ENDOSYMBIOTIC GENE TRANSFER IN THE NUCLEOMORPH CONTAINING ORGANISMS BIGELOWIELLA NATANS AND GUILLARDIA THETA

by

Bruce A. Curtis

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

at

Dalhousie University
Halifax, Nova Scotia
October 2012

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 "ENDOSYMBIOTIC GENE TRANSFER IN THE NUCLEOMORPH CONTAINING ORGANISMS *BIGELOWIELLA NATANS* AND *GUILLARDIA THETA*" by Bruce A. Curtis in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

Dated: October 22, 2012

External Examiner: ___________________________________

Research Supervisor: ___________________________________

Examining Committee: ___________________________________

____

____

Departmental Representative: _____________________________

DALHOUSIE UNIVERSITY

DATE: October 22, 2012

AUTHOR: Bruce A. Curtis

TITLE: ENDOSYMBIOTIC GENE TRANSFER IN THE NUCLEOMORPH CONTAINING ORGANISMS BIGELOWIELLA NATANS AND GUILLARDIA THETA

DEPARTMENT OR SCHOOL: Department of Biochemistry and Molecular Biology

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

____
Signature of Author

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Mitochondria and chloroplasts are eukaryotic organelles that were acquired through endosymbiosis. In the case of the mitochondrion, a heterotrophic cell engulfed and retained an alpha-proteobacterium. The engulfed bacterium, or endosymbiont, underwent extensive cellular and genetic integration with its host, thereby becoming an organelle. Chloroplasts are derived from the engulfment and retention of a photosynthetic cyanobacterium that also experienced loss of cellular functions and genetic material. Although mitochondria and chloroplasts retain their own genomes, most of the proteins that function in these organelles are encoded by genes that were transferred to the nucleus in a process known as Endosymbiotic Gene Transfer (EGT).

Chloroplasts in plants, green algae and red algae are known as primary plastids. Other photosynthetic organisms have secondary plastids that were acquired by engulfing and retaining a primary plastid-bearing alga. In the process, the nucleus of the engulfed alga underwent EGT (and presumably gene loss) to such an extent that it disappeared completely except in two lineages, cryptophytes and chlorarachniophytes, which retain a highly reduced and miniaturized form known as a nucleomorph.

To understand the process of EGT and endosymbiosis in general, the nuclear genomes of the cryptophyte *Guillardia theta* and the chlorarachniophyte *Bigelowiella natans* were sequenced. In the case of *G. theta* its nucleomorph is of red algal origin while the nucleomorph of *B. natans* is derived from a green algal endosymbiont.

Prior to the nuclear genome projects the genomes of the three organelles – plastid, mitochondrion, nucleomorph – had already been sequenced. This allowed investigation of recent transfers of organellar DNA to the nucleus. Mitochondrial transfers to the nucleus are still occurring in both organisms but transfers of plastid and nucleomorph DNA are not. The nucleomorph genomes of *B. natans* and *G. theta* appear 'frozen', unable to undergo EGT and thus unable to disappear as they have in all other lineages with secondary plastids. The creation of a spliceosomal intron from transferred organellar DNA was investigated.

I also investigated nuclear genes whose encoded proteins appear to function in the mitochondrion. 833 putatively mitochondrial targeted proteins were identified in *G. theta* and 720 in *B. natans*.

LIST OF ABBREVIATIONS USED

ATOM	archaic translocase of the outer membrane
EGT	endosymbiotic gene transfer
ER	endoplasmic reticulum
EST	expressed sequence tags
HGT	horizontal gene transfer
IMS	intermembrane mitochondrial space
IMP	inner membrane peptidase
ITS	IMS targeting signal
JGI	Joint Genome Institute
Kb	kilobases
KOG	eukaryotic orthologous groups
LGT	lateral gene transfer
MCP	mitochondrial carrier proteins
MIA	mitochondrial intermembrane space import and assembly
MIP	matrix intermediate peptidase
MPP	mitochondrial processing peptidase
MRO	mitochondrial related organelles
NCGR	National Centre for Genomics Research
NM	nucleomorph
NMD	nonsense mediated decay
NUMT	nuclear mitochondrial DNA
NUNM	nuclear nuclemorph DNA
NUPT	nuclear plastid DNA
ORF	open reading frame
OTU	operational taxonomic unit
PAM	presequence assisted motor
PPC	periplastic compartment
PreP	presequence protease

PTC	premature termination codons
SAR	stramenopiles, alveolates, rhizarians
SVM	support vector machine
TAR	tandem amino acid repeat
TOM	translocase of the outer membrane
TPR	tetratricopeptide repeat

ACKNOWLEDGEMENTS

There are many people I would like to thank for their help and guidance over the last few years. First and foremost I would like to thank my supervisor John Archibald for giving me the opportunity to work on such exciting and overwhelming projects. There has truly been an embarrassment of riches to mine from the nuclear genome projects of *Guillardia theta* and *Bigelowiella natans* and I am sure this document won't be the last time I write something about these organisms. John has been an amazing supervisor. He is generous with his time and expertise and has fostered a relaxed and cooperative lab environment that makes research a pleasure. His skill in shepherding the genome projects to a successful conclusion has been remarkable and I am thankful for the chance to have witnessed it and participated.

Of course the projects would not have achieved anything without a large, dedicated, and deeply knowledgeable team and I would like to list their names here because all of my research truly depended on a collective effort to understand these organisms - Goro Tanifuji, Fabien Burki, Ansgar Gruber, Manuel Irimia, Shinichiro Maruyama, Maria C. Arias, Steven G. Ball, Gillian H. Gile, Yoshihisa Hirakawa, Julia F. Hopkins, Alan Kuo, Stefan A. Rensing, Jeremy Schmutz, Aikaterini Symeonidi, Marek Elias, Robert J. M. Eveleigh, Emily K. Herman, Mary J. Klute, Takuro Nakayama, Miroslav Oborník, Adrian Reyes-Prieto, E. Virginia Armbrust, Stephen J. Aves, Robert G. Beiko, Pedro Coutinho, Joel B. Dacks, Dion G. Durnford, Naomi M. Fast, Beverley R. Green, Cameron Grisdale, Franziska Hempel, Bernard Henrissat, Marc P. Höppner, Ken-Ichiro Ishida, Eunsoo Kim, Ludek Korený, Peter G. Kroth, Yuan Liu, Shehre-Banoo Malik,

Uwe G. Maier, Darcy McRose, Thomas Mock, Jonathon A. D. Neilson, Naoko T. Onodera, Anthony M. Poole, Ellen J. Pritham, Thomas A. Richards, Gabrielle Rocap, Scott W. Roy, Chihiro Sarai, Sarah Schaack, Shu Shirato, Claudio H. Slamovits, David F. Spencer, Shigekatsu Suzuki, Alexandra Z. Worden, Stefan Zauner, Kerrie Barry, Callum Bell, Arvind K. Bharti, John A. Crow, Jane Grimwood, Robin Kramer, Erika Lindquist, Susan Lucas, Asaf Salamov, Geoffrey I. McFadden, Christopher E. Lane, Patrick J. Keeling, Michael W. Gray, Igor V. Grigoriev, John M. Archibald.

I would also like to thank the members of CGEB for providing a stimulating and rewarding place to do graduate studies. I have learned more in the Thursday morning lab seminars than I did in my four years of undergraduate work.

I would like to thank the members of my Supervisory Committee: John Archibald, Andrew Roger; Robert Beiko; and Joe Bielawski. They have provided valuable guidance and insight and were helpful as well in navigating the non-science aspects of being a graduate student.

CHAPTER 1 INTRODUCTION

Eukaryotic lineages have been fashioned by endosymbiosis (Martin, Dagan et al. 2007). Prior to the diversification of eukaryotes into the major groups, the engulfment and retention of a bacterium led to the establishment of the mitochondrion (Gray, Burger et al. 1999). Subsequently, a heterotrophic eukaryotic lineage acquired photosynthetic abilities through an endosymbiotic relationship with cyanobacteria (Figure 1.1) leading to what are known as the primary plastid bearing groups or the Archaeplastida, comprised of red algae, glaucophyte algae and green algae along with their terrestrial descendants (Palmer 2003). In both cases the endosymbiosis resulted in a massive rearrangement of the cellular machinery and processes along with the underlying genetic programming for both the endosymbiont and the host cell. The endosymbiotic bacteria lost most of the functions required of free-living organisms, becoming, in the process, organelles that were fully integrated and dependent upon the host. The reduction in the functional complexity of the endosymbiont was accompanied by an even more drastic reduction of its genetic complexity. Genes for proteins no longer required were lost while many of the genes that encoded proteins found in the organelles were transferred to the nuclear DNA of the host.

The influx of thousands of genes from the protomitochondrion, and in some lineages the protoplastid, had a significant impact on the host cell (Martin 2003; Timmis, Ayliffe et al. 2004). Apart from increasing the genome size and the gene complement by several hundred genes, as well as the genetic rearrangements that were needed to make the new

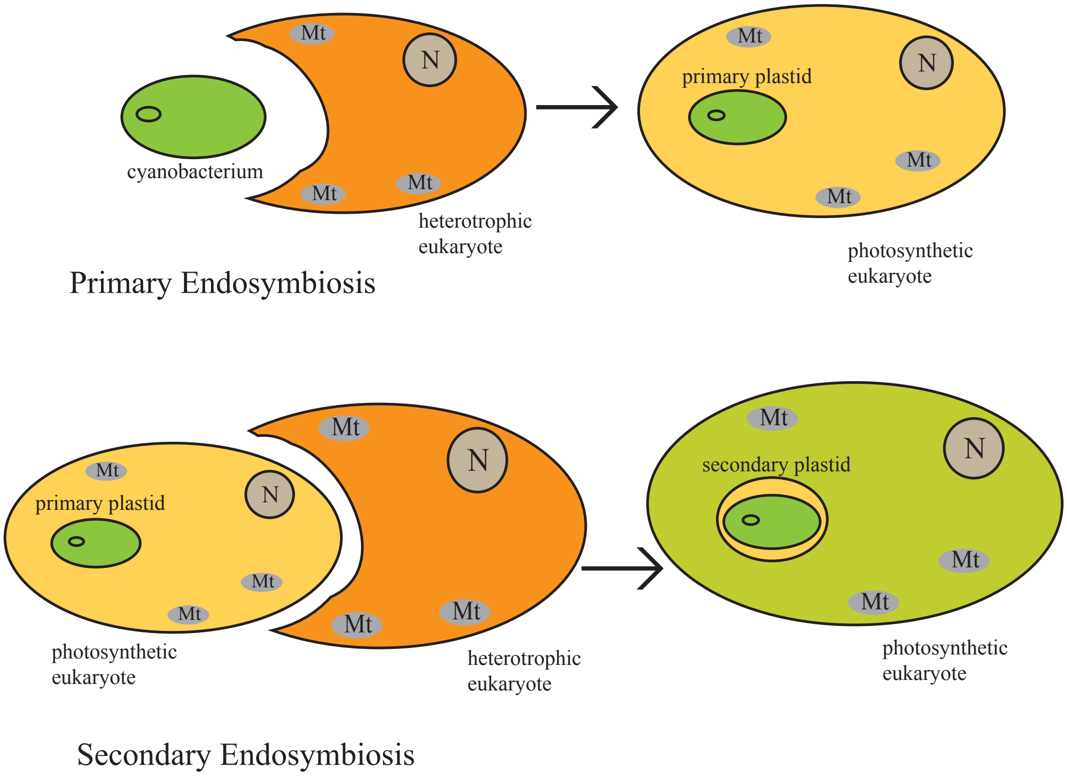


Figure 1.1. Primary and Secondary Endosymbiosis. Primary endosymbiosis involves the engulfment and retention of a cyanobacterium by a heterotrophic eukaryote with the cyanobacterium becoming the primary plastid. Secondary endosymbiosis involves the engulfment and retention of a photosynthetic eukaryote by a heterotrophic eukaryote with the reduced photosynthetic eukaryotic being a secondary plastid. Mt = mitochondrion, N =nucleus.

genes functional, the host cell now had new functions and new proteins. Many of the new functions and proteins were directly related to maintenance and operation of the organelles, but a portion were the result of co-option of endosymbiont-derived genes either to replace existing host ones or acquire new pathways. The developing relationship between the host and the endosymbiont was also a driving force for the “invention” of new proteins from existing ones or acquisition through lateral gene transfer

(LGT)/horizontal gene transfer (HGT) of genes from other lineages to solve the "problems" created by having organelles (Martin and Herrmann 1998; Martin 2003) such as how to import proteins as well as metabolites into the organellar compartments.

Understanding the process by which an endosymbiont becomes an organelle is key to understanding early eukaryotic evolution. To help elucidate aspects of that endosymbiotic transformation I was part of a research team that sequenced the nuclear genomes of two photosynthetic algae, *Guillardia theta* and *Bigelowiella natans*. Why those two organisms were chosen for sequencing requires delving further into the intricacies of endosymbiosis.

A significant portion of the eukaryotic photosynthetic organisms on Earth are not primary plastid lineages (Bhattacharya, Yoon et al. 2003). Instead, they acquired their plastids through further rounds of endosymbiosis. A heterotrophic, unicellular organism engulfed and retained a primary plastid-containing alga (Figure 1.1). As with the original rounds of endosymbiosis this secondary process involved the drastic reduction of the retained free-living cell to a dependent organelle. However, instead of prokaryotic endosymbionts a fully functional photosynthetic eukaryote cell was involved. This engulfed alga had already undergone the reduction of the cyanobacterium to a plastid with the requisite transfer of thousands of prokaryotic genes to the eukaryotic nuclear genome. Consequently, secondary endosymbiosis involved mainly the transfer of genetic material from one eukaryotic genome to another eukaryotic genome, although it should be understood that many of the genes that were transferred originated from the cyanobacterial genome (Bhattacharya, Yoon et al. 2003). This process of eukaryote – eukaryote transfer went to completion, in that the engulfed eukaryotic nuclear genome completely vanished, leaving behind the plastid and a vestigial cytoplasm.

It is believed that the secondary acquisition of plastids has occurred at least three times, once involving a red algal organism and twice involving a green algal endosymbiont. The single red algal engulfment is purported to have led to a large and diverse group of photosynthetic lineages allied with some nonphotosynthetic ones in a supergroup known as Chromalveolata (Cavalier-Smith 1998; Adl, Simpson et al. 2005). It includes stramenopiles, dinoflagellates, ciliates, apicomplexans, haptophytes, and cryptophytes. The creation of the Chromalveolata supergroup was, to a large extent, based on the demonstrably red algal derived plastid present in many of its members and the understandable view that secondary endosymbiosis is a difficult and infrequent phenomenon. The fact that some of the members of Chromalveolata did not have plastids, and yet were clearly related to others that did, was the impetus for much research to try to solidify and "prove" the chromalveolate hypothesis (Patron, Rogers et al. 2004; Tyler, Tripathy et al. 2006; Patron, Inagaki et al. 2007). Recently, however, the chromalveolate hypothesis is looking increasingly unlikely, and new scenarios for the acquisition of their plastids are being suggested (Sanchez-Puerta and Delwiche 2008), and along with it new higher level taxonomic classifications (Hackett, Yoon et al. 2007; Burki, Shalchian-Tabrizi et al. 2008; Hampl, Hug et al. 2009).

One of the lineages in play in the taxonomic restructuring is the supergroup Rhizaria, which includes one of the lineages involved in secondary endosymbiosis with green algae, the chlorarachniophytes. There is increasing evidence that the rhizarians are closely related to the stramenopiles and alveolates (ciliates, apicomplexians, dinoflagellates) and a new supergroup dubbed SAR has been put forward (Hackett, Yoon et al. 2007; Burki, Shalchian-Tabrizi et al. 2008; Hampl, Hug et al. 2009). The other case of secondary endosymbiosis involving a green algal cell is that of euglenids, which are not related at all to the other examples.

It should also be mentioned that other forms of endosymbiosis are known. Dinoflagellates seem particularly prone to variations on secondary endosymbiosis to create mosaic genomes of complex evolutionary histories (Yoon, Hackett et al. 2002; Yoon, Hackett et al. 2005). For example, some dinoflagellates have acquired photosynthesis through tertiary endosymbiosis wherein a secondarily photosynthetic organism, like a haptophyte, has been retained. Quaternary endosymbiosis, involving the engulfment of a tertiary plastid, is also known from dinoflagellates (Hackett, Anderson et al. 2004). Finally, there is the complicating factor of cryptic endosymbiosis. This is when a lineage that is believed to have once had a plastid loses it, only to acquire another plastid at a later date, creating conflicting phylogenetic signals since a gene of EGT origin may have come from the current endosymbiont or from the long lost cryptic endosymbiont (Moustafa, Beszteri et al. 2009).

As mentioned, endosymbiosis involving the engulfment and retention of a primary plastid bearing cell results in the complete disappearance, either through outright loss or transfer, of the eukaryotic genome of the endosymbiont. In two lineages however, the cryptophytes (Douglas, Zauner et al. 2001) and the chlorarachniophytes (Gilson, Su et al. 2006), the reduction of the endosymbiont's eukaryotic genome did not go to completion. Instead, a vestigial nuclear genome known as a nucleomorph remains behind. Eukaryotic genomes generally contain between 10 and 30 thousand genes. Nucleomorphs contain fewer than 500 (Douglas, Zauner et al. 2001; Gilson, Su et al. 2006; Lane, van den Heuvel et al. 2007; Tanifuji, Onodera et al. 2011) (Figure 1.2). Surprisingly, few of the remaining genes are devoted to plastid functions and are, instead, mainly housekeeping genes for the nucleomorph (Gilson and McFadden 2002; Archibald 2007).

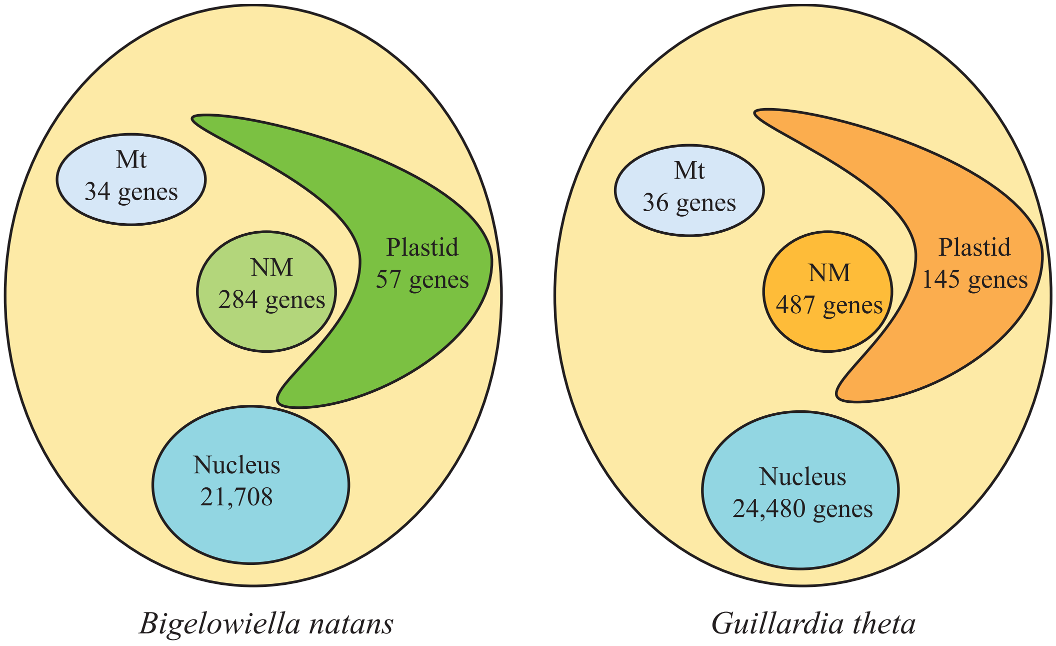


Figure 1.2. Protein coding genes found in the four genomes of *G. theta* and *B. natans*. Mt=mitochondrion, NM=nucleomorph.

The study of nucleomorphs provides us with a window into endosymbiosis since it captures a stage of the reductive process that presumably all secondarily photosynthetic lineages underwent. Consequently, several nucleomorphs have been sequenced, mainly from cryptophytes, such as *G. theta* (Douglas, Zauner et al. 2001), and one from the chlorarachniophytes, *Bigelowiella natans* (Gilson, Su et al. 2006). Undoubtedly there were lineage specific differences in what was lost and when, so by sequencing several nucleomorph genomes comparisons can be made and general trends detected. But having the sequence of the nucleomorph genome is only a small part of the picture, which is why it was decided to sequence the host nuclear genomes as well. Fundamental to the reductive process of endosymbiosis is the transfer of genes from the endosymbiont to the host genome. Having the full host genome would allow us to see what was transferred from the endosymbiont and, just as importantly, what was lost rather than transferred.

The work presented here was part of the nuclear genome projects for *G. theta* and *B. natans* that I was involved in. Because my research was intimately linked with these genome projects, I participated in some fashion with many aspects of them. Consequently, Chapter 2 will give a general overview of the genome projects. It will also provide an organismal context for the research that follows.

One of the goals of the genome projects was to characterize the various proteomes. *G. theta* and *B. natans* are genetically complex unicellular organisms with four genomes each: nuclear; mitochondrial; plastid; and nucleomorph (Archibald 2007). Each of the non-nuclear genomes encodes the basic components of the proteomes associated with their subcellular location, but because of the reductive processes of endosymbiosis their genomes encode a small fraction of the proteins that function in those subcellular compartments. The remainder of the proteins are encoded by the nuclear genome and in the case of the plastid proteome also by the nucleomorph (Gilson, Maier et al. 1997). Fundamental to understanding endosymbiotic gene transfer (EGT) is knowing which genes were transferred and what became of them once they were transferred. Are their proteins now targeted back to the compartment from which they originally came? Have they acquired new functions? And are the various subcellular proteomes comprised of proteins encoded by genes that once resided in the endosymbiont, or are the current proteomes a mosaic of EGT derived proteins alongside host derived proteins, or even proteins encoded in other organellar genomes that have been co-opted for different duties or are dual targeted?

Because of the unique nature of the nucleomorphs considerable effort from the genome teams was directed at predicting the proteome of the periplastidial compartment (PPC) in *G. theta* and *B. natans*, which, being the vestigial cytoplasm of the engulfed algal endosymbiont, is the location of the nucleomorph (Archibald 2007). Intimately linked with the nucleomorph is the plastid that also resides within the PPC, and it too had its proteome for each organism predicted by members of the annotation team. Most of the assignments were done bioinformatically and by manual curation. In the case of *B. natans,* the plastid-PPC complex was successfully isolated and a mass spec study was undertaken that helped confirm and clarify some of the *in silico* predictions (Hopkins, Spencer et al., unpublished).

The final proteome, that of the mitochondrion, was undertaken by me. Using subcellular localization predictions, parsing of automated annotations and literature-directed homology searches, I attempted to determine what proteins were likely to be found in the mitochondrion for both *B. natans* and *G. theta*. The results of this attempt are presented in chapter 3. In describing these mitochondrial proteomes I have chosen not to provide a detailed and descriptive laundry list of each and every protein that potentially resides within the mitochondrion. I have given more attention to certain classes of proteins or organellar subsystems that are of particular interest. I provide more detail on some proteins to give the reader a sense of the investigative process that was sometimes necessary to predict which proteins are targeted where. Where appropriate, I discuss what the phylogenetic profile of certain proteins can tell us about the endosymbiotic process.

While I have characterized EGT as an ancient process that we can only study at a distance, the mechanisms by which genetic material moves from one genome to another are still operational and still active (Adams, Daley et al. 2000; Timmis, Ayliffe et al. 2004). Research in the last 20 years has demonstrated that varying amounts of mitochondrial and plastid DNA are being continuously transferred to the nuclear genome (Martin and Herrmann 1998; Ricchetti, Fairhead et al. 1999; Richly and Leister 2004; Richly and Leister 2004). Much of this transferred DNA is of little consequence, having an ephemeral existence as it gets integrated, degraded or expunged, but occasionally pieces are retained. More significantly some of those pieces can have a meaningful impact on the genome. Whole genes can be transferred and subsequently expressed.

Existing genes can be interrupted and exon/intron boundaries altered by the integration of the organellar DNA (Ricchetti, Tekaia et al. 2004; Noutsos, Kleine et al. 2007; Curtis and Archibald 2010).

Prior to the start of the nuclear genome projects, the nucleomorph and chloroplast genomes of both organisms had been sequenced and published (Douglas and Penny 1999; Douglas, Zauner et al. 2001; Gilson, Su et al. 2006; Rogers, Gilson et al. 2007). The mitochondrial genomes had been sequenced by the Organelle Genome Megasequencing Program (http://www.bch.umontreal.ca/ogmp/) but not published. However, the genome project teams included researchers who were involved in the sequencing of the mitochondrial genomes so I had access to the unpublished mitochondrial sequences for *G. theta* and *B. natans.* As well, during the sequencing of the nuclear genomes, all of the organellar genomes were re-sequenced to some degree due to the presence of contaminating organellar DNA in the purified nuclear DNA samples. Since I had all of the organellar genomes, it became possible to test for the presence of recently transferred organellar DNA in the nuclear genomes. Besides interrogating the nuclear genomes for the presence of mitochondrial and plastid DNA I would also be able to test for the presence of nucleomorph DNA, something that had never been done before since there had never been a nuclear genome sequenced from a nucleomorph-containing organism. My investigations into the presence of organellar DNA in the nuclear genomes of *B. natans* and *G. theta* are detailed in Chapter 4. Besides simply enumerating the transferred pieces I provide context by discussing the frequency of EGTs in other unicellular organisms. I also discuss some of the hypotheses that have been suggested to account for the varying levels of transferred pieces we see in different lineages. During my forays into the literature I also discovered several methodological problems with some of the previous estimates of EGT. I present a case study of some of the issues and what I believe to be best practices for further studies. Finally, I discuss the implications that recent EGT has on the question of why nucleomorphs have persisted in some lineages and what my findings might reveal about their fate.

As mentioned, most of the reductive processes that resulted in the organelles occurred relatively soon after the establishment of the endosymbiont. This assumption is based on the commonality of function and genes displayed by the various organelles. All plastids carry out the same basic operations and many auxiliary pathways are clearly derived and lineage specific. That specificity may be seen at different taxonomic levels. For example, all secondarily acquired red algal plastids share unique genes and pathways and are clearly monophyletic even if no consensus can be reached on the monophyly of their host lineages.

Given that most EGT occurred long ago, to what extent can we study it today? Endosymbiosis involves the massive integration of completely separate genetic lineages to forge new ones that over time have expanded and diversified. Is it possible to look back and untangle the mosaicism of current genomes to see what genes came from where? Or is the imprint of endosymbiosis so degraded and confused that global analysis of EGT and the search for red genes and green genes, for example, is a futile task open to constant revision and reanalysis? I will attempt to address this question from the perspective of *G. theta* and *B. natans*. I will look at a global analysis of these genomes from a blastp profile of top hits standpoint and discuss what a BLAST-based approach can and cannot do for understanding the phylogenetic makeup of genomes. This will be contrasted with a phylogenomic analysis that was in part completed for the purposes of the *B. natans* and *G. theta* genome paper.

CHAPTER 2 THE NUCLEAR GENOME PROJECTS OF *GUILLARDIA THETA* AND *BIGELOWIELLA NATANS*

This chapter includes work published in Bruce A. Curtis, Goro Tanifuji, Fabien Burki, Ansgar Gruber, Manuel Irimia, Shinichiro Maruyama, Maria C. Arias, Steven G. Ball, Gillian H. Gile, Yoshihisa Hirakawa, Julia F. Hopkins, Alan Kuo, Stefan A. Rensing, Jeremy Schmutz, Aikaterini Symeonidi, Marek Elias, Robert J. M. Eveleigh, Emily K. Herman, Mary J. Klute, Takuro Nakayama, Miroslav Oborník, Adrian Reyes-Prieto, E. Virginia Armbrust, Stephen J. Aves, Robert G. Beiko, Pedro Coutinho, Joel B. Dacks, Dion G. Durnford, Naomi M. Fast, Beverley R. Green, Cameron Grisdale, Franziska Hempel, Bernard Henrissat, Marc P. Höppner, Ken-Ichiro Ishida, Eunsoo Kim, Ludek Korený, Peter G. Kroth, Yuan Liu, Shehre-Banoo Malik, Uwe G. Maier, Darcy McRose, Thomas Mock, Jonathon A. D. Neilson, Naoko T. Onodera, Anthony M. Poole, Ellen J. Pritham, Thomas A. Richards, Gabrielle Rocap, Scott W. Roy, Chihiro Sarai, Sarah Schaack, Shu Shirato, Claudio H. Slamovits, David F. Spencer, Shigekatsu Suzuki, Alexandra Z. Worden, Stefan Zauner, Kerrie Barry, Callum Bell, Arvind K. Bharti, John A. Crow, Jane Grimwood, Robin Kramer, Erika Lindquist, Susan Lucas, Asaf Salamov, Geoffrey I. McFadden, Christopher E. Lane, Patrick J. Keeling, Michael W. Gray, Igor V. Grigoriev, John M. Archibald. 2012. Algal nuclear genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature* 492: 59-65.

As a member of the genome project team for both *Bigelowiella natans* and *Guillardia theta* I participated in a number of analyses that were used in the paper describing the two genomes. For contextual reasons this chapter includes a general overview of the methods used for the genome projects. I have highlighted those sections of the methods where my involvement was significant by giving a more detailed description and by indicating the team members who were also involved.

Guillardia theta belongs to the cryptophytes, a group of biflagellated unicellular aquatic organisms (Hoef-Emden and Melkonian 2003). Well over 100 species have been described including freshwater and marine species. The vast majority are photoautotrophs with *Goniomonas* being the only phagotrophic and aplastidial genus (McFadden, Gilson et al. 1994). Although a well-known cryptomonad, *G. theta* had an uncertain taxonomic position within the cryptophytes until 1990 when it was formally described (Hill and Wetherbee 1990). It is a marine photosynthetic organism. Like other cryptophytes it is characterized by a plate-like covering called a periplast. *G. theta* possesses a periplastidial space between the two inner and the two outer plastid membranes. Within this periplastidial compartment can be found the pyrenoid and the nucleomorph (Hill and Wetherbee 1990).

The plastid genome of *G. theta* was sequenced in 1999 (Douglas and Penny 1999) and confirmed that it is of red algal origin, while the nucleomorph genome was sequenced and published a few years later (Douglas, Zauner et al. 2001). The nucleomorph of *G. theta* was reported to consist of three chromosomes with a total size of 551 kbs. A central theme of the nucleomorph paper was the highly compacted nature of the genome with only 17 small spliceosomal introns, 44 overlapping genes and a gene density of 1.07 Kb/gene. Only 30 of the 465 protein coding genes were plastid targeted.

B. natans is a member of the group chlorarachniophytes that are marine amoeboflagellates. They contain chloroplasts with chlorophyll *a* and *b* that are surrounded by four membranes (Ishida, Green et al. 1999). Their most striking features are nucleomorphs, vestigial eukaryotic nuclei, which are located in the periplastidial compartment that lies between the two pairs of membranes that surround the plastid. Nine species have been recognized, scattered among five genera that have been classified based on the location of the pyrenoid and the nucleomorph (Ishida, Green et al. 1999). Based on nucleomorph karyotyping there are probably at least six genera with new species waiting to be characterized (Silver, Koike et al. 2007). The genus *Bigelowiella* currently has two members, *B. natans* and *B. longifila* (Ota, Ueda et al. 2007).

The presence of chlorophyll *a* and *b* suggested early on that the plastid, and by extension the nucleomorph, was of green algal origin (Hibberd and Norris 1984). This was confirmed with the sequencing and analysis of plastid genes (Ishida, Cao et al. 1997) and eventually the entire plastid genome (Rogers, Gilson et al. 2007) (Rogers et al. 2007). The *B. natans* nucleomorph genome sequence was released in 2006 (Gilson, Su et al. 2006) indicating a size of 373 kbp. 331 genes were identified, but only 17 of these encoded proteins thought to function in the plastid, and as with *G. theta* the rest were considered to have housekeeping functions. Unlike *G. theta,* the *B. natans* nucleomorph was replete with introns, 852 in total, but they are also the smallest known, falling within a very narrow range, from 18-21 nt in length. Remarkably, the *B. natans* nucleomorph genome also consists of three chromosomes.

The persistence of the nucleomorphs in these two lineages has, understandably, generated considerable speculation and debate. It seems remarkably inefficient to maintain the whole nucleomorph and PPC apparatus of translation. Early hypotheses centered on the nucleomorph having indispensable roles for the proper functioning and maintenance of the plastids. However, the sequencing of the nucleomorph genomes revealed that few of their genes were actually involved in the plastid. Of the 17 plastid targeted proteins from *B. natans* and 30 from *G. theta* only two overlapped, suggesting that the loss or persistence of any one particular plastid targeted gene from the nucleomorph was essentially down to chance (Gilson, Su et al. 2006). It was also suggested that the number and size of the introns in the *B. natans* nucleomorph might be an impediment to the successful transfer of genes to the host nucleus. The usual eukaryotic spliceosomal machinery would be unable to efficiently recognize and splice out all of the tiny introns (Gilson and McFadden 1996). However, six of the 17 plastid targeted genes lack introns, which would make them available for transfer, while in *G. theta* a tiny fraction of the genes have introns and none of the plastid targeted genes have introns.

There are a number of significant differences between organisms containing primary plastids and those containing secondary plastids. In addition to the double-membrane plastid envelope derived from the cyanobacterium, secondary plastids have one or two additional membranes (McFadden 1999). The first (outermost) membrane appears to be derived from the phagosomal membrane of the host cell, while the second additional membrane is thought to represent the plasma membrane of the algal endosymbiont. Unlike primary plastids, which are found in the cytosol, secondary plastids reside within the lumen of the host cell's endomembrane system (McFadden 1999).

Because secondary plastids are characterized by the presence of three or four plastid membranes, organisms with secondary or tertiary plastids have had to evolve a complex mechanism for protein import (Gould, Somer et al. 2006; Gould, Somer et al. 2006). While a transit peptide can be used to cross the inner plastid membranes, as in primary plastids, proteins targeted to secondary plastids must first cross one or two additional barriers. In heterokonts, haptophytes and cryptophytes this is accomplished via a bipartite N-terminal extension comprised of a signal peptide and a transit peptide (Gould, Sommer et al. 2006). The signal peptide allows the protein to cross the outermost membrane into the ER lumen. The transit peptide is then used to cross the remaining barriers after cleavage of the signal peptide. In some secondary plastid-containing organisms, the cytoplasm of the engulfed eukaryote persists between the 2nd and 3rd plastid membranes. This periplastidial compartment is therefore also a potential destination for nucleus-encoded proteins. Targeting to this compartment appears to be achieved by a minor modification to the transit peptide (Gould, Sommer et al. 2006). A phenylalanine residue upstream of the transit peptide proper causes the protein to continue into the stroma of the plastid, while its absence results in the protein remaining in the periplastidial compartment.

2.1 TAXONOMIC PLACEMENT

B. natans is a chlorarachniophyte, which belongs to the phylum Cercozoa. Along with Formanifera and Radiolaria, Cercozoa has been placed in the supergroup Rhizaria (Cavalier-Smith 1998; Adl, Simpson et al. 2005) in the six supergroup higher level eukaryotic classification system. *G. theta* is a cryptomonad from the phylum Cryptophyta, which along with Haptophyta, Heterokontophyta (stramenopiles) and Alveolata, make up the supergroup Chromalveolata (Adl, Simpson et al. 2005). As mentioned in the introduction, the coherence of the supergroup Chromalveolata has come under increasing doubt. Large scale, multigene phylogenomic analyses tend to split Chromalveolata into two camps, with Cryptophyta and Haptophyta (Patron, Inagaki et al. 2007) closely related, while Stramenopiles and Alveolata are equally related (Harper, Waanders et al. 2005). A surprising result from the last six years is that a number of studies have found a close relationship between Rhizaria, Stramenopiles and Alveolata to the exclusion of the cryptophytes and haptophytes. Some of the studies continue to support the chromalveolates as monophyletic with the inclusion of Rhizaria (Hackett, Yoon et al. 2007). Other studies break the chromalveolata supergroup and, in the process, suggest a new supergroup dubbed SAR (Stramenopiles, Alveolata and Rhizaria) that excludes cryptophytes and haptophytes (Burki, Shalchian-Tabrizi et al. 2007). While most of the recent studies seem to support SAR, the position of the cryptophytes and haptophytes in the eukaryotic tree continues to be unresolved. Some studies have placed haptophytes and cryptophytes with the plants (Burki, Shalchian-Tabrizi et al. 2008) in a highly supported group. Other multigene studies have separated haptophytes and cryptophytes (Yoon, Grant et al. 2008), embedding Cryptophyta within the primary plastid lineages while Haptophyta is a sister group to SAR.

The relationship between the Chromalveolata phyla and the continued belief in the chromalveolate hypothesis is critically important to the study and analysis of plastid endosymbiotic gene transfer (EGT). If cryptophytes and haptophytes are no longer monophyletic with stramenopiles and alveolates then the acquisition of the plastid in the photosynthetic lineages as a single secondary endosymbiotic event becomes untenable. One would have to postulate at least two separate events with one event giving rise to the plastids in cryptophytes and haptophytes, while the second event involved acquisition of a red plastid in the ancestor of the stramenopiles and alveolates. Given that several major lineages from the stramenopiles and alveolates are aplastidic one could even envisage multiple secondary endosymbiotic events rather than an ancestral acquisition. It has also been speculated that the stramenopiles and alveolates acquired their plastids through tertiary endosymbiosis involving the engulfment of a cryptophyte or haptophyte (Bodyl, Stiller et al. 2009). The strong possibility that rhizarians are allied with at least stramenopiles and alveolates also complicates matters since *B. natans* clearly acquired its plastid and nucleomorph from a green algal endosymbiont. If the various chromalveolata lineages acquired their plastids in a single secondary endosymbiotic event then presumably the ancestor of chlorarachniophytes at one point had a red algal endosymbiont that was lost and replaced by the current green algal one. During the period when the chlorarachniophyte ancestor possessed a red algal endosymbiont there may have been EGT. This cryptic endosymbiosis would confound phylogenetic analysis of plastid and PPC targeted proteins. If however, there were multiple instances of secondary or tertiary endosymbiosis among the stramenopiles and alveolates then *B. natans* may not have had a cryptic red algal endosymbiont.

2.2 Methods

Single-cell isolates of *Guillardia theta* CCMP327 and *Bigelowiella natans* CCMP621 were established at the Bigelow Laboratory for Ocean Sciences. Cultures were grown by Chris Lane and Julia Hopkins under a 12h:12H light:dark cycle. Purified nuclear DNA was obtained for both organisms using a Hoechst dye-CsCl density gradient fractionation. Three different-sized libraries, 3kb, 8kb and 34kb fosmids, were generated from the purified DNA and sequenced at the Joint Genome Institute (JGI) using Sanger sequencing. Additional 454 sequencing was done to fill gaps created by hard sequence stops due to high GC areas. In particular, an additional 454 paired end library with 3kb inserts was sequenced for *G. theta* due to the higher number of gaps compared to *B. natans*.

1,393,200 Sanger and 5,145,887 454 reads were generated for *G. theta,* while for *B. natans* 1,127,564 Sanger and 470,629 454 reads were generated. The reads were assembled using a modified version of Arachne (Jaffe, Butler et al. 2003) v. 20071016 with the parameters maxcliq1=150 and BINGE_AND_PURGE=True for *G. theta* and maxcliq1=100, correct1_passes=0 and BINGE_AND_PURGE=True for *B. natans*. Redundant 454 pairs were removed prior to assembly. Those remaining were pre-corrected for indels using the Sanger reads which tend to be of much higher quality.

After the initial assembly, contigs with less than 400 bps and containing only 454 reads were removed after which the assembly was rerun. Contigs were checked against known bacterial and organellar proteins. No bacterial contamination was detected for either genome. Due to the inability to completely separate organellar DNA from nuclear DNA using a Cs-Cl gradient a number of the contigs were identified as being of organellar origin. In *G. theta* one scaffold each was classified as being from the mitochondrial, plastid or the nucleomorph genomes. In *B. natans*, one scaffold was considered mitochondrial, one plastidic and 21 were designated as being from the nucleomorph genome. Scaffolds less than 1000 bps were removed.

The final genome assembly for *G. theta* consisted of 670 scaffolds with a combined size of 83.5 Mb. N50 was 40.4 kb for contigs and 545.8 kb for scaffolds. *B. natans* was considered easier to sequence and assemble and this is reflected in the assembly numbers with 302 scaffolds for a combined size of 94.7 Mb. N50 for contigs was 59.5 kb and 820.0 kb for scaffolds.

To facilitate gene modeling expressed sequence tags (EST) libraries were created using RNA provided by C. Lane, D. Spencer and J. Hopkins and sequenced at JGI for each organism. 40,704 *B. natans* JGI sequenced ESTs were combined with 3,460 previously generated and publicly available ESTs (accession numbers DR038244-DR041707). For *G. theta* 30,720 JGI sequenced ESTs were combined with 18,642 publicly available ones. The combined ESTs were clustered using a JGI in-house program, malign, that uses a kmer=16 based alignment tool and a minimum sequence overlap of 32 with an alignment ID ≥ 98%. Matching sister ESTs were also used to combine clusters where possible. Clusters were then assembled using CAP3 (Huang, Ayliffe et al. 2003) resulting in 14,092 consensus sequences for *B. natans* and 14,142 for *G. theta.*

After assembly, the genomes were annotated using the JGI Annotation Pipeline. The initial step consists of using several gene predictors. At that time JGI used FGENESH (Salamov and Solovyev 2000) and GeneWise (Birney, Clamp et al. 2004). GeneWise was seeded with blastx alignments generated by comparing the genomic sequences against the NCBI non-redundant protein set nr. Publicly available *G. theta* and *B. natans* gene models from previous studies were used to train FGENESH for prediction of *ab initio* gene models. As well, the EST clusters were mapped to the genomic sequences to produce cDNA-based gene models and to guide and improve the identification of exon boundaries for the Genewise models.

2.2.1 AUGUSTUS Modeling

To supplement the JGI gene modeling I, in collaboration with Robert Eveleigh, used the AUGUSTUS software (Stanke, Schoffmann et al. 2006) to predict genes for both *B. natans* and *G. theta*. The prediction software uses a generalized Hidden Markov model employing probability distributions for 47 states. AUGUSTUS can be used to predict genes purely from the genome sequence but is more powerful if there is gene information available for the genome in question that can be used as a training set. The documentation for AUGUSTUS suggests that 100-200 genes are adequate for training purposes. I compiled a 252 gene training set for *B. natans*. 124 of the genes were publicly available, having been sequenced in prior studies (Archibald, Rogers et al. 2003). 104 gene structures were supplied by team members from unpublished data. 24 were generated from searching the EST data generated by JGI and previous studies and finding, based on blastp homology searches, full length transcripts with predicted start and stop codons. AUGUSTUS requires the training set genes to have certain characteristics such as at least one intron and not be more than 70% similar to any other training gene. After assessing the 252 possible training genes I was left with 190.

A run of AUGUSTUS (Stanke, Schoffmann et al. 2006) using the 190 *B. natans* genes as a training/testing set produced 28,236 gene models. A cursory examination of the predicted proteins revealed that a large percentage of them had long runs of identical amino acids. Based on familiarity with other protein sets the high number of tandem amino acid repeats (TAR) in the proteins predicted by AUGUSTUS seemed unusual. Use of an in-house Perl script revealed that 28% of the proteins had at least one occurrence of at least 10 of the same amino acid appearing together (TAR10). To provide context I analyzed the percentage of predicted proteins with TAR10s in other organisms for which annotated genomes are available (Table 2.1).

Table 2.1. Percentage of proteins from genome protein sets with at least one occurrence of 10 identical amino acids in a row.

Species	TAR10 percentage
Emiliania huxleyi	2.71
Phaeodactylum tricornutum	0.26
Arabidopsis thaliana	0.97
Chlamydomonas reinhardtii	3.41
Thalassiosira pseudonana	0.84
Ostreococcus tauri	1.06
Micromonas pusilla	10.15
Micromonas RCC299	2.88
Ostreococcus lucimarinus	1.11
Phytophthora ramorum	0.38
Phytophthora sojae	0.84
Phytophthora infestans	0.19
Aureococcus anophagefferens	1.44
Chlorella vulgaris	5.46
Physcomitrella patens	0.33
Volvox carteri	9.09

All of the other species that I tested had TAR10 percentages far below the 28% from the *B. natans* AUGUSTUS test set. The highest percentage was for *Micromonas pusilla* at

10.15% followed by *Volvox carteri* at 9.09%, but most of them had less than 2%. These results suggested that the AUGUSTUS predicted protein set using the 190 genes as a training/testing set included a number of spurious proteins and/or repetitive areas of the genome like microsatellites that should not be part of proteins.

AUGUSTUS (Stanke, Schoffmann et al. 2006) also has the option to use clustered or raw ESTs. Using 6126 EST clusters with AUGUSTUS generated 21,617 gene models (Table 2.2). The TAR10 percentage using the EST clusters was seven. This result seemed more reasonable when compared to other organisms (Table 2.1). Based on these results I decided that using ESTs as a training set produced a more realistic set of gene models. I tested whether using the EST clusters or a cleaned version of the raw ESTs as the training set made a difference. I also tested several AUGUSTUS options such as a second round of optimization, and using BLAT (Kent 2002) to produce a single transcript from the ESTs or multiple transcripts.

AUGUSTUS (Stanke, Schoffmann et al. 2006) segregates 10% of any training set to use as a test of the final output. The tests reported include sensitivity and specificity scores for the following features: nucleotides; exons; and genes. Sensitivity is defined as the number of correctly predicted features divided by the number of features from the test set, while specificity is the number of correctly predicted features divided by the number of predicted features. A predicted exon is considered ‘correct’ if both predicted splice sites match the splice sites from the test set. A predicted gene is considered correct if all exons are correctly predicted and with no additional exons compared to the test set.

Because 190 genes was such a small training set the test scores for the *B. natans* gene model predictions were deemed to be meaningless. Using the raw, cleaned ESTs made a clear improvement over using the clusters (Table 2.2). Using two rounds of optimization also increased the specificity and sensitivity scores for all the test features. Little or no change was seen in the test scores when using BLAT (Kent 2002) to match the ESTs against the genome sequence regardless of whether one or multiple transcripts were predicted. Ultimately the parameters for Test 4, which included 2 rounds of optimization, were considered the best overall and implemented to create AUGUSTUS (Stanke, Schoffmann et al. 2006) gene models for both *B. natans* and *G. theta*.

Table 2.2. Test scores for *B. natans* generated by AUGUSTUS under several conditions.

	Test1	Test2	Test3	Test4	Test5	Test6
Training set	190 genes	6126 EST clusters	15855 cleaned ESTs	15855 cleaned ESTs	15855 cleaned ESTs	15855 cleaned ESTs
Modifications		PASA	PASA	PASA, 2 rounds of optimization	PASA, BLAT – single hit	PASA, BLAT-single hit Multiple transcripts
Genes predicted	28236	21617	22858	22545	23041	22903
Nucleotide sensitivity		.665	.747	.772	.765	.75
Nucleotide specificity		.478	.487	.496	.502	.481
Exon sensitivity		.361	.496	.531	.528	.524
Exon specificity		.359	.366	.376	.387	.373
Gene sensitivity		.147	.188	.234	.221	.214
Gene specificity		.119	.139	.167	.161	.16
TAR10 %	28	7	9.3	9.1	9.15	9.15

Because of the different gene models that were generated for any one locus a filtered (or 'catalog') set consisting of a single model for each locus was chosen based on homology and EST support. 24,840 catalog models were chosen for *G. theta* and 21,708 for *B. natans*. The catalog models, as well as the alternative models, were functionally annotated using the JGI Annotation Pipeline. This initially consisted of running the models through InterProScan (Zdobnov and Apweiler 2001) and generating hardware-accelerated double affine Smith-Waterman alignments (http://www.timelogic.com) against the following databases: SwissProt (Bairoch, Apweiler et al. 2005); KEGG (Ogata, Goto et al. 1999); PFAM (Bateman, Coin et al. 2004). EC numbers (Bairoch 2000) were mapped to the models using KEGG hits while GO terms (Ashburner, Ball et al. 2000) were generated using InterPro, KEGG and SwissProt hits. SignalP 3.0 (Nielsen, Brunak et al. 1999) was used to generate protein targeting predictions while transmembrane domains were assessed using TMHMM (Krogh, Larsson et al. 2001). Models were also assigned KOG classifications (Koonin, Fedorova et al. 2004). All of the annotation information is stored in genome browsers accessible through the JGI Portals (Grigoriev, Nordberg et al. 2012) for *B. natans* (http://www.jgi.doe.gov/Bnatans) and *G. theta* (http://www.jgi.doe.gov/Gtheta).

2.2.2 RNA-Seq and PASA

Several years after initial sequencing and EST generation, RNA-Seq libraries were constructed from total RNA and sequenced at the National Centre for Genomics Research (NCGR) using an Illumina HiSeq 2000. The sequencing runs generated 100 nt from each paired end. After filtering for poor quality the 11,368,985 reads pairs for *B. natans* and

9,122,347 for *G. theta* were assembled by NCGR using ABySS (Simpson, Wong et al. 2009) resulting in 31,324 contigs over 150 bps in length for *B. natans* and 24,790 for *G. theta*. In collaboration with Eunsoo Kim I reassembled the reads using Trinity (Grabherr, Haas et al. 2011) after error correction using ALLPATHS-LG (Gnerre, Maccallum et al. 2011). A fixed kmer of 25 was used for the Trinity assembly. 98,794 contigs greater than 200 bps were generated for *B. natans* and 37,828 for *G. theta*. The Trinity contigs were used with PASA software (Haas, Delcher et al. 2003) to generate a set of corrected JGI catalog gene models for each genome that were then loaded into the JGI genome browsers as additional alternative models for the purposes of manual annotation. The Trinity contigs and PASA were also used to generate exhaustive alternative splicing reports and alternative transcripts that were used in the evaluation of alternative splicing in both organisms.

2.2.3 Sub-cellular Localization Predictions

Four independent methodologies were used to predict the sub-cellular localization of host encoded proteins to the following compartments: mitochondrion; plastid; PPC; endoplasmic reticulum/ Golgi associated.

2.2.3.1 Mitochondrial Proteomes

Each protein in the complete filtered protein sets for *B. natans* and *G. theta* was assessed for targeting to the mitochondrial compartment in collaboration with Takuro Nakayama using three different sub-cellular targeting prediction methods, TargetP, Predotar and iPSORT. Version 1.1 of TargetP (Emanuelsson, Brunak et al. 2007) was run locally as was iPSORT (Bannai, Tamada et al. 2002)S, while Predotar (Small, Peeters et al. 2004) results were generated through a webserver (http://urgi.versailles.inra.fr/predotar/predotar.html). Although many sub-cellular localization prediction tools are available (Imai and Nakai 2010), TargetP, Predotar and iPSORT were chosen because they are widely used (especially TargetP), are available either for local installation or through a webserver, do not require data transformation, and employ different methods as well as different training sets. Proteins with positive mitochondrial sub-cellular localization predictions for all three programs were retained for downstream analysis. In the case of *G. theta,* 785 proteins were retained while for *B. natans* the number was 612.

2.2.3.2 Plastid Proteomes

Predictions for proteins targeted to the plastid were done by a team in Germany led by Peter Kroth. Their methods focused on what is currently known about plastid targeting in cryptophytes and chlorarachniophytes, especially the sequence and structure of the bipartite presequences that are required to traverse the four membranes that surround the plastids. Because of prior work with diatoms, which share the same plastid origin as cryptophytes, the prediction methods for *G. theta* involved a more sophisticated bioinformatics strategy and allowed a low- and high-confidence set to be inferred (Curtis et al. 2012 In press).

2.2.3.3 ER/Golgi Proteomes

A homology-based approach was developed by team members Robert Eveleigh, Shinichiro Maruyama and me to identify proteins thought to reside in the ER. The database ER-GolgiDB (Wrzeszczynski and Rost 2004) is a curated set of experimentally verified ER/Golgi proteins from *Arabidopsis thaliana*, human and *Saccharomyces cerevisiae*. *B. natans* and *G. theta* protein models were searched against the database using blastp (version 2.2.25+ using an E-value cutoff of 0.01). Predictions were complicated by the fact that prior to insertion into the PPC the proteins must first be directed to the ER. In cryptophytes and stramenopiles the outermost plastid/nucleomorph membrane is continuous with the ER (Gould, Sommer et al. 2006). In chlorarachniophytes the outermost membrane is not connected with any endomembrane and it is believed that the proteins are transported from the ER to the plastid/nucleomorph membrane by vesicular transports (Hirakawa, Gile et al. 2010). In both cases nucleus encoded proteins destined for the PPC or plastid are initially targeted to the ER using an N-terminal signal peptide. Prior to the blastp searches, the signal peptide was removed based on its predicted length from SignalP 3.0. Any sequences with good blastp hits were further examined for the presence of the second part of the bipartite presequence that is required for proteins that leave the ER for the PPC or plastid. The results of the blastp searches were categorized as follows:

(a) If the query-hit BLAST alignment started immediately after the SP cleavage site, the database hit protein was assigned to 'class 1'

(b) If the query-hit BLAST alignment started between 2 and 45 amino acids downstream of the SP cleavage site, the database hit protein was designated 'class 2'.

(c) If the query-hit BLAST alignment started >45 amino acids after SP cleavage site, the database hit protein was designated ‘class 3’

(d) If the top BLAST hit was below the coverage threshold of 50% (i.e., the query-hit match covered <50% of the length of each sequence), the database hit protein was designated ‘class 4’, regardless of whether criteria (a) or (b) were met.

To be considered a putative ER/Golgi protein the query had to have a hit against the ER/Golgi database, was either class 1 or class 2, and had an ER-GolgiDB accuracy score ≥ 75% (Wrzeszczynski and Rost 2004). Using these criteria 650 ER/Golgi proteins were identified in *B. natans* and 763 in *G. theta*. If a more stringent accuracy score of ≥ 95% was used the number of putative ER/Golgi proteins dropped to 233 and 219 in *B. natans* and *G. theta* respectively.

2.2.3.4 PPC proteins

Determining the PPC proteomes for *B. natans* and *G. theta* was initiated as a team effort principally conceived and carried out by Goro Tanifuji, Shinichiro Maruyama, Gillian Gile, John Archibald, and me. The starting point for the identification of proteins targeted to the PPC was the prediction of signal peptides using SignalP 3.0. Because of concerns about the accuracy of the gene models, especially for the N-termini where the bipartite sequences are found, SignalP 3.0 scores were examined for the catalog models as well as all the alternative models. The model with the best positive SignalP 3.0 score was retained for further examination. This resulted in 5,708 putative models for *B. natans* and 7,475 for *G. theta*.

Each model was then given an initial score using the following criteria: 1) SignalP score (Neural network prediction value (0 – 5) plus hidden Markov model predictions, "Yes = 1" or "No = 0"); (2) presence of leader sequence (N-terminal extension) ("Yes = 2, Maybe = 1, No = 0"); (3) a score of "0" or "5" based on the results of a 'support vector machine' (SVM)-based approach to identifying PPC proteins (see below); (4) 'intron score' based on number of introns ("none = 0, 1-10 introns = 1, more than 11 introns = 2"); (5) gene copy number; (6) Sanger EST coverage score (">90% coverage = 0, 33 – 90% coverage = 1, no coverage = 2"); (7) RNA-Seq coverage score (">90% coverage = 0, 33 - 90% coverage = 1, no coverage = 2"); (8) score based on the number of blastp top hits to green algae and Viridiplantae for *B. natans*, or Rhodophyta for *G. theta* ("6-10 hits = 2, 1-5 hits = 1, none =0"); and (9) abundance of D or G/K amino acid residues in the C-terminal region of the candidate proteins in *B. natans* ("Yes = 2, Maybe = 1, No = 0"). Putative PPC proteins with a score of 10 or greater (out of 25 for *B. natans* and 22 for *G. theta*) were retained for further analysis. Any candidates with less than 3 for a SignalP 3.0 score were rejected.

Both plastid and PPC targeted proteins have bipartite presequences (Gould, Sommer et al. 2006; Hirakawa, Gile et al. 2010). The putative target peptide portion that lies within 45 amino acids downstream of the end of the signal peptide was examined for the characteristics that have been found meaningful in predicting whether a protein is retained in the PPC or proceeds to the plastid. This region was first analyzed for the presence of a target peptide using chloroP (Emanuelsson, Nielsen et al. 1999). The mean net charge of the region was determined and those proteins with a score of ≥ 3 were rejected. Specifically in *G. theta*, proteins with the residues F, Y, W or L immediately following the signal peptide cleavage site were rejected as plastid targeted. The putative bipartite area was also assessed using blastp and the NCBI nr database. If the signal peptide portion possessed similarity to the top 10 hits it was rejected.

JGI generated KOG classifications were assigned to the putative PPC targeted proteins where possible. Goro Tanifuji manually verified these assignments using blastp searches and kognitor (http://www.ncbi.nlm.nih.gov/COG/grace/kognitor.html) and moved unverified ones to a functionally ambiguous category. Over half of the putative PPC-targeted proteins did not have KOG assignments. These were placed in conserved or non-conserved protein categories based on them having seven or more blastp hits against the NCBI nr database with an e-value cutoff of 1e-30.

Putative PPC-targeted proteins were also predicted from the PASA corrected gene models using a support vector machine approach. This analysis was done by a team from Germany led by Stefan Rensing and a detailed description can be found in the supplemental materials for Curtis et al. 2012 In press.

2.2.4 Phylogenomics

The phylogenomics analysis for all protein coding genes in *G. theta* and *B. natans* was done as part of the genome project for each species and was formulated and implemented in conjunction with other annotation team members, principally Fabien Burki, Robert Eveleigh, Shinichiro Maruyama, Goro Tanifuji and John Archibald.

A phylogenomic pipeline was instituted to identify genes of putative algal origin in the nuclear genomes of *B. natans* and *G. theta*. Homologs were found by blastp searches against a curated local database of genomic and EST-derived sequences. Specific attention was given to protein sets from chromalveolate lineages and especially red algae because of taxon sampling bias concerns. All catalog protein models from *G. theta* (24,840) and *B. natans* (21,708) were used as queries. The advantages of manual examination of individual trees of interest in large-scale phylogenomic studies prompted us to take steps to reduce the number of operational taxonomic units (OTUs) without sacrificing taxonomic and alignment complexity. OTUs were processed to limit the number of prokaryotic sequences. A first round of alignments and maximum likelihood trees were constructed after which a novel 'de-replication' procedure was employed to further reduce OTU redundancy within highly supported subtrees such as those consisting of multiple prokaryotic sequences from similar genomes. This step also served to reduce the presence of lineage specific recent paralogs from eukaryotic genomes. After de-replication the remaining sequences were realigned using a more rigorous process and ML trees were generated using a local installation of RAxML. For details of programs and program parameters please see Appendices A1-3.

2.2.5 Blastp Analysis

A local protein database was created (Table 2.3). To ensure taxonomic breadth entries were obtained from various sources in addition to NCBI. Whole protein sets were downloaded from the JGI genome portal, the Broad Institute, and protist EST sets from TBestDB (http://amoebidia.bcm.umontreal.ca/pepdb/searches/login.php) (Appendix B local database entries). Since one of the goals was to search for red and green algal signals in the two genomes the paucity of red algal proteins sets was of concern especially since the only complete genome available was from *Cyanidioschyzon merolae*, which has a reduced genome and is probably not typical of most rhodophytes. Consequently, BLAST analysis as well as phylogentic trees were redone when two red algal protein sets were made available by the Bhattacharya lab (http://dbdata.rutgers.edu/data/plantae/) - 36,167 ESTs for *Porphyridium cruentum* and 23,961 AUGUSTUS (Stanke, Schoffmann et al. 2006) gene models for the partial genome of *Calliarthron tuberculosum* (Chan, Yang et al. 2011).

Table 2.3. Taxonomic breakdown of local protein database used for blastp analysis.

Major lineage	Entries
Metazoa	788,787
Amoebozoa	53,352
Fungi	173,360
Rhodophyta	125,954
Viridiplantae	196,555
Bacteria	3,641,103
Archaea	204,473
Chromalveolata	1,331,090
Excavata	162,204
Apusozoa	11,582
Choanoflagellates	25,229
Total	6,551,485

Blastp searches (blastp 2.2.25+, e value cutoff 0.001) of the filtered protein sets from *B. natans* and *G. theta* against the local database were done on the Dalhousie High Performance Computer cluster MOA (https://hpc.dal.ca/wiki/public:system:moa). Results were parsed using custom PERL scripts.

2.3 Results and Discussion

2.3.1 Gene Models

For each genome, gene models were created using FGENESH (Salamov and Solovyev 2000), GeneWise (Birney, Clamp et al. 2004) and AUGUSTUS (Stanke, Schoffmann et al. 2006). From the available models for each locus one model was chosen for the catalog. In *B. natans* 21,708 models were chosen for the catalog while in *G. theta* the catalog contained 24,840 gene models (Table 2.4). The biggest contributor of gene models to the catalog for *B. natans* was FGENESH with 56.9% of the catalog models derived from its predicted set (Table 2.4). AUGUSTUS contributed 28.2% of the catalog models while GeneWise was a distant third at 14.7%. For *G. theta* FGENESH also contributed the most catalog models but its percentage dropped by 13.7% compared to *B. natans* while the percentage of AUGUSTUS models in the catalog increased compared to *B. natans* from 28.2% to 37.4%. The number of GeneWise models in the catalog was higher in *G. theta* compared to *B. natans* but still contributed less than 20% of the models.

During the manual annotation process curators were able to modify existing catalog models as well as promote alternative models to the catalog while at same time demoting the previous catalog model for that particular genome locus. Besides the models based on the modelers FGENESH (Salamov and Solovyev 2000), GeneWise (Birney, Clamp et al. 2004) and AUGUSTUS (Stanke, Schoffmann et al. 2006), annotators also had access to models based on mapping ESTs to the genome sequence and ones based on PASA corrected models derived from RNA-Seq data. 2259 and 1011 catalog models were manually curated in *B. natans* and *G. theta* respectively. During that annotation process the number of genes in the catalog increased by 27 in *B. natans* and by 40 in *G. theta* as annotators split current models and created new catalog loci (Table 2.4). As annotators promoted alternative models the source of the catalog models also changed. In *B. natans* the current catalog contains 76 models based on PASA while in *G. theta* 39 catalog models are derived from PASA results. The number of catalog models based on FGENESH models increased in both *B. natans* and *G. theta* while those based on GeneWise dropped by 169 in *B. natans* and 58 in *G. theta*. The number of catalog gene models based on AUGUSTUS increased slightly in *B. natans* (38) and decreased slightly in *G. theta* (5).

Based on the numbers presented here the PASA corrected models did not have a huge impact on annotation and catalog gene models. However, based on personal experience from looking at hundreds of models for putative mitochondrial targeted protein and from speaking with other annotators, the PASA data was very helpful and informative, particularly for the 5′ regions. The use of PASA models to address questions of gene model inaccuracy is not reflected in the numbers due to the following: rather than promote the PASA model to the catalog, the JGI catalog model could be modified based on the PASA model; time constraints meant that models were not always modified, particularly in cases where none of the models were "correct" but some were more "correct" than others.

Table 2.4. Gene prediction source for catalog models in *G. theta* and *B. natans*.

	B. natans original gene models	*B. natans* gene models after annotation	*G. theta* original gene models	*G. theta* gene models after annotation
FGENESH	12,372 (56.9%)	12,448	10,738 (43.2%)	10,799
GeneWise	3209 (14.7%)	3040	4790 (19.2%)	4732
AUGUSTUS	6127 (28.2%)	6165	9312 (37.4%)	9307
PASA	NA	76	NA	39
ESTs	0	6	0	3
Total	21708	21735	24,840	24,880

2.3.2 Genome Size and Protein Coding Genes

The genome sizes of *G. theta* at 87.2 Mb and *B. natans* at 94.7 Mb puts them near the top end of genome sizes of unicellular eukaryotes, if one ignores the bloated and unusual genomes of dinoflagellates. Among photosynthetic unicellular algae only *Chlamydomonas* has a larger genome (Table 2.5), while among sequenced chromalveolates only *Ectocarpus* at 214 Mb, *Phytophthora infestans* at 240 Mb and *Tetrahymena thermophila* at 104 Mb are bigger. Perhaps more informative is the number of protein coding genes. Although claims of correlation between eukaryotic genome size and gene content (Hou and Lin 2009) have been made, lineage specific analysis shows that genome size is often at the mercy of particular circumstances that belie such a connection. For example *P. infestans* has a genome size of 240 Mb, 4x that of *P. ramorum* and 3x that of *P. sojae* (Table 2.5) yet the number of protein coding genes does not display a similar increase. In fact *P. sojae* has more protein coding genes than *P. infestans* while *P. ramorum* has almost as many protein coding genes as *P. infestans*.

Large, unusual genome sizes are usually attributable to a combination of repetitive noncoding regions like transposons and multiple copies of genes as in some dinoflagellates (Bachvaroff and Place 2008). If one looks at protein coding genes *G. theta* and *B. natans* appear to be very gene rich, much more so than other unicellular photosynthetic algae. With 24,840 protein coding genes *G. theta* has twice as many as the diatoms *Thalassiosira pseudonana* and *Phaeodactylum tricornutum,* and more than any sequenced stramenopile including the multicellular organism *Ectocarpus siliculosus* (Table 2.5).

Table 2.5. Genome size and protein coding genes for select organisms.

Genome	Size (Mb)	Protein coding genes
Guillardia theta	87.2	24,840
Bigelowiella natans	94.7	21,708
Phaeodactylum tricornutum (Bowler, Allen et al. 2008)	27.4	10,402
Thalassiosira pseudonana (Armbrust, Berges et al. 2004)	34	11,242
Aureococcus anophagefferens (Gobler, Berry et al. 2011)	57	11,501
Ectocarpus siliculosus (Cock, Sterck et al. 2010)	214	16,256
Phytophthora infestans (Haas, Kamoun et al. 2009)	240	17,797
Phytophthora ramorum (Tyler, Tripathy et al. 2006)	65	15,743
Phytophthora sojae (Tyler, Tripathy et al. 2006)	95	19,027
Tetrahymena thermophila (Eisen, Coyne et al. 2006)	104	27,424
Chlamydomonas reinhardtii (Merchant, Prochnik et al. 2007)	121	15,143
Micromonas sp RCC299 (Worden, Lee et al. 2009)	20.9	10,056
Arabidopsis thaliana (Arabidopsis Genome Initiative 2000)	115	25,498
Populus trichocarpa (Tuskan, Difazio et al. 2006)	485	45,555

What accounts for the high gene number of *G. theta* and *B. natans*? They do not possess multiple identical copies of genes. Is there something about these organisms that requires a level of protein complexity approaching that seen in metazoans and land plants? An immediate response might be that the presence of the nucleomorph entails additional maintenance and regulatory systems. Undoubtedly this is the case, but in our predictions for the PPC proteomes (Table 2.7), which is where the vast majority of these "additional" proteins would be predicted to be found, the number of genes involved is not sufficient to account for the high gene number of these two species compared to unicellular algae that either do not possess a PPC as in green lineages or have a simplified one as in photosynthetic chromalveolates.

One concern, especially for genomes from lineages with little to no genomic data from closely related species, is that a high percentage of the predicted genes are spurious, artifacts of the gene modeling process. Of the 24,840 protein coding genes in the *G. theta* catalog 10,455 (42%) do not have blastp hits against the NCBI protein database (e-value cutoff 1e-0.001) while for *B. natans* 8431 do not have blastp hits. However, just because a protein lacks a blastp hit does not mean that its corresponding gene is not real, especially for more obscure and poorly sampled lineages. A better measure of the likelihood of a gene being real is whether it is transcribed. Mapping RNA-Seq reads against gene transcripts found that 3824 gene models in *G. theta* and 539 in *B. natans* did not have any RNA-Seq support. However, even this analysis does not tell the whole story. While RNA-seq analysis is superior to traditional EST libraries in its ability to detect genes with low transcriptional activity it nevertheless can miss some. As well, some genes are not constitutively expressed so they may not be represented in total RNA samples depending on the conditions under which the RNA was extracted. In fact, of the 3824 *G. theta* genes that have no detectable transcription products 1965 have blastp hits against the NCBI protein database suggesting that they are real genes (Table 2.6). Therefore, in *G. theta* only 1859 gene models have no RNA-seq support and no matches against the current protein database. Interestingly, *B. natans* has far fewer gene models with no support from RNA-seq. If the 1859 gene models in *G. theta* with no RNA-seq coverage and no database hits are considered spurious this still leaves 22,981 genes that have some level of support as being "real". In *B. natans,* a similar calculation would suggest 21,465 gene models with some level of support.

Table 2.6. Gene models for *B. natans* and *G. theta* without blastp and/or RNA-seq support.

	No blastp hit	No RNA-Seq reads	No RNA-Seq reads with blastp hits	No RNA-seq reads and no blastp hits
G. theta	10,455	3824	1965	1859
B. natans	8,431	539	224	315

2.3.3 Proteome Predictions

Figure 2.1 presents the raw numbers for the proteins that were nucleus encoded and predicted to be targeted and operational in the four subcellular compartments of interest – mitochondrion, plastid, ER/Golgi and PPC/NM. There was concern that because the plastid and PPC targeted proteins must first be directed to the ER that it would be difficult to discern which proteins were targeted to which compartment. An analysis of the overlap between the four proteomes (Figure 2.1) revealed that relatively few proteins were predicted to be targeted to two or more compartments. These results would suggest that the four independent methodologies were reasonably robust in their predictions.

As was expected the most overlap in *B. natans* was between the PPC/NM and plastid proteomes. 152 proteins were independently predicted to be targeted to both the PPC and the plastid. This represents 12.6% of the PPC/NM proteins and 22% of the plastid targeted proteins. In *G. theta* the overlap between these two compartments was even less, only 51 out of 2574 PPC/NM proteins (2%) and 844 plastid proteins (6%). Plastid and PPC targeting is much better understood in those lineages with red algal plastids such as cryptophytes (Kilian and Kroth 2005; Gould, Sommer et al. 2006; Gould, Sommer et al. 2006). The plastid targeted protein predictions were done by a team that has worked extensively on this issue in diatoms, which also have red algal plastids. Consequently, they were able to use predictive methods developed for the diatom systems for *G. theta*. While some work has been done in chlorarachniophytes on plastid and PPC targeting (Hirakawa, Gile et al. 2010; Hirakawa, Burki et al. 2012), the body of knowledge is not as extensive, and what has been done suggests that they do not share the exact same strategies employed by lineages with red algal derived organelles for the importation of proteins to those organelles. So while it was possible to generate high and low confidence predictions for plastid targeting in *G. theta,* the same level of accuracy could not be attained for proteins putatively targeted to the plastid in *B. natans*.

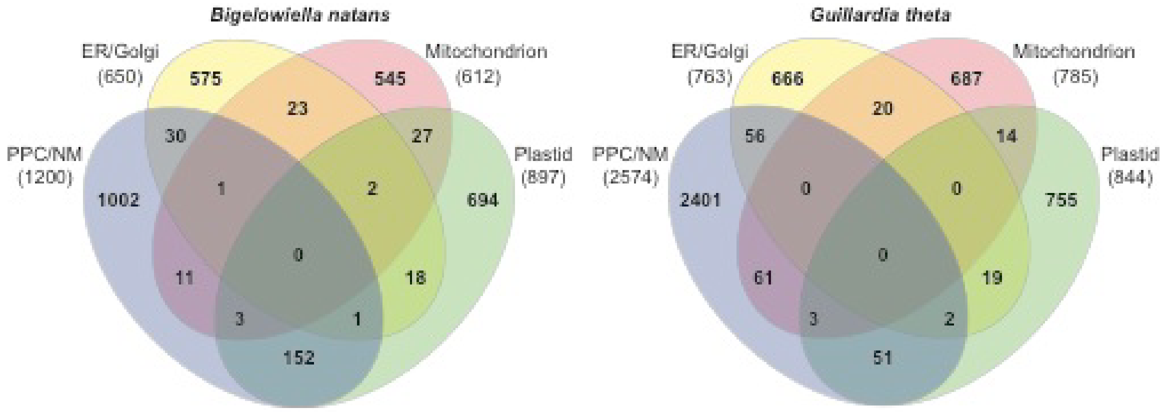


Figure 2.1. Total numbers and extent of overlap between the independently assembled PPC/NM, mitochondrial, plastid and ER/Golgi proteomes of *Bigelowiella natans* and *Guillardia theta*. The ER/Golgi values correspond to the values retrieved with 'ER-GolgiDB accuracy scores' of ≥75%. For *G. theta*, the plastid proteome values correspond to high-confidence proteins only. Figure prepared by Goro Tanifuji (Curtis et al. 2012. Nature. In Press).

Although it is possible for proteins to be dual targeted, the overlapping proteins were removed when reporting the final proteome numbers (Table 2.7) (Curtis et al. 2012 In press). It is likely that the overlaps represent artifacts of the predictive process and, where it was possible to decide, we assigned overlapping proteins to only one of the proteomes. For example, in *G. theta* 61 putative mitochondrial targeted proteins were also predicted to be targeted to the PPC/NM. An examination of the gene models, combined with the predicted functions of their proteins, revealed that all but one of the 61 should in fact be assigned to the PPC proteome.

2041 *G. theta* nucleus encoded proteins were predicted to be targeted to the periplastidial compartment (PPC) and the nucleomorph (NM) (Table 2.7). In *B. natans* the number of PPC/NM targeted proteins was considerably less, only 1001. For a discussion of the functional breakdown of the PPC targeted proteome and why *G. theta* appears to have twice as many PPC targeted proteins as *B. natans* see Curtis et al. 2012 In press. For BLAST analyses of the various targeted proteomes see Chapters 3.

Table 2.7. Predicted sub-cellular localization of nucleus encoded proteins for *B. natans* and *G. theta.*

	Mitochondrion	Plastid (High confidence)	ER/Golgi (> 75%)	PPC/NM
G. theta	687	755	666	2401
B. natans	545	694	575	1001

2.3.4 Blastp Analysis

Of the 24,880 *G. theta* proteins, 12,118 (48.7%) had blastp hits against the local database with an e value cutoff < 1e-10, while 11,384 of *B. natans* proteins had a blastp hit at the same e value cutoff (Table 2.8). Regardless of the e value cutoff, *B. natans* has 2-3% more blastp hits than *G. theta*. However, this slight difference in the percentage of identifiable genes for *B. natans* may be the result of *G. theta* having potentially more spurious genes. An analysis of RNA-Seq support for gene models, particularly those with protein sequences with no blastp hits (Table 2.6), suggested that 1859 of *G. theta* models might be artifactual vs. 315 from *B. natans*. If the total number of proteins are reduced by these amounts *B. natans* and *G. theta* have virtually the same percentage of blastp hits at each e value cutoff. For example, at a cutoff of 1e-40 *G. theta* has 5,332 proteins with hits against a total protein set of 22,981 (23.2%), while for *B. natans* the same calculation gives 23.1% (4966/21465).

Table 2.8. Number and percentage of blastp hits for *B. natans* and *G. theta* at 4 e value cutoffs.

	e-10 cutoff	e-20 cutoff	e-30 cutoff	e-40 cutoff
B. natans	11,384 (52%)	8,435 (39%)	6,403 (29%)	4,966 (23%)
G. theta	12,118 (49%)	8,893 (36%)	6,793 (27%)	5,332 (21%)

The taxonomic distribution of the best hits among major lineages was calculated. Many of the groups showed little change in their percentage of the top hits regardless of the e value cutoff (Table 2.9, Table 2.10). However, as the stringency of the similarity search increased the percentage of top hits to Metazoa and bacteria declined for both *B. natans* and *G. theta,* and increased for stramenopiles and Viridiplantae. In all cases *B. natans* had more top hits to stramenopiles, while in the case of *G. theta* the most top hits were to Viridiplantae . If all the chromalveolate lineages are grouped together *G. theta* does have more top hits to chromalveolate lineages considered as a whole than to Viridiplantae, but only slightly (Table 2.9, Table 2.10), while in *B. natans* the difference is considerable.

Table 2.9. Taxonomic distribution of blastp top hits for *B. natans.*

	-10 cutoff %	-20 cutoff %	-30 cutoff %	-40 cutoff %
Amoebozoa	3.9	3.9	4.0	4.2
Apusozoa	2.8	2.9	3.0	2.8
Metazoa	17.3	16.3	15.8	15.8
Choanoflagellate	4.1	4.3	4.8	4.9
Fungi	6.8	6.7	6.6	6.8
Excavata	3.9	3.7	3.2	2.8
Rhizaria	0.4	0.5	0.5	0.6
Rhodophyta	1.1	1.0	1.0	1.0
Viridiplantae	18.1	19.0	20.0	21.0
Glaucophyta	0.3	0.3	0.2	0.2
Alveolata	4.6	4.2	3.3	2.8
Cryptophyta	4.9	4.9	5.0	5.1
Haptophyta	4.8	4.9	4.6	4.1
Stramenopiles	20.5	21.5	22.5	23.0
Bacteria	9.3	8.9	8.7	8.4
Archaea	0.34	0.2	0.1	0.1

Table 2.10. Taxonomic distribution of blastp top hits for *G. theta.*

	-10 cutoff %	-20 cutoff %	-30 cutoff %	-40 cutoff %
Amoebozoa	3.7	3.7	3.6	3.7
Apusozoa	1.8	1.8	1.8	1.6
Metazoa	18.0	17.0	17.2	16.4
Choanoflagellate	2.8	2.6	2.6	2.7
Fungi	7.0	7.2	7.1	7.5
Excavata	4.1	3.4	3.2	3.0
Rhizaria	2.7	2.7	2.9	3.0
Rhodophyta	3.0	3.3	3.1	2.9
Viridiplantae	20.2	22.2	23.6	24.5
Glaucophyta	0.6	0.6	0.6	0.6
Alveolata	3.0	2.7	2.3	1.9
Cryptophyta	0.1	0.1	0.1	01
Haptophyta	4.8	4.9	4.4	3.8
Stramenopile	16.7	17.6	18.1	18.9
Bacteria	10.3	9.3	8.7	8.5
Archaea	0.5	0.2	0.1	0.10

Top hit BLAST analysis is prone to errors and over interpretation, especially since it is known that the top hit is not always the nearest neighbor (Koski and Golding 2001). A major concern is that the number of top hits for a given group is correlated with the number of entries in the database that was used (Stiller, Huang et al. 2009). I did not find a significant correlation between the taxonomic composition of my database and the top hits for *G. theta* and *B. natans*. Nevertheless it is obvious that at some level the number and type of hits is influenced by the database composition and that having few entries for a particular group like Rhodophyta will decrease the chance of retrieving hits. However, my research suggests that one reaches a saturation point beyond which the addition of more entries makes little difference in the number of hits returned for that particular group. My database included 3.6 million bacterial sequences, making up 55% of the entries, yet their contribution to the top hits was fairly stable at about 9% for both the *G. theta* and *B. natans* protein sequence sets.

Despite the concerns over a BLAST-based analysis, examination of the trends in the BLAST output allowed me to get a sense of the relationship that a particular genome has to the broader eukaryotic lineages. For example *B. natans* shows a greater preference compared to *G. theta* for 'chromalveolate' sequences (Figure 2.2), a result that perhaps is explained by recent shifts in the understanding of the phylogenetic relationship *G. theta* and *B. natans* have to the chromalveolates. As mentioned, the latest large-scale multigene analyses (Burki, Shalchian-Tabrizi et al. 2007; Burki, Shalchian-Tabrizi et al. 2008) seem to support a close relationship between rhizarians, alveolates and stramenopiles to the exclusion of cryptophytes (and possibly haptophytes as well). Some results have suggested that cryptophytes are closer to the primary plastid bearing lineages than was once thought (Burki, Shalchian-Tabrizi et al. 2007; Baurain, Brinkmann et al. 2010; Parfrey, Grant et al. 2010). This is perhaps reflected in *G. theta's* strong preference for viridiplantae sequences in BLAST analyses compared to *B. natans,* even though *B. natans* clearly has a green algal endosymbiont.

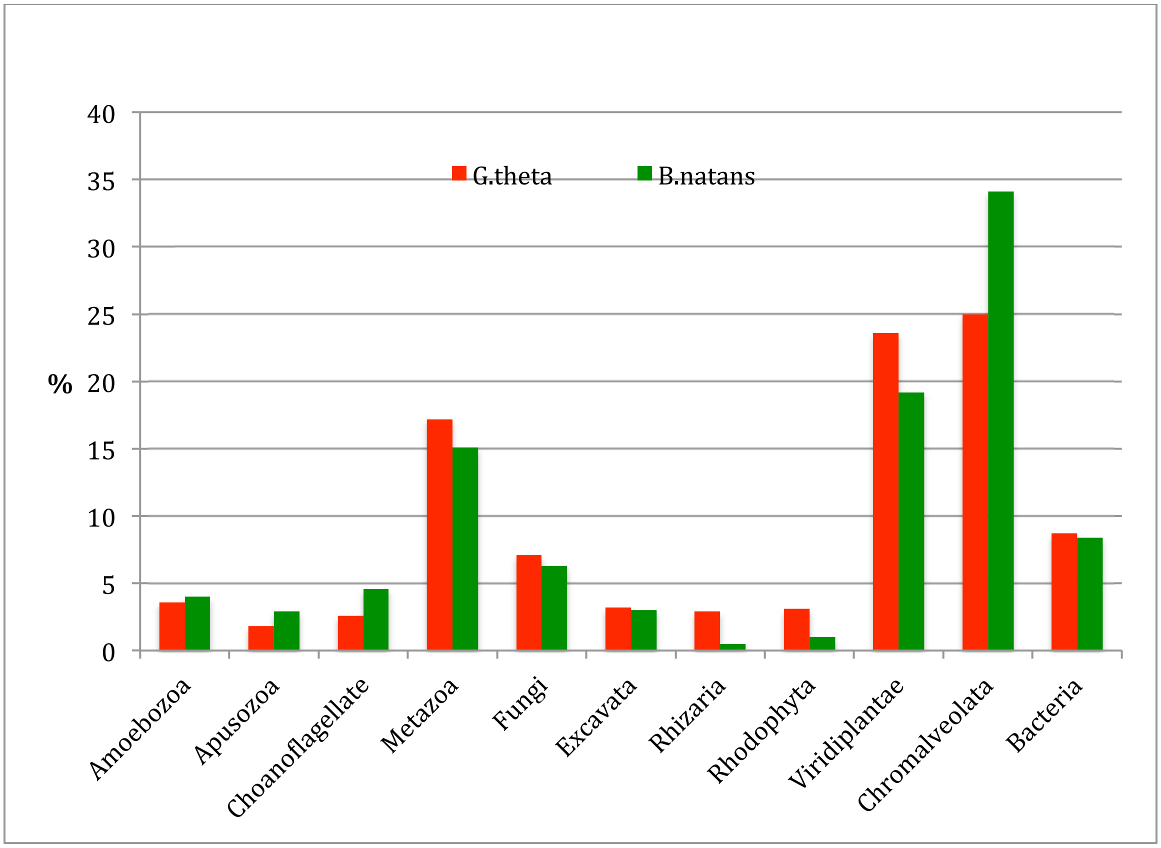


Figure 2.2. Taxonomic distribution of blastp top hits for *B. natans* and *G. theta* with an e value cutoff of 1e-30.

Beyond conveying trends, a BLAST output can be mined for interesting phylogenetic affinities that may be missed using a phylogenomic pipeline. Pipelines generally impose criteria to filter out the sequences and the corresponding trees that will probably not be useful. Most pipelines impose a minimum number of taxa and more importantly a minimum number of major lineages to ensure the tree has the potential to be phylogenetically meaningful. The initial stage of harvesting putatively homologous sequences is also influenced by choices that restrict the ultimate size and content of the tree. In almost all cases, candidate sequences for a particular tree are picked through BLAST analysis. There is considerable leeway in selecting which database to search against, especially if the database is created locally and populated with a curated set of sequences.

Most phylogenomic studies at some point raise the issue of taxon sampling and assume that with more data, especially from certain taxa, the relationships between the major lineages will become clearer and more firmly supported. However, having more data is not necessarily an advantage. The desire to include all of the available data is commendable but it needs to be weighed against practicality. This is why many of the most recent phylogenomic pipelines, including the one for *B. natans* and *G. theta* (Appendix A1-A3 Tree building protocols) (Curtis et al. 2012), restrict the amount of data from what are seen as redundant data sets. For example, our pipeline limited the number of bacterial sequences that would be used to create the initial trees using FastTree (Price, Dehal et al. 2010) to <9 cyanobacteria, <9 alphaproteobacteria and <5 other prokaryotes. The taxa were further pruned in the dereplication stage by limiting the number of entries from highly supported monophyletic subtrees including some eukaryotic ones like streptophytes. The recent analysis of the dinoflagellate *Alexandrium tamarense* (Chan, Soares et al. 2012) also used a protocol that restricted the number of entries from Metazoa and Fungi to <16 each, while the bacterial groups represented was limited to five.

This reduction in the sequences is done to make the creation and curation of trees tractable. The move towards more sophisticated and evolutionarily realistic methods of inferring phylogeny requires more computational power and the time taken to create trees from popular maximum likelihood packages like RAxML is very much dependent on the number of taxa included. Given that entire genome protein sets are analyzed, the desire to restrict the computational load is understandable. The other driving force behind scaling back the size of the entries is that there is an increasing recognition that a purely automated classification of trees from a genome wide analysis is prone to error and that at the very least each tree selected as supporting a conclusion needs to be examined manually. A study re-evaluating the surprising amount of green algal signal in diatom genomes (Moustafa, Beszteri et al. 2009) came to the conclusion that "Unfortunately, the time-consuming visual inspection of phylogenetic trees still remains far more accurate" (Deschamps and Moreira 2012) than automated sorting of massive phylogenomic studies. A similar conclusion was reached in a re-evaluation of the EGT signal in *Chomera velia* (Burki, Flegontov et al. 2012). In both cases a manual examination of the individual trees reduced the number of what they considered legitimate examples of EGT to less than 10% of the original conclusion. The leafier the tree, the more difficult it is to manually inspect, from the purely physical aspects of unwieldy large trees on computer screens to the increased likelihood that the tree will include paralogs or wayward sequences that disrupt monophyletic sub-trees.

Another concern with using phylogenomic studies to investigate EGT and especially LGT is that a large portion of the gene set is not even considered, and much evolutionary information is lost. For example, we used an e value cutoff of -25, which meant that the starting point for the automated portion of the phylogenomic pipeline was 7451 trees *for G. theta* and 6181 for *B. natans* representing about a third of the actual number of genes.

Many of those “missing” trees would be uninformative because they correspond to lineage specific proteins, but at an e value cutoff of e-10 there were 12,118 *G. theta* proteins sequences that had some measure of similarity with other known proteins. Lowering the threshold however generally does not work for phylogenomic studies because it permits the inclusion of sequences that tend to disrupt the phylogenetic signal, such as paralogs or sequences that share a single domain rather than homology across their entire sequence. Even with stricter cutoffs the number of trees that are thrown away because of their unresolved nature is significant, and as recent re-evaluations have indicated, even the winnowed sets of trees are open to different interpretations.

An analysis of EGT must be done against the backdrop of the full evolutionary history of the genome. The extent to which LGT has contributed genetic material can and should influence the assessment of EGT. The propensity with which an organism can uptake exogenous DNA and integrate it into its genome must be taken into account when determining the level of EGT in a particular lineage. For example, the choanoflagellate genomes that have been examined contain over a hundred genes of seemingly algal origin (Sun, Yang et al. 2010). No one, however, is seriously suggesting an algal endosymbiont for choanoflagellates. Instead, LGT is invoked as the likely route of acquisition, made possible because of the lifestyle of choanoflagellates as marine phagotrophs. Stiller suggests that genome-level analysis of EGT include in the experimental design a control genome to measure the extent of background LGT (Stiller 2011). This can be compared with the measure of EGT in the lineage in question to see if the EGT signal rises above the background LGT signal. While this suggestion is warranted, putting it into practice is problematic. Stiller suggests heterotrophic/phagotrophic groups like choanoflagellates as suitable controls because they are thought never to have possessed a plastid. Clearly however the choice of the control can greatly influence the outcome, since certain lineages are more prone to the acquisition of exogenous DNA due to feeding habits and preferences, or because of cellular/genetic conditions that are conducive to eukaryotic-eukaryotic transfers as in certain rotifers (Gladyshev, Meselson et al. 2008). Using the extent of LGT in the control lineage as a measure of LGT in organisms widely divergent evolutionarily seems challenging and prone to error because of lineage specific rates of uptake of exogenous material.

Through an analysis of BLAST results one can perhaps get a sense of the extent of LGT that is not possible for a phylogenomics based approach for some of the reasons outlined above. BLAST results are easy to produce and lend themselves to large scale automated parsing. One of the categories conducive to BLAST analysis but that could be overlooked by other methods is that of taxonomically exclusive hits, i.e., hits that only register against a particular taxonomic group. It is similar to the concept of gene sharing used in recent studies of red algae (Chan, Yang et al. 2011) and dinoflagellates (Chan, Soares et al. 2012). An analysis of the BLAST results for *G. theta* and *B. natans* against a local database (e value cutoff 0.001) generated a sizeable number of taxonomically exclusive hits (Table 2.11).

Apart from the results for Viridiplantae, Rhodophyta and the various chromalveolates, other examples of BLAST hits exclusive to a single lineage can be considered putative examples of LGT. Of course there is always the possibility that the disjunctive taxonomic distribution will disappear with increased taxon sampling. The results for Viridiplantae and Rhodophyta could be either EGT or LGT. The fact that *B. natans* has 214 cases of exclusive sharing with Viridiplantae lineages compared to 169 for *G. theta* implies that at least some of the signal in *B. natans* results from EGT from its green algal endosymbiont.

Table 2.11. Blastp results for *G. theta* and *B. natans* restricted to a single major lineage.

	B. natans	*G. theta*
Alveolata	73	58
Amoebozoa	24	39
Apusozoa	32	14
Archaea	4	8
Bacteria	169	169
Choanoflagellates	31	32
Euglenozoa	8	21
Fungi	52	50
Haptophyta	78	54
Heterolobosea	13	17
Metazoa	181	221
Parabasalia	3	12
Rhodophyta	16	24
Stramenopiles	232	227
Viridiplantae	214	169

Another category that lends itself to BLAST analysis is what I call singlets and doublets, those sequences that generate either one or two hits and thus are not be suitable for building a tree. Singlets are obviously a subset of the taxonomically exclusive set and in many cases, though not always, doublets are as well. To ensure that the BLAST hits have some level of meaning I used a cutoff of 1e-20. I identified 63 and 43 cases of singlets in *B. natans* and *G. theta* respectively while for doublets there were 15 cases for *G. theta* and 33 for *B. natans*. The taxonomic groups that generated these exclusive hits ranged across eukaryotic diversity (Table 2.12) rather than being confined to closely related lineages although even these are possibly the result of LGT and should be examined on a case by case basis.

Table 2.12. Protein sequences with single hits <1e-20 for *B. natans* and *G. theta.*

	B. natans	*G. theta*
Alveolata	4	1
Apusozoa	11	2
Cryptophytes	2	2
Euglenozoa	0	6
Fungi	2	0
Haptophytes	11	8
Planctomycetes	0	1
Chlamydiae	1	0
Proteobacteria	3	0
Fibrobacteres	1	0
Rhizaria	1	4
Rhodophyta	3	2
Stramenopiles	15	13
Viridiplantae	0	3
Choanoflagellates	5	1
Metazoa	2	0
Heterolobosea	2	0

The blastp result for *B. natans* protein 88157 illustrates the point that in conjunction with genome wide phylogenetic analysis a BLAST based approach is very useful. Protein 88157 is 516 amino acids long. Its gene consists of a single exon and based on SignalP 3.0 analysis has a strong signal peptide (HMM 0.943, NN 4/5). Near the 5′ end it has a type III fibronectin domain according to analysis by hmmpfam. The filtered gene model has EST support for the last half of the gene but no RNA-Seq support. In blastp searches against NCBI nr it has only two hits. The best hit is to a hypothetical protein (EGD75730) from the choanoflagellate *Salpingoeca* sp. ATCC 50818 with an e value of 7e-70 and an identity of 38% over 416 positions. The second hit is also to a hypothetical protein (XP_001750601) from the choanoflagellate *Monosiga brevicollis*, but with significantly reduced similarity and an e value of 5e-04.

Because of the low similarity to *Monosiga* this sequence did not pass the threshold for consideration and thus was considered as a singlet. Nor was there a tree generated since a tree based on two sequences is meaningless. Nevertheless, the gene is real and its presence in *B. natans* and the choanoflagellate tells us something about each organism, though what it says is difficult to discern. The strong similarity and highly disjunctive distribution suggests LGT, but the direction of transfer is unclear. Choanoflagellates are known for acquiring algal genes (Sun, Yang et al. 2010) so one would normally assume that the transfer is from *B. natans* to the choanoflagellate. However, the strong differences in similarity shown by the two choanoflagellates would suggest either two independent transfers from *B. natans* with a considerable time difference or transfer of a divergent choanoflagellate gene to *B. natans*. A search of the RNA-Seq data from two other chlorarachniophytes, *Lotharella globosa* and *Lotharella oceanica*, failed to find any similar genes, implying that it is unique to *B. natans,* although it should be noted that *B. natans* did not have any RNA-Seq support for this gene.

A case of lateral gene transfer in *G. theta* underscores the value of searching against all available sequences rather than against limited local databases that for phylogenomic purposes pare down seemingly redundant bacterial sequences. *G. theta* has three similar hypothetical genes -76818, 83219 and 104313. In blastp searches their protein sequences return only four hits, all to marine cyanobacteria, all with strong similarity (Table 2.13). Based on the RAxML tree (Figure 2.3) the most likely scenario is that after lateral gene transfer from cyanobacteria, the *G. theta* gene was duplicated, with one of the copies, G104313, experiencing an increased rate of evolution while the other copy was duplicated again leading to the two very similar genes 76818 and 83219. One of the more interesting aspects of the example is that while NCBI has at least nine *Synechococcus* sp. strains only two of them have this gene. If, to reduce the overwhelming abundance of bacterial sequences, some of the *Synechococcus* strains had been left out of the database used for phylogenetic purposes, the chance that this relationship would be missed is high.

Table 2.13. Blastp results for G. *theta* protein sequences 76818, 83219, 104313.

	Acaryochloris sp. CCMEE 5410	*Acaryochloris marina* MBIC11017	*Synechococcus* sp. PCC7335	*Synechococcus* sp. PCC7002
G76818	3e-106	4e-106	3e-105	4e-99
G83219	6e-107	5e-106	1e-110	3e-103
G104313	4e-37	1e-36	7e-38	2e-35

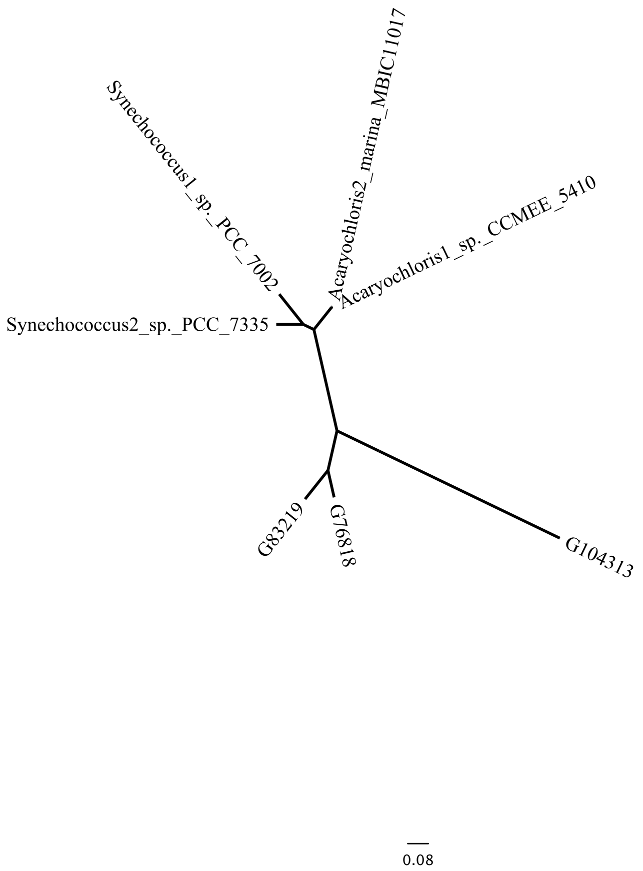

Acaryochloris2_marina_MBIC11017
Synechococcus1_sp._PCC_7002
Acaryochloris1_sp._CCMEE_5410
Synechococcus2_sp._PCC_7335
G83219
G76818
G104313
0.08

Figure 2.3. Unrooted RAxML tree of *G. theta* protein sequences 76818, 83219, 104313.

The final category of LGT signal that could be missed if single gene trees are automatically categorized is what I call chasm results. These are cases where there may be multiple hits to various lineages but the similarity of the top hits to the query sequence is considerably higher than is the case for the rest of the hits. These examples can be deduced by looking for drop-offs in e values of various magnitudes and for various groups (Table 2.14). Some of these results, particular for ‘chasm20’ (i.e., e value differences of 1e-20), may simply be taxon sampling issues. The results for chromalveolates may reflect vertical inheritance. It is interesting to note that *B. natans* has more cases of this type in relation to chromalveolate sequences while *G. theta* has more results for matches with Viridiplantae. For the other groups, such as bacteria, these are potential LGT cases. Even within the plastid bearing groups large gaps in the similarity profile of the hits can indicate LGT. Refer to Figure 3.1 for an example in which a *B. natans* protein sequence is very similar to several diatoms, but based on the tree, it might be interpreted as vertical inheritance.

Table 2.14. Blastp results for *B. natans* and *G. theta* that display large differences in the similarity shown between top hits and lower ranked hits. Chasm20 denotes a drop in similarity as measured by e values of at least 20 (e.g., 1e-60 <-> 1-40).

	Chasm 20		Chasm 30		Chasm 40	
	G. theta	*B. natans*	*G. theta*	*B. natans*	*G. theta*	*B. natans*
Chromalveolates	396	484	213	238	98	130
Viridiplantae	229	174	123	66	78	36
Bacteria	50	61	23	25	14	12
Rhodophyta	11	2	5	1	2	0
Fungi	7	5	0	2	0	2
Excavata	31	19	14	7	7	3
Metazoa	43	69	19	37	10	25
Amoebozoa	12	16	1	5	0	4

2.3.5 Phylogenomic Analysis

As indicated above, a phylogenomic analysis was conducted for *B. natans* and *G. theta*. We conceived and implemented a pipeline that was designed to avoid some of the issues identified in recent re-evaluations of EGT signals in various lineages (Curtis et al. 2012 In press; Appendix A1-A3 Tree building protocols). Although the process of generating the RAxML trees was very much a team effort, the sorting and evaluation of the trees was to a large extent conducted by Fabien Burki. Consequently, I will only briefly discuss the results.

Using an automated tree sorting method (Chan, Yang et al. 2011) 611 trees for *B. natans* were considered potentially indicative of EGT and with high bootstrap support (> 80%) while for *G. theta* 846 trees were identified using the same criteria. After manual examination and classification we were left with 351 trees for *B. natans* and 508 for *G. theta*. These were further classified according to green, green with no red, red, red with no green, glaucophyte, ambiguous, and only Plantae (Curtis et al 2012 In press). The largest category of putatively algal-derived genes for *G. theta* was deemed “ambiguous”, with 147 trees classified as too messy to make any firm conclusions. For *B. natans* 100 trees fell into the “ambiguous” category. These results are similar to what was found for the re-evaluation of the green signal in diatoms. A rigorous manual examination of the trees reduced the putative EGT signal from 1757 trees to 286, of which 104 were classified as unresolved (Deschamps and Moreira 2012).

Such ambiguous trees, both in our study and others, presumably represent a small portion of the truly messy trees that did not even make it through for manual examination and classification because they exhibited little coherence in their phylogenetic signal. Some of the conditions that create these noisy trees even within the winnowed set of putative EGTs have been mentioned already such as presence of paralogs, LGT disrupting monophyletic sub-trees, trees that have too few taxa to be informative, and ‘algal’ with restricted taxonomic distribution.

In the case of *B. natans* and *G. theta* the presence of the nucleomorph probably contributes to the unresolved nature of many of the trees generated for these two organisms. Although other secondary plastid bearing lineages, like diatoms, have PPCs (Gould, Sommer et al. 2006) they are highly reduced compared to the ones in *B. natans* and *G. theta* that putatively have ~1200 proteins in the case of *B. natans* and ~2500 in *G. theta* (Figure 2.1). The "extra" PPC proteomes impose an additional layer of complexity that may result in the production of paralogs to service and maintain two eukaryotic cytosols or the propensity to retain two copies of the same gene, one from the host lineage and one from the algal endosymbiont.

I have deliberately not included many trees in this dissertation because invariably the individual tree is unresolved, or if the major lineages are recovered the bootstrap support for it is woefully inadequate. For example, Deschamps and Moreira present their Figure 2A as a model of a well resolved and well supported tree (Deschamps and Moreira 2012). Their tree does not include the homologous protein sequences for this ER lumen receptor from *B. natans* and *G. theta*. Figure 2.3 shows the corresponding RAxML tree with the inclusion of the *G. theta* sequence and the two *B. natans* paralogs. The major lineages are not resolved, with Viridiplantae interrupted by various chromalveolates while bootstrap support, except for near the terminal nodes, is extremely weak. One of the *B. natans* trees is equally unresolved with the alveolates separated from the stramenopiles, although Viridiplantae is recovered. In fairness it should be noted that the Deschamps and Moreira tree is not as resolved as they claim since *Dictyostelium* is well supported in the chromalveolate subtree and the two red algal sequences are not monophyletic (Deschamps and Moreira 2012).

The *G. theta* and *B. natans* protein sequences will be of interest to researchers focused on questions of higher level eukaryotic diversity because they fill gaps in the eukaryotic tree. Moreover, while the position of rhizarians on the eukaryotic tree of life seems to have stabilized based on recent large-scale studies, cryptophytes and their relationship to the other lineages remains problematic. Unfortunately, based on my experience with the phylogenetic analysis of *G. theta* in particular, the availability of thousands of new protein sequences will not quickly or easily resolve the uncertain taxonomic placement of cryptophytes.

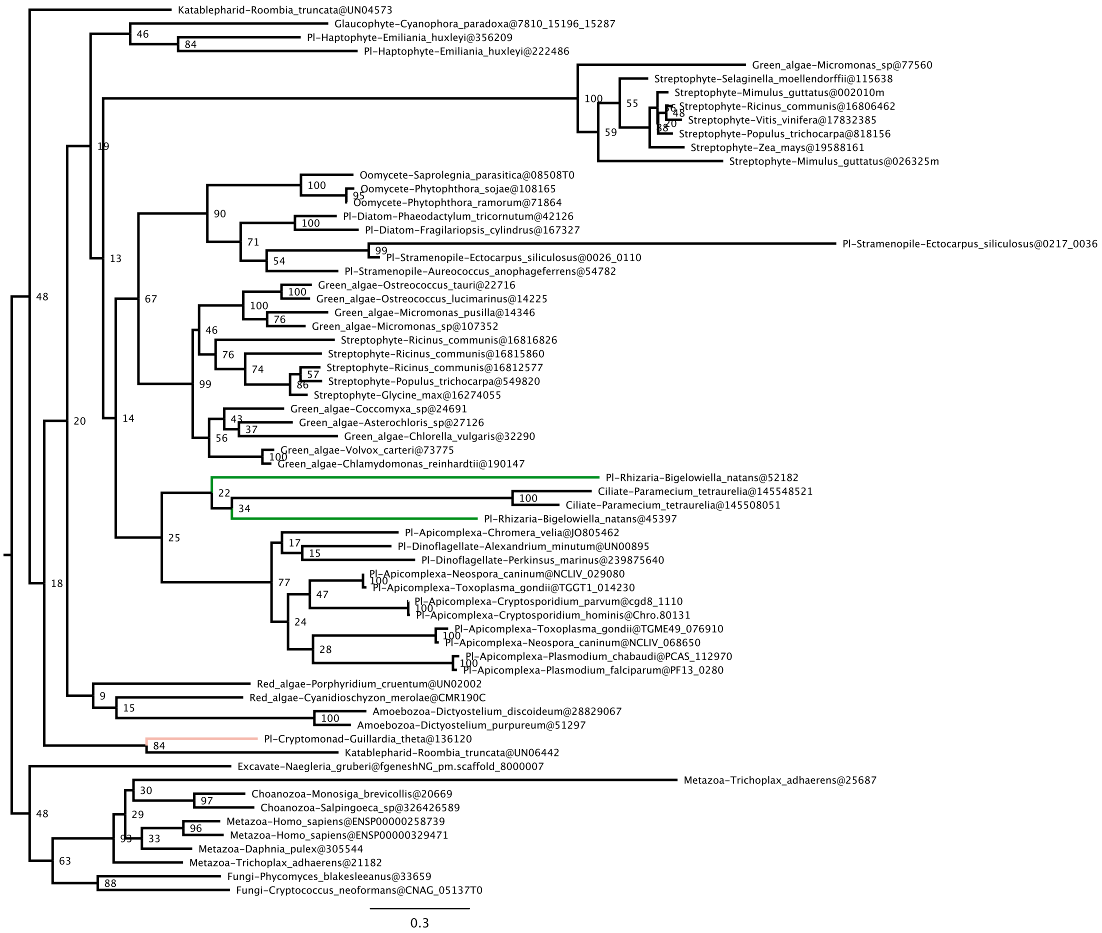

Katablepharid-Roombia_truncata@UN04573
Glaucophyte-Cyanophora_paradoxa@7810_15196_15287
46
Pl-Haptophyte-Emiliania_huxleyi@356209
84
Pl-Haptophyte-Emiliania_huxleyi@222486
Green_algae-Micromonas_sp@77560
Streptophyte-Selaginella_moellendorffii@115638
100
55
Streptophyte-Mimulus_guttatus@002010m
56
Streptophyte-Ricinus_communis@16806462
48
Streptophyte-Vitis_vinifera@17832385
20
38
Streptophyte-Populus_trichocarpa@818156
59
Streptophyte-Zea_mays@19588161
Streptophyte-Mimulus_guttatus@026325m
19
Oomycete-Saprolegnia_parasitica@08508T0
100
Oomycete-Phytophthora_sojae@108165
95
Oomycete-Phytophthora_ramorum@71864
90
Pl-Diatom-Phaeodactylum_tricornutum@42126
100
Pl-Diatom-Fragilariopsis_cylindrus@167327
71
99
Pl-Stramenopile-Ectocarpus_siliculosus@0217_0036
13
Pl-Stramenopile-Ectocarpus_siliculosus@0026_0110
54
Pl-Stramenopile-Aureococcus_anophageferrens@54782
Green_algae-Ostreococcus_tauri@22716
100
48
67
Green_algae-Ostreococcus_lucimarinus@14225
100
Green_algae-Micromonas_pusilla@14346
76
Green_algae-Micromonas_sp@107352
46
Streptophyte-Ricinus_communis@16816826
76
Streptophyte-Ricinus_communis@16815860
74
Streptophyte-Ricinus_communis@16812577
57
Streptophyte-Populus_trichocarpa@549820
86
99
Streptophyte-Glycine_max@16274055
Green_algae-Coccomyxa_sp@24691
43
20
14
Green_algae-Asterochloris_sp@27126
37
56
Green_algae-Chlorella_vulgaris@32290
Green_algae-Volvox_carteri@73775
100
Green_algae-Chlamydomonas_reinhardtii@190147
Pl-Rhizaria-Bigelowiella_natans@52182
22
Ciliate-Paramecium_tetraurelia@145548521
100
Ciliate-Paramecium_tetraurelia@145508051
34
Pl-Rhizaria-Bigelowiella_natans@45397
Pl-Apicomplexa-Chromera_velia@JO805462
25
17
Pl-Dinoflagellate-Alexandrium_minutum@UN00895
15
Pl-Dinoflagellate-Perkinsus_marinus@239875640
Pl-Apicomplexa-Neospora_caninum@NCLIV_029080
77
100
18
Pl-Apicomplexa-Toxoplasma_gondii@TGGT1_014230
47
Pl-Apicomplexa-Cryptosporidium_parvum@cgd8_1110
100
Pl-Apicomplexa-Cryptosporidium_hominis@Chro.80131
24
Pl-Apicomplexa-Toxoplasma_gondii@TGME49_076910
100
Pl-Apicomplexa-Neospora_caninum@NCLIV_068650
28
Pl-Apicomplexa-Plasmodium_chabaudi@PCAS_112970
100
Pl-Apicomplexa-Plasmodium_falciparum@PF13_0280
Red_algae-Porphyridium_cruentum@UN02002
9
Red_algae-Cyanidioschyzon_merolae@CMR190C
15
Amoebozoa-Dictyostelium_discoideum@28829067
100
Amoebozoa-Dictyostelium_purpureum@51297
Pl-Cryptomonad-Guillardia_theta@136120
84
Katablepharid-Roombia_truncata@UN06442
Excavate-Naegleria_gruberi@fgeneshNG_pm.scaffold_8000007
Metazoa-Trichoplax_adhaerens@25687
30
Choanozoa-Monosiga_brevicollis@20669
97
Choanozoa-Salpingoeca_sp@326426589
48
29
Metazoa-Homo_sapiens@ENSP00000258739
96
93
33
Metazoa-Homo_sapiens@ENSP00000329471
Metazoa-Daphnia_pulex@305544
63
Metazoa-Trichoplax_adhaerens@21182
Fungi-Phycomyces_blakesleeanus@33659
88
Fungi-Cryptococcus_neoformans@CNAG_05137T0
0.3

Figure 2.4. RAxML tree for *G. theta* 136120. The *G. theta* branch is indicated in red while the two B. *natans* branches are in green. Bootstrap values are shown at the nodes.

CHAPTER 3 THE MITOCHONDRIAL PROTEOMES OF *BIGELOWIELLA NATANS* AND *GUILLARDIA THETA*

This chapter includes work published in Bruce A. Curtis et al. 2012. Algal nuclear genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature* 492: 59-65.

3.1 INTRODUCTION

3.1.1 Function of the Mitochondrion

Numerous biochemical pathways and processes have been linked to the mitochondrion. These include the metabolism of lipids, nucleotides, amino acids and carbohydrates as well as the synthesis of heme and Fe-S clusters (Burger, Gray et al. 2003). Chief among the roles ascribed to mitochondria is energy production, and they are generally described as the powerhouse of the cell. Energy is produced by oxidative phosphorylation where electrons are transported via the electron transport chain to oxygen, and in the process creating a proton gradient that is used to convert ADP to ATP (van Hellemond, van der Klei et al. 2003). However, aerobic ATP synthesis is not ubiquitous among eukaryotes. Anaerobically functioning mitochondria exist that rely on terminal electron acceptors other than oxygen such as fumarate (Tielens, Rotte et al. 2002). As well, some organisms, such as yeast, can rely on cytosolic fermentation for energy requirements in the absence of oxygen. Hydrogenosomes, which are considered derived mitochondria, can also generate ATP without a membrane bound electron transport chain (van der Giezen 2009).

If oxidative phosphorylation or even energy production are dispensable as mitochondrial functions, is there any role that mitochondria play that appears to be absolutely essential? Mitosomes, which are highly reduced mitochondria without genomes (Embley, van der

Giezen et al. 2003), have reduced metabolic functions. Among the few things that mitosomes do, that is also done by all classical mitochondria, is Fe-S cluster formation (Tovar, Leon-Avila et al. 2003), leading to the speculation that this is truly the only essential function of mitochondria (Tovar 2007).

3.1.2 Origin of Mitochondria

It is now indisputable that present day mitochondria originated from a symbiotic relationship with a bacterium (Margulis 1970). Several lines of evidence also strongly suggest that the symbiotic relationship occurred once. It is possible that multiple instances of bacterial symbiosis took place with varying degrees of reduction and integration, but present day eukaryotic organisms very likely all derive from the same ancient lineage that acquired a bacterial endosymbiont that over time underwent reductive evolution, generating the organelle we call a mitochondrion. Phylogenetic analysis of ribosomal RNA mitochondrial encoded genes consistently support a monophyletic origin (Cedergren, Gray et al. 1988). Although there is considerable variation among eukaryotes in the number of genes that are mitochondrial encoded, the genes that are present all come from a very small group of 107 possible genes (Gray 2012). It is unlikely that losses in independently derived mitochondrial organelles would have converged on such a small set, even smaller than what is seen for plastids. Intriguingly, all of the known gene complements of mitochondrial genomes are a subset of the mitochondrial genome of the obscure jakobid flagellate *Reclinomonas americana*. Its mitochondrial genome is seen as the closest thing we have to the protomitochondrion (Lang, Burger et al. 1997). Indeed, the *R. americana* mitochondrial genome possesses characteristics, like operon-like gene clusters and Shine-Dalgarno motifs, that are bacterial-like (Lang, Burger et al. 1997). Operon-like gene clusters also provide what is perhaps the strongest evidence for a monophyletic origin of mitochondria. Some plant and protist mitochondrial genomes share with bacterial operons the same order of ribosomal genes (Gray 1999).

Phylogenetic analysis of mitochondrial encoded rRNA seemed to pinpoint the likely source of the bacterial endosymbiont as coming from the alpha-proteobacterial class (Schnare and Gray 1982). More specifically, phylogenetic analysis seemed to converge on the order Rickettsiales (Viale and Arakaki 1994; Andersson, Zomorodipour et al. 1998), a group of obligate intracellular parasites, as the modern day descendants of the protomitochondrion. While there seems little doubt that the mitochondrial progenitor was an alpha-proteobacterium some have questioned whether it can or should be narrowed down to Rickettsiales or even further to the family Rickettsiaceae. It has been pointed out that Rickettsiales, as obligate parasites, are prone to high rates of sequence divergence and bias toward AT richness (Esser, Ahmadinejad et al. 2004), genomic conditions that also prevail in mitochondria, leading to long-branch-attraction artifacts in phylogenetic reconstructions. Because of these concerns, researchers have expanded their search for the source of the mitochondrion to include examples of free-living alpha-proteobacteria (Giovannoni, Tripp et al. 2005; Brindefalk, Ettema et al. 2011). Regardless of which order or family is most appropriate they are all alpha-proteobacterial.

The traditional view of the establishment of the mitochondrion is that a eukaryotic cell engulfed and retained an alpha-proteobacterial cell. Over time the bacterial endosymbiont lost functions and genetic material no longer necessary for its permanent role inside a eukaryotic organism (Embley and Martin 2006). This view of mitochondrial symbiosis is known as the archezoan scenario. Recently however, a different scenario for the establishment of the mitochondrion has been gaining ground (Embley and Martin 2006). In this alternate process, rather than a primitive eukaryotic cell engulfing a bacterial cell, an archaeal cell acquired an alpha-proteobacterial endosymbiont that led to not only a mitochondrion but the eukaryotic lineage as well (Martin and Muller 1998). The synergistic relationship between the archaeal host and its bacterial endosymbiont is posited as having been the driving force behind the creation of eukaryotic cells with their complexity and compartmentalization.

The classical view of the acquisition of the mitochondrion has recently met with doubt. The archezoan scenario is predicated upon the existence of an amitochondrial eukaryotic lineage that engulfed and retained an alpha-proteobacterium. The existence of present day amitochondrial lineages such as diplomonads and microsporidians seemed to bolster this scenario especially since phylogenetic analysis placed these lineages without mitochondrion at the base of the eukaryotes (Cavalier-Smith 1987). However, more recent work has demonstrated that these lineages are neither basal to the rest of the eukaryotes nor are they amitochondrial (Embley and Martin 2006). While not possessing classical mitochondria with the typical complement of functions and pathways, these organisms do have structures that are clearly derived from the mitochondrion. These mitochondrial related organelles (MROs), like mitosomes and hydrogenosomes, are reduced mitochondria. Although they do not possess DNA, their structures and biochemical pathways betray their origin. Intermediate forms have also been found that straddle the traditional concept of what a mitochondrion does and what MROs do, along with having a genome (Stechmann, Hamblin et al. 2008) albeit reduced.

Although reports of present day descendants of the amitochondriate eukaryotes have proven incorrect, the fact that they may not actually exist is not necessarily fatal to the classical view. They may yet be discovered. More likely though, these lineages died out, especially if the acquisition of a mitochondrion by related organisms conveyed a substantial evolutionary advantage. It has been suggested (Lane and Martin 2010) that the energy requirements of a typical eukaryotic cell can only be met by mitochondria. If that is the case then it is understandable that no true amitochondriate lineages survived.

Problems also exist for the alternate view that the mitochondrion arose out of an alpha-proteobacterial endosymbiont in an archaebacterial host. In this scenario the bacterial imprint in modern eukaryotes should be overwhelmingly alpha-proteobacterial. While phylogenetic analysis does register a considerable alpha-proteobacterial presence it is not as dominant as one would expect (Pisani, Cotton et al. 2007). One could argue of course that rampant HGT among bacteria diluted the pure alpha-proteobacterial lineage prior to the establishment of the endosymbiont (Richards and Archibald 2011).

In its richest conceptualization the alternate view of mitochondrial acquisition posits that the primitive eukaryotic cell forged by an archaeal host and an alpha-proteobacterial endosymbiont possessed both aerobic and anaerobic respiration (Martin and Muller 1998). This dual energy source could be utilized depending on the particular environmental conditions. Importantly, the anaerobic respiratory pathway of MROs is seen as a vestige of that primitive adaptability. However, phylogenetic analysis of the proteins involved in anaerobic metabolism in MROs does not support a monophyletic origin (Hug, Stechmann et al. 2010) but instead suggests that the MROs scattered across eukaryotic diversity arose independently and any commonalities result from convergent evolution.

3.1.3 Proteomics

Regardless of whether the progenitor of the mitochondrion was engulfed by a primitive eukaryote or a prokaryote soon to be a eukaryote, the creation of the organelle from the endosymbiont resulted in massive gene loss. *Reclinomonas americana* has the largest mitochondrial encoded gene set. Yet its 107 gene set, 67 of which code for proteins, is miniscule compared to a typical bacterium that has thousands of proteins. It is estimated that the free-living alpha-proteobacterial ancestor of the mitochondrion possessed 3000-5000 genes (Boussau, Karlberg et al. 2004). Many of the genes were simply lost but a large portion was transferred to the host genome. Once transferred and rendered transcriptionally active these genes could then generate proteins that could function in the mitochondrial compartment. Because the mitochondrion has significantly reduced functions compared to a free-living bacterium all 3000-5000 genes are not required.

Nevertheless, mitochondrial proteomes are predicted to contain at least 1500 proteins in vertebrate animals (Meisinger, Sickmann et al. 2008).

From the relatively few mitochondrial proteomes that have been characterized we know that not all the mitochondrial targeted proteins originate from the alpha-proteobacterial endosymbiont. Only 10-15% of the yeast proteome displays unequivocal alpha-proteobacterial origins (Karlberg, Canback et al. 2000). A further 40-50% of the proteins appear to be bacterial in nature without any clear indication of where in the bacterial kingdom they come from. The rest appear to be eukaryotic inventions with no discernible bacterial homologs. It is also clear from phylogenetic analysis of genome protein sets that not all the genes transferred from the endosymbiont to the host are targeted back to where they came from. An estimated 800 human genes are derived from the alpha-proteobacterial lineage, presumably mainly from the protomitochondrion, yet only 200 of them encode proteins that operate in the mitochondrion (Szklarczyk and Huynen 2010). The rest have been coopted to function elsewhere in the cell such as fatty acid oxidation in the peroxisome (Gabaldon, Snel et al. 2006).

Not only is there variation in the functions carried out in mitochondria across the breadth of eukaryotic diversity but the proteins and protein complexes that perform the functions, even relatively ubiquitous ones like oxidative phosphorylation, show significant lineage specific differences. Some of the earliest and most detailed work on mitochondrial protein complexes comes from yeast research (Sickmann, Reinders et al. 2003). When those same complexes are studied in other organisms a substantial number of the subunits turn out to be limited to fungi. These results belie the need for mitochondrial proteomic studies that cover the full spectrum of eukaryotic diversity. Without sampling from myriad lineages and doing comparative analyses, speculation about the nature and makeup of the protomitochondrion is likely to be overly conservative and biased. The nuclear genome projects for *Bigelowiella natans* and *Guillardia theta* gave me an opportunity to investigate the nucleus-encoded components of the mitochondrial proteomes from two poorly sampled lineages.

Early investigations of mitochondrial proteomes relied heavily on bioinformatics approaches. The main tools were homology searches against known mitochondrial proteins and assessments of mitochondrial targeting signals using prediction software like TargetP (Emanuelsson, Brunak et al. 2007). Homology searches can do an adequate job identifying a set of proteins that are invariably found in the mitochondrion and reasonably well conserved like those proteins involved in oxidative phosphorylation. However, a high percentage of mitochondrial proteins are novel and thus will fail to be detected. A comparison of the yeast mitochondrial proteome against the human, two of the best and most reliable datasets, found that only 58% of the yeast proteins were present in the human set (Gabaldon and Huynen 2004). What is clear from the proteomes that have been analyzed is that outside a very narrow taxonomic range there are considerable differences between major lineages in their mitochondrial protein sets. Novel pathways have been gained while typical pathways have been lost. If pathways have been transferred to new subcellular locations then they may be falsely identified as being found in the more typical location. Mitochondrial protein complexes also show considerable lineage specific differences in the number and nature of subunits outside the conserved, core proteins and these ancillary subunits will again be undetected in homology searches unless it includes closely related taxa.

The complexity of the eukaryotic cell necessitates the compartmentalization of functions as with oxidative phosphorylation in mitochondria. To ensure that nucleus encoded proteins are directed to the appropriate compartment they possess targeting signals. Mitochondrial associated proteins are no exception. They possess targeting signals that interact with import complexes in the two mitochondrial membranes. The signals are generally of two types, N-terminal presequences and internal motifs. The study and use of presequences has proven more tractable from a bioinformatics standpoint (Imai and Nakai 2010) and has spawned a number of publicly available programs like MitoProt (Claros and Vincens 1996), PSORT (Nakai and Horton 1999), iPSORT (Bannai, Tamada et al. 2002), TargetP (Emanuelsson, Brunak et al. 2007), Predotar (Small, Peeters et al. 2004) and SubLoc (Hua and Sun 2001). Although there is no consensus motif for the presequences they have biochemical characteristics sufficiently similar enough to allow varying levels of detection. These N-terminal peptides that are in most cases cleaved after import consist of 10-90 amino acids and generally have the ability to form positively charged amphiphilic helical structures (Imai and Nakai 2010). Attempts to define a cleavage site consensus have only been moderately successful with a test of the putative motifs R2 and R3 achieving 21 and 33% (Vogtle, Wortelkamp et al. 2009) accuracy using MitoProt (Claros and Vincens 1996) and TargetP respectively (Emanuelsson, Brunak et al. 2007).

Undoubtedly, the identification of N-terminal mitochondrial targeting peptides can be successful. However, the algorithms used for their detection are not sufficient to generate a comprehensive catalog of the mitochondrial proteome. Even for well-studied organisms the rates of false negatives and positives are of concern. Most tend to generate inflated lists of putatively targeted proteins. 4975 proteins were predicted to be mitochondrial targeted in *Arabidopsis* (Heazlewood, Tonti-Filippini et al. 2004) using the predictor iPSORT while TargetP generated a list of 3182. 1940 of the iPSORT predictions were unique to it while 613 of the TargetP predictions were unique. The actual number of mitochondrial-targeted proteins in *Arabidopsis* is estimated to be between 2000 and 3000 (Heazlewood, Tonti-Filippini et al. 2005) but this is only a guess. The mouse mitochondrial proteome has been estimated to contain ~1130 proteins (Pagliarini, Calvo et al. 2008) while others have cast doubt on this low number believing that 1500 is a more realistic estimate (Meisinger, Sickmann et al. 2008). For yeast the best guess is 1000 proteins (Reinders, Zahedi et al. 2006).

All of the current prediction programs have been trained on the typical set of model organisms like human, yeast and *Arabidopsis* and do not necessarily address the specific protein characteristics of more obscure lineages like cryptophytes and chlorarachniophytes. One interesting exception, HECTAR (Gschloessl, Guermeur et al. 2008), which is designed specifically for heterokonts, is unfortunately not amenable in its current web format to global surveys of protein sets. The reliability of the gene models is also of concern since the detection of N-terminal target peptides obviously requires correct and full-length gene predictions. Predicted protein sets from poorly sampled lineages are again at a disadvantage. Gene modeling programs in eukaryotes, with the complications of introns and exons, can do an adequate job at detecting the general presence of a gene but find it more difficult to determine the starts and stops correctly without additional information like homologous genes from closely related organisms. Since target peptides are not highly conserved compared to the rest of the protein the lack of close homologs can make it especially difficult to determine the correct start. Finally, a large portion of mitochondrial targeted proteins do not rely on classical N-terminal target peptides (Smith, Gawryluk et al. 2007; Pagliarini, Calvo et al. 2008) and consequently will fail to be detected in global surveys of protein sequences as being part of the mitochondrial proteome.

Using several different programs can increase the specificity of the predictions, especially when the programs use different protein characteristics and are trained on different protein sets. However, requiring putative positives to meet the thresholds of all the programs used lowers sensitivity since the false negatives will be combined (Heazlewood, Tonti-Filippini et al. 2004).

Because of concerns over the inability to detect novel proteins, poorly conserved proteins or proteins without classical targeting signals, non-bioinformatics approaches have been employed to analyze mitochondrial proteomes. The most popular technique has been tandem mass spectrometry (MS/MS) of peptides isolated from mitochondria. MS/MS has been used to study mitochondrial associated proteins in the typical model organisms like yeast (Sickmann, Reinders et al. 2003; Reinders, Zahedi et al. 2006) and *Arabidopsis* (Heazlewood, Tonti-Filippini et al. 2004). Recent years have also seen MS/MS studies in more unusual lineages or species that may have more relevance to investigating *B. natans* and *G. theta* like the green alga *Chlamydomonas reinhardtii* (Atteia, Adrait et al. 2009) or the ciliate *Tetrahymena thermophila* (Smith, Gawryluk et al. 2007). MS/MS studies are not without their own set of problems that reduce the ability to comprehensively identify mitochondrial proteins. Low abundance proteins are generally missed, as are proteins with few suitable tryptic peptides (Pagliarini, Calvo et al. 2008). Of most concern is the difficulty of isolating pure organellar fractions and the resulting contamination and detection of proteins not actually part of the mitochondrial proteome (Sickmann, Reinders et al. 2003; Smith, Gawryluk et al. 2007). MS/MS studies have reported up to 41% false positive rates (Pagliarini, Calvo et al. 2008).

Another important technique for identifying the subcellular localization of proteins is green fluorescent protein (GFP) tagging (Pagliarini, Calvo et al. 2008). This method allows direct visualization of protein localization using fluorescence microscopy. However, GFP tagging is relatively time consuming and expensive for doing whole proteomes when compared to MS/MS. As well, it requires a genetically tractable system, a condition not often met in less studied organisms such as *B. natans* and *G. theta*.

I have attempted to identify those nuclear genes that code for proteins destined for the mitochondrion in *B. natans* and *G. theta*. The techniques I relied on were purely bioinformatic. Consequently, the results should not be considered definitive or comprehensive but as a starting point. Novel constituents of a proteome, by their very nature, are missed during homology searches. While target signal predictors possess the capacity to detect these unique additions, the uncertainty of gene models reduces one's confidence in these designations. Only through a combination of bioinformatic inference and experimental studies can we hope to gain a firmer grasp on the actual makeup of the mitochondrial proteomes of *B. natans* and *G. theta*.

3.2 Methods and Materials

Mitochondrial target peptides

Each protein in the complete filtered protein sets for *B. natans* and *G. theta* was assessed for targeting to the mitochondrial compartment using three different sub-cellular targeting prediction methods, TargetP, Predotar and iPSORT. Version 1.1 of TargetP (Emanuelsson, Brunak et al. 2007) was run locally as was iPSORT (Bannai, Tamada et al. 2002), while Predotar (Small, Peeters et al. 2004) results were generated through a webserver (http://urgi.versailles.inra.fr/predotar/predotar.html). Although many sub-cellular localization prediction tools are available (Imai and Nakai 2010), TargetP, Predotar and iPSORT were chosen because they are widely used (especially TargetP), are available either for local installation or through a webserver, do not require data transformation, and employ different methods as well as different training sets. Proteins with positive mitochondrial sub-cellular localization predictions for all three programs were retained for downstream analysis.

To assist in the manual curation of the putative mitochondrial targeted proteins a local database was created consisting of mitochondrial proteomes from the following species: mouse; human; *Arabidopsis thaliana*; *Acanthamoeba castellanii*; *Chlamydomonas reinhardtii*; *Tetrahymena thermophila*; *Saccharomyces cerevisiae*. The human and mouse mitochondrial proteomes were obtained from MitoCarta (Pagliarini, Calvo et al. 2008) which is a curated inventory of mammalian mitochondrial genes. The mouse mitochondrial proteome consisted of 1098 protein coding genes while there were 1344 genes for human (Table 3.1). The MitoCarta inventory was created through a combination of mass spectrometry of mitochondria obtained from fourteen different human tissues and assessment of subcellular localization through GFP tagging/microscopy work. The *Arabidopsis* mitochondrial proteome was obtained from SUBA, an *Arabidopsis* subcellular database (Heazlewood, Tonti-Filippini et al. 2005). The subcellular localization predictions in SUBA have been created through various methods including mass spectrometry, GFP tagging, Swiss-Prot annotations and literature searches. To match the stringency of the MitoCarta data, I downloaded from SUBA sequences whose proteins had been identified as being mitochondrial through mass spectrometry (535 entries) and GFP tagging (185 entries). After removing redundancies this left 676 *A. thaliana* mitochondrial sequences.

Table 3.1. Protein sequences from selected mitochondrial proteomes.

	Protein sequences	source
Human	1344	MitoCarta
Mouse	1196	MitoCarta
Arabidopsis thaliana	676	SUBA
Acanthamoeba castellanii	743	R. Gawryluk, Pers. Comm.
Tetrahymena thermophila	573	(Smith, Gawryluk et al. 2007)
Saccharomyces cerevisiae	851	(Reinders, Zahedi et al. 2006)
Chlamydomonas reinhardtii	262	(Atteia, Adrait et al. 2009)

The 573 *Tetrahymena thermophila* protein sequences were obtained from Ryan Gawryluk and correspond to the proteins identified in a tandem mass spectrometry analysis (Smith, Gawryluk et al. 2007). Also obtained from Ryan Gawryluk were 743 protein sequences that were identified from *Acanthamoeba castellanii* using a combination of tandem mass spectrometry and bioinformatic analysis. The yeast mitochondrial proteome is based on LC-MS/MS and MALDI-MS of 1D-SDS-PAGE and 2D-PAGE isolations (Reinders, Zahedi et al. 2006). The mitochondrial proteins for *Chlamydomonas reinhardtii* were also identified by LC-MS/MS (Atteia, Adrait et al. 2009). The putative mitochondrial proteins for *B. natans* and *G. theta* were compared against this local database using blastp (e value cutoff 0.00001).

The putative mitochondrial targeted proteins were also searched against a local protein database that in addition to the non-redundant proteins available through NCBI also had protein datasets derived from genome projects and translated ESTs (Appendix B Local database entries). The top hit for each protein was also added to the spreadsheet for each species. Putative mitochondrial protein IDs were cross-referenced with JGI generated KOG classifications and these assignments were added to the spreadsheets.

Examination of the preliminary results revealed that the putative list of mitochondrial targeted proteins lacked some entries for proteins that are known to be unequivocally mitochondrial targeted and their absence in the mitochondria of *B. natans* or *G. theta* would be highly surprising. Consequently, directed searches of particular genes using a combination of parsing JGI generated annotation results and homology searches were used to identify various proteins and add them to the list of putative proteins. In other cases more detailed searches for particular motifs were employed. Details of the directed searches are interlaced with the results since a description of them divorced from their context would be difficult to follow.

3.3 Results and Discussion

3.3.1 Mitochondrial Carrier Proteins

Mitochondrial carrier proteins are an important and ubiquitous family of proteins targeted to the mitochondrion that do not typically possess presequences amenable to detection by subcellular localization predictors such as TargetP (Emanuelsson, Brunak et al. 2007). Of the 47 annotated mitochondrial carrier proteins in mouse (Pagliarini, Calvo et al. 2008) only 6% have target peptides as detected by TargetP, while in humans only 9% of the 43 carrier proteins appear to have target peptides (Table 3.2). This is in contrast to the entire proteomes for which 62% of the mouse proteins have TargetP detected presequences while in human 63% have TargetP mitochondrial designations.

Table 3.2. Mitochondrial target peptides detected by TargetP from mouse and human mitochondrial proteomes.

	# Mt targeted proteins	# Mt targeted proteins with target peptides	# Mt carrier proteins	# Mt carrier proteins with target peptides
Mouse	1097	679 (62%)	47	3 (6%)
Human	1022	639 (63%)	43	4 (9%)

Mitochondrial carrier proteins are typically identified by searching for conserved motifs and domains combined with homology against known carrier proteins. Classically, mitochondrial carrier proteins (MCPs) possess three unique motifs – P-x-[DE]-x-x-[RK] (proline, any amino acid, either aspartic acid (D) or glutamic acid (E), any amino acid, any amino acid, either arginine (R) or lysine (K)). Because the conserved motifs form cross-links essential to the functioning of the protein it is thought that possessing three motifs is required (Palmieri, Pierri et al. 2011). In particular, the proline is vital since it is the only amino acid with a bend thus allowing the membrane cavity to be opened and closed (Pebay-Peyroula, Dahout-Gonzalez et al. 2003).

The occurrence of the motif in human and mouse protein sets, along with *B. natans* and *G. theta* was investigated using the EMBOSS (Rice, Longden et al. 2000) utility fuzzpro.

Table 3.3. Proteins containing the motif Px[DE]xx[RK] from whole genome protein sets.

# of motifs	Mouse	Human	*B. natans*	*G. theta*
0	20945	23220	16758	19821
1	5748	6951	3901	4025
2	1476	1850	786	768
3	473	506	179	158
4	136	157	53	45
>4	93	115	31	23

The mouse genome has 473 proteins that contain three of the conserved motifs considered diagnostic for MCPs (Table 3.3) but only 47 proteins that are annotated as MCPs. The human genome shows a similar disconnect between the number of motifs present and the actual number of MCPs. Clearly the conserved motif is insufficient to detect *bona fide* MCPs. Moreover, investigations of yeast MCPs reveal that the motif is not always conserved. The proline, which was long considered to be absolutely essential, has been replaced in several yeast proteins by a serine (Nury, Dahout-Gonzalez et al. 2006). It is proposed the defects to the hinge mechanism caused by the mutation of a proline to a serine are compensated for by additional mutations that retained the kink in the structure.

Since the mitochondrial carrier motif was insufficient and misleading when trying to identify MCPs I used the hmmpfam (Krogh, Brown et al. 1994) annotations generated by the JGI annotation pipeline. The hmmpfam (PF00153) Mito_carr family model has an average domain length of 94.4 amino acids. A search of the JGI hmmpfam annotations for proteins with at least one Mito_carr domain found 92 for *G. theta* and 82 for *B. natans*. Although the carriers require three domains to function it was necessary to investigate all the proteins with at least one domain because the models may be incorrect and domains may be missed due to truncated models and/or incorrect exon/intron assignment. Each of the possible MCPs was analyzed for transmembrane regions using TMpred (http://www.ch.embnet.org/software/TMPRED_form.html) and presequence targeting peptides using SignalP 3.0 (Bendtsen, Nielsen et al. 2004) and Targetp 1.1 (nonplant) (Emanuelsson, Brunak et al. 2007). Each possible MCP was searched against the NCBI protein database to assess the following: homology with known MCPs; appropriate length and start sites consistent with homologous proteins; number of mitochondrial carrier domains in homologous proteins; non-canonical carrier motifs.

For potential carrier proteins with less than three PFAM carrier domains, alternative gene models were examined as well as subjecting the intergenic space surrounding the model to blastx analysis against the NCBI protein database to find undetected carrier domains.

Table 3.4. Predicted mitochondrial carrier proteins for *G. theta*.

Protein id	TargetP[1]	Domains	Motif	Type
101972		3	3	
106541		3	3	
107383		3	3	Rim2p/Mrs12p
107986		3	4	S-adenosylmethionine
113675		3	0	
113677		3	3	
116227	M	3	3	S-adenosylmethionine
116850		3	1	
120626		3	3	
121795		3	2	fatty acid anion
133395	M	3	3	
137702	M	3	3	
137950		3	4	S-adenosylmethionine
141004		3	3	
143999		3	2	phosphate
147022		3	2	
147042		3	3	oxoglutarate/malate

Protein id	TargetP[1]	Domains	Motif	Type
152990		3	2	
154373		3	3	
154481	M	3	3	
157347		3	2	oxoglutarate/malate
158030		3	3	
158243		3	2	
158297		3	1	
162408		3	2	carnitine-acylcarnitine
162598		3	1	oxaloacetate
162689		3	2	FAD
163898		3	3	S-adenosylmethionine
45511		3	2	FAD
50187		3	1	
58369		3	2	oxodicarboxylate
62947		3	3	
63299		3	2	phosphate
65536	M	3	3	fatty acid anion
66106		3	1	
68495		3	0	FAD
72577		3	1	
73487	M	3	3	
74274		3	1	S-adenosylmethionine
74767		3	3	
76084		3	2	carnitine-acylcarnitine
77427		3	3	
77466		3	2	oxoglutarate/malate
81659		3	3	carnitine-acylcarnitine
84773		3	3	
85396		3	3	
90757		3	1	phosphate
96015		3	3	
97273		3	3	S-adenosylmethionine

[1] M for TargetP column denotes a predicted mitochondrial target peptide.

Table 3.5. Predicted mitochondrial carrier proteins for *B. natans*.

Protein id	TargetP[1]	Domains	Motif # and non-canonical motif	Type
22199		3	3	carnitine-acylcarnitine
22200		3	3	
33449		3	2	FAD
36786		3	1 PMNYWK PLELVM	carnitine-acylcarnitine
37107		3	3	oxoglutarate/malate
37538		3	3	fatty acid anion
38339	M	3	4	S-adenosylmethionine
40277		3	4	ADP/ATP
41058		3	2	
42229	M	3	2 PADTLL	ADP/ATP
42249		3	2	phosphate
42974		3	2	phosphate
44307		3	2	oxoglutarate/malate
46386		3	3	fatty acid anion
46877		3	2	ADP/ATP
48941	M	3	3	
52610		3	2	
54313		3	4	
54478		3	4	tricarboxylate/dicarboxylate
55259		3	2	MRS3/4
55371		3	3	tricarboxylate/dicarboxylate
56244		3	3	ADP/ATP
56295	M	3	2	
57139		3	3	S-adenosylmethionine
62830	M	3	3	
65246		3	3	
68888	M	3	3	fatty acid anion
70128		3	2 PLDMMQ	
71932		3	2	
74728	M	3	3	S-adenosylmethionine
77004		3	2	oxoglutarate/malate
231832		3	3	
78420		3	2	MRS3/4
79628	M	3	5	PET8
80457		3	2 PLELIM	
81815		3	3	carnitine-acylcarnitine
82136	M	3	2	carnitine-acylcarnitine

Protein id	TargetP[1]	Domains	Motif # and non-canonical motif	Type
82212	M	3	3	
82286		3	3	
82633		3	2	
83248		3	1 PNSVIK PTDIIA	
85988		3	2	FAD
86023		3	2	oxodicarboxylate
87043		3	2	
87389		3	2	
87791	M	3	3	
88309		3	1 PFFVLK PFDLIG	oxaloacetate
89615		3	2	carnitine-acylcarnitine
89988	M/SP	3	3	tricarboxylate/dicarboxylate
91130		3	3	oxaloacetate
91153		3	3	carnitine-acylcarnitine
92708		3	3	phosphate
126166		3	3	
126372		3	4	tricarboxylate/dicarboxylate
128582		3	3	oxoglutarate/malate
133560		3	2	
139013 combined with 139012		3	3	
143618		3	3	carnitine-acylcarnitine
146641	M	3	2 PADTIL	phosphate

[1] M for TargetP column denotes a predicted mitochondrial target peptide.

After examination of the 92 *G. theta* MCP candidates I was left with 49, while of the 82 candidates for *B. natans* I was left with 59 (Table 3.4, 3.5). These numbers are well within the range calculated for other genomes, particularly plants. A recent survey of MCPs (Palmieri, Pierri et al. 2011) found 58 in *Arabidopsis thaliana*, 125 in *Glycine max*, 60 in *Sorghum bicolor*, 73 in *Zea mays*. Of the few photosynthetic unicellular genomes examined, *Ostreococcus lucimarinus* had 38 while *Chlamydomonas reinhardtii* had 37. Six of the *G. theta* candidates, and five in *B. natans*, were rejected because they possessed strong signal peptide predictions. Indeed, one of the complications of MCP studies is establishing subcellular localization for each protein. While the name suggests that they are invariably mitochondrial bound, localization of MCPs has been demonstrated in plastids (Bedhomme, Hoffmann et al. 2005; Bouvier, Linka et al. 2006; Kirchberger, Tjaden et al. 2008; Palmieri and Pierri 2010), peroxisomes (Fukao, Hayashi et al. 2001; Arai, Hayashi et al. 2008; Eubel, Meyer et al. 2008; Linka, Theodoulou et al. 2008) and the ER (Leroch, Neuhaus et al. 2008). In both *B. natans* and *G. theta* any proteins localized to the plastid also require ER signal peptides, so candidates with reasonable signal peptide scores were not kept as mitochondrial MCPs. Several were also rejected because in homology searches they were deemed to be peroxisomal. The rest of the rejections were due to candidates not matching the classical configuration sufficiently. As previously mentioned, proper functioning depends on having three domains that can crosslink so any candidates with less than three were rejected. Often the inability to detect three domains resulted from poor gene models and/or genomic gaps. Sometimes candidates with only two domains as predicted by hmmpfam were retained as legitimate MCPs if the gene model had sufficient room for an additional domain and the important prolines were conserved in homology searches. These cases demonstrated that the motif Px[DE]xx[RK] is not always present (Table 3.5) even in proteins that have been demonstrated to be MCPs in other species. Six of the 49 (12%) and 13 of the 59 (22%) MCPs in *G. theta* and *B. natans* respectively had mitochondrial target peptides predicted. Why these protists should have more MCPs with classical target peptides, especially *B. natans*, than seen in mouse or human MCPs is not clear. In one case (bn89988) the N terminal extension had scores for both a signal peptide and a mitochondrial target peptide opening the possibility of dual targeting. A similar case of ambiguous presequences that generate scores indicating both mitochondrial and plastid targeting was detected in two MCPs from *Arabidopsis thaliana* and *Zea mays* (Bahaji, Ovecka et al. 2011). The study confirmed through GFP fusions as well as immunocytochemical analyses that the two nucleotide transporters were indeed targeted to both organelles.

3.3.1.1 Phylogenetic analysis of mitochondrial carrier proteins

The mitochondrial carrier proteins were compared against a local database comprised of all the publicly available bacterial genomes (2040 taxa) using blastp. None of the *G. theta* MCPs had any hits to alpha-proteobacteria using an e-value threshold of 0.01. This result is not especially surprising since MCPs have long been considered exclusively eukaryotic. Curiously, six of the *B. natans* MCPs had low level hits (4e-04 to 3e-06) to the same hypothetical protein, LLO_3082, from the gamma-proteobacterium *Legionella longbeachae* NSW150. In the 2010 genome paper for this bacterium the authors mention that they found “eukaryotic-like and eukaryotic domain proteins” (Cazalet, Gomez-Valero et al. 2010) without explicitly mentioning carrier proteins. However, in 2012 another group described the surprising secretion of a mitochondrial carrier protein in the related bacterium *Legionella pneumophila* (Dolezal, Aili et al. 2012).

Legionella bacteria are intracellular pathogens that reside within a membrane bound vesicle from which they secrete at least 275 effector proteins that interfere with a number of cell functions (Dolezal, Aili et al. 2012). They also recruit mitochondria to their vesicles. The 2012 paper demonstrated that the putative MCP could be targeted to the yeast mitochondrial membrane and functioned to alter levels of ATP in the cytosol. The authors suggest that *Legionella* acquired these exclusively eukaryotic proteins through lateral gene transfer from a eukaryotic host. Interestingly, in their search for bacterial homologs they also found two putative mitochondrial carrier proteins that are encoded in the *Neorickettsia sennetsu* genome. *N. sennetsu* is also an intracellular pathogen and more importantly, a member of the Rickettsiales, from the order alpha-proteobacteria that has long been considered as the likely source of the protomitochondrion.

What is one to make of this handful of bacterial MCPs? Is it merely a case of LGT promoted by the intracellular lifestyles of the bacteria with these surprising genes, or does it suggest a possible source of this important, varied and ancient protein family in eukaryotes? In my investigation of *B. natans* MCPs I identified two additional putative bacterial MCPs, both with 3 domains, from *Fluoribacter dumoffii* (ZP_10138214) and *Legionella drancourtii* (ZP_09620034), both from the order Legionellales. Clearly, the bacterial MCPs are associated with intracellular parasitism, which is the lifestyle of many alpha-proteobacteria.

If one looks at the RAxML trees (data not shown) for the MCPs they are, with a few exceptions, typical of the trees one would expect for genes that have been vertically inherited. The trees are nevertheless complicated by paralogy. Many trees have multiple entries from *B. natans* and/or *G. theta*, with some of the paralogs grouping together suggesting recent duplications and other paralogs in separate clades suggesting much more ancient duplication. These observations match those seen in a recent review of mitochondrial carriers (Palmieri, Pierri et al. 2011). They concluded that the main MCP families originated prior to the diversification of eukaryotes into the major lineages followed by duplications within each major lineage.

3.3.1.2 LGT of mitochondrial carrier protein

The RAxML tree for bn36786 (Figure 3.1) presents a fairly typical phylogenetic picture for mitochondrial carrier proteins –several paralogs interspersed with green and red algae, and a smattering of stramenopiles. The tree arrangement has bn36786 grouping with two diatoms, which is not particularly unusual. However, the tree does not convincingly portray the extent to which bn36786 is related to the two diatom genes.

The BLAST results reflect more accurately what I believe to be a case of LGT. In a blastp analysis against the NCBI protein database bn36786 had a top hit against *Thalassiosira pseudonana* with an e-value of 2e-117 (Table 3.6). This best hit was followed by two other highly similar matches to diatoms. A *B. natans* paralog (bn80457) is the next best hit but with considerably less similarity to bn36786 than the diatoms show. The paralog is followed by hits to two choanoflagellates at much higher e-values (4e-60, 3e-56) followed by more stramenopiles, green algae and fungi. The large gap in similarity scores shown by bn36786 to the diatom genes versus the other homologs, as well as a paralog, is best explained by LGT. Given the strong hits to two choanoflagellates, known for having an abundance of algal genes present in their DNA (Nedelcu, Miles et al. 2008), the transfer of this gene to other lineages is entirely possible. The most likely scenario is that the paralog bn80457 is the typical gene of this

MCP present in diatoms, oomycetes, haptophytes, green algae and fungi. Since a RNA-Seq contig from another chlorarachniophyte, *Lotharella globosa,* displays an identical similarity pattern, at some point an ancestor of *B. natans* acquired this gene from a diatom. Transfer from a diatom to a chlorarachniophyte is the likely route rather than the reverse since chlorarachniophytes have a mixotrophic lifestyle (Hibberd and Norris 1984). Incidentally, among the blastp hits for bn36786 was one for *Fluoribacter dumoffii* at 9e-26 and *Legionella drancourtii* at 2e-17. These hits are both gamma-proteobacteria, from the order Legionellales (see discussion above).

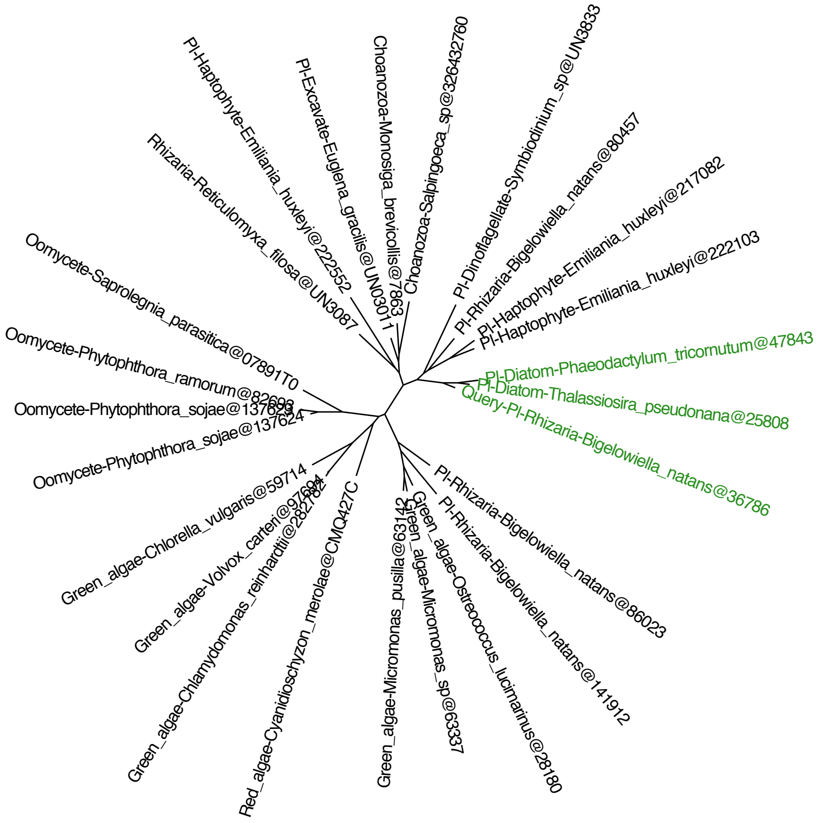

PI-Haptophyte-Emiliania_huxleyi@222552
PI-Excavate-Euglena_gracilis@UN03011
Choanozoa-Monosiga_brevicollis@7863
Choanozoa-Salpingoeca_sp@326432760
PI-Dinoflagellate-Symbiodinium_sp@UN3833
PI-Rhizaria-Bigelowiella_natans@80457
PI-Haptophyte-Emiliania_huxleyi@217082
PI-Haptophyte-Emiliania_huxleyi@222103
PI-Diatom-Phaeodactylum_tricornutum@47843
PI-Diatom-Thalassiosira_pseudonana@25808
Query-PI-Rhizaria-Bigelowiella_natans@36786
PI-Rhizaria-Bigelowiella_natans@86023
PI-Rhizaria-Bigelowiella_natans@141912
Green_algae-Ostreococcus_lucimarinus@28180
Green_algae-Micromonas_sp@63337
Green_algae-Micromonas_pusilla@63142
Red_algae-Cyanidioschyzon_merolae@CMQ427C
Green_algae-Chlamydomonas_reinhardtii@282782
Green_algae-Volvox_carteri@97694
Green_algae-Chlorella_vulgaris@59714
Oomycete-Phytophthora_sojae@137624
Oomycete-Phytophthora_sojae@137629
Oomycete-Phytophthora_ramorum@82693
Oomycete-Saprolegnia_parasitica@07891T0
Rhizaria-Reticulomyxa_filosa@UN3087

Figure 3.1. Unrooted RAxML tree for bn36786. The branch in green contains bn36786.

Table 3.6. Blastp results for the *B. natans* protein bn36786.

Hit species	Gene id	E value	identity
Thalassiosira pseudonana	XP_002294965	2e-117	65%
Phaeodactylum tricornutum	XP_002182348	4e-114	59%
Thalassiosira oceanica	EJK77999	1e-113	65%
Bigelowiella natans	bn80457	3e-72	42%
Monosiga brevicollis	XP_001745291	4e-60	39%
Salpingoeca sp.	EGD78330	3e-56	44%
Phytophthora sojae	EGZ15369	7e-37	34%

3.3.2 Iron Sulfur Cluster Formation Machinery

While the typical mitochondrial function is oxidative phosphorylation it has become increasingly clear that the true indispensible function of mitochondria is the formation of iron-sulfur clusters (Wiedemann, Urzica et al. 2006). A number of organisms have reduced or vestigial mitochondria collectively known as mitochondrial related organelles (MROs). Most MROs lack their own DNA unlike typical mitochondria and do not carry out oxidative phosphorylation. Instead, their primary purpose appears to be the formation of iron-sulfur clusters that are essential to other cellular processes. The formation pathway typically consists of 12 proteins. I was able to identify all 12 genes that code for these proteins in both *B. natans* and *G. theta*. A number of the proteins have dual roles. Erv1, mentioned below, also functions in the import of proteins into the mitochondrion, as does the chaperone Mge1 (see the section on import machinery).

Table 3.7. Mitochondrial Iron-Sulfur cluster proteins for *B. natans.*

Protein	Id	Top hit	E value	notes
Cysteine desulfurase (Nfs1)	55226	*Ostreococcus*	0.0	
	135083	*Aureococcus*	1e-60	
Isd11	49904			
NifU	37883	*Volvox*	1e-57	
	84228	*Danio*	3e-56	
Isu1	53822	*Arabidopsis*	2e-54	
Isa1	48748	*Glaucocystis*	1e-32	
Ferredoxin/Yah1	26430	*Trichoplax*	2e-40	
	141113	*Ectocarpus*	1e-18	
	146787	*Emiliania*	1e-11	
Ferredoxin reductase (Arh1)	70052	*Chlorella*	9e-81	
Frataxin (Yfh1)	63105	*Tetrahymena*	1e-09	
	87672	*Ectocarpus*	1e-20	alternative model
Glutaredoxin (Grx5)	85503	*Aureococcus*	1e-41	
Hsp70 (Ssq1)	92341	*Ochrobactrum*	0.0	
Hsc20 (Jac1)	83963	*Arabidopsis*	2e-10	
GrpE (Mge1)	43108	*Nasonia*	2e-36	
ATM1	73761	*Magnetospirillum*	1e-109	
Erv1	48701	*Taeniopygia*	3e-36	
	57894	*Proterospongia*	1e-25	

Table 3.8. Mitochondrial Iron-Sulfur cluster proteins for *G. theta.*

Protein	Id	Top hit	E value	notes
Cysteine desulfurase (Nsf1)	75525	*Drosophila*	0.0	alternative model
Isd11	40411			
NifU	158274	*Chlamydomonas*	2e-53	
Isu1	92053	*Homo*	1e-54	LPPVK motif
Isa1	154876	*Glaucocystis*	6e-39	
Ferredoxin (Yah1)	72919	*Phytophthora*	2e-39	
	91827	*Phytophthora*	2e-24	
	102078	*Bombus*	7e-07	
	158561	*Fragilariopsis*	6e-38	
Ferredoxin reductase (Arh1)	68742	*Phytophthora*	7e-124	alternative model
Frataxin (Yfh1)	118094	*Debaryomyces*	3e-20	alternative model
Glutaredoxin (Grx5)	91230	*Bigelowiella*	3e-30	
	96730	*Apis*	7e-33	
	158122	*Ectocarpus*	2e-15	
Hsp70 (Ssq1)	79059	*Chlamydomonas*	0.0	
Hsc20 (Jac1)	107346	*Volvox*	3e-26	
GrpE (Mge1)	152067	*Monodelphis*	3e-30	
Atm1	63916	*Physcomitrella*	2e-179	
Erv1	42578	*Aureococcus*	5e-20	

Many of the cellular structures and complexes that require iron-sulfur clusters are found in places other than the mitochondrion. Consequently, after formation, the clusters need to be exported. Proteins dedicated to this export have been identified. I was able to find Atm1, an ABC transporter, in *B. natans* and *G. theta* as well as the protein Erv1.

3.3.3 Identification of Erv1 Genes

The sulfhydryl oxidase, Erv1, found in the mitochondrial intermembrane space is essential for the maturation of iron-sulfur cluster proteins exported to the cytosol from the mitochondrial matrix (Lill and Muhlenhoff 2005). It also plays a role, in concert with Mia40, in the importation of cysteine containing intermembrane space proteins (Chacinska, Koehler et al. 2009). Erv proteins are characterized by a conserved core of ~ 100 AAs that contains two cysteine motifs, CysXXCys and CysX16Cys. Both motifs form disulfide bonds that are essential for stabilization of the protein (CysX16Cys) or the active site (CysXXCys) (Endo, Yamano et al. 2010). Additionally, outside of the core domain, Erv proteins possess an additional disulfide bond that appears to shuttle electrons to the active site disulfide bond (Endo, Yamano et al. 2010). While Erv proteins are found in all mitochondrial membranes they are also found in the ER. In yeast, Erv1p denotes the mitochondrial version while the ER protein is designated Erv2p (Gerber, Muhlenhoff et al. 2001). Although both yeast proteins are small and possess a single conserved domain they are not especially similar (35% identity confined to the core domain).

B. natans has three proteins that contain an Erv domain (43859,48701, 57894). To help determine which are likely to be mitochondrial and which are ER bound (if any) I performed a blastp analysis of all three against the yeast Erv1/2 proteins hoping that one or more of the *B. natans* protein sequences would clearly be more similar to the yeast Erv1 than it was to the Erv2 yeast sequence. Against Erv1 the scores were as follows:

48701 2e-33; 43859 6e-31; 57894 3e-16 (Table 3.9). The scores for Erv2 were equally indeterminate: 43859 4e-25; 48701 4e-23, 57894 7e-10.

A presequence targeting analysis was complicated by the gene models being especially poor. None of the PASA generated models based on RNA-Seq reads matched the filtered models or the alternative models. After taking into account coverage and length of homologous proteins, the presence of necessary motifs, and guided by RNA-Seq mapping, I determined what I believed to be the 'correct' coding sequence for each protein. Using TargetP 1.1 (Emanuelsson, Brunak et al. 2007) 43859 returned a high score for ER targeting, while 57894 generated a high score indicating neither a signal nor a target peptide (Table 3.9). The TargetP 1.1 results for 48701 indicated weak support for mitochondrial targeting. Like most mitochondrial membrane proteins Erv1 proteins tend to return a TargetP evaluation of "other" rather than M (signifying mitochondrial target presequence) since they do not have N-terminal target peptides.

I also looked at the arrangement of the cysteine motifs in the proteins. The yeast Erv1 has two CXXC motifs followed by the longer motif of CX16C. In contrast, yeast Erv2 has the two short motifs flanking the longer one (Ang and Lu 2009). In *B. natans* 48701 and 57894 have similar arrangements to Erv1 while in 43859 the motifs most closely resemble that seen in Erv2 though it should be noted that the C-terminal short motif is CX6C and not CXC, CXXC or CX4C as has been reported in other Erv proteins (Ang and Lu 2009).

It is unclear how diagnostic the cysteine motif arrangement is for Erv1 vs. Erv2. The human homolog of Erv1 along with the available stramenopile examples (*Blastocystis homini* CBK24728, *Phytophthora infestans* EEY69279, *Ectocarpus siliculosus* CBN79383, *Thalassiosira pseudonana* XP_002291504, *Phytophthora sojae* EGZ23753) all have a similar arrangement to yeast Erv1. Unfortunately, the number of annotated Erv2 proteins in Genbank (114) is vastly outnumbered by those for Erv1 (1675). Only three of the annotated Erv2 proteins are nonfungal. *Perkinsus* has two (XP_002774612 (CX4C, CX15C, CXXC); XP_002765828 (CX4C, CXXC, CX16C)) with conflicting arrangements of the cysteine motifs while the *Chlamydomonas* example (XP_001703298) only possess a single short motif followed by the larger motif. Moreover, it is unclear how reliable the assignments are given the similarities in size and structure between Erv1 and Erv2. Is Erv2 exclusively a fungal protein with the few non-fungal annotations in error or are some of the Erv1 annotated proteins mis-annotated due to the nature of automated annotations? Based on the various lines of evidence I conclude that bn43859 is most likely the ER version of Erv while bn48701 and bn57894 both function in the mitochondrial intermembrane space.

Table 3.9. Analysis of *B. natans* Erv domain containing proteins. SP=signal peptide, M= mitochondrial.

Protein id	Erv1 blastp	Erv2 blastp	TargetP	Best hit	Motif arrangement	annotation
43859	6e-31	4e-25	.915 SP	*Volvox*	CXXC,CX16C,CX6C	ER
48701	2e-33	4e-23	.691 M	*Amphimedon*	CXXC,CXXC,CX16C	Mito
57894	3e-16	7e-10	.912 other	*Salpingoeca*	CXXC,CXXC,CX16C	Mito

G. theta also possess three proteins with Erv domains (42578,53205,148797). Blastp results against the yeast Erv1/2 proteins were weak at best. 42578 had an e-value score of 1e-14 against Erv1 and 2e-10 against Erv2. 53205 and 148797 had scores against Erv1 of 3e-7 and 3e-6 respectively and did not have hits (1e-06 cutoff) against Erv2. More importantly, 42578 is a small protein (245 amino acids) with a single domain (Erv1) while 53205 (657 amino acids) and 148797 (627 amino acids) were large proteins possessing thioredoxin domains in addition to the Erv1 domain. Thioredoxin domains, combined with one or more Erv1 domains, is indicative of quiescin sulfhydryl oxidases which are found in the ER (Kodali and Thorpe 2010). Therefore, *G. theta* only has one Erv protein. Whether it is ER or mitochondrial targeted is unclear. The TargetP score suggests mitochondrial (.737 other) while the cysteine motif arrangement (CXXC, CX16C) suggests ER.

3.3.4 Twin CX9C Proteins

Another class of nucleus encoded mitochondrial targeted proteins without presequences are those with twin CX9C motifs (Longen, Bien et al. 2009). These proteins, which are typically small (less than 200 amino acids with most less than 110 (Cavallaro 2010) and are targeted to the intermembrane space, are characterized by two cysteine pairs with 9 residues separating the cysteines in each pair. A search of the yeast (*Saccharomyces cerevisiae)* genome found 86 candidate proteins with the requisite twin CX9C motif (Longen, Bien et al. 2009). From this pool these authors identified 14 proteins that were mitochondrial targeted including Cox17, Cox19, Cox23 and Cmc1. A study of these proteins determined additional characteristics that may be used to identify true twin

CX9C mitochondrial proteins – a) a size between 9 and 18 kDa, b) two helices associated with the cysteine pairs and separated by a loop (HLH), c) hydrophobic residues at positions 4 and 7 of the 9 residues separating the cysteines, particularly in the second CX9C motif. All the 14 candidates in yeast were experimentally shown to be active in the mitochondrion. Following from the yeast study a genome wide analysis of twin CX9C proteins was recently done for a number of eukaryotic organisms (Cavallaro 2010). The only organisms that appeared to lack twin CX9C proteins were *Encephalitozoon cuniculi*, *Entamoeba dispar* and *Entamoeba histolytica* which, the author points out, are all obligate intracellular parasites without classical mitochondria. Instead, these organisms have reduced mitochondria called mitosomes that lack ATP synthesis. The number of twin CX9C proteins identified in the 17 eukaryotic protein sets (Cavallaro 2010) ranged from 39 in *Arabidopsis thaliana* to 10 in *Schizosaccharomyces pombe*. The study also came to the conclusion that proteins with twin Cx9C motifs associated with parallel helices all play a role in the mitochondrial proteome.

Given its apparent importance for mitochondrial functions I undertook a similar bioinformatics search for this class of proteins. Using the EMBOSS utility fuzzpro (Rice, Longden et al. 2000) I identified 246 proteins in *G. theta* and 148 in *B. natans* that contained exactly two CX9C motifs that were separated by no more than 100 amino acid residues. Molecular masses were calculated using the EMBOSS utility PEPSTATS (Rice, Longden et al. 2000). Secondary structures were determined using JPRED3 (Cole, Barber et al. 2008). Logo plots of the two CX9C motifs were created using Weblogo (http://weblogo.berkeley.edu/).

The molecular masses of the 86 candidate proteins in yeast could be divided into three clear groups (Longen, Bien et al. 2009). 16 proteins had a predicted mass between 8 and 20 kDa and of these 14 were determined to be mitochondrial twin CX9C proteins while the other two were characterized as classical zinc finger proteins. Four proteins had a molecular mass below 8 kDa and had previously been identified as metallothioneins involved in metal homeostasis. The rest of the candidate proteins had masses above 30 kDa. None of these larger proteins were considered mitochondrial twin CX9C proteins but instead were generally transcription factors with the conserved cysteine residues as part of DNA-binding zinc fingers. The distribution of masses of the 148 and 246 proteins for *B. natans* and *G. theta* respectively was not as clear-cut. While all of the proteins that I believe to be mitochondrial twin CX9C proteins had masses below 18 kDa, as in yeast, several of them were also below 8 kDa (bn477009, bn55834, gt60979) (Table 3.10), unlike in yeast.

The eukaryotic genome survey also revealed twin CX9C proteins below 8 kDa (Cavallaro 2010). More significantly, yeast had no candidate proteins between 20 and 30 kDa while *B. natans* had 29 (19.5%) and *G. theta* had 38 (15.4%) (Figure 3.2). As well, while almost all of the yeast candidates (14/16) with a mass below 20 kDa turned out to be twin CX9C proteins both *B. natans* and *G. theta* had a large number in this mass range that were not twin CX9C proteins. *G. theta* had 58 candidates below 20 kDa but ultimately only six were considered twin CX9C proteins while in *B. natans* 10 out of 27 were considered twin CX9C proteins. All of the *B. natans* candidates below 20 kDa that were not twin CX9C proteins did not have any similarity with any known protein. In *G. theta*, many of the candidates below 20 kDa that were not considered twin CX9C proteins turned out to be predicted transcription factors. In particular, *G. theta* had an abundance of proteins with a fungal-type DNA-binding domain that involves two zinc atoms bound to six cysteine residues. According to the InterPro entry for this domain (IPR001138) the taxonomic coverage is highly restricted with 15,063 entries for Fungi and only 135 non-fungal entries, half of which are from *Naegleria gruberi* (77). Only one Viridiplantae is recorded as having proteins with this domain – barley with seven. Among chromalveolates, *Thalassiosira* has four proteins with this domain, *Phaeodactylum* has one, the Alveolate *Ichthyophthirius multifillis* also has one and *Ectocarpus* has 33. A text search of the JGI annotations for this domain turned up 99 independent entries in *G. theta* and 34 in *B. natans*. It is unclear whether the highly restricted taxonomic occurrence of this domain is the result of LGT followed by extensive gene family expansion in some of the lineages or merely under reporting. Regardless, these proteins do not appear to be involved in the mitochondrial proteomes of *G. theta or B. natans*.

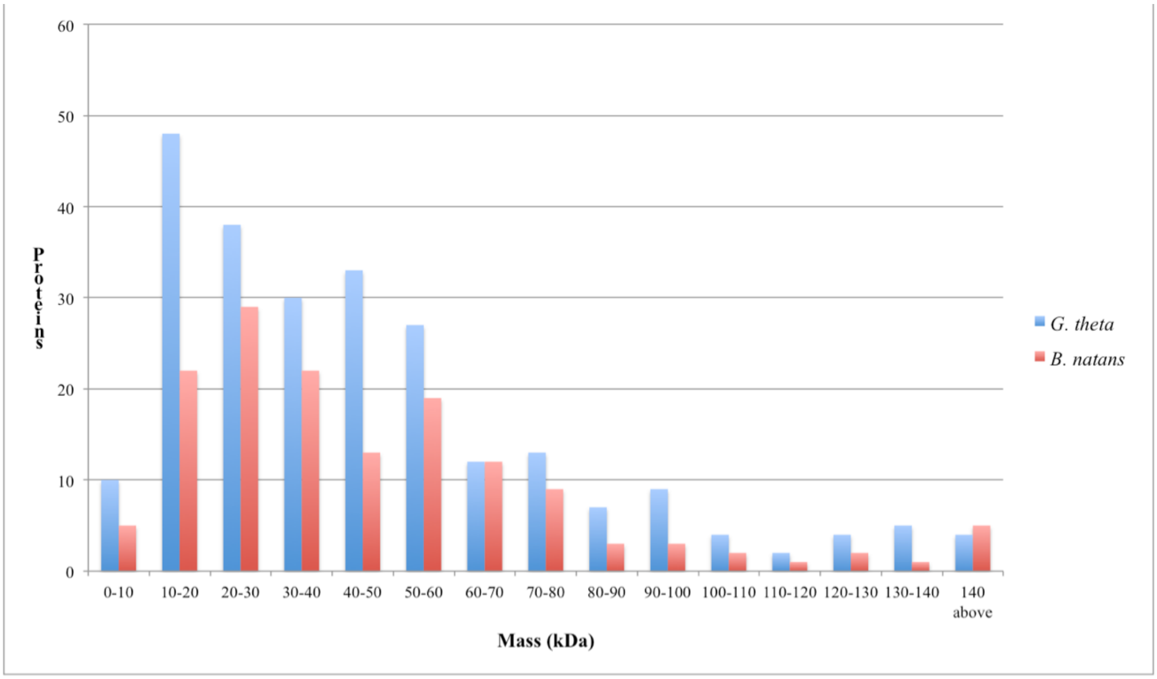


Figure 3.2. Distribution of Cx9C protein masses for *G. theta* and *B. natans.*

An important factor for determining whether the candidates were twin CX9C proteins was the association of a helix-loop-helix arrangement with the CX9C motifs since the conserved cysteines are what stabilize the two antiparallel alpha helices. As with the yeast proteins all of the candidates that were determined to be twin CX9C proteins had the helix-loop-helix arrangement. There were a few proteins, particularly in *G. theta* for which the motifs were partial helices (gt111007, gt122028) but these were determined to be fungal-type DNA binding domains. The eukaryotic nuclear genome-wide study also used secondary structure to weed out non-mitochondrial twin Cx9C proteins (Cavallaro 2010). It should be noted that in that study the discriminating secondary structure was a coiled coil-helix-coiled coil-helix domain (CHCH) which is essentially the same as the helix loop helix arrangement described in the yeast study.

Of less importance for identifying twin CX9C proteins is the presence of hydrophobic residues at certain positions in the motifs. A logo plot of the seven residues in the two motifs (Figure 3.3) reveals, as with the yeast study, a strong preference for hydrophobic amino acids in position seven and to a lesser extent in position three of the second motif. However, bn47709 shows that this criterion is not definitive since it has a hydrophilic histidine in position seven of the second motif. The Cavallaro (Cavallaro 2010) study has a different view with regard to the amino acids between the conserved cysteines that may explain the presence of non-hydrophobic residues at certain positions. Because most of the twin CX9C proteins do not have classical mitochondrial presequences they are imported into the intermembrane mitochondrial space (IMS) via the MIA pathway. Proteins are targeted to the MIA pathway via an IMS targeting signal (ITS) that is composed of nine amino acids upstream or downstream of one of the four conserved cysteines in the CX9C motifs (Sideris, Petrakis et al. 2009). In the study of 385 twin CX9C proteins (Cavallaro 2010) the nine amino acids of the ITS were found downstream of the first Cysteine 36% of the time and downstream of the 3rd Cysteine 34% of the time, effectively making one of the two CX9C motifs synonymous with the ITS. The ITS is further defined by having either two aromatic residues (F,W,H,Y) at positions four and seven or the aromatic at position seven replaced by one that is hydrophobic.

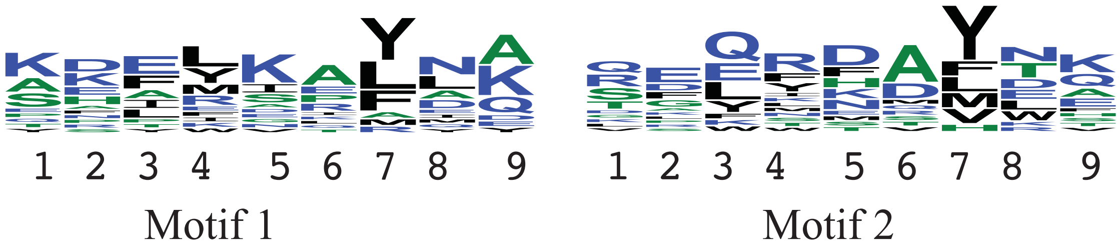


Figure 3.3. Logo plots of CX9C motifs for proteins from both *G. theta* and *B. natans.*

As previously mentioned, I was able to identify 10 proteins in *B. natans* that qualify as twin CX9C proteins and six in *G. theta*. The numbers seem low compared with those identified in the eukaryotic wide genome survey. Several factors may account for the differences in the number of proteins identified. The eukaryotic wide study included proteins that did not strictly conform to the CX9C motif. For example, Cox 6b proteins were considered twin CX9C proteins even though the second motif is CX10C and 10 of the 39 *Arabidopsis* proteins had motifs other than CX9C. Nevertheless, even sticking to classical twin CX9C proteins, *B. natans,* and in particular *G. theta,* had fewer proteins in this class than all other organisms except for *S. pombe* with 10. Unfortunately the eukaryotic wide genome study did not include any photosynthetic algae so low numbers in *G. theta* and *B. natans* may be related to being unicellular algal organisms. All of the protists surveyed were parasites and as mentioned two of them (*E. dispar*, *E. histolytica*) do not have mitochondria. *Trypanosoma cruzi* had 38 proteins but this represented only 20 genes since a number of proteins were products of heterozygous alleles. Taxonomically, the closest species to *G. theta* and *B. natans* is the alveolate *Plasmodium falciparum*. It had 11 proteins, one of which had a CX10C motif so its protein numbers are comparable to *B. natans*.

Table 3.10. Twin CX9C proteins from *B. natans* and *G. theta.* Position seven in the motifs bolded.

Protein id	Mass	First motif	Second motif	function
bn47709	6.6 kDa	CPDTRKL**R**DKC	CKKEIEA**H**KVC	Cox17
bn55834	7.8 kDa	CAAPWKA**Y**LAC	CSGWYFD**Y**WKC	ubiquinol-cytochrome C reductase complex subunit 6
bn63670	9.5 kDa	CAHILIP**L**NAC	CVDLRHA**Y**EAC	NADH-ubiquinone oxidoreductase B18 subunit
bn141972	10.8 kDa	CKDLKAR**Y**DQC	CRALWDT**Y**TSC	Cmc1-like
bn48203	11.4 kDa	CKRFRVA**Y**LKC	CLDKSKS**Y**LQC	Cox19
bn126590	11.5 kDa	CSEIEKK**A**LKC	CQSFFDQ**V**RAC	unknown
bn91213	12.5 kDa	CYSFYEA**Y**NKC	CSPQKSA**V**DKC	CmC1-like
bn141034	14.1 kDa	CKDEMKA**L**AAC	CGEQRTV**Y**TQC	Cmc1-like
bn73620	16.2 kDa	CTKAFDI**M**IYC	CSRQLND**F**NFC	unknown
bn86392	16.7 kDa	CSFELGQ**F**NQC	CQFYFDA**M**NQC	CHCH domain containing protein
Protein id	Mass	First motif	Second motif	function
gt60979	7.8 kDa	CKEAMSR**Y**MAC	CREETKA**Y**LEC	Cox19
gt91152	8.3 kDa	CQKYLKE**L**EAC	CTGQYFD**F**WHC	ubiquinol-cytochrome C reductase complex subunit 6
gt97918	9.1 kDa	CSHLLIP**L**NKC	CTEERHA**Y**EKC	NADH-ubiquinone oxidoreductase B18 subunit
gt115776	11.4 kDa	CKKEINE**F**AQC	CRLQNNA**M**NEC	Cmc1-like
gt155227	11.7 kDa	CANEYST**F**QEC	CQFYMDM**L**TKC	CHC domain containing protein
gt101528	12.7 kDa	CEPFYKA**Y**YDC	CDELRMR**L**DTC	Cmc1-like

3.3.5 Mitochondrial Import Proteins

The enormous reduction in mitochondrial-encoded proteins requires that the vast majority of the mitochondrial proteome is encoded by nuclear genes, synthesized on cytosolic ribosomes and transported into the mitochondrion. The establishment of an import system was key to the success of organellogenesis and in particular facilitated endosymbiotic gene transfer (EGT) (Andersson, Karlberg et al. 2003). It is likely that EGT requires a period when a copy of the gene resides in both the mitochondrion and in the nucleus. During the early stages of endosymbiosis, once it was possible to target the protein encoded by the gene back to the protomitochondrion, the mitochondrial copy of the gene was able to be lost (Andersson, Karlberg et al. 2003). Not all the proteins of the mitochondrial proteome result from EGT. Many have been recruited from the pool of host genes or "invented" but nevertheless still require transport to the mitochondrion (Gray, Burger et al. 2001). Given the central role protein import plays in the successful relationship between the host and the newly evolved organelle the import systems have received considerable attention (see (Chacinska, Koehler et al. 2009) for review).

Initially it was believed that mitochondrial precursor proteins are imported via a single method, the so-called presequence pathway that involved the use of target peptides on the N-termini of the proteins to direct them into mitochondria (Carrie, Murcha et al. 2010). It then became clear that there was a second pathway, generally for carrier proteins, that did not use target peptides but signals internal to the mature protein. Now, we know of at least two other pathways for the import of proteins into the mitochondrion. These other systems involve proteins that are directed to the mitochondrial membranes or the intermembrane space rather than into the matrix (Carrie, Murcha et al. 2010).

Regardless of the eventual pathway used, all proteins destined for some portion of the mitochondrion pass through the translocase of the outer membrane (TOM) (Chacinska, Koehler et al. 2009). The essential portion of TOM is Tom40, a beta barrel protein that forms a channel across the outer membrane. Once through TOM there is additional machinery that determines the final destination. Outer membrane proteins are directed there by the sorting and assembly machinery (SAM), intermembrane space proteins by the mitochondrial intermembrane space import and assembly complex (MIA), inner membrane proteins by Tim22, Tim23 and OXA, and matrix destined proteins by Tim23 (Kutik, Stroud et al. 2009). Besides the localization machinery there are proteins that process the peptides during and after their arrival. Mitochondrial processing peptides (MPP) remove the presequences from proteins that make it to the matrix and in some cases are further processed by the matrix intermediate peptidase (MIP), while some intermembrane space proteins are cleaved by inner membrane peptidases (IMP) (Chacinska, Koehler et al. 2009).

Since all eukaryotic organisms have either a mitochondrion or mitochondrion-related organelles such as mitosomes and hydrogenosomes it is likely that they all possess mitochondrial import proteins (Lithgow and Schneider 2010). An analysis across the breadth of eukaryotic diversity has the potential to elucidate where the importation machineries came from and when they developed. Because the endosymbiotic event that led to mitochondria occurred prior to the acquisition of plastids and possibly prior to the divergence of eukaryotes into the major lineages, the analysis of these proteins may also give us some insight into the taxonomic placement of the major groups. In particular, the taxonomic positions of the chromalveolate and rhizarian lineages are unresolved, perhaps due in part to the transfer of plastid genes and eukaryotic algal genes during secondary (or tertiary) endosymbiosis and the confounding effect these transfers have on large scale phylogenomic studies (Lane and Archibald 2008).

3.3.5.1 TOM complex (Translocase of the Outer Membrane)

The TOM complex is the gateway through which almost all proteins destined for the mitochondrion must pass, and consists of the pore forming protein Tom40 (Lithgow and Schneider 2010). Usually associated with it is Tom7 and in some lineages Tom5 and Tom6. N-terminal targeting signals are recognized by Tom22 or Tom20, while proteins with internal signals are recognized by Tom70.

The minimum requirements for protein import are generally considered to be Tom40, Tom7 and Tom22 (Chacinska, Koehler et al. 2009). All genomes appear to encode Tom40 without exception (Carrie, Murcha et al. 2010; Delage, Leblanc et al. 2011) including *G. theta* and *B. natans* (Table 3.11). *G. theta* also appears to be “typical” in that it has genes for Tom7 and Tom22 as well. *B. natans*, however seems particularly deficient in TOM proteins as I was only able to identify Tom40. Blastp and tblastn searches of the *B. natans* genome using *G. theta*, yeast, *Arabidopsis thaliana*, and diatom homologs failed to generate any meaningful hits for Tom 22 or Tom7. The only other genome to date that is as devoid of TOM components is the pelagophyte *Aureococcus anophagefferens* for which only Tom40 has been identified (Delage, Leblanc et al. 2011). A genome similarly deficient in TOM components is the red alga *Cyanidioschyzon merolae*, which only has Tom40 and Tom22.

Table 3.11. Components of the Mitochondrial Protein Import and Sorting Machineries for *B. natans* and *G. theta*. nf= not found. t=found through text search. h=found or confirmed through homology searches. *=bacterial homolog. **=alpha-proteobacterial homolog. Protein names in bold are considered essential.

Protein Name	***G. theta***	***B. natans***
Translocase of the Outer Mitochondrial Membrane (TOM)		
Tom20	150069 h	nf
Tom70	nf	nf
Tom71	nf	nf
Tom40	166503 th	87802 th
Tom22	118738 th	nf
Tom5	nf	nf
Tom6	nf	nf
Tom7	154676 th	nf
Sorting and Assembly Machinery of the Outer Mitochondrial Membrane (SAM)		
Sam50	110008** h 111251** h	73769** h
Sam37, metaxin1, Tom37	109736* th	nf
Sam35	nf	nf
Mdm10	nf	nf
Mim1	nf	nf
Mitochondrial Intermembrane Space Import and Assembly (MIA)		
Mia40	103258 h	nf
Erv1	42578 h	48701 h 57894 h
Hot13	nf	nf
Intermembrane Space Chaperones		
Tim8	69150 th	86440 th
Tim9	73833 th	50483 th
Tim10	150375 th 153126 h (maybe 12)	130233 h
Tim13	109426 th	76916 h

Protein Name	***G. theta***	***B. natans***
Carrier Translocase of the Inner Mitochondrial Membrane (TIM22)		
Tim22	118542 th 150332 th	88586 th
Tim54	nf	nf
Tim18	nf	nf
Presequence Translocase of the Inner Mitochondrial Membrane (TIM23)		
Tim23	107923 th	84980* th
Tim17	104108 th 144004 th	51633** th
Tim50	150902* h	87562*
Tim21	nf	84917 th 144223 th
Presequence Translocase-Associated Motor (PAM)		
mtHSP70	79059** th	92341** th
Mge1	152067** h	43108** h
Tim44	138039 th	87192** th
Pam18, Tim14	80286** th	36491** h
Pam16	152490** th	87659** h
Pam17	nf	nf
Mitochondrial Processing Peptidases		
MPP alpha	106014** th	91482** th
MPP beta	83986** th	53395** th
MIP	75461* th	73375* th 68741* th
Imp1	88558* th	34144* th 38956* th
Imp2	86641* th	87500* th
Som1	nf	nf
Mitochondrial Export Machinery		
Mba1	nf	nf
Oxa1	117364* h	144244*
Mdm38, LETM1	157808 h 166205 h	34875* h 30247* h

B. natans, A. anophagefferens and *C. merolae* also lack Tom70 (Table 3.11, 3.12; (Delage, Leblanc et al. 2011)) which is the protein required for the initial importation of peptides without presequences. Considering that this is the route that most mitochondrial carrier proteins take, it is surprising that *B. natans* has a full complement of them despite lacking Tom70. However, Tom70 is also absent from *G. theta*, land plants, and green algae (Table 3.11, 3.12; (Delage, Leblanc et al. 2011)). Until recently it was also thought to be absent from stramenopiles but through the use of HMM profiles it has been found in *Blastocystis* and subsequently several other stramenopiles including diatoms and oomycetes (Tsaousis, Gaston et al. 2011) as well as the haptophyte *Emiliania huxleyi*. Since all the lineages that lack a Tom70 homolog demonstrably have mitochondrial-targeted proteins without presequences its absence is puzzling. It may simply be too divergent to find through conventional homology searches as seen with the stramenopile examples (Tsaousis, Gaston et al. 2011), and simply awaits better identification techniques. Alternatively, these lineages seemingly without Tom70 may have a separate and as yet unknown system for detecting and directing proteins with internal signals. This latter explanation would be difficult however to align with the current taxonomic understanding of the relationship between *G. theta*, a cryptophyte, *B. natans* a rhizarian, and the primary photosynthetic lineages. Either the stramenopiles and haptophytes would have to have replaced this unknown system with a Tom70 acquired via HGT from the opisthokonts, or cryptophytes and chlorarachniophytes acquired this unknown system from the Archaeplastida in two separate HGT events, perhaps during the establishment of the photosynthetic endosymbionts. Tom70 is characterized by 11 tetratricopeptide repeat motifs (TPR). Unfortunately TPRs are very common in proteins, which makes finding Tom70 through homology searches difficult. Both *G. theta* and *B. natans* have several large proteins with more than 10 TPRs. This may be fertile ground for an in-depth search of this important but elusive protein.

Table 3.12. Presence of mitochondrial import proteins in select organisms. ni=not identified. x=identified.

Name	***G. theta***	***B. natans***	**Fungi**	**Animal**	**Land Plants**	**Green Algae**	**Red Algae**	**Oomycetes**	**Diatoms**	**Other Stramenopiles**	***Tetrahymena***
Translocase of the Outer Mitochondrial Membrane (TOM)											
Tom20	150069	ni	x	x	x	x	ni	ni	ni	ni	ni
Tom70	ni	ni	x	x	ni	ni	ni	x	x	x	ni
Tom40	166503	87802	x	x	x	x	x	x	x	x	x
Tom22	118738	ni	x	x	x	x	x	x	x	x	x
Tom5	ni	ni	x	x	x	ni	ni	ni	ni	ni	ni
Tom6	ni	ni	x	x	x	ni	ni	ni	ni	ni	ni
Tom7	154676	ni	x	x	x	x	ni	x	x	x	x
Sorting and Assembly Machinery of the Outer Mitochondrial Membrane (SAM)											
Sam50	110008 111251	73769	x	x	x	x	x	x	x	x	ni
Sam37, metaxin1, Tom37	109736	ni	x	x	x	x	ni	x	ni	ni	ni
Sam35	ni	ni	x	ni	x	x	ni	ni	ni	ni	ni
Mdm10	ni	ni	x	ni	ni	ni	ni	ni	ni	ni	ni
Mim1	ni	ni	x	ni	ni	ni	ni	ni	ni	ni	ni
Mitochondrial Intermembrane Space Import and Assembly (MIA)											
Mia40	103258	ni	x	x	x	x	x	x	x	ni	ni
Erv1	42578	48701 57894	x	x	x	x	x	x	x	x	x
Hot13	ni	ni	x	x	ni	ni	x	x	x	x	ni
Intermembrane Space Chaperones											
Tim8	69150	86440	x	x	x	x	x	x	x	x	x
Tim9	73833	50483	x	x	x	x	x	x	x	x	x
Tim10	150375 153126	130233	x	x	x	x	x	x	x	x	x

Name	***G. theta***	***B. natans***	**Fungi**	**Animal**	**Land Plants**	**Green Algae**	**Red Algae**	**Oomycetes**	**Diatoms**	**Other Stramenopiles**	***Tetrahymena***
Tim13	109426	76916	x	x	x	x	x	x	x	x	ni
Carrier Translocase of the Inner Mitochondrial Membrane (TIM22)											
Tim22	118542 150332	88586	x	x	x	x	x	x	x	x	x
Tim54	ni	ni	x	x	ni	ni	ni	ni	ni	Ni	ni
Tim18	ni	ni	x	x	ni	ni	ni	ni	ni	Ni	ni
Presequence Translocase of the Inner Mitochondrial Membrane (TIM23)											
Tim23	107923	84980	x	x	x	x	x	x	x	x	x
Tim17	104108 144004	51633	x	x	x	x	x	x	x	x	x
Tim50	150902	87562	x	x	x	x	x	x	x	x	x
Tim21	ni	84917 144223	x	x	x	x	ni	x	x	x	ni
Presequence Translocase-Associated Motor (PAM)											
mtHsp70	79059	92341	x	x	x	x	x	x	x	x	x
Mge1	152067	43108	x	x	x	x	x	x	x	x	x
Tim44	138039	87192	x	x	x	x	x	x	x	x	x
Pam18, Tim14	80286	36491	x	x	x	x	x	x	x	x	x
Pam16	152490	87659	x	x	x	x	x	x	x	x	x
Pam17	ni	ni	x	ni	ni	ni	x	ni	ni	ni	ni

Data other than *B. natans*, *G. theta* and *Tetrahymena* adapted from (Delage, Leblanc et al. 2011) Fungi=*Saccharomyces cerevisiae.*
Animals=*Homo sapiens*. Land plants=*Arabidopsis thaliana.* Green algae= *Ostreococcus lucimarinus* and *Chlamydomonas reinhardtii*.
Red algae= *Cyanidioschyzon merolae*. Oomycetes = *Phytophthora infestans* and *Albugo laibachii*. Diatoms=*Phaeodactylum tricornutum* and
Thalassiosira pseudonana. Other stramenopiles=*Aureococcus anophagefferns* and *Ectocarpus siliculosus*. *Tetrahymena thermophila* data from (Smith, Gawryluk et al. 2007).

As mentioned, Tom22 is considered part of the minimum machinery required to import cytosolic proteins and as such is thought to have developed early in the establishment of the mitochondrion. *B. natans,* however, lacks a Tom22 homolog, or even a Tom20, which is thought to be similar enough to Tom22 to replace its function (Lithgow and Schneider 2010). Again *A. anophagefferens* has the only other genome that lacks both *tom22* and *tom20* (Table 3.11, 3.12; (Delage, Leblanc et al. 2011)). Other stramenopiles possess a *tom22* but not a *tom20*. *G. theta,* on the other hand, has both *tom22* and *tom20*, a condition shared by green lineages (Table 3.11, 3.12; (Delage, Leblanc et al. 2011)).

3.3.5.2 SAM complex (Sorting and Assembly Machinery)

Similar to the TOM complex, SAM is characterized by a central, channel forming protein - Sam50 (Chacinska, Koehler et al. 2009). This protein appears to be homologous to the bacterial protein Omp85/BamA (Paschen, Neupert et al. 2005). In yeast Sam50 has two ancillary proteins, Sam35 and Sam37, whose functions in animals are replaced by the distantly related orthologs Metaxin1 and Metaxin2 (Kutik, Stroud et al. 2009). In green plants and green algae Sam35 is present but Sam37 appears to be replaced by a metaxin (Carrie, Murcha et al. 2010). The animal and plant metaxins have an arrangement that clearly distinguishes them from yeast37, their distantly related ortholog. The metaxins and Sam37 have two Glutathione S-transferase domains as well as a transmembrane domain. In Sam37, the transmembrane domain is located at the N terminus while in metaxins it is found at the C terminus (Carrie, Murcha et al. 2010). Further distinctions can be made between plant and animal metaxins by the presence of two motifs in the plant version that are not found in either animal version.

Once again, *B. natans* lacks the diversity in SAM subunits displayed by fungi, animals and green lineages (Table 3.11, 3.12; (Delage, Leblanc et al. 2011)). It has the core subunit Sam50 but none of the ancillary ones. It shares this simplicity with red algae, *A. anophagefferens* and diatoms. Oomycetes appear to have a single metaxin, as does

Ectocarpus (Table 3.11, 3.12; (Delage, Leblanc et al. 2011)). As mentioned, animals have two metaxins to replace Sam35 and Sam37. In phylogenetic trees the stramenopile metaxins group with metaxin1s from animals. *G. theta* also has a single metaxin that groups with animal metaxin1. Surprisingly, *G. theta* has two copies of Sam50. They appear to be paralogs, consistently grouping together in trees.

Sam50 is essential since it is found in every nuclear genome examined to date. This suggests that its appearance predates the diversification of early eukaryotes. Beyond that the taxonomic distribution of the various subunits is confusing and probably not informative. Based on having a single SAM subunit, Sam50, *B. natans* is allied with red algae, diatoms and *Aureococcus*. However, other stramenopiles like *Ectocarpus* and the oomycetes have, in addition to Sam50, animal-like metaxins, as does *G. theta* (Table 3.11, 3.12; (Delage, Leblanc et al. 2011)). The divide between animal and plant metaxins would suggest that *G. theta*, along with oomycetes and *Ectocarpus* are more closely related to animals than green plants which is completely at odds with all believable higher order eukaryotic trees. The most plausible explanation is that an “animal-like” metaxin1 was the ancestral state for all eukaryotes. In fungi the domains were switched to produce Sam37 while in green plants and green algae divergence from the original metaxin occurred after they split from other eukaryotes. It should be noted, that in animals, metaxin is not found in a complex with Sam50 (Kozjak-Pavlovic, Ross et al. 2007) while in plants it is (Carrie, Murcha et al. 2010). *G. theta*, oomycetes and *Ectocarpus* retained the original animal-like metaxin while in *B. natans*, *Aureococcus* and diatoms it was lost or has diverged enough to be unrecognizable.

3.3.5.3 Tiny Tims (Translocase of the inner mitochondrial membrane)

The intermembrane space contains two small complexes of two subunits each. Tim9/Tim10 act as a chaperone to direct hydrophobic proteins to the SAM or TIM22 complexes while Tim8/Tim13 direct proteins to the TIM23 complex (Lithgow and Schneider 2010). Apart from some organisms with reduced mitochondria like *Trichomonas vaginalis* all eukaryotes have these two small complexes and almost always have the full complement of two subunits for each. The only exception appears to be *Aureococcus* for which Tim9 is the only subunit that has been identified (Carrie, Murcha et al. 2010). Curiously, *Blastocystis*, which has a mitochondrial related organelle, is also a stramenopile and only has Tim9 (Stechmann, Hamblin et al. 2008) suggesting that this subunit is more essential than the others. *B. natans* has all four tiny tims while *G. theta* has all four plus an additional Tim10 (Table 3.11, 3.12). Additional Tim10s are also seen in animals (Gentle, Perry et al. 2007) as well as additional Tim8s.

3.3.5.4 MIA (mitochondrial intermembrane space assembly machinery)

The import of small intermembrane space proteins like Tim8, 9, 10, and 13 has been characterized in yeast. The process involves two subunits, Mia40 and Erv1 with Mia40 acting as a receptor in the intermembrane space (Hell 2008). As discussed in a previous section, I was able to identify Erv1 genes in both *B. natans* and *G. theta* and it has been identified from all the other genomes and major lineages suggesting that it is essential and ancestral (Table 3.11, 3.12; (Delage, Leblanc et al. 2011)). The *G. theta* Mia40 gene was found through homology searches using the yeast gene (EEU09187). However, in *B.*

natans no Mia40 could be identified through homology searches using the yeast gene, the *G. theta* gene, a *Micromonas* gene (ACO63035), *Thalassiosira pseudonana* (XP_002292670) or the *Arabidopsis* version (AED93159). A study of Mia40 in *Arabidopsis thaliana* (Carrie, Murcha et al. 2010) showed that the gene was not essential for the successful import of the Tim proteins into the intermembrane space and it was suggested that Erv1 is sufficient. Mia40 is present in oomycetes and diatoms but absent from *Ectocarpus* and *Aureococcus* (Table 3.12; (Delage, Leblanc et al. 2011)). It is also absent from a number of parasitic protists with mitochondrial related organelles. Given its near ubiquity Mia40 also appears to be ancestral to present-day eukaryotes but has been lost in some of the stramenopiles and rhizarians.

3.3.5.5 TIM23 complex

The largest of the mitochondrial import complexes is TIM23 with 9 subunits in yeast (Lithgow and Schneider 2010). It is responsible for the import of proteins with presequences into or across the inner membrane. The barest essentials are the three subunits embedded in the inner membrane, Tim23, Tim17 and Tim50. The rest of the subunits are located on the matrix side. In yeast, incoming proteins are bound by HSP70 which is recruited to the complex by Tim44 (Hell 2008). HSP70 binding is regulated by MgeI, a nucleotide exchange factor, and the J-complex that consists of Tim14 and Tim16 (Lithgow and Schneider 2010). These ancillary components have also been described as the PAM complex or presequence assisted motor and consist of 5 subunits – Tim44, HSP70, Pam16 or Tim16 , Pam17, and Pam 18 or Tim14 (Lithgow and Schneider 2010).

The final subunit is Tim21, which is embedded in the inner membrane but does not interact directly with proteins.

The core Tim subunits 17, 23 and 50 are found in all genomes including *G. theta* and *B. natans*, indicating they are ancestral and essential (Table 3.11, 3.12; (Delage, Leblanc et al. 2011)). *G. theta* has 2 copies of Tim17, something that is also seen in animals and green plants, though not green algae. Plant Tim17s have a characteristic C-terminal motif that is not present in either of the *G. theta* versions or for that matter in the *B. natans* version. Among the non-core subunits *B. natans* and *G. theta* have mtHSP70, Tim14, Tim44 and Mge1 which is consistent with the rest of the genomes from organisms that have fully functional mitochondria (Table 3.11, 3.12; (Delage, Leblanc et al. 2011)). Also found across the board are Pam16/Tim16 and Pam18/Tim14 except in *Aureococcus* which appears to lack Pam16/Tim16. Also consistent across the genomes, including *B. natans* and *G. theta*, is the lack of Pam17 except in fungi, suggesting that it is lineage specific. The only difference between *B. natans* and *G. theta* with regard to TIM23 subunits is that *B. natans* has two Tim21s while *G. theta* appears to have none. Tim21 is also absent from red algae and *Aureococcus* (Table 3.11, 3.12; (Delage, Leblanc et al. 2011)).

3.3.5.6 TIM22 Complex

The TIM22 complex is responsible for importing those carrier proteins with internal signals into the inner membrane (Lithgow and Schneider 2010). In yeast, the complex consists of Tim22 and three associated proteins - Tim18, 54 and 12. Non-fungal genomes only encode Tim22 so the associated proteins are lineage specific (Delage, Leblanc et al. 2011). *B. natans* contains a single *tim22* gene which is consistent with all other lineages except green plants which contain three, and *G. theta* which contains two (Table 3.11, 3.12; (Delage, Leblanc et al. 2011)).

3.3.5.7 MPP (Mitochondrial Processing Peptidases)

After being imported, those proteins with N-terminal presequences are processed by peptidases that cleave the targeting signals (Chacinska, Koehler et al. 2009). The Inner Membrane Peptidase (IMP) complex is a heterodimer of Imp1 and Imp2. The proteins are similar in sequence and function but can be distinguished by an Rx5P motif in Imp1 and a Nx5S motif in Imp2 (Burri, Strahm et al. 2005). Delage et al. (Delage, Leblanc et al. 2011) suggest a further refinement of the discriminating motifs with Imp1 having GDNx7Rx5P and Imp2 EGDx8Nx5[S/P]. *G. theta* has one Imp1 while *B. natans* appears to have two (Table 3.11). The *G. theta* Imp1 has the longer motif (GDNx7Rx5P) as does the *B. natans* protein 34144 but the second Imp1 for *B. natans* (38956) only has the shorter motif Rx5P. Both *B. natans* and *G. theta* have a single copy of Imp2. In both organisms the motif is not perfectly conserved. *B. natans* has EGDx8N but not the last portion (x5P) while the *G. theta* protein sequence matches even less with just EGD at the beginning. In blastp searches against the NCBI protein database they both consistently hit Imp2s. The *G. theta* gene model is supported by EST data.

The mitochondrial processing peptidase (MPP) cleaves the N-terminal targeting sequences from proteins as they are transported into the organelle (Lithgow and

Schneider 2010). It exists as a heterodimeric enzyme composed of an alpha and a beta subunit. The MPP also functions in complex III of the mitochondrial respiratory chain (see below) where the two subunits are known as Core1 (alpha) and Core2 (beta). In some lineages, such as animals and yeast, the functions are spatially separated and there are four individual genes, while in plants the peptidase enzyme is part of complex III so there are only two genes, an alpha/Core1 and a beta/Core2 (Carrie, Murcha et al. 2010). *B. natans* and *G. theta* appear to be of the plant type with only a single gene for alpha/Core1 and a single gene for beta/Core2 (Table 3.11). Because these genes are highly conserved it is unlikely that the other copies were simply not identified. Blastp searches of the alpha subunit against the genome it came from only returns hits for itself and the beta and vice versa for the beta subunit. The beta subunit in both organisms has the motif HxxEHx76E that is indicative of enzymatic activity (Delage, Leblanc et al. 2011). The two subunits separate out into two main eukaryotic clades, along with a bacterial clade, in a phylogenetic tree (data not shown).

Even after being cleaved by MPP, proteins are further processed by the mitochondrial intermediate peptidase (MIP) (Lithgow and Schneider 2010). MIP consists of a single protein belonging to a large family of zinc metalloproteases that includes several oligopeptidases, two of which function in the cytosol. *G. theta* has a single copy of MIP while *B. natans* appears to have two (68741 and 73375) (Table 3.11). Both *B. natans* versions hit against MIPs in NCBI blastp analyses and contain the domain M3A_MIP. If the MIP annotated *Arabidopsis thaliana* protein (AT5G51540) is compared against the full *B. natans* protein set both of the *B. natans* putative MIPs have similar scores (73375

2e-48, 68741 1e-44) while the third best hit 71630 (4e-28) is the protein thimet oligopeptidase, which is a member of the zinc metalloproteases. Neither putative MIP has targeting information that would be useful in determining whether they are both found in mitochondria. For now, both *B. natans* proteins will be considered putative MIPs.

The presequences that are cleaved by MPP and MIP need to be removed from the mitochondrial matrix since their accumulation has been reported to adversely affect the integrity and function of the organelle membranes (Wieprecht, Apostolov et al. 2000; Stahl, Moberg et al. 2002). A similar situation occurs in the chloroplast after the target signals are removed. In both organelles the 'cleanup' of presequences is accomplished by a zinc-metalloprotease (Stahl, Moberg et al. 2002) that in the case of mitochondria is known as a presequence protease (PreP). Investigations in *Arabidopsis thaliana* demonstrated that one of the PrePs is dual targeted to the mitochondria and chloroplasts and moreover has been shown to be active in both organelles (Bhushan, Lefebvre et al. 2003). Dual targeting can be achieved by altering the protein in some fashion, such as alternative splicing or post-translational modification, so that proteins with different presequences are generated. It can also be achieved through a target signal that is ambiguous enough to be recognized by both plastid and mitochondrial translocons. The latter strategy appears to be the case with the dual targeted *Arabidopsis* PreP (Bhushan, Lefebvre et al. 2003). A similar case of dual targeting for PreP using ambiguous presequences seems to be the case in *B. natans* and *G. theta* as well. TargetP 1.1 (Emanuelsson, Brunak et al. 2007) predicts mitochondrial targeting with a high score of 0.904 for the *B. natans* protein 54816 when using the non-plant setting. If the plant setting is used the scores are 0.751 for plastid targeting and 0.712 for mitochondrial targeting. Moreover, a SignalP 3.0 (Bendtsen, Nielsen et al. 2004) analysis of 54816 generates a high HMM score (0.978) indicating a signal peptide is present within the first 35 amino acids. Sequence similarity with proteins from green plants does not start until ~80 amino acids in which provides ample room for the bipartite targeting signal required *for B. natans* proteins destined for the plastid. In *G. theta* the possibility of dual targeting is not as obvious from the predictions made by subcellular localization predictors. Using TargetP 1.1 the cryptophyte protein generates a mitochondrial score of 0.812 using the non-plant setting and a plastid score of 0.887 using the plant setting. SignalP 3.0 did not suggest a signal peptide.

3.3.5.8 Bacterial Ancestry of Import Proteins

The mitochondrion was originally a bacterium that was engulfed and retained by a eukaryotic cell (Andersson, Karlberg et al. 2003). Consequently, the mitochondrial membranes are the structural equivalent of the two cellular membranes while the matrix is the reduced periplasmic space. As more and more of the genes were transferred to the host, and consequently the encoded proteins targeted back to the endosymbiont, the ability to get those proteins from the cytoplasm into the protomitochondrion became increasingly important. To what extent do the various mitochondrial import machineries and processing complexes derive from the bacterial endosymbiont? Or are they primarily eukaryotic inventions made necessary by the increasing reliance on the host to house the genetic material required by the mitochondrion?

All mitochondrion-destined proteins must pass through the TOM complex. Given its central role, one might expect at least some of the subunits of this complex to have an endosymbiotic origin, especially Tom40, which is found in virtually all eukaryotic organisms with mitochondria. However, I was not able to detect any bacterial signal for any of the Tom subunits in *G. theta* or *B. natans,* including Tom40. This seems to be the general consensus in the literature (Carrie, Murcha et al. 2010; Delage, Leblanc et al. 2011) i.e., that the TOM complex is not of bacterial origin. Recently however, a study in *Trypanosoma brucei* (Pusnik, Schmidt et al. 2011) determined that this organism lacks Tom40 and instead has a protein dubbed ATOM for archaic translocase of the outer mitochondrial membrane. The study concluded that ATOM is not related to Tom40 but is related to a class of bacterial outer membrane proteins and furthermore represents the ancestral state of the protein transport system in the last common ancestor of eukaryotes. The paper presents three scenarios for the evolutionary history of Tom40 and ATOM. In two of these scenarios ATOM is replaced by a Tom40, of unknown origin, in all lineages except Trypanosomatids. ATOM may be derived from the alpha-proteobacterial endosymbiont or it may have arisen from LGT from other bacteria and replaced the endosymbiont's main translocase. In the third scenario Tom40 evolved from ATOM whose origin is either from the endosymbiont or from LGT from other bacteria. So, depending on which scenario is correct Tom40 may or may not have been derived from bacteria. However, another group (Zarsky, Tachezy et al. 2012) re-examined the *T. brucei* ATOM protein and concluded that it is in fact a divergent Tom40 and that neither Tom40 nor ATOM have a bacterial origin.

The SAM complex, which is located in the outer membrane and where it is involved in the correct folding of beta-barrel proteins, has clear homology with the outer membrane protein assembly complex from alpha-proteobacteria and as such is most likely derived from the endosymbiont (Delage, Leblanc et al. 2011). It is also found in all eukaryotes with mitochondria, which supports its ancient derivation prior to any major splits in eukaryotic lineages.

The other major component of the SAM complex is metaxin/Sam35/Sam37. As previously discussed the most likely scenario for its current taxonomic distribution is an ancestral state with subsequent lineage specific losses and divergence. Metaxin contains a glutathione S-transferase domain. In blastp analysis against bacterial genomes the *G. theta* metaxin protein had reasonably good hits (2e-19) to proteobacterial glutathione S-transferases. Among the top hits were several alpha-proteobacterial species including *Parvibaculum lavamentivorans* (5e-18, 28% identity) suggesting that metaxin is endosymbiotically derived. There is no indication that any of the tiny tims (8,9,10,13) have similarity with bacterial proteins. Although they seem to be ubiquitous among eukaryotes and thus presumably ancient they appear to be eukaryotic inventions.

Some of the subunits of TIM22 and TIM23 may have bacterial antecedents. The clearest case is for the chaperone mtHSP70. Both the *G. theta* and *B. natans* versions have very high similarity to alpha-proteobacterial molecular chaperones DnaK. The *G. theta* protein has 71% identity over 584 positions with a protein from *Pseudovibrio* sp. while the *B. natans* equivalent has 70% identity over 634 positions with a DnaK chaperone from the alpha-proteobacteria *Ochrobactrum anthropi*. Equally convincing is the case for Mge1, which, for both *B. natans* and *G. theta* versions, has reasonably strong hits to alpha-proteobacterial heat shock protein GrpE. The *G. theta* protein has 36% identity across 214 positions with an *Agrobacterium radiobacter* GrpE while the best hit for the *B. natans* protein is once again from *Ochrobactrum anthropi* with 48% across 143 positions.

Tim14 also appears to be derived from an alpha or gamma proteobacterial chaperone, DnaJ. This small protein (71 aas in *B. natans*) displays 49% identity across 55 positions in *B. natans* versus HSP DnaJ from *Methylomicrobium alcaliphilum*, while in *G. theta* the same protein has 52% identity across 66 positions against the gamma-proteobacterium *Marinobacter hydrocarbonoclasticus*. All three of these proteins, Tim14, mtHSP70, and Mge1 are universally present in eukaryotes and it seems very likely they derive from the bacterial endosymbiont. Tim16/Pam16 also displays similarity to an HSP protein though not as convincingly as mtHSP70, Mge1 or Tim14. For the *B. natans* Tim16 the best bacterial hit is for a protein from the alpha-proteobacteria *Ehrlichia ruminantium* with 33% identity across 58 positions while in *G. theta* the best hit is against the gamma-proteobacteria *Allochromatium vinosum* with an identity of 43% across 56 positions, though the next best hit is to an alpha-proteobacterial protein. Tim16 is only missing from *Aureococcus* (Delage, Leblanc et al. 2011) so it too seems likely derived from the endosymbiont.

Some of the TIM subunits have weak similarity to bacterial transport proteins raising the possibility that these have been co-opted to function in the import systems of eukaryotes.

The Tim23 in *B. natans*, though not in *G. theta,* has weak similarity to an archaeal and alpha-proteobacterial oligopeptide transporter. Tim22 in *G. theta* has weak similarity to alpha-proteobacterial transport proteins, and in *B. natans* to a protein of the inner membrane transport system from Firmicutes. The *B. natans* Tim44 also has weak similarity to an inner membrane transport system, this time from alpha-proteobacteria. Finally, Tim50 has weak similarity to alpha and gamma proteobacterial NLI interacting domain containing proteins. Tim50 has been linked with a family of LIM interactor-interacting factor-like (NLI-IF) phosphatases in eukaryotes (Satow, Chan et al. 2002).

The only TIM22/23 subunit found in either *B. natans* or *G. theta* that does not display some level of similarity to a proteobacterial protein of similar function is Tim17. The *B. natans* version has very weak similarity to a few hypothetical proteins from alphaproteobacteria, while the two versions from *G. theta* have very weak similarity to an acetyltransferase from proteobacteria in the case of one of the versions, while the other has equally weak similarity to a Firmicute protein. Although Tim17 is ubiquitous in eukaryotes the evidence is not sufficient to ascribe to it a bacterial origin.

All of the mitochondrial peptidases in *B. natans* and *G. theta* have bacterial origins. Both subunits of MPP, the complex that is principally involved in cleaving the presequence once the protein has reached the matrix, have strong hits to alpha proteobacterial peptidases. The MIP proteins in *B. natans* and *G. theta* also have convincing hits against bacterial oligopeptidase but in neither case are the top hits to alphaproteobacteria. Instead, the hits are scattered among the various major bacterial lineages suggesting that while MIP is bacterial in origin it was not acquired from the endosymbiont. This conclusion is supported by (Delage, Leblanc et al. 2011) who also believe the IMP subunits have an alpha-proteobacterial origin. However, in local blastp analysis against 2002 bacterial genomes and against bacterial proteins at NCBI the IMP subunits from both *B. natans* and *G. theta* consistently matched firmicute bacteria and other bacterial lineages in preference to alpha-proteobacteria.

3.3.6 Oxidative Phosphorylation

3.3.6.1 Complex I

The largest and most involved of the respiratory complexes is NADH:ubiquinone oxidoreductase. A central core of 14 subunits is found in both eukaryotes and bacteria (Gabaldon, Rainey et al. 2005). Beyond the key bacterial derived elements, eukaryotic lineages have a host of additional subunits that function both structurally and as assembly factors. Some of these additional subunits may or may not be considered part of complex I, depending on the author. As well, new subunits are constantly being identified with more sophisticated proteomic approaches. 46 subunits have been identified in *Bos taurus* (Carroll, Fearnley et al. 2003; Hirst, Carroll et al. 2003), 45 in humans (Smeitink, Sengers et al. 2001), 35 in the fungus *Neurospora crassa* (Videira and Duarte 2002), 41 in the fungus *Pichia pastoris* (Bridges, Fearnley et al. 2010), 30 in *Arabidopsis thaliana* in 2003 (Heazlewood, Howell et al. 2003) and 49 in 2010 (Klodmann, Sunderhaus et al. 2010), and 42 in *Chlamydomonas reinhardtii* (Cardol, Gonzalez-Halphen et al. 2005). It should be noted that complex I suffers from a severe case of nomenclature balkanization with at least five separate naming schemes. Except for the 12 potentially mitochondrial encoded subunits (Nad 1, 2, 3, 4, 4L, 5, 6, 7, 8, 9, 10, 11) I have chosen to use SwissProt nomenclature (but see Table 3.13 for equivalent names).

A significant portion of the core subunits is encoded on mitochondrial DNA though which ones remain, rather than being transferred to the nucleus, depends on the lineage. In humans, and other mammals, only seven (*nad 1, 2, 3, 4, 4L, 5, 6*) are found in the mitochondrial DNA while in *Reclinomonas americana*, an excavate protist, with the largest mitochondrial genome known, 12 remain (*nad 1, 2, 3 ,4, 4L, 5, 6, 7 ,8, 9, 10, 11*) (Lang, Burger et al. 1997). *B. natans* displays an intermediate level of EGT of complex I subunits with nine in the mtDNA (*nad 1, 2, 3, 4, 4L, 5, 6, 7, 9*) while *G. theta* retains the same subunits as *R. americana* in its mtDNA as do the other cryptophytes *Hemiselmis andersenii* (Kim, Lane et al. 2008) and *Rhodomonas salina* (Hauth, Maier et al. 2005). It is unusual to retain *nad8*. Besides the cryptophytes and *Reclinomonas* only *Naegleria* is recorded as having *nad8* in its mtDNA. The presence of *nad10* in mtDNA is almost as rare, present in several ciliate lineages (*Tetrahymena* (Smith, Gawryluk et al. 2007) and *Paramecium* (Barth and Berendonk 2011)) and two green algae, *Chlorokybus atmophyticus* ((Turmel, Otis et al. 2007) and *Nephroselmis olivacea* (Turmel, Lemieux et al. 1999)). While unusual, the retention of *nad10* is indicative of random processes since the taxonomic distribution of its retention is disjunctive.

Using a combination of parsing JGI automated annotations and homology searches using known complex I subunits I was able to identify 21 nucleus encoded subunits in *G. theta* for a total of 33 and 25 in *B. natans* for a total of 34. *B. natans* and *G. theta* share a similar complement of subunits (Table 3.13). *B. natans* has a version of all those found in *G. theta* except for NIPM. Nor could a homolog be detected in RNA-Seq data from *B. natans* or several other chlorarachniophytes (*Lotharella globosa*, *Lotharella oceanica*). The one subunit found in *B. natans* that was not identified in *G. theta* was NB2M (B12). The *B. natans* version of NB2M was found in an area of the genome with no predicted gene models using the *Chlamydomonas reinhardtii* copy (EDP07174) in a tblastn search. The similarity was very weak so the match may be spurious.

Gabaldon et al. (Gabaldon, Rainey et al. 2005) proposed that beyond the 14 core subunits that came from the mitochondrial ancestor, 21 additional core subunits are universally found in eukaryotic organisms that have a complex I (grey in Table 3.13). I was unable to identify through bioinformatic means eight of these 35 subunits. *B. natans* was missing seven (NIPM, NUYM, NESM, NIMM, NUJM, NUXM, CI84) while *G. theta* was also missing seven (NUYM, NESM, NIMM, NUJM, NUXM, CI84, NB2M). Is the inability to detect these eight genes a consequence of sequence divergence making it impossible to find them through homology searches or does it represent lineage specific losses that do not counter the claims of a core eukaryotic assemblage? It should be noted that 10 of the "core" subunits were not found in a tandem mass spectrometry study of the ciliate *Tetrahymena thermophila* (Smith, Gawryluk et al. 2007) and of these, four were also not found in *G. theta* and *B. natans* (NESM, NIMM, CI84, NUJM) while NB2M was not found in *G. theta*. Mitochondrial proteomic surveys in protists lag far behind plant and metazoan studies so claims of eukaryotic universality may change. For example, it has been suggested that the two versions of N7BM found in all species except *C. reinhardtii* (Gabaldon, Rainey et al. 2005) resulted from an ancient duplication. However, only one version of N7BM was detected in *G. theta*, *B. natans*, *Tetrahymena* (Smith, Gawryluk et al. 2007) or in blastp searches of the protist genomes of *Thalassiosira pseudonana*, *Phytophthora ramorum* and *Ostreococcus lucimarinus*.

Table 3.13. Complex I genes from *B. natans* and *G. theta.* Gene names in grey are considered part of a universal eukaryotic core. ME=mitochondrion encoded. nf=not found.

Gene name	*G. theta*	*B. natans*
Nad1	ME	ME
Nad2	ME	ME
Nad3	ME	ME
Nad4	ME	ME
Nad4L	ME	ME
Nad5	ME	ME
Nad6	ME	ME
Nad7, NUCM	ME	ME
Nad8, NDUFS8, NUIM	ME	55021
Nad9, NUGM	ME	ME
Nad10, NUKM, NDUFS7	ME	49058
NUAM, Nad11, 75 kDa, NDUFS1	ME	46295
NUHM, NDUFV2, 24 kDa	151242	55826
NUBM, NDUFV2, 51 kDa	89568	85575
NUEM, 39 kDa, NDUFA9	160957	85967
N5BM, B14.7, NDUFA11	111845	88586
NIPM, 15 kDa, NDUFS5	70946	nf
NUMM, 13 kDa, NDUFS6	165716	145867
NUYM, AQDQ, NDUFS4	nf	nf
NESM, ESSS, NDUFB11	nf	nf
NIMM, MWFE, NDUFA1	nf	nf
NIDM, PDSW, NDUFB	153600	89203
NUJM, PGIV, NDUFA8	nf	nf
ACPM, SDAP, NDUFAB1	110892	43207
NI2M, B22, NDUFB9	155205	135978
NB8M, B18, NDUFB7	97918	63670
N7BM, B17.2, DAP13	112529	86859
NB6M, B16.6, NDUFA12	146129	88478
NB4M, B14, NDUFA6	150588	137303

Gene name	*G. theta*	*B. natans*
NUFM, B13,NDUFA5	154274	56530
NB2M, B12, NDUFB3	nf	Found in putative gene-poor region using *C. reinhardtii* homolog as query
NI8M, B8, NDUFA2	150904	39560
NUXM	nf	nf
CI30, CIA30	62892	nf
CI84, CIA84	nf	nf
CA1	152983 98506 155008 112464	86584
NUVM, B15, NDUFB4	nf	nf
NIGM, AGGG	nf	nf
NUUM	nf	nf
NUML, MLRQ, NDUFA4	99492	51310
MidA	133463	68792
Ndufaf3	72009	139677
Foxred1	nf	38746

Proteomic studies always reveal additional complex I subunits that are novel or limited in their taxonomic distribution. Both *G. theta* and *B. natans* appear to have a NUML subunit which was thought to be exclusively metazoan (Gabaldon, Rainey et al. 2005). This subunit was also detected in a proteomic study of *Acanthamoeba castellanii* (Gawryluk, Chisholm et al. 2012). Another intriguing gene that was found in *A. castellanii* and that also appears to be present in *G. theta* and *B. natans* is carbonic anhydrase (Gawryluk and Gray 2010). Until the 2010 study, gamma class carbonic anhydrases were only known from green plants and green algae as well as bacteria*.* *G. theta* has four genes that code for this class of carbonic anhydrases while *B. natans* has a single gene. All of the genes except for gt152983 have presequences that strongly support mitochondrial targeting and it should be noted that some legitimate mitochondrial associated gamma CAs do not have predicted mitochondrial targeting peptides (Gawryluk and Gray 2010).

The widespread appearance of gamma CAs across the breadth of Eukaryota (Gawryluk and Gray 2010) was seen as an indication that the ancestral eukaryotic core of complex I should be expanded beyond the 35 already identified (Gabaldon, Rainey et al. 2005). Given the bacterial homologs of gamma CAs it also suggests that the protomitochondrial contribution to the mitochondrial proteome is greater than the traditional 14 subunits. Indeed, at least six of the additional 21 subunits have been found to have alpha-proteobacterial homologs (ACMP, NUEM, NI8M, CI30, NUMM and N7BM) (Gabaldon, Rainey et al. 2005). All six of these subunits are found in *B. natans* and *G. theta* as well as *Tetrahymena* (Smith, Gawryluk et al. 2007) suggesting that they are not only part of the eukaryotic core of complex I but products of EGT from the primitive mitochondrion prior to the expansion of eukaryotic diversity.

3.3.6.2 Complex II

Complex II or succinate dehydrogenase, though simple in structure, plays a dual role in the mitochondrion, acting as the second complex in the electron transport chain and as a component in the tricarboxylic acid cycle. It is composed of four subunits, SdhA, SdhB, SdhC and SdhD. SdhA and SdhB are both hydrophilic subunits bound together in the mitochondrial matrix while SdhC and SdhD are hydrophobic molecules bound to the inner mitochondrial membrane. Other subunits have been noted in some organisms but these additional elements are lineage specific (Huang, Taylor et al. 2010).

SdhA and SdhB are large, highly conserved proteins, generally with over 70% identity between homologs of major lineages and are universally nucleus encoded and were easily identified in both *B. natans* and *G. theta* (Table 3.14). The other subunits, SdhC and SdhD are small and poorly conserved even within major lineages (Burger, Lang et al. 1996). In *G. theta*, both SdhD and SdhC are encoded in the mitochondrial genome which mirrors the situation in two other cryptophytes, *Hemiselmis andersenii* (Kim, Lane et al. 2008) and *Rhodomonas salina* (Hauth, Maier et al. 2005). Interestingly, SdhC and SdhD are next to each other in the mitochondrial genome, an arrangement also seen in *R. salina*, *H. andersenii*, as well as the excavate *Reclinomonas americana* and bacteria (Burger 1996). This conservation of gene order in mitochondrial DNA and bacterial DNA has been interpreted as indicative of the monophyly of mitochondria (Burger, Lang et al. 1996). As an example of the poor amino acid conservation, SdhD homologs for *G. theta* and *Rhodomonas* share only 23 positions out of 91 (25%). In *B. natans* SdhC and SdhD are both nucleus encoded. The *B. natans* SdhC (52072) has a presequence that generates both strong SignalP scores (HMM 0.953 NN 4/5) and mitochondrial targeting TargetP scores (0.907 M non-plant).

Table 3.14. Complex II subunits in *G. theta* and *B. natans*. ME=mitochondrial encoded. t=found through text search. h=found or confirmed through homology searches.

	G. theta	*B. natans*
SdhA	159138 h	53042 th
SdhB	159012 h	87410 th
SdhC	ME	52072 h
SdhD	ME	35639 h

3.3.6.3 Complex III

Complex III, also known as Ubiquinol-cytochrome c oxidoreductase or cytochrome bc1 complex is the third major component in the respiratory electron transport chain. In mitochondria it is embedded in the inner membrane. A structure similar in function, known as the cytochrome b6f complex, is located in the thylakoid membrane of plastids. Eukaryotic complex IIIs consist of the 3 catalytic core subunits (Smith, Fox et al. 2012) – cytochrome b (COB), cytochrome c1 (CYT1), Rieske iron-sulfur protein (RIP1)– that are also found in the prokaryotic equivalent. However, unlike bacterial complex IIIs, those found in eukaryotes consist of additional subunits. The most studied examples of complex III are from bovine heart cells (Iwata, Lee et al. 1998) with 11 additional subunits and yeast (Schagger and Pfeiffer 2000) with 10 subunits.

Typically, cytochrome b is mitochondrion encoded with the rest of the structural subunits nucleus encoded (O'Brien, Zhang et al. 2009) as are any assembly factors. This is the case in *B. natans* and *G. theta* as well. A search for these subunits showed that *G. theta* and *B. natans* have very similar complexes (Table 3.15). Structurally, they possess the same subunits and both are missing Qcr8 and Qcr10. Of special interest are the two largest subunits Cor1 and Cor2. It has been observed that Cor1 and Cor2 are very similar, if not in fact homologous, to the mitochondrial processing peptidase (MPP), which has two subunits, alpha and beta (Gakh, Cavadini et al. 2002). Animals and yeast have Cor1 and Cor2 proteins as well as alpha and beta subunits of the MPP. In plants Cor1 is identical to the MPP beta subunit while Cor2 plays the role of the alpha subunit of the MPP. In fact, in plants the MPP is part of complex III. It would appear that *G. theta* and

B. natans have a plant-like Cor1/beta-MPP, Cor2/alpha-MPP arrangement. Blastp searches using the four yeast homologs only returned two hits for each of the genomes. What is not clear in *B. natans* and *G. theta* is whether these bifunctional proteins are only physically present as part of complex III. In *Neurospora crassa* there is only one gene that codes for the proteins Cor1 and beta-MPP but complex III and the MPP are separate physical entities (Gakh, Cavadini et al. 2002).

Table 3.15. Complex III proteins in *B. natans* and *G. theta.* ME=mitochondrial encoded. nf=not found. t=found through text search. h=found or confirmed through homology searches.

	Structure Assembly	*G. theta*	*B. natans*
Cor1	S	83986 h	53395 h
Cor2	S	106014 h	91482 h
COB	S	ME	ME
CYT1	S	164178 th	56213 t
RIP1	S	67785 th	90651 h
QCR6	S	91152 th	55834 t
QCR7	S	156324 th	90551 t
QCR8	S	nf	nf
QCR9	S	153289 th	128137 th
QCR10	S	nf	nf
Cbs1	A	nf	nf
Cbs2	A	nf	nf
Cbp1	A	nf	nf
Cbp2	A	nf	nf
Cbp3	A	101987 th	81100 t
Cbp4	A	nf	nf
Cbp6	A	nf	nf
Cyt2	A	73289 h 65435 h	36033 t h
Cyc2	A	nf	nf
Bca1	A	nf	nf
TTC19	A	nf	nf
Bcs1	A	74509 h 153155 h 86537 h	86319 h 86666 h
Mtzm1	A	nf	nf

As with the other complexes there are a number of assembly factors. Most of the assembly factors identified in yeast do not appear to have homologs in *B. natans* or *G. theta* (Table 3.15). Only three out of the 13 were identified in both *B. natans* and *G. theta* (Cyt2, Cbp3 and Bcs1). Of the seven factors (Cbs1, Cbs2, Cbp1, Cbp2, Cbp3, Cbp4, Cbp6) in yeast that are responsible for translational activation of COB mRNA (Smith, Fox et al. 2012) only Cbp3 was detected in *B. natans* and *G. theta*. This inability to detect homologs may be due to divergence of the genes. The blastp results for Cbp3 were quite weak. In *G. theta* the top hit for Cbp3 was *Micromonas* with an e-value of 2e-20 while in *B. natans* the best hit was for *Thalassiosira* with an e-value of 1e-11. However, the same assembly factors are absent from humans, *Arabidopsis thaliana* and *Chlamydomonas reinhardtii* (Cardol, Gonzalez-Halphen et al. 2005) suggesting that the missing ones are lineage specific to fungi. Cyt2, a heme lyase responsible for the covalent attachment of heme c to apo-cytochrome c1 (Cyt1) is present in both genomes. *G. theta* appears to have two versions while *B. natans* has a single mitochondrial heme lyase.

3.3.6.3.1 Assembly factor Bcs1

The assembly factor Bcs1, which is responsible for the insertion of Rieske Fe/S (Rip1) into the bc1 complex is an AAA-ATPase protein (Wagener and Neupert 2012) whose phylogenetic profile reveals a deep division between metazoans/fungi and plants (Frickey and Lupas 2004). Bcs1 proteins share a basic structure that can be divided into two halves (Wagener and Neupert 2012). The first half contains an N-terminal region that is exposed to the intermembrane space followed by a transmembrane region next to a domain that contains mitochondrial targeting signals. The second half contains an AAA domain, with a Walker A motif, a Walker B motif and an Arginine finger (RxxR). The second half is very similar in both plants and metazoans/fungi while the first half is different with plants having an AAA associated domain and metazoans/fungi having a unique domain called BCS1_N. Blastp searches using the yeast BCS1 identified three candidate genes in the *G. theta* genome (74509, 153155, 86537). All three code for proteins that contain an AAA domain in the second half with the requisite motifs (Walker A, Walker B, arginine finger). All three also possess a transmembrane domain in the first half and a targeting signal domain that is of the fungal/animal type. In blastp results the three *G. theta* proteins have best hits against fungal and amoeboid BCS1 sequences (best blastp hit for 86537 – *Dictyostelium discoideum* 3e-73). In RAxML trees (data not shown) the three *G. theta* Bcs1 protein sequences consistently group together, indicating that they are paralogs.

The situation in *B. natans* is somewhat less clear. Three candidates were also identified through homology searches (86319, 89502, 86666). 86319 is clearly a metazoan/fungal type BCS1 possessing all of the necessary motifs and domains including the BCS1_N domain in the first half. Its top hit is *Trichoplax* with an e-value of 2e-145. While the protein model for 89502 has a Bcs1 type second half, including an arginine finger, its first half is truncated and does not possess a domain. The genomic space upstream of 89502 is empty and thus has the potential for accommodating an N-terminal with a domain but attempts to find additional coding regions, through tblastn searches and interrogating the RNA-Seq data, failed. In blastp searches 89502 has top hits that are fungal (*Clavispora* 7e-58) suggesting that it is a metazoan/fungal type Bcs1. The gene model for 86666 is very poor and includes three genomic gaps. The encoded protein possesses a BSC1_N domain in the first half and an AAA domain in the second but without an arginine finger. Blastp hits, as with the others, consistently match fungal and amoeboid BCS1 proteins (top hit *Dictyostelium purpureum* 5e-40) but suggest that the protein model starts too early by 100-200 amino acids. Given the presence of the unique BCS1_N domain in 86319 and 86666 it is reasonable to assume that they are metazoan/fungal type Bsc1 proteins but the assignment of 89502 remains unresolved. A search of the contigs generated from RNA-Seq data for the chlorarachniophyte *Lotharella globosa* produced three Bsc1 candidates all of which had BSC1_N domains in the first half suggesting that it is entirely possible for *B. natans* to have three metazoan/fungal type Bcs1 proteins.

The taxonomic distribution of the BCS1_N domain is curious. Besides being present in animals and fungi it turns up in a single stramenopile (*Aureococcus anophagefferens*), various amoeboid protozoa, the excavates *Leishmania* and *Trypanosoma*, a single green alga (*Micromonas* RCC299) and a handful of alveolates including various *Plasmodium* species, *Toxoplasma gondii*, the dinoflagellate *Karlodinium micrum*, *Perkinsus marinus* and *Neospora caninum*. These results however do not imply that all other eukaryotes have a plant type Bcs1 since that arrangement of domains seems limited strictly to plants. Green algae, stramenopiles (other than *Aureococcus*), haptophytes, and red algae appear not to possess Bcs1 proteins or at least ones that conform to either metazoan/fungal or plant types. Blastp searches in these lineages that apparently lack Bcs1 homologs invariably turn up proteins that contain the AAA domain, often in a 26S protease regulatory subunit.

3.3.6.4 Complex IV

Cytochrome c oxidase or complex IV contains up to 13 subunits in mammals (Tsukihara, Aoyama et al. 1996) with three (COX 1, 2, 3) encoded in the mitochondrion with the rest nucleus encoded. Bacteria typically have four subunits. Yeast, with a different nomenclature, has at least 11 subunits while *Arabidopsis* has 14 (Heazlewood, Tonti-Filippini et al. 2004).

Table 3.16. Complex IV subunits in *G. theta* and *B. natans.* ME=mitochondrial encoded. nf=not found. t=found through text search. h=found or confirmed through homology searches.

	Structure or Assembly	*G. theta*	*B. natans*
Subunit 1 (Cox1)	S	ME	ME
Subunit 2 (Cox2)	S	ME	ME
Subunit 3 (Cox3)	S	ME	ME
Subunit 4 (Cox4)	S	nf	nf
Subunit 5b (Cox5b)	S	160714 t	86246 t
Subunit 6a (Cox6a)	S	102457 t	136989 t
Subunit 6b (Cox6b)	S	149927 t	86518 t
Subunit 10 (Cox10)	A	73866 h	142491 th
Assembly protein Yah1	A	72919 h	26430 h
Assembly protein (Cox11)	A	107319 th	57639 h
Assembly protein (Cox15)	A	104486 t	85155 t
Copper chaperone (Cox17)	A	74478 h	47709 t
Assembly protein (Cox19)	A	60979 t	48203 t
Sco1/Sco2 family protein	A	116071 t	57828 th
Sco1/Sco2 family protein	A	134392 th	nf
Pet191	A	108676 t	65300 h
Shy1	A	120094 t	86772 h
CmC1-like	A	115776	141972
CmC1-like	A	101528	91213
CmC1-like	A		141034

As with all eukaryotic organisms cox1 is encoded in the mitochondrial genome in *B. natans* and *G. theta*. However, as discussed in the chapter on NUMTs a small portion of cox1 from the mitochondrial genome of *G. theta* has been integrated into the host nuclear genome. Its truncated nature suggests that it is nonfunctional. Cox2 and cox3 are also mitochondrial encoded in *B. natans* and *G. theta*.

Homolog searches using *Thalassiosira* (XP_002296945) and *Albugo* (CCA17261) versions of cox4 against all models returned no hits in *B. natans* or *G. theta*. Cox4, which is exclusively eukaryotic, does not appear to be present in plants or algae other than *Thalassiosira* and some oomycetes. Nor does *G. theta* or *B. natans* have two copies of subunit 6b like in plants, fungi and animals. Interestingly, it has been suggested that the duplication of subunit 6b occurred prior to the divergence between plants and metazoans (Cavallaro 2010). However, green algae, like other protists, have only one version of subunit 6b.

A number of the complex IV subunits identified in *Arabidopsis* (Heazlewood, Tonti-Filippini et al. 2004) appear to be specific to the green algal/green plant lineages. Homology searches (blastn, tblastn) using *Arabidopsis* proteins cox x1-x6 did not turn up any candidates in *B. natans* or *G. theta*. Similarly, no homologs were detected for *Arabidopsis* complex IV subunits 5c (AEC10834) or 6c (ABD38862).

Besides the subunits that actually comprise complex IV a number of proteins are considered essential for the complex's assembly. The complement of assembly proteins has been studied most in yeast where at least 32 have been identified (Mick, Fox et al. 2011) with roles such as translational regulation, membrane insertion and processing, heme a synthesis, copper transport and insertion, and chaperone-like functions. Many of these are specific to yeast or more likely so divergent that the homologs are impossible to detect in organisms less studied than yeast. Many of these assembly proteins are quite short which further reduces the ability to detect homologs. Through text searches of the JGI annotations and homology searches using blastp and tblastn I have identified a number of these complex IV assembly proteins in *B. natans* and *G. theta*. Homologs for all of the yeast assembly proteins involved in Heme a synthesis (Cox10, Cox15, Yah1) were found in both genomes. Homologs for these are also found in mammals (Mick, Fox et al. 2011). Most of the proteins involved in copper transport and insertion were also found including Cox11, Cox17, Cox19 and Sco1. *G. theta* appears to have a homolog for Sco2, unlike *B. natans* which is similar to the situation in mammals (Mick, Fox et al. 2011). Neither genome had a definitive homolog for Cox23, which is involved in copper trafficking. However, both organisms have several proteins of unknown function that share a domain linked with Cox23 as well as several other assembly proteins like Cox17 and Cox19. An investigation of these proteins (gt115776, gt101528, bn141972, bn91213, bn141034) and their domains (see section above on Twin CX9C proteins) strongly suggest that they are mitochondrial proteins and potentially involved in the biogenesis of complex IV.

No homologs were found in *B. natans* and *G. theta* for any of the yeast assembly proteins involved in translational regulation or membrane insertion and processing though it is unlikely that these roles have been abandoned in these organisms. Homologs for two assembly factors (Shy1, Pet191) with chaperone-like function were detected through blastp searches for both organisms.

3.3.6.5 Complex V

ATP synthases are transmembrane proteins found in bacteria, mitochondria, plastids and vacuoles that act as proton pumps. In mitochondria the primary role is to couple the proton gradient across the inner membrane to the synthesis of ATP. In this role it is the last complex (V) of the oxidative phosphorylation pathway. While ATP synthases are associated with numerous organelles and pump a diverse range of ions they share a similar overall structure consisting of the catalytic core (F1) and the proton translocating portion (FO) (Hong and Pedersen 2003). The F1 portion is highly conserved and invariably, regardless of species, is comprised of several alpha and beta subunits alternating in a cap-like formation with a gamma subunit protruding from it and connecting it to the FO portion. Also part of F1 is an epsilon subunit that attaches to the end of the gamma subunit while the delta subunit caps the alpha/beta cap. The F1 portion, which is embedded in the membrane, is less conserved and the number of subunits varies. The mitochondrial version of F1 is particularly rich in subunits compared with its counterpart in bacteria which typically has three subunits (a, b, c) or chloroplasts with four (a, b, b′,c). Mitochondrial F1 in yeast has been reported to have 14 subunits while animals typically have 10-12 (Hong and Pedersen 2003).

For the most part *G. theta* and *B. natans* have a similar and typical complement of mitochondrial ATP synthase subunits (Table 3.17) with some mitochondrial encoded and others nucleus encoded and subsequently targeted. All of the highly conserved F1 subunits were found in both species. Since the FO subunits are less conserved and vary in number across lineages it was not surprising that there were some differences between *B. natans* and *G. theta*. Unlike *B. natans, G. theta* has a FO d chain gene. In *B. natans* it was not possible to detect a homolog of the FO b subunit either through text searches of the JGI annotation or through blastp or tblastn searches of the full genome. Although the b subunit is indispensible its sequence is poorly conserved so failure to identify a version should not be taken as an absence. In *G. theta* the b subunit is mitochondrial encoded.

Table 3.17. Complex V subunits in *G. theta* and *B. natans.* ME=mitochondrial encoded. nf=not found.

Protein	Structure or Assembly	*G. theta*	*B. natans*
F1 alpha	S	ME	ME 89115 78004
F1 beta	S	64866	92352
F1 gamma	S	164186	48015
FO OSCP	S	150228	89534
F1 delta	S	151115	53251
F1 epsilon	S	152682	61318
F1 assembly factor 1 (Atp11)	A	117769	146631
FO subunit 6(a)	S	ME	ME
FO subunit 6(b)	S	ME	nf
FO subunit 9 (c)	S	ME	ME
FO subunit 8 (A6L)	S	ME	ME
Chaperone (Atp12)	A	103155	87321
FO d chain	S	160877	nf

In both *B. natans* and *G. theta* the mitochondrial and plastid genomes encode an alpha subunit. *B. natans* has two additional alpha subunits that are nucleus encoded and that are targeted to the mitochondrion (Table 3.17) based on TargetP scores and homology. Why *B. natans* has three mitochondrial alpha subunits is unclear. Although the F1 portion is composed of three alpha subunits alternating with three beta subunits in all known cases the subunits are identical. *Homo sapiens* have three alpha isoforms but these are products of alternative splicing rather than from separate genes. *G. theta* has a nucleus encoded alpha subunit (102682) that is unusual in that SignalP returns a strong SP signature (HMM .995 NN 5/5) suggesting that it is plastid targeted while TargetP scores suggest that it is mitochondrial targeted (0.716 M M 3). In Blastp searches 102682 retrieves plastid versions (either encoded or targeted) of the alpha subunit including from stramenopiles and perhaps most significantly from red algae. The best hit is for a nucleus encoded alpha subunit with a signal peptide prediction (0.817 TargetP) from *Phaeodactylum tricornutum*. Incidentally, *P. tricornutum* also has a plastid encoded alpha subunit. The evidence certainly suggests that 102682 is plastid targeted, and the product of ancient EGT, but dual targeting to the mitochondrion does exist as a possibility given the TargetP scores and that mitochondria also possess large quantities of this subunit.

3.3.7 Mitochondrial Genome Comparison

A partial mitochondrial genome sequence of *B. natans* (HQ840955) was acquired from NCBI. It consists of a single contig of 36,375 bases. During the course of the *B. natans* nuclear genome project a contig was created from sequencing reads generated from contaminating mitochondrial DNA. This contig is 37,351 bases in length. The two contigs were compared using GAP4 from the Staden package (Bonfield, Smith et al. 1995). 21 single nucleotide differences were found in the overlap between the two sequences. Examination of the Sanger reads from the nuclear genome project confirmed that the NCBI version is correct. The extra 976 bases from the JGI assembly did not contain any genes.

As with *B. natans* a *G. theta* mitochondrial genome was sequenced in the course of the nuclear genome project. It consisted of two contigs, one 37512 bps, the other 3110 bps. Open reading frames equal to or greater than 50 amino acids were generated using the Artemis package (Rutherford, Parkhill et al. 2000). The resulting open reading frames (ORFs) were compared against a local database of 36,006 protein coding genes from 2490 mitochondrial genomes obtained from the NCBI FTP:Genome site (ftp://ftp.ncbi.nlm.nih.gov/genomes/). Among the mitochondrial genomes searched against were those from two cryptophytes, *Hemiselmis andersenii* (Kim, Lane et al. 2008) and *Rhodomonas salina* (Hauth, Maier et al. 2005) as well as the gene-rich genome of the excavate *Reclinomonas americana* (Lang, Burger et al. 1997). The *G. theta* ORFs that did not show homology with known mitochondrial genes and did not lie within a region already designated as coding for a gene were searched against the NCBI protein database.

Reclinomonas americana has the largest complement of protein coding genes of any known mitochondrial genome (Lang, Burger et al. 1997). It has 67 protein coding genes in 69 kb. Plants often have larger mitochondrial genomes in terms of base pairs but much of the DNA in these bloated genomes is derived from small repetitive sequences, duplicated genes or foreign DNA. For example, *Vitis vinifera* has a mitochondrial genome size of 773 kb yet only 37 genes that code for proteins (Goremykin 2009). It has 12 mitochondrial pseudogenes. Interestingly, 42.2% of the chloroplast genome has been incorporated into the *Vitis* mitochondrial genome, including 27 plastid genes with intact ORFs and 41 plastid pseudogenes (Goremykin, Salamini et al. 2009). On the opposite end of the spectrum all animal mitochondrial genomes analyzed to date have 14 or fewer protein coding genes with most having 13 (Burger, Gray et al. 2003).

The *Reclinomonas* mitochondrial genome, with its 67 protein coding genes has long been seen as the genome most closely resembling the ancestral protomitochondrion (Lang, Burger et al. 1997). Indeed, all of the protein sets from all of the mitochondrial genomes sequenced to date appear to be a subset of that identified in *Reclinomonas*, apart from the unique ORFs with no known function, or genes that appear to have been transferred from the nucleus or the chloroplast (Gray, Lang et al. 2004). In other cases of unique genes, the annotation is incorrect. *Pythium ultimum* has a gene annotated as the ribosomal protein *l3* (ACZ44473) that would be a first for mitochondrial genomes. However, the gene is clearly a ribosomal *s13*, which is a frequent constituent of mitochondrial genomes. The remarkable consistency of mitochondrial genome coding capacity has been seen as indicative of a single origin of this organelle. Moreover, it suggests that the reduction of the engulfed alpha-proteobacterium, in terms of function and gene complement, was relatively rapid, occurring prior to the diversification of eukaryotes into the major lineages we now have. Clearly there were lineage specific losses over time such as the highly reduced genomes of animals but those losses were all from a small pool of common genes.

Obviously mitochondrial genomes in isolation are inadequate for inferring complete mitochondrial proteomes consisting of hundreds of proteins. As part of the initial reduction most of the necessary genes were transferred from the alpha-proteobacterial endosymbiont to the host nucleus. Since most of the mitochondrial proteins are derived from nucleus encoded genes the proteins are targeted back to the mitochondrion. Most mitochondria share the basic function of respiration and coupled oxidative phosphorylation. This is mirrored in their genomes as well since, apart from ribosomal proteins, the vast majority of genes that still reside on mitochondrial DNA are involved in the structure or assembly of the complexes that make up the electron transport and ATP synthesis machinery. Again, however, the mitochondrial genomes do not code for all the necessary proteins required for respiration and ATP synthesis. The complexes are a mosaic of nuclear and mitochondrial encoded proteins. Though the number of subunits that are encoded on the mitochondrial DNA may vary across lineages, which subunits are present is limited to a small consistent collection that is essentially a subset of the proteins encoded by the *Reclinomonas* mitochondrial genome and that reflects the bacterial origins of the respiration machinery. Elaborations on the basic bacterial complexes result from exclusively eukaryotic inventions or the recruitment of bacterial subunits not from the alpha-proteobacterial ancestor (Lithgow and Schneider 2010).

Table 3.18. Mitochondrial genomes with the largest protein coding gene sets.

	# protein coding genes	Lineage
Reclinomonas americana	67	Excavate
Malawimonas jakobiformis	47	Excavate
Chlorokybus atmophyticus	46	Viridiplantae
Chaetosphaeridium globosum	43	Viridiplantae
Naegleria gruberi	42	Excavate
Chara vulgaris	42	Viridiplantae
Hemiselmis andersenii	42	Cryptophyte
Rhodomonas salina	41	Cryptophyte
Physcomitrella patens	41	Viridiplantae
Marchantia polymorpha	40	Viridiplantae

How do the mitochondrial genomes of *G. theta* and *B. natans* compare with already sequenced genomes particularly *Reclinomonas*? I identified 39 protein-coding genes in *G. theta* and 27 in *B. natans*. *G. theta*'s 39 protein coding genes are comparable to that seen in the two other cryptophytes that have had their mitochondrial genomes sequenced, *Hemiselmis andersenii* with 42 and *Rhodomonas salina* with 41 (Table 3.18). Among mitochondrial genomes cryptophytes have some of the most gene rich (Table 3.18), being in the top ten with several excavates and green algae/primitive plants. *B. natans*, with 27 protein coding genes, occupies the middle ground between the relatively sparse mitochondrial genomes of animals and fungi and the more complex plant, cryptophyte and excavate ones. The gene complement of *B. natans* is comparable to that seen in Rhodophyta (23-33) (Gray, Lang et al. 2004) and slightly less than for most stramenopiles (30-35).

A gene-by-gene comparison of the three sequenced mitochondrial genomes from cryptophytes reveals few differences. *Hemiselmis* has a SecY-independent protein translocase component (TatA) that was not identified in *G. theta* or *Rhodomonas* (Table 3.19). *Hemiselmis* also has two ribosomal proteins (L10, L31) that *G. theta* lacks while *Rhodomonas* lacks only L31. The only gene that *Rhodomonas* has that *Hemiselmis* does not is a group II intron reverse transcriptase. *G. theta* also lacks this reverse transcriptase. Since the *G. theta* mitochondrial genome is partial some of the missing genes may be present on the un-sequenced portion. It should be noted that the missing *G. theta* genes were not found in the nuclear genome.

When compared to *Reclinomonas*, *G. theta* and the other cryptophytes have the typical smattering of mitochondrial encoded complex I genes, cytochrome c oxidase subunits, large and small ribosomal proteins and succinate:ubiquinone oxidoreductase subunits (Table 3.19). The only category completely missing is the RNA polymerase subunits, an absence shared by all other mitochondrial genomes. *B. natans* has a similar distribution of typical mitochondrial encoded protein encoded genes and again lacks the multiple RNA polymerase subunits. The evolutionary history of mitochondrial RNA polymerase in *Reclinomonas* is intriguing. The four components of the *Reclinomonas* RNA polymerase are demonstrably eubacterial (Lang, Burger et al. 1997) suggesting that they constitute the original protomitochondrial RNA polymerase. In other organisms the mitochondrial RNA polymerase is a single polypeptide homologous to phage RNA polymerases and typically encoded in the nucleus (Gray, Lang et al. 2004). At some point, clearly early on, the bacterial RNA polymerase found in *R. americana* was replaced by a different system. *G. theta* clearly has a single subunit RNA polymerase (182930) that is nucleus encoded rather than the multiple subunits found on the

Reclinomonas mitochondrial DNA. *B. natans* probably also has a single subunit RNA polymerase but the genomic area that contains weak matches to other single subunit RNA polymerases is riddled with gaps and a useful gene model cannot be generated. Curiously, RNA-Seq data for *B. natans* did not cover the area in question. However, RNA-Seq data from the chlorarachniophyte *Lotharella globosa* yielded a contig that clearly corresponds to a single subunit phage type mitochondrial RNA polymerase.

If one assumes that *Reclinomonas* represents a protomitochondrial baseline against which one can compare other mitochondrial genomes, then missing genes may have been transferred to the host nucleus in other lineages, particularly if they code for proteins that are core constituents of the mitochondrial machinery. This certainly appears to be the case for the respiration and ATP synthesis complexes. The *G. theta* mitochondrial genome encodes five of the six *Reclinomonas* ATP synthesis subunits with the missing one, gamma, encoded in the nucleus, all 12 of the NADH dehydrogenase subunits, five of six cytochrome c oxidase subunits with the missing one nucleus-encoded and two of three succinate:ubiquinone oxidoreductase subunits, again with missing subunit 2 nucleus encoded (Table 3.19). *B. natans* with its reduced gene complement compared to *G. theta* has even more genes for core subunits that have been transferred to the host nucleus. Nad 10, 11 and 8 are products of EGT as are subunits 2, 3, and 4 of succinate:ubiquinone oxidoreductase. For ATP synthase, as with *G. theta*, the gamma subunit has been transferred. Curiously, subunit b was not identified in the *B. natans* nuclear genome.

A comparison of *G. theta* and *B. natans* ribosomal genes against those found in the *Reclinomonas* mitochondrial genome yields a mixed pattern. Some of the missing genes have been transferred to the host and have clear mitochondrial targeted versions (L1, L11, L2, L20, L27, L34) in both species. In other cases I was unable to detect any nucleus encoded versions that had clear mitochondrial transit peptides (L10, L18, L31, S10) in either genome suggesting that these ribosomal proteins are not present in the mitochondrion. In other cases there was a mitochondrial version, either nuclear or mitochondrial encoded in one species but not the other (Table 3.19). For example, in *G. theta* S1, S2 and S13 are mitochondrial encoded while in *B. natans* no obvious mitochondrial versions were found through homology searches. Curiously, *G. theta* has a propensity to transfer large subunit genes with eight of the 15 *Reclinomonas* mitochondrial encoded proteins nucleus encoded while none of the 12 small subunit genes have been transferred. *B. natans* exhibits a less striking bias in the type of ribosomal gene transferred. Five of the large subunit genes have been transferred to the nucleus but only two of the small subunit genes.

One complicating factor is that both organisms have four independent ribosomal systems, one for each of the main compartments – cytosol, PPC, plastid and mitochondrion - that have protein subunits encoded in the nuclear genome. This opens up the possibility of dual targeting, especially for ribosomal proteins destined for either the plastid or the mitochondrion since the ribosomal structures in both organelles are derived from bacterial ancestors. In several cases where a mitochondrial targeted protein gene was missing a plastid version was present, as well as cytoplasmic versions. No clear cases of dual targeting were identified for organellar ribosomal proteins, but the possibility exists since not all mitochondrial targeted ribosomal proteins have classical presequences. Eukaryotic and bacterial ribosomal proteins are sufficiently different that they can be differentiated.

Table 3.19. Comparison of protein coding genes from select mitochondrial genomes. x=present. NE=nucleus encoded. nf=not found.

Reclinomonas	*G. theta*	*Hemiselmis**	*Rhodomonas**	*B. natans*
ATP synthase FO subunit 6	x	x	x	x
ATP synthase FO subunit 8	x	x	x	x
ATP synthase FO subunit 9	x	x	x	x
ATP synthase F1 subunit alpha	x	x	x	x
ATP synthase F1 subunit gamma	NE			NE
ATP synthase subunit b	x	Orf7	Orf2	nf
NADH dehydrogenase subunit 1	x	x	x	x
NADH dehydrogenase subunit 10	x	x	x	NE
NADH dehydrogenase subunit 11	x	x	x	NE
NADH dehydrogenase subunit 2	x	x	x	x
NADH dehydrogenase subunit 3	x	x	x	x
NADH dehydrogenase subunit 4	x	x	x	x
NADH dehydrogenase subunit 4L	x	x	x	x
NADH dehydrogenase subunit 5	x	x	x	x
NADH dehydrogenase subunit 6	x	x	x	x
NADH dehydrogenase subunit 7	x	x	x	x
NADH dehydrogenase subunit 8	x	x	x	NE

Reclinomonas	*G. theta*	*Hemiselmis**	*Rhodomonas**	*B. natans*
NADH dehydrogenase subunit 9	x	x	x	x
Orf169	nf			nf
Orf717	nf			nf
RNA polymerase subunit alpha	nf			nf
RNA polymerase subunit beta	nf			nf
RNA polymerase subunit beta'	nf			nf
Sec-independent protein translocase component TatC	x	x	x	nf
SecY-independent protein translocase component TatA	nf	Orf60		nf
SecY-type transporter protein	nf			nf
apocytochrome b	x	x	x	x
component involved in Haem biosynthesis	NE			NE
cytochrome c oxidase subunit 1	x	x	x	x
cytochrome c oxidase subunit 2	x	x	x	x
cytochrome c oxidase subunit 3	x	x	x	x
elongation factor Tu	NE			NE
heme lyase	NE			NE
ribosomal protein L1	NE			NE
ribosomal protein L10	nf	Orf17	Orf166	nf
ribosomal protein L11	NE			NE
ribosomal protein L14	x	x	x	x
ribosomal protein L16	x	x	x	x
ribosomal protein L18	nf			nf
ribosomal protein L19	NE			nf
ribosomal protein L2	NE			NE
ribosomal protein L20	NE			NE
ribosomal protein L27	NE			NE
ribosomal protein L31	nf	Orf18		nf
ribosomal protein L32	NE			nf
ribosomal protein L34	NE			NE
ribosomal protein L5	x	x	x	x
ribosomal protein L6	x	x	x	x
ribosomal protein S1	x	Orf15	Orf207	nf
ribosomal protein S10	nf			nf

Reclinomonas	*G. theta*	*Hemiselmis**	*Rhodomonas**	*B. natans*
ribosomal protein S11	x	x	x	x
ribosomal protein S12	x	x	x	x
ribosomal protein S13	x	x	x	nf
ribosomal protein S14	x	x	x	x
ribosomal protein S19	x	x	x	NE
ribosomal protein S2	x	x	x	nf
ribosomal protein S3	x	x	x	x
ribosomal protein S4	x	x	x	x
ribosomal protein S7	x	x	x	x
ribosomal protein S8	x	x	x	NE
succinate:ubiquinone oxidoreductase subunit 2	NE			NE
succinate:ubiquinone oxidoreductase subunit 3	x	x	x	NE
succinate:ubiquinone oxidoreductase subunit 4	x	x	x	NE
transcription initiation factor sigma	nf			nf
	nf	Orf10		nf
	nf	Orf33		nf
Group II intron reverse transcriptase	nf		ORF621 ORF762	nf
	nf		ORF172	nf
Reclinomonas	*G. theta*	*Hemiselmis**	*Rhodomonas**	*B. natans*
	nf		ORF72	nf

*Since the nuclear genomes of *Rhodomonas* and *Hemiselmis* are not available it is not possible to indicate if the missing gene is nucleus encoded.

3.3.8 Functional Classification of Proteomes

Protein sets from the *B. natans* and *G. theta* nuclear genomes were analyzed with three mitochondrial targeting predictors. Only those proteins that had as their top prediction mitochondrial targeting for all three programs were retained for further analysis. While this strategy of requiring consensus undoubtedly removed legitimate proteins that are part of the mitochondrial proteome, accepting all those with mitochondrial predictions would have resulted in an inflated and inaccurate list (Table 3.20). TargetP 1.1 (Emanuelsson, Brunak et al. 2007), one of most widely used programs, predicted 3,635 mitochondrial targeted proteins for *G. theta* and 3,707 for *B. natans*. These predictions are very likely gross overestimates. The other two predictors in isolation also generated unrealistic numbers. Inflated numbers for predictions are not unusual for mitochondrial proteomic surveys. For the *Arabidopsis thaliana* study TargetP generated 3,182 mitochondrial hits while 4,975 and 2,417 were generated for IPSORT and Predotar, respectively (Heazlewood, Tonti-Filippini et al. 2004).

Table 3.20. Proteins with mitochondrial targeting predicted by three subcellular localization predictors.

	G. theta	*B. natans*
TargetP	3635	3707
Predotar	1678	2118
IPSORT	3723	4142
3 way overlap	785	612

The overlap between the three prediction programs generated a list of 785 putatively mitochondrial targeted proteins for *G. theta* and 612 for *B. natans*. These lists were reduced further by removal of proteins predicted to be targeted to the other subcellular compartments investigated as part of the genome projects –the PPC, plastid, and ER/Golgi (Figure 2.1). While it is possible that some of the removed proteins are dual targeted, particularly if they were plastid targeted, in the absence of experimental results like GFP tagging, I felt it was prudent to remove them from consideration. Entries were also removed during the course of manual annotation as it became clear that they were unlikely to be mitochondrial targeted based on function and/or incorrect gene models creating spurious N-terminal target peptides. Proteins/genes were also added to the list during manual annotation as outlined above. For example, most of the mitochondrial carrier proteins did not have classical targeting presequences but nevertheless are part of the mitochondrial proteome. Ultimately I was left with a list of 833 putatively mitochondrial targeted proteins for *G. theta* and 720 for *B. natans*.

The number of putative mitochondrial targeted proteins for *G. theta* and *B. natans* seems reasonable when compared with other proteomic studies that have been done. The first survey for *Arabidopsis thaliana* found 416 mitochondrial proteins (Heazlewood, Tonti-Filippini et al. 2004), 496 for *Chlamydomonas reinhardtii* (Atteia, Adrait et al. 2009), 615 for human (Taylor, Fahy et al. 2003), 573 for *Tetrahymena thermophila* (Smith, Gawryluk et al. 2007) and 751 for yeast (Sickmann, Reinders et al. 2003). In subsequent years further studies, using different techniques, have added proteins to these lists. The human mitochondrial proteome is now at 1344 proteins while for yeast the number of identified proteins targeted to the mitochondrion has been increased to 851 (Reinders, Zahedi et al. 2006).

The list of proteins putatively targeted to the mitochondria of *B. natans* and *G. theta* provided herein should not be construed as definitive. Based on other mitochondrial proteomes it is likely that several hundred additional proteins should be added to the lists. Moreover, undoubtedly the lists very likely contain false positives due to incorrect gene models and poor or inadequate curation by me, as well as errors in other mitochondrial proteomic analyses. Without experimental evidence for subcellular localization these proteins that I have identified should be considered putative.

Table 3.21. KOG classifications and categories for putative mitochondrial targeted proteins in *G. theta* and *B. natans*. The largest number of proteins for each category between *B. natans* and *G. theta* is highlighted in grey.

	G. theta	*B. natans*
CELLULAR PROCESSES AND SIGNALING		
Cytoskeleton	21	12
Posttranslational modification, protein turnover, chaperones	38	45
Signal transduction mechanisms	25	14
Nuclear structure	3	4
Intracellular trafficking, secretion, and vesicular transport	23	16
Defense mechanisms	6	4
Cell wall/membrane/envelope biogenesis	11	6
INFORMATION STORAGE AND PROCESSING		
Chromatin structure and dynamics	9	2
Replication, recombination and repair	5	11
RNA processing and modification	6	19
Transcription	8	5
Translation, ribosomal structure and biogenesis	29	33
METABOLISM		
Amino acid transport and metabolism	28	55
Carbohydrate transport and metabolism	9	15
Cell cycle control, cell division, chromosome partitioning	8	16
Coenzyme transport and metabolism	7	13
Energy production and conversion	109	127
Inorganic ion transport and metabolism	20	8
Lipid transport and metabolism	13	17
Nucleotide transport and metabolism	3	14
Secondary metabolites biosynthesis, transport and catabolism	3	6
POORLY CHARACTERIZED		
Function unknown	35	34
General function prediction only	56	60
TOTAL	475	536

Of the 833 putatively mitochondrial targeted proteins in *G. theta* 475 could be assigned KOG categories according to the JGI annotations, while in *B. natans* 536 out of 720 were assigned KOG categories. In terms of raw numbers of proteins, differences can be discerned between the two proteomes. Of the nine categories under the classification Metabolism eight of them have more *B. natans* proteins assigned to them compared to *G. theta* (Table 3.21). The only category in Metabolism that has more *G. theta* proteins is inorganic ion transport and metabolism. *B. natans* has a lot more proteins assigned to the category amino acid transport and metabolism (55 vs. 28) (Table 3.21). *B. natans* also has more energy production and conversion proteins (127 vs. 109). It should be noted that mitochondrial carrier proteins (MCPs) fall into the energy production and conversion category and *B. natans* had 10 more MCPs than *G. theta*. *G. theta* appears to have more proteins assigned to the classification cellular processing and signaling, particularly to the categories signal transduction (25 vs. 14) and intracellular trafficking, secretion, and vesicular transport (23 vs. 16) (Table 3.21).

In terms of percentages, energy production and conversion proteins are about equal in *G. theta* and *B. natans* with 25% in *B. natans* (Figure 3.4) and 24% in *G. theta* (Figure 3.5). The proteomes also have roughly equal percentages for translation, ribosomal structure and biogenesis (6%), post translational modification, protein turnover and chaperones (8%), as well as function unknown (6-7%) and general function (11-12%). *B. natans* has a higher percentage of amino acid, transport and metabolism proteins (10% vs. 6%) while *G. theta* has higher percentages of intracellular trafficking, secretion, and vesicular transport (5% vs. 3%) and signal transduction mechanisms (5% vs. 2.5% (not shown)).

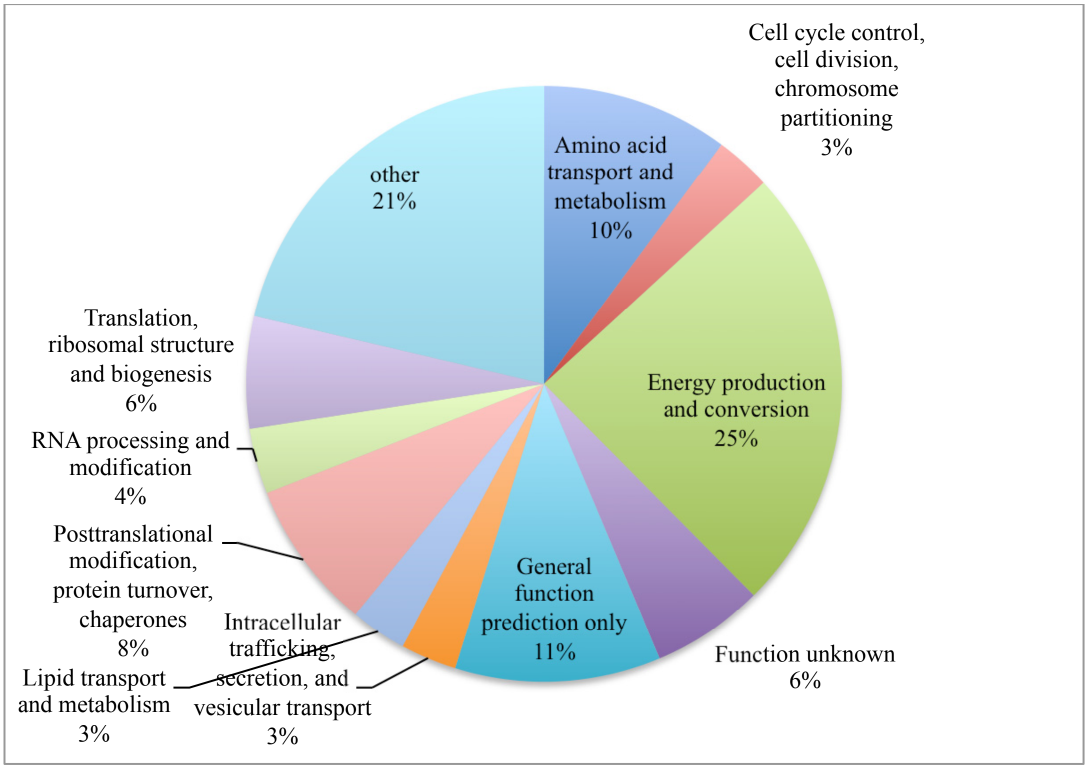


Figure 3.4. Top ten KOG categories for mitochondrial targeted proteins in *B. natans*.

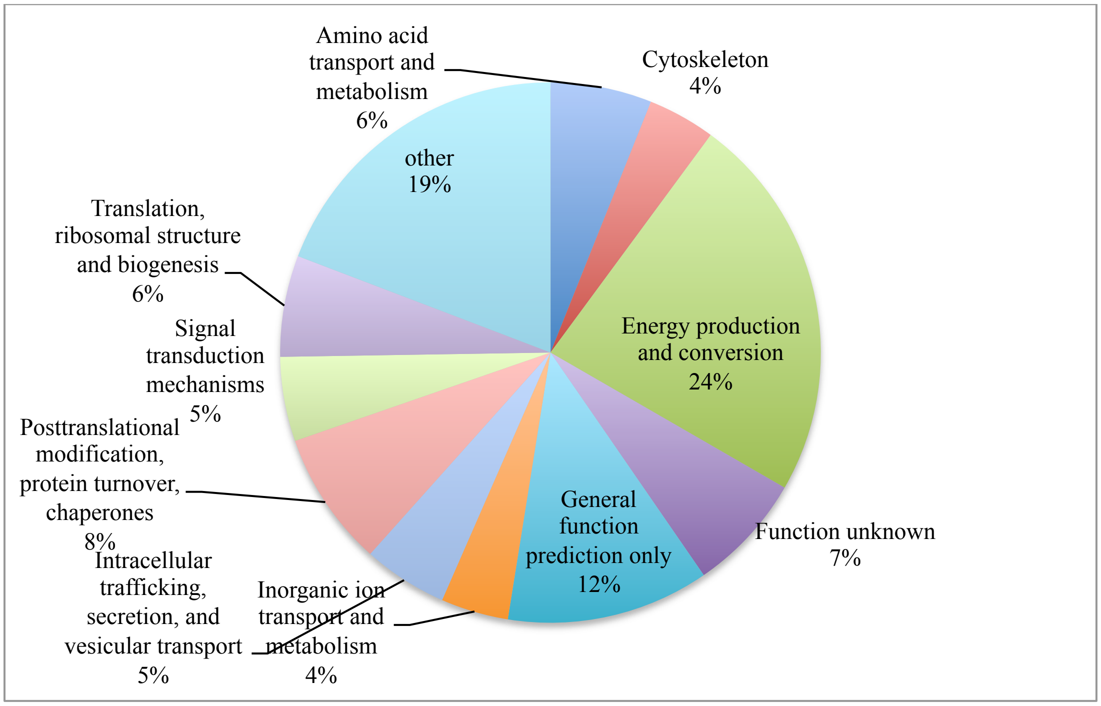


Figure 3.5. Top ten KOG categories for mitochondrial targeted proteins in *G. theta*.

Functional categorizations of mitochondrial proteomes in other organisms have found that energy generation contributes the most proteins. In *Arabidopsis thaliana* ~22% of the proteins were assigned to an energy category (Heazlewood, Tonti-Filippini et al. 2004) while in *Tetrahymena thermophila* 20% where energy related proteins (Smith, Gawryluk et al. 2007). The percentage of general function and function unknown (17% *B. natans*, 18% *G. theta*) is also comparable to other proteomic studies with 17.3% in *Arabidopsis* (Heazlewood, Tonti-Filippini et al. 2004), 18.6% in the human proteome (Taylor, Fahy et al. 2003), 17.3% in *Chlamydomonas* (Atteia, Adrait et al. 2009) and 18.3% in yeast (Sickmann, Reinders et al. 2003). In general, the *B. natans* and *G. theta* mitochondrial proteomes appear to be fairly typical with a predominance of proteins devoted to energy metabolism and translation/post translational modification.

3.3.9 BLAST Profiles

Along with determining KOG classifications for putative mitochondrial targeted proteins I also analyzed their taxonomic affiliations using blastp searches against a local database (Appendix B Local database entries) and calculating the number of top hits that could be assigned to major taxonomic groups. As a comparison, to see if the taxonomic profiles would change, I did a similar blastp search using the protein lists generated for the other 3 proteomes investigated – PPC, plastid, ER.

Of the 833 mitochondrial targeted proteins for *G. theta*, 474 had blastp hits at an e value cutoff of 1e-10 while *B. natans* had 684 hits from 720 protein sequences (Table 3.22). Nearly all of the putative ER proteins had blastp hits unlike for the PPC proteomes for which about 40% of the proteins had blastp hits. For the plastid proteomes 67% of the predicted proteins for *G. theta* had blastp hits while in *B. natans* only 57% had hits. The differences between the proteomes in terms of percentages with blastp hits is to some extent a reflection of the methods used to predict them. For example, putative ER proteins were predicted based on their homology with known ER proteins. The difference in percentage of mitochondrial proteins with blastp hits can be attributed to *B. natans* having better PASA gene models than *G. theta* in the 5′ area. This allowed me to make decisions based on presequence targeting information in more cases.

Table 3.22. Proteome entries with blastp hits (e value cutoff 1e-10) for *G. theta.*

	Mito	ER	PPC	Plastid
Total genes	833	689	2461	774
Genes with hits	474 (56%)	644 (93%)	954 (38%)	522 (67%)

Table 3.23. Proteome entries with blastp hits (e value cutoff 1e-10) for *B. natans.*

	Mito	ER	PPC	Plastid
Total genes	720	597	1012	723
Genes with hits	684 (95%)	560 (93%)	440 (43%)	416 (57%)

If EGT occurred between the algal endosymbiont and the host one would expect that the blastp top hit taxonomic profiles of the proteins targeted to the different compartments would be different. One would expect the plastid proteome to reflect its algal origins and lean towards hits against primary plastid lineages. One would also expect that many proteins would have top hits to the major eukaryotic group that they are from. These assumptions are borne out for the most part by examination of the taxonomic profiles of the blastp top hits (Figure 3.6, Figure 3.7).

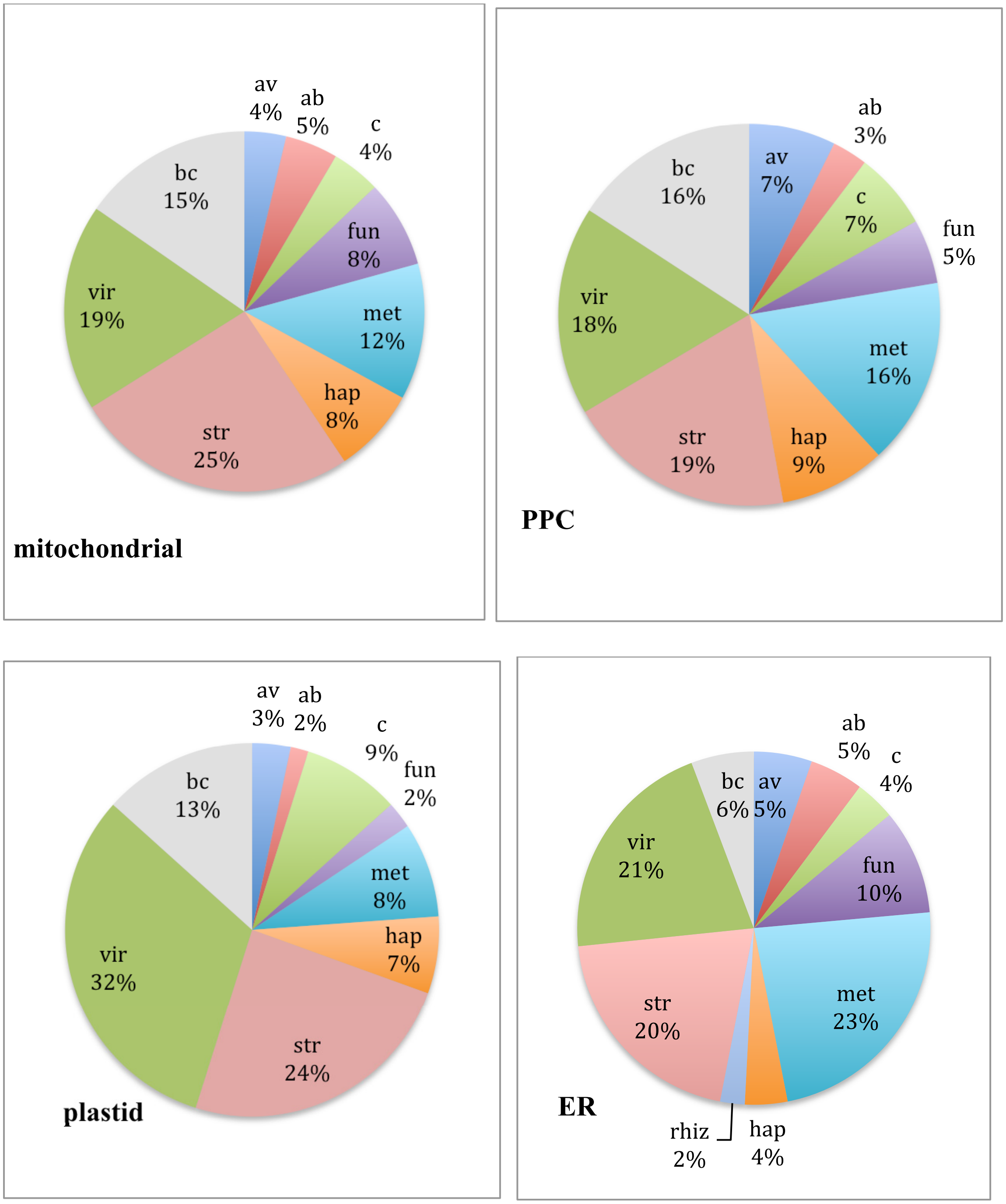


Figure 3.6. Taxonomic profiles of top blastp (1e-10) hits for *B. natans* proteomes. Av=alveolates, ab=Amoebozoa, c=cryptophytes, fun=fungi, met= metazoans, hap= haptophytes, str=stramenopiles, vir= Viridiplantae, bc=bacteria.

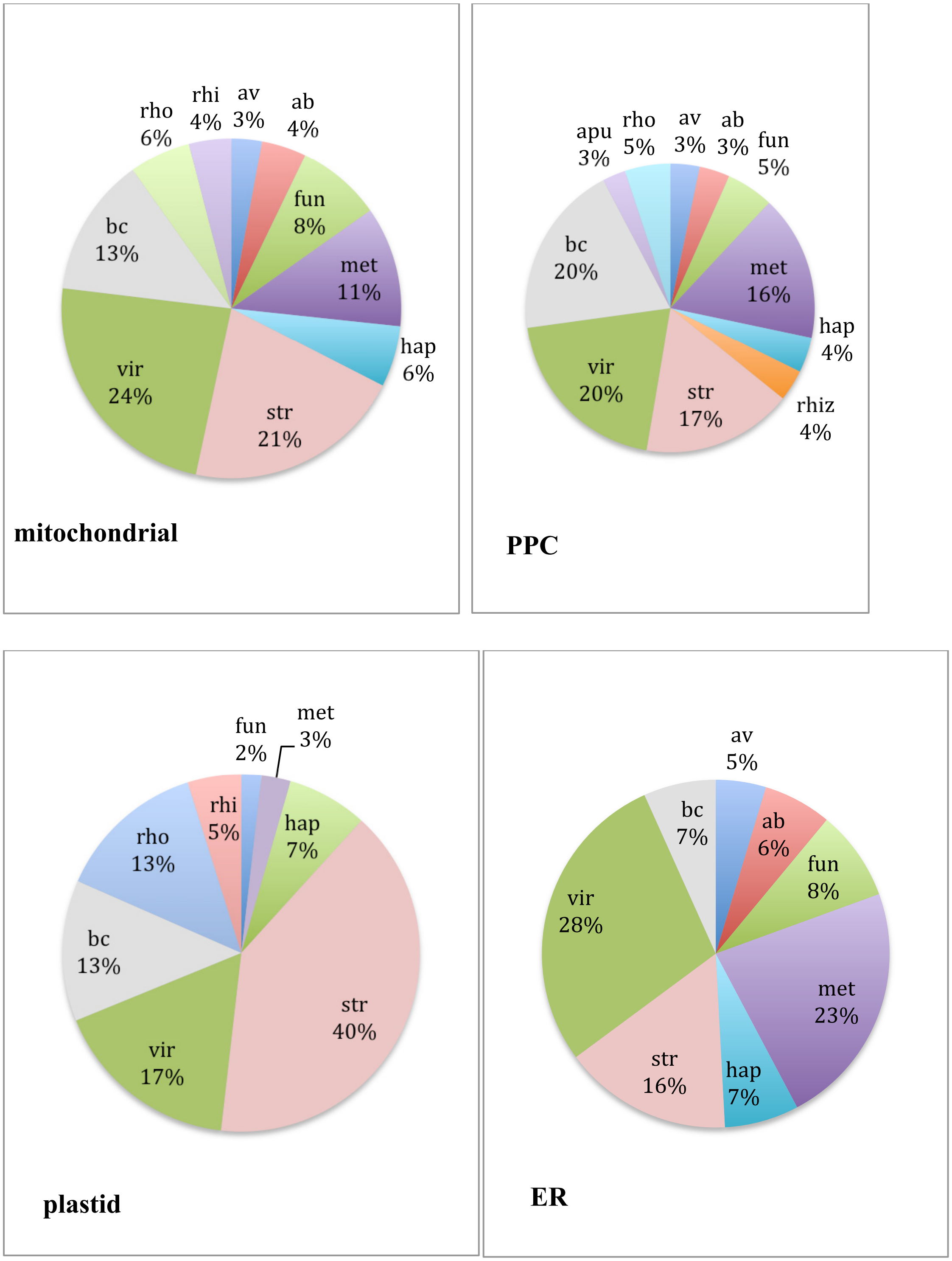


Figure 3.7. Taxonomic profiles of top blastp (1e-10) hits for *G. theta* proteomes. Av=alveolates, ab=Amoebozoa, c=cryptophytes, fun=fungi, met= metazoans, hap= haptophytes, str=stramenopiles, vir= Viridiplantae, bc=bacteria, rhiz=rhizarians, rho=rhodophytes, apu=apusomonads.

In *G. theta* 13% of the top hits are red algal proteins as would be expected for an organism with a red algal plastid (Figure 3.7). Stramenopile hits make up 40% of the plastid hits. This is a large increase in the percentage of top stramenopile hits compared to that observed for the other three proteomes. This increase in stramenopile hits for the plastid proteome of *G. theta* is understandable given that stramenopiles and cryptophytes share the same red algal plastid regardless of whether stramenopiles acquired this plastid in the same secondary endosymbiotic event as cryptophytes or individual lineages acquired a plastid via tertiary endosymbiosis from a cryptophyte. Also of note is the increase in the top hits to metazoan proteins in the ER proteome compared to the other proteomes. The reason for this increase is unclear but may reflect the composition of the ER database that was used in part to predict the ER proteome. The same increase is seen in the *B. natans* ER proteome taxonomic profile. For the other lineages the percentages tend to remain the same regardless of the proteome, reflecting a steady state of background BLAST "noise." Only when there are differences in the percentages between proteomes and/or between the same proteome for different organism are the profiles indicating something of interest.

In *B. natans*, as would be expected with its green algal plastid, the percentage of top hits to Viridiplantae sequences is higher for the plastid proteome (32%) compared to the ~20% seen in the other proteomes (Figure 3.6). The other observation of note is the percentage of top hits against cryptophyte proteins. For the mitochondrial and ER proteomes the percentage of top hits that are from cryptophytes is 4%. This increases to 7% for the PPC and 9% for the plastid proteome. The differences could be simply

BLAST "noise" and one hesitates to over analyze but there have been suggestions of an ancestral red algal plastid in rhizarians that was lost (Dorrell and Smith 2011). The increase in the plastid and PPC cryptophyte signal could be a reflection of red algal proteins that were acquired via EGT, retained, and subsequently used with the newly acquired green algal endosymbiont in chlorarachniophytes.

Comparisons between the two organisms are generally what one would expect. For the plastid proteome *B. natans* proteins have more top hits to Viridiplantae sequences than does *G. theta*. *B. natans* also has slightly higher percentages for top hits against stramenopiles except for the plastid proteome. This probably reflects the current thinking that rhizarians are allied with stramenopiles while the relationship between cryptophytes and stramenopiles is perhaps restricted to the proteins derived via EGT from the red algal endosymbiont (Sanchez-Puerta and Delwiche 2008; Bodyl, Stiller et al. 2009).

3.4 Summary

As part of the genome projects for *B. natans* and *G. theta* I investigated their mitochondrial proteomes. Overall, these two organisms have typical, unremarkable mitochondria and possess the usual pathways and functions that are generally associated with this subcellular compartment. The analyses were bioinformatic in nature, relying on the identification of mitochondrial targeting presequences, homology searches using known mitochondrial protein sequences, and parsing of automated annotation results. A combination of techniques was necessary since some types of analysis are not appropriate for various classes of proteins and required specialized and unique bioinformatics investigations.

One such class was mitochondrial carrier proteins. Typically these proteins do not possess targeting presequences and in *G. theta* and *B. natans* were initially found by searching for a characteristic domain. Further analysis was required to eliminate false positives and ultimately 49 mitochondrial carrier protein genes were identified in *G. theta* and 59 in *B. natans*. These results are in the range displayed by other lineages. The possible unexpected bacterial origin of this class of genes was also discussed and while no conclusion was reached I did identify mitochondrial carrier protein-like genes in bacteria.

Another class of mitochondrial targeted protein that required an in-depth analysis of its characteristics to identify its constituents was twin CX9C proteins. I found 10 genes that code for this type of protein in *B. natans* and 6 in *G. theta*. While the number of genes found was low compared to other lineages the results were not abnormal and may reflect bias in taxon sampling combined with the difficulties of identification.

Other typical and well known mitochondrial systems were investigated through a combination of literature and homology searches. For example, all components of the ubiquitous iron sulfur cluster formation machinery were found in both organisms. Genes for all of the known import complexes were also found for both organisms but differences were seen in which subunits were present when comparing *B. natans* to *G.*

theta as well as to other major lineages. In general, *G. theta* appears to be richer in protein import machinery subunits compared to *B. natans* and seems similar in its subunit complexity to green plants and green algae while *B. natans* seems most similar to some stramenopiles and red algae in the makeup of its import complexes. Another typical mitochondrial function is aerobic respiration or oxidative phosphorylation. I searched for homologous genes for the five complexes involved in respiration. Both *G. theta* and *B. natans* had standard subunit complements with no surprises.

In the course of the nuclear genome projects the mitochondrial genome was also sequenced for both *B. natans* and *G. theta*. The sets of protein coding genes of the mitochondrial genomes were compared to that seen in other organisms particularly the excavate *Reclinomonas americana*. *G. theta* has 39 protein coding genes, somewhat below the number for other cryptomonad mitochondrial genomes but within the range typically seen for plants. *B. natans* has transferred more of its mitochondrial genome to its nucleus and only 27 protein coding genes were found, placing it below most stramenopiles but comparable to red algal Neither *B. natans* or *G. theta* had any unusual mitochondrial encoded genes.

The nucleus encoded mitochondrial targeted proteins for *G. theta* and *B. natans* were classified according to the KOG functional categories. As expected, more *B. natans* and *G. theta* genes belonged to the energy production and conversion category than to any other category. The number of genes for each category was fairly typical. A comparison between *B. natans* and *G. theta* revealed that *B. natans* has a mitochondrial proteome more weighted towards amino acid transport and metabolism than does *G. theta* while *G. theta* is richer in genes devoted to signal transduction mechanisms. In other KOG categories they were virtually identical.

The final analysis was a taxonomic profiling of the top BLAST hits for the nucleus encoded mitochondrial proteins as well as for the other three proteomes investigated for the genome projects: plastid; ER/Golgi; and nucleomorph/PPC. The results were generally what one would expect. Plastid targeted genes tended to have hits against other photosynthetic lineages and *B. natans* had more hits to green plant and green algal lineages than did *G. theta*. The mitochondrial proteomes of *G. theta* and *B. natans* had very similar taxonomic profiles although *B. natans* had a slightly higher percentage of top hits to stramenopile lineages than did *G. theta* while the reverse was seen for hits to "green" lineages. The overall profiles tend to support the hypothesis that rhizarians, like *B. natans*, are more closely related to stramenopiles and alveolates than is *G. theta* whose position on the eukaryotic tree continues to be problematic.

CHAPTER 4 DETECTION OF ORGANELLE-TO-NUCLEUS GENE TRANSFER

This chapter includes work published in Curtis BA, Archibald JM. A spliceosomal intron of mitochondrial DNA origin. Curr Biol. 2010 Nov 9;20(21):R919-20.

4.1 Introduction

The first indications that organellar DNA was being transferred and integrated into nuclear genomes came from hybridization studies of mouse mitochondrial DNA and nuclear DNA (du Buy and Riley 1967). However, it was not until the genomics era that we had the ability to study the extent of such transfers in diverse lineages. We now know that nuclear mitochondrial DNA (NUMTs) are a common feature of the nuclear genomic landscape, being found in virtually all sequenced eukaryotes (Huang, Ayliffe et al. 2004; Hazkani-Covo, Zeller et al. 2010). In photosynthetic organisms the plastid acts as a second source of organellar DNA that can be transferred to the host nuclear genome (Cullis, Vorster et al. 2008).

The vast majority of nuclear plastid DNA (NUPTs) and NUMTs appear to be small pieces, although some large stretches that include entire genes have been detected, particularly in plants (Noutsos, Richly et al. 2005; Liu, Zhuang et al. 2009). There is no preference for coding or noncoding regions of the organellar DNA to be transferred, nor any particular gene (Leister 2005). This randomness bolsters the view that NUMTs and NUPTs are not derived from back transcribed mRNA but pieces of DNA from degraded organellar chromosomes (Henze and Martin 2001).

There appear to be hotspots of integration in the host nucleus in that sometimes pieces from different regions of the organellar genome are found next or close to each other in the nuclear genome (Richly and Leister 2004; Richly and Leister 2004). A recent study (Lloyd and Timmis 2011) using tobacco was able to sequence a nuclear region before and after multiple insertions of plastid DNA fragments, demonstrating that the pieces are indeed discrete pieces of DNA rather than a single transferred piece that has experienced deletions post integration. The pieces also appear to be integrated in the same event instead of being inserted at different times.

Most of the integration sites are in noncoding regions and appear to have little impact on the nuclear genome (Timmis, Ayliffe et al. 2004). However NUMTs and NUPTs do have the capacity to alter the nuclear genome in a meaningful way (Noutsos, Kleine et al. 2007). Some of the transferred pieces have been found in introns while others appear to have created new introns. Several human diseases have been linked to mutants caused by the insertion of mitochondrial DNA (Noutsos, Kleine et al. 2007; He, Tao et al. 2010).

Perhaps most significantly, entire organellar genes have been integrated into the host nuclear genome and subsequently expressed (Huang, Ayliffe et al. 2004; Liu, Zhuang et al. 2009). This transfer of entire genes presumably mirrors the early stages of endosymbiosis and the transformation of the endosymbiont to an organelle with reduced functions and a reduced DNA complement. Researchers believe that the transformation happened relatively quickly and involved large and repeated transfers of DNA from the endosymbiont as well as tremendous loss, as the endosymbiont dispensed with genes for functions that were no longer necessary (Martin 2003). However, the presence of large NUMTs and NUPTs spanning entire genes demonstrates that organisms retain the capacity to reduce their organellar genomes and transfer functions to the host.

As mentioned, it is no longer thought that NUPTs and NUMTs are derived from mRNA but are bits of DNA from degraded organellar chromosomes. The availability of mitochondrial DNA has been linked to mitophagy wherein damaged or abnormal mitochondria are broken down by vacuoles leading to free floating bits of DNA (Abeliovich 2007). Others have postulated direct physical links via fusions between the nuclear membrane and mitochondrial membrane that would facilitate the transfer of DNA (Mota 1963). Genomic fragments in the cytoplasm may result if a mitochondrion is lyzed by vacuoles or lysosome, or experiences transient breaks in its membrane integrity such as when morphological changes like budding or fusing occur (Brennicke, Grohmann et al. 1993; Thorsness and Weber 1996; Berg and Kurland 2000). Cellular stress also appears to disrupt organellar membranes (Cullis, Vorster et al. 2008) permitting the release of nucleic acids into the cytoplasm.

While most of the nucleus encoded mitochondrial derived (NUMT) portions as well as the nucleus encoded plastid derived (NUPT) pieces appear to be small fragments with no functional significance, the successful transfer and expression of a gene from the chloroplast to the nucleus in “real” time has been demonstrated (Huang, Ayliffe et al. 2003). In this elegant study it was demonstrated that the transfer and stable inheritance of a plastid gene in the nucleus had occurred. A screen of kanamycin-resistant seedlings found 16 out 250,000 progeny had transferred the gene and more importantly expressed it. Given the randomness of the transfer process in terms of pieces and size, the actual number of nonfunctional integrations would be substantially higher. Studies in maize (Lough, Roark et al. 2008) and rice (Matsuo, Ito et al. 2005) also concluded that NUMTs and NUPTs were being created continuously in their respective nuclear genomes. Nor is the high incidence of organellar DNA transfer limited to plants. Similar rates of NUMTs have been seen in rodents (Triant and DeWoody 2007), honeybees (Pamilo, Viljakainen et al. 2007) and fungi (Sacerdot, Casaregola et al. 2008). Such results suggest that NUMTs and NUPTs could have a far greater role in shaping the nucleus than previously thought, either by the introduction of new genes or by the interruption of existing ones.

One of the stumbling blocks for a successful gene transfer from the organelles is that the transferred gene needs to acquire suitable regulatory elements and, if its encoded protein is to function in the organelle from whence it came, targeting information. For mitochondrial proteins the targeting signal is generally a presequence as is the case for plastid target signals in primary photosynthetic organisms. However, in lineages with secondarily acquired plastids a bipartite presequence is required, the first part being a signal peptide to direct the protein to the secretory system and then a target sequence to retain the protein in the PPC or send it on to the plastid (McFadden 1999; Gould, Sommer et al. 2006). While acquisition of the requisite targeting signals seems onerous, various studies have uncovered numerous methods for transferred genes to overcome this hurdle. Among the strategies to gain functionality is to co-opt existing organelle targeted genes. The mitochondrial *rps10* gene in carrot was inserted into a nuclear copy of the mitochondrial targeted *hsp22* gene and utilizes the presequence already there to target its protein to the mitochondrion (Adams, Daley et al. 2000). Exon shuffling type processes can also generate the necessary elements for expression and targeting (Nugent and Palmer 1991; Daley, Adams et al. 2002). Finally, some nucleus encoded proteins destined for the mitochondrion rely on internal targeting signals rather than presequences for entry into the organelle (Adams, Daley et al. 2000).

Several hypotheses have been proposed for why organellar genomes lose material to the host nucleus. Perhaps the most prominent and widely discussed hypothesis is Muller's ratchet which, in the case of the mitochondrion, suggests that the accumulation of deleterious mutations in the mitochondrial genome is reduced when a gene moves to the nucleus (Adams and Palmer 2003). This may very well be the case for animal mitochondria that experience much higher rates of nucleotide substitution than the nucleus. In plants however, with significantly larger mitochondrial genomes than animals, the nucleotide substitution rates for mitochondrial genomes are lower than for the nuclear DNA and they appear to have highly efficient mechanisms for minimizing mutations (Wolfe, Li et al. 1987).

Another proposal is that by relocating genes to the nucleus beneficial mutations can be retained unlike in the mitochondrion where they would be suppressed. Once a 'better copy' is functional in the host nucleus the original in the organelle can be lost (Blanchard and Lynch 2000). Other proposals hinge on the efficiency and cost of maintaining organellar genomes. Smaller genomes would presumably be at an advantage in intraorganellar competition and tend to be favoured, resulting in genome streamlining (Selosse, Albert et al. 2001). Again however, plants have relatively bloated mitochondria compared to animals and the propensity with which plant organelles, particularly plastids, pick up exogenous DNA would seem to argue against streamlining as a pervasive and vital process, at least in plants. Another popular and widely cited hypothesis is that the production of toxic free radicals during normal organellar processes puts selective pressure on the genomes to move genes to the less toxic environment of the host nucleus (Allen and Raven 1996).

The transfer of endosymbiont DNA to the host genome played a central role in the successful establishment of both mitochondria and chloroplasts. These transfers however, continue to this day. They have the ability to alter the genomic landscape of both the host and organelle and by studying aspects of the current transfer process such as rate, mechanisms and fate of the integrated pieces, we can learn much about the ancient process that generated the organelles. With the sequencing of the host genomes for *Bigelowiella natans* and *Guillardia theta* I was able to investigate NUMTs and NUPTs for the first time from two major eukaryotic lineages. Because these organisms also contain nucleomorphs this was also the first opportunity to look for nuclear nucleomorph DNA (NUNMs) and glimpse the process of transfer from the captured eukaryotic nucleus to the host nucleus, an operation that in all other lineages has gone to completion.

4.2 Materials and Methods

Assembled scaffolds for the nuclear and mitochondrial genomes of *B. natans* and *G. theta* were downloaded in FASTA format from the JGI web portal. The plastid (chloroplast) and nucleomorph genomes were downloaded from NCBI (*B. natans* nucleomorph: NC_010004.1, NC_010005.1, NC_010006.1; *B natans* plastid: NC_008408.1) (*G. theta* nucleomorph: AF165818.4, AJ010592.2, AF083031.2; *G. theta* plastid: NC_000926.1). Local databases were created from the nuclear scaffolds using formatdb.

To detect recent organelle-to-nucleus gene transfer events, each organellar genome sequence was used as a query in a BLAST search against the corresponding local nuclear database. Other than defaults, blastn (version 2.2.17) parameters used were r=2 and e=0.001.

The sequences of potential organellar fragments, along with 500 base pairs (bp) on either side, were extracted from the nuclear scaffolds and used in blastn and blastx searches against the NCBI nucleotide collection and non-redundant protein sequence database, respectively (default parameters for version 2.2.21 were used). The results were used to determine whether putative organellar-derived fragments were located within or close to coding regions, and to rule out the possibility that such fragments in fact corresponded to highly conserved and ubiquitous genes, such as the small subunit ribosomal RNA gene and HSP70, which are typically found in organellar and nuclear genomes. Blastx and blastn searches were also used to determine whether *bona fide* organellar-derived DNA fragments in the *B. natans* nuclear genome were derived from coding or non-coding regions of the organellar genomes.

To determine whether putative organellar-derived fragments were artifactual chimeras, all relevant sequence reads were downloaded from the JGI portals and used to establish a local database using formatdb. The organellar fragment, along with 50 bp on either side, was extracted from the nuclear scaffolds and using blastn searched against the local database of sequence reads. Organellar fragments with at least two independent reads containing identical arrangements of organellar/nuclear or nuclear/organellar sequence were considered real organellar fragments integrated into the *B. natans* and/or *G. theta* nuclear genomes.

4.2.1 Cell Culture and RNA Extractions

B. natans CCMP2755 was grown in 2L of f/2-Si medium made with artificial seawater for 28 days. RNA was extracted as follows. Pelleted cells were re-suspended in a 10X volume of a Tris-borate buffer [150 mM Tris-base, 50 mM sodium tetraborate, 50 mM EDTA, SDS to 2%, and beta-mercaptoethanol to 1% added just before use], vortexed, heated in a 50^{o} C water bath for 10 min and passed through a French press at 8,000 psi. An equal volume of 24:1 chloroform: isoamyl alcohol was added. The slurry was vortexed and centrifuged at 10,000 rpm for 15 min. The top layer was retained and mixed well with 0.5 volumes of room temp 100% ethanol, and then centrifuged at 10,000 rpm for 5 min. The supernatant was removed and 1/9 volume of 5M potassium acetate was added and mixed. An equal volume of phenol- chloroform was added. Two rounds of phenol-chloroform extraction were performed followed by a chloroform extraction. The resulting aqueous layer was combined with 2 volumes of 100% ethanol and stored at -20^o C overnight. The tube was then centrifuged at 10,000 rpm for 5 min, the supernatant discarded and the insides of the tube washed with 3 mL of 80% ethanol. The tube was then centrifuged at 10,000 rpm for 5 min. The supernatant was again discarded and the pellet dried in a vacuum desiccator for 10 min. The pellet was dissolved in 200 ul of water. 1/10 volume of 10M LiCl was added, mixed well and stored at 4^o C overnight. The tube was then centrifuged at 12,000 rpm for 15 min, with the resulting pellet washed with 80% ethanol and spun again. The pellet was dried in a vacuum desiccator and dissolved in 100 ul water. 1/10 volume of 3M sodium acetate was added and one round of phenol-chloroform extraction was performed. To the top layer two volumes of 100% ethanol was added and the mixture was stored at -20^o C overnight. Precipitated material was centrifuged (12,000 rpm for 15 min) and the resulting pellet washed twice with 80% ethanol, dried in a vacuum desiccator for 10 min and re-suspended in 75 ul of water.

4.2.2 Amplification using Degenerate Primers

Universal primers were designed to amplify a portion of the guanine nucleotide-binding protein gene from chlorarachniophyte species. The forward primer was based on an alignment of the region of interest from *B. natans* and the closest matching paralog from the diatom *Thalassiosira pseudonana* . The reverse primer was based on an alignment with the *B. natans* gene and the closest matching gene from the rhizarian *Reticulomyxa filosa*. The forward primer (956F CTMCTMGGAGCTGGAGARTC) had 8 fold degeneracy while the reverse had 16 fold degeneracy (956R

CKTTGKCCACCRACATCRAATA) and would generate from *B. natans* a PCR product of ~900 bps.

The primers were used with the following genomic DNA: CCMP623; CCMP1259; CCMP1481 (close relative of *Bigelowiella longifila* U03479); CCMP2314 (close relative *of Lotharella globosa* (AF076169); CCMP2755 (single isolate strain from *B. natans* CCMP621). A step-down PCR protocol was used using the following parameters: 95^o C for 5 min; 14 cycles of 30 sec at 94^o C, 1 min at 63^o C with a 1^o degree decrease each cycle, 90 sec at 72^o C; followed by 30 cycles of 94^o C for 30 sec, 1 min at 49^o C and 90 sec at 72^o C; ending with a final 7 min extension at 72^o C.

4.2.3 RT-PCR

Using the *B. natans* genome sequence, the following two primers were determined to be specific to the guanine nucleotide-binding protein alpha subunit gene that contained the putative intron of mitochondrial origin: Bnintron.R1 AGAGAAAATGGGCGCAGACC; Bnintron.F1 TAGAAGGCGGGCTGAATCTGT. Prior to RT-PCR, RNA was treated with Invitrogen's DNase I (Amplification Grade). RT-PCR products were amplified using 1.2 ng of RNA template, the gene-specific primers and Qiagen's Omniscript Reverse Transcription kit in a two-step process (reverse transcription followed by PCR). The final PCR stage used a step-down protocol using the following conditions: 95^o C for 5 min; 14 cycles of 30 sec at 94^o C, 1 min at 63^o C with a 1^o degree decrease each cycle, 90 sec at 72^o C; followed by 30 cycles of 94^o C for 30 sec, 1 min at 49^o C and 90 sec at 72^o C; ending with a final 7 min extension at 72^{o} C.

4.2.4 Cloning and Sequencing

RT-PCR products were gel purified using the Qiagen MinElute Gel Extraction Kit and cloned with Promega's pGEM-T Easy Vector System. 10 clones were fully sequenced with universal primers M13F, M13R on a Beckman CEQ8000. Sequences were edited and assembled using the Staden Package.

4.3 Results and Discussion

4.3.1 NUMTs

A blastn analysis (Altschul, Gish et al. 1990) of the *B. natans* mitochondrial genome against the final nuclear genome assembly identified nine possible pieces of mitochondrial DNA that had been inserted into the host nuclear DNA. To ensure that the possible NUMTs (nuclear mtDNA) were not part of the final genome assembly due to chimeric reads (mitochondrial DNA/nuclear DNA) the ends were checked against all individual reads. This resulted in two candidate NUMTs being rejected as miss-assemblies. The seven confirmed NUMTs ranged from 56 bps (bn8-2) to 306 bps (bn38-1) (Table 4.1) with an average length of 134 bps. The sequence identity ranged from a low of 80% (bn80-1) to two NUMTs that were 100% identical to their counterparts in the mitochondrial genome (bn38-1 and bn72-1 are 306 and 196 bp, respectively).

Table 4.1. Nuclear sequences of mitochondrial origin in *B. natans.*

NUMT name	Identical bases/length	Mitochondrial Genome Coordinates	Mitochondrial Gene	Nuclear Scaffold Coordinates	NUMT Insertion Site
bn8-1	133/154	13677-13828	50S rpl16	1283161-1283314	5′ hypothetical protein 3′ bn6-2
bn8-2	52/56	3016-3071	non-coding DNA	1283315-1283370	5′ bn6-1 3′ intergenic
bn38-1	306/306	1-306	non-coding DNA	509974-510279	5′ hypothetical protein 3′ CAx16 (intergenic)
bn43-1	70/86	334-417	non-coding DNA	570424-570340	intergenic
bn72-1	196/196	23664-23859	23s rDNA	188513-188708	intergenic
bn80-1	53/68	21784-21837	23s rDNA	108027-108094	intergenic
bn164-1	64/72	32873-32944	COXI	5973-6044	Guanine-binding protein, alpha subunit: protein 155542

Two of the *B. natans* NUMTs, bn8-1 and bn8-2, are located on the same scaffold and lie next to each other but are 10600 bps apart in the mitochondrial genome (Figure 4.1). The location of the pieces in the nuclear genome does not appear to be consistently linked with any particular genomic feature. Four of the seven mitochondrial DNA fragments are derived from genes: bn8-1 from the gene for 50S ribosomal protein L16; bn80-1 and bn72-1 from 23S rDNA; bn164-1 from COXI. In contrast, Bn43-1, bn72-1 and bn80-1 reside entirely within intergenic regions, while bn8-1 and bn38-1 overlap with a protein gene on one end (Table 4.1). In the case of bn8-1, the first 19 bps of the NUMT correspond to the last bases of the coding region for protein 129278, a hypothetical protein with no obvious homology to any proteins currently in the Genbank protein database and with no support from RNA-Seq data. The first 202 bp of bn38-1 are part of exon 2 in the hypothetical protein 76016 while the rest of the NUMT is part of the first intron and is bounded by a short CA repeat (16x). As with 129278, protein 76016 has no homology with any known protein and does not have RNA-Seq support. The most interesting NUMT in *B. natans* is bn164-1, a fragment of mtDNA whose recent insertion created an intron in the alpha subunit of a guanine binding protein (bn155542) (Curtis and Archibald 2010) (see below).

Analysis of mitochondrial genome-derived fragments in the nuclear genome of *G. theta* identified 13 potential NUMTs. All were deemed to be real integrations after checking individual reads for chimeras. The sizes range from 53 bp (gt27-1, gt238-2) to 221 bp (gt238-1) (Table 4.2), with an average length of 104 bps. The degree of similarity ranges from 75.6% (gt21-1) to 100% (gt269-1). Three instances were identified in which two of the *G. theta* NUMTs reside on the same genomic scaffold. As with *B. natans*, the NUMT-containing contigs within the scaffold are very close to each other (Figure 4.2) but their scaffold proximity does not match their placement in the mitochondrial genome. The NUMTs Gt21-1 and gt21-2 are one base apart on the scaffold but 7989 bases apart in the *G. theta* mitochondrial genome. As well, two other pairs [(gt238-1 and gt238-1) (gt460-1, 460-2)] are separated by 33 bp on their respective scaffolds but are 16384 bps and 2564 bps apart in the mitochondrial genome. Eight of the 13 pieces derive from mitochondrial coding regions (Table 4.2). COX II contributed two NUMTs (gt2-1, gt116-1), as did COXI (gt238-1, gt460-1). Portions of the coding regions for rps19 (gt21-2), rps12 (gt238-2), nad11(460-2) and NADH dehydrogenase subunit 5 (gt361-1) were also transferred to the host nuclear genome. All of the *G. theta* NUMTs reside entirely within intergenic regions, except for gt7-1 and gt238-1. Gt7-1 is in the third intron of the gene for protein gt101478. Unlike bn164-1, whose insertion created a new intron in the *B. natans* genome, gt7-1 represents a small portion of a 668 bp intron in a hypothetical protein gene with no RNA-Seq or homology support. Gt238-1, which is derived from an internal portion of the mitochondrial gene COXI, was assigned protein ID 48929 by the gene-modeling pipeline. Since the NUMT is far too small to encode a functional mitochondrial-targeted protein (it would encode only 73 amino acids of the 530 in the mitochondrial COXI) and has no RNA-Seq support, it is likely that protein 48929 is an artifact of the automated gene finding process.

Table 4.2. Nuclear sequences of mitochondrial origin in *G. theta.*

NUMT name	Identical bases/length	Mitochondrial Genome Coordinates	Mitochondrial Gene	Nuclear Scaffold Coordinates	NUMT Insertion Site
gt2-1	138/150	16100-16249	COXII	1290361-1290507	intergenic
gt7-1	89/90	32102-32191	non-coding DNA	899850-899939	intronic
gt21-1	87/115	18536-18642	Rps19	491486-491600	5′ gt21-2 3′ intergenic
gt21-2	59/66	10482-10547	non-coding DNA	491419-491484	5′ intergenic 3′ gt21-1
gt27-1	44/53	23562-23614	non-coding DNA	361265-361317	intergenic
gt116-1	147/166	16094-16249	COXII	205952-206117	intergenic
gt130-1	57/65	19290-43177	non-coding DNA	43113-43177	intergenic
gt238-1	184/221	1320-1540	COXI	3796-4016	5′ gt238-2 3′ intergenic
gt238-2	51/53	17925-17977	Rps12	3710-3762	5′ intergenic 3′ gt238-1
gt269-1	59/59	9166-9224	non-coding DNA	52855-52913	intergenic
gt361-1	114/116	33254-33369	NADH dehydrogenase subunit 5	3641-3756	intergenic
gt460-1	100/110	1087-1196	COXI	6169-6278	5′ gt460-2 3′ intergenic
gt460-2	83/91	3761-3851	Nad11	6045-6135	5′ intergenic 3′ gt460-1

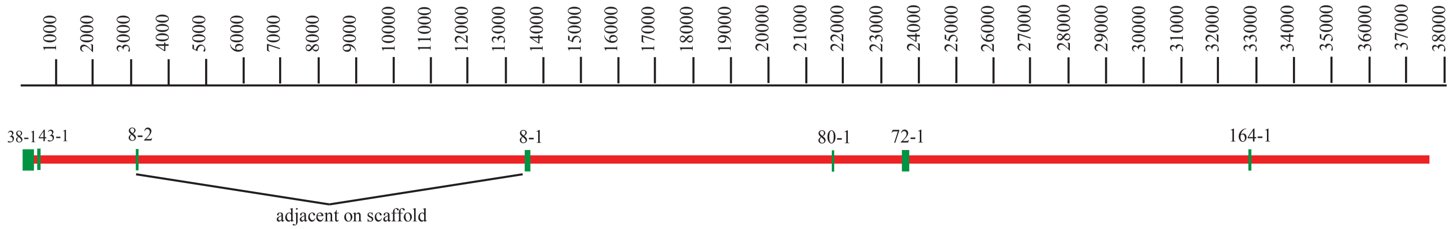


Figure 4.1. Mitochondrial genome position of *Bigelowiella natans* NUMTs. The thick red line represents a linear mitochondrial molecule with an arbitrary start point. The NUMTs are represented as green bars and are proportional to the NUMT size. NUMT labels above the bars indicate the nuclear scaffold where the NUMT is found. Angle brackets are used to indicate when nuclear scaffolds contain more than one NUMT and the distance between the NUMTs on the scaffold.

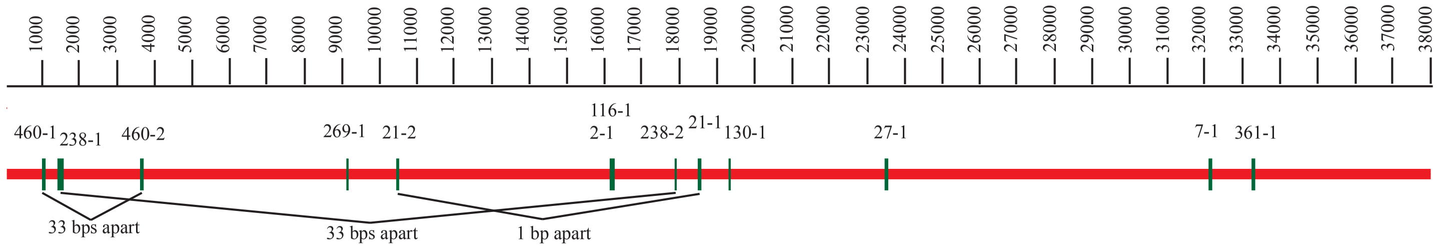


Figure 4.2. Mitochondrial genome position of *Guillardia theta* NUMTs. The thick red line represents a linear mitochondrial molecule with an arbitrary start point. The NUMTs are represented as green bars and are proportional to the NUMT size. NUMT labels above the bars indicate the nuclear scaffold where the NUMT is found. Angle brackets are used to indicate when nuclear scaffolds contain more than one NUMT and the distance between the NUMTs on the scaffold.

4.3.2 NUPTs and NUNMs

Possible candidates of transfer from the plastid genome (NUPTs) and nucleomorph genome (NUNMs) to the host nuclear genomes of *G. theta* and *B. natans* were identified from blastn analysis. All candidate transfers were rejected as representing miss-assemblies after examination of the individual reads.

4.3.3 Comparison and Context

To put the number of organellar DNA pieces that have been transferred to the host genomes of *B. natans* and *G. theta* in context I evaluated the presence of NUMTs and NUPTs in all the single celled photosynthetic protist species for which nuclear and organellar genomes were available (Table 4.3). I also included some multicellular photosynthetic organisms such as *Volvox carteri, Ectocarpus siliculosus* and *Physcomitrella patens* as well as a few nonphotosynthetic stramenopiles (*Phytophthora* sps).

A third of the photosynthetic unicellular algae have no NUPTs while six of 11 have no NUMTs. If they do have transferred pieces most of the photosynthetic unicellular algae have more NUPTs than NUMTs. This is in contrast to both *B. natans* and *G. theta* that have some NUMTs but no NUPTs. Only assembly four of *Chlamydomonas reinhardtii* has more NUMTs than NUPTs (51 vs. 48) among unicellular algae. The number of transferred pieces also tends to be low for the photosynthetic unicellular algae, which reflects what was observed for *B. natans* and *G. theta*. The highest number of transferred pieces is 51 NUMTs for *Chlamydomonas*. In contrast, the multicellular green photosynthetic organisms *Volvox* and *Physcomitrella* have hundreds of NUMTs and NUPTs. These higher numbers of transferred and retained pieces mirror what is seen in the literature for green plants (Hazkani-Covo, Zeller et al. 2010; Smith, Crosby et al. 2011), where for example *Oryza sativa* has been reported to have in excess of 2000 NUPTs and 1985 NUMTs. The only protists that have numbers of transferred pieces approaching that seen in green plants are two nonphotosynthetic stramenopiles, *Phytophthora infestans* with 552 NUMTs and *Phytophthora sojae* with 163 NUMTs. The only non-green multicellular photosynthetic alga in my database, *Ectocarpus siliculosus*, had more NUMTs than NUPTs (66 vs. 28). These are low numbers compared to green multicellular organisms that tend to have hundreds of NUPTs and NUMTs (Table 4.3) (Smith, Crosby et al. 2011).

Table 4.3. NUMTs and NUPTs detected in select organisms.

Genome	Classification	Assembly Version	# NUMTs	# NUPTs
*Aureococcus anophagefferens**	Stramenopile	1	NA	0
*Chlamydomonas reinhardtii**	Chlorophyta	1	44	53
*Chlamydomonas reinhardtii**	Chlorophyta	3	49	53
*Chlamydomonas reinhardtii**	Chlorophyta	4	51	48
*Chlorella variabilis**	Chlorophyta	1	2	55
Cyanidioschyzon merolae	Rhodophyta	final	0	2
Ectocarpus siliculosus	Stramenopile	2	66	28
*Emiliania huxleyi**	Haptophyta	1	6	36
Micromonas strain RCC299*	Chlorophyta	3	0	0
*Micromonas pusilla**	Chlorophyta	2	0	18
*Ostreococcus lucimarinus**	Chlorophyta	1	0[1]	0
Ostreococcus RCC809*	Chlorophyta	1	0	0
*Ostreococcus tauri**	Chlorophyta	2	2	4
*Phaeodactylum tricornutum**	Stramenopile	2	3	3
*Physcomitrella patens**	Streptophyta	1.1	385	72
Phytophthora infestans	Stramenopile	1	552	-
*Phytophthora ramorum**	Stramenopile	1.1	14	-
*Phytophthora sojae**	Stramenopile	1.1	163	-
*Thalassiosira pseudonana**	Stramenopile	3	0	0
*Volvox carteri**	Chlorophyta	2	428	927

* Nuclear and organellar genomes obtained from the JGI Genome Portal http://genome.jgi.doe.gov/
Cyanidioschyzon merolae nuclear genome obtained from http://merolae.biol.s.u-tokyo.ac.jp/download/
Cyanidioschyzon chloroplast NC_004799, mitochondrion NC_000887
Ectocarpus siliculosus genomes obtained from https://bioinformatics.psb.ugent.be/gdb/ectocarpus/
Phytophthora infestans nuclear and mitochondrion genomes obtained from http://www.broadinstitute.org
[1]Used *Ostreococcus tauri* mitochondrial genome

What accounts for the differences in the number of recently transferred pieces of organellar DNA? Suggestions include variation in the size of the nuclear genome that the pieces are being inserted into (Hazkani-Covo, Zeller et al. 2010).

There is no correlation between the number of NUPTs and the organelle genome size (Table 4.4) for the unicellular photosynthetic organisms (Spearman non-parametric rho=0.32). The size range for plastid genomes is considerable, from 41 Kbs in *Micromonas* to 204 Kbs in *Chlamydomonas*. Similarly, no correlation was found between the size of the mitochondrial genome and the number of NUMTs identified either in unicellular algae (Table 4.5) (Spearman non-parametric rho=0.44) or in a more comprehensive study of 85 organisms that included fungi, plants and animals (Hazkani-Covo, Zeller et al. 2010). Curiously, *Chlamydomonas* has the largest plastid genome of the unicellular organisms evaluated but the smallest mitochondrial genome.

While no correlation was observed between the number of organellar DNA pieces transferred and the organelle genome size there does appear to be a correlation with the nuclear genome size and the number of NUMTs (Table 4.5)(Spearman non-parametric rho=0.77 p 0.0048). However, no correlation is seen between nuclear genome size and NUPTs (Table 4.4) (Spearman non-parametric rho=0.11) (if instances of no EGT are excluded rho=0.64 p=0.1). The 2010 study (Hazkani-Covo, Zeller et al. 2010), which was restricted to NUMTs, did find a correlation between NUMTs and genome size but only for genomes larger than 200 Mb. Prior studies of NUMTS and genome size did not explicitly test for correlation (Richly and Leister 2004) but suggestions were made that larger genomes would have more transferred organellar pieces (Bensasson, Zhang et al. 2001). The explanation for the observed correlation is that larger genomes contain more noncoding regions that are able to absorb insertions with little deleterious effect. In smaller genomes the likelihood of insertion into a gene is greater and the pressure to purge those pieces is consequently of greater importance.

Why is there correlation between NUMTs and nuclear genome size but not for NUPTs? Lack of correlation may be due to the small sample size as suggested for earlier studies (Hazkani-Covo, Zeller et al. 2010). As well, all of the genomes for the unicellular photosynthetic organisms are relatively small, ranging between 12 Mb and 167 Mb. The 2010 study (Hazkani-Covo, Zeller et al. 2010) indicated that the correlation between NUMTs and genome size did not hold for genomes under 200 Mb suggesting that there is a fundamental difference between large and small genomes when it comes to integration of organellar DNA and its persistence in the nuclear genome. However, I was able to demonstrate a correlation between NUMTs and genome size using virtually the same small genome species as for the NUPTs. The main difference between the two results is that three *Phytophthora* species were included in the NUMT study but not in the NUPT study because they lack plastids. If the *Phytophthora* results are excluded the correlation is no longer significant (rho=0.71 p=0.02).

The *Phytophthora* results in isolation would seem to strongly support a correlation between the number of transferred pieces and nuclear genome size. *P. ramorum* has the smallest genome (65 Mbps) as well as the fewest number of NUMTs (Table 4.5) at 14. As the genome size increases (*P. sojae* 186 Mbps, *P. infestans* 228 Mbps) the number of NUMTs increases (*P. sojae* 163, *P. infestans* 552). Since all three species belong to the same genus one would expect that lineage specific differences other than genome size that might contribute to the number of NUMTS would be greatly reduced compared to the differences between an animal and a plant. However, the strong correlation does not hold when one examines NUMT numbers in other cases where several genomes from the same genus have been investigated. For example, the Hazkani-Covo et al. study (Hazkani-Covo, Zeller et al. 2010) includes results from three *Aspergillus* species. The genome size range is only 8 Mb (29, 34, 37) but the number of NUMTs differs widely (11, 6, 64). The four *Drosophila* species and the four *Schizosaccharomyces* species (Hazkani-Covo, Zeller et al. 2010) also exhibit no correlation between NUMT numbers and genome size. Since all of these genomes are less than 200 Mb it may be that the lack of intraspecific correlation reflects a fundamental difference between "small" and "large" genomes.

Another popular explanation given for lineage specific differences in the number of transferred organellar pieces is the number of organelles per cell. More specifically, species with a single organelle should have little or no transfers (Barbrook, Howe et al. 2006). Since the source of organellar DNA for integration is believed to be lysed organelles, having the only organelle disintegrate would be cell suicide. This hypothesis was tested recently for both NUPTs and NUMTs (Smith, Crosby et al. 2011) and in both cases it was concluded that organisms with single organelles had significantly less transfers than organisms with multiple organelles per cell. Polyplastidic organisms had, on average 80 times more transferred plastid DNA than monoplastidic organisms.

While these results seem intuitive, closer examination of individual cases suggests that the limited transfer window hypothesis is not the whole story. The diatom *Thalassiosira pseudonana* has multiple mitochondria and plastids (Misumi, Yoshida et al. 2008) yet I was not able to detect any NUMTS or NUPTs, while other unicellular algae with one plastid, including another diatom, *Phaeodactylum tricornutum*, had at least a few NUPTs (Table 4.4) (Smith, Crosby et al. 2011). *Volvox*, which has a single plastid, has over 1000 NUPTs which is more than the number of NUMTs (802) even though it has multiple mitochondria (Smith, Crosby et al. 2011). *Chlamydomonas* has roughly the same number of NUMTs as it does NUPTs yet has multiple mitochondria and only one plastid (Smith, Crosby et al. 2011).

The authors explain these exceptions to the limited window hypothesis by suggesting that the final fate of the potential NUPTs and NUMTs is determined by the genome's ability to detect and purge these exogenous sequences (Smith, Crosby et al. 2011). If indeed the capacity of the nuclear genome to tolerate transferred DNA is the determining factor for the number of NUMTs and NUPTS then whether the cell has one or many organelles should not really matter.

Table 4.4. NUPTs, organellar genome size and organelle number per cell of selected organisms.

	NUPTs	Organelle genome size	Genome size	Organelle #
Micromonas pusilla	18	41,811	21.9	1
Bigelowiella natans	0	69,166	94.7	1
Ostreococcus tauri	4	71,666	12.5	1
Micromonas RCC299	0	72,585	20.0	1
Aureococcus anophagefferens	0	89,599	56.6	1
Emiliania huxleyi	36	105,309	167.7	1
Phaeodactylum tricornutum	2	117,369	26.1	1
Guillardia theta	0	121,524	87.2	1
Thalassiosira pseudonana	0	128,814	28.7	Many
Cyanidioschyzon merolae	2	149,887	16.5	1
Chlorella variabilis	55	150,613	49.0	1
Chlamydomonas reinhardtii	48	204,159	112.3	1

Table 4.5. NUMTs, organellar genome size and organelle number per cell of selected organisms.

	NUMTs	Organelle genome size	Genome size	Organelle #
Chlamydomonas reinhardtii	51	15,758	112.3	many
Emiliania huxleyi	6	29,013	167.7	1
Cyanidioschyzon merolae	0	32,211	16.5	1
Bigelowiella natans	6	36,946	94.7	1
Phytophthora infestans	552	37,957	228.5	Unknown
Phytophthora ramorum	14	39,314	65	Unknown
Guillardia theta	13	40,622	87.2	1
Micromonas pusilla	0	41,691	21.9	1
Phytophthora sojae	163	42,977	86.0	Unknown
Thalassiosira pseudonana	0	43,827	28.7	Many
Ostreococcus tauri	2	44,237	12.5	1
Micromonas RCC299	0	47,425	20.0	1
Ostreococcus RCC809	0	48,593	13.2	1
Phaeodactylum tricornutum	3	77,356	26.1	many

Determination of organelle numbers per cell is difficult at best. Cell biologists were startled to discover that the many mitochondria of yeast were in fact a single, large, branched mitochondrion (Hoffmann and Avers 1973). However, recent 3D studies using ultrathin-sectioning electron microscopy have now determined that yeast have 1-4 mitochondria depending on the cell and the physiological state (Yamaguchi, Namiki et al. 2011). Given the uncertainty surrounding organellar numbers in intensively studied organisms like *Saccharomyces cerevisiae* it begs the question of how reliable estimates are for more obscure species with only a few ultrastructure investigations. Most of the taxonomic monographs for algal species are hopelessly vague when it comes to organelle numbers, particularly for mitochondria. It is unclear whether this vagueness results from indifference, assumptions about organelle numbers, or an implicit acknowledgement of the difficulties both biological and technical of determining precise numbers. Consequently, statements about organelle numbers in studies on NUMTs and NUPTs in protists, especially for mitochondria, should be treated with caution.

Given that the limited window hypothesis (Barbrook, Howe et al. 2006) does not state that no transfers will occur in species with a single organelle, only that they will be limited, this implies that there are occasions during which the single organelle is able to transfer DNA to the host nucleus. What those other routes are is not mentioned.

One of the more significant sources for lysed organellar DNA, even for cells with a single organelle is sexual reproduction. The vast majority of eukaryotic organisms, at least those studied, exhibit uniparental inheritance of organelles (Birky 2008; Takano, Onoue et al. 2010). The precise mechanisms are varied but at some stage during the sexual process some of the organelles are lost. For example in *Chlamydomonas* (Nishimura and Stern 2010) the zygote contains two chloroplasts, one from each mating type. During gamete formation the chloroplast of one of the types is protected from degradation while the other is subsequently digested in the zygote by mate specific DNase that results in the release of plastid DNA into the cytoplasm. In other organisms the degradation occurs during gametogenesis resulting in gametes without any organelles (Kuroiwa and Uchida 1996). Potentially, however, during the process of creating cells without organelles some of the gametes may have experienced integration of organellar DNA, which is then passed on during fertilization. Whatever the precise method by which uniparental inheritance of organelles occurs ample amounts of lyzed organellar DNA are available in the cytosol for uptake into the nucleus. Indeed, it has even been hypothesized that the mechanisms for degradation of a certain portion of the organelles evolved to provide a source of nucleotides during periods of starvation (Sears and VanWinkle-Swift 1994).

The sexual life cycle or even whether they experience sexual reproduction is sometimes an open question for protists. However, asexual reproduction also provides opportunities for the production of NUMTs or NUPTs. During cell division the organelles are replicated but if something goes wrong and the process is aborted or a faulty organelle is created the damaged organelles will be lysed.

The limited window hypothesis seems less than convincing as an explanation for the low number of NUPTs and/or NUMTs in certain lineages. While it is tempting to ascribe a single and relatively simple cause to a widespread phenomenon the truth seems more likely to be messy and particular to each lineage as a diverse set of unique cellular conditions and lifestyles interpenetrate to control the flow and integration of exogenous DNA into the host genome and ultimately whether it will be retained.

Whatever the reasons for the extent to which a lineage experiences the transfer of organellar DNA to the host genome, *B. natans* and *G. theta* do not have any NUPTs or for that matter NUNMs. Since they both have some NUMTs one can assume that the nuclear genome is perfectly capable of integrating organellar DNA. To my knowledge, it has not been suggested that the DNA from the mitochondrion might be somehow different or easier to integrate than plastid or nucleomorph DNA. In fact because of unique mitochondrial genetic codes some NUMTs would be even harder to maintain and express than NUPTs (Adams and Palmer 2003). Since *B. natans* and *G. theta* have nucleomorphs and plastids that are sequestered behind two additional membranes compared to the mitochondrion, one might speculate that there is a physical barrier to the transfer of plastid and nucleomorph DNA. However, several observations tend to negate this suggestion. While nucleomorphs are relatively rare, secondarily acquired plastids are not and a number of organisms with four membranes around their plastids, like *Emiliania huxleyi* and *Ectocarpus siliculosus* (Table 4.3) have NUPTs. The very fact that few lineages have nucleomorphs illustrates the pervasive and widespread transfer of organellar DNA from secondarily acquired plastids and their original hosts.

Can the lack of NUNMs in *B. natans* and *G. theta* tell us anything about the fate of nucleomorphs in these lineages? They also lack NUPTs but the transfer or lack thereof of plastid DNA to the host genome should be seen as a separate phenomenon. As has already been stated green plants exhibit a tremendous rate of transfer and integration of plastid DNA yet their plastids remain and with sizeable genomes. The persistence of plastids and the reason(s) for it has been the subject of considerable speculation and debate (Race, Herrmann et al. 1999; Keeling and Slamovits 2005) which I believe is not pertinent to the general course of evolution experienced by engulfed algal hosts. There appear to be legitimate regulatory and biochemical reasons for the persistence of plastids and a plastid genome but seemingly no plausible rationale for the vestigial remains of the captured algal host and the end result in almost all cases has been the disappearance of that captured nucleus.

Every secondarily photosynthetic lineage at some point had a nucleomorph or its structural and genomic equivalent. It is believed that the vast majority of the endosymbiotic gene transfer occurred relatively quickly (Timmis, Ayliffe et al. 2004) reducing the captured photosynthetic alga to a ghost of itself. But the transfers continued until the vestigial remains could be dispensed with without killing the cell. How gradual the decline of the “nucleomorph” was is unclear but presumably the process mirrored, albeit on a lesser scale, the more massive and ancient transfers. Random pieces of nucleomorph DNA would be integrated into the host genome. Since the nucleomorph retained a viable copy of the gene the transferred version was not immediately required and could acquire over time the necessary regulatory elements and targeting information. Once both copies were fully functional, the nucleomorph version could be lost without harm.

This process of transfer and decay appears to have been arrested in *B. natans* and *G. theta*. The lack of NUNMs demonstrates that for whatever reason these organisms are no longer able to transfer portions of the nucleomorph DNA to the host genome. It is very difficult for them to lose genes from the nucleomorph because there is no backup copy in the host nucleus. If the gene codes for a protein vital to the plastid or the nucleomorph, the loss of the gene would be highly detrimental if not fatal to the existence of the organism. Any genes that are lost from the nucleomorph would simply vanish rather than being transferred to the host nucleus. This appears to be the case in *G. theta*. Because three other nucleomorph genomes have been fully sequenced from other cryptophytes a comparative analysis of their gene complements was possible. 17 genes were identified in the nucleomorph genomes of *Cryptomonas paramecium* and *Hemiselmis andersenii* that are not present in the nucleomorph of *G. theta* (Lane, van den Heuvel et al. 2007; Tanifuji, Onodera et al. 2011). A search of the host genome failed to find copies of these 17 genes that were demonstrably the result of EGT. In several cases the required nucleomorph protein appears to be supplied by a duplication of a host derived gene accompanied by retargeting to the PPC. In *B. natans* the abundance of alternative splicing (Curtis et al. 2012 In press) may serve a similar compensatory role by allowing a host derived protein to function in several places including a PPC that has lost its native version of the protein.

4.3.4 Methodological Differences and Errors in Previous Studies

During the course of my research it became clear that the methodology and criteria for identifying NUMTs and NUPTs can result in different numbers from study to study. Here I summarize the most important factors.

A recent study of NUMTs in sequenced genomes (Hazkani-Covo, Zeller et al. 2010) included a number of protists in their study, some of which were the same as those analyzed by me herein. Interestingly, in all cases where NUMTs were detected the numbers differed between the 2010 analysis and mine (Table 4.6). The discrepancies between the number of NUMTs or NUPTs identified for an organism from different studies occurs for several reasons that are outlined below.

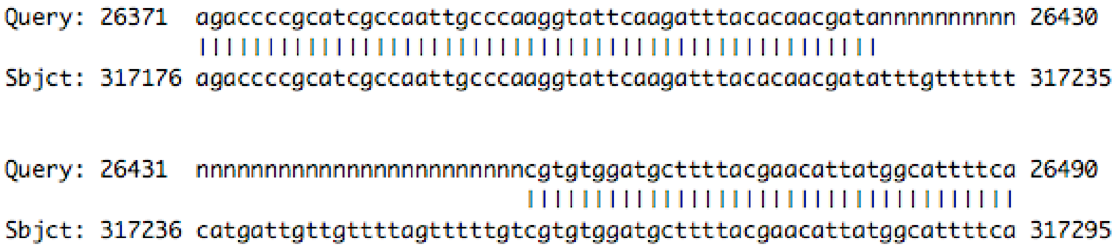
Table 4.6. NUMTs identified in protist nuclear genomes from (Hazkani-Covo, Zeller et al. 2010) and from current study.

	Hazkani-Covo et al. analysis	My analysis
Chlamydomonas reinhardtii	45	51
Ostreococcus tauri	7	2
Cyanidioschyzon merolae	0	0
Emiliania huxleyi	1	6
Thalassiosira pseudonana	0	0

The most obvious cause is that different versions of the genome have been used. With each new draft of the genome assembly possible NUMTs and NUPTs may appear or disappear as the genome is refined. For example in *Chlamydomonas reinhardtii* 44 NUMTs were detected using version one of the genome, 49 using version three and 51 using version four (Table 4.3). A similar situation occurred with *G. theta*. Two NUPTs were detected in the 4X coverage version as well as the preliminary 8X version. However, the final version removed these two areas as unsupported since they were bounded on all sides by a series of Ns.

NUMTs and NUPTs are detected through blastn searches of the organellar genome against the host genome. Consequently, the ability to identify pieces and the number of pieces is heavily dependent on blastn parameters. First and foremost is e value cutoff. In some organisms large chunks over 1 kb in length of organellar DNA are inserted into the host genome making it relatively easy to detect them through blastn. However, most of the pieces identified in protist genomes tend to be small. If the e value cutoff is too stringent small pieces may be missed. As well, NUMTs and NUPTs are usually not functional and thus prone to decay that can hide their origins if too strict an e value is used. Conversely, a high e value cutoff (e.g., 1) will tend to produce an abundance of spurious results, generally short areas of low complexity. Most studies have used an e value of 0.0001, which is seen as a compromise between capturing too many false hits and excluding the small and/or decayed pieces. I used an e value cutoff of 0.001.

Another blastn parameter that tends to be overlooked is r or the match reward. Prior to BLAST 2.2.18 the default r was one while subsequent versions have used a reward match value of two. Since few if any NUMT or NUPT analyses report their blastn parameters apart from e (the e value cutoff) one can assume that the default values have been employed. Unfortunately, the difference between a match reward of one and two when analyzing NUMTs or NUPTs can be substantial. A reward of two results in longer continuous matching areas. Because of the propensity for NUMTs and NUPTs to decay over time what may originally have been a single large transfer has the potential to be identified as two or more independent pieces. Additionally, an r=2 parameter is better at identifying single pieces that in an analysis using r=1 is reported as two pieces due to areas of low complexity. In the (Hazkani-Covo, Zeller et al. 2010) study seven NUMTs were identified in *Ostreococcus tauri*. Although the only blastn parameter they provide is an e value cutoff of 0.0001 it is clear from replicating their results that an r value of one was used. Two of those NUMTs (OT1, OT2) were identified from the same area of chromosome 17 (Table 4.7) and were 35 bps apart. The NUMTs also originated from the same area of the mitochondrial genome and were also 35 bps apart. However, if an r value of two is used the two pieces from chromosome 17 are identified as a single NUMT in a blastn analysis. The 35 bps separating the two pieces (26420-26455) are designated as areas of low complexity by blastn (Figure 4.3). Using an r value of two can also find pieces that are completely missed when using r=1 due to decay of the transferred piece. In the NUMT analysis of *G. theta* a 694 bp piece was found with an e value of 4e-65 when using r=2 but no portion of the piece was identified when using r=1. Only 68% (475/694) of the bases were identical between the current mitochondrial genome sequence and the transferred piece with most of the stretches of identical bases less than 10 bps.



```
Query: 26371  agaccccgcatcgccaattgcccaaggtattcaagatttacacaacgatannnnnnnnnn 26430
              ||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 317176 agaccccgcatcgccaattgcccaaggtattcaagatttacacaacgatatttgtttttt 317235

Query: 26431  nnnnnnnnnnnnnnnnnnnnnnnncgtgtggatgcttttacgaacattatggcattttca 26490
                                      ||||||||||||||||||||||||||||||||||||
Sbjct: 317236 catgattgttgttttagtttttgtcgtgtggatgcttttacgaacattatggcattttca 317295
```

Figure 4.3. Blastn alignment of *Ostreococcus tauri* genomic sequence (Sbjct) with a portion of its mitochondrial genome sequence (Query).

The content and structure of the organellar genome searched against the nuclear genome using BLAST can also affect how many transferred pieces are identified. Many organellar genomes have large areas of duplication and/or several copies of a gene. In automated screens for NUMTs or NUPTs the same transferred piece may hit two or more areas of the organellar genome and thus be counted several times. For example, in the *Ostreococcus tauri* NUMT study two pieces were counted twice. A 29 bp piece from chromosome 9 with the coordinates 182866-182894 was counted twice (OT4, OT5) since it matched two independent areas of the mitochondrial genome (21671-21699 and 34335-34363) (Table 4.7). Similarly, the same 43 bp piece from Chromosome 12 was counted twice (OT6, OT7) because it matched the same portion from the two mitochondrial copies of the ribosomal large subunit (Table 4.7).

Table 4.7. NUMTs identified in *Ostreococcus tauri*. Coordinates, size and chromosome number taken from (Hazkani-Covo, Zeller et al. 2010).

NUMT id	Chromosome	Mito start	Mito end	NUMT start	NUMT end	Size	conclusion
OT1	17	26011	26420	316816	317225	410	Same transferred piece
OT2	17	26455	26619	317260	317424	165	
OT3	0	9550	9610	406	466	61	NUPT
OT4	9	21671	21699	182866	182894	29	duplicate
OT5	9	34335	34363	182866	182894	29	
OT6	12	18709	18751	100437	100479	43	Duplicate LSU
OT7	12	37283	37325	100437	100479	43	

Another frequent error in NUMT and NUPT analyses is the failure to remove blastn hits against highly conserved genes that are found in the organellar genomes as well as the nuclear genome. Again, the *Ostreococcus* results illustrate this problem. As mentioned OT6 and OT7 are actually the same nuclear piece and should have been counted only once. An examination of the genomic context of OT6 and OT7 would reveal that this potential NUMT was actually a portion of the nuclear version of the ribosomal LSU which is a highly conserved and ubiquitous gene found in all organellar and nuclear genomes and consequently was not transferred from the mitochondrion. Similarly, OT3 is actually a NUPT and derives from a plastid version of an alpha subunit of ATP synthase CF1. Against the mitochondrial genome OT3 shares 88% of its bases but is 100% identical against the chloroplast genome. Other highly conserved genes that can create spurious hits are HSP70, EF-Tu and the rRNA small subunit.

The final hurdle to surmount prior to declaring a NUMT or NUPT as real is an examination of the individual reads that were used in the assembly for that particular region. It is extremely difficult to isolate pure nuclear genomic DNA. Consequently, organellar DNA invariably contaminates the nuclear DNA to be sequenced. During the creation of the individual pieces to be sequenced chimeric products comprised of organellar and nuclear DNA can be generated. This can lead to assembled portions of organellar DNA being integrated into the nuclear genome. These miss-assemblies can be identified by examining the reads that cross the boundary between NUPT or NUMT and the host nuclear genome. If there are two or more independent reads (not from the same clone or piece) that share the same boundary arrangement between NUMT or NUPT and the nuclear genome it is likely a real transferred piece. In a number of cases, including from *G. theta* and *B. natans,* I have rejected potential NUPTs or NUMTs because the read arrangements strongly suggested chimeric sequences leading to spurious integrations (Table 4.8). Unfortunately it is not always possible to obtain the individual reads for eukaryotic genome projects.

Table 4.8. Potential NUPTs rejected after examination of individual reads.

	# of potential NUPTs	# of NUPTs after read analysis
Cyanidioschyzon merolae	2	1
Emiliania huxleyi	36	21
Micromonas pusilla	18	18
Phaeodactylum tricornutum	2	2
Chlamydomonas reinhardtii	48	48

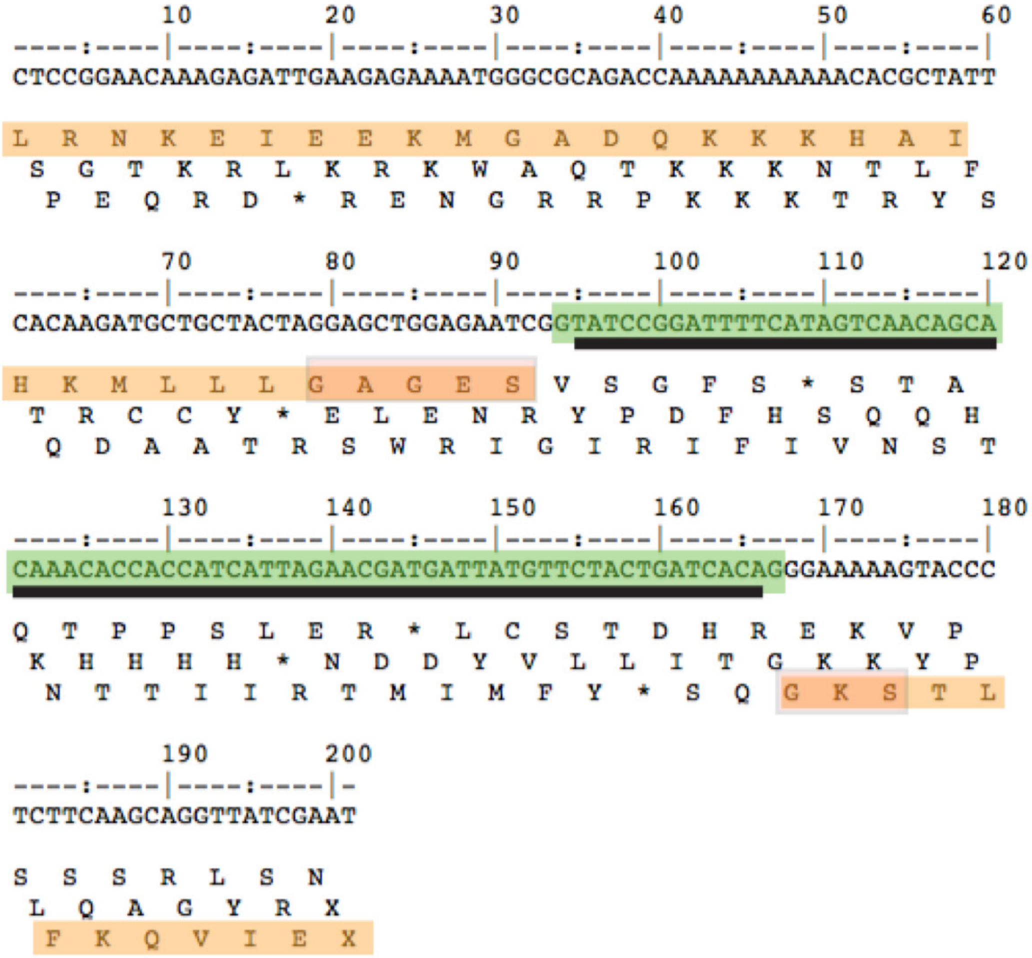
In conclusion, of the seven NUMTS identified in *Ostreococcus tauri* by the 2010 study only the two found by me and corresponding to OT1/OT2 and OT4/OT5 can be considered real after taking into account the various difference and errors that are common in NUMT and NUPT analyses. Similar problems and methodological differences with the 2010 data account for the rest of the discrepancies between the number of NUMTs they identified and the numbers found by me (Table 4.6). The same holds true for a study of NUMTs in *Phytophthora ramorum* and *Phytophthora sojae* (Krampis, Tyler et al. 2006).

4.3.5 A Spliceosomal Intron of Mitochondrial DNA Origin

The discovery that eukaryotic genes are composed of coding regions (exons) interspersed with non-coding segments (introns) that must be spliced out to create a functioning protein was surprising (Berget, Moore et al. 1977). Since then debate has raged about the origin of spliceosomal introns, their function and how new ones are created (Roy and Irimia 2009). At least six mechanisms have been suggested to account for novel introns (Roy and Irimia 2009) but the lack of examples of genuine recent gains coupled with the rapid rate at which introns evolve has obscured their origins. Significantly, the identified mechanisms relate to new introns derived from pre-existing ones or internal modifications of the nuclear genome rather than introns arising from the incorporation of exogenous material. The acquisition of novel introns via DNA insertion was examined in *Drosophila* (Farlow, Meduri et al. 2010) but there was no discussion of where the DNA might have come from.

One possible source of DNA that may create new introns is NUMTs and NUPTs (or NUNMs). Previous analyses of these organellar fragments in nuclear DNA often include information on their new genomic environment including their proximity to existing genes. The vast majority of NUMTs and NUPTs are found in noncoding areas, but occasionally the fragments interrupt existing exons or introns and there are 2 reports of NUMTS creating new introns. The first case involved a 74 bp mitochondrial DNA insertion into a human gene (Ricchetti, Tekaia et al. 2004). The second report came from a detailed examination of regions from various *Daphnia* subpopulations (Li, Tucker et al. 2009).

In the course of examining the nuclear genomic context of the NUMTS in *B. natans* I noted that one of them, bn164-1, appeared to be in an intron of a gene for an alpha subunit of a guanine binding protein. More to the point, the NUMT appeared to be the entire intron. It consisted of a 72 bp piece of the *cox*1 gene and was 88% identical to the mitochondrial DNA (Figure 4.4). To demonstrate that it is indeed a spliceosomal intron I conducted reverse transcriptase experiments. Ten clones of a PCR fragment that includes the intron of interest were sequenced. Six of the 10 clones had the NUMT/intron spliced out. The cloned PCR fragment had two additional introns and examination of the entire sequenced region showed high levels of alternative splicing with almost every combination of spliced/not spliced for the three introns (Figure 4.5). These results provided one of the first indications that *B. natans* displays an unprecedented level of alternative splicing, particularly intron retention, that was later examined in greater detail for the *B. natans* and *G. theta* genome paper (Curtis et al. 2012 In press)



```
         10        20        30        40        50        60
----:----|----:----|----:----|----:----|----:----|----:----|
CTCCGGAACAAAGAGATTGAAGAGAAAATGGCGCAGACCAAAAAAAAAAACACGCTATT

L  R  N  K  E  I  E  E  K  M  G  A  D  Q  K  K  K  H  A  I
 S  G  T  K  R  L  K  R  K  W  A  Q  T  K  K  K  N  T  L  F
  P  E  Q  R  D  *  R  E  N  G  R  R  P  K  K  K  T  R  Y  S

         70        80        90       100       110       120
----:----|----:----|----:----|----:----|----:----|----:----|
CACAAGATGCTGCTACTAGGAGCTGGAGAATCGGTATCCGGATTTTCATAGTCAACAGCA

H  K  M  L  L  L  G  A  G  E  S  V  S  G  F  S  *  S  T  A
 T  R  C  C  Y  *  E  L  E  N  R  Y  P  D  F  H  S  Q  Q  H
  Q  D  A  A  T  R  S  W  R  I  G  I  R  I  F  I  V  N  S  T

        130       140       150       160       170       180
----:----|----:----|----:----|----:----|----:----|----:----|
CAAACACCACCATCATTAGAACGATGATTATGTTCTACTGATCACAGGGAAAAAGTACCC

Q  T  P  P  S  L  E  R  *  L  C  S  T  D  H  R  E  K  V  P
 K  H  H  H  H  *  N  D  D  Y  V  L  L  I  T  G  K  K  Y  P
  N  T  T  I  I  R  T  M  I  M  F  Y  *  S  Q  G  K  S  T  L

        190       200
----:----|----:----|-
TCTTCAAGCAGGTTATCGAAT

S  S  S  R  L  S  N
 L  Q  A  G  Y  R  X
  F  K  Q  V  I  E  X
```

Figure 4.4. NUMT bn164-1. Black line is the NUMT, green box is the intron, dark tan is the highly conserved walker A motif, light tan boxes are the amino acid sequence of the alpha subunit of the guanine binding protein.

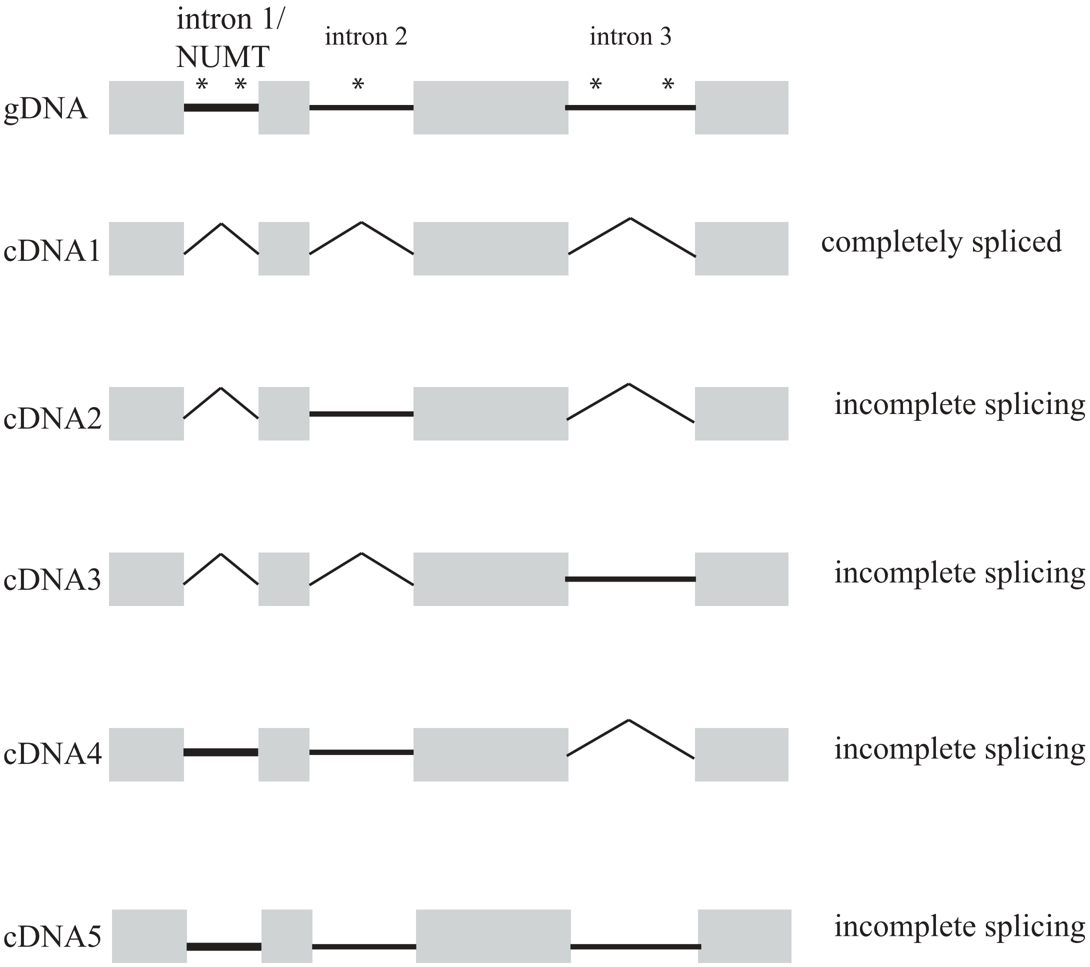


Figure 4.5. Pattern of intron splicing and retention for introns 1-3 of the *B. natans* protein 155542. Straight lines denote retained introns, angled lines indicate the intron has beenspliced, * denote stop codons.

While the intron contains canonical donor /acceptor sites (GT/AG) the corresponding region in the mitochondrial genome does not, instead being TT for GT and AA for AG. An analysis of 30,875 introns from *B. natans* found 15 with a TT donor site and 11 with an AA acceptor site. The vast majority of introns in *B. natans* have canonical splice sites

(Table 4. 9). It is unclear whether the original mitochondrial DNA or the NUMT after integration has mutated. RNA-Seq data from other chlorarachniophytes were queried but this paralog of the alpha subunit of a guanine binding protein was not found. Attempts to PCR the region from closely related chlorarachniophyte using degenerate primers and genomic DNA also failed to shed any light on the arrangement of the gene in other species. The region of interest was successfully amplified from three strains of *B. natans* (CCMP2755 (genomic project strain), CCMP1259, CCMP623). The sequenced products were identical from all three strains. No amplified products were generated for the *Lotharella globosa* strain (CCMP2314) or the *Bigelowiella longifila* strain (CCMP1481). The most likely scenario is that after integration the NUMT was subject to mutational pressure to produce a viable transcript with canonical donor/acceptor sites. It has been shown that novel introns in *Drosophila* are often not fully functional in that they have weak splice sites that can be missed by the spliceosome (Farlow 2010) and can result in intron retention. The aberrant mRNAs are subsequently destroyed by the Nonsense Mediated Decay (NMD) pathway due to the presence of premature termination codons (PTCs) (Sayani, Janis et al. 2008; Hansen, Lareau et al. 2009). Consequently, new introns without classical intronic features are shielded from selective pressure thus permitting these new introns time to develop into regions with strong splice sites. This very well may be the case for NUMT bn164-1 in gene 155542. All of the introns show some level of intron retention and they all have stops that would induce the NMD to destroy these truncated transcripts (Figure 4.5). Furthermore, bn155542 belongs to a large paralogous gene family with at least 35 members. During the early stages of the NUMT becoming a proper intron the reduced levels of correctly coded proteins may be compensated for by paralogs.

Table 4.9. Donor and acceptor dinucleotide splice sites in introns of *B. natans.*

Site type	Nucleotide pair	# of occurrences	Percentage
Donor	AA	1	
Donor	AC	2	
Donor	AG	0	
Donor	AT	25	
Donor	CA	0	
Donor	CC	0	
Donor	CG	0	
Donor	CT	56	
Donor	GA	27	
Donor	GC	158	0.5
Donor	GG	940	3
Donor	GT	29639	95.9
Donor	TA	15	
Donor	TC	0	
Donor	TG	0	
Donor	TT	0	
acceptor	AA	11	
acceptor	AC	61	0.19
acceptor	AG	30599	99.1
acceptor	AT	44	
acceptor	CA	2	
acceptor	CC	0	
acceptor	CG	34	
acceptor	CT	54	0.17
acceptor	GA	8	
acceptor	GC	0	
acceptor	GG	0	
acceptor	GT	0	
acceptor	TA	2	
acceptor	TC	0	
acceptor	TG	0	
acceptor	TT	48	

To what extent is the finding that one intron in *B. natans* appears to be the result of the integration of mitochondrial DNA into the host nuclear DNA significant? Is this the answer to where introns come from? It would appear that it is at least a partial answer since we now have three examples, two from metazoans and now from a protist. The metazoan examples both come from intensely studied model organisms, which may be an important factor in their discovery. Both organisms have genomic sequences from closely related species as well as regional sequences from sub populations, all of which increases the likelihood of finding novel introns that are new enough that their origins have not been disguised through time and mutation of either the intron or the DNA source. Because most NUMTs and NUPTs are small, non-functional pieces they are free to collect mutations that can quickly disguise their organellar ancestry and make them introns of unknown origin. In the case of *B. natans* it was simply chance that an intron was new enough to betray its source. I would predict that with increased sequencing capacity and the ability to look at sequencing differences in sub populations more examples of NUMTs, and NUPTs, as nascent or fully functional introns will be discovered.

CHAPTER 5 CONCLUSIONS

5.1 EGT – RECENT AND ANCIENT

During the initial stages of my graduate work I was given the opportunity to study EGT in two fascinating but highly complex and relatively unknown algal systems. Armed with the naiveté of a new student I believed that the genome projects for *Bigelowiella natans* and *Guillardia theta* would prove to be the ideal setting for investigating EGT and the transformative role it has for both the endosymbiont and the host. Four years later I am only now beginning to grasp the true complexity of this research goal and the problems associated with studying something that happened so long ago. Fortunately, the study of recent and ongoing EGT to the nucleus is much easier. It requires a significantly smaller toolkit of procedures and theoretical underpinnings. Despite its relative simplicity, the study of organellar DNA transfers to the nucleus can inform our understanding and conceptualization of the endosymbiotic processes that resulted in the establishment of the mitochondrion and the various plastid lineages. It can also provide insight into the evolving nature of genomes at the individual or species level.

My research into the transfer of organellar DNA to the nucleus in *G. theta* and *B. natans* yielded two significant results. First, I demonstrated that a piece of the mitochondrial genome that had been transferred to the nuclear genome in *B. natans* created a new spliceosomal intron. One of the puzzling aspects of eukaryotic genes is the origin of the introns that interrupt the protein coding sequence. Early debate on introns revolved around whether they were of ancient or recent origin. It is now clear that the introduction of introns to exonic stretches of the genome is an ongoing process. What still puzzles though is where the introns come from, the actual DNA. The creation of introns from existing ones has been demonstrated but until recently there were no known cases of *de novo* introns. My study of *B. natans* adds to a handful of examples of introns created by the integration of organellar DNA suggesting that, at least for some introns, the puzzle of their origin has been solved.

My study of transfers of mitochondrial (NUMT), plastid (NUPT) and nucleomorph DNA (NUNM) to the nucleus has also yielded insight into the puzzle of the persistence of nucleomorphs in certain lineages like *B. natans* and *G. theta*. While I was able to demonstrate the presence of NUMTs in both lineages, no NUPTs or NUNMs were found for either species. Organellar DNA in the nucleus tends to quickly become indistinguishable from the genomic landscape that it has been inserted into so it is possible that my failure to detect any transfers of plastid or nucleomorph DNA to the nucleus results from this propensity to "blend in." However, the detection of NUMTs in both organisms suggests that transfer from plastids and nucleomorphs has ceased as an ongoing process. The lack of transfers is not terribly significant from the standpoint of plastids since they persist in all photosynthetic organisms. Nucleomorphs, however, are highly unusual, because in the vast majority of lineages with secondary plastids the engulfed eukaryotic nucleus has vanished. We believe that the lack of NUNMs in the genomes of *B. natans* and *G. theta* is part of the answer to why nucleomorphs persist in these lineages. The reductive process of endosymbiosis is facilitated by EGT in that the loss of important genes from the endosymbiont can be compensated for by the gene copy transferred to the nucleus. If the endosymbiont, or in this case the organelle, cannot transfer genes to the nucleus any important gene losses that occur would be detrimental. In essence, the nucleomorph is stuck. It cannot lose genes without harming itself.

5.2 Mitochondrial proteomes

Because of the unusual nature of nucleomorphs and the insights that can be gained from studying them, much of the focus for *G. theta* and *B. natans* has been on plastids and secondary endosymbiosis. However, both organisms are the product of endosymbiotic processes even more ancient than the establishment of photosynthesis. The mitochondrion is a relict of the first endosymbiotic event that radically changed the eukaryotic lineage and as such warrants investigation if we wish to understand the impact EGT has had. My analysis of the mitochondrial proteomes for *B. natans* and *G. theta* is part of that investigation, a laying of the groundwork so to speak, to determine which genes remain in the mitochondrion, which genes have transferred to the host and target their proteins back to the mitochondrion, and which proteins derive from genes that are not related to the endosymbiont. The catalogue of mitochondrial targeted proteins that I have identified for *B. natans* and *G. theta* must be considered provisional and merely a first pass that needs to be tested and expanded upon through experimental means.

APPENDIX A1: Tree building protocol

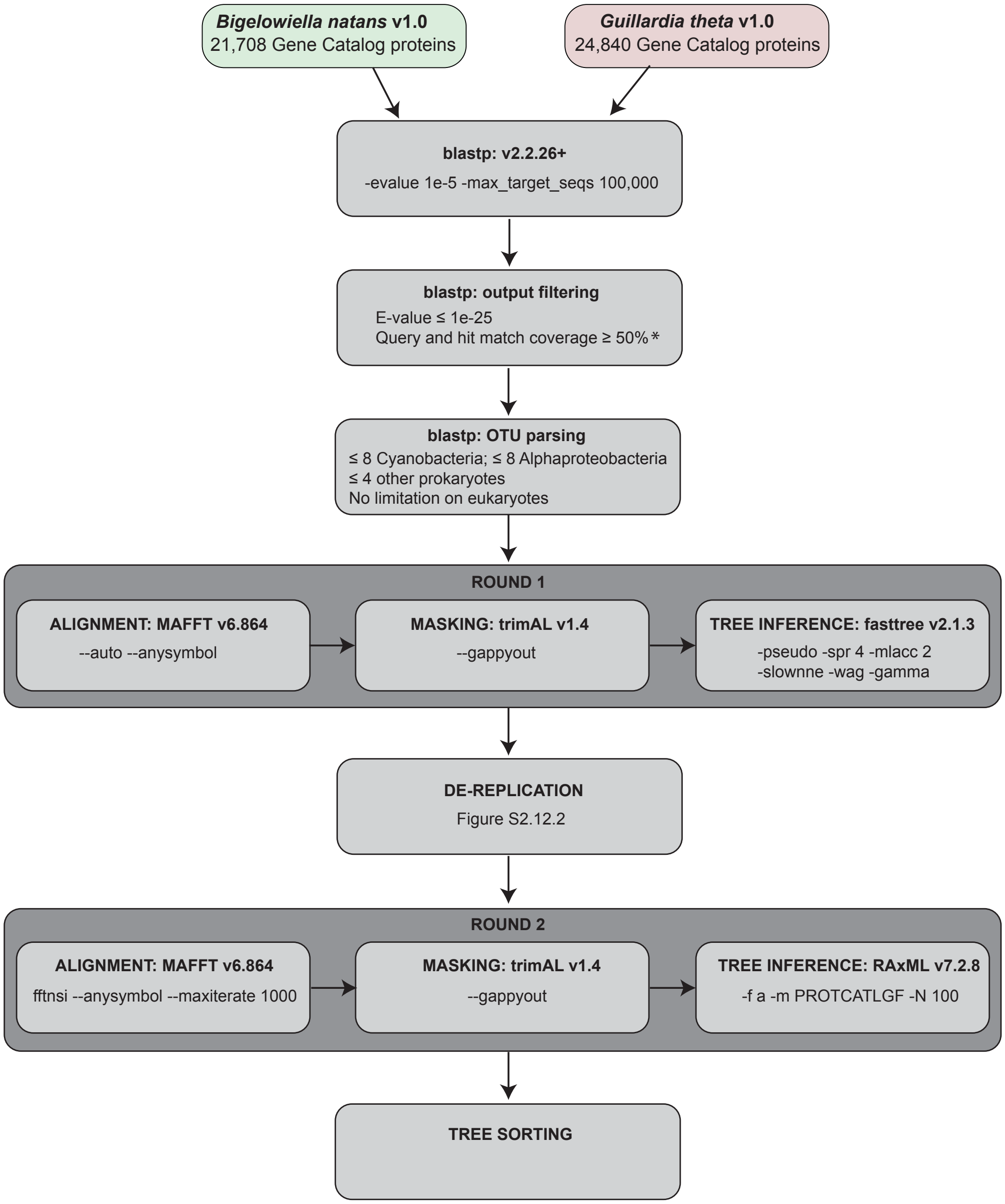

Bigelowiella natans v1.0
21,708 Gene Catalog proteins
Guillardia theta v1.0
24,840 Gene Catalog proteins
blastp: v2.2.26+
-evalue 1e-5 -max_target_seqs 100,000
blastp: output filtering
E-value ≤ 1e-25
Query and hit match coverage ≥ 50% *
blastp: OTU parsing
≤ 8 Cyanobacteria; ≤ 8 Alphaproteobacteria
≤ 4 other prokaryotes
No limitation on eukaryotes
ROUND 1
ALIGNMENT: MAFFT v6.864
--auto --anysymbol
MASKING: trimAL v1.4
--gappyout
TREE INFERENCE: fasttree v2.1.3
-pseudo -spr 4 -mlacc 2
-slownne -wag -gamma
DE-REPLICATION
Figure S2.12.2
ROUND 2
ALIGNMENT: MAFFT v6.864
fftnsi --anysymbol --maxiterate 1000
MASKING: trimAL v1.4
--gappyout
TREE INFERENCE: RAxML v7.2.8
-f a -m PROTCATLGF -N 100
TREE SORTING

Figure A1. Phylogenomics workflow for *Bigelowiella natans* and *Guillardia theta.*
*Not applied to proteins derived from EST / RNA-Seq data. Figure prepared by Fabien Burki (Curtis et al. 2012 under revision)

APPENDIX A2: De-replication

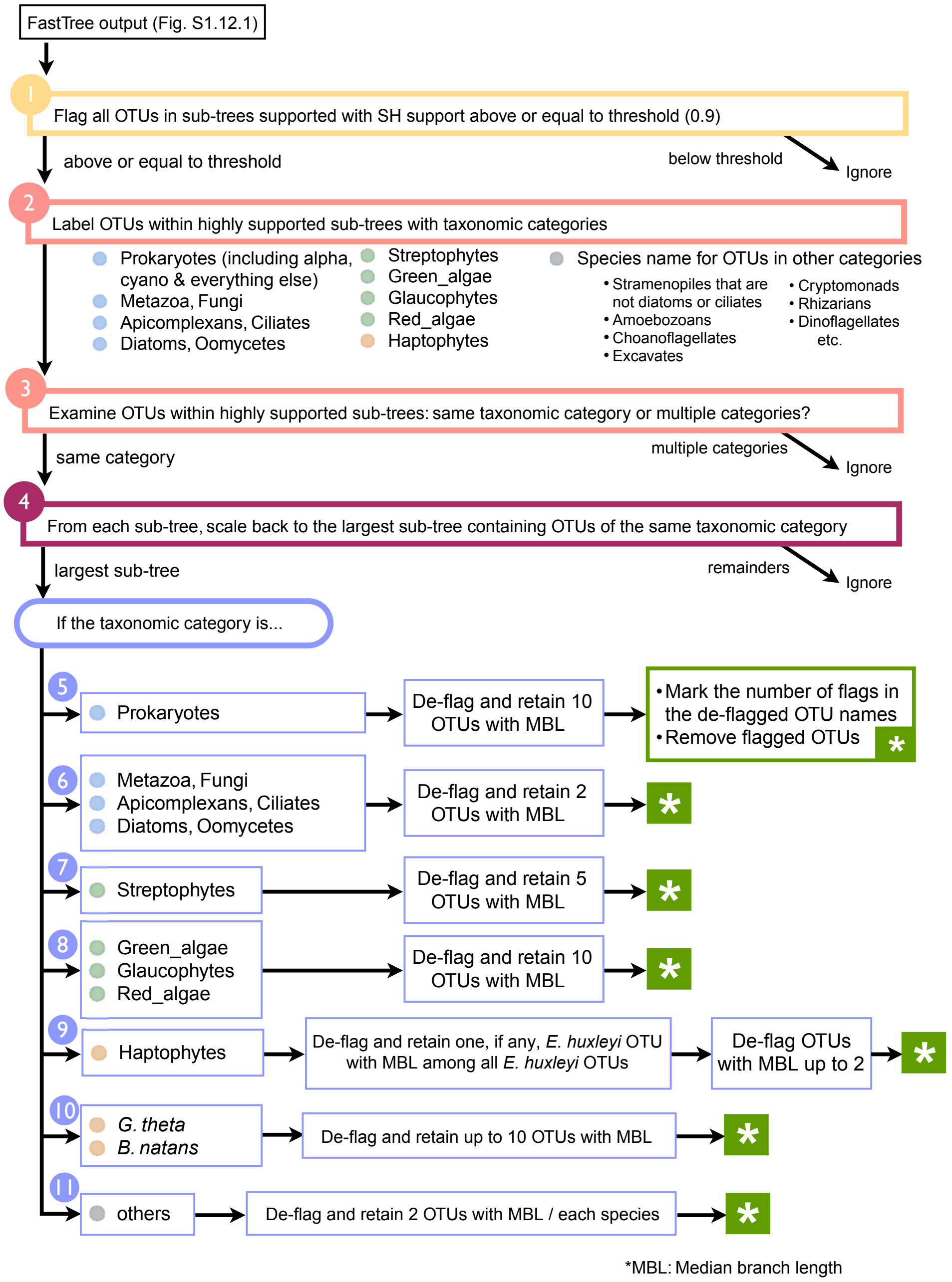

FastTree output (Fig. S1.12.1)
1
Flag all OTUs in sub-trees supported with SH support above or equal to threshold (0.9)
above or equal to threshold
below threshold
Ignore
2
Label OTUs within highly supported sub-trees with taxonomic categories
● Prokaryotes (including alpha, cyano & everything else)
● Metazoa, Fungi
● Apicomplexans, Ciliates
● Diatoms, Oomycetes
● Streptophytes
● Green_algae
● Glaucophytes
● Red_algae
● Haptophytes
● Species name for OTUs in other categories
• Stramenopiles that are not diatoms or ciliates
• Amoebozoans
• Choanoflagellates
• Excavates
• Cryptomonads
• Rhizarians
• Dinoflagellates
etc.
3
Examine OTUs within highly supported sub-trees: same taxonomic category or multiple categories?
same category
multiple categories
Ignore
4
From each sub-tree, scale back to the largest sub-tree containing OTUs of the same taxonomic category
largest sub-tree
remainders
Ignore
If the taxonomic category is...
5
● Prokaryotes
De-flag and retain 10 OTUs with MBL
•Mark the number of flags in the de-flagged OTU names
•Remove flagged OTUs
*
6
● Metazoa, Fungi
● Apicomplexans, Ciliates
● Diatoms, Oomycetes
De-flag and retain 2 OTUs with MBL
*
7
● Streptophytes
De-flag and retain 5 OTUs with MBL
*
8
● Green_algae
● Glaucophytes
● Red_algae
De-flag and retain 10 OTUs with MBL
*
9
● Haptophytes
De-flag and retain one, if any, *E. huxleyi* OTU with MBL among all *E. huxleyi* OTUs
De-flag OTUs with MBL up to 2
*
10
● *G. theta*
● *B. natans*
De-flag and retain up to 10 OTUs with MBL
*
11
● others
De-flag and retain 2 OTUs with MBL / each species
*
*MBL: Median branch length

Figure A2. De-replication workflow for reducing OTU complexity in phylogenetic trees of *Bigelowiella natans* and *Guillardia theta* proteins.Figure prepared by Shinichiro Maruyama (Curtis et al. 2012 under revision).

APPENDIX A3: Tree sorting

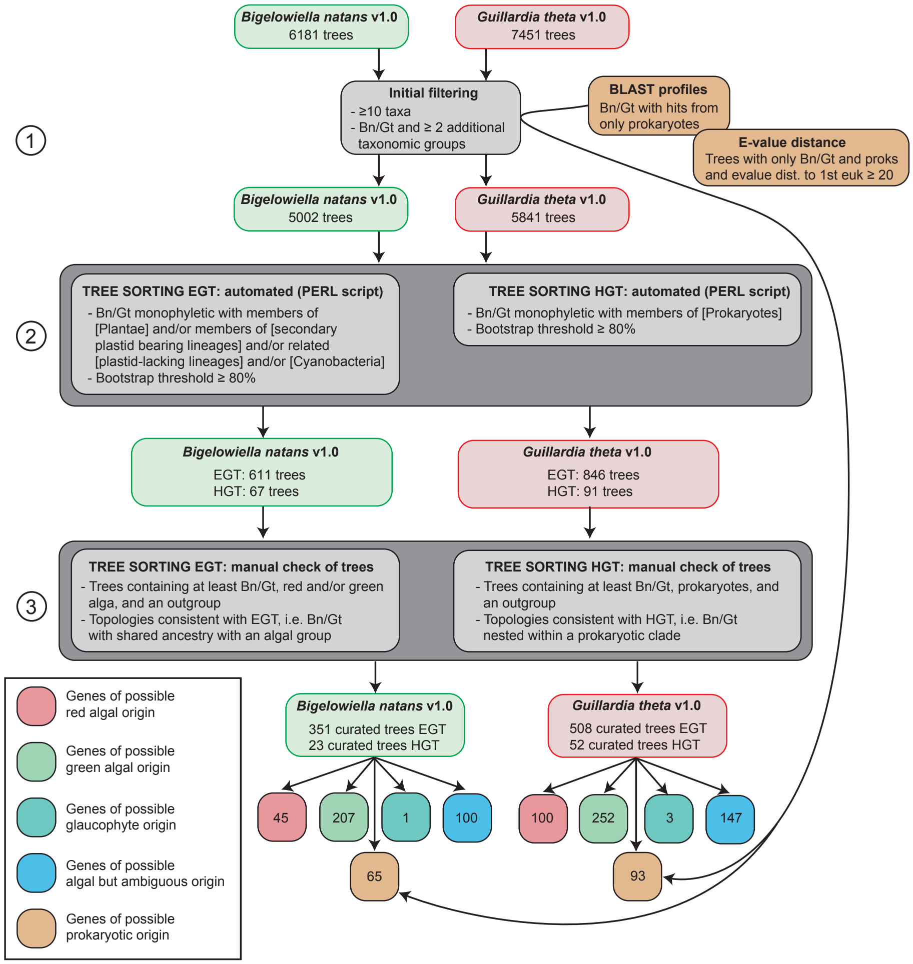

Bigelowiella natans v1.0
6181 trees
Guillardia theta v1.0
7451 trees
1
Initial filtering
- ≥10 taxa
- Bn/Gt and ≥ 2 additional taxonomic groups
BLAST profiles
Bn/Gt with hits from only prokaryotes
E-value distance
Trees with only Bn/Gt and proks and evalue dist. to 1st euk ≥ 20
Bigelowiella natans v1.0
5002 trees
Guillardia theta v1.0
5841 trees
2
TREE SORTING EGT: automated (PERL script)
- Bn/Gt monophyletic with members of [Plantae] and/or members of [secondary plastid bearing lineages] and/or related [plastid-lacking lineages] and/or [Cyanobacteria]
- Bootstrap threshold ≥ 80%
TREE SORTING HGT: automated (PERL script)
- Bn/Gt monophyletic with members of [Prokaryotes]
- Bootstrap threshold ≥ 80%
Bigelowiella natans v1.0
EGT: 611 trees
HGT: 67 trees
Guillardia theta v1.0
EGT: 846 trees
HGT: 91 trees
3
TREE SORTING EGT: manual check of trees
- Trees containing at least Bn/Gt, red and/or green alga, and an outgroup
- Topologies consistent with EGT, i.e. Bn/Gt with shared ancestry with an algal group
TREE SORTING HGT: manual check of trees
- Trees containing at least Bn/Gt, prokaryotes, and an outgroup
- Topologies consistent with HGT, i.e. Bn/Gt nested within a prokaryotic clade
Genes of possible red algal origin
Genes of possible green algal origin
Genes of possible glaucophyte origin
Genes of possible algal but ambiguous origin
Genes of possible prokaryotic origin
Bigelowiella natans v1.0
351 curated trees EGT
23 curated trees HGT
Guillardia theta v1.0
508 curated trees EGT
52 curated trees HGT
45
207
1
100
65
100
252
3
147
93

Figure A3. Tree-sorting procedure for detecting endosymbiotic gene transfer (EGT) and horizontal gene transfer (HGT) in the *Bigelowiella natans* and *Guillardia theta* genomes. Figure prepared by Fabien Burki (Curtis et al. 2012 under revision)

Appendix B: local database entries

Alveolate	Apicomplexa	Cryptosporidium_hominis	Genome
Alveolate	Apicomplexa	Cryptosporidium_parvum	Genome
Alveolate	Apicomplexa	Neospora_caninum	Genome
Alveolate	Apicomplexa	Plasmodium_berghei	Genome
Alveolate	Apicomplexa	Plasmodium_chabaudi	Genome
Alveolate	Apicomplexa	Plasmodium_falciparum	Genome
Alveolate	Apicomplexa	Toxoplasma_gondii	Genome
Alveolate	Chromerid	Chromera_velia	EST
Alveolate	Ciliate	Paramecium_tetraurelia	Genome
Alveolate	Ciliate	Tetrahymena_thermophila	Genome
Alveolate	Dinoflagellate	Alexandrium_catenella	EST
Alveolate	Dinoflagellate	Alexandrium_minutum	EST
Alveolate	Dinoflagellate	Alexandrium_ostenfeldii	EST
Alveolate	Dinoflagellate	Alexandrium_tamarense	EST
Alveolate	Dinoflagellate	Amphidinium_carterae	EST
Alveolate	Dinoflagellate	Heterocapsa_triquetra	EST
Alveolate	Dinoflagellate	Karenia_brevis	EST
Alveolate	Dinoflagellate	Karlodinium_micrum	EST
Alveolate	Dinoflagellate	Oxyrrhis_marina	EST
Alveolate	Dinoflagellate	Perkinsus_marinus	Genome
Alveolate	Dinoflagellate	Symbiodinium_sp	EST
Amoebozoa		Dictyostelium_discoideum	Genome
Amoebozoa		Dictyostelium_purpureum	Genome
Choanozoa		Monosiga_brevicollis	Genome
Choanozoa		Salpingoeca_sp	Genome
Cryptomonad		Goniomonas_pacifica	EST
Cryptomonad		Guillardia_theta	Genome
Cryptomonad		Rhodomonas_salina	EST
Excavate		Naegleria_gruberi	Genome
Excavate	Euglenid	Euglena_gracilis	EST
Excavate	Euglenid	Euglena_mutabilis	EST
Excavate	Euglenid	Euglena_longa	EST
Fungi		Aspergillus_niger	Genome
Fungi		Batrachochytrium_dendrobatidis	Genome
Fungi		Cryptococcus_neoformans	Genome
Fungi		Laccaria_bicolor	Genome
Fungi		Neurospora_crassa	Genome
Fungi		Phycomyces_blakesleeanus	Genome
Fungi		Ustilago_maydis	Genome
Haptophyte		Emiliania_huxleyi	Genome

Haptophyte		Isochrysis_galbana	EST
Haptophyte		Pavlova_lutheri	EST
Haptophyte		Prymnesium_parvum	EST
Katablepharid		Roombia_truncata	EST
Metazoa		Branchiostoma_floridae	Genome
Metazoa		Ciona_intestinalis	Genome
Metazoa		Danio_rerio	Genome
Metazoa		Daphnia_pulex	Genome
Metazoa		Gallus_gallus	Genome
Metazoa		Homo_sapiens	Genome
Metazoa		Lottia_gigantea	Genome
Metazoa		Mus_musculus	Genome
Metazoa		Nematostella_vectensis	Genome
Metazoa		Rattus_norvegicus	Genome
Metazoa		Trichoplax_adhaerens	Genome
Plantae	Glaucophyte	Cyanophora_paradoxa	Genome
Plantae	Glaucophyte	Glaucocystis_nostochinearum	EST
Plantae	Green algae	Asterochloris_sp	Genome
Plantae	Green algae	Chlamydomonas_reinhardtii	Genome
Plantae	Green algae	Chlorella_vulgaris	Genome
Plantae	Green algae	Coccomyxa_sp	Genome
Plantae	Green algae	Micromonas_pusilla	Genome
Plantae	Green algae	Micromonas_sp	Genome
Plantae	Green algae	Ostreococcus_lucimarinus	Genome
Plantae	Green algae	Ostreococcus_tauri	Genome
Plantae	Green algae	Volvox_carteri	Genome
Plantae	Red algae	Calliarthron_tuberculosum	Genome survey
Plantae	Red algae	Chondrus_crispus	EST
Plantae	Red algae	Eucheuma_denticulatum	EST
Plantae	Red algae	Furcellaria_lumbricalis	EST
Plantae	Red algae	Galdieria_sulphuraria	EST
Plantae	Red algae	Gracilaria_sp	EST
Plantae	Red algae	Griffithsia_okiensis	EST
Plantae	Red algae	Porphyra_haitanensis	EST
Plantae	Red algae	Porphyra_yezoensis	EST
Plantae	Red algae	Porphyridium_cruentum	EST
Plantae	Red algae	Cyanidioschyzon_merolae	Genome
Plantae	Streptophyte	Arabidopsis_thaliana	Genome
Plantae	Streptophyte	Brachypodium_distachyon	Genome
Plantae	Streptophyte	Glycine_max	Genome
Plantae	Streptophyte	Medicago_truncatula	Genome
Plantae	Streptophyte	Mimulus_guttatus	Genome
Plantae	Streptophyte	Oryza_sativa	Genome

Plantae	Streptophyte	Physcomitrella_patens	Genome
Plantae	Streptophyte	Populus_trichocarpa	Genome
Plantae	Streptophyte	Ricinus_communis	Genome
Plantae	Streptophyte	Selaginella_moellendorffii	Genome
Plantae	Streptophyte	Vitis_vinifera	Genome
Plantae	Streptophyte	Zea_mays	Genome
Prokaryotes		Actinobacteria	Genome
Prokaryotes		Aquificae	Genome
Prokaryotes		Bacteroidetes	Genome
Prokaryotes		Chlamydiae	Genome
Prokaryotes		Chlorobi	Genome
Prokaryotes		Chloroflexi	Genome
Prokaryotes		Crenarchaeota	Genome
Prokaryotes		Cyanobacteria	Genome
Prokaryotes		Deferribacteres	Genome
Prokaryotes		Deinococci	Genome
Prokaryotes		Euryarchaeota	Genome
Prokaryotes		Firmicutes	Genome
Prokaryotes		Fusobacteria	Genome
Prokaryotes		Nitrospirae	Genome
Prokaryotes		Planctomycetes	Genome
Prokaryotes		Alphaproteobacteria	Genome
Prokaryotes		Proteobacteria-nonalpha	Genome
Prokaryotes		Spirochaetes	Genome
Prokaryotes		Synergistetes	Genome
Prokaryotes		Tenericutes	Genome
Prokaryotes		Thaumarchaeota	Genome
Prokaryotes		Thermotogae	Genome
Prokaryotes		Unclassified Bacteria	Genome
Prokaryotes		Verrucomicrobia	Genome
Rhizaria		Bigelowiella_natans	Genome
Rhizaria		Gromia_sphaerica	EST
Rhizaria		Gymnochlora_stellata	EST
Rhizaria		Paracercomonas_marina	EST
Rhizaria		Reticulomyxa_filosa	EST
Stramenopile	Diatom	Fragilariopsis_cylindrus	Genome
Stramenopile	Diatom	Phaeodactylum_tricornutum	Genome
Stramenopile	Diatom	Thalassiosira_pseudonana	Genome
Stramenopile	Oomycete	Phytophthora_ramorum	Genome
Stramenopile	Oomycete	Phytophthora_sojae	Genome
Stramenopile	Oomycete	Saprolegnia_parasitica	Genome
Stramenopile		Aureococcus_anophageferrens	Genome
Stramenopile		Ectocarpus_siliculosus	Genome

APPENDIX C1: Mitochondrial targeted proteins in *Bigelowiella natans*

53491;KOG1615;Phosphoserine phosphatase
66650;KOG0430;Xanthine dehydrogenase
86309;KOG1441;Glucose-6-phosphate/phosphate and phosphoenolpyruvate/phosphate antiporter
87240;KOG0613;Projectin/twitchin and related proteins
143160;KOG1399;Flavin-containing monooxygenase
37883;KOG2358;NifU-like domain-containing proteins
38594;KOG0460;Mitochondrial translation elongation factor Tu
39560;KOG3446;NADH:ubiquinone oxidoreductase NDUFA2/B8 subunit
43207;KOG1748;"Acyl carrier protein/NADH-ubiquinone oxidoreductase, NDUFAB1/SDAP subunit"
46295;KOG2282;"NADH-ubiquinone oxidoreductase, NDUFS1/75 kDa subunit"
46458;KOG0571;Asparagine synthase (glutamine-hydrolyzing)
48748;KOG1120;Fe-S cluster biosynthesis protein ISA1 (contains a HesB-like domain)
49058;KOG1687;"NADH-ubiquinone oxidoreductase, NUFS7/PSST/20 kDa subunit"
49157;KOG0225;"Pyruvate dehydrogenase E1, alpha subunit"
49337;KOG3007;Mu-crystallin
49865;KOG2467;Glycine/serine hydroxymethyltransferase
49904;KOG3801;Uncharacterized conserved protein BCN92
50431;KOG0786;3-isopropylmalate dehydrogenase
52072;KOG0449;"Succinate dehydrogenase, cytochrome b subunit"
52185;KOG1715;Mitochondrial/chloroplast ribosomal protein L12
53042;KOG2403;"Succinate dehydrogenase, flavoprotein subunit"
53109;KOG1454;Predicted hydrolase/acyltransferase (alpha/beta hydrolase superfamily)
53251;KOG1758;"Mitochondrial F1F0-ATP synthase, subunit delta/ATP16"
53292;KOG1255;"Succinyl-CoA synthetase, alpha subunit"
53395;KOG0960;"Mitochondrial processing peptidase, beta subunit, and related enzymes (insulinase superfamily)"
53822;KOG3361;Iron binding protein involved in Fe-S cluster formation
54214;KOG1706;Argininosuccinate synthase
55226;KOG1549;Cysteine desulfurase NFS1
55266;KOG3078;Adenylate kinase
55480;KOG0571;Asparagine synthase (glutamine-hydrolyzing)
56108;KOG0453;Aconitase/homoaconitase (aconitase superfamily)
56191;KOG2794;Delta-aminolevulinic acid dehydratase
56530;KOG3365;"NADH:ubiquinone oxidoreductase, NDUFA5/B13 subunit"
57542;KOG0374;"Serine/threonine specific protein phosphatase PP1, catalytic subunit"
57639;KOG2540;Cytochrome oxidase assembly factor COX11
58151;KOG1680;Enoyl-CoA hydratase
68985;KOG3957;Predicted L-carnitine dehydratase/alpha-methylacyl-CoA racemase
69569;KOG0454;3-isopropylmalate dehydratase (aconitase superfamily)
72652;KOG0452;RNA-binding translational regulator IRP (aconitase superfamily)
84228;KOG2358;NifU-like domain-containing proteins
85155;KOG2725;Cytochrome oxidase assembly factor COX15
85465;KOG1569;50S ribosomal protein L1
85575;KOG2658;"NADH:ubiquinone oxidoreductase, NDUFV1/51kDa subunit"
85967;KOG2865;"NADH:ubiquinone oxidoreductase, NDUFA9/39kDa subunit"
86246;KOG3352;"Cytochrome c oxidase, subunit Vb/COX4"
86255;KOG0899;Mitochondrial/chloroplast ribosomal protein S19
86464;KOG0846;Mitochondrial/chloroplast ribosomal protein L15/L10
86518;KOG3057;"Cytochrome c oxidase, subunit VIb/COX12"
86772;KOG1563;"Mitochondrial protein Surfeit 1/SURF1/SHY1, required for expression of cytochrome oxidase"
86859;KOG3382;"NADH:ubiquinone oxidoreductase, B17.2 subunit"
87192;KOG2580;"Mitochondrial import inner membrane translocase, subunit TIM44"
87321;KOG3015;F1-ATP synthase assembly protein
87410;KOG3049;"Succinate dehydrogenase, Fe-S protein subunit"
87500;KOG1568;"Mitochondrial inner membrane protease, subunit IMP2"
87562;KOG2832;"TFIIF-interacting CTD phosphatase, including NLI-interacting factor (involved in RNA polymerase II regulation)"
87786;KOG1071;"Mitochondrial translation elongation factor EF-Tsmt, catalyzes nucleotide exchange on EF-Tumt"
89065;KOG0785;"Isocitrate dehydrogenase, alpha subunit"
90568;KOG2707;Predicted metalloprotease with chaperone activity (RNAse H/HSP70 fold)
90651;KOG1671;"Ubiquinol cytochrome c reductase, subunit RIP1"
91482;KOG2067;"Mitochondrial processing peptidase, alpha subunit"
91840;KOG0454;3-isopropylmalate dehydratase (aconitase superfamily)
92341;KOG0102;"Molecular chaperones mortalin/PBP74/GRP75, HSP70 superfamily"
92352;KOG1350;"F0F1-type ATP synthase, beta subunit"
92898;KOG0139;Short-chain acyl-CoA dehydrogenase
126152;KOG1641;Mitochondrial chaperonin
128068;KOG1680;Enoyl-CoA hydratase
135083;KOG1549;Cysteine desulfurase NFS1
135978;KOG3466;"NADH:ubiquinone oxidoreductase, NDUFB9/B22 subunit"

```
140293;KOG0975;"Branched chain aminotransferase BCAT1, pyridoxal phosphate enzymes type IV superfamily"
141113;KOG3309;Ferredoxin
144244;KOG1239;Inner membrane protein translocase involved in respiratory chain assembly
146631;KOG3281;Mitochondrial F1-ATPase assembly protein
26107;KOG1864;Ubiquitin-specific protease
33309;KOG2335;tRNA-dihydrouridine synthase
33503;KOG0054;"Multidrug resistance-associated protein/mitoxantrone resistance protein, ABC superfamily"
34567;KOG1440;CDP-diacylglycerol synthase
38633;KOG1481;Cysteine synthase
41304;KOG0876;Manganese superoxide dismutase
41505;KOG0619;FOG: Leucine rich repeat
42229;KOG0749;Mitochondrial ADP/ATP carrier proteins
44956;KOG3139;N-acetyltransferase
45150;KOG0480;"DNA replication licensing factor, MCM6 component"
45701;KOG1537;Homoserine kinase
46305;KOG0911;Glutaredoxin-related protein
46484;KOG0440;Cell cycle-associated protein Mob1-1
46835;KOG1504;Ornithine carbamoyltransferase OTC/ARG3
47084;KOG4073;Pterin carbinolamine dehydratase PCBD/dimerization cofactor of HNF1
47086;KOG0192;Tyrosine kinase specific for activated (GTP-bound) p21cdc42Hs
47320;KOG1696;60s ribosomal protein L19
48529;KOG0082;G-protein alpha subunit (small G protein superfamily)
49415;KOG1944;Peroxisomal membrane protein MPV17 and related proteins
50565;KOG0327;"Translation initiation factor 4F, helicase subunit (eIF-4A) and related helicases"
50797;KOG1641;Mitochondrial chaperonin
51026;KOG0024;Sorbitol dehydrogenase
51111;KOG3293;Small nuclear ribonucleoprotein (snRNP)
51505;KOG0239;Kinesin (KAR3 subfamily)
51525;KOG1721;FOG: Zn-finger
51736;KOG0634;Aromatic amino acid aminotransferase and related proteins
51785;KOG0558;Dihydrolipoamide transacylase (alpha-keto acid dehydrogenase E2 subunit)
51796;KOG1885;Lysyl-tRNA synthetase (class II)
51931;KOG0524;"Pyruvate dehydrogenase E1, beta subunit"
52064;KOG0725;Reductases with broad range of substrate specificities
52126;KOG3277;Uncharacterized conserved protein
52988;KOG3039;Uncharacterized conserved protein
53052;KOG0990;"Replication factor C, subunit RFC5"
53364;KOG1448;Ribose-phosphate pyrophosphokinase
53472;KOG1411;Aspartate aminotransferase/Glutamic oxaloacetic transaminase AAT1/GOT2
53893;KOG3954;"Electron transfer flavoprotein, alpha subunit"
54028;KOG3419;Mitochondrial/chloroplast ribosomal protein S16
54463;KOG2674;Cysteine protease required for autophagy - Apg4p/Aut2p
54652;KOG1790;60s ribosomal protein L34
54784;KOG2481;Protein required for normal rRNA processing
54827;KOG0660;Mitogen-activated protein kinase
55026;KOG4208;Nucleolar RNA-binding protein NIFK
55271;KOG1390;Acetyl-CoA acetyltransferase
55582;KOG3505;Mitochondrial/chloroplast ribosomal protein L33-like
55797;KOG0619;FOG: Leucine rich repeat
56044;KOG2319;Vacuolar assembly/sorting protein VPS9
56207;KOG3464;60S ribosomal protein L44
56574;KOG1915;Cell cycle control protein (crooked neck)
56650;KOG1119;Mitochondrial Fe-S cluster biosynthesis protein ISA2 (contains a HesB-like domain)
57056;KOG1422;"Intracellular Cl- channel CLIC, contains GST domain"
57687;KOG0438;Mitochondrial/chloroplast ribosomal protein L2
57958;KOG1956;DNA topoisomerase III alpha
58532;KOG0258;Alanine aminotransferase
58792;KOG3331;Mitochondrial/chloroplast ribosomal protein L4/L29
59059;KOG2834;"Nuclear pore complex, rNpl4 component (sc Npl4)"
60053;KOG0817;Acyl-CoA-binding protein
60462;KOG1721;FOG: Zn-finger
60737;KOG2933;Uncharacterized conserved protein
64543;KOG4362;Transcriptional regulator BRCA1
65978;KOG0787;Dehydrogenase kinase
66429;KOG1686;Mitochondrial/chloroplast ribosomal L21 protein
66486;KOG1809;Vacuolar protein sorting-associated protein
66534;KOG1404;Alanine-glyoxylate aminotransferase AGT2
67649;KOG4197;FOG: PPR repeat
68074;KOG0257;"Kynurenine aminotransferase, glutamine transaminase K"
68375;KOG1252;Cystathionine beta-synthase and related enzymes
```

```
68474;KOG0588;Serine/threonine protein kinase
69042;KOG0601;Cyclin-dependent kinase WEE1
69049;KOG0296;Angio-associated migratory cell protein (contains WD40 repeats)
69347;KOG2635;Medium subunit of clathrin adaptor complex
70006;KOG4308;LRR-containing protein
70289;KOG4198;RNA-binding Ran Zn-finger protein and related proteins
70324;KOG1501;Arginine N-methyltransferase
70855;KOG2855;Ribokinase
71113;KOG4762;DNA replication factor
71338;KOG0450;"2-oxoglutarate dehydrogenase, E1 subunit"
71479;KOG1031;Predicted Ca2+-dependent phospholipid-binding protein
71916;KOG1721;FOG: Zn-finger
72174;KOG0787;Dehydrogenase kinase
72204;KOG1802;RNA helicase nonsense mRNA reducing factor (pNORF1)
72434;KOG0550;Molecular chaperone (DnaJ superfamily)
72748;KOG0646;WD40 repeat protein
72972;KOG2986;Uncharacterized conserved protein
73352;KOG2992;Nucleolar GTPase/ATPase p130
73456;KOG1267;"Mitochondrial transcription termination factor, mTERF"
74137;KOG1875;Thyroid hormone receptor-associated coactivator complex component (TRAP170)
74342;KOG0379;Kelch repeat-containing proteins
74828;KOG2992;Nucleolar GTPase/ATPase p130
74969;KOG1600;Fatty acid desaturase
75072;KOG0057;"Mitochondrial Fe/S cluster exporter, ABC superfamily"
75085;KOG0260;"RNA polymerase II, large subunit"
75108;KOG0433;Isoleucyl-tRNA synthetase
75433;KOG1870;Ubiquitin C-terminal hydrolase
75621;KOG1154;Gamma-glutamyl kinase
75680;KOG1869;"Splicing coactivator SRm160/300, subunit SRm300"
75875;KOG4197;FOG: PPR repeat
76059;KOG4424;Predicted Rho/Rac guanine nucleotide exchange factor/faciogenital dysplasia protein 3
76108;KOG3696;Aspartyl beta-hydroxylase
76612;KOG1238;Glucose dehydrogenase/choline dehydrogenase/mandelonitrile lyase (GMC oxidoreductase family)
76903;KOG0671;LAMMER dual specificity kinases
77205;KOG2470;Similar to IMP-GMP specific 5'-nucleotidase
77339;KOG1721;FOG: Zn-finger
77442;KOG0975;"Branched chain aminotransferase BCAT1, pyridoxal phosphate enzymes type IV superfamily"
77561;KOG0061;"Transporter, ABC superfamily (Breast cancer resistance protein)"
77784;KOG4249;Uncharacterized conserved protein
78191;KOG4393;Predicted pseudouridylate synthase
78736;KOG1257;NADP+-dependent malic enzyme
78862;KOG1971;Lysyl hydroxylase
78898;KOG1134;Uncharacterized conserved protein
79126;KOG0619;FOG: Leucine rich repeat
79172;KOG2420;Phosphatidylserine decarboxylase
79330;KOG1816;Ubiquitin fusion-degradation protein
79406;KOG2123;Uncharacterized conserved protein
79628;KOG0768;Mitochondrial carrier protein PET8
79749;KOG0335;ATP-dependent RNA helicase
79783;KOG0619;FOG: Leucine rich repeat
80042;KOG1426;FOG: RCC1 domain
80071;KOG4342;Alpha-mannosidase
80093;KOG0053;Cystathionine beta-lyases/cystathionine gamma-synthases
80189;KOG3078;Adenylate kinase
80198;KOG1854;Mitochondrial inner membrane protein (mitofilin)
80456;KOG1402;Ornithine aminotransferase
80500;KOG0197;Tyrosine kinases
80604;KOG1870;Ubiquitin C-terminal hydrolase
81109;KOG4197;FOG: PPR repeat
81161;KOG1237;H+/oligopeptide symporter
81590;KOG0922;DEAH-box RNA helicase
81725;KOG1304;Amino acid transporters
81883;KOG0920;ATP-dependent RNA helicase A
81923;KOG0733;Nuclear AAA ATPase (VCP subfamily)
81996;KOG1643;Triosephosphate isomerase
82124;KOG1944;Peroxisomal membrane protein MPV17 and related proteins
82136;KOG0758;Mitochondrial carnitine-acylcarnitine carrier protein
82264;KOG3326;Uncharacterized conserved protein
82285;KOG2743;Cobalamin synthesis protein
82371;KOG1253;tRNA methyltransferase
```

```
82429;KOG2430;"Glycosyl hydrolase, family 47"
82715;KOG3756;Pinin (desmosome-associated protein)
82840;KOG2825;Putative arsenite-translocating ATPase
83113;KOG0613;Projectin/twitchin and related proteins
83145;KOG0244;Kinesin-like protein
83245;KOG1320;Serine protease
83567;KOG1635;Peptide methionine sulfoxide reductase
83971;KOG2633;Hismacro and SEC14 domain-containing proteins
84110;KOG4589;Cell division protein FtsJ
84171;KOG0787;Dehydrogenase kinase
84289;KOG1579;Homocysteine S-methyltransferase
84830;KOG1502;Flavonol reductase/cinnamoyl-CoA reductase
85031;KOG3822;Succinyl-CoA:alpha-ketoacid-CoA transferase
85128;KOG2844;Dimethylglycine dehydrogenase precursor
85156;KOG1426;FOG: RCC1 domain
85211;KOG1423;Ras-like GTPase ERA
85241;KOG0118;FOG: RRM domain
85347;KOG0235;Phosphoglycerate mutase
85359;KOG3433;Protein involved in meiotic recombination/predicted coiled-coil protein
85396;KOG0981;DNA topoisomerase I
85625;KOG3029;Glutathione S-transferase-related protein
85758;KOG0803;Predicted E3 ubiquitin ligase
85922;KOG1235;Predicted unusual protein kinase
85953;KOG1550;Extracellular protein SEL-1 and related proteins
86157;KOG0619;FOG: Leucine rich repeat
86167;KOG0619;FOG: Leucine rich repeat
86248;KOG4297;C-type lectin
86300;KOG2759;"Vacuolar H+-ATPase V1 sector, subunit H"
86301;KOG3033;Predicted PhzC/PhzF-type epimerase
86313;KOG2112;Lysophospholipase
86430;KOG0523;Transketolase
86678;KOG0043;"Uncharacterized conserved protein, contains DM10 domain"
86684;KOG0614;cGMP-dependent protein kinase
86864;KOG4067;Uncharacterized conserved protein
87020;KOG4203;Armadillo/beta-Catenin/plakoglobin
87126;KOG2260;"Cell division cycle 37 protein, CDC37"
87130;KOG2641;Predicted seven transmembrane receptor - rhodopsin family
87211;KOG1630;Growth hormone-induced protein and related proteins
87220;KOG3370;dUTPase
87421;KOG4267;Predicted membrane protein
87523;KOG3434;60S ribosomal protein L22
87597;KOG1361;Predicted hydrolase involved in interstrand cross-link repair
87611;KOG4297;C-type lectin
87653;KOG2299;Ribonuclease HI
87668;KOG2815;Mitochondrial/choloroplast ribosomal protein S15
87703;KOG2130;"Phosphatidylserine-specific receptor PtdSerR, contains JmjC domain"
87745;KOG3373;Glycine cleavage system H protein (lipoate-binding)
87747;KOG3869;Uncharacterized conserved protein
87791;KOG0752;Mitochondrial solute carrier protein
87811;KOG2965;Arginase
87827;KOG0692;Pentafunctional AROM protein
87832;KOG2793;"Putative N2,N2-dimethylguanosine tRNA methyltransferase"
87862;KOG1494;NAD-dependent malate dehydrogenase
88010;KOG0852;"Alkyl hydroperoxide reductase, thiol specific antioxidant and related enzymes"
88026;KOG0331;ATP-dependent RNA helicase
88161;KOG3710;EGL-Nine (EGLN) protein
88257;KOG1371;UDP-glucose 4-epimerase/UDP-sulfoquinovose synthase
88444;KOG1112;"Ribonucleotide reductase, alpha subunit"
88458;KOG1018;Cytosine deaminase FCY1 and related enzymes
88651;KOG4529;Uncharacterized conserved protein
88810;KOG3019;Predicted nucleoside-diphosphate sugar epimerase
88985;KOG1104;"Nuclear cap-binding complex, subunit NCBP1/CBP80"
89069;KOG1216;von Willebrand factor and related coagulation proteins
89294;KOG1576;Predicted oxidoreductase
89415;KOG3005;GIY-YIG type nuclease
89462;KOG4061;DMQ mono-oxygenase/Ubiquinone biosynthesis protein COQ7/CLK-1/CAT5
89482;KOG1336;Monodehydroascorbate/ferredoxin reductase
89530;KOG4197;FOG: PPR repeat
89553;KOG1618;Predicted phosphatase
89572;KOG2944;Glyoxalase
```

```
89685;KOG0688;Peptide chain release factor 1 (eRF1)
89744;KOG1158;NADP/FAD dependent oxidoreductase
89754;KOG2659;LisH motif-containing protein
89788;KOG4432;Uncharacterized NUDIX family hydrolase
89898;KOG1575;"Voltage-gated shaker-like K+ channel, subunit beta/KCNAB"
90117;KOG0672;Halotolerance protein HAL3 (contains flavoprotein domain)
90153;KOG1155;"Anaphase-promoting complex (APC), Cdc23 subunit"
90596;KOG1577;Aldo/keto reductase family proteins
90623;KOG2672;Lipoate synthase
90645;KOG0466;"Translation initiation factor 2, gamma subunit (eIF-2gamma
90719;KOG0106;Alternative splicing factor SRp55/B52/SRp75 (RRM superfamily)
90809;KOG0975;"Branched chain aminotransferase BCAT1, pyridoxal phosphate enzymes type IV superfamily"
91083;KOG1752;Glutaredoxin and related proteins
91191;KOG0504;FOG: Ankyrin repeat
91211;KOG2506;SpoU rRNA Methylase family protein
91248;KOG2413;Xaa-Pro aminopeptidase
91378;KOG3373;Glycine cleavage system H protein (lipoate-binding)
91444;KOG0089;Methylenetetrahydrofolate dehydrogenase/methylenetetrahydrofolate cyclohydrolase
91603;KOG1294;Apurinic/apyrimidinic endonuclease and related enzymes
91706;KOG1535;Predicted fumarylacetoacetate hydralase
91890;KOG2517;Ribulose kinase and related carbohydrate kinases
91920;KOG4129;Exopolyphosphatases and related proteins
91991;KOG0731;AAA+-type ATPase containing the peptidase M41 domain
92383;KOG0166;Karyopherin (importin) alpha
92455;KOG2004;Mitochondrial ATP-dependent protease PIM1/LON
92458;KOG1876;"Actin-related protein Arp2/3 complex, subunit ARPC4"
92505;KOG0440;Cell cycle-associated protein Mob1-1
92593;KOG3886;GTP-binding protein
92630;KOG1708;Mitochondrial/chloroplast ribosomal protein L24
92664;KOG2599;Pyridoxal/pyridoxine/pyridoxamine kinase
92689;KOG1422;"Intracellular Cl- channel CLIC, contains GST domain"
92714;KOG2831;ATP phosphoribosyltransferase
92788;KOG0235;Phosphoglycerate mutase
92883;KOG2157;Predicted tubulin-tyrosine ligase
125673;KOG2964;Arginase family protein
125849;KOG1408;WD40 repeat protein
126103;KOG3857;"Alcohol dehydrogenase, class IV"
126551;KOG1919;RNA pseudouridylate synthases
126606;KOG0383;Predicted helicase
126618;KOG2992;Nucleolar GTPase/ATPase p130
126868;KOG3713;Voltage-gated K+ channel KCNB/KCNC
127488;KOG4157;"beta-1,6-N-acetylglucosaminyltransferase, contains WSC domain"
127691;KOG1550;Extracellular protein SEL-1 and related proteins
127911;KOG4022;Dihydropteridine reductase DHPR/QDPR
128780;KOG0856;Predicted pilin-like transcription factor
128833;KOG2433;Uncharacterized conserved protein
129339;KOG2486;Predicted GTPase
129467;KOG2763;Acyl-CoA thioesterase
129469;KOG2108;3'-5' DNA helicase
129723;KOG1311;DHHC-type Zn-finger proteins
129857;KOG2450;Aldehyde dehydrogenase
131248;KOG0331;ATP-dependent RNA helicase
131295;KOG1562;Spermidine synthase
131414;KOG4471;Phosphatidylinositol 3-phosphate 3-phosphatase myotubularin MTM1
131519;KOG0800;FOG: Predicted E3 ubiquitin ligase
131686;KOG4283;"Transcription-coupled repair protein CSA, contains WD40 domain"
131801;KOG4197;FOG: PPR repeat
131828;KOG4197;FOG: PPR repeat
131834;KOG4197;FOG: PPR repeat
132297;KOG0160;Myosin class V heavy chain
132612;KOG0472;Leucine-rich repeat protein
132714;KOG1577;Aldo/keto reductase family proteins
132777;KOG2245;Poly(A) polymerase and related nucleotidyltransferases
133077;KOG4569;Predicted lipase
133282;KOG1944;Peroxisomal membrane protein MPV17 and related proteins
133569;KOG4701;Chitinase
133789;KOG2368;Hydroxymethylglutaryl-CoA lyase
133928;KOG2890;Predicted membrane protein
134601;KOG2992;Nucleolar GTPase/ATPase p130
134949;KOG1237;H+/oligopeptide symporter
```

```
135253;KOG3699;Cytoskeletal protein Adducin
135279;KOG4197;FOG: PPR repeat
135599;KOG4254;Phytoene desaturase
136009;KOG4197;FOG: PPR repeat
136010;KOG1369;Hexokinase
136247;KOG1337;N-methyltransferase
136549;KOG1591;Prolyl 4-hydroxylase alpha subunit
136602;KOG4308;LRR-containing protein
136620;KOG4177;Ankyrin
137282;KOG0254;Predicted transporter (major facilitator superfamily)
137816;KOG1623;Multitransmembrane protein
138357;KOG0346;RNA helicase
138616;KOG4064;Cysteine dioxygenase CDO1
138785;KOG1582;UDP-galactose transporter related protein
139095;KOG0029;Amine oxidase
140231;KOG1591;Prolyl 4-hydroxylase alpha subunit
140491;KOG2117;Uncharacterized conserved protein
140501;KOG2383;Predicted ATPase
140529;KOG1516;Carboxylesterase and related proteins
140732;KOG1605;"TFIIF-interacting CTD phosphatase, including NLI-interacting factor (involved in RNA polymerase II regulation)"
140819;KOG1285;"Beta, beta-carotene 15,15'-dioxygenase and related enzymes"
141176;KOG3945;Uncharacterized conserved protein
141182;KOG3637;"Vitronectin receptor, alpha subunit"
141401;KOG2323;Pyruvate kinase
141698;KOG3140;Predicted membrane protein
141739;KOG4759;Ribosome recycling factor
141750;KOG2546;"Abl interactor ABI-1, contains SH3 domain"
141802;KOG2002;TPR-containing nuclear phosphoprotein that regulates K(+) uptake
141934;KOG1308;Hsp70-interacting protein Hip/Transient component of progesterone receptor complexes and an Hsp70-binding protein
142494;KOG0776;Geranylgeranyl pyrophosphate synthase/Polyprenyl synthetase
142710;KOG1394;3-oxoacyl-(acyl-carrier-protein) synthase (I and II)
143322;KOG1540;Ubiquinone biosynthesis methyltransferase COQ5
143381;KOG2844;Dimethylglycine dehydrogenase precursor
144128;KOG2304;3-hydroxyacyl-CoA dehydrogenase
145261;KOG0327;"Translation initiation factor 4F, helicase subunit (eIF-4A) and related helicases"
145383;KOG1576;Predicted oxidoreductase
145781;KOG1176;Acyl-CoA synthetase
145832;KOG4563;Cell cycle-regulated histone H1-binding protein
145913;KOG4197;FOG: PPR repeat
146362;KOG1270;Methyltransferases
146395;KOG4694;Predicted membrane protein
146608;KOG4424;Predicted Rho/Rac guanine nucleotide exchange factor/faciogenital dysplasia protein 3
146751;KOG0890;"Protein kinase of the PI-3 kinase family involved in mitotic growth, DNA repair and meiotic recombination"
147422;KOG0186;Proline oxidase
86464;KOG0846;Mitochondrial/chloroplast ribosomal protein L15/L10
46835;KOG1504;Ornithine carbamoyltransferase OTC/ARG3
41839;KOG1360;5-aminolevulinate synthase
53042;KOG2403;"Succinate dehydrogenase, flavoprotein subunit"
49058;KOG1687;"NADH-ubiquinone oxidoreductase, NUFS7/PSST/20 kDa subunit"
146631;KOG3281;Mitochondrial F1-ATPase assembly protein
36033;KOG3996;Holocytochrome c synthase/heme-lyase
55826;KOG3196;"NADH:ubiquinone oxidoreductase, NDUFV2/24 kD subunit"
77004;KOG0759;Mitochondrial oxoglutarate/malate carrier proteins
55371;KOG0756;Mitochondrial tricarboxylate/dicarboxylate carrier proteins
87791;KOG0752;Mitochondrial solute carrier protein
50431;KOG0786;3-isopropylmalate dehydrogenase
56108;KOG0453;Aconitase/homoaconitase (aconitase superfamily)
72652;KOG0452;RNA-binding translational regulator IRP (aconitase superfamily)
146787;KOG3309;Ferredoxin
26430;KOG3309;Ferredoxin
43108;KOG3003;Molecular chaperone of the GrpE family
58085;KOG0399;Glutamate synthase
83963;KOG3192;Mitochondrial J-type chaperone
92952;KOG1799;Dihydropyrimidine dehydrogenase
55021;KOG3256;"NADH:ubiquinone oxidoreductase, NDUFS8/23 kDa subunit"
73761;KOG0057;"Mitochondrial Fe/S cluster exporter, ABC superfamily"
89209;KOG0430;Xanthine dehydrogenase
70052;KOG1800;Ferredoxin/adrenodoxin reductase
```

```
87672;KOG3413;"Mitochondrial matrix protein frataxin, involved in Fe/S protein biosynthesis"
48701;KOG3355;Mitochondrial sulfhydryl oxidase involved in the biogenesis of cytosolic Fe/S proteins
57894;KOG3355;Mitochondrial sulfhydryl oxidase involved in the biogenesis of cytosolic Fe/S proteins
22199;KOG0758;Mitochondrial carnitine-acylcarnitine carrier protein
22200;KOG0752;Mitochondrial solute carrier protein
33449;KOG0764;Mitochondrial FAD carrier protein
36786;KOG0758;Mitochondrial carnitine-acylcarnitine carrier protein
37107;KOG0759;Mitochondrial oxoglutarate/malate carrier proteins
37538;KOG0753;Mitochondrial fatty acid anion carrier protein/Uncoupling protein
38339;KOG0768;Mitochondrial carrier protein PET8
40277;KOG0749;Mitochondrial ADP/ATP carrier proteins
41058;KOG0767;Mitochondrial phosphate carrier protein
42229;KOG0749;Mitochondrial ADP/ATP carrier proteins
42249;KOG0767;Mitochondrial phosphate carrier protein
42974;KOG0767;Mitochondrial phosphate carrier protein
44307;KOG0759;Mitochondrial oxoglutarate/malate carrier proteins
46386;KOG0753;Mitochondrial fatty acid anion carrier protein/Uncoupling protein
46430;KOG0769;Predicted mitochondrial carrier protein
46877;KOG0749;Mitochondrial ADP/ATP carrier proteins
48941;KOG0762;Mitochondrial carrier protein
52610;KOG0750;Mitochondrial solute carrier protein
54313;KOG0762;Mitochondrial carrier protein
54478;KOG0756;Mitochondrial tricarboxylate/dicarboxylate carrier proteins
55259;KOG0760;Mitochondrial carrier protein MRS3/4
55371;KOG0756;Mitochondrial tricarboxylate/dicarboxylate carrier proteins
56244;KOG0749;Mitochondrial ADP/ATP carrier proteins
56295;KOG0752;Mitochondrial solute carrier protein
57139;KOG0768;Mitochondrial carrier protein PET8
65246;KOG0762;Mitochondrial carrier protein
68888;KOG0753;Mitochondrial fatty acid anion carrier protein/Uncoupling protein
71932;KOG0762;Mitochondrial carrier protein
74728;KOG0768;Mitochondrial carrier protein PET8
77004;KOG0759;Mitochondrial oxoglutarate/malate carrier proteins
78420;KOG0760;Mitochondrial carrier protein MRS3/4
79628;KOG0768;Mitochondrial carrier protein PET8
80457;KOG0750;Mitochondrial solute carrier protein
81815;KOG0758;Mitochondrial carnitine-acylcarnitine carrier protein
82136;KOG0758;Mitochondrial carnitine-acylcarnitine carrier protein
82212;KOG0752;Mitochondrial solute carrier protein
82286;KOG0770;Predicted mitochondrial carrier protein
82633;KOG0765;Predicted mitochondrial carrier protein
83248;KOG0765;Predicted mitochondrial carrier protein
85988;KOG0764;Mitochondrial FAD carrier protein
86023;KOG0754;Mitochondrial oxodicarboxylate carrier protein
87043;KOG0036;Predicted mitochondrial carrier protein
87389;KOG0036;Predicted mitochondrial carrier protein
87791;KOG0752;Mitochondrial solute carrier protein
88309;KOG0755;Mitochondrial oxaloacetate carrier protein
89039;KOG0758;Mitochondrial carnitine-acylcarnitine carrier protein
89615;KOG0758;Mitochondrial carnitine-acylcarnitine carrier protein
89988;KOG0756;Mitochondrial tricarboxylate/dicarboxylate carrier proteins
91130;KOG0755;Mitochondrial oxaloacetate carrier protein
91153;KOG0758;Mitochondrial carnitine-acylcarnitine carrier protein
92708;KOG0767;Mitochondrial phosphate carrier protein
126166;KOG0752;Mitochondrial solute carrier protein
126372;KOG0756;Mitochondrial tricarboxylate/dicarboxylate carrier proteins
128582;KOG0759;Mitochondrial oxoglutarate/malate carrier proteins
133560;KOG0762;Mitochondrial carrier protein
139012;KOG0036;Predicted mitochondrial carrier protein
139013;KOG0752;Mitochondrial solute carrier protein
141912;KOG0750;Mitochondrial solute carrier protein
143618;KOG0758;Mitochondrial carnitine-acylcarnitine carrier protein
146641;KOG0767;Mitochondrial phosphate carrier protein
41237;KOG1182;"Branched chain alpha-keto acid dehydrogenase complex, alpha subunit"
72023;KOG1684;Enoyl-CoA hydratase
84989;KOG1682;Enoyl-CoA isomerase
55169;KOG0369;Pyruvate carboxylase
41402;KOG2617;Citrate synthase
87802;KOG3296;"Translocase of outer mitochondrial membrane complex, subunit TOM40"
40390;KOG2663;"Acetolactate synthase, small subunit"
```

56780;KOG1276;Protoporphyrinogen oxidase
85182;KOG1276;Protoporphyrinogen oxidase
66428;KOG2799;"Succinyl-CoA synthetase, beta subunit"
92245;KOG2799;"Succinyl-CoA synthetase, beta subunit"
38920;KOG0465;Mitochondrial elongation factor
57828;KOG2792;Putative cytochrome C oxidase assembly protein
47709;KOG3496;Cytochrome c oxidase assembly protein/Cu2+ chaperone COX17
48203;KOG3477;"Putative cytochrome c oxidase, subunit COX19"
142491;KOG1380;Heme A farnesyltransferase
55834;KOG4763;Ubiquinol-cytochrome c reductase hinge protein
56213;KOG3052;Cytochrome c1
81100;KOG2873;Ubiquinol cytochrome c reductase assembly protein CBP3
90551;KOG3440;"Ubiquinol cytochrome c reductase, subunit QCR7"
128137;KOG2739;Leucine-rich acidic nuclear protein
89115;KOG1353;"F0F1-type ATP synthase, alpha subunit"
23686;KOG1353;"F0F1-type ATP synthase, alpha subunit"
48015;KOG1531;"F0F1-type ATP synthase, gamma subunit"
63670;KOG3468;"NADH:ubiquinone oxidoreductase, NDUFB7/B18 subunit"
141972;KOG3481;Uncharacterized conserved protein
91212;KOG4441;"Proteins containing BTB/POZ and Kelch domains, involved in regulatory/signal transduction processes"
141034;KOG4624;Uncharacterized conserved protein
86392;KOG4090;Uncharacterized conserved protein
89534;KOG1662;"Mitochondrial F1F0-ATP synthase, subunit OSCP/ATP5"
86319;KOG0743;AAA+-type ATPase
86666;KOG0743;AAA+-type ATPase
88586;KOG3225;"Mitochondrial import inner membrane translocase, subunit TIM22"
145867;KOG3456;"NADH:ubiquinone oxidoreductase, NDUFS6/13 kDa subunit"
88478;KOG3300;"NADH:ubiquinone oxidoreductase, B16.6 subunit/cell death-regulatory protein"
137303;KOG3426;"NADH:ubiquinone oxidoreductase, NDUFA6/B14 subunit"
51310;KOG3389;"NADH:ubiquinone oxidoreductase, NDUFS4/18 kDa subunit"
68792;KOG2901;Uncharacterized conserved protein
139677;KOG3363;Uncharacterized conserved nuclear protein
38746;KOG2853;Possible oxidoreductase
39504;KOG3257;Mitochondrial/chloroplast ribosomal protein L11
92176;KOG1870;Ubiquitin C-terminal hydrolase
38576;KOG4707;Mitochondrial/chloroplast ribosomal protein L20
89541;KOG4600;Mitochondrial ribosomal protein MRP7 (L2)
43603;KOG4612;Mitochondrial ribosomal protein L34
73769;KOG2602;Predicted cell surface protein homologous to bacterial outer membrane proteins
86440;KOG3489;"Mitochondrial import inner membrane translocase, subunit TIM8"
50483;KOG3479;"Mitochondrial import inner membrane translocase, subunit TIM9"
51633;KOG1652;"Mitochondrial import inner membrane translocase, subunit TIM17"
144223;KOG4836;Uncharacterized conserved protein
36491;KOG0723;Molecular chaperone (DnaJ superfamily)
87659;KOG3442;Uncharacterized conserved protein
73375;KOG2090;Metalloendopeptidase family - mitochondrial intermediate peptidase
68741;KOG2090;Metalloendopeptidase family - mitochondrial intermediate peptidase
34144;KOG0171;"Mitochondrial inner membrane protease, subunit IMP1"
38956;KOG0171;"Mitochondrial inner membrane protease, subunit IMP1"
34875;KOG1043;Ca2+-binding transmembrane protein LETM1/MRS7
30247;KOG1043;Ca2+-binding transmembrane protein LETM1/MRS7

APPENDIX C2: Mitochondrial targeted proteins in *Guillardia theta*

100123;KOG0633;Histidinol phosphate aminotransferase
100192;KOG0032;"Ca2+/calmodulin-dependent protein kinase, EF-Hand protein superfamily"
100288;KOG1550;Extracellular protein SEL-1 and related proteins
100311;KOG4034;Uncharacterized conserved protein NOF (Neighbor of FAU)
100415;KOG2301;"Voltage-gated Ca2+ channels, alpha1 subunits"
100418;KOG4707;Mitochondrial/chloroplast ribosomal protein L20
100553;KOG0619;FOG: Leucine rich repeat
100701;KOG0516;"Dystonin, GAS (Growth-arrest-specific protein), and related proteins"
100763;KOG3752;Ribonuclease H

```
100801;KOG1239;Inner membrane protein translocase involved in respiratory chain assembly
100929;KOG0504;FOG: Ankyrin repeat
101972;KOG0770;Predicted mitochondrial carrier protein
101987;KOG2873;Ubiquinol cytochrome c reductase assembly protein CBP3
102204;KOG0161;Myosin class II heavy chain
102743;KOG4701;Chitinase
102856;KOG1981;SOK1 kinase belonging to the STE20/SPS1/GC kinase family
102862;KOG1981;SOK1 kinase belonging to the STE20/SPS1/GC kinase family
102886;KOG2992;Nucleolar GTPase/ATPase p130
102887;KOG1981;SOK1 kinase belonging to the STE20/SPS1/GC kinase family
103155;KOG3015;F1-ATP synthase assembly protein
103258;KOG4149;Uncharacterized conserved protein
103622;KOG4415;Uncharacterized conserved protein
103878;KOG1236;Predicted unusual protein kinase
104108;KOG1652;"Mitochondrial import inner membrane translocase, subunit TIM17"
104175;KOG1550;Extracellular protein SEL-1 and related proteins
104196;KOG3599;Ca2+-modulated nonselective cation channel polycystin
104403;KOG2900;Biotin synthase
104486;KOG2725;Cytochrome oxidase assembly factor COX15
104499;KOG2369;Lecithin:cholesterol acyltransferase (LCAT)/Acyl-ceramide synthase
104682;KOG4701;Chitinase
104723;KOG0768;Mitochondrial carrier protein PET8
104989;KOG4297;C-type lectin
104991;KOG2958;Galactose-1-phosphate uridylyltransferase
105061;KOG4308;LRR-containing protein
105115;KOG1490;GTP-binding protein CRFG/NOG1 (ODN superfamily)
105173;KOG4456;"Inner centromere protein (INCENP), C-terminal domain"
105786;KOG1203;Predicted dehydrogenase
105791;KOG1003;Actin filament-coating protein tropomyosin
105896;KOG3676;Ca2+-permeable cation channel OSM-9 and related channels (OTRPC family)
106014;KOG2067;"Mitochondrial processing peptidase, alpha subunit"
106252;KOG4297;C-type lectin
106510;KOG1383;Glutamate decarboxylase/sphingosine phosphate lyase
106541;KOG0762;Mitochondrial carrier protein
106612;KOG2763;Acyl-CoA thioesterase
107050;KOG4797;Transcriptional regulator
107319;KOG2540;Cytochrome oxidase assembly factor COX11
107346;KOG3192;Mitochondrial J-type chaperone
107364;KOG1649;"SWI-SNF chromatin remodeling complex, Snf5 subunit"
107383;KOG0757;Mitochondrial carrier protein - Rim2p/Mrs12p
107429;KOG2624;Triglyceride lipase-cholesterol esterase
107458;KOG0331;ATP-dependent RNA helicase
107712;KOG2661;Peptidase family M48
107760;KOG2030;Predicted RNA-binding protein
107824;KOG0241;Kinesin-like protein
107923;KOG3324;"Mitochondrial import inner membrane translocase, subunit TIM23"
107986;KOG0768;Mitochondrial carrier protein PET8
108167;KOG1878;"Nuclear receptor coregulator SMRT/SMRTER, contains Myb-like domains"
108676;KOG4114;Cytochrome c oxidase assembly protein PET191
108697;KOG0766;Predicted mitochondrial carrier protein
108737;KOG4224;"Armadillo repeat protein VAC8 required for vacuole fusion, inheritance and cytosol-to-vacuole protein targeting"
109069;KOG3277;Uncharacterized conserved protein
109426;KOG1733;"Mitochondrial import inner membrane translocase, subunit TIM13"
109633;KOG4372;Predicted alpha/beta hydrolase
109652;KOG3610;Plexins (functional semaphorin receptors)
109736;KOG3028;"Translocase of outer mitochondrial membrane complex, subunit TOM37/Metaxin 1"
109800;KOG1217;Fibrillins and related proteins containing Ca2+-binding EGF-like domains
109998;KOG3528;FOG: PDZ domain
110008;KOG2602;Predicted cell surface protein homologous to bacterial outer membrane proteins
110014;KOG2435;Uncharacterized conserved protein
110516;KOG3954;"Electron transfer flavoprotein, alpha subunit"
110892;KOG1748;"Acyl carrier protein/NADH-ubiquinone oxidoreductase, NDUFAB1/SDAP subunit"
110990;KOG4287;Pectin acetylesterase and similar proteins
111251;KOG2602;Predicted cell surface protein homologous to bacterial outer membrane proteins
111297;KOG4415;Uncharacterized conserved protein
111335;KOG3710;EGL-Nine (EGLN) protein
111524;KOG1164;Casein kinase (serine/threonine/tyrosine protein kinase)
111617;KOG2301;"Voltage-gated Ca2+ channels, alpha1 subunits"
111813;KOG4197;FOG: PPR repeat
112016;KOG3029;Glutathione S-transferase-related protein
```

```
112464;KOG4042;"Dynactin subunit p27/WS-3, involved in transport of organelles along microtubules"
112468;KOG2360;Proliferation-associated nucleolar protein  (NOL1)
112512;KOG2689;Predicted ubiquitin regulatory protein
112529;KOG3382;"NADH:ubiquinone oxidoreductase, B17.2 subunit"
112701;KOG4297;C-type lectin
112917;KOG1591;Prolyl 4-hydroxylase alpha subunit
112959;KOG2301;"Voltage-gated Ca2+ channels, alpha1 subunits"
112991;KOG2382;Predicted alpha/beta hydrolase
113021;KOG1149;Glutamyl-tRNA synthetase (mitochondrial)
113138;KOG1550;Extracellular protein SEL-1 and related proteins
113578;KOG0137;Very-long-chain acyl-CoA dehydrogenase
113675;KOG0769;Predicted mitochondrial carrier protein
113677;KOG0752;Mitochondrial solute carrier protein
113712;KOG1012;"Ca2+-dependent lipid-binding protein CLB1/vesicle protein vp115/Granuphilin A, contains C2 domain"
113769;KOG2553;Pseudouridylate synthase
113813;KOG1337;N-methyltransferase
113864;KOG0382;Carbonic anhydrase
113913;KOG4495;"RNA polymerase II transcription elongation factor Elongin/SIII, subunit elongin B"
113965;KOG4589;Cell division protein FtsJ
114052;KOG0381;HMG box-containing protein
114270;KOG0516;"Dystonin, GAS (Growth-arrest-specific protein), and related proteins"
114287;KOG1282;Serine carboxypeptidases (lysosomal cathepsin A)
114482;KOG2266;"Chromatin-associated protein Dek and related proteins, contains SAP DNA binding domain"
114545;KOG4548;Mitochondrial ribosomal protein L17
114715;KOG3429;Predicted peptidyl-tRNA hydrolase
114770;KOG1982;"Nuclear 5'-3' exoribonuclease-interacting protein, Rai1p"
114808;KOG0053;Cystathionine beta-lyases/cystathionine gamma-synthases
115085;KOG2644;3'-phosphoadenosine 5'-phosphosulfate sulfotransferase (PAPS reductase)/FAD synthetase and related enzymes
115097;KOG3260;Calcyclin-binding protein CacyBP
115265;KOG0516;"Dystonin, GAS (Growth-arrest-specific protein), and related proteins"
115300;KOG2920;Predicted methyltransferase
115603;KOG0158;Cytochrome P450 CYP3/CYP5/CYP6/CYP9 subfamilies
115680;KOG4775;Uncharacterized protein SFI1 involved in G(2)-M transition
115839;KOG4415;Uncharacterized conserved protein
115864;KOG1477;SPRY domain-containing proteins
116071;KOG2792;Putative cytochrome C oxidase assembly protein
116227;KOG0768;Mitochondrial carrier protein PET8
116783;KOG1236;Predicted unusual protein kinase
116850;KOG0769;Predicted mitochondrial carrier protein
116893;KOG4701;Chitinase
117050;KOG4600;Mitochondrial ribosomal protein MRP7 (L2)
117185;KOG1208;Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases)
117364;KOG1239;Inner membrane protein translocase involved in respiratory chain assembly
117501;KOG1812;Predicted E3 ubiquitin ligase
117769;KOG3281;Mitochondrial F1-ATPase assembly protein
118067;KOG0619;FOG: Leucine rich repeat
118073;KOG2683;Sirtuin 4 and related class II sirtuins (SIR2 family)
118094;KOG3413;"Mitochondrial matrix protein frataxin, involved in Fe/S protein biosynthesis"
118098;KOG0762;Mitochondrial carrier protein
118542;KOG3225;"Mitochondrial import inner membrane translocase, subunit TIM22"
118650;KOG0516;"Dystonin, GAS (Growth-arrest-specific protein), and related proteins"
118715;KOG0266;WD40 repeat-containing protein
118738;KOG4111;"Translocase of outer mitochondrial membrane complex, subunit TOM22"
119889;KOG2726;Mitochondrial polypeptide chain release factor
119893;KOG1420;"Ca2+-activated K+ channel Slowpoke, alpha subunit"
120094;KOG1563;"Mitochondrial protein Surfeit 1/SURF1/SHY1, required for expression of cytochrome oxidase"
120149;KOG3969;Uncharacterized conserved protein
120208;KOG2458;"Endoplasmic reticulum protein EP58, contains filamin rod domain and KDEL motif"
120283;KOG3430;Dynein light chain type 1
120362;KOG1343;"Histone deacetylase complex, catalytic component HDA1"
120492;KOG1067;Predicted RNA-binding polyribonucleotide nucleotidyltransferase
120626;KOG0770;Predicted mitochondrial carrier protein
120774;KOG1944;Peroxisomal membrane protein MPV17 and related proteins
120923;KOG0161;Myosin class II heavy chain
121103;KOG2152;Sister chromatid cohesion protein
121162;KOG0516;"Dystonin, GAS (Growth-arrest-specific protein), and related proteins"
121170;KOG1596;Fibrillarin and related nucleolar RNA-binding proteins
121390;KOG3803;Transcription factor containing C2HC type Zn finger
121394;KOG3012;Uncharacterized conserved protein
121468;KOG4297;C-type lectin
```

```
121646;KOG0161;Myosin class II heavy chain
121795;KOG0753;Mitochondrial fatty acid anion carrier protein/Uncoupling protein
122273;KOG0557;Dihydrolipoamide acetyltransferase
122609;KOG3212;Uncharacterized conserved protein related to IojAP
123019;KOG0597;Serine-threonine protein kinase FUSED
131776;KOG0027;Calmodulin and related proteins (EF-Hand superfamily)
131845;KOG1484;Putative Zn2+ transporter MSC2 (cation diffusion facilitator superfamily)
131936;KOG2237;Predicted serine protease
132133;KOG0513;Ca2+-independent phospholipase A2
132251;KOG0933;"Structural maintenance of chromosome protein 2 (chromosome condensation complex Condensin, subunit E)"
132543;KOG0161;Myosin class II heavy chain
132731;KOG2301;"Voltage-gated Ca2+ channels, alpha1 subunits"
133007;KOG4626;O-linked N-acetylglucosamine transferase OGT
133395;KOG0751;Mitochondrial aspartate/glutamate carrier protein Aralar/Citrin (contains EF-hand Ca2+-binding domains)
133463;KOG2901;Uncharacterized conserved protein
133540;KOG0696;Serine/threonine protein kinase
133553;KOG1840;Kinesin light chain
134392;KOG2792;Putative cytochrome C oxidase assembly protein
135045;KOG2344;Exocyst component protein and related proteins
135376;KOG2615;Permease of the major facilitator superfamily
135385;KOG1221;Acyl-CoA reductase
135651;KOG2992;Nucleolar GTPase/ATPase p130
135750;KOG2301;"Voltage-gated Ca2+ channels, alpha1 subunits"
135998;KOG4172;Predicted E3 ubiquitin ligase
136206;KOG0856;Predicted pilin-like transcription factor
136509;KOG4701;Chitinase
137267;KOG0274;Cdc4 and related F-box and WD-40 proteins
137494;KOG1426;FOG: RCC1 domain
137512;KOG1015;"Transcription regulator XNP/ATRX, DEAD-box superfamily"
137702;KOG0760;Mitochondrial carrier protein MRS3/4
137950;KOG0768;Mitochondrial carrier protein PET8
138039;KOG2580;"Mitochondrial import inner membrane translocase, subunit TIM44"
138058;KOG0507;CASK-interacting adaptor protein (caskin) and related proteins with ankyrin repeats and SAM domain
138126;KOG4100;Uncharacterized conserved protein
138301;KOG0768;Mitochondrial carrier protein PET8
138485;KOG3935;Predicted glycerate kinase
139208;KOG0516;"Dystonin, GAS (Growth-arrest-specific protein), and related proteins"
139735;KOG4308;LRR-containing protein
140012;KOG1187;Serine/threonine protein kinase
140222;KOG4161;Methyl-CpG binding transcription regulators
140223;KOG4235;Mitochondrial thymidine kinase 2/deoxyguanosine kinase
140265;KOG2058;Ypt/Rab GTPase activating protein
140534;KOG4821;Predicted Na+-dependent cotransporter
140568;KOG0239;Kinesin (KAR3 subfamily)
140581;KOG0498;"K+-channel ERG and related proteins, contain PAS/PAC sensor domain"
140689;KOG4771;Nucleolar protein (NOP16) involved in 60S ribosomal subunit biogenesis
141004;KOG0752;Mitochondrial solute carrier protein
141727;KOG0450;"2-oxoglutarate dehydrogenase, E1 subunit"
141811;KOG2726;Mitochondrial polypeptide chain release factor
142044;KOG0516;"Dystonin, GAS (Growth-arrest-specific protein), and related proteins"
142613;KOG4162;Predicted calmodulin-binding protein
142790;KOG2266;"Chromatin-associated protein Dek and related proteins, contains SAP DNA binding domain"
143556;KOG1878;"Nuclear receptor coregulator SMRT/SMRTER, contains Myb-like domains"
143567;KOG1236;Predicted unusual protein kinase
143999;KOG0767;Mitochondrial phosphate carrier protein
144004;KOG1652;"Mitochondrial import inner membrane translocase, subunit TIM17"
144409;KOG0632;Phytochelatin synthase
144412;KOG1971;Lysyl hydroxylase
144527;KOG1536;Biotin holocarboxylase synthetase/biotin-protein ligase
144780;KOG0401;"Translation initiation factor 4F, ribosome/mRNA-bridging subunit (eIF-4G)"
144911;KOG3614;Ca2+/Mg2+-permeable cation channels (LTRPC family)
145230;KOG1516;Carboxylesterase and related proteins
145238;KOG4674;Uncharacterized conserved coiled-coil protein
145404;KOG3528;FOG: PDZ domain
145603;KOG1346;Programmed cell death 8 (apoptosis-inducing factor)
146129;KOG3300;"NADH:ubiquinone oxidoreductase, B16.6 subunit/cell death-regulatory protein"
146135;KOG4415;Uncharacterized conserved protein
147022;KOG0762;Mitochondrial carrier protein
147042;KOG0759;Mitochondrial oxoglutarate/malate carrier proteins
147086;KOG4364;Chromatin assembly factor-I
```

```
147559;KOG0725;Reductases with broad range of substrate specificities
148122;KOG0927;Predicted transporter (ABC superfamily)
148130;KOG4652;HORMA domain
148131;KOG0027;Calmodulin and related proteins (EF-Hand superfamily)
148797;KOG3355;Mitochondrial sulfhydryl oxidase involved in the biogenesis of cytosolic Fe/S proteins
149131;KOG1144;Translation initiation factor 5B (eIF-5B)
149782;KOG0540;"3-Methylcrotonyl-CoA carboxylase, non-biotin containing subunit/Acetyl-CoA carboxylase carboxyl transferase,
subunit beta"
149798;KOG0725;Reductases with broad range of substrate specificities
149884;KOG2860;"Uncharacterized conserved protein, contains TraB domain"
149920;KOG0715;Molecular chaperone (DnaJ superfamily)
149927;KOG3057;"Cytochrome c oxidase, subunit VIb/COX12"
150228;KOG1662;"Mitochondrial F1F0-ATP synthase, subunit OSCP/ATP5"
150266;KOG3435;Mitochondrial/chloroplast ribosomal protein L54/L37
150356;KOG1154;Gamma-glutamyl kinase
150375;KOG3480;"Mitochondrial import inner membrane translocase, subunits TIM10/TIM12"
150554;KOG4411;Phytoene/squalene synthetase
150588;KOG3426;"NADH:ubiquinone oxidoreductase, NDUFA6/B14 subunit"
150683;KOG4367;Predicted Zn-finger protein
150764;KOG0053;Cystathionine beta-lyases/cystathionine gamma-synthases
150902;KOG2832;"TFIIF-interacting CTD phosphatase, including NLI-interacting factor (involved in RNA polymerase II
regulation)"
150973;KOG2495;NADH-dehydrogenase (ubiquinone)
151115;KOG1758;"Mitochondrial F1F0-ATP synthase, subunit delta/ATP16"
151185;KOG0198;MEKK and related serine/threonine protein kinases
151242;KOG3196;"NADH:ubiquinone oxidoreductase, NDUFV2/24 kD subunit"
151328;KOG4195;Transient receptor potential-related channel 7
151400;KOG3326;Uncharacterized conserved protein
152067;KOG3003;Molecular chaperone of the GrpE family
152490;KOG3442;Uncharacterized conserved protein
152876;KOG1119;Mitochondrial Fe-S cluster biosynthesis protein ISA2 (contains a HesB-like domain)
152983;KOG3121;"Dynactin, subunit p25"
152990;KOG0757;Mitochondrial carrier protein - Rim2p/Mrs12p
153024;KOG4722;Zn-finger protein
153028;KOG0745;Putative ATP-dependent Clp-type protease (AAA+ ATPase superfamily)
153126;KOG0082;G-protein alpha subunit (small G protein superfamily)
153155;KOG0743;AAA+-type ATPase
153600;KOG4009;"NADH-ubiquinone oxidoreductase, subunit NDUFB10/PDSW"
153736;KOG1641;Mitochondrial chaperonin
154023;KOG0868;Glutathione S-transferase
154060;KOG2770;Aminomethyl transferase
154135;KOG2862;Alanine-glyoxylate aminotransferase AGT1
154188;KOG1494;NAD-dependent malate dehydrogenase
154369;KOG2815;Mitochondrial/choloroplast ribosomal protein S15
154373;KOG0752;Mitochondrial solute carrier protein
154481;KOG0751;Mitochondrial aspartate/glutamate carrier protein Aralar/Citrin (contains EF-hand Ca2+-binding domains)
154527;KOG0813;Glyoxylase
154876;KOG1120;Fe-S cluster biosynthesis protein ISA1 (contains a HesB-like domain)
155008;KOG4750;Serine O-acetyltransferase
155205;KOG3466;"NADH:ubiquinone oxidoreductase, NDUFB9/B22 subunit"
155227;KOG4090;Uncharacterized conserved protein
155239;KOG0187;40S ribosomal protein S17
155357;KOG0787;Dehydrogenase kinase
156001;KOG1182;"Branched chain alpha-keto acid dehydrogenase complex, alpha subunit"
156042;KOG2502;Tub family proteins
156324;KOG3440;"Ubiquinol cytochrome c reductase, subunit QCR7"
156592;KOG1698;Mitochondrial/chloroplast ribosomal protein L19
156771;KOG2250;Glutamate/leucine/phenylalanine/valine dehydrogenases
157086;KOG1321;Protoheme ferro-lyase (ferrochelatase)
157094;KOG1696;60s ribosomal protein L19
157347;KOG0759;Mitochondrial oxoglutarate/malate carrier proteins
157415;KOG2524;"Cobyrinic acid a,c-diamide synthase"
157808;KOG1043;Ca2+-binding transmembrane protein LETM1/MRS7
158019;KOG1257;NADP+-dependent malic enzyme
158030;KOG0752;Mitochondrial solute carrier protein
158122;KOG1752;Glutaredoxin and related proteins
158129;KOG3280;Mitochondrial/chloroplast ribosomal protein L17
158239;KOG2311;NAD/FAD-utilizing protein possibly involved in translation
158243;KOG0749;Mitochondrial ADP/ATP carrier proteins
158274;KOG2358;NifU-like domain-containing proteins
```

```
158561;KOG3309;Ferredoxin
158965;KOG3257;Mitochondrial/chloroplast ribosomal protein L11
159012;KOG3049;"Succinate dehydrogenase, Fe-S protein subunit"
159138;KOG2403;"Succinate dehydrogenase, flavoprotein subunit"
159352;KOG0857;60s ribosomal protein L10
159505;KOG1159;NADP-dependent flavoprotein reductase
159564;KOG3886;GTP-binding protein
159585;KOG0959;"N-arginine dibasic convertase NRD1 and related Zn2+-dependent endopeptidases, insulinase superfamily"
159725;KOG2195;Transferrin receptor and related proteins containing the protease-associated (PA) domain
159960;KOG1395;Tryptophan synthase beta chain
160137;KOG2635;Medium subunit of clathrin adaptor complex
160293;KOG1389;3-oxoacyl CoA thiolase
160714;KOG3352;"Cytochrome c oxidase, subunit Vb/COX4"
160877;KOG3366;"Mitochondrial F1F0-ATP synthase, subunit d/ATP7"
160884;KOG1426;FOG: RCC1 domain
160930;KOG1236;Predicted unusual protein kinase
160957;KOG2865;"NADH:ubiquinone oxidoreductase, NDUFA9/39kDa subunit"
161126;KOG0621;Phospholipid scramblase
161318;KOG1401;Acetylornithine aminotransferase
161512;KOG1840;Kinesin light chain
161569;KOG0552;FKBP-type peptidyl-prolyl cis-trans isomerase
161614;KOG4809;Rab6 GTPase-interacting protein involved in endosome-to-TGN transport
161810;KOG1683;Hydroxyacyl-CoA dehydrogenase/enoyl-CoA hydratase
162358;KOG4609;Predicted phosphoglycerate mutase
162408;KOG0758;Mitochondrial carnitine-acylcarnitine carrier protein
162598;KOG0755;Mitochondrial oxaloacetate carrier protein
162662;KOG1196;Predicted NAD-dependent oxidoreductase
162689;KOG0764;Mitochondrial FAD carrier protein
162753;KOG1399;Flavin-containing monooxygenase
162858;KOG1012;"Ca2+-dependent lipid-binding protein CLB1/vesicle protein vp115/Granuphilin A, contains C2 domain"
163170;KOG2352;Predicted spermine/spermidine synthase
163218;KOG0915;Uncharacterized conserved protein
163859;KOG1191;Mitochondrial GTPase
163898;KOG0768;Mitochondrial carrier protein PET8
163902;KOG0167;FOG: Armadillo/beta-catenin-like repeats
164178;KOG3052;Cytochrome c1
164186;KOG1531;"F0F1-type ATP synthase, gamma subunit"
164335;KOG1071;"Mitochondrial translation elongation factor EF-Tsmt, catalyzes nucleotide exchange on EF-Tumt"
164581;KOG4297;C-type lectin
164588;KOG4674;Uncharacterized conserved coiled-coil protein
164872;KOG0700;Protein phosphatase 2C/pyruvate dehydrogenase (lipoamide) phosphatase
164927;KOG1647;"Vacuolar H+-ATPase V1 sector, subunit D"
165205;KOG0450;"2-oxoglutarate dehydrogenase, E1 subunit"
165245;KOG0498;"K+-channel ERG and related proteins, contain PAS/PAC sensor domain"
165280;KOG2486;Predicted GTPase
165716;KOG3456;"NADH:ubiquinone oxidoreductase, NDUFS6/13 kDa subunit"
165783;KOG3716;Carnitine O-acyltransferase CPTI
165899;KOG0399;Glutamate synthase
165999;KOG2830;Protein phosphatase 2A-associated protein
166205;KOG1043;Ca2+-binding transmembrane protein LETM1/MRS7
166234;KOG2992;Nucleolar GTPase/ATPase p130
166368;KOG2747;Histone acetyltransferase (MYST family)
166503;KOG3296;"Translocase of outer mitochondrial membrane complex, subunit TOM40"
166657;KOG0950;"DNA polymerase theta/eta, DEAD-box superfamily"
166830;KOG1624;Mitochondrial/chloroplast ribosomal protein L4
166834;KOG4297;C-type lectin
166939;KOG0019;Molecular chaperone (HSP90 family)
40411;KOG3801;Uncharacterized conserved protein BCN92
42578;KOG3355;Mitochondrial sulfhydryl oxidase involved in the biogenesis of cytosolic Fe/S proteins
45511;KOG0764;Mitochondrial FAD carrier protein
46987;KOG0438;Mitochondrial/chloroplast ribosomal protein L2
49807;KOG3860;Acyltransferase required for palmitoylation of Hedgehog (Hh) family of secreted signaling proteins
50187;KOG0765;Predicted mitochondrial carrier protein
53205;KOG3355;Mitochondrial sulfhydryl oxidase involved in the biogenesis of cytosolic Fe/S proteins
58369;KOG0754;Mitochondrial oxodicarboxylate carrier protein
60979;KOG3477;"Putative cytochrome c oxidase, subunit COX19"
62892;KOG1203;Predicted dehydrogenase
62947;KOG0751;Mitochondrial aspartate/glutamate carrier protein Aralar/Citrin (contains EF-hand Ca2+-binding domains)
63299;KOG0767;Mitochondrial phosphate carrier protein
63916;KOG0057;"Mitochondrial Fe/S cluster exporter, ABC superfamily"
```

```
64866;KOG1350;"F0F1-type ATP synthase, beta subunit"
65244;KOG4442;"Clathrin coat binding protein/Huntingtin interacting protein HIP1, involved in regulation of endocytosis"
65435;KOG3996;Holocytochrome c synthase/heme-lyase
65536;KOG0753;Mitochondrial fatty acid anion carrier protein/Uncoupling protein
65587;KOG1294;Apurinic/apyrimidinic endonuclease and related enzymes
66106;KOG0765;Predicted mitochondrial carrier protein
66600;KOG1442;GDP-fucose transporter
67785;KOG1671;"Ubiquinol cytochrome c reductase, subunit RIP1"
68305;KOG1898;"Splicing factor 3b, subunit 3"
68495;KOG0764;Mitochondrial FAD carrier protein
68742;KOG1800;Ferredoxin/adrenodoxin reductase
68815;KOG1235;Predicted unusual protein kinase
69150;KOG3489;"Mitochondrial import inner membrane translocase, subunit TIM8"
69155;KOG2644;3'-phosphoadenosine 5'-phosphosulfate sulfotransferase (PAPS reductase)/FAD synthetase and related enzymes
69464;KOG0060;"Long-chain acyl-CoA transporter, ABC superfamily (involved in peroxisome organization and biogenesis)"
69596;KOG0285;Pleiotropic regulator 1
70145;KOG1944;Peroxisomal membrane protein MPV17 and related proteins
70371;KOG0192;Tyrosine kinase specific for activated (GTP-bound) p21cdc42Hs
70571;KOG2551;Phospholipase/carboxyhydrolase
70823;KOG1032;"Uncharacterized conserved protein, contains GRAM domain"
70870;KOG3614;Ca2+/Mg2+-permeable cation channels (LTRPC family)
72009;KOG3363;Uncharacterized conserved nuclear protein
72211;KOG3419;Mitochondrial/chloroplast ribosomal protein S16
72577;KOG0765;Predicted mitochondrial carrier protein
72919;KOG3309;Ferredoxin
73289;KOG3996;Holocytochrome c synthase/heme-lyase
73476;KOG2738;Putative methionine aminopeptidase
73487;KOG0760;Mitochondrial carrier protein MRS3/4
73833;KOG3479;"Mitochondrial import inner membrane translocase, subunit TIM9"
73866;KOG1380;Heme A farnesyltransferase
74106;KOG1947;"Leucine rich repeat proteins, some proteins contain F-box"
74274;KOG0765;Predicted mitochondrial carrier protein
74509;KOG0743;AAA+-type ATPase
74767;KOG0768;Mitochondrial carrier protein PET8
74910;KOG0950;"DNA polymerase theta/eta, DEAD-box superfamily"
75326;KOG0465;Mitochondrial elongation factor
75461;KOG2090;Metalloendopeptidase family - mitochondrial intermediate peptidase
75525;KOG1549;Cysteine desulfurase NFS1
75595;KOG2556;Leishmanolysin-like peptidase (Peptidase M8 family)
76084;KOG0758;Mitochondrial carnitine-acylcarnitine carrier protein
76661;KOG0266;WD40 repeat-containing protein
77318;KOG0852;"Alkyl hydroperoxide reductase, thiol specific antioxidant and related enzymes"
77427;KOG0768;Mitochondrial carrier protein PET8
77466;KOG0759;Mitochondrial oxoglutarate/malate carrier proteins
77590;KOG1593;Asparaginase
78640;KOG2844;Dimethylglycine dehydrogenase precursor
79059;KOG0102;"Molecular chaperones mortalin/PBP74/GRP75, HSP70 superfamily"
79815;KOG0381;HMG box-containing protein
80286;KOG0723;Molecular chaperone (DnaJ superfamily)
81033;KOG0619;FOG: Leucine rich repeat
81165;KOG4347;GTPase-activating protein VRP
81659;KOG0758;Mitochondrial carnitine-acylcarnitine carrier protein
82398;KOG0759;Mitochondrial oxoglutarate/malate carrier proteins
83445;KOG1023;"Natriuretic peptide receptor, guanylate cyclase"
83901;KOG0752;Mitochondrial solute carrier protein
83930;KOG2249;3'-5' exonuclease
83986;KOG0960;"Mitochondrial processing peptidase, beta subunit, and related enzymes (insulinase superfamily)"
84137;KOG2965;Arginase
84773;KOG0763;Mitochondrial ornithine transporter
85396;KOG0761;Mitochondrial carrier protein CGI-69
85736;KOG0460;Mitochondrial translation elongation factor Tu
85780;KOG1282;Serine carboxypeptidases (lysosomal cathepsin A)
85967;KOG1569;50S ribosomal protein L1
86282;KOG0975;"Branched chain aminotransferase BCAT1, pyridoxal phosphate enzymes type IV superfamily"
86293;KOG2302;"T-type voltage-gated Ca2+ channel, pore-forming alpha1I subunit"
86388;KOG3581;Creatine kinases
86502;KOG1402;Ornithine aminotransferase
86537;KOG0743;AAA+-type ATPase
86641;KOG1568;"Mitochondrial inner membrane protease, subunit IMP2"
87185;KOG3980;RNA 3'-terminal phosphate cyclase
```

```
87417;KOG2617;Citrate synthase
87533;KOG2599;Pyridoxal/pyridoxine/pyridoxamine kinase
87575;KOG0307;"Vesicle coat complex COPII, subunit SEC31"
87979;KOG0725;Reductases with broad range of substrate specificities
88190;KOG1530;Rhodanese-related sulfurtransferase
88345;KOG1680;Enoyl-CoA hydratase
88558;KOG0171;"Mitochondrial inner membrane protease, subunit IMP1"
88722;KOG2743;Cobalamin synthesis protein
88734;KOG1197;Predicted quinone oxidoreductase
89568;KOG2658;"NADH:ubiquinone oxidoreductase, NDUFV1/51kDa subunit"
89623;KOG4197;FOG: PPR repeat
89880;KOG0235;Phosphoglycerate mutase
90066;KOG0137;Very-long-chain acyl-CoA dehydrogenase
90757;KOG0767;Mitochondrial phosphate carrier protein
91152;KOG4763;Ubiquinol-cytochrome c reductase hinge protein
91782;KOG0876;Manganese superoxide dismutase
91827;KOG3309;Ferredoxin
92053;KOG3361;Iron binding protein involved in Fe-S cluster formation
92151;KOG0707;Guanylate kinase
92563;KOG3857;"Alcohol dehydrogenase, class IV"
93546;KOG0831;Acyl-CoA:diacylglycerol acyltransferase (DGAT)
93556;KOG4308;LRR-containing protein
94019;KOG2451;Aldehyde dehydrogenase
94276;KOG0558;Dihydrolipoamide transacylase (alpha-keto acid dehydrogenase E2 subunit)
94796;KOG3331;Mitochondrial/chloroplast ribosomal protein L4/L29
95428;KOG0225;"Pyruvate dehydrogenase E1, alpha subunit"
95515;KOG2542;Uncharacterized conserved protein (YdiU family)
95798;KOG0454;3-isopropylmalate dehydratase (aconitase superfamily)
96015;KOG0749;Mitochondrial ADP/ATP carrier proteins
96154;KOG1231;Proteins containing the FAD binding domain
96730;KOG0911;Glutaredoxin-related protein
97273;KOG0768;Mitochondrial carrier protein PET8
97605;KOG1261;Malate synthase
97799;KOG1411;Aspartate aminotransferase/Glutamic oxaloacetic transaminase AAT1/GOT2
97833;KOG0370;"Multifunctional pyrimidine synthesis protein CAD (includes carbamoyl-phophate synthetase, aspartate transcarbamylase, and glutamine amidotransferase)"
97918;KOG3468;"NADH:ubiquinone oxidoreductase, NDUFB7/B18 subunit"
98165;KOG1550;Extracellular protein SEL-1 and related proteins
98427;KOG0141;Isovaleryl-CoA dehydrogenase
98434;KOG4694;Predicted membrane protein
98506;KOG3121;"Dynactin, subunit p25"
98704;KOG1575;"Voltage-gated shaker-like K+ channel, subunit beta/KCNAB"
99024;KOG1413;N-acetylglucosaminyltransferase I
99154;KOG0891;DNA-dependent protein kinase
99492;KOG2216;Conserved coiled/coiled coil protein
99515;KOG0487;"Transcription factor Abd-B, contains HOX domain"
99773;KOG1695;Glutathione S-transferase
99859;KOG0613;Projectin/twitchin and related proteins
99904;KOG4197;FOG: PPR repeat
```

APPENDIX D: Copyrights and Permissions

Portions of the following papers were included in some of the chapters. As a contributing author I retain the right to use these papers in various ways as indicated below without requiring permission.

Chapters 2 and 3

Curtis, B. A., G. Tanifuji, et al. (2012). "Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs ." Nature **492**: 59-65.

http://www.nature.com/authors/policies/license.html
"NPG author licence policy
This publishers' policy applies to all journals published by the Nature Publishing Group (NPG), including the Nature journals.
NPG does not require authors of original (primary) research papers to assign copyright of their published contributions. Authors grant NPG an exclusive licence to publish, in return for which they can reuse their papers in their future printed work without first requiring permission from the publisher of the journal. "

Also,
http://www.nature.com/reprints/permission-requests.html
"If you are the author of this content (or his/her designated agent) please read the following. Since 2003, ownership of copyright in in original research articles remains with the Authors*, and provided that, when reproducing the Contribution or extracts from it, the Authors acknowledge first and reference publication in the Journal, the Authors retain the following non-exclusive rights:

1. To reproduce the Contribution in whole or in part in any printed volume (book or thesis) of which they are the author(s)."

Chapter 4

Curtis, B. A. and J. M. Archibald (2010). "A spliceosomal intron of mitochondrial DNA origin." Curr Biol **20**(21): R919-920.

http://www.cell.com/current-biology/authors
"Authors' rights
As an author you (or your employer or institution) may do the following:

- include the article in full or in part in a thesis or dissertation (provided that this is not to be published commercially);"

Also,
http://www.elsevier.com/wps/find/authorsview.authors/rights
"How authors can use their own journal articles

Authors publishing in Elsevier journals have wide rights to use their works for teaching and scholarly purposes without needing to seek permission."

REFERENCES

Abeliovich, H. (2007). "Mitophagy: the life-or-death dichotomy includes yeast." Autophagy **3**(3): 275-277.

Adams, K. L., D. O. Daley, et al. (2000). "Repeated, recent and diverse transfers of a mitochondrial gene to the nucleus in flowering plants." Nature **408**(6810): 354-357.

Adams, K. L. and J. D. Palmer (2003). "Evolution of mitochondrial gene content: gene loss and transfer to the nucleus." Mol Phylogenet Evol **29**(3): 380-395.

Adl, S. M., A. G. Simpson, et al. (2005). "The new higher level classification of eukaryotes with emphasis on the taxonomy of protists." J Eukaryot Microbiol **52**(5): 399-451.

Allen, J. F. and J. A. Raven (1996). "Free-radical-induced mutation vs redox regulation: costs and benefits of genes in organelles." J Mol Evol **42**(5): 482-492.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

Andersson, S. G., O. Karlberg, et al. (2003). "On the origin of mitochondria: a genomics perspective." Philos Trans R Soc Lond B Biol Sci **358**(1429): 165-177; discussion 177-169.

Andersson, S. G., A. Zomorodipour, et al. (1998). "The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria." Nature **396**(6707): 133-140.

Ang, S. K. and H. Lu (2009). "Deciphering structural and functional roles of individual disulfide bonds of the mitochondrial sulfhydryl oxidase Erv1p." J Biol Chem **284**(42): 28754-28761.

Arai, Y., M. Hayashi, et al. (2008). "Proteomic analysis of highly purified peroxisomes from etiolated soybean cotyledons." Plant Cell Physiol **49**(4): 526-539.

Archibald, J. M. (2007). "Nucleomorph genomes: structure, function, origin and evolution." Bioessays **29**(4): 392-402.

Archibald, J. M., M. B. Rogers, et al. (2003). "Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga *Bigelowiella natans*." Proc Natl Acad Sci U S A **100**(13): 7678-7683.

Armbrust, E. V., J. A. Berges, et al. (2004). "The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism." Science **306**(5693): 79-86.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-29.

Atteia, A., A. Adrait, et al. (2009). "A proteomic survey of *Chlamydomonas reinhardtii* mitochondria sheds new light on the metabolic plasticity of the organelle and on the nature of the alpha-proteobacterial mitochondrial ancestor." Mol Biol Evol **26**(7): 1533-1548.

Bachvaroff, T. R. and A. R. Place (2008). "From stop to start: tandem gene arrangement, copy number and trans-splicing sites in the dinoflagellate *Amphidinium carterae*." PLoS One **3**(8): e2929.

Bahaji, A., M. Ovecka, et al. (2011). "Dual targeting to mitochondria and plastids of AtBT1 and ZmBT1, two members of the mitochondrial carrier family." Plant Cell Physiol **52**(4): 597-609.

Bairoch, A. (2000). "The ENZYME database in 2000." Nucleic Acids Res **28**(1): 304-305.

Bairoch, A., R. Apweiler, et al. (2005). "The Universal Protein Resource (UniProt)." Nucleic Acids Res **33**(Database issue): D154-159.

Bannai, H., Y. Tamada, et al. (2002). "Extensive feature detection of N-terminal protein sorting signals." Bioinformatics **18**(2): 298-305.

Barbrook, A. C., C. J. Howe, et al. (2006). "Why are plastid genomes retained in non-photosynthetic organisms?" Trends Plant Sci **11**(2): 101-108.

Barth, D. and T. U. Berendonk (2011). "The mitochondrial genome sequence of the ciliate *Paramecium caudatum* reveals a shift in nucleotide composition and codon usage within the genus *Paramecium*." BMC Genomics **12**: 272.

Bateman, A., L. Coin, et al. (2004). "The Pfam protein families database." Nucleic Acids Res **32**(Database issue): D138-141.

Baurain, D., H. Brinkmann, et al. (2010). "Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles." Mol Biol Evol **27**(7): 1698-1709.

Bedhomme, M., M. Hoffmann, et al. (2005). "Folate metabolism in plants: an *Arabidopsis* homolog of the mammalian mitochondrial folate transporter mediates folate import into chloroplasts." J Biol Chem **280**(41): 34823-34831.

Bendtsen, J. D., H. Nielsen, et al. (2004). "Improved prediction of signal peptides: SignalP 3.0." J Mol Biol **340**(4): 783-795.

Bensasson, D., D. Zhang, et al. (2001). "Mitochondrial pseudogenes: evolution's misplaced witnesses." Trends Ecol Evol **16**(6): 314-321.

Berg, O. G. and C. G. Kurland (2000). "Why mitochondrial genes are most often found in nuclei." Mol Biol Evol **17**(6): 951-961.

Berget, S. M., C. Moore, et al. (1977). "Spliced segments at the 5' terminus of adenovirus 2 late mRNA." Proc Natl Acad Sci U S A **74**(8): 3171-3175.

Bhattacharya, D., H. S. Yoon, et al. (2003). "Photosynthetic eukaryotes unite: Endosymbiosis connects the dots." Bioessays **26**: 50-60.

Bhushan, S., B. Lefebvre, et al. (2003). "Dual targeting and function of a protease in mitochondria and chloroplasts." EMBO Rep **4**(11): 1073-1078.

Birky, C. W., Jr. (2008). "Uniparental inheritance of organelle genes." Curr Biol **18**(16): R692-695.

Birney, E., M. Clamp, et al. (2004). "GeneWise and Genomewise." Genome Res **14**(5): 988-995.

Blanchard, J. L. and M. Lynch (2000). "Organellar genes: why do they end up in the nucleus?" Trends Genet **16**(7): 315-320.

Bodyl, A., J. W. Stiller, et al. (2009). "Chromalveolate plastids: direct descent or multiple endosymbioses?" Trends Ecol Evol **24**(3): 119-121; author reply 121-112.

Bonfield, J. K., K. Smith, et al. (1995). "A new DNA sequence assembly program." Nucleic Acids Res **23**(24): 4992-4999.

Boussau, B., E. O. Karlberg, et al. (2004). "Computational inference of scenarios for alpha-proteobacterial genome evolution." Proc Natl Acad Sci U S A **101**(26): 9722-9727.

Bouvier, F., N. Linka, et al. (2006). "*Arabidopsis* SAMT1 defines a plastid transporter regulating plastid biogenesis and plant development." Plant Cell **18**(11): 3088-3105.

Bowler, C., A. E. Allen, et al. (2008). "The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes." Nature **456**(7219): 239-244.

Brennicke, A., L. Grohmann, et al. (1993). "The mitochondrial genome on its way to the nucleus: different stages of gene transfer in higher plants." FEBS Lett **325**(1-2): 140-145.

Bridges, H. R., I. M. Fearnley, et al. (2010). "The subunit composition of mitochondrial NADH:ubiquinone oxidoreductase (complex I) from *Pichia pastoris*." Mol Cell Proteomics **9**(10): 2318-2326.

Brindefalk, B., T. J. Ettema, et al. (2011). "A phylometagenomic exploration of oceanic alphaproteobacteria reveals mitochondrial relatives unrelated to the SAR11 clade." PLoS One **6**(9): e24457.

Burger, G., M. W. Gray, et al. (2003). "Mitochondrial genomes: anything goes." Trends Genet **19**(12): 709-716.

Burger, G., B. F. Lang, et al. (1996). "Genes encoding the same three subunits of respiratory complex II are present in the mitochondrial DNA of two phylogenetically distant eukaryotes." Proc Natl Acad Sci U S A **93**(6): 2328-2332.

Burki, F., P. Flegontov, et al. (2012). "Re-evaluating the green versus red signal in eukaryotes with secondary plastid of red algal origin." Genome Biol Evol **4**(6): 626-635.

Burki, F., K. Shalchian-Tabrizi, et al. (2007). "Phylogenomics reshuffles the eukaryotic supergroups." PLoS One **8**: e790.

Burki, F., K. Shalchian-Tabrizi, et al. (2008). "Phylogenomics reveals a new 'megagroup' including most photosynthetic eukaryotes." Biology Letters.

Burri, L., Y. Strahm, et al. (2005). "Mature DIABLO/Smac is produced by the IMP protease complex on the mitochondrial inner membrane." Mol Biol Cell **16**(6): 2926-2933.

Cardol, P., D. Gonzalez-Halphen, et al. (2005). "The mitochondrial oxidative phosphorylation proteome of *Chlamydomonas reinhardtii* deduced from the Genome Sequencing Project." Plant Physiol **137**(2): 447-459.

Carrie, C., M. W. Murcha, et al. (2010). "An in silico analysis of the mitochondrial protein import apparatus of plants." BMC Plant Biol **10**: 249.

Carroll, J., I. M. Fearnley, et al. (2003). "Analysis of the subunit composition of complex I from bovine heart mitochondria." Mol Cell Proteomics **2**(2): 117-126.

Cavalier-Smith, T. (1987). "Eukaryotes with no mitochondria." Nature **326**(332-333).

Cavalier-Smith, T. (1998). "A revised six-kingdom system of life." Biol. Rev. Camb. Philos. Soc. **73**: 203-266.

Cavallaro, G. (2010). "Genome-wide analysis of eukaryotic twin CX9C proteins." Mol Biosyst **6**(12): 2459-2470.

Cazalet, C., L. Gomez-Valero, et al. (2010). "Analysis of the *Legionella longbeachae* genome and transcriptome uncovers unique strategies to cause Legionnaires' disease." PLoS Genet **6**(2): e1000851.

Cedergren, R., M. W. Gray, et al. (1988). "The evolutionary relationships among known life forms." J Mol Evol **28**(1-2): 98-112.

Chacinska, A., C. M. Koehler, et al. (2009). "Importing mitochondrial proteins: machineries and mechanisms." Cell **138**(4): 628-644.

Chan, C. X., M. B. Soares, et al. (2012). "Analysis of *Alexandrium tamarense* (Dinophyceace) genes reveals the complex evolutionary history of a microbial eukaryote." Journal of Phycology **47**: 1-13.

Chan, C. X., E. C. Yang, et al. (2011). "Red and green algal monophyly and extensive gene sharing found in a rich repertoire of red algal genes." Curr Biol **21**(4): 328-333.

Claros, M. G. and P. Vincens (1996). "Computational method to predict mitochondrially imported proteins and their targeting sequences." Eur J Biochem **241**(3): 779-786.

Cock, J. M., L. Sterck, et al. (2010). "The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae." Nature **465**(7298): 617-621.

Cole, C., J. D. Barber, et al. (2008). "The Jpred 3 secondary structure prediction server." Nucleic Acids Res **36**(Web Server issue): W197-201.

Cullis, C. A., B. J. Vorster, et al. (2008). "Transfer of Genetic Material Between the Chloroplast and Nucleus: How is it Related to Stress in Plants?" Ann Bot (Lond).

Curtis, B. A. and J. M. Archibald (2010). "A spliceosomal intron of mitochondrial DNA origin." Curr Biol **20**(21): R919-920.

Curtis, B. A., G. Tanifuji, et al. (2012). "Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs ." Nature 492:59-65.

Daley, D. O., K. L. Adams, et al. (2002). "Gene transfer from mitochondrion to nucleus: novel mechanisms for gene activation from Cox2." Plant J **30**(1): 11-21.

Delage, L., C. Leblanc, et al. (2011). "In silico survey of the mitochondrial protein uptake and maturation systems in the brown alga *Ectocarpus siliculosus*." PLoS One **6**(5): e19540.

Deschamps, P. and D. Moreira (2012). "Reevaluating the green contribution to diatom genomes." Genome Biol Evol **4**(7): 683-688.

Dolezal, P., M. Aili, et al. (2012). "*Legionella pneumophila* secretes a mitochondrial carrier protein during infection." PLoS Pathog **8**(1): e1002459.

Dorrell, R. G. and A. G. Smith (2011). "Do red and green make brown?: perspectives on plastid acquisitions within chromalveolates." Eukaryot Cell **10**(7): 856-868.

Douglas, S. E. and S. L. Penny (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**(2): 236-244.

Douglas, S. E., S. Zauner, et al. (2001). "The highly reduced genome of an enslaved algal nucleus." Nature **410**: 1091-1096.

du Buy, H. G. and F. L. Riley (1967). "HYBRIDIZATION BETWEEN THE NUCLEAR AND KINETOPLAST DNA'S OF *Leishmania enriettii* AND BETWEEN NUCLEAR AND MITOCHONDRIAL DNA'S OF MOUSE LIVER." Proc Natl Acad Sci U S A **57**(3): 790-797.

Eisen, J. A., R. S. Coyne, et al. (2006). "Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote." PLoS Biol **4**(9): e286.

Emanuelsson, O., S. Brunak, et al. (2007). "Locating proteins in the cell using TargetP, SignalP and related tools." Nat Protoc **2**(4): 953-971.

Emanuelsson, O., H. Nielsen, et al. (1999). "ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites." Protein Science **8**: 978-984.

Embley, T. M. and W. Martin (2006). "Eukaryotic evolution, changes and challenges." Nature **440**(7084): 623-630.

Embley, T. M., M. van der Giezen, et al. (2003). "Hydrogenosomes, mitochondria and early eukaryotic evolution." IUBMB Life **55**(7): 387-395.

Endo, T., K. Yamano, et al. (2010). "Structural basis for the disulfide relay system in the mitochondrial intermembrane space." Antioxid Redox Signal **13**(9): 1359-1373.

Esser, C., N. Ahmadinejad, et al. (2004). "A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes." Mol Biol Evol **21**(9): 1643-1660.

Eubel, H., E. H. Meyer, et al. (2008). "Novel proteins, putative membrane transporters, and an integrated metabolic network are revealed by quantitative proteomic analysis of *Arabidopsis* cell culture peroxisomes." Plant Physiol **148**(4): 1809-1829.

Farlow, A., E. Meduri, et al. (2010). "Nonsense-mediated decay enables intron gain in *Drosophila*." PLoS Genet **6**(1): e1000819.

Frickey, T. and A. N. Lupas (2004). "Phylogenetic analysis of AAA proteins." J Struct Biol **146**(1-2): 2-10.

Fukao, Y., Y. Hayashi, et al. (2001). "Developmental analysis of a putative ATP/ADP carrier protein localized on glyoxysomal membranes during the peroxisome transition in pumpkin cotyledons." Plant Cell Physiol **42**(8): 835-841.

Gabaldon, T. and M. A. Huynen (2004). "Shaping the mitochondrial proteome." Biochim Biophys Acta **1659**(2-3): 212-220.

Gabaldon, T., D. Rainey, et al. (2005). "Tracing the evolution of a large protein complex in the eukaryotes, NADH:ubiquinone oxidoreductase (Complex I)." J Mol Biol **348**(4): 857-870.

Gabaldon, T., B. Snel, et al. (2006). "Origin and evolution of the peroxisomal proteome." Biol Direct **1**: 8.

Gakh, O., P. Cavadini, et al. (2002). "Mitochondrial processing peptidases." Biochim Biophys Acta **1592**(1): 63-77.

Gawryluk, R. M., K. A. Chisholm, et al. (2012). "Composition of the mitochondrial electron transport chain in *Acanthamoeba castellanii*: Structural and evolutionary insights." Biochim Biophys Acta **1817**(11): 2027-2037.

Gawryluk, R. M. and M. W. Gray (2010). "Evidence for an early evolutionary emergence of gamma-type carbonic anhydrases as components of mitochondrial respiratory complex I." BMC Evol Biol **10**: 176.

Gentle, I. E., A. J. Perry, et al. (2007). "Conserved motifs reveal details of ancestry and structure in the small TIM chaperones of the mitochondrial intermembrane space." Mol Biol Evol **24**(5): 1149-1160.

Gerber, J., U. Muhlenhoff, et al. (2001). "Yeast ERV2p is the first microsomal FAD-linked sulfhydryl oxidase of the Erv1p/Alrp protein family." J Biol Chem **276**(26): 23486-23491.

Gilson, P. R., U. G. Maier, et al. (1997). "Size isn't everything: lessons in genetic miniaturisation from nucleomorphs." Curr. Opin. Genet. Dev. **7**(6): 800-806.

Gilson, P. R. and G. I. McFadden (1996). "The miniaturized nuclear genome of a eukaryotic endosymbiont contains genes that overlap, genes that are cotranscribed, and the smallest known spliceosomal introns." Proc. Natl. Acad. Sci. USA **93**(15): 7737-7742.

Gilson, P. R. and G. I. McFadden (2002). "Jam packed genomes--a preliminary, comparative analysis of nucleomorphs." Genetica **115**(1): 13-28.

Gilson, P. R., V. Su, et al. (2006). "Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus." Proc Natl Acad Sci U S A **103**(25): 9566-9571.

Giovannoni, S. J., H. J. Tripp, et al. (2005). "Genome streamlining in a cosmopolitan oceanic bacterium." Science **309**(5738): 1242-1245.

Gladyshev, E. A., M. Meselson, et al. (2008). "Massive horizontal gene transfer in bdelloid rotifers." Science **320**(5880): 1210-1213.

Gnerre, S., I. Maccallum, et al. (2011). "High-quality draft assemblies of mammalian genomes from massively parallel sequence data." Proc Natl Acad Sci U S A **108**(4): 1513-1518.

Gobler, C. J., D. L. Berry, et al. (2011). "Niche of harmful alga *Aureococcus anophagefferens* revealed through ecogenomics." Proc Natl Acad Sci U S A **108**(11): 4352-4357.

Goremykin, V. V., F. Salamini, et al. (2009). "Mitochondrial DNA of *Vitis vinifera* and the issue of rampant horizontal gene transfer." Mol Biol Evol **26**(1): 99-110.

Gould, S. B., M. S. Sommer, et al. (2006). "Protein targeting into the complex plastid of cryptophytes." J Mol Evol **62**(6): 674-681.

Gould, S. B., M. S. Sommer, et al. (2006). "Nucleus-to-Nucleus Gene Transfer and Protein Retargeting into a Remnant Cytoplasm of Cryptophytes and Diatoms." Mol Biol Evol **23**(12): 2413-2422.

Grabherr, M. G., B. J. Haas, et al. (2011). "Full-length transcriptome assembly from RNA-Seq data without a reference genome." Nat Biotechnol **29**(7): 644-652.

Gray, M. W. (1999). "Evolution of organellar genomes." Curr Opin Genet Dev **9**(6): 678-687.

Gray, M. W. (2012). "Mitochondrial evolution." Cold Spring Harb Perspect Biol **4**(9).

Gray, M. W., G. Burger, et al. (1999). "Mitochondrial evolution." Science **283**(5407): 1476-1481.

Gray, M. W., G. Burger, et al. (2001). "The origin and early evolution of mitochondria." Genome Biol **2**(6): REVIEWS1018.

Gray, M. W., B. F. Lang, et al. (2004). "Mitochondria of protists." Annu Rev Genet **38**: 477-524.

Grigoriev, I. V., H. Nordberg, et al. (2012). "The genome portal of the Department of Energy Joint Genome Institute." Nucleic Acids Res **40**(Database issue): D26-32.

Gschloessl, B., Y. Guermeur, et al. (2008). "HECTAR: a method to predict subcellular targeting in heterokonts." BMC Bioinformatics **9**: 393.

Haas, B. J., A. L. Delcher, et al. (2003). "Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies." Nucleic Acids Res **31**(19): 5654-5666.

Haas, B. J., S. Kamoun, et al. (2009). "Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*." Nature **461**(7262): 393-398.

Hackett, J. D., D. M. Anderson, et al. (2004). "Dinoflagellates: A remarkable evolutionary experiment." Am. J. Bot. **91**: 1523-1534.

Hackett, J. D., H. S. Yoon, et al. (2007). "Phylogenomic analysis supports the monophyly of cryptophytes and haptophytes and the association of rhizaria with chromalveolates." Mol Biol Evol **24**(8): 1702-1713.

Hampl, V., L. Hug, et al. (2009). "Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups"." Proc Natl Acad Sci U S A **106**(10): 3859-3864.

Hansen, K. D., L. F. Lareau, et al. (2009). "Genome-wide identification of alternative splice forms down-regulated by nonsense-mediated mRNA decay in *Drosophila*." PLoS Genet **5**(6): e1000525.

Harper, J. T., E. Waanders, et al. (2005). "On the monophyly of chromalveolates using a six-protein phylogeny of eukaryotes." Int J Syst Evol Microbiol **55**(Pt 1): 487-496.

Hauth, A. M., U. G. Maier, et al. (2005). "The *Rhodomonas salina* mitochondrial genome: bacteria-like operons, compact gene arrangement and complex repeat region." Nucleic Acids Res **33**(14): 4433-4442.

Hazkani-Covo, E., R. M. Zeller, et al. (2010). "Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes." PLoS Genet **6**(2): e1000834.

He, X., S. Tao, et al. (2010). "The most redundant sequences in human CpG island library are derived from mitochondrial genome." Genomics Proteomics Bioinformatics **8**: 81-91

Heazlewood, J. L., K. A. Howell, et al. (2003). "Mitochondrial complex I from *Arabidopsis* and rice: orthologs of mammalian and fungal components coupled with plant-specific subunits." Biochim Biophys Acta **1604**(3): 159-169.

Heazlewood, J. L., J. Tonti-Filippini, et al. (2005). "Combining experimental and predicted datasets for determination of the subcellular location of proteins in *Arabidopsis*." Plant Physiol **139**(2): 598-609.

Heazlewood, J. L., J. S. Tonti-Filippini, et al. (2004). "Experimental analysis of the *Arabidopsis* mitochondrial proteome highlights signaling and regulatory components, provides assessment of targeting prediction programs, and indicates plant-specific mitochondrial proteins." Plant Cell **16**(1): 241-256.

Hell, K. (2008). "The Erv1-Mia40 disulfide relay system in the intermembrane space of mitochondria." Biochim Biophys Acta **1783**(4): 601-609.

Henze, K. and W. Martin (2001). "How do mitochondrial genes get into the nucleus?" Trends Genet **17**(7): 383-387.

Hibberd, D. J. and R. E. Norris (1984). "Cytology and ultrastructure of *Chlorarachnion reptans* (Chlorarachniophyta divisio nova, Chlorarachniophyceae classis nova)." J. Phycol. **20**: 310-330.

Hill, D. R. A. and R. Wetherbee (1990). "*Guillardia theta* gen. et sp.nov. (Cryptophyceae)." Can. J. Bot. **68**(9): 1873-1876.

Hirakawa, Y., F. Burki, et al. (2012). "Genome-based reconstruction of the protein import machinery in the secondary plastid of a chlorarachniophyte alga." Eukaryot Cell **11**(3): 324-333.

Hirakawa, Y., G. H. Gile, et al. (2010). "Characterization of periplastidal compartment-targeting signals in chlorarachniophytes." Mol Biol Evol **27**(7): 1538-1545.

Hirst, J., J. Carroll, et al. (2003). "The nuclear encoded subunits of complex I from bovine heart mitochondria." Biochim Biophys Acta **1604**(3): 135-150.

Hoef-Emden, K. and M. Melkonian (2003). "Revision of the genus *Cryptomonas* (Cryptophyceae): a combination of molecular phylogeny and morphology provides insights into a long-hidden dimorphism." Protist **154**: 371-409.

Hoffmann, H. P. and C. J. Avers (1973). "Mitochondrion of yeast: ultrastructural evidence for one giant, branched organelle per cell." Science **181**(4101): 749-751.

Hong, S. and P. L. Pedersen (2003). "ATP synthases: insights into their motor functions from sequence and structural analyses." J Bioenerg Biomembr **35**(2): 95-120.

Hopkins, J.F., D. F. Spencer, et al. (2012). " Proteomics reveals plastid- and periplastid-targeted proteins in the chlorarachniophyte alga *Bigelowiella natans.*" Unpublished.

Hou, Y. and S. Lin (2009). "Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for dinoflagellate genomes." PLoS One **4**(9): e6978.

Hua, S. and Z. Sun (2001). "Support vector machine approach for protein subcellular localization prediction." Bioinformatics **17**(8): 721-728.

Huang, C. Y., M. A. Ayliffe, et al. (2003). "Direct measurement of the transfer rate of chloroplast DNA into the nucleus." Nature **422**(6927): 72-76.

Huang, C. Y., M. A. Ayliffe, et al. (2004). "Simple and complex nuclear loci created by newly transferred chloroplast DNA in tobacco." Proc Natl Acad Sci U S A **101**(26): 9710-9715.

Huang, S., N. L. Taylor, et al. (2010). "Functional and composition differences between mitochondrial complex II in *Arabidopsis* and rice are correlated with the complex genetic history of the enzyme." Plant Mol Biol **72**(3): 331-342.

Hug, L. A., A. Stechmann, et al. (2010). "Phylogenetic distributions and histories of proteins involved in anaerobic pyruvate metabolism in eukaryotes." Mol Biol Evol **27**(2): 311-324.

Imai, K. and K. Nakai (2010). "Prediction of subcellular locations of proteins: where to proceed?" Proteomics **10**(22): 3970-3983.

Ishida, K., Y. Cao, et al. (1997). "The origin of chlorarachniophyte plastids, as inferred from phylogenetic comparisons of amino acid sequences of EF-Tu." J. Mol. Evol. **45**(6): 682-687.

Ishida, K., B. R. Green, et al. (1999). "Diversification of a chimaeric algal group, the chlorarachniophytes: phylogeny of nuclear and nucleomorph small-subunit rRNA genes." Mol. Biol. Evol. **16**(3): 321-331.

Iwata, S., J. W. Lee, et al. (1998). "Complete structure of the 11-subunit bovine mitochondrial cytochrome bc1 complex." Science **281**(5373): 64-71.

Jaffe, D. B., J. Butler, et al. (2003). "Whole-genome sequence assembly for mammalian genomes: Arachne 2." Genome Res **13**(1): 91-96.

Karlberg, O., B. Canback, et al. (2000). "The dual origin of the yeast mitochondrial proteome." Yeast **17**(3): 170-187.

Keeling, P. J. and C. H. Slamovits (2005). "Causes and effects of nuclear genome reduction." Curr. Op. Gen. Dev. **15**: 601-608.

Kent, W.J. (2002). "BLAT--the BLAST-like alignment tool." Genome Res. 12: 656-664.

Kilian, O. and P. G. Kroth (2005). "Identification and characterization of a new conserved motif within the presequence of proteins targeted into complex diatom plastids." Plant J **41**(2): 175-183.

Kim, E., C. E. Lane, et al. (2008). "Complete sequence and analysis of the mitochondrial genome of *Hemiselmis andersenii* CCMP644 (Cryptophyceae)." BMC Genomics **9**: 215.

Kirchberger, S., J. Tjaden, et al. (2008). "Characterization of the *Arabidopsis* Brittle1 transport protein and impact of reduced activity on plant metabolism." Plant J **56**(1): 51-63.

Klodmann, J., S. Sunderhaus, et al. (2010). "Internal architecture of mitochondrial complex I from *Arabidopsis thaliana*." Plant Cell **22**(3): 797-810.

Kodali, V. K. and C. Thorpe (2010). "Oxidative protein folding and the Quiescin-sulfhydryl oxidase family of flavoproteins." Antioxid Redox Signal **13**(8): 1217-1230.

Koonin, E. V., N. D. Fedorova, et al. (2004). "A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes." Genome Biol **5**(2): R7.

Koski, L. B. and G. B. Golding (2001). "The closest BLAST hit is often not the nearest neighbor." J Mol Evol **52**(6): 540-542.

Kozjak-Pavlovic, V., K. Ross, et al. (2007). "Conserved roles of Sam50 and metaxins in VDAC biogenesis." EMBO Rep **8**(6): 576-582.

Krampis, K., B. M. Tyler, et al. (2006). "Extensive variation in nuclear mitochondrial DNA content between the genomes of *Phytophthora sojae* and *Phytophthora ramorum*." Mol Plant Microbe Interact **19**(12): 1329-1336.

Krogh, A., M. Brown, et al. (1994). "Hidden Markov models in computational biology. Applications to protein modeling." J Mol Biol **235**(5): 1501-1531.

Krogh, A., B. Larsson, et al. (2001). "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes." J Mol Biol **305**(3): 567-580.

Kuroiwa, T. and H. Uchida (1996). "Organelle divisions and cytoplasmic inheritance." Bioscience **46**: 827-835.

Kutik, S., D. A. Stroud, et al. (2009). "Evolution of mitochondrial protein biogenesis." Biochim Biophys Acta **1790**(6): 409-415.

Lane, C. E. and J. M. Archibald (2008). "The eukaryotic tree of life: endosymbiosis takes its TOL." Trends Ecol Evol **23**(5): 268-275.

Lane, C. E., K. van den Heuvel, 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. USA **104**: 19908-19913.

Lane, N. and W. Martin (2010). "The energetics of genome complexity." Nature **467**(7318): 929-934.

Lang, B. F., G. Burger, et al. (1997). "An ancestral mitochondrial DNA resembling a eubacterial genome in miniature." Nature **387**(6632): 493-497.

Leister, D. (2005). "Origin, evolution and genetic effects of nuclear insertions of organelle DNA." Trends Genet **21**(12): 655-663.

Leroch, M., H. E. Neuhaus, et al. (2008). "Identification of a novel adenine nucleotide transporter in the endoplasmic reticulum of *Arabidopsis*." Plant Cell **20**(2): 438-451.

Li, W., A. E. Tucker, et al. (2009). "Extensive, recent intron gains in *Daphnia* populations." Science **326**(5957): 1260-1262.

Lill, R. and U. Muhlenhoff (2005). "Iron-sulfur-protein biogenesis in eukaryotes." Trends Biochem Sci **30**(3): 133-141.

Linka, N., F. L. Theodoulou, et al. (2008). "Peroxisomal ATP import is essential for seedling development in *Arabidopsis thaliana*." Plant Cell **20**(12): 3241-3257.

Lithgow, T. and A. Schneider (2010). "Evolution of macromolecular import pathways in mitochondria, hydrogenosomes and mitosomes." Philos Trans R Soc Lond B Biol Sci **365**(1541): 799-817.

Liu, S. L., Y. Zhuang, et al. (2009). "Comparative analysis of structural diversity and sequence evolution in plant mitochondrial genes transferred to the nucleus." Mol Biol Evol **26**(4): 875-891.

Lloyd, A. H. and J. N. Timmis (2011). "Endosybiotic evolution in action: Real-time observations of chloroplast to nucleus gene transfer." Mob Genet Elements **1**(3): 216-220.

Longen, S., M. Bien, et al. (2009). "Systematic analysis of the twin cx(9)c protein family." J Mol Biol **393**(2): 356-368.

Lough, A. N., L. M. Roark, et al. (2008). "Mitochondrial DNA transfer to the nucleus generates extensive insertion site variation in maize." Genetics **178**(1): 47-55.

Margulis, L. (1970). Origin of eukaryotic cells. Yale University Press.

Martin, W. (2003). "Gene transfer from organelles to the nucleus: frequent and in big chunks." Proc Natl Acad Sci U S A **100**(15): 8612-8614.

Martin, W., T. Dagan, et al. (2007). "The evolution of eukaryotes." Science **316**(5824): 542-543; author reply 542-543.

Martin, W. and R. G. Herrmann (1998). "Gene transfer from organelles to the nucleus: how much, what happens, and Why?" Plant Physiol **118**(1): 9-17.

Martin, W. and M. Muller (1998). "The hydrogen hypothesis for the first eukaryote." Nature **392**(6671): 37-41.

Matsuo, M., Y. Ito, et al. (2005). "The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast-nuclear DNA flux." Plant Cell **17**(3): 665-675.

McFadden, G. I. (1999). "Plastids and protein targeting." J. Eukaryot. Microbiol. **46**(4): 339-346.

McFadden, G. I., P. R. Gilson, et al. (1994). "*Goniomonas*: rRNA sequences indicate that this phagotrophic flagellate is a close relative of the host component of cryptomonads." Europ. J. Phycol. **29**(1): 29-32.

Meisinger, C., A. Sickmann, et al. (2008). "The mitochondrial proteome: from inventory to function." Cell **134**(1): 22-24.

Merchant, S. S., S. E. Prochnik, et al. (2007). "The *Chlamydomonas* genome reveals the evolution of key animal and plant functions." Science **318**(5848): 245-250.

Mick, D. U., T. D. Fox, et al. (2011). "Inventory control: cytochrome c oxidase assembly regulates mitochondrial translation." Nat Rev Mol Cell Biol **12**(1): 14-20.

Misumi, O., Y. Yoshida, et al. (2008). "Genome analysis and its significance in four unicellular algae, *Cyanidioschyzon merolae*, *Ostreococcus tauri*, *Chlamydomonas reinhardtii*, and *Thalassiosira pseudonana.*" J Plant Res **121**(1): 3-17.

Mota, M. (1963). "Electron microscope study of relationshp between nucleus and mitochondria in *Chlorophytum capense* (L.) Kuntze." Cytologia **28**: 409-416.

Moustafa, A., B. Beszteri, et al. (2009). "Genomic footprints of a cryptic plastid endosymbiosis in diatoms." Science **324**(5935): 1724-1726.

Nakai, K. and P. Horton (1999). "PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization." Trends Biochem Sci **24**(1): 34-36.

Nedelcu, A. M., I. H. Miles, et al. (2008). "Adaptive eukaryote-to-eukaryote lateral gene transfer: stress-related genes of algal origin in the closest unicellular relatives of animals." J Evol Biol **21**(6): 1852-1860.

Nielsen, H., S. Brunak, et al. (1999). "Machine learning approaches for the prediction of signal peptides and other protein sorting signals." Protein Eng **12**(1): 3-9.

Nishimura, Y. and D. B. Stern (2010). "Differential replication of two chloroplast genome forms in heteroplasmic *Chlamydomonas reinhardtii* gametes contributes to alternative inheritance patterns." Genetics **185**(4): 1167-1181.

Noutsos, C., T. Kleine, et al. (2007). "Nuclear insertions of organellar DNA can create novel patches of functional exon sequences." Trends Genet **23**(12): 597-601.

Noutsos, C., E. Richly, et al. (2005). "Generation and evolutionary fate of insertions of organelle DNA in the nuclear genomes of flowering plants." Genome Res **15**(5): 616-628.

Nugent, J. M. and J. D. Palmer (1991). "RNA-mediated transfer of the gene coxII from the mitochondrion to the nucleus during flowering plant evolution." Cell **66**(3): 473-481.

Nury, H., C. Dahout-Gonzalez, et al. (2006). "Relations between structure and function of the mitochondrial ADP/ATP carrier." Annu Rev Biochem **75**: 713-741.

O'Brien, E. A., Y. Zhang, et al. (2009). "GOBASE: an organelle genome database." Nucleic Acids Res **37**(Database issue): D946-950.

Ogata, H., S. Goto, et al. (1999). "KEGG: Kyoto Encyclopedia of Genes and Genomes." Nucleic Acids Res **27**(1): 29-34.

Ota, S., K. Ueda, et al. (2007). "*Norrisiella sphaerica* gen. et sp. nov., a new coccoid chlorarachniophyte from Baja California, Mexico." J Plant Res **120**(6): 661-670.

Pagliarini, D. J., S. E. Calvo, et al. (2008). "A mitochondrial protein compendium elucidates complex I disease biology." Cell **134**(1): 112-123.

Palmer, J. D. (2003). "The symbiotic birth and spread of plastids: how many times and whodunnit?" J. Phycol. **39**: 4-11.

Palmieri, F. and C. L. Pierri (2010). "Structure and function of mitochondrial carriers - role of the transmembrane helix P and G residues in the gating and transport mechanism." FEBS Lett **584**(9): 1931-1939.

Palmieri, F., C. L. Pierri, et al. (2011). "Evolution, structure and function of mitochondrial carriers: a review with new insights." Plant J **66**(1): 161-181.

Pamilo, P., L. Viljakainen, et al. (2007). "Exceptionally high density of NUMTs in the honeybee genome." Mol Biol Evol **24**(6): 1340-1346.

Parfrey, L. W., J. Grant, et al. (2010). "Broadly sampled multigene analyses yield a well-resolved eukaryotic tree of life." Syst Biol **59**(5): 518-533.

Paschen, S. A., W. Neupert, et al. (2005). "Biogenesis of beta-barrel membrane proteins of mitochondria." Trends Biochem Sci **30**(10): 575-582.

Patron, N. J., Y. Inagaki, et al. (2007). "Multiple gene phylogenies support the monophyly of cryptomonad and haptophyte host lineages." Curr. Biol. **17**: 887-891.

Patron, N. J., M. B. Rogers, et al. (2004). "Gene replacement of fructose-1,6-bisphosphate aldolase supports the hypothesis of a single photosynthetic ancestor of chromalveolates." Eukaryot Cell **3**(5): 1169-1175.

Pebay-Peyroula, E., C. Dahout-Gonzalez, et al. (2003). "Structure of mitochondrial ADP/ATP carrier in complex with carboxyatractyloside." Nature **426**(6962): 39-44.

Pisani, D., J. A. Cotton, et al. (2007). "Supertrees disentangle the chimerical origin of eukaryotic genomes." Mol Biol Evol **24**(8): 1752-1760.

Price, M. N., P. S. Dehal, et al. (2010). "FastTree 2--approximately maximum-likelihood trees for large alignments." PLoS One **5**(3): e9490.

Pusnik, M., O. Schmidt, et al. (2011). "Mitochondrial preprotein translocase of trypanosomatids has a bacterial origin." Curr Biol **21**(20): 1738-1743.

Race, H. L., R. G. Herrmann, et al. (1999). "Why have organelles retained genomes?" Trends Genet **15**(9): 364-370.

Reinders, J., R. P. Zahedi, et al. (2006). "Toward the complete yeast mitochondrial proteome: multidimensional separation techniques for mitochondrial proteomics." J Proteome Res **5**(7): 1543-1554.

Ricchetti, M., C. Fairhead, et al. (1999). "Mitochondrial DNA repairs double-strand breaks in yeast chromosomes." Nature **402**(6757): 96-100.

Ricchetti, M., F. Tekaia, et al. (2004). "Continued colonization of the human genome by mitochondrial DNA." PLoS Biol **2**(9): E273.

Rice, P., I. Longden, et al. (2000). "EMBOSS: the European Molecular Biology Open Software Suite." Trends Genet **16**(6): 276-277.

Richards, T. A. and J. M. Archibald (2011). "Cell evolution: gene transfer agents and the origin of mitochondria." Curr Biol **21**(3): R112-114.

Richly, E. and D. Leister (2004). "NUMTs in sequenced eukaryotic genomes." Mol Biol Evol **21**(6): 1081-1084.

Richly, E. and D. Leister (2004). "NUPTs in sequenced eukaryotes and their genomic organization in relation to NUMTs." Mol Biol Evol **21**(10): 1972-1980.

Rogers, M. B., P. R. Gilson, et al. (2007). "The complete chloroplast genome of the chlorarachniophyte *Bigelowiella natans*: evidence for independent origins of chlorarachniophyte and euglenid secondary endosymbionts." Mol Biol Evol **24**(1): 54-62.

Roy, S. W. and M. Irimia (2009). "Mystery of intron gain: new data and new models." Trends Genet **25**(2): 67-73.

Rutherford, K., J. Parkhill, et al. (2000). "Artemis: sequence visualization and annotation." Bioinformatics **16**(10): 944-945.

Sacerdot, C., S. Casaregola, et al. (2008). "Promiscuous DNA in the nuclear genomes of hemiascomycetous yeasts." FEMS Yeast Res **8**(6): 846-857.

Salamov, A. A. and V. V. Solovyev (2000). "Ab initio gene finding in *Drosophila* genomic DNA." Genome Res **10**(4): 516-522.

Sanchez-Puerta, M. V. and C. F. Delwiche (2008). "A hypothesis for plastid evolution in chromalveolates." J. Phycol. **44**(5): In Press.

Satow, R., T. C. Chan, et al. (2002). "Molecular cloning and characterization of dullard: a novel gene required for neural development." Biochem Biophys Res Commun **295**(1): 85-91.

Sayani, S., M. Janis, et al. (2008). "Widespread impact of nonsense-mediated mRNA decay on the yeast intronome." Mol Cell **31**(3): 360-370.

Schagger, H. and K. Pfeiffer (2000). "Supercomplexes in the respiratory chains of yeast and mammalian mitochondria." EMBO J **19**(8): 1777-1783.

Schnare, M. N. and M. W. Gray (1982). "3'-Terminal sequence of wheat mitochondrial 18S ribosomal RNA: further evidence of a eubacterial evolutionary origin." Nucleic Acids Res **10**(13): 3921-3932.

Sears, B. B. and K. VanWinkle-Swift (1994). "The salvage/turnover/repair (STOR) model for uniparental inheritance in *Chlamydomonas*: DNA as a source of sustenance." J Hered **85**(5): 366-376.

Selosse, M., B. Albert, et al. (2001). "Reducing the genome size of organelles favours gene transfer to the nucleus." Trends Ecol Evol **16**(3): 135-141.

Sickmann, A., J. Reinders, et al. (2003). "The proteome of *Saccharomyces cerevisiae* mitochondria." Proc Natl Acad Sci U S A **100**(23): 13207-13212.

Sideris, D. P., N. Petrakis, et al. (2009). "A novel intermembrane space-targeting signal docks cysteines onto Mia40 during mitochondrial oxidative folding." J Cell Biol **187**(7): 1007-1022.

Silver, T. D., S. Koike, et al. (2007). "Phylogeny and nucleomorph karyotype diversity of chlorarachniophyte algae." J Eukaryot Microbiol **54**(5): 403-410.

Simpson, J. T., K. Wong, et al. (2009). "ABySS: a parallel assembler for short read sequence data." Genome Res **19**(6): 1117-1123.

Small, I., N. Peeters, et al. (2004). "Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences." Proteomics **4**(6): 1581-1590.

Smeitink, J., R. Sengers, et al. (2001). "Human NADH:ubiquinone oxidoreductase." J Bioenerg Biomembr **33**(3): 259-266.

Smith, D. G., R. M. Gawryluk, et al. (2007). "Exploring the mitochondrial proteome of the ciliate protozoon *Tetrahymena thermophila*: direct analysis by tandem mass spectrometry." J Mol Biol **374**(3): 837-863.

Smith, D. R., K. Crosby, et al. (2011). "Correlation between nuclear plastid DNA abundance and plastid number supports the limited transfer window hypothesis." Genome Biol Evol **3**: 365-371.

Smith, P. M., J. L. Fox, et al. (2012). "Biogenesis of the cytochrome bc(1) complex and role of assembly factors." Biochim Biophys Acta **1817**(2): 276-286.

Stahl, A., P. Moberg, et al. (2002). "Isolation and identification of a novel mitochondrial metalloprotease (PreP) that degrades targeting presequences in plants." J Biol Chem **277**(44): 41931-41939.

Stanke, M., O. Schoffmann, et al. (2006). "Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources." BMC Bioinformatics **7**: 62.

Stechmann, A., K. Hamblin, et al. (2008). "Organelles in Blastocystis that blur the distinction between mitochondria and hydrogenosomes." Curr Biol **18**(8): 580-585.

Stiller, J. W. (2011). "Experimental design and statistical rigor in phylogenomics of horizontal and endosymbiotic gene transfer." BMC Evol Biol **11**: 259.

Stiller, J. W., J. Huang, et al. (2009). "Are algal genes in nonphotosynthetic protists evidence of historical plastid endosymbioses?" BMC Genomics **10**: 484.

Sun, G., Z. Yang, et al. (2010). "Algal genes in the closest relatives of animals." Mol Biol Evol **27**(12): 2879-2889.

Szklarczyk, R. and M. A. Huynen (2010). "Mosaic origin of the mitochondrial proteome." Proteomics **10**(22): 4012-4024.

Takano, H., K. Onoue, et al. (2010). "Mitochondrial fusion and inheritance of the mitochondrial genome." J Plant Res **123**(2): 131-138.

Tanifuji, G., N. T. Onodera, et al. (2011). "Complete nucleomorph genome sequence of the nonphotosynthetic alga *Cryptomonas paramecium* reveals a core nucleomorph gene set." Genome Biol Evol **3**: 44-54.

Taylor, S. W., E. Fahy, et al. (2003). "Characterization of the human heart mitochondrial proteome." Nat Biotechnol **21**(3): 281-286.

Thorsness, P. E. and E. R. Weber (1996). "Escape and migration of nucleic acids between chloroplasts, mitochondria, and the nucleus." Int Rev Cytol **165**: 207-234.

Tielens, A. G., C. Rotte, et al. (2002). "Mitochondria as we don't know them." Trends Biochem Sci **27**(11): 564-572.

Timmis, J. N., M. A. Ayliffe, et al. (2004). "Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes." Nat Rev Genet **5**(2): 123-135.

Tovar, J. (2007). Mitosomes of parasitic protozoa: biology and evolutionary significance. Origin of mitochondria and hydrogenosomes. W. F. Martin and M. Muller. Berlin, Springer-Verlag**:** 277-300.

Tovar, J., G. Leon-Avila, et al. (2003). "Mitochondrial remnant organelles of *Giardia* function in iron-sulphur protein maturation." Nature **426**(6963): 172-176.

Triant, D. A. and J. A. DeWoody (2007). "Extensive mitochondrial DNA transfer in a rapidly evolving rodent has been mediated by independent insertion events and by duplications." Gene **401**(1-2): 61-70.

Tsaousis, A. D., D. Gaston, et al. (2011). "A functional Tom70 in the human parasite Blastocystis sp.: implications for the evolution of the mitochondrial import apparatus." Mol Biol Evol **28**(1): 781-791.

Tsukihara, T., H. Aoyama, et al. (1996). "The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 A." Science **272**(5265): 1136-1144.

Turmel, M., C. Lemieux, et al. (1999). "The complete mitochondrial DNA sequences of *Nephroselmis olivacea* and *Pedinomonas minor.* Two radically different evolutionary patterns within green algae." Plant Cell **11**(9): 1717-1730.

Turmel, M., C. Otis, et al. (2007). "An unexpectedly large and loosely packed mitochondrial genome in the charophycean green alga *Chlorokybus atmophyticus*." BMC Genomics **8**: 137.

Tuskan, G. A., S. Difazio, et al. (2006). "The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray)." Science **313**(5793): 1596-1604.

Tyler, B. M., S. Tripathy, et al. (2006). "Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis." Science **313**(5791): 1261-1266.

van der Giezen, M. (2009). "Hydrogenosomes and mitosomes: conservation and evolution of functions." J Eukaryot Microbiol **56**(3): 221-231.

van Hellemond, J. J., A. van der Klei, et al. (2003). "Biochemical and evolutionary aspects of anaerobically functioning mitochondria." Philos Trans R Soc Lond B Biol Sci **358**(1429): 205-213; discussion 213-205.

Viale, A. M. and A. K. Arakaki (1994). "The chaperone connection to the origins of the eukaryotic organelles." FEBS Lett **341**(2-3): 146-151.

Videira, A. and M. Duarte (2002). "From NADH to ubiquinone in *Neurospora* mitochondria." Biochim Biophys Acta **1555**(1-3): 187-191.

Vogtle, F. N., S. Wortelkamp, et al. (2009). "Global analysis of the mitochondrial N-proteome identifies a processing peptidase critical for protein stability." Cell **139**(2): 428-439.

Wagener, N. and W. Neupert (2012). "Bcs1, a AAA protein of the mitochondria with a role in the biogenesis of the respiratory chain." J Struct Biol **179**(2): 121-125.

Wiedemann, N., E. Urzica, et al. (2006). "Essential role of Isd11 in mitochondrial iron-sulfur cluster synthesis on Isu scaffold proteins." EMBO J **25**(1): 184-195.

Wieprecht, T., O. Apostolov, et al. (2000). "Interaction of a mitochondrial presequence with lipid membranes: role of helix formation for membrane binding and perturbation." Biochemistry **39**(50): 15297-15305.

Wolfe, K. H., W. H. Li, et al. (1987). "Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs." Proc Natl Acad Sci U S A **84**(24): 9054-9058.

Worden, A. Z., J. H. Lee, et al. (2009). "Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*." Science **324**(5924): 268-272.

Wrzeszczynski, K. O. and B. Rost (2004). "Annotating proteins from endoplasmic reticulum and Golgi apparatus in eukaryotic proteomes." Cell Mol Life Sci **61**(11): 1341-1353.

Yamaguchi, M., Y. Namiki, et al. (2011). "Structome of *Saccharomyces cerevisiae* determined by freeze-substitution and serial ultrathin-sectioning electron microscopy." J Electron Microsc (Tokyo) **60**(5): 321-335.

Yoon, H. S., J. Grant, et al. (2008). "Broadly sampled multigene trees of eukaryotes." BMC Evol Biol **8**: 14.

Yoon, H. S., J. D. Hackett, et al. (2002). "A single origin of the peridinin- and fucoxanthin-containing plastids in dinoflagellates through tertiary endosymbiosis." Proc. Natl. Acad. Sci. USA **99**(18): 11724-11729.

Yoon, H. S., J. D. Hackett, et al. (2005). "Tertiary Endosymbiosis Driven Genome Evolution in Dinoflagellate Algae." Mol Biol Evol. **22**(5): 1299-308.

Zarsky, V., J. Tachezy, et al. (2012). "Tom40 is likely common to all mitochondria." Curr Biol **22**(12): R479-481; author reply R481-472.

Zdobnov, E. M. and R. Apweiler (2001). "InterProScan--an integration platform for the signature-recognition methods in InterPro." Bioinformatics **17**(9): 847-848.