Organelle Genome Evolution Within The Green Algal Genus Dunaliella

Michael Del Vasto
The University of Western Ontario

Supervisor
David Smith
The University of Western Ontario

Graduate Program in Biology
A thesis submitted in partial fulfillment of the requirements for the degree in Master of Science
© Michael Del Vasto 2015
ORGANELLE GENOME EVOLUTION WITHIN THE GREEN ALGAL GENUS DUNALIELLA

(Thesis format: Integrated Article)

by

Michael Del Vasto

Graduate Program in Biology

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science

The School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

© Michael Del Vasto 2015
Abstract

A large amount of diversity in genomic size and content exists within the mitochondrial and plastid genomes of green algae. However, there are still many un-sequenced green algal organelle genomes. In Smith et al. 2010, the green algal species Dunaliella salina, isolated from Western Australia (strain CCAP 19/18), had its organelle genomes fully sequenced. The genomes of this organism were found to contain large amounts of noncoding DNA. The lack of sequenced organelle genomes of green algae prevented the comparison of genomic architectures in other closely related species. In this study, I expanded on the information from the 2010 study by sequencing the organelle genomes from two Dunaliella species (D. salina CONC-001 and D. viridis CONC-002) isolated from a saline pond in Chile and compared their genomic architectures to that of D. salina CCAP 19/18. Sequencing, assembly, and bioinformatics analyses of the Chilean strains revealed genomic expansion within their mitochondrial and plastid genomes. Accompanying the increase in organelle genome size were highly inflated noncoding regions and an excess of introns. Upon further investigation, D. salina CCAP 19/18 and D. salina CONC-001 were found to be more similar to each other than to D. viridis CONC-002. Nevertheless, the two D. salina strains, given their major differences in organelle genome size, appear to represent distinct species.

Keywords

Organelle genomes, green algae, genome expansion, noncoding DNA, genome mapping, Chlamydomonadales, substitution rates, genome architecture, genome diversity, mutation rates
Co-Authorship Statement

Chapter 2 of this thesis has been published in Genome Biology and Evolution. It involved the collaboration of: Francisco Figueroa-Martinez, Jonathan Featherston, Mariela Gonzalez, Adrian Reyes-Prieto, Pierre Durand, David Smith me (Michael Del Vasto). All co-authors contributed to chapter 2 of this thesis: Francisco Figueroa-Martinez and Adrian Reyes-Prieto isolated, grew, sequenced and mapped the organelle genomes of Polytoma uvella and Chlamydomonas leiostraca, Mariela Gonzalez isolated both Dunaliella salina and Dunaliella viridis, Jonathan Featherston and Pierre Durand grew and sequenced the genomic data of D. salina and D. viridis, David Smith helped in the writing and editing as well as the editing of other chapters of the thesis, I analyzed the raw genomic data of D. salina CONC-001 and D. viridis CONC-002 from Chile and mapped the organelle genomes of D. salina CONC-001 and D. viridis CONC-002.
Acknowledgments

I thank my supervisor, Dr. David Smith, for all of his support and hard work during my master’s degree. He never stopped believing in me and had always inspired me to go above and beyond in everything. As soon as I met with David I knew that he was the perfect supervisor for me. He has taught me so many important things in these past two years that I will carry with me for the rest of my life. I am truly grateful and consider myself extremely lucky to have worked with such an inspiring person as David.

I would also like to thank my advisors Dr. Susanne Kohalmi and Dr. Denis Maxwell for all of their support. They provided me with excellent advice on important aspects of my results, and their opinions helped improve my research.

A portion of my research involved a collaboration of work that helped made this research possible. I thank my collaborators: Francisco Figueroa-Martinez, Jonathan Featherston, Mariela Gonzalez, Adrian Reyes-Prieto, and Pierre Durand, for all of their hard work.

I thank my lab mates, Sara Asmail and Yao Tian, for all of their help and support, and for the amazing bond we formed together, we will forever be the original 3. I also thank fellow graduate students, who played a huge role in making my graduate experience the best it could be. I’d also like to thank Jennifer McDonald and Laura Garofalo. Jennifer, I thank you for all of your helpful advice. Laura, the friendship we have formed throughout our graduate career is something that will stick forever. You have helped me in so many ways, I am truly thankful for everything.

Finally, I thank the most important people in my life, my parents. Mom and Dad, I honestly do not know what I would do without the continued love and support you both have always showed me, and the encouragement the both of you have always given me to allow me to reach my full potential. I only hope that I make you two as proud of me as I am to be your son.
Table of Contents.

Abstract .......................................................................................................................... ii
Co-Authorship Statement ............................................................................................... iii
Acknowledgments ........................................................................................................... iv
Table of Contents .......................................................................................................... v
List of Tables .................................................................................................................. viii
List of Figures ............................................................................................................... ix
List of Appendices ......................................................................................................... x
List of Abbreviations ..................................................................................................... xi

1 Green algae and organelle genomes ................................................................. 1
   1.1 Green algae ...................................................................................................... 1
   1.2 *Dunaliella* .................................................................................................. 3
   1.3 Species concept .............................................................................................. 5
   1.4 Endosymbiosis ............................................................................................... 6
   1.5 Organelle genome architecture ..................................................................... 8
   1.6 Chlamydomonadalean algae ......................................................................... 10
   1.7 Mutations and mutation rates ....................................................................... 11
   1.8 Thesis rational ............................................................................................... 12
   1.9 References ..................................................................................................... 13

2 Massive and widespread organelle genomic expansion in the green algal genus
*Dunaliella* ................................................................................................................ 18
   2.1 Introduction .................................................................................................... 18
   2.2 Materials and methods ................................................................................ 19
   2.3 Results and discussion ................................................................................ 20
      2.3.1 Sequencing new chlamydomonadalean mitochondrial genomes. ......... 20
4.4 Concluding remarks ........................................................................................................ 63
4.5 References ....................................................................................................................... 64
Appendices .......................................................................................................................... 67
Curriculum Vitae ................................................................................................................ 76
List of Tables

Table 2-1: *Dunaliella salina* CONC-001 and *D. viridis* CONC-002 plastid genome assembly statistics .................................................................28

Table 2-2: Mitochondrial and plastid DNA (mtDNA and ptDNA) substitution rates for two geographically distinct isolates of *Dunaliella salina*: CONC-001 (Chile) and CCAP 19/18 (Australia).................................................................29

Table 3-1: Location and number of introns in *D. salina* CCAP 19/18, *D. salina* CONC-001 and *D. viridis* CONC-002 .............................................................................................................44

Table 3-2: Pairwise identity of all coding regions in the *D. salina* CCAP 19/18, *D. salina* CONC-001 and *D. viridis* CONC-002 mtDNAs.......................................................46

Table 3-3: Pairwise Identity at the protein and DNA level for the seven protein coding genes ................................................................................................................49
List of Figures

Figure 1-1: Tree of life..................................................................................................................2

Figure 1-2: *Dunaliella* cells. Phase-contrast microscopy showing the flagella and cellular structures of *Dunaliella*. ........................................................................................................4

Figure 2-1: Mitochondrial genomic architecture and expansion within the Chlamydomonadales.................................................................................................................................23

Figure 2-2: Mitochondrial genomic upheaval in *Dunaliella*. ....................................................25

Figure 3-1: Gene order differences between *D. salina* CONC-001 and *D. salina* CCAP 19/18. ........................................................................................................................................42

Figure 3-2: Gene order of *D. salina* CCAP 19/18 and *D. viridis* CONC-002 .........................43

Figure 3-3: Protein alignment of a portion of cox1 and nad6......................................................49

Figure 3-4: Predicted tertiary protein structure of *cox1*. ......................................................52

Figure 3-5: Predicted tertiary protein structure of *nad6*. ......................................................53
List of Appendices

Appendix A: *Dunaliella salina* nucleotide substitution rates for individual protein-coding loci. Mitochondrial located genes. ........................................67

Appendix B: *Dunaliella salina* nucleotide substitution rates for individual protein-coding loci. Plastid located genes. ..........................................................68

Appendix C: Synonymous substitution rates in plastid and mitochondrial genomes of various plastid-bearing lineages.................................................................71

Appendix D: Maximum-likelihood protein phylogeny of chlamydomonadalean algae....72

Appendix E: References for Appendices A-D ..................................................................72

Appendix F: Permission for use of publication ..............................................................75
List of Abbreviations

CCAP 19/18: Culture Collection of Algae and Protozoa 19/18, the strain of *D. salina* isolated from Hutt Lagoon in Western Australia.

**cob**: Cytochrome b

**CONC-001**: Microalgal Culture Collection, Universidad de Concepción the strain of *D. salina* isolated from La Rinconada hypersaline pond in the Atacama Desert in northern Chile

**CONC-002**: Microalgal Culture Collection, Universidad de Concepción, the strain of *D. viridis* isolated from La Rinconada hypersaline pond in the Atacama Desert in northern Chile

**cox1**: Cytochrome c oxidase I

**mtDNA**: Mitochondrial genome

**nad1**: NADH dehydrogenase subunit 1

**nad2**: NADH dehydrogenase subunit 2

**nad4**: NADH dehydrogenase subunit 4

**nad5**: NADH dehydrogenase subunit 5

**nad6**: NADH dehydrogenase subunit 6

**nucDNA**: Nuclear genome

**ptDNA**: Plastid genome

**rrnL1 or L1**: Large subunit ribosomal RNA fragment 1

**rrnL2 or L2**: Large subunit ribosomal RNA fragment 2

**rrnL3 or L3**: Large subunit ribosomal RNA fragment 3
rrnL4 or L4: Large subunit ribosomal RNA fragment 4

rrnL5 or L5: Large subunit ribosomal RNA fragment 5

rrnL6 or L6: Large subunit ribosomal RNA fragment 6

rrnS1 or S1: Small subunit ribosomal RNA fragment 1

rrnS2 or S2: Small subunit ribosomal RNA fragment 2

rrnS3 or S3: Small subunit ribosomal RNA fragment 3

trnM or M: tRNA representing methionine

trnQ or Q: tRNA representing glutamine

trnW or W: tRNA representing tryptophan
Chapter 1

1 Green algae and organelle genomes

1.1 Green algae

Green algae are eukaryotic organisms that belong to the Viridiplantae lineage (or the “green lineage”), which includes green algae and land plants (Leliaert et al. 2012) (Figure 1-1), the latter of which evolved from a green algal ancestor (Kenrick and Crane 1997). This is why green algae and land plants share many of the same characteristics (Van den Hoek et al. 1995), including a double-membrane bound plastid with chlorophyll a and b, as well as other accessory pigments (Leliaert et al. 2012). Green organisms can be divided into two phyla, Chlorophyta and Streptophyta (Figure 1-1 B) (Leliaert et al. 2012). The majority of green algal organisms are found in Chlorophyta, whereas some green algal organisms and all land plants are found among Streptophyta. Chlorophyta can be divided into four classes: Chlorophyceae, Prasinophyceae, Ulvophyceae, and Trebouxiophyceae (Lewis and McCourt 2004). Chlorophyceae is composed of five orders: Sphaeropleales, Chaetophorales, Chaetopeltidales, Oedogoniales, and Chlamydomonadales (Brouard et al. 2010). Chlamydomonadales contain three well-studied genera of green algae, *Chlamydomonas*, *Volvox*, and *Dunaliella*. This thesis research focuses on the evolution of organelle genome architecture within the Chlamydomonadales.

Green algae live in various types of environments around the world. They can be found in terrestrial areas, as well as both marine and freshwater habitats (Graham 2009). In addition to living in different environments, green algae come in different thallus types: single cell (unicellular) or multicellular flagellates, unicellular and multicellular nonflagellates, branched and unbranched filaments, and coenocytes (cells with multiple nuclei) (Bold and Wynne 1978). The structural body type of green algal organisms is not an indication of relatedness, as organisms with similar thallus structures may not be closely related, whereas organisms with different body structures can be closely related (Graham 2009).
Figure 1-1: Tree of life A) Representation of the three domains of life (modified from Ciccarelli et al. 2006; Adl et al. 2012 using creative commons licensing). B) A cladogram of the domain Eukaryota, showing the classes that make up Viridiplantae lineage, made in MEGA 6.06 using a gene sequence from an organism in each class.
1.2 *Dunaliella*

*Dunaliella* is a genus of unicellular flagellate green algae whose cells lack a cell wall and can reproduce both sexually and asexually (Oren 2005) (Figure 1-2). In addition to allowing the cell to be motile, the flagella aids in sexual reproduction. Sexual reproduction begins by the flagella of two cells touching, causing a bridge to form and then the fusing of cells (Oren 2005). The likelihood of sexual reproduction is increased as the level of salt in an environment decreases (Martinez *et al.* 1995). *Dunaliella* cells are found in saline rich habitats throughout the world (Seckbach and Oren 2007). The *Dunaliella* genus contains organisms with varying degrees of salt tolerance (Seckbach and Oren 2007). For example, *Dunaliella salina* can grow best in an environment with a salt (NaCl) concentration of 2.053M, whereas *Dunaliella viridis* grows best at 1.027M of NaCl (Seckbach and Oren 2007).

To cope with the fluctuation in salinity levels, *Dunaliella* cells use two methods. The first method is the use of ion pumps located in their plasma membranes (Lee 2008). When the cells enter a higher salt environment the proteins in the plasma membrane pump out sodium from inside the cell (Lee 2008). The second method is the production and accumulation of glycerol in the cell (Seckbach and Oren 2007). The absence of a cell wall causes osmotic levels to have a huge influence on the shape of cells (Oren 2005). Varying salinity levels cause the cells to shrink when there is an increase and swell when there is a decrease (Lee 2008). *D. salina* is able to survive in fluctuating salt levels because of glycerol in the cells (Oren 2005). The role of glycerol in *Dunaliella* cells was first discovered by Craigie and McLachlan, they found that there was an increase in glycerol when *Dunaliella tertiolecta* was incubated in increasing NaCl conditions (Craigie and McLachlan 1964).

*Dunaliella* has received an increase in attention by researchers, in particular the species *D. salina*. *D. salina* is known for being a β-carotene-producing factory, where up to 10% of the cell’s dry weight is β-carotene (Ben-Amotz *et al.* 1982). β-carotene is a pigment in the apparatus known as the eyespot for green algae (Kreimer 2009). The eyespot contains photoreceptors that single the cell to move towards or away from light. (Kreimer 2009). In the eyespot, β-carotene, as well as other carotenoids, protect the cell from
Figure 1-2: *Dunaliella* cells. Phase-contrast microscopy showing the flagella and cellular structures of *Dunaliella* (http://www.scienceimage.csiro.au/image/7595 (picture on the left) and https://www.flickr.com/photos/63319497@N00/2844629276/ (picture on right) using creative commons licensing.)
The production of β-carotene is responsible for the deep red colour of *D. salina* and other *Dunaliella* species (Oren 2005).

Not only is *D. salina* halotolerant and rich in β-carotene, it has also been earmarked as an excellent candidate for generating biofuel (Weldy and Huesemann 2007). Potential application of *D. salina* for biofuel has gained much attention from the research community. Decreased levels of nitrogen causes oil production in green algae, this oil can then be used to make biofuels (http://jgi.doe.gov/why-sequence-chlamydomonas-and-chlorella/). Four reasons why *D. salina* is an excellent choice for biofuel production: 1) it produces higher levels of oil compared to other green algae, 2) the culturing of *D. salina* is low in cost and simple, 3) mass cultures of *D. salina* are already being grown in various countries for its β-carotene production, and 4) therefore it is also economically efficient to use it as a source of biofuel (Weldy and Huesemann 2007). The Department of Energy Joint Genome Institute is attempting to sequence the nuclear genome because of the oils associated with *D. salina* and their use as a potential fuel source for transportation vehicles (Smith *et al.* 2010). Sequencing the nuclear genome will allow researchers to gain an understanding for genes as well as pathways involved in the production of oils (http://jgi.doe.gov/why-sequence-chlamydomonas-and-chlorella/).

### 1.3 Species concept

Using the biological concept when defining a species, two organisms are usually considered the same species if they are able to produce fertile offspring. A problem with this definition is that many microbial organisms do not reproduce sexually. Other methods of identifying species are based on morphology and genetic data. However, this can be a problem for green algae as they are not well studied and often genetic information is missing. *D. salina*, for example, has genetic data available from isolates collected from countries, such as Spain, Egypt, Iran, and Australia (AlgaeBase). Unfortunately genetic data required to confirm that these isolates belong to the same species is missing. Suggesting that many of the isolates have been incorrectly identified as the same species when actually they are members of distinct species. I faced the issue of misidentified species directly by exploring two distinct geographical isolates of *D. salina*. 
1.4 Endosymbiosis

Algae and land plants contain three different genetic compartments: the nucleus, the mitochondrion, and the plastid. The genome that resides inside the nucleus is called the nuclear genome, whereas the genomes located inside the mitochondrion and the plastid are referred to as the mitochondrial and plastid genomes, respectively, or collectively the organelle genomes.

The ancestors of both organelles were two separate free-living organisms. Approximately 1.5 (for the plastid) and 2 billion (for the mitochondrion) years ago a host cell engulfed a cyanobacterium and α-proteobacterium, respectively. As time went on the former evolved into the plastid and the latter evolved into the mitochondrion of today (Gray et al. 1999; McBride et al. 2006). The engulfing of a free-living organism by a host is known as endosymbiosis. The endosymbiotic theory is a proposal that the mitochondrion and plastid arose from endosymbiosis. The theory gets its support from studies done on the mitochondrial genome, allowing researchers to trace back its evolutionary roots to an ancestor α-proteobacterium (Gray et al. 1999). Comparative genome analysis on plastid genomes has shown homology of sequences to cyanobacteria, therefore supporting the endosymbiotic theory for plastids (Leister and Pesaresi 2005). Since both the α-proteobacterium and cyanobacterium were free-living organisms they each had their own genome, and retained some of it after the endosymbiotic event. The genome of the α-proteobacterium eventually became the mitochondrial genome, mtDNA, while the genome of the cyanobacterium became the plastid genome, ptDNA. As these free-living organisms evolved into their modern day organelle genomes, a reduction in genome size and gene number had occurred (Lynch 2007). This loss of genes resulted from the migration of genes to the host’s nuclear genome called inter-genomic gene transfer (IGT) (Kleine et al. 2009). It is estimated that as much as 75% of the eukaryotic nuclear genome came from the α-proteobacterium endosymbiont (Lynch 2007). Due to IGT only some of the genes required for the proper function of the organelles are encoded for by the organelle genomes, whereas the majority of the genes required for organelle functions are encoded in the nuclear genome and post-translationally imported into the organelles (Peeters and Small 2001). Nuclear-encoded, organelle-targeted proteins have a special
N-terminal signal sequence allowing them to be sent to the organelles (Pujol et al. 2007). If the majority of organelle genes are nuclear encoded why do organelle genomes still exist, and why have not all of the organelle genes transferred to the nucleus? There are some hypotheses as to why organelles have retained their genome. One hypothesis postulates that certain organelle proteins are too hydrophobic to pass through the organelle membrane, therefore the proteins have to be encoded by the organelle genomes. The majority of genes that have remained within organelle DNA generally have functions associated with the organelle, such as cell respiration for the mitochondrion and photosynthesis for the plastid (Gray 2004). Both the mitochondrion and plastid genomes are commonly, but not always transferred to future generations via uniparental inheritance.

Mitochondrial genomes, and organelle genomes in general, are smaller than their nuclear counterparts, but they can still contain a very complex architecture. mtDNAs can have variation in their AT vs GC nucleotide ratios, topology, chromosome number, amount of noncoding DNA, gene number and genome size (Smith and Keeling 2015). To date the smallest mtDNA is 5.9 kilobases (Kb), whereas the largest is 1,685 Kb (Allen et al. 2007; Mach 2011). A feature of mitochondrial genomes (and most genomes for that matter) is that size is not necessarily proportional to the number of genes encoded in the genome. For example, even though the mtDNA of the cucumber is millions of nucleotides long, it contains fewer genes than the jakobid Reclinomonas americana mtDNA, which is only 69 kb long (Lang et al. 1997). In fact, of all eukaryotes, the mitochondrial gene content of R. americana (97 genes) most closely resembles that of a modern day α-proteobacterium, which codes for almost 4,000 genes (Lynch 2007).

Unlike mitochondria, plastids are only found in a subset of eukaryotes, including land plants, green algae and red algae. The word “plastid” encompasses different types of cellular compartments that store and produce compounds used by the cell (Wise 2007), including the leukoplast, the etioplast, the chromoplast, and the chloroplast, which carries out photosynthesis. Aside from photosynthesis, plastids perform other crucial cellular process, such as fatty-acid biosynthesis, nitrogen assimilation, amino acid biosynthesis, and starch biosynthesis (Tetlow et al. 2005). With regard to genome size, plastid genomes
rarely exceed 225 Kb in length and are on average about 150 Kb in land plants, there are exceptions that exist (Lynch 2007). In most plastid-containing eukaryotes, approximately 2500 proteins found within the chloroplast, only about 100 of which are coded for by the chloroplast genome (Wise 2007).

1.5 Organelle genome architecture

One widely studied attribute of organelle genomes is their architecture (size, shape and contents of a genome). It is evident that organelle genomes are smaller in size than nuclear genomes. Organelle genomes are composed of a double-stranded DNA, which is typically arranged in one of the three topologies: linear, truly circular, or circular mapping but linear in structure (Gray et al. 2004). The third type refers to genomes whose ends join and form a circle when being mapped, but in reality the structure is linear inside the organelle. Bendich (1996) investigated the shape of mtDNA by using conventional and pulsed-field gel electrophoresis studies. The gel revealed more linear mtDNA for the organisms tested as opposed to circular mtDNA (Bendich 1996). Oldenburg and Bendich (2004) further investigated the shape of organelle genomes, by studying corn chloroplast genomes using restriction digests. They found that the chloroplast genome was likely a concatemer of linear molecules, since portions of the genome found could not have arose form circular forms (Oldenburg and Bendich 2004).

The contents of both nuclear and organelle genomes are similar; a genome can contain coding as well as noncoding regions, and the relative ratio of these regions can vary from one organism to the next. Coding regions in organelle genomes are sequences of DNA that give rise to mRNA, rRNA, and tRNA. The number of genes encoded in the mitochondrial genomes can range from around 5-100 (Gray et al. 1999), where the most typical cases of reduced genomic content are found in alveolates and various other microbial eukaryotes, including Polytomella, which are close relatives of Dunaliella (Gray and Boer 1988). Noncoding regions include introns and intergenic spacers, not all organelle genomes contain introns, and the presence/absence of such can vary from one species to the next. The mtDNAs of animals usually have very little if any noncoding DNA whereas those of land plants are usually large and contain lots of noncoding DNA.
Introns are described as the noncoding area between the exons within a single gene, but they are often much more complex and can contain their own genes (intron encoded genes), which are involved in intron migration and other processes. There are four different classes of introns, but only two of the four, group I and group II, occur in organelle genomes (Gray et al. 2004). The two types of introns found in organelle genomes can be grouped together under the title of “autocatalytic introns”. As the name suggests, an autocatalytic intron is an intron that is able to splice itself from a sequence without the aid of any proteins. It is the secondary structures made by both group I and group II introns that allow for the intron to be spliced. Group II introns are known for their stem-loop structure, which was first established based on phylogenetic data known and then later confirmed by studying the thermodynamic prosperities of the intron (Kvakman et al. 1990; Lehmann and Schmidt 2003).

Plastid genomes display a similar array of architectural components and variations as mitochondrial genomes (Smith and Keeling 2015). However, unlike mtDNAs, plastid genomes often contain two inverted repeats, which divide the genome into two single-copy regions: a small single-copy region (SSR) and a large single-copy region (LSR). Shared characteristics, like proteins of chlorophyll a/b antenna complexes (Wolfe et al. 1994), among plastids allow researchers to make the assumption that these characteristics were acquired from the original cyanobacterium that gave rise to modern-day plastids (Leister and Pesaresi 2005).

Much has been said about the similarities and differences between mitochondrial and plastid genomes (Smith and Keeling 2015). In general, mitochondrial genomes display more architectural diversity and reach more “extremes” in size than plastid genomes. In land plants, the plastid genome is smaller than the mitochondrial genome, but in algae the opposite is usually true. Some of the few species with gigantic plastid genomes (greater than 250 kb) are chlamydomonadalean algae. Indeed, for a relatively small slice of the eukaryotic tree, chlamydomonadaleans harbour an impressive amount of organelle genome diversity.
1.6 Chlamydomonadalean algae

Chlamydomonadalean green algal species, the group that contains the widely studied genera *Chlamydomonas*, *Volvox*, and *Dunaliella*, can have very bizarre organelle genomic architectures. The first chlamydomonadalean green alga to have its mtDNA completely sequenced was *Chlamydomonas reinhardtii* (Remacle and Matagne 1998), which has a linear mtDNA of 15.8 kb, encoding just 13 genes, distributed on both strands of DNA (Remacle and Matagne 1998). The mtDNA of this organism was first characterized in 1978, and was fully sequenced in 1993 (Ryan et al. 1978; Remacle and Matagne 1998). Today, there exist sequence data for a wide range of chlamydomonadalean organelle genomes, among that is *D. salina* (Smith et al. 2010).

In 2010, Smith *et al.* sequenced the organelle genomes of a *D. salina* strain collected from Western Australia, CCAP 19/18, which gave significant insights into organelle genome size diversity of chlamydomonadalean green algae. Not only were the mtDNA and ptDNA of *D. salina* surprisingly large (28 kb and 269 kb, respectively), but both of the genomes also contained an unprecedented proportion of noncoding and intronic DNA (greater than 50%) (Smith *et al.* 2010). So far the only other chlamydomonadalean known to have a noncoding content in their mtDNA of over 50% is *Volvox carteri* (Smith *et al.* 2010).

Although largely composed of noncoding nucleotides, the *D. salina* organelle genomes do contain genes; following in the pattern of other chlamydomonadalean species, *D. salina* has a reduced mtDNA coding for only 12 essential mitochondrial genes, 7 of them are protein coding, 3 tRNA genes and 2 rRNA genes (Smith *et al.* 2010). An interesting feature of *D. salina* and some other chlamydomonadalean mtDNAs is that the rRNA genes are fragmented and scrambled throughout the genome. Also all 12 genes are encoded on the same strand, which is not unique to *D. salina*, and found in certain other chlamydomonadalean green algae (Smith *et al.* 2010). The *D. salina* plastid genome contains 102 genes, 66 of which represent proteins.

One of the biggest variations that occur in chlamydomonadalean algae is the amount of noncoding DNA. Based on the available data from nine chlamydomonadalean algae,
noncoding content in the mtDNA ranges from 18% to over 60% (Smith et al. 2010). Along with a variation in content, this class of organisms also experience a wide variation in their size of their mitochondrial genome. There is also variation seen in the plastid genome; however, due to the lack of fully sequenced plastid genomes for chlamydomonean green algae it is not as defined. From the latest data available we see extremely large sizes of plastid genomes, as high as 525 Kb in size (Smith et al. 2010).

1.7 Mutations and mutation rates

Mutation is a heritable change that can affect the function of a gene. Along with mutations, other processes like: gene duplication, whole genome duplication, transposable elements, genome reduction and gene loss, are mechanisms that can cause diversity of genomes (Bridges 1936; Wolfe and Shields 1997; Hamer et al. 2001; Boscaro et al. 2013). The process that is the main focus of my thesis is mutation and that rates at which they occur, however, these rates can be difficult to estimate (Smith et al. 2012). Looking at relative substitution rates of silent-sites between two closely related organisms can give an idea about mutation rates (Smith et al. 2012). Mutations can fall into one of three categories: detrimental, neutral, or beneficial. Detrimental mutations cause a decrease in an organism’s ability to survive and reproduce (fitness), neutral mutations have no effect on the organism’s fitness while beneficial mutations increase an organisms fitness.

Mutation, migration, natural selection and genetic drift, are factors that cause evolution to occur over time, and mutation is a huge factor for the evolution of genomes (Lynch et al. 2006).

High mutation frequencies can be especially problematic for larger genomes, because they are composed of more nucleotides. Large genomes contain more targets for mutations than small genomes. Consequently, given the same mutation and replication rates, large genomes will incur more mutations than small genomes. The mutational-hazard hypothesis argues that having excessive amounts of noncoding DNA increases an organism’s risk of acquiring potentially deleterious mutations (Lynch et. al 2006). Having a bloated genome can therefore be a burden for organisms. For example, organelle genomes contain self-splicing introns for which the secondary structure is crucial for the excision of the intron. If a mutation occurs that prevents the proper formation of the
secondary structure then the intron might not be removed from the gene, which could be lethal for the cell. This is why having a genome composed of large amounts of noncoding DNA—in this case lots of introns—can poses a risk to acquire more mutations as compared to compact genomes with very little noncoding DNA.

The burden of having a large genome is alleviated by a low mutation rate within the genome. Indeed, some of the largest organelle genomes ever observed have some of the lowest mutations rates, in fact land plant mtDNAs are actually considered to be the “slowest evolving cellular genomes” known (Scheffler 2008). Animal mtDNAs on the other hand, are usually smaller and more compact with little to no noncoding DNA, and tend to have a high mutation rate. Since nuclear genomes of both animals and plants have similar mutation rates, the differences that arises in organelle genomic architecture between these two groups is thought to be due to mutation (Lynch et al. 2006). Some factors that can cause a mutation rate to increase in the mtDNA are free oxygen radicals produced by the electron transport chain, faster replication than nucDNA causing an increase in the probability of mismatch, and the lack of repair proteins being encoded by the mtDNA (Lynch et al. 2006).

1.8 Thesis rational

This thesis explores the diversity of organelle genomic diversity within the green algal genus *Dunaliella* and attempts to place this diversity in context to what is known about organelle DNA evolution within the Chlamydomonadales as a whole. Specifically, I want to understand the processes responsible for organelle genome expansion, and investigate the role of mutation, if any, in influencing mitochondrial and plastid genome size. As highlighted above, *Dunaliella*—given its propensity for highly expanded organelle DNAs—is an excellent candidate for addressing these issues. Before now the only species of *Dunaliella* that has had its organelles completely sequenced is *D. salina* CCAP 19/18. It is not known if other *Dunaliella* species or other *D. salina* strains have expanded organelle DNAs.

To gain further insight into the evolution of organelle genomes, I, with the help of my supervisor and collaborating laboratories, sequenced the mitochondrial and plastid
genomes from two Chilean isolates of *Dunaliella*: *D. viridis* strain CONC-002 and *D. salina* strain CONC-001. Sequencing and mapping these two additional organisms will allow for the comparison of both, two Chilean isolates, and different geographical isolates. These comparisons will allow for the levels of diversity to be measured, potentially allowing for support of species identification. There is such great diversity within the organelle genomes of chlamydomonadalean green algae that it makes them the perfect model organisms for studies of genome evolution. Comparing the same species, if they are, from two different geographical locations tests the accuracy of species identification. If the two geographical isolates of *D. salina* (CCAP 19/18 and CONC-001) are both identified properly than there should be a higher similarity between the two isolates than there is between each isolate and *D. viridis* CONC-002, regardless of being from two different geographical environments.

1.9 References


Chapter 2

2 Massive and widespread organelle genomic expansion in the green algal genus *Dunaliella*


2.1 Introduction

The mitochondrial genomes of chlamydomonadalean green algae (Chlorophyta, Chlorophyceae) are somewhat of a contradiction (Leliaert *et al.* 2012). On the one hand, they have the smallest gene contents of any known organelle genomes from the Archaeplastida (Plantae *sensu lato*), encoding 7–8 proteins, 2 rRNAs, and 1–3 tRNAs (Smith *et al.* 2013^a^; Smith *et al.* 2013^b^), and they can also be very small (<13.5 kb) and compact (>80% coding) (Smith *et al.* 2010^a^). On the other hand, certain chlamydomonadalean mitochondrial DNAs (mtDNAs) are distended with repeats and introns, and composed almost entirely of noncoding nucleotides (Smith and Lee 2010).

One species with a particularly bloated mitochondrial genome is *D. salina*—a unicellular biflagellate, which lives in hypersaline environments, can accumulate large amounts of β-carotene, and is a prime candidate for biofuel production (Oren 2005). Complete mtDNA sequencing of *D. salina* CCAP 19/18, isolated from the Hutt Lagoon in Western Australia, revealed an unprecedentedly high intron density for a green alga (~1.5 intron per gene) as well as vast, repeat-rich intergenic regions (Smith *et al.* 2010^b^). An equally expanded genome was also uncovered in the plastid, implying that similar forces are shaping both organelle DNAs (Smith *et al.* 2010^b^).

Various studies have used chlamydomonadalean algae, including the model organism *C. reinhardtii* and its close multicellular relative *V. carteri*, to explore the evolution of genome size (e.g., Smith and Lee 2010). These investigations suggest that mutation rate,
DNA maintenance machineries, and random genetic drift have a major role in fashioning organelle chromosomes (Smith and Lee 2010; Hua et al. 2012; Smith et al. 2013b). Such studies, however, have yet to be applied to *Dunaliella* species, and it is still unknown if other members of the genus have inflated organelle genomes.

Here, we survey mitochondrial genome size and content within and outside the *Dunaliella* lineage. We show that although the mtDNA gene repertoire is nearly fixed across the Chlamydomonadales, there is an approximately four-fold variation in genome size and an eighteen-fold variation in intron content, with *Dunaliella* species having among the most expanded mtDNAs of all explored green algae. The same is likely true for their plastid genomes as well. The levels of organelle DNA divergence between distinct *D. salina* strains are used to investigate the potential forces underpinning such massive levels of genomic expansion.

### 2.2 Materials and methods

*D. salina* CONC-001 and *D. viridis* CONC-002 (Microalgae Culture Collection, Universidad de Concepción, Chile; updated acronyms CCM-UDEC 001 and CCM-UDEC 002, respectively) were grown in J/1 medium, supplemented with 15% NaCl w/v, at 18°C under a 14-h light/10-hr dark cycle, and harvested as previously described (González et al. 1999; Gómez and González 2005). Total DNA from each isolate was extracted using the Qiagen DNeasy Plant Mini Kit (Qiagen, Venlo, Limburg, NL) with liquid nitrogen disruption. Illumina sequencing libraries were prepared with the Nextera DNA Sample Preparation Kit (Illumina, San Diego, CA, USA) and sequenced on an Illumina MiSeq platform (v2 chemistry; 250x250 paired-end sequencing reads).

*Chlamydomonas leiostraca* SAG 11-49 (Culture Collection of Algae at the University of Göttingen, Germany) was grown in *Volvox* medium (Provasoli and Pintner 1960) at 18°C under a 14-h light/10-hr dark cycle and with constant shaking (200 rpm). *Polytoma uvella* UTEX 964 (Culture Collection of Algae at the University of Texas at Austin) was cultured in *Polytoma* medium under the same conditions, but without shaking. Cells were harvested in logarithmic growth by centrifugation at 4500 rpm for 10 minutes, and whole genomic DNA from each strain was extracted by standard phenol-chloroform methods.
and ethanol precipitation. Library preparation and Illumina sequencing (HiSeq 2000) were performed at the Roy J. Carver Center for Genomics of the University of Iowa (100x100 paired-end sequencing reads).

The organelle genomes of all four algae were assembled de novo with Ray v2.2.0 (Boisvert et al. 2010), using k-mers of 21, 27, 31, and 37, and separately with CLC Genomics Workbench v6.0.4 (Qiagen, Prismet, DK), using a word size of 20, bubble size of 50, and paired-end scaffolding. The resulting Ray and CLC contigs were scanned for organelle sequences using BLAST-based methods and the mitochondrial and plastid genomes of chlamydomonadalean algae as queries. Hits to organelle DNA were assembled into larger contigs using read-mapping approaches with Geneious v7.1.4 (Biomatters Ltd., Auckland, NZ). Organelle introns were identified with RNAweasel (http://megasun.bch.umontreal.ca/RNAweasel/) and through alignments with other chlamydomonadalean organelle DNAs.

Organelle genes were aligned with MUSCLE (Edgar 2004), implemented through Geneious, using default settings. Synonymous and nonsynonymous substitutions were measured with the CODEML program of PAML v4.3 (Yang 2007), employing the maximum likelihood method and the F3x4 codon model. Substitutions in non-protein-coding regions were estimated with BASEML of PAML, using the HKY85 model. The mitochondrial genome data described here are deposited in GenBank under accession numbers KP691601 (D. salina CONC-001), KP691602 (D. viridis CONC-002), KP696389 (C. leiostraca), and KP696388 (P. uvella). The D. salina CONC-001 protein-coding ptDNA genes used to measure substitution rates are in Appendix C.

2.3 Results and discussion

2.3.1 Sequencing new chlamydomonadalean mitochondrial genomes.

As part of an ongoing, collaborative initiative, we have been sequencing and characterizing organelle genomes from diverse chlamydomonadalean species (Hamaji et al. 2013; Smith et al. 2013b) (www.volvocales.org). Among these species are two distinct Dunaliella isolates, which were collected from the La Rinconada hypersaline pond in the
Atacama Desert in northern Chile: *D. salina* CONC-001 and *D. viridis* CONC-002 (Gómez-Silva *et al.* 1990; González *et al.* 1999; Gómez and González 2005). Closely related to *Dunaliella* are two other algae that we have also been investigating: the freshwater flagellate *C. leiostraca* SAG 11-49 and the free-living, nonphotosynthetic unicell *P. uvella* UTEX 964, which are a model duo for studying the loss of photosynthesis (Figueroa-Martinez *et al.* 2015).

Next-generation sequencing of total cellular DNA from these algae followed by mitochondrial genome assembly yielded both expected and unexpected results. At first glance, all four mtDNAs appear similar to one another and to those of various other chlamydomonadaleans: they map as single circular chromosomes (Bendich 1993), have identical gene compliments (representing 7 proteins, 3 tRNAs, and 2 rRNAs), and contain sections of overlapping gene order (Figure 2-1). Moreover, in each of the algae the mitochondrial large and small subunit rRNA genes are fragmented and scrambled into six and three coding modules, respectively (Figure 2-1), which is a common theme throughout the order, with species from the Reinhardtinia clade (Nakada *et al.* 2008) displaying even greater levels of rRNA gene fragmentation (Smith *et al.* 2013b). Three of the four genomes also have notably high guanine and cytosine contents: 39% (*C. leiostraca*), 46.7% (*D. salina* CONC-001), 47.1% (*D. viridis* CONC-002), and 55% (*P. uvella*). These elevated GC values are not entirely unexpected: the Chlamydomonadales is known to harbour species with exceptionally high mtDNA GC compositions, including the colorless alga *Polytomella capuana* (57.2%) as well as some members of the *Lobochlamys* genus (~50-65%) (Smith 2012). That said, *D. salina* CCAP 19/18, unlike its Chilean counterparts, has a low mitochondrial GC content (34.4%), underscoring that organelle nucleotide content can differ drastically even among closely related species and strains.

Further inspection revealed even more differences among the mitochondrial genomes. The *D. viridis* CONC-002 and *D. salina* CONC-001 mtDNAs, with respective lengths of ~46 kb and ~50 kb, are around 3–4-times larger than those of *C. leiostraca* (14 kb) and *P. uvella* (17.4 kb), and on average >2-times larger than other available chlamydomonadalean mtDNAs, despite mitochondrial gene content being almost
identical across the entire lineage. What’s more, the *C. leiostraca* and *P. uvella* mtDNAs contain no introns, are densely packed (≤30% noncoding), and have matching gene orders, whereas *D. viridis* CONC-002 has 13 introns and *D. salina* CONC-001 has 17, and both species are distended with noncoding mtDNA (>70%) and have differing gene orders (Figure 2-1). In fact, the mtDNA of the Chilean *D. salina* described here is almost twice as large as the previously reported mtDNA of an Australian *D. salina* CCAP 19/18 (~50 kb vs 28.3 kb), which, like the Chilean strain, also has an abundance of introns (18) (Figure 2-1) (Smith *et al*. 2010b).

In all, 48 putative introns are distributed among the three sequenced *Dunaliella* mitochondrial genomes, representing ~65% of all identified mitochondrial introns from the Chlamydomonadales. When looking at the location of the *Dunaliella* introns, 28 have unique insertion sites (situated within four different protein-coding genes and five different rRNA-coding modules), 17 are found in at least two of the isolates, 11 are located in only a single isolate, and eight contain an intronic open reading frame (ORF) (Figure 2-1). All but one of the introns appear to be of group I affiliation, and the decaying remnants of intronic ORFs were uncovered in the intergenic DNA of both *D. salina* CONC-001 and *D. viridis* CONC-002 (Figure 2-1), suggesting a complex history of intron loss and gain throughout the evolution of *Dunaliella* mitochondria.

### 2.3.2 Genomic upheaval within *dunaliella* mitochondria.

The mitochondrial genomes of *Dunaliella* algae seem to be in a state of upheaval (Figure 2-2). They are the most bloated and intron-rich mtDNAs observed from the Chlorophyta (Figure 2-2A), have undergone substantial genomic rearrangements, and are riddled with short, simple repeats, which have spread throughout the intergenic, intronic and, in some cases, coding regions (Figures 2-2B and 2-2C). These repeats differ in sequence among the three *Dunaliella* mtDNAs, but can be folded into similar hairpin structures (Smith *et al*. 2010b), and in some respects resemble the palindromic organelle repeats from *V. carteri* (Smith and Lee 2009). Repeat-like insertions were also uncovered within coding regions of the *Dunaliella* mtDNAs, resulting in elongated exonic sequences
Figure 2-1: Mitochondrial genomic architecture and expansion within the Chlamydomonadales. Tree of chlamydomonadalean algae, showing mitochondrial genome conformation, size, intron content, and expansion (red); branching order based on the phylogenetic analyses of Nakada et al. (2008), González et al. (2009), Smith et al. (2013a), and Figueroa-Martínez et al. (2015), as well as that in Appendix D. Venn diagram highlighting shared and unique introns (based on insertion sites) among the three available Dunaliella mtDNAs. Mitochondrial genome maps for Dunaliella salina CONC-001 (outer) and C. leiostraca (inner). D. salina CONC-001 and D. salina CCAP 19/18 have identical mtDNA gene orders and contents (not including introns, intronic open reading frames, or pseudogenes), and so do C. leiostraca and P. uvella. Breakpoints in mitochondrial gene synteny between D. salina and D. viridis are marked with a double-diamond symbol (red). Note: the mitochondrial genome size and intron number for C. reinhardtii and V. carteri can vary due to optional introns in some strains (Smith et al. 2013b). Superscript 1 indicates Chlorogonium capillatum SAG 12-2e was formerly called Chlorogonium elongatum SAG 12-2.
relative to those from other closely related mtDNAs (Figure 2-2B). All of this is in stark contrast to the compact and “ordered” mtDNAs of *C. leiostraca* and *P. uvella*, which are devoid of introns and have few repeats (Figures 2-1, 2-2A and 2-2C).

Given what we know about the Chlamydomonadales, the most recent common ancestor of the group likely had a compact, circular-mapping, intron-poor mtDNA (Figure 2-1). However, at some point after the divergence of the *Dunaliella* and *Polytoma/C. leiostraca* lineages, the former experienced severe mtDNA inflation (Figure 2-1), characterized by the proliferation of intronic and repetitive DNA (Figures 2-2A and 2-2C). Mitochondrial genomic expansion is also observed in other chlorophyte lineages, including the volvocine line (Smith et al. 2013b) and certain members of the Sphaeropleales (Fučíková et al. 2014), but it is not as pronounced as that within *Dunaliella*, and is mostly a product of increases in intergenic sequence rather than a combination of repeats, introns, and other kinds of genomic embellishments.

Although all three of the available *Dunaliella* mitochondrial genomes are expanded, there is, nonetheless, an impressive amount of variation in size, intron, and noncoding content among them (28–50 kb; 13–18 introns per genome; ~60–75% noncoding DNA) (Figures 2-1 and 2-2A). As is apparent from self-similarity dot plots (Figure 2-2C), repeat content scales positively with mitochondrial genome size across the *Dunaliella* genus, and the order as a whole (Smith et al. 2013b). Other types of genomic embellishments also go up in abundance relative to mtDNA size. For example, the Chilean *D. salina* has more nonstandard ORFs and pseudogenes than its Australian counterpart (Figure 2-1). The same, however, cannot be said for overall intron number, which again is highest in the Australian isolate, reinforcing that intron abundance alone does not account for the inflated architectures of the *Dunaliella* mitochondrial genomes.

### 2.3.3 Co-expansion of the chloroplast genomes.

Using the same datasets employed for the mitochondrial genome assemblies, we explored the plastid genomic architectures of *D. salina* CONC-001 and *D. viridis* CONC-002
Figure 2-2: Mitochondrial genomic upheaval in *Dunaliella*. (A) Noncoding content (x-axis) vs intron abundance (y-axis) for chlamydomonadalean mtDNAs. Noncoding statistics were calculated following the methods of Smith et al. (2010). *C. leiostraca*, *P. uvella*, and the three *Dunaliella* isolates are marked on plot. (B) Insertions within *D. salina* CONC-001 (red) and *D. viridis* (blue) mtDNA protein-coding genes relative to *D. salina* CCAP 19/18. These insertions are also absent from the *C. leiostraca* mtDNA. (C) Dot plot similarity matrices of chlamydomonadalean mitochondrial genomes. Each matrix contains an mtDNA sequence plotted against itself (size of the genome is marked in the bottom right corner). Dots within the matrix highlight regions of nucleotide sequence similarity. The main diagonal represents the mtDNA on the x-axis matching against its partner on the y-axis. Dots adjacent to the main diagonal correspond to repetitive DNA. Plots were generated with JDotter (Brodie et al. 2004), using a plot size of 1,000 bases/pixel and a sliding window size of 50.
(those of *C. leiostraca* and *P. uvella*, which are relatively compact, will be described elsewhere). The plastid DNAs (ptDNAs) of CONC-001 and CONC-002 appear to be equally or even more expanded than the neighboring mitochondrial genomes. *De novo* assemblies of paired-end Illumina reads from each of the algae gave dozens of short (~0.5–17 kb) ptDNA contigs (Table 2-1). For both *Dunaliella* isolates, the plastid contigs were AT-rich, had one to a few genes apiece, harbored many introns, and contained extensive repeats, which were found in almost all of the identified intronic and intergenic regions (Table 2-1). These repeats prevented the assembly of larger contigs and the bridging of smaller ones, which is a recurring problem in green algal plastid genomics—and one that is hampering the assembly of the *D. salina* CCAP 19/18 nuclear genome (United States Department of Energy Joint Genome Institute). For example, palindromic repeats hindered the assembly of the ~525 kb plastid chromosome of *V. carteri* (Smith and Lee 2009), and the recently sequenced ptDNA of the ulvophyte *Acetabularia acetabulum*, which is >1 Mb and repeat-rich, resulted in a highly fragmented assembly (>60 contigs) (de Vries et al. 2013).

Based on the number, proportion, and density of genes identified on the ptDNA contigs as well as the accumulative size of these contigs (Table 2-1), we estimate that the Chilean *D. salina* and *D. viridis* plastid genomes are at least 370 kb and 280 kb, respectively, and are possibly much larger, which makes them giants among all available ptDNAs. This is consistent with the previously published *D. salina* CCAP 19/18 ptDNA (assembled with Sanger sequencing reads) (Smith et al. 2010b), which at ~265 kb and >65% noncoding, is one of the largest ptDNAs ever observed. Like the corresponding mitochondrial genomes, the CONC-001 and CONC-002 ptDNAs are intron-rich (CONC-001 likely has >30 introns), and repeats have infiltrated both the intronic and the intergenic regions. But unlike the mtDNAs, the plastid introns are mostly of group II affiliation and the plastid repeats do show sequence similarities across the three ptDNAs.

Together, these data are an excellent illustration of how mitochondrial and plastid genomes can arrive at similar extremes in a single organism or cell—in this case, the co-expansion of mtDNA and ptDNA within members of the *Dunaliella* genus. Similar observations have come from other eukaryotes, including *V. carteri* for which both the
mtDNA and the ptDNA have uncharacteristically long intergenic regions and large amounts of repetitive DNA (Smith and Lee 2009; Smith and Lee 2010). The trend can also go in the opposite direction. For instance, the mitochondrial and plastid genomes of many prasinophyte algae, such as Ostreococcus tauri, are paragons of compactness (Robbens et al. 2007), as are those of the red alga Cyanidioschyzon merolae (Ohta et al. 2003). Convergent evolution between mtDNA and ptDNA can be seen throughout the eukaryotic tree, and in many cases both organelle genomes have independently evolved the same features and taken on similar genomic embellishments. However, when this is observed, the intensity of genomic embellishment is typically more pronounced in mitochondria than in plastids (Barbrook et al. 2010; Smith and Keeling 2015). But this does not appear to be true for Dunaliella: the ptDNA has been pushed to an equivalent or greater extreme than the mtDNA, at least in terms of noncoding content.

2.3.4 Genetic divergence: high in the mitochondrion, low in the plastid

To better understand the evolution of expanded Dunaliella organelle DNAs, we studied the levels of genetic divergence between them. Using a maximum-likelihood approach, we measured the rates of organelle nucleotide substitution between D. salina CONC-001 and CCAP 19/18, which revealed huge, unparalleled differences in mtDNA versus ptDNA divergence for green algae (Table 2-2; Appendix A and B).

Substitution rates between CONC-01 and CCAP 19/18 were 2–14 times greater in the mtDNA than in the ptDNA (Table 2-2). The average number of substitutions per synonymous site ($d_S$) for mtDNA protein-coding genes (~1.2) was about 13 times that of ptDNA proteins (0.09). Concatenated gene datasets gave an identical trend (Table 2-2; Appendix A and B). The rates of substitution at nonsynonymous sites ($d_N$) followed a similar pattern: $d_N$ of the mtDNA (0.04) was approximately an order of magnitude greater than that of the ptDNA (0.005); however, both organelle genomes had low $d_N/d_S$ ratios (<0.1) (Table 2-2), indicating strong purifying selection at nonsynonymous sites. Substitution rates for rRNA genes were also greater for the mtDNA than for the ptDNA (0.1 vs 0.06) (Table 2-2).
Table 2-1: *Dunaliella salina* CONC-001 and *D. viridis* CONC-002 plastid genome assembly statistics.

<table>
<thead>
<tr>
<th>Plastid DNA Contig Statistics</th>
<th><em>D. salina</em> CONC-001</th>
<th><em>D. viridis</em> CONC-002</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number</td>
<td>80</td>
<td>72</td>
</tr>
<tr>
<td>Size range (kb)</td>
<td>0.6–8.9</td>
<td>0.5–16.8</td>
</tr>
<tr>
<td>Avg. size (kb)</td>
<td>2.2</td>
<td>2.4</td>
</tr>
<tr>
<td>Avg. read coverage/base</td>
<td>43.5</td>
<td>31.2</td>
</tr>
<tr>
<td>Avg. number genes/contig&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.8</td>
<td>1.2</td>
</tr>
<tr>
<td>Proportion of genes identified (%)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>65</td>
<td>88</td>
</tr>
<tr>
<td>Overall intron count</td>
<td>32</td>
<td>13</td>
</tr>
<tr>
<td>Accumulative size (kb)</td>
<td>178.3</td>
<td>171.8</td>
</tr>
<tr>
<td>Predicted plastid genome size (kb)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>&gt;370 kb</td>
<td>&gt;280 kb</td>
</tr>
</tbody>
</table>

<sup>a</sup>Number is <1 because in some cases a single gene is distributed over multiple contigs.

<sup>b</sup>Percentage of plastid-encoded genes identified in the CONC-001 and CONC-002 assemblies relative to the genes found in the completely sequenced *D. salina* CCAP 19/18 ptDNA. Does not include non-standard genes, such as intronic open reading frames; duplicate genes, such as the rRNAs, were counted only once.

<sup>c</sup>Assuming missing genes are found on single ~2 kb contigs and an average gap between contigs of ~1 kb.
Table 2-2: Mitochondrial and plastid DNA (mtDNA and ptDNA) substitution rates for two geographically distinct isolates of *Dunaliella salina*: CONC-001 (Chile) and CCAP 19/18 (Australia).

<table>
<thead>
<tr>
<th></th>
<th>Substitutions per site</th>
<th>Substitution rate ratios (pt : mt)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ptDNA</td>
<td>mtDNA</td>
</tr>
<tr>
<td>Synonymous sites</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average (SD)</td>
<td>0.09 (0.32)</td>
<td>1.164 (0.52)</td>
</tr>
<tr>
<td>Concatenation</td>
<td>0.074</td>
<td>0.989</td>
</tr>
<tr>
<td>Nonsynonymous sites</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average (SD)</td>
<td>0.005 (0.02)</td>
<td>0.043 (0.02)</td>
</tr>
<tr>
<td>Concatenation</td>
<td>0.004</td>
<td>0.041</td>
</tr>
<tr>
<td>$d_S/d_S$ (SD)</td>
<td>0.084 (0.18)</td>
<td>0.042 (0.02)</td>
</tr>
<tr>
<td>rRNAs$^a$</td>
<td>0.056</td>
<td>0.107</td>
</tr>
<tr>
<td>Introns</td>
<td>&gt;&gt;0.1</td>
<td>Unalignable</td>
</tr>
<tr>
<td>Intergenic regions</td>
<td>&gt;&gt;0.1</td>
<td>Unalignable</td>
</tr>
</tbody>
</table>

SD: standard deviation; $d_S/d_S$: ratio of nonsynonymous to synonymous substitutions per site, based on averages of individual loci not concatenated datasets. The substitution rate statistics for the individual loci within mitochondrial and plastid genomes are shown in Supplementary Table S1.

$a$For mtDNA and ptDNA includes the concatenation of all rRNA-coding regions.
We tried measuring substitutions within noncoding DNA as well. The mitochondrial introns and intergenic regions from CONC-001 and CCAP 19/18 were unalignable, implying very high levels of substitution (>>1 per site), much higher than those observed for the mtDNA synonymous sites. Although many of the plastid noncoding regions were also unalignable, we were able to align ten complete ptDNA intergenic spacers, which when concatenated (4.6 kb) harbored ~0.15 substitutions per site (Table 2-2), which is greater than that observed for plastid synonymous sites. Like with the coding data, these findings are consistent with a higher rate of nucleotide substitution in the mitochondrion as compared to the plastid. Similar overall conclusions come from substitution rate analyses of *D. salina* versus *D. viridis*, but the levels of substitution are saturated in both compartments making it difficult to gauge the relative rates of substitution between the mtDNA and ptDNA, and between synonymous and noncoding sites.

Organelle substitution rate statistics are available for a number of plastid-bearing eukaryotes, including various green algae (Appendix C) (Hua *et al.* 2012). Compared to other species, the levels of synonymous-site divergence between CONC-001 and CCAP 19/18 are high for the mtDNA and low for the ptDNA. For example, the average synonymous-site divergence between *Chlamydomonas globosa* SAG 7.73 (formerly called *C. incerta*) and *C. reinhardtii* is ~0.30 for both mitochondrial- and plastid-located genes (Hua *et al.* 2012), which contrast sharply with the values from *D. salina*: ~1.2 for the mtDNA and 0.09 for the ptDNA. Moreover, the relative levels of *dS* between the mitochondrial and plastid compartments of *D. salina* (13:1) are among the highest yet observed from green, red, or glaucophyte algae (Appendix C).

What do these extreme differences in *dS* mean? At the very least, the high levels of substitution within the mtDNA suggest that *D. salina* CONC-001 and CCAP 19/18 represent distinct populations or “species” (González *et al.* 2009). They also point towards major differences in the organelle mutation rates. If synonymous nucleotide positions are assumed to be neutrally evolving, then the synonymous-site divergence between species or distinct populations can provide an entrée into mutation rate (Kimura 1983). For *D. salina*, there is a 13-fold difference in *dS* between mitochondrial-located versus plastid-located genes (Table 2-2), indicating that there is much higher mutation
rate in the mitochondrion than the plastid. These findings could also be a sign of high and low absolute mutation rates within mitochondrion and plastid, respectively, but this is speculative as we do not know how long ago CONC-001 and CCAP 19/18 shared a common ancestor. If these two compartments do have drastically different mutational patterns, then it would, on the face of it, conflict with their similarly expanded genomic architectures.

2.3.5 Unraveling the roots of organelle genomic expansion.

In many respects, the *Dunaliella* organelle genomes have parallel architectures to land plant mtDNAs, which are renowned for their expansive intergenic regions, large densities of introns and repeats, and overall high levels of sequence upheaval (Sloan *et al.* 2012). Similar to the *Dunaliella* organelle genomes, land plant mtDNAs boast impressive variations in synonymous substitution rates, both within and among genomes (Sloan *et al.* 2012; Richardson *et al.* 2013; Zhu *et al.* 2014). For instance, the enormous, multichromosomal mtDNAs of *Silene conica* and *S. noctiflora* have extraordinarily high synonymous substitution rates (Sloan *et al.* 2012), whereas the tulip tree has one of the most mutationally quiescent mitochondrial genomes of any eukaryote (Richardson *et al.* 2013). Land plant mtDNAs, as with *Dunaliella*, also show vastly different rates of substitution in coding versus noncoding regions, a feature that has provided major insights into the process of organelle genome expansion (Christensen 2013).

After finding that the modes of molecular evolution differ between coding and noncoding regions in *Arabidopsis* mtDNA, Christensen (2013) proposed that land plants employ two types of mtDNA repair, each of which has shaped mitochondrial genomic architecture: “Within genes, a bias toward gene conversion would keep measured mutation rates low, whereas in noncoding regions, break-induced replication (BIR) explains the expansion[s] and rearrangements. Both processes are types of double-strand break repair, but enhanced second-strand capture in transcribed regions versus BIR in non-transcribed regions can explain the two seemingly contradictory features of plant mitochondrial genome evolution—the low mutation rates in genes and the striking expansions of noncoding sequences”.
The same argument can be made for the Dunaliella mitochondrial and plastid chromosomes. Both organelle DNAs have much lower substitution rates within synonymous sites as compared to the noncoding regions. This is especially apparent for the ptDNA, which has an average $d_s$ of only 0.09 but for which most of the intergenic regions are unalignable (Table 2-2). If the intergenic regions in the Dunaliella organelle genomes are repaired via BIR it would help explain the large amounts of genomic expansion and rearrangements observed between them (Figure 2-1) as well as the widespread intergenic turmoil (Figures 2-2B and 2-2C). Indeed, BIR within organelle systems is known to be inaccurate and cause rearrangements, chimeric genes, and expansions (Davila et al. 2011). Conversely, accurate repair of coding organelle DNA in Dunaliella, by homologous recombination or gene conversion, for example, would account for the comparatively low synonymous substitution rates, particularly in the plastid.

Nevertheless, there is still an order of magnitude variation in $d_s$ for the mitochondrial versus plastid protein-coding genes of D. salina, indicating that the efficiency of DNA repair, be it by BIR or gene conversion or homologous recombination, differs greatly between these compartments. In plants and algae, virtually all of the organelle DNA maintenance machineries are nuclear encoded, and their proficiency are known to vary between species and compartments (Sloan and Taylor 2012). A Dunaliella nuclear genome sequence is not yet available, but work is presently underway by the DOE JGI to generate one. Investigations of nuclear-encoded, organelle-targeted DNA repair proteins will likely give further insights into the evolution of the Dunaliella organelle genomes. There is also the potential that the extremely salty habitats in which many Dunaliella species reside is in some way impacting the molecular evolution of their organelle DNAs. Whatever the root cause of their inflated architectures, the Dunaliella mitochondrial and plastid genomes are veritable heavyweights among green algal organelle genomes.

2.4 References


Christensen AC. 2013. Plant mitochondrial genome evolution can be explained by DNA repair mechanisms. Genome Biol Evol. 5:1079–1086.


Chapter 3

3 Sequence diversity within the mitochondrial genes of the green algal genus *Dunaliella*

3.1 Introduction

3.1.1 Genomic diversity

Diversity can occur on at least two different levels: the large scale, which includes genomic diversity (e.g., differences in genome size), and the small scale, which includes genetic diversity (e.g., nucleotide polymorphisms). Throughout evolution, changes occur within and among organisms that give rise to genomic and genetic diversity, not only between different species or populations, but even within populations (e.g., heteroplasmy). However, the rate of genetic and genomic change varies from one organism to the next.

Mutation along with other processes, such as genetic drift, migration and natural selection, drive evolution and diversity of organelle genomes, but it can occur at different rates (Lynch 2007). For instance, mutation rates are very low in the mitochondria of land plants, whereas animal mitochondria experience high rates of mutation (Lynch *et al.* 2006). The major differences in mutation between land plant and animal mtDNAs have been used to explain very different genomic architectures between these two groups in what is known as the mutational hazard hypothesis (Lynch *et al.* 2006). This theory predicts that genomes with large amounts of noncoding DNA—bloated genomes—will have lower levels of genetic diversity because of a lower underlying mutation rate (Lynch *et al.* 2006; Lynch 2007). Indeed, according to this theory, large amounts of noncoding DNA are a mutational burden for organisms because noncoding DNA represent targets for potential deleterious mutations (Lynch *et al.* 2006). Noncoding DNA are regions of a genome composed of introns and the spaces between protein coding-, tRNA- and rRNA-genes, referred to as intergenic DNA. Having a lower mutation rate is a way of compensating for this burden. It may seem counterintuitive that noncoding as opposed to coding DNA can be a burden. Introns, which are a type of noncoding DNA, are a good
example. Mitochondrial introns are typically autocatalytic and rely on their secondary structure for proper splicing. A mutation within a mitochondrial intron that interfered with the secondary structure it could alter the splicing of the intron and cause the expression of a non-functional protein. In support of the mutational hazard hypothesis, land plant mitochondrial genomes have high numbers of introns and experience some of the lowest levels of diversity (Wolfe et al. 1987). Conversely, animal mtDNAs, which are renowned for having elevated rates of mutation, are almost devoid of introns.

Diversity in a sequence, whether it is at the protein or DNA level, can have an influence on the product based on that sequence. Throughout evolution, diversity has occurred in DNA sequences by: simple point mutations, insertions, deletions and rearrangements (Arodź and Płonka 2013). Some mutations do not affect the protein made by a gene due to the redundancy of the genetic code, multiple codons code for the same amino acid. However, mutations causing the coding for different amino acids happen as well. Diversity at the protein level is a result of an amino acid substitution in the protein sequence. Based on the side chains of amino acids, amino acids are categorized into four types: polar, nonpolar, acidic and basic. An amino acid substitution can cause either the same or different type of amino acid to be replaced in the sequence. A change for a different type of amino acid might cause a large effect on the tertiary structure of the protein, but studies have shown that while amino acid sequences do differ between organisms, portions of the protein structure are conserved (Orengo and Thornton 2005).

3.1.2 Research objectives

To measure sequence diversity between different Dunaliella isolates, I compared the mitochondrial gene and protein sequences of two distinct geographical isolates of D. salina—one from Chile (CONC-001) and one from Australia (CCAP 19/18)—as well as a Chilean strain of D. viridis (CONC-002) (see Chapter 2 for details on these strains). My objective was to gain insights into the relationships between genetic diversity and geographical origin of these three algae. More specifically, I wanted to see if the two strains of D. salina (CCAP 19/18 and CONC-001), whose origins are thousands of kilometers apart, are genetically similar to one another, or instead if D. salina CONC-001 and D. viridis CONC-002, which were isolated from the same pond (the La Rinconada
hypersaline pond in the Atacama Desert) on the same day are more similar to each other than to *D. salina* CCAP 19/18, even though they have been identified as distinct species. Some hesitancy of identification and relatedness occur when dealing with these particular organisms because no previous genetic analyses comparing these three strains exist and that the classification of CONC-001 and CONC-002 as *D. salina* and *D. viridis*, respectively, is based primarily on morphological and physiological features (Gómez-Silva et al. 1990; González et al. 1999; Gómez and González 2005). Therefore one might expect the two geographical strains of *D. salina* (CONC-001 and CCAP 19/18) to be more similar to each other than to *D. viridis* CONC-002, but if misidentification occurred this result may not be true.

### 3.2 Methods

#### 3.2.1 Mitochondrial genome sequencing, assembly, and annotation.

Please refer to Chapter 2 section 2.2.

#### 3.2.2 Sequence diversity

Pairwise DNA alignment of mitochondrial gene sequences of the three *Dunaliella* species was carried out with ClustalW (Thompson *et al.* 1994) implemented through Geneious v7.0 (Biomatters Ltd., Auckland, NZ). The parameters were set as follows: a gap open cost of 15 and a gap extend cost of 6.66. All introns were excluded from the alignment as they caused incorrect alignment with the gene sequences to occur. Multiple alignment was performed on all three organisms as well as a pairwise alignment was performed on the following pairs: *D. salina* CONC-001 vs. *D. viridis* CONC-002, *D. salina* CCAP 19/18 vs. *D. salina* CONC-001, and *D. salina* CCAP 19/18 vs. *D. viridis* CONC-002. Multiple and pairwise alignments align the sequences, highlighting where differences in sequence occur and gives a percentage of how identical the sequences are. For the mitochondrial protein alignments, the DNA sequences were first translated into protein sequences in Geneious 7.0 using the standard genetic code aligned using ClustalW via Geneious 7.0 using a gap open cost of 10 and a gap extend cost of 0.1.
3.2.3 Tertiary protein structure

To determine if protein sequence diversity has an effect on the structure and folding of the proteins, the server RaptorX was used (Peng and Xu 2011a; Peng and Xu 2011b; Källberg et al. 2012; Ma et al. 2013). The protein sequences of all seven protein-coding genes were submitted to the web server RaptorX, which predicts the 3D structure of a protein. RaptorX aligns the query to a template or multiple templates, giving each template a value of the quality of the template. Tests preformed on each template to assign quality are: p-value, representing the quality of the template, uGDT and GDT, unnormalized Global Distance Test, which measures the absolute model quality (Peng and Xu 2011a). The smaller p-value means a higher quality template, whereas the higher the uGDT and GDT value the better the template, where GDT ranges from 1-100 (Peng and Xu 2011a). It then ranks the templates from highest to lowest based on the p-value and uGDT/GDT value. These templates come from known crystalized protein structures that are in the RaptorX database. In addition to alignment, RaptorX also assigns a NEFF value to a protein (Peng and Xu 2011b). "The NEFF value ranges from 1-20 and can be interpreted as the expected number of amino acid substitutions at each position" (Peng and Xu 2011b). The NEFF value is useful because it gives information about the sequence of a protein, and shows the importance of structural information over homologous information (Peng and Xu 2011b). RaptorX web server is found at http://raptorx.uchicago.edu/. This web server does not require the user to set any parameters or special settings, for the input sequences.

3.3 Results

3.3.1 Sequence diversity

The mtDNA sequences of *D. salina* CONC-001 and *D. viridis* CONC-002 were described in Chapter 2. But details on their mitochondrial gene order were only briefly highlighted and examined. Therefore, I examined in closer detail the arrangement of genes within the mitochondrial genomes of these two taxa. I found that the two *D. salina* strains (CONC-001 and CCAP 19/18) have almost identical mitochondrial gene orders, with only a minor rearrangement involving the location of large subunit ribosomal RNA
fragment 4, *rrnL4*. In the mtDNA of *D. salina* CONC 001 *rrnL4* is positioned between large subunit ribosomal RNA fragment 3, *rrnL3* and large subunit ribosomal RNA fragment 2, *rrnL2*, whereas in *D. salina* CCAP 19/18 *rrnL4* is positioned between small subunit ribosomal RNA fragment 1, *rrnS1*, and *trnQ*, tRNA representing glutamine, (Figure 3-1). It is important to note within the chlamydomonadalean mtDNAs the *rrnL4* gene is known for being hard to identify and there remains the possibility that this gene was incorrectly annotated in *D. salina* CCAP 19/18. Conversely, the overall mitochondrial gene arrangement of *D. viridis* CONC-002 was significantly different than those of *D. salina* CONC-001 and CCAP 19/18, but some similarities were observed among the three strains (Figure 3-2).

By using colour-coded blocks to differentiate regions of genes, Figure 3-2 outlines and compares the mitochondrial gene order of *D. salina* CCAP 19/18 and CONC-001 (*rrnL4* notwithstanding) to that of *D. viridis* CONC-002. The figure highlights four large sections of genes (each shown in a different colour) that have are rearranged between the *D. salina* and *D. viridis* CONC-002 mitochondrial genomes. Interestingly, the gene order within each of the four segments remains the same between *D. salina* and *D. viridis* CONC-002 (Figure 3-2). Ultimately, only two inversion events separate the gene order of *D. salina* CCAP 19/18 and *D. viridis* CONC-002, with a potential relocation of *rrnL4* in CONC-001 (Figure 3-2).

### 3.3.2 Noncoding content

As noted earlier, noncoding mtDNA regions can be separated into introns and intergenic regions. The majority of the *Dunaliella* mitochondrial genomes are made up of noncoding DNA. For *D. viridis*, 34.45 kb (75%) of its mitochondrial genome is noncoding, with introns making up 8.47 kb (18%), whereas the *D. salina* CONC-001 contains 34.98 kb (70%) of noncoding DNA, with introns contributing 15.75 kb (45%) of the noncoding DNA. Although there is an increase in the amount of noncoding mtDNA in both of the Chilean strains (CONC-001 and CONC-002) relative to *D. salina* CCAP 19/18, it does not correlate with an increase in intron number (Table 3-1).
The mtDNA of *D. viridis* CONC-002 only contains 13 introns, five fewer than *D. salina* CCAP 19/18 (Table 3-1). All 13 of the *D. viridis* CONC-002 introns are group I introns, and two of the 13 introns contain ORFs encoding for maturase-like proteins. Eleven of the mitochondrial introns in *D. viridis* CONC-002 are shared by *D. salina* CCAP 19/18.

**D. salina Chile**

![Gene order of D. salina Chile](image1)

**D. salina Australia**

![Gene order of D. salina Australia](image2)

Figure 3-1: Gene order differences between *D. salina* CONC-001 and *D. salina* CCAP 19/18. Single letter abbreviation of genes is used as follows: S3 (small subunit ribosomal RNA fragment 3), L3 (Large subunit ribosomal RNA fragment 3) L4 (large subunit ribosomal RNA fragment 4), L2 (large subunit ribosomal RNA fragment 2), W (tRNA representing tryptophan), Q (tRNA representing glutamine) and L1 (large subunit ribosomal RNA fragment 1). Only a portion of the gene order is shown and the genome is portrayed as a linear molecular instead of circular for simplicity. Size of blocks is not to scale with the length of the genes, used to highlight the only difference in gene order. The gene *rrnL4* is coloured blue to emphasize where the difference in gene order occurs.
Figure 3-2: Gene order of *D. salina* CCAP 19/18 and *D. viridis* CONC-002. Gene sizes are not to scale and genome is shown as linear for simplicity. Same size coloured blocks are used to visualize the difference in gene order between the two organisms, *D. salina* CCAP 19/18 and *D. viridis* CONC-002. Gene abbreviations are used as follows: *rrnL3* (large subunit ribosomal RNA fragment 3), *rrnL2* (large subunit ribosomal RNA fragment 2), *trnW* (tRNA representing tryptophan) *trnQ* (tRNA representing glutamine), *rrnL4* (large subunit ribosomal RNA fragment 4), *rrnS1* (small subunit ribosomal RNA fragment 1), *rrnL1* (large subunit ribosomal RNA fragment 1), *rrnL6* (large subunit ribosomal RNA fragment 6), *cob* (cytochrome b), *rrnS2* (small subunit ribosomal RNA fragment 2), *nad6* (NADH dehydrogenase subunit 6), *nad5* (NADH dehydrogenase subunit 5), *trnM* (tRNA representing methionine), *rrnL5* (large subunit ribosomal RNA fragment 5), *nad1* (NADH dehydrogenase subunit 1), *nad4* (NADH dehydrogenase subunit 4), *cox1*: (cytochrome c oxidase I), *nad2* (NADH dehydrogenase subunit 2) *rrnS3* (small subunit ribosomal RNA fragment 3).
Table 3-1: Location and number of introns in *D. salina* CCAP 19/18, *D. salina* CONC-001 and *D. viridis* CONC-002.

<table>
<thead>
<tr>
<th>Species</th>
<th>Number of introns per gene</th>
<th>Total number of introns</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cob</td>
<td>cox1</td>
</tr>
<tr>
<td><em>D. salina</em> CCAP 19/18</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td><em>D. salina</em> CONC-001</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td><em>D. viridis</em> CONC-002</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>
and two are unique. The two unique introns are located in cytochrome c oxidase I, cox1, and small subunit ribosomal RNA fragment 3, rrnS3 (Table 3-1). For cox1 there are six introns in D. viridis CONC-002, whereas there are only five in D. salina CCAP 19/18, and an addition of an intron in rrnS3, which is a gene that does not contain any introns in D. salina CCAP 19/18.

D. salina CONC-001 contains 17 mitochondrial introns—one fewer than its Australian counterpart (Table 3-1). Sixteen of the 17 introns are group I introns and one is a group II intron. Of the 17 introns, four of them contain intronic ORFs, encoding for maturase-like proteins. Fourteen introns are shared between both strains of D. salina (CCAP 19/18 and CONC-001) and three are unique to D. salina CONC-001. rrnS3 and large subunit ribosomal RNA fragment 4, rrnL4, contain the three unique introns, 2 and 1 introns respectively, where no introns are present in these genes for D. salina CCAP 19/18 (Table 3-1).

3.3.3 Sequence diversity

Sequence diversity among the Dunaliella mtDNAs can be observed on various levels, including the lengths of the genes. For example, a common theme for almost all of the 19 coding regions (representing 12 genes) in the Dunaliella mtDNAs is that the two Chilean species have longer coding regions than their Australian counterpart. However, there are exceptions to this pattern. Six of the 19 coding regions in D. viridis CONC-002 are shorter than those of both D. salina strains (CCAP 19/18 and CONC-002). Also, three of the coding regions in D. salina CCAP 19/18 (all of them rRNAs) are longer than the corresponding regions from the Chilean species. Diversity was also observed in the sequences themselves.

To assess the levels of Dunaliella mtDNA sequence diversity, four different alignment and comparison tests were performed: 1) each gene from all three organisms, 2) each gene from the two D. salina strains (CCAP 19/18 and CONC-001), 3) each gene from D. salina CONC-001 and D. viridis CONC-002 and 4) each gene from D. salina CCAP/19/18 and D. viridis CONC-002 (Table 3-2). Similarity was determined for all
Table 3-2: Pairwise identity of all coding regions in the *D. salina* CCAP 19/18, *D. salina* CONC-001 and *D. viridis* CONC-002 mtDNAs.

<table>
<thead>
<tr>
<th>Coding region</th>
<th>Pairwise Identity (%) of <em>D. salina</em> CONC-001 and <em>D. viridis</em> CONC-002</th>
<th>Pairwise Identity (%) of <em>D. salina</em> CCAP 19/18 and <em>D. salina</em> CONC-001</th>
<th>Pairwise Identity (%) of <em>D. salina</em> CCAP 19/18 and <em>D. viridis</em> CONC-002</th>
<th>Multiple alignment (%) of all three organisms</th>
</tr>
</thead>
<tbody>
<tr>
<td>cob</td>
<td>74.3</td>
<td>88.1</td>
<td>72.6</td>
<td>78.4</td>
</tr>
<tr>
<td>cox1</td>
<td>80.7</td>
<td>88.3</td>
<td>78.6</td>
<td>82.5</td>
</tr>
<tr>
<td>rrnL1</td>
<td>78.4</td>
<td>95.0</td>
<td>75.6</td>
<td>82.9</td>
</tr>
<tr>
<td>rrnL2</td>
<td>66.6</td>
<td>87.4</td>
<td>73.9</td>
<td>75.9</td>
</tr>
<tr>
<td>rrnL3</td>
<td>79.7</td>
<td>93.7</td>
<td>79.7</td>
<td>84.3</td>
</tr>
<tr>
<td>rrnL4</td>
<td>59.9</td>
<td>75.5</td>
<td>62.0</td>
<td>64.4</td>
</tr>
<tr>
<td>rrnL5</td>
<td>75.8</td>
<td>85.6</td>
<td>73.8</td>
<td>77.7</td>
</tr>
<tr>
<td>rrnL6</td>
<td>73.1</td>
<td>81.4</td>
<td>71.5</td>
<td>74.7</td>
</tr>
<tr>
<td>rrnS1</td>
<td>80.7</td>
<td>75.3</td>
<td>79.5</td>
<td>77.1</td>
</tr>
<tr>
<td>rrnS2</td>
<td>71.3</td>
<td>88.1</td>
<td>73.2</td>
<td>77.4</td>
</tr>
<tr>
<td>rrnS3</td>
<td>67.2</td>
<td>80.2</td>
<td>62.3</td>
<td>69.5</td>
</tr>
<tr>
<td>trnM</td>
<td>88.2</td>
<td>98.7</td>
<td>89.5</td>
<td>92.1</td>
</tr>
<tr>
<td>trnQ</td>
<td>77.0</td>
<td>95.9</td>
<td>77.0</td>
<td>82.9</td>
</tr>
<tr>
<td>trnW</td>
<td>80.8</td>
<td>79.5</td>
<td>86.3</td>
<td>82.2</td>
</tr>
<tr>
<td>Coding region</td>
<td>Pairwise Identity (%) of D. salina CONC-001 and D. viridis CONC-002</td>
<td>Pairwise Identity (%) of D. salina CCAP 19/18 and D. viridis CONC-002</td>
<td>Pairwise Identity (%) of D. salina CONC-001 and D. salina CCAP 19/18</td>
<td>Multiple alignment (%) of all three organisms</td>
</tr>
<tr>
<td>---------------</td>
<td>-------------------------------------------------</td>
<td>-------------------------------------------------</td>
<td>-------------------------------------------------</td>
<td>------------------------------------------</td>
</tr>
<tr>
<td>nad1</td>
<td>75.0</td>
<td>88.1</td>
<td>76.7</td>
<td>79.9</td>
</tr>
<tr>
<td>nad2</td>
<td>71.7</td>
<td>88.6</td>
<td>68.6</td>
<td>76.2</td>
</tr>
<tr>
<td>nad4</td>
<td>67.2</td>
<td>81.0</td>
<td>70.0</td>
<td>72.5</td>
</tr>
<tr>
<td>nad5</td>
<td>73.8</td>
<td>88.6</td>
<td>73.5</td>
<td>78.6</td>
</tr>
<tr>
<td>nad6</td>
<td>65.5</td>
<td>78.8</td>
<td>67.1</td>
<td>70.3</td>
</tr>
</tbody>
</table>
genes at both the DNA and where applicable, the amino acid level. A multiple alignment test yielded $trnM$, tRNA representing methionine, as the highest sequence similarity for all organisms (92.1%), while the lowest similarity was found in the $rrnS3$ gene sequence (69%). To look further into gene similarities, all 12 genes (distributed over 19 coding regions) were also compared between two of the three organisms at a time. Gene sequences between both geographical isolates of $D. salina$ (CCAP 19/18 and CONC-001) had higher percentages of similarity than the comparison of gene sequences between each isolate and $D. viridis$ CONC-002 (Table 3-2). The similarity between the two $D. salina$ strains (CCAP 19/18 and CONC-001) was high; in fact, 16 of the 19 coding regions were more than 80% identical at the nucleotide level. In contrast, when the gene sequences between $D. salina$ CONC-001 and $D. viridis$ CONC-002 were compared only 5 genes were more than 80% similar at the nucleotide level. These same trends were observed at the protein level as well.

Multiple and pairwise alignment was conducted on seven protein-coding genes in the $Dunaliella$ mtDNAs. As expected, this comparison yielded higher levels of sequence similarity than those observed at the DNA level for all seven gene sequences (Table 3-3). When considering both the DNA and protein-level alignments, a pattern emerges: $D. salina$ CCAP 19/18 and $D. salina$ CONC-001 strains are more similar to one another than they are to $D. viridis$ CONC-002. At the amino acid level, $cox1$ is the only gene that does not follow this pattern, showing a higher sequence similarity between $D. salina$ CONC-001 and $D. viridis$ CONC-002 (96.9%) than in the other species/strain comparisons. However, of all the protein-coding genes, $cox1$ shows the highest levels of sequence similarity among all three $Dunaliella$ isolates (95.2%). At the other end of spectrum are NADH dehydrogenase subunit 4, $nad4$, and NADH dehydrogenase subunit 6, $nad6$, which show relatively low levels of amino-acid sequence conservation among the Chilean and Australian organisms (76.6% for both genes).

Looking at the amino acid substitutions for all seven proteins I found substitutions for the same type of amino acid, ex. a polar amino acid being replaced by another polar amino acid, and substitutions for a different type of amino acid, ex. a polar amino acid being replaced by a nonpolar amino acid. Figure 3-3 shows the alignments of a portion of two
Table 3-3: Pairwise Identity at the protein and DNA level for the seven protein coding genes

<table>
<thead>
<tr>
<th>Gene</th>
<th>Pairwise Identity (%) of D. salina CONC-001 vs D. viridis CONC-002</th>
<th>Pairwise Identity (%) of D. salina CCAP 19/18 vs D. salina CONC-001</th>
<th>Pairwise Identity (%) of D. salina CCAP 19/18 vs D. viridis CONC-002</th>
<th>Multiple alignment (%) of all three organisms</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA</td>
<td>Protein</td>
<td>DNA</td>
<td>Protein</td>
<td>DNA</td>
</tr>
<tr>
<td>cob</td>
<td>74.3</td>
<td>86.8</td>
<td>88.1</td>
<td>94.4</td>
</tr>
<tr>
<td>cox1</td>
<td>80.7</td>
<td>96.9</td>
<td>88.3</td>
<td>95.1</td>
</tr>
<tr>
<td>nad1</td>
<td>75.0</td>
<td>91.6</td>
<td>88.1</td>
<td>97.0</td>
</tr>
<tr>
<td>nad2</td>
<td>71.7</td>
<td>80.1</td>
<td>88.6</td>
<td>90.5</td>
</tr>
<tr>
<td>nad4</td>
<td>67.2</td>
<td>70.5</td>
<td>81.0</td>
<td>83.4</td>
</tr>
<tr>
<td>nad5</td>
<td>73.8</td>
<td>81.2</td>
<td>88.6</td>
<td>92.3</td>
</tr>
<tr>
<td>nad6</td>
<td>65.5</td>
<td>73.7</td>
<td>78.8</td>
<td>79.4</td>
</tr>
</tbody>
</table>
Figure 3-3: Protein alignment of a portion of cox1 and nad6. Only a portion of the amino acid sequence for each protein is shown. Here single letter amino acids are shown. Boxes and arrows highlight amino acid substitutions. Red boxes and arrows represent a substitution for the same type of amino acid, whereas dark purple boxes and arrows represent a substitution for a different type of amino acid.
protein sequences, *cox1* and *nad6*, and highlights where an amino acid substitution occurs, and if it is the same or different type of amino acid that was replaced. In Figure 3-3 the second amino acid in the alignment of *cox1* is an example of a substitution for the same type of amino acid. Serine (S), a polar amino acid, is present in the protein sequence for *D. salina* CONC-001 and *D. viridis* CONC-002, but in *D. salina* CCAP 19/18 serine is replaced with tyrosine (Y), a polar amino acid. A substitution for a different type of amino acid can be seen in the second amino acid in the alignment of *nad6*. *D. salina* CONC-001 and *D. viridis* CONC-002 have tyrosine (Y) as the second amino acid in the *nad6* protein sequence, whereas *D. salina* CCAP 19/18 has isoleucine (I), a nonpolar amino acid (Figure 3-3). While there were substitutions for different types of amino acids, the majority of the substitutions occurring are substitutions for the same type of amino acid. Since it had been determined that diversity within the protein sequences does occur between these organisms, the next step was to investigate the effect of these changes on the tertiary structure of the proteins.

The predicted protein folding of *cox1* and *nad6* for all three organisms is shown in Figures 3-4 and 3-5, respectively. Tertiary structure prediction was performed on all seven protein sequences, but only the tertiary structures of the highest and lowest similarity, *cox1* and *nad6* respectively, are shown. For *cox1* the protein sequences from each of the three organisms aligned with two templates. Both strains of *D. salina* (CCAP 19/18 and CONC-001) aligned with the same two templates and the two templates were ranked in the same order. *D. viridis* CONC-002 also aligned with two templates, but only one template was the same among all three organisms. While this template was ranked number one for *D. salina* CCAP 19/18 and CONC-001, it was ranked number two for *D. viridis* CONC-002. It is important to note that all the templates matched were for *cox1* proteins in different organisms. Since only one template was consistent among all three organisms that template, which is the template of a bovine heart cytochrome c oxidase complex, it was used for comparison of predicted protein structure. P-values for this template are: 3.33e-17, 1.07e-17 and 1.09e-17 for *D. salina* CCAP 19/18, *D. salina* CONC-001 and *D. viridis* CONC-002, respectively. The GDT value of this template for all three organisms is 90. The predicted tertiary protein structures of *cox1* for all three organisms
Figure 3-4: Predicted tertiary protein structure of *cox1* for *D. salina* CCAP 19/18 (A), *D. salina* CONC-001 (B) and *D. viridis* CONC-002 (C). Red boxes show differences in folding structure.
Figure 3-5: Predicted tertiary protein structure of \textit{nad6} for \textit{D. salina} CCAP 19/18 (A), \textit{D. salina} CONC-001 (B) and \textit{D. viridis} CONC-002 (C). Red boxes highlight differences in tertiary protein structure.
are almost identical (Figure 3-4), with small differences at the middle and the C-terminus of the protein (Figure 3-4 red box). It is also important to know that the amino acid sequence that makes up this cox1 protein for all three organisms is the same size, 510 amino acids, which could aid in the similarity of the folding. The whole cox1 protein sequence length for all three Dunaliella organisms were modeled in the structural predictions. A pairwise identity test between the cox1 protein sequence of the template and the cox1 protein sequence of *D. salina* CCAP 19/18, *D. salina* CONC-001 and *D. viridis* CONC-002 yielded similarities of 55.8%, 57.4% and 56.8%, respectively. As mentioned above, cox1 is the protein that has the highest similarity between all three organisms so it is not a surprise that similar tertiary protein structures are found between the organisms. However, similar structures were also found in the least similar protein sequence.

Although nad6 is the most divergent protein (at the nucleotide and amino acid levels) among the Dunaliella isolates (Figure 3-5 red boxes), it nonetheless has a highly conserved tertiary structure. Nad6 is a different size for each of the three organisms: 157 amino acids for *D. salina* CCAP 19/18, 176 amino acids for *D. salina* CONC-001, and 163 amino acids for *D. viridis* CONC-002. The nad6 protein sequences for both strains of *D. salina* (CCAP 19/18 and CONC-001) aligned with five templates, but only four are the same templates and the template ranked first is in the only template that has same rank between the two strains. The nad6 protein sequence of *D. viridis* CONC-002 only aligned to one template, which is the same template that was ranked first for the two *D. salina* strains (CCAP 19/18 and CONC-001). All matched templates of the two *D. salina* strains (CCAP 19/18 and CONC-001) were from NADH dehydrogenase complex of other organisms. The template used for the prediction of nad6 is from the NADH dehydrogenase complex of *Escherichia coli*. p-values of *D. salina* CCAP 19/18, *D. salina* CONC-001 and *D. viridis* CONC-002 for this template are: 3.77e-3, 2.86e-3 and 3.15e-3, respectively. GDT values for this template are: 85, 77 and 84 for *D. salina* CCAP 19/18, *D. salina* CONC-001 and *D. viridis* CONC-002, respectively. The whole nad6 protein sequence length for all three Dunaliella organisms were modeled in the structural predictions. A pairwise identity test between the nad6 protein sequence of the template
and the nad6 protein sequence of *D. salina* CCAP 19/18, *D. salina* CONC-001 and *D. viridis* CONC-002 yielded similarities of 24.4%, 21.5% and 23%, respectively.

3.4 Discussion

3.4.1 Sequence diversity

As predicted, an overall higher sequence similarity between the two *D. salina* strains (CCAP 19/18 and CONC-001) compared to the similarity between each strain and *D. viridis* CONC-002. When the protein sequences were compared, the percentages of similarity for each protein were higher than the respective gene sequence similarities. This increase in similarity at the protein level is from synonymous mutations that have occurred in the DNA sequence. This phenomenon occurs because of the degeneracy in the genetic code, multiple codons coding for the same amino acid. Further analysis of the protein sequences revealed not only substitutions, but also additions and deletions of amino acids had occurred. Substitutions that have taken place involved the same type of amino acid, ex. a polar amino acid being replaced by another polar amino acid, as well as substitutions between different types of amino acids, ex. a polar amino acid being replaced by a nonpolar amino acid. To visualize the effect of amino acid substitutions, additions and deletions on the tertiary protein structure each protein sequence had a hypothetical structure made using the web server RaptorX.

Using RaptorX to predict the 3D structure of the proteins for each of the three organisms, *D. salina* CCAP 19/18, *D. salina* CONC-001 and *D. viridis* CONC-002, found that the structures were overall very similar. In a 1986 study done by Chothia and Lesk investigating the protein structures of various protein families they found an increase in the diversity of a protein sequence resulted in more changes in protein structure to occur. A protein model that has at least 50% sequence similarity to another protein sequence was found to be a good model for the structure (Chorthia and Lesk 1986). If the sequence similarity is less than 20%, major structural differences between the protein structures are present (Chorthia and Lesk 1986). The alignments between the template used and the cox1 protein sequence of *D. salina* CCAP 19/18 *D. salina* CONC-001 and *D. viridis* CONC-002 was more than 50% similar, therefore according to Chorthia and Lesk (1986)
the template used for this protein was a good model for predicting the protein structure of cox1 for all three *Dunaliella* organisms. However, the similarity between the template used for nad6 and nad6 protein sequence of the three *Dunaliella* organisms was less than 25%, meaning the template used is not a good representation of the tertiary structure of nad6 for the three *Dunaliella* organisms, and major structural differences are present between the structure of the template nad6 and the nad6 of the three *Dunaliella* organisms.

### 3.4.2 Species similarity

Major similarities are present in the coding regions of the three *Dunaliella* organisms. Firstly, all three organisms contain the same 12 mitochondrial genes. In addition, all three organisms have their large and small ribosomal RNA genes broken up into the same number of pieces, 9 large and 3 small. The three *Dunaliella* organisms start to differ in the order of genes within their genome, but even within this difference some similarity between all three organisms still exists. The gene order between the two strains of *D. salina* (CCAP 19/18 and CONC-001) is nearly identical and the gene order between *D. salina* CCAP 19/18 and *D. viridis* CONC-002 may seem different, but upon further investigation some consistencies exist. While areas of similarity among all three *Dunaliella* organisms are present, a higher amount of similarity between *D. salina* CCAP 19/18 and *D. salina* CONC-001 is evident. With the exception of genome size, *D. salina* CONC-001 has more in common with its Australian counterpart than with *D. viridis* CONC-002; this was especially noticeable when investigating gene sequence conservation.

More similarities can be found when comparing the genomes at a genetic level. Comparing gene and protein sequences resulted in a closer look into the conservation of sequences and tertiary protein structure. The results obtained from gene comparison further supported the idea that even though their geographical location is different, both strains of *D. salina* (CCAP 19/18 and CONC-001) were more alike than each of them were to *D. viridis* CONC-002, supporting the identification of both strains as the same species.
Throughout evolution various mechanisms, ex. mutation, gene duplication, gene loss, have been acting on genomes causing them to be diverse. While there has been geographical isolation separating the two *D. salina* strains (CCAP 19/18 CONC-001) allowing different process to act on their genomes causing diversity, some aspects of their genome has retained similar.

The mutational hazard hypothesis proposed by Lynch *et al.* (2006) is the idea that bloated genomes have low levels of diversity. However, even though the mtDNA of *D. salina* is bloated with noncoding DNA, my results indicate that there is high diversity between the mitochondrial genomes of the two strains of *D. salina* (CCAP 19/18 and CONC-001). To start, there is a huge difference in size of the genomes, in which *D. salina* CONC-001 is almost double the size of *D. salina* CCAP19/18. In Chapter 2 I discovered that the synonymous substitution rates of the mtDNA of *D. salina* (CCAP 19/18 and CONC-001) are one of the highest rates seen in green algae. As mentioned earlier, mutation is a contributor to the diversity seen in organelle genomes. Since the mutation rate is high in the mtDNAs, it is no surprise that we see diversity in the two *D. salina* strains (CCAP 19/18 and CONC-001). Typically the mtDNA of land plants experience low levels of diversity, whereas animal mitochondrial genomes tend to experience higher levels. The synonymous substitution rate found between the mitochondrial genomes of *D. salina* CCAP 19/18 and *D. salina* CONC-001 (see chapter 2 section 2.3.4) is lower than what is found when comparing the average synonymous substitution rates of animal mtDNA, but it is similar to that seen when comparing monocots and dicots (Wolfe et al. 1987). Therefore, even though the mitochondrial genomes of the *D. salina* strains (CCAP 19/18 and CONC-001) experience high mutation rates they are still closer to those of land plants than animals.

### 3.5 References


Chapter 4

4 Geographical and population-genetic effects on organelle genomic architecture

4.1 Genomic architecture and the mutational hazard hypothesis

The organelle genomes of green algae, particularly chlamydomonadalean algae, can have different architectures (Leliaert et al. 2012), making them excellent models for studying genome evolution (Nedelcu 1998; Popescu and Lee 2006; Smith and Lee 2010; Smith et al. 2013). This thesis focuses on organelle genomic diversity within a specific chlamydomonadalean genus: Dunaliella. Within this genus, the most striking example of organelle genomic diversity was the variation of size observed among the mitochondrial and plastid genomes. Indeed, the sizes of mitochondrial genomes of the Chilean Dunaliella isolates (CONC-001 and CONC-002) were almost double that of the Australian strain (CCAP 19/18), and similar observations were made for the plastid genome data. This finding was not completely unexpected—various other studies have uncovered large variation in mitochondrial genome size, even among closely related species (e.g., Smith et al. 2010; Fučíková et al. 2014; Smith and Keeling 2015). The variation in size of the Dunaliella organelle DNAs was mostly due to differences in amount of noncoding DNA, which is consistent with other studies on green algal organelle genomes (Smith et al. 2013). The majority of noncoding DNA within the Dunaliella organelle genomes is repetitive DNA, making it challenging to assemble organelle genomes, especially the plastid genome—a reoccurring theme in plastid genomics (Smith and Lee 2009).

It has been argued—in what is often called the mutational hazard hypothesis—that the accumulation of noncoding DNA is primarily due to two factors: 1) the mutation rate, which defines the burden of harbouring excess DNA, and 2) the effective population size, which defines the ability of natural selection to perceive and eliminate this burden, whereby large effective population sizes are better at eliminating and/or avoiding noncoding DNA (Lynch 2007). Supporting this hypothesis is the observation that
organisms with low mutation rates and small effective population sizes often have very bloated genomes whereas those with high mutation rates and large effective population sizes have compact genomes (Lynch and Conery 2003).

When placing the data described in this thesis on *Dunaliella* organelle genomes alongside the mutational hazard hypothesis, some general trends emerge. First, the level of genetic diversity among the three *Dunaliella* species was relatively low when compared to the inter-species mtDNA and ptDNA diversity observed in other groups (Appendix C) (Smith 2015). Low diversity could be a reflection of a very low mutation rate and/or a low effective population size, as discussed in Chapter 2, and is ultimately in line with the mutational hazard hypothesis. Conversely, the mtDNA diversity among the *Dunaliella* isolates, although quite low as well, was ~10-times greater than that within the plastid compartment (Table 2-2). Given that both the mtDNAs and ptDNAs from the *Dunaliella* species have similarly expanded genomes, this finding is not consistent with the mutational hazard hypothesis. In other words, one would have expected equally low levels of diversity in both the mitochondrial and plastid compartments of these species.

Further insights into the processes driving organelle genome evolution within the *Dunaliella* genus could come from within-population organelle DNA diversity studies. Indeed, the genetic diversity within a population, as opposed to the divergence between populations or species, is primarily a product of the effective population size and the mutation rate. One might expect, therefore, that the organelle genetic diversity within *Dunaliella* populations, particularly those representing CONC-001 and CONC-002, to be very low.

### 4.2 *Dunaliella* mitochondrial and chloroplast genome architecture is similar but distinct from those in other green algae.

Green algal mitochondrial genomes tend to fall into one of two categories: reduced-derived genome architecture or ancestral (Nedelcu 1998). Reduced-derived mitochondrial genomes are typically very small in size (15-25 kb), have low gene contents, and fragmented rRNA genes (Nedelcu *et al.* 2000). Reduced-derived mitochondrial genomes
are also often referred to as “Chlamydomonas-like,” because the characteristics of this type of genome are found in C. reinhardtii (Turmel et al. 1999), which was the first chlamydomonadalean mtDNA to be described in detail. Conversely, an “ancestral” mitochondrial genome architecture is typically large (45-55 kb), contains many protein coding genes (>10), almost, if not all, the tRNA genes, the 5S rRNA gene, and has continuous rRNA genes (Nedekcu 2000). In other words, the “ancestral” architectures reflect more closely what the original ancestral green algal mtDNA was believed to look like.

With regards to these two types of mitochondrial genome architectures, the Chilean Dunaliella species, CONC-001 and CONC-002 as well as the Australian CCAP 19/18 have characteristics of both types. Even though some ancestral traits exist, the mtDNA architectures have more characteristics shared by Chlamydomonas species. The size of the mitochondrial genomes of the two Chilean organisms is larger than 45 kb, which is a reminiscent of the ancestral type. However, the low number of genes and the fragmentation of rRNA genes located within the mitochondrial genomes of these two Chilean organisms are consistent with Chlamydomonas –like genomes (Kroymann and Zetsche 1997). It is important to emphasize that the large sizes of the Dunaliella mtDNAs are likely not an ancestral trait, but instead evolved quite recently in evolutionary time, since the divergence between the common ancestor or Dunaliella spp. and C. reinhardtii.

It is noteworthy that the chlorophycean green alga Scenedesmus obliquus was also found to be an intermediate and straddle the line of both types of green algal mitochondrial genomes. S. obliquus has a mitochondrial genome that is similar to size and gene content to that of ancestral green algae, but contains fragmented rRNA genes, which is a characteristic of Chlamydomonas –like genomes (Nedelcu et al. 2000). However, S. obliquus tends to fall more in the middle of the two types of mitochondrial genomes compared to those of the Dunaliella species from Chile. While there are differences between the mitochondrial genomes of two Dunaliella Chilean species with that of the Australian Dunaliella species, the main foundations of the genomes are still similar to that of other Chlamydomonas species.
4.3 Species diversity and classification

Organelle genomes have been used in evolutionary studies to gain information about the phylogenetic relatedness of organisms (Jansen et al. 2007; Lemieux et al. 2014). From sequencing the mitochondrial genomes of the two Chilean organisms, *D. salina* CONC-001 and *D. viridis* CONC-002, as expected, I found that there was a higher similarity between *D. salina* CONC-001 and *D. salina* CCAP 19/18. Not only did these two strains have an almost identical gene order, but their gene and protein sequences had a higher similarity than *D. viridis* CONC-002 had to *D. salina* CCAP 19/18.

In addition to nuclear small subunit rDNA (18S), both plastid and mitochondrial sequences have been used to aid in the identification and taxonomic placement of organisms (Lewis and McCourt 2004; Rodríguez-Ezpeleta et al. 2007). My findings on organelle genomic architecture support the identification of *D. salina* CONC-001 as being similar to *D. salina* CCAP 19/18. However, there is no doubt that these two isolates are members of distinct populations, and likely represent distinct but closely related species. As shown from my research, substitution rates of organelle genomes can aid in the verification that organisms are being identified properly, or as correctly as possible based on the available data.

Smith et al. (2010) noted that among the *D. salina* sequences available on GenBank a high amount of sequence diversity that occurs. Reasoning for this high level of diversity is due to improper and incorrect identification of *D. salina* strains (Smith et al. 2010). Using additional data like organelle genome sequences could help alleviate the issue of misidentification of species.

4.4 Concluding remarks

In this thesis, I further demonstrate that the mitochondrial genomes of green algal organisms contain a diverse genomic architecture. Studying mitochondrial genomic architecture demonstrated that a genomic expansion event has occurred within the mitochondrial genomes of the green algal genus *Dunaliella*. This genomic expansion was due to the accumulation of noncoding DNA resulting in larger genomes that were inflated
with noncoding DNA. In a hypothetical protein-folding test, I found that while there was diversity within the amino acid sequence between the mitochondrial genes of three organisms, the folding of the proteins were not very different. By gaining additional information of mitochondrial genomic architecture it can reveal situations in evolution where dramatic events have occurred within a closely related group of organisms. A better understanding of the identification of species can be gained when comparing the mitochondrial genomes as it gives further genetic support in the accuracy of the identification. This research on organelle genomes adds another piece to the puzzle of just how bizarre organelle genomes can be, and while they stand in the shadow of nuclear genomes they should not be overlooked.

4.5 References


Smith DR, Lee RW. 2009. The mitochondrial and plastid genomes of *Volvox carteri*: bloated molecules rich in repetitive DNA. BMC Genomics 10:132


Appendices

Appendix A: *Dunaliella salina* nucleotide substitution rates for individual protein-coding loci. Mitochondrial located genes.

<table>
<thead>
<tr>
<th>MtDNA locus</th>
<th># codons</th>
<th>(d_N / d_S)</th>
<th>(d_N)</th>
<th>(d_S)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Cob</em></td>
<td>389</td>
<td>0.0445</td>
<td>0.0319</td>
<td>0.716</td>
</tr>
<tr>
<td><em>cox1</em></td>
<td>509</td>
<td>0.0311</td>
<td>0.027</td>
<td>0.8686</td>
</tr>
<tr>
<td><em>nad1</em></td>
<td>295</td>
<td>0.008</td>
<td>0.0179</td>
<td>2.2332</td>
</tr>
<tr>
<td><em>nad2</em></td>
<td>434</td>
<td>0.0436</td>
<td>0.0439</td>
<td>1.0067</td>
</tr>
<tr>
<td><em>nad4</em></td>
<td>482</td>
<td>0.0505</td>
<td>0.0704</td>
<td>1.3961</td>
</tr>
<tr>
<td><em>nad5</em></td>
<td>557</td>
<td>0.0444</td>
<td>0.0415</td>
<td>0.9355</td>
</tr>
<tr>
<td><em>nad6</em></td>
<td>156</td>
<td>0.0707</td>
<td>0.0703</td>
<td>0.9939</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>403.14</strong></td>
<td><strong>0.042</strong></td>
<td><strong>0.043</strong></td>
<td><strong>1.164</strong></td>
</tr>
<tr>
<td><strong>Standard deviation</strong></td>
<td><strong>138.46</strong></td>
<td><strong>0.019</strong></td>
<td><strong>0.020</strong></td>
<td><strong>0.515</strong></td>
</tr>
</tbody>
</table>
Appendix B: *Dunaliella salina* nucleotide substitution rates for individual protein-coding loci. Plastid located genes.

<table>
<thead>
<tr>
<th>PtDNA locus</th>
<th># codons</th>
<th>$d_S / d_S$</th>
<th>$d_S$</th>
<th>$d_S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>atpA</td>
<td>503</td>
<td>0.001</td>
<td>0</td>
<td>0.0139</td>
</tr>
<tr>
<td>atpB</td>
<td>479</td>
<td>0.0447</td>
<td>0.0113</td>
<td>0.2531</td>
</tr>
<tr>
<td>atpE</td>
<td>134</td>
<td>---</td>
<td>0.0029</td>
<td>0</td>
</tr>
<tr>
<td>atpF</td>
<td>173</td>
<td>0.001</td>
<td>0</td>
<td>0.0166</td>
</tr>
<tr>
<td>atpH</td>
<td>81</td>
<td>0.001</td>
<td>0.0001</td>
<td>0.085</td>
</tr>
<tr>
<td>atpI</td>
<td>238</td>
<td>0.0255</td>
<td>0.0143</td>
<td>0.5616</td>
</tr>
<tr>
<td>ccsA</td>
<td>396</td>
<td>0.1162</td>
<td>0.0042</td>
<td>0.0362</td>
</tr>
<tr>
<td>cemA</td>
<td>593</td>
<td>0.2312</td>
<td>0.0054</td>
<td>0.0234</td>
</tr>
<tr>
<td>chlL</td>
<td>290</td>
<td>0.025</td>
<td>0.0568</td>
<td>2.2727</td>
</tr>
<tr>
<td>petA</td>
<td>287</td>
<td>0.0617</td>
<td>0.0027</td>
<td>0.0444</td>
</tr>
<tr>
<td>petB</td>
<td>214</td>
<td>0.001</td>
<td>0</td>
<td>0.0424</td>
</tr>
<tr>
<td>petD</td>
<td>159</td>
<td>0.001</td>
<td>0</td>
<td>0.024</td>
</tr>
<tr>
<td>petG</td>
<td>36</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>petL</td>
<td>31</td>
<td>0.001</td>
<td>0</td>
<td>0.0404</td>
</tr>
<tr>
<td>psaA</td>
<td>711</td>
<td>0.0121</td>
<td>0.0006</td>
<td>0.0472</td>
</tr>
<tr>
<td>psaB</td>
<td>734</td>
<td>0.0436</td>
<td>0.0006</td>
<td>0.013</td>
</tr>
<tr>
<td>psaC</td>
<td>80</td>
<td>0.001</td>
<td>0</td>
<td>0.0349</td>
</tr>
<tr>
<td>PtDNA locus</td>
<td># codons</td>
<td>$d_N / d_S$</td>
<td>$d_N$</td>
<td>$d_S$</td>
</tr>
<tr>
<td>-------------</td>
<td>----------</td>
<td>-------------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td><em>psaJ</em></td>
<td>40</td>
<td>0.4914</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><em>psbA</em></td>
<td>352</td>
<td>0.0331</td>
<td>0.0038</td>
<td>0.1157</td>
</tr>
<tr>
<td><em>psbB</em></td>
<td>507</td>
<td>0.001</td>
<td>0</td>
<td>0.0083</td>
</tr>
<tr>
<td><em>psbC</em></td>
<td>461</td>
<td>0.0213</td>
<td>0.0018</td>
<td>0.0848</td>
</tr>
<tr>
<td><em>psbD</em></td>
<td>351</td>
<td>0.001</td>
<td>0</td>
<td>0.0276</td>
</tr>
<tr>
<td><em>psbE</em></td>
<td>81</td>
<td>0.001</td>
<td>0</td>
<td>0.0356</td>
</tr>
<tr>
<td><em>psbF</em></td>
<td>43</td>
<td>0.001</td>
<td>0</td>
<td>0.0408</td>
</tr>
<tr>
<td><em>psbH</em></td>
<td>82</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><em>psbI</em></td>
<td>36</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><em>psbJ</em></td>
<td>41</td>
<td>0.001</td>
<td>0.0002</td>
<td>0.1716</td>
</tr>
<tr>
<td><em>psbK</em></td>
<td>45</td>
<td>0.0507</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><em>psbL</em></td>
<td>37</td>
<td>0.9961</td>
<td>0.1247</td>
<td>0.1251</td>
</tr>
<tr>
<td><em>psbM</em></td>
<td>36</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><em>psbN</em></td>
<td>43</td>
<td>0.4303</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><em>psbT</em></td>
<td>30</td>
<td>---</td>
<td>0.0138</td>
<td>0</td>
</tr>
<tr>
<td><em>psbZ</em></td>
<td>61</td>
<td>---</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><em>rpl2</em></td>
<td>274</td>
<td>0.053</td>
<td>0.0014</td>
<td>0.027</td>
</tr>
<tr>
<td><em>rpl5</em></td>
<td>180</td>
<td>0.001</td>
<td>0</td>
<td>0.0123</td>
</tr>
<tr>
<td>PtDNA locus</td>
<td># codons</td>
<td>$d_N / d_S$</td>
<td>$d_N$</td>
<td>$d_S$</td>
</tr>
<tr>
<td>-------------</td>
<td>----------</td>
<td>-------------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>rpl14</td>
<td>118</td>
<td>0.001</td>
<td>0</td>
<td>0.0441</td>
</tr>
<tr>
<td>rpl16</td>
<td>136</td>
<td>0.41</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>rpl20</td>
<td>111</td>
<td>0.0627</td>
<td>0.0035</td>
<td>0.056</td>
</tr>
<tr>
<td>rpl23</td>
<td>115</td>
<td>0.0726</td>
<td>0.0034</td>
<td>0.047</td>
</tr>
<tr>
<td>rpl36</td>
<td>36</td>
<td>0.0943</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>rps4</td>
<td>251</td>
<td>---</td>
<td>0.0015</td>
<td>0</td>
</tr>
<tr>
<td>rps8</td>
<td>143</td>
<td>0.3676</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>rps9</td>
<td>195</td>
<td>0.001</td>
<td>0</td>
<td>0.0159</td>
</tr>
<tr>
<td>rps11</td>
<td>129</td>
<td>0.1006</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>rps12</td>
<td>128</td>
<td>0.001</td>
<td>0</td>
<td>0.0427</td>
</tr>
<tr>
<td>rps18</td>
<td>139</td>
<td>0.001</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>rps19</td>
<td>91</td>
<td>0.0351</td>
<td>0.0045</td>
<td>0.1284</td>
</tr>
<tr>
<td>tufA</td>
<td>417</td>
<td>0.001</td>
<td>0</td>
<td>0.013</td>
</tr>
<tr>
<td>ycf3</td>
<td>170</td>
<td>0.001</td>
<td>0</td>
<td>0.0178</td>
</tr>
<tr>
<td>ycf4</td>
<td>219</td>
<td>0.144</td>
<td>0.0038</td>
<td>0.0266</td>
</tr>
<tr>
<td>ycf12</td>
<td>32</td>
<td>0.001</td>
<td>0.0001</td>
<td>0.0551</td>
</tr>
<tr>
<td>Average</td>
<td>201.353</td>
<td>7.842</td>
<td>0.005</td>
<td>0.090</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>182.918</td>
<td>26.859</td>
<td>0.019</td>
<td>0.324</td>
</tr>
</tbody>
</table>
Appendix C: Synonymous substitution rates in plastid and mitochondrial genomes of various plastid-bearing lineages.

<table>
<thead>
<tr>
<th></th>
<th>Substitutions per synonymous site (SD)</th>
<th>Substitution rate ratios (pt : mt)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ptDNA</td>
<td>mtDNA</td>
</tr>
<tr>
<td><strong>ARCHAEPLASTIDA</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Glaucohytes</strong>a</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Cyanophora</em></td>
<td>1.01 (1.22)</td>
<td>5.29 (3.17)</td>
</tr>
<tr>
<td><strong>Green algae</strong>b</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Dunaliella</em></td>
<td>0.09 (0.32)</td>
<td>1.16 (0.52)</td>
</tr>
<tr>
<td><em>Chlamydomonas</em></td>
<td>0.30 (0.11)</td>
<td>0.29 (0.05)</td>
</tr>
<tr>
<td><em>Mesostigma</em></td>
<td>0.11 (0.06)</td>
<td>0.17 (0.11)</td>
</tr>
<tr>
<td><strong>Seed plants</strong>c</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Angiosperms</td>
<td>0.39 (0.01)</td>
<td>0.13 (0.01)</td>
</tr>
<tr>
<td>Gymnosperms</td>
<td>0.61 (0.03)</td>
<td>0.28 (0.02)</td>
</tr>
<tr>
<td><strong>Red algae</strong>d</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Porphyra</em></td>
<td>0.47 (0.22)</td>
<td>1.76 (0.58)</td>
</tr>
<tr>
<td><strong>HAPTOPHYTA</strong>e</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Phaeocystis</em></td>
<td>0.25 (0.16)</td>
<td>2.41 (0.97)</td>
</tr>
</tbody>
</table>

SD: standard deviation. Synonymous-site substitution rates are based on averages among loci, not concatenations. aData from Smith et al. (2014). bData from present study, Popescu and Lee (2007) and Hua et al. (2012). cData from Drouin et al. (2008). dData from Smith et al. (2012). eData from Smith et al. (2014).
Appendix D: Maximum-likelihood protein phylogeny of chlamydomonadalean algae. Conceptual protein translations of seven mitochondrial genes (cob, cox1, nad1, nad2, nad4, nad5, and nad6) were aligned with MAFFT v7 (Katoh and Standley 2013) and manually refined using Se-Al v2.0a11 (http://tree.bio.ed.ac.uk/software/seal/). Multiple alignments are available upon request. The best-fit amino acid substitution model for the seven-protein concatenated multiple alignment (1931 amino acids long) was selected with ProtTest3 using the AIC selection criterion (Darriba et al. 2011). Maximum likelihood trees were estimated with RAxML v7.2.6 (Stamatakis 2006) using the selected WAG+I substitution model. Branch support was assessed with 500 non-parametric bootstrap replicates. Bayesian posterior probabilities were calculated with MrBayes 3.2.1 (Ronquist et al., 2012) using the WAG+I+Γ substitution model running a Metropolis-coupled Markov Chains Monte Carlo (MC3) for 1.5 million generations. Two independent MC3 runs were performed simultaneously starting from different random trees. Trees were sampled every 100th generation. Final posterior probabilities were calculated after discarding trees sampled from the first 500,000 generations. Note: Branching order of Figure 1 from the main text is based on phylogenetic analyses of Nakada et al. (2008), González et al. (2009), Smith et al. (2013), and Figueroa-Martinez et al. (2015).

Appendix E: References for Appendices A-D


Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses
Permission Request

Dear Michael Del Vasto


Thank you for your recent email requesting permission to reuse all or part of your article in a thesis.

As part of your copyright agreement with Oxford University Press you have retained the right, after publication, to use all or part of the article and abstract, in the preparation of derivative works, extension of the article into a booklength work, or in another works collection, provided that a full acknowledgement is made to the original publication in the journal. As a result, you should not require direct permission from Oxford University Press to reuse your article. However, reuse of your article must be in line with the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/) your article has been published under.

For full details of our publication and rights policy please see the attached link to our website:


If you have any other queries, please feel free to contact us.

Yours sincerely

Emma Thornton

Appendix F: Permission for use of publication
Curriculum Vitae

Name: Michael Del Vasto

Post-secondary
Education and
Degrees:
The University of Western Ontario
London, Ontario, Canada
2008-2012 B.Sc.

The University of Western Ontario
London, Ontario, Canada
2013-Present M.Sc.

Honours and
Awards:
Department of Biology TA Award
2015

Related Work
Teaching Assistant
Experience
The University of Western Ontario
2013-2015

Publications: