COMPARATIVE GENOMICS OF ENDOSYMBIOTICALLY-DERIVED ORGANELLES IN CRYPTOPHYTE ALGAE

by

Christa E. Moore

Submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
June 2013

DALHOUSIE UNIVERSITY

DEPARTMENT OF BIOCHEMISTRY AND MOLECULAR BIOLOGY

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled “COMPARATIVE GENOMICS OF ENDOSYMBIOTICALLY-DERIVED ORGANELLES IN CRYPTOPHYTE ALGAE” by Christa E. Moore in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

Dated: June 6, 2013

External Examiner: ______________________________

Research Supervisor: ______________________________

Examining Committee: ______________________________

____

Departmental Representative: ______________________________

DALHOUSIE UNIVERSITY

DATE: June 6, 2013

AUTHOR: Christa E. Moore

TITLE: COMPARATIVE GENOMICS OF ENDOSYBIOTICALLY-DERIVED ORGANELLES IN CRYPTOPHYTE ALGAE

DEPARTMENT OR SCHOOL: Department of Biochemistry and Molecular Biology

DEGREE: Ph.D. CONVOCATION: October YEAR: 2013

____

Signature of Author

DEDICATION PAGE

This thesis is dedicated to my husband, my loving family, and to my supervisor for pushing me to see just how far I could go. While I’ve enjoyed the ride, I’m happy to say that I’ve reached the end and am ready for the next adventure.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The cryptophytes are an enigmatic group of unicellular algae that acquired their plastids through the process of secondary endosymbiosis, which involved the uptake and retention of a red algal endosymbiont. There are several eukaryotic lineages that contain red algal-derived secondary plastids, but the cryptophytes and chlorarachniophytes are unique among them in that along with the plastid, they have retained the vestigial endosymbiont nucleus, called the nucleomorph. The ability to study cryptophyte nucleomorph genome biology and evolution provides a unique opportunity to investigate the processes of genome reduction, gene loss, and endosymbiotic gene transfer that followed the establishment of red secondary plastids. Here, I present the complete nucleomorph and plastid genome sequences of the photosynthetic marine cryptophyte *Chroomonas mesostigmatica* CCMP1168. My comparative analysis of cryptophyte nucleomorph genomes shows that there is a highly conserved core gene set, including an ultra-conserved set of plastid-related genes, but that there is also lineage-specific gene loss. The presence of pseudogenes and relict open reading frames in areas of gene order conservation indicate that nucleomorph genome reduction via gene loss is still occurring. My comparative analysis of cryptophyte plastid genomes reveals that the plastid genome architecture of photosynthetic species has been incredibly slowly evolving compared to other red secondary plastids. In addition, the identification of several group II introns in the *C. mesostigmatica* plastid genome, possibly acquired through lateral gene transfer from more than one source, suggests that cryptophyte plastid genomes may be more affected by exogenous genetic material than previously thought. Using phylogenetic analyses of expanded cryptophyte nucleomorph and plastid gene sets, I attempt to resolve the tree of cryptophytes and to identify the closest modern day red algal ancestor of the cryptophyte plastid and nucleomorph. Although the nucleomorph and plastid gene phylogenies are not congruent, I highlight meaningful inferences made by combining the comparative genomic and phylogenomic data. In terms of gene content and genome architecture, the cryptophyte plastid and nucleomorph genomes are the most similar to modern-day red algae of all known red algal secondary plastid-containing lineages. Study of these organellar genomes continues to shed light on the evolution of photosynethic eukaryotes.

LIST OF ABBREVIATIONS USED

aa	amino acids
ANOVA	analysis of variance
bp	base pairs
CCMP	Culture Collection of Marine Phytoplankton
cDNA	complementary deoxyribonucleic acid
CsCl	cesium chloride
DNA	deoxyribonucleic acid
EGT	endosymbiotic gene transfer
EST	expressed sequence tag
gDNA	genomic deoxyribonucleic acid
IEP	intron-encoded protein
ITS	internal transcribed spacer
Kbp	kilobase pairs
L	litre
LGT	lateral gene transfer
Mbp	megabase pairs
mL	millilitre
nORF	conserved nucleomorph hypothetical open reading frame
nr	non-redundant
ORF	open reading frame
ORFan	unique nucleomorph hypothetical open reading frame
PCR	polymerase chain reaction

PFGE	pulsed-field gel electrophoresis
PPC	periplastidial compartment
RCC	Roscoff Culture Collection
rDNA	ribosomal deoxyribonucleic acid
rRNA	ribosomal ribonucleic acid
SAR	stramenopiles, alveolates, rhizarians
snRNA	small nuclear ribonucleic acid
RT	reverse transcriptase
RT-PCR	reverse transcription polymerase chain reaction
tRNA	transfer ribonucleic acid
μg	microgram
ycf	conserved chloroplast hypothetical reading frame

ACKNOWLEDGEMENTS

I would like to thank several people for their help and support during my time as a graduate student. First and foremost, I would like to thank my supervisor, Dr. John Archibald, for giving me this opportunity and being nothing but supportive and encouraging throughout my entire program, from challenges in my research, to those in my personal life. I have certainly learned a great deal from him about the kind of mentor I hope to be. I would also like to thank various people in the Archibald Lab, past and present, who have helped me along the way, starting with Dr. Chris Lane and Natalie Donaher for taking me under their wings and helping me get my project off to a successful start. I would like to thank Dr. Bruce Curtis for helping me to acquire the bioinformatics skills needed for this work, for his contributions to my project, for his constant support and words of encouragement along the way, and for always taking the time to answer any question I had, regardless of how busy he was. For technical guidance over the years, I thank Drs Murray Schnare, David Spencer, and Eunsoo Kim. I would like to thank Naoko Tanifuji for assistance with computer-related issues, and for helping me to learn some of the more complicated molecular biology techniques required for this work. I also thank Dr. Goro Tanifuji for always so willingly sharing his knowledge and expertise on cryptophytes. I would also like to thank all of the other Archibald lab members, members of the CGEB labs, departmental administrators, Paul Briggs, and Heidi MacKinnon, for their support, useful discussions, and for making the department such an enjoyable place to work. I would like to extend a special thank you to the members of my advisory committee, Drs Andrew Roger, Alastair Simpson, and Ford

Doolittle, for all of the useful suggestions and feedback, and for their support and encouragement. Finally, I would like to thank my friends and family for their love, encouragement, and unwavering support throughout this long and challenging endeavour. I am incredibly fortunate to have so many cheerleaders.

CHAPTER 1 INTRODUCTION

Cryptophytes are biflagellate unicellular algae that are found extensively in aquatic habitats. The majority of cryptophytes are photosynthetic and live in marine environments, although there are some species that are non-photosynthetic, and some that live in fresh water. Aside from being important primary producers, cryptophytes are incredibly interesting from an evolutionary perspective because they evolved via a process of successive endosymbioses. In the first, or ‘primary’ endosymbiosis (Figure 1.1), a non-photosynthetic heterotrophic host engulfed and retained a cyanobacterial cell, which became established as the organelle known as the plastid (or chloroplast) (Cavalier-Smith 2000; Reyes-Prieto et al. 2007). Primary plastids are bounded by two membranes. These membranes are believed to correspond to the two membranes of the engulfed cyanobacterial cell (Cavalier-Smith1982), and that the phagosomal membrane, if indeed phagocytosis was the method of uptake, has been lost. Conversion of the cyanobacterium from endosymbiont to organelle was a complicated process involving genome reduction by way of gene loss and the massive transfer of genes from the cyanobacterial genome to that of the host, a process known as endosymbiotic gene transfer (EGT). While some protein products of transferred genes acquired new functions within host-derived compartments (Reyes-Prieto et al. 2006), a dedicated protein-targeting system was established to direct the protein products for a number of transferred genes back into the cyanobacterial compartment (Soll and Schleiff 2004; Gould et al. 2008). The resulting cyanobacterial-derived genome became orders of magnitude smaller compared to that of its free-living relatives. Modern-day plastids encode only a couple hundred proteins; however, to function properly the plastid requires import of at least one thousand more.

The establishment of the plastid and acquisition of photosynthesis was a pivotal event in eukaryote evolution. From this cellular union arose the glaucophytes, red algae, and green algae (from which the land plants are descended) (see review by Reyes-Prieto et al. 2007 and references therein). Red and green algae are very diverse and species-rich groups that have managed to infiltrate a wide range of marine, fresh water, and even extreme habitats, such as hypersaline or hyperthermal environments (Ciniglia et al. 2004; Lewis and McCourt 2004; Oren 2005, Yoon et al. 2006). Although the primary plastid-bearing red and green algae represent a great deal of algal diversity, the majority of marine phototrophs evolved as a result of secondary endosymbiosis.

In secondary endosymbiosis, a heterotrophic eukaryote engulfs and retains a photosynthetic red or green alga (Figure 1.1). The process of secondary endosymbiosis has resulted in a vast array of photosynthetic, and secondarily non-photosynthetic, eukaryotes of immense ecological and biological importance, including harmful bloom-causing algae and some human parasites. The uptake and retention of a green algal endosymbiont has occurred twice independently: once in the chlorarachniophytes, and once in the euglenids (Rogers et al. 2007; Takahashi et al. 2007). The plastids of cryptophytes, haptophytes, stramenopiles, dinoflagellates, and apicomplexans are derived from a red algal endosymbiont (Cavalier-Smith 1986; Kowallik et al. 1995; Wilson et al. 1996; Delwiche and Palmer 1997; Douglas et al. 2001; Yoon et al. 2002; Green 2004;

Sánchez Puerta et al. 2005), although at present it is unclear whether it was only one endosymbiotic event that gave rise to these lineages (e.g., Burki et al. 2007; Baurain et al. 2010; Burki et al. 2012).

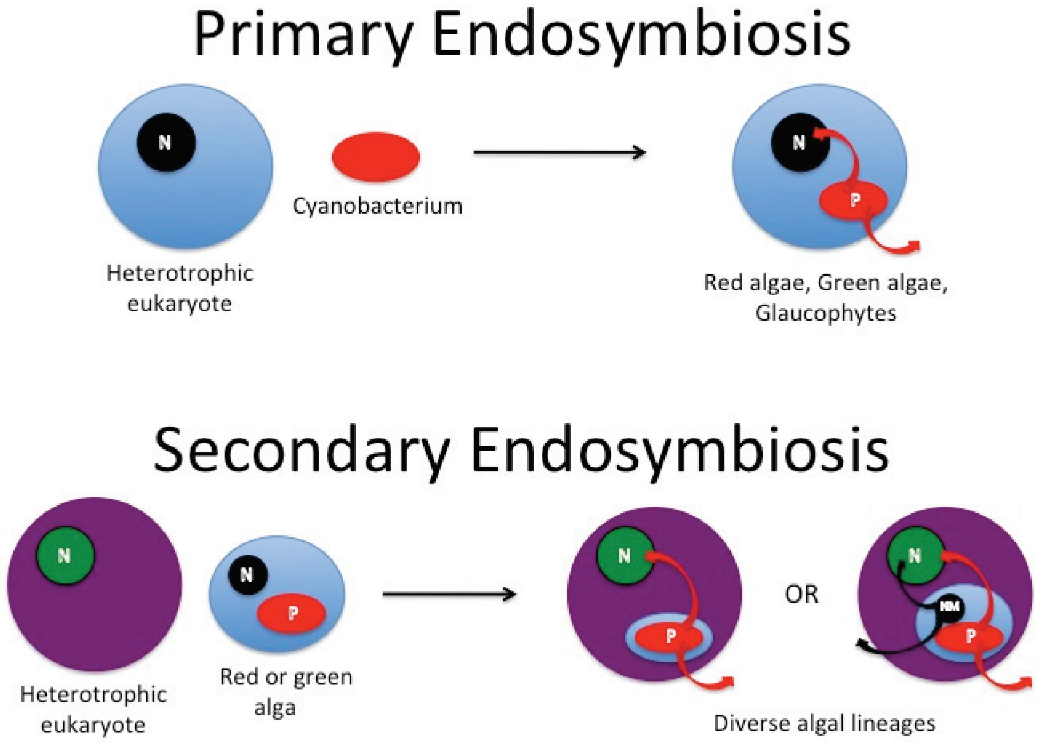


Figure 1.1 Primary and secondary endosymbiosis. In the primary endosymbiosis, a heterotrophic eukaryote engulfs and retains a cyanobacterium that becomes fixed as an organelle called a plastid. This endosymbiosis results in the evolution of red algae, green algae, and the glaucophytes. In secondary endosymbiosis, a heterotrophic eukaryote engulfs and retains either a red or green alga, resulting in the evolution of diverse algal lineages. In most lineages, the algal nucleus is lost. In two lineages, the engulfed algal nucleus remains in miniaturized form and is called the nucleomorph. Arrows represent the transfer of genes either from one genetic compartment to another, or gene loss. For simplicity, mitochondria are not shown. N = nucleus, P = plastid, NM = nucleomorph.

As a result of secondary endosymbiosis, the endosymbiont-turned-organelle is surrounded by four membranes: the inner two corresponding to the double membrane of the primary plastid, the third corresponding to the plasma membrane of the endosymbiont, and the outer membrane corresponding to the phagosomal membrane of the host (McFadden 1999). There are, however, a few variations in membrane topology. In the cryptophytes, haptophytes, and stramenopiles, the outermost membrane of the red-algal endosymbiont is contiguous with the host endomembrane and nuclear envelope. In the euglenids and dinoflagellates, the secondary plastid is surrounded by only three membranes, as one membrane, presumably the endosymbiont plasma membrane, has been lost (Cavalier-Smith 1999; Sulli et al. 1999; Nassoury et al. 2003).

Like primary endosymbiosis, the process of secondary endosymbiosis involves gene loss as well as the transfer of a large number of genes from the endosymbiont to the host nucleus. Evolution of a sophisticated protein targeting system was required to ensure that organelle-destined proteins, whose genes are now present in the host genome, are targeted across up to four membranes and into the appropriate compartment (Gould et al. 2008). In most secondary plastid-containing lineages, the endosymbiont has been reduced to the point that all that remains is the plastid; however, in two lineages, the cryptophytes and chlorarachniophytes, relicts of the endosymbiont nucleus, called the 'nucleomorph', and cytosol, called the periplastidial compartment (PPC), remain (Figure 1.2). Why nucleomorphs persist in these two lineages but have been completely lost in others remains an open question. Could it be that nucleomorphs are an evolutionary intermediate and that simply given more time, they will eventually disappear? Or perhaps there are factors that inhibit complete reduction, such as the inability of some plastid-related genes to be transferred to the host nucleus and their protein products imported back into the organelle.

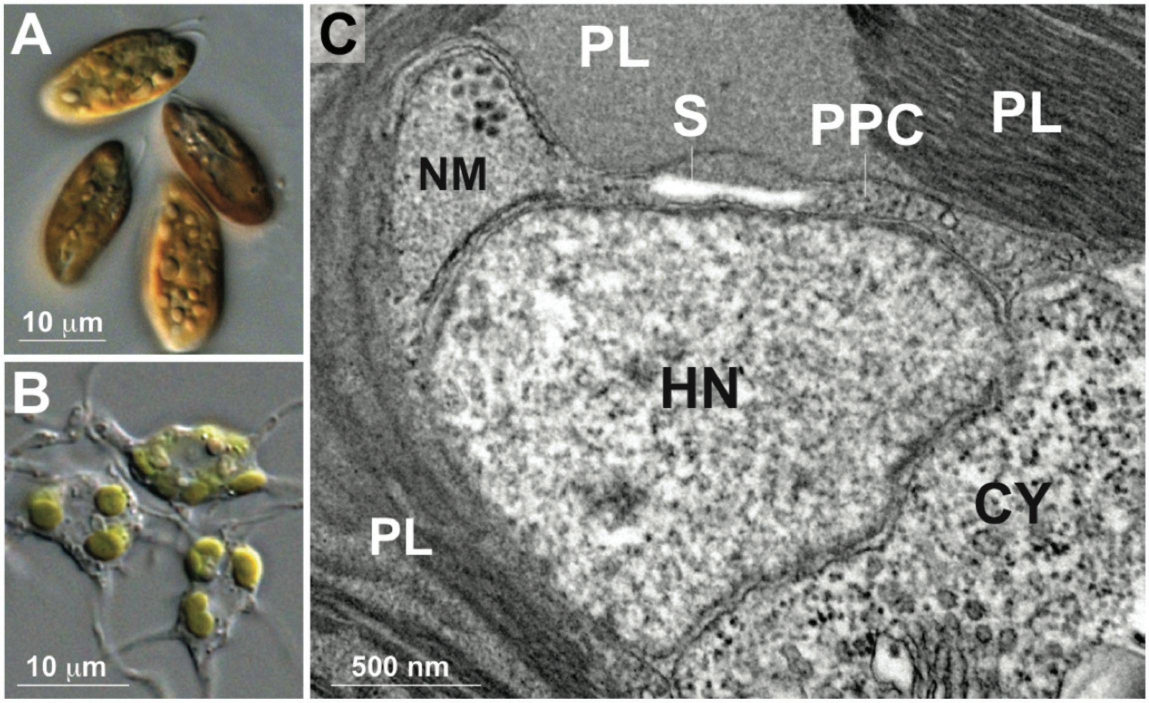


Figure 1.2 A) A light micrograph of the cryptophyte *Storeatula* sp. CCMP1868. Cryptophytes harbour a single nucleomorph-plastid complex derived from a red alga. B) A light micrograph of the chlorarachniophyte *Chlorarachnian reptans* CCMP238. Chlorarachniophytes harbour multiple nucleomorph-plastid complexes derived from a green alga. C) A transmission electron micrograph of the cryptophyte alga *Guillardia theta* CCMP2712 showing the host nucleus (HN), host cytosol (CY), and red algal-derived plastid (PL), nucleomorph (NM) and residual endosymbiont cytosol, the periplastidial compartment (PPC), which is the site of starch (S) production and storage. Image originally published in Christa E. Moore and John M. Archibald. 2009. Nucleomorph genomes. Annu Rev Genet. 43:251-264.

At the time that I began my research, there were only three completely sequenced nucleomorph genomes: one from the chlorarachniophyte *Bigelowiella natans* (Gilson et al. 2006), and two from the cryptophytes *Guillardia theta* (Douglas et al. 2001) and

Hemiselmis andersenii (Lane et al. 2007). About halfway through my research, the nucleomorph genome of the secondarily non-photosynthetic *Cryptomonas paramecium* was published (Tanifuji et al. 2011), raising the total number of nucleomorph genome sequences available to four. Although the cryptophyte and chlorarachniophyte nucleomorphs are derived from completely independent endosymbiotic events involving a red and green algal endosymbiont, respectively, they share many features in common, including very A+T rich (~75%), highly reduced (~330-1,030 Kbp) and compact genomes (~0.88-1.09 genes/Kbp) organized in a three chromosome architecture, with a small repertoire of protein genes (284-505) that are mostly housekeeping in nature, only a small number of which (17-30) are required for plastid functioning (Douglas et al. 2001; Gilson et al. 2006; Lane et al. 2007; Moore and Archibald 2009; Ishida et al. 2011; Tanifuji et al. 2011; Moore et al. 2012). The retention of genes encoding proteins targeted to the plastid was believed to be one of the main reasons why nucleomorphs persist in these two lineages, however, only two out of 47 retained plastid genes (out of presumably >1,000 ancestral cyanobacterial genes) are present in both cryptophyte and chlorarachniophyte nucleomorph genomes. Furthermore, the two genes in common, *hsp60* and *clpP*, have been transferred to the host nucleus in diatoms and apicomplexans, suggesting that there is not any one particular plastid protein-coding gene that is unable to be transferred (Gilson et al. 2006).

Another interesting feature of all sequenced nucleomorph genomes to date is the presence of sub-telomeric rRNA operons. The number of operons and their orientation varies from species to species (Douglas et al. 2001; Gilson et al. 2006; Lane et al. 2007; Silver et al.

2010; Tanifuji et al. 2011). Cryptophyte and chlorarachniophyte nucleomorph genomes also possess an abundance of hypothetical open reading frames that at the sequence level, show no detectable sequence similarity to any known gene in any database. These 'ORFan' genes comprise up to 30% of the total gene complement in nucleomorph genomes (Moore and Archibald 2009). Nucleomorph genes are also notoriously divergent in sequence relative to homologs in other organisms, which could explain why so many ORFs are unidentifiable, yet appear to be 'bona fide' genes as they are present in expressed sequence tag (EST) surveys (Patron et al. 2006), and are often greater than 1,000 bp in length (Lane et al. 2007).

Convergent evolution has resulted in a number of commonalities between the nucleomorph genomes of cryptophytes and chlorarachniophytes, but there are also some distinct differences. The *B. natans* nucleomorph genome is enriched in ultrashort (18-21 bp) spliceosomal introns; there are a total of 852 introns found in 240 protein-coding genes, resulting in an intron density of 3.1 introns/gene, which is on par with the intron density observed in modern green algal nuclear genomes (Gilson et al. 2006). Partial nucleomorph genomic data from other chlorarachniophytes show that the presence of numerous small spliceosomal introns is a common feature of chlorarachniophyte nucleomorph genomes (Slamovits and Keeling 2009). In stark contrast, there are only 17 spliceosomal introns in *G. theta* (Douglas et al. 2001), two in *C. paramecium* (Tanifuji et al. 2011), and none at all in *H. andersenii* (Lane et al. 2007). Another notable difference between the nucleomorph genome of *B. natans* and those of the cryptophytes is that *B. natans* lacks genes for proteasome-mediated protein degradation (Gilson et al. 2006).

These genes are still present in the nucleomorph genomes of cryptophytes (Douglas et al. 2001; Lane et al. 2007; Tanifuji et al. 2011). With only one chlorarachniophyte nucleomorph genome sequence currently available, it is difficult to identify any unifying characters that may explain the persistence of nucleomorphs in both the cryptophytes and chlorarachniophytes. Indeed, the independent nature of their origin may result in different reasons for their retention.

Comparative analyses of the three cryptophyte nucleomorph genome sequences have provided further insight into the genome reductive processes associated with secondary endosymbiosis. The *C. paramecium* nucleomorph genome contains a much smaller set of plastid-associated genes (17) than *G. theta* or *H. andersenii* (30). Given that *C. paramecium* has secondarily lost the ability to photosynthesize, a reduced set of plastid-associated protein-coding genes is unsurprising. Interestingly, photosynthesis-independent plastid-associated genes were also found to be missing in the *C. paramecium* nucleomorph genome. Similarly, the *H. andersenii* nucleomorph genome is completely devoid of spliceosomal introns, and as such, is deficient in spliceosome-associated protein-coding genes. Another notable difference between the three cryptophyte nucleomorph genomes is their size: the *C. paramecium* nucleomorph genome is the smallest, at 485.9 Kbp, followed by *G. theta* at 550.5 Kbp, and then *H. andersenii* at 571.4 Kbp. Although there is up to an 85.5 Kbp size difference, genome size is not correlated with the number of genes present, but rather is affected by the lengths of the genes themselves, especially the mysterious ORFan genes (Tanifuji et al. 2011).

Although there are distinct differences between the three cryptophyte nucleomorph genomes, there are also some important similarities in addition to those present in all nucleomorph genomes described earlier (gene density, three chromosome architecture, etc.), the most notable of which is a conserved set of core genes. Tanifuji et al. (2011) identified a set of 230 conserved proteins from *C. paramecium* that could be compared with the set of 234 conserved proteins from *G. theta*, and the set of 245 conserved proteins from *H. andersenii*. These protein sets excluded plastid-associated and spliceosome-related proteins. Out of the 230 *C. paramecium* proteins, 217 (94.3%) were found to be present in all three species. While there is clearly a highly conserved set of genes that primarily function in nucleomorph maintenance and replication, there are also a substantial number of genes for which functions cannot be ascribed, many of which are ORFan genes and unique to each species, but others that are conserved amongst the cryptophyte nucleomorphs. These 'nORF's are hypothetical protein-coding genes that are found to show homology only to other nucleomorph hypothetical protein-coding genes. Despite retaining no, or very little, similarity at the sequence level, several of these nORFs do retain positional conservation as they are found within syntenic regions (regions of gene order conservation) in the cryptophyte nucleomorph genomes (Lane et al. 2007; Tanifuji et al. 2011).

Although our knowledge of nucleomorph genome biology and evolution has improved over the past several years with the sequencing of additional cryptophyte nucleomorph genomes, many questions still remain. What are the functions of the mysterious ORFans and nORFs? Are there other factors that contribute to the nucleomorph genome size diversity observed? How and when were spliceosomal introns lost from the cryptophyte nucleomorph genome? To answer some of these questions, I sequenced and annotated the nucleomorph genome of *Chroomonas mesostigmatica* CCMP1168. In Chapter 2 I describe the main features of the *C. mesostigmatica* nucleomorph genome and compare it to the other cryptophyte nucleomorph genomes of *H. andersenii*, *G. theta*, and the non-photosynthetic *Cryptomonas paramecium.*

A thorough understanding of nucleomorph genome evolution, however, is incomplete without considering its accompanying plastid. Based on the red algal plastid genome sequences available, it appears as though the diversity of red algae is not reflected in their plastid genome sequences. Sequenced red algal plastid genomes range in size from ~150-191 Kbp and contain 223-250 genes, with essentially no unique hypothetical ORFs. The majority of protein-coding genes are conserved amongst all the plastid genomes (Reith and Munholland 1995; Glöckner et al. 2000; Ohta et al. 2003; Hagopian et al. 2004; Smith et al. 2012; Janouškovec et al. 2013). Red algal plastid genomes are the most gene rich plastid genomes and contain the largest complement of cyanobacterial genes found in primary or secondary plastid genomes. In addition, red algal plastid genomes are relatively slow evolving compared to those of other primary or secondary plastids (Janouškovec et al. 2013). The plastid genomes of cryptophytes are reduced in size and content relative to primary red algal plastids. The smallest cryptophyte plastid genome currently sequenced is that of the non-photosynthetic *C. paramecium*. At 77.7 Kbp in size (Donaher et al. 2009), it is substantially smaller than the plastid genomes of the photosynthetic *G. theta* (121.5 Kbp; Douglas and Penny 1999) and *Rhodomonas salina*

(135.9 Kbp; Khan et al. 2007) and is missing 71 protein-coding genes present in the photosynthetic cryptophyte plastid genomes. Unsurprisingly many of the missing genes encode photosynthesis-related proteins. Cryptophyte plastid genomes are also found to contain a larger number of unique hypothetical protein-coding genes than red algal plastid genomes, and pseudogenes, which have not been found in red algal or other red algal-derived secondary plastid genomes. Another unique feature of cryptophyte plastid genomes is the presence of group II introns. Group II introns were found in the *groEL* gene of two *Rhodomonas* species, in the *psbN* gene of *R. salina*, and in the *chlB* gene of *Chroomonas pauciplastida* and *H. andersenii* (Khan and Archibald 2008), but they are not present in the plastid genomes of *C. paramecium* and *G. theta*. A phylogeny of cryptophyte group II intron-encoded proteins (IEPs) suggests that group II introns were laterally transferred from a euglenid (Khan and Archibald 2008). Group II introns have not been found in other red algal-derived secondary plastid genomes, and have only recently been identified in the tRNA-Met genes of some red algal plastid genomes. Whether the group II introns of cryptophyte plastids were acquired both through lateral gene transfer (LGT) or vertical inheritance is unclear. In Chapter 3 I describe the plastid genome of *C. mesostigmatica* CCMP1168 and compare it to the cryptophyte and red algal plastid genomes currently available. In the same chapter I re-examine the distribution and origin of group II introns in cryptophyte plastid genomes, as well as compare gene content and synteny in the context of ‘red’ plastid genome evolution.

In addition to complete plastid and nucleomorph genome sequences across the diversity of cryptophytes and red algae, a well resolved phylogeny of cryptophytes is needed to answer questions about the extent of nucleomorph gene loss and whether there is a pattern for the retention of genes. Previous attempts to reconstruct the tree of cryptophytes have focused primarily on the use of 18S ribosomal protein genes, both nuclear and nucleomorph-encoded, as there are so few nucleomorph genomes, and until recently, not a single cryptophyte nuclear genome was available. From these single-gene trees, several clades of cryptophytes have been identified, although the relationships between the various clades remain unclear (Hoef-Emden 2008). Furthermore, the identity of the red algal ancestor of the cryptophyte nucleomorph is currently unknown. Like the cryptophytes, there is very little red algal nuclear data presently available. There are only two completely sequenced red algal nuclear genomes: one from *Cyanidioschyzon merolae*, a unicellular alga that inhabits acidic hot springs (Matsuzaki et al. 2004), and one from the seaweed known as “Irish moss”, *Chondrus crispus*, which was only recently completely sequenced (Collén et al. 2012). Identification of the closest living relative to the red alga that gave rise to the secondary plastid and nucleomorph of cryptophytes will provide a reference point for more accurately assessing the reductive processes that have shaped modern-day nucleomorphs and plastids. In Chapter 4 I attempt to better resolve the tree of cryptophytes and to elucidate the identity of the red algal endosymbiont that gave rise to the nucleomorphs of cryptophyte algae by using multi-protein data sets that contain nucleomorph and plastid protein sequences from *C. mesostigmatica* combined with those of previously sequenced cryptophyte plastid and nucleomorph genomes, and recently available red algal complete genomic data and EST surveys.

CHAPTER 2 NUCLEOMORPH COMPARATIVE GENOMICS

This chapter includes work published in Christa E. Moore, Bruce Curtis, Tyler Mills, Goro Tanifuji, and John M. Archibald. 2012. Nucleomorph genome sequence of the cryptophyte alga *Chroomonas mesostigmatica* CCMP1168 reveals lineage-specific gene loss and genome complexity. *Genome Biology and Evolution* 4(11):1162-1175. The original article was published by Oxford University Press. I performed the majority of the methods and analyses described, and composed most of the manuscript.

2.1 Cryptophyte Nucleomorph Genome Biology and Evolution

Nucleomorph genomes are the smallest nuclear genomes known and range in size from ~330 – 1,030 kilobase pairs (Kbp) (Silver et al. 2007; Phipps et al. 2008; Tanifuji et al. 2010; Ishida et al. 2011), orders of magnitude smaller than even the most reduced genomes of eukaryotic parasites, such as the 2.9 megabase pair (Mbp) genome of the microsporidian *Enchephalitozoon cuniculi* (Katinka et al. 2001). The genomes of these miniature nuclei have shrunk dramatically in size and content over millions of years to ~1 Mbp or less and with only several hundred genes. The process of genome reduction has resulted in most of the genes being lost or transferred to the host nucleus, streamlining of the intergenic spacers, and almost complete elimination of repetitive sequence. To date, three cryptophyte nucleomorph genomes have been sequenced, those of *Guillardia theta* (Douglas et al. 2001)*, Hemiselmis andersenii* (Lane et al. 2007), and the secondarily non-photosynthetic *Cryptomonas paramecium* (Tanifuji et al. 2011), which are 550.5 Kbp, 571.4 Kbp, and 485.9 Kbp in size, respectively. A single chlorarachniophyte nucleomorph genome has also been sequenced, the 373 Kbp nucleomorph genome of

Bigelowiella natans (Gilson et al. 2006). With this limited sampling, nucleomorph genome comparisons within the chlorarachniophyte lineage are impossible, and between the chlorarachniophyte and cryptophyte lineages limited. Consequently, we know little about the evolutionary forces that have shaped these genomes and why nucleomorphs persist in chlorarachniophytes and cryptophytes but have been lost in other secondary plastid-bearing algae (reviewed by Moore and Archibald 2009).

Comparative studies of the three sequenced cryptophyte nucleomorph genomes reveal striking similarities with respect to genome architecture and composition. All three genomes (and in fact all nucleomorph genomes examined to date) have three small chromosomes with ribosomal deoxyribonucleic acid (rDNA) operons on the chromosome ends and with one of two types of unusual telomere sequences: GA_n (GA_{17} for *H. andersenii* and GA_9 for *C. paramecium*) and $[AG]_7AAG_6A$ for *G. theta* (Douglas et al. 2001; Lane et al. 2007; Silver et al. 2007; Tanifuji et al. 2010; Tanifuji et al. 2011). These genomes display a similar degree of nucleotide composition bias (~75% A+T) and have similar coding capacities (518-548 genes). This latter point is interesting given that their total genome sizes differ by up to 64 Kbp, yet they have very similar gene densities (0.98-1.09 gene/Kbp). Approximately 60% of the genes annotated in these genomes encode proteins involved in core eukaryotic processes, such as transcription, translation, and protein folding, but the remaining ~40% cannot be ascribed a particular function based on sequence similarity as they either show homology only to other cryptophyte nucleomorph genes of unknown function, or they show no similarity whatsoever to any known gene in current databases. Essentially nothing is known about the latter 'ORFan' genes except that their transcripts have been observed in EST surveys (e.g., Patron et al. 2006), and they tend to encode proteins rich in amino acids specified by A+T-rich codons (Lane et al. 2007).

Nucleomorph gene sequences are notoriously divergent compared to their homologs in free-living organisms, and are often shorter as a result of internal deletions and the whittling away of amino and carboxy terminal-coding regions (Lane et al. 2007). In addition, spliceosomal introns are rare in cryptophyte nucleomorph genomes and are in fact completely absent in the case of *H. andersenii*, the first described instance of complete spliceosomal intron loss from a nuclear genome (Lane et al. 2007). Of the known genes present in nucleomorph genomes, there are very few whose protein products function in the plastid, which is surprising given that nucleomorph-encoded ‘plastid’ genes are often touted as the primary reason for nucleomorph persistence (Zauner et al. 2000; Gilson and McFadden 2002; Archibald 2007). To gain a better understanding of the evolution and ultimate fate of nucleomorphs in cryptophyte algae, I completely sequenced the nucleomorph genome of *Chroomonas mesostigmatica* CCMP1168. Members of the genus *Chroomonas* have predicted nucleomorph genome sizes that are >200 Kbp larger than those currently sequenced (Tanifuji et al. 2010; Lane et al. 2006). The nucleomorph genome of *C. mesostigmatica* is the largest and the most complex of its kind, with numerous repetitive regions and multi-copy genes, features that are rare in nucleomorph genomes sequenced to date. Comparative analyses provide insight into the identity of some of the mysterious ORFan genes, evidence for a more highly conserved core set of genes than previously thought, and further support for the notion that nucleomorphs have yet to reach an endpoint in their reductive evolution.

2.2 METHODS

2.2.1 Cell Culture and Pulsed-Field Gel Electrophoresis

Chroomonas mesostigmatica CCMP1168 was grown in f/2 media at room temperature on a 12-hour light/dark cycle with constant aeration using a stir bar and plate (Figure 2.1).

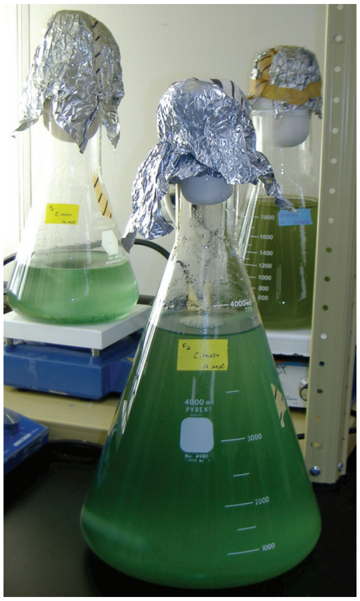


Figure 2.1 *Chroomonas mesostigmatica* CCMP1168 cultures grown in f/2 media with constant aeration in 4 L (front center) and 1 L (back left) flasks.

A previous karyotype analysis by Lane et al. (2006) predicted the *C. mesostigmatica* nucleomorph genome size to be ~805 Kbp. To verify this estimate, I created agarose plugs for pulsed-field gel electrophoresis (PFGE) from 400 mL of log-phase culture following the method described in Eschbach et al. (1991) for three approximate cell counts of 5x10^6, 1x10^7, and 5x10^7 cells per plug. The agarose plugs were run in a CHEF-DR III Pulsed-Field Electrophoresis System (BioRad Laboratories, Hercules, CA, USA) on a 1% agarose gel dissolved in 0.5% TBE buffer (TRIS, boric acid, EDTA) at 14ºC for 22 hours using a voltage of 6.0 V/cm and 0.2-22 second switch time. Pulsed-field gel electrophoresis allows for the separation of intact chromosomes from the various genetic compartments (Figure 2.2A) based on size. To visualize the nucleomorph chromosomes under ultraviolet light, the gel was stained with ethidium bromide (Figure 2.2B).

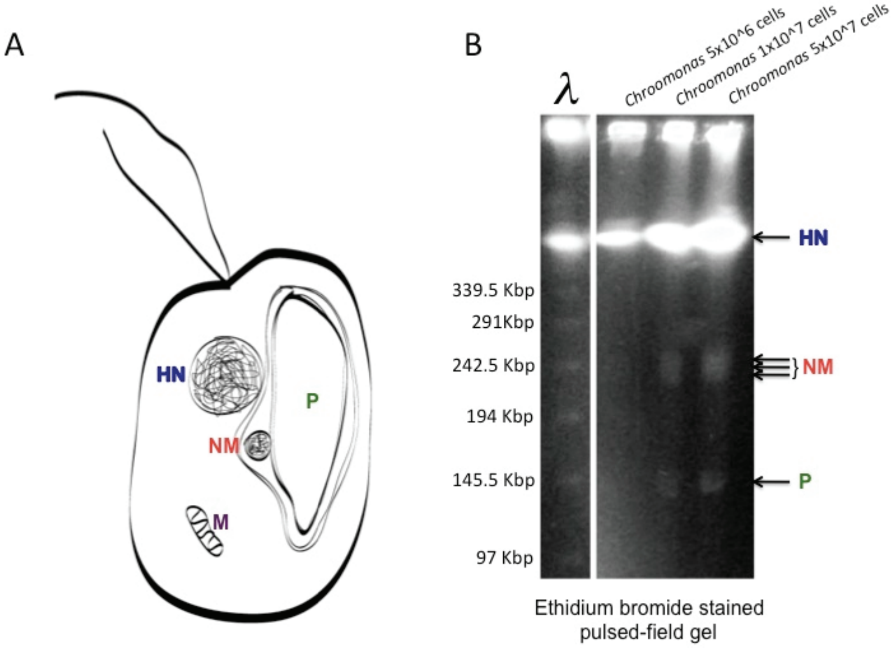


Figure 2.2 A) Schematic of the various genetic compartments in cryptophyte cells. HN = host nucleus, P = plastid, NM = nucleomorph, and M = mitochondrion. B) Ethidium bromide stained pulsed-field gel of *C. mesostigmatica* CCMP1168 agarose plugs at three different cell counts: 5x10^6, 1x10^7, and 5x10^7 cells per plug. Lambda DNA is used as a size ladder. The host nuclear chromosomes can be seen as a large band near the top. The plastid genome is also present near the 145 Kbp size marker. Using a standard logarithmic curve, the three nucleomorph chromosomes were found to have approximate sizes of 267.5 Kbp, 262.5 Kbp, and 257.5 Kbp for a total genome size of 787.5 Kbp. The actual genome size is likely to differ from this estimate, however, because the amount of DNA in the plug affects the migration rate, and the chromosomes are not well resolved.

2.2.2 Nucleomorph DNA Isolation

To obtain a DNA sample enriched in nucleomorph DNA for sequencing, total DNA was extracted from a total of 120 L of dense culture using a standard phenol/chloroform extraction procedure. Total DNA was then fractionated by Hoechst dye (No 33258, Sigma-Aldrich, St. Louis, MO, USA)-cesium chloride (CsCl) density gradient centrifugation (Figure 2.3A).

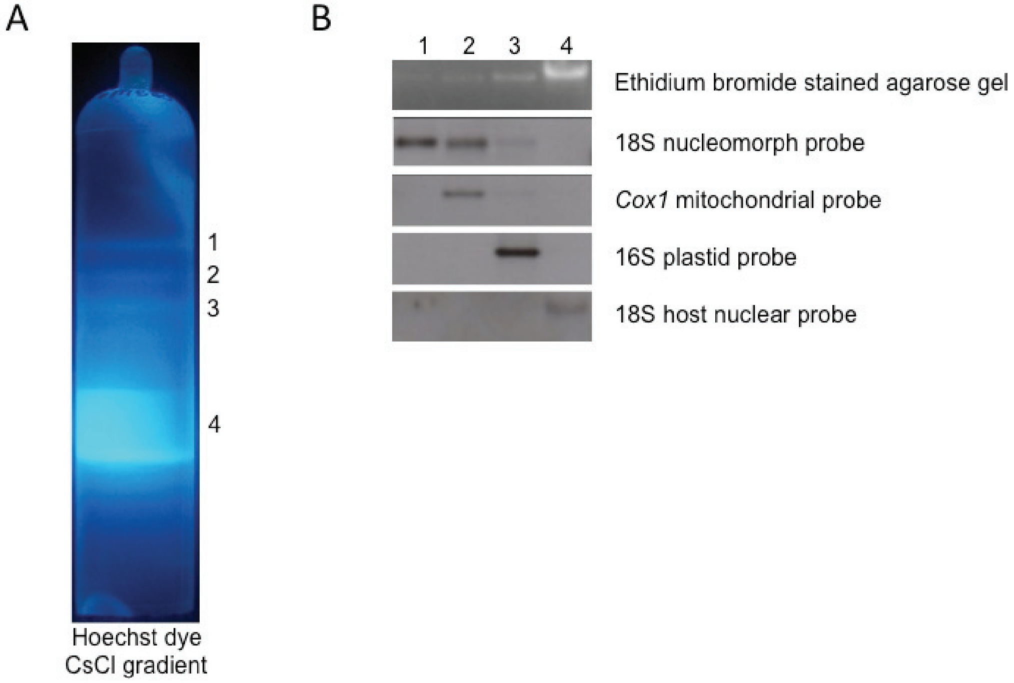


Figure 2.3 A) Separation of *C. mesostigmatica* CCMP1168 total DNA by Hoechst dye-CsCl density gradient centrifugation. B) Samples from fractions 1, 2, 3, and 4 of the gradient were resolved by agarose gel electrophoresis (top). The DNA was transferred to a nylon membrane for Southern blot hybridizations using species-specific 18s rDNA host nuclear and nucleomorph probes, a species-specific *cox1* mitochondrial probe, and a 16S rDNA plastid probe (*Rhodomonas* sp. CCMP1178).

The resulting fractions were analyzed by Southern blot hybridizations using organelle genome-specific probes to identify fractions of predominantly mitochondrial (species-specific *cox*1), plastid (*Rhodomonas* sp. CCMP1178 16S rDNA), nuclear (species-specific host 18S rDNA), and nucleomorph (species-specific endosymbiont 18S rDNA) origin (Figure 2.3B). Hoechst dye-CsCl density gradient centrifugation and fraction purification, and Southern blot hybridizations were performed as described in Lane and Archibald (2006) and Lane et al. (2006). The top fraction of the Hoechst dye-CsCl gradient was found to be enriched in nucleomorph DNA. To obtain enough nucleomorph DNA for 454 pyrosequencing, the top fractions from four Hoechst dye-CsCl gradients were combined.

2.2.3 DNA Sequencing and Genome Assembly

A sample containing approximately 5 micrograms (μg) of nucleomorph genome-enriched DNA was 454 pyrosequenced (GS FLX titanium series, McGill University and Génome Québec Innovation Centre, Montréal, QC, Canada) to a depth of ~100X coverage, generating over 110 Mbp of raw sequence data. Reads over 300 base pairs (bp) in length were assembled using the GAP4 program (Staden package, v4.11; Bonfield et al. 1995). As the DNA sample contained contaminating organellar and bacterial DNAs, nucleomorph-derived contigs were identified by blastx searches against the GenBank non-redundant (nr) database (National Center for Biotechnology Information, Bethesda, MD, USA), and contigs were manually refined.

In order to determine which contigs belonged to which chromosome, I repeated the PFGE experiment described above using plugs with 5 x 10^7 cells, a voltage of 4.1 V/cm, and a 30-10 second switch time for 60 hours. This resulted in better resolution of the nucleomorph chromosomes (Figure 2.4). The contigs were assigned to their respective chromosomes using Southern blot hybridizations with probes (see Appendix A for primer sequences) that were either gene specific, or contained repetitive DNA (Figure 2.4).

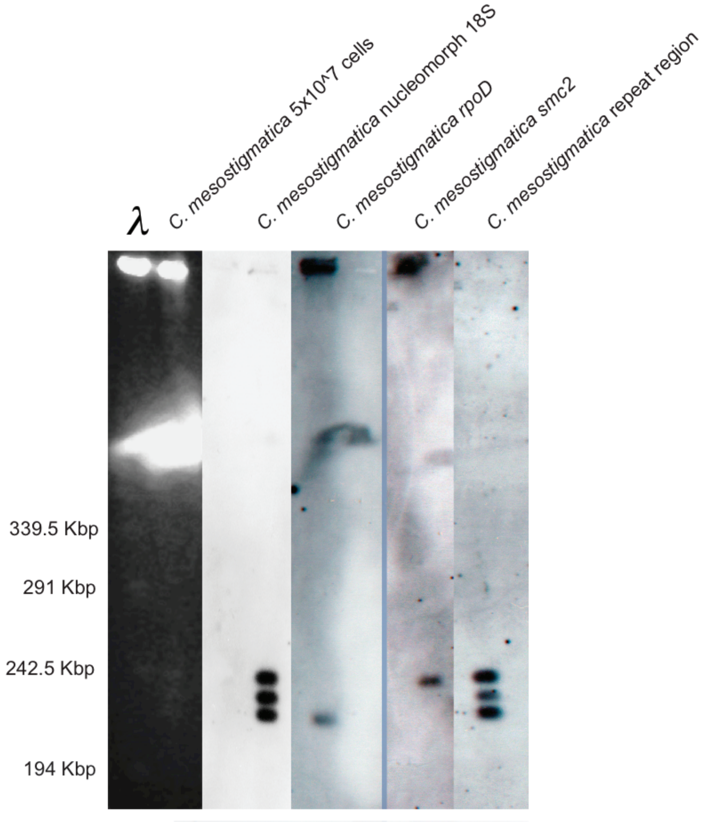

λ
C. mesostigmatica 5x10^7 cells
C. mesostigmatica nucleomorph 18S
C. mesostigmatica rpoD
C. mesostigmatica smc2
C. mesostigmatica repeat region
339.5 Kbp
291 Kbp
242.5 Kbp
194 Kbp

Figure 2.4 Ethidium bromide-stained *C. mesostigmatica* nucleomorph chromosomes separated by PFGE and lambda DNA as a ladder (left). To the right of the pulsed-field gel, a southern blot hybridization using a *C. mesostigmatica* CCMP1168 nucleomorph-specific 18S rDNA probe shows the positions and resolution of the nucleomorph chromosomes. Southern blot hybridizations using *rpoD* and *smc2* probes were used to assign the contigs containing these genes to chromosomes 3 and 1, respectively. A probe containing a repeat region shows that the repeat is present in all three chromosomes.

Based on the contig mapping, the remaining gaps were closed using polymerase chain reaction (PCR) with site-specific primers. PCR products were cloned into pGEM-T Easy vectors (Promega, Madison, WI, USA) and Sanger sequenced (GENEWIZ, Cambridge, MA, USA).

An additional 1 μg of nucleomorph-enriched DNA was sequenced on an Illumina GAIIx sequencer (Cofactor Genomics, St. Louis, MO, USA), producing 833,000 reads, 90% of which were integrated into the 454-based assembly to aid in frameshift and homopolymer correction.

2.2.4 RNA Extraction and Transcriptome Sequencing

For transcriptome sequencing, total RNA was extracted from ~3.3x10^7 cells using Trizol (Invitrogen, Burlington, ON, Canada), followed by standard phenol/chloroform precipitation, and subsequent precipitation using lithium chloride. A 10 μg sample of RNA was sequenced using Illumina RNA-Seq (National Center for Genome Resources, Santa Fe, NM, USA), generating 2.37 gigabase pairs of raw sequence data. Raw reads from the RNA-Seq data were mapped onto the contigs using the Burrows-Wheeler Aligner (v0.5.9 with default settings; Li and Durbin 2009) to further verify the assembly and aid in spliceosomal intron boundary identification.

2.2.5 Genome Annotation

Open reading frames (ORFs) larger than 50 amino acids (aa) in size were predicted using Artemis (v13.0; Rutherford et al. 2000) and genes were manually annotated based on blastx and blastp (e value < 0.001; Altschul et al. 1990) searches against the GenBank nr database (web interface: http://blast.ncbi.nlm.nih.gov/Blast.cgi) as well as a local database of red algal (*Cyanidioschyzon merolae*) and cryptophyte nucleomorph genomic data (*G. theta, H. andersenii,* and *C. paramecium*). Pfam searches were also performed (web interface: http://pfam.sanger.ac.uk/) (Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, UK). Gene annotations followed the conventions of Douglas et al. (2001), Lane et al. (2007), and Tanifuji et al. (2011). For ease of comparison with other cryptophyte nucleomorph ORFs, I categorized each of the *C. mesostigmatica* nucleomorph ORFs either as a 'conserved ORF', a 'nORF', or an 'ORFan'. A conserved ORF is a protein-coding gene with annotated homologs in other nuclear (and in most cases, other cryptophyte nucleomorph) genomes. A nORF is a hypothetical protein-coding gene with annotated homologs only in other cryptophyte nucleomorph genomes. An ORFan is a hypothetical protein-coding gene with no significant sequence similarity to any gene in current databases.

Transfer RNAs (tRNA) were identified using tRNAScan-SE (v1.21; Lowe and Eddy 1997) (http://lowelab.ucsc.edu/tRNAscan-SE/), and ribosomal RNAs (rRNA) were identified using blastn searches against the GenBank nr database. One small nuclear RNA (snRNA) was identified using a blastn search against my local database. Spliceosomal introns were identified manually using canonical GT/AG intron boundary searches and alignments of homologous protein sequences.

2.2.6 Reverse-Transcription PCR and Intron Verification

RNA for reverse-transcription (RT) PCR experiments to verify the predicted spliceosomal intron boundaries was obtained using the RNeasy® Mini and RNeasy® MinElute™ Cleanup kits (Qiagen, Toronto, ON, Canada). The quality and quantity of RNA extracted was assessed by gel electrophoresis. Two site-specific primers per gene, one upstream of the 5’ intron splice site and one downstream of the 3’ intron splice site, were designed for RT-PCR verification of predicted spliceosomal introns in the following genes: *rps*16 (Cmeso_rps16_F1: 5'-CATAGTCCAAGTATTCGGAAAAA-3' and Cmeso_rps16_R1: 5'-GCTCCTTTTCCTCCTGCTTT-3'), *rps*23 (Cmeso_rps23_F1: 5'-TGTTTAAATAAAAGAATGGGATCAG-3' and Cmeso_rps23_R1: 5'-TCAAAGAAACCCCTGCTACC-3'), *rps*24 (Cmeso_rps24_F1: 5'-GGAAGAAATCAAAATTACCACCA-3' and Cmeso_rps24_R1: 5'-TGTTAAAACGAGTTTTTCCTCTGA-3'), and *rpl*9 (Cmeso_rpl9_F1: 5'-GCATGAAACCAATTCTAACAAACA-3' and Cmeso_rpl9_R1: 5'-CGGCACCAGCAGATACAAG-3'). Reverse transcription (RT) reactions using the site-specific primers were performed by previous honours student Tyler Mills for his honours research project, under my supervision, using the Omniscript™ RT kit (Qiagen, Toronto, ON, Canada), followed by PCR amplification of the resulting cDNA. The same site-specific primers were also used in PCR reactions with total genomic DNA template. Amplicon sizes were determined and compared using gel electrophoresis (Figure 2.5). Purified cDNA PCR products for each of the genes were then cloned using the TOPO-XL vector (Invitrogen, Burlington, ON, Canada) and Sanger-sequenced using a Beckman-Coulter CEQ 8000 capillary DNA sequencer.

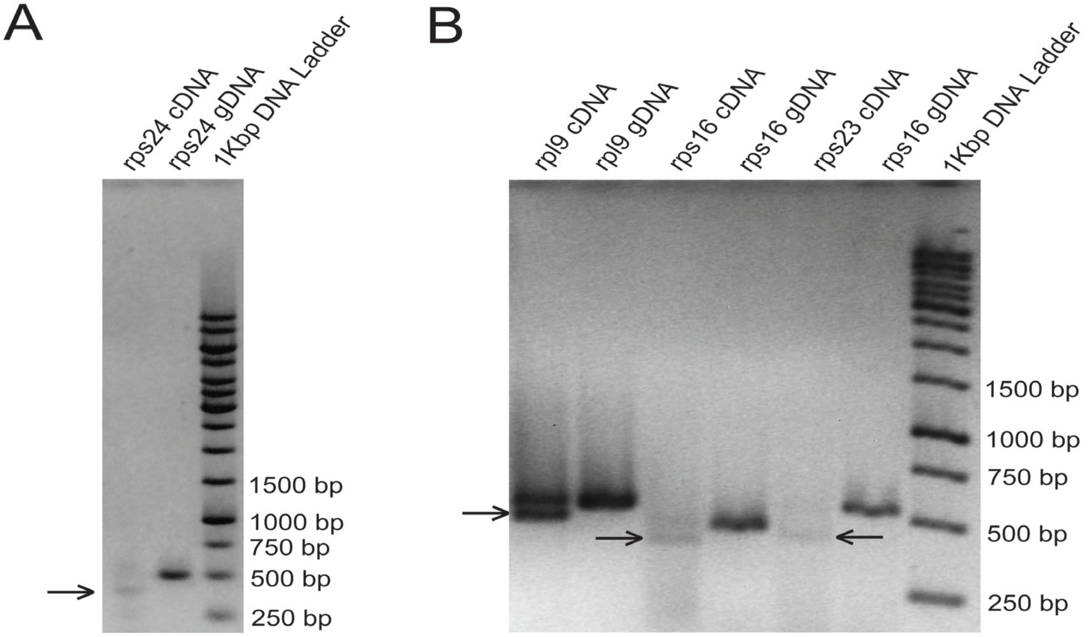


Figure 2.5 Verification of spliceosomal intron removal in the *Chroomonas mesostigmatica* nucleomorph genome. The figure shows agarose gel electrophoresis of PCR amplicons generated using cDNA and genomic DNA template and site-specific primers. Genes examined were (A) rps24, and (B) rpl9, rps16, and rps23. The cDNA amplicons (indicated by arrowheads) for each gene are shorter in length compared to their respective PCR-generated genomic DNA amplicons. Intron removal was verified by sequencing.

2.2.7 Significance Testing and Data Deposition

The statistical significance of protein and intergenic spacer size differences was determined for one-way analysis of variance (ANOVA) and multiple comparisons between *C. mesostigmatica*, *H. andersenii*, *C. paramecium*, and *G. theta* nucleomorph data using AnalystSoft Inc., StatPlus:mac – statistical analysis program for Mac OS, version 2009 (www.analystsoft.com/en/).

The complete nucleomorph genome sequence of *C. mesostigmatica* CCMP1168 has been deposited in GenBank using the following accession numbers: CP003680, CP003681 and CP003682.

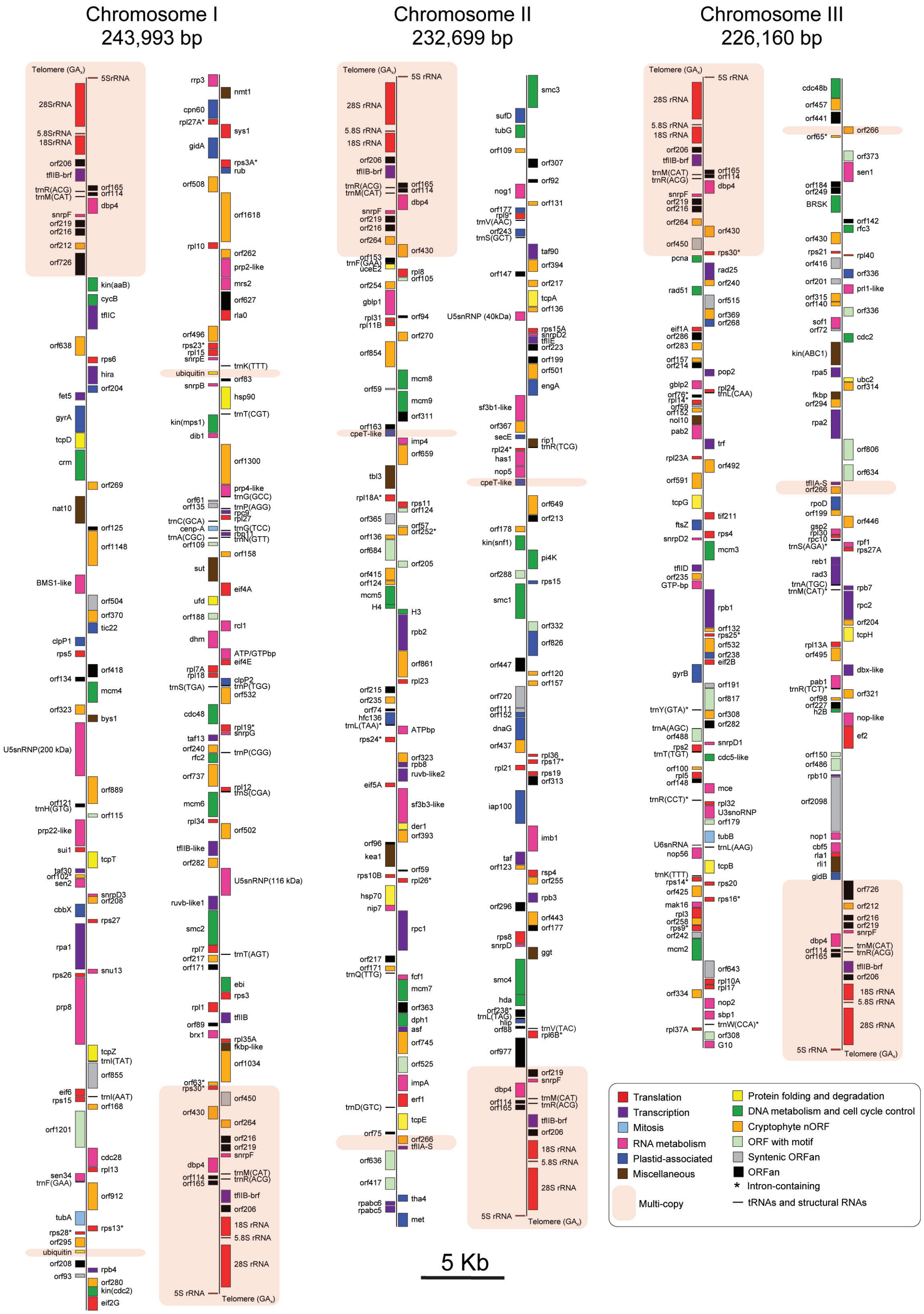
2.3 Results and Discussion

2.3.1 Genome Architecture and Size Variation

The nucleomorph genome of *C. mesostigmatica* CCMP1168 is comprised of three linear chromosomes of ~244 Kbp, 233 Kbp, and 226 Kbp, with a total genome size of 702.9 Kbp (Figure 2.6). A previous karyotyping analysis of the *C. mesostigmatica* nucleomorph showed three similarly sized chromosomes totaling approximately 805 Kbp (Lane et al. 2006). My independent analysis suggested that the three nucleomorph chromosomes are smaller than the original size estimates, which is supported by the total genome size determined by sequencing. Although smaller than initial PFGE-based estimates, the *C. mesostigmatica* nucleomorph genome is still the largest nucleomorph genome sequenced to date. The G+C content is 25.94%, similar to that seen in other cryptophyte nucleomorph genomes (Table 2.1). The telomere sequence is GA_{13}, similar to the GA_{17} and GA_9 nucleomorph telomere sequences of *H. andersenii* (Lane et al. 2007) and *C. paramecium* (Tanifuji et al. 2011), respectively. Sub-telomeric rDNA operons exist on all six chromosome ends, followed by a long stretch (up to 13 Kbp) of repeated sequence consisting of several ORFs (for both hypothetical proteins and proteins of known function) and repetitive sequence that has presumably been homogenized through recombination. There are several other multi-copy genes that appear on more than one chromosome (two copies of ubiquitin, cpeT-like and tfIIA-S genes, and three copies of orf266), and in each case the duplicates are essentially identical to one another. The majority of the genes are, however, present in single copy.

Unlike all other nucleomorph genomes sequenced to date, the *C. mesostigmatica* nucleomorph genome is rich in simple, repetitive sequence, consisting primarily of innumerable A and T homopolymer runs of varying length (in coding and non-coding sequence), short sequence repeats (in the intergenic regions as well as in the internal transcribed spacer (ITS) regions of the rDNA operon), and a 12-bp repeat ($TA_2GA_2TA_5$, 4-25 copies) on five of the six chromosome ends. In addition to repetitive sequence in the intergenic spacers, there is also repetitive sequence present within both protein-coding genes and ribosomal RNA genes. The variable regions of the 28S large subunit ribosomal RNA gene contain lengthy A and T homopolymer runs (up to 37 bp), as well as short sequence repeats. Repetitive elements in the variable regions of rDNA genes have been observed in other organisms, and can mimic the base composition of the ITS sequences (see Gray and Schnare 1990 and references therein).

Figure 2.6 *Chroomonas mesostigmatica* nucleomorph genome map. The genome is comprised of three linear chromosomes, shown broken artificially at their midpoints, with genes on the left indicating transcription from bottom to top, and genes on the right indicating transcription from top to bottom. Colors of the blocks correspond to assigned functional categories and multi-copy genes are highlighted in pink. An asterisk beside the gene name indicates the gene contains an intron. Genes for which there are currently no known homologs (ORFans) are shown in black, genes that have homologs only in other cryptophyte nucleomorphs (nORFs) are shown in orange, and motif-containing genes whose identity cannot be determined with confidence are shown in light green. ORFan genes that retain conserved positions within syntenic regions between one or more other cryptophyte nucleomorphs are shown in grey (syntenic ORFans).


Chromosome I
243,993 bp
Chromosome II
232,699 bp
Chromosome III
226,160 bp
Telomere (GA$_n$)
5SrRNA
28SrRNA
5.8SrRNA
18SrRNA
orf206
tfIIB-brf
trnR(ACG)
trnM(CAT)
orf165
orf114
dbp4
snrpF
orf219
orf216
orf212
orf726
kin(aaB)
cycB
tfIIC
orf638
rps6
hira
orf204
fet5
gyrA
tcpD
crm
orf269
nat10
orf125
orf1148
BMS1-like
orf504
orf370
tic22
clpP1
rps5
orf418
orf134
mcm4
orf323
bys1
U5snRNP(200 kDa)
orf889
orf121
trnH(GTG)
orf115
prp22-like
sui1
tcpT
taf30
orf102*
sen2
snrpD3
orf208
cbbX
rps27
rpa1
rps26
snu13
prp8
tcpZ
trnI(TAT)
orf855
eif6
rps15
trnI(AAT)
orf168
orf1201
cdc28
rpl13
sen34
trnF(GAA)
orf912
tubA
rps28*
rps13*
orf295
ubiquitin
orf208
rpb4
orf93
orf280
kin(cdc2)
eif2G
rrp3
nmt1
cpn60
rpl27A*
sys1
gidA
rps3A*
rub
orf508
orf1618
rpl10
orf262
prp2-like
mrs2
orf627
rla0
orf496
rps23*
rpl15
snrpE
trnK(TTT)
ubiquitin
orf83
snrpB
hsp90
trnT(CGT)
kin(mps1)
dib1
orf1300
prp4-like
trnG(GCC)
orf61
orf135
trnP(AGG)
rpc9
rpl27
trnC(GCA)
cenp-A
trnG(TCC)
trnA(CGC)
rbp11
trnN(GTT)
orf109
orf158
sut
eif4A
ufd
orf188
rcl1
dhm
ATP/GTPbp
eif4E
rpl7A
rpl18
clpP2
trnP(TGG)
trnS(TGA)
orf532
cdc48
rpl19*
snrpG
taf13
orf240
trnP(CGG)
rfc2
orf737
rpl12
trnS(CGA)
mcm6
rpl34
orf502
tfIIB-like
orf282
U5snRNP(116 kDa)
ruvb-like1
smc2
rpl7
trnT(AGT)
orf217
orf171
ebi
rps3
rpl1
tfIIB
orf89
brx1
rpl35A
fkbp-like
orf1034
orf63*
rps30*
orf450
orf430
orf264
orf216
orf219
snrpF
dbp4
trnM(CAT)
trnR(ACG)
orf114
orf165
tfIIB-brf
orf206
18S rRNA
5.8S rRNA
28S rRNA
5S rRNA
Telomere (GA$_n$)
Telomere (GA$_n$)
5S rRNA
28S rRNA
5.8S rRNA
18S rRNA
orf206
tfIIB-brf
trnR(ACG)
trnM(CAT)
orf165
orf114
dbp4
snrpF
orf219
orf216
orf264
orf430
orf153
trnF(GAA)
uceE2
rpl8
orf105
orf254
gblp1
orf94
rpl31
rpl11B
orf270
orf854
mcm8
orf59
mcm9
orf311
orf163
cpeT-like
imp4
orf659
tbl3
rpl18A*
rps11
orf124
orf365
orf57
orf252*
orf136
orf684
orf205
orf415
orf124
mcm5
H4
H3
rpb2
orf861
rpl23
orf215
orf235
orf74
hfc136
trnL(TAA)*
ATPbp
rps24*
orf323
rpb8
ruvb-like2
eif5A
sf3b3-like
der1
orf393
orf96
kea1
orf59
rpl26*
rps10B
hsp70
nip7
rpc1
orf217
orf171
trnQ(TTG)
fcf1
mcm7
orf363
dph1
asf
orf745
orf525
impA
erf1
trnD(GTC)
tcpE
orf75
orf266
tfIIA-S
orf636
orf417
tha4
rpabc6
rpabc5
met
smc3
sufD
tubG
orf109
orf307
orf92
nog1
orf131
orf177
rpl9*
trnV(AAC)
orf243
trnS(GCT)
taf90
orf394
orf147
orf217
tcpA
orf136
U5snRNP (40kDa)
rps15A
snrpD2
tfIIE
orf223
orf199
orf501
engA
sf3b1-like
orf367
secE
rip1
trnR(TCG)
rpl24*
has1
nop5
cpeT-like
orf649
orf213
orf178
kin(snf1)
pi4K
orf288
rps15
smc1
orf332
orf826
orf447
orf120
orf157
orf720
orf111
orf152
dnaG
orf437
rpl36
rps17*
rpl21
rps19
orf313
iap100
imb1
taf
orf123
rsp4
orf255
rpb3
orf296
orf443
orf177
rps8
snrpD
ggt
smc4
hda
orf238*
trnL(TAG)
hlip
orf88
trnV(TAC)
rpl6B*
orf977
orf219
snrpF
dbp4
trnM(CAT)
trnR(ACG)
orf114
orf165
tfIIB-brf
orf206
18S rRNA
5.8S rRNA
28S rRNA
5S rRNA
Telomere (GA$_n$)
Telomere (GA$_n$)
5S rRNA
28S rRNA
5.8S rRNA
18S rRNA
orf206
tfIIB-brf
trnM(CAT)
trnR(ACG)
orf165
orf114
dbp4
snrpF
orf219
orf216
orf264
orf430
orf450
rps30*
pcna
rad25
orf240
rad51
orf515
orf369
orf268
eif1A
orf286
orf283
orf157
orf214
pop2
gblp2
orf79*
rpl14*
orf59
orf152
rpl24
trnL(CAA)
nol10
pab2
trf
rpl23A
orf492
orf591
tcpG
tif211
ftsZ
rps4
snrpD2
mcm3
tfIID
orf235
GTP-bp
rpb1
orf132
rps25*
orf532
orf238
eif2B
gyrB
orf191
orf817
trnY(GTA)*
orf308
orf282
trnA(AGC)
orf488
rps2
snrpD1
trnT(TGT)
cdc5-like
orf100
rpl5
orf148
mce
trnR(CCT)*
rpl32
U3snoRNP
orf179
U6snRNA
tubB
nop56
trnL(AAG)
tcpB
trnK(TTT)
rps14*
rps20
orf425
mak16
rps16*
rpl3
orf258
rps9*
orf242
mcm2
orf643
rpl10A
orf334
rpl17
nop2
sbp1
rpl37A
trnW(CCA)*
orf308
G10
cdc48b
orf457
orf441
orf266
orf65*
orf373
sen1
orf184
orf249
BRSK
orf142
rfc3
orf430
rps21
rpl40
orf416
orf336
orf201
prl1-like
orf315
orf140
orf336
sof1
orf72
cdc2
kin(ABC1)
rpa5
ubc2
orf314
fkbp
orf294
rpa2
orf806
orf634
tfIIA-S
orf266
rpoD
orf199
orf446
gsp2
rpl30
rpc10
rpf1
trnS(AGA)*
rps27A
reb1
rad3
trnA(TGC)
trnM(CAT)*
rpb7
rpc2
orf204
tcpH
rpl13A
orf495
dbx-like
pab1
trnR(TCT)*
orf98
orf227
h2B
orf321
nop-like
ef2
orf150
orf486
rpb10
orf2098
nop1
cbf5
rla1
rli1
gidB
orf726
orf212
orf216
orf219
snrpF
dbp4
trnM(CAT)
trnR(ACG)
orf114
orf165
tfIIB-brf
orf206
18S rRNA
5.8S rRNA
28S rRNA
5S rRNA
Telomere (GA$_n$)
5 Kb
🟥 Translation
🟪 Transcription
🟦 Mitosis
RNA metabolism
Plastid-associated
🟫 Miscellaneous
Multi-copy
🟨 Protein folding and degradation
🟩 DNA metabolism and cell cycle control
🟧 Cryptophyte nORF
ORF with motif
Syntenic ORFan
■ ORFan
* Intron-containing
— tRNAs and structural RNAs

The *C. mesostigmatica* nucleomorph genome harbours 580 genes: 505 protein-coding genes (453 unique genes), 50 tRNA genes (all tRNAs present), and 25 other non-messenger RNA genes (specifying ribosomal RNAs and a U6 snRNA) (Table 2.1). At 703 Kbp, the genome is more than 100 Kbp larger than any of the other cryptophyte nucleomorph genomes sequenced to date, yet it contains a similar number of genes. There are only 61 more genes in *C. mesostigmatica* than in *C. paramecium* (whose genome is 217 Kbp smaller), 32 more genes than in *G. theta* (which is 152.4 Kbp smaller), and 55 more genes than *H. andersenii* (which is 131.5 Kbp smaller) (see section 2.3.2 below for further description of gene content differences). The gene density of the *C. mesostigmatica* nucleomorph genome is 0.83 genes/Kbp, which is notably lower than *H. andersenii*, *G. theta*, or *C. paramecium* at 1.09, 0.98, and 1.07 genes/Kbp, respectively. Overall, however, while the number of genes in the *C. mesostigmatica* nucleomorph genome is higher than in the other sequenced nucleomorph genomes, gene number is not strictly correlated with nucleomorph genome size. The larger size of the *C. mesostigmatica* nucleomorph genome cannot be attributed solely to the presence of more genes.

Previous work (Lane et al. 2007; Tanifuji et al. 2011) observed size differences for homologous nucleomorph-encoded proteins, which could account for some of the genome size variation observed. I tested this hypothesis with a four-way comparison. A trend towards a decrease in gene/protein size with decreasing nucleomorph genome size was observed, but this trend is not strict and not statistically significant, whether I compare the average size of all the proteins encoded in each of the cryptophyte nucleomorph genomes (p > 0.05) or a 227-protein subset that is shared among all four cryptophyte nucleomorphs (p > 0.05) (Table 2.1). If the sizes of these 227 homologous proteins are examined individually, I do see a net gain in amino acids as compared to the total number of amino acids for these proteins from the smallest nucleomorph genome to the largest (74,684, 75,098, 79,289, and 80,142 for *C. paramecium*, *G. theta*, *H. andersenii*, and *C. mesostigmatica*, respectively). An increase of 853 amino acids for the 227 proteins in *C. mesostigmatica* compared to *H. andersenii* does increase the genome size, but there is a 130 Kbp size difference between these two genomes, so the increase due to protein size alone is minimal.

Interestingly, a significant difference in protein size is observed (p < 0.01) when the average sizes of the ORFan genes are compared across all four genomes (Table 2.1). It was previously observed that the smallest sequenced cryptophyte nucleomorph genome, that of *C. paramecium*, encodes ORFan proteins that are on average much smaller in size compared to those in the other nucleomorph genomes, and so it was hypothesized that nucleomorph genome size diversity may be largely influenced by size variation in these ORFan genes (Tanifuji et al. 2011). However, the *C. mesostigmatica* nucleomorph genome does not encode larger ORFan proteins on average; ORFan gene size is thus not a contributing factor to increased nucleomorph genome size in *C. mesostigmatica*, and cannot account for the nucleomorph genome size variation observed within the cryptophytes.

Table 2.1 Comparison of cryptophyte nucleomorph genome features.

Genome Feature	*Chroomonas mesostigmatica*	*Hemiselmis andersenii*	*Guillardia theta*	*Cryptomonas paramecium*
Genome size (Kbp)[a]	702.9	571.4	550.5	485.9
G+C content (%)	25.94	25.18	26.43	26.05
Number of genes[b]:				
Protein-coding	505	472	487	466
Total	580	525	548	519
Gene density (genes/Kbp)	0.83	1.09	0.977	1.07
rRNAs	24	15	24	18
tRNAs	50	38	37	34
Number of overlapping genes	20	44	11	33
Average protein length (aa):				
All proteins	357	338	312	289
227 shared proteins	353	349	329	331
ORFans	264	190	268	190
Average intergenic spacer (bp):				
Syntenic spacers	91	77	41	62
All spacers	200	132	93	102
Number of ORFan genes (% of protein-coding genes)	94 (19)	74 (16)	155 (32)	133 (29)
Number of spliceosomal introns	24	0	17	2

[a] Telomere sequences are not included in the total genome size.

[b] Includes current data from GenBank, the gene analysis of Tanifuji et al. (2011) and the previously unannotated *gidB* in *G. theta*.

It is also not the case that the *C. mesostigmatica* nucleomorph genome is significantly enriched in longer genes. If the distribution of ORFs according to their size is compared across the four cryptophyte nucleomorph genomes, a striking trend is seen when the shortest ORFs (<150 aa) are considered, i.e., the percentage of ORFs in this size range is negatively correlated with genome size such that the smallest genome contains the highest percentage of small ORFs, and as genome size increases, the percentage of ORFs of that size decreases (Figure 2.7).

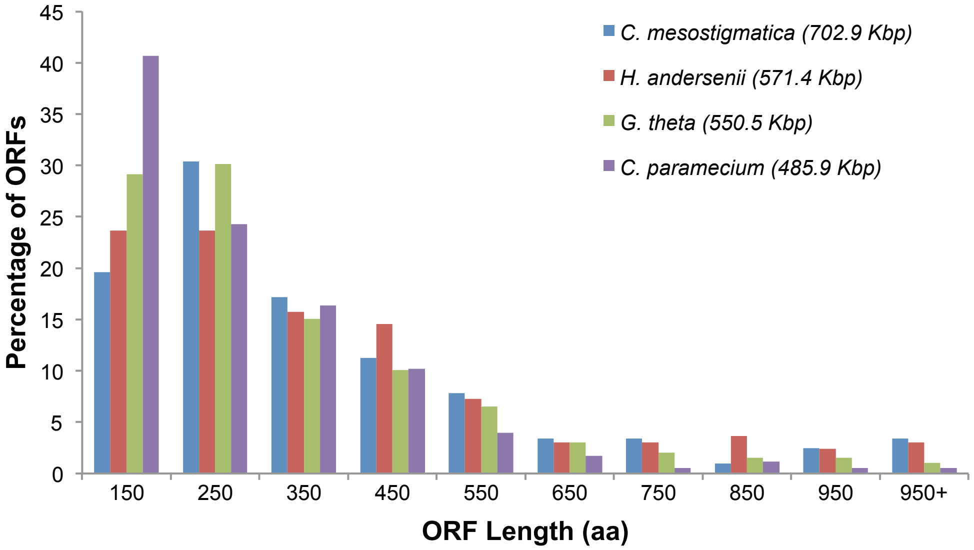


Figure 2.7 Percentage of all cryptophyte nucleomorph ORFs per genome as a function of length. Each of the four nucleomorph genomes examined in this study has a different distribution of ORF sizes. The smaller nucleomorph genomes are enriched in shorter ORFs, and as the size of the ORF increases, the percentage of those ORFs decreases. Larger nucleomorph genomes are slightly enriched in longer ORFs.

As ORF size increases, the trend is not perfectly linear. Nevertheless, I do observe that for ORFs longer than 550 amino acids, there is a slight trend towards the larger genomes containing more of these longer genes than the smaller genomes. This trend is, however, not enough to account for most of the genome size variation observed.

The single largest contributing factor to the larger nucleomorph genome size in *C. mesostigmatica* is the amount of non-coding sequence. The average intergenic spacer size is significantly larger ($p < 0.01$) in the *C. mesostigmatica* nucleomorph genome compared to the nucleomorph genomes of *H. andersenii*, *C. paramecium*, and *G. theta* when the average size of all the intergenic spacers are compared (200 bp for *C. mesostigmatica*, 132 bp for *H. andersenii*, 93 bp for *G. theta*, and 102 bp for *C. paramecium*), and when the average size of the intergenic spacers within syntenic regions are compared (91 bp for *C. mesostigmatica*, 77 bp for *H. andersenii*, 41 bp for *G. theta*, and 62 bp for *C. paramecium*) (Table 2.1). Interestingly, differences were observed in intergenic spacer sizes depending on the relative orientation of the bounding genes: head-to-head, head-to-tail, or tail-to-tail (Figure 2.8).

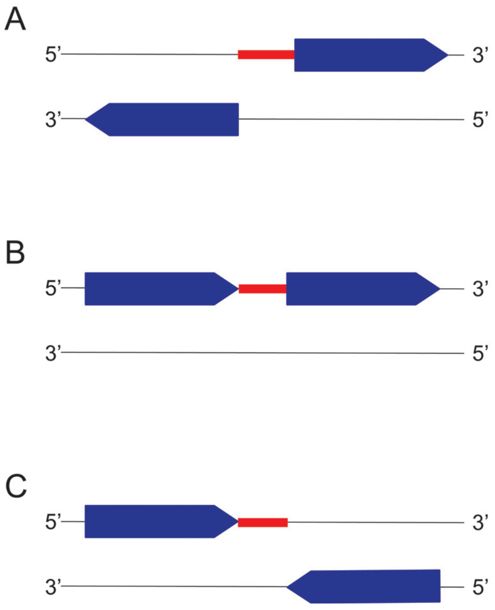


Figure 2.8 Intergenic spacers relative to gene orientation. Spacers, shown in red, were categorized by bounding genes, shown in blue, according to whether they are oriented A) head-to-head, B) head-to-tail (or tail-to-head), or C) tail-to-tail.

For all four species examined, the intergenic spacers are the smallest when the bounding genes are oriented tail-to-tail (Table 2.2). The size difference is statistically significant when compared to intergenic spacers bounded by genes that are oriented head-to-head, or head-to-tail, for *H. andersenii*, *G. theta*, and *C. paramecium*. In fact, the highest prevalence of overlapping genes, or genes that have no spacer between them, occurs when the genes are oriented tail-to-tail (Table 2.2).

Table 2.2 Comparison of average intergenic spacer size for different gene orientations.

Gene Orientation	Average Intergenic Spacer Size (bp)			
	Chroomonas mesostigmatica	*Hemiselmis andersenii*	*Guillardia theta*	*Cryptomonas paramecium*
Head-Head	203.4	130.4	103.0	119.8
Head-Tail	217.9	152.2	106.3	115.5
Tail-Tail	166.5	95.7	64.3	63.7
One-way ANOVA p-level	p = 0.07216	p = 0.00562	p = 0.04407	p = 0.00006
Number of 0 bp spacers/Total				
Head-Head	8/151 (5.3%)	2/135 (1.5%)	7/137 (5.1%)	10/127 (7.9%)
Head-Tail	6/271 (2.2%)	4/249 (1.6%)	20/273 (7.3%)	7/259 (2.8%)
Tail-Tail	10/154 (6.5%)	8/138 (5.8%)	31/140 (22.1%)	18/130 (13.8%)

A previous analysis of 32 nucleomorph-derived ESTs for *G. theta* showed that 31 out of 32 transcripts terminated within downstream genes (Williams et al. 2005). Thus, it appears as though many terminator sequences have moved within downstream genes, allowing for reduction of the intergenic space at the 3' end of genes. This could explain the shorter intergenic spacers found between genes oriented tail-to-tail. Conversely, in the same EST analysis, transcripts were rarely found to initiate within upstream genes, suggesting that the intergenic regions at the 5' end of genes still retain transcriptional regulatory elements, which could explain why I observe larger intergenic spacers between genes oriented head-to-head and head-to-tail.

In sum, the observed diversity in cryptophyte nucleomorph genome size can be attributed primarily to differences in the amount of non-coding DNA, as well as the number of genes (in particular the presence/absence of multi-copy genes), together with minor variation in the length of homologous genes.

2.3.2 Proteins of Known Function

Of the 505 putative protein genes in the *C. mesostigmatica* nucleomorph genome, 235 encode proteins predicted to have core eukaryotic 'housekeeping' functions. These include transcription, translation, DNA metabolism and cell cycle control, RNA metabolism, protein folding, protein degradation, and mitosis (Appendix B). The *C. mesostigmatica* nucleomorph genome contains the identical set of 31 plastid-associated genes found in the nucleomorph genomes of both *H. andersenii* and *G. theta*. The gene for the plastid-targeted glucose-inhibited division protein B, *gidB*, was initially presumed missing from the plastid and nucleomorph genomes of *G. theta* (Douglas et al. 2001). However, our comparative analyses of gene order conservation led to the identification of a single copy of *gidB* in the nucleomorph genome of *G. theta*. Apart from the non-photosynthetic species *C. paramecium*, which has lost many photosynthesis-related genes, it is not clear why precisely the same set of 31 genes for plastid-associated proteins are retained in *C. mesostigmatica*, *G. theta,* and *H. andersenii* and have not been differentially transferred to the host nuclear genome. Nonetheless, their presence in the four species examined (a sub-set of which is present in the non-photosynthetic species *C. paramecium*) suggests that this particular suite of plastid-associated genes was 'locked in' prior to the radiation of the major cryptophyte lineages.

A comparison of gene content for biological functions conserved across the four species reveals a high degree of overlap (Figure 2.9). Out of the 311 genes examined, 216 (69.5%) are present in the nucleomorph genomes of all four species. In the case of genes that are missing from two or three species, there are no clear patterns to account for their loss; a punctate distribution of gene loss is observed across all of the functional categories examined, which include transcription, translation, mitosis, cell cycle control, protein folding, protein degradation, DNA metabolism, and RNA metabolism (Appendix B). The only notable exception is that the nucleomorph genome of *C. mesostigmatica* contains more genes whose protein products function in spliceosomal intron removal (discussed below). There are, however, three very distinct patterns of gene loss for those genes lost from only a single species.

As previously reported, *C. paramecium* has lost photosynthesis capability and as a result, has a reduced set of nucleomorph-encoded photosynthesis-related genes (Tanifuji et al. 2011). As expected, the set of 24 genes that are present in *C. mesostigmatica*, *H. andersenii* and *G. theta* but missing from *C. paramecium* is comprised primarily of plastid-associated genes. Similarly, the nucleomorph genome of *H. andersenii* has been shown to be completely devoid of spliceosomal introns and deficient in splicing-related genes, thus it is unsurprising that genes required for spliceosomal intron removal make up the set of seven genes missing from *H. andersenii*. There is no obvious functional explanation to account for the 11 genes that are absent from the nucleomorph genome of *G. theta*. Previous comparative studies of cryptophyte nucleomorph genes have shown that *G. theta* genes tend to be more divergent compared to their homologs in other nucleomorph genomes (Tanifuji et al. 2011). Our data support this observation, and having additional nucleomorph genome data from a close relative of *G. theta* would help in determining whether these genes are indeed missing, or are present but have diverged beyond detection by sequence similarity.

The most surprising observation from the four-way comparison is that in the set of 27 genes that are shared between *G. theta*, *H. andersenii,* and *C. paramecium,* but absent from *C. mesostigmatica*, 22 are involved in protein degradation. All 21 genes encoding subunits of the proteasome are missing from the nucleomorph genome of *C. mesostigmatica*, as well as the E2 ubiquitin conjugating enzyme gene *ubc4* (Appendix B). The significance of this observation is unclear. I examined RNA-Seq data from *C. mesostigmatica* for nuclear genes that encode proteasome subunits that could potentially be targeted into the PPC, i.e., the residual cytoplasm of the endosymbiont in which proteasome-mediated protein degradation would presumably take place. However, only a single, apparently host-derived copy of each proteasome subunit gene was found, each of which encodes a protein with no obvious amino-terminal extensions reminiscent of the bipartite leader sequences required for such targeting (Gould et al. 2006[a,b]). It is thus unclear whether canonical protein degradation pathways exist within the PPC of *C. mesostigmatica* and if so, which proteins are involved. It is entirely possible that some of the mysterious 'ORFan' genes, which constitute 20% of the protein-coding genes in the *C. mesostigmatica* nucleomorph genome, and of which we know nothing about, play a role in protein degradation.

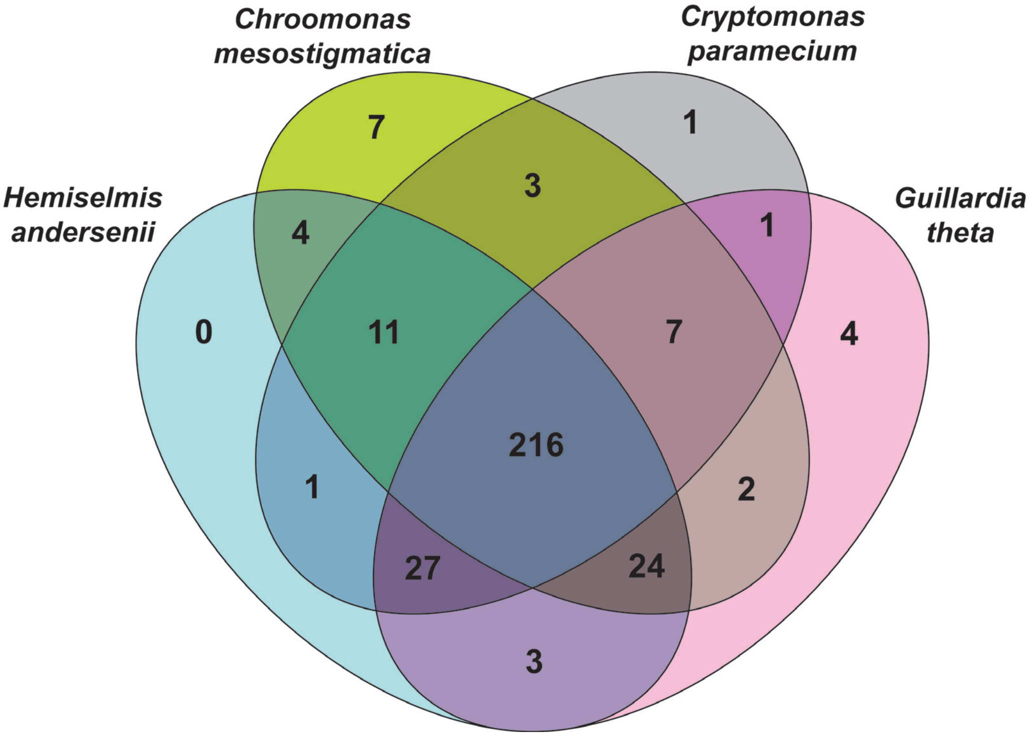


Figure 2.9 Four-way cryptophyte nucleomorph gene content comparison. There are 311 genes of known or predicted function annotated in cryptophyte nucleomorph genomes, 216 of which (~70%) are present in all four cryptophyte nucleomorph genomes presently sequenced, forming a highly conserved core gene set. Aside from the lineage-specific photosynthesis-related, spliceosome, and proteasome gene loss, the distribution of missing genes appears to be random with respect to each species and functional gene category.

Interestingly, there are some hallmark genes of the ubiquitin-proteasome degradation pathway that are present in the *C. mesostigmatica* nucleomorph genome, such as ubiquitin (2 copies), the ubiquitin-fusion degradation protein (*ufd*), and two ubiquitin-conjugating enzymes (*uceE2* and *ubc2*). However, most of these genes are known to be involved in other cellular processes besides protein degradation, such as protein import via SELMA, the pre-protein translocator located in the second outermost membrane of complex plastids in cryptophytes (*ufd* as well as *cdc28* and *der1*, which are also nucleomorph-encoded in *C. mesostigmatica*) (Bolte et al. 2011); DNA damage repair (*ubc2*) (Jentsch et al. 1987); disassembly of the mitotic spindle (*ufd*) (Cao et al. 2003); and chromatin structural maintenance, gene expression, and stress response (ubiquitin) (Gardner et al. 2005; Conaway et al. 2002; and Finley and Chau 1991, respectively). Furthermore, in the absence of a complete *C. mesostigmatica* nuclear genome sequence, it is presently not possible to conclude with certainty that the host nuclear genome does not possess at least some of the missing proteasome genes. However, it is interesting that the nucleomorph genome of the chlorarachniophyte *B. natans* is also devoid of obvious proteasome subunit genes (Gilson et al. 2006). Analysis of the recently released nuclear genome sequence of *B. natans* (Curtis et al. 2012) revealed a complete lack of nuclear-encoded, proteasome-mediated degradation proteins targeted to the PPC, suggesting that protein degradation in the PPC of *B. natans* is not performed using the canonical proteasome-degradation pathway. While the nuclear genome project of *B. natans* has given tremendous insight into the PPC “parts list”, further study of the PPC proteome is needed to elucidate the mechanism of protein degradation in *B. natans*, insight that could prove useful in characterizing the protein degradation system for nucleomorph-encoded proteins in *C. mesostigmatica*.

2.3.3 Proteins of Unknown Function

A substantial proportion of the protein-coding genes in the *C. mesostigmatica* nucleomorph genome (~30%) are hypothetical in nature. One-third of these genes are cryptophyte nucleomorph-specific ORFs, or 'nORFs', meaning they have clear homologs in other cryptophyte nucleomorph genomes but not in other known genomes. The remaining two-thirds, however, are true 'ORFan' genes, meaning they show no obvious sequence based homology to any gene in known databases, nucleomorph-derived or otherwise. Some of these genes are present in syntenic blocks, i.e., they occupy the same position within a block of genes conserved between different nucleomorph genomes. These 'syntenic ORFans', as first described by Lane et al. (2007), not only exhibit positional conservation, but often their size is also conserved. Despite showing no detectable sequence similarity, based on their size and positional information they are presumed to be homologous.

Interestingly, many of the *C. mesostigmatica* nORFs show significant sequence similarity to those of *H. andersenii*, but are noticeably less similar to those of *C. paramecium* or *G. theta*, a pattern that is also seen for genes of known function. This gives further support to phylogenies inferred from host and nucleomorph 18S rDNA, which suggest that members of the genus *Chroomonas* and the genus *Hemiselmis* are more closely related to each other than they are to other known cryptophytes (e.g., Hoef-Emden 2008). Upon close inspection, many of the hypothetical protein-coding genes in the *C. mesostigmatica* and *H. andersenii* nucleomorph genomes do in fact show sequence similarity to each other. The high degree of sequence similarity and the more conservative nature of the *C. mesostigmatica* ORFs allowed me to ascribe predicted functions to nine previously un-annotated hypothetical protein-coding genes based on homology and synteny. These include the plastid-associated gene *gidB*, the DNA-directed RNA polymerase II subunit gene *rpb4*, and the highly-conserved mini-chromosome maintenance genes *mcm8* and *mcm9* in *G. theta*; the mRNA splicing factor gene *prl1-like* in *H. andersenii*; the acidic ribosomal protein gene *rla1* in both *H. andersenii* and *C. paramecium*; and the nucleolar protein gene *nop-like* and spliceosomal genes *prp4-like* and U5 snRNP (40 kDa) in *C. paramecium*. In addition, the observed sequence similarity between hypothetical *C. mesostigmatica* ORFs to previously annotated ORFan genes in *H. andersenii, C. paramecium*, and *G. theta* allowed me to reclassify 58 of these ORFan genes as nORFs. I was also able to reclassify 73 ORFan genes as syntenic ORFs.

In terms of the total proportion of nORF genes relative to ORFans, the addition of *C. mesostigmatica* nucleomorph protein genes resulted in a reduction from 76% ORFan genes in a 3-way analysis (i.e., *G. theta*, *H. andersenii* and *C. paramecium*) to 55% in a 4-way comparison (Figure 2.10). This significant reduction can be mostly attributed to the close relatedness of *C. mesostigmatica* to *H. andersenii*. Despite these improvements, many questions remain surrounding the functions of the nORF and ORFan genes that can only be answered with additional nucleomorph genomic data from other closely related cryptophyte species, nuclear genomic data from red algae, and more detailed biochemical knowledge of the processes taking place within the periplastidial compartment.

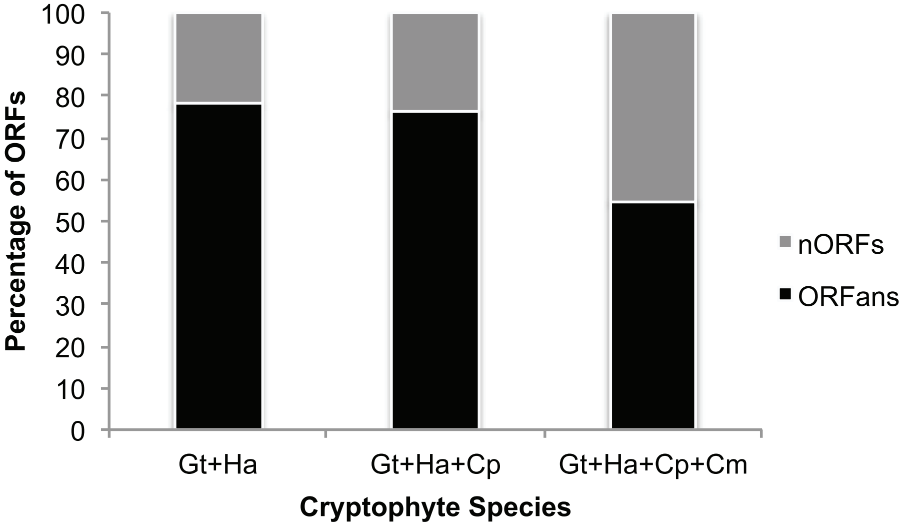


Figure 2.10 Hypothetical proteins inferred from complete cryptophyte nucleomorph genomes. The graph shows the proportion of cryptophyte nucleomorph-specific hypothetical protein-coding genes (nORFs) and hypothetical protein-coding genes unique to each individual nucleomorph genome (ORFans) relative to the total number of hypothetical protein-coding genes as additional nucleomorph genomic sequences become available. The leftmost bar compares the proportions of the two types of hypothetical proteins for *G. theta* (Gt) and *H. andersenii* (Ha). The second bar compares these proportions with the addition of *C. paramecium* (Cp). The proportions do not change substantially until the addition of the fourth genome, *C. mesostigmatica* (Cm), shown in the third bar, where the proportion of ORFan genes drops by 21%.

2.3.4 Spliceosomal Introns

Unlike the intron-rich nucleomorph genome of the chlorarachniophyte *B. natans,* there are relatively few spliceosomal introns present in cryptophyte nucleomorph genomes: 24 in *C. mesostigmatica*, 17 in *G. theta*, 2 in *C. paramecium*, and none in *H. andersenii* (Table 2.1). With only two red algal nuclear genome sequences in hand, one from the unicellular extremophile *Cyanidioschyzon merolae* (Matsuzaki et al. 2004), and the other from the seaweed *Chondrus crispus* (Collén et al. 2013), it is difficult to say whether nucleomorphs were intron-rich at one time; however, both red-algal nuclear genomes are intron poor, suggesting that the cryptophyte nucleomorph ancestral genome was also intron poor.

The number of genes related to spliceosomal intron removal varies across the four cryptophyte nucleomorph genomes. Most notably, the nucleomorph genome of *H. andersenii* is completely devoid of spliceosomal introns and genes dedicated to intron removal (Lane et al. 2007). Nevertheless, the *H. andersenii* genome retains divergent homologs of a few genes (*prl1-like*, *snu13*, *cdc28*, *snrpD*, and *snrpD2*) whose protein products function in mRNA splicing in other organisms. The nucleomorph genomes of *C. paramecium*, which contains only two spliceosomal introns (62 bp and 100 bp in length), and *G. theta*, which contains 17 spliceosomal introns (42-52 bp in length), possess 17 and 15 'spliceosomal' genes, respectively (Appendix B). In comparison, the *C. mesostigmatica* nucleomorph genome is predicted to possess 24 spliceosomal introns (Figure 2.6; Table 2.3), seven more than *G. theta*, and possesses 28 splicing-related protein-coding genes (Appendix B), almost double the number found in *G. theta*. The lengths of the *C. mesostigmatica* nucleomorph spliceosomal introns are longer on average, ranging from 50-211 bp in size (Table 2.3).

Together with honours student Tyler Mills, I used RT-PCR, cloning and cDNA sequencing to confirm the splicing of four nucleomorph introns in *C. mesostigmatica* and to verify their predicted intron-exon boundaries: *rps23*, *rps24*, *rpl9*, and *rps16* were shown to have introns of 114 bp, 114bp, 77 bp, and 57 bp, respectively (Figure 2.5). In addition, I used RNA-Seq data to verify the correct boundaries and removal of spliceosomal introns from *rpl18A*, *rpl19*, *rpl24*, *rpl26*, *rpl27A*, *rps9*, *rps13*, *rps25*, *rps28*, *orf65*, and *orf102*. Interestingly, I observed a case of miss-splicing of a 57 bp intron in the ribosomal protein gene *rpl26* via an alternate 3' intron boundary. Use of the alternate 3' splice site results in a substantially truncated, and presumably non-functional, protein of only 53 amino acid residues (the full length protein is predicted to be 124 amino acid residues). A second RNA-Seq-derived contig shows the correct removal of the intron using the predicted GT/AG intron boundaries producing a full-length *rpl26* mRNA. Like the spliceosomal introns of *G. theta* (Douglas et al. 2001), most of the introns are present at the extreme 5' end of the genes, several immediately following the start codon, and contain a 5'- GTAAGT consensus motif.

Table 2.3 Spliceosomal intron-containing genes and intron sequences in the nucleomorph genome of *C. mesostigmatica.*

Gene	Intron	Length (nt)
rpl6B	5'-GTAAGTACTAAAAAGTTATTGATTAAAAAAAAAATTTTTTGAAAAATATAATTAATTTATATTTTTTTAAAG-3'	72
rpl9^{+}	5'-GTAAGTTTCGAAATAAAATTTTTTCTTTATCAGATTTTTTTAATTTTTTTACCAGTTAAAATAAAAATAAACGCTAG-3'	77
rpl14	5'-GTAAGAATCTCATTTGAAAAATAATGCTTTAAAAAGTAACTAGTTAATCAATCATTAAAAAATTTAAG-3'	68
rpl18A$*^{+}$	5'-GTAAGTATCCAATCAAAAGAAAAGTTTAATGTTAAAAATTTTAATTTTTTATATTTTTAG-3'	60
rpl19$*^{+}$	5'-GTAAGTATATTTAATTTTTAAAAAATTTAAAAAAATATATTTTAATAATTTTAATTATTTTTTTTAG-3'	67
rpl24^{+}	5'-GTAAGTATAGAAATGTGAATTTATTCGATAATTTAAACCTACAACAAATTTAG-3'	53
rpl26^{+}	5'-GTAAGTATACAAAAAAATTTTATTCATTATTCAAAAATTTATCTTATTTATTTTTAG-3'	57
rpl27A$*^{+}$	5'-GTAAGAACTTAAAAACTTTTATAATTGCGTTTCGTTTTTTTTCACCAAGGAAAGAATACTTATCTGAAAAAAG-3'	73

Gene	Intron	Length (nt)
rps3A*	5’-GTAAGTTGTGTGACTTATATAAAATTTTTTGGATTTCTTTTCTTATTTTCCTAAAAAAAG-3’	60
rps9*$^{+}$	5’-GTAAGAAAAAAATATTTTTTTTAAAATTAAATCAATTTATATTTTTTTAGATGAACTATAGCTCATGAATATTAAAAG-3’	78
rps13*$^{+}$	5’-GTTAGTTTTCTGAACCATTTATTTATTTTTTTTAAAAAATTTAGATTTTTGACCTTTAAATATATTATAAAATAG-3’	75
rps14	5’-GTAAGAATTTTGACTAAATAAAACGTTTTTGTAAAAAATATAATTTATTTGAAATAGCCCGAG-3’	63
rps16*$^{+}$	5’-GTAAGTATCAAAAAAATTAAGTATCAAAAAAATTCAATAACAACTAATTTTTAAAAG-3’	57
rps17*	5’-GTAAGAATTTAAACTAATGGAAAATAAATTTGTTAATGTAAAATTTTTTTTTTTTGAAAAAATACTAACAAAAAAG-3’	77
rps23^{+}	5’-GTAAGTTTAAAAAAAGACATTTTTTCTAATGGTCGCTAAGAAAGAAATTATTTTTTTTTTAGGCATGGATCATTTTTTTTTAAAAAATTTAGAAAAACTCTAACAAATTAAAAG-3’	114
rps24^{+}	5’-GTAAGTTTTTATATTTATTTTAAAAAAAACTTTGATTGAAATTTAATTTTCAGTTTTCAAAGAATTATTTTTTTTTGTAGTATGAACTTTTTGATTACAAAATTATAAAAAAAG-3’	114
rps25^{+}	5’-GTAAGTTTGAAAAAATTTCTTTAACTTTTAATTTATTTTAAATTTATTTATAATTATTTTTGATATAG-3’	68
rps28*$^{+}$	5’-GTAAGTAAGAAAAAAATTTTTGATTACCTTTATCAAAGATTGTGGATAATAATTAAACTTTAG-3’	64
rps30	5’-GTAAGAAATTTTATATAATTATTAAAAATAAATTTCTTTTTTGTTTTATTTTTATTTTCTCAAAAAAAAAAAAATAAAATAAAAATAAAAAAAAAAAGACATTGGCTAACACGCTTTTTTTTTTTTTTTTTATATTTTTATATTTTTATATTTTTTCCAAAAAAAAATAAAATAAAATAAAAAAAAAAAAAAAAAAAAATTTTTACCCTTCAG-3’	211
orf63	5’-GTAAGTAGAATAGCAGAATATTGTTCTTTTTTATGTAATCACAGTCGATTTTTTAATCTTAAAACCAAAAATATAG-3’	76
orf65^{+}	5’-GTAAGATAAATTTTCTTTTAAAATTTTTTTTTAATTTTTTAATTTTTAAG-3’	50
orf76	5’-GTAAAAAAATATTTTTCATTTTAAATCCACATGTAAATCTTTATTTTTGGTTCAAATTAAG-3’	61
orf102^{+}	5’-GTAAGTATACATTTAAAAATTTAAATTAAGATTTATTCATCTGATTTTTTTTTCAG-3’	56
orf252	5’-GTAAATAATAACATAATTAAATTTATGTAAATATTTTTATTTAAATAAAATTGAAG-3’	56

* Intron present in gene in *G. theta.*

+ Intron removal verified experimentally by RT-PCR and/or RNA-Seq data.

Examination of the distribution of nucleomorph spliceosomal introns across the tree of cryptophytes reveals a very clear pattern. The cryptophytes branch into five major clades: (1) *Chroomonas, Hemiselmis,* and *Komma*, with the *Hemiselmis* species branching from within the *Chroomonas* clade; (2) *Guillardia* and *Hanusia*; (3) *Cryptomonas*; (4) *Geminigera*, *Plagioselmis*, and *Teleaulax*; and (5) *Rhinomonas*, *Rhodomonas*, and

Storeatula (Hoef-Emden 2008; Figure 2.11). Complete nucleomorph genome sequences are now available for members of clades 1, 2, and 3, and so spliceosomal introns can be inferred to be present in all three clades. There are no nucleomorph genome sequences available for members of clade 4, which is a poorly understood group. Most sequences in current databases for members of this group represent rRNA gene sequences amplified from environmental samples, and few members are available in culture.

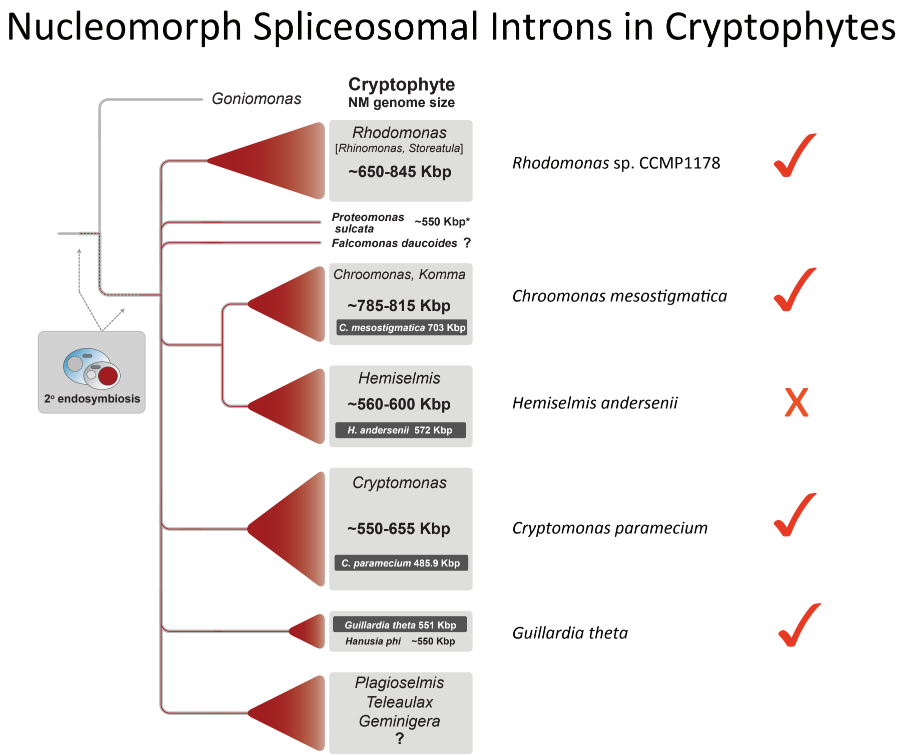


Figure 2.11 Distribution of nucleomorph spliceosomal introns across the tree of cryptophytes. A schematic phylogenetic tree shows the diversity of cryptophytes and nucleomorph genome sizes. The presence of nucleomorph spliceosomal introns is indicated by a check mark beside the name of the member of the clade in which they were found. Species with complete nucleomorph genomes are highlighted. The dimensions of the red triangles indicate the relative number of described taxa per lineage and its depth. The dashed arrows and line indicate that it is presently unknown whether the cryptophyte red algal plastid was gained before or after the divergence of *Goniomonas*. Figure adapted from Moore and Archibald (2009).

Partial nucleomorph genomic data generated by the Archibald Lab for a member of clade 5, *Rhodomonas* sp. CCMP1178, show that spliceosomal introns and the machinery required for their splicing are indeed present. These include a 51 bp spliceosomal intron present in a gene for one of the proteasome subunits, *prsA7* (JX515791), as well as a 76 bp spliceosomal intron present in the regulator of epidermal growth factor gene, *ebi* (JX515790). The 5' splice sites for these two introns have the 5'- GTAAGT consensus motif observed in the spliceosomal introns in the *G. theta, C. paramecium,* and *C. mesostigmatica* nucleomorph genomes, and additionally, these two genes in the *G. theta* nucleomorph genome also contain spliceosomal introns. Also present in the *Rhodomonas* sp. CCMP1178 nucleomorph genomic data is the large and highly conserved spliceosomal protein gene *prp8* (JX515789)*,* whose protein product performs a key role in the catalytic core of the spliceosome (Grainger and Beggs 2005) and is present in all spliceosomal intron-containing nucleomorph genomes sequenced to date (Douglas et al. 2001; Gilson et al. 2006; Lane et al. 2007; Tanifuji et al. 2011).

Given that spliceosomal introns and the machinery required for their removal are present in the nucleomorphs of all clades examined thus far (Figure 2.11), coupled with the fact that spliceosomal introns are absent in the nucleomorph genome of *H. andersenii* yet present in the nucleomorph genome of its close relative, *C. mesostigmatica*, it seems likely that the complete loss of spliceosomal introns occurred somewhere within the *Hemiselmis* clade.

2.3.5 Synteny and Recombination

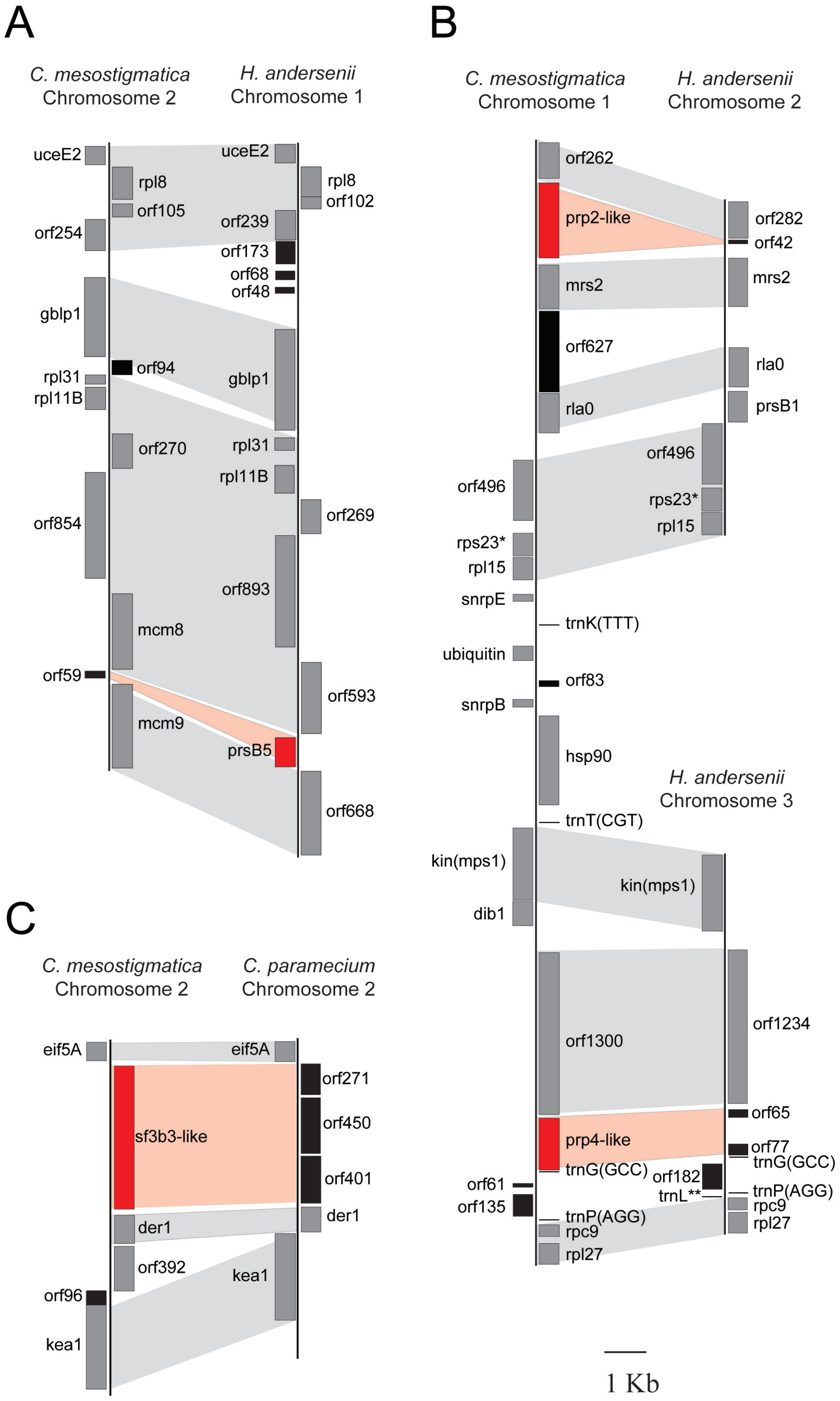
The analyses described in sections 2.3.2 and 2.3.3 above have shown that gene order conservation, or synteny, can be a helpful feature for annotating nucleomorph genomes. For example, the existence of an ORFan or nORF in one genome in the same position as an evolutionarily conserved ORF in another genome allows us to predict the identity of the ORFan/nORF. As in other reduced eukaryotic genomes such as those of microsporidian parasites (e.g., Slamovits et al. 2004), the highly compact nature of nucleomorph genomes has been suggested to represent a barrier to the frequent recombination seen in 'typical' nuclear genomes; intergenic regions are very short and most genes are single copy, making recombination-mediated disruption of an ORF probable and likely to be deleterious (Archibald and Lane 2009). I have found that while the degree of synteny is certainly highest between the *C. mesostigmatica* nucleomorph genome and that of its close relative, *H. andersenii*, the length of the syntenic blocks are noticeably shorter compared to those shared between *H. andersenii*, *C. paramecium*, and *G. theta*. The average number of genes within a syntenic region, defined as a stretch of four or more homologous genes (not including nORFs), between *C. mesostigmatica* and *H. andersenii, C. paramecium,* and *G. theta* is 9.0 (n=36), 7.1 (n=34), and 6.7 (n=29), respectively. In comparison, the average number of genes in syntenic regions between *H. andersenii* and *C. paramecium* is 19.4 (n=18), 9.4 (n=34) between *H. andersenii* and *G. theta*, and 10.1 (n=27) between *C. paramecium* and *G. theta*.

The more highly 'scrambled' nature of the *C. mesostigmatica* nucleomorph genome is presumably due to the fact that it contains longer intergenic regions, making viable intra-

and inter-chromosomal recombination events more likely. Interestingly, many apparent disruptions of syntenic blocks in the *C. mesostigmatica* genome occur where proteasome subunit genes are presumed to have been present, based on their conserved gene order in the other nucleomorph genomes. This observation suggests a recombination-mediated mechanism for gene loss, at least in the case of the missing proteasome subunit genes.

There is one instance where a small ‘mystery’ ORF of 59 aa is present in *C. mesostigmatica* within a syntenic region whose counterpart in the other three nucleomorph genomes are ‘large’ ORFs encoding the prsB5 proteasome subunit: 219 aa in *H. andersenii* (Figure 2.12A), 234 aa in *C. paramecium*, and 205 aa *in G. theta*. Even more prevalent are instances of a single large ORF in the *C. mesostigmatica* nucleomorph genome showing ‘positional synteny’ with one or more small ORFs in *H. andersenii, C. paramecium,* and *G. theta*. For example, there is one ORF of 42 amino acids in length and two small ORFs of 77 and 65 amino acids in length in the *H. andersenii* genome that occupy the same positions as the putative spliceosomal genes *prp2-like* and *prp4-like*, respectively, in the *C. mesostigmatica* genome (Figure 2.12B). Similarly, there are small ORFs of 55 aa, 61 aa, 62 aa, and 57 aa in *C. paramecium* in syntenic position with the conserved plastid-associated gene *orf268*, hypothetical protein-coding gene *orf425*, the histone chaperone gene *hira*, and the thylakoid assembly protein gene *tha4* (there is also a 75 aa ORF occupying this position in the *G. theta* nucleomorph genome), respectively, in *C. mesostigmatica* (data not shown).

Figure 2.12 ORF degradation in cryptophyte nucleomorph genomes. Schematic shows degenerating ORFs in syntenic regions between *C. mesostigmatica*, *H. andersenii* and *C. paramecium*. Homologous genes are shown in gray, with gray highlights indicating the syntenic positions of the genes on the chromosome of each species. Genes shown in black are ORFan genes. Genes shown in red are those where one or more ORFan genes occupy the same syntenic position in a stretch of genes that have conserved order in another nucleomorph genome, which are highlighted in red. A) An ORFan gene of 59 amino acids in *C. mesostigmatica* occupies the same syntenic position as the proteasome subunit gene *prsB5* in *H. andersenii*. B) ORFan genes occupy the same syntenic positions in *H. andersenii* as the spliceosomal genes *prp2-like* and *prp4-like* in *C. mesostigmatica*. C) Three ORFan genes on chromosome two in *C. paramecium* occupy the same syntenic position and sum to be a similar size as the splicing factor gene *sf3b3-like* in *C. mesostigmatica*.


A
C. mesostigmatica
Chromosome 2
H. andersenii
Chromosome 1
uceE2
rpl8
orf105
orf254
uceE2
rpl8
orf102
orf239
orf173
orf68
orf48
gblp1
orf94
rpl31
rpl11B
gblp1
rpl31
rpl11B
orf270
orf854
orf269
orf893
mcm8
orf59
orf593
mcm9
prsB5
orf668
B
C. mesostigmatica
Chromosome 1
H. andersenii
Chromosome 2
orf262
prp2-like
orf282
orf42
mrs2
mrs2
orf627
rla0
rla0
prsB1
orf496
orf496
rps23*
rpl15
rps23*
rpl15
snrpE
trnK(TTT)
ubiquitin
orf83
snrpB
hsp90
H. andersenii
Chromosome 3
trnT(CGT)
kin(mps1)
kin(mps1)
dib1
orf1300
orf1234
orf65
prp4-like
orf77
trnG(GCC)
trnG(GCC)
orf182
orf61
trnL**
trnP(AGG)
orf135
trnP(AGG)
rpc9
rpc9
rpl27
rpl27
1 Kb
C
C. mesostigmatica
Chromosome 2
C. paramecium
Chromosome 2
eif5A
eif5A
orf271
sf3b3-like
orf450
orf401
der1
der1
orf392
kea1
orf96
kea1

There are also a few instances where the corresponding syntenic position in one of the other cryptophyte nucleomorph genomes to that of an ORF in *C. mesostigmatica* is occupied by several ORFs whose sizes sum to be similar in length to the single syntenic ORF in *C. mesostigmatica*. For example, there are three ORFan genes that total 1122 aa (orf450, orf271, and orf401) in *C. paramecium* that occupy the same syntenic position of the 1156aa mRNA splicing factor *sf3b3-like* in *C. mesostigmatica* (Figure 2.12C). Similarly, there are some smaller ORFs occupying the same syntenic position in *C. paramecium* as the splicing factor gene *sf3b1-like* in *C. mesostigmatica* (data not shown). The *C. paramecium* nucleomorph genome only contains two spliceosomal introns and it appears as though the splicing factor genes are deteriorating: the small syntenic hypothetical protein-coding genes are presumably remnants of genes that are no longer functional, as they have been reduced through mutation and purging of intergenic sequence. Whether or not these genes have been transferred to the host nucleus or simply lost can only be determined with complete nuclear genomes for these organisms.

2.4 CHAPTER SUMMARY

In conclusion, the complete nucleomorph genome sequence of *C. mesostigmatica* has provided valuable insight into the factors contributing to cryptophyte nucleomorph size diversity, genome biology, and their evolutionary fate. I have identified several factors that contribute to the size variation in the nucleomorph genomes observed, including slight differences in the lengths of protein-coding genes and the total number of genes, and most notably, differences in the lengths of the intergenic spacers. In contrast to the other cryptophyte nucleomorph genomes, the nucleomorph genome of *C. mesostigmatica* contains numerous (and larger) spliceosomal introns, more multi-copy genes, a lower degree of synteny, and repetitive regions. These features make the *C. mesostigmatica* nucleomorph genome the most 'complex' nucleomorph genome studied to date, exhibiting features more characteristic of its presumed free-living red algal relatives. The presence of spliceosomal introns in the *C. mesostigmatica* nucleomorph genome, yet their absence in the nucleomorph genome of its close relative *H. andersenii*, means that the complete loss of nucleomorph spliceosomal introns can be pinpointed to somewhere within the genus *Hemiselmis*. Additional nucleomorph genome data from *Chroomonas* and *Hemiselmis* species will help to determine the 'when' and 'how' of spliceosomal intron loss, a seemingly rare event that has only been observed in one other eukaryotic genome: the highly reduced nuclear genome of the intracellular microsporidian parasite *Enterocytozoon bieneusi* (Akiyoshi et al. 2009; Keeling et al. 2010).

Although nucleomorph genes tend to be highly divergent compared to their counterparts in free-living relatives, the nucleomorph genes of *C. mesostigmatica* are more conservative in nature than those in the other cryptophyte nucleomorph genomes sequenced thus far, allowing for additional functional annotation of previously annotated hypothetical protein-coding genes. Furthermore, my comparative analyses show that there is a more highly conserved core of genes present in cryptophyte nucleomorph genomes than previously thought, including an ultra-conserved set of plastid-associated genes. Beyond this highly conserved core, however, there is lineage-specific gene loss: spliceosomal genes in *Hemiselmis,* plastid-associated genes in *Cryptomonas*, and proteasome genes in *Chroomonas*. My synteny analysis has shown the apparent decay of some of these genes, indicating that nucleomorph genome reduction in the cryptophytes has not yet reached an endpoint. Nucleomorph genomes are valuable models for studying genome reduction and compaction in eukaryotes. As I have shown, much can be learned from nucleomorph comparative genomics studies of more closely related cryptophyte species.

CHAPTER 3 PLASTID COMPARATIVE GENOMICS

3.1 Cryptophyte Plastid Genome Biology and Evolution

As described in detail in Chapter 2, the engulfed endosymbiont ancestor of cryptophytes was a red alga. Along with its nucleus, which has been drastically reduced to what we now refer to as the nucleomorph, the red algal endosymbiont contributed a valuable organelle that conferred photosynthetic capability to its host: the plastid. Although red algae are economically and environmentally important organisms, and much is known about their diversity (Yoon et al. 2006), relatively little is known about their plastid genomes. Recently, a manuscript describing three new complete plastid genome sequences from the red algae *Calliarthron tuberculosum*, *Chondrus crispus*, and *Grateloupia lanceola*, and the partial plastid genome of *Cruoria* sp., was published (Janouškovec et al. 2013) The former two genome sequences are now available in GenBank, thus raising the number of complete red algal plastid genome sequences to nine (in contrast, there are presently 27 completely sequenced green algal plastid genomes available in GenBank). The other publicly available red algal plastid genome sequences belong to *Cyanidium caldarium* (Glöckner et al. 2000), *Cyanidioschyzon merolae* (Ohta et al. 2003), *Porphyra purpurea* (Reith and Munholland 1995), *Porphyra umbilicalis* (Smith et al. 2012), *Gracilaria tenuistipitata* var. *liui* (Hagopian et al. 2004), *Pyropia yezoensis* (unpublished), and *Pyropia haitanensis* (unpublished). Given that there are over 6,000 described species and 700 genera of red algae (Guiry and Guiry 2013) that range in habitat from the most extreme aquatic conditions to coastal marine environments, and that there are less than a dozen completely sequenced nuclear and plastid genomes combined, more red algal genomic data are needed in order to understand the evolution of this incredibly complex and diverse group of organisms. Such data will aid in improving our understanding of the process of secondary endosymbiosis.

Red algal plastid genomes are gene dense and contain a much higher number of genes (~220-250) compared to green algal plastids (~100) (Brouard et al. 2011; Janouškovec et al. 2013). Based on recent data, it appears as though red algal plastid genomes are slowly evolving as their gene content is very highly conserved across a large species diversity, with relatively few species-specific ORFs, as well as a high degree of synteny, suggesting that modern red-algal plastid genomes may best approximate the ancestral state of all plastid genomes (Janouškovec et al. 2013). This, however, is not the case for secondary red plastids. Red algal plastid genomes acquired through secondary endosymbiosis differ greatly in size and content. The most extremely reduced red algal-derived plastid genomes are the apicoplast genomes of a group of animal parasites that have secondarily lost photosynthesis, the apicomplexans. Apicoplast genomes are ~35 Kbp in size and encode fewer than 50 proteins, which are mainly used to maintain the apicoplast itself (Wilson et al. 1996). Many nucleus-encoded proteins must be imported into the organelle, as it serves as the site of fatty acid and isoprenoid biosynthesis (Lim and McFaddden 2010). The plastid genomes of peridinin-containing dinoflagellates are also highly reduced relative to red algal plastid genomes, and are unique amongst all plastids derived from secondary endosymbiosis because their genomes are composed of minicircles, small DNA molecules ranging in size from 2-10 Kbp that contain 1-3 genes

(Barbrook et al. 2006a). The peridinin plastids of dinoflagellates are believed to be the ancestral plastid in these organisms; some dinoflagellates have replaced this plastid through the process of tertiary endosymbiosis with plastids derived from a range of endosymbionts, including cryptophytes, haptophytes, and diatoms (which themselves contain secondary red plastids), and through serial secondary endosymbiosis with green algae (Schnepf and Elbrachter 1988; Watanabe et al. 1990; Chesnick et al. 1996; Tengs et al. 2000; Shalchian-Tabrizi et al. 2006; Minge et al. 2010).

The plastid genomes of cryptophytes, haptophytes, and stramenopiles are also reduced relative to red algal primary plastids, although to different extents, and not to the extreme degree of reduction seen in the apicomplexans and peridinin dinoflagellates. As mentioned in Chapter 1, it is not clear whether all secondary red plastids are derived from a single endosymbiotic event or multiple independent secondary and/or tertiary endosymbioses (Cavalier-Smith 1982; Delwiche and Palmer 1997; Hackett et al. 2007; Burki et al. 2007; Burki et al. 2009; Burki et al. 2010; Baurain et al. 2010; Janouškovec et al. 2010; Burki et al. 2012). However, parallels can still be drawn from studying and comparing the process of plastid genome reduction in these lineages.

Before initiating this research, there were only three completely sequenced cryptophyte plastid genomes. Two of the genomes belong to the photosynthetic cryptophytes *R. salina* (Khan et al. 2007) and *G. theta* (Douglas and Penny 1999), and the other belongs to the secondarily non-photosynthetic *C. paramecium* (Donaher et al. 2009). Unsurprisingly, the *C. paramecium* plastid genome contains a reduced set of photosynthesis-related genes relative to the other cryptophyte plastid genomes. In the photosynthetic cryptophytes, plastid gene content appears to be relatively stable, although with only two genome sequences to compare, it is impossible to say whether this trend will hold true across the diversity of cryptophytes. The plastid genome of *R. salina* CCMP1319 was found to possess several unique features that have not been observed in any other non-cryptophyte secondary red plastid to date, such as group II introns and a gene encoding the tau/gamma subunit of DNA polymerase III (*dnaX*), both of which are believed to have been acquired through lateral gene transfer (Khan et al. 2007; Khan and Archibald 2008; Fong and Archibald 2008). Pseudogenes were also found. It is presently unclear how common these features are to cryptophyte plastid genomes, and when the lateral gene transfers occurred.

Contigs corresponding to the *C. mesostigmatica* CCMP1168 plastid genome were identified during assembly of the nucleomorph genome described in Chapter 2. From these contigs I assembled and annotated the plastid genome so that I could compare it to the other red algal and cryptophyte plastid genomes currently sequenced. The addition of the *C. mesostigmatica* plastid genome to the small collection of cryptophyte plastid genomes, as well as other red algal-type secondary plastid genomes, provides additional sequence data from which to infer evolutionary relationships between red algae and secondary red plastid-containing organisms, as well as the processes behind plastid genome reduction in these lineages.

3.2 METHODS

3.2.1 Genome Assembly and Annotation

C. mesostigmatica plastid genome-derived contigs from the assembly described in section 2.2.3 were identified using blastx searches against the GenBank nr database (National Center for Biotechnology Information, Bethesda, MA, USA), and the contigs were manually refined. From these contigs, site-specific primers were designed to close a 38 bp gap between the *rps6* and 5S rRNA genes (CmesoPL_rps6_F2: 5'-TCCTAGTAGAACGAGGGGCTAA-3' and CmesoPL_23S-3'_F1: TTATCGTGCCAACGGTACAC-3'). The PCR product was cloned into a pGEM-T Easy vector (Promega, Madison, WI, USA) and Sanger sequenced (GENEWIZ, Cambridge, MA, USA).

ORFs larger than 30 aa in size were predicted using Artemis (v13.0; Rutherford et al. 2000). Protein-coding genes and pseudogenes were manually annotated based on blastp (e value <0.001; Altschul et al. 1990) searches against the GenBank nr database as well as by comparison to the *R. salina* (Khan et al. 2007) and *G. theta* (Douglas and Penny 1999) plastid genomes. A gene was considered a pseudogene, and thus non-functional, if it was severely truncated due to a premature stop codon, or if the reading frame was interrupted by frameshifts and there was no evidence of introns. rRNA genes were identified by blastn searches against the GenBank nr database. tRNA genes were identified using tRNAscan-SE v1.21 (Lowe and Eddy 1997; http://lowelab.ucsc.edu/tRNAscan-SE/).

3.2.2 Phylogenetic Analyses

A dataset containing 57 group II intron-encoded proteins (IEPs) and reverse transcriptases from bacteria and eukaryotes was created. Sequences in the dataset were collected from GenBank based on (1) top blastp hits (e-value cutoff of 1e-30) of cryptophyte and red algal IEP sequences against the GenBank nr database, (2) the previous analysis of Khan et al. (2008), and (3) specific searches through organellar and bacterial genomes to ensure a comprehensive sampling of red algal IEPs and other eukaryotic IEPs, as well as a broad sampling of bacterial reverse transcriptases. Phylogenetic inferences were made using the complete dataset as well as a smaller subset that excluded mitochondrial IEPs. Protein sequences were aligned using MUSCLE v3.6 (Edgar 2004) with default settings, and the alignments were automatically trimmed using GBlocks v0.91b (Castresana 2000) with the less stringent setting for shorter alignments. The trimmed alignments contained 57 taxa and 165 sites for the full dataset, and 42 taxa and 175 sites for the partial dataset. Phylogenetic inferences were made using RAxML v7.2.8 (Stamatakis 2006) under the following parameters: algorithm = rapid bootstrap (-f a), number of bootstraps = 100 (-N 100), model of evolution = PROTCAT with the LG substitution matrix (-m PROTCATLGF), and independently with PhyML v3.0 (Guidon et al. 2010) under the following parameters: number of bootstraps = 100 (-b 100), substitution model = LG (-m LG), amino acid frequencies = model-given (-f m), and search operation = best of nearest-neighbour interchange (NNI) and subtree pruning and regrafting (SPR) (-s BEST).

3.3 Results and Discussion

3.3.1 Architectural Features of Cryptophyte Plastid Genomes

The *C. mesostigmatica* plastid genome is 139,403 bp in size, has a G+C content of 36%, and contains 149 protein-coding genes, 6 rRNA genes, and 32 tRNAs (Figure 3.1; Table 3.1). Like the other sequenced cryptophyte plastid genomes, the *C. mesostigmatica* plastid genome is very compact: 81.7% of the genome is coding, and there are three instances of overlapping genes. The overlap between the *psbD* and *psbC* genes is a common feature in all red algal-type secondary plastids of photosynthetic species. The overlap between *rpl4* and *rpl23*, and between *atpD* and *atpF*, are also present in the plastid genomes of the photosynthetic cryptophytes *R. salina* (Khan et al. 2007) and *G. theta* (Douglas and Penny 1999), and in diatoms (Oudot-Le Secq et al. 2007). The *C. mesostigmatica* plastid genome is the largest cryptophyte plastid genome sequenced to date, and like the *R. salina* and *G. theta* plastid genomes, it has an inverted repeat containing the rRNA operon (5S, 16S, and 26S rRNA genes) as well as two tRNAs. The plastid genome of the non-photosynthetic species *C. paramecium* contains only a single copy of the rRNA operon; the other copy is presumed to have been lost during the process of genome reduction and compaction that was associated with the loss of photosynthesis, as the area surrounding the inverted repeat is highly conserved in other cryptophyte plastid genomes, yet in *C. paramecium* this area has experienced inversions and gene loss (Donaher et al. 2009). The presence of inverted repeats is likely an ancestral feature, as they are found in red algal plastid genomes (Janouškovec et al. 2013), as well as the secondary red plastid genomes of haptophytes (Sánchez Puerta et al. 2005) and diatoms (Oudot-Le Secq et al. 2007).

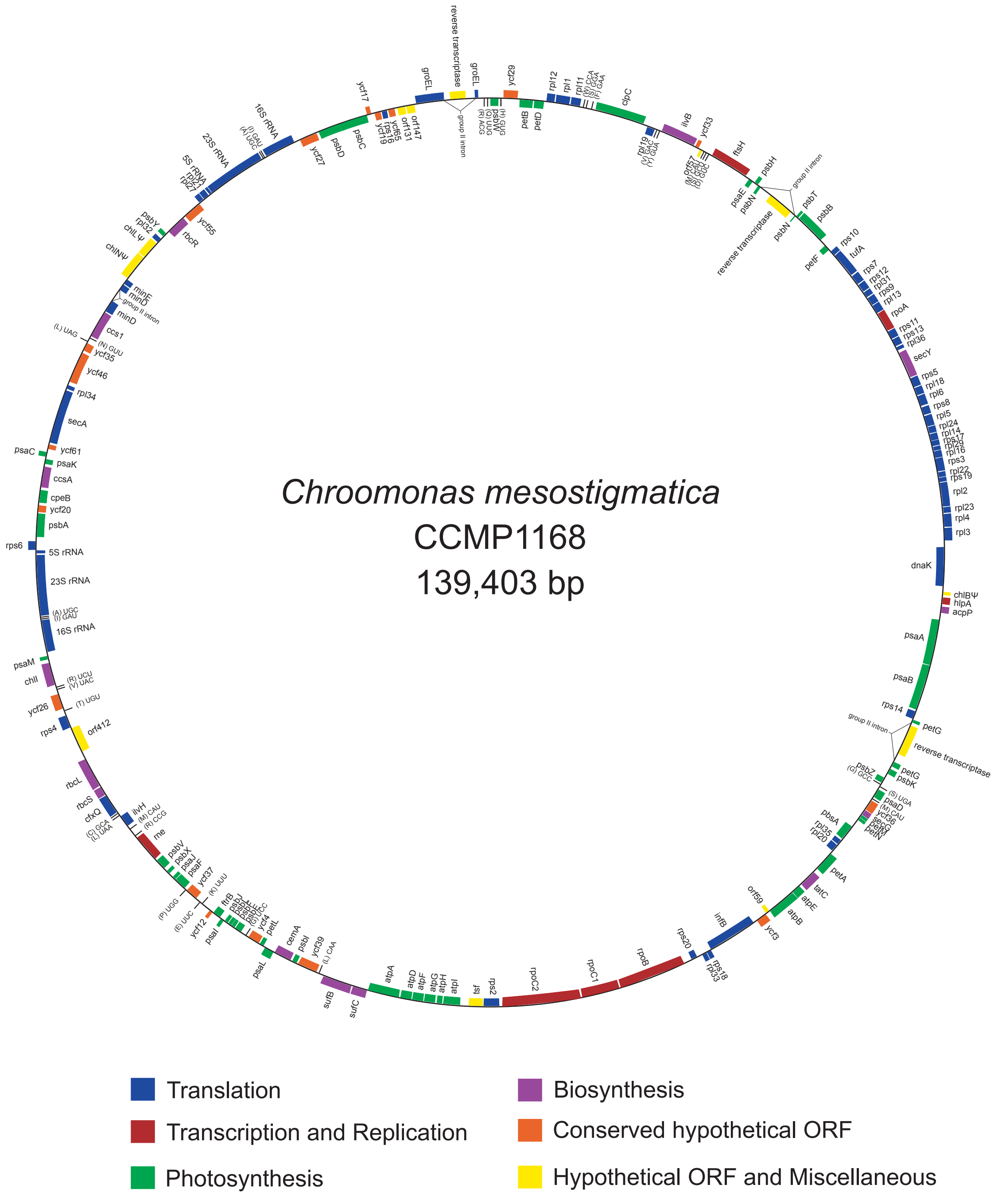


Figure 3.1 Plastid genome map of *Chroomonas mesostigmatica* CCMP1168. Genes located on the outside of the circle are transcribed counterclockwise, and those inside the circle are transcribed clockwise. Genes are coloured according to their functional category.

Comparative analysis of the plastid genomes of photosynthetic cryptophytes (*C. mesostigmatica*, *R. salina*, and *G. theta*) shows that they all have a very similar genome size and G+C content, and they contain a very similar number of protein-coding genes and tRNAs (Table 3.1). The *C. mesostigmatica* plastid genome contains the tRNA (V)-GAC, which is not present in the plastid genome of either *G. theta* or *R. salina*.

Table 3.1 A comparison of the main features of cryptophyte plastid genomes.

Organism	Size (bp)	G+C content (%)	Protein-coding genes	tRNAs	Inverted Repeat
Chroomonas mesostigmatica	139,403	36	149	32	Yes
Rhodomonas salina	135,854	34	146	31	Yes
Guillardia theta	121,524	32	147	30	Yes
Cryptomonas paramecium	77,717	38	82	29	No

Strikingly, 100% gene order conservation was observed between the plastid genomes of the photosynthetic cryptophyte species. The only differences between the three genomes in terms of gene order are the rare instances of a gene or tRNA being present in one species, but absent in one or both of the others (Table 3.2). Of particular note, the *C. mesostigmatica* plastid genome does not contain the *dnaX* gene that is present in the plastid genome of *R. salina.* Khan et al. (2007) found that the *dnaX* gene was possibly laterally transferred from a firmicute bacterium. Given that *dnaX* is not present in the plastid genomes of *G. theta*, *C. paramecium*, or *C. mesostigmatica*, yet is found in several members of the genus *Rhodomonas* (Khan et al. 2007), it is likely that the gene was laterally transferred into an ancestral *Rhodomonas* species. The *C. mesostigmatica* plastid genome contains two other genes not found in any other cryptophyte plastid genome. One gene, *ycf26,* is a conserved hypothetical chloroplast reading frame that is not found in other sequenced secondary red plastid genomes, but is present in most red algal plastid genomes. The gene encodes a 232 aa protein that is similar to the cyanobacterial multi-sensor signal transduction histidine kinase. The other gene is a small (57 aa), unique hypothetical protein-coding gene found between the *ycf33* and *ftsH* genes. This area in the *R. salina* plastid genome was not amenable to sequencing. The authors predict that there are about 100 bp that remain unsequenced in the gap between the two genes, based on restriction enzyme digests, and so there are not likely to be any protein-coding genes present in that area (Khan et al. 2007). The same intergenic region in the *G. theta* plastid genome is 80 bp, so it appears as though the hypothetical protein-coding gene *ORF57* is indeed unique to *C. mesostigmatica*. There are also instances of similarly-sized hypothetical proteins that do not show detectable sequence similarity to one another, but do exhibit gene order conservation in synteny comparisons, as described in the nucleomorph genomes in Chapter 2 (page 46). Again, these genes are presumably homologous, but due to the increased rates of evolution observed in some cryptophyte plastid genomes (Hoef-Emden et al. 2005; Donaher et al. 2009) they have diverged beyond the point of recognition at the sequence level.

The high degree of synteny and gene content conservation suggests that reduction of the cryptophyte plastid genome occurred prior to the diversification of the various cryptophyte lineages for which we have complete plastid genome sequences, and that gene loss and genome rearrangement rarely or never occur, unless there is a major driving factor, such as the loss of photosynthesis. Indeed, recent plastid gene transfers were not detected in the nuclear genome of *G. theta* (Curtis et al. 2012). One possible explanation could be that cryptophytes have only one plastid per cell. Plastid lysis, which would provide genomic material for incorporation into the host nucleus, would likely be a fatal event, and so the window of opportunity for endosymbiotic gene transfer is limited (Barbrook et al. 2006b; Smith et al. 2011).

Table 3.2 Presence/absence of protein genes and tRNAs in the plastid genomes of photosynthetic cryptophyte species. A gene name in the table indicates that the gene occupies the corresponding genomic location, but that homology between the two genes is not detectable.

Gene/tRNA	*Chroomonas mesostigmatica*	*Rhodomonas salina*	*Guillardia theta*
ycf20	+	-	+
chlNΨ	+	+	-
chlLΨ	+	+	-
ycf55	+	-	orf252
orf131	+	orf146	orf65
orf147	+	orf142	orf125
dnaX	-	+	-
trnS(GGA)	+	+	-
dnaB	-	+	+
trnV(GAC)	+	-	-
chlBΨ	+	+	-
orf59	+	-	-
orf62	-	RTΨ	+
orf53	-	-	+
orf412	+	orf403	orf282
ycf26	+	+	-

3.3.2 Pseudogenes in Cryptophyte Plastid Genomes

Another interesting feature of cryptophyte plastid genomes is the presence of pseudogenes (Table 3.3). The three light-independent protochlorophyllide oxidoreductase (LIPOR) subunits *chlB*, *chlL*, and *chlN*, are pseudogenes in *C. mesostigmatica* and *R. salina*. A search for LIPOR genes in the available *C. mesostigmatica* transcriptome data (described in Chapter 2) was unsuccessful. These genes are not present in the plastid genomes of *G. theta* and *C. paramecium*, but are present in other cryptophyte plastid genomes, including *Chroomonas pauciplastida* and *H. andersenii* (Fong and Archibald 2008), which are close relatives of *C. mesostigmatica*, as well as the plastid genomes of the red algae *P. purpurea* and *P. yezoensis*. LIPOR genes are also found in other unrelated plastid genomes, including those of the green algae *Chlamydomonas reinhardtii* (Choquet et al. 1992) and *Chlorella protothecoides* (Shi and Shi 2006), and in the glaucophyte *Cyanophora paradoxa* (Stirewalt et al. 1995), but are missing from the plastid genomes of the red algae *C. merolae*, *C. caldarium*, *G. tenuistipitata*, and *C. crispus*. The presence of these genes in the plastid genomes of some red algae and cryptophytes, yet their absence in others, suggests that the light-independent chlorophyll biosynthesis pathway is non-essential.

There is also a reverse transcriptase (RT) pseudogene present in the plastid genome of *R. salina*, but not in any other fully sequenced cryptophyte plastid genome (Table 3.3). Without complete plastid genome sequences for *C. pauciplastida* or *H. andersenii* it is impossible to determine whether or not there is a reverse transcriptase pseudogene present in those genomes. None of the sequenced red algal plastid genomes contain stand-alone reverse transcriptase genes (Table 3.3). There are, however, group II introns with intron-encoded proteins (IEPs) containing reverse transcriptase domains that are present in several cryptophyte plastid genomes (see below), as well as in the plastid genomes of the florideophytes *Chondrus crispus*, *Calliarthron tuberculosum*, *Gracilaria tenuistipitata*, and *Grateloupia lanceola* (Janouškovec et al. 2013). Given that the *R. salina* plastid genome contains a group II intron with an encoded protein, it is likely that the stand-alone reverse transcriptase pseudogene is a relict group II IEP gene.

Table 3.3 Presence/absence of pseudogenes in cryptophyte and red algal plastid genomes. Full length gene present (+), gene not present (-), pseudogene present (Ψ), or data missing (?) are indicated. RT = reverse transcriptase.

Organism	*chlB*	*chlL*	*chlN*	RT
Cryptophytes				
Chroomonas mesostigmatica	Ψ	Ψ	Ψ	-
Rhodomonas salina	Ψ	Ψ	Ψ	Ψ
Chroomonas pauciplastida	+	+	+	?
Hemiselmis andersenii	+	+	+	?
Guillardia theta	-	-	-	-
Cryptomonas paramecium	-	-	-	-
Red Algae				
Chondrus crispus	-	-	-	-
Cyanidioschyzon merolae	-	-	-	-
Cyanidium caldarium	-	-	+	-
Gracilaria tenuistipitata	-	-	-	-
Pyropia yezoensis	+	+	+	-
Porphyra purpurea	+	+	+	-

3.3.3 Cryptophyte and Red Algal Group II IEPs

Group II introns are a class of self-splicing retroelements from which spliceosomal introns and long non-terminal repeat retrotransposons are thought to originate (Dai and Zimmerly 2002; Lambowitz and Zimmerly 2004; Koonin 2006; Rogozin et al. 2012). The structure of a typical group II intron consists of two parts: the intron RNA and an intron-encoded protein (IEP). A typical IEP consists of four domains: an RT domain (with eight sub-domains), an X domain (maturase), a D domain (DNA binding), and an En domain (endonuclease) (Robart and Zimmerly 2005). Intron splicing requires the maturase activity of the X domain, and all four domains are required for intron mobility. Prior to this study, introns in red algal and secondary red plastids were reported to be rare. Group II introns were initially reported in the cryptophyte plastid genomes of *Rhodomonas salina* (formerly *Pyrenomonas salina*) (Maier et al. 1995) and another *Rhodomonas salina* strain, CCMP1319 (Khan et al. 2007). Maier et al. (1995) identified a unique twintron (a group II intron nested within a group II intron) in the *groEL* plastid gene of *R. salina*. The complete plastid genome sequence of *R. salina* CCMP1319 revealed the presence of a group II intron in the *psbN* gene, but not in the *groEL* gene (Khan et al. 2007). This provoked a survey of group II introns in cryptophyte plastid genomes by Khan et al. (2008), who identified group II introns in the *groEL* gene of some other *Rhodomonas* species, but not all surveyed, as well as in the *chlB* gene of *C. pauciplastida* and *H. andersenii* (initially identified in the study of Fong and Archibald 2008) (Table 3.4). A phylogenetic analysis of the cryptophyte IEP sequences suggested that the cryptophyte group II introns were laterally transferred from a euglenid (Khan et al. 2008).

Table 3.4 Distribution of group II introns in red algal and cryptophyte plastid genomes. Presence is indicated by a '+', absence is indicated by a '-', and question marks indicate missing data. An asterisk indicates that the group II intron does not contain an intron-encoded protein. The *Rhodomonas* sp. CCMP2045 *groEL* gene contains two group II introns, one with an intron-encoded protein, and one without.

Organism	Gene					
	chlB	*groEL*	*minD*	*petG*	*psbN*	tRNA -Met
Cryptophytes						
C. mesostigmatica	-	+	+*	+	+	-
C. pauciplastida	+*	?	?	?	?	-
H. andersenii	+	?	?	?	?	-
R. salina (Maier)	?	+	?	?	?	-
R. salina CCMP1319	-	-	-	-	+	-
R. baltica RCC350	?	-	?	?	?	-
Rhodomonas sp. CCMP1170	?	-	?	?	?	-
Rhodomonas sp. CCMP1178	?	+	?	?	?	-
Rhodomonas sp. CCMP2045	?	+*,+	?	?	?	-
G. theta	-	-	-	-	-	-
C. paramecium	-	-	-	-	-	-
Red Algae						
C. tuberculosum	+	-	-	-	-	+
C. crispus	-	-	-	-	-	+
G. tenuistipitata	-	-	-	-	-	+
G. lanceola	-	-	-	-	-	+
P. purpurea	-	-	-	-	-	-
P. yezoensis	-	-	-	-	-	-
C. caldarium	-	-	-	-	-	-
C. merolae	-	-	-	-	-	-

Whereas other intron-containing cryptophyte plastid genomes possess a single group II intron in a single gene (an exception is the *groEL* gene of *Rhodomonas* sp. CCMP2045, which contains two group II introns), the *C. mesostigmatica* plastid genome is unique in that it contains four group II intron-containing genes: *groEL*, *minD*, *petG*, and *psbN* (Table 3.4). The *minD* group II intron is the only intron that does not contain an IEP. Like the previously identified cryptophyte group II IEPs, the *C. mesostigmatica* group II IEPs lack the D and En domain-encoding regions that are involved in intron mobility. The *petG* group II IEP contains the RT and X domains, the *psbN* group II IEP contains only the RT domain, and the *groEL* group II IEP contains only the N-terminal domain of RT.

A recent paper by Janouškovec et al. (2013) identified group II introns in the plastid genomes of some red algae (Table 3.4), suggesting that the group II introns observed in cryptophyte plastid genomes may have multiple origins, through LGT and vertical inheritance. The florideophyte plastid genomes were found to possess a group II intron with an ORF in the tRNA-Met gene. In the *C. tuberculosum* plastid genome, a second group II intron with an ORF was found in the *chlB* gene. To determine the origin(s) of group II introns in cryptophyte and red algal plastid genomes, I constructed a phylogeny of the group II intron IEP sequences from cryptophyte, red algal (the *chlB* group II IEP from *C. tuberculosum* was too divergent to include in the analysis), and green algal plastids, as well as reverse transcriptases from diverse bacteria (Figure 3.2). The phylogenetic analysis supports a relationship between the group II intron IEPs of red algae and cryptophytes, which branch sister to cyanobacterial reverse transcriptases. Although the relationship between the cryptophyte plastid group II introns and those of euglenids is recovered, as seen by Khan et al. (2007), the statistical support is very weak. Interestingly, the *C. mesostigmatica petG* group II IEP branches with moderate statistical support with the reverse transcriptases of *Escherichia coli*, *Polaromons* sp., firmicute bacteria, and the plastid group II IEP of the green alga *Pyramimonas parkeae* (Figure 3.2). To extend the breadth of IEP sequences, I added group II IEP sequences from the mitochondrial genomes of red algae, green algae, and other eukaryotes (Figure 3.3). The addition of mitochondrial IEP sequences produced a phylogeny that supports the original relationships observed, although with lower statistical support.

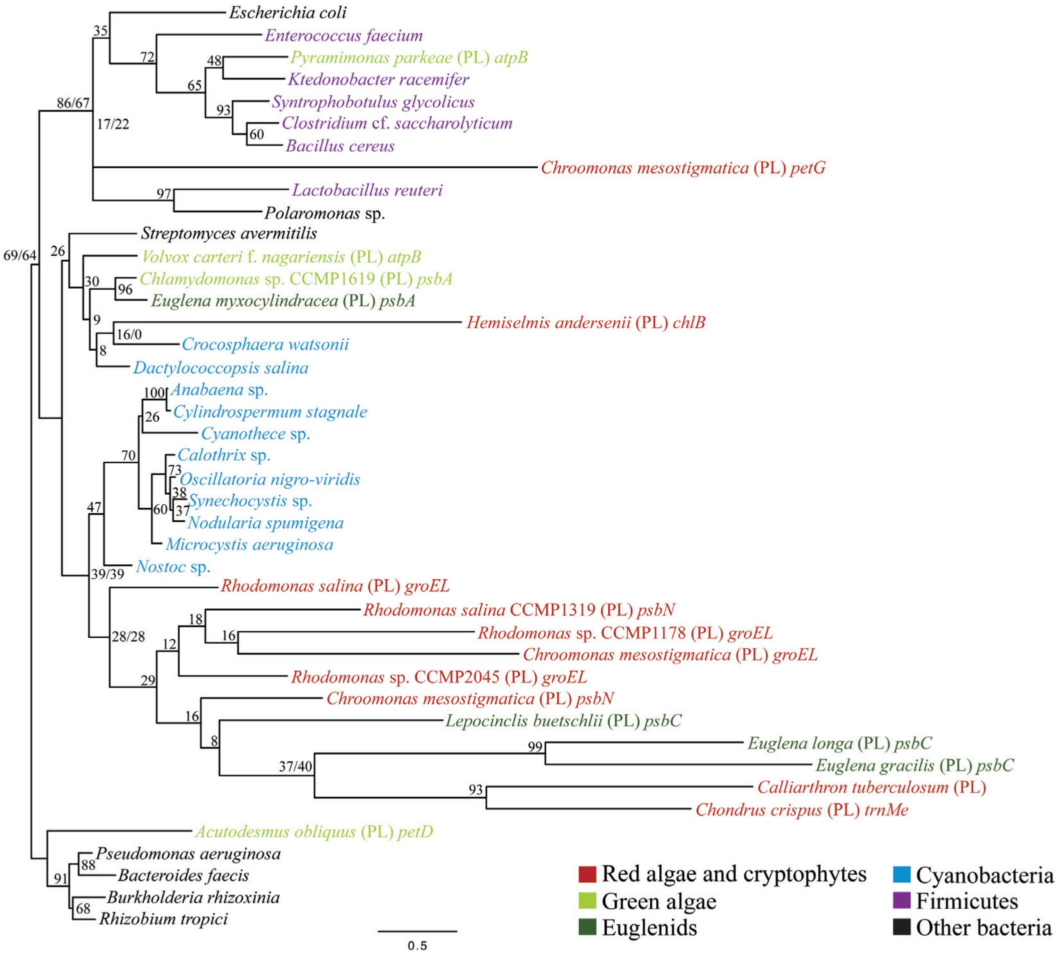

Escherichia coli
Enterococcus faecium
Pyramimonas parkeae (PL) atpB
Ktedonobacter racemifer
Syntrophobotulus glycolicus
Clostridium cf. saccharolyticum
Bacillus cereus
Chroomonas mesostigmatica (PL) petG
Lactobacillus reuteri
Polaromonas sp.
Streptomyces avermitilis
Volvox carteri f. nagariensis (PL) atpB
Chlamydomonas sp. CCMP1619 (PL) psbA
Euglena myxocylindracea (PL) psbA
Hemiselmis andersenii (PL) chlB
Crocosphaera watsonii
Dactylococcopsis salina
Anabaena sp.
Cylindrospermum stagnale
Cyanothece sp.
Calothrix sp.
Oscillatoria nigro-viridis
Synechocystis sp.
Nodularia spumigena
Microcystis aeruginosa
Nostoc sp.
Rhodomonas salina (PL) groEL
Rhodomonas salina CCMP1319 (PL) psbN
Rhodomonas sp. CCMP1178 (PL) groEL
Chroomonas mesostigmatica (PL) groEL
Rhodomonas sp. CCMP2045 (PL) groEL
Chroomonas mesostigmatica (PL) psbN
Lepocinclis buetschlii (PL) psbC
Euglena longa (PL) psbC
Euglena gracilis (PL) psbC
Calliarthron tuberculosum (PL)
Chondrus crispus (PL) trnMe
Acutodesmus obliquus (PL) petD
Pseudomonas aeruginosa
Bacteroides faecis
Burkholderia rhizoxinia
Rhizobium tropici
35
72
48
65
93
60
86/67
17/22
97
69/64
26
30
96
9
16/0
8
100
26
70
73
38
60
37
47
39/39
18
28/28
16
12
29
16
8
99
37/40
93
88
91
68
0.5
🟥 Red algae and cryptophytes
🟩 Green algae
🟩 Euglenids
🟦 Cyanobacteria
🟪 Firmicutes
⬛ Other bacteria

Figure 3.2 Maximum likelihood phylogeny of group II IEPs from algal plastid genomes and reverse transcriptases from bacteria. The concatenated alignment contained 43 taxa and 175 sites. Scale bar indicates inferred number of amino acid substitutions per site. PL = plastid. RAxML bootstrap values are shown. PhyML bootstrap values are also indicated at relevant nodes (RAxML/PhyML).

Interestingly, the *C. mesostigmatica petG* group II IEP along with the mitochondrial IEPs of the red alga *P. yezoensis*, the green alga *Marchantia polymorpha*, and the fungus *Allomyces macrogynus*, as well as the *atpB* plastid IEP of the green alga *Pyramimonas* *parkeae* group sister to a group of firmicute RT sequences with moderate statistical support after the addition of the mitochondrial IEP sequences. This suggests that the organellar genomes could have acquired a group II intron laterally from firmicute bacteria.

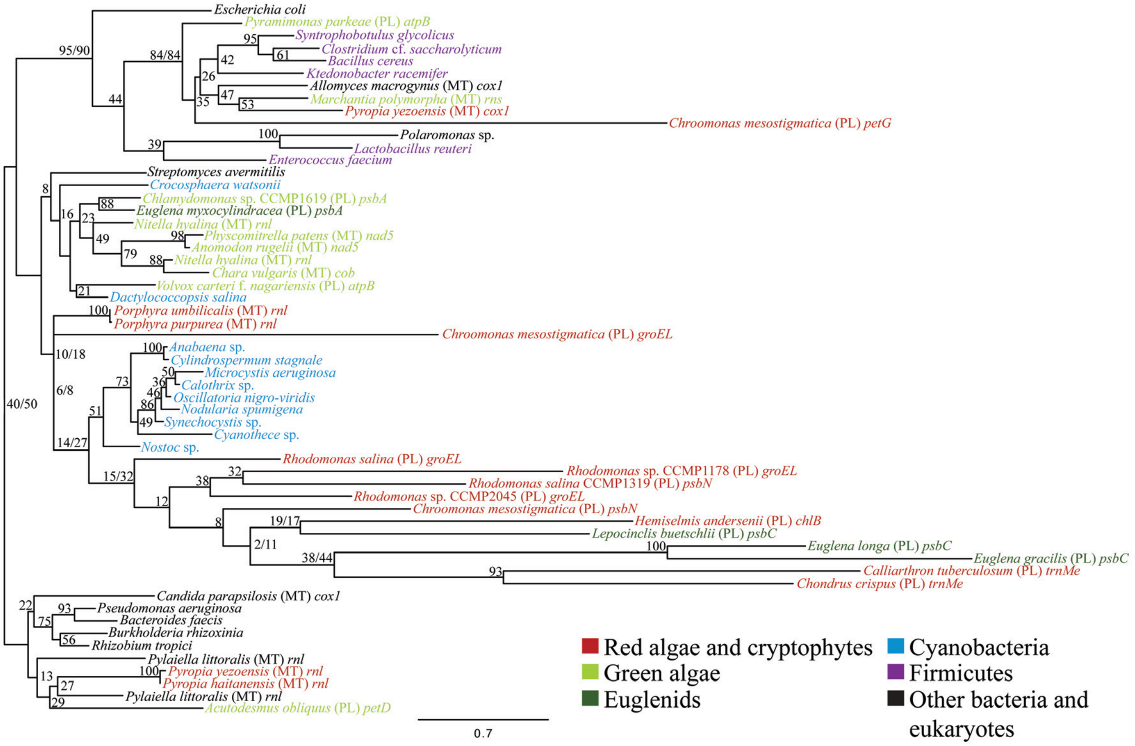

Escherichia coli
Pyramimonas parkeae (PL) atpB
Syntrophobotulus glycolicus
Clostridium cf. saccharolyticum
Bacillus cereus
Ktedonobacter racemifer
Allomyces macrogynus (MT) cox1
Marchantia polymorpha (MT) rns
Pyropia yezoensis (MT) cox1
Chroomonas mesostigmatica (PL) petG
Polaromonas sp.
Lactobacillus reuteri
Enterococcus faecium
Streptomyces avermitilis
Crocosphaera watsonii
Chlamydomonas sp. CCMP1619 (PL) psbA
Euglena myxocylindracea (PL) psbA
Nitella hyalina (MT) rnl
Physcomitrella patens (MT) nad5
Anomodon rugelii (MT) nad5
Nitella hyalina (MT) rnl
Chara vulgaris (MT) cob
Volvox carteri f. nagariensis (PL) atpB
Dactylococcopsis salina
Porphyra umbilicalis (MT) rnl
Porphyra purpurea (MT) rnl
Chroomonas mesostigmatica (PL) groEL
Anabaena sp.
Cylindrospermum stagnale
Microcystis aeruginosa
Calothrix sp.
Oscillatoria nigro-viridis
Nodularia spumigena
Synechocystis sp.
Cyanothece sp.
Nostoc sp.
Rhodomonas salina (PL) groEL
Rhodomonas sp. CCMP1178 (PL) groEL
Rhodomonas salina CCMP1319 (PL) psbN
Rhodomonas sp. CCMP2045 (PL) groEL
Chroomonas mesostigmatica (PL) psbN
Hemiselmis andersenii (PL) chlB
Lepocinclis buetschlii (PL) psbC
Euglena longa (PL) psbC
Euglena gracilis (PL) psbC
Calliarthron tuberculosum (PL) trnMe
Chondrus crispus (PL) trnMe
Candida parapsilosis (MT) cox1
Pseudomonas aeruginosa
Bacteroides faecis
Burkholderia rhizoxinia
Rhizobium tropici
Pylaiella littoralis (MT) rnl
Pyropia yezoensis (MT) rnl
Pyropia haitanensis (MT) rnl
Pylaiella littoralis (MT) rnl
Acutodesmus obliquus (PL) petD
95/90
84/84
95
61
42
26
47
53
35
44
100
39
8
88
16
23
49
98
79
88
21
100
10/18
100
50
36
46
86
73
6/8
49
51
40/50
14/27
32
38
15/32
12
19/17
8
2/11
100
38/44
93
22
93
75
56
100
13
27
29
0.7
🟥 Red algae and cryptophytes
🟩 Green algae
🟩 Euglenids
🟦 Cyanobacteria
🟪 Firmicutes
■ Other bacteria and eukaryotes

Figure 3.3 Maximum likelihood phylogeny of group II IEPs from the plastid and mitochondrial genomes of algae and other eukaryotes, and reverse transcriptases from bacteria. The concatenated alignment contained 57 taxa and 165 sites. Scale bar indicates inferred number of amino acid substitutions per site. MT = mitochondrial. PL = plastid. RAxML bootstrap values are shown. PhyML bootstrap values are also indicated at relevant nodes (RAxML/PhyML).

The relationship between the *C. mesostigmatica petG* IEP and firmicute bacteria is interesting given that the *dnaX* gene in the plastid genome of *R. salina* is also suspected to be laterally transferred from firmicute bacteria (Khan et al. 2007). It has been previously suggested that the mitochondrial group II introns of the red alga *P. purpurea* were acquired laterally from cyanobacteria (Burger et al. 1999). The group II IEPs sequences from *P. purpurea* and *P. umbilicalis* branch sister to the cyanobacterial RT sequences in my analyses as well, but there is no statistical support for this relationship.

Lateral gene transfer into red algal and red algal-derived plastid genomes was once thought to be rare, however, as more plastid genomes from the red lineage are sequenced, more instances of laterally transferred genes are being reported. For example, the genes encoding the large and small subunits of RuBisCO, *rbcL* and *rbcS*, were suggested to have been laterally transferred to red algal plastid genomes from proteobacteria (Delwiche and Palmer 1996; Rice and Palmer 2006). Also, the *rpl36* gene of haptophyte and cryptophyte plastid genomes was found not to be an ortholog of the *rpl36* gene found in all other plastids; rather it was acquired from an undefined lineage of bacteria (Rice and Palmer 2006). In describing group II introns in red algal tRNA genes, Janouškovec et al. (2013) also proposed that the plastid *leuC* and *leuD* genes of *G. tenuistipitata* were laterally acquired from proteobacteria. Although my group II IEP phylogenies only weakly support lateral gene transfer for the *C. mesostigmatica petG* plastid group II intron, and the mitochondrial *rnl* group II introns found in red algae, the possibility cannot be ignored in light of the evidence that both primary and secondary red plastid genomes have taken up foreign DNA. In addition, a maximum likelihood phylogenetic analysis of the group II intron DNA sequences produced a similar result, showing that the *C. mesostigmatica petG* intron branches with the *aptB* intron of *Pyramimonas parkeae* with high statistical support (data not shown). While the support for the *C.*

mesostigmatica petG IEP grouping with the *P. parkeae atpB* IEP, some mitochondrial IEPs, and firmicute reverse transcriptases is low (Figure 3.3), searches of the *C. mesostigmatica petG* group II IEP against the GenBank nr database do show that pairwise comparisons of the IEP sequences are significant. The top blastp hit is the *P. parkeae atpB* group II IEP (1e-68), followed by hits to the firmicute *Ktedonobacter racemifer* reverse transcriptase (7e-41), and then the mitochondrial *cox1* group II IEP of *Allomyces macrogynus* (1e-29). Additionally, a blastp search of the *C. mesostigmatica psbN* group II IEP against the GenBank nr database shows the top blast hit to be the *groEL* group II IEP of *R. salina* CCMP2045 (1e-82), followed by hits to reverse transcriptases of several cyanobacteria (5e-57). The phylogenetic relationship of the *C. mesostigmatica psbN* IEP to the *R. salina* CCMP2045 *groEL* IEP, other cryptophyte, red algal, and euglenid IEPs, and cyanobacterial reverse transcriptases is not well supported in either tree (Figures 3.2 and 3.3). The group II intron DNA phylogeny also shows the *C. mesostigmatica psbN* group II intron branching with the IEP-containing *groEL* group II intron of *R. salina* CCMP2045 with high statistical support (data not shown). This relationship is particularly interesting given that the *R. salina* CCMP2045 *groEL* gene contains two group II introns, one containing an IEP and the other without, and in the group II intron DNA phylogeny, the two introns do not branch together. There is thus reason to suspect that some of the cryptophyte group II introns may have been acquired independently.

The plastid group II IEP protein sequences tend to be divergent and shorter than their counterparts in other genomes, making phylogenetic inferences very difficult. Also, secondary structures of the introns themselves have been shown to differ greatly, and can be extremely difficult to predict (Khan et al. 2008). Additional organellar group II intron IEP sequences from more closely related cryptophytes and red algae are needed to confirm the origin of group II introns in red algal and red algal-derived plastid genomes.

3.3.4 Synteny in Cryptophyte and Red Algal Plastid Genomes

As discussed in section 3.3.1 above, the plastid genomes of photosynthetic cryptophytes have retained 100% gene order conservation. Synteny analysis can be a useful tool for inferring relatedness between species. The recent red algal plastid synteny analysis of Janouškovec et al. (2013) revealed that three orthologous gene clusters (syntenic blocks) account for all of the genes in the plastid genomes of the florideophytes *C. tuberculosum*, *C. crispus*, *G. tenuistipitata*, and *G. lanceola*. When the plastid genomes of *P. purpurea* and *P. yezoensis*, members of the Bangiales, are compared with the florideophyte plastid genomes, the number of syntenic blocks increases to 11, although only five rearrangements are required to align the most divergent genome pair. The authors concluded that red algal plastid genome architecture is slowly evolving relative to other plastid genomes.

Using the 11 orthologous gene clusters defined by Janouškovec et al. (2013), I examined levels of synteny between the *C. mesostigmatica* plastid genome and the plastid genomes of *C. crispus*, a representative florideophyte, and *P. purpurea*, a representative member of the Bangiales. The plastid genome of *P. purpurea* was arbitrarily chosen, as it is collinear with the plastid genome of *P. yezoensis*. *C. crispus* was chosen because its genome arrangement best approximates the ancestral state of florideophyte plastid genomes, as only three inversions are required to align the genome with that of *P. purpurea* (Janouškovec et al. 2013). The *C. mesostigmatica* plastid genome shares 10 syntenic blocks with the plastid genome of *C. crispus* and 11 syntenic blocks with the plastid genome of *P. purpurea* (Figure 3.4).

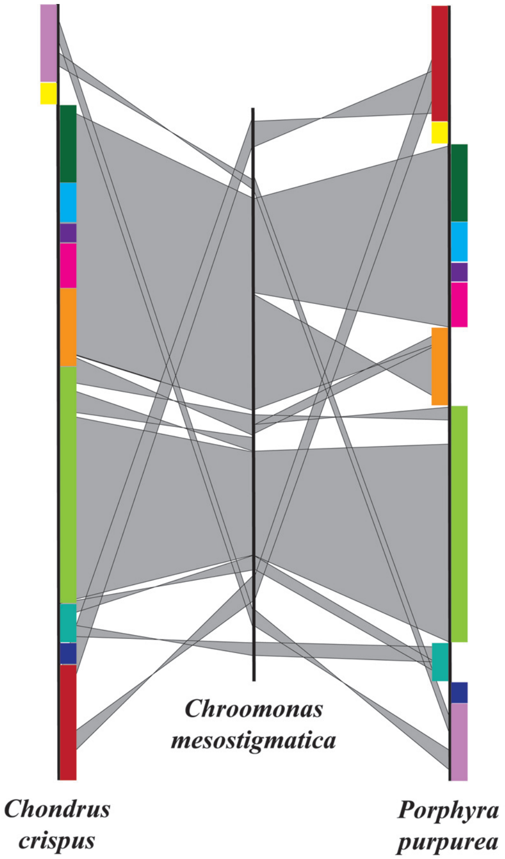


Figure 3.4 Synteny between the *C. mesostigmatica* plastid genome and the red algal plastid genomes of *C. crispus* (Florideophyceae) and *P. purpurea* (Bangiales). Black vertical lines represent the complete genomes, which have been linearized for comparison. Coloured bars represent blocks of synteny between *C. crispus* and *P. purpurea*. Inversions between the *C. crispus* and *P. purpurea* plastid genomes are represented by the same coloured bar being present on different sides (left versus right) of the vertical black lines. The grey shaded areas represent syntenic regions with *C. mesostigmatica*, showing areas of genome rearrangement and inversions.

The number of inversions and the boundaries of the syntenic regions vary between the genomes. There are six inversions between the *C. mesostigmatica* and *C. crispus* plastid genomes, and four inversions between the *C. mesostigmatica* and *P. purpurea* plastid genomes.

The largest syntenic block is between *C. mesostigmatica* and *C. crispus* and spans 51.3 Kbp (~37%) of the *C. mesostigmatica* genome. A 28.3 Kbp subsection of the largest syntenic block is shared between *C. mesostigmatica* and *P. purpurea.* The second largest syntenic block is 31.4 Kbp (~23% of the *C. mesostigmatica* genome) and is present between *C. mesostigmatica* and both red algal plastid genomes. Together, these two syntenic regions encompass 60% of the *C. mesostigmatica* plastid genome. The remaining syntenic blocks are significantly smaller in size; however, when all the syntenic blocks are combined, they comprise ~85% of the protein-coding genes in *C. mesostigmatica*.

In comparison, a synteny analysis between the *C. mesostigmatica* plastid genome and the plastid genome of *C. merolae*, a representative member of the Cyanidiales, shows a greater degree of rearrangement between the two genomes. There are 14 syntenic blocks between the plastid genomes of *C. mesostigmatica* and *C. merolae* and six inversions (Figure 3.5). The syntenic regions between *C. mesostigmatica* and *C. merolae* are significantly smaller than those between *C. mesostigmatica* and *C. crispus* or *P. purpurea*. There are five syntenic blocks between *C. mesostigmatica* and *C. merolae* that are larger than 10 Kbp, ranging in size from 18.1 Kbp to 11.6 Kbp. Together, these five syntenic blocks only contain ~52% of the *C. mesostigmatica* plastid genome. However, when all the syntenic regions are combined, 90.6% of the protein-coding genes present in the *C. mesostigmatica* plastid genome are present in syntenic regions with the *C. merolae* plastid genome. The plastid genome of *C. merolae* is more than 30 Kbp smaller than the other currently sequenced plastid genomes (excluding *C. caldarium*, which is also a member of the Cyanidiales), but contains a similar number of genes. The smaller genome size is attributed to its impressive degree of compaction; the median intergenic spacer size is 10 bp, about 6-8 times smaller than the median intergenic distances observed in the other red algal plastid genomes (Janouškovec et al. 2013).

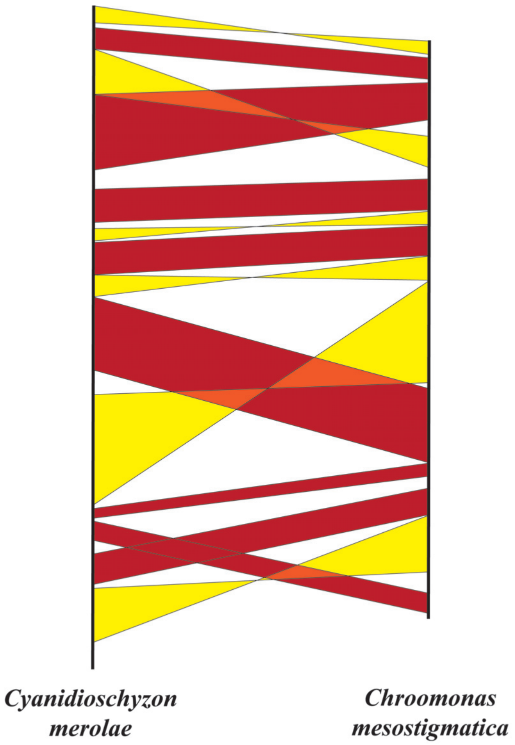


Figure 3.5 Synteny map showing areas of gene order conservation between the *C. mesostigmatica* and *C. merolae* plastid genomes. The black vertical lines represent the complete genomes, which have been linearized for comparison. Syntenic regions are coloured in red and yellow, with yellow indicating an inversion.

Another interesting comparison is the number of rRNA operons, their orientation, and the particular genes that bound them. All of the florideophyte and cyanidiophycean plastid genomes currently sequenced possess only a single copy of the rRNA operon, unlike the plastid genomes of the Bangiales and secondary red plastid-containing lineages, like the cryptophytes (excluding *C. paramecium*), which possess repeats containing the rRNA operon. In the bangialean plastid genomes, the repeats are direct repeats, whereas in the cryptophyte plastid genomes, the repeats are inverted. The inverted repeat is presumed to be an ancestral feature, as it is found in green- and red-algal primary and secondary plastids, implying that a copy has been lost in the florideophyte and cyanidiophycean plastid genomes (Janouškovec et al. 2013). A comparison of the genes surrounding the *C. mesostigmatica* plastid rRNA operons to those surrounding the rRNA operon(s) in red algal plastids shows a variety of arrangements (Figure 3.6).

In one *C. mesostigmatica* repeat, *chlI-psaM* precedes the rRNA operon, with *rps6* immediately downstream. This exact arrangement is not observed in any of the analyzed red algal plastid genomes, but one of the repeats in *P. purpurea* has *chlI-psaM* preceding the operon, and the other repeat has *rps6* immediately following the operon. Several of the red algal genomes, including *C. crispus, G. tenuistipitata*, and *C. caldarium* also have *rps6* downstream of the operon. The other *C. mesostigmatica* repeat has *psbD-ycf27* preceding the operon, and *rpl21-rpl27* immediately following. Again, this exact arrangement is not observed in any of the red algal plastid genomes, however, *C. crispus*, *G. tenuistipitata*, *P. purpurea*, and *C. tuberculosum* do have *psbD-ycf27* preceding the operon. The only red algal plastid genome to have precisely the same downstream arrangement of *rpl21-rpl27* is that of *C. merolae.*

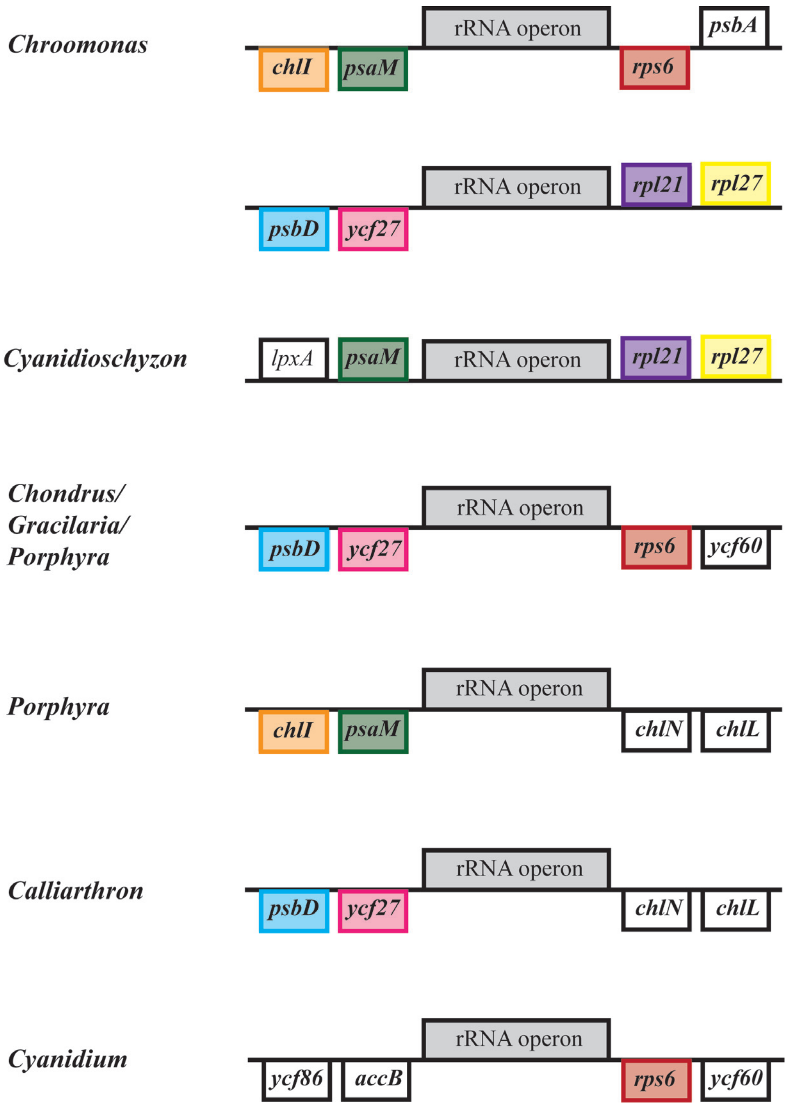


Figure 3.6 Gene order conservation in the rRNA operon-containing repeat of cryptophyte and red algal plastid genomes. Genes immediately upstream and downstream of the plastid rRNA operons are shown, and those flanking the *C. mesostigmatica* plastid genome that are also found flanking the rRNA operon(s) of red algal plastid genomes are coloured. Genes on the top of the black line are transcribed left to right, and those on the bottom are transcribed right to left.

Whether comparing cryptophyte plastid gene order with that of the reduced cyanidiophycean plastid genomes, or that of the more conserved bangialean or florideophyte plastid genomes, gene loss and genome rearrangements have made it difficult to infer gene order in the ancestral red algal plastid genome. Based on the genome-wide synteny analysis, the cryptophyte plastid genome is most similar in architecture to the plastid genomes of the Florideophyceae. However, the additional analysis of genes bounding the *C. mesostigmatica* rRNA operons does indicate that there are particular shared architectural features with the plastid genome of the cyanidiophycean *C. merolae.*

Douglas and Penny (1999) noted the presence of tRNA genes at the boundaries of the syntenic blocks between the cryptophyte *G. theta* and the red alga *P. purpurea* plastid genomes, as well as adjacent to *P. purpurea* genes that have been lost from the *G. theta* plastid genome. The authors suggest that tRNAs may play a role in gene loss and genome rearrangement.

3.3.5 Gene Content of 'Red' Plastids

A comparison of the complement of protein-coding genes in red algal and red algal-derived plastid genomes reveals a spectrum in terms of gene content (Figure 3.7). The degree of genome reduction in red algal-derived plastids, and even within those of the red algae, varies dramatically. As a result, coding capacity also varies between the lineages. At the top end of the gene content spectrum are the red algae themselves. Sequenced red algal plastid genomes range in size from ~150-192 Kbp and contain 189-209 protein-coding genes (Reith and Munholland, 1995; Glöckner et al. 2000; Ohta et al. 2003; Hagopian et al. 2004; Smith et al. 2012; Janouškovec et al. 2013). The plastid-encoded proteins are used primarily in photosynthesis, biosynthesis, and in basic housekeeping functions, such as transcription and translation. At the bottom end of the spectrum are the highly reduced plastid genomes of apicomplexans and the peridinin-containing dinoflagellates. The 35 Kbp apicoplast genome contains 30 protein-coding genes: 17 ribosomal proteins, 3 subunits of RNA polymerase, *sufB*, *clpC*, *tufA*, and 7 hypothetical ORFs (Wilson et al. 1996). The plastid genome of peridinin dinoflagellates is composed of 2-10 Kbp minicircles that contain usually one, but up to three genes, and some are empty (Barbrook et al. 2006a). The 12 protein-coding genes contained in these genomes encode components of photosystems I and II, ATP synthase subunits, and genes involved in electron transfer. There is absolutely no overlap between the protein sets of apicomplexan and peridinin dinoflagellate plastids (Figure 3.7).

In the middle of the gene content spectrum are the secondary red plastids of cryptophytes, haptophytes, and stramenopiles. The plastid genomes in these lineages contain genes whose proteins function primarily in photosynthesis and maintenance of the organelle. Some genes for biosynthetic pathways, such as iron-sulfur cluster formation, have been retained in cryptophytes and diatoms (and also in apicomplexans), but many of the biosynthesis-related genes present in red algal plastid genomes have been lost (Douglas and Penny 1999; Sánchez Puerta et al. 2005; Khan et al. 2007; Oudot-Le Secq et al. 2005; Donaher et al. 2009).

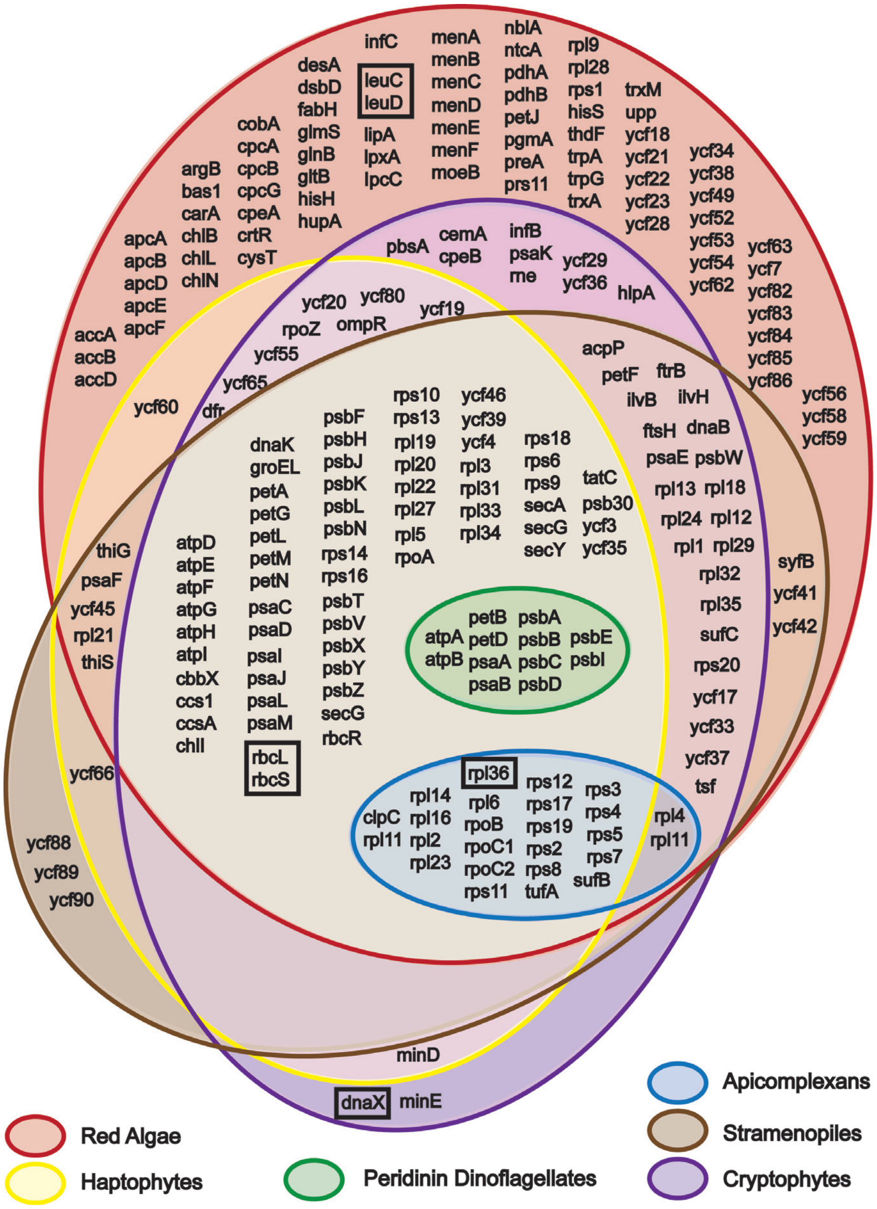


Figure 3.7 Red algal and red algal-derived secondary plastid gene content comparison. Coloured circles represent the total protein-coding gene complement present in completely sequenced plastid genomes for members of that group. Boxed genes have been acquired through LGT. Red algae = *Cyanidium caldarium*, *Cyanidioschyzon merolae*, *Gracilaria tenuistipitata*, *Porphyra purpurea*, and *Pyropia yezoensis*. Cryptophytes = *C. mesostigmatica*, *G. theta*, and *R. salina*. Haptophytes = *Emiliania huxleyi* and *Pavlova lutheri*. Stramenopiles = the diatoms *Phaeodactylum tricornutum*, *Odontella sinensis*, *Synedra acus*, and *Thalassiosira pseudonana* and the brown algae *Ectocarpus siliculosus* and *Fucus vesiculosus.* Apicomplexans = *Plasmodium falciparum* and *Toxoplasma gondii*. Peridinin dinoflagellates = minicircles of *Amphidinium carterae*, *Amphidinium operculatum*, and *Heterocapsa triquetra*.

The total protein-coding gene complement of red algal primary and secondary plastid genomes, excluding the 8 genes acquired through LGT (discussed in section 3.3.3 above) is 234 genes (Figure 3.7). Out of these 234 protein genes, 97 (41.5%) are shared amongst the primary and secondary plastid genomes (excluding the apicomplexans and peridinin dinoflagellates). There are very few protein genes present in the secondary plastid genomes that are not present in red algal plastid genomes, namely *minD* and *minE*, which are involved in plastid division, and four conserved hypothetical ORFs (ycfs). Haptophytes share a single gene with red algae that is not found in other red plastids, *ycf60*. Stramenopiles share three genes with red algae that are not found in other red plastids: the phenylalanine tRNA-synthetase gene *syfB* and two ycfs.

Cryptophyte plastid genomes share the most genes in common with red algae to the exclusion of other red plastids, including the genes for heme oxygenase (*pbsA*), an envelope membrane protein (*cemA*), subunit B of phycoerythrin (*cpeB*), translation initiation factor 2 (*infB*), a histone-like DNA-binding protein (*hlpA*), ribonuclease E (*rne*), photosystem I subunit X (*psaK*), and two ycfs. Cryptophyte plastid genomes contain 60.1% (139 out of 228) of the genes found in red algal plastid genomes, the highest proportion of retained genes out of all the red algal-derived secondary plastid-containing lineages (stramenopiles = 57%, haptophytes = 48.7%). Although cryptophytes and stramenopiles retain the most similar gene set, there is no clear pattern to which genes are retained in the plastid genomes of the two lineages. Perhaps as more red algal primary and secondary plastid genomes are sequenced and the identities of the ycfs are established, gene loss patterns will emerge.

3.4 Chapter Summary

With the addition of the *C. mesostigmatica* plastid genome sequence, there are now four complete cryptophyte plastid genome sequences available for comparison. These plastid genome sequences come from diverse cryptophytes, yet they possess many similarities, particularly within the photosynthetic species. With only one plastid genome sequence available from a non-photosynthetic cryptophyte, it is difficult to infer the processes involved in gene loss and genome reduction associated with the loss of photosynthesis in this lineage. Comparison of plastid genomes from the photosynthetic species, however, has given a great deal of insight into cryptophyte plastid evolution. Notably, gene order is conserved across the entire plastid genome, and there are very few unique and/or missing genes. This finding suggests that large-scale reduction of the cryptophyte plastid following secondary endosymbiosis occurred prior to diversification of the cryptophytes.

Gene order is also conserved in several large syntenic regions of the cryptophyte plastid genome with those of red algae, although the degree of synteny is dependent upon the red algal lineage being considered. The largest region of gene order conservation is with the plastid genome of the florideophyte *C. crispus*. The *C. mesostigmatica* plastid genome shares the least amount of synteny with the plastid genomes of the extremophilic Cyanidiales, which themselves display the least amount of synteny with other red algal plastid genomes. Interestingly, although the *C. mesostigmatica* plastid genome is least similar in terms of synteny to the plastid genome of the cyanidiophycean *C. merolae*, one of the *C. mesostigmatica* inverted repeats exhibits gene order conservation downstream of the rRNA operon with *C. merolae*, an arrangement of genes that is not present downstream of the rRNA operon in any of the other red algal plastid genomes. The upstream gene region has gene order conservation with the Bangiales and florideophytes as do the flanking regions of the rRNA operon in the other repeat.

In addition to gene order conservation, cryptophyte plastid genomes also share the largest proportion of genes with red algal plastids out of the secondary red plastid-containing lineages. In the scheme of reductive evolution following secondary endosymbiosis, the cryptophyte plastids are most similar to modern-day red-algal plastids, and so are useful for gaining insight into the architecture and gene repertoire of the ancestral secondary red plastid.

The most surprising feature of the *C. mesostigmatica* plastid genome is that it contains four group II introns, each located in a distinct gene. In comparison, the completely sequenced *R. salina* CCMP1319 plastid genome contains two group II introns found within the same gene, and the completely sequenced *G. theta* plastid genome contains only a single group II intron. Group II introns have recently been reported in the tRNA-Met gene of florideophyte plastid genomes, as well as in the *chlB* gene of *C. tuberculosum*, warranting an investigation into the origin(s) and relationship(s) between the group II introns of red algal and cryptophyte plastid genomes. Although the origin(s) of the group II introns are not clear from phylogenies of the group II IEP sequences, the placement of the *C. mesostigmatica petG* IEP within the group containing firmicutes and other eukaryotes is significant given that this grouping is well-supported and that the position of the *C. mesostigmatica petG* IEP sequence does not appear to be an artefact of long-branch attraction, unlike the other cryptophyte group II IEP sequences. Furthermore, pair-wise comparisons of the *C. mesostigmatica petG* IEP and firmicute reverse transcriptase sequences show considerable similarity, suggesting that cryptophyte plastid genomes may have acquired group II introns more than once, and from different bacterial donors. If this is the case, then in light of other evidence that suggests a number of plastid genes have non-cyanobacterial origins, it appears as though secondary plastids may be affected by LGT to a higher degree than previously thought.

CHAPTER 4 ORIGIN OF THE CRYPTOPHYTE PLASTID

4.1 INTRODUCTION

Cryptophytes are a diverse group of algae that are mostly photosynthetic (some *Cryptomonas* species have secondarily lost the ability to photosynthesize) and that inhabit both marine and freshwater environments (Shalchian-Tabrizi et al. 2008). Resolving the tree of cryptophytes has proven difficult as there are very few complete genome sequences available (one nuclear, four nucleomorph, and four plastid genomes) for concatenated multi-protein phylogenies, and analyses using nuclear or nucleomorph rRNA sequences have not provided enough resolution to determine the relative branching order of the highly supported subgroups (clades) with certainty (Hoef-Emden 2008). As more cryptophyte plastid and nucleomorph genome sequences have become available, there have been attempts to elucidate the branching order in the tree of cryptophytes using concatenated protein alignments. For example, Donaher et al. (2009) constructed a maximum-likelihood (ML) phylogenetic tree from 22 plastid proteins and 5,076 amino acid positions (sites), derived from 16 taxa including the cryptophytes *C. paramecium*, *H. andersenii*, and *G. theta*, and from red algae, stramenopiles, the glaucophyte *C. paradoxa*, green algae, and rooted using the cyanobacterium *Synechocystis*. In their tree, *R. salina* and *G. theta* grouped together to the exclusion of *C. paramecium* with maximum statistical support, although the *C. paramecium* branch was more than double the length of the *R. salina* and *G. theta* branches. This relationship is not observed in the phylogenetic analysis of Tanifuji et al. (2010) in which host nuclear 18S rRNA genes were used to infer relationships between the cryptophyte clades. In their analysis, *Cryptomonas* species group with *G. theta* to the exclusion of *Rhodomonas* species, although with less than 70% bootstrap support. Unfortunately, attempts using concatenated nucleomorph protein datasets have produced ambiguous results, due in large part to the divergent nature of nucleomorph-encoded proteins (Goro Tanifuji, personal correspondence).

Another lingering uncertainty related to cryptophyte evolution is the identity of the red algal ancestor that gave rise to the cryptophyte plastid and nucleomorph. The previous phylogenetic approaches described above have not given any real indication of which red algal species is most closely related to cryptophytes when either nucleomorph or plastid genes are compared. Results from Khan et al. (2007), who analyzed a concatenation of 45 plastid proteins (9,081 sites) from cryptophytes, haptophytes, stramenopiles, green algae, plants, glaucophytes, and cyanobacteria, showed the cryptophytes (together with the haptophytes and stramenopiles) branching with *G. tenuistipitata* and *P. purpurea* to the exclusion of the Cyanidiales, *C. merolae* and *C. caldarium*, although with weak statistical support (posterior probability of 0.61). Support for this relationship improved dramatically, however, when fast evolving sites were removed. Donaher et al. (2009) reported a similar relationship in their phylogenetic analysis of 22 concatenated plastid proteins (described above) with the cryptophytes branching (along with the haptophyte *E. huxleyi*) with *G. tenuistipitata* and *P. purpurea* to the exclusion of the Cyanidiales, although again with very weak statistical support (posterior probability of 0.56).

Identification of the red algal ancestor of cryptophyte nucleomorphs and plastids may help to resolve the even more contentious issue of whether all secondary red plastids are derived from the same endosymbiotic event, or whether red secondary plastids were acquired multiple times. There are presently two main, although not mutually exclusive, hypotheses regarding the evolution of red secondary plastids: the chromalveolate hypothesis and the hacrobian hypothesis. The chromalveolate hypothesis posits that there was a single endosymbiotic event that gave rise to all red algal-derived secondary plastid-containing lineages, namely the cryptophytes, haptophytes, stramenopiles (known collectively as the chromists), and the alveolates, which consists of the ciliates, dinoflagellates, and apicomplexans (Cavalier-Smith 1999). The basic tenet of the chromalveolate hypothesis is that secondary plastid acquisition is a complex event that involved extensive gene transfer and cell biological innovation. The number of secondary endosymbioses invoked should thus be as few as possible. Some members of the 'chromalveolate' group are plastid-lacking (e.g., ciliates), and so if the chromalveolate hypothesis is true, this means that they have lost their plastid secondarily.

In support of the chromalveolate hypothesis is a uniting feature of all red algal-derived secondary plastids (excluding apicoplasts): the presence of chlorophyll-*c* (Cavalier-Smith 1986). In addition, the plastids of cryptophytes, haptophytes, and stramenopiles are bounded by four membranes and share a unique membrane topology whereby the outermost membrane is contiguous with the host endomembrane and nuclear envelope (Cavalier-Smith 1999; Gould et al. 2008). The plastids of peridinin dinoflagellates and apicomplexans are highly derived, and their genomes highly reduced, thus making them difficult to compare with other secondary plastids of the red lineage and for their gene sequences to be included in phylogenetic analyses. The recent discovery of a photosynthetic organism related to the apicomplexans, *Chromera velia* (Moore et al. 2008), has shed light on the relationship of apicomplexans and dinoflagellates to other chromalveolates. Phylogenetic analyses of plastid proteins including those from *C. velia* support the monophyly of alveolate plastids and a relationship to red algal plastids, as well as a common origin of the plastids in alveolates and stramenopiles, however a single origin for chromalveolates is not well supported (Janouškovec et al. 2010). Nuclear gene phylogenies have added an additional layer of complexity to the chromalveolate hypothesis by supporting the inclusion of the Rhizaria. The Rhizaria are a supergroup composed primarily of amoebae and amoeboflagellates (Nikolaev et al. 2004; Burki et al. 2007). Numerous phylogenetic analyses support a relationship between **s**tramenopiles, **a**lveolates and **r**hizarians, informally called the **SAR** clade (Hackett et al. 2003; Burki et al. 2007; Burki et al. 2009; Burki et al. 2010; Burki et al. 2012; Brown et al. 2012). Recent large-scale analyses of nuclear genes have not supported the chromalveolate hypothesis. In one analysis, the haptophytes branch with the SAR clade, but the cryptophytes do not (Burki et al. 2012), and in another analysis, neither the cryptophytes nor haptophytes go with the SAR clade (Baurain et al. 2010).

The hacrobian hypothesis posits that the cryptophytes and haptophytes are each other's closest relatives and share a common plastid (Okamoto et al. 2009). This hypothesis is based on a culmination of evidence along different lines, including the presence of the laterally-transferred *rpl36* gene in the plastids of cryptophytes and haptophytes, as well as support from plastid and nuclear phylogenies (Hackett et al. 2007; Patron et al. 2007; Janouškovec et al. 2010). However, recent evidence suggests that the cryptophyte and haptophyte host lineages may not be monophyletic (Burki et al. 2012).

Due to the differing signals observed in plastid and host nuclear gene phylogenies, determining the position of the cryptophytes is pivotal in our understanding of plastid evolution in those lineages containing red plastids derived through secondary endosymbiosis. With the addition of the *C. mesostigmatica* nucleomorph and plastid genomes described in Chapters 2 and 3, I was able to expand the cryptophyte nucleomorph and plastid protein dataset. From this new dataset, I constructed concatenated protein phylogenies to try and better resolve the tree of cryptophytes and to identify the red algal ancestor of the cryptophyte nucleomorph and plastid with the aim of gaining new insight into the relationship of cryptophytes to other red plastid-containing lineages.

4.2 METHODS

A local database containing plastid and nuclear-encoded protein sequences from red algal and red algal secondary plastid-containing organisms (Table 4.1) was created using the makeblastdb program from BLAST (Altschul et al. 1990). Nucleomorph and plastid protein sequences from the cryptophytes *C. mesostigmatica*, *H. andersenii* (nucleomorph proteins only), *G. theta*, *C. paramecium*, and *R. salina* (plastid proteins only) were added to the database.

Table 4.1 List of taxa used for phylogenetic analyses. The number of entries represents the number of gene or expressed sequence tag (EST) sequences present in the database for each organism, and for which genome. Sequences were retrieved from the data sources listed. NCBI = National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). JGI Genome Portal = United States Department of Energy Joint Genome Institute Genome Portal (http://genome.jgi.doe.gov).

Organism	Genome	# of entries	Source
Arabidopsis thaliana	Nuclear	35,130	http://www.arabidopsis.org/
Arabidopsis thaliana	Plastid	85	NCBI
Aureococcus anophagefferens	Plastid	105	JGI Genome Portal
Calliarthron tuberculosum	Nuclear	23,961	http://dbdata.rutgers.edu/data /plantae/
Calliarthron tuberculosum	Plastid	201	NCBI
Chlamydomonas reinhardtii	Nuclear	19,595	JGI Genome Portal
Chlamydomonas reinhardtii	Plastid	69	NCBI
Chondrus crispus ESTs translated in-house	Nuclear	4114	NCBI
Chondrus crispus	Plastid	204	NCBI
Chroomonas mesostigmatica	Nucleomorph	452	This study
Chroomonas mesostigmatica	Plastid	159	This study
Cryptomonas paramecium	Nucleomorph	466	NCBI
Cryptomonas paramecium	Plastid	82	NCBI
Cyanidioschyzon merolae	Nuclear	5014	http://merolae.biol.s.utokyo. ac.jp/download
Cyanidioschyzon merolae	Plastid	207	NC_004799
Cyanidium caldarium	Plastid	197	NCBI
Cyanophora paradoxa	Nuclear	32,167	http://cyanophora.rutgers.edu /cyanophora/blast.php
Cyanophora paradoxa	Plastid	149	NCBI
Ectocarpus siliculosus	Plastid	148	https://bioinformatics.psb. ugent.be/gdb/ectocarpus
Emiliania huxleyi	Plastid	119	NCBI
Gracilaria sp. ESTs	Nuclear	8347	NCBI
Gracilaria tenuistipitata	Plastid	203	NCBI

Organism	Genome	# of entries	Source
Guillardia theta	Nucleomorph	485	NCBI
Guillardia theta	Plastid	147	NCBI
Hemiselmis andersenii	Nucleomorph	515	NCBI
Micromonas pusilla	Nuclear	10,475	JGI Genome Portal
Micromonas pusilla	Plastid	57	JGI Genome Portal
Ostreococcus tauri	Nuclear	7651	JGI Genome Portal
Ostreococcus tauri	Plastid	61	JGI Genome Portal
Phaeodactylum tricornutum	Plastid	132	JGI Genome Portal
Porphyra sp. ESTs translated in-house	Nuclear	28,104	NCBI
Porphyra purpurea	Plastid	209	NCBI
Porphyridium cruentum ESTs translated in-house	Nuclear	21,702	http://dbdata.rutgers.edu/data/plantae/
Porphyridium cruentum	Plastid	36	NCBI
Rhodomonas salina	Plastid	146	NCBI
Thalassiosira pseudonana	Plastid	141	NCBI

The *C. mesostigmatica* nucleomorph and plastid protein sets were compared against the local database using blastp v2.2.26 (Altschul et al. 1990) with default settings. The blastp results were parsed to identify highly similar proteins using an e-value cutoff of 1e-40. A protein was selected for phylogenetic analysis if highly similar proteins were identified from at least three red algae. If more than one protein sequence from the same organism was considered highly similar to the corresponding *C. mesostigmatica* protein, then the protein with the best hit was retained. A total of 175 *C. mesostigmatica* protein sequences were selected for phylogenetic analysis.

Individual multiple sequence alignments were made for each of the 175 protein sequences using MUSCLE v3.8.31 (Edgar 2004). Ambiguously aligned positions were removed from the alignments in an automated fashion using BMGE v1.1 (Criscuolo and Gribaldo 2010) with default settings. Phylogenetic inferences were made using the trimmed alignments with RAxML v7.2.5 (Stamatakis 2006) under the following parameters: algorithm = rapid bootstrap (-f a), number of bootstraps = 100 (-N 100), model of evolution = PROTCAT with the LG substitution matrix (-m PROTCATLGF). Each of the 175 individual trees was examined for sufficient taxonomic breadth as well as problematic taxa, such as those producing long branches. A number of proteins were removed from further analysis, resulting in a total of 57 nucleomorph and 58 plastid proteins.

The individual alignments for the 57 selected nucleomorph proteins were concatenated into a single multiple-protein alignment consisting of 15 taxa and 13,230 sites. The individual alignments for the 58 selected plastid proteins were concatenated into a single multiple-protein alignment consisting of 21 taxa and 17,211 sites. Phylogenetic inferences were made using the concatenated alignments with RAxML as described above.

4.3 Results and Discussion

4.3.1 Phylogeny of Cryptophyte and Red Algal Plastid Genes

To assess the relative branching order of the cryptophytes and the closest red algal relative of the cryptophyte plastid, I constructed a phylogenetic tree from a concatenation of 58 conserved plastid proteins from cryptophytes, stramenopiles, the haptophyte *E. huxleyi*, red algae, and from green algae and the land plant *Arabidopsis thaliana*, which were used as outgroup taxa (Figure 4.1).

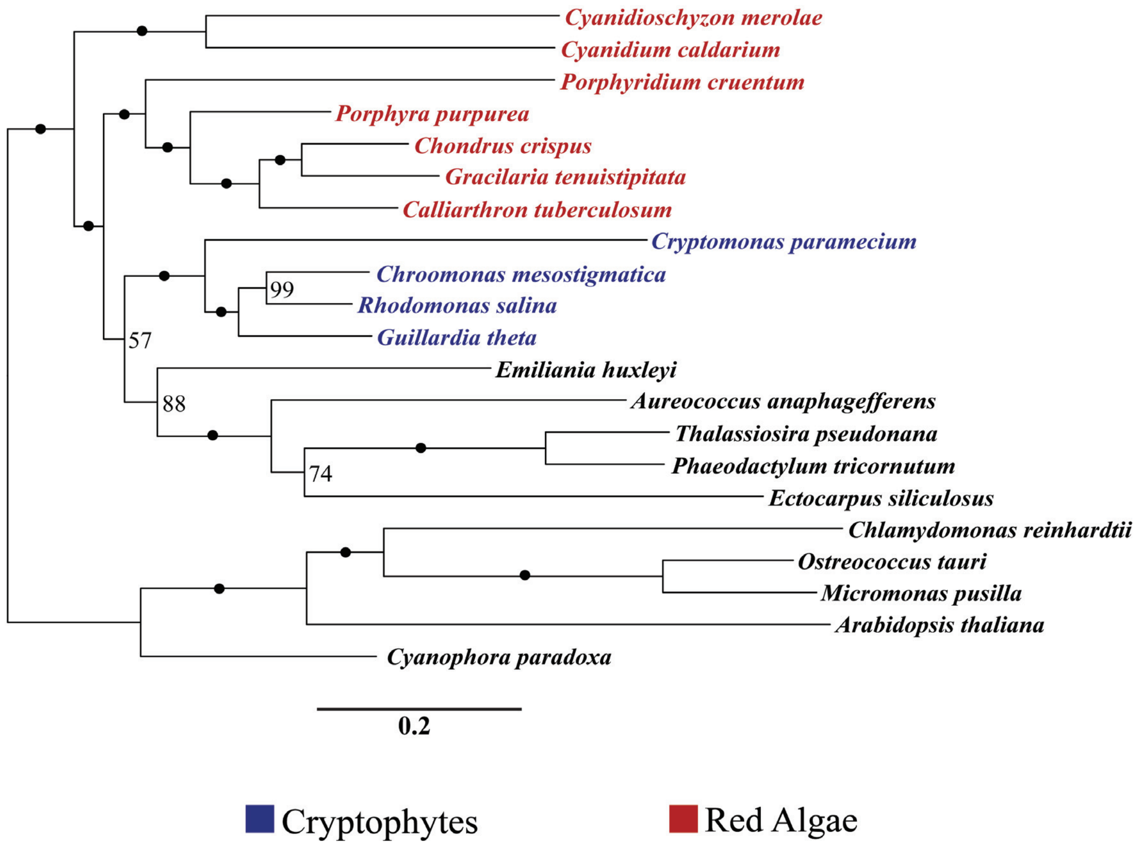


Figure 4.1 Maximum likelihood phylogeny of 58 concatenated cryptophyte and red-algal plastid proteins from 21 taxa (17,211 sites) Black dots represent maximum bootstrap support. Scale bar represents the number of amino acid substitutions per site.

In the phylogeny the branching order of the cryptophyte taxa relative to each other is highly supported, with maximum bootstrap support for almost all nodes. *C. mesostigmatica* and *R. salina* are the most closely related taxa, and branch sister to *G. theta*. In the phylogeny, these three taxa are more closely related to each other than to *C. paramecium*. This exact relationship was previously reported in a host nuclear 18S rRNA gene phylogeny, although with weak statistical support (Tanifuji et al. 2010). In another phylogeny constructed from 22 plastid proteins from the plastid genomes of *R. salina*, *G. theta*, and *C. paramecium*, *R. salina* and *G. theta* branch together to the exclusion of *C. paramecium* (Donaher et al. 2009). Although there is no plastid genome sequence for a member of the genus *Hemiselmis*, previous phylogenies have consistently shown that relative to other cryptophyte clades, members of the genus *Chroomonas* and the genus *Hemiselmis* are each others closest relatives and that *Hemiselmis* species may actually have evolved from within the *Chroomonas* clade (Lane et al. 2005; Hoef-Emden et al. 2008; Tanifuji et al. 2010). Lack of sequence data from a *Hemiselmis* species should have little to no impact on the branching order observed.

In my analysis, the cryptophytes and other red plastid-containing taxa form a monophyletic group, but with weak statistical support. The red secondary plastid-containing group branches sister to the red algal group containing the florideophytes and Bangiophyceae, to the exclusion of the Cyanidiophyceae. The synteny analysis in Chapter 3 (section 3.3.4) supports this relationship, as cryptophyte plastid genomes were found to share longer regions of gene order conservation with the plastid genomes of the Florideophyceae and Bangiales than to those of the Cyanidiophyceae. In previously published phylogenies of concatenated datasets of 34+ conserved plastid proteins, as well as subsets of functionally-related proteins (e.g., photosystem apparatus, transcription and translation) the same relationship was observed (Hagopian et al. 2004; Janouškovec et al. 2010). Interestingly, the phylogenetic analysis of Janouškovec et al. (2010) that supports the relationship between cryptophytes and Florideophyceae/Bangiophyceae also supports the hacrobian grouping, as cryptophytes and haptophytes were shown to branch together to the exclusion of the stramenopiles and alveolates. In my analysis, however, the haptophytes do not branch with the cryptophytes to the exclusion of other chromalveolates, and thus the hacrobian grouping is not supported.

4.3.2 Nucleomorph and Red Algal Nuclear Gene Phylogeny

To gain another perspective on the evolutionary relationships of the cryptophyte clades relative to each other, and collectively to red algae, I performed a phylogenetic analysis using a concatenated protein alignment of 57 conserved proteins from cryptophyte nucleomorph genomes, red algal nuclear genomes, and from the nuclear genomes of green algae, the land plant *Arabidopsis thaliana*, and the glaucophyte *Cyanophora paradoxa*, which were used as outgroup taxa (Figure 4.2). The nucleomorph protein phylogeny shows a different branching order of the cryptophyte taxa compared to the plastid phylogeny. In the nucleomorph protein tree, *C. mesostigmatica* and *H. andersenii* group together, and are sister to the group containing *G. theta* and *C. paramecium*. As discussed earlier, members of the genus *Chroomonas* and the genus *Hemiselmsis* are known to be closely related, and more closely related to each other than to other cryptophytes. The *G. theta* and *C. paramecium* branches are twice as long as the *C. mesostigmatica* and *H. andersenii* branches. Nucleomorph genes are known to be highly divergent in sequence (Hoef-Emden et al. 2002; Lane et al. 2006; Phipps et al. 2008, Tanifuji et al. 2010) and so additional nucleomorph protein sequences from closely related taxa will be required to determine the branching order of the cryptophyte clades independently of, and to verify the findings from, phylogenies using plastid proteins.

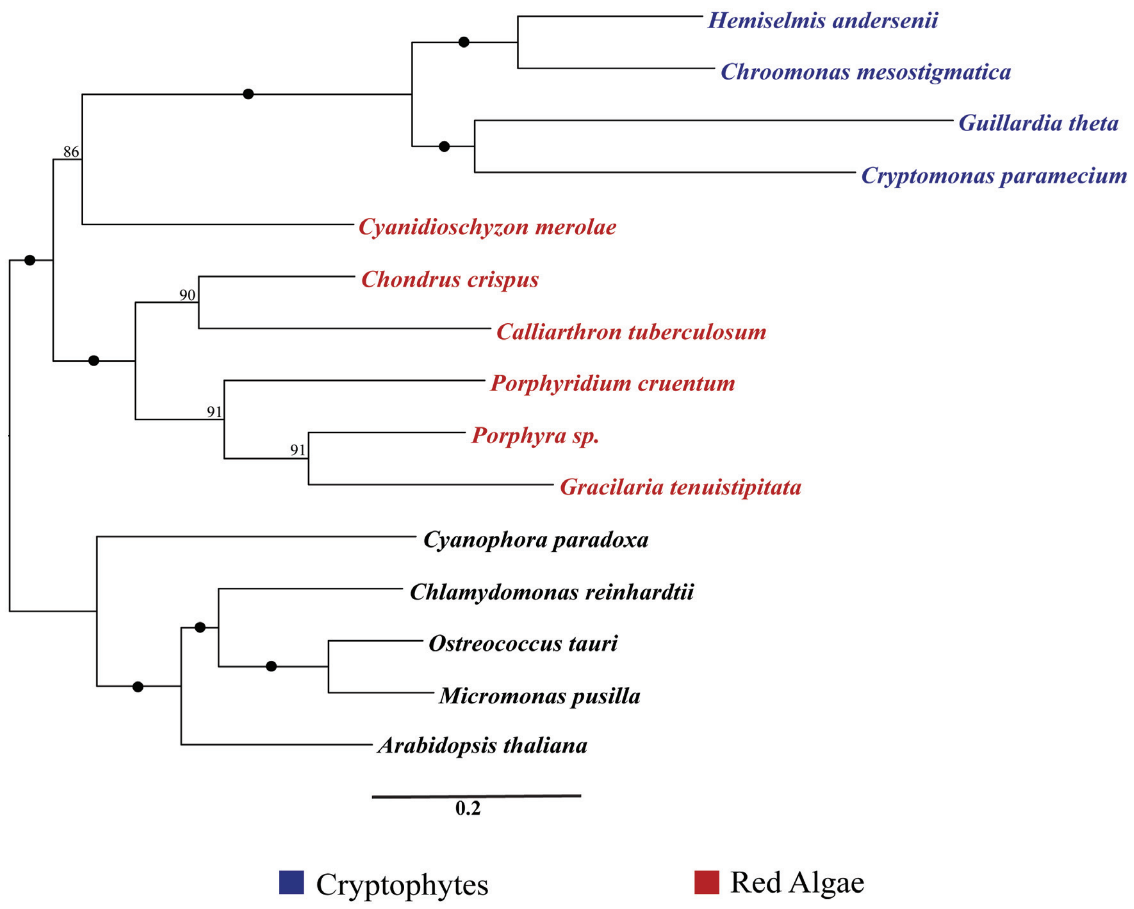


Figure 4.2 Maximum likelihood phylogeny of 57 concatenated cryptophyte nucleomorph and red-algal nuclear proteins from 15 taxa (13,320 sites). Black dots represent maximum bootstrap support. Scale bar represents the number of amino acid substitutions per site.

Interestingly, in the nucleomorph protein phylogeny the cryptophytes branch sister to *C. merolae* to the exclusion of all other red algae, albeit with only moderate statistical support. This relationship is not observed in the plastid protein phylogeny described in section 4.3.1. In fact, in the plastid phylogeny the opposite relationship exists whereby the cryptophytes branch with the florideophytes and Bangiophyceae to the exclusion of the Cyanidiophyceae with maximum statistical support. Although the dataset of red algal nuclear genes is expanding, thanks to the recent genome and EST data that have been made publicly available, the breadth of coverage of red algal diversity is still very low. Additional nucleomorph and nuclear genome data from red algae is necessary to determine the ancestor of the cryptophyte nucleomorph.

4.4 Chapter Summary

The plastid and nucleomorph protein phylogenies produce incongruent evolutionary histories of the cryptophytes and the origin of their red algal-derived organelles. The nucleomorph phylogeny recovers the previously highly supported grouping of *Chroomonas* and *Hemiselmis*, and show this pair branching sister to a clade containing *Cryptomonas* and *Guillardia*. The phylogeny also shows the closest red-algal relative of the cryptophytes to be the extremophile *Cyanidioschyzon merolae*. This relationship is not observed in my phylogeny of plastid proteins, nor in other phylogenies using concatenated protein datasets. In contrast, my plastid protein phylogeny does support the close relationship of cryptophytes to Florideophyceae and Bangiophyceae that is observed in other plastid protein phylogenies. Furthermore, the branching order of the cryptophytes, for which there are complete plastid genome sequences, is highly supported in the plastid phylogeny and supports previous phylogenies based on host rRNA gene sequences. The plastid phylogeny also weakly recovers chromalveolate taxa as a monophyletic group, but does not specifically support the hacrobian relationship of the cryptophyte and haptophytes, a relationship that has recently been challenged by a large-scale phylogeny using host nuclear gene sequences (Burki et al. 2012). Future analyses could employ Bayesian phylogenetic inference methods that account for heterogeneity across sites using models such as GAMMA+CAT, and/or the removal of fast-evolving sites in the alignment. Another strategy could be amino acid recoding, as codon-bias is known to be a feature of compositionally biased genomes like those of nucleomorphs and plastids.

It is interesting to note that the Cyanidiophyceae are long-branching in the plastid protein tree, but not in the nucleomorph protein tree. The higher rate of genome 'scrambling' in the Cyanidiales plastid (discussed in section 3.3.4) combined with the observed differences in branching order between the nucleomorph and plastid protein trees makes similarities between cryptophyte and red algal plastids difficult to interpret. Furthermore, nucleomorph gene sequences tend to be divergent relative to their counterparts in free-living organisms. It was hoped that with expanded taxonomic sampling from more cryptophytes, phylogenetic artefacts might be reduced such that congruence between nucleomorph and plastid gene phylogenies could be achieved, and new knowledge about the origin and evolution of the cryptophyte organelle obtained. The data in hand are still clearly insufficient to address this problem. Additional data from red algal nuclear and cryptophyte nucleomorph genomes, in particular the identification and use of 'short' branching taxa that can be used as surrogates for related ‘long’-branching members, should help improve the resolution of nucleomorph protein-based phylogenies.

CHAPTER 5 CONCLUSIONS

The process of secondary endosymbiosis has had a profound impact on the genomes of photosynthetic eukaryotes. Substantial gene loss and EGT following endosymbiosis has resulted in varying degrees of genome reduction and compaction, which in some cases has been to the extent that the endosymbiont-derived genomes have been completely lost. The cryptophytes represent an intermediate stage in this process, as the nucleus of the red algal endosymbiont is still present, albeit in a miniaturized form called a nucleomorph. The complete nucleomorph genome sequence of *C. mesostigmatica* CCMP1168 is the fourth cryptophyte nucleomorph genome to be completely sequenced, and has added to our depth of knowledge of nucleomorph genome biology and evolution by providing information about nucleomorph genome architecture in members of the previously unsampled *Chroomonas* clade, which possess some of the largest nucleomorph genomes known (Lane et al. 2005; Phipps et al. 2008; Tanifuji et al. 2010). Indeed, the *C. mesostigmatica* CCMP1168 nucleomorph genome is the largest nucleomorph genome currently sequenced, and by comparing it to the other cryptophyte nucleomorph genomes, I have shown that the largest contributing factor to the diversity of genome sizes observed within the cryptophytes is not differences in the number of genes, or the length of the genes, but rather the size of the intergenic regions, which influences the rate at which synteny is reduced. In addition, I have shown that of all the sequenced cryptophyte nucleomorph genomes, the *C. mesostigmatica* nucleomorph genome most closely resembles the ancestral red algal nuclear genome in terms of basic genome architecture, as it exhibits the lowest degree of genome reduction, compaction and gene loss, and possesses numerous repetitive regions, multi-copy genes, and spliceosomal introns, features that are much less prevalent or completely absent in the other sequenced cryptophyte nucleomorph genomes. Furthermore, identification of spliceosomal introns in all major cryptophyte clades except *Hemiselmis*, coupled with confirmation of the close relationship of *C. mesostigmatica* to *H. andersenii* through gene sequence comparisons, synteny analyses, and phylogenetic analyses, means that future analyses examining the 'when' and 'how' of spliceosomal intron loss, a rare phenomenon that has only been reported in one other nuclear genome (Akiyoshi et al. 2009; Keeling et al. 2010), can be focused directly on members of the genus *Hemiselmis*.

What is the fate of the cryptophyte nucleomorph? Will it eventually disappear? My nucleomorph genome comparative analyses show that although there is a more highly conserved core set of genes conserved amongst cryptophyte nucleomorph genomes than previously thought, including an ultra-conserved set of plastid-associated genes, there is also lineage-specific gene loss: *H. andersenii* has lost spliceosomal introns and the genes required for their removal (Lane et al. 2007), the non-photosynthetic *C. paramecium* has lost plastid-associated genes (Tanifuji et al. 2011), and I have shown that *C. mesostigmatica* has lost all genes encoding subunits of the proteasome. My analysis of gene order conservation amongst the cryptophyte nucleomorph genomes has revealed the apparent decay of some of these genes. This finding indicates that cryptophyte nucleomorph genome reduction has not yet reached an endpoint. Are the genes simply being lost? Or are they being transferred to the host nucleus? Without a complete nuclear genome sequence for *C. mesostigmatica* it is impossible to say, however, the recently

sequenced nuclear genome of *G*. *theta* has given insight into these questions.
A survey of EGT-derived genes in the nuclear genomes of *G. theta* and *B. natans* revealed no evidence of recent nucleomorph-to-nucleus gene transfer (Curtis et al. 2012). In addition, several genes that are missing from the nucleomorph genome of *G. theta*, but that are present in the nucleomorph genomes of other cryptophytes, were found to have host-derived versions (the products of host gene duplications), whose protein products are predicted to be imported into the PPC (Curtis et al. 2012). These observations suggest that nucleomorph DNA is no longer being incorporated into the host genome, and that genes lost from nucleomorph genomes are either being accompanied by loss of function within the endosymbiont-derived compartments, or are having the functions of their protein products supplemented by host-derived proteins. The presence of pseudogenes and relict ORFs in syntenic regions between the cryptophyte nucleomorph genomes does suggest that genome reduction is still occurring, but the apparent inability of these organisms to transfer nucleomorph DNA to the host genome means that nucleomorphs are likely to persist.

Cryptophyte plastid genomes provide an additional line of evidence that can be used to trace the evolutionary history and origin of red algal-derived plastids. The *C. mesostigmatica* plastid genome is the third plastid genome to be sequenced from the photosynthetic cryptophytes. My comparison of cryptophyte plastid genomes (excluding the non-photosynthetic *C. paramecium*) reveals that their architectures are evolving incredibly slowly, as gene order is 100% conserved and there are very few unique and/or missing genes. In addition, my comparative analyses of cryptophyte and red algal plastid genomes have shown that they share architectural similarity, a large gene repertoire, and features not present in other secondary red algal-derived plastids, such as group II introns. These features make cryptophyte plastid genomes ideal candidates for exploring the evolution of red secondary plastids. The identification of four group II introns in the *C. mesostigmatica* plastid genome, and the evidence presented here that at least some of them may have been acquired through LGT from difference sources, combined with previous reports of several plastid genes derived from LGT events, show that red plastid genomes and their derivatives are perhaps more dynamic than previously thought.

The identity of the red algal ancestor that gave rise to the cryptophyte plastid and nucleomorph remains a mystery. The *C. mesostigmatica* CCMP1168 plastid and nucleomorph genome sequences presented here have extended the dataset of red nuclear and plastid proteins. Using this extended dataset, I have attempted to resolve the tree of cryptophytes and gain insight into the closest living red algal relative of the cryptophyte plastid and nucleomorph. However, the incongruent phylogenetic signals in the nucleomorph and plastid gene trees, and lack of resolution of the cryptophyte sequences relative to those of red algae, indicate that current taxonomic sampling is insufficient. The current collection of red algal nuclear and plastid genome sequences represents only a tiny fraction of red algal diversity, which limits our ability to infer relationships between the plastids of chromalveolates and their ancestor. In spite of this, I have shown that the cryptophytes are of pivotal importance to our understanding of the process of plastid evolution through secondary endosymbiosis, and ultimately, that there is still much to be learned from the study of their endosymbiotically-derived organelles.

REFERENCES

Akiyoshi DE, Morrison HG, Lei S, Feng X, Zhang Q, et al. 2009. Genomic survey of the non-cultivable opportunistic human pathogen, *Enterocytozoon bieneusi*. PLoS Pathog. 5:e1000261.

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

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

Archibald JM, Lane CE. 2009. Going, going, not quite gone: nucleomorphs as a case study in nuclear genome reduction. J Hered. 100:582-590.

Barbrook AC, Santucci N, Plenderleith LJ, Hiller RG, Howe CJ. 2006a. Comparative analysis of dinoflagellate chloroplast genomes reveals rRNA and tRNA genes. BMC Genomics. 7:297.

Barbrook AC, Howe CJ, Purton S. 2006b. Why are plastid genomes retained in non-photosynthetic organisms? Trends Plant Sci. 11:101-108.

Baurain D, Brinkmann H, Petersen J, Rodriguez-Ezpeleta N, Stechmann A, et al. 2010. Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles. Mol Biol Evol. 27:1698-1709.

Bolte K, Gruenheit N, Felsner G, Sommer MS, Maier UG, et al. 2011. Making new out of old: Recycling and modification of an ancient protein translocation system during eukaryotic evolution. Bioessays. 33:368-376.

Bonfield JK, Smith KF, Staden R. 1995. A new DNA sequence assembly program. Nucleic Acids Res. 23:4992-4999.

Brouard J-S, Otis, C, Lemieux C, Turmel M. 2011. The chloroplast genome of the green alga *Schizomeris leibleinii* (Chlorophyceae) provides evidence for bidirectional DNA replication from a single origin in Chaetophorales. Genome Biol Evol. 3:505-511.

Brown MW, Kolisko M, Silberman JD, Roger AJ. 2012. Aggregative multicellularity evolved independently in the eukaryotic supergroup Rhizaria. Curr Biol. 22:1123-1127.

Burger G, Saint-Louis D, Gray MW, Lang BF. 1999. Complete sequence of the mitochondrial DNA of the red alga *Porphyra purpurea*. Cyanobacterial introns and shared ancestry of red and green algae. Plant Cell. 11:1675-1694.

Burki F, Shalchian-Tabrizi K, Minge M, Skjaeveland A, Nikolaev SI, et al. 2007. Phylogenomics reshuffles the eukaryotic supergroups. PLoS ONE. 2:e790.

Burki F, Inagaki Y, Brate J, Archibald JM, Keeling PJ, et al. 2009. Large-scale phylogenomic analyses reveal that two enigmatic protist lineages, Telonemia and Centroheliozoa, are related to photosynthetic chromalveolates. Genome Biol Evol. 1:231-238.

Burki F, Kudryavtsev A, Matz MV, Aglyamova GV, Bulman S, et al. 2010. Evolution of Rhizaria: new insights from phylogenomic analysis of uncultivated protists. BMC Evol Biol. 10:377.

Burki F, Okamoto N, Pombert JF, Keeling PJ. 2012. The evolutionary history of haptophytes and cryptophytes: phylogenomic evidence for separate origins. Proc R Soc B. 279:2246-2254.

Cao K, Nakajima R, Meyer HH, Zheng Y. 2003. The AAA-ATPase Cdc48/p97 regulates spindle disassembly at the end of mitosis. Cell. 115:355-367.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17:540-552.

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

Cavalier-Smith T. 1999. Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and eukaryote family tree. J Euk Microbiol. 46:246-366.

Cavalier-Smith T. 2000. Membrane heredity and early chloroplast evolution. Trends Plant Sci. 5:174-182.

Cavalier-Smith T. 2002. Nucleomorphs: enslaved algal nuclei. Curr Opin Microbiol. 5:612-619.

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

Ciniglia C, Yoon HS, Pollio A, Pinto G, Bhattacharya D. 2004. Hidden biodiversity of the extremophilic Cyanidiales red algae. Mol Ecol. 13:1827-1838.

Collén J, Porcel B, Carré W, Ball SG, Chaparro C, et al. 2013. Genome structure and metabolic features in the red seaweed *Chondrus crispus* shed light on evolution of the Archaeplastida. Proc Natl Acad Sci U S A. 110:5247-5252.

Conaway RC, Brower CS, Conaway JW. 2002. Emerging roles of ubiquitin in transcription regulation. Science. 296:1254-1258.

Criscuolo A, Gribaldo S. 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol. 10:210.

Curtis BA, Tanifuji G, Burki F, Gruber A, Irimia M, et al. 2012. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. Nature. 492:59-65.

Dai L, Zimmerly S. 2002. Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behaviour. Nucleic Acids Res. 30:1091-1102.

Delwiche CF, Palmer JD. 1996. Rampant horizontal transfer and duplication of rubisco genes in eubacteria and plastids. Mol Biol Evol. 13:873-882.

Delwiche CF, Palmer JD. 1997. The origin of plastids and their spread via secondary endosymbiosis. Plant Syst Evol. 11:S53-S86.

Donaher N, Tanifuji G, Onodera NT, Malfatti SA, Chain PSG, et al. 2009. The complete plastid genome sequence of the secondarily non-photosynthetic alga *Cryptomonas paramecium*: reduction, compaction, and accelerated evolutionary rate. Genome Biol Evol. 1:439-448.

Douglas SE, Penny SL. 1999. The plastid genome of the cryptophyte alga *Guillardia theta*: complete sequence and conserved synteny groups confirm its common ancestry with red algae. J Mol Evol. 48:236-244.

Douglas S, Zauner S, Fraunholz M, Beaton M, Penny S, et al. 2001. The highly reduced genome of an enslaved algal nucleus. Nature. 410:1091-1096.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792-1797.

Eschbach S, Hofmann CJ, Maier UG, Sitte P, Hansmann P. 1991. A eukaryotic genome of 660 kb: electrophoretic karyotype of nucleomorph and cell nucleus of the cryptomonad alga *Pyrenomonas salina*. Nucleic Acids Res. 19:1779-1781.

Finley D, Chau V. 1991. Ubiquitination. Annu Rev Cell Biol. 7:25-69.

Fong A, Archibald JM. 2008. Evolutionary dynamics of light-independent protochlorophyllide oxidoreductase genes in the secondary plastids of cryptophyte algae. Eukaryotic Cell. 7:550-553.

Gardner RG, Nelson ZW, Gottschling DE. 2005. Ubp10/Dot4p regulates the persistence of ubiquitinated histone H2B: distinct roles in telomeric silencing and general chromatin. Mol Cell Biol. 25:6123-6139.

Gilson PR, McFadden GI. 2002. Jam packed genomes – a preliminary, comparative analysis of nucleomorphs. Genetica. 115:13-28.

Gilson PR, Su V, Slamovits CH, Reith ME, Keeling PJ, et al. 2006. Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus. Proc Natl Acad Sci U S A. 103:9566-9571.

Glöckner G, Rosenthal A, Valentin K. 2000. The structure and gene repertoire of an ancient red algal plastid genome. J Mol Evol. 51:382-390.

Gould SB, Sommer MS, Hadfi K, Zauner S, Kroth PG, et al. 2006a. Protein targeting into the complex plastids of cryptophytes. J Mol Evol. 62:674-681.

Gould SB, Sommer MS, Kroth PG, Gile GH, Keeling PJ, et al. 2006b. Nucleus-to-nucleus gene transfer and protein retargeting into a remnant cytoplasm of cryptophytes and diatoms. Mol Biol Evol. 23:2413-2422.

Gould SB, Waller RF, McFadden GI. 2008. Plastid evolution. Annu Rev Plant Biol. 59:491-517.

Green BR. 2004. The chloroplast genome of dinoflagellates – a reduced instruction set? Protist. 155:23-31.

Guidon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology. 59:307-321.

Guiry MD, Guiry GM. 2013. AlgaeBase: Worldwide electronic publication, National University of Ireland, Galway. http://www.algaebae.org; searched on 11 April 2013.

Hackett JD, Yoon H-S, Soares MB, Bonaldo MF, Casavant TL, et al. 2007. Phylogenomic analysis supports the monophyly of cryptophytes and haptophytes and the association of Rhizaria with chromalveolates. Mol Biol Evol. 24:1702-1713.

Hagopian JC, Reis M, Kitajima JP, Bhattacharya D, de Oliveira MC. 2004. Comparative analysis of the complete plastid genome sequence of the red alga *Gracilaria tenuistipitata* var. *liui* provides insights into the evolution of rhodoplasts and their relationship to other plastids. J Mol Evol. 59:464-477.

Hoef-Emden K, Marin B, Melkonian M. 2002. Nuclear and nucleomorph SSU rDNA phylogeny in the Cryptophyta and the evolution of cryptophyte diversity. J Mol Biol. 55:161-179.

Hoef-Emden K, Tran HD, Melkonian M. 2005. Lineage-specific variations of congruent evolution among DNA sequences from three genomes, and relaxed selective constraints on *rbcL* in *Cryptomonas* (Cryptophyceae). BMC Evol Biol. 5:56.

Hoef-Emden K. 2008. Molecular phylogeny of the phycocyanin-containing cryptophytes: evolution of biliproteins and geographical distribution. J Phycol. 44:985-993.

Ichiyanagi K, Beauregard A, Belfort M. 2003. A bacterial group II intron favours retrotransposition into plasmid targets. Proc Natl Acad Sci U S A. 100:15742-15747.

Ishida K, Endo H, Koike S. 2011. *Partenskyella glossopodia* (Chlorarachniophyceae) possesses a nucleomorph genome of approximately 1 Mbp. Phycol Res. 59:120-122.

Janouškovec J, Horak A, Obornik M, Lukes J, Keeling PJ. 2010. A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. Proc Natl Acad Sci U S A. 107:10949-10954.

Janouškovec J, Liu S-L, Martone P, Carré W, Leblanc C, et al. 2013. Evolution of red algal plastid genes: Ancient architectures, introns, horizontal gene transfer, and taxonomic utility of plastid markers. PLoS ONE. 8:e59001.

Jentsch S, McGrath JP, Varshavsky A. 1987. The yeast DNA repair gene RAD6 encodes a ubiquitin-conjugating enzyme. Nature. 329:131-134.

Katinka MD, Duprat S, Cornillot E, Méténier G, Thomarat F, et al. 2001. Genome sequence and gene compaction of the eukaryotic parasite *Encephalitozoon cuniculi*. Nature. 414:450-453.

Keeling PJ, Corradi N, Morrison HG, Haag KL, Ebert D, et al. 2010. The reduced genome of the parasitic microsporidian *Enterocytozoon bieneusi* lacks genes for core carbon metabolism. Genome Biol Evol. 2:304-309.

Khan H, Parks N, Kozera C, Curtis BA, Parsons BJ, et al. 2007. Plastid genome sequence of the cryptophyte alga *Rhodomonas salina* CCMP1319: lateral transfer of putative DNA replication machinery and a test of chromist phylogeny. Mol Biol Evol. 24:1832-1842.

Khan H, Archibald JM. 2008. Lateral transfer of introns in the cryptophyte plastid genome. Nucleic Acids Res. 36:3043-3053.

Koonin EV. 2006. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? Biol Direct. 1:12.

Kowallik KV, Stoebe B, Schaffran I, Kroth-Pancic P, Freier U. 1995. The chloroplast genome of a chlorophyll c-containing alga, *Odontella sinensis*. Plant Mol Biol Rep. 13:336-342.

Lambowitz AM, Zimmerly S. 2004. Mobile group II introns. Annu Rev Genet. 38:1-35.

Lane CE, Khan H, MacKinnon M, Fong A, Theophilou S, et al. 2005. Insight into the diversity and evolution of the cryptomonad nucleomorph genome. Mol Biol Evol. 23:856-865.

Lane CE, Archibald JM. 2006. Novel nucleomorph genome architecture in the cryptomonad genus *Hemiselmis*. J Eukaryot Microbiol. 53:515-521.

Lane CE, Khan H, MacKinnon M, Fong A, Theophilou S, et al. 2006. Insight into the diversity and evolution of the cryptomonad nucleomorph genome. Mol Biol Evol. 23:856-865.

Lane CE, van den Heuvel K, Kozera C, Curtis BA, Parsons BJ, et al. 2007. Nucleomorph genome of *Hemiselmis andersenii* reveals complete intron loss and compaction as a driver of protein structure and function. Proc Natl Acad Sci U S A. 104:19908-19913.

Lewis LA, McCourt RM. 2004. Green algae and the origin of land plants. Am J Bot. 91:1535-1556.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25:1754-1760.

Lim L, McFadden GI. 2010. The evolution, metabolism and functions of the apicoplast. Proc R Soc B. 365:749-763.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25:955-964.

Matsuzaki M, Misumi O, Shin-i T, Maruyama S, Takahara M, et al. 2004. Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. Nature. 428:653-657.

McFadden GI. 1999. Plastids and protein targeting. J Euk Microbiol. 46:339-346.

Minge MA, Shalchian-Tabrizi K, Tørresen OK, Takishita K, Probert I, et al. 2010. A phylogenetic mosaic plastid proteome and unusual plastid-targeting signals in the green-coloured dinoflagellate *Lepidodinium chlorophorum*. MBC Evol Biol. 10:168.

Moore RB, Obornik M, Janouškovec J, Chrudimký T, Vancová M, et al. 2008. A photosynthetic alveolate closely related to apicomplexan parasites. Nature. 451:959-963.

Moore CE, Archibald JM. 2009. Nucleomorph genomes. Annu Rev Genet. 43:251-264.

Nassoury N, Cappadocia M, Morse D. 2003. Plastid ultrastructure defines the protein import pathway in dinoflagellates. J Cell Sci. 116:2867-2874.

Nikolaev SI, Berney C, Fahrni JF, Bolivar I, Polet S, et al. 2004. The twilight of Heliozoa and the rise of Rhizaria, an emerging supergroup of amoeboid eukaryotes. Proc Natl Acad Sci U S A. 101:8066-8071.

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

Okamoto N, Chantangsi C, Horak A, Leander BS, Keeling PJ. 2009. Molecular phylogeny and description of the novel katablepharid *Roombia truncata* gen. et sp. nov., and establishment of the Habrobian taxon nov. PLoS ONE. 4:e7080.

Oren A. 2005. A hundred years of *Duniella* research: 1905-2005. Saline Systems. 1:2.

Oudot-Le Secq MP, Grimwood J, Shapiro H, Armbrust EV, Bowler C, et al. 2007. Chloroplast genomes of the diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*: comparison with other plastid genomes of the red lineage. Mol Genet Genomics. 277:427-439.

Patron NJ, Rogers, MB, Keeling PJ. 2006. Comparative rates of evolution in endosymbiotic nuclear genomes. BMC Evol Biol. 6:46.

Patron NJ, Inagaki Y, Keeling PJ. 2007. Multiple gene phylogenies support the monophyly of cryptomonad and haptophyte host lineages. Curr Biol. 17:887-891.

Phipps KD, Donaher NA, Lane CE, Archibald JM. 2008. Nucleomorph karyotype diversity in the freshwater cryptophyte genus *Cryptomonas*. J Phycol. 44:11-14.

Reith M, Munholland J. 1995. Complete nucleotide sequence of the Porphyra purpurea chloroplast genome. Plant Mol Biol Reporter. 13:333-335.

Reyes-Prieto A, Hackett JD, Soares MB, Bonaldo MF, Bhattacharya D. 2006. Cyanobacterial contribution to algal nuclear genomes is primarily limited to plastid functions. Curr Biol. 16:2320-2325.

Reyes-Prieto A, Weber APM, Bhattacharya D. 2007. The origin and establishment of the plastid in algae and plants. Annu Rev Genet. 41:147-168.

Rice DW, Palmer JD. 2006. An exceptional horizontal gene transfer in plastids: gene replacement by a distance bacterial paralog and evidence that haptophyte and cryptophyte plastids are sisters. BMC Biol. 4:31.

Robart AR, Zimmerly S. 2005. Group II intron retroelements: function and diversity. Cytogenet Genome Res. 110:589-597.

Rogers, MB, Gilson PR, Su V, McFadden GI, Keeling PJ. 2007. The complete chloroplast genome of the chlorarachniophyte *Bigelowiella natans*: evidence for independent origins of chlorarachniophyte and euglenid secondary endosymbionts. Mol Biol Evol. 24:54-62.

Rogozin IB, Carmel L, Scuros M, Koonin EV. 2012. Origin and evolution of spliceosomal introns. Biol Direct. 7:11.

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, et al. 2000. Bioinformatics. 16:944-945.

Sánchez Puerta MV, Bachvaroff TR, Delwiche CF. 2005. The complete plastid genome sequence of the haptophyte *Emiliania huxleyi*: a comparison to other plastid genomes. DNA Res. 12:151-156.

Shalchian-Tabrizi K, Minge AM, Cavalier-Smith T, Nedreklepp JM, Klaveness D et al. 2006. Combined heat shock protein 90 and ribosomal RNA sequence phylogeny supports multiple replacements of dinoflagellate plastids. J Eukaryot Microbiol. 53:217-224.

Shalchian-Tabrizi K, 2008. Diversification of unicellular eukaryotes: cryptomonad colonizations of marine and fresh waters inferred from revised 18S rRNA phylogeny. Environ Microbiol. 10:2635-2644.

Schnepf E, Elbrachter M. 1988. Cryptophycean-like double membrane-bound chloroplast in the dinoflagellate *Dinophysis Ehrenb*. – Evolutionary, Phylogenetic and Toxicological Implications. Botanica Acta. 101:196-203.

Silver TD, Koike S, Yabuki A, Kofuji R, Archibald JM, et al. 2007. Phylogeny and nucleomorph karyotype diversity of chlorarachniophyte algae. J Eukaryot Microbiol. 54:403-410.

Slamovits CH, Fast NM, Law JS, Keeling PJ. 2004. Genome compaction and stability in microsporidian intracellular parasites. Curr Biol. 14:891-896.

Slamovits CH, Keeling PJ. 2009. Evolution of ultra-small spliceosomal introns in highly reduced nuclear genomes. Mol Biol Evol. 26:1699-1705.

Smith DR, Crosby K, Lee RW. 2011. Correlation between nuclear plastid DNA abundance and plastid number supports the limited transfer window hypothesis. Genome Biol Evol. 3:365-371.

Smith DR, Hua J, Lee RW, Keeling PJ. 2012. Relative rates of evolution among the three genetic compartments of the red alga Porphyra differ from those of green plants and do not correlate with genome architecture. Mol Phylogenet Evol. 65:464-477.

Soll J, Schleiff E. 2004. Protein import into chloroplasts. Nat Rev Mol Cell Biol. 5:198-208.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22:2688-2690.

Stirewalt VL, Michalowski CB, Loffelhardt W, Bohnert HJ, Bryant DA. 1995. Nucleotide sequence of the cyanelle genome from *Cyanophora paradoxa*. Plant Mol Biol Rep. 13:327-332.

Sulli C, Fang Z, Muchhal U, Schwartzbach SD. 1999. Topology of *Euglena* chloroplast protein precursors within endoplasmic reticulum to Golgi to chloroplast transport vesicles. J Biol Chem. 274:457-463.

Takahashi F, Okabe Y, Nakada T, Sekimoto H, Ito M, et al. 2007. Origins of the secondary plastids of Euglenophyta and Chlorarachniophyta as revealed by an analysis of the plastid-targeting, nuclear-encoded gene *psbO*. J Phycol. 43:1302-1309.

Tanifuji G, Onodera NT, Hara Y. 2010. Nucleomorph genome diversity and its phylogenetic implications in cryptomonad algae. Phycol Res. 58:230-237.

Tanifuji G, Onodera NT, Wheeler TJ, Dlutek M, Donaher N, et al. 2011. Complete nucleomorph genome sequences of the nonphotosynthetic alga *Cryptomonas paramecium* reveals a core nucleomorph gene set. Genome Biol Evol. 3:44-54.

Tengs T, Dahlberg OJ, Shalchian-Tabrizi K, Klaveness D, Rudi K, et al. 2000. Phylogenetic analyses indicate that the 19’ hexanoyloxy-fucoxanthin-containing dinoflagellates have tertiary plastids of haptophyte origin. Mol Biol Evol. 17:718-729.

Watanabe MM, Suda S, Inouye I, Sawaguchi T, Chihara M. 1990. *Lepidodinium viride* gen. et sp. nov. (Gymnodiniales, Dinophyta), a green dinoflagellate with a chlorophyll A- and B-containing endosymbiont. J Phycol. 26:741-751.

Williams BAP, Slamovits C, Patron NJ, Fast NM, Keeling PJ. 2005. A high frequency of overlapping gene expression in compacted eukaryotic genomes. Proc Natl Acad Sci U S A. 102:10936-10941.

Wilson RJM, Denny PW, Preiser PR, Rangachari K, Roberts K, et al. 1996. Complete gene map of the plastid-like DNA of the malaria parasite *Plasmodium falciparum*. J Mol Biol. 261:155-172.

Yoon H-S, Hackett JD, Pinto G, Bhattacharya D. 2002. The single, ancient origin of Chromist plastids. Proc Natl Acad Sci U S A. 99:15507-15512.

Yoon H-S, Müller KM, Sheath RG, Ott FD, Bhattacharya D. 2006. Defining the major lineages of red algae (Rhodophyta). J Phycol. 42:482-492.

Zauner S, Fraunholz M, Wastl J, Penny S, Beaton M, et al. 2000. Chloroplast protein and centrosomal genes, a tRNA intron, and odd telomeres in an unusually compact eukaryotic genome, the cryptomonad nucleomorph. Proc Natl Acad Sci U S A. 97:200-205.

APPENDIX A   LIST OF PRIMERS USED FOR PROBES

The following primers were used to make probes for the Southern blot hybridizations described in Chapter 2. NM = nucleomorph, HN = host nuclear, MT = mitochondrial. An asterisk indicates universal primers. All other primers are specific to *C. mesostigmatica*.

Gene	Forward Primer Sequence	Reverse Primer Sequence
cycB	CCCATCCGTTTTGTCATTTT	GGCAAAGGGCAGCATATTT
gidA	CCCAAATACGGGGGTTTTAT	CAGGCCTTACCATCTTGGAA
kin(mps1)	GCGAGTTCGAATTTTACCACA	TGATGATGATGATGGATTTGG
kin(snf2)	AAATGATTGGCGGTCTTGAA	TTTCCCATCGAATGGCTTAG
mcm7	TCCGACCGTATCGTTCTTTT	AGCACCGAGCAAAGAGATTG
rpb2	TAAGGAATGCCCTTTTGACG	CCAAATTCTGCAACGGATCT
rpoD	TAGGCTTAATCCGTGCTGCT	TCTCTTTCTCTGGGGCTCAA
smc1	TCGAGAATGATGACCAACCA	GGGAGATCGTCCGAGTGAAA
smc2	CGGCATTTCAAGCTATTTTTG	AAGCAGCATCGATTTCATCC
sut	TTCGGAATTTCCATCCAAAA	GCCATTGGAAAAGCACTCAT
tcpA	AGAATAGCCATGGCAACAGG	ATTCAGTTCCTCCACGCAAA
tfIIB-brf	CATTTGTTGGTCGATTTTTCG	CATCATCTCCTCCTCCTCCA
repeat	CCTATCCATGCCACAAGAGG	TTAAACGGAGGATGCTTTCG
NM 18S	TGTAGAGATGACGATGCTG	GCCCTTTCGGCCTGCCATG
HN 18S	TGTCGGGGTCGGGAGGACTG	AGCAAAAGCCTTCTTTGAACG
MT _cox1_*	TCAACAAATCATAAAGATATTGG	ACTTCTGGATGTCCAAAAAAYCA
PL 16S*	GGCTCAGGATGAACGCTGGC	CCTCACGCGGTATTGCTCCG

APPENDIX B CONSERVED CRYPTOPHYTE NUCLEOMORPH GENES

Presence/absence of conserved cryptophyte nucleomorph protein-coding genes in *C. mesostigmatica* (Cm), *H. andersenii* (Ha), *C. paramecium* (Cp), and *G. theta* (Gt), organized by functional category. Numbers indicate gene copy number.

Functional Category	Gene	Nucleomorph Genome			
		Cm	Ha	Cp	Gt
Translation	*cbp*	0	0	0	1
	ef2	1	1	1	1
	eif1A	1	1	1	1
	eif2B	1	1	1	1
	eif2G	1	1	1	1
	eif4A	1	1	1	1
	eif4E	1	1	1	1
	eif5A	1	1	1	1
	eif6	1	1	1	1
	erf1	1	1	1	1
	ncbP2	0	1	0	1
	rla0	1	1	1	1
	rla1	1	1	1	1
	rpl1	1	1	1	1
	rpl10	1	1	1	1
	rpl10A	1	1	1	1
	rpl11B	1	1	1	1
	rpl12	1	1	1	1
	rpl13	1	1	1	1
	rpl13A	1	1	1	1
	rpl14	1	1	1	1
	rpl15	1	1	1	1
	rpl17	1	1	1	1
	rpl18	1	1	1	1
	rpl18A	1	1	1	1
	rpl19	1	1	1	1
	rpl21	1	1	1	1
	rpl23	1	1	1	1
	rpl23A	1	1	1	1
	rpl24	1	1	1	1
	rpl26	1	1	1	1
	rpl27	1	1	1	1
	rpl27A	1	1	1	1
	rpl3	1	1	1	1
	rpl30	1	1	1	1

Functional	Gene	Nucleomorph Genome			
Category		Cm	Ha	Cp	Gt
	rpl31	1	1	1	1
	rpl32	1	1	1	1
	rpl34	1	1	1	1
	rpl35A	1	1	1	1
	rpl36	1	1	1	1
	rpl37A	1	1	1	1
	rpl40	1	1	2	1
	rpl5	1	1	1	1
	rpl6B	1	1	1	1
	rpl7	1	1	1	1
	rpl7A	1	1	1	1
	rpl8	1	1	1	1
	rpl9	1	2	1	1
	rps10B	1	1	1	1
	rps11	1	1	1	1
	rps13	1	1	1	1
	rps14	1	1	1	1
	rps15	1	1	1	1
	rps15A	1	1	1	1
	rps16	1	1	1	1
	rps17	1	1	1	1
	rps19	1	1	1	1
	rps2	1	1	1	1
	rps20	1	1	1	1
	rps21	1	1	0	1
	rps23	1	1	1	1
	rps24	1	1	1	1
	rps25	1	1	1	1
	rps26	1	1	1	1
	rps27	1	1	1	1
	rps27A	1	1	1	1
	rps28	1	1	1	1
	rps29A	0	1	1	1
	rps3	1	1	1	1
	rps30	2	1	1	1
	rps3A	1	1	1	1
	rps4	1	1	1	1
	rps5	1	1	1	1
	rps6	1	1	1	1
	rps8	1	1	1	1
	rps9	1	1	1	1
	rsp4	1	1	1	1
	sui1	1	1	1	1
	sys1	1	1	1	1

Functional Category	Gene	Nucleomorph Genome			
		Cm	Ha	Cp	Gt
	tif211	1	1	1	1
	Yrpl24	1	1	0	1
Transcription	*asf*	1	1	1	1
	dbx-like	1	1	1	0
	fet5	1	1	1	1
	hira	1	1	0	1
	hsf	0	1	1	1
	pop2	1	1	1	1
	rad25	1	1	1	1
	rad3	1	1	1	1
	reb1	1	1	1	1
	rpa1	1	1	1	1
	rpa2	1	1	1	1
	rpa5	1	1	1	1
	rpabc5	1	1	1	1
	rpabc6	1	1	1	1
	rpb1	1	1	1	1
	rpb10	1	1	1	1
	rpb11	1	1	1	0
	rpb2	1	1	1	1
	rpb3	1	1	1	1
	rpb4	1	1	1	1
	rpb7	1	1	1	1
	rpb8	1	1	1	1
	rpbY	0	0	0	1
	rpc1	1	1	1	1
	rpc10	1	1	1	1
	rpc2	1	1	1	1
	rpc9	1	1	1	1
	ruvB-like1	1	1	1	0
	ruvB-like2	1	1	1	0
	taf	1	1	1	0
	taf13	1	1	1	1
	taf30	1	1	1	1
	taf90	1	1	1	1
	tfIIA-S	2	1	1	1
	tfIIB	1	1	1	1
	tfIIB-brf	6	1	1	1
	tfIIB-like	1	1	1	1
	tfIID	1	1	1	3
	tfIIE	1	1	1	1
	tfIIC	1	1	1	1
	trf	1	1	1	1

Functional	Gene	Nucleomorph Genome			
Category		Cm	Ha	Cp	Gt
Protein folding and degradation	*der1*	1	1	1	1
	hsp70	1	1	1	1
	hsp90	1	1	1	1
	prsA1	0	1	1	1
	prsA2	0	1	1	1
	prsA3	0	1	1	1
	prsA5	0	1	1	1
	prsA6	0	1	1	1
	prsA7	0	1	1	1
	prsB1	0	1	1	1
	prsB3	0	1	1	1
	prsB4	0	1	1	1
	prsB5	0	1	1	1
	prsB6	0	1	1	1
	prsB7	0	1	1	1
	prsS1	0	1	1	1
	prsS10B	0	1	1	1
	prsS12	0	1	1	1
	prsS13	0	1	1	1
	prsS4	0	1	1	1
	prsS6A	0	1	1	1
	prsS6B	0	1	1	1
	prsS7	0	1	1	1
	prsS8	0	1	1	1
	rbp1	0	1	1	1
	tcpA	1	1	1	1
	tcpB	1	1	1	1
	tcpD	1	1	1	1
	tcpE	1	1	1	1
	tcpG	1	1	1	1
	tcpH	1	1	1	1
	tcpT	1	1	1	1
	tcpZ	1	1	1	1
	ubc2	1	1	0	1
	ubc4	0	3	4	5
	ubiquitin	2	1	0	0
	uceE2	1	1	1	1
	ufd	1	1	1	1
Mitosis	*cdc5-like*	1	0	0	0
	cenp-A	1	1	1	1
	ranbpm	0	1	0	1
	tubA	1	1	1	1
	tubB	1	1	1	1
	tubG	1	1	1	1

Functional Category	Gene	Nucleomorph Genome			
		Cm	Ha	Cp	Gt
DNA metabolism and cell cycle control	*BRSK*	1	1	3	1
	cdc2	1	1	0	1
	cdc48a	1	1	1	1
	cdc48b	1	1	1	1
	crm	1	1	0	1
	cycB	1	1	1	1
	dph1	1	1	1	1
	ebi	1	1	1	1
	h2B	1	1	1	1
	h3	1	1	1	1
	h4	1	1	1	1
	hat	0	1	0	1
	hda	1	1	1	1
	kin(aaB)	1	1	1	1
	kin(cdc)	0	1	1	0
	kin(cdc2)	1	1	1	1
	kin(gs)	0	1	1	1
	kin(mps1)	1	1	0	1
	kin(snf1)	1	1	1	1
	mcm2	1	1	1	1
	mcm3	1	1	1	1
	mcm4	1	1	1	0
	mcm5	1	1	1	1
	mcm6	1	1	1	1
	mcm7	1	1	1	1
	mcm8	1	1	0	1
	mcm9	1	1	0	1
	pcna	1	1	1	1
	pi4K	1	1	1	1
	pp1	0	1	1	1
	rad51	1	1	1	1
	rfc2	1	1	1	1
	rfc3	1	1	1	0
	smc1	1	1	1	1
	smc2	1	1	0	0
	smc3	1	1	1	1
	smc4	1	1	0	0
	ste4	0	0	0	1
	trithorax-like	1	0	0	0
RNA metabolism	*ATPbp*	1	1	1	1
	ATP/GTP-bp	1	1	1	1
	brx1	1	1	1	0

Functional	Gene	Nucleomorph Genome			
Category		Cm	Ha	Cp	Gt
	bsm1-like	1	1	1	1
	cbf5	1	1	1	1
	cdc28	1	1	1	1
	dbp4	6	1	1	1
	dhm	1	1	1	1
	dib1	1	0	1	0
	fcf1	1	1	1	1
	G10	1	0	1	1
	gblp1	1	1	1	1
	gblp2	1	1	1	1
	gsp2	1	1	1	1
	GTP-bp	1	1	1	1
	has1	1	1	1	1
	imb1	1	1	1	1
	imp4	1	1	1	1
	impA	1	1	1	1
	mak16	1	1	1	1
	mce	1	1	1	1
	mrs2	1	1	1	1
	nip7	1	1	1	1
	nog1	1	1	1	1
	nop1	1	1	1	1
	nop2	1	1	1	1
	nop5	1	1	1	1
	nop56	1	1	1	1
	nop-like	1	1	1	0
	pab1	1	1	1	1
	pab2	1	1	1	1
	prl1-like	1	1	1	1
	prp2-like	1	0	0	0
	prp22-like	1	0	0	0
	prp4-like	1	0	1	0
	prp8	1	0	1	1
	rcl1	1	1	1	1
	rpf1	1	1	1	1
	rrp3	1	1	1	1
	sbp1	1	1	1	1
	sen1	1	1	1	1
	sen2	1	1	1	1
	sen34	1	1	1	1
	sf3b1-like	1	0	0	0
	sf3b3-like	1	0	0	0
	ski2	0	0	0	1
	snrpB	1	0	0	0

Functional	Gene	Nucleomorph Genome			
Category		Cm	Ha	Cp	Gt
	snrpD	1	1	1	1
	snrpD1	1	0	1	1
	snrpD2	1	1	1	1
	snrpD3	1	0	1	1
	snrpE	1	0	1	1
	snrpE-like	0	0	1	0
	snrpF	6	0	0	1
	snrpG	1	0	1	1
	snu13	1	1	1	1
	sof1	1	1	1	1
	ste13	0	0	1	1
	U3snoRNP	1	1	1	0
	U5snRNP (40kDa)	1	0	1	0
	U5snRNP (116kDa)	1	0	1	1
	U5snRNP (200kDa)	1	0	0	1
Plastid-associated	*cbbX*	1	1	1	1
	clpP1	1	1	1	1
	clpP2	1	1	1	1
	cpeT-like	2	1	0	1
	cpn60	1	1	1	1
	dnaG	1	1	1	1
	eng	1	1	1	1
	ftsZ	1	1	1	1
	gidA	1	1	1	1
	gidB	1	1	1	1
	orf152	1	1	1	1
	orf177	1	1	0	1
	orf204	1	1	1	1
	orf238	1	1	0	1
	orf243	1	1	0	1
	orf268	1	1	0	1
	orf336	1	1	0	1
	orf826	1	1	1	1
	gyrA	1	1	0	1
	gyrB	1	1	0	1
	hfc136	1	1	0	1
	hlip	1	1	0	1
	iap100	1	1	1	1
	met	1	1	0	1
	rpoD	1	1	1	1
	rps15	1	1	1	1

Functional	Gene	Nucleomorph Genome			
Category		Cm	Ha	Cp	Gt
	rub	1	1	0	1
	secE	1	1	1	1
	sufD	1	1	1	1
	tha4	1	1	0	1
	tic22	1	1	1	1
Miscellaneous	*bystin-like*	1	1	1	1
	fkbp	1	1	0	1
	fkbp-like	1	1	0	0
	ggt	1	1	1	1
	kea1	1	1	1	1
	kin(ABC)	1	1	1	0
	nat10	1	1	1	1
	nmt1	1	1	1	1
	nol10	1	1	1	1
	rip1	1	1	1	1
	rli1	1	1	1	1
	sut	1	1	0	1
	tbl3	1	1	1	1

APPENDIX C COPYRIGHT PERMISSION

Figure 1.2 in Chapter 1 and Figure 2.11 in Chapter 2 were originally published in the following manuscript:

Moore CE, Archibald JM. 2009. Nucleomorph genomes. Annu Rev Genet. 43:251-264.

As an author I have the right to use any or all material from this paper in my thesis without acquiring copyright permission, as described in Annual Reviews' publication rights policies, found here: http://www.annualreviews.org/page/authors/author-instructions/distributing/copyright_mandate

The relevant policy states:

"What rights do I enjoy as an author?
After manuscript acceptance and copyright transfer, substantial rights are granted to the author(s) by Annual Review:
2. The nonexclusive right to use, reproduce, distribute, perform, update, create derivatives, and make copies of the work (electronically or in print) in connection with the author's teaching, conference presentations, lectures, and publications, provided proper attribution is given
8. The right to include the work, whole or in part, in a dissertation or thesis.

Must I obtain permission from Annual Reviews to reuse the material in my review?
No."

Most of the work described in Chapter 2 was published in the following manuscript:

Moore CE, Curtis BA, Mills T, Tanifuji G, Archibald JM. 2012. Nucleomorph genome sequence of the cryptophyte alga *Chroomonas mesostigmatica* CCMP1168 reveals lineage-specific gene loss and genome complexity. Genome Biol Evol 4:1162-1175.

An an author, I have the right to use any or all material from this paper in my thesis without acquiring copyright permission, as described in Oxford Journals' publication rights policies, found here: http://www.oxfordjournals.org/access_purchase/publication_rights.html

The relevant policy states:

"Rights retained by ALL Oxford Journal Authors

- The right, after publication by Oxford Journals, to use all or part of the Article and Abstract for their own personal use, including their own classroom teaching purposes;
- The right, after publication by Oxford Journals, to use all or part of the Article and Abstract in the preparation of derivative works, extensions of the article into book-length or in other works, provided that a full acknowledgement is made to the original publication in the journal;
- The right to include the article in full or in part in a thesis or dissertation, provided that it is not published commercially;

For the uses specified here, please note that there is no need for you to apply for written permission from Oxford University Press in advance. Please go ahead with the use ensuring that a full acknowledgement is made to the original source of the material including the journal name, volume, issue, page numbers, year of publication, title of article, and to Oxford University Press and/or the learned society."