel A; lanes 6, 12), and by circular RT-PCR (Table 4.7; Fig. 4.6, panel B). None of the transcripts sequenced for either species contained any evidence of editing. Thus, poly(U) tail addition and editing are found only in dinoflagellates that possess the ancestral peridinin plastid, or the fucoxanthin replacement lineage.

Discussion

My data show that 3' terminal poly(U) tails are added during RNA processing in the plastids of the fucoxanthin dinoflagellate *Karenia mikimotoi*, as seen in the ancestral plastids of peridinin dinoflagellates and *Chromera velia* (Figs. 4.1-4.3). I additionally show that transcripts from *K. mikimotoi* plastids are subject to extensive base editing, as observed in some peridinin dinoflagellate plastids (Table 4.3). Subsequent to the experiments discussed in this chapter, the presence of editing has been independently reported in the related fucoxanthin dinoflagellate species *Karlodinium veneficum* (Jackson *et al*., 2013).

As there is no evidence for plastid transcript poly(U) tail addition or editing in either the haptophyte *Emiliania huxleyi* or the diatom *Phaeodactylum tricornutum*, the most parsimonious explanation is that these transcript processing pathways arose within the alveolates (Figs. 4.4, 4.5). These transcript processing pathways were therefore retained

from the peridinin plastid and applied to the replacement fucoxanthin plastid lineage, dramatically altering its RNA metabolism. These pathways have not, however, been acquired by other the plastids of other dinoflagellates that have acquired replacement plastids, such as dinotoms and green dinoflagellates (Fig. 4.6).

Since their origin in the fucoxanthin plastid, poly(U) tail addition and editing appear to have become widespread features of plastid transcript processing, as inferred by the large number of polyuridylylated transcripts identified by circular RT-PCR (Fig. 4.3) and the broad distribution of editing sites across the coding sequences studied (Table 4.3). Certain editing events, such as the removal of premature in-frame termination codons from the *K. mikimotoi psaA* transcripts, may have important roles in enabling the functional expression of plastid genes. Similar events, in which editing enables the translation of a complete open reading frame, have previously been documented in plastid transcript editing events in plants, and in peridinin dinoflagellates (Hoch *et al.*, 1991; Yoshinaga *et al.*, 1996; Zauner *et al.*, 2004).

The application of the pathways to the *K. mikimotoi* plastid *rbcL* transcript is particularly striking, as the *rbcL* gene of peridinin dinoflagellates and *C. velia* is located in the nucleus and its transcripts do not receive a poly(U) tail (Janouškovec *et al.*, 2010; Morse *et al.*, 1995). Moreover, the *K. mikimotoi rbcL* gene encodes a form ID enzyme (comprising 8 large and 8 small subunits), as is found in haptophytes and most other plastids descended from red algae (Tabita *et al.*, 2008; Takishita *et al.*, 2000). In contrast, peridinin dinoflagellates and *C. velia* possess a form II gene (comprising 2 large subunits only), which is believed to have been acquired by a lateral gene transfer from a bacterial donor, and replaced the ancestral ID form gene (Janouškovec *et al.*, 2010; Morse *et al.*, 1995; Tabita et al., 2008). Thus, poly(U) tails and editing can be successfully applied to plastid transcripts that do not have direct orthologues in peridinin-containing dinoflagellates.

Overall, my observations suggest an important addition to conventional models of plastid evolution. My data prove that host lineages can retain plastid-associated pathways from prior symbioses and apply them to replacement plastids, in which they may confer novel functions. These pathways might enhance the stability of the replacement plastid in the host cell, or customise the metabolic or regulatory pathways found within the plastid to the physiological requirements of the host (Dorrell and Howe, 2012b; Howe *et al.*, 2008a; Larkum *et al.*, 2007). In the light of recent data indicating that serial endosymbiosis may have occurred extensively across the eukaryotes (Dorrell and Howe, 2012b; Huang and Gogarten, 2007; Moustafa *et al.*, 2009), I propose that the biology of many prominent plastid lineages in eukaryotes may have been altered by functions derived from preceding endosymbioses.

Chapter Five- Plastid genome sequences and transcript processing pathways have evolved together in the fucoxanthin dinoflagellate *Karlodinium veneficum*

Introduction

Plastid gene expression involves a complex set of post-transcriptional processing events, including transcript cleavage, splicing, substitutional editing, and 3' end modification (Barkan, 2011). Many of these transcript processing have evolved independently in specific plastid lineages. For example, trans-splicing of transcripts of fragmented plastid genes has emerged independently in plants, green algae and one red algal species (*Rhodella*), but has not been documented in plastids acquired via the secondary endosymbiosis of a red alga (Glanz and Kück, 2009; Richaud and Zabulon, 1997). Similarly, post-transcriptional editing of transcript sequences has evolved in plants since their divergence from green algae (Fujii and Small, 2011; Yoshinaga *et al.*, 1996).

The presence of specific transcript processing pathways in individual plastid lineages may influence the evolution of the underlying genome sequence. For example, transcript editing in plant plastids, which is predominantly involved in cytosine deamination, is believed to have coevolved with an enrichment in the GC-content of the underlying genome sequence, relative to the plastids of closely related green algae (Fujii and Small, 2011). Studying the coevolution of plastid genes and genome sequences, however, is complicated by the fact that many of the major plastid lineages are evolutionarily ancient, and cannot provide direct insight into the events that occur shortly following endosymbiotic plastid acquisition (Parfrey *et al.*, 2011; Wellman and Gray, 2000).

The plastids of fucoxanthin dinoflagellates present an ideal system for exploring the coevolution of plastid genomes and transcript processing pathways. As I have previously discussed, the fucoxanthin plastid was acquired through a serial endosymbiotic replacement of the ancestral peridinin-containing dinoflagellate plastid with one derived from a haptophyte alga (Dorrell and Howe, 2012b). This must have occurred after the Gymnodiniaceae, the lineage containing the fucoxanthin dinoflagellates, diverged from other dinoflagellate species (Bachvaroff *et al.*, 2014). This is believed to have occurred- from molecular and fossil evidence- no earlier than 250 million years ago (Medlin, 2011; Parfrey *et al.*, 2011). Thus, the fucoxanthin plastid represents one of the most recently acquired plastid lineages known.

A near-complete plastid genome sequence has been determined for the fucoxanthin dinoflagellate *Karlodinium veneficum* (Gabrielsen *et al.*, 2011). This genome is highly divergent, having lost over forty of the genes associated with the plastid genomes of free-living haptophytes, having undergone extensive rearrangement, and containing large regions of coding sequence that have little conservation to previously studied plastid genes (Gabrielsen *et al.*, 2011). It has been suggested that certain genes within the *K. veneficum* plastid genome (e.g. *rbcL, dnaK*) are located on small episomal elements in addition to on the main chromosomal plastid genome, (Espelund *et al.*, 2012). These may constitute an independently evolved population of plastid minicircles to those observed in peridinin dinoflagellate plastids (Howe *et al.*, 2008b; Zhang *et al.*, 1999), although the complete sequence of a fucoxanthin plastid minicircle has yet to be obtained. Overall, it therefore appears that the fucoxanthin plastid genome has undergone rapid post-endosymbiotic evolution.

Previously, I have shown that two pathways associated with the peridinin plastid, 3' terminal poly(U) tail addition and sequence editing, also occur in plastids of the fucoxanthin dinoflagellate *Karenia mikimotoi* (Dorrell and Howe, 2012a). Since the publication of these data, sequence editing has also been demonstrated to occur in plastid transcripts from *Karlodinium veneficum*, suggesting that these pathways were acquired by a common ancestor of extant fucoxanthin dinoflagellates, although poly(U) tails have not yet been reported in this species (Jackson *et al.*, 2013). I have additionally shown that poly(U) tail addition and editing are not found in free-living haptophytes, and are thus likely to be remnants of the ancestral peridinin-containing plastid symbiosis, applied to the fucoxanthin plastid following its endosymbiotic acquisition by the dinoflagellate host (Dorrell and Howe, 2012a). As these pathways are therefore very recently acquired (and have been applied to transcripts of a fast-evolving genome), they provide a unique opportunity to explore the coevolution of plastid genes and gene expression pathways.

This project was conceived to investigate how poly(U) tail addition and editing have been adapted to function across the entire genome of a fucoxanthin dinoflagellate plastid. I wished to determine whether poly(U) tail addition and editing are associated with transcripts of every gene in fucoxanthin plastids, or are associated with transcripts of some plastid genes and not others. I accordingly surveyed the distribution of poly(U) tail addition and editing sites across the entire published plastid genome sequence of the fucoxanthin dinoflagellate *Karlodinium veneficum* (Espelund *et al.*, 2012; Gabrielsen *et al.*, 2011). This represents the first genome-wide study of transcript processing in an algal plastid lineage. I demonstrate that almost every gene in the *K. veneficum* plastid can give rise to polyuridylylated and edited transcripts,

as occurs in peridinin plastids (Barbrook *et al.*, 2012; Howe *et al.*, 2008b), suggesting that each pathway has become a widespread feature of fucoxanthin plastid transcript processing since the serial endosymbiosis event.

In addition, I wished to investigate whether transcript processing events in the *K. veneficum* plastid have been influenced by the extremely unusual evolution of the plastid genome sequence. I have identified unusual roles for poly(U) tail addition and editing on transcripts of highly divergent regions of the *K. veneficum* plastid genome. Poly(U) tail addition may enable the differentiation of mRNAs generated from functional genes, from transcripts of pseudogenes that have arisen through recent genome rearrangements. Editing events are particularly associated with fast-evolving sequences and in-frame insertions that have arisen recently in fucoxanthin dinoflagellate plastids, and might constrain the phenotypic consequences of these highly divergent sequences on plastid protein function. I additionally present evidence that these pathways may have indirectly contributed to the evolution of highly divergent sequences, such as a novel 3' extension to the *atpA* coding sequence (CDS) that is generated through transcript editing.

Finally, I wished to confirm whether any genes in the *K. veneficum* plastid genome are located on episomal minicircles. I present the first complete sequence of an episomal minicircle in a serially acquired dinoflagellate plastid, which contains a complete *dnaK* gene, and has evolved convergently to the minicircles found in peridinin dinoflagellate plastids. Transcripts of this minicircle receive poly(U) tails and are edited, indicating that the pathways underpinning these processing events have adapted to the fragmentation of the *K. veneficum* plastid genome. Overall, my data reveal extensive and complex coevolutionary trends between the plastid genome sequence and transcript processing machinery of fucoxanthin dinoflagellates.

Results

Poly(U) tail addition was established in a common ancestor of extant fucoxanthin plastids

I wished to determine whether poly(U) tail addition, as has been documented to occur in *Karenia mikimotoi*, occurs in plastids of the fucoxanthin dinoflagellate *Karlodinium veneficum* (Dorrell and Howe, 2012a). As these two species are distantly related within the fucoxanthin dinoflagellates, this would indicate that poly(U) tail addition was acquired by a common ancestor of all extant fucoxanthin plastid lineages, following its endosymbiotic acquisition (Bergholtz *et al.*, 2006). cDNA was synthesised from *Karlodinium veneficum* total cellular

Table 5.1. Primers for oligo-d(A) primed RT-PCR of *K. veneficum* transcripts.

Genes in bold text are those shown in Fig. 5.1. The terminus positions of transfer RNA and novel ORF genes are given relative to the published *K. veneficum* plastid genome sequence (Gabrielsen *et. al.,* 2011).

Oligo-d(A) primer	GGGACTAGTCTCGAGAAAAAAAAAAAAAAAAA

1. mRNAs

Gene	**PCR forward primer**	**Gene**	**PCR forward primer**
atpA	TAGCACAAGCGAACGCACTA	rpl14	CGACGTGAATTAAAACGC
atpB	ATCCCTTCTCGCAGTTCAGC	rpl16	CGACGGGGCAATTTCTTCAC
atpE	CAAACTCCAGCACGAAAATAG	rpl19	CTTTTGAAGGCAACATTATTGC
atpF-1	GCAGCGCCTTTGTAGTAGG	rpl2	CACCTGGAACACGAGGCAAA
atpF-2	TTACTTAGGAGAGTATTTGTACTAAAATTAC	rpl27	TGGAAGAGATTCTATAGCAAAACGA
atpG	GCCAACAATGGCAATTC	rpl3	GCCATTCAGGTAGCATATTC
atpH	GCGTCAGTTTTAGCTGCTG	rpl31	GCTCGAGTTTATCTTGATGACC
atpI	CGAATATGCTGCTCCAAC	rpl33	CGATTTTTTGCCATAAATTC
cbbX	GCAGCCTTACTTCTTATTCAACGA	rpl36	TGAGAGTTGTTAGTTCTTTTGGTAAC
chlI	AAGACAGAGTTTGCGGGAC	rpl5	TGAACAGCTGCCTAAGATACGA
clpC	GTTACGAGGAAGGGGGAC	**rpl6**	**TCGTTATCCGCACCAAGGAG**
groEL	TCGGCGCTTCCACATTAACT	rpoA	GCCAAGGAAAGATTATACCGTC
ORF1	**ATGCTTCTGGCGATACCG**	rpoB	AACAGGATAAACAACAAAGCGT
petA	CGCTGCAACTGGTAAATCCG	rpoC1	CGAGCCCCTACATTACATCGT
petB	GGGGCATACTCGTGGTTTCA	rpoC2	GCAATGGTTTTATAGTGCGG
petD	AATGATTACGGAGAACCGGCA	rps10	TCAGTGTGTTCGTACTACG
petG	CAGTGATTACGGGTAACTTATGTG	rps11	GAAGAGGGAAAAGTTGCG
psaA	**AACCTGGGCACTTTTCGACT**	rps12	CGCAACGACGAGGAG
psaB	GCAGATTTAGTAGATCCGTCAACA	rps13	TGCAGGAGTAAGACTACCCAG
psaC	GTTCTTGAAATGGTACCGTG	rps14	ACGACGTGAATTAAAACGCGA
psaF-psaJ	GGATGGTTTACCACGAATTGA	rps16	CATTAACCCAGATGAGATAAATTTAGTC
psbA	**AGAGAGCCAGTTGCAGGTTC**	rps18	AGACGTGATCTTAGATTTAAGGC
psbB	ACGAACCACTAGGATTTGTTCCA	rps19	TCTTATTTGGACATGGTCACGTT
psbC	**TACTTGGGTTAGGCGGG**	rps2	CAACTGCTCTTATTAAAGTGTCG
psbD	**AATTGCGCGCTTAGTGG**	rps20	TTACGAAAGGTCAAATGATGAAAGC
psbE	TGGTTCAACAGGTGAAAGGC	rps3	CCAACAGGATTTCGATTGGGT
psbF	GAAGAAGAAGAAATAATCTATGAAAG	rps4	TCTCGTTATACAGGTCCGAAG
psbH	GATCCGCAAAGATACAAACC	**rps5**	**CGAGCTCGAAATGTTCGCAT**
psbI	CGGATGCTGCGGTAC	rps7	CGTAATAAGTATCGGCGTCG
psbK	ACCTGAATCATACCGTTTATTTCGT	rps8	ACATGTATCCGCAATGCTAGTT
psbL	AGAGAAAAAGAGCATCCTTGGG	rps9	CACGCACATATTCGGC
psbN	CAAATACGTCATACAGTGGG	secA	GCTGATCTATTTCTGGAAGAATC
psbT	GAGAATTCGGTTTATCACTAACG	secY	AGGACGAAGGTCAATATGGTCA
psbV	ATCCTGCGATAGCATTAGACTTGA	tufA	GGGCAATGCCTCAAAC
rbcL	**ACTGGGGCAACCATGG**	ycf3	GGGCTGGTCTATCTGCAC
rbcS-1	CCGCAAAACATATTATATTAAACG	ycf39	GGTGTTTAAAGCAGTGGG
rbcS-2	CACATATTAAATGTAGATTCGATTTG	ycf4	ACTATGGTGGGAGGAAAAGGTG

2. gene-specific cDNA primers

Gene	**cDNA primer sequence**	**Gene**	**cDNA primer sequence**
atpF-1	CCTACTACAAAGGCGCTGC	**rpoC1**	AAACATTTTCGTTCTCCCC
atpF-2	GACAAGGCTAATCAAGTTGAGG	**rps13**	CTGTTCGTTGACCTCG
psbA	**TCATAGCTAGGTTACAGAGGGG**	**rps18**	GATCATACCTTACCCCCG
psbN	TTATGTGAAATTATTTTCATGTTCTTC	**rps2**	CCTCCCTTCCACTTGTG
rbcS-1	GGAACGGGAGAGCTACG	**secY**	CCACTGATAGACTGCTAGTGCAA
rpl19	TGGTAATATAGGGGTTCGATG	**ycf39**	AAGTAAAGCTAACTTCCTTTTCCAA

Table 5.1 (continued)

2. ribosomal RNAs	PCR forward primer 1	PCR forward primer 2	cDNA primer sequence
rrf	GACGTTAGATAGCATAGTTGTTCC	GACGTTAGATAGCATAGTTGTTC	CAGCGTTCATCCTGAGCCA
rrl	GGTGGTCAGTTTGACTGG	TCCATATCGACGGGGAGGTT	ACTACCCTCCTAAAAACTCTTCACA
rrs	CCCCTGTAGTCCTAGCCG	GGGAGCGAAAGGGATTAG	ACCTTCCAGTACGGCTACCT

3. novel ORFs	ORF start	ORF end	PCR forward primer
ORF1	**104346**	**104951**	**ATGCTTCTGGCGATACCG**
ORF2	6197	5478	CCTGCGTCTCAAAAAGG
ORF3	14770	14237	ACACCCTCTTGTGTTGCTG
ORF4	83027	83437	ACTACTTTTACAACGAACTTGTCC
ORF5	105383	105045	GGTGTTTCGCCGAGAAGAAG

4. transfer RNAs	tRNA start	tRNA end	PCR forward primer
Arg^{TCT}	108690	108618	TTACTAATGTTACTTTGTTAGCGC
Asn^{GTT}	107484	107555	TCACATGCGTAGGTTTGATGG
Asp^{ACG}	71421	71349	TAGTGAGTGACAGGGCTACTG
Cys^{GCA}	42095	42165	ATCCCTATATCAAAGATGGTGAC
Glu^{TTC}	47468	47540	CCCGTCGTCTAGCGAAC
Gly^{TCC}	4197	4127	CGAACGTTACAGATGTGATACTG
His^{GTG}	33007	33079	TGCAATTGATCGATAAGGCG
Ile^{GAT}	137407	137335	AAAGGTGAGATGTTGTCGTGG
Leu^{TAG}	24493	24415	TGTCGACTTTTCCAAACGATC
Lys^{TTT}	59580	59508	TGAGTCCGATGCCTTCAACC
Met^{CAT}	118317	118245	CTACCGCTGAGCTATCTGGG
Phe^{GAA}	118638	118709	TTTTTCTCCTCCAAATTGCCTTC
Pro^{TGG}	77210	77283	ACCGTTTCTTCCTGGGGAC
Ser^{TGA}	87315	87399	TGACGGAGTGAGGGATGAG
Tyr^{GTA}	15608	15688	TCCGTTCTTACAAACTCAACG

RNA using an oligo-d(A) primer, as described previously (Dorrell and Howe, 2012a). PCR was then performed using the same oligo-d(A) primer as the PCR reverse primer, and forward primers specific to a representative selection of genes across the *Karlodinium veneficum* plastid genome (Table 5.1) (Gabrielsen *et al.*, 2011). These included five photosynthesis genes (*psbA, psbC, psbD, psaA, rbcL)* shown to contain poly(U) sites in *Karenia mikimotoi* (Fig. 5.1, lanes 1-5) (Dorrell and Howe, 2012a), as well as two plastid housekeeping genes (*rpl6, rps5*) that are not located on the plastid genomes of peridinin dinoflagellates (Fig. 5.1, lanes 6-7), and a 603 bp ORF located in a 1636 bp previously unannotated region between the *Karlodinium veneficum chlI* and *psbL* genes that shows no homology to any previously annotated nucleotide or protein sequence, henceforth termed *ORF1* (Fig. 5.1, lane 8) (Gabrielsen *et al.*, 2011).

Each gene tested through this strategy yielded high abundance products of between 400 and 900 bp. These products were directly sequenced using the gene-specific PCR forward primer

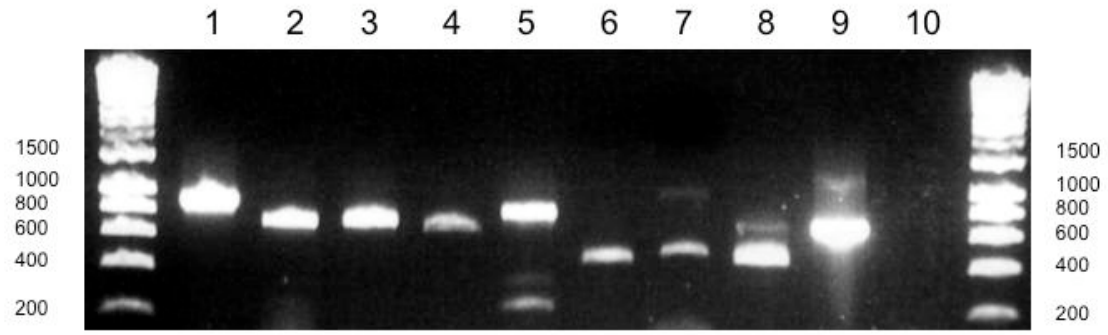


Fig. 5.1: Presence of poly(U) tails in *Karlodinium veneficum* plastid transcripts.

The gel photo shows the result of a series of representative oligo-d(A) RT-PCRs for specific transcripts from the *Karlodinium veneficum* plastid genome. **Lanes 1-5**: oligo-d(A) RT-PCRs of transcripts that have previously been shown to receive poly(U) tails in *Karenia mikimotoi* (*psbA, psbC, psbD, psaA, rbcL*). Lanes **6-7**: oligo-d(A) RT-PCRs of representative transcripts that have not previously been investigated in *Karenia mikimotoi* (*rpl6, rps5*). Lane **8**: oligo-d(A) RT-PCR of the previously unannotated *ORF1*. Lane **9**: RT-PCR of *Karlodinium veneficum psbA* using a cDNA template generated using an internal gene specific cDNA synthesis and PCR reverse primer, and the same *psbA* forward primer used in Lane 1. Lane **10**: PCR using the same primers as Lane 9, under template negative conditions. The faint secondary band at approximately 1000 bp in lane 6 corresponds to a dicistronic polyuridylylated *rpl6-rps5* transcript. The secondary bands visible in lanes 5, 8 and 9 are PCR chimeras.

as before, and found to correspond to transcripts containing poly(U) sequences, located within the 3' UTR of the gene concerned. A 900 bp band in the *rpl6* RT-PCR, which was substantially lower in abundance than the 500 bp monocistronic, polyuridylylated *rpl6* transcript, was found to correspond to a polyuridylylated dicistronic *rpl6-rps5* transcript, with the same poly(U) site as identified in the *rps5* RT-PCR (Fig. 5.1, lanes 6-7). Lower abundance products were identified in the *rbcL* RT-PCR (200 bp band) and in the ORF1 PCR (700 bp band), but these could not be confirmed to correspond to polyuridylylated plastid transcripts through sequencing (Fig. 5.1, lanes 5, 8). The poly(U) sequences identified through oligo-d(A) RT-PCR were not at positions that corresponded to poly(T) tracts in the *Karlodinium veneficum* plastid genome, and hence are post-transcriptional modifications to the transcript sequence.

Table 5.2. Primers for circular RT-PCR of *Karlodinium veneficum* transcripts.

Gene	cDNA synthesis primer	PCR reverse primer	PCR forward primer
psbA	TCATAGCTAGGTTACAGAGGGG	AACCTGCAACTGGCTCTC	GGGTGTGTCAACAATGGC
psbC	CATCTCCACCCCCTGG	CCACCAGGCAAAACC	GCAGCAGCAGGTTTTGAG

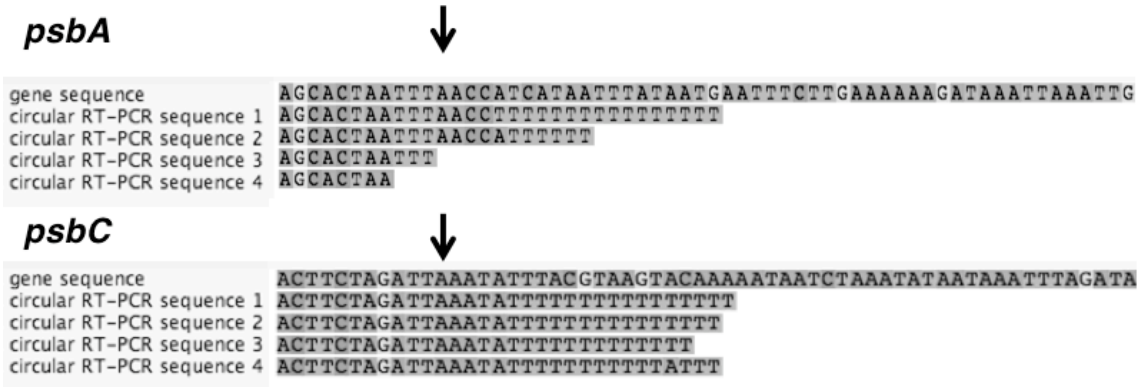


Fig. 5.2: Circular RT-PCRs of *Karlodinium veneficum psbA* and *psbC* transcripts.

This figure shows the 3' termini of *psbA* and *psbC* transcripts, as identified by circular RT-PCR, aligned with the corresponding genomic sequences. For each gene, the final 10 nt of coding sequence and the first 50 nt of the 3' UTR is shown. A vertical arrow corresponds to the TAA-STOP codon employed by each gene. For both genes, transcripts could be identified that clearly terminated at the 3' end in the 3' UTR, and possessed a poly(T) sequence that did not correspond to the underlying genomic sequence, confirming that transcripts in the *K. veneficum* plastid receive a post-transcriptional 3' poly(U) tail. All of the poly(U) tails identified were homopolymeric. Although two of the *psbA* transcripts identified did not possess a 3' poly(U) tail, these transcripts terminate within the CDS, upstream of the STOP codon, and are therefore likely to represent the degradation products of polyuridylylated transcripts.

To confirm that the oligo-d(A) RT-PCR products corresponded to 3' terminal transcript poly(U) tails, as opposed to internal sequence insertions, or to artifacts generated by mispriming of the oligo-d(A) primer, RT-PCRs were performed using circular RNA, and cDNA and PCR synthesis primers specific to the *Karlodinium veneficum psbA* and *psbC* genes (Table 5.2). This technique has previously been employed to confirm the presence of polyuridylylated *psbA* and *psbC* transcripts in *Karenia mikimotoi* (Fig. 5.2) (Dorrell and Howe,

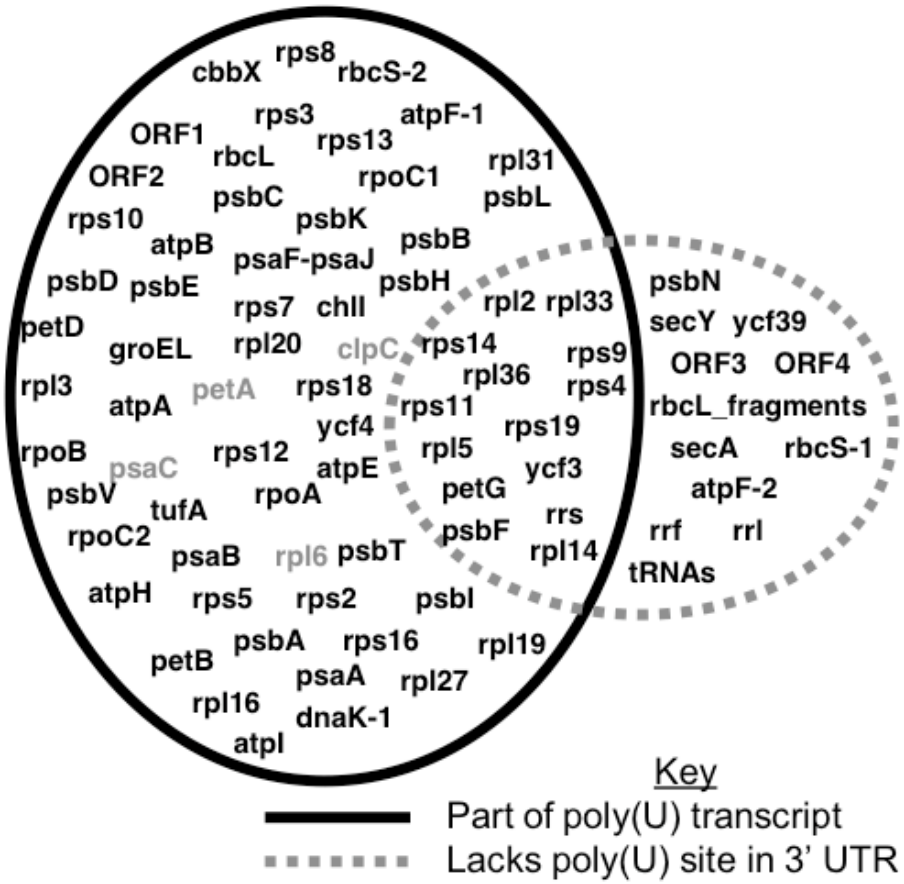


Fig. 5.3: Extent of transcript polyuridylylation across the *Karlodinium veneficum* plastid.

The Venn diagram shows the transcript polyuridylylation state of every gene within the *K. veneficum* plastid. Genes in the overlap sector between the two circles lack poly(U) sites in their respective 3' UTR sequences, but were retrieved as part of polyuridylylated polycistronic transcripts, with the poly(U) site positioned in the 3' UTR of a downstream gene. The poly(U) tails of genes shaded in grey may be generated by the transcription of genomic poly(T) tracts. These data were obtained with the assistance of an undergraduate student, Elisabeth Richardson.

2012a). Using this approach, transcripts of both genes were identified that possessed homopolymeric poly(U) tails on the 3' end (Fig. 5.2). Although non-polyuridylylated *psbA* transcripts were also identified by this approach, all of these transcripts terminated at the 3' end within the CDS, hence are likely to represent transcript degradation products, as opposed to mature transcripts generated by a poly(U)-independent pathway (Fig. 5.2). Thus, poly(U) tails are added to a wide variety of plastid transcripts in *Karlodinium veneficum*, as

Table 5.3: 3' UTR characteristics within the *Karlodinium veneficum* plastid

For genes that possess poly(U) sites, the first 10 bp of the 3' UTR, and the first 10 bp downstream of the poly(U) site are listed. For genes that lack poly(U) sites, the first 10 bp of the 3' UTR is shown, alongside whether polyuridylylated polycistronic transcripts were identified by oligo-d(A) primed RT-PCR. This table was assembled with the assistance of an undergraduate student, Elisabeth Richardson.

1. Poly(U) genes	UTR length	3' UTR	post-Poly(U)	Poly(U) length	Notes
atpA	9	AATCAAAAA	AGAATAAATT...	16	
atpB	8	AAAATTTA	AAATATTTGA...	16	
atpE	48	TATTTTATAA...	TTATTAATGA...	23	Previously unannotated gene
atpF-1	15	TAATAATAAT...	CGTAGATTTA...	20	
atpH	12	TGATATTAT...	ATGAATTACC...	19	
atpI	14	AATTTGTAAT...	AATATCATTA...	27	
cbbX	34	TTTATTATAG...	AAAAAAGGTT...	17	
chlI	138	TTGATGACTA...	AAGATTAATT...	20	
clpC	32	GATCGAATTA...	GCCCTTATCG...	7	Poly(U) site positioned within T16 tract
dnaK-1	77	ATAGTCTTAT...	GCCTCCATCG...	11	Poly(U) site positioned within T12 tract
groEL	3	TT	ACAAGAAATT...	22	
ORF1	50	TAATATTTTA...	ATTAATGATA...	20	
ORF2	23	ATGGATTTTT...	TTATAAATCT...	21	
petA	20	ATAATTTTTT...	GCCCTCATCG...	7	Poly(U) site positioned within T17 tract
petB	5	ATTAA	ATACAATAAT...	20	
petD	4	GTTT	AGAGATAAAA...	21	
psaA	91	TATATTTTAT...	GTGATTAAAA...	22	
psaB	10	TTAAGTTTAA	ATAATAATAT...	19	
psaC	11	ATCTTTTTTT...	ACGTAAAAAA...	12	Poly(U) site positioned within T8 tract
psaF-psaJ	19	ATTAAAAAAA...	TATAAGAAAA...	19	
psbA	4	CCAT	CATAATTTAT...	20	
psbB	9	ATTTATTT	AAAAGCATTT...	19	
psbC	7	ATATTTA	CGTAAGTACA...	20	
psbD	4	TTTA	AAAATAAATT...	21	
psbE	17	GTAAGAGTTA...	GTAAGAGTTA...	22	
psbH	24	TTTTTTGTAT...	AAAATCGCAT...	19	
psbI	5	TTTAT	CGACCAAATA...	17	
psbK	29	AAATATAATT...	CATTTGTTAT...	20	
psbL	31	AGAGATTTCA...	CCATAGGTTA...	16	
psbT	32	AATTTGGATT...	CAATGTACTT...	23	
psbV	6	TATTTT	ATTATATCTT...	21	
rbcL	3	TCT	TCTATGAAAA...	28	
rbcS-2	16	CAGTAATTGA...	ATAATTAGTT...	16	
rpl3	60	ATATTTTAAT...	GCCTGTTCGT...	19	Poly(U) site positioned 2bp into rpl2 CDS
rpl6	18	AAATATTAAA...	AAAAATGCTG...	19	Poly(U) site positioned within T7 tract
rpl16	74	GAAAAATATA...	AAATAGATTA...	22	
rpl19	91	AGTATTTTAT...	GTAAAGGTAG...	18	
rpl20	29	TACACTCGCT...	AACTTTTCCA...	23	
rpl27	69	TTCAAAACCT...	TCAATCAAAT...	20	
rpl31	16	TAATAATAAT...	ACCCACGATA...	18	
rpoA	2	AT	AGAAAATATA...	21	
rpoB	11	ATTATTGTAA...	GTATGGTATG...	26	Poly(U) site positioned 2bp into rpoC1 CDS
rpoC1	6	CATATA	ATATTAGAAT...	21	
rpoC2	9	TTTCTAAAA	AATATGATTA...	22	
rps2	93	ATAAAAATAA...	ACAAAAAGTA...	24	
rps3	33	AGTTTGGGT...	AATAAGATAT...	17	
rps5	4	TAAT	ATTCTTTAAA...	19	
rps7	6	ATTTAG	TATATTCATA...	20	
rps8	6	ATACTA	n/a	19	Poly(U) site positioned 48bp within rpl6 CDS
rps10	17	ATCTAAATAT...	ATATAATTTA...	17	Previously unannotated gene
rps12	3	TTT	TAATATTTAA...	21	
rps13	103	TTAAAAGAAT...	CTCGATATAC...	20	

Table 5.3 (continued)

1. Poly(U) genes	UTR length	3' UTR	post-Poly(U)	Poly(U) length	Notes
rps16	39	AATCGTATAA...	AATTCTCGTG...	20	
rps18	42	ATTTTTTTTT...	TGGAGATCAA...	24	
tufA	28	ATATAGGAAA...	GTAATAAAGT...	18	
ycf4	44	TTTTAAGATA...	TAAAATTTTT...	20	

2. Non-poly(U) genes	3' UTR length	3' UTR	Forms polycistronic poly(U) transcript?	Notes
atpF-2	n/a	n/a	NO	
ORF3	64	AGATGTAATC...	NO	
ORF4	>1000	AAATGAAATT...	NO	
petG	18	TAATAATAAT...	YES	Previously unannotated gene
psbF	63	CAATTGCAAT...	YES	
psbN	388	ATTGTATAGT...	NO	
rbcL fragment 1	n/a	n/a	NO	
rbcL fragment 2	n/a	n/a	NO	
rbcS-1	34	CATTAAACTG...	NO	
rpl14	23	AGTGTTAATT...	YES	
rpl2	n/a	n/a	YES	Possesses internal poly(U) site; no 3' UTR as rps19 CDS overlaps with 3' end
rpl5	n/a	n/a	YES	No 3' UTR as rps8 CDS overlaps with 3' end
rpl33	n/a	n/a	YES	No 3' UTR as rps18 CDS overlaps with 3' end
rpl36	1	T	YES	
rps4	30	GATAACTTAA...	YES	
rps9	82	TTTATTTTAT...	YES	
rps11	66	AAAAATTATT...	YES	
rps14	392	TATATTAAGG...	YES	
rps19	51	TATTTATATT...	YES	
rrf	122	TCAACTAAAA...	NO	
rrl	699	CCAAAAATTA...	NO	
rrs	299	TGTAAAAGAT...	YES	
secA	718	AAAAATTATA...	NO	
secY	717	AAAATTATATA..	NO	
ycf3	91	TTTAATGTAT...	YES	
ycf39	>1000	TCCAGTAATT...	NO	

with *Karenia mikimotoi*. This confirms that poly(U) tail addition, similarly to transcript editing, occurred in the common ancestor of extant fucoxanthin dinoflagellates (Jackson *et al.*, 2013).

Extent of poly(U) tail addition within the *Karlodinium veneficum* plastid

I wished to determine how widespread poly(U) tail addition was across the *Karlodinium veneficum* plastid transcriptome. To do this, oligo-d(A) RT-PCRs were performed for every annotated protein-coding and ribosomal RNA gene within the plastid genome, including previously unannotated *atpE, petG* and *rps10* genes (Fig. 5.3; Tables 5.1, 5.3). Oligo-d(A) RT-PCRs were also performed for fifteen predicted tRNA genes in the *K. veneficum* plastid genome, and three further predicted ORFs of more than 300 bp length that bear no sequence homology to any previously identified plastid gene (Table 5.1) (Gabrielsen *et al.*, 2011). 54 of the 75 protein-coding genes, and two of the four novel ORFs surveyed, were found to possess poly(U) sites in the associated 3' UTR (Fig. 5.3, Table 5.3). Four of the 56

poly(U) sites observed were positioned within genomic poly(T) tracts (Fig. 5.3; Table 5.3), so it is possible they have arisen through primer mis-annealing. However, the remaining 52 were not, and are likely to correspond to post-transcriptional modifications.

For several of the oligo-d(A) RT-PCRs, products were obtained consistent with the presence of polyuridylylated polycistronic transcripts. Some of the genes that were found to possess associated poly(U) sites in the 3' UTR, such as *rpl6*, also gave rise to polycistronic polyuridylylated transcripts (Fig. 5.1; Table 5.3). In addition, several of the genes tested by oligo-d(A) RT-PCR gave rise only to polycistronic polyuridylylated transcripts, indicating that they lack poly(U) sites in the adjacent 3' UTR. In total, 13 of the 21 genes that were found not to possess associated poly(U) sites in the 3' UTR were found to give rise to polycistronic polyuridylylated transcripts (Fig. 5.3; Table 5.3). Thus, 69 genes within the *Karlodinium veneficum* plastid were identified to give rise to a polyuridylylated transcript of some form, indicating that poly(U) tail addition is a widespread feature of plastid transcript processing.

A small number of the protein-coding genes and unannotated ORFs in the *K. veneficum* plastid failed to yield visible products in any oligo-d(A) RT-PCR attempted (Fig. 5.3, Table 5.2). In each case, products could not be detected even following a nested reamplification of the primary PCR product, with the same oligo-d(A) primer and a second gene specific primer positioned downstream of the first (Table 5.1). None of these genes was positioned directly upstream of a gene that possessed a poly(U) site and was in the same transcriptional orientation, suggesting that they are unlikely to give rise to polycistronic polyuridylylated transcripts (Fig. 5.4; Table 5.2). Similarly, products corresponding to monocistronic polyuridylylated could not be obtained in oligo-d(A) RT-PCRs using primers specific to any of the ribosomal RNA subunits or tRNA genes, although a tricistronic polyuridylylated *rrs-petG-atpF-1* transcript was identified (Fig. 5.2; Table 5.1). To confirm that the failure to obtain products for these genes was not due to the PCR primer performed, RT-PCRs were performed using cDNA synthesis primers internal to the CDS of each of the protein coding genes, and ribosomal RNA genes found not to give rise to polyuridylylated transcripts, and the same PCR forward primer as used for oligo-d(A) primed RT-PCR (Table 5.1). In each case, products could be obtained, indicating that the gene in question is likely to give rise only to non-polyuridylylated transcripts (Tables 5.1, 5.3).

Poly(U) sites are associated with alternative processing events

It has been shown in peridinin dinoflagellates that some poly(U) sites are positioned within the mature mRNA of the downstream gene (Barbrook *et al.*, 2012). These poly(U) sites are believed to be involved in alternative end cleavage events, in which the processing of a

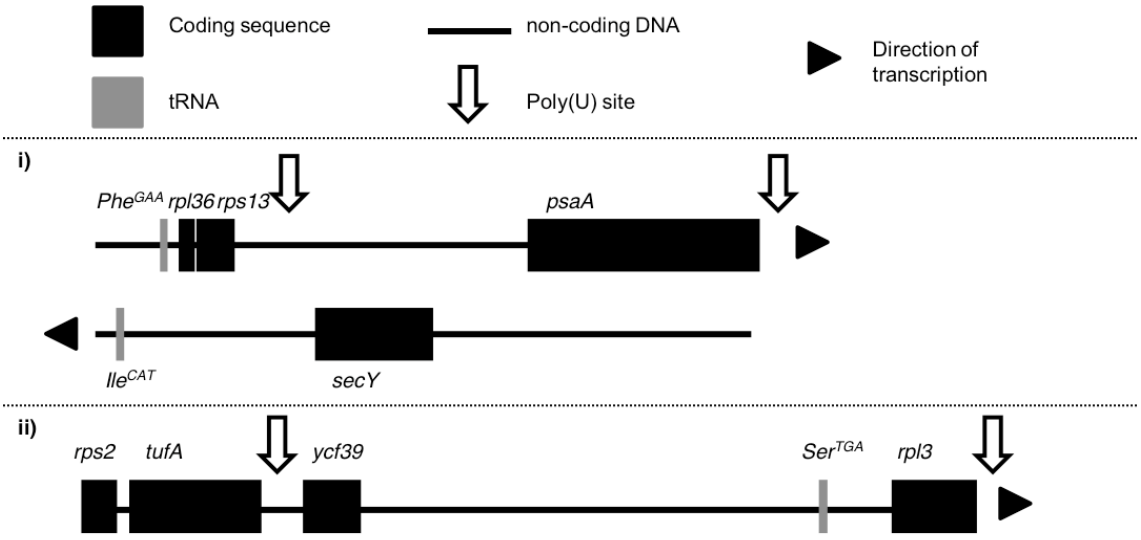


Fig. 5.4: Genomic contexts of non-polyuridylylated protein-coding genes in the *Karlodinium veneficum* plastid.

These diagrams show the order of genes surrounding *secY* **(i)** and *ycf39* **(ii)**, neither of which gives rise to polyuridylylated transcripts, in the *K. veneficum* plastid. The coding content of each strand is shown separately. *secY* is located between two genes (*rps13, psaA*) of opposing transcriptional orientation, and is located over 15kbp upstream of the *psaF-psbJ* fusion gene poly(U) site. *ycf39* is located immediately upstream of a large region of DNA with no annotated function, and the nearest poly(U) site (associated with *rpl3*) is nearly 6kbp downstream.

specific poly(U) site from a polycistronic precursor transcript prevents the processing of a translationally functional transcript of the gene located downstream. Similar events involving alternative transcript cleavage have been identified in plant plastids (Barkan *et al.*, 1994; Rock *et al.*, 1987).

Several of the genes in the *K. veneficum* plastid were found to possess an associated poly(U) site located within the CDS of a downstream gene (Table 5.3). Most dramatically, within the ten-gene ribosomal protein operon extending from *rpl3* through to *rps5*, four genes (*rpl3, rpl16, rps8, rpl6*) were identified to possess an overlapping poly(U) site, whereas only gene, *rps5*, was found to possess an associated poly(U) site in its 3' UTR (Fig. 5.5) (Gabrielsen *et al.*, 2011). Using a forward primer specific to *rpl2,* a further poly(U) site was detected 296bp within the *rpl2* CDS, although this poly(U) site could not be identified using a forward primer specific to the upstream *rpl3* gene (Fig. 5.5). Thus, the poly(U) tail addition

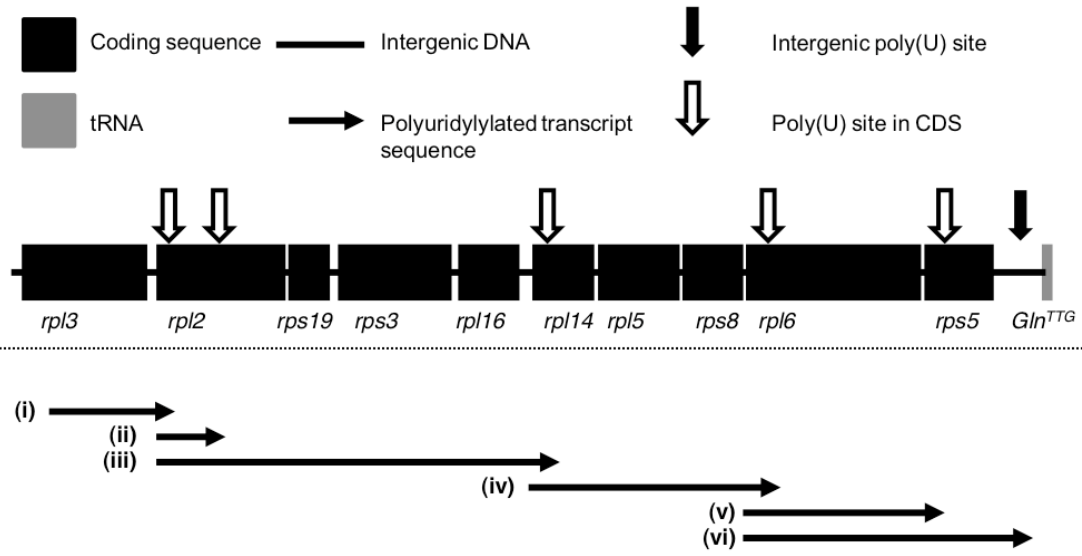


Fig. 5.5: Overlapping poly(U) sites within the *Karlodinium veneficum* ribosomal protein superoperon.

This diagram shows the array of 3' termini associated with polyuridylylated transcripts, identified by oligo-d(A) RT-PCR over ten genes extending from *rpl3* to *rps5* in the *K. veneficum* plastid genome. **(i)**, *rpl3* transcript with poly(U) site positioned in *rpl2*; **(ii)**, internally polyuridylylated *rpl2* transcript; **(iii)**, polycistronic polyuridylylated *rpl2-rps19-rps3-rpl16* transcript with poly(U) site positioned in *rpl14*; **(iv)**, polycistronic polyuridylylated *rpl14-rpl5-rps8* transcript with poly(U) site positioned in *rpl6*; **(v)**, polyuridylylated *rpl6* transcript with poly(U) site positioned in *rps5*; and **(vi)**, polycistronic polyuridylylated *rpl6-rps5* transcript. These data were obtained with the assistance of an undergraduate student, Elisabeth Richardson.

machinery in fucoxanthin plastid may play an important role in transcript processing, for example by directing alternative end cleavage of polycistronic transcripts.

Distribution of poly(U) sites within the *K. veneficum* plastid

I wished to determine whether there were any factors that determined why certain genes possessed a poly(U) site, and others did not. Other than the absence of poly(U) sites associated with ribosomal and transfer RNA genes, which suggests that poly(U) tail addition is associated with the processing of transcripts derived from protein coding genes, there were no clear trends underpinning which genes possessed poly(U) sites (Fig. 5.3, Tables 5.3). While there was a weak enrichment in the number of poly(U) sites associated with genes that encode products directly involved in photosynthesis reactions (photosynthetic

Table 5.4. Genome rearrangements associated with genes that possess unusual poly(U) sites in the *Karlodinium veneficum* plastid.

This table lists all of the genes that lack poly(U) sites, or possess overlapping poly(U) sites, in the *K. veneficum* plastid genome. The genes positioned downstream of each gene are listed for *K. veneficum*, for haptophyte plastid genome sequences (*Emiliania huxleyi, Phaeocystis globosa, Pavlova lutheri*, and the uncultured prymnesiophyte C19847), and for diatom plastid genome sequences (*Phaeodactylum tricornutum, Thalassiosira pseudonana*). "n/c" implies that there is not a consistent gene order between different members of the same group. Genes likely to have undergone a rearrangement specifically within the *K. veneficum* plastid genome are highlighted in bold. For these genes, a different downstream gene is present in *K. veneficum* to haptophyte genomes, there is no evidence for lineage-specific recombination events within the haptophytes, and the gene order found in haptophytes is also found in diatom plastids and is thus likely to represent the ancestral state. Although some of the genes that lack poly(U) sites or possess unusual poly(U) sites have undergone recent rearrangement events (e.g. *rpl3, secY*), many others in the *K. veneficum* plastid have not (e.g. *rpl14, rrl).*

		Downstream gene			
Gene	**Poly(U) site**	***K. veneficum***	**Haptophytes**	**Diatoms**	**Genome rearrangement?**
rpl5	overlapping	rps8	rps8	rps8	No
rpl3	**overlapping**	**rpl2**	**rpl23; rpl2**	**rpl4; rpl23; rpl2**	**Loss of rpl23**
rps8	overlapping	rpl6	rpl6	rpl6	No
rpoB	overlapping	rpoC1	rpoC1	rpoC1	No
rps9	**absent**	**rpl31**	**rpl31; rps12**	**rpl31; rps12**	**Loss of rpl31**
rrf	absent	rrs	n/c	psbY	Uncertain
rrl	absent	rrf	rrf	rrf	No
rrs	absent	rrl	n/c	rrl	Uncertain
secY	**absent**	**anti-rps13; rpl36**	**rpl36**	**rpl36**	**Insertion of antisense rps13**
ycf3	absent	psbT	psbD	atpB	Uncertain
ycf39	absent	trnS	n/c	ycf41; psbI	Uncertain
psbF	**absent**	**psbH**	**psbL**	**psbL**	**Recombination**
psbN	**absent**	**trnL**	**anti-psbT**	**anti-psbT**	**Recombination**
rpl14	absent	rpl5	rpl5	rpl24; rpl5	No
rpl33	absent	rps18	rps18	rps18	No

electron transfer and Calvin Cycle genes), over genes of non-photosynthesis function (encoding components of other biochemical pathways or the plastid housekeeping machinery), this was judged not to be statistically significant (chi-squared, P= 0.07).

The *K. veneficum* plastid genome has undergone extensive rearrangement events since its divergence from free living haptophytes (Gabrielsen *et al.*, 2011; Puerta *et al.*, 2005). I wished to determine whether these rearrangement events may have influenced the distribution of poly(U) sites in the *K. veneficum* plastid genome. To do this, the gene order of the *K. veneficum* plastid genome was compared to the plastid genomes of the free-living haptophyte species *Emiliania huxleyi, Phaeocystis globosa, Pavlova lutheri,* the uncultured prymnesiophyte C19847 (Table 5.4) (Baurain *et al.*, 2010; Cuvelier *et al.*, 2010; Puerta *et al.*, 2005). Plastid genomes of the diatom algae *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* were used as an evolutionary outgroup, to confirm the likely ancestral gene organisation state (Table 5.4) (Green, 2011; Oudot-le-Secq *et al.*, 2007). No consistent relationship could be observed between the distribution of poly(U) sites, and inferred recombination events in the *K. veneficum* plastid (Table 5.4). Overall, it appears that gene function and genome rearrangements do not underpin the distribution of poly(U) sites across the *K. veneficum* plastid.

The initial report of transcript poly(U) tail addition in peridinin dinoflagellates suggested that certain A/T rich motifs might be specifically associated with poly(U) sites in dinoflagellate 3' UTRs (Wang and Morse, 2006), although subsequent reports have not been able to detect similar motifs in other dinoflagellate species (Barbrook *et al.*, 2012; Nelson *et al.*, 2007). I wished to determine whether there might be specific sequences associated with poly(U) sites in the *K. veneficum* plastid.

To do this, the sequences of the 3' UTR of each gene shown to possess an associated poly(U) site were compared to each other to identify conserved primary sequence motifs, changes to GC and purine/ pyrimidine content, or predicted RNA secondary structures that were specifically associated with the presence of a poly(U) site (Table 5.3). A similar comparison was made with the sequences located in the first 100 bp downstream of each poly(U) site, to identify potential downstream poly(U)-associated motifs (Table 5.3). To confirm that any sequence features identified were not a general feature of 3' UTRs in the *K. veneficum* plastid, similar searches were made within the first 100 bp of the 3' UTR sequences associated with genes that do not possess poly(U) sites (Table 5.3).No sequence features were identified through this strategy that were significantly associated with the presence of a poly(U) site. It is therefore likely that instead of poly(U) tail addition being associated with common sequence motifs, poly(U) sites in fucoxanthin plastids are defined by motifs specific to individual genes.

Table 5.5. Editing events within the *Karlodinium veneficum* plastid transcriptome.

This table presents an overview of the editing observed in *K. veneficum* in comparison to those observed in previous studies of *K. veneficum* (Jackson *et al.*, 2013) and of the related fucoxanthin dinoflagellate *Karenia mikimotoi*. Detailed editing information for each transcript studied is given below the overview. This data was obtained with the assistance of an undergraduate student, Elisabeth Richardson.

Overview	Karenia	Karlodinium-Jackson	Karlodinium extended
Sequence length	5473	7373	36084
A-C	26	0	15
A-G	59	131	789
A-U	0	0	16
C-A	1	0	8
C-G	0	0	4
C-U	17	8	49
G-A	15	15	99
G-C	24	0	8
G-U	0	0	1
U-A	0	0	11
U-C	116	67	540
U-G	2	0	11
Total	260	221	1539
% bases edited	4.75	2.86	4.27
% transitions	79.6	100.0	96.0
% transversions	20.4	0.0	4.0
% GC-enrich	78.1	89.6	88.0
% GC-deplete	12.7	10.4	10.0
% GC-neutral	9.2	0.0	1.9
% non-synonymous	58.5	95	87.1
% synonymous	41.5	5	12.9

Gene	Sequence length (bp)		Editing observed												% edited		
1. Poly(U) genes	CDS	UTR	A-C	A-G	A-U	C-A	C-G	C-U	G-A	G-C	G-U	U-A	U-C	U-G	CDS	UTR	overall
atpA	1071	9		29				3	7	1		1	18		5.2	37.5	5.5
atpB	438	8		19				1	3				6		6.6	0.0	6.5
atpE	594	48		22				1	7				29		9.9	0.0	9.2
atpF-1	381	15	1	16		1		4	1				8		7.6	14.3	7.8
atpH	183	12		4									3		3.8	0.0	3.6
atpI	300	14		10									2		4.0	0.0	3.8
cbbX	670	34		12								2	9		3.4	0.0	3.3
chlI	678	138		3					1				9		1.8	0.7	1.6
clpC	637	32		9									7		2.5	0.0	2.4
dnaK-1	1303	77	1	31	2	2		7	6	1	1	1	17		4.8	7.9	5.0
groEL	384	3		10					1				8		4.9	0.0	4.9
ORF1	299	50		11					6				23		13.4	0.0	11.5
ORF2	512	23		8									5	1	2.7	0.0	2.6
petA	677	20		34				2	7	0			21		9.5	0.0	9.2
petB	427	5		16				2	5				13		8.4	0.0	8.3
petD	325	4		15					4				14		10.2	0.0	10.0
psaA	2265	91		67		1		1	3				22		4.0	4.4	4.0
psaB	372	10		11					2				7		5.4	0.0	5.2
psaC	108	11													0.0	0.0	0.0
psaF-psaJ	370	19		8				1					13		5.9	0.0	5.7
psbA	1041	4	1	1				2						1	0.5	0.0	0.5
psbB	858	9		16				1	1				4		2.6	0.0	2.5
psbC	599	7		16				1	2				10		4.8	0.0	4.8
psbD	293	4	6	8	4		2			3			12	6	14.0	0.0	13.8
psbE	204	17		7									3		4.9	0.0	4.5
psbH	297	24		11				1	1				4		5.7	0.0	5.3
psbI	111	5		7									4		9.9	0.0	9.5

Table 5.5 (continued)

Gene	Sequence length (bp) CDS	UTR	Editing observed A-C	A-G	A-U	C-A	C-G	C-U	G-A	G-C	G-U	U-A	U-C	U-G	% edited CDS	UTR	overall
psbK	50	29		1											2.0	0.0	1.3
psbL	192	31	1	6				1		1			4		6.8	0.0	5.8
psbT	126	32		4					3				2		7.1	0.0	5.7
psbV	372	6		7									3		2.7	0.0	2.6
rbcL	681	3		7				2					2		1.6	0.0	1.6
rbcS-2	317	16		7				2	1				5		4.7	0.0	4.5
rpl3	637	60		6					1				14		3.3	0.0	3.0
rpl6	361	18		6					1				6		3.6	0.0	3.4
rpl16	257	74		9									6		5.8	0.0	4.5
rpl19	221	91		6				1					3		4.5	0.0	3.2
rpl20	233	29		7									1		3.4	0.0	3.1
rpl27	233	69		3				1	2				2		3.0	1.5	2.6
rpl31	201	16	1	8									3		6.0	0.0	5.5
rpoA	665	2		9				1					12		3.3	0.0	3.3
rpoB	275	11		2									10		4.4	0.0	4.2
rpoC1	406	6		10				3					9		5.4	0.0	5.3
rpoC2	636	9		14				1	2				5		3.5	0.0	3.4
rps2	702	93		27	1				1				19		6.7	1.1	6.0
rps3	689	33		17				1	2				7		3.9	0.0	3.7
rps5	348	4		9						1			9		5.5	0.0	5.4
rps7	259	6		9									2		4.2	0.0	4.2
rps8	335	56		4					2				2		2.4	0.0	2.0
rps10	380	17		5					1				3		2.4	0.0	2.3
rps12	372	3		2					1			1			1.1	0.0	1.1
rps13	346	103	1	9					1			2	10		5.5	2.9	5.1
rps16	245	39		5									2		2.9	0.0	2.5
rps18	584	42	2	17	1			2	6	1		1	5	2	5.7	7.3	5.9
tufA	1314	28		39					1				12		4.0	0.0	3.9
ycf4	483	44		15					1			1	10		5.6	0.0	5.1

2. Non-poly(U) genes	CDS	UTR	A-C	A-G	A-U	C-A	C-G	C-U	G-A	G-C	G-U	U-A	U-C	U-G	CDS	UTR	overall
atpF-2	227	n.d.													0.0	x	0.0
ORF3	319	n.d.		1											0.3	x	0.3
ORF4	178	n.d.		1											0.6	x	0.6
petG	101	18		13				1					10		23.8	0.0	20.2
psbF	101	63		1											1.0	0.0	0.6
psbN	201	n.d.		1										1	1.0	x	1.0
rbcL fragment 1	626	n.d.						1							0.2	x	0.2
rbcL fragment 2	252	n.d.													0.0	x	0.0
rbcS-1	330	n.d.		1											0.3	x	0.3
rpl14	303	23		15					1						5.3	0.0	4.9
rpl2	193	n/a		8									1		4.7	0.0	4.7
rpl5	424	n/a		15									8		5.4	0.0	5.4
rpl33	210	n/a		1											0.5	0.0	0.5
rpl36	152	1		5				1					5		7.2	x	7.2
rps4	539	30		13					3				7		4.1	3.4	4.0
rps9	169	82		4									2		2.4	2.5	2.4
rps11	124	66		2					1				3		4.8	0.0	3.2
rps14	316	392	1	1									2		1.3	0.0	0.6
rps19	149	51		3									2		3.4	0.0	2.5
rrf	200	n.d.					1								0.5	x	0.5
rrl	639	n.d.							6			1	5		1.9	x	1.9
rrs	433	299							1						0.2	0.0	0.1
secA	824	n.d.		2								1	9		1.5	x	1.5
secY	753	n.d.		20				1	2				24		6.2	x	6.2
ycf3	482	91		3				3	1				7		2.9	0.0	2.4

Global trends in editing across the *Karlodinium veneficum* plastid transcriptome

A recent study by Jackson et al. has profiled editing events in the *Karlodinium veneficum* plastid, by comparing transcript and genomic sequences for regions of 14 different genes (Jackson *et al.*, 2013). Four different forms of editing were observed, all of which were transitions, consisting predominantly of A to G and U to C editing events, as well as small numbers of G to A and C to U conversions (Jackson *et al.*, 2013). This is in contrast to the situation in *Karenia mikimotoi*, in which transversion substitutions were also identified (Dorrell and Howe, 2012a).

I wished to determine what forms of editing occur in the *Karlodinium veneficum* plastid, and in particular whether transversion substitutions are present. To do this, the complete plastid transcript dataset sequenced in this study, which is more extensive than that obtained by Jackson et al. (Jackson *et al.*, 2013), was compared to the published *Karlodinium veneficum* plastid genome sequence (Table 5.5). Editing could be detected in the majority of transcript sequences, regardless of whether the underlying gene possessed an associated poly(U) site or not, indicating that it is a widespread feature of plastid transcript processing in *K. veneficum* (Table 5.5).

Approximately 4.3% of the sites studied across the *K. veneficum* plastid transcriptome were edited, slightly higher than previous estimates (Jackson *et al.*, 2013). For some genes, higher frequencies of editing were edited, extending to 14% of positions for the *Karlodinium veneficum psbD* gene, and 24% of residues in the highly divergent *petG* sequence (Table 5.5). Editing sites were situated predominantly within gene sequences, although a low level of editing (1.6%) was detected in polyuridylylated transcript 3'UTR sequences (Table 5.2), as previously seen in *Karenia mikimotoi* (Dorrell and Howe, 2012a). Most (88%) of the editing events lead to an increase in transcript GC-content, consistent with previous studies (Dorrell and Howe, 2012a; Jackson *et al.*, 2013) (Fig. 5.6). Although the majority (96%) of editing events observed were transition events, seven different transversion events were detected in the *Karlodinium veneficum* transcriptome (Fig. 5.8).

Many (87%) of the editing events in the *Karlodinium veneficum* plastid are predicted to have non-synonymous effects on the corresponding protein sequence (Table 5.5). Some of these editing events may be required for the correct function of the encoded protein. For example, eleven of the genes in the *Karlodinium veneficum* plastid contain premature in-frame termination codons, which would prevent the translation of the complete protein sequence. Correction of premature termination codons through editing has previously been reported for *Karlodinium veneficum rpoB, rps13, psaA* and *secY* transcripts, and *psaA* in *Karenia*

Table 5.6. Premature termination codons removed by editing of *Karlodinium veneficum* plastid transcripts.

Gene	genomic	transcript	edited to
atpA	TGA	CAA	glutamine
cbbX	TAA	CAA	glutamine
petA-1	TGA	CAA	glutamine
petA-2	TAG	CGG	arginine
psaA	TAG	TGG	tryptophan
rpl14	TAG	TGG	tryptophan
rpoB	TAG	CGG	arginine
rps10	TAG	CAG	glutamine
rps13	TAA	CAA	glutamine
rps19	TAG	CAG	glutamine
secY	TAA	TGG	tryptophan
ycf39	TAG	CAG	glutamine

mikimotoi (Dorrell and Howe, 2012a; Jackson *et al.*, 2013). All of the remaining premature termination codons in the *Karlodinium veneficum* genome were confirmed to be removed from the corresponding polyuridylylated transcript sequences by editing (Table 5.6). Consistent with previous reports, edited *Karlodinium veneficum* transcripts were additionally found to show an increase in sequence similarity, relative to the genomic sequence, to the corresponding sequences from the haptophytes *Emiliania huxleyi* and *Phaeocystis globosa* (Table 5.7) (Jackson *et al.*, 2013). Editing in the *Karlodinium veneficum* plastid therefore appears to reduce the effects of divergent mutations on plastid protein sequence.

Editing of sequences unique to the *Karlodinium veneficum* plastid

Not all of the non-synonymous editing events observed within the *Karlodinium veneficum* plastid have readily inferred effects on plastid protein function. While more than one in ten of the transcript codons sequenced in this study appear to have undergone a non-synonymous change due to editing, this leads to a net increase of only 1.6% in sequence conservation between the *K. veneficum* and haptophyte protein sequences (Table 5.7). The other editing events may have selectively neutral or disadvantageous effects, or affect sequences that are not found in free-living haptophytes. Many of the genes in the *K. veneficum* plastid genome contain sequence insertions, or fast-diverging regions that bear no homology to haptophyte sequences (Gabrielsen *et al.*, 2011). Editing events associated with these sequences might play important roles in permitting the function of plastid proteins without changing the overall identity of the sequence to those found in other plastid lineages. I wished to determine whether highly divergent regions of the *K. veneficum* plastid genome were edited more

Table 5.7. Effect of editing on sequence conservation between *K. veneficum* transcript and orthologous haptophyte sequences.

This table lists the proportion of residues in *K. veneficum* transcripts altered by editing, and the net effect editing on protein sequence similarity to orthologues from haptophyte species (*Emiliania huxleyi* and *Phaeocystis globosa*), as identified by BLASTx sequence alignments. "x" implies that the change in sequence identity could not be calculated due to poor alignment of the *K. veneficum* and haptophyte sequences. This data was obtained with the assistance of an undergraduate student, Elisabeth Richardson.

Gene	Sequence length aligned (bp)	% substitutions generated by transcript editing		% change in protein sequence identity with editing	
		nucleotide	amino acid	Emiliania	Phaeocystis
atpA	1070	5.33	14.02	2.98	2.98
atpB	384	7.55	17.97	1.68	1.60
atpE	594	10.27	28.28	3.57	4.00
atpF-1	153	7.19	21.57	x	x
atpH	141	1.42	4.26	-4.44	-4.35
atpI	300	4.00	11.00	2.04	1.02
cbbX	636	3.62	10.85	4.33	4.81
chlI	678	1.77	5.31	0.46	0.46
clpC	742	2.56	7.28	0.40	0.77
dnaK-1	1306	4.44	10.11	2.17	1.90
groEL	384	4.95	13.28	3.17	2.38
petA	681	9.25	22.47	3.39	1.69
petB	81	12.35	22.22	11.11	11.11
petD	246	10.98	25.61	6.33	6.33
psaA	1854	4.05	10.03	2.05	1.48
psaB	372	5.38	11.29	0.00	0.84
psaC	108	0.00	0.00	0.00	0.00
psaF-psaJ	369	5.96	16.26	0.00	3.08
psbA	783	0.38	0.77	0.78	0.78
psbB	1199	3.92	9.76	0.00	0.00
psbC	597	0.84	2.01	2.02	2.02
psbD	612	0.49	1.47	0.99	0.99
psbE	186	6.45	16.13	1.64	4.92
psbF	108	1.85	2.78	0.00	0.00
psbH	297	6.06	14.14	8.62	6.90
psbI	111	9.91	18.92	0.00	0.00
psbK	48	2.08	6.25	6.67	6.67
psbN	189	1.59	3.17	2.94	2.94
psbT	126	7.14	16.67	7.14	x
psbV	372	2.69	7.26	1.67	1.67
rbcL	678	1.92	4.87	0.00	0.00
rbcS-2	300	5.00	13.00	4.26	4.26
rpl14	285	6.32	13.68	2.06	2.06
rpl16	234	5.13	14.10	-1.37	1.30
rpl19	216	4.17	11.11	6.38	3.57
rpl2	161	3.11	9.32	-1.89	1.72
rpl20	196	3.57	10.71	3.57	2.27
rpl27	156	4.49	11.54	0.00	0.00
rpl3	636	3.30	9.43	0.69	1.43
rpl31	201	5.97	16.42	0.00	1.52
rpl33	210	0.48	1.43	1.79	1.79
rpl36	150	3.33	10.00	4.44	4.44
rpl5	249	6.43	18.07	4.62	4.55
rpl6	348	3.74	10.34	4.35	3.48
rpoA	663	3.17	8.14	-1.41	-1.54

Table 5.7 (continued)

Gene	Sequence length aligned (bp)	% substitutions generated by transcript editing		% change in protein sequence identity with editing	
		nucleotide	amino acid	Emiliania	Phaeocystis
rpoB	273	4.40	12.09	0.00	2.25
rpoC1	405	5.43	14.07	0.62	0.00
rpoC2	615	3.25	9.27	1.37	1.53
rps10	171	3.51	10.53	0.00	0.00
rps11	123	2.44	7.32	x	x
rps12	223	1.35	2.69	2.74	2.74
rps13	279	4.66	13.98	2.27	2.22
rps14	315	1.59	4.76	0.00	0.00
rps16	181	3.87	11.60	2.08	0.00
rps18	180	4.44	13.33	x	x
rps19	129	3.88	11.63	0.00	0.00
rps2	729	6.45	17.28	1.35	0.91
rps3	687	3.78	10.92	1.82	0.00
rps4	537	4.10	10.61	2.82	5.23
rps5	357	5.32	14.29	0.00	-0.99
rps7	237	3.38	8.86	x	0.00
rps8	333	2.40	7.21	3.60	3.57
rps9	168	2.38	5.36	0.00	0.00
secA	824	1.46	4.37	-3.57	-1.59
secY	441	6.12	15.65	5.13	3.03
tufA	920	4.67	12.39	0.97	-0.65
ycf3	480	2.92	8.75	2.78	2.78
ycf39	411	5.35	14.60	3.13	3.45
ycf4	390	5.64	13.85	2.46	4.10
TOTAL	**24750**	**4.02**	**10.51**	**1.69**	**1.64**

frequently relative to other sequences. The *K. veneficum psaA* and *tufA* genes were selected as models to explore editing site distribution. To test whether these genes contain highly edited regions, the entire polyuridylylated transcript sequence of each gene was identified, and the entire sequence of each gene was confirmed by sequencing PCR products generated using a gDNA template. To ensure that the sequence comparison was accurate, the transcript and gene sequences were each amplified and sequenced twice. Editing frequencies were calculated across each sequence using a sliding 60 bp window (Fig. 5.6, panel A). To test whether editing was preferentially associated with particularly divergent regions in each sequence, the predicted sequence conservation was calculated between the predicted *K. veneficum psaA* and *tufA* transcript translation products, and the corresponding *E. huxleyi* protein sequences (Fig. 5.6, panel A), over the same sliding window as before.

Both genes were found to contain regions of in which >15% of residues are edited, compared to an average editing rate across each gene of approximately 4% (Fig. 5.6, panel A; Table 5.5). In both genes, editing was negatively correlated to sequence conservation with *E. huxleyi* (Pearson correlation= -0.56 for *psaA,* -0.67 for *tufA*; P < E-07 for both genes). Over a

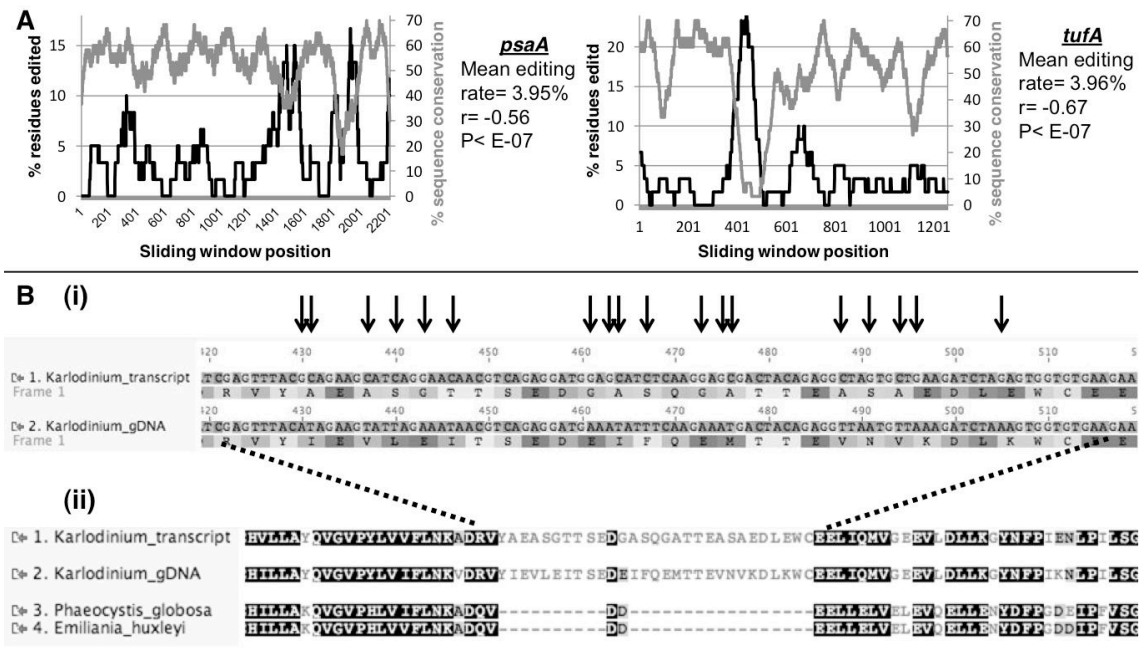


Fig. 5.6: Editing is preferentially associated with highly divergent regions of *Karlodinium veneficum* plastid genes.

Panel A compares the frequency of editing with sequence conservation in a 60 bp sliding window over the entire lengths of the *K. veneficum psaA* and *tufA* genes. The horizontal axis shows the starting position of each window within each gene sequence. The left hand vertical axis of each graph (black line) depicts the total percentage of nucleotide positions within each window that are edited within the transcript sequence. The right hand vertical axis (grey line) shows the proportion of residues within the predicted translation product of the transcript sequence of this window that are conserved with the predicted orthologous protein sequence in *Emiliania huxleyi*. A table to the right hand side of each graph shows the total proportion of sites that are edited over the entire CDS, the Pearson coefficient and the significance value of the correlation between sequence conservation and editing frequency.

Panel B shows nucleotide **(i)** and protein sequence alignments **(ii)** covering a highly edited 84 bp region of the *K. veneficum* plastid *tufA* gene. This region contains 18 editing sites, as inferred by the comparison of polyuridylylated transcript and genetic sequence (labelled with vertical arrows). This region encodes a 28 aa in-frame insertion, which is not found in haptophyte TufA sequences. The data in this figure was generated with the assistance of an undergraduate student, Elisabeth Richardson.

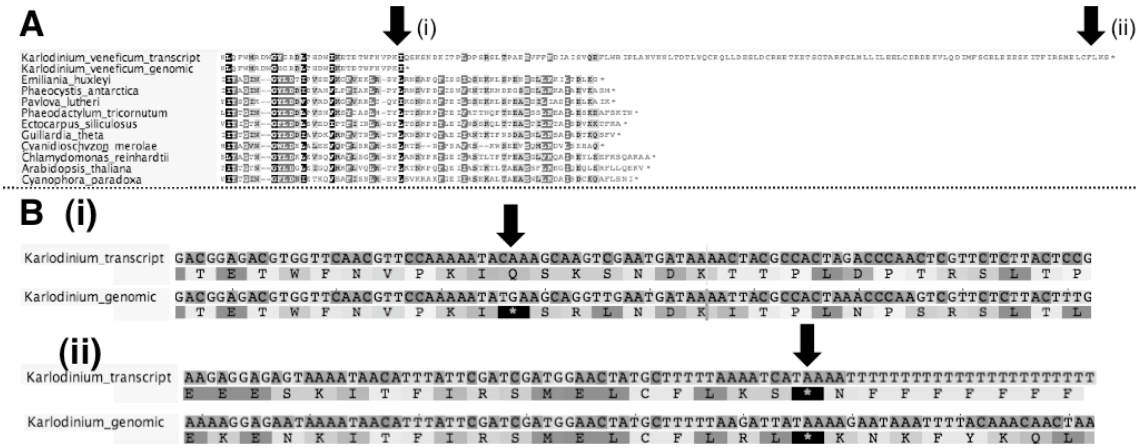


Fig. 5.7: Generation of a novel C-terminal sequence extension by editing of *Karlodinium veneficum atpA* transcripts.

Panel A shows an alignment of the predicted translation products of the genomic and transcript sequences of *K. veneficum atpA* with protein sequences from other plastid lineages. **Panel B** shows a nucleotide sequence alignment, and predicted translation products of two regions of the *K. veneficum* genomic and transcript sequence in detail. Residues important for defining the size of the predicted translation product of each *K. veneficum* sequence are labelled in each alignment as **(i, ii)**.

The *K. veneficum* gene contains an in-frame TGA STOP codon within the predicted CDS. This is altered by editing to a CAA-Gln codon **(i)** in the transcript, enabling the translation of the complete AtpA C-terminus. However, the *atpA* transcript sequence does not possess a termination codon at the same position as orthologous sequences. Instead, it encodes an 85aa C-terminal extension that is not conserved with other AtpA sequences, which terminates in an unedited TAA STOP codon **(ii)**. These data were obtained with the assistance of an undergraduate student, Elisabeth Richardson.

third of the editing events within *tufA* occur within an 84 bp region, which forms less than one-twelfth of the entire gene, and is significantly more highly edited than the rest of the sequence (chi-squared: $P< 0.05$). This region corresponds to an in-frame insertion that is not found in any species other than *K. veneficum* (Fig. 5.6, panel B). Thus, editing events are associated with regions of sequence that are recently acquired or are highly divergent. Editing might be involved in modifying the translation products of these sequences to facilitate protein function.

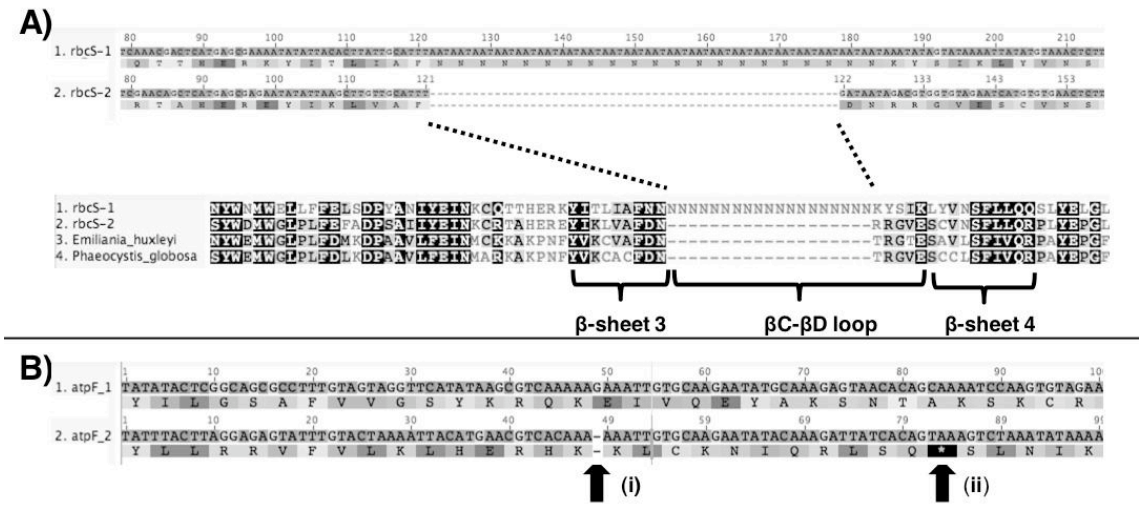


Fig. 5.8: Aligned protein and transcript sequences of paralogous copies of the *rbcS* and *atpF-2* sequences in the *Karlodinium veneficum* plastid.

Panel A shows the aligned sequences of *rbcS-1* and *rbcS-2*, as sequenced from RT-PCR products generated with a gene-specific internal cDNA primer. The *rbcS-1* transcript contains a 66 bp AAT repeat insertion, which would be translated in-frame as a poly(N) sequence. This sequence is not found in RbcS protein sequences from the representative free-living haptophytes *Emiliania huxleyi* and *Phaeocystis globosa*, and is predicted to be positioned immediately downstream of β-sheet 3, extending into the βC-βD loop domain in the *K. veneficum* RbcS-1 protein sequence. The expression product of *rbcS-2*, by contrast, aligns well with the haptophyte RbcS sequences.

Panel B shows the aligned sequences of *K. veneficum atpF-1* and *atpF-2*, as obtained from RT-PCR products as before. The *atpF-1* and *atpF-2* sequences are similar to each other, except in the presence of a single nucleotide deletion **(i)** in the *atpF-2* transcript. This causes a frame-shift that is predicted to lead to the translation of a premature termination codon located 36 bp downstream of the deletion site **(ii)**, preventing the expression of a full-length AtpF protein sequence from *atpF-2*.

Editing-facilitated divergent C-terminal evolution of *Karlodinium veneficum* AtpA

For the *Karlodinium veneficum atpA* gene, editing appears to be involved in the generation of a novel 3' extension on the conventional CDS (Fig. 5.7). The *K. veneficum atpA* gene contains a premature in-frame TGA codon, which is edited to form a CAA-glutamine codon in the mature transcript sequence. However, the *K. veneficum atpA* gene does not contain a STOP codon at the consensus 3' end position found in other plastid sequences. The

translation product of the *K. veneficum atpA* transcript is similar in sequence up to the final six amino acids in the *E. huxleyi* plastid AtpA protein, where it diverges to contain a 95aa C-terminal extension that bears no homology to any other known sequence (Fig. 5.7). The expression of this extension would be possible only from edited transcript sequences, and therefore transcript editing may have allowed divergent evolution to have occurred within the ATP synthase complex of the *K. veneficum* plastid.

Differential recognition of pseudogenes by the *Karlodinium veneficum* plastid transcript processing machinery

The application of editing to novel sequence insertions, such as those found in *tufA* and *atpA*, suggests that the editing machinery may recognise highly divergent regions of the *Karlodinium veneficum* genome. I wished to determine whether there were sequences within the *K. veneficum* genome that are highly divergent and specifically do not interact with either the editing or poly(U) tail addition machinery. Notably, several of the genes in the *K. veneficum* plastid are present in multiple copies, some of which appear to be functional, while others are likely to be pseudogenes (Gabrielsen *et al.*, 2011). For example, two copies of the *rbcS* gene are present: *rbcS-2*, which is likely to encode a functional protein, and *rbcS-1*, which contains an in-frame insertion within the region encoding the βC-βD loop domain of the RuBisCo small subunit, which if expressed would be likely to interfere with its function (Fig. 5.8, panel A) (Larson *et al.*, 1997; Li *et al.*, 2005; Tabita *et al.*, 2008). Similarly, two copies of the *atpF* gene are present: a previously annotated gene (*atpF-1*), and a previously unannotated pseudogene (*atpF-2*), positioned downstream of and in reverse orientation to *psbB,* which contains an internal frame-shift sequence deletion that would prevent the translation of the complete protein sequence (Fig. 5.8, panel B).

I wished to determine whether transcripts of the *rbcS-1* and *atpF-2* pseudogenes receive poly(U) tails and are edited. Polyuridylylated *rbcS-2* and *atpF-1* transcripts were detected by oligo-d(A) RT-PCR, using PCR forward primers specific to each sequence (Fig. 5.9, lanes 2, 5). However, polyuridylylated *rbcS-1* and *atpF-2* transcripts could not be detected through the same approach (Fig. 5.9, lanes 1, 6). RT-PCRs using cDNA synthesis primers specific to the *rbcS-1* and *atpF-2* gene sequences generated products, suggesting that transcripts of each gene that did not possess poly(U) tails (and thus were not detectable by oligo-d(A) RT-PCR) were present (Fig. 5.9, lanes 3-4, 7-8). The products of these RT-PCRs were sequenced, and were confirmed to contain the in-frame insertion (in *rbcS-1*) and the frame-shift deletion (in *atpF-2*) inferred from the published plastid genome sequence (Gabrielsen *et al.*, 2011). No editing sites were identified within the *atpF-2* transcript, and only one editing

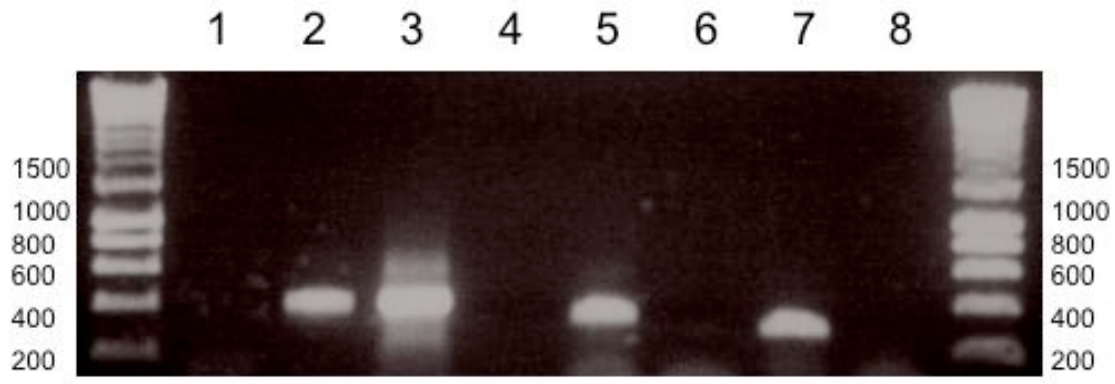


Fig. 5.9: Specific addition of poly(U) tails to transcripts of functional gene paralogues in the *Karlodinium veneficum* plastid.

This gel photo shows the result of a series of RT-PCRs to identify whether transcripts of functional and pseudogenic copies of the *rbcS* and *atpF* genes in the *K. veneficum* plastid receive poly(U) tails. **Lanes 1-2:** oligo-d(A) RT-PCR of *rbcS-1* (pseudogene) and *rbcS-2* (functional). Lanes **3-4:** RT-PCR of *rbcS-1* with a gene-specific internal cDNA synthesis primer under template positive (lane 3) and negative (lane 4) conditions. Lanes **5-6:** oligo-d(A) RT-PCR of *atpF-1* (functional) and *atpF-2* (pseudogene). Lanes **7-8:** RT-PCR of the *atpF-2* region with a gene-specific cDNA synthesis primer under template positive (lane 7) and negative (lane 8) conditions.

event was detected on the *rbcS-1* transcript, which is significantly fewer than the fifteen editing events observed over the same region of the *rbcS-2* transcript sequence (Table 5.5; binomial test, P< E-05). The absence of either poly(U) tail addition or editing from pseudogene transcripts indicates that both transcript processing events are preferentially associated with functional genes in the *K. veneficum* plastid.

Presence of minicircles in the *Karlodinium veneficum* plastid

Certain genes within the *Karlodinium veneficum* plastid genome, such as *rbcL* and *dnaK,* are enriched in sequencing libraries relative to others (Espelund *et al.*, 2012). These genes have been shown not only to be located on the chromosomal *K. veneficum* plastid genome sequence, but also on multiple small elements, containing fragments of individual genes, that do not assemble onto the plastid genome (Espelund *et al.*, 2012).The episomal elements

Table 5.8. Primers for amplification of episomal *Karlodinium veneficum rbcL* and *dnaK* sequences.

Oligo-d(A) primer	GGGACTAGTCTCGAGAAAAAAAAAAAAAAAAA	
1. Oligo-d(A) RT-PCR of rbcL	**PCR forward primer**	**cDNA synthesis primer**
chromosomal rbcL	ACTGGGGCAACCATGG	
rbcL_fragment-1	GGGGCATAGGGAAATGG	CCTAGCTTTTTCCGTGAAAG
rbcL_fragment-2	CTATAATGAATCTCGACCCATTT	GAAGATGGTACCCGTGC
2. Oligo-d(A) RT-PCR of dnaK	**PCR forward primer**	
1	GCAGGGAAAATTGCAGG	
2	CGGAGATACACAGTTAGGTGG	
3	CAAGGATAAAGGATGCTGC	
4	GCTGCAGATAATCAACCTG	
3. TAiL-PCR of dnaK	**Arbitrary degenerate primer**	**Gene specific primers**
1	TTNTCGASTWTSGWGTT	CGGAGATACACAGTTAGGTGG
2	TTWGTGNAGWANCANAGA	CAAGGATAAAGGATGCTGC
3	CCTTNTWGAWTWTWGWWTT	GCAAGGGGAACGAGAG
4	CCTTWGTGNAWWANCANAWA	
5	GGAACWACNTWTWNGTNTTW	
6	TTACWACANGWWGNTGNTWT	
7	GGAANACTWAWAWCWWAWA	
8	TTAANCWAGWCWCWAWWAA	
4. Circular RT-PCR of dnaK		
cDNA primer	GTGATCTCCGGAAGTTGC	

have been suggested to correspond to a population of plastid-located minicircles, which have arisen independently of those found in peridinin dinoflagellates (Espelund *et al.*, 2012; Zhang *et al.*, 1999). However, it is not known whether these episomal elements are located in the *K. veneficum* plastid, nor has a complete episomal element yet been sequenced and confirmed to form a minicircle.

I wished to determine whether episomal fragments in *K. veneficum* may give rise to polyuridylylated transcripts. Poly(U) tail addition has not been identified in dinoflagellate nuclei or mitochondria, and would accordingly confirm localisation of the elements to the *K. veneficum* plastid (Dorrell and Howe, 2012a). Initially, primers were designed that were specific to two episomal copies of *rbcL*, and oligo-d(A) RT-PCRs were performed for each gene copy (Table 5.8). Neither of the episomal *rbcL* genes was found to give rise to polyuridylylated transcripts (Fig. 5.10). Non-polyuridylylated transcripts of each gene were identified via gene-specific RT-PCRs, as previously described, but only one putative editing site was identified across 878 bp episomal transcript sequence (Table 5.5). In contrast, transcripts of the chromosomal *rbcL* gene receive poly(U) tails, and are highly edited (Fig.

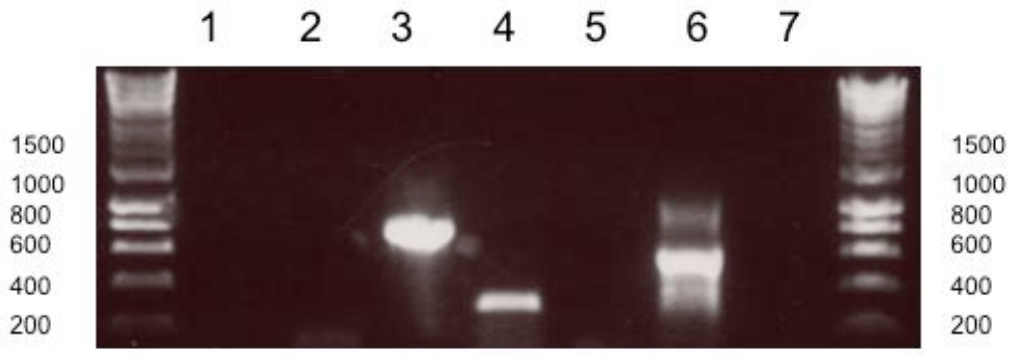


Fig. 5.10: Absence of polyuridylylated transcripts from episomal fragment copies of the *Karlodinium veneficum rbcL* gene.

This gel photo shows the result of a series of RT-PCRs to detect poly(U) tails on copies of *rbcL* located either on the chromosomal plastid genome, or on separate episomal elements. **Lane 1-2**: oligo-d(A) RT-PCR using forward primers specific to two episomal *rbcL* sequences (*rbcL* fragments 1, 2; GBID: 185572.1, 185573.1), demonstrating the absence of polyuridylylated transcripts of either gene. Lane **3**: oligo-d(A) RT-PCR using a forward primer specific to the complete *rbcL* gene located on the chromosomal plastid genome, confirming that this gene gives rise to polyuridylylated transcripts. Lanes **4-5**: RT-PCR of *rbcL* fragment 1 using a gene-specific cDNA synthesis primer, under template positive and negative conditions, and lanes **6-7**: RT-PCR of *rbcL* fragment 2 using a gene-specific cDNA synthesis primer, under template positive and negative conditions, confirming the presence of non-polyuridylylated transcripts of each episomal fragment.

5.10; Table 5.5). Thus, poly(U) tail addition and editing are preferentially associated with transcripts of the chromosomal *rbcL* gene.

I additionally investigated the origin and identity of polyuridylylated *dnaK* transcripts in *K. veneficum*. Whereas there is a complete copy of the *rbcL* gene within the *K. veneficum* plastid genome, the chromosomal *dnaK* genes lack consensus terminal regions, and contain frame-shift mutations, suggesting that they do not give rise to translationally functional *dnaK* transcripts (Fig. 5.11, panel A) (Espelund *et al.*, 2012; Gabrielsen *et al.*, 2011). Polyuridylylated transcripts could not be identified for either chromosomal *dnaK* gene. Instead, each *dnaK* primer amplified the same polyuridylylated transcript, which will henceforth be termed *dnaK-1* (Fig. 5.11, panel B; Table 5.8). The *dnaK-1* transcript encodes

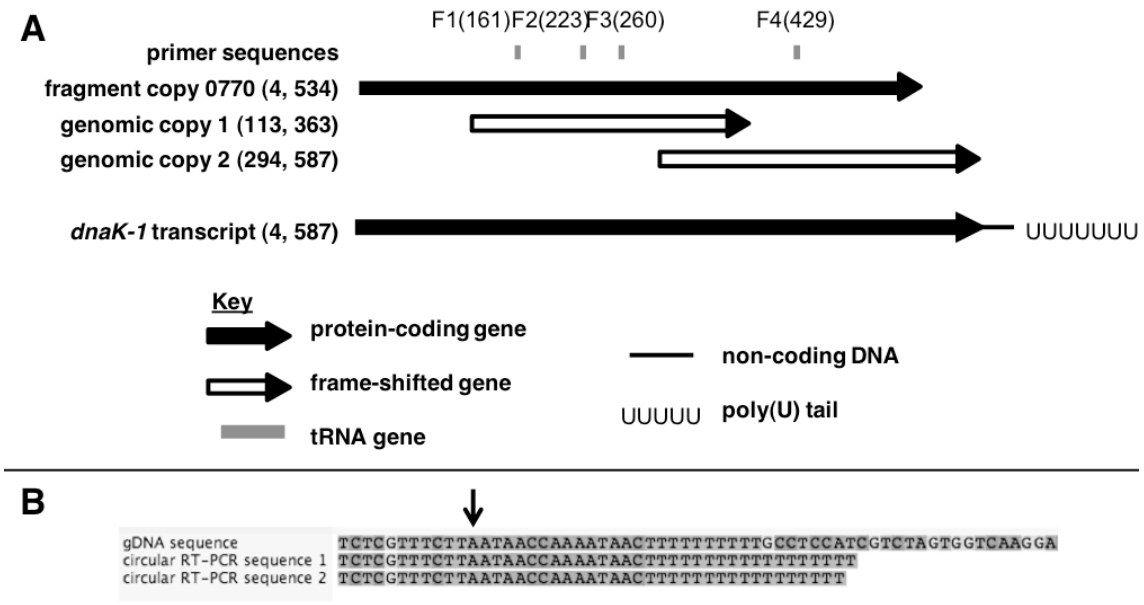


Fig. 5.11: Alignments of *Karlodinium veneficum dnaK* sequences.

Panel A shows a schematic diagram of the two copies of *dnaK* located on the *K. veneficum* plastid genome, the extensive copy of *dnaK* located on episomal fragment 0770, the four primers generated from this sequence to perform oligo-d(A) RT-PCR, and the complete *dnaK-1* CDS. Numbers in parentheses after each gene name correspond to the beginning and end of the regions of the *E. huxleyi* plastid Hsp70 protein sequence to which each gene product is homologous. Numbers in parentheses correspond to the equivalent position of the primer on the *E. huxleyi* protein sequence.

Panel B shows an alignment of the 3' termini of *dnaK-1* transcripts as obtained by circular RT-PCR with the underlying genomic sequence, displayed as in Fig. 5.2. Although the *dnaK-1* poly(U) site corresponds to a poly(T) tract in the genomic sequence, this is shorter (12 nucleotides) than the poly(U) tails identified through circular RT-PCR (18-19 nucleotides). This demonstrates that the *dnaK-1* transcripts are subject to post-transcriptional 3' poly(U) tail addition, confirming their localisation to the *K. veneficum* plastid.

a complete plastid Hsp70, and does not contain any frame-shifts or align with either chromosomal *dnaK* gene, suggesting that it is expressed from an episomal element.

To identify what genetic elements might give rise to the *dnaK-1* transcript, thermal asymmetric interlaced PCR (TAiL-PCR) was performed (Liu *et al.*, 1995), using combinations of primers derived from the *dnaK-1* transcript sequence (Table 5.8). This was supported with

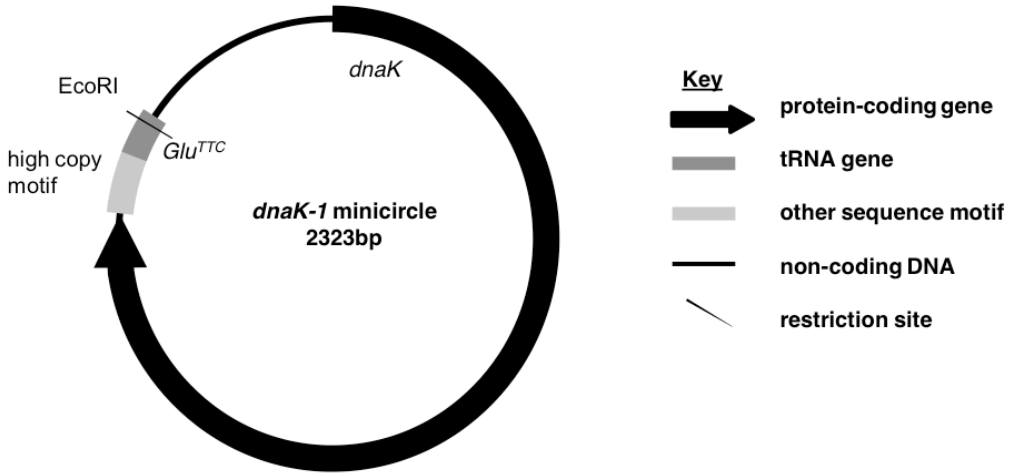


Fig. 5.12: Schematic diagram of the *Karlodinium veneficum dnaK-1* minicircle.

The 2323 bp *dnaK-1* minicircle contains a complete *dnaK-1* positioned directly upstream of the predicted "high copy element", and a Glu^{TCC} tRNA gene in the same transcriptional orientation. A single EcoRI restriction site is present on the template strand of the tRNA gene.

circular RT-PCRs using primers specific to *dnaK-1*, to confirm the full length of the *dnaK-1* transcript sequence (Table 5.8). A single genetic element was identified through TAiL-PCR that covered the entire *dnaK-1* CDS, and extended past the poly(U) site in the 3' UTR. The *dnaK-1* poly(U) site coincides with a genomic T12 motif. However, *dnaK-1* transcripts were identified through circular RT-PCR with poly(U) tails of up to 19 nt length, implying that they are generated through post-transcriptional sequence modification (Fig. 5.11, panel B). In addition, the *dnaK-1* transcript sequence contained extensive evidence of editing, as inferred by comparison to the underlying genetic sequence (Table 5.5). Thus, *dnaK-1* is transcribed from a single contiguous genetic element, located within the *Karlodinium veneficum* plastid, but separate from the chromosomal genome sequence.

Surprisingly, the *dnaK-1* 3' UTR sequence obtained by TAiL-PCR was found to be positioned immediately upstream of a sequence identical to the 5' end of the *dnaK-1* gene. This is consistent with the *dnaK-1* gene being located on a plastid minicircle (Fig. 5.12). The *dnaK-1* minicircle is 2323 bp long, and contains a single *Eco*RI restriction site, which is consistent with a 2.3 kbp band containing the *dnaK* gene identified through Southern blotting of *Eco*RI-digested *K. veneficum* gDNA (Fig. 5.12) (Espelund *et al.*, 2012). In addition to a complete *dnaK* gene, this minicircle contains a Glu^{TTC} tRNA gene, and a single "high copy" region that

is conserved with other episomal sequences previously identified from *K. veneficum* (Fig. 13) (Espelund *et al.*, 2012). This is the first complete plastid minicircle identified in a fucoxanthin dinoflagellate, confirming that the fucoxanthin plastid genome has undergone a similar fragmentation to that observed in peridinin dinoflagellates. Furthermore, it appears that in the case of *dnaK,* the poly(U) and editing machinery may recognise transcripts of genes located on minicircles (such as *dnaK-1*) over genes located on the chromosomal plastid genome.

Discussion

I have characterised the distribution of editing and poly(U) tail addition sites across the plastid genome of the fucoxanthin dinoflagellate *Karlodinium veneficum*. This represents the first genome-wide study of transcript processing in an algal plastid lineage. I have demonstrated that poly(U) tails are added to plastid transcripts in *Karlodinium*, as in *Karenia mikimotoi* (Figs. 5.1-5). I also found extensive sequence editing events, including transversion substitutions that have not previously been detected in *Karlodinium veneficum* but do occur in *Karenia mikimotoi* (Table 5.5) (Dorrell and Howe, 2012a; Jackson *et al.*, 2013). Notably, these transversion substitutions were identified in transcripts of *psaA* and *dnaK-1*, for which we generated multiple transcript sequences, and directly sequenced the underlying genetic elements, and thus are unlikely to be artifacts generated by sequencing errors within individual oligo-d(A) RT-PCRs, or sequence errors within the original *K. veneficum* plastid genome assembly (Table 5.5) (Jackson *et al.*, 2013). As *Karlodinium* and *Karenia* are distantly related within the fucoxanthin dinoflagellates, these transcript processing events are likely to have occurred in the common ancestor of all fucoxanthin plastid lineages (Bergholtz *et al.*, 2006; Gabrielsen *et al.*, 2011)

The distribution of poly(U) tail addition and editing sites in *Karlodinium veneficum* mirrors what has previously been documented in peridinin dinoflagellates. Poly(U) tail addition in the *K. veneficum* plastid is associated principally with mRNAs, and transcripts of ORFs of unannotated function, whereas ribosomal and transfer RNA genes do not possess poly(U) sites (Figs. 5.1, 5.3). Both ribosomal and mRNA sequences are edited (Table 5.5). In peridinin dinoflagellates, poly(U) tails are added to a wide variety of mRNA and ORF transcripts, but are not added to transfer RNAs, and have been inferred not to be added to certain ribosomal RNAs (e.g. 23S rRNA in *Lingulodinium polyedrum*) (Barbrook *et al.* , 2012; Nelson *et al.*, 2007; Wang and Morse, 2006). Likewise, in some peridinin plastids, both ribosomal and mRNAs have been shown to be edited (Dang and Green, 2009; Zauner *et al.*, 2004). However, poly(U) tail addition and editing are associated with a greater diversity of transcripts in *K. veneficum* than in peridinin dinoflagellate plastids. Transcripts of many genes that are not retained in peridinin plastids (e.g. protein-coding genes that do not encode

components of the plastid photosynthesis machinery) receive poly(U) tails, and are edited (Bachvaroff *et al.*, 2004; Howe *et al.*, 2008b). Thus, the poly(U) tail addition and editing machinery have been adapted to the greater coding content of the fucoxanthin plastid genome (Gabrielsen *et al.,* 2011).

I have additionally identified evidence that poly(U) tail addition and editing have coevolved with the *K. veneficum* plastid genome. In certain cases, poly(U) tail addition and editing may constrain the effects of highly divergent sequences on plastid physiology. For example, poly(U) tails and editing are not associated with transcripts of pseudogenes such as *rbcS-1* and *atpF-2* (Figs. 5.9, 5.10). The differential processing of pseudogene transcripts has not previously been reported in peridinin dinoflagellates, in which at least some pseudogene transcripts are extensively edited (Iida *et al.*, 2009). Poly(U) tail addition and editing might have a role in discriminating functional genes from non-functional gene fragments generated by recent rearrangements in fucoxanthin dinoflagellate plastid genomes. Similarly, the association of editing sites with fast-evolving sequences, such as the in-frame insertion in *tufA*, has not been described in other dinoflagellates (Fig. 5.6). This contrasts with plastid editing in plants, which is predominantly associated with slowly-evolving sites within the genome sequence (Fujii and Small, 2011; Hayes *et al.*, 2012). Editing and poly(U) tail addition might therefore neutralise the effects of fast-diverging sequences and recently acquired insertions on protein function in fucoxanthin plastids.

I have additionally reported the presence of a novel 3' sequence extension to the *K. veneficum atpA* CDS transcripts, which is formed as a result of editing (Fig. 5.7). The extension of a transcript sequence through editing into non-conserved sequence has not previously been reported in any plastid lineage. It is possible that the extension sequence associated with *K. veneficum atpA* gene evolved first (i.e. by the loss of the consensus termination codon on an otherwise conventionally organised gene), before the evolution of the premature termination codon and the associated editing events. Alternatively, the premature termination codon might have evolved first, leading to relaxed selection pressure and loss of the downstream consensus termination codon. The subsequent application of editing to this system would have caused the translation of a novel region of sequence. Thus, sequence editing may have indirectly facilitated divergent sequence evolution in fucoxanthin dinoflagellate plastids.

Finally, I have identified one plastid gene- *dnaK-1*- that is located on an episomal minicircle, (Fig. 5.12). This represents the first complete plastid minicircle sequence from a fucoxanthin dinoflagellate, and suggests convergent evolution in the organisation of the fucoxanthin and peridinin plastid genomes (Espelund *et al.*, 2012; Zhang *et al.*, 1999). I have additionally

shown that the *dnaK-1* minicircle gives rise to polyuridylylated and edited transcripts (Fig. 5.11). Thus, the poly(U) tail addition and editing machinery have adapted to the fragmentation of the *Karlodinium veneficum* plastid genome. Overall, it appears that poly(U) tail addition and editing in *K. veneficum* has evolved dynamically alongside the underlying genome, reducing the effects of mutations on plastid function, and adapting to- and potentially enabling the evolution- of divergently organised sequences. Further studies of dinoflagellates that have undergone serial endosymbiosis may provide important insights into the coevolution of plastid genomes and gene expression pathways.

Chapter Six- Poly(U) tail addition plays a central role in plastid transcript processing in the dinoflagellate alga *Karenia mikimotoi*

Introduction

Plastid gene expression involves complex transcript processing events. In plants, these include terminal end cleavage, *cis*- and *trans*-splicing, 3' poly(A) tail addition, and substitutional base editing (Barkan, 2011). Transcripts that do not have coding functions, such as antisense transcripts, which are generated by transcription of the non-template strands of plastid genes, may be degraded by terminal nucleases (Sharwood *et al.*, 2011; Stern *et al.*, 2010). The functional consequences of different plastid transcript processing events in plants have been explored, and some have been shown to be functionally connected to one another. For example, poly(A) tail addition allows specific plastid transcripts to be degraded from the 3' end (Kudla *et al.*, 1996; Yehudai-Resheff *et al.*, 2001). Similarly, sequence editing may be required for the generation of specific recognition motifs for the plastid splicing machinery (Asano *et al.*, 2013; Georg *et al.*, 2010).

Less is known about the roles of plastid transcript processing events that occur in lineages other than plants. One such event is the addition of a 3' poly(U) tail to plastid transcripts in peridinin dinoflagellates, and other alveolate lineages (Dorrell and Howe, 2012; Janouškovec *et al.*, 2010; Wang and Morse, 2006). Previous studies of transcript processing in peridinin dinoflagellates have suggested that poly(U) tail addition is associated with the maturation of plastid transcripts. Poly(U) tails have been inferred in these studies to protect transcripts from 3' end degradation (Barbrook *et al.*, 2012), and to enable other transcript processing events, such as the substitutional editing of transcript sequences (Dang and Green, 2009). I have additionally demonstrated that poly(U) tail addition is specifically associated with sense transcripts over antisense transcripts of peridinin dinoflagellate minicircles.

Previously, I have shown that the plastids in fucoxanthin dinoflagellate species such as *Karenia mikimotoi* and *Karlodinium veneficum*, which arose through the serial endosymbiotic replacement of the ancestral peridinin plastid lineage, possess the same poly(U) tail addition pathway as the peridinin plastid (Dorrell and Howe, 2012a; Richardson *et al.*, 2014). I have additionally shown that the plastid genome of the fucoxanthin dinoflagellate *Karlodinium veneficum,* which is known to retain many more genes than are found in peridinin plastid lineages (Gabrielsen *et al.,* 2011), gives rise to a wide range of polyuridylylated transcripts, even including transcripts of genes that are not retained in the peridinin plastid (Richardson *et al.*, 2014). The diversity of polyuridylylated transcripts produced in fucoxanthin

dinoflagellates renders them a valuable system in which to investigate the roles of poly(U) tail addition in plastid transcript processing.

This project was conceived to investigate the significance of poly(U) tail addition for transcript processing in the fucoxanthin plastid lineage, using the species *Karenia mikimotoi* as a model system. I firstly wished to identify the range of transcripts that constitute the *Karenia mikimotoi* plastid transcriptome. I identified polyuridylylated transcripts of plastid origin in *Karenia mikimotoi* via a novel next generation sequencing protocol, using a double-stranded cDNA template generated with an oligo-d(A) synthesis primer. I demonstrate that the *Karenia mikimotoi* plastid gives rise to a diverse range of polyuridylylated transcripts, many of which are also found in *Karlodinium veneficum,* despite highly divergent evolution in fucoxanthin plastid genome content and organisation. I additionally wished to determine whether poly(U) tail addition is associated with other processing events in *Karenia mikimotoi*, as has been suggested for the ancestral peridinin dinoflagellate plastid. To do this, I sequenced multiple transcript processing intermediates from the *K. mikimotoi rpl36-rps13-rps11* and *psbD-Met^{CAT}-ycf4* loci. I demonstrate that poly(U) tail addition is associated with multiple transcript processing events in fucoxanthin plastids, including the completion of editing, and alternative end cleavage events. Finally, I wished to determine whether antisense transcripts were present in the *K. mikimotoi* plastid, and identify whether these antisense transcripts receive poly(U) tails or are edited. I demonstrate that antisense plastid transcripts are present in *K. mikimotoi*, and these transcripts generally do not receive poly(U) tails, similar to in peridinin dinoflagellate plastids. Surprisingly, the antisense transcripts in *K. mikimotoi* show complementary editing patterns to the corresponding sense transcripts. Editing of plastid antisense transcripts has not previously been reported, and may indicate a previously unidentified role for antisense transcripts in plastid transcript processing. Overall, my data suggest that the poly(U) tail may have a central role in plastid transcript processing in fucoxanthin dinoflagellates. Poly(U) tail addition may function both directly, by enabling other processing events to occur, and indirectly, by regulating other transcripts (e.g. antisense transcripts) that play important roles in plastid transcript processing.

Results

Oligo-d(A) cDNA sequencing reveals the plastid transcriptome of *Karenia mikimotoi*

I wished to characterise the diversity of transcripts produced in plastids of the fucoxanthin dinoflagellate *Karenia mikimotoi.* To do this, double stranded cDNA was generated from *K. mikimotoi* total cellular RNA. An oligo-d(A) primer, previously shown to anneal to poly(U) tails

Table 6.1. Primers used for oligo-d(A) RT-PCR of *K. mikimotoi* plastid NGS contigs

oligo-d(A) GGGACTAGTCTCGAGAAAAAAAAAAAAAAAAA

Gene	PCR forward primer	Gene	PCR forward primer
atpA	GTGATTTAAGTGCCTTTTTACCC	**rbcS**	CTCATGATCCGCACCTTC
atpB	GGATTAATGGGATATCAACCG	**rpl16**	AAAATGAGTGGAAATTTATTCATG
atpH	GGGACATTATTACTTTCATTAGCC	**rpl2**	GTCCGACTGGTATTTAAAAGTTG
atpI	GGGCCAACTACCTAATTCG	**rpl22**	CAAACCCCGTCTCAAAGG
cbbX	CATACGGCACCCAGAAC	**rpl23**	GTTTCGCGGTAGCGTAG
chlI	GACGCAGTAAAAACCGC	**rpl3**	CGTCCCGCAGATAGACATC
clpC	CACCACCGGGTTATGTTG	**rpl31**	GATGGGCAACTTATTTTGAAAG
dnaK	TCTCGGCACGACAAATAG	**rpl36**	CTTTTAGGATAAAATATCAAGGTTACAAC
groEL	GCGACCGAGACTGAGATG	**rpl5**	TTCAGACGTGCCTTTCC
petA	GACCCATCAGACGAGCAAC	**rpl6**	CCTACCGGAATAGAAGTGACTG
petB	TTGTCAGGGCTGGATG	**rpoA**	TGCTAGAACAACTCCCGC
petD	TGTCTACGGCTTTCCTGC	**rpoB**	TGGTCGTACTGGACTTGC
psaA	GCCGGTCTAGTTCTAGCAG	**rpoC1**	AATGGTATTCCTGTTGGAGAG
psaB	TCTCAAAGGGGAGGGG	**rpoC2**	CGGGCGCCAACTTTAC
psaC	CGTGTCCTTGTGATGTTCTTG	**rps10**	TGCGTTCTTATAGTCCAGCC
psaD	GTGGTTACGTGGGTTGC	**rps11**	TTAGCAATACAATTGCAACACTTAC
psaF	CGCGATCTGTATAGGTTGG	**rps12**	GGTTCGAGGAGGTCGTG
psaL	GGCTGCACCTATTTCGAG	**rps13**	GCTCTCGAAAACGGAAATC
psbA	GCTATCAGGCTCACTTTTATATGC	**rps14**	CTCGTAATTCTGCACCGAC
psbB	ATCGCGGCAGGAACTC	**rps17**	CAGGTTGATTCGGCAAC
psbC	CGACGGCTGCTGAAG	**rps19**	GGTCACGGAGCTCAACTATAC
psbD	TATCAGTGGGAGGTTGGTTAAC	**rps3**	TGAGATTCGTCGGCAG
psbE	AACAGGCGAACGTCCG	**rps4**	CTGCAGGCTAAACAACGAC
psbH	CAGCAGGTTCTCCCGAC	**rps5**	CTACTAACACGCTCACGCTC
psbI-1	AAGGCTCATGATTTTGGG	**rps7**	GGTTCGAGGAGGTCGTG
psbI-2	GCTTAGCTAACGGATTTACAATC	**rps8**	TTCAGACGTGCCTTTCC
psbI-3	TCAGGGAAGGTTGATGGTAAAG	**secA**	GGTATGGGCAAAAGTAAACC
psbL	AGATAACCCATTCGAAGGC	**secY**	TTGGGCTTTATTGCTGC
psbN	CGAGTTCTCTACATATAGCTGGTG	**tatC**	CAGGTCAACTAAGCGGG
psbT	TCTTACTTTTTGGTACTCTAGGGG	**tufA**	AGTTCCAGAAGCGGACG
psbV	TTCTGAGTTACACCCTTGTCC	**ycf3**	GTCTCTCAGCCCAAACATC
psbX	CTTCATGCAAAGTCTCGTTC	**ycf39**	CAACAATACTAGTGGCCGG
rbcL	GCGGAGTTAGAAAGCCC	**ycf4**	CAGAATTCATACCTCAAGGGTTAG

in fucoxanthin plastid lineages (Dorrell and Howe, 2012a; Richardson *et al.*, 2014), was used for the cDNA synthesis step, to enrich for polyuridylylated transcripts. Illumina sequences were generated from the double-stranded cDNA library using a MiSeq platform. Contigs of probable plastid origin were identified by reciprocal BLAST searches against plastid protein sequences, as inferred from the translation products of plastid transcripts from the related fucoxanthin dinoflagellate *Karlodinium veneficum* (Jackson *et al.*, 2013; Richardson *et al.*, 2014), and four free-living haptophyte species (*Emiliania huxleyi, Phaeocystis globosa, Pavlova lutheri,* and the uncultured haptophyte C19847) (Baurain *et al.*, 2010; Cuvelier *et al.*, 2010; Puerta *et al.*, 2005). PCR forward primers were designed for each contig identified through this approach, and an individual RT-PCR, using an oligo-d(A) cDNA synthesis and

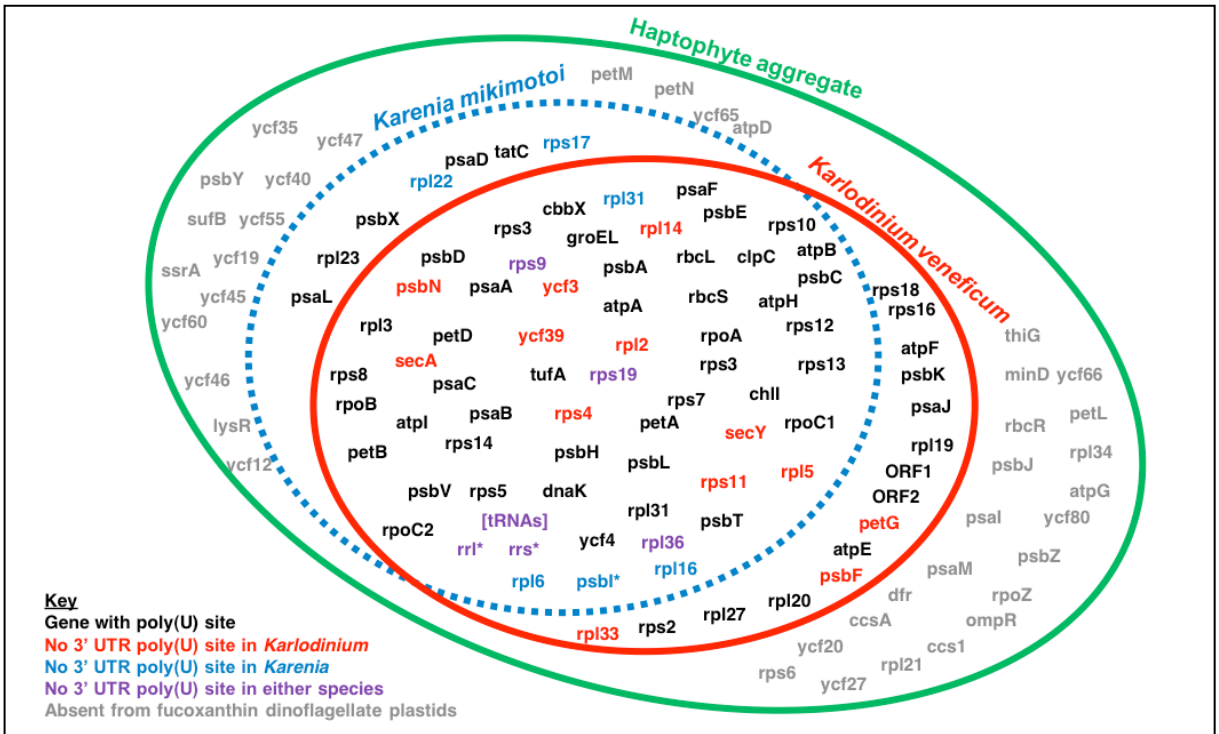


Fig. 6.1: The *Karenia mikimotoi* plastid transcriptome.

The Venn diagram shows the total polyuridylylated transcripts that could be assigned to the *Karenia mikimotoi* plastid (blue, dashed circle), overlaid with the coding content of the plastid genome of the related fucoxanthin dinoflagellate *Karlodinium veneficum* (red circle) (Gabrielsen *et al.*, 2011). Genes that are located in the plastids of other haptophyte species (*Emiliania huxleyi, Phaeocystis globosa, Pavlova lutheri,* and the uncultured species C19847), which were not identified in the *Karenia mikimotoi* plastid and are not plastid-encoded in *Karlodinium veneficum*, are contained within the larger green circle.

Genes are coloured according to the presence of poly(U) sites. Genes in black possess a poly(U) site in the associated 3' UTR in all fucoxanthin dinoflagellate species in which they are plastid-located. Genes in blue, red, and purple respectively lack associated poly(U) sites in the 3' UTR of the gene concerned in *Karenia mikimotoi*, in *Karlodinium veneficum*, and in both fucoxanthin dinoflagellate species, although these genes may give rise to polyuridylylated polycistronic transcripts. Genes labelled with an asterisk (*psbI, rrl, rrs*) are those that could not be identified to give rise to polyuridylylated polycistronic or monocistronic transcripts in *Karenia mikimotoi*, but are inferred to be located in the plastid genome. Genes surrounded with square brackets were not identified from the assembly of next generation sequencing data, and were identified using alternative strategies (e.g. TAiL-PCR).

Table 6.2: Tabulated plastid transcript sequences from *Karenia mikimotoi*.

Gene	Extent	Read coverage	Contig length (bp)	Poly(U) transcript?	Poly(U) site in 3' UTR?	3' UTR length (bp)	Poly(U) length (bp)
atpA	5' Partial	171.29	876	Y	Y	16	22
atpB	5' Partial	99.20	783	Y	Y	29	28
atpH	Complete	16.08	336	Y	Y	5	54
atpI	5' Partial	12.98	336	Y	Y	21	34
cbbX	Complete	5.86	1005	Y	Y	13	21
chlI	5' Partial	20.44	772	Y	Y	3	23
clpC	5' Partial	2.88	1150	Y	Y	18	19
dnaK	5' Partial	11.65	1363	Y	Y	9	16
groEL	5' Partial	5.92	522	Y	Y	12	28
petA	5' Partial	12.69	905	Y	Y	27	22
petB	5' Partial	45.95	511	Y	Y	7	29
petD	5' Partial	18.26	497	Y	Y	10	23
psaA	Complete	165.94	2383	Y	Y	21	24
psaB	Complete	96.31	2369	Y	Y	10	21
psaC	Complete	7.58	313	Y	Y	20	28
psaD	Complete	7.89	652	Y	Y	2	33
psaF	Complete	11.91	590	Y	Y	28	23
psaL	5' Partial	2.09	284	Y	Y	5	23
psbA	Complete	3442.63	1127	Y	Y	21	20
psbB	5' Partial	44.74	606	Y	Y	0	18
psbC	Complete	123.42	1489	Y	Y	11	20
psbD	Complete	257.93	1066	Y	Y	9	6
psbE	Complete	3.06	374	Y	Y	6	37
psbH	Complete	9.66	389	Y	Y	6 to 28	30
psbI	Complete	5.93	445	N	N	x	x
psbL	Complete	1.91	212	Y	Y	4	37
psbN	Complete	3.12	373	Y	Y	32 to 104	29
psbT	5' Partial	1.93	164	Y	Y	6	83
psbV	Complete	12.88	545	Y	Y	3	24
psbX	5' Partial	2.05	356	Y	Y	3	22
rbcL	5' Partial	237.82	1326	Y	Y	7	23
rbcS	5' Partial	312.49	439	Y	Y	4	99
rpl14	Complete	6.52	1559	Y	Y	21	44
rpl16	Complete	6.00	409	Y	N	x	x
rpl2	Complete	5.80	1033	Y	Y	40	62
rpl22	Complete	9.69	1507	Y	N	x	x
rpl23	Complete	7.24	493	Y	Y	20	65
rpl3	Complete	1.51	718	Y	Y	23	12
rpl31	Complete	5.92	1770	Y	N	x	24
rpl36	Complete	0.40	1475	Y	N	x	x
rpl5	Complete	8.72	1559	Y	Y	8	36
rpl6	5' Partial	32.95	895	Y	N	x	x
rpoA	5' Partial	4.00	717	Y	Y	33	24
rpoB	5' Partial	9.78	494	Y	Y	20	25
rpoC1	5' Partial	4.68	1009	Y	Y	3	26
rpoC2	5' Partial	2.51	575	Y	Y	4	94
rps10	5' Partial	6.29	490	Y	Y	50	59
rps11	Complete	3.97	1475	Y	Y	31	12
rps12	Complete	9.15	1770	Y	Y	59	x
rps13	Complete	5.96	1475	Y	Y	27	15
rps14	5' Partial	2.15	362	Y	Y	18	33
rps17	5' Partial	7.44	1559	Y	N	x	x
rps19	5' Partial	2.76	1507	Y	N	x	x

Table 6.2 (continued)

Gene	Extent	Read coverage	Contig length (bp)	Poly(U) transcript?	3' UTR poly(U) site?	3' UTR length (bp)	Poly(U) length
rps3	Complete	16.90	1507	Y	Y	99	27
rps4	5' Partial	13.60	471	Y	Y	49	25
rps5	Complete	12.33	895	Y	Y	55	31
rps7	Complete	15.03	1770	Y	Y	3	25
rps8	Complete	13.79	1559	Y	Y	43	22
rps9	5' Partial	6.81	1770	Y	N	x	x
rrl	Internal	3298.86	767	N	N	x	x
rrs	Complete	2381.44	1304	N	N	x	x
secA	5' Partial	6.15	1446	Y	Y	37	24
secY	Complete	4.83	1025	Y	Y	54	19
tatC	5' Partial	12.24	683	Y	Y	25	21
tufA	5' Partial	70.32	327	Y	Y	3	32
ycf3	Complete	2.29	556	Y	Y	5	23
ycf39	Complete	7.07	1061	Y	Y	20	21
ycf4	Complete	9.46	772	Y	Y	3	13
Total			63322				
Average		185.7				19.7	30.4

PCR reverse primer, and a PCR forward primer specific to the contig, was performed to confirm that the contig formed part of a polyuridylylated transcript (Table 6.1).

63322 bp sequence, covering 68 genes, were identified from the next generation sequencing data that are likely to be located in the *Karenia mikimotoi* plastid genome (Fig. 6.1, genes inside blue circle except those in parentheses; Table 6.2). This was very similar to the number of genes (76) identified in the *Karlodinium veneficum* plastid genome (Fig. 6.1, genes inside red circle), but was much smaller than the number of genes found in the plastids of free-living haptophytes (Fig. 6.1, genes inside green circle) (Gabrielsen *et al.*, 2011; Green, 2011; Puerta *et al.*, 2005; Richardson *et al.*, 2014). Many of the genes found in haptophyte plastids were not identified in either fucoxanthin dinoflagellate (Fig. 6.1; genes in grey).

Widespread distribution of poly(U) sites in fucoxanthin plastid genomes

Of the 68 genes identified within the *Karenia mikimotoi* plastid, 65 gave rise to polyuridylylated products identifiable by oligo-d(A) RT-PCR (Fig. 6.1, genes in blue circle, except those asterisked or in parentheses). Three further genes (*psbI, rrl, rrs*) could not be identified as part of a polyuridylylated transcript, but were assigned to the *Karenia mikimotoi* plastid for other reasons, and will be discussed subsequently (Fig. 6.1, asterisked genes).

Table 6.3. Gene clusters identified in *Karenia mikimotoi*

This table lists the gene clusters identified in *K. mikimotoi* via the direct assembly of transcriptome data, and via thermal asymmetric interlaced PCR. Primers for the thermal asymmetric interlaced PCRs that yielded multigene contigs are listed at the bottom. (anti) denotes a gene in a reverse transcriptional orientation relative to the remainder of the contig.

1. Gene clusters	Method of assembly	Poly(U) genes
psbC-Met^{CAT}	TAiL-PCR	psbC
psbD-Met^{CAT}-ycf4-(anti)rpoA	TAiL-PCR	psbD; ycf4; rpoA
rbcL-Phe^{AAA}	TAiL-PCR	rbcL
rpl16-rps17-rpl14-rpl5-rps8	Assembled from transcriptome data	rpl14; rpl5; rps8
rpl31-rps12-rps7	Assembled from transcriptome data	rps12; rps7
rpl36-rps13-rps11-(anti)atpI	Assembled from transcriptome data (rpl36-rps13-rps11) and TAiL-PCR (atpI)	rps13; rps11; atpI
rpl6-rps5	Assembled from transcriptome data	rps5
rps19-rpl22-rps3	Assembled from transcriptome data	rps3
(anti)Tyr^{GTA}-psbI	TAiL-PCR	none
tufA-psaA	TAiL-PCR	tufA; psaA

2. Primers for TAiL-PCR

Contig	gene-specific primer 1	gene-specific primer 2
psbC-Met^{CAT}	AATAGATGATTACTAGTAATAAATATAAAGAGGC	TAATCAACAACATTTTTAATTTAATCG
psbD-Met^{CAT}-ycf4	GCTATTCACGGAGCGAC	CAAACGGTGGTTACACTTCTTC
ycf4-(anti)rpoA	AAAACTAACGGTACATAATTATGCTAGAC	GCTCAGTTAGCCAATGGG
rbcL-Phe^{AAA}	CAACGATACTCCAGATGATCAAC	CCGCTAATAAAAATAGAAACTTATCC
rrps11-(anti)atpI	TTAGCAATACAATTGCAACACTTAC	ACGAGGTGGAATACTAAAGAGG
(anti)Tyr^{GTA}-psbI	AACATACCTTACTCTATAGCCTTTCG	TTTGGGTTTCGCGATG
tufA-psaA	CTAGCGGAATCAAATAAACGAC	CACGTTGTGCCAATTCC

	gene-specific primer 3		Arbitrary degenerate primers
psbC-Met^{CAT}	GCTACTTCCTTTAACTTTGAGGC	1	TTNTCGASTWTSGWGTT
psbD-Met^{CAT}-ycf4	TGGTAATGGTCTCTAACACGTC	2	TTWGTGNAGWANCANAGA
ycf4-(anti)rpoA	TGTAATCTCGAAGTCCTCG	3	CCTTNTWGAWTWTWGWWTT
rbcL-Phe^{AAA}	CCCTTTCTAAATTTTTAGAGTCG	4	CCTTWGTGNAWWANCANAWA
rps11-(anti)atpI	CCGTCGAAGACAACATTCTTAG	5	GGAACWACNTWTWNGTNTTW
(anti)Tyr^{GTA}-psbI	GTAGGGAAGCAGGTGTTGG	6	TTACWACANGWWGNTGNTWT
tufA-psaA	CGACAAAAGACCAAATACAAAAAG	7	GGAANACTWAWAWCWWAWA
		8	TTAANCWAGWCWCWAWWAA

57 of the genes tested by oligo-d(A) RT-PCR were found to give rise to a monocistronic polyuridylylated product, i.e. the poly(U) site was located in the associated 3' UTR of the gene concerned (Fig. 6.1; genes in black and in red). In the remaining 8 cases, oligo-d(A) RT-PCR generated a polycistronic product, containing the gene from which the PCR forward primer was designed, and one or more genes located downstream, terminating in a 3' poly(U) tail (Fig. 6.1; genes shaded in purple and in blue, except those asterisked or in parentheses; Tables 6.2, 6.3). These genes are unlikely to possess associated 3' UTR poly(U) sites, but can give rise to polyuridylylated polycistronic transcripts, as have previously

been identified in *Karlodinium veneficum*, and in the plastids of peridinin dinoflagellates (Barbrook *et al.*, 2012; Dang and Green, 2010; Richardson *et al.*, 2014).

I wished to determine whether there was any clear association between the distribution of poly(U) sites in the *Karlodinium veneficum* and the *Karenia mikimotoi* plastids. Comparing the distribution of poly(U) sites in the *Karenia mikimotoi* plastid transcriptome with data previously obtained for the *Karlodinium veneficum* plastid genome, only a small number of genes were found that did not possess an associated poly(U) site in the 3' UTR in either species (Fig. 6.1; genes in purple) (Richardson *et al.*, 2014). This association was not statistically significant (chi-squared test, P=0.15). Several genes were found that lack an associated 3' UTR poly(U) site in *Karlodinium veneficum*, but possess one in *Karenia mikimotoi* (Fig. 6.1; genes in red, within blue circle). Other genes were found to lack an associated 3' UTR poly(U) site in *Karenia mikimotoi*, but possess one in *Karlodinium veneficum* (Fig. 6.1; genes in blue, within red circle). Overall, it appears that poly(U) sites are associated with the majority of genes in fucoxanthin dinoflagellate plastids. This strongly suggests that poly(U) tail addition is a widespread feature of transcript processing in fucoxanthin plastids.

Identification of non-polyuridylylated transcripts in the *Karenia mikimotoi* plastid

In addition to the polyuridylylated transcripts identified, a small number of genes were identified from next generation sequencing data that are likely to be located in the *K. mikimotoi* plastid, but do not give rise to polyuridylylated transcripts. Sequences corresponding to plastid 16S and 23S rRNA genes were additionally identified from the next generation sequencing data, but polyuridylylated ribosomal RNA transcripts were not detectable by oligo-d(A) RT-PCR (Fig. 6.1). It is likely that the ribosomal RNA genes are plastid-located, as these genes are not known to have been relocated to the nucleus in any photosynthetic eukaryote (Green, 2011). The next generation sequencing dataset did not contain any identifiable tRNA genes. To confirm that tRNA genes are present in the *K. mikimotoi* plastid genome, bidirectional thermal asymmetric interlaced PCR (TAiL-PCR) extensions were performed of five representative plastid genes (*psbA, psbC, psbD, psaA, rbcL*) for which the underlying 3' UTR sequences in *Karenia mikimotoi* had previously been obtained (Table 6.3) (Dorrell and Howe, 2012a; Takishita *et al.*, 1999). TAiL-PCRs were additionally performed for one representative multigene contig (*rpl36-rps13-rps11*) assembled directly from the next generation sequencing data (Table 6.3). tRNA genes were identified that were adjacent to the *psbC, psbD, psbI* and *rbcL* genes (Tables 6.1, 6.3). These tRNA genes were found to lack associated poly(U) sites by oligo-d(A) RT-PCR (Fig. 6.1, Table 6.1). The absence of poly(U) sites from ribosomal and transfer RNA genes is

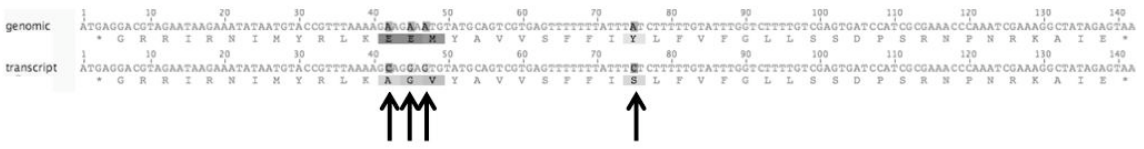


Fig. 6.2: Presence of editing on *K. mikimotoi psbI* transcripts.

This figure shows the aligned genomic and transcript sequences of *K. mikimotoi psbI.* Four discrepancies between the *psbI* genetic and transcript sequences, labelled with arrows, indicate that *psbI* transcripts are subject to editing.

consistent with previous reports from other dinoflagellate species that poly(U) tails are principally associated with plastid mRNA processing (Nelson *et al.*, 2007; Richardson *et al.*, 2014; Wang and Morse, 2006).

In addition to the transfer and ribosomal RNA genes identified, a contig was found within the next generation sequencing dataset was identified that corresponded to the plastid *psbI* gene. Oligo-d(A) primed RT-PCR against the *psbI* gene did not generate any products, indicating that *psbI* does not give rise to polyuridylylated monocistronic or polycistronic transcripts. The absence of poly(U) tails from *psbI* transcripts was confirmed independently using three alternative RT-PCR forward primers, and repeating each RT-PCR with RNA samples isolated from different source cultures (Table 6.1).The genetic element underpinning the *psbI* transcript was sequenced by TAiL-PCR, and found to be adjacent to a predicted tyrosyl-tRNA gene (Table 6.3). Four sites were found to differ to the corresponding transcript sequence, suggesting that the *psbI* transcript is edited (Fig. 6.2). Editing is associated with plastid transcripts, but not nuclear transcripts in fucoxanthin dinoflagellates, indicating that the *psbI* gene is located on the *K. mikimotoi* plastid genome (Dorrell and Howe, 2012a; Jackson *et al.*, 2013; Richardson *et al.*, 2014).

Independent gene transfer events in fucoxanthin plastid genomes

The *Karlodinium veneficum* plastid genome retains far fewer genes than the plastid genomes of free-living haptophyte relatives (Gabrielsen *et al.*, 2011; Richardson *et al.*, 2014). The genes that have been lost from the *Karlodinium veneficum* plastid have been suggested to have been relocated to the nucleus, although none of these genes has previously been identified in nuclear EST libraries of fucoxanthin dinoflagellate species (Burki *et al.*, 2014; Dorrell and Howe, 2012b; Gabrielsen *et al.*, 2011). I wished to determine whether plastid genes have been transferred to the nuclei of fucoxanthin dinoflagellates. I additionally wished

Table 6.4: Genes of probable plastid origin identified from *Karlodinium veneficum* nuclear EST libraries.

This table lists contigs assembled from sequences identified *Karlodinium veneficum* nuclear ESTs by reciprocal tBLASTn/ BLASTx searches with *Karenia mikimotoi* plastid transcript sequences. The first 50 aa of the predicted translation product of the contig is shown, alongside (for rpl22, and rpl23) targeting predictions obtained using HECTAR and/ or TargetP (Emmaneulson *et al.*, 1999; Gschloessl *et al.*, 2009). Complete nucleotide sequences of each contig are shown in Appendix 3.

	Coverage	CDS interval	Constituent accessions	Targeting prediction
1. psaD	Internal	<1-end	AmSd244SL1 , Am2d85SL1	
Sequence	FIRDGEVEKYVMTWSSKSEQIIELPTGGAASMKQGENLMYFRKKEQALAL...			
2. rpl22	Complete	328-780	KME00004684, KME00008386	[20 aa SP + 24 aa TP] (HECTAR)
Sequence	MWRTSMIVAHLASSIYAVSPPLSYRAGSEMSSGVAMRRLADALMNNNRIR...			
3. rpl23	Complete	8-529	AmSd316SL1	SP (HECTAR) / 78 aa TP (TargetP)
Sequence	MALRVLVSIALACLAREAHTENEETEKLASLLFALAPQHPQMKVATSGQ...			

to determine whether independent gene transfer events have occurred in different fucoxanthin dinoflagellate species since their divergence.

Seven of the genes inferred to be located in the *Karenia mikimotoi* plastid are not present in the *Karlodinium veneficum* plastid genome (Fig. 6.1). Of these seven genes, transcripts corresponding to three genes (*psaD, rpl22,* and *rpl23*) were identified in published nuclear EST libraries from *Karlodinium veneficum* (Table 6.4). The complete N-termini of the *Karlodinium veneficum* Rpl22 and Rpl23 protein sequences were assembled from the EST data (Table 6.4). These were found to contain a predicted plastid targeting sequence, consisting of an N-terminal signal peptide, and a downstream transit peptide sequence (Table 6.4). The targeting sequences identified were consistent in structure with plastid targeting sequences that have previously been characterised in fucoxanthin dinoflagellates, confirming that these genes have been relocated to the nucleus (Patron and Waller, 2007; Yokoyama *et al.*, 2011).Thus, independent gene transfer events have occurred in individual fucoxanthin dinoflagellate species since their divergence from each other.

Independent changes to fucoxanthin plastid gene order and content

In addition to having undergone extensive reduction, fucoxanthin plastid genomes are highly divergently organised relative to other plastid lineages. The *Karlodinium veneficum* plastid

genome contains evidence for extensive recombination relative to free-living haptophytes, and many of the individual genes contain insertions or deletions unique to this species (Gabrielsen *et al.*, 2011; Richardson *et al.*, 2014). I wished to determine whether changes to gene order and structure have occurred independently in different fucoxanthin plastid genomes since their divergence.

To determine when recombination events have occurred in fucoxanthin plastids, the gene order of polycistronic loci in *Karenia mikimotoi* identified by assembly of NGS read data and by TAiL-PCR was compared to the plastid genomes of *Karlodinium veneficum* and free-living haptophytes. Species-specific recombination events were found in both fucoxanthin dinoflagellates. For example, the *rpl36-rps13-rps11* operon, found in the plastids of most algae, including haptophytes and *Karenia mikimotoi,* has been disrupted in *Karlodinium veneficum*, with *rps13* located upstream of and in opposing orientation to *secY*, and *rps11* located downstream of and in opposing orientation to a prolyl-tRNA gene (Gabrielsen *et al.*, 2011; Green, 2011). Similarly, the *Karenia mikimotoi psbD* gene is located upstream of a methionyl-tRNA gene (Met^{CAT}), and the photosystem I assembly factor gene *ycf4.* This locus is not known in any other plastid lineage including in free-living haptophytes, as *psbD* is typically located upstream of *psbC* (Green, 2011; Oudot-Le Secq *et al.*, 2007). The *Karlodinium veneficum psbD* gene, and is located upstream of and in opposing orientation to the *Karlodinium veneficum atpA* gene (Gabrielsen *et al.*, 2011).

To determine whether sequence insertions and deletions have occurred independently in individual fucoxanthin plastid lineages, 9179 aa plastid protein sequence, generated by the conceptual translation of 54 plastid transcript sequences in *Karenia mikimotoi*, was aligned to the equivalent predicted plastid protein sequences from *Karlodinium veneficum* and from free-living haptophytes. Within this dataset, 109 sequence insertions or deletions were identified that were present in either *Karenia mikimotoi* or *Karlodinium veneficum,* but were not present in haptophytes (Table 6.5). Of these, only 10 were conserved between both fucoxanthin dinoflagellates, while the remaining 99 were unique to either *Karenia mikimotoi* or to *Karlodinium veneficum* (Table 6.5). Overall, it appears that the plastid genomes of fucoxanthin dinoflagellates have diverged substantially in content and organisation since their endosymbiotic acquisition.

Table 6.5. Tabulated indels identified across 9179 aa aligned fucoxanthin dinoflagellate plastid protein sequence

Indels are listed by form (insertion, deletion, and N- and C-terminal extension) and by gene. Indels were identified by alignment against orthologous plastid protein sequences from 17 different species of algae, including the haptophytes *Emiliania huxleyi, Phaeocystis globosa,* and *Pavlova lutheri*, as well as representatives of stramenopiles, cryptomonads, red algae, green algae and glaucophytes (listed in Chapter Two). Indels are only counted if they were not found in any species other than *Karenia mikimotoi* or *Karlodinium veneficum.* These data were obtained with the assistance of an undergraduate student, George Hinksman.

		Evolutionary distribution		
1. By form	**Total**	**Both taxa**	***Karenia mikimotoi***	***Karlodinium veneficum***
Insertions	59	8	21	30
Deletions	25	0	9	16
N-terminal extensions	9	1	4	4
C-terminal extensions	16	1	9	6
Total	109	10	43	56

		Evolutionary distribution		
2. By gene	**Alignment length**	**Both taxa**	***Karenia mikimotoi***	***Karlodinium veneficum***
atpA	245	0	0	3
atpB	236	0	0	0
atpH	61	0	0	1
atpI	61	0	0	0
cbbX	291	0	4	1
chlI	128	0	0	0
clpC	275	0	2	0
dnaK	443	1	6	0
groEL	57	0	0	0
petA	284	1	3	3
petB	151	0	0	0
petD	141	0	1	0
psaA	767	0	1	3
psaB	468	0	0	3
psaC	82	0	1	0
psaF	185	2	3	7
psbA	206	0	0	0
psbB	199	0	0	0
psbC	472	0	0	1
psbD	199	0	0	0
psbE	85	0	2	1
psbH	67	0	0	0
psbI	39	0	0	0
psbL	39	0	1	1
psbN	44	1	0	1
psbT	28	0	0	0
psbV	165	0	1	0
rbcL	302	0	0	0
rbcS	112	0	0	0
rpl14	122	0	0	1
rpl16	127	1	1	0
rpl2	176	0	0	2

Table 6.5 (continued)

	Alignment length	Evolutionary distribution		
		Both taxa	*Karenia mikimotoi*	*Karlodinium veneficum*
rpl31	71	0	1	0
rpl36	49	0	1	0
rpl5	98	0	0	0
rpl6	85	0	0	0
rpoA	191	0	3	2
rpoC1	283	1	1	5
rps11	131	0	2	1
rps12	90	0	0	0
rps13	125	0	2	0
rps14	49	0	0	1
rps19	59	0	0	1
rps3	217	0	1	4
rps4	77	0	1	0
rps5	160	1	0	0
rps7	157	1	0	4
rps8	140	1	0	2
secA	57	0	0	0
secY	161	0	1	2
tufA	95	0	0	0
ycf3	172	0	2	1
ycf39	220	0	1	2
ycf4	128	0	1	1
Total	9179			

Relationships between poly(U) tail addition and cleavage of polycistronic transcripts

Previous studies of peridinin dinoflagellates have identified the presence of polycistronic polyuridylylated transcripts (Barbrook *et al.*, 2012). However, the majority of transcripts in peridinin dinoflagellate plastids, as identified by northern blotting studies, are monocistronic (Dang and Green, 2009; Nisbet *et al.,* 2008). It is possible that the polycistronic polyuridylylated transcripts identified might represent the mature transcripts produced from specific loci (Barbrook *et al.*, 2012). Alternatively, poly(U) tails might be added to transcripts early in processing, prior to the cleavage of the mature transcript 5' end (Dang and Green, 2010; Nisbet *et al.*, 2008). There is even evidence at certain plastid loci that poly(U) tail addition might be involved in alternative cleavage events, specifying which mature mRNAs are produced from polycistronic precursors containing multiple poly(U) sites (Barbrook *et al.*, 2012). Previously, I have presented evidence that poly(U) sites may similarly be involved in alternative end cleavage in *Karlodinium veneficum* (Richardson *et al.*, 2014).

I wished to determine whether poly(U) tail addition was related to other transcript cleavage events in *Karenia mikimotoi*. I wished to determine whether polycistronic polyuridylylated transcripts were abundant in *Karenia mikimotoi*, or whether the majority of the

Table 6.6. Northern blot probes to detect *Karenia mikimotoi* plastid transcripts.

This table lists the sequence of the T7 arm of the pGEM-T Easy vector alongside the first 50 bp of each probe sequence complementary to *K. mikimotoi* plastid gene sequences. The range of sequence covered by each probe is given relative to the underlying CDS, as identified by PCR.

T7 arm TAATACGACTCACTATAGGGCGAATTGGGCCCGACGTCGCATGCTCCCGGCCGCCATGGCCGCGGGATT

Probe	start	end	Sequence
rps13	402	303	GTTGTCCTCGAGTTGGAAGACCCGCGTTGCGTCTTTTTCCCCTCAGGGTT...
rps11	466	232	CCCCAACCTGCACCAGCTACAGTCACTTGTACCTCAGTTATGTTGAACAT...
psbD	393	70	CTATAGGACCAGAAAAAGCAATCGCATTGTAAGGTCTAATACCGACCAAC...
ycf4	665	493	TTGACCAATTCGCATATAATATTTTACACTAATTTAGTTGTCAACTGTCA...

polyuridylylated transcripts were monocistronic. I additionally wished to determine whether plastid transcripts in *Karenia mikimotoi* undergo alternative end processing.

The *rpl36-rps13-rps11* and *psbD-Met^{CAT}-ycf4* loci were selected as models in which to investigate transcript processing events. *rpl36-rps13-rps11* was one of the multigene loci that could be directly assembled from next generation sequencing data, and polyuridylylated dicistronic *rpl36-rps13* and tricistronic *rpl36-rps13-rps11* transcripts could be directly amplified by oligo-d(A) RT-PCR using a primer specific to *rpl36* (Tables 6.1, 6.3). In contrast, the *psbD-Met^{CAT}-ycf4* locus was assembled from TAiL-PCR data, and oligo-d(A) RT-PCRs for *psbD* and *ycf4* only yielded monocistronic transcripts (Table 6.3).

To quantify polycistronic transcripts at each locus, northern blots of *K. mikimotoi* RNA were hybridised to probes specific to *rps13, rps11, psbD* and *ycf4* (Fig. 6.3; Table 6.6). Northern blots were not probed for *rpl36*, as the coding sequence was too short to design a reliable probe. The terminus positions associated with transcripts at each locus were identified by performing RT-PCRs using circularised RNA and cDNA primers specific to the *rps13, rps11, psbD* and *ycf4* genes (Table 6.1). To determine the full diversity of transcripts generated from the *rpl36-rps13-rps11* and *psbD-Met^{CAT}-ycf4* loci, PCR primers were designed against different regions of each locus, and employed in different combinations (Table 6.7). For example, for *psbD* cDNA, two PCR reverse were primers designed to anneal to the *psbD* CDS, and ten forward primers were used, of which three were designed to anneal within *psbD* to detect monocistronic transcripts, two were designed within the intergenic region containing *Met^{CAT}*, and five were designed to anneal within *ycf4* to detect polycistronic

Table 6.7. Primers used for circular RT-PCR of *Karenia mikimotoi* plastid transcripts.

1. rpl36-rps13-rps11	rps13	rps11
cDNA primer 1	GTTGTCCTCGAGTTGGAAG (rps13 3' end)	GCAATTGTATTGCTAAAGTTAGCTAATATATG (rps11 5' end)
cDNA primer 2	CCCTTTTCGTTTTACAATTTG (rps13 5'end)	
Reverse primer 1	CCCTTTTCGTTTTACAATTTG (rps13 5' end)	CCCTTTTCGTTTTACAATTTG (rps13 5' end)
Reverse primer 2	ATCGTTTACGAAGCGAACTC (rps13 5' end)	ATCGTTTACGAAGCGAACTC (rps13 5' end)
Reverse primer 3		GTTGTCCTCGAGTTGGAAG (rps13 3' end)
Reverse primer 4		TTTAATTAAATACCTAGGAAATATCAACTGTAAC (rps13 3' end)
Reverse primer 5		CTAGGAAATATCAACTGTAACTCTTGC (rps11 5' end)
Reverse primer 6		CGAAATCCCTTCCAATTTTG (rps11 5' end)
Forward primer 1	GCTCTCGAAAACGGAAATC (rps13 5' end)	
Forward primer 2	AACGTTATTGAAGATCCCAAAC (rps13 3' end)	
Forward primer 3	CGGAAGCGGTATTAAGGC (rps13 3' end)	
Forward primer 4	AAGTTCAAATGAAGTAAGACTCAAAAG (intergenic)	
Forward primer 5	TTAGCAATACAATTGCAACACTTAC (rps11 5' end)	
Forward primer 6	ACGAGGTGGAATACTAAAGAGG (rps11 3' end)	ACGAGGTGGAATACTAAAGAGG (rps11 3' end)
Forward primer 7	CCGTCGAAGACAACATTCTTAG (rps11 3' end)	CCGTCGAAGACAACATTCTTAG (rps11 3' end)

2. psbD-Met^{CAT}-ycf4	psbD	ycf4
cDNA primer 1	CCTCCTAGTTCAAGCCACC (psbD 5' end)	TCTGGAATTGACAGTTGACAG
Reverse primer 1	AAGTAATCCTGACCAACCAATG (psbD 5' end)	AAGTAATCCTGACCAACCAATG (psbD 5' end)
Reverse primer 2	GTGTGGAAACGGCTGC (psbD 5' end)	GTGTGGAAACGGCTGC (psbD 5' end)
Reverse primer 3		CCTCCTAGTTCAAGCCACC (psbD 5' end)
Reverse primer 4		CAACCGTGCTATTTCAAACTG (psbD 5' end)
Reverse primer 5		GTTTTCATGAGGTTGATCTTGG (psbD 3' end)
Reverse primer 6		GCGACCTTGGGCTTATG (Met^{CAT})
Reverse primer 7		TATATTTCTTTGTCCCAAACTGAG (ycf4 5' end)
Reverse primer 8		CGTTAACAAATACTTCGCCAG (ycf4 5' end)
Forward primer 1	GCTATTCACGGAGCGAC (psbD 3' end)	
Forward primer 2	CAAACGGTGGTTACACTTCTTC (psbD 3' end)	
Forward primer 3	TGGTAATGGTCTCTAACACGTC (psbD 3' end)	
Forward primer 4	CATAAGCCCAAGGTCGC (Met^{CAT})	
Forward primer 5	CGTTCAATCTTCTCCTCAAC (Met^{CAT})	
Forward primer 6	CAGAATTCATACCTCAAGGGTTAG (ycf4 5' end)	
Forward primer 7	AAAACTAACGGTACATAATTATGCTAGAC (ycf4 3' end)	AAAACTAACGGTACATAATTATGCTAGAC (ycf4 3' end)
Forward primer 8	GCTCAGTTAGCCAATGGG (ycf4 3' end)	GCTCAGTTAGCCAATGGG (ycf4 3' end)
Forward primer 9	TCTGGAATTGACAGTTGACAG (ycf4 3' end)	TCTGGAATTGACAGTTGACAG (ycf4 3' end)
Forward primer 10	TTGACAGCTGACAACTAAATTAGTG (ycf4 3' end)	TTGACAGCTGACAACTAAATTAGTG (ycf4 3' end)

transcripts covering all three genes (Table 6.7). Each possible combination of PCR reverse and forward primer (e.g. for *psbD,* 20 different combinations) was tested, and each RT-PCR was repeated three times, using cDNA templates generated from independently isolated and circularised RNA samples.

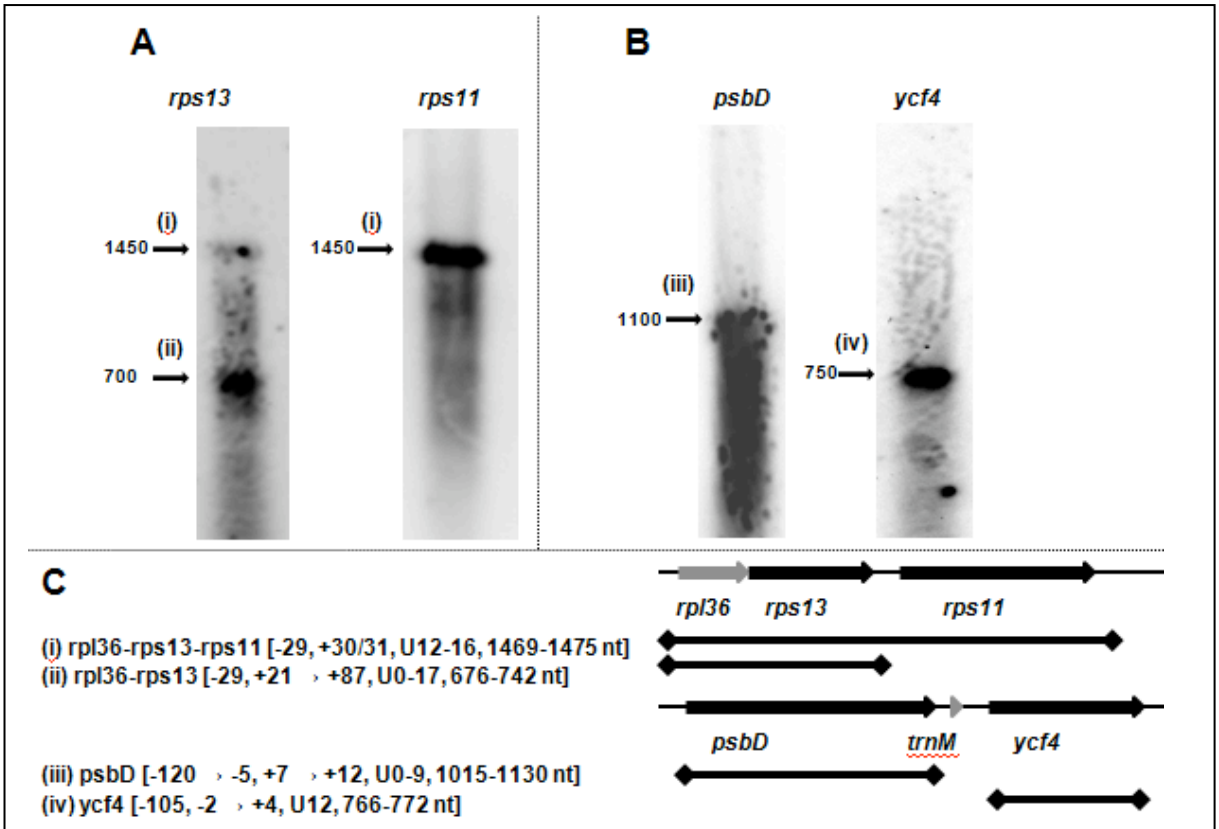


Fig. 6.3: Transcript processing at the *K. mikimotoi rpl36-rps13-rps11* and *psbD-Met^{CAT}-ycf4* loci.

Panels A and **B** show the results of northern blots to identify transcripts covering the *rps13* and *rps11* **(A)**, and *psbD* and *ycf4* genes **(B)**. A representative lane of the DIG-labelled RNA molecular weight marker used to identify each transcript is shown to the left of each blot, and each band is labelled with the expected size of the corresponding transcript.

Panel C shows schematic diagrams of the *rpl36-rps13-rps11* and *psbD-Met^{CAT}-ycf4* loci. Genes that possess associated poly(U) sites are shown in black, and genes that lack poly(U) sites in their associated 3' UTR are shown in grey. Thin black lines correspond to non-coding DNA. The transcripts identified by circular RT-PCR that correspond to visible bands in each northern blot are shown for each locus. Labels in round brackets correspond to the labels above each band in the northern blot, and labels in square brackets give the 5' end position, 3' end position, poly(U) tail length, and lengths of the transcripts identified.

At the *rpl36-rps13-rps11* locus, the overwhelming majority of mature transcripts identified were polycistronic. For the *rps13* northern blot, two bands were identified, one of approximately 1450 nt length, and one of 700 nt length (Fig. 6.3, panel A). Polycistronic *rpl36-rps13-rps11* transcripts and *rpl36-rps13* of equivalent length, respectively, to the 1450 nt and 700 nt bands, were amplified by circular RT-PCR (Fig. 6.3, panel C; Table 6.8). Monocistronic *rps13* transcripts could not be identified using either technique. While monocistronic *rps11* transcripts were detectable by circular RT-PCR, only the 1450 nt *rpl36-rps13-rps11* band was visible in the *rps11* blot, suggesting that monocistronic *rps11* transcripts are low in abundance (Fig. 6.3, panel A; Table 6.8).

All of the *rpl36-rps13-rps11* transcripts identified contained a 3' poly(U) tail that did not

Table 6.8. Circular RT-PCR data for the *K. mikimotoi rpl36-rps13-rps11* and *psbD-Met^{CAT}-ycf4* loci.

This table lists all of the circular RT-PCR products obtained for sense transcripts over each locus, and the PCR primers used to identify them. Terminus positions are given for relative to the underlying CDS. PCR primer numbers correspond to those given in Table 6.7. Transcripts of a length equivalent to bands identified in northern blots are highlighted in bold, and the corresponding band number is listed as per Fig. 6.3.

1. rpl36-rps13-rps11	Transcript dimensions				Primers		Northern band	Notes
rpl36-rps13	5' end	3' end	Poly(U)	Length	R	F		
Non-poly(U) transcript 1	**-29**	**21**	**0**	**676**	**2**	**1**	**ii**	
Non-poly(U) transcript 2	**-29**	**54**	**0**	**709**	**2**	**3**	**ii**	**3' end extends through poly(U) site**
Non-poly(U) transcript 3	**-29**	**87**	**0**	**742**	**2**	**2**	**ii**	**3' end extends through poly(U) site**
Non-poly(U) transcript 4	**-29**	**87**	**0**	**742**	**2**	**2**	**ii**	**3' end extends through poly(U) site**
Non-poly(U) transcript 5	-12	-15	0	623	2	1		
Non-poly(U) transcript 6	-11	-11	0	626	2	1		
Non-poly(U) transcript 7	5	23	0	644	2	1		
Non-poly(U) transcript 8	5	-170	0	451	2	1		
Non-poly(U) transcript 9	13	-194	0	419	2	1		
Non-poly(U) transcript 10	22	14	0	618	2	3		
Non-poly(U) transcript 11	22	32	0	636	2	1		
poly(U) transcript 1	**-29**	**32**	**5**	**687**	**2**	1	**ii**	
poly(U) transcript 2	**-29**	**38**	**17**	**710**	**2**	**3**	**ii**	
rps11								
Non-poly(U) transcript 1	-115	-8	0	838	4	6		5' end extends into rps13
poly(U) transcript 1	-50	30	8	819	4	6		5' end extends into rps13
poly(U) transcript 2	52	31	12	722	6	7		
poly(U) transcript 3	52	31	12	722	6	7		
rpl36-rps13-rps11								
poly(U) transcript 1	**-29**	**30**	**14**	**1475**	**2**	**6**	**i**	
poly(U) transcript 2	**-29**	**31**	**13**	**1475**	**2**	**5**	**i**	
poly(U) transcript 3	**-29**	**31**	**12**	**1474**	**2**	**5**	**i**	
poly(U) transcript 4	**-29**	**31**	**16**	**1478**	**2**	**6**	**i**	

Table 6.8 (continued)

2. psbD-Met^{CAT}-ycf4	Transcript dimensions				Primers		Northern band	Notes
	5' end	3' end	Poly(U)	Length	R	F		
psbD								
Non-poly(U) transcript 1	-139	278	0	1415	2	4		3' end extends into ycf4
Non-poly(U) transcript 2	-132	130	0	1260	1	3		3' end extends into Met^{CAT}
Non-poly(U) transcript 3	-131	43	0	1172	2	4		3' end extends into Met^{CAT}
Non-poly(U) transcript 4	-129	92	0	1219	2	4		3' end extends into Met^{CAT}
Non-poly(U) transcript 5	**-74**	**8**	**0**	**1080**	**1**	**3**	**iii**	
Non-poly(U) transcript 6	-69	-171	0	896	2	3		
Non-poly(U) transcript 7	**-60**	**7**	**0**	**1065**	**2**	**3**	**iii**	
Non-poly(U) transcript 8	6	-88	0	904	2	3		
Non-poly(U) transcript 9	22	-178	0	798	2	3		
Non-poly(U) transcript 10	24	-33	0	941	2	3		
Non-poly(U) transcript 11	129	40	0	909	2	3		
poly(U) transcript 1	**-120**	**12**	**8**	**1130**	**1**	**3**	**iii**	
poly(U) transcript 2	**-118**	**11**	**9**	**1127**	**1**	**3**	**iii**	
poly(U) transcript 3	**-118**	**12**	**8**	**1128**	**1**	**3**	**iii**	
poly(U) transcript 4	**-53**	**10**	**1**	**1061**	**2**	**3**	**iii**	
poly(U) transcript 5	**-53**	**10**	**1**	**1061**	**2**	**3**	**iii**	
poly(U) transcript 6	**-5**	**12**	**6**	**1015**	**2**	**3**	**iii**	
ycf4								
Non-poly(U) transcript 1	-191	274	0	1128	7	9		5' end in Met^{CAT}; extends through poly(U) site
Non-poly(U) transcript 2	-191	274	0	1128	7	9		5' end in Met^{CAT}; extends through poly(U) site
Non-poly(U) transcript 3	-191	274	0	1128	7	9		5' end in Met^{CAT}; extends through poly(U) site
Non-poly(U) transcript 4	-191	274	0	1128	7	9		5' end in Met^{CAT}; extends through poly(U) site
Non-poly(U) transcript 5	-146	-141	0	668	8	8		
Non-poly(U) transcript 6	-130	-248	0	545	8	8		
Non-poly(U) transcript 7	-129	-179	0	613	8	8		
Non-poly(U) transcript 8	-105	-233	0	535	7	7		
Non-poly(U) transcript 9	-47	-195	0	515	8	8		
Non-poly(U) transcript 10	-29	-56	0	636	7	7		
Non-poly(U) transcript 11	-25	-55	0	633	8	8		
Non-poly(U) transcript 12	-21	-56	0	628	7	7		
Non-poly(U) transcript 13	-21	-56	0	628	7	7		
Non-poly(U) transcript 14	-20	-4	0	679	7	9		
Non-poly(U) transcript 15	5	-76	0	582	8	8		
poly(U) transcript 1	**-105**	**-2**	**12**	**766**	**7**	**9**	**iv**	
poly(U) transcript 2	**-105**	**3**	**12**	**771**	**7**	**9**	**iv**	
poly(U) transcript 3	**-105**	**4**	**12**	**772**	**7**	**9**	**iv**	
psbD-Met^{CAT}-ycf4								
Non-poly(U) transcript 1	6	118	0	2036	2	9		

correspond to a poly(T) tract in the underlying genetic sequence. In contrast, only two of the *rpl36-rps13* transcripts were polyuridylylated (Table 6.8). Several of the non-polyuridylylated transcripts identified terminated upstream of the corresponding poly(U) site, and might correspond to degradation products of polyuridylylated transcripts, although the majority of

these were of substantially less than 700 nt length, and are thus unlikely to form an abundant component of the *K. mikimotoi* plastid transcriptome (as inferred from northern blot hybridisation) (Fig 6.4; Table 6.8). Three of the non-polyuridylylated transcripts, however, extended past the *rps13* poly(U) site into the 5' end of the *rps11* CDS, suggesting that they were generated through the alternative cleavage of a common precursor transcript covering all three genes. These transcripts were all of greater than 700 nt length and thus might correspond to hybridisation within the *rps13* northern blot (Fig. 6.3; Table 6.8). Notably, all of the *rpl36-rps13* transcripts of greater than 700 nt length, and all of the *rpl36-rps13-rps11* transcripts identified terminated at the same 5' end position, 29 nt upstream of the *rpl36* CDS (Table 6.8). As these are likely to be the most abundant transcripts produced from the *rpl36-rps13-rps11* locus (as inferred by northern blotting), this indicates that all of the mature transcripts produced from this locus share a consensus 5' processing site. Thus, the only terminus processing event that varies over the *rpl36-rps13-rps11* locus is that certain transcripts terminate within the *rps13* 3' UTR, while others extend to the *rps11* poly(U) site. Poly(U) tail addition might therefore play a specific role in alternative end processing at the *rpl36-rps13-rps11* locus.

In contrast to the situation for *rpl36-rps13-rps11*, the majority of transcripts covering the *psbD-Met*^{CAT}*-ycf4* locus were monocistronic (Fig. 6.3, panels B, C). The *psbD* northern blot yielded hybridisation of less than 1100 nt, and the *ycf4* northern blot a single band at 750 nt (Fig. 6.3, panel B), both of which were of equivalent size to monocistronic polyuridylylated transcripts obtained by circular RT-PCR (Fig. 6.3, panel C; Table 6.8). Polycistronic *psbD-Met*^{CAT} and *psbD-Met*^{CAT}*-ycf4* transcripts of 1200-2000 nt length, as well as a 1100 nt *Met*^{CAT}*-ycf4* transcript were identified through circular RT-PCR. However, hybridisation corresponding to these transcripts could not be identified in either blot (Fig. 6.3, panel B; Table 6.8). None of the polycistronic transcripts identified through circular RT-PCR were polyuridylylated. To determine whether polycistronic polyuridylylated transcripts are produced from this locus, an RT-PCR was performed using oligo-d(A) cDNA, a PCR forward primer positioned within *psbD,* and a PCR reverse primer positioned within *ycf4* (Table 6.9). Products were obtained, indicating that some polycistronic polyuridylylated transcripts are present, although given the absence of a corresponding northern signal, and the absence of poly(U) tails from the few polycistronic transcripts identified through circular RT-PCR, these transcripts are likely to only be present at extremely low abundance.

Table 6.9. Primers used to sequence non-polyuridylylated and polycistronic transcripts from the rpl36-rps13-rps11 and psbD-Met^{CAT}-ycf4 loci

Where a PCR reverse primer sequence is not specifically provided, the cDNA synthesis primer was used as the PCR reverse primer.

Transcript	cDNA synthesis/ PCR reverse primer	PCR forward primer
Poly(U) psbD-Met^{CAT}-ycf4	cDNA primer: GGGACTAGTCTCGAGAAAAAAAAAAAAAAAAA reverse primer: GACCAATTCGCATATAATATTTTAC	TATCAGTGGGAGGTTGGTTAAC
Non-poly(U) rpl36-rps13	CGAAATCCCTTCCAATTTTG	GCTCTCGAAAACGGAAATC
Non-poly(U) rpl36-rps13-rps11	GCTTTTTTTAAAGATGACTGCG	GCTCTCGAAAACGGAAATC
Non-poly(U) rps11	GCTTTTTTTAAAGATGACTGCG	TTAGCAATACAATTGCAACACTTAC
Non-poly(U) psbD	GCGACCTTGGGCTTATG	TATCAGTGGGAGGTTGGTTAAC
Non-poly(U) ycf4	CTTAAAAGCTAACGTAATGAAACTTC	CAGAATTCATACCTCAAGGGTTAG
Non-poly(U) psbD-Met^{CAT}-ycf4	CTTAAAAGCTAACGTAATGAAACTTC	TATCAGTGGGAGGTTGGTTAAC

Relationships between poly(U) tail addition and transcript editing

Previous studies have shown that for several plastid genes in peridinin dinoflagellates, transcripts that extend downstream of the poly(U) site are less extensively edited than the corresponding polyuridylylated transcript (Dang and Green, 2009). This may suggest that poly(U) tail addition is associated with transcript editing. Alternatively, editing may occur independently of poly(U) tail addition during transcript processing. In this scenario, transcripts that possess poly(U) tails may still be more highly edited than those that do not (if they represent more mature intermediates in transcript processing, and as a result of having been present in the plastid for longer periods of time have been more likely to subject to editing). However, if poly(U) tail addition is not associated with editing, there may also be transcripts present that have undergone poly(U)-independent 3' terminal cleavage events that are highly edited, or equally there may be certain polyuridylylated transcripts that are not edited.

I wished to determine whether poly(U) tail addition was directly associated with editing in *Karenia mikimotoi*. To do this, editing events were compared for different transcript processing intermediates from the *K. mikimotoi rpl36-rps13-rps11* and *psbD-Met^{CAT}-ycf4* loci (Tables 6.9, 6.10). If poly(U) tail addition is directly associated with editing, polyuridylylated transcripts should be more highly edited than non-polyuridylylated transcripts covering the same sequence. If instead editing occurs progressively during transcript processing, but is not directly dependent on poly(U) tail addition, monocistronic transcripts should be more highly edited than polycistronic transcripts, regardless of whether these transcripts possess a poly(U) tail.

Table 6.10: Editing data for the *K. mikimotoi rps13-rps11* and *psbD-Met^{CAT}-ycf4* loci.

This table presents an overview the editing events observed in polyuridylylated and non-polyuridylylated transcripts of different lengths identified over the *rpl36-rps13-rps11* and *psbD-Met^{CAT}-ycf4* loci. ".n.d." indicates that the transcript did not cover the corresponding region of sequence.

1. rpl36-rps13-rps11			Transcript sequence					
Region	Length (bp)	Editing	rpl36-rps13 poly(U)	rpl36-rps13 non-poly(U)	rpl36-rps13-rps11 poly(U)	rpl36-rps13-rps11 non-poly(U)	rps11 poly(U)	rps11 non-poly(U)
rpl36	164	total	9	9	9	3	n.d.	n.d.
		%	5.49	5.49	5.49	1.83	n.d.	n.d.
rps13	462	total	29	29	29	18	n.d.	n.d.
		%	6.28	6.28	6.28	3.90	n.d.	n.d.
intergenic	43	total	n.d.	0	0	0	n.d.	0
		%	n.d.	0.00	0.00	0.00	n.d.	0.00
rps11	732	total	n.d.	n.d.	30	4	30	4
		%	n.d.	n.d.	4.10	0.55	4.10	0.55

2. psbD-Met^{CAT}-ycf4			Transcript sequence					
Region	Length (bp)	Editing	psbD poly(U)	psbD non-poly(U)	psbD-Met^{CAT}-ycf4 poly(U)	psbD-Met^{CAT}-ycf4 non-poly(U)	ycf4 poly(U)	ycf4 non-poly(U)
5' UTR	132	total	3	0	n.d.	n.d.	n.d.	n.d.
		%	2.27	0.00	n.d.	n.d.	n.d.	n.d.
psbD	999	total	22	17	13	9	n.d.	n.d.
		%	2.20	1.70	1.30	0.90	n.d.	n.d.
Met^{CAT} intergenic	262	total	n.d.	0	0	0	n.d.	0
		%	n.d.	0.00	0.00	0.00	n.d.	0.00
ycf4	664	total	n.d.	n.d.	0	0	37	11
		%	n.d.	n.d.	0.00	0.00	6.07	1.80

Editing events were tabulated for polyuridylylated monocistronic *rps11, psbD* and *ycf4* transcripts, and polycistronic *rpl36-rps13-rps11, rpl36-rps13* and *psbD-Met^{CAT}-ycf4* transcripts by comparing the oligo-d(A) RT-PCR products and circular RT-PCR sequences obtained for each transcript, to gene sequences amplified by PCR (Tables 6.9, 6.10). To confirm that the sequences generated were correct, each transcript was resequenced twice, using oligo-d(A) primed cDNA synthesised from independently isolated RNA samples (Table 6.9). Consistent with data previously obtained for *Karenia mikimotoi* transcripts, the polyuridylylated *psbD, ycf4, rpl36-rps13* and *rpl36-rps13-rps11* transcripts were extensively edited (Table 6.10) (Dorrell and Howe, 2012a). Between 2.3 and 6.3% of the residues for each CDS were edited. Editing sites were much less frequent in non-coding sequence, and could not be identified on the *Met^{CAT}* tRNA gene between *psbD* and *ycf4* (Table 6.10).

To quantify editing on non-polyuridylylated transcripts, cDNA was synthesised using primers that were positioned downstream of *psbD, ycf4, rps13* and *rps11* poly(U) sites. PCRs were

performed using the cDNA synthesis primer as a reverse primer, and the same PCR forward primers previously used for oligo-d(A) primed RT-PCR of each gene (Table 6.9). As before, each transcript was sequenced three times, using RNA isolated from separate RT-PCR was repeated three times, and the assembled sequences of these transcripts, alongside the terminal regions of non-polyuridylylated transcripts identified by circular RT-PCR, were compared to the gene sequences as before (Table 6.10).

Many of the editing events found on polyuridylylated transcripts were not found on the corresponding non-polyuridylylated transcripts. Only four of the thirty sites (13.3%) within the *rps11* CDS that were edited on polyuridylylated transcripts were also edited on transcripts that extended past the *rps11* poly(U) site (Table 6.10). Similar differences in editing state were observed for polyuridylylated versus non-polyuridylylated *psbD* and *ycf4* transcripts (Table 6.10). Thus, for *rps11, psbD,* and *ycf4,* poly(U) tail addition is associated with the completion of transcript editing. In contrast, there were no differences in editing between polyuridylylated *rpl36-rps13* transcripts, and *rpl36-rps13* transcripts identified by direct or by circular RT-PCR to extend downstream of the *rps13* poly(U) site (Table 6.10). Across all of the genes studied, no editing events were found specifically on non-polyuridylylated transcripts, and not found on the corresponding polyuridylylated transcript sequence (Table 6.10).

Different patterns of editing were observed on polycistronic transcripts from the *rpl36-rps13-rps11* and *psbD-Met^{CAT}-ycf4* loci. The polyuridylylated *rpl36-rps13-rps11* transcripts appeared to be edited to completion, showing identical patterns of editing to lower molecular weight transcripts (Table 6.10). This indicates that editing at the *rpl36-rps13-rps11* locus occurs on polycistronic transcripts covering all three genes. In contrast, polycistronic transcripts within the *psbD-Met^{CAT}-ycf4* locus were less extensively edited than either the corresponding polyuridylylated or non-polyuridylylated *psbD* and *ycf4* transcripts, suggesting that editing is associated with transcript cleavage (Table 6.10). No editing events at all were detected within the *ycf4* region of polycistronic *psbD-Met^{CAT}-ycf4* transcripts, even for the transcript amplified from oligo-d(A) cDNA (Table 6.10). This was confirmed in three independent replicates of the *psbD-Met^{CAT}-ycf4* RT-PCRs, using separately isolated RNA samples. Thus, editing of *ycf4* is specifically associated with monocistronic transcripts, whereas polycistronic *ycf4* transcripts are not edited even if they are polyuridylylated. Overall, there are complex relationships between poly(U) tail addition and editing in fucoxanthin plastids. Although for certain genes (*rps11, psbD*) poly(U) tail addition is connected to editing, for others (*rps13*) it is not. In certain cases, (*ycf4*) editing may be

Table 6.11. Primers for RT-PCRs to detect antisense plastid transcripts in *Karenia mikimotoi*.

This table lists the primers used to (i) identify antisense transcripts and (ii) confirm specificity of the antisense transcript cDNA primers used. Sense cDNA template sequences (used as positive controls for the cDNA specificity tests) were generated using primer (2) of the corresponding gene.

i. cDNA synthesis primers

Gene	Primer 1- Antisense transcripts	Primer 2- Sense transcripts
psbA	GCTATCAGGCTCACTTTTATATGC	CCATCGTAGAAACTCCCATAG
psbD	TATCAGTGGGAGGTTGGTTAAC	GTTTTCATGAGGTTGATCTTGG
psaA	CACGTAGTTCAGCTCTGATACC	CACGTTGTGCCAATTCC
rbcL	GATGCGTATGGCAGGTG	GTTGATCATCTGGAGTATCGTTG
ycf4	CAGAATTCATACCTCAAGGGTTAG	GACCAATTCGCATATAATATTTTAC
rps13	GCTCTCGAAAACGGAAATC	GTTGTCCTCGAGTTGGAAG
rps11	TTAGCAATACAATTGCAACACTTAC	CCCCAACCTGCACCAG

ii. Primers to confirm the specificity of the antisense cDNA synthesis primer

	Primer 3- upstream PCR forward primer	Primer 4- PCR reverse primer
psbA	ATCACAGCAGACAACACCCG	TACCCCCATTGTAAAGCC
psbD	ACGACTGGCTAAAACGAGAC	AAAATATTAGCTATGTTTATTCAAGTACAAC
psaA	GCCGGTCTAGTTCTAGCAG	CACGTTGTGCCAATTCC
rbcL	GCGGAGTTAGAAAGCCC	GTTGATCATCTGGAGTATCGTTG
ycf4	TGGTAATGGTCTCTAACACGTC	GACCAATTCGCATATAATATTTTAC
rps13	CTTTTAGGATAAAATATCAAGGTTACAAC	GTTGTCCTCGAGTTGGAAG
rps11	ATCGTTTACGAAGCGAACTC	CCCCAACCTGCACCAG

dependent both on the addition of a poly(U) tail and the cleavage of polycistronic precursors into monocistronic mRNAs.

Antisense transcripts are present in fucoxanthin plastids

Previously, I have shown that antisense transcripts containing regions of minicircle are present in peridinin dinoflagellates, similar to the antisense transcripts proposed to be produced in the plastids of plants (Georg *et al.*, 2010; Hotto *et al.*, 2012) and apicomplexans (Bahl *et al.*, 2010; Kurniawan, 2013). I wished to determine whether antisense transcripts were likewise present in fucoxanthin dinoflagellate plastids.

A series of RT-PCRs were performed to detect antisense transcripts of seven genes (*psbA, psbD, psaA, rbcL, rps13, rps11, ycf4*). cDNA was generated using primers with the same sequence as the non-template strand of each gene, i.e. the same sequence as the sense transcript, and complementary to the antisense transcript of each gene (Table 6.11; Fig. 6.4,

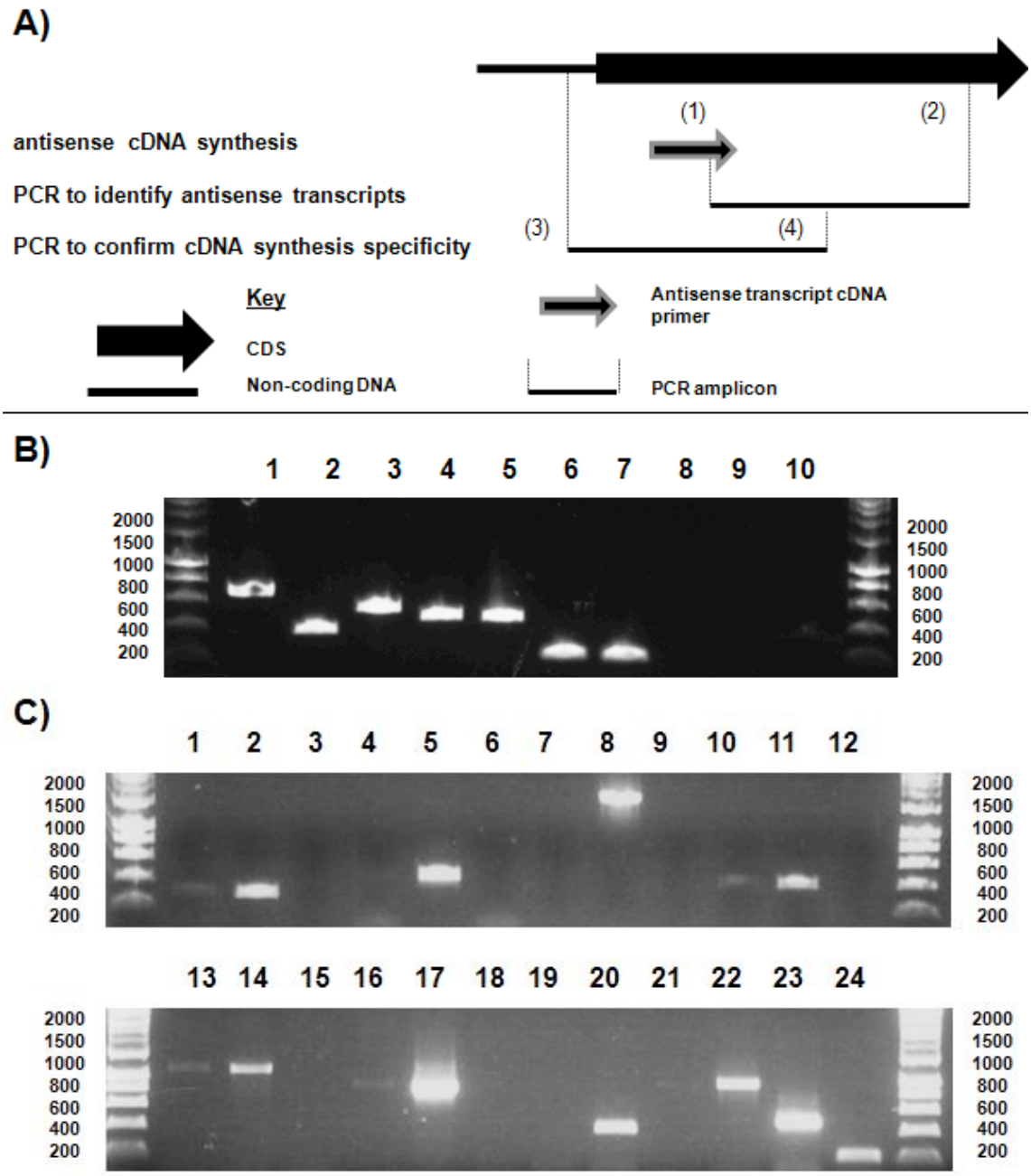


Fig. 6.4. RT-PCR identification of antisense plastid transcripts in *Karenia mikimotoi*.

Panel A shows a diagram of the RT-PCRs performed to identify antisense transcripts in *K. mikimotoi*. Antisense transcripts were detected using a cDNA synthesis primer designed against the non-template strand of plastid sequence (primer 1), and a complementary PCR reverse primer (primer 2). To confirm that the cDNA synthesis primers was specific to antisense transcripts, the cDNA template was amplified using PCR primers flanking the predicted cDNA synthesis primer annealing site (primers 3, 4). If the cDNA synthesis primer was specific to antisense transcripts, products would not be obtained (as PCR primer 3 is positioned upstream of the cDNA synthesis site), whereas products should be obtained in reactions using cDNA templates generated with primers specific to sense transcripts (e.g. primers 2, 4).

Fig. 6.4 (continued)

Panel B shows the results of RT-PCRs to detect antisense transcripts for seven plastid genes. Lanes **1-7**: RT-PCRs for antisense *psbA, psbD, psaA, rbcL, ycf4, rps13, rps11*. Lanes **8-10**: template negative controls for *psbA, ycf4, rps11*.

Panel C shows the results of RT-PCRs to confirm specificity of cDNA synthesis. Lanes **1-3**: PCR using primers flanking the predicted *psbA* antisense cDNA synthesis site, and antisense (**1**) and sense (**2**) cDNA templates, and template negative conditions (**3**). Lanes **4-6**: the same reactions, for *psbD*; **7-9**: *psaA***; 10-12**: *rbcL*; **13-15**: *ycf4*; **16-18**: *rps13*; **19-21**: *rps11*. Lanes **22-24**: RT-PCRs for antisense *psbA, ycf4, rps11* transcripts using PCR primers positioned downstream of the cDNA synthesis site, as in Panel B (antisense transcript positive controls).

panel A). Each cDNA synthesis primer was confirmed by BLAST not to be similar to any sequence identified on the template strand of the corresponding gene, thus should preferentially anneal to antisense transcripts. PCRs were then performed using cDNA generated with each synthesis primer, and PCR primers positioned within each gene, downstream of the cDNA synthesis site (Table 6.11; Fig. 6.4, panel A).

For every gene tested, products were identified (Fig. 6.4, panel B). To confirm that these products specifically corresponded to antisense transcripts (rather than the result of the cDNA synthesis primer annealing promiscuously to sense transcripts), and additional PCR was performed for each gene, using the same cDNA template as previously used to amplify antisense transcripts, and a PCR forward primer positioned upstream of the antisense transcript cDNA synthesis site (Table 6.11; Fig. 6.4, panel A). If the cDNA synthesis primer had promiscuously annealed to sense strand transcripts of the gene, products would be detected, whereas products would not be if the cDNA primer was specific to antisense transcripts (Fig. 6.4, panel A). Each of the reactions performed with antisense cDNA templates, generated negligible products (Fig. 6.4, panel C). In contrast, RT-PCRs using each combination of PCR primers, and cDNA synthesised with a primer similar to the template strand of the gene (which would anneal to sense transcripts) identified abundant products (Fig. 6.4, panel C). Although for a few of the genes tested (e.g. *psbA, ycf4*) generated low abundance products in the antisense cDNA amplification, which may be the result of promiscuous annealing, these products were much less abundant than the corresponding products from the sense cDNA amplification, or the antisense transcripts

Table 6.12. Primers used for 5' RACE of *Karenia mikimotoi* antisense transcripts.

RNA adapter	GCUGAUGGCGAUGAGCACUGGGUUGCAA	
Adapter-specific PCR primer 1	GCTGATGGCGATAGC	
Adapter-specific PCR primer 2	GATGAGCACTGGGTTGC	
Gene	**rps13**	**rps11**
cDNA synthesis primer	GCTCTCGAAAACGGAAATC	TTAGCAATACAATTGCAACACTTAC
Gene-specific PCR primer 1	GCTCTCGAAAACGGAAATC	ACGAGGTGGAATACTAAAGAGG
Gene-specific PCR primer 2	CGGAAGCGGTATTAAGGC	CCGTCGAAGACAACATTCTTAG
Gene	**psbD**	**ycf4**
cDNA synthesis primer	TTGAACTAGGAGGCTTGTGG	CAGAATTCATACCTCAAGGGTTAG
Gene-specific PCR primer 1	GCTATTCACGGAGCGAC	AAAACTAACGGTACATAATTATGCTAGAC
Gene-specific PCR primer 2	CAAACGGTGGTTACACTTCTTC	GCTCAGTTAGCCAATGGG

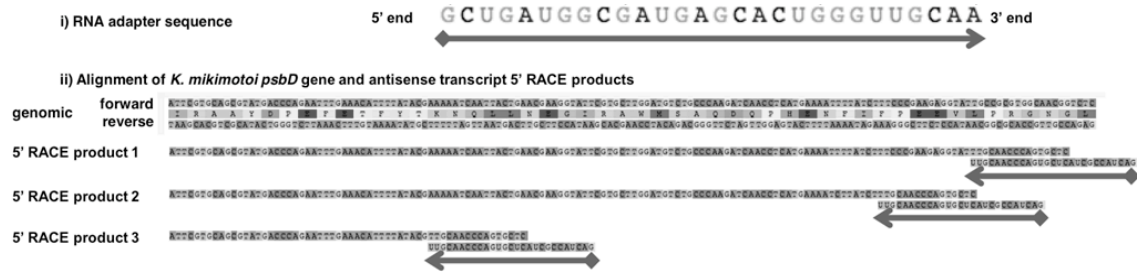


Fig. 6.5. 5' RACE of *Karenia mikimotoi* plastid antisense transcripts.

This figure shows the RNA adapter sequence used for 5' RACE **(i)**, and an alignment of the 3' end of the *K. mikimotoi psbD* gene, and three sequences obtained using the 5' RACE protocol using primers specific to antisense *psbD* transcripts **(ii)**. Each sequence aligns with the *psbD* template strand, followed by a region that aligns with the 3' end of the RNA adapter used (shown below each product in reverse complemented form). The ligation site confirms that the transcript terminus identified is a 5' end, and given the orientation of this terminus relative to the *psbD* CDS must be derived from an antisense transcript.

previously identified using PCR primers positioned downstream of the antisense cDNA synthesis site, suggesting that they only represent a minority of the cDNA templates generated by the corresponding cDNA synthesis primer (Fig. 6.4, panel C; compare lanes 1, 2, 22; lanes 13, 14, 23). Thus, the highly abundant products visible in Fig. 6.4 correspond to plastid antisense transcripts.

To generate an independent line of evidence for the presence of plastid antisense transcripts RNA ligase-mediated 5' RACE was performed, as previously described for antisense transcripts in *Amphidinium carterae*. Using the same cDNA synthesis primers as before, and

Table 6.13. Primers for RT-PCRs to detect polyuridylylated antisense transcripts.

Primers used for the RT-PCRs shown in Panel A of Fig. 6.6 are shown in bold text.

oligo-d(A) GGGACTAGTCTCGAGAAAAAAAAAAAAAAAAA

Gene	PCR reverse primer	Gene	PCR reverse primer
atpA	GAAGAAGCATGTCGTCGC	rpl16	GTAATTTATGCGAAGCTAATCG
atpB	CGCAGGGACGTATATTGC	rpl2	CCCCAATGCAACTTTACC
atpH	ACACAATGCAACAACAAGACC	rpl22	CTGAAGTGCTTTCCCCG
atpI	AGATTCAGCAATGTACGAACAAG	rpl23	CTATTTTCGCATCACCTGC
cbbX	GCCAAATAGCGGACGTAGAG	rpl3	GGCAACGAACCTTTGAGG
chlI	AGAACGGGAGACCTGGG	rpl31	GTCGTGATCCCAACCG
clpC	AAAGCCGGGGTGAGTAAG	rpl36	GTAAAGTCGCCCTCTTCG
dnaK	GCATCCAATGTAGCCCG	rpl5	CGCTGATGACGACGAG
groEL	GTAGACGCATCGTAGCCAC	rpl6	CGAAGCAAAGTGACCTACCC
petA	GCAGAAGGCGTACCTAACG	rpoA	GGTAGCGGTTGGAGTTG
petB	TCCACCACGAAGTAACGC	rpoB	GGTATCCCCGGTTTTGG
petD	GCCGAAGCAGAAATCAAC	rpoC1	CCAGTACTCGGCGACC
psaA	**GTAGGGAAGCAGGTGTTGG**	rpoC2	CGAACCCAAACGAAGG
psaB	AATGGAACCAACCTGCG	rps10	AAACTGGTCTCGGGAGG
psaC	CCTGCTAATACGCCAGACC	**rps11**	**GCAATTGTATTGCTAAAGTTAGCTAATATATG**
psaD	GAGGCGAGCGCATTC	rps12	CGATCTTTCACCGGCAC
psaF	ACGGTTGTAGAAGCCTTCC	**rps13**	**GTTGTCCTCGAGTTGGAAG**
psaL	TTTGTCACTTTCTGCGTCAG	rps14	CCTAACTACCAACTTGAGCG
psbA	**TACCCCCATTGTAAAGCC**	rps17	TCGTGGTTGCGCTTG
psbB	CACCCTTGTGCGATACC	rps19	GTATAGTTGAGCTCCGTGACC
psbC	CCAGCGCCTAGAACGG	rps3	CGTTCCAGTAATTGCGC
psbD	**AAAATATTAGCTATGTTTATTCAAGTACAAC**	rps4	GACAAGGCGAACAAAACC
psbE	GCACCAAAACGTTCGG	rps5	AACAAAACCGCTCGTGC
psbH	GTTGATCCCCAGGCAG	rps7	CAGCAGCAGCTACAATCC
psbI	AACATACCTTACTCTATAGCCTTTCG	rps8	CGACCTCCTCCATAGGC
psbL	GTCTGACACACTCTTAGTTCAAAAAAATAAC	secA	AGCTGTCGACTTCGCTCC
psbN	GGAAAGGATCGCGGAG	secY	CCAGCTATCACGACCCC
psbT	TCGCAATTCTTGGGCTATC	tatC	CTTGGACAAAGCAGGGG
psbV	TCCACCCCATTTTTCACC	tufA	CAGTAACAGTCCCAGCCC
psbX	GATCCTATCGAGAGAGCTAACCGG	ycf3	CGTAACCATAACCGCGTG
rbcL	**GTTCCCGCATGGATATG**	ycf39	CTCAGACGACGGTGTAGC
rbcS	GGTTCTCCCACGTGCTTC	**ycf4**	**GAGAAATAATCCTAATATTATTCCGATG**
rpl14	TTCGCGGTGTGCTTG		

combinations of PCR primers specific to antisense transcripts for each gene, sequences were amplified that were similar to the template strands of the *psbD, ycf4, rps13* and *rps11* genes that terminated in a 5' end adaptor ligation site (Table 6.12; Fig. 6.5). None of the adaptor ligation sites corresponded to regions of genomic sequence similar to either adaptor PCR primer, and similar products could not be identified in control 5' RACE reactions performed without T4 RNA ligase, indicating that these products were not the result of promiscuous hybridisation of the adaptor PCR primers to *cis*-encoded sequence in each

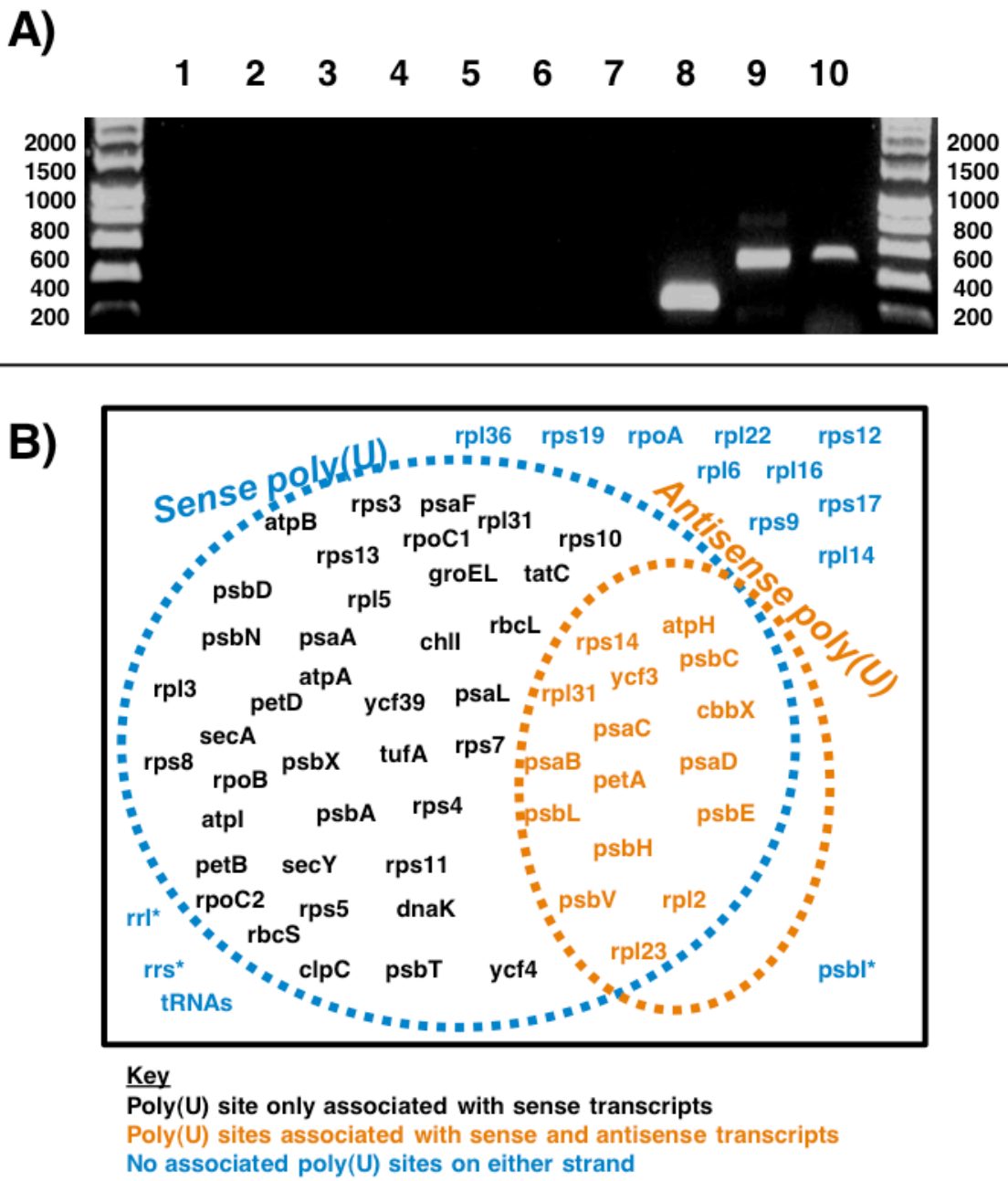


Fig. 6.6: Absence of poly(U) tails from antisense *K. mikimotoi* plastid transcripts.

Panel A shows the gel photograph of a series of RT-PCRs performed with oligo-d(A) cDNA to detect polyuridylylated sense and antisense transcripts of seven genes (*psbA, psbD, psaA, rbcL, ycf4, rps13, rps11*) in the *Karenia mikimotoi* plastid. Lanes **1-7**: PCRs performed with an oligo-d(A) cDNA template and PCR primer, and PCR primers with the same sequence as the template strands of seven genes (*psbA, psbD, psaA, rbcL, ycf4, rps13, rps11*), indicating that polyuridylylated antisense transcripts are absent. Lanes **8-10**: RT-PCRs performed with oligo-d(A) cDNA and an oligo-d(A) PCR primer together with PCR primers with the same sequence as the non-template strands of three genes (*psbA, psbD, rps13*),confirming the presence of polyuridylylated sense transcripts.

Fig. 6.6 (continued)

Panel B shows the number of genes identified across the entire *K. mikimotoi* plastid to give rise to polyuridylylated sense and antisense transcripts. Genes are colour-coded according to the identification of poly(U) sites associated with sense or antisense transcripts. Genes marked with an asterisk are ones that did not possess a poly(U) site in the associated 3' UTR of the non-template strand, but were identified as part of a polyuridylylated polycistronic sense transcript. These data were obtained with the assistance of an undergraduate student, George Hinksman.

gene (Fig. 6.5). Thus, the 5' RACE products correspond to the 5' ends of plastid antisense transcripts (Fig. 6.5).

Strand-specific transcript poly(U) tail addition in fucoxanthin dinoflagellates

Previously, I have shown that poly(U) tails are specifically associated with sense transcripts in peridinin dinoflagellates, and are not added to antisense transcripts. This may indirectly allow antisense transcripts to be recognised during processing, and to be removed from transcript pools, similarly as has been documented to occur in plant plastids (Sharwood *et al.*, 2011). I wished to determine whether poly(U) tail addition was likewise a strand-specific transcript processing event in fucoxanthin dinoflagellates.

To test for polyuridylylated antisense transcripts, cDNA was generated using an oligo-d(A) cDNA synthesis primer as before. PCRs were then performed using the same oligo-d(A) primer, and primers with the same sequence as the template strands of genes from the *Karenia mikimotoi* plastid (Table 6.13). Seven genes (*psbA, psbC, psbD, rbcL, ycf4, rps13,* rps11) were initially tested. None of these genes was found to give rise to polyuridylylated antisense products (Fig. 6.6, panel A; lanes 1-7). Polyuridylylated sense *psbA, ycf4* and *rps11* transcripts could be amplified from the same oligo-d(A) cDNA preparations, using PCR primers with the same sequence as the non-template strand of each gene, confirming that the cDNA synthesis reaction had been successful (Fig. 6.6, panel A; lanes 8-10).

The initial RT-PCR was extended, using primers designed against the template strands of every gene identified in the *K. mikimotoi* plastid (Table 6.13). 52 of the 68 genes failed to give any products in the antisense oligo-d(A) RT-PCR (Fig. 6.6, panel B). The remaining 16 genes did yield products, but these were low in abundance and generally could not be identified in independent replicates of the same PCR (Fig. 6.6, panel B). No polycistronic

Table 6.14. Antisense *rpl36-rps13-rps11* and *psbD-Met^{CAT}-ycf4* transcript termini as identified by circular RT-PCR and 5' RACE.

This table lists the 5' and 3' termini of antisense transcripts from the *rpl36-rps13-rps11* and *psbD-Met^{CAT}-ycf4* loci, as per Table 6.7. The terminus positions of the antisense transcripts are shown relative to the terminus positions of the corresponding CDS. Note that as the antisense transcripts are in opposing orientation to the CDS, the antisense transcript 5' terminus is given relative to the 3' terminus of the CDS, and vice versa. The PCR primers used to identify each circular RT-PCR product correspond to those listed in Table 6.6; primers used to amplify 5' RACE products are listed in Table 6.12. Unless specifically noted, the antisense termini identified terminate within the CDS of the gene, and do not extend through residues complementary to the poly(U) site.

	Transcript dimensions				PCR primers		
1. rpl36-rps13-rps11	**5' end**	**3' end**	**Poly(U)**	**Length**	**R**	**F**	**Notes**
rps13 circular RT-PCR							
Antisense transcript 1	14	-51	0	618	F3	R2	3' end extends into rpl36
Antisense transcript 2	14	-51	0	618	F3	R2	3' end extends into rpl36
rps11 circular RT-PCR							
Antisense transcript 1	19	-497	0	953	F6	R3	3' end extends through rps13 poly(U) site
Antisense transcript 2	19	-497	0	953	F6	R3	3' end extends through rps13 poly(U) site
Antisense transcript 3	19	-493	0	957	F6	R3	3' end extends through rps13 poly(U) site
Antisense transcript 4	19	-493	0	957	F6	R3	3' end extends through rps13 poly(U) site
5' RACE							
Antisense rps13 transcript 5' end 1	331	n/a	n/a	n/a	n/a	n/a	5' end extends through rps13 poly((U) site into rps11
Antisense rps11 transcript 5' end 1	252	n/a	n/a	n/a	n/a	n/a	5' end extends through rps11 poly(U) site
	Transcript dimensions				**PCR primers**		
2. psbD-Met^{CAT}-ycf4	**5' end**	**3' end**	**Poly(U)**	**Length**	**R**	**F**	**Notes**
psbD circular RT-PCR							
Antisense transcript 1	316	22	0	1292	F1	R2	5' end extends through poly(U) site into ycf4
Antisense transcript 2	-12	149	0	837	F1	R2	
Antisense transcript 3	-34	-124	0	1088	F1	R2	
Antisense transcript 4	-45	103	0	850	F1	R2	
Antisense transcript 5	-94	45	0	859	F1	R2	
Antisense transcript 6	-100	-25	0	923	F1	R2	
ycf4 circular RT-PCR							
Antisense transcript 1	-56	-17	0	624	F7	R8	
Antisense transcript 2	-56	-21	0	628	F8	R8	
Antisense transcript 3	-56	-21	0	628	F7	R8	
Antisense transcript 4	-56	-21	0	628	F7	R8	
5' RACE							
Antisense psbD transcript 5' end 1	-19	n/a	n/a	n/a	n/a	n/a	
Antisense psbD transcript 5' end 2	-33	n/a	n/a	n/a	n/a	n/a	
Antisense psbD transcript 5' end 3	-108	n/a	n/a	n/a	n/a	n/a	
Antisense psbD transcript 5' end 4	-108	n/a	n/a	n/a	n/a	n/a	
Antisense ycf4 transcript 5' end 1	72	n/a	n/a	n/a	n/a	n/a	5' end extends through ycf4 poly(U) site
Antisense ycf4 transcript 5' end 2	-18	n/a	n/a	n/a	n/a	n/a	

antisense transcripts were identified through this approach. Across the entire plastid genome, the total frequency of poly(U) sites associated with antisense transcripts was significantly lower than for sense transcripts (chi-squared, P< E^-12). Thus, antisense transcripts in fucoxanthin plastids generally do not receive poly(U) tails.

I have previously shown that antisense transcripts in peridinin dinoflagellates do not terminate at positions complementary to the non-template strand poly(U) site. This indicates that the cleavage of the poly(U) site is a feature specifically associated with the processing of sense transcripts. I wished to determine whether a similar situation were true for fucoxanthin dinoflagellates. To do this, circular RT-PCRs were performed to identify antisense transcripts from the *psbD-Met*^{CAT}*-ycf4* and *rpl36-rps13-rps11* loci. cDNA was synthesised using the same primers used to initially identify antisense transcripts of each gene, and the same combinations of PCR primers used for circular RT-PCRs of sense transcripts at each locus (Tables 6.7, 6.11).

None of the antisense transcripts identified through circular RT-PCR possessed poly(U) tails or any other form of 3' terminal modification (Table 6.14). Antisense transcripts were identified that terminated within the CDS, or within the 3' UTR of each gene, but did not extend to the non-template strand poly(U) site (Table 6.14). Antisense transcripts were additionally identified, either through circular RT-PCR or through the previous 5' RACE reactions, which extended through residues complementary to the non-template strand poly(U) site of all four genes (Table 6.14). However, no antisense transcripts were identified either through circular RT-PCR or 5' RACE that terminated at a position complementary to a poly(U) site at either end (Table 6.14). Thus, poly(U) tail addition, and processing of the poly(U) site are specifically associated with sense transcripts in fucoxanthin dinoflagellate plastids. The specific addition of poly(U) tails to sense transcripts might allow them to be discriminated from antisense transcripts during transcript processing.

Sense and antisense transcripts undergo complementary editing events

Previous studies have shown that some antisense transcripts in plant plastids extend through residues that correspond to editing sites on the sense transcripts (Georg *et al.*, 2010). It has not, however, been determined in these cases whether the antisense transcripts are themselves edited.

All of the antisense transcripts sequenced in this study contained evidence for extensive sequence editing (Fig. 6.7). Editing was detected in antisense transcripts amplified by direct RT-PCR, circular RT-PCR, 5' RACE, and in the small number of antisense transcripts

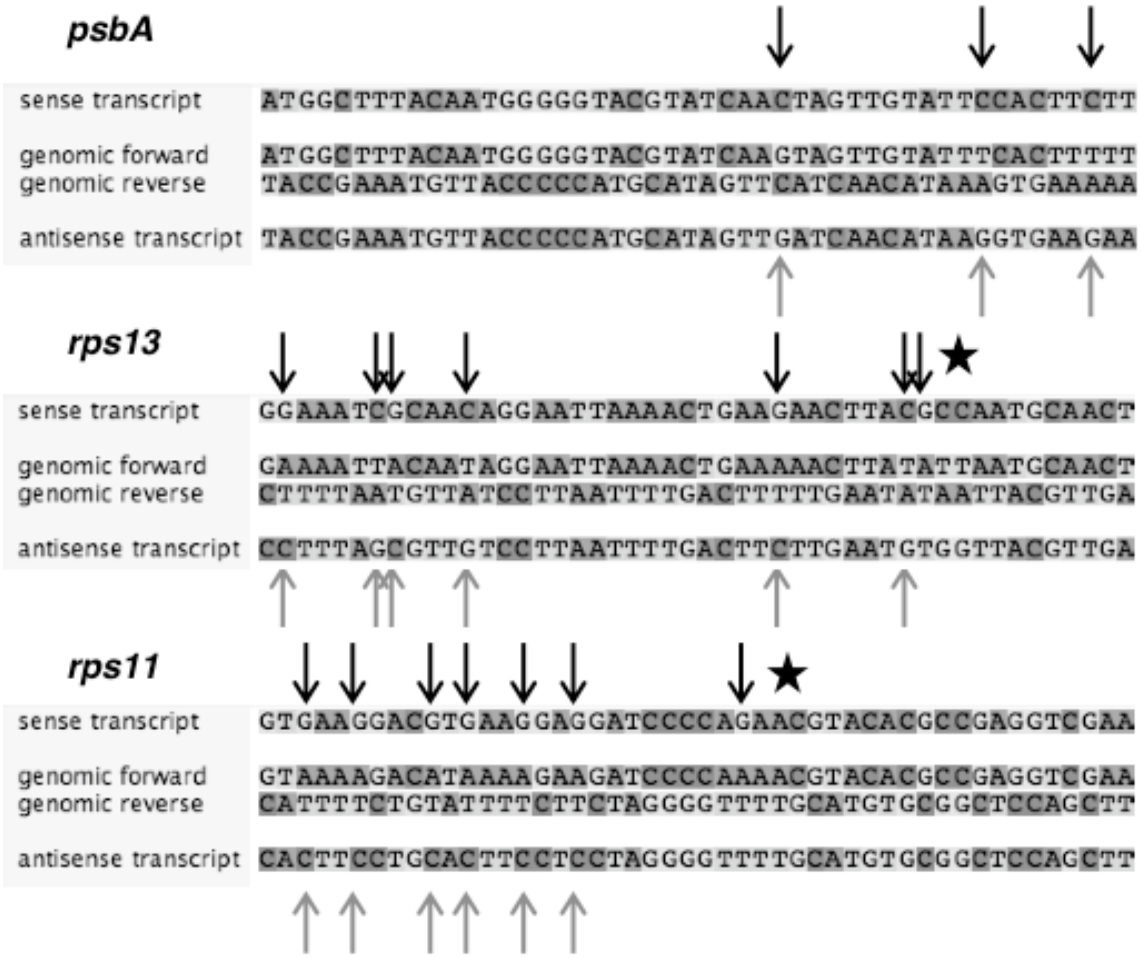


Fig. 6.7: Complementary editing of sense and antisense transcripts.

This figure shows alignments of 50nt regions of sense and antisense *psbA*, *rps13* and *rps11* transcript sequences against the corresponding gene sequences. Arrows correspond to editing sites in each transcript sequence. In each case, broadly complementary patterns of editing were observed between the sense and the corresponding antisense transcripts. Asterisks label one editing site on *rps13*, and one editing site on *rps11*, that are specific to sense transcripts.

detected by oligo-d(A) primed RT-PCR. The editing observed in each antisense transcript was complementary to that of the corresponding sense transcripts, For example, an A to G editing event on a sense transcript would be matched by a U to C editing event at the complementary site on the antisense transcript (Fig. 6.7). A few of the antisense transcripts were not edited at all of the sites identified as edited in the corresponding polyuridylylated sense transcript (Fig. 6.7).

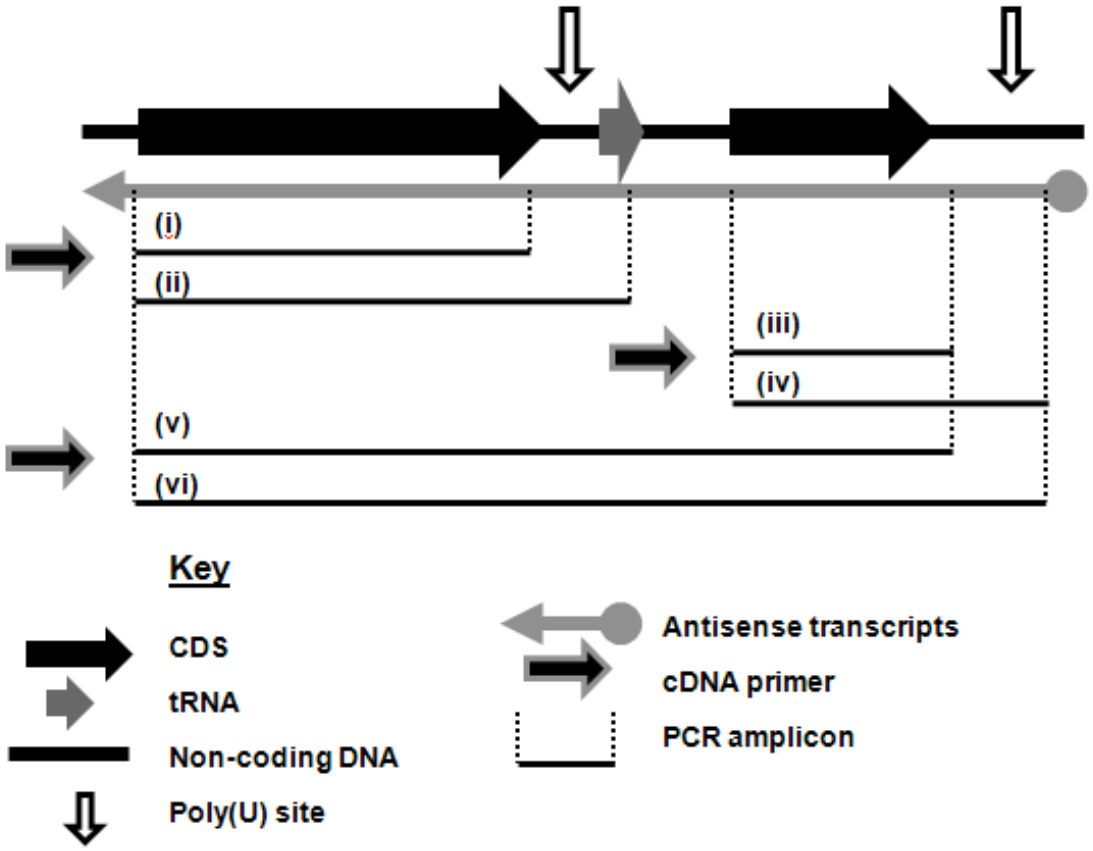


Fig. 6.8: RT-PCRs to identify the relationships between editing and cleavage of antisense transcripts.

This schematic diagram shows six RT-PCRs used to determine the relationship between editing and cleavage of antisense transcripts over the *rpl36-rps13-rps11* and *psbD-Met*^{CAT}*-ycf4* loci. For example, for the *psbD-Met*^{CAT}*-ycf4* locus, antisense transcripts were amplified from cDNA preparations using cDNA synthesis primers derived from the non-template strands of the *psbD* and *ycf4* genes. Antisense *psbD* transcripts were then amplified using the *psbD* antisense cDNA synthesis primer, and a complementary PCR primer positioned either within the 3' end of the *psbD* CDS **(i)**, or within the *Met*^{CAT} region, after the *psbD* poly(U) site **(ii)**. Antisense *ycf4* transcripts were identified using the *ycf4* antisense cDNA synthesis primer, and complementary PCR primers positioned within the 3' end of the *ycf4* CDS **(iii)**, or after *ycf4* poly(U) site **(iv)**. Finally, antisense transcripts covering the complete *psbD-Met*^{CAT}*-ycf4* locus were identified using the *psbD* cDNA primer, and complementary *ycf4* PCR primers as before **(v, vi)**. Similar RT-PCRs were used to sequence different antisense transcripts covering the *rpl36-rps13-rps11* locus.

I wished to determine whether individual nucleotides were edited in the same order on sense and antisense transcripts. The *rpl36-rps13-rps11* and *psbD-Met*^{CAT}*-ycf4* loci, for which

Table 6.15. Editing states of different terminal processing intermediates of antisense *rpl36-rps13-rps11* and *psbD-Met^{CAT}*-ycf4 transcripts

The first table lists the different combinations of PCR primers used to identify different regions of antisense transcripts from the *rpl36-rps13-rps11* and *psbD-Met^{CAT}-ycf4* loci. PCR reactions are numbered corresponding to those shown in Fig. 6.8. The second table compares specific relationships between editing and terminal processing observed for sense and antisense transcripts over each locus. For both sense and antisense transcripts, cleavage of the *rps11*, *psbD* and *ycf4* 3' UTR is associated with editing, whereas cleavage of the *rps13* UTR is not. In addition, polycistronic sense and antisense *psbD-Met^{CAT}-ycf4* locus are only edited within the *psbD* region.

1. PCR primers

	rpl36-rps13-rps11	cDNA primer	PCR forward primer
i	**rps13-rpl36 (5' end in CDS)**	GCTCTCGAAAACGGAAATC	GTTGTCCTCGAGTTGGAAG
ii	**rps13-rpl36 (5' end in UTR)**	GCTCTCGAAAACGGAAATC	CGAAATCCCTTCCAATTTTG
iii	**rps11 (5' end in CDS)**	TTAGCAATACAATTGCAACACTTAC	CCCCAACCTGCACCAG
iv	**rps11 (5' end in UTR)**	TTAGCAATACAATTGCAACACTTAC	GCTTTTTTTAAAGATGACTGCG
v	**rps11-rps13-rpl36 (5' end in CDS)**	GCTCTCGAAAACGGAAATC	CCCCAACCTGCACCAG
vi	**rps11-rps13-rpl36 (5' end in UTR)**	GCTCTCGAAAACGGAAATC	GCTTTTTTTAAAGATGACTGCG

	psbD-Met^{CAT}-ycf4	cDNA primer	PCR forward primer
i	**psbD (5' end in CDS)**	TATCAGTGGGAGGTTGGTTAAC	GTTTTCATGAGGTTGATCTTGG
ii	**psbD (5' end in UTR)**	TATCAGTGGGAGGTTGGTTAAC	GCGACCTTGGGCTTATG
iii	**ycf4 (5' end in CDS)**	CAGAATTCATACCTCAAGGGTTAG	GACCAATTCGCATATAATATTTTAC
iv	**ycf4 (5' end in UTR)**	CAGAATTCATACCTCAAGGGTTAG	CTTAAAAGCTAACGTAATGAAACTTC
v	**ycf4-Met^{CAT}-psbD (5' end in CDS)**	TATCAGTGGGAGGTTGGTTAAC	GACCAATTCGCATATAATATTTTAC
vi	**ycf4-Met^{CAT}-psbD (5' end in UTR)**	TATCAGTGGGAGGTTGGTTAAC	CTTAAAAGCTAACGTAATGAAACTTC

2. editing events

	sense	antisense
Editing events associated with 3' UTR cleavage		
rps11 transcripts that terminate within the CDS (iii) versus UTR (iv)	30 (iii); 4 (iv)	27 (iii); 4 (iv)
psbD transcripts that terminate within the CDS (i) versus UTR (ii)	22 (i); 17 (ii)	20 (i); 13 (ii)
ycf4 transcripts that terminate within the CDS (iii) versus UTR (iv)	37 (i); 11 (ii)	25 (i); 4 (ii)
Editing events independent of 3' UTR cleavage		
rpl36 region of rpl36-rps13 transcripts that terminate within the CDS (i) versus UTR (ii)	38 (i); 38 (ii)	36 (i); 36 (ii)
Asymmetric editing of psbD-Met^{CAT}-ycf4		
psbD and ycf4 regions of psbD-Met^{CAT}-ycf4 transcripts that terminate within the CDS (v)	13 (psbD); 0 (ycf4)	10 (psbD); 0 (ycf4)
psbD and ycf4 regions of psbD-Met^{CAT}-ycf4 transcripts that terminate within the UTR (vi)	9 (psbD); 0 (ycf4)	7 (psbD); 0 (ycf4)

I had already obtained detailed editing data regarding the processive order of editing on sense transcripts, were selected as model systems for exploring the progression of editing

on antisense transcripts. Antisense transcripts were amplified from each locus using the same cDNA preparations as before, and different combinations of PCR primers, to detect different cleavage intermediates (Fig. 6.8; Table 6.15). For example, antisense transcripts were amplified that extended through the poly(U) sites of each gene using a PCR reverse primer positioned past the poly(U) site in the 3' UTR of the non-template strand (Fig. 6.8; Table 6.15). Antisense transcripts covering more than one gene were amplified using a similar approach (Fig. 6.8; Table 6.15). Each RT-PCR was repeated three times with cDNA generated from independently isolated RNA samples, and the consensus sequence for each reaction was compared to the corresponding genomic sequence to infer editing.

Similar relationships were identified between editing and transcript cleavage for both sense and antisense transcripts (Table 6.15). In the same way that polyuridylylated and sense transcripts are more highly edited than non-polyuridylylated equivalents, antisense *rps11, psbD* and *ycf4* transcripts that extended into the non-template strand 3' UTR were less highly edited than antisense transcripts amplified using combinations of PCR primers internal to the CDS (Table 6.15). In contrast, both sense and antisense transcripts that extend past the *rps13* poly(U) site are highly edited, indicating that cleavage of the *rps13* 3' UTR is not specifically associated with editing of either strand (Table 6.15). Notably, antisense transcripts that cover the entire *psbD-Met*^{CAT}*-ycf4* locus were only edited in the region complementary to *psbD* and not edited at all in the *ycf4* region, similarly to the polycistronic sense transcripts previously amplified for this locus (Table 6.15). Thus, editing events on sense and antisense transcripts in the *K. mikimotoi* plastid occur in the same order. This might indicate that sense and antisense transcripts are processed together, with individual editing events on one strand being matched by complementary events on the other.

Discussion

I have investigated the role of transcript poly(U) tail addition in plastid gene expression in the fucoxanthin dinoflagellate *Karenia mikimotoi*. I used a novel next generation sequencing technique, using double-stranded cDNA synthesised with an oligo-d(A) primer, to identify polyuridylylated transcripts of plastid origin. The *Karenia mikimotoi* plastid transcriptome, and thus the underlying genome, is highly divergent from that of the related fucoxanthin dinoflagellate *Karlodinium veneficum*. Over one in ten of the 68 genes inferred from the transcriptome generated to be located within the *Karenia mikimotoi* plastid have been lost from the *Karlodinium veneficum* plastid genome (Gabrielsen *et al.*, 2011; Richardson *et al.*, 2014), and it is likely that other genes have been lost independently from the *Karenia mikimotoi* plastid (Fig. 6.1; Table 6.4). This is surprising as fucoxanthin dinoflagellates are believed to have diverged from each other less than 250 million years ago, based on

molecular and fossil estimates, and potentially much more recently (John *et al.*, 2003; Parfrey *et al.*, 2011). In contrast, the plastid genomes of tobacco and the liverwort *Marchantia polymorpha*, which diverged approximately 450 million years ago, are identical in coding content, except for the loss of *rps16* from *M. polymorpha*, and the loss of *rpl21* and an arginyl-tRNA gene from the tobacco plastid (Dorrell and Smith, 2011; Parfrey *et al.*, 2011; Shimada and Sugiura, 1991). Further differences in gene structure and order between the two fucoxanthin dinoflagellates (Tables 6.3, 6.5), indicate that the plastid genomes of different fucoxanthin dinoflagellates are extremely divergent from each other, despite their recent endosymbiotic origin.

The widespread distribution of poly(U) sites across the otherwise fast-evolving fucoxanthin plastid genome indicates that poly(U) tail addition has important roles in transcript processing. My data provide insights into possible roles for poly(U) tail addition at the *rpl36-rps13-rps11* and *psbD-Met*^{CAT}*-ycf4* loci. Polycistronic transcripts are generated from both loci, but transcripts from each locus undergo different cleavage events (Fig. 6.3). Polycistronic *rpl36-rps13-rps11* and *rpl36-rps13* transcripts are high in abundance, and may frequently receive poly(U) tails, suggesting that they represent the mature transcripts generated from the *rpl36-rps13-rps11* locus, whereas for *psbD-Met*^{CAT}*-ycf4* locus only monocistronic transcripts are abundant (Fig. 6.3). Notably, the *rpl36-rps13-rps11* and *rpl36-rps13* transcripts possess identical 5' ends to one another (Fig. 6.3; Table 6.8). Thus, at the *rpl36-rps13-rps11* locus, variation in 3' end processing and poly(U) tail addition may determine the coding content of each transcript, determining whether individual transcripts terminate within the *rps13* 3' UTR or contain a complete *rps11* CDS (Fig. 6.3). This is similar to data previously obtained from peridinin dinoflagellates and *Karlodinium veneficum*, which implies a potential role for poly(U) tail addition in alternative end processing of plastid transcripts (Barbrook *et al.*, 2012; Richardson *et al.*, 2014).

I additionally identify complex relationships between poly(U) tail addition and editing in the *K. mikimotoi* plastid. Poly(U) tail addition to transcripts covering *rps11, psbD* and *ycf4* is associated with high levels of editing (Table 6.10). This may be due to differences in the longevity of polyuridylylated versus non-polyuridylylated transcripts of these genes (for example, if polyuridylylated transcripts are more stable than non-polyuridylylated equivalents, they might persist in the plastid for longer periods of time and undergo greater degrees of editing). Alternatively, poly(U) tail addition may occur on these transcripts concurrent to the completion of plastid transcript editing, as has previously been suggested to occur (Dang and Green, 2009). Some of the editing events may also depend on transcript cleavage. For example, editing of the *ycf4* CDS does not occur on polycistronic *ycf4* transcripts, even if the

transcript possesses a poly(U) tail (Table 6.10). As certain editing events are specific to the polyuridylylated monocistronic *ycf4* transcripts, it seems unlikely that poly(U) tail addition is irrelevant for editing at this locus (Table 6.10). Poly(U) tail addition and transcript cleavage may, however, have a cooperative role in enabling editing, with editing occurring only on transcripts that possess mature 5' and 3' termini.

Finally, I have demonstrated that transcripts containing regions of antisense plastid sequence are present in fucoxanthin dinoflagellates, similarly to in peridinin dinoflagellates (Figs. 6.4, 6.5). Similarly to the situation in *Amphidinium carterae*, the antisense transcripts are generally not polyuridylylated; however, they are additionally highly edited (Figs. 6.6, 6.7). The presence of editing, which is not associated with nuclear gene expression in fucoxanthin dinoflagellates, strongly indicates that the antisense transcripts are generated from within the plastid itself, rather than from NUPTs. Although it is possible that there are sequences within the *K. mikimotoi* nucleus that are derived from plastid transcripts, as opposed to plastid gene sequences, which might have been relocated to the nucleus via a reverse transcriptase-mediated transfer event, studies of plastid-to-nucleus gene transfer in plants have indicated that RNA-mediated gene transfer only occurs at extremely low levels (Sheppard *et al.*, 2011). It seems unlikely that a sufficiently wide range of RNA-mediated gene transfer events have occurred in *K. mikimotoi* to give rise to a sufficiently diverse array of nuclear-located template sequences to account for the wide diversity of different editing states inferred to be present on the *K. mikimotoi* antisense transcripts (Fig. 6.7; Table 6.15). Thus, assuming that sense and antisense transcripts are generated alongside one another in the fucoxanthin plastid, the preferential application of poly(U) tails to sense transcripts may have an important role in discriminating sense transcripts from complementary antisense transcripts (Fig. 6.6).

The high levels of editing identified on the *K. mikimotoi* plastid transcripts is surprising, as editing of antisense transcripts has not previously been reported in any plastid lineage. Moreover, sense and antisense transcripts show complementary patterns of editing to one another (Fig. 6.7; Table 6.15). The complementary editing of sense and antisense transcripts may be evidence of the self-priming and extension of plastid sense transcripts *in vivo* by an RNA-dependent RNA polymerase, similarly to as has been suggested to be present in plants (Zandueta-Criado and Bock, 2004). Alternatively, plastid antisense transcripts might initially be transcribed from genomic templates, and undergo complementary patterns of editing to sense transcripts during processing. For example, completely unedited sense and antisense transcripts might anneal together early during processing, as has previously been suggested to occur in plant plastids, (Sharwood *et al.*, 2011; Zghidi-Abouzid *et al.*, 2011) and be edited

together as a dimer. This is supported by the similar processive relationships observed for editing on sense and antisense transcripts, with editing on polycistronic sense and antisense transcripts covering the *psbD-Met*^{CAT}*-ycf4* locus, for example, solely being confined to the *psbD* CDS (Table 6.15). The precise biochemical relationships between sense and antisense transcripts, and the transcript editing and poly(U) tail addition machineries in fucoxanthin plastids remain to be characterised. That notwithstanding, my data indicate that poly(U) tail addition has complex effects on transcript processing in fucoxanthin plastids, either through directly influencing other processing events, or by distinguishing functional transcripts from antisense transcripts, which may themselves have important roles in processing . More detailed investigation of poly(U) tail addition in *Karenia mikimotoi* may provide valuable insights into this unusual plastid gene expression system, and into the diversity of plastid physiology outside the plants.

Chapter Seven- Evolution and function of plastid transcript processing in algal relatives of malaria parasites

Introduction

The transition from a photosynthetic to a parasitic lifestyle has occurred many times across the eukaryotes, having been documented in members of the plants and the green, red and brown algae (Blouin and Lane, 2012; Gornik *et al.*, 2012; Tillich and Krause, 2010; Walker *et al.*, 2011). Typically, parasites that are descended from photosynthetic species retain non-photosynthetic plastids, with associated genomes. This genome is often highly reduced in content, with many previously functional genes either converted into pseudogenes, or lost completely (Randle and Wolfe, 2005; Tillich and Krause, 2010). In particular, genes that encode components of the photosynthetic electron transport machinery, which I will henceforth term "photosynthesis genes", are frequently lost from the plastid genomes of parasitic eukaryotes.

One of the most important parasitic lineages to be descended from photosynthetic ancestors is the apicomplexans. The apicomplexans include major pathogens of humans (*Plasmodium, Toxoplasma, Cryptosporidium*) and of livestock (*Theileria, Babesia*) (Walker *et al.*, 2011). All studied apicomplexans- apart from *Cryptosporidium* - retain a non-photosynthetic plastid that is fundamental to parasite viability and pathology (Fichera and Roos, 1997; McFadden *et al.*, 1996; Walker *et al.*, 2011). This plastid, termed the "apicoplast", is of red algal derivation and shares a common ancestry with the plastids found in peridinin dinoflagellates (Janouškovec *et al.*, 2010). The apicoplast has lost all photosynthesis genes from its genome (Cai *et al.*, 2003; Wilson *et al.*, 1996). This stands in contrast to the peridinin dinoflagellate plastid genome, which only retains photosynthesis genes, along with genes for ribosomal and transfer RNAs (Barbrook *et al.* 2013; Howe *et al.*, 2008b). The gene expression machinery of the apicoplast is also different from that of dinoflagellate plastids: for example, whereas dinoflagellate plastid transcripts receive 3' poly(U) tails, apicoplast transcripts do not (R.E.R. Nisbet, pers. comm.) (Dorrell *et al.*, 2014; Wang and Morse, 2006).

In the past decade, two fully photosynthetic species have been identified that form sister-groups of apicomplexans, to the exclusion of dinoflagellates. These are the "chromerid" algae *Chromera velia* and *Vitrella brassicaformis* (Janouškovec *et al.*, 2010; Moore *et al.*, 2008; Oborník *et al.*, 2012). The plastid genomes of *C. velia* and *V. brassicaformis* have been sequenced, and are of the same endosymbiotic derivation as the apicoplast, and the peridinin dinoflagellate plastid (Janouškovec *et al.*, 2010). However, unlike either the apicoplast or the peridinin dinoflagellate plastid, chromerid plastid genomes retain genes of

both photosynthetic and non-photosynthetic function, as well as open reading frames that are specific to either species and do not encode proteins of recognisable function (Janouškovec *et al.*, 2010; Janouškovec *et al.*, 2013b).

The chromerid algae represent an appealing model system to reconstruct the nature of the ancestor of what has became the apicoplast. Some of the biochemical pathways associated with the chromerid plastid, such as the use of a glutamate-independent pathway for tetrapyrrole synthesis, are also found in apicoplast lineages, but are not found in other algae (Koreny *et al.*, 2011). This suggests that certain plastid metabolism pathways associated with apicomplexans evolved prior to the loss of photosynthesis genes from the apicoplast. Other features of chromerid plastids, such as the sterol biosynthesis pathways employed, and the use of galactolipids in the plastid membranes, are similar to what is found in other photosynthetic plastid lineages, but have been lost or have been functionally modified in apicomplexans (Botté *et al.*, 2011; Botté *et al.*, 2013; Leblond *et al.*, 2012). Changes to these features may have occurred concurrent to the transition of apicomplexans from photosynthesis to parasitism.

A study performed prior to the work in this chapter demonstrated that poly(U) tails are applied to transcripts of three photosynthesis genes (*psbA, psbC, psaA*) in the *Chromera velia* plastid (Janouškovec *et al.*, 2010) This implies that poly(U) tail addition also occurred in the plastids of the photosynthetic ancestors of apicomplexans, but has been lost in their parasitic descendants. However, this study did not investigate whether poly(U) tails were added to plastid non-photosynthesis gene transcripts, or to plastid transcripts in *Vitrella brassicaformis*, or what functional role poly(U) tail addition plays in chromerid plastid gene expression. It is therefore not clear what consequences the loss of this pathway may have had on early apicomplexans.

This project was conceived to investigate the functional role of poly(U) tail addition in chromerid plastid gene expression. I wished to investigate which genes in chromerid plastids give rise to polyuridylylated transcripts, and determine whether it is a pathway that is broadly applied to every transcript of the plastid genome, or whether it is specifically associated with photosynthesis genes. I additionally wished to determine what functional roles poly(U) tail addition plays in plastid transcript processing in chromerids. From this, I wished to infer whether changes to or loss of the transcript poly(U) tail addition machinery may have played a role in the transition of early apicomplexans from a photosynthetic to a parasitic lifestyle.

I demonstrate that in both *Chromera velia* and *Vitrella brassicaformis,* poly(U) tails are principally added to transcripts of photosynthesis genes. Conversely, transcripts of other

genes in chromerid plastids tend not to be polyuridylylated. This is the first characterised plastid transcript processing pathway that differentially recognises a particular functional category of genes. I additionally demonstrate that poly(U) tail addition plays an important role in plastid photosynthesis gene expression, targeting transcripts of functional genes over pseudogene transcripts, and may influence other events in plastid transcript processing. As poly(U) tail addition appears to function principally in the expression of plastid photosynthesis genes, its loss may have played an important role in the loss of photosynthesis from early apicomplexans.

Results

Poly(U) tails are principally associated with photosynthesis gene transcripts in *Chromera velia* and *Vitrella brassicaformis*

I wished to determine whether plastid transcripts in *V. brassicaformis* receive poly(U) tails, similarly to in *C. velia* and in dinoflagellates (Dorrell and Howe, 2012a; Janouškovec et al., 2010; Wang and Morse, 2006). I additionally wished to test whether transcripts of plastid genes that are not directly involved in photosynthesis receive poly(U) tails in either species. To do this, cDNA was generated from each species using an oligo-d(A) primer, as previously shown to anneal to transcript poly(U) tails (Dorrell and Howe, 2012a; Richardson *et al.*, 2014). The oligo-d(A) primed cDNA was used as template for a series of PCR reactions using the same oligo-d(A) primer and a series of forward primers specific to plastid genes from each species (Table 7.1).

RT-PCRs performed against photosynthesis genes from each species (*psbA, C. velia atpB-2, V. brassicaformis atpB*) generated products of between 400 and 900 bp (Fig. 7.1, panel A; lanes 1-2, 7-8). These were of a size consistent with monocistronic polyuridylylated transcripts, with the poly(U) site in the 3' UTR of the gene concerned (considering gene size, and the positions of the RT-PCR primers employed).The identity of each oligo-d(A) primed RT-PCR product was confirmed by direct sequencing, using the gene-specific PCR primer. Similar products were observed with a control transcript from a dinoflagellate plastid (*Amphidinium carterae psbA*) that is known to be polyuridylylated (Fig. 7.1, panel B; lane 1) (Barbrook *et al.*, 2012). Analogous RT-PCRs against representative plastid genes from both species that do not encode products directly involved in photosynthesis (*rps11, rrs*) failed to generate products (Fig. 7.1, panel A; lanes 3-4, 9-10). Products could not be identified for these genes even after a second successive round of PCR amplification, using the primary

Table 7.1: Primers used for oligo-d(A) RT-PCR

Primer sequences used in the RT-PCRs shown in fig. 7.1 are shown in bold text.

Oligo-d(A)	GGGACTAGTCTCGAGAAAAAAAAAAAAAAAAA		
	1. Chromera velia		
Gene	**PCR forward primer**	**Gene**	**PCR forward primer**
acsF	GGTTGGTGTCAGGATGAG	psbE	CTGGAGGATCTACTGGCG
atpA	TCCAGGTCGCGAAGC	psbH	GGGGACCCTACACCTGTAAC
atpB-1	GTTGTTCGTCTTCAATACATAGC	psbJ	GCGTTCCTCTTTGATTTG
atpB-2	**AAAAGAGAAAGCGCAGATC**	psbK	TCTCCAAGCTTCCTCTTG
atpH-1	AAGGAATTGTAGCAGCGTG	psbN	GTTGGTCGGCTCGATTAG
atpH-2	GAAAGCAATCGAGCCTTG	psbT	TGTACGTTACTTTTCGGGAC
atpI	TGTAATTTTACATTTCATGGAGAAAC	psbV	CCGATTCAATCCGAACTG
ccsA	TGGAGATCCAGCACACTTC	rpl11	CCCATCTCCACCCGTC
clpC-1	CAAAATGCGATGAACGAC	rpl14	AGCGGGGCGTTAAGG
clpC-2	GGCGTATGGTCATGCAAC	rpl16	ACAGGCTCCCGAAAGAC
clpC-3	CCACGGCCAGTACATCC	rpl2	TGGTGGTGCTGTGTTTG
ORF115	CCTTTTGAACGTGGGG	rpl20	CGCGTTCCACTGAGTTTAC
ORF1173	CCGCACTTTAGCCTGACTC	rpl3	GGCTTTTTAGGTGTGGTCG
ORF122	TTTGAGGCTCGGTTGG	rpl31	CGTATTTTGTGACGGGAC
ORF128	TGGTTGGGGTCTTCTACCC	rpl36	GCGATGCCACTCCAC
ORF137	GAGTTGGTGAGTAGGGCGG	rpl4	CATTCTCCGGGGTCG
ORF147	AGGTGGTGAAGTTGGGGTTC	rpl5	GCTGCAACAAATTACCGG
ORF157	CACCCTTGCAGCGGATTTTC	rpl6	CAGCTTGAGTGAGCCGAC
ORF175	GGACTTGGAGGAAGCATC	rpoA	GTTACAAAGCTCTGGTCCC
ORF201	TGGTTGGGGTCTTCTACC	rpoB	GACGGAACCTCTCCAGAC
ORF207	GGGGTGCTTACAGTTTCG	rpoC1	TTGCCAGTGTTACCTCCTG
ORF230	TGGAATGAAGAACCTCGCCC	rpoC2	GTGCGTGTACTTGATCCG
ORF247	TCAGCAGGGCCCAAAG	**rps11**	**TGATCAAACCTGCCCAAC**
ORF264 3' end	GCACTGGGACAACTCAAC	rps12	CCCCTGCTTTACGTGG
ORF264 5' end	GCTGCGAAAAGTCTAGCG	rps13	ATTTAGTCGTTGACGTGGG
ORF325	GGGTCAAAAAGCGAGGAC	rps14	CAAAAATCTCTAGAGCGAATAAAG
ORF389	CTGCTAGACTGGTGCTACG	rps17	AGCTATTTTTCGGGCTTTG
ORF391	CAAGGTTTCGTTGGGC	rps18	CGCAAAGAAAGGTTCCTG
ORF634	GGAGCGTACGTTAAAGGG	rps19	CATCTTCAACCGCGTAATG
petA	CCCGGTAAGATCACCTG	rps2	TGCCGTTATTCGTGC
petB	CGGCCCAGTTTTAGTCC	rps3	GAAACCTACGACCGCTC
petD	CCTTCTCCGTTTAATTCCG	rps4	GCGAAAGACCAACGATG
petG	AACGAACCTCTTTTGTTTGG	rps7	TCCCTGTTCCACGAGC
psaA-1	CAACGTAGTGGCTTGGTC	rps8	AAGCGATGTTCTGAGCG
psaA-2	CGCCCAAGCGTAAGTAATC	rrl	CAGTCGCCTCCTAAAAGG
psaB	GGGTCGCGGTTTATGC	**rrs**	**TAGATGTTGGACGCACG**
psaC	GTGTTCGGGCTTGTCC	secA	GCAGGTCGTAAAGCAAGTG
psbA	**ATGGAATTCGTGAGCCAG**	secY	GCTTCATTTGCGGAAATAC
psbB	GCTTCCATGTGGTACGG	tatC	GCAGGGCATTAAATTCTTG
psbC	TATGGACCGACGGGG	tufA	CCTTTTGTGTCCGGGTC
psbD	CGTTGGCTTCATTTCTTTATG	ycf3	GGGCAGCATGGCAAG
	2. Vitrella brassicaformis		
Gene	**PCR forward primer**	**Gene**	**PCR forward primer**
acsF	CATAAGCTGCTTGGTCTCC	clpC	GACTGGAGCACGAGCAG
atpB	**CAGATGACTTGACGGACC**	ORF136	GGGGCATTTTCTGGC
atpH	GGGTTGGCAAACAAAAGAG	ORF3	GCTCCATTTGAAGCCC
atpI	GGAATACGACGCAAAGG	ORF87	GGATGGGAGCAGACTGG
ccsA	TGTGCGCTGAACTTGC	petB	CACCGGTGTGATCATGTC
ccsl	GCCAGGACAATCAAACC	petD	CCTTTCGCGACTCCAC
chlB	ACATATTGCCAACACGACC	petG	GGCATGTCTAGCGTTTCG
chlL	GAGGCTATGTGGTGGGAG	petN	TGTGCGCTGAACTTGC
chlN	GAGGGTATGCGAAATAGGG	psaA	GGAGTCTTGTACGCTCGC

Table 6.1 (continued)

Vitrella brassicaformis			
Gene	**PCR forward primer**	**Gene**	**PCR forward primer**
psaC	TTTACGACACCTGCATCG	**rps11**	**CCCGTTTCGGAAAAAC**
psaD	GAGGACAGGGGGTGAAG	rps14	GCGCGTGACACATTTAAC
psbA	**CCGTCTAGGTATGCGTCC**	rps16	CGAACCTATTCTCCAGCC
psbB	TCTTTGCGGCCTTCG	rps17	TTGTGACCAGTGTGTCAATG
psbD	CGTTGGCTTCATTTCTTTATG	rps18	CCGCGCTTTAGCTTTAGAG
psbH	GAGCACGCGTCTAACG	rps19	AACTCTAGCGCACCTGC
psbJ	CGCGTGCCTTTATGG	rps2	ACCGGGTGTTTCTCTGG
psbK	TGTGCGCTGAACTTGC	rps4	GCTCTTCTCTGTGAGTTGGC
psbN	TCGTCGCTTCAATGCC	**rrs**	**GCGTCTGTAGGTGGTTTG**
psbT	CGCGTGCCTTTATGG	secA	CTGCACGTGTGGATCAG
psbV	CGAAGACCAACCAAAACG	sufB	GGATGTGCGTAGGATCAC
rpl2	TTAGCTACAGCTCGCGG	ycf3	GGCTCTTACTTGGAGGCG
rpl3	CCCGTCTGGACACCTTC	ycf4	CGTGGTTCTCGTGAGTGG
rpoC1	GATTTTGATGGCGACCAG		

3. Gene-specific cDNA primers	
Gene	**Primer sequence**
Chromera atpH-1	GAAAAAGCTGAGCACGC
***Chromera* psbA**	**CAACTACCGGTCAAATTGC**
***Chromera* rps11**	**GTCTAAAGGCAGGATACGC**
***Chromera* rrs**	**GGTTTGACGTGGACGAG**
***Vitrella* psbA**	**GGCATACAGCAAGGAAG**
***Vitrella* rps11**	**CATTGTGCGGAGCTAAAGTC**
***Vitrella* rrs**	**GTGTACAAGGCTCGGGAAC**

4. Control RT-PCRs		
Gene	**PCR forward primer**	**Gene-specific cDNA primer**
***Chromera* Hsp90**	**CCGGTGAGGACTTGATCT**	**CTCCATGGTCTTCTTGCTC**
***Phaeodactylum* psbA**	**GCGGTTTTTGTGGTTGGATTAC**	**TAAAGCACGAGAGTTGTTAAATGAAG**
***Amphidinium* psbA**	**CTTCTAACGCAATCGGTGTCC**	**GATACCAATTACAGGCCAAGC**

PCR product as a reaction template. Products were detected for these genes by RT-PCR using a gene-specific cDNA and PCR reverse primer (Fig. 7.1, panel A; lanes 5-6, 11-12), implying that transcripts of each gene are present, but do not receive a poly(U) tail. Similar results were observed for a nuclear transcript (*C. velia Hsp90*) as well as a transcript from a diatom plastid (*Phaeodactylum tricornutum psbA*) which has previously been shown not to receive a poly(U) tail (Fig. 7.1, panel B; lanes 3-6) (Dorrell and Howe, 2012a).

To determine whether poly(U) sites are significantly associated with photosynthesis genes in chromerids, similar oligo-d(A) RT-PCRs were performed for every annotated gene and open reading frame in the *C. velia* plastid genome (n=78) and over half the genes in the *V. brassicaformis* plastid genome (n=43, out of 74 total) (Tables 7.2, 7.3). Each of the products were confirmed by direct sequencing, and negative results were confirmed with a second

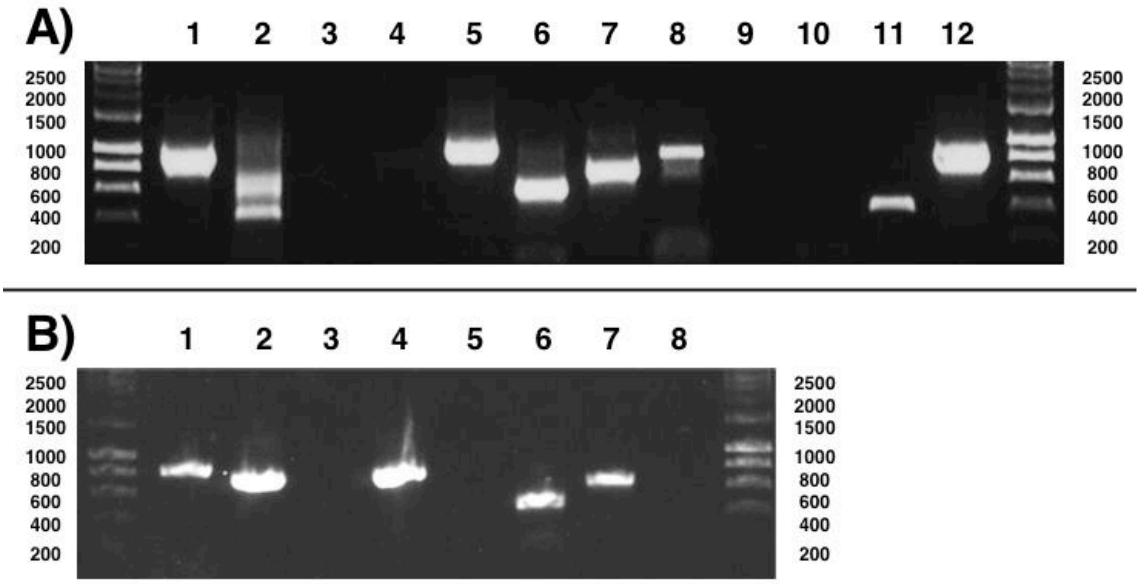


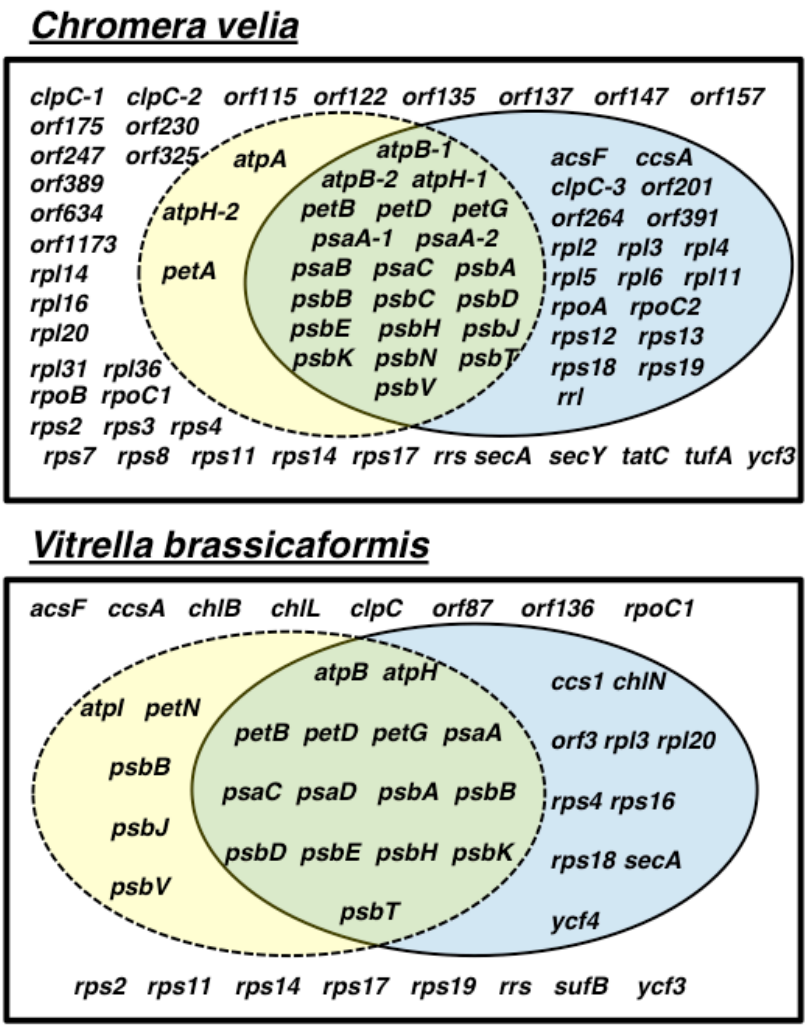
Fig. 7.1: Differential polyuridylylation of plastid transcripts from *C. velia* and *V. brassicaformis*.

This gel photo shows the distribution of poly(U) sites across four representative genes in chromerid plastids. Hyperladder I (Bioline) was used as a size marker, with the positions of size bands given to the side of each gel photo. These data were obtained with the assistance of an undergraduate student, James Drew.

Panel A: RT-PCRs for *C. velia* (**lanes 1-6**) and *V. brassicaformis* (**lanes 7-12**). Lanes **1-2, 7-8**: Oligo-d(A) RT-PCRs for the photosynthesis genes *psbA* and *atpB-2* (*C. velia)/atpB* (*V. brassicaformis*); lanes **3-4, 9-10:**, oligo-d(A) RT-PCRs for the non-photosynthesis genes *rps11* and *rrs*; lanes **5-6, 11-12**: RT-PCRs using an internal, gene-specific cDNA primer for *rps11* and *rrs* for both species. The multiple bands observed for *C. velia atpB-2* (lane 2) correspond to different *atpB-2* transcripts containing alternative poly(U) sites.

Panel B: control RT-PCRs. **Lanes 1-2**: oligo-d(A) and internal gene-specific RT-PCRs for *A. carterae psbA; lane*s **3-4**: oligo-d(A) and internal gene-specific RT-PCRs for *P. tricornutum psbA; lane*s **5-6**: oligo-d(A) and internal gene-specific RT-PCRs for *C. velia Hsp90*, lanes **7-8**: PCR positive (DNA template) and negative controls (no template) for *C. velia psbA*.

round of PCR amplification using the primary product as template, as before. None of the transcripts sequenced from either species contained any evidence of post-transcriptional



Chromera

	Poly(U)	Non-poly(U)
PS	22	3
Non-PS	16	22

P < 0.0005

Vitrella

	Poly(U)	Non-poly(U)
PS	15	5
Non-PS	9	14

P < 0.05

Fig. 7.2: The total distribution of poly(U) sites across chromerid plastids.

The Venn diagrams show the total results of oligo-d(A) RT-PCRs for genes from *C. velia* and *V. brassicaformis*. Chi-squared distributions and P values for the significance of association between photosynthetic function and presence of an associated poly(U) site are shown to the right of each diagram. These data were obtained with the assistance of an undergraduate student, James Drew.

sequence editing. The absence of editing from *C. velia* has since been confirmed in an independent study (Janouškovec et al., 2013b).

In both species, poly(U) tail addition was significantly associated with transcripts of photosynthesis genes (chi-squared: *C. velia* $P < 0.005$; *V. brassicaformis* $P< 0.05$). While some genes were found to contradict general patterns - i.e. photosynthesis genes that do not possess associated poly(U) sites, or non-photosynthesis genes that give rise to polyuridylylated transcripts- most of these exceptions were specific to one species (Fig. 7.2). Only two non-photosynthesis genes (*rpl3* and *rps18*) were found to possess poly(U) sites in both species. Furthermore, none of the non-polyuridylylated photosynthesis genes was conserved between *C. velia* and *V. brassicaformis* (Fig. 7.2). Poly(U) sites are therefore strongly associated with genes that function in photosynthesis in chromerid plastids.

Poly(U) sites are highly variable in chromerid plastids

Each poly(U) site identified was aligned with the *Chromera velia* and *Vitrella brassicaformis* plastid genomes (Janouškovec *et al.*, 2010; Janouškovec *et al.*, 2013b). None of the poly(U) sites identified was predicted to lie within poly(T) tracts of more than 6 bp in either genomic sequence, suggesting that the poly(U) tails identified correspond to genuine post-transcriptional modifications (Janouškovec *et al.*, 2010). The poly(U) sites identified were in very variable positions. Poly(U) sites in *C. velia* were located an average of 145 nt downstream of the associated gene, but in one case (*ORF264*) a poly(U) site was detected 50 nt upstream of the stop codon, and in another (*psbH*) a poly(U) site was detected 584 nt into the 3' UTR (Table 7.2). Similarly, while the poly(U) sites in *V. brassicaformis* were located an average of 55 nt downstream of the stop codon, in one case (*petG*) a 3' UTR of 277 nt was recorded (Table 7.3).

There even appeared to be variation in the poly(U) sites associated with different transcripts from specific genes. For example, an oligo-d(A) RT-PCR for *C. velia atpB-2* produced multiple bands that were visible on gel electrophoresis (Fig. 7.1, panel A; lane 2). Similar results were found for a number of other *C. velia* plastid genes (Table 7.2). To assess the variation in the *atpB-2*, the products of the oligo-d(A) RT-PCR were cloned, and individual colonies, corresponding to individual transcripts, were sequenced (Fig. 7.3). Across twenty clones, eleven different poly(U) sites were observed, ranging from 60 to 467 nt into the 3' UTR (Fig. 7.3, Table 7.2), which broadly corresponded to the different band sizes visible on oligo-d(A) RT-PCR (Fig. 7.1, panel A; lane 3).

Table 7.2. Features of the poly(U) sites present in *Chromera velia*

This table shows the result of every diagnostic oligo-d(A) RT-PCR reaction performed for *C. velia* plastid genes. Each gene is listed either as having a direct function in photosynthetic electron transfer, having a defined function that is not directly associated with photosynthesis, or as being an ORF with no defined function. Genes for which a range of UTR length values are given were found to produce multiple polyuridylylated products; the most extreme positions and values identified are given. In genes where multiple potential poly(U) sites were identified, the average of the most extreme values observed were used for calculations of total species mean values. These data were obtained with the assistance of an undergraduate student, James Drew.

Gene	Photosynthesis function?	Poly(U) site	3' UTR length (bp)	Poly(U) tail length	Notes
acsF	**N**	Y	145	20	Poly(U) site 137nt into 5' end of ORF391
atpA	**Y**	-			
atpB-1	**Y**	Y	15 to 75	18 to 22	
atpB-2	**Y**	Y	60 to 437	23 to 44	
atpH-1	**Y**	Y	116	74	Poly(U) site adjacent to 5' end of Lys^{TTT} tRNA
atpH-2	**Y**	-			
atpI	**Y**	Y	146 to 162	15 to 18	Poly(U) site 18 to 30nt into 5' end of rpl11
ccsA	**N**	Y	141	19	Poly(U) site 6nt upstream of Pro^{TGG} tRNA
clpC-1	**N**	-	297 to 492	19 to 27	
clpC-2	**N**	-			
clpC-3	**N**	Y	43	16	
ORF115	**ORF**	-			
ORF1173	**ORF**	-			
ORF122	**ORF**	-			
ORF135	**ORF**	-			
ORF137	**ORF**	-			
ORF147	**ORF**	-			
ORF157	**ORF**	-			
ORF175	**ORF**	-			
ORF201	**ORF**	Y	67 to 107	17 to 36	Poly(U) site 41nt into clpC
ORF230	**ORF**	-			
ORF247	**ORF**	-			
ORF264	**ORF**	Y	-50 to 181	20	Generated using primer against 3' end of gene
ORF325	**ORF**	-			
ORF389	**ORF**	-			
ORF391	**ORF**	Y	350	17	Poly(U) site 296nt into 5' end of ORF157
ORF634	**ORF**	-			
petA	**Y**	-			
petB	**Y**	Y	62 to 197	33	Poly(U) site up to 27nt into 5' end of psbH
petD	**Y**	Y	86 to 177	14 to 40	
petG	**Y**	Y	301 to 327	20 to 46	
psaA-1	**Y**	Y	226	19	
psaA-2	**Y**	Y	56	19	
psaB	**Y**	Y	158	15	
psaC	**Y**	Y	116 to 346	18 to 26	
psbA	**Y**	Y	11 to 75	27	
psbB	**Y**	Y	94	20	Poly(U) site 69nt into 5' end of psaA-1
psbC	**Y**	Y	99 to 411	13 to 18	Poly(U) site up to 274nt into 5' end of clpC-2
psbD	**Y**	Y	93	20	
psbE	**Y**	Y	89 to 299	33 to 45	Poly(U) site 185nt into 5' end of psaA-2

Table 7.2 (continued)

Gene	Photosynthesis function?	Poly(U) site	3' UTR length (bp)	Poly(U) tail length	Notes
psbH	**Y**	Y	76 to 580	28 to 36	Poly(U) site 10nt into 5' end of atpA
psbJ	**Y**	Y	136	20	
psbK	**Y**	Y	176	18	Poly(U) site 127nt into 5' end of psbV
psbN	**Y**	Y	110	34	
psbT	**Y**	Y	115	43	Poly(U) site 22nt into 5' end of psaC
psbV	**Y**	Y	128 to 275	20 to 36	Poly(U) site 8nt into 5' end of rpl4
rpl11	**N**	Y	53	15	Poly(U) site 11nt into 5' end of rpoB
rpl14	**N**	-			
rpl16	**N**	-			
rpl2	**N**	Y	131	18	Poly(U) site 53nt into 5' end of ORF634
rpl20	**N**	-			
rpl3	**N**	Y	67	14	
rpl31	**N**	-			
rpl36	**N**	-			
rpl4	**N**	Y	84	18	Poly(U) site 15nt into 5' end of ORF1173
rpl5	**N**	Y	84	21	Poly(U) site 25nt into 5' end of rps8
rpl6	**N**	Y	37	27	Poly(U) site 4nt into 5' end of secY
rpoA	**N**	Y	14	38	Poly(U) site 9nt into 5' end of ORF115
rpoB	**N**	-			
rpoC1	**N**	-			
rpoC2	**N**	Y	83	17	Poly(U) site 73nt into 5' end of rps3
rps11	**N**	-			
rps12	**N**	Y	84	26	
rps13	**N**	Y	72	27	Poly(U) site 42nt into 5' end of ORF135
rps14	**N**	-			
rps17	**N**	-			
rps18	**N**	Y	351 to 372	18	Poly(U) site 13nt into 5' end of rps11
rps19	**N**	Y	189	18	Poly(U) site 72nt into 5' end of rpl2
rps2	**N**	-			
rps3	**N**	-			
rps4	**N**	-			
rps7	**N**	-			
rps8	**N**	-			
rrl	**N**	Y	2	20	
rrs	**N**	-			
secA	**N**	-			
secY	**N**	-			
tatC	**N**	-			
tufA	**N**	-			
ycf3	**N**	-			
Mean			144.5	24.5	

Poly(U) sites are not associated with other sequence features in chromerid plastid genomes

I wished to determine whether there were any other sequence features, beyond a gene function in photosynthesis, that were associated with poly(U) sites in chromerid plastids. Certainly, there were no clear trends underpinning the locations of genes that possess associated poly(U) sites in the plastid genomes of either species. Poly(U) sites were

Table 7.3. Features of the poly(U) sites present in V*itrella brassicaformis.*

Poly(U) site information for *V. brassicaformis* is shown as per in Table 7.2. These data were obtained with the assistance of an undergraduate student, James Drew.

Gene	Photosynthesis function?	Poly(U) site	3' UTR length (bp)	Poly(U) tail length	Notes
acsF	**N**	-			
atpB	**Y**	Y	263	19	Poly(U) site adjacent to 5' end of Tyr^{GTA} tRNA
atpH	**Y**	Y	20	18	
atpI	**N**	-			Forms dicistronic atpI-atpH transcript
ccsA	**N**	-			Forms tetracistronic rps14-psbV-ccsA-psbK transcript
ccsI	**N**	Y	5	16	Poly(U) site adjacent to 5' end of Asn^{GTT} tRNA
chlB	**N**	-			
chlL	**N**	-			
chlN	**N**	Y	16	17	Poly(U) site adjacent to 5' end of Val^{TAC} tRNA
clpC	**N**	-			
ORF136	**ORF**	-			
ORF3	**ORF**	Y	36	16	
ORF87	**ORF**	-			
petB	**Y**	Y	21	17	
petD	**Y**	Y	30	18 to 23	Poly(U) site adjacent to 5' end of Gly^{TCC} tRNA
petG	**Y**	Y	33 to 277	16 to 19	3' UTR extends over antisense region containing psbN
petN	**Y**	-			
psaA	**Y**	Y	86	15	
psaC	**Y**	Y	50	18	
psaD	**Y**	Y	14	18	Poly(U) site adjacent to 5' end of Leu^{CAA} tRNA
psbA	**Y**	Y	17	13 to 18	
psbB	**Y**	-			
psbD	**Y**	Y	18	18	
psbE	**Y**	Y	72	16	Forms dicistronic ycf4-psbE transcript
psbH	**Y**	Y	92	15	
psbJ	**Y**	-			Forms dicistronic psbJ-psbT transcript
psbK	**Y**	Y	27	20	Poly(U) site adjacent to 5' end of Trp^{CCA} tRNA
psbT	**Y**	Y	32	20	Poly(U) site adjacent to 5' end of Met^{CAT} tRNA
psbV	**Y**	-			Forms tetracistronic rps14-psbV-ccsA-psbK transcript
rpl20	**N**	Y	75	15	
rpl3	**N**	Y	150	18	3' UTR extends over antisense region containing Phe^{GAA}
rpoC1	**N**	-			
rps11	**N**	-			
rps14	**N**	-			Forms tetracistronic rps14-psbV-ccsA-psbK transcript
rps16	**N**	Y	28	17	Poly(U) site adjacent to 5' end of Ser^{TGA} tRNA
rps17	**N**	-			
rps18	**N**	Y	42	20	Poly(U) site 24nt into 5' end of ORF3
rps19	**N**	-			
rps2	**N**	-			
rps4	**N**	Y	27	17	Poly(U) site adjacent to 5' end of His^{GTG} tRNA
rrs	**N**	-			
secA	**N**	Y	7	16	Poly(U) site adjacent to 5' end of Asn^{GTT} tRNA
sufB	**N**	-			
ycf3	**N**	-			
ycf4	**N**	Y	27	17	Additionally forms dicistronic ycf4-psbE transcript
Mean		54.6	17.4		

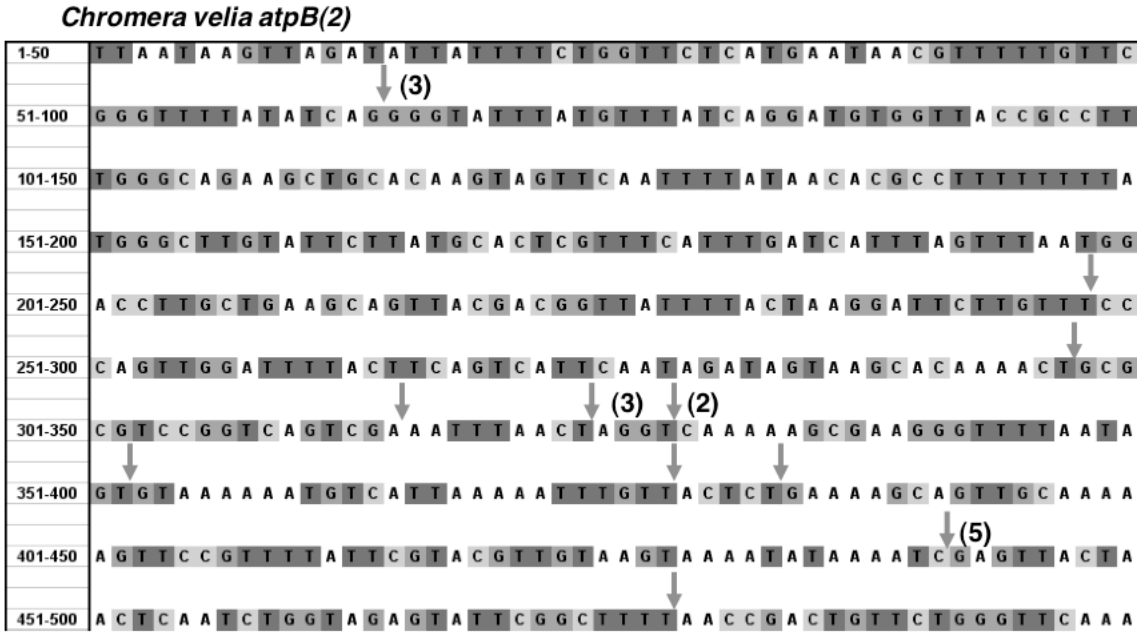


Fig. 7.3: Associated poly(U) sites of the *C. velia atpB-2* gene.

This alignment shows the first 500 bp downstream of the *Chromera velia atpB-2* gene. Grey arrows correspond to the different poly(U) sites, identified from the sequences of twenty randomly selected separate, individual cloned oligo-d(A) RT-PCR products using a gene-specific forward PCR primer against *C. velia atpB-2*. Numbers indicate that multiple colonies gave rise to the same poly(U) site.

identified on genes located at the 5' end (*C. velia petG*), in the interior (*C. velia atpB-2*) and 3' end (*C. velia psbA*) of clusters of photosynthesis genes, as well as on photosynthesis genes located between genes of non-photosynthetic function (e.g. *C. velia atpI*, positioned between *rps14* and *rps11*) (Janouškovec *et al.*, 2010). There was likewise no clear association between the presence of poly(U) sites and genes that were located either at the start or end of potential operons in *C. velia*, with operons defined as clusters of genes that are in the same transcriptional orientation to each other that are not interrupted by genes of opposing orientation (Table 7.4, chi-squared: P>0.35) (Janouškovec *et al.*, 2010).

Plant plastids utilise two RNA polymerases: a nucleus-encoded polymerase, related to the phage-type mitochondrial polymerase, and a bacterial-type, plastid-encoded polymerase (Hedtke *et al.*, 1997; Liere *et al.*, 2011). Each of these polymerases is able to transcribe the majority of the genes in the chloroplast genome (Krause *et al.*, 2000; Zhleyazkova *et al.*,

Table 7.4: Bioinformatic analysis of the distribution of poly(U) sites in *Chromera velia*.

This table presents an overview of possible associations between the presence of poly(U) sites in the *C. velia* plastid genome alongside different gene features. These features include the function of the protein encoded by the gene, the position of each gene within predicted operons are listed, as defined from the genomic sequence (Janouškovec *et al.*, 2010), and the presence of bacterial promoter sites within the gene 5' UTR, as predicted via a Neural Network server with a threshold value of 0.8 (Reese, 2001). Chi-squared values are calculated for each possible association over the genome as a whole, and excluding ORFs of unannotated function. In addition, the transcript abundance of genes that possess and lack associated poly(U) sites are compared. Transcript abundance levels are taken from the mean read coverage obtained in a previous study, in logarithmic and in ranked terms (Janouškovec *et al.*, 2013b). Transcript abundance is calculated against the genome as a whole and against each functional category of genes (photosynthesis genes, non-photosynthesis genes and unannotated ORFs). The individual sequence features and transcript abundance values for each individual gene are presented below the overview.

1. Overview

Chi-squared association between poly(U) site	All genes	Excluding ORFs
Gene function in photosynthesis	0.000	0.001
Start of predicted operon	0.368	0.178
End of predicted operon	0.799	0.623
Predicted bacterial promoter in 5' UTR	0.026	0.237

Read coverage (Janouškovec et al., 2013)	Log_{10} value		Rank value	
	Poly(U)	Non-poly(U)	Poly(U)	Non-poly(U)
All genes	4.49	4.26	30	51
Photosynthesis genes	4.56	4.06	17	33
Non-photosynthesis genes	4.47	4.44	43	53
ORFs	3.27	3.11	50	51

2. Individual Gene Features

			Transcription features			Read coverage (Janouškovec *et al.*, 2013)	
Gene	Poly(U)	Photosynthesis	Operon Start	Operon End	Promoter	Log_{10}	Rank
acsF	Y	N	N	N	Y	3.21	41
atpA	N	Y	N	N	N	3.00	47
atpB-1	N	Y	N	Y	Y	3.21	39
atpB-2	Y	Y	N	N	N	3.60	27.5
atpH-1	Y	Y	N	Y	N	1.16	78
atpH-2	N	Y	N	N	N	3.21	40
atpI	Y	Y	N	N	N	4.20	15
ccsA	Y	N	Y	N	N	3.18	45
clpC-1	Y	N	N	N	Y	3.26	38
clpC-2	N	N	N	N	Y	3.37	35
clpC-3	N	N	Y	N	N	2.08	73
orf115	N	N	N	N	N	2.43	69
orf1173	N	N	N	N	Y	3.41	31
orf135	N	N	N	N	N	1.77	77
orf137	N	N	N	N	N	2.79	58

Table 7.4 (continued)

Gene	Poly(U)	Photosynthesis	Transcription features			Read coverage (Janouškovec et al., 2013)	
			Operon Start	Operon End	Promoter	Log_{10}	Rank
orf264	Y	N	N	Y	Y	4.32	11
orf325	N	N	N	N	N	2.86	53
orf389	N	N	N	N	N	4.24	13
orf391	N	N	N	N	N	2.30	71
orf634	N	N	N	N	N	4.23	14
petA	N	Y	Y	N	Y	2.64	61
petB	Y	Y	N	N	Y	4.04	21
petD	Y	Y	N	Y	Y	4.62	6
petG	Y	Y	N	N	Y	3.42	30
psaA-1	Y	Y	N	N	Y	4.19	17
psaA-2	Y	Y	N	N	N	4.45	9
psaB	Y	Y	N	N	Y	4.11	20
psaC	Y	Y	N	N	Y	3.93	23
psbA	Y	Y	N	N	N	4.69	4
psbB	Y	Y	N	N	N	4.16	18
psbC	Y	Y	N	N	Y	4.29	12
psbD	Y	Y	N	N	Y	5.58	3
psbE	Y	Y	N	N	Y	4.36	10
psbH	Y	Y	N	N	Y	4.51	8
psbJ	Y	Y	N	Y	Y	4.58	7
psbK	Y	Y	Y	N	Y	4.62	5
psbN	Y	Y	N	N	Y	4.19	16
psbT	Y	Y	N	N	Y	3.90	24
psbV	Y	Y	N	N	N	3.97	22
rpl11	Y	N	N	N	Y	2.81	55
rpl14	N	N	N	N	Y	2.62	63
rpl16	N	N	N	N	N	2.63	62
rpl2	Y	N	N	N	N	2.83	54
rpl3	Y	N	N	N	Y	3.47	29
rpl31	N	N	N	N	N	3.09	46
rpl36	N	N	N	N	N	1.78	76
rpl4	Y	N	N	N	Y	3.36	36
rpl5	Y	N	N	N	Y	2.87	51
rpl6	Y	N	N	N	Y	2.96	48
rpoA	Y	N	N	N	N	1.99	75
rpoB	N	N	N	N	N	2.49	67
rpoC1	N	N	N	N	Y	2.07	74
rpoC2	Y	N	N	N	Y	3.34	37
rps11	N	N	N	N	Y	2.75	60
rps12	Y	N	N	N	Y	3.39	34
rps13	Y	N	N	N	Y	2.95	49
rps14	N	N	N	N	Y	2.80	56
rps17	N	N	N	N	Y	2.76	59
rps18	Y	N	N	N	Y	2.92	50
rps19	Y	N	N	N	N	5.68	2
rps2	N	N	Y	N	Y	2.55	66
rps3	N	N	N	N	N	3.82	25
rps4	N	N	Y	N	Y	2.46	68
rps7	N	N	N	N	Y	2.34	70
rps8	N	N	N	N	N	2.56	64
rrf	N	N	N	N	N	3.18	44
rrl	Y	N	Y	N	Y	2.86	52
rrs	N	N	N	Y	N	5.76	1
secA	N	N	Y	N	Y	2.19	72
secY	Y	N	N	N	N	3.19	43
tatC	N	N	N	N	Y	3.19	42
tufA	N	N	N	N	Y	4.12	19
ycf3	N	N	N	Y	N	3.60	27.5

2013). However, while the nucleus-encoded polymerase is active in developmentally inactive tissue, where it is involved in basal transcription of the plastid genome, the plastid-encoded polymerase is principally active in photosynthetic tissue, and thus is predominantly involved in the transcription of photosynthesis genes (Liere *et al.*, 2011; Williams-Carrier *et al.*, 2014). I wished to determine whether poly(U) sites in chromerid plastids are associated with genes that are transcribed via a specific plastid RNA polymerase. While there is no evidence for the presence of a phage-type polymerase in algal plastids (Teng *et al.*, 2013; Yin *et al.*, 2010), α, β and β' subunits of a bacterial-type plastid polymerase are encoded in the plastid genomes of *C. velia and V. brassicaformis* (Janouškovec *et al.*, 2010). To test whether this polymerase might preferentially transcribe genes that contain poly(U) sites, bacterial-type promoter sequences were identified across the 5' UTR of every gene in the *C. velia* plastid using a Neural Network Promoter Prediction server (Reese, 2001).

Similar to what has been reported in plants (Liere *et al.*, 2011; Zhelyazkova *et al.*, 2012), candidate promoters were identified at a wide range of positions, including upstream of photosynthesis and non-photosynthesis genes in the *C. velia* plastid (Table 7.4). Across the entire genome, bacterial promoters were weakly enriched upstream of genes that possess poly(U) sites (chi-squared: $P < 0.05$; Table 7.4). However, this was almost entirely due to the fact that bacterial promoters were generally not found upstream of ORFs of unknown function, which are also less likely to possess poly(U) sites than photosynthesis genes. Excluding ORFs of unknown function, there was not a significant association between the presence of predicted bacterial promoters and poly(U) sites (chi-squared: $P > 0.2$; Table 7.4). There is therefore not a convincing association between the activity of a bacterial RNA polymerase and the distribution of poly(U) sites in chromerid plastids.

The initial report of poly(U) tails in dinoflagellate plastids suggested that specific sequence motifs might be associated with poly(U) sites (Wang and Morse, 2006), although subsequent studies in other dinoflagellate species could not detect similar motifs (Howe *et al.*, 2008b; Nelson *et al.*, 2007; Richardson *et al.*, 2014). I wished to determine whether specific poly(U) associated motifs were present in chromerid plastids. To do this, the 3' UTR sequences of every polyuridylylated transcript were compared to one another. To identify possible poly(U)-associated motifs located downstream of each poly(U) site, the first 100 bp after each poly(U) site were similarly compared to one another. To ensure that any motifs identified were specifically associated with polyuridylylated transcripts, rather than a general feature of 3' UTR sequences in chromerid plastid genomes, the 3' UTR sequences of the plastid genes identified to lack an associated poly(U) site were independently searched for conserved motifs. Across each of the alignments inspected, there were no conserved sequence motifs,

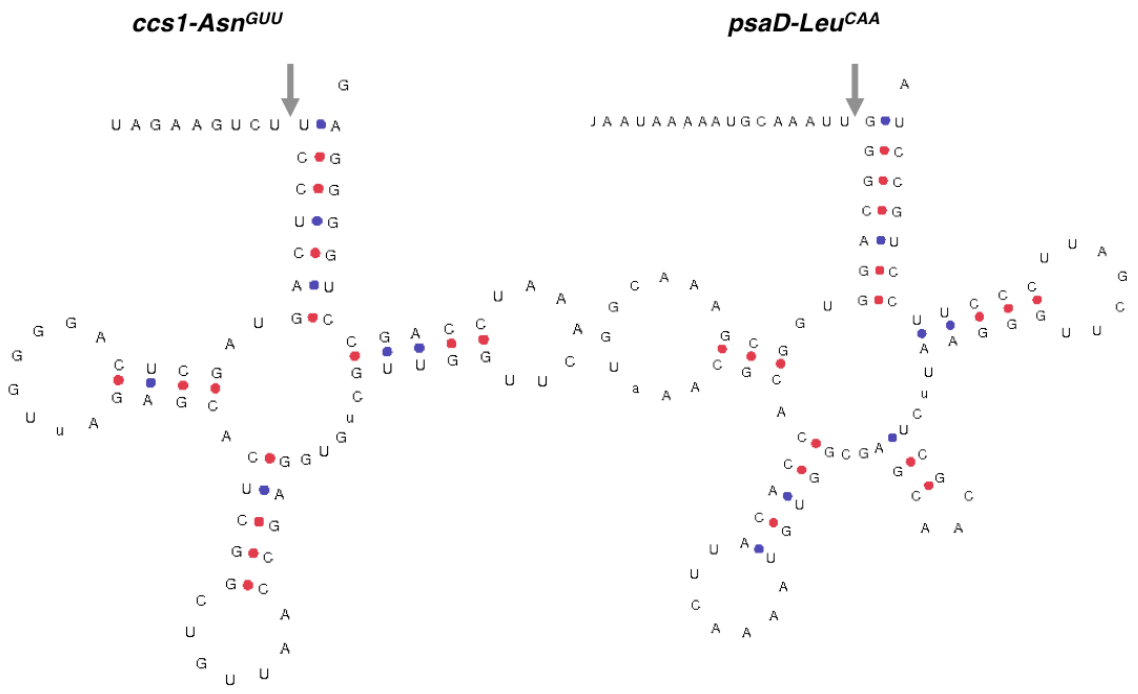


Fig. 7.4: tRNA-associated poly(U) sites in *Vitrella brassicaformis*.

These diagrams show the 3' UTRs of transcripts for the *V. brassicaformis ccs1* and *psaD* genes, as defined by oligo-d(A) RT-PCR. Grey arrows show the associated poly(U) sites for each gene; the poly(U) tail is not directly shown. In both genes, the poly(U) site is positioned immediately upstream of a tRNA gene (respectively Asn^{GUU} and Leu^{CAA}). The position and structure of each tRNA, as predicted by the tRNAscan-SE server (Lowe and Eddy, 1997) is shown for each transcript sequence.

changes to GC or purine/pyrimidine content, or predicted secondary structures that were universally associated with the presence of poly(U) sites (data not shown).

Ten of the twenty-four poly(U) sites in *V. brassicaformis*, including four sites associated with non-photosynthesis genes (*ccs1, chlN, rps4, rps16*) were immediately adjacent to the 5' end of predicted tRNAs (Fig. 7.4, Table 7.3). This suggests that some of the poly(U) sites in *V. brassicaformis* are generated by the cleavage of downstream tRNAs from precursor transcripts. However, many of the poly(U) sites identified in *V. brassicaformis* were tRNA-independent, and only one poly(U) site in *C. velia* (associated with *atpH-1*) was adjacent to a tRNA gene, indicating that this feature is not likely to have been ancestrally associated with poly(U) tail addition in chromerid plastids (Table 7.2). Overall, other than proximity to a

photosynthesis gene, there are no sequence features that are universally associated with poly(U) sites in the plastid genomes of either chromerid species.

Poly(U) tail addition is associated with high levels of transcript abundance in *Chromera velia*

It has been suggested that the poly(U) tails found in peridinin dinoflagellate plastids may facilitate plastid gene expression, either by protecting transcripts from 3' end degradation (Barbrook *et al.*, 2012), or by enabling other transcript processing events, such as editing, that allow translation of a functional protein sequence (Dang and Green, 2009). I wished to determine whether poly(U) tail addition might therefore be associated with highly expressed genes in chromerid plastids.

A recent next generation sequencing study has identified substantial variation in the abundance of different transcripts produced from the *Chromera velia* plastid genome (Janouškovec *et al.*, 2013b). This variation in abundance even extended to transcripts of different genes from within individual operons, indicating it is at least in part dependent on differences in transcript processing over different genes (Janouškovec *et al.*, 2013b). Calculating from the quantitative read coverage data obtained in this study (Janouškovec *et al.*, 2013b), genes that possess poly(U) sites are significantly more highly expressed than those that do not (Table 7.4, Mann-Whitney test, $P < E\text{-}04$).

While there may be a general association between poly(U) tail addition and high levels of expression, other gene-specific factors are also likely to influence transcript abundance. Many of the most abundant transcripts in the *C. velia* plastid encode photosystem subunits (Janouškovec *et al.*, 2013b). The same situation has previously been identified in plant plastids, in which photosynthesis genes are generally more highly expressed than genes of non-photosynthetic function (Krause *et al.*, 2000; Nakamura *et al.*, 2003). The high abundance of polyuridylylated transcripts in *C. velia* might therefore be due to the fact that many of these transcripts encode photosynthesis proteins, as opposed to being directly due to the presence of a poly(U) tail (Fig. 7.2, Table 7.4).

Notably, several of the photosynthesis genes in the *C. velia* plastid genome are present in multiple copies, or as multiple gene fragments (Janouškovec *et al.*, 2010; Janouškovec *et al.*, 2013b). These multiple copy genes present an ideal system in which to investigate differences in transcript processing between genes of closely related function. The *psaA* and *atpB* genes are each split into two functional units, which encode separate parts of the mature protein sequence (Janouškovec *et al.*, 2010; Janouškovec *et al.*, 2013b). Each of the

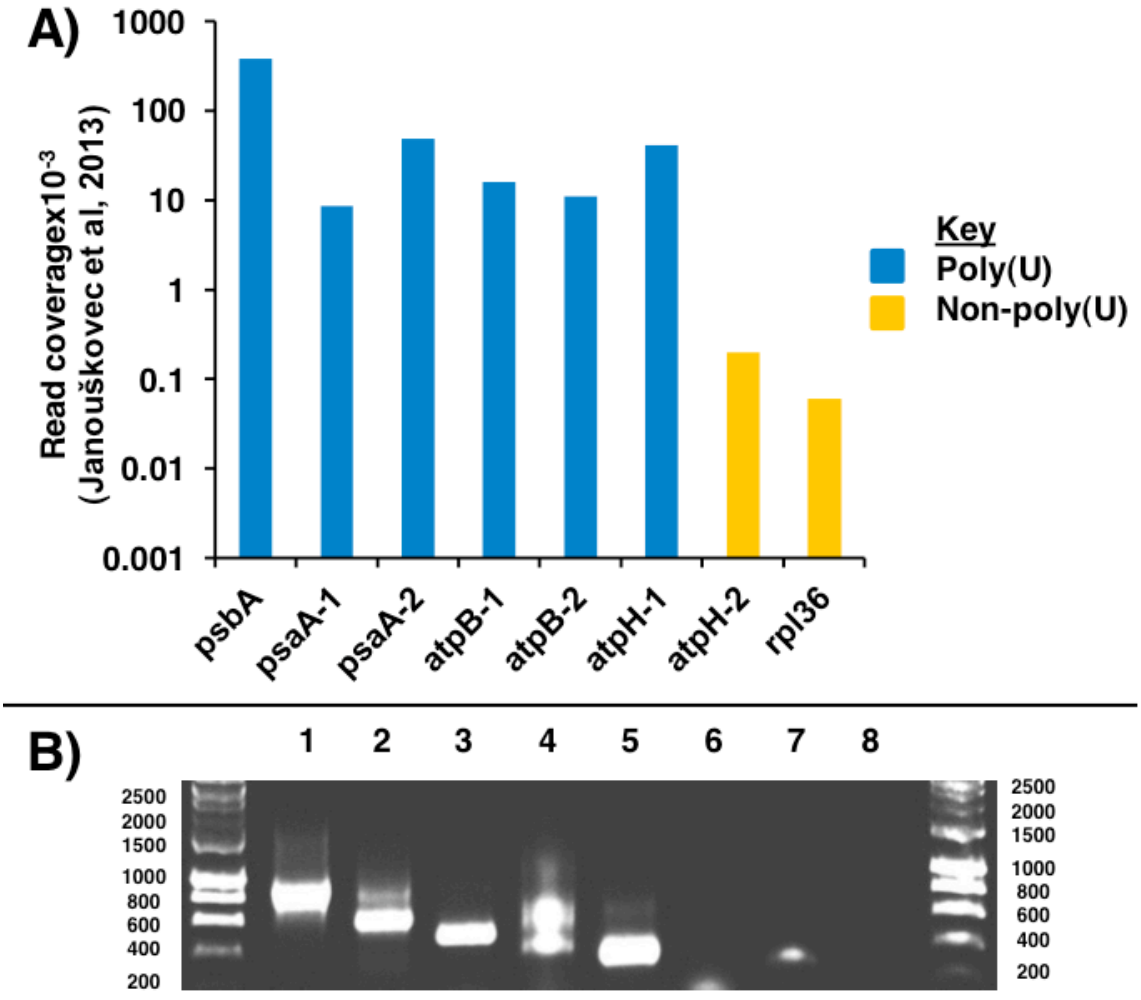


Fig. 7.5: Polyuridylylation of duplicated photosynthesis gene transcripts in the *C. velia* plastid.

Panel A shows the abundance of transcripts for duplicated genes in the *C. velia* plastid in quantitative read data obtained by Janouškovec et al. (Janouškovec *et al.*, 2013b). The most (*psbA)* and least abundant transcripts encoding recognisable proteins *(rpl36)* identified in this study are also shown. Polyuridylylated transcripts are shaded in blue and non-polyuridylylated transcripts in orange.

Panel B shows a gel photo of oligo-d(A) RT-PCR products for (**lanes 1-6**) *psaA-1*, *psaA-2*, *atpB-1, atpB-2, atpH-1* and *atpH-2.* Poly(U) tails were found on transcripts of both *psaA* genes, both *atpB* genes and *atpH-1*, but polyuridylylated *atpH-2* transcripts were not found. Lane **7:** gene-specific RT-PCR for *atpH-2*, demonstrating that non-polyuridylylated *atpH-2* transcripts are present. Lane **8**: template negative control.

Table 7.5. Primers for circular RT-PCR of *Chromera velia* plastid transcripts

Reaction	cDNA primer	PCR reverse primer	PCR forward primer
1. Monocistronic			
atpB(2)	CACCAAGTGCTACCCTCAT	TGCGGTACGGGGTTAC	AAAAGAGAAAGCGCAGATC
atpH(2)	TCCTTGTTCTTGGTGCG	TCCTTCCAGTTGACGGC	TCTGGTATCTCCAACTTATTTGG
atpI	TGCGGTAGCAGCTGTTAAG	GTTTCTCCATGAAATGTAAAATTAC	GCGGATGAGTTGACAGG
petB	ACTGGGCCGATGAAGG	TATTTAAGCCTATTCTTTTGTTTAACC	GTCATTGCCCTGTTAGCAC
psbA	CCAGCCATGTGGAAAGG	GCGGCTACAAAAGCAGC	GTAACGCGCACAATTTCC
psbH	CCTTGAAAAAAATCACTAAACG	CAGCATCGTTGGTACTCC	TTATCACTGTCGACTGAGCC
rps14	CACGCGAATAGTTTGAACC	CCTCATGAAGGTTGTATTGC	AAAACTTACCGGGTTTTCG
rps18-A	TGCCCCGACAGTGTC	CCAACTCGTCTAATGCAGC	CCAAGCTTTGAAAACGC
rps18-B	TGCCCCGACAGTGTC	CCAACTCGTCTAATGCAGC	CCGTCGTAAATATGAAGAACG
2. Dicistronic			
petB-psbH	ACTGGGCCGATGAAGG	TATTTAAGCCTATTCTTTTGTTTAACC	TTATCACTGTCGACTGAGCC
rps14-atpI	CACGCGAATAGTTTGAACC	CCTCATGAAGGTTGTATTGC	GCGGATGAGTTGACAGG

psaA and *atpB* gene fragments give rise to highly abundant transcripts, which are not *trans*-spliced together, and are instead separately translated to form distinct and presumably functionally active proteins (Fig. 7.5, panel A) (Janouškovec *et al.*, 2013b). Consistent with the high levels of transcript abundance, poly(U) tails were detected on both *psaA* and both *atpB* transcripts (Fig. 7.5, panel B; lanes 1-4).

The *atpH* gene is present in two paralogous copies on the *C. velia* plastid genome, with very different expression levels. *atpH-1* encodes a relatively conventional ATP synthase CF0 c subunit, and transcripts of this gene are abundant (Fig. 7.5, panel A) (Janouškovec *et al.*, 2013b). In contrast, *atpH-2* contains a 3' in-frame insertion, encoding a novel 89 aa C-terminal extension not found in any other annotated sequence, which is likely to render it non-functional (Janouškovec *et al.*, 2013b). Transcripts of *atpH-2* are the least abundant photosynthesis gene transcripts within the *C. velia* plastid and are only marginally more abundant than *rpl36*, the least abundant transcript of recognisable protein-coding function (Fig. 7.5, panel A) (Janouškovec *et al.*, 2013b). Notably, while transcripts of *atpH-1* receive a poly(U) tail, transcripts of *atpH-2* do not (Fig. 7.5, panel B; lanes 5-8). The loss of a poly(U) site from the *atpH-2* gene, which is associated with a much lower level of transcript abundance than the polyuridylylated *atpH-1* gene, strongly indicates that the presence of a poly(U) tail is associated with high levels of gene expression in chromerid plastids.

Relative extent of poly(U) tail addition to chromerid plastid transcripts

It has previous been shown that while effectively every gene in dinoflagellate plastids can give rise to polyuridylylated transcripts, low levels of non-polyuridylylated transcripts are also present (Barbrook *et al.*, 2012; Dorrell and Howe, 2012a). I wished to determine the relative

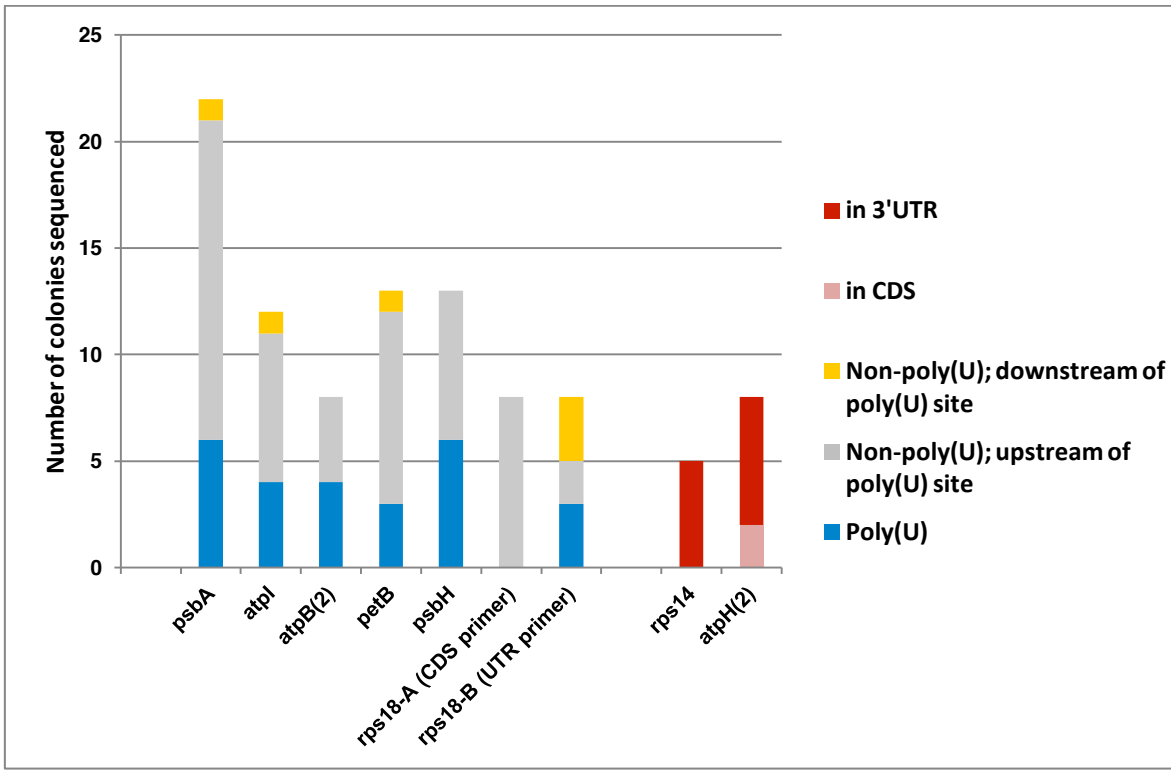


Fig. 7.6: 3' terminus positions of *Chromera velia* circular RT-PCR products.

This graph shows the different 3' terminus positions identified for transcripts of six genes identified to possess poly(U) sites (*psbA, atpI, atpB(2), petB, psbH,* and *rps18*) and two that do not (*rps14, atpH(2)*) as previously inferred by oligo-d(A) RT-PCR. Detailed terminus positions data corresponding to these circular RT-PCRs is given in Table 7.5. *rps18* circular RT-PCRs performed using PCR primers internal to the CDS (*rps18-A*) are shown separately to those performed using a PCR primer positioned immediately adjacent to the *rps18* poly(U) site (*rps18-B*). Circular RT-PCR products from genes that possess adjacent poly(U) sites are shaded blue if they terminate in a poly(U) tail, grey if they do not possess a poly(U) tail but terminate upstream of polyuridylylated transcripts identified in the same circular RT-PCR (hence may correspond to the 3' end degradation products of polyuridylylated transcripts), and orange if they extend through the associated poly(U) site (hence are likely to have been generated via a poly(U)-independent 3' end maturation pathway). Circular RT-PCR products that lack adjacent poly(U) sites are shaded pale red if they terminate within the 3' end of the CDS (i.e. are not translationally competent at the 3' end), and dark red if they extend in the 3' UTR (i.e. the processing state of the 3' end should enable translation of the CDS).

extent of poly(U) tail addition in chromerid plastids. In particular, I wished to identify whether any of the polyuridylylated non-photosynthesis gene transcripts identified by RT-PCR were abundant components of the chromerid plastid transcriptome (Fig. 7.2).

To quantify the relative abundance of polyuridylylated plastid transcripts, RT-PCRs were performed using circularised RNA for a range of plastid genes in *Chromera velia.* For each gene, cDNA was generated using a gene-specific cDNA synthesis primer positioned internal to the coding sequence (CDS), an outward-directed PCR reverse primer that annealed to the CDS 5' end, and a PCR forward primer that annealed to the CDS 3' end (Table 7.5). Six genes known to possess poly(U) sites were tested; five photosynthesis genes (*psbA, petB, psbH, atpB-2* and *atpI*) and *rps18*, one of only two non-photosynthesis genes found for which polyuridylylated transcripts were identified by oligo-d(A) primed RT-PCR in both *C. velia* and *Vitrella brassicaformis* (Fig. 7.2). In addition, circular RT-PCRs were performed for two genes (*rps14, atpH-2)* that were not found to give rise to polyuridylylated transcripts by oligo-d(A) RT-PCR (Fig. 7.2).

Consistent with the oligo-d(A) RT-PCR data, transcripts of *C. velia psbA, atpB-2, atpI, petB* and *psbH* that possessed 3' terminal, homopolymeric poly(U) tails were identified through this approach (Fig. 7.6; Table 7.6). Only two of the polyuridylylated transcripts identified by circular RT-PCR, out of a total of 27 sequenced, contained any nucleotides other than uridine within the 3' tail, indicating that heteropolymeric tails are extremely rare in chromerid plastids, and no other forms of 3' terminal modification were identified on any other transcript (Table 7.6). Non-polyuridylylated transcripts were identified for several photosynthesis genes, but almost all of these transcripts terminated either within the CDS or upstream of the poly(U) site, hence they may be the degradation products of previously polyuridylylated transcripts (Fig. 7.6; Table 7.6). Within the five polyuridylylated photosynthesis genes, only three transcripts were identified (one transcript each for *psbA, petB* and *atpI*) that extended past the corresponding poly(U) site (Table 7.6). Thus, the majority of photosynthesis gene transcripts in the *C. velia* plastid are likely to receive 3' poly(U) tails during processing.

None of the *atpH-2* and *rps14* transcripts identified through circular RT-PCR were found to possess poly(U) tails or any other form of terminal modifications, although many of these transcripts extended into the 3' UTR of the gene (Fig. 7.6; Table 7.6). Surprisingly, a circular RT-PCR using primers internal to the *rps18* gene failed to identify any polyuridylylated transcripts (although their existence was indicated by oligo-d(A) RT-PCR), but instead recovered large numbers of transcripts that terminated within the 3' UTR, upstream of the previously identified consensus poly(U) site (Fig. 7.6; Table 7.6). Polyuridylylated *rps18* transcripts could only be identified through circular RT-PCR by using a PCR forward primer

Table 7.6. Tabulated circular RT-PCR data for *Chromera velia*.

Terminus positions of each transcript are given relative to the corresponding CDS. *rps18* transcripts are separated into those amplified using PCR primers internal to the *rps18* CDS (Series A) and those amplified using a PCR forward primer positioned within the 3' UTR, immediately upstream of the poly(U) site (Series B). Transcripts that are of an equivalent length to hybridisation visible in northern blots (Figs. 7.8, 7.10) are shown in bold text.

Transcript	5' end	3' end	Poly(U)	Length (bp)	Notes
1. poly(U) genes					
psbA					
Non-poly(U) transcript 1	-39	-20	0	1050	
Non-poly(U) transcript 2	-33	-6	0	1058	
Non-poly(U) transcript 3	27	15	0	1019	
Non-poly(U) transcript 4	33	-18	0	980	
Non-poly(U) transcript 5	37	36	0	1030	
Non-poly(U) transcript 6	37	11	0	1005	
Non-poly(U) transcript 7	51	38	0	1018	
Non-poly(U) transcript 8	65	-22	0	944	
Non-poly(U) transcript 9	100	-1	0	930	
Non-poly(U) transcript 10	101	-1	0	929	
Non-poly(U) transcript 11	105	78	0	1004	3' end extends through poly(U) region
Non-poly(U) transcript 12	271	11	0	771	
Non-poly(U) transcript 13	279	0	0	752	
Non-poly(U) transcript 14	285	-6	0	740	
Non-poly(U) transcript 15	340	-19	0	672	
Non-poly(U) transcript 16	373	-9	0	649	
Poly(U) transcript 1	**-42**	**71**	**3**	1147	
Poly(U) transcript 2	**-32**	**75**	**11**	1149	
Poly(U) transcript 3	34	65	4	1066	
Poly(U) transcript 4	34	43	31	1071	
Poly(U) transcript 5	105	69	17	1012	
Poly(U) transcript 6	161	75	11	956	
atpI					
Non-poly(U) transcript 1	-173	175	0	1070	3' end extends through poly(U) region
Non-poly(U) transcript 2	-35	-117	0	640	
Poly(U) transcript 1	-35	-35	0	722	
Poly(U) transcript 2	**-35**	**156**	**6**	**913**	**3' end extends 28bp into rpl11**
Non-poly(U) transcript 5	-35	83	6	840	
Non-poly(U) transcript 4	-34	-120	0	636	
Poly(U) transcript 3	**-12**	**150**	**7**	**884**	**3' end extends 22bp into rpl11**
Poly(U) transcript 4	**-12**	**150**	**2**	**884**	**3' end extends 22bp into rpl11**
atpB(2)					
Poly(U) transcript 1	**-56**	**252**	**6**	**1759**	
Non-poly(U) transcript 1	-47	-256	0	1242	
Non-poly(U) transcript 2	-17	-82	0	1386	
Poly(U) transcript 2	**3**	**135**	**8**	**1583**	
Non-poly(U) transcript 3	27	3	0	1427	
Poly(U) transcript 3	**88**	**145**	**9**	**1508**	
Poly(U) transcript 4	**94**	**374**	**6**	**1731**	
Non-poly(U) transcript 4	603	99	0	947	

Table 7.6 (continued)

Transcript	5' end	3' end	Poly(U)	Length (bp)	Notes
petB					
Non-poly(U) transcript 1	-747	2	0	1387	5' end extends 149bp into petG
Non-poly(U) transcript 2	-18	339	0	995	3' end extends through poly(U) region, 168bp into psbH
Non-poly(U) transcript 3	4	85	0	719	3' end may extend through poly(U) region
Non-poly(U) transcript 4	7	69	0	700	3' end may extend through poly(U) region
Non-poly(U) transcript 5	26	-22	0	590	
Non-poly(U) transcript 6	38	-24	0	576	
Non-poly(U) transcript 7	64	53	0	627	
Non-poly(U) transcript 8	81	3	0	560	
Non-poly(U) transcript 9	124	17	0	531	
Non-poly(U) transcript 10	125	169	0	682	
Poly(U) transcript 1	123	180	6	695	3' end extends 9bp into psbH
Poly(U) transcript 2	-9	188	9	835	3' end extends 17bp into psbH
Poly(U) transcript 3	**-498**	**62**	**10**	**1198**	**Poly(U) tail contains G**
psbH					
Non-poly(U) transcript 1	-432	-13	0	673	5' end extends 261bp into petB
Poly(U) transcript 1	**-398**	**79**	**11**	**731**	**5' end extends 227bp into petB; 3' end 13bp into atpA**
Non-poly(U) transcript 2	-40	-99	0	195	
Non-poly(U) transcript 3	-40	-90	0	204	
Non-poly(U) transcript 4	27	-81	0	146	
rps18	**RT-PCR A**				
Non-poly(U) transcript 1	-130	-13	0	-117	
Non-poly(U) transcript 2	-129	320	0	-449	
Non-poly(U) transcript 3	144	-58	0	202	
Non-poly(U) transcript 4	274	150	0	124	
Non-poly(U) transcript 5	337	271	0	66	
Non-poly(U) transcript 6	337	271	0	66	
Non-poly(U) transcript 7	344	203	0	141	
Non-poly(U) transcript 8	396	150	0	246	
	RT-PCR B				
Non-poly(U) transcript 1	-131	369	0	-500	3' end extends 10bp into rps11
Poly(U) transcript 1	-128	372	7	-500	3' end extends 13bp into rps11
Poly(U) transcript 2	-131	351	11	-482	
Non-poly(U) transcript 2	114	409	0	-295	3' end extends through poly(U) region, 50bp into rps11
Poly(U) transcript 3	-55	369	8	-424	3' end extends 10bp into rps11
Non-poly(U) transcript 3	112	361	0	-249	
Non-poly(U) transcript 4	114	409	0	-295	3' end extends through poly(U) region, 50bp into rps11
Non-poly(U) transcript 5	-133	397	0	-530	3' end extends through poly(U) region, 38bp into rps11
2. non-poly(U) genes					
atpH-(2)					
Non-poly(U) transcript 1	-47	77	0	633	
Non-poly(U) transcript 2	-35	-44	0	500	
Non-poly(U) transcript 3	**-35**	**368**	**0**	**912**	**3' end extends 83bp into psbA**
Non-poly(U) transcript 4	-35	-44	0	500	
Non-poly(U) transcript 5	-35	77	0	621	
Non-poly(U) transcript 6	-35	249	0	793	
Non-poly(U) transcript 7	**-35**	**1023**	**0**	**1567**	**3' end extends 738bp into psbA**
Non-poly(U) transcript 8	**-35**	**1023**	**0**	**1567**	**3' end extends 738bp into psbA**
rps14					
Non-poly(U) transcript 1	-20	170	0	489	3' end extends 102bp into atpI
Non-poly(U) transcript 2	4	52	0	347	
Non-poly(U) transcript 3	69	4	0	234	
Non-poly(U) transcript 4	77	5	0	227	
Non-poly(U) transcript 5	86	115	0	328	3' end extends 47bp into atpI

Table 7.6 (continued)

Transcript	5' end	3' end	Poly(U)	Length (bp)	Notes
3. dicistronic transcripts					
rps14-atpI					
Poly(U) transcript 1	-10	149	13	1248	3' end extends 21bp into rpl11
Non-poly(U) transcript 1	4	-12	0	1073	
Non-poly(U) transcript 2	4	-12	0	1073	
Non-poly(U) transcript 3	58	-22	1009	0	
petB-psbH					
Poly(U) transcript 1	**-297**	**243**	**5**	**1603**	**3' end extends 177bp into atpA; Poly(U) tail contains C**
Non-poly(U) transcript 1	-270	-114	0	1219	
Non-poly(U) transcript 2	-229	-92	0	1200	
Poly(U) transcript 2	-3	4	7	1070	
Poly(U) transcript 3	3	79	7	1139	3' end extends 13bp into atpA
Non-poly(U) transcript 3	24	28	0	1067	
Poly(U) transcript 4	53	94	5	1104	3' end extends 28bp into atpA
Poly(U) transcript 5	183	580	19	1460	

that annealed directly upstream of the *rps18* poly(U) site, thus biasing the PCR for transcripts that extended at least as far as the poly(U) site (Table 7.5). However, using this primer, equal numbers of non-polyuridylylated transcripts were identified that extended past the consensus poly(U) site (Fig. 7.6; Table 7.6). This suggests that the effective concentration of polyuridylylated *rps18* transcripts was very low. Therefore, while *rps18* and some other non-photosynthesis genes may possess poly(U) sites that are detectable by oligo-d(A) RT-PCR, the vast majority of transcripts of these genes do not receive poly(U) tails. Thus, poly(U) tails are preferentially associated with the processing of photosynthesis gene transcripts in chromerid plastids.

Presence of polycistronic polyuridylylated transcripts in chromerid plastids

Polycistronic polyuridylylated transcripts have been identified in a wide range of dinoflagellate plastids (Barbrook *et al.*, 2012; Dang and Green, 2010; Richardson *et al.*, 2014). Many of the plastid genes investigated by oligo-d(A) RT-PCR in *Vitrella brassicaformis*, were found to give rise to polycistronic polyuridylylated transcripts (Table 7.3). Some genes, for which monocistronic products were not detected by oligo-d(A) RT-PCR, were found instead to produce polycistronic polyuridylylated products, with the poly(U) site in the 3' UTR of the gene furthest downstream. These polycistronic polyuridylylated products extended over two genes (e.g. *ycf4-psbE*) and in one case, even over four genes (*rps14-psbV-ccsA-psbK*) (Table 7.2).

I wished to determine whether polyuridylylated polycistronic transcripts are also present in *Chromera velia*. To do this, three plastid loci were selected *(atpH2-psbA, ORF207-atpB2,*

Table 7.7. Primers for RT-PCRs of polyuridylylated dicistronic transcripts in *C. velia*

Oligo-d(A)	GGGACTAGTCTCGAGAAAAAAAAAAAAAAAAA	
Transcript	**PCR forward primer**	**PCR reverse primer**
atpH(1)-psbA	GAAAGCAATCGAGCCTTG	CAACTGGTGCAGAGAAAGC
rps14-atpI	CAAAAATCTCTAGAGCGAATAAAG	CTCCAAATAAAAGCTTCACCC
ORF230-atpB(2)	TCAGCAGGGCCCAAAG	TGCGGTACGGGGTTAC
petG-petB	AACGAACCTCTTTTGTTTGG	ACTGGGCCGATGAAGG
petB-psbH	GACAGGAGCAGCAATGAC	AGTTACAGGTGTAGGGTCCC

Table 7.8. Northern probes for *C. velia* plastid transcripts.

This table lists the sequence of the T7 arm of the pGEM-T Easy vector, alongside the first 50 bp of each probe sequence complementary to plastid genes. The terminus positions of each probe are given relative to the corresponding CDS, except for the intergenic probe, where the positions are given relative to the 3' end of *rps14*.

T7 arm TAATACGACTCACTATAGGGCGAATTGGGCCCGACGTCGCATGCTCCCGGCCGCCATGGCCGCGGGATT

Probe	Start	End	Probe sequence
psbA	943	178	CTCGAAATTACACGTCCTTCTGCATCAATGATTGATTGGTTGAAATTGAA...
atpI	282	46	TCCAAATAAAAGCTTCACCCATATTACAAGGATGTAGCAAATTAAGATTTGTGTTC...
atpB-2	402	-226	CACCAAGTGCTACCCTCATTCGAGCCGCAGGGGTTTCATTCATTTGTCCA...
petB	159	415	ACTGGGCCGATGAAGGGAATAACTTCGGGCACACCCGTTACAATTTTACA...
psbH	251	28	TAGGCTCAGTCGACAGTGATAAAGTCCACTTGTACCGTTTGATTGCAAAT...
atpH-2	379	112	TCCGGACAGAGAACATCGATAATATCAATAACAGAGGCATATGAAAAAGC...
rps14	283	11	AAAACCCGGTAAGTTTTGGGTATGAATCATTTTTTTTAGATAATGTCGAG...
intergenic rps14-atpI	140	-108	GTTTCTCCATGAAATGTAAAATTACAAATATTTCAACAATAATGAACGGC...

rps14-atpI), and RT-PCRs specific to polyuridylylated polycistronic transcripts were performed using oligo-d(A) primed cDNA, and PCR primers that would amplify a region spanning the upstream and downstream genes of each locus (Table 7.7). In each case, products were obtained (Fig. 7.7). In addition, polyuridylylated dicistronic transcripts for *rps14-atpI* were identified by circular RT-PCR using cDNA synthesis and PCR reverse primers specific to *rps14*, and a PCR forward primer specific to *atpI* (Tables 7.1, 7.5). Thus, cotranscription is extensive across the chromerid plastid genome, and many polycistronic transcripts can receive poly(U) tails. This has subsequently been confirmed by an independently conducted study (Janouškovec *et al.*, 2013b).

Poly(U) tail addition is associated with transcript cleavage

A recent northern blotting study of *Chromera velia psaA-1* and *psaA-2* detected abundant polycistronic transcripts (Janouškovec *et al.*, 2013b). I wished to determine whether this was

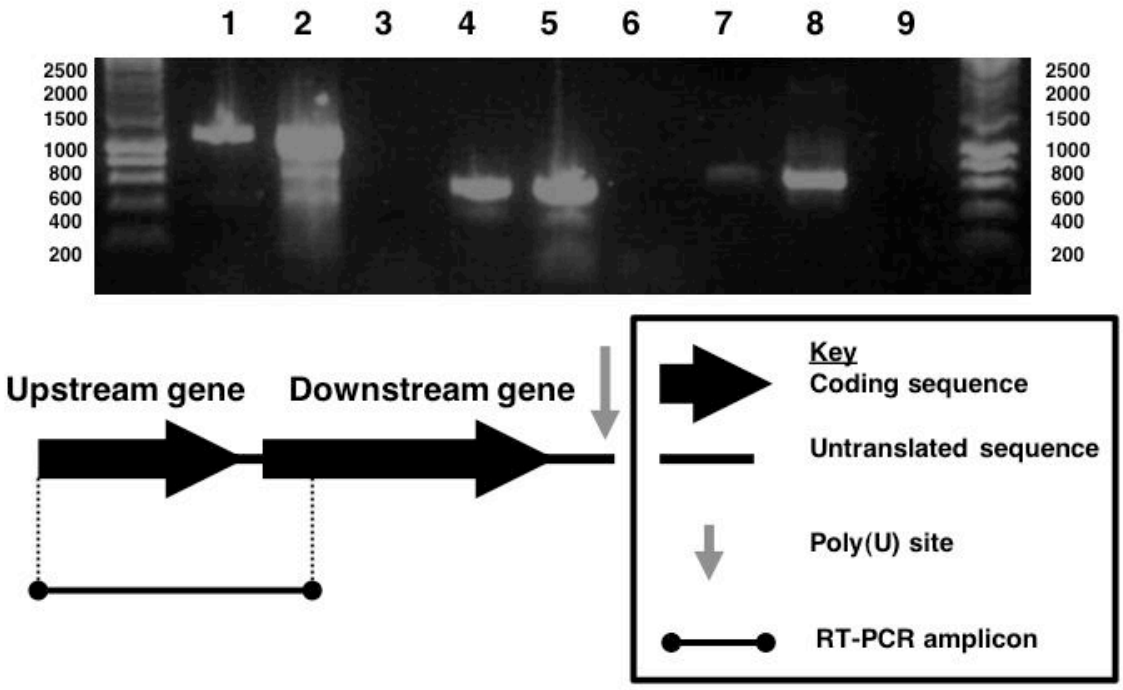


Fig. 7.7: Polycistronic polyuridylylated transcripts in *C. velia*.

This gel photo shows RT-PCRs to detect polyuridylylated dicistronic *atpH(2)-psbA, ORF230-atpB(2) and rps14-atpI* transcripts. A diagram of the RT-PCRs performed is shown beneath the gel photo. Each locus consists of an upstream gene that lacks an associated poly(U) site and a downstream gene that contains an associated poly(U) site. For each locus, PCRs were performed using a forward primer against the 5' end of the upstream gene and a reverse primer internal to the downstream gene. **Lanes 1-3**: PCR over the *atpH(2)-psbA* intergenic region using lane **1**, oligo-d(A) cDNA; lane **2**, gDNA; lane **3**, template negative conditions. Lanes **4-6**: as lanes 1-3, with *ORF230-atpB(2)*. Lanes **7-9**: as lanes 1-3, with *rps14-atpI*.

a universal feature of chromerid plastid transcript processing, or whether, for certain chromerid plastid genes, the predominant transcripts were monocistronic. I also wished to determine whether there were any differences in the transcript cleavage events associated with genes that possessed poly(U) sites, and those that did not. To test this, northern blots of *C. velia* RNA were analysed using probes specific to transcripts of three genes that possess poly(U) sites (*psbA, atpI, atpB-2*) as well as two that do not (*atpH-2, rps14*) (Table 7.8). The *atpH-2* probe sequence was designed to cover the 3' end insertion unique to *atpH-2* gene, and therefore was not expected to cross-hybridise substantially with *atpH-1* transcripts.

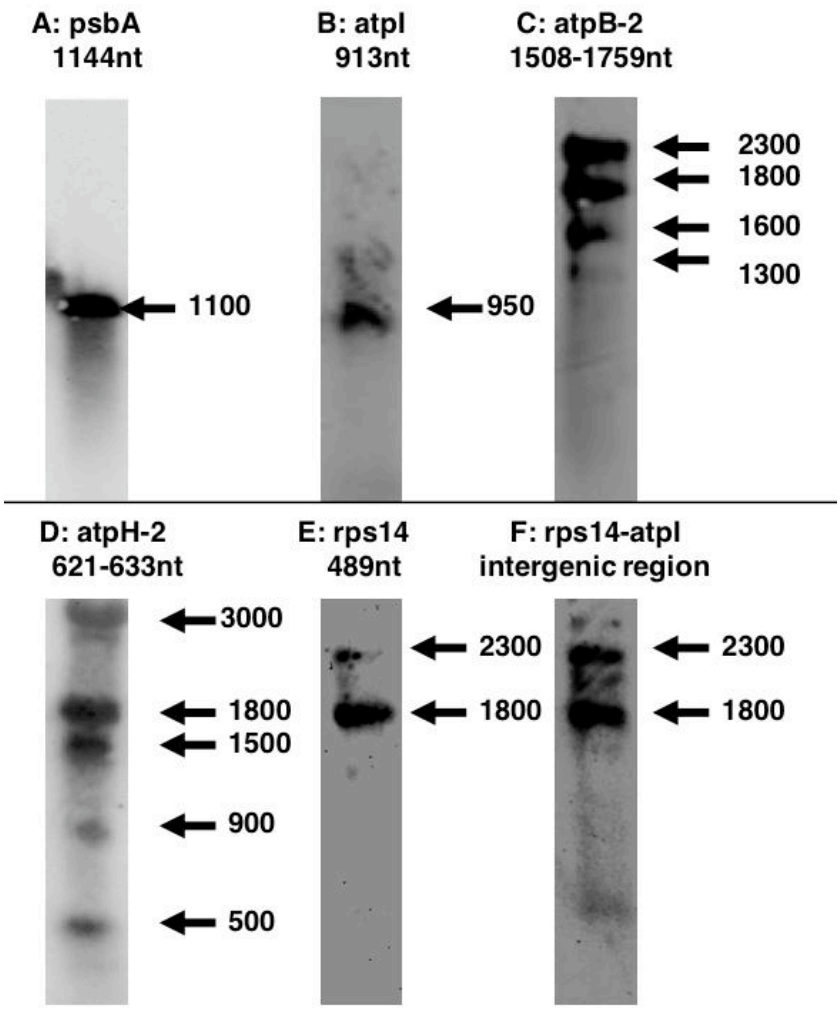


Fig. 7.8: Northern blots of *C. velia* plastid transcripts.

This figure shows the results of northern blots against a representative series of transcripts in the *C. velia* plastid. The sizes of monocistronic polyuridylylated transcripts (panels A-C), or of monocistronic non-polyuridylylated transcripts that cover the entire CDS (panels D-E), as obtained by circular RT-PCR (Table 7.6), are listed above the corresponding blot. **Panels A-C**: northern blots probed for *psbA, atpB-2* and *atpI*. Bands are broadly equivalent to the size of monocistronic transcripts as obtained by circular RT-PCR. **Panel D**: northern blot probed for *atpH-2*. Although a low abundance 500 nt band is present, the most intense bands are likely to correspond to polycistronic precursors, as obtained by circular RT-PCR, at ≥900 nt. **Panel E**: *rps14*, which lacks an associated poly(U) site. Bands of an equivalent size to a monocistronic *rps14* transcript are not detectable and instead, two higher molecular weight bands are observed, at 1700 nt and at 2000 nt. **Panel F**: northern blot probed with a probe overlapping the 3' end of *rps14* and the *rps14* 3' UTR, recovering bands of the same size as those in Panel D, indicating that *rps14* transcripts extend through this region.

For each of the polyuridylylated genes studied, the predominant bands in the northern blot corresponded to monocistronic transcripts. For *psbA*, a single band was observed corresponding to a 1100nt transcript, while for *atpI* a high intensity band was observed corresponding to a 950 nt transcript (Fig. 7.8, panels A, B). These agree with the sizes of monocistronic, polyuridylylated transcripts of each gene identified by circular RT-PCR (Table 7.6). It is possible that non-polyuridylylated *psbA* and *atpI* transcripts may also have been present in these bands. However, no non-polyuridylylated transcripts of either gene were identified by circular RT-PCR that were of an equivalent length to the bands visible in the northern blots (Table 7.6). For *atpB-2*, multiple bands were identified. The two high intensity bands at 1600 and 1800 nt are of equivalent size to monocistronic, polyuridylylated transcripts obtained by circular RT-PCR (Fig. 7.8, panel C; Table 7.6). A band of 2000 nt was additionally observed in the *atpB-2* northern blot (Fig. 7.8, panel C). Although transcripts of a similar length were not detected by circular RT-PCR, this band may correspond to monocistronic, polyuridylylated transcripts that extend to the most distant poly(U) site associated with *atpB-2*, positioned 467 nt into the 3' UTR (Fig. 7.3, Table 7.2). No non-polyuridylylated *atpB-2* transcripts of greater than 1500 nt length were identified by circular RT-PCR. However, circular RT-PCR did reveal the presence of non polyuridylylated *atpB-2* transcripts with 3' ends internal to the *atpB-2* CDS, which might correspond to a faint 1300 nt band detected in the northern blot (Fig. 7.8, panel C; Table 7.6).

Bands of a size consistent with polycistronic transcripts were not detected in either the *psbA* or *atpB-2* blots (Fig. 7.8, panel A). A 1400 nt band in the *atpI* northern blot might correspond to a polycistronic precursor, but this band was of much lower intensity than the band corresponding to the monocistronic transcript (Fig. 7.8, panel B). Overall, it appears that while polycistronic transcripts may be produced from many loci in chromerid plastids, the majority of the polyuridylylated transcripts present are monocistronic.

In contrast to the situation for *psbA, atpI* and *atpB-2*, no hybridisation was identified that corresponded to monocistronic transcripts in either the *atpH-2* or *rps14* northern blots (Fig. 7.8; panels D-E). The predominant bands in the *atpH-2* northern blot were 900 nt in length or greater (Fig. 8, panel D). The 900 and 1500 nt bands correspond in size to polycistronic *atpH-2* transcripts obtained by circular RT-PCR that extended well into the *psbA* CDS (Table 7.6). The *atpH-2* blot did not contain any hybridisation at a size (600 nt) corresponding to the monocistronic transcripts identified by circular RT-PCR. A low intensity band at 500 nt was identified, but the only transcripts identified of similar size through circular RT-PCR terminated at the 3' end within the *atpH-2* CDS. This band is therefore likely to correspond to

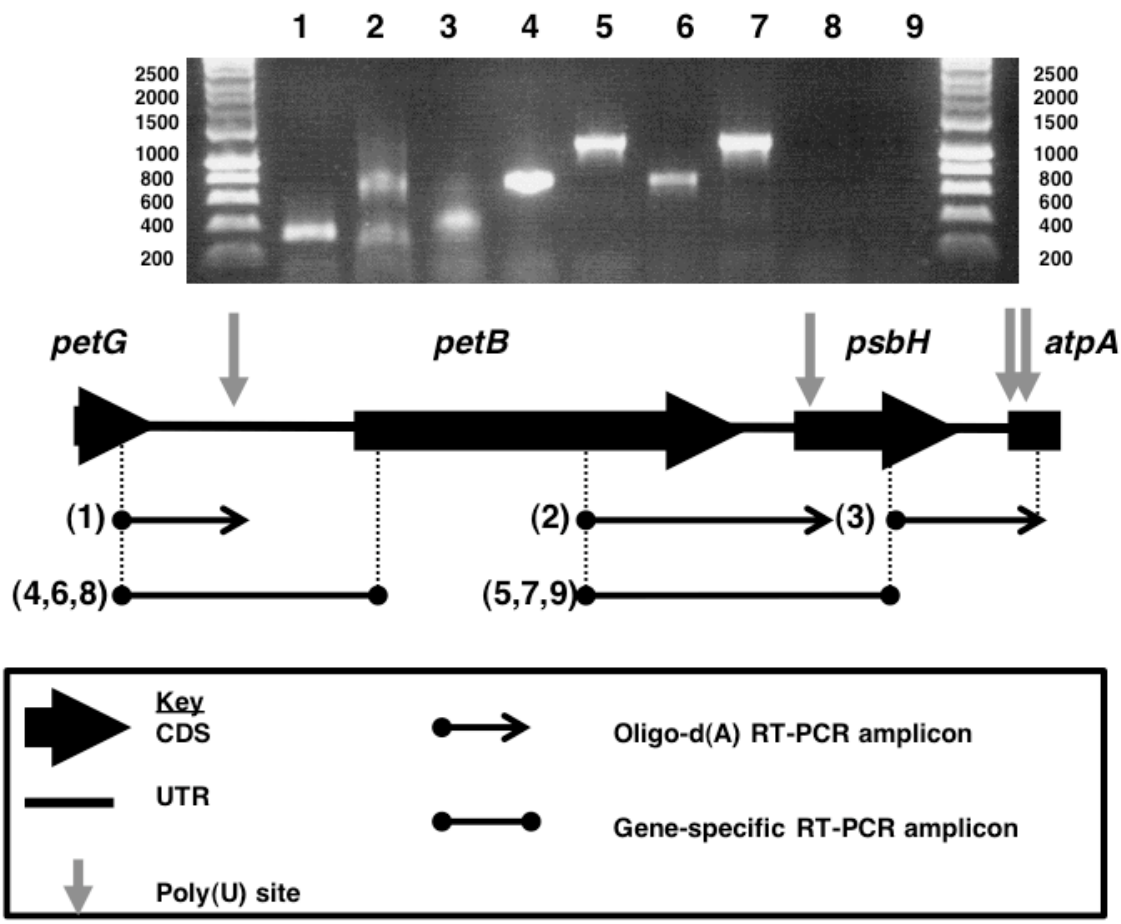


Fig. 7.9. Cotranscription of the *Chromera velia petG-petB-psbH* locus.

This gel photo shows the result of a series of RT-PCRs to detect monocistronic and polycistronic transcripts over the *C. velia petG-petB-psbH* locus. As per Fig. 7.7, a transcript diagram with each of the PCR amplicons tested is shown beneath the gel photo. **Lanes 1-3**: oligo-d(A) RT-PCR for *psbH, petB* and *petG* transcripts (all polyuridylylated). The poly(U) sites associated with the *petB* and *psbH* genes are positioned respectively inside the 5' ends of the *psbH* CDS and the *atpA* CDS, hence mature *petB, psbH* and *atpA* mRNAs cannot be generated from the same transcript. lanes **4-5**: oligo-d(A) RT-PCR for the intergenic *petG-petB* and *petB-psbH* regions, lanes **6-7**: PCR for the same intergenic regions using DNA template, lanes **8-9**: PCR for the same intergenic region using template negative conditions. The positive results for lanes 4-5 indicate that polycistronic transcripts are present covering this locus.

degraded transcripts, as opposed to translationally functional monocistronic mRNAs (Fig. 7.8, panel D; Table 7.6).

In the case of *rps14*, only bands at 1700 and 2000 nt were observed, far larger than the c. 500 nt monocistronic transcripts obtained by circular RT-PCR (Fig. 7.8, panel E). It appears that *rps14* transcripts may possess a lengthy 3' UTR, as similar bands were also recovered by a probe that spanned the 3' end of the *rps14* CDS and the downstream non-coding region between *rps14* and the adjacent *atpI* gene (Fig. 7.8; panel F). In total, whereas many of the polyuridylylated transcripts in chromerid plastids are monocistronic, the majority of the translationally functional transcripts containing non-photosynthesis and non-polyuridylylated genes are polycistronic. Thus, transcripts of polyuridylylated genes undergo more extensive terminal cleavage events than transcripts of genes that lack poly(U) sites. Poly(U) tail addition might accordingly be associated with directing further transcript cleavage events in chromerid plastids.

Transcripts in the *C. velia* plastid are subject to alternative processing

It has been suggested that poly(U) tail addition is involved in alternative processing events at several loci in dinoflagellate plastids (Barbrook *et al.*, 2012; Richardson *et al.*, 2014). At these loci, the poly(U) site associated with one gene is located within the mature transcript sequence of the gene located downstream. Processing of the poly(U) site would thus prevent the generation of a translationally functional transcript of the downstream gene from a common polycistronic precursor. Several polyuridylylated transcripts identified through oligo-d(A) RT-PCR, in both chromerid species, were found to extend at the 3' end into the downstream CDS (Tables 7.2, 7.3). I wished to determine whether these polyuridylylated transcripts are generated from the cleavage of longer, polycistronic precursors through alternative processing events.

The *C. velia petG-petB-psbH* locus was selected as a model system in which to investigate alternative processing events (Fig. 7.9). Each gene within this locus possesses a poly(U) site, as indicated by oligo-d(A) primed RT-PCR (Fig. 7.9, lanes 1-3; Table 7.2). The poly(U) sites associated with *petB* extend up to 27nt within the 5' end of *psbH*, hence it would be impossible to generate complete *psbH* transcripts from a polycistronic precursor that had already yielded a polyuridylylated *petB* transcript (Table 7.2). Similarly, the poly(U) sites associated with *psbH* are located up to 510 nt into *atpA*, which would prevent the production of *psbH* and *atpA* transcripts from the same precursor molecule (Table 7.2). Dicistronic *petG-petB* and *petB-psbH* transcripts were identified using similar nested oligo-d(A) primed RT-PCRs as before (Fig. 7.9, lanes 4-9; Table 7.7). In addition, polyuridylylated dicistronic *petB-psbH* transcripts were identified by circular RT-PCR using cDNA synthesis and PCR reverse primers specific to *petB*, and a PCR forward primer specific to *psbH* (Tables 7.5, 7.6). This indicates that *petB* and *psbH* are cotranscribed in the *Chromera velia* plastid.

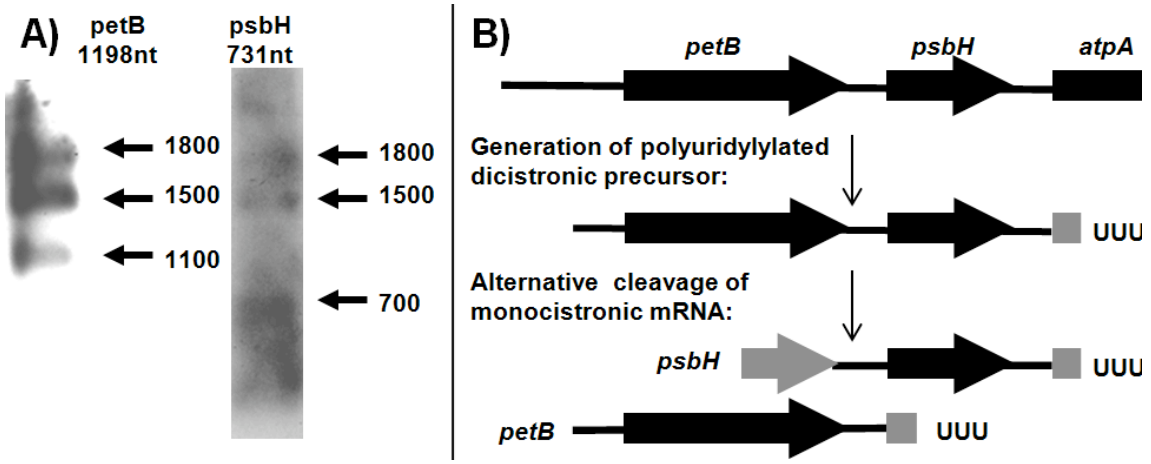


Fig. 7.10. Alternative processing events at the *C. velia petB-psbH* locus.

Panel A shows northern blots analysed using probes against the *petB* and *psbH* genes. Left DIG-labelled RNA ladder I (Roche) with sizes indicated. The sizes of monocistronic polyuridylylated transcripts as obtained by circular RT-PCR are listed above the corresponding blot. In each blot, two conserved higher molecular weight bands are present at 1600 and 1800 nt, which are likely to represent polycistronic precursors covering both the *petB* and *psbH* genes. In addition, lower molecular weight bands unique to either the *petB* (1100 nt) or *psbH* blots are observed (700 nt), consistent with monocistronic transcripts as recovered by circular RT-PCR.

Panel B shows a possible model for the alternative processing of transcripts over the *C. velia petB-psbH* locus. Each CDS is shown with a thick black arrow, and non-coding DNA is shown by thin black lines. Thick grey lines show incomplete CDS regions on transcript ends that have been generated by alternative processing. Vertical arrows show the likely progression of transcript processing events. The *petB* and *psbH* genes are initially cotranscribed from a promoter element located upstream of the *petB* gene, as part of a long polycistronic transcript that may also extend over the *petG* and *atpA* genes. The initial primary transcript generated is processed to form shorter precursors, such as a dicistronic polyuridylylated *petB-psbH* transcript that extends from the *petB* 5' UTR to a poly(U) site positioned downstream of *psbH*, within the *atpA* CDS . This dicistronic transcript may be cleaved to form monocistronic polyuridylylated *petB* or *psbH* transcripts. As the *petB* poly(U) site is positioned within the *psbH* CDS and the *psbH* 5' end is positioned within the *petB* CDS, mature *petB* and *psbH* transcripts cannot be generated from the same precursor, and thus are cleaved from different precursors via mutually exclusive processing steps.

To determine what cleavage events are associated with transcripts from this locus, northern blots were hybridised with probes for *C. velia petB* and *psbH* (Fig. 7.10). The *psbH* probe was positioned downstream of the *petB* poly(U) site and the *petB* probe was positioned at the 5' end of the CDS to minimise any potential overlap between probe sequences (Table 7.5). In contrast to previous observations for *psbA, atpI* and *atpB-2*, 1600 and 1800 nt bands were identified in both the *petB* and *psbH* blots that are likely to correspond to polycistronic transcripts covering both genes (Fig. 7.10, panel A). The 1600 nt band was of an equivalent size to polyuridylylated, dicistronic *petB-psbH* transcripts obtained by circular RT-PCR (Table 7.4, panel D). Lower molecular weight bands were also identified that were specific to either the *petB* or the *psbH* blots (Fig. 7.10, panel A). The 1100 nt band seen when probed for *petB* is similar in size to a monocistronic transcript, which possess a poly(U) site located in the *psbH* CDS, as obtained by circular RT-PCR (Fig. 7.10, panel A; Table 7.4, panel D). Similarly, the 700 nt band seen when probed for *psbH* is similar in size to a monocistronic polyuridylylated transcript sequenced by circular RT-PCR (Fig. 7.10, panel A; Table 7.4, panel D). This transcript overlaps at the 5' end with the *petB* CDS, as well as possessing a poly(U) site located in the *atpA* CDS, suggesting that both ends are alternatively processed. Thus, polycistronic precursors covering the *petB-psbH* locus may be cleaved into monocistronic mRNAs.

It is possible that, instead of being generated by the processing of common, polycistronic precursors, mRNAs in chromerid plastids that have overlapping terminus regions might be separately transcribed from different promoter sites in the 5'UTR of each gene, and accumulate as independent populations of transcripts. Although bacterial promoters were identified within the 5' UTR of both the *petB* and *psbH* genes (Table 7.4), the promoter located within the *psbH* 5' UTR would not give rise to the predominant monocistronic mRNAs identified by circular RT-PCR, as these extend into the *petB* CDS (Table 7.6). While it is possible that *psbH* transcripts are generated from a promoter internal to the *petB* CDS, internal promoter sites are uncommon for protein-coding genes in plant plastids, and generally appear to give rise only to very low levels of transcripts (Liere *et al.*, 2011; Vera *et al.*, 1992; Zhelyazkova *et al.*, 2012). Thus, the *petB* and *psbH* transcripts are most likely to be cotranscribed from a common promoter element upstream of the *petB* 5' end, and form dicistronic polyuridylylated precursors. In at least some cases, these precursors undergo alternative 5' and 3' cleavage events to generate monocistronic polyuridylylated *petB* and *psbH* mRNAs (Fig. 7.10, panel B). Poly(U) tail addition might thus direct alternative transcript processing events in chromerid plastids, specifying which mature mRNAs are produced from common polycistronic precursors.

Discussion

I have characterised the distribution (Figs. 7.1-7.6) and function (Figs 7.5, 7.7-7.10) of poly(U) tail addition across the plastid transcriptomes of the chromerid algae *Chromera velia* and *Vitrella brassicaformis*, which are closely related to parasitic apicomplexans. The poly(U) tail addition events found in chromerid plastids share some degree of similarity with those of dinoflagellates. The presence of multiple, alternative poly(U) sites within the 3' UTR of one gene has previously been observed in many dinoflagellate species (Fig. 7.3) (Barbrook *et al.*, 2012; Dorrell and Howe, 2012a; Nelson *et al.*, 2007; Wang and Morse, 2006). The association between poly(U) sites and tRNA cleavage in *V. brassicaformis* has also been identified in the dinoflagellate *Heterocapsa triquetra* (Fig. 7.4) (Dang and Green, 2009; Nelson *et al.*, 2007).

However, unlike in dinoflagellates, poly(U) tail addition occurs only on some of the transcripts in chromerid plastids. To date, only one protein-coding gene that lacks an associated poly(U) site has been identified in a peridinin dinoflagellate plastid - *petD* in *Amphidinium carterae* (Barbrook *et al.*, 2012). I have previously shown that the overwhelming majority of genes in fucoxanthin-containing dinoflagellate plastids, which have acquired the poly(U) tail addition machinery following their endosymbiotic replacement of the ancestral peridinin plastid lineage, likewise possess associated poly(U) sites in their 3' UTR (Dorrell and Howe, 2012a; Richardson *et al.*, 2014). Conversely, many of the protein-coding genes in both the *C. velia* and *V. brassicaformis* plastids lack an associated poly(U) site and these principally encode products that do not directly function in photosynthetic electron transfer (Figs. 7.1, 7.2). While a few photosynthesis genes were identified in either *C. velia* or *V. brassicaformis* that lacked poly(U) sites, and a few polyuridylylated non-photosynthesis gene transcripts were identified, very few of these exceptions were conserved between both species (Fig. 7.2). Furthermore, my *rps18* circular RT-PCR data suggest that at least some of the poly(U) sites associated with non-photosynthesis genes may not be processed on the majority of transcripts of these genes (Fig. 7.6; Table 7.6). Thus, poly(U) tail addition on chromerid plastid transcripts appears to be dependent on a photosynthetic function of the translation product. This is the first characterised plastid transcript processing pathway to preferentially target a particular functional category of genes.

With this in mind, the function of transcript poly(U) tail addition in chromerid plastids is particularly intriguing. Previously, I have shown that the poly(U) tail addition machinery in the fucoxanthin dinoflagellate *Karlodinium veneficum* can similarly discriminate between functional and pseudogene transcripts (Richardson *et al.*, 2014). Transcript processing complexes are known to be involved in negative regulation of non-functional transcripts in

other organelle lineages (such as transcripts of pseudogenes associated with cytoplasmic male sterility phenotypes in plant mitochondria, and antisense transcripts in plant plastids) (Chase, 2007; Sharwood *et al.*, 2011). Here, I demonstrate that poly(U) tail addition is not only associated with transcripts of functional photosynthesis genes, but poly(U) tail addition is associated with high levels of transcript abundance. For example, transcripts of a functional plastid photosynthesis gene (*atpH-1*), which are highly abundant in the *C. velia* plastid, receive poly(U) tails, whereas transcripts of an equivalent pseudogene (*atpH-2*), which are much less abundant, are not polyuridylylated (Fig. 7.5). Notably, the *C. velia atpH-2* CDS does not contain premature termination codons, or other features that would prevent its expression, and many of the *atpH-2* transcripts detected by circular RT-PCR covered the complete CDS, i.e. would be translationally competent (Table 7.6). The loss of a poly(U) site on the *atpH-2* transcript and consequent reduction in transcript abundance (Janouškovec *et al.*, 2013b), could minimise expression of *atpH-2* without inactivation of the underlying gene sequence. Similarly, the high expression level of photosynthesis gene transcripts in chromerid plastids, which has been suggested to enable rapid photo-physiological adaptation to changing light conditions, may be facilitated by poly(U) tail addition (Janouškovec *et al.*, 2013b; Quigg *et al.*, 2012). Overall, it appears that poly(U) tails is likely to be involved in the functional expression of photosynthesis genes in chromerid plastids.

One possible means for the poly(U) tail to facilitate plastid gene expression would be to direct specific cleavage events on precursor transcripts such as the polycistronic polyuridylylated transcripts identified in both species (Table 7.3; Figs. 7.7, 7.9). Studies in peridinin dinoflagellates have indicated that the addition of a poly(U) tail may be directly associated with the cleavage of polycistronic transcripts into monocistronic mRNAs (Dang and Green, 2010; Nisbet *et al.*, 2008), and I have previously presented evidence that poly(U) tail addition is associated with transcript 5' end cleavage in the fucoxanthin dinoflagellate *Karenia mikimotoi*. In the *C. velia* plastid, it appears that polyuridylylated transcripts undergo greater degrees of cleavage events, beyond the addition of a poly(U) tail, to non-polyuridylylated transcripts (Figs. 7.8). At loci such as *petG-petB-psbH* that contain multiple poly(U) sites, poly(U) tail addition might additionally be involved in specifying which mature mRNAs are produced through the alternative processing of common polycistronic precursors (Fig. 7.10). Poly(U) tail addition might thus play a similar role to alternative 3' end polyadenylylation in nuclear transcript processing, which may substantially alter the coding capacity and regulatory properties of nuclear transcripts (Wang *et al.*, 2008; Wu *et al.*, 2011).

Overall, my data indicate that poly(U) tail addition plays an important role in processing photosynthesis gene transcripts in chromerid plastids. This pathway presumably played a

similar role in early ancestors of apicomplexans, and its loss might underline key events in the evolution of parasitism in this lineage. It remains to be determined whether poly(U) tail addition was lost from early apicomplexans following the transition to parasitism, or whether the loss poly(U) tail addition may have preceded, and even facilitated the loss of photosynthesis pathways. Further analysis of the gene expression machinery of chromerids may provide important insights into the evolutionary steps that converted a photosynthetic alga into non-photosynthetic apicomplexan parasites.

Chapter Eight- Thesis Conclusions

Summary of thesis results

During my PhD, I have investigated the evolution and function of transcript processing in the extremely diverse plastids found within the alveolates. I have characterised plastid transcript processing events in the peridinin-containing dinoflagellate *Amphidinium carterae*, and the chromerid algae *Chromera velia* and *Vitrella brassicaformis,* as representatives of alveolate species possessing the ancestral, red algal derived plastid lineage. I have additionally characterised transcript processing events in representatives of all three dinoflagellate lineages documented to have replaced the ancestral plastid lineages with ones of alternative phylogenetic derivation, by serial endosymbiosis. These are the fucoxanthin-containing species *Karenia mikimotoi* and *Karlodinium veneficum* (possessing haptophyte-derived plastids), the "dinotom" *Kryptoperidinium foliaceum* (possessing diatom-derived plastids), and *Lepidodinium chlorophorum* (possessing green algal-derived plastids).

I have investigated the phylogenetic distribution of two unusual transcript processing pathways, 3' poly(U) tail addition, and sequence editing, across alveolate plastid lineages. From this, I wished to infer whether dinoflagellate plastids derived through serial endosymbiosis are supported by transcript processing pathways originating from the ancestral peridinin plastid lineage. I have also investigated which transcripts in each alveolate plastid lineage studied receive poly(U) tails. In particular, I wished to determine whether poly(U) tail addition is specifically associated with photosynthesis gene transcripts in chromerid algae, such that its absence from apicomplexan parasites, which are closely related to chromerids, may have occurred concurrent with the loss of photosynthesis genes from the apicoplast.

In addition, I have investigated whether the plastid genomes of individual alveolate lineages have undergone divergent evolutionary events since their endosymbiotic acquisition. For example, I wished to determine whether the plastid genomes of fucoxanthin dinoflagellates contain minicircle elements that have arisen independently to those observed in peridinin dinoflagellates. As a corollary of this, I have considered how divergent evolutionary events in individual alveolate plastid genomes, such as genome fragmentation into minicircles, the gain of novel in-frame sequence insertions, and the generation of pseudogenes, may have affected the associated transcript processing events observed. Finally, I have investigated the function of poly(U) tail addition in transcript processing in each alveolate plastid lineage. I have also examined the processing events associated with non-coding transcripts, such as antisense transcripts, present in alveolate plastids.

From my data, I can draw the following conclusions, which are discussed in more detail below:

1. **Poly(U) tail addition is preferentially associated with transcripts of photosynthesis genes in chromerid plastids**. This is the first plastid transcript processing pathway documented to preferentially target transcripts that are involved in a particular biochemical pathway.

2. **Poly(U) tail addition and editing occur in fucoxanthin dinoflagellate plastids.** This provides definitive proof that the biology of serially acquired plastids may be affected by pathways retained from previous symbioses.

3. **Fucoxanthin plastid genomes are highly divergently organised.** For example, I have provided the first complete evidence that the fucoxanthin plastid genome contains episomal minicircles, which have arisen in parallel to the minicircles in peridinin dinoflagellate plastids.

4. **Poly(U) tail addition and editing have been adapted to the divergent evolution of alveolate plastid genomes.** For example, I have provided the first evidence that poly(U) tail addition is preferentially associated with transcripts of functional genes, over transcripts of pseudogenes in alveolate plastid genomes.

5. **Poly(U) tail addition has complex and interconnected relationships to other events in plastid transcript processing.** I have identified potential functions for poly(U) tail addition in directing sequence editing and terminal processing events for specific alveolate plastid transcripts.

6. **Highly edited, but non-polyuridylylated antisense transcripts are present in dinoflagellate plastids.** These are the first documented antisense transcripts in algal plastid lineages.

In this chapter, I will discuss the significance of each of these discoveries for understanding the biochemical processes that underpin alveolate plastid gene expression, and the broader consequences of these discoveries for current theories of plastid evolution. I will additionally present schematic diagrams showing the taxonomic distribution of poly(U) tail addition and editing in the alveolates (Fig. 8.1), and their inferred functional roles in transcript processing (Figs. 8.2-6). I will conclude by providing a brief overview of experimental work that might further our understanding of transcript processing in alveolate plastid lineages.

Conclusion 1: poly(U) tail addition is preferentially associated with photosynthesis gene expression in chromerid plastids (Chapter Seven)

I have shown that poly(U) tail addition in chromerid algae is predominantly associated with transcripts that encode photosystem subunits, whereas transcripts of other genes (e.g. those encoding components of other biochemical pathways, or components of the plastid housekeeping machinery) do not give rise to abundant polyuridylylated transcripts. Although poly(U) sites have been lost from a small number of photosynthesis genes, and gained by a few non-photosynthesis genes, these may constitute only a small number of the total transcripts produced in chromerid plastids. Furthermore, while there are a large number of photosynthesis genes that possess associated poly(U) sites in both *Chromera velia* and *Vitrella brassicaformis*, there appears to be only a limited degree of overlap between the photosynthesis genes that lack poly(U) sites, or the non-photosynthesis genes that possess poly(U) sites, between the two chromerid species. Thus, poly(U) tail addition in chromerids is biased towards photosynthesis genes.

The preferential application of poly(U) tails to photosynthesis gene transcripts in chromerids may provide valuable insights into the mechanisms that allow the differentiation of different functional categories of plastid genes. Previous studies of the mechanisms that enable the recognition of specific plastid genes have focussed on gene expression in plants. In plants, two RNA polymerases are used: a polymerase of bacterial origin, encoded in the plastid, and a polymerase of phage origin, encoded in the nucleus, which is believed to have evolved from the phage-type, nucleus-encoded polymerase that operates in mitochondria (Kapoor *et al.*, 1997; Liere *et al.*, 2011; McBride *et al.*, 1994). The phage-type plastid polymerase is not associated with non-plant plastid lineages, in which only bacterial-type polymerases have been characterised (Teng *et al.*, 2013).

Initial studies of the function of each plant plastid RNA polymerase suggested that the plastid-encoded polymerase might play a specific role in the transcription of plastid photosynthesis genes (Allison *et al.*, 1996; Hajdukiewicz *et al.*, 1997). Suppression of the bacterial-type polymerase typically prevents the development of photosynthetically functional tissue (Zhelyazkova *et al.*, 2012). In addition, tissue in which the plastid-encoded polymerase is not expressed only produce limited quantities of transcripts of plastid photosynthesis genes (which are otherwise typically highly abundant), whereas the levels of transcripts of plastid non-photosynthesis genes (which are typically only produced at low levels) are not substantially affected by the absence of the plastid-encoded polymerase (Allison *et al.*, 1996; Hajdukiewicz *et al.*, 1997). However, more recent studies that have characterised the promoter sequences recognised by each plant plastid RNA polymerase has revealed that the

majority of plastid genes have indicated that the majority of plastid genes may be transcribed by both types of polymerase, although a small number of non-photosynthesis genes solely possess promoters for the nucleus-encoded polymerase, and a few genes of both photosynthesis and non-photosynthesis function may be specifically transcribed by the plastid-encoded polymerase (Swiatecka-Hagenbruch *et al.*, 2007; Zhelyazkova *et al.*, 2012). The plastid-encoded polymerase appears to be required for photosynthetic function not because it solely transcribes photosynthesis genes, but because it is itself only expressed in significant quantities in photosynthetically active tissue (Liere *et al.*, 2011). Thus, the exact significance of transcription in the differential regulation of photosynthesis versus non-photosynthesis genes in plant plastids remains uncertain.

Similarly to the situation for transcription, it is not clear whether there is any clear division of labour between the targets of different transcript processing events in plants. Certain events, such as transcript cleavage, appear to be ubiquitous features of plastid transcript processing (Barkan, 2011; Stern *et al.*, 2010). Other processing events (such as editing, and *cis*- and *trans*-splicing), while not universal features, occur on a wide functional range of plastid transcripts (Fujii and Small, 2011; Glanz and Kück, 2009; Tillich and Krause, 2010). The predominant application of poly(U) tails to photosynthesis gene transcripts in chromerid algae represents the first transcript processing event documented in any plastid lineage to preferentially target a particular functional category of genes. It remains to be determined whether this feature is particularly unusual, or whether other photosynthetic eukaryote lineages also utilise a transcript processing machinery that preferentially targets specific functional components of the plastid transcriptome.

The distribution of poly(U) sites in chromerid plastids additionally may provide insights into the evolution of their non-photosynthetic apicomplexan relatives. The previous identification of poly(U) tail addition in peridinin dinoflagellates (Wang and Morse, 2006) and in *Chromera velia* (Janouskovec et al., 2010) indicates that poly(U) tail addition occurred in a common ancestor of the dinoflagellate, chromerid and apicomplexan lineages (Fig. 8.1, points A-C). Thus, the absence of poly(U) tail addition in apicomplexans is likely to be a result of secondary loss of the associated pathway (Fig. 8.1, point D) (R.E.R. Nisbet, pers. comm.) (Dorrell *et al.*, 2014). The identification of poly(U) tails in *Vitrella brassicaformis*, as presented in this thesis, which has been suggested to be the closest related characterised photosynthetic lineage to the apicomplexans pinpoints the loss of poly(U) tails to occurring concurrent with the loss of photosynthesis genes from the apicomplexan plastid (Fig. 8.1, point D) (Janouskovec *et al.*, 2010; Janouškovec *et al.*, 2012a).

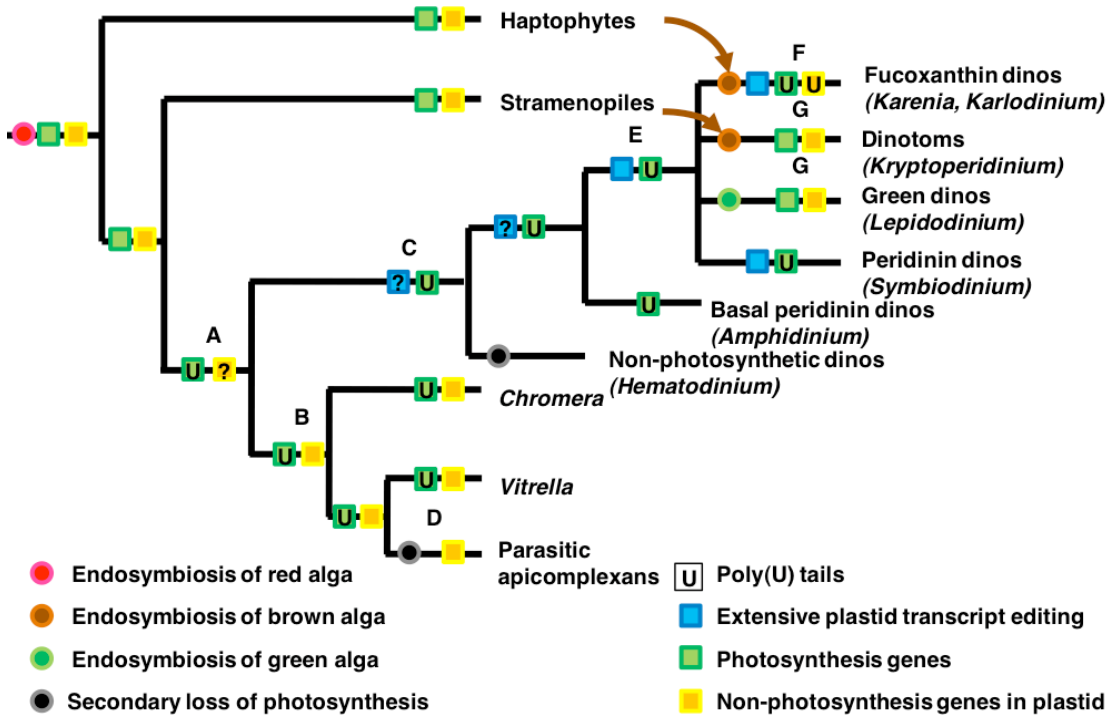


Fig. 8.1: Taxonomic distribution of poly(U) tail addition and editing across the alveolates.

This diagram shows the evolutionary relationships between different alveolate groups and their closest plastid-bearing relatives, and the presence of poly(U) tails and editing on plastid transcripts. Key points in alveolate evolution are marked with letters. Events that are debated are labelled with question marks. Poly(U) tail addition evolved in a common ancestor of dinoflagellate, apicomplexan and chromerid plastids (A). This pathway was specialised towards photosynthesis genes at least in the common ancestor of chromerids and apicomplexans, which retained both photosynthesis genes non-photosynthesis genes in its plastid (B). In the peridinin dinoflagellates, non-photosynthesis genes were relocated to the nucleus (C), whereas poly(U) tail addition was lost from the parasitic apicomplexans, concurrent with the loss of photosynthesis genes (D).

Transcript editing arose within the dinoflagellate lineage. This may have occurred in a common ancestor of all extant dinoflagellates, following their divergence from chromerids (C), or occurred within a subset of peridinin dinoflagellates, as the some basally divergent lineages (e.g. *Amphidinium*) do not contain evidence for extensive plastid transcript editing (Bachvaroff *et al.*, 2014; Barbrook *et al.*, 2012). Poly(U) tail addition and editing has been retained in fucoxanthin dinoflagellates from the ancestral peridinin symbiosis, and applied to the incoming replacement plastid (F), but is not found in other serially acquired dinoflagellate plastid lineages (G).

Chromera and *Vitrella* appear to have diverged in separate events from the apicomplexans (Janouskovec *et al.*, 2010; Janouškovec *et al.*, 2012a). Thus, the preferential addition of poly(U) tails to transcripts encoding photosystem proteins in both genera indicates that the poly(U) machinery was also associated with photosynthesis genes in the common ancestor of chromerids and apicomplexans (Fig. 8.1, point B). It was not possible to infer this from previous studies of chromerid plastid transcript processing, which solely focussed on photosynthesis genes (Janouskovec *et al.*, 2010; Janouskovec *et al.*, 2013), or from studies of peridinin dinoflagellates, as all protein-coding genes of non-photosynthetic function in these species have been relocated to the nucleus (Bachvaroff *et al.*, 2004; Hackett *et al.*, 2004; Howe *et al.*, 2008b) (Fig. 8.1, point C). Thus, apicomplexans have lost an important transcript processing event associated with the expression of photosystem proteins, alongside the transition from photosynthesis to parasitism (Fig. 8.1, point D).

It is possible that an early ancestor of apicomplexans changed from a photosynthetic to a non-photosynthetic lifestyle, and the poly(U) machinery was subsequently lost due to a lack of selective pressure for its retention. Equally, if the poly(U) tail addition were essential for the expression of photosystem proteins, the loss of this pathway might have been a key step in the transition of early apicomplexans from photosynthesis towards parasitism. Examples are known in parasitic plants where genes that should encode otherwise functional proteins have lost associated sequences required for transcription, or transcript processing. For example, the parasitic plant *Harveya huttonii* retains an *rbcL* gene that encodes a complete rubisco large subunit protein (Randle and Wolfe, 2005). However, copies of this gene have highly divergent associated promoters, Shine-Dalgarno sequences, and 3' terminal processing sites, which prevent their functional expression (Randle and Wolfe, 2005). Similarly, the loss of consensus transcript editing sites, which may affect the function of the proteins encoded, has been inferred to have occurred in plastid genes of the parasitic plants *Cuscuta reflexa* and *C. gronovii* (Funk et al., 2007; Tillich and Krause, 2010). It is possible that, in ancestors of early apicomplexans, changes to an entire plastid transcript processing pathway might have underpinned plastid gene loss, and the transition from photosynthesis to parasitism.

Conclusion 2: poly(U) tail addition and editing occur in fucoxanthin dinoflagellate plastids (Chapters Four, Five)

I have shown that plastid transcripts in the fucoxanthin dinoflagellates *Karenia mikimotoi* and *Karlodinium veneficum* receive poly(U) tails, and are edited (Fig. 8.1, point F). These pathways appear to be unique to the fucoxanthin plastid lineage, as the serially acquired plastids found in dinotom algae and *Lepidodinium* do not possess either transcript

processing pathway (Fig. 8.1, point G). Poly(U) tail addition and editing appear to play important roles in the functional expression of fucoxanthin plastid genes. The presence of transcript editing, have been reported independently in *Karlodinium veneficum* (Jackson *et al.*, 2013).

I have additionally demonstrated that editing and poly(U) tail addition are not associated with the plastids of free-living haptophyte relatives of the fucoxanthin lineage (e.g. *Emiliania huxleyi*), or with stramenopiles (e.g. *Phaeodactylum tricornutum*). Plastid poly(U) tail addition presumably therefore evolved within the alveolates following their divergence from other eukaryotes (Fig. 8.1, point A). Editing, which has been shown in this thesis and in independently performed studies not to occur in chromerid algae, is likely to have evolved within the dinoflagellates (although the exact point at which it originated remains debated) (Fig. 8.1, points C, E) (Howe *et al.*, 2008b; Janouskovec *et al.*, 2013). Presumably, the poly(U) tail addition and editing pathways were applied to the incoming fucoxanthin plastid following its serial endosymbiotic acquisition, and may be derived from pathways already present in peridinin dinoflagellate host.

The origin of transcript poly(U) tail addition and editing in fucoxanthin dinoflagellates may provide valuable insights into the processes underpinning plastid evolution. Previous studies have shown that some nuclear genes in fucoxanthin dinoflagellates that encode putative plastid proteins (e.g. cysteine synthase, glyceraldehyde-3-phosphate dehydrogenase isoform C1) have been retained from the earlier peridinin symbiosis, although it has not been shown definitively that these encode proteins that function in the fucoxanthin plastid (Nosenko *et al.*, 2006; Patron *et al.*, 2006). In peridinin dinoflagellates, transcript editing is not only known in plastids but in mitochondria, and mitochondrial RNA editing events have been identified in fucoxanthin lineages, raising the question of whether the peridinin plastid or mitochondria gave rise to the editing events now found in fucoxanthin plastids (Jackson *et al.*, 2007; Nash *et al.*, 2007). In contrast, poly(U) tail addition is not known to occur in any other dinoflagellate organelle other than the peridinin and fucoxanthin plastids. Thus, the most parsimonious explanation for the presence of poly(U) tail addition in fucoxanthin plastids is that RNA processing pathways from the ancestral peridinin plastid lineage was retained following serial endosymbiosis, and applied to the incoming fucoxanthin plastid. My data represent the first biochemical proof that plastids acquired through serial endosymbiosis are supported by pathways inherited from their predecessors. This has previously been predicted as part of the “shopping bag” model for plastid evolution, which states that plastids may be supported by genes obtained from different donor lineages, which were acquired prior to the plastid endosymbiosis event (Dorrell and Howe, 2012b; Larkum *et al.*, 2007).

One outstanding question is how a pathway derived from the peridinin dinoflagellate plastid was retained and applied to the incoming replacement fucoxanthin lineage. It is possible that the ancestral peridinin plastid was initially lost, and that fucoxanthin dinoflagellates are descended from secondarily non-photosynthetic ancestors that subsequently acquired plastids from a novel phylogenetic source. This raises the question of how long genes associated with poly(U) tail addition could remain in the nucleus of the non-photosynthetic ancestor, in presumably vestigial form, before they would be lost through drift or purifying selection. That said, it is well understood that functionally redundant regions of sequence (e.g. NUPTs) can be retained within individual nuclear lineages for millions of years before they are lost (Rousseau-Gueutin *et al.*, 2011), and examples are known of secondarily non-photosynthetic dinoflagellates that may retain "footprints" of genes retained from the peridinin plastid symbiosis (Matsuzaki *et al.*, 2007; Wisecaver and Hackett, 2010).

An alternative scenario is that the fucoxanthin plastid was acquired before the peridinin plastid was lost. Thus, early ancestors of fucoxanthin dinoflagellates may have simultaneously harboured plastids of two different endosymbiotic derivations, one of which, derived from the peridinin plastid, utilised a functional poly(U) tail addition pathway, which could then be readily applied to the incoming replacement lineage. This hypothesis would therefore avoid an intermediate stage in which the genes associated with poly(U) tail addition were non-functional, and thus vulnerable to deselection. However, while there are a large number of examples known of otherwise non-photosynthetic eukaryotes that are able to form productive relationships with photosynthetic symbionts, hence represent possible intermediates in the endosymbiotic acquisition of plastids, there is only very limited evidence that lineages of eukaryotes that already possess their own endogenous chloroplasts may supplement these with further photosynthetic endosymbionts (Johnson, 2011; Prechtl *et al.*, 2004; Stoecker *et al.*, 2009). Thus, the possibility that a lineage of dinoflagellates arose that simultaneously possessed two plastid lineages seems less ecologically plausible than an initial loss of the original plastid, followed by acquisition of the replacement. Ultimately, distinguishing between these two scenarios will require characterisation of plastid evolution and transcript processing pathways in a greater range of dinoflagellate species than have currently been investigated.

A further question concerns the extent to which the acquisition of pathways from prior symbionts occurs in the evolution of other plastid lineages. This largely depends on how many additional serial endosymbioses have occurred in other photosynthetic eukaryotes, beyond the well-characterised examples within the dinoflagellates (Dorrell and Smith, 2011). As previously discussed, there is evidence from genomic data that diatom algae, and other lineages that possess secondary, red algal plastids, historically possessed a green algal

endosymbiont (Dorrell and Smith, 2011; Moustafa, 2009). Many of the green algal genes documented in these lineages appear to have related biochemical functions. For example, the genes of green algal origin identified include genes that encode components of the xanthophyll cycle (violaxanthin de-epoxidase, and zeaxanthin epoxidase), which is an important component of the plant and green algal photoprotective machinery, but is not known in red algae (Frommolt *et al.*, 2008; Goss and Jakob, 2010). Notably, xanthophyll cycle intermediates including violaxanthin accumulate in diatoms cultured under high light conditions, and the suppression of the violaxanthin de-epoxidase gene of the model diatom species *Phaeodactylum tricornutum* with an antisense construct reduces non-photochemical quenching capacity (Lavaud *et al.,* 2012; Lohr and Wilhelm, 1999). Thus, it is possible that the retention of green algal genes has allowed diatoms to tolerate more extreme light regimes than if they had only utilised pathways native to the extant red plastid lineage. It has similarly been suggested that the genes of chlamydiobacterial origin found in archaeplastid lineages may have been retained from a previous endosymbiont (Huang and Gogarten, 2007). Some of these genes have functions in carbohydrate metabolism (e.g. isoamylase, and ADP-glucose starch synthase), and it has been suggested that the chlamydiobacterial genes have enabled the more efficient metabolism of fixed carbon exported from the plastid (Ball *et al.*, 2013; Price *et al.*, 2012). More extensive and systematic explorations of the extent of serial endosymbiosis across the eukaryotes may confirm whether the biological activities of extant plastids have been optimised by pathways retained from historical symbioses.

Conclusion 3: Fucoxanthin plastid genomes are highly divergently organised (Chapters Five, Six)

I have documented evidence for extremely divergent genome evolution in fucoxanthin dinoflagellate plastids. I have shown that the plastid genomes of *Karenia mikimotoi* and *Karlodinium veneficum* have undergone different gene loss events and changes to gene structure and order from each other. Previous studies of fucoxanthin dinoflagellate plastid genomes have documented extremely rapid sequence evolution, which has impeded the deduction of the phylogenetic affinity of the plastid itself (Inagaki *et al.*, 2004; Yoon *et al.*, 2002). It appears that some of this divergent sequence evolution has occurred following the endosymbiotic acquisition of the fucoxanthin plastid by its dinoflagellate host.

I have additionally demonstrated that the *Karlodinium veneficum* plastid genome has undergone a parallel fragmentation event to that of the peridinin plastid lineage. I have generated the complete sequence of an episomal minicircle containing the *Karlodinum veneficum dnaK* gene. This is the first complete minicircle sequence obtained from a non-

peridinin plastid lineage. A previous study, using next generation sequencing and Southern blotting data, inferred the presence of minicircles in the *Karlodinium veneficum* plastid, although did not obtain a complete minicircle sequence (Espelund *et al.*, 2012). Although my data could be explained by tandemly repeated copies of *dnaK* , the Southern blots presented by Espelund et al. only contained evidence for single *dnaK* sequence copies, and did not contain any visible bands that would correspond to multimeric copies of *dnaK* sequence (Espelund *et al.*, 2012). It remains to be determined whether further minicircles are present in *K. veneficum* (for example, minicircles containing the episomal *rbcL* fragments). It additionally remains to be determined whether similar minicircles are present in *Karenia mikimotoi*, or whether the fragmentation of the *Karlodinium veneficum* plastid genome occurred following the divergence of the fucoxanthin dinoflagellates. Nevertheless, my data demonstrate unusual convergence in the organisation, as well as the expression machinery associated with fucoxanthin and peridinin dinoflagellate plastid genomes.

Conclusion 4: poly(U) tail addition and editing have been adapted to the divergent evolution of alveolate plastid genomes (Chapters Five, Seven)

My data provide insights into how the transcript processing machinery has responded to changes to the content and organisation of alveolate plastid genomes. I have shown that in fucoxanthin dinoflagellates, the overwhelming majority of genes give rise to polyuridylylated and edited transcripts. Notably, many of the genes that possess associated poly(U) sites in fucoxanthin dinoflagellate plastids have been relocated to the nucleus in peridinin dinoflagellates (Bachvaroff *et al.*, 2004; Howe *et al.*, 2008b). These include genes that encode plastid proteins with non-photosynthesis functions. Even genes that have been lost or were never present in the peridinin dinoflagellate lineage, such as the genes encoding a form ID rubisco, and ORFs within the *Karlodinium veneficum* plastid that have no similarity to any previously annotated sequence, give rise to polyuridylylated and edited transcripts (Morse *et al.*, 1995; Takishita *et al.*, 2000).

The wide variety of polyuridylylated transcripts generated in fucoxanthin plastids contrasts with the situation in chromerids, in which the majority of polyuridylylated transcripts encode proteins involved in photosynthesis (Fig. 8.1; compare points B, F). The widespread distribution of poly(U) sites in fucoxanthin plastids may reflect the fact that in peridinin dinoflagellates, as a consequence of almost every gene of non-photosynthesis function having been relocated to the nucleus, poly(U) tails are applied to effectively every transcript of the plastid genome (Fig. 8.1, point C). Thus, while the poly(U) machinery in chromerids was associated with photosynthesis genes, it has been converted into a pathway involved in

general expression of the fucoxanthin plastid genome, as a consequence of the unusual genome reduction events observed in the peridinin plastid lineage.

In addition, transcript processing pathways may have important roles in constraining the phenotypic consequences of divergent sequence evolution in fucoxanthin dinoflagellates. For example, I have demonstrated that transcript editing removes in-frame termination codons that would prevent the complete translation of certain fucoxanthin transcript sequences (e.g. *Karenia mikimotoi psaA*). Similar results have been found in independently conducted studies of fucoxanthin dinoflagellate plastid transcripts (Jackson *et al.*, 2013). Furthermore, editing within the *Karlodinium veneficum* plastid genome appears to be especially frequent on highly divergent sequences such as recently acquired sequence insertions. Editing may therefore play an important role in correcting the effects of divergent mutations in fucoxanthin plastid genomes that might otherwise prove deleterious.

It remains to be determined whether poly(U) tail addition has a similar role to editing in limiting the effects of divergent evolution in alveolate plastid genomes. Notably, however, transcripts of pseudogenes in the *Karlodinium veneficum* (e.g. *rbcS-1, atpF-2*) and *Chromera velia* (*atpH-2*) plastids do not receive poly(U) tails. The absence of poly(U) tail addition from pseudogenes has not previously been reported in alveolate plastids. It is possible that poly(U) tail addition may have a role in discriminating between transcripts of functional genes, and transcripts of pseudogenes generated by recent rearrangements in alveolate plastid genomes. The precise significance of this for alveolate plastid gene expression awaits further characterisation.

It remains to be determined how poly(U) tail addition and editing are targeted to specific sites in alveolate plastid genomes. Identifying the mechanisms by which this occurs may explain why poly(U) addition and editing have become associated with effectively every gene in fucoxanthin plastids, and why these transcript processing events remain preferentially associated with functional genes in alveolate plastids. Notably, I could not identify any motifs that were universally associated with poly(U) sites either in the published plastid genome sequences of chromerids (*Chromera velia*, *Vitrella brassicaformis*) or fucoxanthin dinoflagellates (*Karlodinium veneficum*) (Gabrielsen *et al.*, 2011; Janouskovec *et al.*, 2010; Janouskovec *et al.*, 2013). Although the initial report of poly(U) tail addition, in the peridinin dinoflagellate *Lingulodinium polyedrum*, identified two A/T-rich motifs, positioned upstream of the poly(U) sites of multiple genes in the plastid genome (Wang and Morse, 2006), similar motifs have not been reported in subsequent studies of poly(U) tail addition in other peridinin species (Barbrook *et al.*, 2012; Dang and Green, 2009). As large numbers of editing events may occur on dinoflagellate plastid transcripts, as reported here and elsewhere (Zauner *et*

al., 2004), it similarly seems unlikely that there are conserved sequences, located adjacent to each editing site, which are essential for each editing event.

It is therefore probable that poly(U) tail addition and editing are not dependent on universally conserved *cis*-acting sequences within target genes. This contrasts with poly(A) tail addition in nuclear transcript processing, which occurs adjacent to specific motifs that are conserved between the 3' UTRs of different nuclear genes (Fitzgerald and Shenk, 1981; Sheets *et al.*, 1990). Instead, the poly(U) addition machinery might bind to individual adapter proteins that are then recruited to specific sites on each transcript. Nucleus-encoded adapter proteins that recognise individual sites in transcript sequence (e.g. pentatricopeptide repeat proteins, also referred to as PPR proteins), are known to direct transcript processing events in plant plastids (Barkan, 2011; Schmitz-Linneweber and Small, 2008). PPR proteins are highly diversified in fucoxanthin dinoflagellate nuclear genomes, and are subject to rapid transcriptional regulation under varying environmental conditions (Morey *et al.*, 2011; Van Dolah *et al.*, 2007). A plastid-targeted PPR protein has recently been identified in the apicomplexan *Plasmodium falciparum*, suggesting that PPR proteins function in a wide range of alveolate plastid lineages (J. McKenzie, R.E.R. Nisbet, pers. comm.). If these PPR proteins were involved in poly(U) tail addition or editing , they could either be rapidly diversified through sexual recombination to enable the processing of transcripts in fucoxanthin dinoflagellate plastids, or be subject to selection so that only transcripts of functional plastid genes were processed.

Conclusion 5: poly(U) tail addition has complex and interconnected relationships to other events in plastid transcript processing (Chapters Four, Six, Seven)

I have investigated the functional roles of poly(U) tail addition in plastid transcript processing events in chromerids, and in the peridinin and fucoxanthin dinoflagellate plastid lineages. Previous studies of peridinin dinoflagellates have indicated that polycistronic transcripts are produced, as a result of rolling circle transcription and the transcription of minicircles that contain multiple genes, and that polycistronic transcripts may receive poly(U) tails (Barbrook *et al.*, 2012; Dang and Green, 2010; Nelson *et al.*, 2007; Nisbet *et al.*, 2008). The addition of a poly(U) tail has furthermore been inferred to enable other processing events, such as cleavage of the transcript 5' end, and editing (Barbrook *et al.*, 2012; Dang and Green, 2009; Dang and Green, 2010). My data demonstrate that the functional significance of poly(U) tail addition varies for different genes in alveolate plastids. For certain genes, other processing events alongside poly(U) tail addition may potentially also facilitate the final events in transcript maturation.

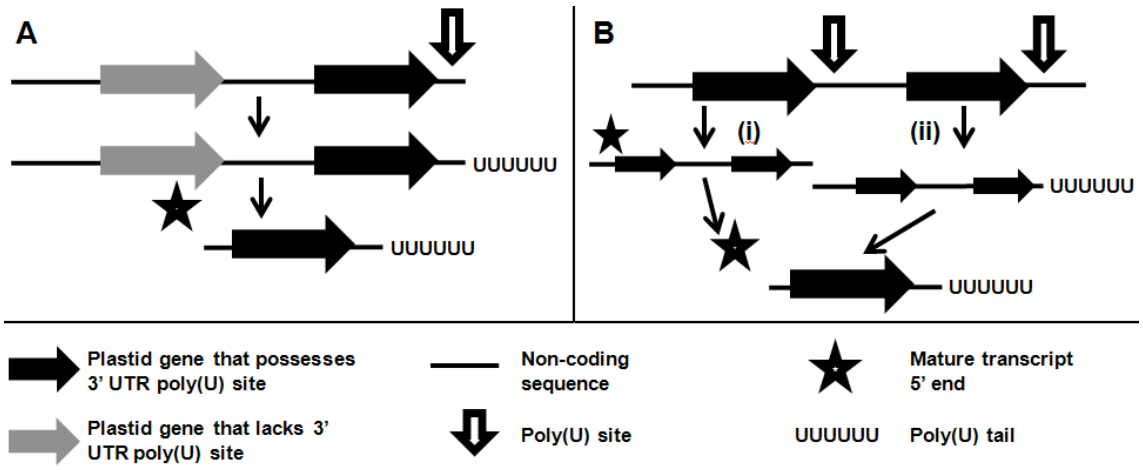


Fig. 8.2: Relationships between poly(U) tail addition and cleavage.

This diagram shows the transcript cleavage events associated with different polycistronic transcripts in alveolate plastids. **Panel A** shows a polycistronic transcript derived from an upstream gene that lacks a 3' UTR poly(U) site, and a downstream gene that possesses a poly(U) site (e.g. *Chromera velia rps14-atpI*). Transcripts of the non-polyuridylylated, upstream gene *(rps14*) are present to significant levels only as polycistronic transcripts, suggesting that monocistronic *rps14* transcripts are not generate in significant levels. In contrast, monocistronic, polyuridylylated transcripts of the downstream gene (*atpI*) are highly abundant, suggesting that following transcription the dicistronic *rps14-atpI* precursors may undergo further cleavage events to generate mature *atpI* transcripts.

Panel B shows a polycistronic transcript containing multiple polyuridylylated genes (e.g. multi-copy *Amphidinium carterae atpA;* minicircle core region and *petB* gene not shown). These transcripts either undergo maturation of the 5' end first **(i)**, prior to 3' end cleavage and poly(U) tail addition, or receive a poly(U) tail **(ii)** before cleavage of the 5' end.

I have identified transcript processing events associated with multi-copy transcripts generated by rolling circle transcription in peridinin dinoflagellates, and transcripts in fucoxanthin dinoflagellate and chromerid plastids. I have confirmed that polycistronic transcripts are present in chromerids and in fucoxanthin dinoflagellates, and these polycistronic transcripts may possess poly(U) tails, as occurs in the peridinin plastid lineage (Fig. 8.2). The cotranscription of plastid genes, and the addition of poly(U) tails to polycistronic transcripts, has also been independently reported in *Chromera velia* (Janouskovec et al., 2013). I have additionally shown that transcripts of genes that lack

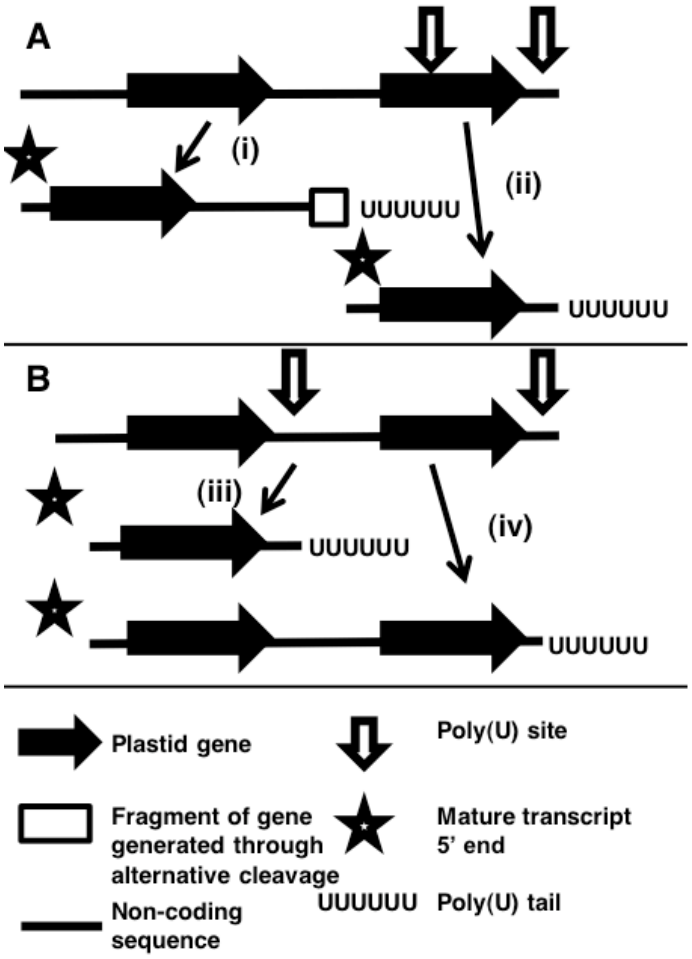


Fig. 8.3: Roles of poly(U) tail addition in alternative end processing.

This diagram shows two forms of alternative end processing observed in alveolate plastids. **Panel A** shows a polycistronic transcript derived from two plastid genes, both of which possess poly(U) sites (e.g. *Chromera velia petB-psbH*). The poly(U) site associated with the upstream gene (*petB*) is positioned within the CDS of the downstream gene (*psbH*). Thus, processing of the *petB* poly(U) site prevents the generation of a transcript containing a complete *psbH* CDS from the same polycistronic precursor.

Panel B shows a different form of alternative processing, which occurs as a result of variable 3' end processing at certain loci (e.g. *Karenia mikimotoi rps13-rps11*; *rpl36* gene upstream of *rps13* not shown). All of the mature transcripts generated from this locus possess the same consensus 5' end position and only differ in terms of whether they terminate at the 3' end at the poly(U) site of the upstream gene (*rps13*) or of the downstream gene (*rps11*). Thus, the addition of a poly(U) tail to either position implicitly determines whether an individual transcript extends over the *rps11* CDS, and thus affects the relative abundance of *rps13* and *rps11* transcripts.

associated 3' UTR poly(U) sites, such as *C. velia rps14*, and *Karenia mikimotoi rpl36*, are frequently retained on polycistronic transcripts, while monocistronic transcripts of these genes are only present at low abundance (Fig. 8.2, panel A). In contrast, genes that possess associated 3' UTR poly(U) sites, including genes positioned downstream of others that lack poly(U) sites (e.g. *C. velia psbA, atpI*) accumulate as monocistronic transcripts (Fig. 8.2, panel A). Thus, poly(U) tail addition may be associated with further transcript cleavage events in alveolate plastids.

I additionally studied the terminal cleavage events associated with transcripts that containing multiple polyuridylylated genes (e.g. multi-copy *Amphidinium carterae* transcripts containing tandem copies of plastid sequence). For *A. carterae atpA,* I identified transcripts that extended past the poly(U) site, but possessed a mature 5' end, as well as polycistronic, polyuridylylated transcripts (that implicitly possess mature 3' ends but have yet to undergo 5' end maturation). It is likely poly(U) tail addition and 5' end maturation occur independently to each other on transcripts of these genes, with some transcripts undergoing 5' end maturation before the poly(U) tail is added, and others receiving a poly(U) tail prior to cleavage of the 5' end (Fig. 8.2, panel C).

I have finally shown that poly(U) tail addition is involved in the alternative processing of transcript ends in chromerids and in fucoxanthin dinoflagellates. Alternative transcript processing events have previously been identified in the plastids of peridinin dinoflagellates, and of plants (Barbrook *et al.*, 2012; Pfalz *et al.*, 2009; Rock *et al.*, 1987). Some of the genes studied (e.g. *C. velia petB*) possess poly(U) sites that are located within the downstream coding sequences, such that poly(U) tail addition would prevent the maturation of transcripts of the downstream gene from the same precursor transcript (Fig. 8.3, panel A). A different form of alternative end processing occurs at the *Karenia mikimotoi rpl36-rps13-rps11* locus. At this locus, two major transcripts are produced: transcripts that extend from a consensus 5' end site upstream of *rpl36* to a poly(U) site downstream of *rps11*, and transcripts that extend from the same consensus 5' site, but terminate at the 3' end in the *rps13* 3' UTR, which possesses its own associated poly(U) site. Thus, 3' end processing and selection of either the *rps13* or *rps11* poly(U) sites determines whether individual transcripts extend into the *rps11* CDS (Fig. 8.3, panel B). This alternative processing event may determine the relative abundance of *rps13* and *rps11* transcripts in the *Karenia mikimotoi* plastid.

Finally, I have observed complex relationships between poly(U) addition and editing in the *Karenia mikimotoi* plastid. For certain genes (e.g. *psbD, rps11*), the processing of the poly(U) site is associated with the completion of editing, as polyuridylylated transcripts are highly edited, whereas transcripts that extend through the poly(U) site are not. For others

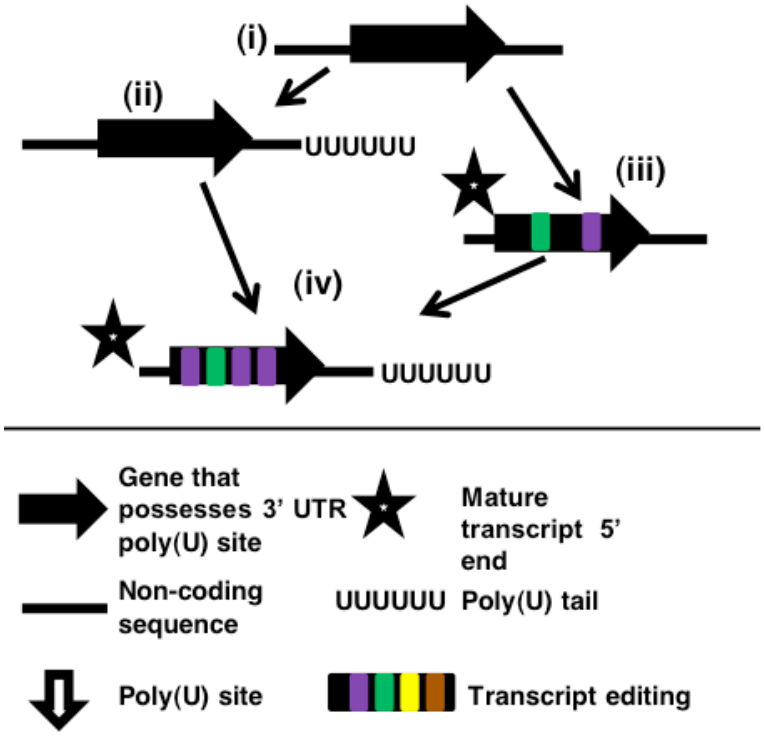


Fig. 8.4: Cooperative relationships between poly(U) tail addition, editing and 5' end cleavage.

This diagram indicates editing events that occur on different processing intermediates of *Karenia mikimotoi ycf4* transcripts. *ycf4* is initially cotranscribed with other plastid genes, and the polycistronic transcripts lack mature 5' ends, poly(U) tails or editing (**i**; upstream genes not shown). The addition of a poly(U) tail, and cleavage of the transcript 5' end (**ii, iii**) occur subsequently, and are required for the editing of the mature transcript sequence (**iv**). Neither poly(U) tail addition, or 5' end cleavage, is sufficient to enable the complete editing of *ycf4*, as transcripts that possess a poly(U) tail but have immature 5' ends are unedited (**ii**), and transcripts that possess 5' ends but lack poly(U) tails are only partially edited (**iii**). Thus, poly(U) tail addition and 5' end cleavage have a cooperative role in enabling editing of *ycf4* transcripts.

(e.g. *rps13*), non-polyuridylylated transcripts are highly edited, indicating that processing of the poly(U) site is not essential for editing. Most surprising are the editing events associated with transcripts of the *Karenia mikimotoi ycf4* gene. Although non-polyuridylylated *ycf4* transcripts are edited much less extensively than polyuridylylated transcripts, polycistronic transcripts that extend upstream of the consensus 5' end position are not edited at all, regardless of whether they possess a poly(U) tail or not (Fig. 8.4). Thus, for certain

transcripts, both cleavage of the 5' end and processing of the poly(U) site are associated with the completion of editing in the transcript sequence.

Overall, my data indicate that poly(U) tail addition has important roles in directing transcript cleavage, editing, and accumulation. However, poly(U) tail addition has different functional consequences for transcripts of different alveolate plastid genes. For certain genes, such as *Karenia mikimotoi ycf4*, poly(U) tail addition may even be functionally interconnected to other transcript processing events, such as 5' end cleavage, with both having to occur to permit further processing events such as transcript editing. The interconnected relationships between different alveolate transcript processing events is similar to the functional relationships between different events in nuclear transcript processing pathways, in which, for example, the poly(A) tail addition machinery recruits the spliceosomal machinery to act on precursor transcripts, and the spliceosomal machinery may in turn license poly(A) tail addition (Kyburz *et al.*, 2006; Rigo and Martinson, 2009).

Conclusion 6: highly edited antisense transcripts are present in dinoflagellate plastids (Chapters Three, Six)

I have identified transcripts containing antisense plastid sequence in both peridinin and fucoxanthin dinoflagellates. These represent the first characterised antisense transcripts of any dinoflagellate genome. The antisense transcripts in *Karenia mikimotoi* undergo processing events that are specifically associated with the fucoxanthin plastid (e.g. editing, and, more rarely, poly(U) tail addition), indicating that these are genuine plastid antisense transcripts, as opposed to transcripts of plastid sequences located within the nuclear genome. While antisense transcripts have been detected in the plastids of plants and apicomplexans, and in cyanobacteria, these represent the first reported antisense transcripts in an algal plastid lineage (Bahl *et al.*, 2010; Georg *et al.*, 2010; Hotto *et al.*, 2010; Kurniawan, 2013; Sakurai *et al.*, 2012).

Although it is clear that the overaccumulation of antisense transcripts in plant plastids is deleterious, it is not known whether antisense transcripts in plants play other roles in plastid gene expression (Hotto *et al.*, 2010; Zghidi-Abouzid *et al.*, 2011). I have identified a potential functional role for antisense transcripts in directing editing events in fucoxanthin plastids. In the *Karenia mikimotoi* plastid, antisense transcripts are not only highly edited, but undergo complementary patterns of editing to sense transcripts. Although antisense transcripts have been shown to extend over residues complementary to editing sites in plant plastids, the direct editing of an antisense plastid transcript sequence has not previously been reported (Georg *et al.*, 2010).

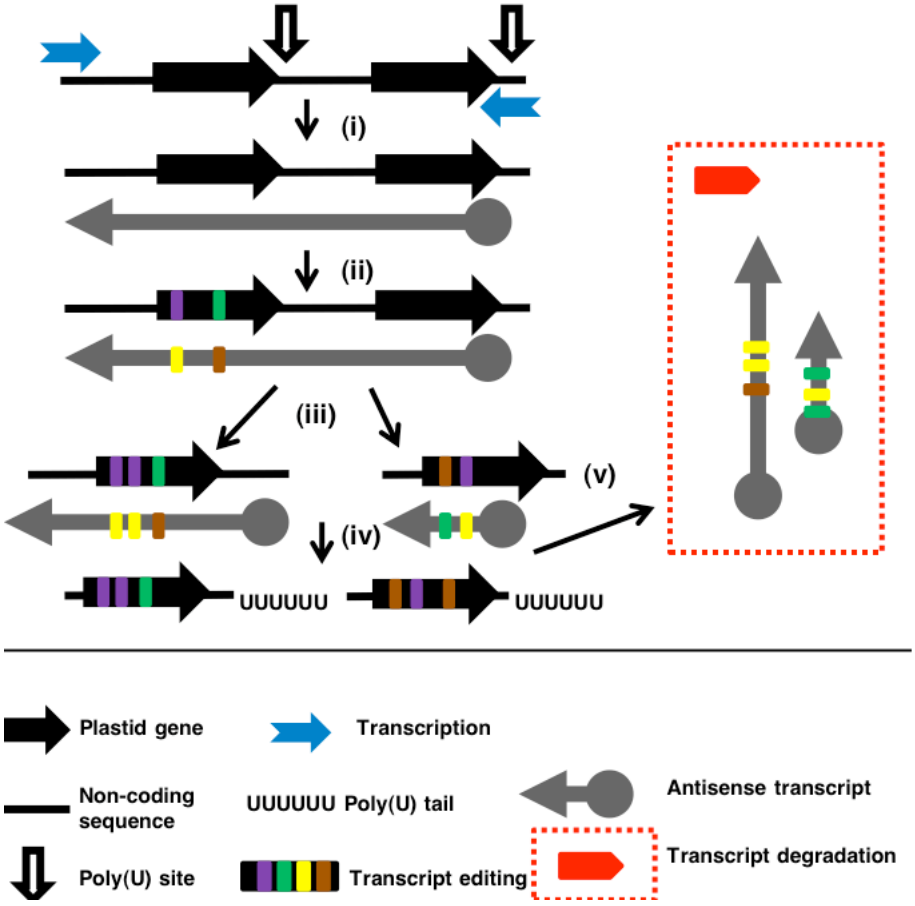


Fig. 8.5: A model for the processing of sense and antisense transcripts in dinoflagellate plastids.

This figure shows the possible events associated with the generation, processing and degradation of antisense transcripts in dinoflagellate plastids. Antisense transcripts are initially generated via the transcription of the non-template strand of plastid genes. Complementary sense and antisense transcripts then anneal together, forming dimers **(i)**.

Dimers of sense and antisense transcripts undergo complementary editing events, such that the substitution of a base on the sense transcript occurs alongside a substitution at the complementary position on the antisense transcript that maintains Watson-Crick Base pairing **(ii)**. In the diagram, purple bars depict editing events that are complementary to the editing events depicted by yellow bars, and green bars are complementary to the editing events depicted by brown bars. Editing events then occur in the same processive order on sense and antisense transcripts **(iii)**.

During the completion of editing, a poly(U) tail is added to the sense transcript **(iv)**. This enables the sense transcript to be discriminated from the non-polyuridylylated antisense transcripts, which may subsequently be targeted for degradation **(v).**

It is possible that antisense transcripts present in the *Karenia mikimotoi* are generated from a previously edited template, for example plastid sense transcripts, via a plastid-located RNA-dependent RNA polymerase (Zandueta-Criado and Bock, 2004). If so, it is surprising that many of the antisense transcripts identified in *K. mikimotoi* extend through poly(U) sites into regions of non-coding sequence that are poorly represented in sense transcript populations, and relatively few possess terminal features that would indicate they have been generated from the highly abundant mature mRNAs present in dinoflagellate plastids (e.g. 5' poly(A) sequences complementary to, and potentially generated via the reverse transcription of 3' poly(U) tails). Similarly, some of the most abundant antisense transcripts within the *A. carterae* plastid cover regions of non-coding sequence with relatively low sense transcript coverage (e.g. the *psbA* minicircle 5' UTR).

Antisense transcripts in dinoflagellate plastids might instead be transcribed in a completely unedited form from plastid gene sequences, and subsequently undergo complementary editing events to the corresponding sense transcripts. One possibility is that a completely edited antisense transcript anneals to an unedited sense transcript, and acts as a template for the editing events that occur. A completely edited sense transcript might then act as a template for editing of unedited antisense transcripts. Similar RNA-mediated storage and transfer of information is observed in certain nuclear lineages, such as the programmed deletion of genomic sequence from ciliate micronuclei during the transition to a macronuclear organisation, which is mediated by small RNA templates that are generated from the nuclei of previously differentiated macronuclei (Eisen *et al.*, 2006; Mochizuki and Gorovsky, 2004), and in telomerase-mediated telomere extension, in which the RNA template incorporated into the telomerase holoenzyme defines the telomeric repeat sequence subsequently generated through reverse transcription (Greider and Blackburn, 1985; Yu *et al.*, 1990). Alternatively, completely unedited sense and antisense transcripts might anneal together early during processing and be edited together as a dimer (Fig. 8.5). This is supported by the fact that editing events on sense and antisense transcripts in *Karenia mikimotoi* occur in the same processive order, and show similar relationships to transcript cleavage (Fig. 8.5).

The complementary editing of sense and antisense transcripts is particularly intriguing given the diversity of editing events observed in dinoflagellate plastids. Previous studies of peridinin dinoflagellate species have identified 8 of the 12 different nucleotide interconversions that can theoretically occur as a result of substitutional editing (Table 8.1) (Howe *et al.*, 2008b; Zauner *et al.*, 2004). I have identified 9 different types of editing interconversion in the *Karenia mikimotoi* plastid, and all twelve different possible types of editing event in the *Karlodinium veneficum* plastid (Table 8.1). An extremely diverse array of

Table 8.1: Transcript editing events observed in peridinin and fucoxanthin plastid lineages.

This table shows the total range of different nucleotide interconversions observed in previous studies of plastid transcripts in four peridinin dinoflagellate species (*Alexandrium tamarense, Ceratium horridum, Heterocapsa triquetra, Lingulodinium polyedrum, Symbiodinium minutum*) (Dang and Green, 2009; Iida *et al.*, 2009; Mungpakdee *et al.,* 2014; Wang and Morse, 2006; Zauner *et al.*, 2004), and in the fucoxanthin dinoflagellates *Karenia mikimotoi* and *Karlodinium veneficum*, as determined in this thesis and since published (Dorrell and Howe, 2012a; Richardson *et al.*, 2014), and in independently conducted studies (Jackson *et al.*, 2013).

Species	Publication	Editing events	Total
Peridinin species		**A ↔ G; A ↔ U; C ↔ U; A → C; G → C; U → G**	**9**
Ceratium horridum	Zauner *et al.*, 2004	A ↔ G; C ↔ U; A → C; G → C; U → G	7
Alexandrium tamarense	Iida *et al.*, 2009	A ↔ G; C → U; G → C	4
Lingulodinium polyedrum	Wang and Morse, 2006	A ↔ G; C ↔ U; G → C; U → G	6
Heterocapsa triquetra	Dang and Green, 2009	A ↔ G; A ↔ U; C ↔ U; A → C; G → C	8
Symbiodinium minutum	Mungpakdee *et al.*, 2014	A → C; A ↔ G; A → U; C ↔ U; G → C; G → U	8
Fucoxanthin species		**A ↔ C; A ↔ G; A ↔ U; C ↔ G; C ↔ U; G ↔ U**	**12**
Karenia mikimotoi	Dorrell and Howe, 2012a; this thesis	A ↔ C; A ↔ G; C ↔ U; A→ U; G → C; U → G	9
Karlodinium veneficum	Jackson *et al.*, 2013; Richardson *et al.*, 2014; this thesis	A ↔ C; A ↔ G; A ↔ U; C ↔ G; C ↔ U; G ↔ U	12

editing interconversions has also been documented in dinoflagellate mitochondrial RNA processing (Jackson *et al.*, 2007; Lin *et al.*, 2002). However, this diversity of editing events has not been identified on any other branch of the tree of life, and it is not clear how dinoflagellates are able to perform such a diverse range of editing events (Knoop, 2011).

It is possible that editing of complementary sites on antisense transcripts might enable editing events on the sense transcript that would not otherwise be directly possible. For example, a particular nucleotide substitution on a sense transcript could be generated by performing a complementary substitution at the corresponding position on an antisense transcript. A reciprocal substitution could then be made at the desired site on the sense transcript via a nucleotide exchange and repair pathway. Mismatch-dependent nucleotide exchange events have been documented in mitochondrial tRNA processing pathways in several lineages, such as the chytrid fungus *Spizellomyces punctatus*, the amoeba *Acanthamoeba castellanii*, and the land snail *Euhadra herklotsi*, although these events are specifically associated with the 5' termini of tRNA sequences, rather than on residues located in internal regions of transcript sequence, as are observed to be edited in dinoflagellates (Bullerwell and Gray, 2005; Knoop, 2011; Yokobori and Paabo, 1995). It remains to be determined whether the antisense transcripts in peridinin dinoflagellate lineages that perform

extensive editing of plastid transcript sequences likewise undergo editing. If complementary editing of sense and antisense transcripts were a widespread feature in dinoflagellate plastid transcripts, it might provide some indication that antisense transcripts play a specific functional role in dinoflagellate plastid gene expression.

It additionally remains to be determined whether antisense transcripts are preferentially removed from dinoflagellate plastids, either during transcript processing or following its completion. In plant plastids, antisense transcripts are preferentially degraded (Sharwood *et al.*, 2011). Certainly, antisense transcripts in both peridinin and fucoxanthin plastids appear to be present at much lower abundance than the complementary sense transcripts. This might be because antisense transcripts are simply transcribed at lower levels than the complementary sense transcripts, as has previously been documented to be the case in plant plastids (Zhelyazkova *et al.*, 2012). Equally, processing events that occur subsequent to transcription may bias dinoflagellate plastid transcript pools in favour of sense transcripts.

With this in mind, the fact that poly(U) tails are generally not added to antisense transcripts in both peridinin and fucoxanthin dinoflagellates is intriguing (Fig. 8.5). Previous studies have suggested that the poly(U) tail confers 3' end stability to dinoflagellate plastid transcripts, and may enable the formation of secondary structures in plastid transcripts that would prevent nucleolytic degradation (Barbrook *et al.*, 2012; Dang and Green, 2009). The specific addition of a poly(U) tail to sense transcripts during transcript processing might therefore enable antisense transcripts to be preferentially degraded, leaving a plastid transcript pool enriched in mature mRNAs (Fig. 8.5).

Future directions

The most major unresolved questions regarding alveolate plastid transcript processing are the identities of the effector proteins involved, and the sites they recognise. In particular, it will be valuable to identify which proteins are involved in poly(U) tail addition and editing in alveolate plastids. Identifying these proteins, and confirming that they are present in fucoxanthin dinoflagellates, would provide definitive proof that the transcript processing machinery of fucoxanthin dinoflagellates has originated from the ancestral peridinin symbiosis.

Given the limited coding capacity of peridinin dinoflagellate and chromerid genomes, and the absence of any ORFs that are of unknown function but are conserved across alveolate plastids, it is likely that the proteins involved in transcript processing are nucleus-encoded (Howe *et al.*, 2008a; Janouskovec *et al.*, 2013; Zhang *et al.*, 1999). However, identifying the proteins required for poly(U) tail addition and editing will be a major task. The enzymes that

directly perform the nucleotide interconversions observed in editing (e.g. cytosine deaminases) remain unknown even in plants (Barkan, 2011; Fujii and Small, 2011). To date, only a small number of proteins have been identified that are individually required for the generation of multiple editing events in plant organelles, which might constitute general components of the editing machinery (Bentolila *et al.*, 2012; Takenaka *et al.*, 2012; Zhang *et al.*, 2014). Similarly, poly(U) polymerases involved in nuclear and mitochondrial RNA metabolism of other lineages appear to have arisen independently to each other, from the extremely diverse family of proteins that constitute poly(A) polymerases, and are difficult to identify based on sequence similarity alone (Aphasizhev, 2005; Lange *et al.*, 2009).

One alternative to identifying the proteins involved in poly(U) tail addition and editing would be to identify whether adapters are required for recruiting the poly(U) tail addition and editing machinery to specific sites in alveolate plastids. As stated previously, transcript processing in plant plastids is dependent on adapters such as PPR proteins, which bind to specific sites within the transcript sequence and either recruit or occlude effector components of the processing machinery, and PPR proteins also appear to be highly diversified in certain alveolate lineages (Barkan, 2011; Morey *et al.*, 2011; Schmitz-Linneweber and Small, 2008). Determining whether PPR proteins or similar adapter proteins are required for editing or poly(U) tail addition in alveolate plastids may provide valuable insights into the evolution of alveolate transcript processing pathways.

It will additionally be interesting to test whether the presence of a poly(U) tail directly influences the accumulation and expression of polyuridylylated transcripts. To date, no reliable and straightforward transformation strategies have yet been developed for any photosynthetic alveolate lineage, which makes it difficult to explore directly the consequences of poly(U) tail addition via genetic manipulation (Qin *et al.*, 2012; ten Lohuis and Miller, 1998). One alternative might be to test the effects of poly(U) tail presence *in vitro*. For example, the relative stability of synthetic polyuridylylated and non-polyuridylylated transcripts could be compared when incubated in enzymatically active dinoflagellate plastid preparations. Alternatively, the effects of poly(U) tail addition on transcript stability could be inferred *in vivo*, by comparing the stability of transcripts that frequently receive a poly(U) tail (e.g. plastid mRNAs) to those that do not (e.g. antisense transcripts) following the disruption of plastid transcription. Similarly, it could be inferred whether the poly(U) tail is associated with translation by determining whether polyuridylylated mRNAs are more highly associated with polysomal fractions than mRNAs that lack poly(U) tails.

A greater understanding of the taxonomic distribution of poly(U) tail addition and editing will be vital to understand the relationships between changes to the plastid transcript processing

machinery, and other evolutionary transitions in alveolate lineages. For example, it is not clear whether poly(U) tail addition was lost from early apicomplexans at the same point that photosynthesis genes were lost from the apicoplast, or whether photosynthesis genes were lost before or after the transition towards a parasitic lifestyle. These questions might be resolved by isolating species that are even closer relatives of parasitic apicomplexans than are chromerids, and determining whether these species are photosynthetic, and retain a poly(U) tail addition pathway. A particularly interesting model for exploration would be a member of the currently uncultured ARL-IV and ARL -V clades, identified from metagenomic sequences from coralline environments, which are believed to be the closest sampled relatives to the apicomplexans (Janouškovec *et al.*, 2012a, b).

Similarly, it remains to be determined whether other pathways to poly(U) tail addition and editing derived from the peridinin plastid symbiosis function in serially acquired dinoflagellate plastids. It will be particularly interesting to determine whether dinotoms or *Lepidodinium*, which do not perform poly(U) tail addition or editing, retain any other components of the ancestral peridinin plastid gene expression machinery, such as PPR proteins (Burki *et al.*, 2014; Minge *et al.*, 2010).

A final and important evolutionary question that remains to be resolved is why the alveolates possess such distinctive plastid genomes and plastid transcript processing pathways. Chromerids, peridinin dinoflagellates and fucoxanthin dinoflagellates have highly unusual plastid genomes (Gabrielsen *et al.*, 2011; Howe *et al.*, 2008b; Janouskovec *et al.*, 2013), and poly(U) tail addition is found in all of these lineages. In contrast, the plastid genomes of diatoms and haptophytes, which do not possess the poly(U) tail addition pathway, are more conventionally organised (Imanian *et al.*, 2010; Puerta *et al.*, 2005; Ruck *et al.*, 2014). The unusual transcript processing pathways observed in alveolate plastids may have originated after the extremely fast sequence evolution commenced. Equally, alveolate plastid genomes and transcript processing pathways might have a more tightly interconnected evolutionary history. Poly(U) tail addition and editing might enable the host to correct or moderate the accumulation of divergent transcript sequences, which would otherwise compromise the function of the plastid, and thus indirectly enable divergent mutations to persist over evolutionary timescales. Ultimately, investigating why alveolate plastids have undergone such divergent and dramatic evolutionary events might provide valuable insights into the coevolution of plastid genomes and biochemistry, and the events that surround major evolutionary transitions in plastid lineages across the eukaryotes.

Appendix I- Glossary of Abbreviations Used

ARL- Apicomplexan-related lineage (c.f. Janouškovec *et al.* 2012a, 2012b)

ATP- Adenosine triphosphate

CCTH- Centrohelids, Cryptomonads, Haptophytes and Telonemids, alternatively collectively termed "Hacrobia" (c.f. Burki *et al.*, 2009; Okamoto *et al.*, 2009)

cDNA- Complementary DNA

CDS- Coding sequence

Chromerid- Paraphyletic photosynthetic lineages related to apicomplexans, containing *Chromera velia* and *Vitrella brassicaformis* (c.f. Janouškovec *et al.*, 2010; Oborník *et al.*, 2012)

CPD-star- Disodium-2-chloro-5-(4-methoxyspiro[1,2-dioxetane-3,2'-5-chlorotricyclo[3.3.1.1^{3,7}]decan])-4-yl]-1-phenyl phosphate; chemiluminescent substrate for HRP

DIG- Digoxigenin

Dino- Dinoflagellate

Dinotom- Dinoflagellate containing diatom-derived serially acquired plastids (c.f. Imanian *et al.*, 2010; Imanian *et al.*, 2012)

DMSO- Dimethyl sulphoxide

dNTP- Deoxynucleotide triphosphate

EST- Expressed sequence tag

Foram- Foraminiferan

Fucoxanthin dinoflagellate- Dinoflagellate containing haptophyte-derived serially acquired plastids, which harbour the light-harvesting pigment fucoxanthin (c.f. Gabrielsen *et al.*, 2011; Takishita *et al.*, 1999)

Green dinoflagellate- Dinoflagellate containing green algal-derived serially acquired plastids

HRP- Horseradish peroxidase

IPTG- Isopropyl β-D-1-thiogalactopyranoside; galactose analogue and transcriptional inducer of galactose metabolism genes

LB- Lysogeny broth

MES- 2-(*N*-morpholino) ethanesulphonic acid

mRNA- Messenger RNA

NADP- Nicotinamide adenine dinucleotide triphosphate

NGS- Next generation sequencing

ORF- Open reading frame

PCR- Polymerase chain reaction

PEG- Polyethylene glycol

Peridinin dinoflagellate- Dinoflagellate containing the ancestral plastid lineage shared with chromerids and apicomplexans, purported to be of red algal origin, and harbouring the accessory light harvesting pigment peridinin (c.f. Howe *et al.*, 2008b)

Poly(U) tail- 3' terminal homo poly(uridylyl) tail

5' RLM-RACE- (RNA-ligase mediated) 5' rapid amplification of capped ends

RT-PCR- Reverse transcription-PCR

TAiL-PCR- Thermal asymmetric interlaced PCR

Tris- Tris(hydroxymethyl)aminomethane buffer

tRNA- Transfer RNA

UTR- Untranslated region

X-Gal- 5-bromo-4-chloro-3-indolyl-β-galactopyranoside

Appendix II- Bibliography

Allen JF. 1993. Control of gene expression by redox potential and the requirement for chloroplast and mitochondrial genomes. *Journal of Theoretical Biology* **165**, 609-631.

Allen JF. 2003. The function of genomes in bioenergetic organelles. *Philosophical Transactions of the Royal Society* **358**, 19-37.

Allison LA, Simon LD, Maliga P. 1996. Deletion of *rpoB* reveals a second distinct transcription system in plastids of higher plants. *EMBO Journal*. **15,** 2802-2809.

Aphasizhev R. 2005. RNA uridylyltransferases. *Cellular and Molecular Life Sciences* **62**, 2194-2203.

Asano T, Miyao A, Hirochika H, Kikuchi S, Kadowaki K. 2013. A pentatricopeptide repeat gene of rice is required for splicing of chloroplast transcripts and RNA editing of *ndhA*. *Plant Biotechnology* **30**, 57-63.

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

Bachvaroff TR, Gornik SG, Concepcion GT, Waller RF, Mendez GS, Lippmeier JC, Delwiche CF. 2014. Dinoflagellate phylogeny revisited: Using ribosomal proteins to resolve deep branching dinoflagellate clades. *Molecular Phylogenetics and Evolution* **70**, 314-322.

Bahl A, Davis PH, Behnke M, Dzierszinski F, Jagalur M, Chen F, Shanmugam D, White MW, Kulp D, Roos DS. 2010. A novel multifunctional oligonucleotide microarray for *Toxoplasma gondii*. *BMC Genomics* **11**, 603.

Baker JR. 1994. The origins of parasitism in the protists. *International Journal of Parasitology* **24**, 1131-1137.

Ball SG, Subtil A, Bhattacharya D, Moustafa A, Weber APM, Gehre L, Colleoni C, Arias M-C, Cenci U, Dauvillee D. 2013. Metabolic effectors secreted by bacterial pathogens: essential facilitators of plastid endosymbiosis? *Plant Cell* **25**, 7-21.

Barbrook AC, Dorrell RG, Burrows J, Plenderleith LJ, Nisbet RER, Howe CJ. 2012. Polyuridylylation and processing of transcripts from multiple gene minicircles in chloroplasts of the dinoflagellate *Amphidinium carterae*. *Plant Molecular Biology* **79**, 347-357.

Barbrook AC, Howe CJ. 2000. Minicircular plastid DNA in the dinoflagellate *Amphidinium operculatum*. *Molecular and General Genetics* **263**, 152-158.

Barbrook AC, Howe CJ, Kurniawan DP, Tarr SJ. 2010. Organization and expression of organellar genomes. *Philosophical Transactions of the Royal Society* **365**, 785-797.

Barbrook AC, Santucci N, Plenderleith LJ, Hiller RG, Howe CJ. 2006. Comparative analysis of dinoflagellate chloroplast genomes reveals rRNA and tRNA genes. *BMC Genomics* **7,** 297.

Barbrook AC, Symington H, Nisbet RER, Larkum A, Howe CJ. 2001. Organisation and expression of the plastid genome of the dinoflagellate *Amphidinium operculatum*. *Molecular Genetics and Genomics* **266**, 632-638.

Barbrook AC, Voolstra CR, Howe CJ. 2013. The chloroplast genome of a *Symbiodinium* sp. clade C3 isolate. *Protist* **165**, 1-13.

Barkan A. 2011. Expression of plastid genes: organelle-specific elaborations on a prokaryotic scaffold. *Plant Physiology* **155**, 1520-1532.

Barkan A, Walker M, Nolasco M, Johnson D. 1994. A nuclear mutation in maize blocks the processing and translation of several chloroplast messenger RNAs and provides evidence for the differential translation of alternative messenger RNA forms. *EMBO Journal* **13**, 3170-3181.

Baurain D, Brinkmann H, Petersen J, Rodriguez-Ezpeleta N, Stechmann A, Demoulin V, Roger AJ, Burger G, Lang BF, Philippe H. 2010. Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles. *Molecular Biology and Evolution* **27**, 1698-1709.

Becker B, Hoef-Emden K, Melkonian M. 2008. Chlamydial genes shed light on the evolution of photoautotrophic eukaryotes. *BMC Evolutionary Biology* **8**, 203.

Bentolila S, Heller WP, Sun T, Babina AM, Friso G, van Wijk KJ, Hanson MR. 2012. RIP1, a member of an *Arabidopsis* protein family, interacts with the protein RARE1 and broadly affects RNA editing. *Proceedings of the National Academy of Sciences USA* **109**, 1453-1461.

Berends Sexton T, Jones JT, Mullet JE. 1990. Sequence and transcriptional analysis of the barley ctDNA region upstream of *psbD-psbC* encoding *trnK(UUU), rps16, trnQ(UUG), psbK, psbI* and *trnS(GCU). Current Genetics* **17,** 445-454

Bergholtz T, Daugbjerg N, Moestrup O, Fernandez-Tejedor M. 2006. On the identity of *Karlodinium veneficum* and description of *Karlodinium armiger* sp nov (Dinophyceae), based on light and electron microscopy, nuclear-encoded LSU rDNA, and pigment composition. *Journal of Phycology* **42**, 170-193.

Berney C, Pawlowski J. 2006. A molecular time-scale for eukaryote evolution recalibrated with the continuous microfossil record. *Proceedings of the Royal Society* **273**, 1867-1872.

Blouin NA, Lane CE. 2012. Red algal parasites: models for a life history evolution that leaves photosynthesis behind again and again. *Bioessays* **34,** 226-235.

Botté CY, Yamaryo-Botté Y, Janouškovec J, Rupasinghe T, Keeling PJ, Crellin P, Coppel RL, Maréchal E, McConville MJ, McFadden GI. 2011. Identification of plant-like galactolipids in *Chromera velia*, a photosynthetic relative of malaria parasites. *Journal of Biological Chemistry* **286**, 29893-29903.

Botté CY, Yamaryo-Botté Y, Rupasuinghe TWT, Mullin KA, MacRae JI, Spurck TP, Kalanon M, Shears MJ, Coppel RL, Crellin PK, Maréchal E, McConville MJ, McFadden GI. 2013. Atypical lipid composion in the purified relict plastid (apicoplast) of malaria parasites. *Proceedings of the National Academy of Sciences USA* **110,** 7506-7511.

Brand LE, Campbell L, Bresnan E. 2012. *Karenia*: The biology and ecology of a toxic genus. *Harmful Algae* **14**, 156-178.

Brown MW, Kolisko M, Silberman JD, Roger AJ. 2012. Aggregative multicellularity evolved independently in the eukaryotic supergroup Rhizaria. *Current Biology* **22**, 1123-1127.

Bullerwell CE, Gray MW. 2005. In vitro characterization of a tRNA editing activity in the mitochondria of *Spizellomyces punctatus*, a chytridiomycete fungus. *Journal of Biological Chemistry* **280**, 2463-2470.

Burki F, Corradi N, Sierra R, Pawlowski J, Meyer GR, Abbott CL, Keeling PJ. 2013. Phylogenomics of the intracellular parasite *Mikrocytos mackini* reveals evidence for a mitosome in Rhizaria. *Current Biology* **23**, 1541-1547.

Burki F, Flegontov P, Oborník M, Cihlár J, Pain A, Lukes J, Keeling PJ. 2012. Re-evaluating the green versus red signal in eukaryotes with secondary plastid of red algal origin. *Genome Biology and Evolution* **4,** 626-635.

Burki F, Imanian B, Hehenberger E, Hirakawa Y, Maruyama S, Keeling PJ. 2014. Endosymbiotic gene transfer in tertiary plastid-containing dinoflagellates. *Eukaryotic Cell* **13,** 246-255.

Burki F, Inagaki Y, Bråte J, Archibald JM, Keeling PJ, Cavalier-Smith T, Sakaguchi M, Hashimoto T, Horák A, Kumar S, Klaveness D, Jakobsen KS, Pawlowski J, Shalchian-Tabrizi K. 2009. Large-scale phylogenomic analyses reveal that two enigmatic protist lineages, *Telonemia* and Centroheliozoa, are related to photosynthetic chromalveolates. *Genome Biology and Evolution* **1,** 231-238.

Cai XM, Fuller AL, McDougald LR, Zhu G. 2003. Apicoplast genome of the coccidian *Eimeria tenella. Gene* **321,** 39-46.

Cazenave C, Uhlenbeck OC. 1994. RNA template-directed RNA synthesis by T7 RNA polymerase. *Proceedings of the National Academy of Science USA* **91**, 6972-6976.

Chase CD. 2007. Cytoplasmic male sterility: a window to the world of plant mitochondrial-nuclear interactions. *Trends in Genetics* **23**, 81-90.

Cock JM, Sterck L, Rouze P, Scornet D, Allen AE, Amoutzias G, Anthouard V, Artiguenave F, Aury JM, Badger JH, Beszteri B, Billiau K, Bonnet E, Bothwell JH, Bowler C, Boyen C, Brownlee C, Carrano CJ, Charrier B, Cho GY, Coelho SM, Collen J, Corre E, Da Silva C, Delage L, Delaroque N, Dittami SM, Doulbeau S, Elias M, Farnham G, Gachon CMM, Gschloessl B, Heesch S, Jabbari K, Jubin C, Kawai H, Kimura K, Kloareg B, Kupper FC, Lang D, Le Bail A, Leblanc C, Lerouge P, Lohr M, Lopez PJ, Martens C, Maumus F, Michel G, Miranda-Saavedra D, Morales J, Moreau H, Motomura T, Nagasato C, Napoli CA, Nelson DR, Nyvall-Collen P, Peters AF, Pommier C, Potin P, Poulain J, Quesneville H, Read B, Rensing SA, Ritter A, Rousvoal S, Samanta M, Samson G, Schroeder DC, Segurens B, Strittmatter M, Tonon T, Tregear JW, Valentin K, von Dassow P, Yamagishi T, Van de Peer Y, Wincker P. 2010. The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. *Nature* **465**, 617-621.

Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. 2008. The archaebacterial origin of eukaryotes. *Proceedings of the National Academy of Sciences USA* **105**, 20356-20361.

Cumbo VR, Baird AH, Moore RB, Negri AP, Neilan BA, Salih A, van Oppen MJ, Wang Y, Marquis CP. 2013. *Chromera velia* is endosymbiotic in larvae of the reef corals *Acropora digitifera* and *A. tenuis. Protist* **164,** *237-244*.

Cummins CA, McInerney JO. 2011. A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. *Systematic Biology* **60,** 833-844.

Curtis BA, Tanifuji G, Burki F, Gruber A, Irimia M, Maruyama S, Arias MC, Ball SG, Gile GH, Hirakawa Y, Hopkins JF, Kuo A, Rensing SA, Schmutz J, Symeonidi A, Elias M, Eveleigh RJM, Herman EK, Klute MJ, Nakayama T, Obornik M, Reyes-Prieto A, Armbrust EV, Aves SJ, Beiko RG, Coutinho P, Dacks JB, Durnford DG, Fast NM, Green BR, Grisdale CJ, Hempel F, Henrissat B, Hoeppner MP, Ishida K-I, Kim E, Koreny LK, Kroth PG, Liu Y, Malik S-B, Maier UG, McRose D, Mock T, Neilson JAD, Onodera NT, Poole AM, Pritham EJ, Richards TA, Rocap G, Roy SW, Sarai C, Schaack S, Shirato S,

Slamovits CH, Spencer DF, Suzuki S, Worden AZ, Zauner S, Barry K, Bell C, Bharti AK, Crow JA, Grimwood J, Kramer R, Lindquist E, Lucas S, Salamov A, McFadden GI, Lane CE, Keeling PJ, Gray MW, Grigoriev IV, Archibald JM. 2012. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature* **492**, 59-65.

Cuvelier ML, Allen AE, Monier A, McCrow JP, Messie M, Tringe SG, Woyke T, Welsh RM, Ishoey T, Lee JH, Binder BJ, DuPont CL, Latasa M, Guigand C, Buck KR, Hilton J, Thiagarajan M, Caler E, Read B, Lasken RS, Chavez FP, Worden AZ. 2010. Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. *Proceedings of the National Academy of Sciences USA* **107**, 14679-14684.

Dacks JB, Marinets A, Doolittle WF, Cavalier-Smith T, Logsdon JM. 2002. Analyses of RNA polymerase II genes from free-living protists: phylogeny, long branch attraction, and the eukaryotic Big Bang. *Molecular Biology and Evolution* **19,** 830-840.

Dang Y, Green BR. 2009. Substitutional editing of *Heterocapsa triquetra* chloroplast transcripts and a folding model for its divergent chloroplast 16S rRNA. *Gene* **442**, 73-80.

Dang Y, Green BR. 2010. Long transcripts from dinoflagellate chloroplast minicircles suggest "rolling circle" transcription. *Journal of Biological Chemistry* **285**, 5196-5203.

de Mendoza A, Sebe-Pedros A, Sestak MS, Matejcic M, Torruella G, Domazet-Loso T, Ruiz-Trillo I. 2013. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *Proceedings of the National Academy of Sciences USA* **110**, 4858-4866.

Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M, Claverie JM, Gascuel O. 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Research* **36**, 465-469.

Deschamps P, Moreira D. 2012. Re-evaluating the green contribution to diatom genomes. *Genome Biology and Evolution* **4**, 683-688.

Deusch O, Landan G, Roettger M, Gruenheit N, Kowallik KV, Allen JF, Martin W, Dagan T. 2008. Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. *Molecular Biology and Evolution* **25**, 748-761.

Dorrell RG, Drew J, Nisbet RE, Howe CJ. 2014. Evolution of chloroplast transcript processing in *Plasmodium* and its chromerid algal relatives. *PLoS Genetics* **10**, 1004008.

Dorrell RG, Howe CJ. 2012a. Functional remodeling of RNA processing in replacement chloroplasts by pathways retained from their predecessors. *Proceedings of the National Academy of Sciences USA* **109,** 18879-18884.

Dorrell RG, Howe CJ. 2012b. What makes a chloroplast? Reconstructing the establishment of photosynthetic symbioses. *Journal of Cell Science* **125**, 1865-1875.

Dorrell RG, Smith AG. 2011. Do red and green make brown?: Perspectives on plastid acquisitions within chromalveolates. *Eukaryotic Cell* **10**, 856-868.

Eisen JA, Coyne RS, Wu M, Wu DY, Thiagarajan M, Wortman JR, Badger JH, Ren QH, Amedeo P, Jones KM, Tallon LJ, Delcher AL, Salzberg SL, Silva JC, Haas BJ, Majoros WH, Farzad M, Carlton JM, Smith RK, Garg J, Pearlman RE, Karrer KM, Sun L, Manning G, Elde NC, Turkewitz AP, Asai DJ, Wilkes DE, Wang YF, Cai H, Collins K, Stewart A, Lee SR, Wilamowska K, Weinberg Z, Ruzzo WL, Wloga D, Gaertig J, Frankel J, Tsao CC, Gorovsky MA, Keeling PJ, Waller RF, Patron NJ, Cherry JM, Stover NA, Krieger CJ, del Toro C, Ryder HF, Williamson SC, Barbeau RA, Hamilton EP, Orias E. 2006. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biology* **4**, 1620-1642.

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

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* **8**, 978-984.

Embley TM, Martin W. 2006. Eukaryotic evolution, changes and challenges. *Nature* **440**, 623-630.

Escalera L, Reguera B, Takishita K, Yoshimatsu S, Koike K, Koike K. 2011. Cyanobacterial endosymbionts in the benthic dinoflagellate *Sinophysis canaliculata* (Dinophysiales, Dinophyceae). *Protist* **162**, 304-314.

Espelund M, Minge MA, Gabrielsen TM, Nederbragt AJ, Shalchian-Tabrizi K, Otis C, Turmel M, Lemieux C, Jakobsen KS. 2012. Genome fragmentation is not confined to the peridinin plastid in dinoflagellates. *PLoS One* **7**, 38809.

Fichera ME, Roos DS. 1997. A plastid organelle as a drug target in apicomplexan parasites. *Nature* **390**.

Fisk JC, Ammerman ML, Presnyak V, Read LK. 2008. TbRGG2, an essential RNA editing accessory factor in two *Trypanosoma brucei* life cycle stages. *Journal of Biological Chemistry* **283**, 23016-23025.

Fitzgerald M, Shenk T. 1981. The sequence 5'-AAUAAA-3' forms part of the recognition site for polyadenylation of late SV40 messenger RNAs. *Cell* **24**, 251-260.

Frommolt R, Werner S, Paulsen H, Goss R, Wilhelm C, Zauner S, Maier UG, Grossman AR, Bhattacharya D, Lohr M. 2008. Ancient recruitment by chromists of green algal genes encoding enzymes for carotenoid biosynthesis. *Molecular Biology and Evolution* **25**, 2653-2667.

Fujii S, Small I. 2011. The evolution of RNA editing and pentatricopeptide repeat genes. *New Phytologist* **191**, 37-47.

Fujiwara S, Iwahashi H, Someya J, Nishikawa S, Minaka N. 1993. Structure and cotransciption of the plastid-encoded *rbcL* and *rbcS* genes of *Pleurochrysis carterae* (Prymnesiophyta). *Journal of Phycology* **29**, 347-355.

Funk HT, Berg S, Krupinska K, Maier UG, Krause K. 2007. Complete DNA sequences of the plastid genomes of two parasitic flowering plant species, *Cuscuta reflexa* and *Cuscuta gronovii*. *BMC Plant Biology* **7**, 45.

Gabrielsen TM, Minge MA, Espelund M, Tooming-Klunderud A, Patil V, Nederbragt AJ, Otis C, Turmel M, Shalchian-Tabrizi K, Lemieux C, Jakobsen KS. 2011. Genome evolution of a tertiary dinoflagellate plastid. *PLoS One* **6**, 19132.

Gachon CMM, Heesch S, Kuepper FC, Achilles-Day UEM, Brennan D, Campbell CN, Clarke A, Dorrell RG, Field J, Gontarek S, Menendez CR, Saxon RJ, Veszelovszki A, Guiry MD, Gharbi K, Blaxter M, Day JG. 2013. The CCAP KnowledgeBase: linking protistan and cyanobacterial biological resources with taxonomic and molecular data. *Systematics and Biodiversity* **11**, 407-413.

Garcia-Cuetos L, Moestrup O, Hansen PJ, Daugbjerg N. 2010. The toxic dinoflagellate *Dinophysis acuminata* harbours permanent chloroplasts of cryptomonad origin, not kleptochloroplasts. *Harmful Algae* **9,** 25-38.

Georg J, Honsel A, Voss B, Rennenberg H, Hess WR. 2010. A long antisense RNA in plant chloroplasts. *New Phytologist* **186**, 615-622.

Gile GH, Slamovits C. 2014. Transcriptomic analysis reveals evidence for a cryptic plastid in the colpodellid *Voromonas pontica*, a close relative of chromerids and apicomplexan parasites. *PLoS One* **9**, 96258.

Glanz S, Kück U. 2009. *Trans*-splicing of organelle introns- a detour to continuous RNAs. *Bioessays* **31**, 921-934.

Gornik SG, Ford KL, Mulhern TD, Bacic A, McFadden GI, Waller RF. 2012. Loss of nucleosomal DNA condensation coincides with appearance of a novel nuclear protein in dinoflagellates. *Current Biology* **22**, 2303-2312.

Goss R, Jakob T. 2010. Regulation and function of xanthophyll cycle-dependent photoprotection in algae. *Photosynthesis Research* **106**, 103-122.

Green BR. 2011. Chloroplast genomes of photosynthetic eukaryotes. *Plant Journal* **66**, 34-44.

Greider CW, Blackburn EH. 1985. Identification of a specific telomere terminal transferase activity in *Tetrahymena* extracts. *Cell* **43**, 405-413.

Gschloessl B, Guermeur Y, Cock JM. 2008. HECTAR: a method to predict subcellular targeting in heterokonts. *BMC Bioinformatics* **9**, 393.

Hackett JD, Yoon HS, Soares MB, Bonaldo MF, Casavant TL, Scheetz TE, Nosenko T, Bhattacharya D. 2004. Migration of the plastid genome to the nucleus in a peridinin dinoflagellate. *Current Biology* **14**, 213-218.

Hajdukiewicz PT, Allison LA, Maliga P. 1997. The two RNA polymerases encoded by the nuclear and the plastid compartments transcribe distinct groups of genes in tobacco plastids. *EMBO Journal* **16,** 4041-4048.

Hampl V, Hug L, Leigh JW, Dacks JB, Lang BF, Simpson AGB, Roger AJ. 2009. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". *Proceedings of the National Academy of Science USA* **106**, 3859-64.

Haxo FT, Kycia JH, Somers GF, Bennett A, Siegelman HW. 1976. Peridinin-chlorophyll A proteins of dinoflagellate *Amphidinium carterae* (Plymouth 450). *Plant Physiology* **57**, 297-303.

Hayes ML, Giang K, Mulligan RM. 2012. Molecular evolution of pentatricopeptide repeat genes reveals truncation in species lacking an editing target and structural domains under distinct selective pressures. *BMC Evolutionary Biology* **12**, 13.

Hedtke B, Börner T, Weihe A. 1997. Mitochondrial and chloroplast phage-type RNA polymerases in Arabidopsis. *Science* **277**, 809-811.

Herrmann KM, Weaver LM. 1999. The shikimate pathway. *Annual Review of Plant Physiology and Plant Molecular Biology* **50**, 473-503.

Hiller RG. 2001. 'Empty' minicircles and *petB/atpA* and *psbD/psbE* (cytb(559) alpha) genes in tandem in *Amphidinium carterae* plastid DNA. *FEBS Letters* **505**, 449-452.

Hjort K, Goldberg AV, Tsaousis AD, Hirt RP, Embley TM. 2010. Diversity and reductive evolution of mitochondria among microbial eukaryotes. *Philosophical Transactions of the Royal Society* **365**, 713-727.

Hoch B, Maier RM, Appel K, Igloi GL, Kossel H. 1991. Editing of a chloroplast messenger RNA by creation of an initiation codon. *Nature* **353**, 178-180.

Hoefnagel MHN, Atkin OK, Wiskich JT. 1998. Interdependence between chloroplasts and mitochondria in the light and the dark. *Biochimica Et Biophysica Acta* **1366**, 235-255.

Horiguchi T, Takano Y. 2006. Serial replacement of a diatom endosymbiont in the marine dinoflagellate *Peridinium quinquecorne* (Peridiniales, Dinophyceae). *Phycological Research* **54**, 193-200.

Hotto AM, Germain A, Stern DB. 2012. Plastid non-coding RNAs: emerging candidates for gene regulation. *Trends in Plant Science* **17**, 737-744.

Hotto AM, Huston ZE, Stern DB. 2010. Overexpression of a natural chloroplast-encoded antisense RNA in tobacco destabilizes 5S rRNA and retards plant growth. *BMC Plant Biology* **10,** 213.

Howe CJ, Barbrook AC, Nisbet RER, Lockhart PJ, Larkum AWD. 2008a. The origin of plastids. *Philosophical Transactions of the Royal Society* **363**, 2675-2685.

Howe CJ, Nisbet RER, Barbrook AC. 2008b. The remarkable chloroplast genome of dinoflagellates. *Journal of Experimental Botany* **59**, 1035-1045.

Huang CY, Ayliffe MA, Timmis JN. Simple and complex nuclear loci created by newly transferred chloroplast DNA in tobacco. *Proceedings of the National Academy of Science USA* **101,** 9710-9715.

Huang JL, Gogarten JP. 2007. Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? ***Genome Biology* 8, 99.**

Hwang SR, Tabita FR. 1991. Cotranscription, deduced primary structure, and expression of the chloroplast-encoded *rbcL* and *rbcS* genes of the marine diatom *Cylindrotheca* sp. strain N1. *Journal of Biological Chemistry* **266**, 6271-6279.

Igamberdiev AU, Lea PJ. 2006. Land plants equilibrate O_2 and CO_2 concentrations in the atmosphere. *Photosynthesis Research* **87**, 177-194.

Iida S, Kobiyama A, Ogata T, Murakami A. 2009. Identification of transcribed and persistent variants of the *psbA* gene carried by plastid minicircles in a dinoflagellate. *Current Genetics* **55**, 583-591.

Iida S, Kobiyama A, Ogata T, Murakami A. 2010. Differential DNA rearrangements of plastid genes, *psbA* and *psbD*, in two species of the dinoflagellate *Alexandrium*. *Plant and Cell Physiology* **51**, 1869-1877.

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,** 43763.

Imanian B, Pombert JF, Keeling PJ. 2010. The complete plastid genomes of the two 'dinotoms' *Durinskia baltica* and *Kryptoperidinium foliaceum*. *PLoS One* **5**, 10711.

Inagaki Y, Simpson AGB, Dacks JB, Roger AJ. 2004. Phylogenetic artifacts can be caused by leucine, serine, and arginine codon usage heterogeneity: dinoflagellate plastid origins as a case study. *Systematic Biology* **53**, 582-593.

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. *Proceedings of the National Academy of Sciences USA* **99**, 9294-9299.

Jackson CJ, Gornik SG, Waller RF. 2013. A tertiary plastid gains RNA editing in its new host. *Molecular Biology and Evolution* **30**, 788-792.

Jackson CJ, Norman JE, Schnare MN, Gray MW, Keeling PJ, Waller RF. 2007. Broad genomic and transcriptional analysis reveals a highly derived genome in dinoflagellate mitochondria. *BMC Biology* **5**, 41.

Janouškovec J, Horák A, Oborník M, Lukes J, Keeling PJ. 2010. A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. *Proceedings of the National Academy of Sciences USA* **107**, 10949-10954.

Janouškovec J, Liu S-L, Martone PT, Carre W, Leblanc C, Collen J, Keeling PJ. 2013a. Evolution of red algal plastid genomes: ancient architectures, introns, horizontal gene transfer, and taxonomic utility of plastid markers. *PLoS One* **8,** 59001.

Janouškovec J, Sobotka R, Lai DH, Flegontov P, Koník P, Komenda J, Ali S, Prásil O, Pain A, Oborník M, Lukes J, Keeling PJ. 2013b. Split photosystem protein, linear-mapping topology, and growth of structural complexity in the plastid genome of *Chromera velia*. *Molecular Biology and Evolution* **30**, 2447-2462.

Janouškovec J, Horák A, Barott KL, Rohwer FL, Keeling PJ. 2013c. Environmental distribution of coral-associated relatives of apicomplexan parasites. *ISME Journal* **7,** 444-447.

Janouškovec J, Horák A, Barott KL, Rohwer FL, Keeling PJ. 2012. Global analysis of plastid diversity reveals apicomplexan-related lineages in coral reefs. *Current Biology* **22**, 518-519.

Janouškovec J, Tikhonenkov D, Mikhailov K, Simdyanov T, Aleoshin V, Mylnikov A, Keeling P. 2013d. Colponemids represent multiple ancient alveolate lineages. *Current Biology* **23,** 2546-2552.

John U, Fensome RA, Medlin LK. 2003. The application of a molecular clock based on molecular sequences and the fossil record to explain biogeographic distributions within the A*lexandrium tamarense* "species complex" (Dinophyceae). *Molecular Biology and Evolution* **20**, 1015-1027.

Johnson, MD. 2011. Acquired phototrophy in ciliates: a review of cellular interactions and structural adaptations. *Journal of Eukaryotic Microbiology* **58**, 185-195.

Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M, Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M, Tabata S. 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. *DNA research* **3**, 109-136.

Kapoor S, Suzuki JY, Sugiura M. 1997. Identification and functional significance of a new class of non-consensus-type plastid promoters. *Plant Journal* **11**, 327-337.

Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research* **33**, 511-518.

Katoh T, Mimuro M, Takaichi S. 1989. Light-harvesting particles isolated from a brown alga, *Dictyota dichotoma*- a supramolecular assembly of fucoxanthin-chlorophyll protein complexes. *Biochimica et Biophysica Acta* **976**, 233-240.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647-1649.

Kleine T, Maier UG, Leister D. 2009. DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. *Annual Review of Plant Biology* **60**, 115-138.

Kneip C, Voss C, Lockhart PJ, Maier UG. 2008. The cyanobacterial endosymbiont of the unicellular algae *Rhopalodia gibba* shows reductive genome evolution. *BMC Evolutionary Biology* **8**, 30.

Knoop V. 2011. When you can't trust the DNA: RNA editing changes transcript sequences. *Cellular and Molecular Life Sciences* **68**, 567-586.

Koreny L, Sobotka R, Janouškovec J, Keeling PJ, Oborník M. 2011. Tetrapyrrole synthesis of photosynthetic chromerids is likely homologous to the unusual pathway of apicomplexan parasites. *Plant Cell* **23**, 3454-3462.

Krause K, Maier RM, Kofer W, Krupinska K, Herrmann RG. 2000. Disruption of plastid-encoded RNA polymerase genes in tobacco: expression of only a distinct set of genes is not based on selective transcription of the plastid chromosome. *Molecular and General Genetics* **263**, 1022-1030.

Kudla J, Hayes R, Gruissem W. 1996. Polyadenylation accelerates degradation of chloroplast mRNA. *EMBO Journal* **15**, 7137-7146.

Kurniawan, D. 2013. Transcription in the malaria parasite *Plasmodium falciparum*. *PhD dissertation*, University of Cambridge.

Kyburz A, Friedlein A, Langen H, Keller W. 2006. Direct interactions between subunits of CPSF and the U2 snRNP contribute to the coupling of pre-mRNA 3 ' end processing and splicing. *Molecular Cell* **23**, 195-205.

Laatsch T, Zauner-S, Stobe-Maier B, Kowallik KV, Maier UG. 2004. Plastid-derived single gene minicircles of the dinoflagellate *Ceratium horridum* are localised in the nucleus. *Molecular Biology and Evolution* **21**, 1318-1322.

Lahr DJG, Katz LA. 2009. Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *Biotechniques* **47**, 857-863.

Lange H, Sement FM, Canaday J, Gagliardi D. 2009. Polyadenylation-assisted RNA degradation processes in plants. *Trends in Plant Science* **14**, 497-504.

Larkum AWD, Lockhart PJ, Howe CJ. 2007. Shopping for plastids. *Trends in Plant Science* **12**, 189-195.

Larson EM, Obrien CM, Zhu GH, Spreitzer RJ, Portis AR. 1997. Specificity for activase is changed by a Pro-89 to Arg substitution in the large subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase. *Journal of Biological Chemistry* **272**, 17033-17037.

Laslett D, Canback B. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research* **32**, 11-16.

Lavaud J, Materna AC, Sturm S, Vugrinec S, Kroth PJ, Silencing of the violaxanthin de-epoxidase gene in the diatom *Phaeodactylum tricornutum* reduces diatoxanthin synthesis and non-photochemical quenching. *PLoS One* **7,** 36806.

Leblond JD, Dodson J, Khadka M, Holder S, Seipelt RL. 2012. Sterol composition and biosynthetic genes of the recently discovered photosynthetic alveolate, *Chromera velia* (chromerida), a close relative of apicomplexans. *Journal of Eukaryotic Microbiology* **59**, 191-197.

Leung SK, Wong JTY. 2009. The replication of plastid minicircles involves rolling circle intermediates. *Nucleic Acids Research* **37**, 1991-2002.

Li CH, Salvucci ME, Portis AR. 2005. Two residues of rubisco activase involved in recognition of the rubisco substrate. *Journal of Biological Chemistry* **280**, 24864-24869.

Liere K, Weihe A, Börner T. 2011. The transcription machineries of plant mitochondria and chloroplasts: composition, function, and regulation. *Journal of Plant Physiology* **168**, 1345-1360.

Lim L, McFadden GI. 2010. The evolution, metabolism and functions of the apicoplast. *Philosophical Transactions of the Royal Society* **365**, 749-763.

Lin SJ. 2011. Genomic understanding of dinoflagellates. *Research in Microbiology* **162**, 551-569.

Lin SJ, Zhang HA, Spencer DF, Norman JE, Gray MW. 2002. Widespread and extensive editing of mitochondrial mRNAS in dinoflagellates. *Journal of Molecular Biology* **320**, 727-739.

Lin SJ, Zhang HA, Zhuang YY, Tran B, Gill J. 2010. Spliced leader-based metatranscriptomic analyses lead to recognition of hidden genomic features in dinoflagellates. *Proceedings of the National Academy of Sciences USA* **107**, 20033-20038.

Lindmark DG, Muller M. 1973. Hydrogenosome, a cytoplasmic organelle of the anaeobic flagellate *Tritichomonas foetus*, and its role in pyruvate metabolism. *Journal of Biological Chemistry* **248**, 7724-7728.

Liu X, Brutlag DL, Liu JS. 2001. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pacific Symposium on Biocomputing* **6**, 127-138.

Liu YG, Mitsukawa N, Oosumi T, Whittier RF. 1995. Efficient isolation and mapping of *Arabidopsis thaliana* T-DNA insert junctions by thermal asymmetric interlaced PCR. *Plant Journal* **8**, 457-463.

Lohr, M, Wilhelm C. 1999. Algae displaying the diadinoxanthin cycle also possess the violaxanthin cycle. *Proceedings of the National Academy of Science USA* **96**, 8784-8789.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **25**, 955-964.

Magee AM, Aspinall S, Rice DW, Cusack BP, Semon M, Perry AS, Stefanovic S, Milbourne D, Barth S, Palmer JD, Gray JC, Kavanagh TA, Wolfe KH. 2010. Localized hypermutation and associated gene losses in legume chloroplast genomes. *Genome Research* **20**, 1700-1710.

Marin B, Nowack ECM, Melkonian M. 2005. A plastid in the making: evidence for a second primary endosymbiosis. *Protist* **156**, 425-432.

Matsumoto T, Ishikawa SA, Hashimoto T, Inagaki Y. 2011a. A deviant genetic code in the green alga-derived plastid in the dinoflagellate *Lepidodinium chlorophorum*. *Molecular Phylogenetics and Evolution* **60**, 68-72.

Matsumoto T, Shinozaki F, Chikuni T, Yabuki A, Takishita K, Kawachi M, Nakayama T, Inouye I, Hashimoto T, Inagaki Y. 2011b. Green-colored plastids in the dinoflagellate genus *Lepidodinium* are of core chlorophyte origin. *Protist* **162**, 268-276.

Matsuzaki M, Kuroiwa H, Kuroiwa T, Kita K, Nozaki H. 2008. A cryptic algal group revealed: a plastid biosynthesis pathway in the oyster parasite *Perkinsus marinus*. *Molecular Biology and Evolution* **25,** 1167-1179.

McBride KE, Schaaf DJ, Daley M, Stalker DM. 1994. Controlled expression of plastid transgenes in plants based on a nuclear DNA-encoded and plastid-targeted T7 RNA polymerase. *Proceedings of the National Academy of Sciences USA* **91**, 7301-7305.

McFadden GI, Reith ME, Munholland J, Lang-Unnasch N. 1996. Plastid in human parasites. *Nature* **381**.

Medlin LK. 2011. The Permian-Triassic mass extinction forces the radiation of the modern marine phytoplankton. *Phycologia* **50**, 684-693.

Minge MA, Shalchian-Tabrizi K, Torresen OK, Takishita K, Probert I, Inagaki Y, Klaveness D, Jakobsen KS. 2010. A phylogenetic mosaic plastid proteome and unusual plastid-targeting signals in the green-colored dinoflagellate *Lepidodinium chlorophorum*. *BMC Evolutionary Biology* **10,** 191.

Mochizuki K, Gorovsky MA. 2004. Small RNAs in genome rearrangement in *Tetrahymena. Current Opinion in Genetics and Development* **14,** 181-187.

Moore RB, Ferguson KM, Loh WKW, Hoegh-Guldberg C, Carter DA. 2003. Highly organized structure in the non-coding region of the *psbA* minicircle from clade C

Symbiodinium. *International Journal of Systematic and Evolutionary Microbiology* **53**, 1725-1734.

Moore RB, Obornik M, Janouškovec J, Chrudimsky T, Vancova M, Green DH, Wright SW, Davies NW, Bolch CJS, Heimann K, Slapeta J, Hoegh-Guldberg O, Logsdon JM, Carter DA. 2008. A photosynthetic alveolate closely related to apicomplexan parasites. *Nature* **451**, 959-963.

Morey JS, Monroe EA, Kinney AL, Beal M, Johnson JG, Hitchcock GL, Van Dolah FM. 2011. Transcriptomic response of the red tide dinoflagellate, *Karenia brevis*, to nitrogen and phosphorus depletion and addition. *BMC Genomics* **12,** 346.

Morse D, Salois P, Markovic P, Hastings JW. 1995. A nuclear-encoded form II RuBisCO in dinoflagellates. *Science* **268**, 1622-1624.

Moszczynski K, Mackiewicz P, Bodyl A. 2012. Evidence for horizontal gene transfer from Bacteroidetes bacteria to dinoflagellate minicircles. *Molecular Biology and Evolution* **29**, 887-892.

Moustafa A. 2009. Evolutionary and functional genomics of photosynthetic eukaryotes. *PhD dissertation*, University of Iowa.

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

Nakamura T, Furuhashi Y, Hasegawa K, Hashimoto H, Watanabe K, Obokata J, Sugita M, Sugiura M. 2003. Array-based analysis on tobacco plastid transcripts: preparation of a genomic microarray containing all genes and all intergenic regions. *Plant and Cell Physiology* **44**, 861-867.

Nash EA, Barbrook AC, Edwards-Stuart RK, Bernhardt K, Howe CJ, Nisbet RER. 2007. Organization of the mitochondrial genome in the dinoflagellate *Amphidinium carterae*. *Molecular Biology and Evolution* **24**, 1528-1536.

Nelson MJ, Dang YK, Filek E, Zhang ZD, Yu VWC, Ishida K, Green BR. 2007. Identification and transcription of transfer RNA genes in dinoflagellate plastid minicircles. *Gene* **392**, 291-298.

Nelson MJ, Green BR. 2005. Double hairpin elements and tandem repeats in the non-coding region of *Adenoides eludens* chloroplast gene minicircles. *Gene* **358**, 102-110.

Nisbet RER, Hiller RG, Barry ER, Skene P, Barbrook AC, Howe CJ. 2008. Transcript analysis of dinoflagellate plastid gene minicircles. *Protist* **159**, 31-39.

Nisbet RER, Koumandou VL, Barbrook AC, Howe CJ. 2004. Novel plastid gene minicircles in the dinoflagellate *Amphidinium operculatum*. *Gene* **331**, 141-147.

Norbury CJ. 2010. 3' uridylation and the regulation of RNA function in the cytoplasm. *Biochemical Society Transactions* **38**, 1150-1153.

Nosenko T, Lidie KL, Van Dolah FM, Lindquist E, Cheng JF, Bhattacharya D. 2006. Chimeric plastid proteome in the florida "red tide" dinoflagellate *Karenia brevis*. *Molecular Biology and Evolution* **23**, 2026-2038.

Oborník M, Modrý D, Lukeš M, Cernotíková-Stříbrná E, Cihlář J, Tesařová M, Kotabová E, Vancová M, Prášil O, Lukeš J. 2012. Morphology, ultrastructure and life cycle of *Vitrella brassicaformis* n. sp., n. gen., a novel chromerid from the Great Barrier Reef. *Protist* **163**, 306-323.

Oborník M, Vancová M, Lai DH, Janouškovec J, Keeling PJ, Lukeš J. 2011. Morphology and ultrastructure of multiple life cycle stages of the photosynthetic relative of Apicomplexa, C*hromera velia*. *Protist* **162**, 115-130.

Okamoto N, Chantangsi C, Horák A, Leander BS, Keeling PJ. 2009. Molecular phylogeny and description of the novel katablepharid *Roombia truncata* gen. et sp. nov., and establishment of the Hacrobia taxon nov. *PLoS One* **4**, 7080.

Okamoto N, Keeling PJ. 2014. The 3D Structure of the Apical Complex and association with the flagellar apparatus revealed by serial TEM tomography in *Psammosa pacifica*, a distant relative of the Apicomplexa. *PLoS One* **9**, 84653.

Oldenburg DJ, Bendich AJ. 2004. Most chloroplast DNA of maize seedlings in linear molecules with defined ends and branched forms. *Journal of Molecular Biology* **335**, 953-970.

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

Owari S, Hayashi A, Ishida KI. 2014. Subcellular localisation of minicircle DNA in the dinoflagellate *Amphidinium massartii. Phycological Research*. **62,** 1-8.

Palmer JD. 1987. Chloroplast DNA evolution and biosystematic uses of chloroplast DNA variation. *American Naturalist* **130**, S6-S29.

Parfrey LW, Lahr DJ, Knoll AH, Katz LA. 2011. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proceedings of the National Academy of Sciences USA* **108**, 13624-13629.

Patron NJ, Waller RF. 2007. Transit peptide diversity and divergence: a global analysis of plastid targeting signals. *Bioessays* **29**, 1048-1058.

Patron NJ, Waller RF, Keeling PJ. 2006. A tertiary plastid uses genes from two endosymbionts. *Journal of Molecular Biology* **357**, 1373-1382.

Pfalz J, Bayraktar OA, Prikryl J, Barkan A. 2009. Site-specific binding of a PPR protein defines and stabilizes 5 ' and 3 ' mRNA termini in chloroplasts. *EMBO Journal* **28**, 2042-2052.

Phillips MJ, Delsuc F, Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. *Molecular Biology and Evolution* **21**, 1455-1458.

Prechtl J, Kneip C, Lockhart P, Wenderoth K, Maier UG. 2004. Intracellular spheroid bodies of *Rhopalodia gibba* have nitrogen-fixing apparatus of cyanobacterial origin. *Molecular Biology and Evolution* **21**, 1477-1481.

Price DC, Chan CX, Yoon HS, Yang EC, Qiu H, Weber APM, Schwacke R, Gross J, Blouin NA, Lane C, Reyes-Prieto A, Durnford DG, Neilson JAD, Lang BF, Burger G, Steiner JM, Loeffelhardt W, Meuser JE, Posewitz MC, Ball S, Arias MC, Henrissat B, Coutinho PM, Rensing SA, Symeonidi A, Doddapaneni H, Green BR, Rajah VD, Boore J, Bhattacharya D. 2012. The *Cyanophora paradoxa* genome elucidates the origin of photosynthesis in algae and plants. *Science* **335**, 843-847.

Prihoda J, Tanaka A, de Paula WBM, Allen JF, Tirichine L, Bowler C. 2012. Chloroplast-mitochondria cross-talk in diatoms. *Journal of Experimental Botany* **63**, 1543-1557.

Puerta MVS, Bachvaroff TR, Delwiche CF. 2005. The complete plastid genome sequence of the haptophyte *Emiliania huxleyi*: a comparison to other plastid genomes. *DNA Research* **12**, 151-156.

Qin S, Lin H, Jiang P. 2012. Advances in genetic engineering of marine algae. *Biotechnology Advances* **30**, 1602-1613.

Quigg A, Kotabová E, Jarešová J, Kaňa R, Setlík J, Sedivá B, Komárek O, Prášil O. 2012. Photosynthesis in *Chromera velia* represents a simple system with high efficiency. *PLoS One* **7**, 47036.

Randle CP, Wolfe AD. 2005. The evolution and expression of *rbcL* in holoparasitic sister-genera *Harveya* and *Hyobanche* (Orobanchaceae). *American Journal of Botany* **92**, 1575-1585.

Reese MG. 2001. Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. *Computers and Chemistry* **26**, 51-56.

Reyes-Prieto A, Moustafa A, Bhattacharya D. 2008. Multiple genes of apparent algal origin suggest ciliates may once have been photosynthetic. *Current Biology* **18**, 956-962.

Richardson E, Howe CJ, Dorrell RG. 2014. Genome-wide transcript profiling reveals the coevolution of chloroplast gene sequences and transcript processing pathways in the fucoxanthin dinoflagellate *Karlodinium veneficum*. *Molecular Biology and Evolution,* **in press**.

Richaud C, Zabulon G. 1997. The heme oxygenase gene (*pbsA*) in the red alga *Rhodella violacea* is discontinuous and transcriptionally activated during ironlimitation. *Proceedings of the National Academy of Sciences USA* **94,** 11736-11741.

Rigo F, Martinson HG. 2009. Polyadenylation releases mRNA from RNA polymerase II in a process that is licensed by splicing. *RNA* **15**, 823-836.

Rock CD, Barkan A, Taylor WC. 1987. The maize plastid *psbB-psbF-petB-petD* gene cluster- spliced and unspliced *petB* and *petD* RNAs encode alternative products. *Current Genetics* **12**, 69-77.

Rombel IT, Sykes KF, Rayner S, Johnston SA. 2002. ORF-FINDER: a vector for high-throughput gene identification. *Gene* **282**, 33-41.

Rott R, Drager RG, Stern DB, Schuster G. 1996. The 3' untranslated regions of chloroplast genes in *Chlamydomonas reinhardtii* do not serve as efficient transcriptional terminators. *Molecular and General Genetics* **252**, 676-683.

Rousseau-Gueutin M, Ayliffe MA, Timmis JN. 2011. Conservation of plastid sequences in the plant nuclear genome for millions of years facilitates endosymbiotic evolution. *Plant Physiology* **157,** 2181-2193.

Ruck EC, Nakov T, Jansen RK, Theriot EC, Alverson AJ. 2014. Serial gene losses and foreign DNA underlie size and sequence variation in the plastid genomes of diatoms. *Genome Biology and Evolution* **6,** 644-654.

Saffo MB, McCoy AM, Rieken C, Slamovits CH. 2010. *Nephromyces*, a beneficial apicomplexan symbiont in marine animals. *Proceedings of the National Academy of Sciences USA* **107**, 16190-16195.

Sagan L. 1967. On the origin of mitosing cells. *Journal of Theoretical Biology* **14**, 255-274.

Sakurai I, Stazic D, Eisenhut M, Vuorio E, Steglich C, Hess WR, Aro E-M. 2012. Positive regulation of *psbA* gene expression by *cis*-encoded antisense RNAs in *Synechocystis* sp. PCC 6803. *Plant Physiology* **160**, 1000-1010.

Schmitz-Linneweber C, Small I. 2008. Pentatricopeptide repeat proteins: a socket set for organelle gene expression. *Trends in Plant Science* **13**, 663-670.

Scotto-Lavino E, Du G, Frohman MA. 2006. Amplification of 5 ' end cDNA with 'new RACE'. *Nature Protocols* **1**, 3056-3061.

Shalchian-Tabrizi K, Skanseng M, Ronquist F, Klaveness D, Bachvaroff TR, Delwiche CF, Botnen A, Tengs T, Jakobsen KS. 2006. Heterotachy processes in rhodophyte-derived secondhand plastid genes: implications for addressing the origin and evolution of dinoflagellate plastids. *Molecular Biology and Evolution* **23**, 1504-1515.

Sharwood RE, Halpert M, Luro S, Schuster G, Stern DB. 2011. Chloroplast RNase J compensates for inefficient transcription termination by removal of antisense RNA. *RNA* **17**, 2165-2176.

Sheets MD, Ogg SC, Wickens MP. 1990. Point mutation in AAUAAA and the poly(A) addition site- effects on the accuracy and efficiency of cleavage and polyadenylatoin *in vitro*. *Nucleic Acids Research* **18**, 5799-5805.

Sheppard AE, Madesis P, Lloyd AH, Day A, Ayliffe AM, Timmis JN. 2011. Introducing an RNA editing requirement into a plastid-localised transgene reduces but does not eliminate functional gene transfer to the nucleus. *Plant Molecular Biology* **76,** 299-309.

Shimada H, Sugiura M. 1991. Fine structural features of the chloroplast genome-comparision of the sequenced chloroplast genomes. *Nucleic Acids Research* **19**, 983-995.

Shoguchi E, Shinzato C, Kawashima T, Gyoja F, Mungpakdee S, Koyanagi R, Takeuchi T, Hisata K, Tanaka M, Fujiwara M, Hamada M, Seidi A, Fujie M, Usami T, Goto H, Yamasaki S, Arakaki N, Suzuki Y, Sugano S, Toyoda A, Kuroki Y, Fujiyama A, Medina M, Coffroth MA, Bhattacharya D, Satoh N. 2013. Draft assembly of the *Symbiodinium minutum* nuclear genome reveals dinoflagellate gene structure. *Current Biology* **23**, 1399-1408.

Siddharthan R, Siggia ED, van Nimwegen E. 2005. PhyloGibbs: A Gibbs sampling motif finder that incorporates phylogeny. *PLoS Computational Biology* **1**, 534-556.

Siemeister G, Hachtel W. 1990. Organisation and nucleotide sequence of ribosomal RNA genes on a circular 73 kbp DNA from the colourless flagellate *Astasia longa*. *Current Genetics* **17**, 433-438.

Slamovits C, Keeling PJ. 2008. Plastid-derived genes in the non-photosynthetic alveolate *Oxyrrhis marina*. *Molecular Biology and Evolution* **25,** 1297-1306.

Smyth RP, Schlub TE, Grimm A, Venturi V, Chopra A, Mallal S, Davenport MP, Mak J. 2010. Reducing chimera formation during PCR amplification to ensure accurate genotyping. *Gene* **469**, 45-51.

Stern DB, Goldschmidt-Clermont M, Hanson MR. 2010. Chloroplast RNA metabolism. *Annual Review of Plant Biology* **61**, 125-155.

Stern DB, Gruissem W. 1987. Control of plastid gene expression- 3' inverted repeats act as messenger RNA processing and stabilising elements, but do not terminate transcription. *Cell* **51**, 1145-1157.

Stoecker D, Johnson MD, de Vargas C, Not F. 2009. Acquired phototrophy in aquatic protists. *Aquatic Microbial Ecology* **57,** 279-310.

Stiller JW, Huang JL, Ding Q, Tian J, Goodwillie C. 2009. Are algal genes in nonphotosynthetic protists evidence of historical plastid endosymbioses? *BMC Genomics* **10,** 484.

Strittmatter G, Kössel H. 1984. Cotranscription and processing of 23S, 4.5S and 5S rRNA in chloroplasts from *Zea mays*. *Nucleic Acids Research* **12**, 7633-7647.

Suzuki K, Miyagishima S. 2010. Eukaryotic and eubacterial contributions to the establishment of the plastid proteome estimated by large-scale phylogenetic analyses. *Molecular Biology and Evolution* **27**, 581-590.

Swiatecka-Hagenbruch M, Liere K, Borner T. 2007. High diversity of plastidial promoters in *Arabidopsis thaliana*. *Molecular Genetics and Genomics* **277**, 725-734.

Tabita FR, Satagopan S, Hanson TE, Kreel NE, Scott SS. 2008. Distinct form I, II, III, and IV RuBisCo proteins from the three kingdoms of life provide clues about RuBisCo evolution and structure/ function relationships. *Journal of Experimental Botany* **59**, 1515-1524.

Takenaka M, Zehrmann A, Verbitskiy D, Kugelmann M, Härtel B, Brennicke A. 2012. Multiple organellar RNA editing factor (MORF) family proteins are required for RNA editing in mitochondria and plastids of plants. *Proceedings of the National Academy of Sciences USA* **109,** 5104-5109.

Takishita K, Ishida KI, Ishikura M, Maruyama T. 2005. Phylogeny of the *psbC* gene, coding a photosystem II component CP43, suggests separate origins for the peridinin- and fucoxanthin derivative-containing plastids of dinoffagellates. *Phycologia* **44**, 26-34.

Takishita K, Kawachi M, Noel M-H, Matsumoto T, Kakizoe N, Watanabe MM, Inouye I, Ishida K-I, Hashimoto T, Inagaki Y. 2008. Origins of plastids and glyceraldehyde-3-phosphate dehydrogenase genes in the green-colored dinoflagellate *Lepidodinium chlorophorum*. *Gene* **410**, 26-36.

Takishita K, Nakano K, Uchida A. 1999. Preliminary phylogenetic analysis of plastid-encoded genes from an anomalously pigmented dinoflagellate *Gymnodinium mikimotoi* (Gymnodiniales, Dinophyta). *Phycological Research* **47**, 257-262.

Takishita K, Nakano K, Uchida A. 2000. Origin of the plastid in the anomalously pigmented dinoflagellate *Gymnodinium mikimotoi* (Gymnodiniales, Dinophyta) as inferred from phylogenetic analysis based on the gene encoding the large subunit of form I-type RuBisCo. *Phycological Research* **48**, 85-89.

ten Lohuis MR, Miller DJ. 1998. Genetic transformation of dinoflagellates (*Amphidinium* and *Symbiodinium*): expression of GUS in microalgae using heterologous promoter constructs. *Plant Journal* **13**, 427-435.

Teng CY, Dang Y, Danne JC, Waller RF, Green BR. 2013. Mitochondrial genes of dinoflagellates are transcribed by a nuclear-encoded single-subunit RNA polymerase. *PLoS One* **8**, 65387.

Tillich M, Krause K. 2010. The ins and outs of editing and splicing of plastid RNAs: lessons from parasitic plants. *New Biotechnology* **27**, 256-266.

Torarinsson E, Lindgreen S. 2008. WAR: webserver for aligning structural RNAs. *Nucleic Acids Research* **36**, 79-84.

Van Dolah FM, Lidie KB, Morey JS, Brunelle SA, Ryan JC, Monroe EA, Haynes BL. 2007. Microarray analysis of diurnal- and circadian-regulated genes in the Florida red-tide dinoflagellate Karenia brevis (Dinophyceae). *Journal of Phycology* **43,** 741-752.

Vera A, Matsubayashi T, Sugiura M. 1992. Active transcription from a promoter positioned within the coding region of a divergently oriented gene: the tobacco chloroplast *rpl32* gene. *Molecular and General Genetics* **233**, 151-156.

Walker G, Dorrell RG, Schlacht A, Dacks JB. 2011. Eukaryotic systematics: a user's guide for cell biologists and parasitologists. *Parasitology* **138**, 1638-1663.

Wang D, Qu ZP, Adelson DL, Zhu JK, Timmis JN. 2014. Transcription of nuclear organellar DNA in a model plant system. *Genome Biology and Evolution,* **in press.**

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**.

Wang YL, Morse D. 2006. Rampant polyuridylylation of plastid gene transcripts in the dinoflagellate *Lingulodinium*. *Nucleic Acids Research* **34**, 613-619.

Weber APM, Osteryoung KW. 2010. From endosymbiosis to synthetic photosynthetic life. *Plant Physiology* **154**, 593-597.

Wellman CH, Gray J. 2000. The microfossil record of early land plants. *Philosophical Transactions of the Royal Society* **355**, 717-731.

Williams-Carrier R, Zoschke R, Belcher S, Pfalz J, Barkan A. 2014. A major role for the plastid-encoded RNA polymerase complex in the expression of plastid transfer RNAs. *Plant Physiology* **164**, 239-248.

Wilson RJM, Denny PW, Preiser PR, Rangachari K, Roberts K, Roy A, Whyte A, Strath M, Moore DJ, Moore PW, Williamson DH. 1996. Complete gene map of the plastid-like DNA of the malaria parasite *Plasmodium falciparum*. *Journal of Molecular Biology* **261**, 155-172.

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 Biology and Evolution* **3**, 1220-1230.

Wu X, Liu M, Downie B, Liang C, Ji G, Li QQ, Hunt AG. 2011. Genome-wide landscape of polyadenylation in *Arabidopsis* provides evidence for extensive alternative polyadenylation. *Proceedings of the National Academy of Sciences USA* **108,** 12533-12538.

Yehudai-Resheff S, Hirsh M, Schuster G. 2001. Polynucleotide phosphorylase functions as both an exonuclease and a poly(A) polymerase in spinach chloroplasts. *Molecular and Cellular Biology* **21**, 5408-5416.

Yin C, Richter U, Boerner T, Weihe A. 2010. Evolution of plant phage-type RNA polymerases: the genome of the basal angiosperm *Nuphar advena* encodes two mitochondrial and one plastid phage-type RNA polymerases. *BMC Evolutionary Biology* **10**, 379.

Yokobori S, Paabo S. 1995. Transfer RNA editing in land snail mitochondria. *Proceedings of the National Academy of Sciences USA* **92**, 10432-10435.

Yokoyama A, Takahashi F, Kataoka H, Hara Y, Nozaki H. 2011. Evolutionary analyses of the nuclear-encoded photosynthetic gene *psbO* from tertiary plastid-containing algae in Dinophyta. *Journal of Phycology* **47**, 407-414.

Yoon HS, Hackett JD, Bhattacharya D. 2002. A single origin of the peridinin- and fucoxanthin-containing plastids in dinoflagellates through tertiary endosymbiosis. *Proceedings of the National Academy of Sciences USA* **99**, 11724-11729.

Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattacharya D. 2004. A molecular timeline for the origin of photosynthetic eukaryotes. *Molecular Biology and Evolution* **21**, 809-818.

Yoshinaga K, Iinuma H, Masuzawa T, Uedal K. 1996. Extensive RNA editing of U to C in addition to C to U substitution in the *rbcL* transcripts of hornwort chloroplasts and the origin of RNA editing in green plants. *Nucleic Acids Research* **24**, 1008-1014.

Yu GL, Bradley JD, Attardi LD, Blackburn EH. 1990. In vivo alteration of telomere sequences and senescence caused by mutated Tetrahymena telomerase RNAs. *Nature* **344,** 156-162.

Záhonová K, Hadariová L, Vacula R, Yurchenko V, Eliáš M, Krajčovič J, Vesteg M. 2014. A small portion of plastid transcripts is polyadenylated in the flagellate *Euglena gracilis*. *FEBS Letters* **588**, 783-788.

Zandueta-Criado A, Bock R. 2004. Surprising features of plastid *ndhD* transcripts: addition of non- encoded nucleotides and polysome association of mRNAs with an unedited start codon. *Nucleic Acids Research* **2004,** 542-550.

Zauner S, Greilinger D, Laatsch T, Kowallik KV, Maier UG. 2004. Substitutional editing of transcripts from genes of cyanobacterial origin in the dinoflagellate *Ceratium horridum*. *FEBS Letters* **577**, 535-538.

Zghidi-Abouzid O, Merendino L, Buhr F, Ghulam MM, Lerbs-Mache S. 2011. Characterization of plastid *psbT* sense and antisense RNAs. *Nucleic Acids Research* **39**, 5379-5387.

Zhang F, Tang WJ, Hedtke B, Zhong LL, Liu L, Peng LW, Lu CM, Grimm B, Lin RC. 2014. Tetrapyrrole biosynthetic enzyme protoporphyrinogen IX oxidase 1 is required for plastid RNA editing. *Proceedings of the National Academy of Sciences USA***, in press.**

Zhang ZD, Cavalier-Smith T, Green BR. 2002. Evolution of dinoflagellate unigenic minicircles and the partially concerted divergence of their putative replicon origins. *Molecular Biology and Evolution* **19**, 489-500.

Zhang ZD, Green BR, Cavalier-Smith T. 2000. Phylogeny of ultra-rapidly evolving dinoflagellate chloroplast genes: a possible common origin for sporozoan and dinoflagellate plastids. *Journal of Molecular Evolution* **51**, 26-40.

Zhang ZD, Green BR, Cavalier-Smith T. 1999. Single gene circles in dinoflagellate chloroplast genomes. *Nature* **400**, 155-159.

Zhelyazkova P, Sharma CM, Förstner KU, Liere K, Vogel J, Börner T. 2012. The primary transcriptome of barley chloroplasts: numerous noncoding RNAs and the dominating role of the plastid-encoded RNA polymerase. *Plant Cell* **24**, 123-136.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**, 3406-3415.

Appendix III- Additional Transcript Sequences

This appendix lists transcript sequences generated during my PhD that were not deposited in GenBank due to sequence length (all plastid transcript sequences) or due to being third party annotations (*Karlodinium veneficum* EST assemblies). Accession numbers for all sequences deposited are provided in Chapter Two.

Chapter Five

> *Karlodinium veneficum* polyuridylylated *psaC* transcript, partial sequence
GATTGTATTGGTTGTAAGAGATGTGAAACAGTATGTCCAACAGATTTTATAAGTATAAGG
GTTTATCTTGGATGTGAAAATTCTCGTAGTTTAGGTTTAACCTATTGAATCTTTTTTTTTTT
TTTTTTTTT

>*Karlodinium veneficum* polyuridylylated *psbI* transcript, partial sequence
GCGGACCCAAATTGATGCGTAATTAATTCTTAAATTGAAATAGTAGTATGTTTGGATTAAA
AGTTGTAGTCTACGGAGTTGTGACTTTTTTTATATCAATTTTTGTATTTGGATTTCTTTCAG
GCGATACATCACGAGTATCTAATAAGCCTGCATAATTTATTTTTTTTTTTTTTTT

>*Karlodinium veneficum* polyuridylylated *psbK* transcript, partial sequence
ATTTTTTTACTACTTGCAGTTGTTTGGCAAGCAGCCGTTGGTTTTAGATAAAAATATAATT
ACATTTTATATAAATTCTATTTTTTTTTTTTTTTTTTTT

>*Karlodinium veneficum* non-polyuridylyated *ORF4* transcript, partial sequence
AAACTCTATATTCCAAAACCTTAAAAATATTGATGAAGAAGTAGCTCATCATGAGATTTAT
CCCTCACTGAAGTATCTACAAGATAACATCTTGTGGAGTAATACTTACTACTACTTTTACA
ACGAACTTGTCCATTTTTTTACCAATATCGGATTTAAATCCGAAGGTTTTGGAATG

Chapter Six

>*Karenia mikimotoi* polyuridylylated *psbT* transcript, partial sequence
TTATACATGTTCTTACTTTTTGGTACTCTAGGGGTTATATTTTTTGCTATATTTTTCCGTGA
TAGCCCAAGAATTGCGACATAGTCGAAATTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT

>*Karenia mikimotoi* polyuridylylated antisense *petA* transcript, partial sequence
AGGGGATGTTATCGCAAGAGACGCACTGGCAAGGTGACAGTTCGAACATTTTTTTTTTTT
TTTTTTTT

>*Karenia mikimotoi* polyuridylylated antisense *psbE* transcript, partial sequence
TTCCTGATACATCATACGCTAATCCTGTCAATACAAATAAAAATCCAGAAATGAAAAGTG
AAGGAATAGTAATAGTATGAATTAACCAGTATCGGACACTCGTGAGTATGTCGGTAAAC
GGACGTTCACCTGTTGAACCGCCAGCCATTCTGGTATATACTTTTATTTTTTTTTTTTTT
TTTTTT

>*Karenia mikimotoi* polyuridylylated antisense *psbH* transcript, partial sequence
CCCAGCTTGGTTCTAATACCGCTACCTTTCAGGGCGTACGCATCGCGTTGGTCGGGAG
AACCTGCTGTAGACATATTTTTTTTACGCCGCACTTTTATTTTTTTTTTTTTT

>*Karenia mikimotoi* polyuridylylated antisense *psbL* transcript, partial sequence
CCCCAAAACAAGGAAGTTCGATTTAGTTCGACAGGTATGCCTTCGAATGGGTTATCTTGT
AACCCTAAACGAGTACCCTTTAGGTAATTTATTTCCATCAATAGATGGAATTCCACTTTTT
TTTTTTTTTTTTT

>*Karlodinium veneficum psaD* transcript (assembled from EST data), partial sequence
TTCATAAGAGATGGCGAGGTTGAAAAATATGTTATGACGTGGTCCAGCAAAAGCGAGCA
AATCATAGAACTGCCGACGGGCGGCGCGGCTTCAATGAAACAAGGCGAGAATCTGATG
TACTTCAGGAAAAAGGAACAAGCCCTCGCTCTCAGTAGATACCTTAAGACAAACTTTAAG
ATCGCAGATTTCAAGGTTTACCGCATCTACCCAGGCGGAGAGGTACAATTCATTCATCC
TGCAGATGGTGTTCCTTCTGAGAAGGTTAATGCGGGACGCATTGGAGTCGGCAACGTA
CCATGGTCCATTGGCAAGAATCCGAGGATTGGAAAATTCGAAAAATCAAACCCTACTAA
C

>*Karlodinium veneficum rpl22* transcript (assembled from EST data), partial sequence
GGTGCGCTTCGCTACAGAGCGTCTTCGCGAACAACACGAGCTGTAGCAAAGCTCTGAA
TTCCCAAGAAAGTAAAAGAACTAGAGGCCAACAGACTTGACCATGATGTAGCAAGATAA
TGATTATCAACATAACAGAAACACAAGCAATCCGCCTTCTACACAGAAGAAAAGTAGAG
GGCTGCATTAACTGTACTGCATTCAGTGCATTGCATTAAGACGATCTGAGGACCGCAGA
GGGCTCTCAGACTGCTAAGGCGTACCTCTTTTCTGGGTGAAGCCAAGGCGCTGAAGCT
TGCTTGAGATAGTTTCCATTGATCCAACGCCAACATGTGGAGAACCTCGATGATTGTTG
CTCACTTGGCTTCTTCGATTTATGCGGTCTCACCACCACTCTCATACAGAGCAGGCTCC
GAGATGAGCAGTGGAGTAGCCATGCGACGACTTGCAGATGCGTTAATGAATAACAATCG
CATCAGAGATCCCAGCATTGCCGGTTACGCAAAAGCTAGGAATGTCAGAATGTCACCGA
CCAAAGTGAGACGCCCCATAAACGAAATCAGGGGTAAATCATATGCGGAAGCTTTGACG
CTACTCGAATACATGCCATATCATTCTTGCATGCCCATCGCTAAAGTTGTGAAATCTGCA

GCAGCAAATGCAGTGAACAATCATGGCTACGATAACATTGCTGATCTGTACGTGGCAGC
GGCCTATGTGGATCAAGGCCCCACGCTGAAGAGGATGAGACCACGCGCGCAAGGAAG
AGCGTATTCGATACAAAAGAAAACCTGTACGATTACGATCGAAATGAAAGAAAAGGAGA
AGCCTAAAGAGGAAGCCTAAGCCCTAAGCTCAGGGCTAGCGTTTTGGCCCAGCGCTCA
CTTGAGTCGCGCGCAGTTGCAGAGCCTTAGGTAACCATCCTGATATTCAGGCATTGTCT
TGTGTGCTCAGTGAGAAAAAAA

>*Karlodinium veneficum rpl23* transcript (assembled from EST data), partial sequence
TTCATCCATGGCGCTTCGTGTGCTGGTTTCGATAGCTCTCGCTTGCTTGGCTCGCGAGG
CTCACACAGAGAATGAGGAGACAGAAAAGTTAGCATCATTGCTTTTCGCACTCGCGCCC
CAACACCCCCAGATGAAGGTCGCGACGTCTGGACAACCTGAGATGAAGGCGAGGACCC
ACCTCAAACCGCCACCAAAGAAGGGAAATCCAAGACAACCCCGAGAGACTTATTACAGA
AACAACCCGATTCTCGATTACGATCTCATAAAGTATCCCGTACTCACAGAGAAATCGATC
AAGAACATTGAAAACCATCAGACCTATACTTTCGCGGTTGCCAAAGATGCAGACAAGCC
TGAGATCAAGGCAGCGATTGAGGGTCTCTTCAATGTATCGGTCAAGAAGTTGAATACGC
TGAATGCACCACCGAAGAGGCGCCGCGTCGGAAAGACCACCGGTAAGGCGCGCCAAT
ACAAGCGCGCGTTTGTGCGCGTTAAGGAAGGGGATTCTATCACTCTGTTCGAGGAGGA
ATGAAATTGATGCGAAGCATTTGAGTGTGGCTTCTGTTTTAGCAGCCGACCGAAATGGA
AATGGCGTTATTGAGGGACATGAGTGGTGCCAATGTTATCAACGTTTATTCGTGTGTTCA
TGTCTCTGTTGGCTTCATGCATGGCCTTCTTGTACTTTTTCTCTTTGGCGAAGGTCTTGC
AGGCAAGCCTTCTCTAGAGTCATGCTGCCGCTAGAGCGTAGTTGCCAGCTTGGGTGCT
ATTGGTTTCTTCTTGCCCCCGATGAAGCTGAGCTTCCTCGTTTCTTCTTGTTTGAAAAAA
AAAAAAA

Appendix IV- Journal Publications Arising

This section consists of the cover page of every paper published over the course of my thesis to which I have contributed data, as listed below. Papers are listed in the order of publication. Papers that are underlined contain material that has been included in my thesis.

- **Dorrell RG, Smith AG.** 2011. Do red and green make brown?: Perspectives on plastid acquisitions within chromalveolates. *Eukaryotic Cell* **10**, 856-868.
- **Walker G, Dorrell RG, Schlacht A, Dacks JB.** 2011. Eukaryotic systematics: a user's guide for cell biologists and parasitologists. *Parasitology* **138**, 1638-1663.
- **Dorrell RG, Howe CJ.** 2012. What makes a chloroplast? Reconstructing the establishment of photosynthetic symbioses. *Journal of Cell Science* **125**, 1865-1875.
- **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,** 43763.
- **Barbrook AC, Dorrell RG, Burrows J, Plenderleith LJ, Nisbet RER, Howe CJ**. 2012. Polyuridylylation and processing of transcripts from multiple gene minicircles in chloroplasts of the dinoflagellate *Amphidinium carterae*. *Plant Molecular Biology* **79**, 347-357.
- **Dorrell RG, Howe CJ**. 2012. Functional remodeling of RNA processing in replacement chloroplasts by pathways retained from their predecessors. *Proceedings of the National Academy of Science USA* **109,** 18879-18884.
- **Gachon CMM, Heesch S, Kuepper FC, Achilles-Day UEM, Brennan D, Campbell CN, Clarke A, Dorrell RG, Field J, Gontarek S, Menendez CR, Saxon RJ, Veszelovszki A, Guiry MD, Gharbi K, Blaxter M, Day JG**. 2013. The CCAP KnowledgeBase: linking protistan and cyanobacterial biological resources with taxonomic and molecular data. *Systematics and Biodiversity* **11**, 407-413.
- **Dorrell RG, Butterfield ER, Nisbet RER, Howe CJ**. 2013. Evolution: unveiling early alveolates. *Current Biology* **23**, 1093-1096.
- **Dorrell RG, Drew J, Nisbet RE, Howe CJ**. 2014. Evolution of chloroplast transcript processing in *Plasmodium* and its chromerid algal relatives. *PLoS Genetics* **10**, 1004008.