THE EVOLUTION OF ORGANELLE GENOME ARCHITECTURE

by

David Roy Smith

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

at

Dalhousie University
Halifax, Nova Scotia
August 2010

DALHOUSIE UNIVERSITY

DEPARTMENT OF BIOLOGY

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "THE EVOLUTION OF ORGANELLE GENOME ARCHITECTURE" by David Roy Smith in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Dated: August 13, 2010

External Examiner: _____

Research Supervisor: _____

Examining Committee: _____

_____

Departmental Representative: _____

DALHOUSIE UNIVERSITY

DATE: 13 August 2010

AUTHOR: David Roy Smith

TITLE: THE EVOLUTION OF ORGANELLE GENOME ARCHITECTURE

DEPARTMENT OR SCHOOL: Department of Biology

DEGREE: PhD CONVOCATION: October YEAR: 2010

Permission is herewith granted to Dalhousie University to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.

_____

Signature of Author

For my sister, Poppi

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Genomic sequence data from the three domains of life have revealed a remarkable diversity of genome architectures. The relative contributions of adaptive versus non-adaptive processes in shaping this diversity are poorly understood and hotly debated. This thesis investigates the evolution of genome architecture in the Chloroplastida (i.e., green algae and land plants), with a particular focus on the mitochondrial and plastid genomes of chlamydomonadalean algae (Chlorophyceae, Chlorophyta). Much of the work presented here describes unprecedented extremes in: i) genome compactness (i.e., the fraction of noncoding DNA in a genome), ii) genome conformation (e.g., circular vs. linear vs. linear fragmented genomes), iii) intron and repeat content; and iv) nucleotide-composition landscape (e.g., GC-rich vs. AT-rich genomes). These data are then combined with intra-population nucleotide diversity data to explore the degree to which non-adaptive forces, such as random genetic drift and mutation rate, have shaped the organelle and nuclear genomes of the Chloroplastida. The major conclusions from this dissertation are that chlamydomonadalean algae show a much greater variation in organelle genome architecture than previously thought — this group boasts some of the most unusual mitochondrial and plastid genomes from all eukaryotes — and that the majority of this variation can be explained in non-adaptive terms.

ACKNOWLEDGEMENTS

I thank my supervisor and close friend Bob Lee — you're a great mentor! I will remember our lunches at King's and the laughs in the lab. Many high fives to Tudor Borza, Ananda Venkatachalam, and Jimeng Hua. And special thanks to Santhosh Karanth and Kate Crosby. My love to Greg, Claire, Mom, and Dad. Finally, I thank all of the people — friends, reviewers, committee and Dalhousie faculty members, and especially my supervisor — who graciously gave their time to read, comment, correct, and improve the various chapters presented in this thesis. Because I did not always take their advice, any errors or failours in this work are completely my own.

The work presented in this thesis was supported by a grant to Robert W. Lee from the Natural Sciences and Engineering Research Council (NSERC) of Canada. David Roy Smith is an Izaak Walton Killam Memorial Scholar and holds a Canada Graduate Scholarship from NSERC.

CHAPTER 1: INTRODUCTION

The primary objectives of this thesis are to uncover and describe extremes in genome architecture and then address the evolutionary forces governing these extremes, particularly the relative contributions of adaptive versus non-adaptive processes. To do this, I have focused on the mitochondrial and plastid genomes from chloroplastidial species (i.e., green algae and land plants [Adl et al. 2005]), especially those found in the Chlamydomonadales (Chlorophyceae, Chlorophyta). By employing a variety of different chlamydomonadalean algae, as well as the land plant *Selaginella*, I argue that most of the observed diversity in organelle genome architecture within the Chlamydomonadales, and the Chloroplastida in general, is the result of non-adaptive processes. I am not the first to take a non-adaptive approach to explaining genome evolution, but I am arguably in the minority. Thus, a primary theme throughout this thesis, and the major take-home message, is this: One need not invoke adaptation to explain the magnificent array of genomic architectures observed throughout our natural world, but rather one can call upon neutral processes, which, if thoroughly considered, can elegantly account for much of the genomic diversity described here and elsewhere.

I have divided the thesis into the following sections, each of which focuses on particular aspects of genome architecture. Chapters 2, 3, and 4 explore the evolution of nucleotide landscape and genome conformation in the organelle DNAs of the non-photosynthetic, green algal group *Polytomella* and the lycophyte genus *Selaginella*. More specifically, they attempt to explain the forces driving GC enrichment of organelle DNA, the evolution of linear mitochondrial genomes and their telomeres, and how linear, monomeric mitochondrial DNAs can become fragmented into linear, bipartite molecules. Chapters 5, 6, and 7 address the evolution of organelle and nuclear genome size, and test a contemporary theory on genome evolution called the mutational-hazard hypothesis (Lynch 2007); these three chapters present complete genome sequences and silent-site nucleotide diversity values for the model green algae *Chlamydomonas reinhardtii* and *Volvox carteri*. Together, these data are used to gain insights into fundamental population genetic parameters, such as mutation rate and effective genetic population size, and how these parameters influence the evolution of genome size. The latter parts of the thesis (Chapter 8 and Appendix E) describe some of the extremes in intron and repeat content in

chlamydomonadalean organelle DNAs; these data are then placed in context to the findings presented in earlier chapters.

This Ph.D. dissertation is a publication-based thesis, which means that all of the primary chapters have been published in peer reviewed journals and that these chapters are in "journal format," i.e., they each contain their own abstract, introduction, methods and materials, results, and discussion sections. All of the citations have been placed in a single References section, beginning on page 183. In total, this thesis is based on eight publications (Smith and Lee 2008a, 2008b, 2009a, 2009b, 2010; Smith 2009; Smith et al. 2010a, 2010b), which have been incorporated into seven thesis chapters (Chapters 2 to 8) and one appendix (Appendix E). The full citations for these eight publications are located at the beginning of the given chapter or appendix that each publication represents. Copyright permission letters for the chapters that were not published in open-access journals (Smith and Lee 2008a, 2010; Smith 2009, Smith et al. 2010a) are found in Appendix A. I am the first, or only, author on all of the publications presented in this thesis; and I declare that I analyzed and, when applicable, collected the data that are described in these papers, and wrote the manuscripts — for further details see Student Contribution to Manuscripts in Thesis.

I have structured and formatted the thesis so that each of the main chapters can, for the most part, stand alone. However, to reduce repetition, especially in the various materials and methods sections and with respect to certain figures, I sometimes refer the reader to earlier chapters. Moreover, within the thesis, when I mention work that is either discussed in later chapters or that was already presented in a preceding chapter, I cite both the chapter number as well as the formal reference for the given data (e.g., Smith and Lee 2008a; Chapter 2). For clarity, in each chapter I redefine scientific terms and acronyms, and give the full scientific name when I first refer to a species. Given that this thesis and the publications from which it was constructed are the product of the combined efforts of my supervisor and myself — and in some instances, our collaborators — I have chosen to use the first person plural throughout much of the writing rather than the singular. Finally, in some instances the thesis chapters differ slightly from their corresponding publications. This is because i) I altered the published manuscripts so that they meet Dalhousie thesis-formatting requirements, ii) I cut sections from some of the

publications (as discussed above) to reduce repetition, and iii) I corrected typos and errors that were present in the published manuscripts.

CHAPTER 2: MITOCHONDRIAL GENOME OF THE COLORLESS GREEN ALGA *POLYTOMELLA CAPUANA*: A LINEAR MOLECULE WITH AN UNPRECEDENTED GC CONTENT.

Published as:

Smith DR, Lee RW (2008) Mitochondrial genome of the colorless green alga *Polytomella capuana*: a linear molecule with an unprecedented GC content. Mol Biol Evol 25:487-496.

Abstract

One common observation concerning mitochondrial genomes is that they have a low guanine and cytosine (GC content); of the complete mitochondrial genome sequences currently available at the National Center for Biotechnology Information (NCBI) as of July, 2007, the GC content ranges from 13.3% to 53.2% and has an average value of 38%. This study presents the GC-rich mitochondrial genome (57% GC) of the colorless green alga *Polytomella capuana*. The disproportion of G and C among the different regions of the *P. capuana* mitochondrial DNA (mtDNA) suggests that a neutral process is responsible for the GC bias — potentially biased gene conversion. In addition, DNA sequencing and gel electrophoresis analyses indicate that the *P. capuana* mitochondrial genome is a single 13 kilobase (kb), linear molecule with "hairpin loop" telomeres: a novel terminal structure among described linear, green algal mtDNAs. Furthermore, using a series of GC-rich inverted repeats found within the *P. capuana* mitochondrial genome, we describe recombination-based scenarios of how intact linear mtDNA conformations can be converted into the fragmented forms found in other *Polytomella* taxa.

Introduction

One of the most distinguishing characteristics of mitochondrial genomes as compared to nuclear genomes, is their low GC content; of the 1,125 complete mtDNA sequences available at NCBI as of July, 2007, the GC content ranges from 13.3% to 53.2% and has an average value of 38%. Although sampling is highly biased towards animal mitochondrial genomes (1015 out of the 1125), the trend of having a low GC content is seen throughout other major groups, including fungi (17.1-43.2% GC), the Archaeplastida (Adl et al. 2005) (22.2-45.2% GC), and the heterogeneous group of eukaryotic unicells negatively defined as not belonging to animals, fungi, or the Archaeplastida (14.0-41.2% GC). Various hypotheses for explaining why mtDNA is GC poor have been proposed. For example, it has been suggested that low levels of G and C are due to a mutational bias caused by the loss of DNA repair genes in the endosymbiotic genome that gave rise to the mitochondrial genome (Glass et al. 2000; Moran 2002; Burger and Lang 2003). Others contend that a low GC content correlates with adaptation

to an intracellular lifestyle where high levels of adenosine and uridine triphosphate (ATP and UTP) relative to cytidine and guanosine triphosphate (CTP and GTP) make replication and transcription of an AT-rich genome more efficient (Howe et al. 2003; Rocha and Danchin 2002). Convergent evolution towards a low GC content is also seen in the genomes of plastids, symbionts, and endocellular parasites (Dybvig and Voelker 1996; Ogata et al. 2001; Kusumi and Tachida 2006).

The mitochondrial genome of the colorless green alga *P. capuana* attracted our interest because of its potential for having a particularly high GC content relative to other available mitochondrial genome sequences, not only for green algae but also for eukaryotes in general. This view was motivated by the high GC content of a 768 nucleotide (nt) segment of *cox1* from the mtDNA of *P. capuana* (GenBank accession number DQ221113; Mallet and Lee 2006), especially at sites expected to be under low selective constraint, such as four-fold degenerate sites (defined as third-position codon sites that can tolerate any of the four nucleotides without altering the amino acid specified).

Polytomella is a group of wall-less, colorless unicells (Pringsheim 1955) belonging to the "*Reinhardtinia* clade" (Nakada et al. 2008) of chlorophycean green algae. Available *Polytomella* taxa fall into three lineages (Mallet and Lee 2006). The members of at least one of these lineages, represented by *Polytomella parva,* possess fragmented mitochondrial genomes (Fan and Lee 2002; Mallet and Lee 2006) whereas *P. capuana*, the only known member of earliest branching lineage, appears to have an intact, linear mitochondrial genome (Mallet and Lee 2006). Little is known about the mtDNA conformation of the third lineage, represented by *Polytomella* strain SAG 63-10; however, unpublished data (Mallet and Lee personal communication) indicate that it is linear fragmented. Substantial DNA sequence data exist only for the mitochondrial genome of *P. parva*, which is comprised of two linear fragments with estimated sizes of 13.5 and 3.5 kb (Fan and Lee 2002). Sequence data spanning 97% and 86%, of the 13.5 and 3.5 kb fragments, respectively, are available. The 3.5 kb fragment encodes only one gene (*nad6*), which is missing from the 13.5 kb fragment. The telomeres of both fragments contain virtually identical inverted repeats that are at least 1.3 kb in length; however, the extreme termini of both fragments still remain to be sequenced.

Chlamydomonas reinhardtii, a close relative of *Polytomella* taxa (Nakayama et al. 1996; Pröschold et al. 2001; Gerloff-Elias et al. 2005), also has a linear mitochondrial genome. The telomeres of this completely sequenced 16 to 19 kb linear genome are each composed of an ~600 nt inverted repeat with an ~40 nt, non-complementary 3′ extension. (Gray and Boer 1988; Michaelis et al. 1990). Based on the potential interaction of the telomeres with internal repeats, two models of replication have been proposed for the *C. reinhardtii* mtDNA (Vahrenholz et al. 1993). There are no apparent similarities between the telomeric sequences of *P. parva* and *C. reinhardtii* (Fan and Lee 2002), and a model describing how the *P. parva* mtDNA may replicate has not yet been proposed.

Taken as a whole, sequence data from the *P. capuana* mtDNA will provide useful information on nucleotide composition biases in mitochondrial genomes as well as knowledge about the evolution of mitochondrial telomeres and the mechanisms by which intact, linear mtDNA conformations were converted into fragmented forms. With these motives in mind, we sequenced to completion the *P. capuana* mitochondrial genome.

Materials and Methods

Strain, Culture Conditions, Mitochondrial Enrichment, and DNA Extraction

We used a stock of *P. capuana* [Sammlung von Algenkulturen Göttingen (SAG) strain 63-5] made axenic by Mallet and Lee (2006). Cells were cultured at 22°C in *Polytomella* medium (0.1% tryptone, 0.2% yeast extract, 0.2% sodium acetate) and harvested in the late logarithmic phase of growth ($OD_{750\ nm} = 0.4$; determined with a Bausch and Lomb Spectronic 20 spectrophotometer) by centrifugation (1,000 g) at 4°C. Cells were disrupted with a Dounce homogenizer. The mitochondria-enriched fraction was prepared and treated with DNase I following procedure B of Ryan et al. (1978). Isolation of DNA followed the method of Ryan et al. (1978), with the exception that there was no further DNA purification step employing preparative CsCl gradient centrifugation.

DNA Amplification

PCR experiments were performed in High Fidelity Platinum SuperMix (Invitrogen, Carlsbad, CA) using DNA from a mitochondria-enriched fraction as the template. DNA was initially denatured at 94°C for 3 min, then amplified by 35 cycles of denaturation at 94°C for 45 s, annealing at 50°C-60°C (depending on the melting temperature of the primers) for 30 s, and extension at 72°C for 3 min; there was a final extension at 72°C for 10 min. The telomeric regions of the *P. capuana* mtDNA were amplified using: i) the long walk PCR method of Katz et al. (2000), ii) terminal deoxynucleotidyl transferase (TdT) tailing as described by Förstemann et al. (2000) and Bah et al. (2004), and iii) standard PCR amplification (as described above).

DNA Blotting and Hybridization

Blotting of agarose gels onto Hybond N+ membranes (GE Healthcare, Buckinghamshire, UK) was performed using the Vacublot XL system (GE Healthcare). Probes used in this study were labeled and hybridized to samples with the AlkPhos Direct Labelling and Detection System (GE Healthcare) following the manufacturer's instructions. Label was detected by exposing the membranes to Fuji Super RX medical X-ray film (Fuji Photo Film Co., Tokyo, Japan).

Cloning and Sequencing of DNA Fragments

PCR, long walk PCR, and TdT tailing products were separated by agarose gel electrophoresis, purified with the QIAquick Gel Extraction Kit (Qiagen, Germantown, MD), then cloned using the TOPO TA Cloning Kit (Invitrogen). Plasmid DNA was extracted with the QIAquick Spin Miniprep Kit (Qiagen). PCR products and the clones derived from PCR products were sequenced on both strands at the Centre for Applied Genomics, Hospital for Sick Children, Toronto, Canada.

Sequence Analysis

Sequences were edited and assembled using CodonCode Aligner (Version 1.5.2; CodonCode Corporation, Dedham, MA). The boundaries of the mitochondrial rRNA-

coding modules were estimated by sequence comparisons with their counterparts in the *P. parva* and *C. reinhardtii* mitochondrial genomes. MtDNA repeats were identified using REPuter (Kurtz et al. 2001). Secondary structures and folding energies of the inverted repeat sequences were predicted with Mfold (Zuker 2003). The equation used for calculating the GC-skew value was: (G – C)/(G + C); that for the AT-skew was: (A – T)/(A + T).

Nucleotide Sequence Accession Number

The complete sequence of the *P. capuana* mtDNA is deposited in GenBank under the accession number EF645804.

Results

General Features of the P. capuana *MtDNA*

The *P. capuana* mitochondrial genome is a single linear molecule of 12,998 nt with terminal inverted repeats (i.e., the sequence of one terminus is present in an inverted orientation relative to the other terminus) (Figure 2.1). The size and conformation of the *P. capuana* mtDNA were confirmed by pulse-field gel electrophoresis and restriction endonuclease digestion patterns (data not shown). The coding regions of this mtDNA are arranged into two unequally sized clusters with opposite transcriptional polarities, which proceed outwards toward the ends of the genome. The region separating the two transcriptional orientations shows no similarity to the potential promoter sequence identified in the corresponding region of the *P. parva* (Fan and Lee 2002) and *C. reinhardtii* mtDNAs (Duby et al. 2001). The complement and arrangement of genes in the *P. capuana* mtDNA parallels that of *P. parva*, except for *nad6*, which is located internal to the left terminal repeat in *P. capuana* (Figure 2.1) but on a separate chromosome in *P. parva* (Figure 2.1). The sole tRNA gene, *trnM*(*cau*), has characteristics that are consistent with a role in elongation rather than initiation, like the mitochondrial-encoded *trnM*(*cau*) of other *Reinhardtinia*-clade taxa (Boer and Gray 1988, Denovan-Wright et al. 1998).

Of the 12,998 nt comprising the *P. capuana* mtDNA, 10,662 nt (82%) code for proteins and functional RNAs, and 2,336 nt (18%) represent noncoding DNA. The latter can be subdivided into intergenic regions and terminal repeats, which constitute 555 nt and 1,780 nt (890 nt at each terminus), respectively. The 19 intergenic regions identified in the *P. capuana* mtDNA range in size from 3 to 62 nt and have an average length of 29 nt.

The *P. capuana* has a genome-wide GC bias; the overall GC content is 57.2%. The allocation of G versus C (GC-skew) on the main sense strand (the strand encoding the gene for *cox1*) is negligible with a value of only 0.006. The distribution of A versus T is slightly more skewed at -0.09, reflecting a slight tendency towards T on the main sense strand.

GC-bias by Region

The GC content of the *P. capuana* mtDNA differs considerably among the various regions of the genome. Table 2.1 shows that the GC values of the *P. capuana* mtDNA exceed those of *P. parva* and *C. reinhardtii* for all defined regions.

The nucleotide composition of the coding mtDNA in *P. capuana* has an average GC content of 56.4% (Table 2.1). This value reflects both rRNA- and protein-coding regions evenly. In opposition, the *trnM* gene is slightly GC poor (47.9%). The inflated GC content of the protein-coding genes in the *P. capuana* mtDNA comes from a large number of codons ending in G or C (76%) (Table 2.1); when considering only four-fold degenerate sites, this value is even higher (85%). In both of these cases, the occurrence of G versus C is approximately equal. At the more functionally constrained first and second codon positions, the nucleotide compositions are less GC rich, with values of 52% and 41%, respectively (Table 2.1). The protein-coding genes in the mitochondrial genomes of *P. parva* and *C. reinhardtii* show a tendency towards A and T at all three codon positions (Table 2.1). The *P. capuana* protein-coding genes, in spite of being GC rich, show a derived amino acid composition similar to that of the *P. parva* and *C. reinhardtii* protein-coding genes. Alanine is the only significant exception to this trend, encoded 308 times in the *P. capuana* mtDNA but only 257 and 215 times in that of *C. reinhardtii* and *P. parva*, respectively. It is noteworthy that alanine is encoded by the GC-rich codon family GCN.

Noncoding DNA in the *P. capuana* mitochondrial genome is more GC rich than the coding DNA (Table 2.1); the average GC content of the telomeres is 58.7%, and that of the intergenic spacer regions is 68% (Table 2.1), with the individual intergenic regions ranging from 55% to 100%. Noncoding DNA in both the *P. parva* and *C. reinhardtii* mitochondrial genomes is GC poor (<50%), except for the telomeres of *C. reinhardtti*, which are slightly GC rich (54%) (Table 2.1).

Repetitive Elements

Repetitive elements in the *P. capuana* mtDNA can be divided into two categories: short inverted repeats, which can be folded into stem-loop structures, and long terminal inverted repeats, which make up the telomeric regions.

Short inverted repeat elements: Sixteen pairs of inverted repeats were identified throughout the coding and noncoding regions of the *P. capuana* mtDNA (Figure 2.2). The inverted repeats vary in length from 5 to 27 nt, and all 16 pairs can be folded into stem-loop (hairpin) structures (Figure 2.2). Although there is no sequence identity between pairs of inverted repeats, and though the size of stems and loops among their different predicted secondary structures varies considerably, two trends are apparent: i) the stems are GC rich (> 60%) while the loops are less so (~50%), and ii) the location of the stem-loop structures correlate with the start and end of coding regions.

Twelve of the inverted-repeat pairs have an arrangement where one inverted repeat is found in intergenic DNA and the other (matching) inverted repeat is located in an adjacent coding region thereby resulting in a stem-loop structure that spans both coding and noncoding DNA. In 9 of 12 cases where this occurs, the "loop" portion of the hairpin contains the start of a coding region, and the "stem" component is adjacent to the end of a coding region (Figure 2.2). Every protein-coding gene and all but two of the rRNA-coding modules found in the *P. capuana* mtDNA are bordered by potential stem-loop structures.

A few of the inverted repeats share sequence identity with other parts of the genome. For example, a 20 nt perfect-match sequence corresponding to the stem of the *cox1*/*nad4* hairpin structure (Figure 2.2) was found inserted into *nad6*. Similarly, a 12 nt portion of the *cob*/*nad6* stem (Figure 2.2) was found inserted into *nad5.* The location of

the inverted repeat between the *nad6* and *cob* genes (comprising one half of the stem in the *cob*/*nad6* stem-loop structure) corresponds to the region that is fragmented in other *Polytomella* lineages.

The "global minimum" of a cumulative GC-skew plot (a plot that measures the change in G versus C over a moving window) of the main-sense strand in the *P. capuana* mtDNA occurs at the apex (nucleotide 9,934) of the largest stem-loop structure in the genome — that between the regions encoding the *rrnl*-L1 and *rrns*-S4 gene fragments (Figure 2.2). The global minimum of a GC-skew plot is often used to predict the origin of replication in bacterial and mitochondrial genomes (Grigoriev 1998). A similar but much weaker stem-loop structure was found in the mitochondrial genome of *P. parva* at the corresponding region.

Telomeric Repeats: The terminal regions (telomeres) of the *P. capuana* mitochondrial genome proved unamenable to standard cloning techniques. Sequencing of the telomeres was thus achieved using terminal deoxynucleotidyl transferase (TdT) tailing (Förstemann et al. 2000; Bah et al. 2004) and long walk PCR (Katz et al. 2000). The nature of these protocols is that the TdT tailing method allows access to the 3′ end of a telomere, whereas the long walk PCR approach works outwards on the strand containing the 5′ end; by using each of these techniques, one can sequence both strands of a telomere independently.

Sequencing results from TdT tailing and long walk PCR (Figure 2.3) suggest that the terminal regions of the *P. capuana* mtDNA exist (*in vitro*) in two separate conformations: a closed (hairpin loop) conformation and an open (nicked loop) conformation. The nucleotide sequence from both these conformations appears to be identical; however, that from the closed conformation seems to terminate with a 220 nt single-stranded loop, whereas in the open conformation this loop appears nicked (Figure 2.3). The location of the nick was shown to vary but was most often observed at the apex of the loop. Further experiments using standard PCR techniques were performed to confirm the sequence of the telomeric regions. Gel electrophoresis analyses and restriction digest results also support the idea that the *P. capuana* mtDNA telomeres can exists in either a open or a closed conformation (Figures 2.4 and 2.5).

Discussion

Nucleotide-composition Bias in the P. capuana *MtDNA*

The *P. capuana* mtDNA has the highest GC content (57.2%) of any completely sequenced mitochondrial (or organelle) genome currently deposited in the NCBI organelle genome data bank (http://www.ncbi.nlm.nih.gov/genomes/static/euk_o.html; July, 2007). Heretofore, all completely sequenced green algal mitochondrial genomes have had a GC content below 46%, which leads to the question: why is the mtDNA of *P. capuana* GC rich? Or alternatively, why is it so AT poor? Hypotheses about biased nucleotide composition fall into two categories, which we will call: selectionist- and neutralist-based. The former suggest that nucleotide bias is the result of natural selection; for example, this approach is often used to argue that GC richness is an adaptation to homeothermy (Jabbari and Bernardi 2004) or UV tolerance (Singer and Ames 1970; Kellogg and Paul 2002). In opposition, neutralist-based hypotheses posit that nucleotide inequities arise by either biased mutation pressure or biased gene conversion (BGC). Biased mutational pressure is often applied to the low GC content of organelle and endocellular parasite genomes (Dybvig and Voelker 1996; Ogata et al. 1998; Kusumi and Tachida 2006), whereas BGC, which results from biased repair of mismatches in heteroduplexed recombination intermediates (Eyre-Walker 1993; Holmquist 1992; Galtier et al. 2001; Galtier and Duret 2007), has been correlated to GC richness. Furthermore, gene conversion, which is believed to occur in organelle genomes (Birky and Walsh 1992; Walsh 1992; Khakhlova and Bock 2006; Tatarenkvo and Avise 2007), may be GC biased in mitochondrial systems: a positive correlation between the GC content of repeat sequences and their recombinogenic properties in mitochondrial genomes has been observed in fungi and plants (Dieckmann and Gandy 1987; Clark-Walker 1989; Weiller et al. 1991; Nakazono et al. 1995) and has been suggested for certain *Reinhardtinia*-clade green algae (Boer and Gray 1991; Nedelcu 1997, 1998; Nedelcu and Lee 1998).

Within the *P. capuana* mtDNA, the GC bias is highest at what are typically regarded as among the most neutrally evolving positions (intergenic and four-fold degenerate sites) and it is lowest at what are generally considered more functionally constrained positions (first and second codon sites). This disproportion of GC content

among the different regions of the *P. capuana* mitochondrial genome is best explained by the negative selection principle of the neutral theory of molecular evolution (Kimura 1983). If a neutral process is responsible for the GC bias of *P. capuana* mtDNA, is that process the effect of biased mutation pressure or BGC_{GC}? — or both? Any attempt to choose between these possibilities is purely speculative; however, we tend to favor a predominant role of BGC_{GC}, for reasons discussed below.

Duret et al. (2006) proposed that in the nuclear DNA of mammals the GC content of a given region reflects a balance between an AT-biased mutation process and BGC_{GC} (a GC-biased fixation process). They further suggest (following the theory of Galtier 2004) that GC-rich regions form rapidly during times when the recombination rate is high enough for BGC_{GC} to be effective; then, once the recombination rate decreases, the GC content declines slowly as a result of AT-mutation pressure. If we suppose that mitochondrial genomes have an AT-biased mutation process and that gene conversion in mtDNA is GC biased, we can apply Duret's model to mitochondrial systems. Under this premise, the history of most mitochondrial genomes would reflect an AT-biased mutation process and a recombination rate where BGC_{GC} is ineffective, thereby giving rise to AT-rich mitochondrial genomes. For the mtDNA of *P. capuana*, however, we suggest a recent history with high levels of recombination, thus, shifting the nucleotide composition towards G and C. According to this hypothesis, any new mutation in one of the multiple copies of the *P. capuana* mtDNA that results in heteroplasmy, where one allele is A•T at a given site and the other allele is G•C at the same position, should be preferentially converted to the G•C allele via BGC_{GC}. Reasons why the *P. capuana* mtDNA may have undergone an increased rate of recombination would have to be entirely suppositional at this time. One intriguing observation, however, is that the isolate of *P. capuana* used for this research came from Italy (Capua), whereas the available mtDNA sequences of *P. parva* come from a strain isolated in the UK (Cambridge). Sun exposure between these two regions differs substantially, invoking the possibility that an elevated recombination rate may be a repair response to UV damage in the *P. capuana* mtDNA; but because little is known about the full habitat range of either *P. capuana* or *P. parva*, little weight can be placed on this observation. Moreover, we have been unable

to get *P. capuana* to form cysts in the laboratory, unlike the other known species of *Polytomella*, which may make it more susceptible to UV damage.

The P. capuana *MtDNA in Relation to Other Linear Mitochondrial Genomes*

The *P. capuana* mitochondrial genome is one of several examples of linear mtDNA from the *Reinhardtinia* clade of chlorophycean green algae (for a compilation see Laflamme and Lee 2002; Mallet and Lee 2006; Popescu and Lee 2006). When comparing the telomeres of the *P. capuana* mtDNA with those of other *Reinhardtinia*-clade algae for which telomere data are available (namely *C. reinhardtii* and *P. parva*), no universal themes are apparent. Both the length and sequence of the telomeres in the *P. capuana* mtDNA differ substantially from those of *P. parva* and *C. reinhardtii.* Furthermore, the terminal structures of the *P. capuana* mtDNA, which appear to exist in either closed (hairpin loop) or open (nicked loop) conformations, are different from those of *C. reinhardtii*, which are made up of 3´ single-stranded extensions (Vahrenholz et al. 1993). We are unable to exclude the possibility that the open telomeric conformation is the result of nicking during the DNA extraction process and does not normally exist *in vivo*. *P. capuana* is not the first example of a linear mtDNA with terminal hairpins; they are found in the linear mitochondrial genome of the yeast *Pichia* (Dinouël et al. 1993), at one end of the *Paramecium* mtDNA (Pritchard and Cummings 1981), and also in the mitochondrial plasmid of the plant pathogenic fungus *Rhizoctonia solani* (Miyashita et al. 1990). Other examples of this telomeric structure come from viruses of eukaryotic cells (Baroudy et al. 1982; Gonzalez et al. 1986), including a virus that infects certain species of the green-algal genus *Chlorella* (Rohozinski et al. 1989), and from the bacterial plasmids of the genus *Borrelia* (Hinnebusch and Barbour 1991).

All linear genomes must develop a strategy to overcome the end-replication problem, as defined by Olovnikov (1971) and Watson (1972). For the *C. reinhardtii* mitochondrial genome, two replication models have been proposed (Vahrenholz et al. 1993). One model involves reverse transcription of an internal repeat via a putative mitochondrial encoded reverse transcriptase (*rtl*), whereas the second model takes into account that the reverse transcriptase gene may be non-functional. No open reading frames resembling a reverse transcriptase-like gene were found in the *P. capuana* (or *P.*

parva) mitochondrial genome, and the fact that the structure of its telomeres depart from that of *C. reinhardtii* suggest that it uses a different replication strategy. Although in the case of *P. capuana* no strategy is apparent, many replication models for linear genomes with terminal hairpins have been suggested (Cavalier-Smith 1974; Bateman 1975; Pritchard and Cummings 1981; Baroudy et al. 1983; Dinouël et al. 1993; Traktman 1996).

Origin of MtDNA Fragmentation in the Polytomella *Genus*

Although linear, fragmented mitochondrial genomes have been observed in other eukaryotic lineages, including four classes of *Cnidaria* (Warrior and Gall 1985; Bridge et al. 1992; Ender and Schierwater 2003) and the ichthyosporean *Amoebidium parasiticum* (Burger et al. 2003), the *Polytomella* genus represents a unique example in that substantial mtDNA sequence data exist for both a linear fragmented and a linear intact genome from two closely related taxa, thus, allowing for comparative analyses.

Assuming that the linear bipartite mitochondrial genome of *P. parva* was derived from an ancestral unfragmented, linear molecule, we can posit that the ancestral mtDNA conformation may have been similar to that of *P. capuana*: a single linear chromosome with terminal inverted repeats. Furthermore, since the gene arrangement of the unfragmented *P. capuana* mitochondrial genome is parallel to that of the fragmented *P. parva* mitochondrial genome — fragmentation notwithstanding — we can consider the ancestral *Polytomella* gene arrangement to be equivalent to that of *P. capuana*, where the gene encoding *nad6* is found internal to the left telomere. Under these premises, the *P. capuana* mitochondrial genome can act as a model for understanding fragmentation of the ancestral *Polytomella* mtDNA.

Several features of the *P. capuana* mtDNA suggest that the *nad6* gene is in an unstable region. Its terminal position lends itself to recombination and possible fragmentation more readily than other internally located genes: recombination rates have been shown to be higher at the termini of linear chromosomes as compared to their more centrally located regions (Eichler and Sankoff 2003; See et al. 2006;). Also, the intergenic sequence between the *nad6* and *cob* genes is comprised of a potentially unstable GC-rich inverted repeat: in the mitochondrial genomes of *Neurospora cerevisiae*

and *Saccharomyces cerevisiae* GC-rich inverted repeats were shown to have inflated rates of recombination causing genome rearrangements or deletion mutations that were maintained in the population of mtDNAs (Clark-Walker 1989; Almasan and Mishra 1988). In fact, portions of the *P. capuana nad6*/*cob* intergenic sequence were found inserted into both *nad5* and the telomeres, indicating that this sequence may have mobile properties, perhaps similar to those of the GC clusters found in the mtDNA of certain *Saccharomyces* species (de Zamaroczy and Bernardi 1986). By using a portion of the short inverted repeat sequences in the *nad6*/*cob* intergenic region and homologous sequences in the telomere regions, we were able to outline a scenarios involving illegitimate recombination between *P. capuana* mitochondrial genomes, which produce products structurally similar to the 3.5 kb and 13.5 kb mtDNAs of *P. parva*, respectively (Figure 2.6A). *Polytomella* lineages with fragmented mtDNA forms (as described in Figure 2.6A) may have become fixed by random genetic drift, especially if the populations went through a bottleneck. According to this possibility, one might expect an ongoing, low-level production of such fragmented mtDNA forms from the intact mtDNA structure in *P. capuana*. Although we were not able to detect these forms by a PCR approach, we were able to reliably detect PCR products, using a wide range of primer combinations, that are consistent with other illegitimate recombination events involving the short inverted repeat sequences in the telomere regions as shown in Figure 2.6B.

Short Inverted Repeats in MtDNA

The short GC-rich inverted repeat sequences in the *P. capuana* mtDNA evoke several questions regarding their evolution and function, such as: i) do they play a role in gene expression, ii) do they have mobile properties, and iii) are they related to the inverted repeats found in other mitochondrial genomes. Inverted repeat sequences capable of forming stem-loop structures have been described in the mitochondrial genomes of animals, fungi, plants, and a series of *Reinhardtinia*-clade algae including *C. reinhardtii* (Boer and Gray 1986; Boer and Gray 1991), and two species of *Volvox* (Aono et al. 2002). In the above cases the inverted repeat sequences are generally restricted to intronic or intergenic regions, and in many cases they have been implicated in mobility or RNA processing (Boer and Gray 1986; Boer and Gray 1991; Nedelcu and Lee 1998;

Aono et al. 2002). What distinguishes the inverted repeats of the *P. capuana* mtDNA from those of most other mitochondrial systems is that they span both coding and noncoding DNA, often resulting in stem-loop structures that contain the start or the end of a gene; this arrangement suggests that the inverted repeats and their putative secondary structures may have a role in gene expression, perhaps akin to mammalian mitochondrial systems, where large poly-cistronic transcripts are processed by cleavage at the boundaries of tRNA sequences, which flank almost every gene (Clayton 1984). We are aware of only one other mitochondrial genome with a similar orientation of inverted repeats: the dinoflagellate *Amphidinium carterae* (Nash et al. 2007); however, the utility of the repeats within this taxon are also unknown. The absence of inverted repeat sequences from the *P. parva* mitochondrial genome indicates that the inverted repeats in the *P. capuana* mtDNA may be invasive elements, conceivably appearing in *P. capuana* after it diverged from its common ancestor with *P. parva*; but we can not eliminate the possibility that the elements were lost from the *P. parva* mtDNA. Two of the inverted repeats share sequence identity with other parts of the genome; 20 nt of the *cox1*/*nad4* hairpin structure was found inserted into *nad6*, and 12 nt of the *cob/nad6* stem was inserted into *nad5*. This suggests that some of the inverted repeat sequences in the *P. capuana* mitochondrial genome may have mobile properties, but the lack of conservation in primary sequence among the different inverted repeats implies that if there is mobility it may be dependent on secondary structure rather than a specific nucleic acid sequence.

Table 2.1. Nucleotide Composition by Region of the *P. capuana* MtDNA

	Overall			Coding[a]			Intergenic			Telomeres			Codon Site Position								
													Site 1			Site 2			Site 3		
	Pc	*Pp*	*Cr*	*Pc*	*Pp*	*Cr*	*Pc*	*Pp*	*Cr*	*Pc*	*Pp*	*Cr*	*Pc*	*Pp*	*Cr*	*Pc*	*Pp*	*Cr*	*Pc*	*Pp*	*Cr*
%A	19.5	27.2	27.6	18.2	25.9	22.7	15.0	38.4	30.0	25.6	24.9	26.7	22.9	26.2	24.7	17.5	18.9	16.4	8.1	27.5	15.5
%T	23.3	31.8	27.2	25.4	34.4	32.2	17.1	32.7	30.5	15.7	31.2	18.8	24.8	29.5	27.2	41.2	41.8	42.8	15.9	37.4	38.4
%C	28.4	19.9	22.4	28.4	19.4	21.9	35.0	15.0	20.7	27.8	22.9	31.2	20.3	16.9	17.5	22.8	20.9	21.5	42.8	20.9	26.3
%G	28.8	21.1	22.8	28.0	20.3	23.2	33.0	13.9	18.8	30.9	21.1	23.3	31.9	27.3	30.5	18.5	18.4	19.3	33.2	14.2	19.8
%GC	57.2	41.0	45.2	56.4	39.7	45.1	68.0	28.9	39.5	58.7	44.0	54.5	52.2	44.2	48.0	41.3	39.3	40.8	76.0	35.1	46.1

Note: *Pc – Polytomella capuana*; *Pp – Polytomella parva*; *Cr – Chlamydomonas reinhardtii*.

[a]Based on protein-, rRNA-, and tRNA-coding regions.

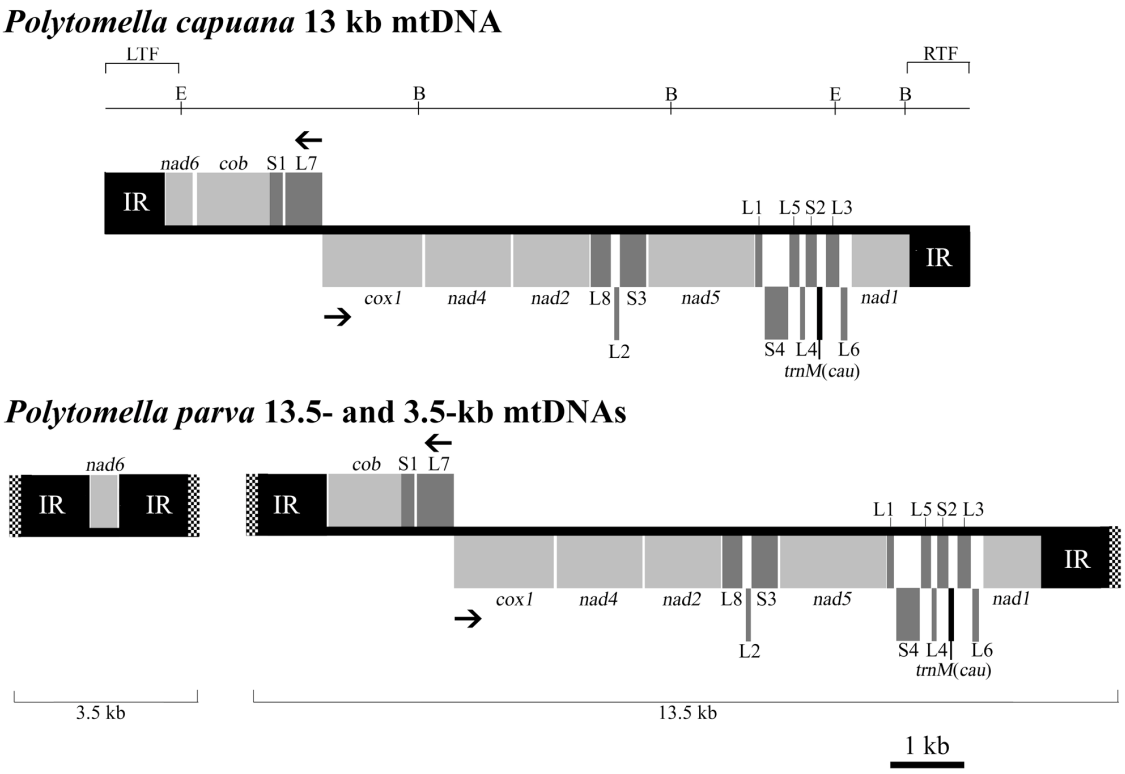

Figure 2.1. Genetic Maps of the *P. capuana* and the *P. parva* MtDNAs

Protein-coding, rRNA-coding, and terminal inverted repeat (IR) regions are shown in light gray, dark gray, and black, respectively. S1-S4 and L1-L8 denote the small subunit and large subunit rRNA-coding modules, respectively; *trnM* (*cau*) signifies the gene for $tRNA^{met}$. Thick solid arrows indicate transcriptional polarities. Restriction sites for *Eco*RI (E) and *Bam*H1 (B) are shown on the *P. capuana* mtDNA map, and the fragments identified by brackets above this restriction map (left terminal restriction fragment [LTF] and right terminal restriction fragment [RTF]) represent the location of molecular probes used in this work. The checkered regions on these maps are at present unsequenced.

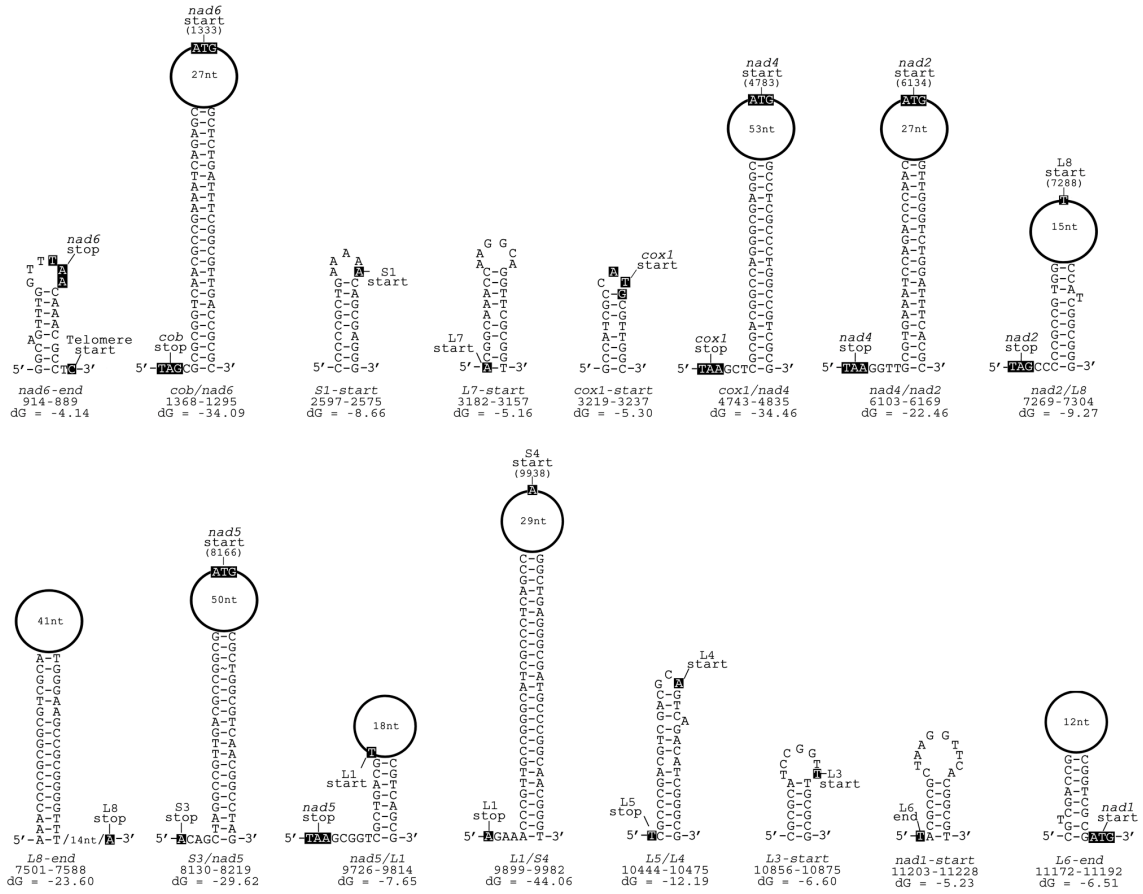

Figure 2.2. Putative Stem-loop Structures Inferred From the *P. capuana* MtDNA Sequence

Name (based on genome location), genome coordinates (nt), and folding energy (dG) at 37°C are given beneath each secondary structure, respectively. If the "loop" component of the secondary structure is larger than 10 nt, it is depicted by a hollow black circle with its size shown in the center. The predicted start and stop sites of coding regions are shaded; when these sites occur within one of the depicted loop structures their position in the genome is shown.

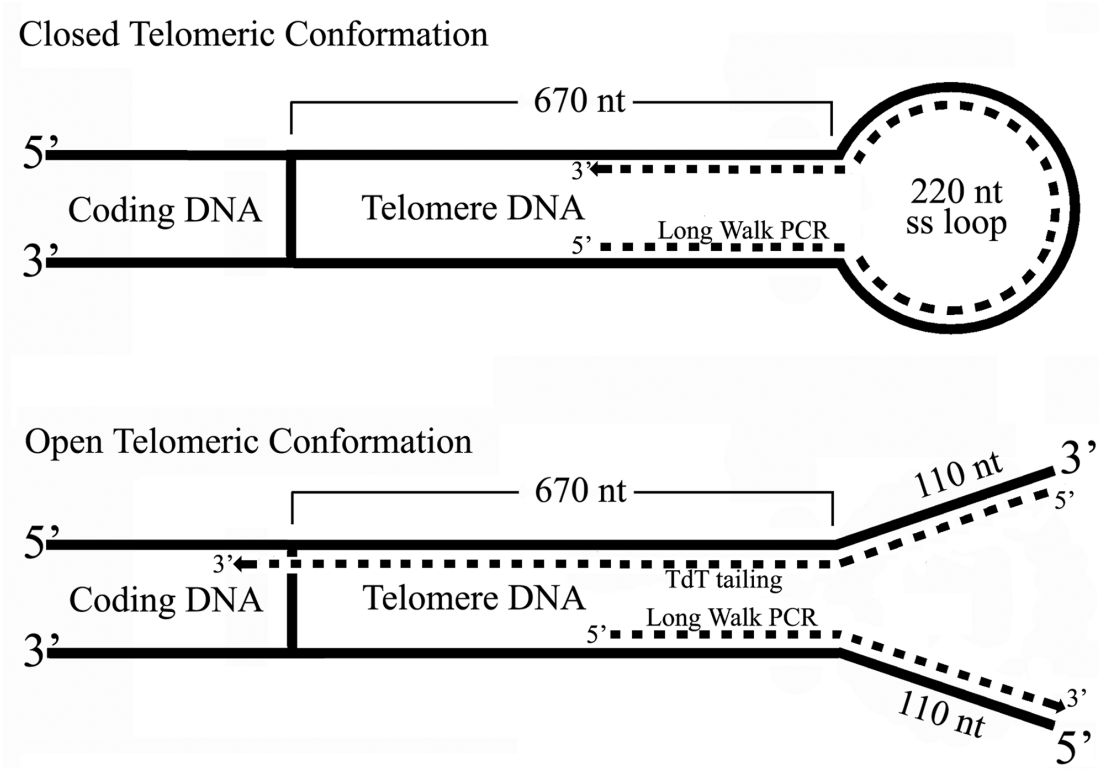

Figure 2.3. MtDNA Telomere Conformation of *P. capuana*

TdT tailing and long walk PCR reactions were independently performed on terminal restriction fragments (LTF and RTF from Figure 2.1) coming from the left and right ends of the *P. capuana* mtDNA. Sequencing of the long walk PCR and TdT tailing products suggest that the telomeres can exist in either a closed (hairpin loop) or an open (nicked loop) conformation. Products obtained by either long walk PCR or TdT tailing are depicted with a dashed line.

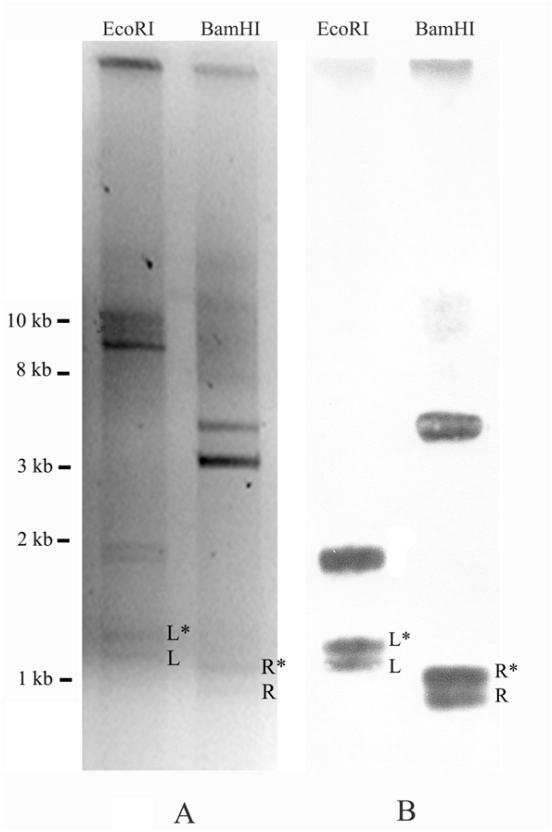

Figure 2.4. Southern Blot Analysis of the *P. capuana* MtDNA-enriched Fraction with Telomere Probes

(A) When the mtDNA of *P. capuana* is digested with *Eco*RI or *Bam*HI and run on an ethidium-bromide-stained agarose (1%) gel each of the terminal fragments, as defined by the restriction map (Figure 2.1), appear as two bands: the left terminal fragment (L [main band] and L* [shadow band]) and the right terminal fragment (R [main band] and R* [shadow band]). (B) The Southern blot of the gel probed with a labeled 700 nt PCR-amplified portion of the telomeres shows that both the main bands and shadow bands are efficiently labeled. These results are consistent with telomere structures where the main band represents a closed (hairpin loop) conformation and the shadow band represents an open (nicked loop) conformation (Dinouël et al. 1993).

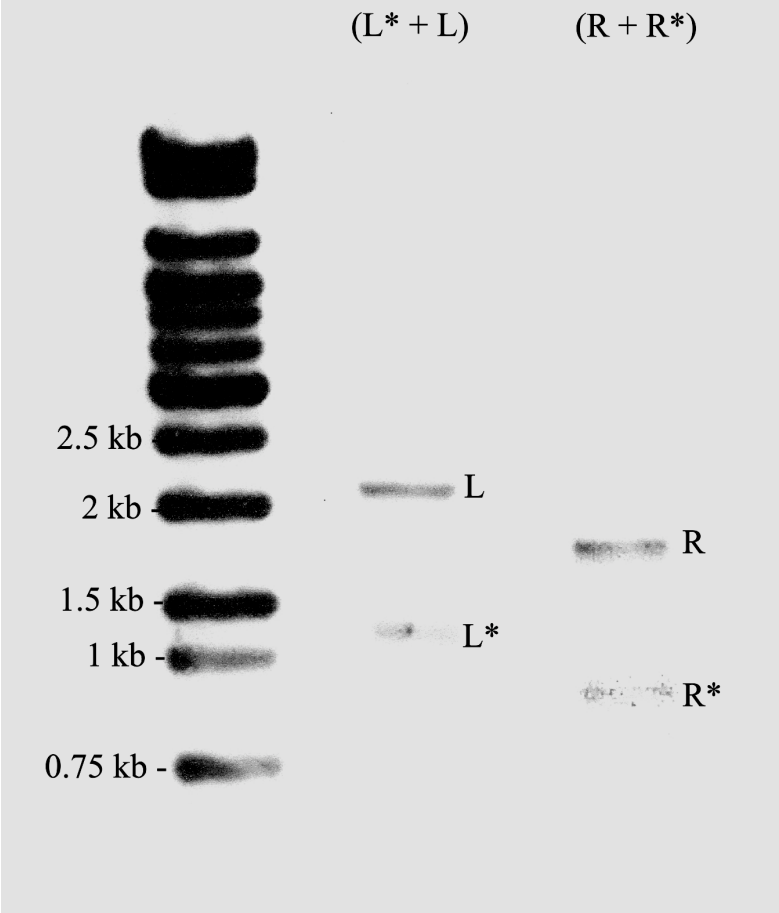

Figure 2.5. Denaturing Gel Electrophoresis and Southern Blot Analysis of the *P. capuana* MtDNA Terminal Restriction Fragments

The *Eco*R1-generated left terminal main band (L) and shadow band (L*) and *Bam*H1-generated right terminal main band (R) and shadow band (R*) restriction fragments were excised from a non-denaturing agarose gel and run under denaturing conditions in a formamide agarose (1%) gel. Based on their migration pattern under nondenaturing conditions (Figure 2.4), the main-band restriction fragments (L and R) migrated as molecules of twice their expected lengths while the shadow-band restriction fragments (L* and R*) migrated as molecules of their expected lengths. This suggests that the main-band restriction fragments are held together by a terminal loop structure. The initial lengths (kb) of the nondenatured fragments were as follows: L (1.1), L* (1.2), R (0.9), R* (1.0). In the denaturing gel, L and R gave bands of approximately 2.1 kb and 1.8 kb, respectively. The labeled probe was a 700 nt PCR-amplified portion of the telomeres. The Fermentas GeneRuler 1 kb DNA ladder is shown in the left-most lane.

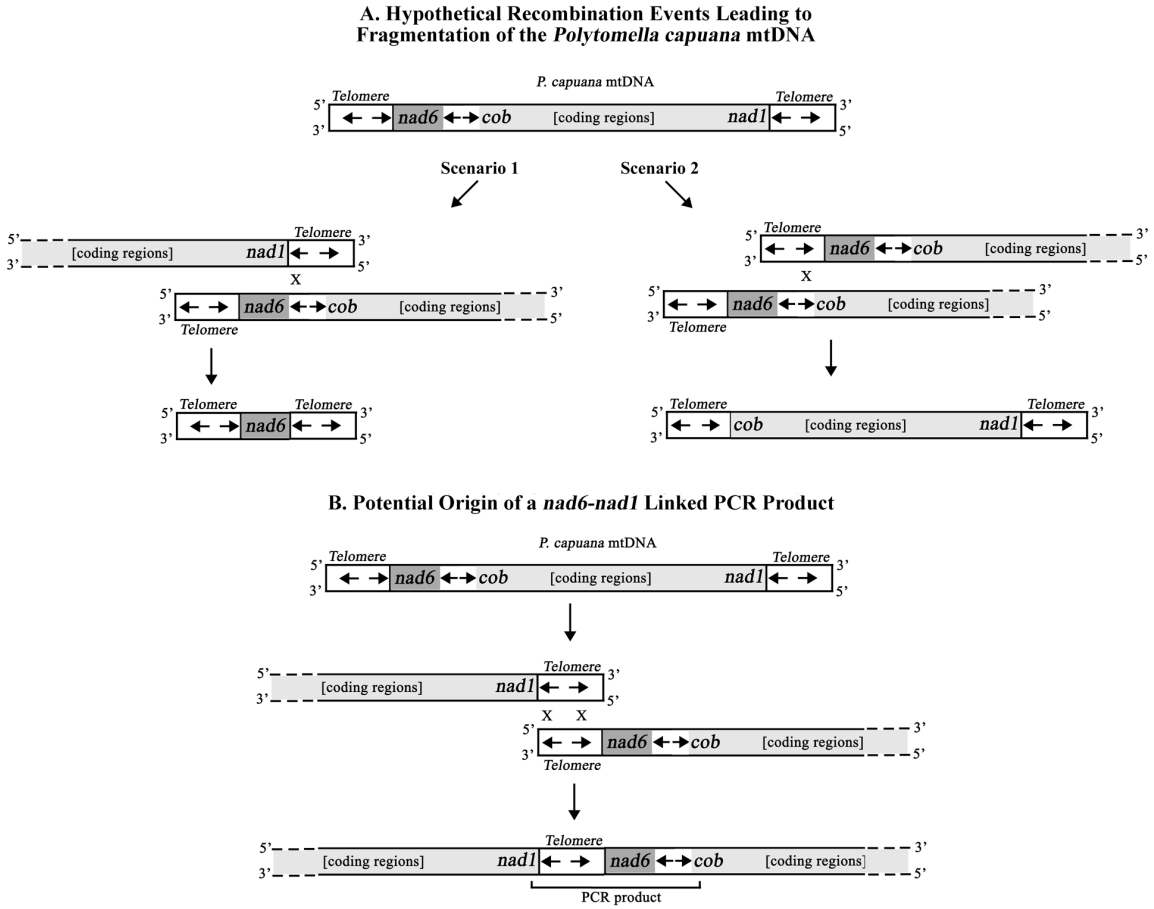

Figure 2.6. Fragmentation of *Polytomella* MtDNA

(A) Outline of two scenarios involving illegitimate recombination between short inverted repeats in the *P. capuana* mtDNA that produce products structurally similar to the 3.5 kb and 13.5 kb mtDNAs of *P. parva*. (B) PCR product consistent with an illegitimate recombination event involving the short inverted repeat sequences in the telomere regions resulting in a *nad6*/*nad1*-linked fragment. Arrows represent the short inverted repeats.

CHAPTER 3: EVOLUTION OF MITOCHONDRIAL DNA IN THREE KNOWN LINEAGES OF *POLYTOMELLA*

Published as:

Smith DR, Jimeng Hua, Lee RW (2010) Evolution of linear mitochondrial DNA in three known lineages of *Polytomella*. Curr Genet *In press*.

Abstract

Although DNA sequences of linear mitochondrial genomes are available for a wide variety of species, sequence and conformational data from the extreme ends of these molecules (i.e., the telomeres) are limited. Data on telomeres are important because they can provide insights into how linear genomes overcome the end-replication problem. This study explores the evolution of linear mitochondrial DNAs (mtDNAs) in the green-algal genus *Polytomella* (Chlorophyceae, Chlorophyta), the members of which are non-photosynthetic. Earlier works analyzed the linear and linear, fragmented mitochondrial genomes of *Polytomella capuana* and *Polytomella parva*. Here we present the mtDNA sequence for *Polytomella* strain SAG 63-10 (also known as *Polytomella piriformis* [Pringsheim 1963]), which is the only known representative of a mostly unexplored *Polytomella* lineage. We show that the *P. piriformis* mtDNA is made up of two linear fragments of 13 and 3 kilobases (kb). The telomeric sequences of the large and small fragments are terminally inverted, and appear to end *in vitro* with either closed (hairpin loop) or open (nicked loop) structures as also shown here for *P. parva* and shown earlier for *P. capuana*. The structure of the *P. piriformis* mtDNA is more similar to that of *P. parva*, which is also fragmented, than to that of *P. capuana*, which is contained in a single chromosome. Phylogenetic analyses reveal high substitution rates in the mtDNA of all three *Polytomella* species relative to other chlamydomonadalean algae. These elevated rates could be the result of a greater number of vegetative cell divisions and/or small population sizes in *Polytomella* species as compared with other chlamydomonadalean algae.

Introduction

Although it was once assumed that all mitochondrial genomes are circular molecules (or at least circularly-mapping molecules), it is now well established that linear and linear fragmented mitochondrial genomes have evolved numerous times in diverse eukaryotic lineages (for examples and references see Table 3.1). Linear mitochondrial DNAs (mtDNAs) are defined not only by the fact that they migrate as linear molecules in gel electrophoresis analyses but also by the presence of specialized terminal structures called telomeres (Nosek and Tomáska 2002, 2003). Nucleotide sequence data of linear

mitochondrial genomes are available for a wide variety of species (reviewed by Nosek et al. [2004a]); however, in many cases sequence data from the extreme ends (i.e., telomeres) of linear mtDNAs are unavailable. This is primarily because the telomeres of linear mitochondrial genomes often have terminal conformations that are unamenable to PCR and cloning, such as 3′ or 5′ overhangs or covalently-closed ends (Nosek and Tomáska 2002, 2003). Indeed, many of the linear mtDNAs that are listed in GenBank as "complete genome sequences" are missing sequence data from their telomeres. This is unfortunate because by knowing the sequence and end conformation of telomeres, one can gain insights into how linear mitochondrial genomes overcome the end-replication problem, as defined by Olovnikov (1971) and Watson (1972), a problem which is not only interesting from a genome-evolution perspective but one which has important implications for cancer research (Nosek et al. 2004a).

One eukaryotic lineage that has many species with linear mtDNAs is the "*Reinhardtinia* clade," *sensu* Nakada et al. (2008), a monophyletic group of green algae found within the chlorophycean class of the Chlorophyta (Lewis and McCourt 2004). Almost all of the *Reinhardtinia*-clade mtDNAs that have been examined appear to be either linear or linear fragmented (Laflamme and Lee 2003) with the exception of the *Volvox carteri* mtDNA, which assembles as a circular molecule (Smith and Lee 2010; Chapter 7). Conversely, all of the characterized green algal mitochondrial genomes from outside this clade map as circular molecules, save for the mtDNAs of some *Lobochlamys* taxa, which may be linear fragmented (Borza et al. 2009).

The first mtDNA from the *Reinhardtinia* clade to be completely sequenced was that of the model organism *Chlamydomonas reinhardtii* (Gray and Boer 1988; Michaelis et al. 1990). This 16 to 19 kilobase (kb) linear genome has inverted-repeat telomeres of ~600 nucleotide (nt) that each contain an ~40 nt, non-complementary 3′ extension. The only other well-studied *Reinhardtii*-clade linear mtDNAs come from algae within the *Polytomella* genus — a group of non-photosynthetic, naturally-wall-less unicells (Pringsheim 1955).

Fan and Lee (2002) sequenced the *Polytomella parva* mtDNA and reported that it is made up of two linear fragments of ~13.5 and ~3.5 kb with nine and one gene, respectively. Like the *C. reinhardtii* mtDNA, the telomeres of both fragments form ~1.3

kb inverted repeats; however, it was assumed that the DNA sequence of the extreme telomere ends was not determined and no potential telomere conformation was proposed. Further work by Mallet and Lee (2006), using a *cox1* phylogeny, revealed that there are at least three distinct lineages of *Polytomella*: two that are closely related and represented by *P. parva* and *Polytomella* strain SAG 63-10 (referred to hereafter by its original *nomen nudum* name of *Polytomella piriformis* [Pringsheim 1963]), and a third, deep-branching lineage represented by *Polytomella capuana*. Gel-electrophoresis- and Southern-blot-analyses suggested that the *P. capuana* mtDNA is contained in a single linear chromosome (Mallet and Lee 2006), whereas unpublished data (Mallet MA and Lee RW, personal communication) seemed to indicate that the mitochondrial genome of *P. piriformis*, like that of *P. parva*, is a linear bipartite molecule.

Smith and Lee (2008a; Chapter 2) confirmed the linear monomeric conformation of the *P. capuana* mitochondrial genome by obtaining its complete DNA sequence. They found the *P. capuana* mtDNA to be a 13-kb linear molecule with ~0.9 kb inverted-repeat telomeres that terminate *in vitro* with either a closed (hairpin-loop) or an open (nicked-loop) conformation. It was also shown that the gene located on the ~3.5 kb mtDNA chromosome in *P. parva* (i.e., *nad6*) is found internal to the left telomere on the *P. capuana* mtDNA; based on the position of this gene, Smith and Lee (2008a; Chapter 2) proposed a recombination-driven scenario of how the linear intact mtDNA of *P. capuana* could be converted into a linear fragmented form like that of *P. parva*. An additional and unexpected observation was that the *P. capuana* mtDNA, unlike that of *P. parva*, is rich in guanine and cytosine (GC rich), which is an unusual and, at that time, unprecedented trait for an organelle genome (Smith and Lee 2008a; Chapter 2).

This study examines mitochondrial-genome evolution in *P. piriformis*, the only known representative of a third and mostly unexplored *Polytomella* lineage (Mallet and Lee 2006). The mtDNA sequence of *P. piriformis*, including the sequence and structure of the telomeres, is described and compared with those from other *Polytomella* species. Moreover, to complement our *P. piriformis* mtDNA sequence data, we confirmed the sequence of the *P. parva* mtDNA telomeres, which were reported to be only partially sequenced (Fan and Lee 2002). The ultimate aims of this study are to: i) confirm if the mtDNA of *P. piriformis* is fragmented like that of *P. parva* or linear-intact like that of *P.*

capuana; ii) gain insight into mtDNA telomere evolution and how linear mitochondrial genomes overcome the end-replication problem; and iii) provide complete mtDNA sequence data, including the structure of the telomeres, for members from each of the three known *Polytomella* lineages.

Materials and Methods

Polytomella *Strains*

Polytomella strain SAG 63-10 was obtained in 2004 from the Sammlung von Algenkulturen in Göttingen, Germany, whereas *P. parva* (UTEX L 193) was obtained from the Culture Collection of Algae at the University of Texas at Austin around 1994. SAG 63-10 was originally isolated by E.G. Pringsheim at the Göttingen Botanical Gardens sometime before 1963 (Pringsheim 1963; SAG website [http://sagdb.uni-goettingen.de/]). When Pringsheim first described strain 63-10 he named it *Polytomella piriformis* (Pringsheim 1963), a name we will use here; however, this name is currently considered a *nomen nudum* because a taxonomically valid description of *P. piriformis* was never published.

Sequencing the P. piriformis *and* P. parva *MtDNAs*

Culturing, mitochondrial enrichment, DNA extraction and amplification, mtDNA sequencing, and Southern blot analyses of *P. piriformis* and *P. parva* followed the same protocols as previously described for *P. capuana* (Smith and Lee 2008a; Chapter 2). Reverse transcriptase (RT) PCR reactions were performed with the SuperScript III One-Step RT-PCR System (Invitrogen, Carlsbad, CA) following the manufacturer's protocol. DNA sequencing was performed on both strands using a 3730xl DNA Analyzer (Applied Biosystems, Foster City, CA) at the Macrogen Sequencing Facility in Rockville, MD.

Sequence Analyses

Sequences were assembled using CodonCode Aligner Version 3.0.1 (CodonCode Corporation, Dedham, MA), which employs the Phred, Cross-match, and Phrap algorithms for base calling, sequence comparison, and sequence assembly, respectively.

Assemblies were performed with a minimum-percent-identity score of 98, a minimum-overlap length of 500 nt, a match score of 1, a mismatch penalty of -2, a gap penalty of -2, and an additional first-gap penalty of -3. The boundaries of the *P. piriformis* mitochondrial rRNA-coding modules were estimated by sequence comparisons with their counterparts in the *P. parva* and *P. capuana* mtDNAs. Multiple alignments of DNA sequences were performed with CLUSTAL W (Thompson et al. 1994) using the default settings. The maximum-likelihood phylogeny was constructed with PAUP* version 4.0b10 (Swofford 2003) using the general time reversible model (GTR + I + G), which was chosen by Modeltest (version 3.7 [Posada and Crandal 1998]) as the best-fit model of nucleotide substitution for our given dataset. The tree was built using the first and second codon positions of the mtDNA-encoded proteins shared among all completely sequenced chlorophycean mtDNAs.

Nucleotide Sequence Accession Numbers

The nucleotide sequence of the *P. piriformis* mitochondrial genome is deposited in GenBank under the accession numbers GU108480 (13 kb fragment) and GU108481 (3 kb fragment).

Results and Discussion

P. parva *MtDNA Telomeres*

To complement our *P. piriformis* mtDNA sequence data, we attempted to finish sequencing the *P. parva* mtDNA telomeres. Fan and Lee (2002), using agarose gel electrophoresis data, estimated that the *P. parva* mitochondrial genome is separated into two fragments of 13.5 and 3.5 kb. Thus, when Fan and Lee (2002) cloned, sequenced, and assembled the two *P. parva* mtDNA fragments and ended up with contigs of 13 and 3 kb, they believed that they were missing approximately 250 nt of telomeric sequence from the ends of each chromosome. Here, using DNA isolated from a mitochondria-enriched fraction of *P. parva* (UTEX L 193), we performed terminal deoxynucleotidyl transferase (TdT) tailing experiments (Förstemann et al. 2000 and Bah et al. 2004) on the large and small mtDNA fragments. Analyses of the resulting telomeric sequence data

indicate that Fan and Lee (2002) completely sequenced the *P. parva* mitochondrial genome — we found no evidence that the telomeres extend beyond the sequence that they generated. Therefore, the two *P. parva* mtDNA fragments appear to have lengths of ~13 and 3 kb. The reason why these mtDNAs migrate as ~13.5 and 3.5 kb molecules in agarose gel electrophoresis analyses may be linked to the conformation of their telomeres (discussed further below).

Characterization of the P. piriformis *Mitochondrial Genome*

When DNA was isolated from a mitochondria-enriched fraction of *P. piriformis* and run on agarose gels, two distinct bands with estimated sizes of 13.5 and 3.5 kb (referred to hereafter as the large and small fragments) were consistently observed after staining with ethidium bromide (Figure 3.1). Both of these bands co-migrated with linear DNA markers under various concentrations of agarose (data not shown), supporting the hypothesis that they represent linear molecules. DNA sequencing confirmed that both fragments are components of the *P. piriformis* mitochondrial genome. This conclusion was supported by Southern blot hybridization experiments using *P. piriformis* mtDNA probes (Figure 3.2). Southern blot analyses also identified a *P. piriformis*-mtDNA-hybridizing band of ~1.8 kb, which was not visible on ethidium bromide stained gels but was consistently observed in Southern blots using different DNA isolates of *P. piriformis* (Figure 3.2). Probes corresponding to the *P. piriformis* mtDNA telomeres as well as those coming from the gene found on the small fragment (*nad6*) hybridized to the 1.8 kb component, whereas those designed from genes on the large fragment did not hybridize to this component. For these reasons, we believe that the 1.8 kb component is derived from the small fragment but does not represent a primary constituent of the *P. piriformis* mitochondrial genome. Similar findings were observed for *Polytomella* strains coming from the *P. parva* lineage, where in addition to the two main mtDNA bands, Southern blot analyses identified mtDNA-hybridizing components with sizes of 1.8 and 2.1 kb (Fan and Lee 2002; Mallet and Lee 2006); moreover, *Polytomella papillata* and *Polytomella magna* (both of the *P. parva* lineage) also showed a mtDNA-hybridizing band of 5 kb.

General Features of the P. piriformis *Mitochondrial Genome*

Complete genetic maps of the *P. piriformis* large- and small-mtDNA fragments are shown in Figure 3.3A. For comparison, this figure also contains the mtDNA genetic maps of *P. capuana* (Smith and Lee 2008a; Chapter 2), *P. parva* (Fan and Lee 2002), and *C. reinhardtii* (Gray and Boer 1988). Note that unlike the mtDNAs of *P. piriformis* and *P. parva*, which are linear bipartite genomes, those of *P. capuana* and *C. reinhardtii* are linear monomeric molecules (Figure 3.3A). The sequenced lengths of the two *P. piriformis* mtDNA fragments are 13,004 and 3,079 nt (accumulative length = 16,083 nt), which is about the same as that of the two *P. parva* mtDNAs (3,018 and 13,135 nt; accumulative length = 16,153 nt) and around 3 kb larger than the linear monomeric mtDNA of *P. capuana,* which, at 12,998 nt, is the smallest archaeplastidial mitochondrial genome observed to date. As was observed for *P. parva* and *P. capuana*, the sequenced lengths of the *P. piriformis* mtDNAs are approximately 500 nt smaller than what agarose gel electrophoresis results would suggest. We believe that the conformation of the mtDNA telomeres is retarding migration of these molecules in agarose gels (discussed below). In addition to DNA sequencing, the sizes of the *P. piriformis* mtDNAs were confirmed by restriction endonuclease digestion experiments (Figure 3.1). We were unable to PCR amplify circular forms of either the large or small fragments, nor could we recover products that linked the two molecules together.

Overall, the *P. piriformis* mitochondrial genome contains ten genes, nine of which are found on the large fragment (representing six proteins, one tRNA, and two rRNAs) and one on the small fragment (*nad6*). The genes on the large fragment are organized into two unequally sized clusters with opposing transcriptional polarities that proceed outwards toward the ends of the chromosome. The sole tRNA-coding gene, *trnM*, has characteristics that are consistent with a role in elongation rather than initiation, as is true for the *trnM* found in other sequenced *Reinhardtinia*-clade mtDNAs. The two ribosomal-RNA-coding genes (*rrnl* and *rrns*), like those of other chlorophycean mitochondrial genomes, are fragmented and scrambled (Figure 3.3A). The degree to which the *rrnl* and *rrns* regions are fragmented (four and eight coding modules, respectively) is the same as that of the *C. reinhardtii* and *V. carteri* mitochondrial genomes. The mtDNA gene complement and gene arrangement for *P. piriformis* are identical to those of *P. parva*

(which also has a bipartite mtDNA with *nad6* isolated on the smallest fragment) and *P. capuana*, except that the *P. capuana* mtDNA is contained in a single chromosome with *nad6* located internal to the left terminal repeat (Figure 3.3A). With only ten mtDNA-encoded genes, none of which harbour introns, these three *Polytomella* strains have the most reduced mitochondrial genomes observed from the Archaeplastida. The only species known to have even smaller mtDNA gene repertoires are found in the phyla Apicomplexa and Dinoflagellata, and arguably some species within the supergroup Excavata (Kairo et al. 1994; Nash et al. 2007).

Of the 16,083 nt that comprise the *P. piriformis* mitochondrial genome, 10,590 nt (66%) are coding and 5,493 nt (34%) are noncoding. The latter category can be subdivided into telomeres and intergenic spacers, which represent 5,233 nt (32.5%) and 260 nt (1.6%) of the genome, respectively. If the telomeres are ignored, the *P. piriformis* mtDNA is only 1.6% noncoding, which is inordinately compact for a green-algal mitochondrial genome (see Figure 7.1 [Chapter 7] for a schematic compilation). Relative to that of *P. piriformis*, the *P. parva* mtDNA has a comparable distribution of coding, intergenic and telomeric DNA, whereas the *P. capuana* mtDNA has more intergenic DNA (555 vs. 260 nt) and the telomeres constitute a smaller fraction of the genome (13.7% vs. 32.5%); this is because they are shorter than those of *P. piriformis* and *P. parva* (~0.9 kb vs. ~1.3 kb) and because there are only two of them rather than four as in the *P. piriformis* and *P. parva* mitochondrial genomes. Unlike the *P. capuana* mtDNA, where inverted repeats, which can folded into stem-loop structures, punctuate all but two of the genes, the *P. piriformis* mitochondrial genome contains no obvious inverted-repeat elements in its intergenic regions. Re-analyses of the *P. parva* mtDNA reveal the same conclusion. We cannot say if the lack of short inverted repeats in the *P. piriformis* and *P. parva* mtDNAs is due to an alternative mode of mitochondrial-transcript processing relative to *P. capuana*. However, RT-PCR analyses with either *P. piriformis*, *P. parva*, or *P. capuana* RNA can generate products that span more than one mitochondrial-encoded gene (up to at least five genes), suggesting that polycistronic mitochondrial transcripts are being generated in all three of these algae. We were also able to generate RT-PCR products that spanned most of the telomeric regions from these three *Polytomella* species.

The GC content of the *P. piriformis* mtDNA (42% overall; 41% and 43% for the large and small fragments, respectively) is unremarkable in comparison to that of *P. capuana*, which is one of the more GC-rich mtDNAs observed to date (57%), but comparable to that of *P. parva* (41% overall, 40% large fragment, and 42% small fragment). Among the different regions of the *P. piriformis* mitochondrial genome, the telomeres have a higher GC content (45%) than the coding segments (40%) and the intergenic spacers (37%). This trend is also observed for the *P. parva* mtDNA. The GC contents of the different mtDNA codon positions for *P. piriformis* are 45% (1st position), 39% (2nd position), 36% (3rd position), and 25% (3rd-position synonymous sites), which is similar to that of *P. parva* (44%$_{(GC1)}$, 39%$_{(GC2)}$, 35%$_{(GC3)}$, and 24%$_{(GCsyn)}$) but different than that of *P. capuana* (52%$_{(GC1)}$, 41%$_{(GC2)}$, 76%$_{(GC3)}$, and 85%$_{(GCsyn)}$). The fact that the nucleotide composition of the *P. piriformis* mtDNA is most biased at what are typically regarded as among the most neutrally evolving positions in a genome (intergenic and synonymous sites) and is least biased at the more functionally constrained first- and second-codon positions, suggests that a neutral process, such as mutation pressure or biased gene conversion, is driving the nucleotide composition towards A and T. Similarly, it has been argued that a neutral process is biasing the nucleotide content of the *P. capuana* mtDNA; however, unlike for the *P. piriformis* and *P. parva* mtDNAs, this process appears to be skewed towards G and C (Smith and Lee 2008a; Chapter 2).

Another notable feature of the *P. piriformis* mtDNA is that the *nad5* gene has an alternative start codon: it uses "GTG" instead of "ATG." The *nad5* start codon of *P. parva* is also "GTG" but that of *P. capuana* is the canonical "ATG." Our RT-PCR analyses did not reveal any indication that the *P. piriformis nad5* "GTG" start codon is converted to "AUG" via RNA editing. And although it remains to be confirmed, it is reasonable to assume that the *P. parva nad5* "GTG" start codon also remains unedited in RNA transcripts. These data support the generally held belief that there is no RNA editing in green-algal mitochondria (Nedelcu et al. 2000; Steinhauser 1999; Turmel et al. 2010). To the best of our knowledge, this is the first recorded example of a chlorophycean mitochondrial genome with an alternative start codon. There are, however, many examples of Metazoan mtDNAs that use "GTG" start codons

(http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=cgencodes).

P. piriformis *MtDNA Telomeres*

We obtained the *P. piriformis* mtDNA telomeric sequences using terminal TdT tailing (Förstemann et al. 2000; Bah et al. 2004) and long walk PCR (Katz et al. 2000). The former allowed us to amplify the 3′ ends of the telomeres whereas the latter approach amplified the 5′ telomeric ends. See Smith and Lee (2008a; Chapter 2) for a more detailed description of how these techniques can be used to amplify telomeric regions.

The *P. piriformis* mitochondrial telomeric sequences form an inverted repeat (i.e., the sequence of one terminus is the reverse complement of the other terminus). Inverted-repeat telomeres are a common feature among species with linear mtDNAs, and are found in *P. parva*, *P. capuana*, *C. reinhardtii*, and *Pandorina morum* as well as in certain apicomplexans, ciliates, cnidarians, and fungi (Table 3.1). Moreover, linear mitochondrial plasmids typically end in inverted repeats (Handa 2007). Each of the *P. piriformis* mtDNA telomeres is ~1.3 kb in length, which is about the same size as the *P. parva* telomeres and 0.4 kb longer than those of *P. capuana* (Table 3.1). Within each of the *P. piriformis* mtDNA fragments, the two telomeric sequences are identical, but the telomeres between fragments differ slightly (pairwise identity = 93.7%) — for *P. parva*, alignments of the mtDNA telomeres of the different fragments gives pairwise identity values of ~99.5%. We do not know why there is more sequence divergence between the telomeric sequences of the two *P. piriformis* mtDNA fragments as compared to those of *P. parva*. One explanation is that the mtDNA of *P. piriformis* has a higher mutation rate and/or a lower rate of intermolecular gene conversion relative to that of *P. parva*. Alignments of the *P. piriformis* telomeres with those of *P. parva* gave pairwise identity values of ~54% (Figure 3.4). These alignments reveal long stretches of unrelated sequences punctuated with short blocks of what appears to be conserved sequence elements (Figure 3.4). We were unable to align either the *P. piriformis* or *P. parva* mtDNA telomeres with those of *P. capuana* or *C. reinhardtii*.

TdT tailing and long walk PCR analyses suggest that the *P. piriformis* telomeres

exist *in vitro* in two separate conformations: a closed (hairpin loop) conformation and an open (nicked loop) conformation (Figure 3.1; Figure 3.3C). The nucleotide sequences of these two telomeric conformations appear to be identical to one another; however, that from the closed conformation seems to terminate with an ~150 nt single-stranded loop, whereas in the open conformation this loop is nicked. The location of the nick was shown to vary but was most often observed at the apex of the loop. Standard-PCR techniques confirmed the sequence of the telomeric regions. Agarose gel electrophoresis and restriction digest results provide further support for the idea that the *P. piriformis* mtDNA telomeres can have an open or closed conformation (Figure 3.1). These findings are consistent with data on the *P. capuana* mtDNA telomeres, which were shown to be either hairpin loop or nicked loop structures, but for *P. capuana* the loop portion of the hairpin is ~220 nt in length. Preliminary analyses (Smith DR and Lee RW, unpublished data) suggest that the mitochondrial telomeres of *P. parva* are also found in either open or closed conformations. It should be emphasized that at present we cannot rule out the possibility that the nicked loop telomeric conformation observed for the *P. piriformis* mtDNA is the result of nicking during the DNA extraction process. The telomeric conformations may help explain why the *P. piriformis* mtDNA fragments migrate as 13.5- and 3.5-kb molecules in agarose gel electrophoresis analyses instead of their sequenced sizes of 13 and 3 kb. One would expect a linear genome with a 150-nt single-stranded hairpin loop at each end (or with ~75 nt of noncomplementary DNA at each end — i.e., nicked loops) to migrate slower in an agarose gel than a linear DNA molecule of the same size but with blunt ends. This same argument applies to the *P. parva* and *P. capuana* mtDNAs, which also migrate in agarose gels at sizes larger than what DNA sequencing would suggest (Mallet and Lee 2006; Smith and Lee 2008a; Chapter 2).

Hairpin loop telomeres are not uncommon. In addition to *Polytomella* mtDNA, they are found on the mitochondrial genomes of *Pichia*, *Williopsis*, and *Paramecium* (Pritchard and Cummings 1981; Dinouël et al. 1993), the mitochondrial plasmid of the plant pathogenic fungus *Rhizoctonia solani* (Miyashita et al. 1990), the genomes of certain viruses (Baroudy et al. 1982; González et al. 1986; Rohozinski et al. 1989), and the bacterial plasmids of *Borrelia burgdorferi* (Hinnebusch and Barbour 1991). Their wide distribution suggests that hairpin loop telomeres may provide an effective strategy

for overcoming the end-replication problem.

All linear genomes must overcome the end-replication problem, as defined by Olovnikov (1971) and Watson (1972). Unlike eukaryotic nuclear genomes, which mostly employ telomerase to overcome this problem, it is generally accepted that linear mtDNAs do not use telomerase. A variety of telomerease-independent replication models for linear genomes with terminal hairpins have been suggested (Cavalier-Smith 1974; Bateman 1975; Pritchard and Cummings 1981; Baroudy et al. 1983; Dinouël et al. 1993; Traktman 1996). Many of these models center on the fact that a DNA molecule with closed ends is essentially a single-stranded circle, meaning that conventional DNA replication should be sufficient to amplify the entire molecule. Based on the nucleotide sequences of *P. piriformis*, no obvious mode of replication is apparent. However, the fact that we observe both closed and nicked loop terminal conformations is indicative of a Cavalier-Smith, Bateman-type model (1975), where conventional DNA replication proceeds through the terminal hairpins and generates a double-stranded circular molecule, which is then cut at the palindromic sequences by endonucleases followed by unpairing of the two strands and finally self-pairing and ligation of the ends.

Mitochondrial Genome Evolution of Reinhardtinia*-clade Algae*

Completion of the *P. piriformis* mitochondrial genome sequence allows for a more thorough phylogenetic evaluation of *Polytomella* species. Previously, Mallet and Lee (2006) performed a phylogenetic analysis using *cox1* sequence data from seven *Polytomella* strains, including *P. piriformis*, and showed that these strains form three distinct lineages, which are represented by *P. capuana*, *P. parva*, and *P. piriformis*, with the *P. parva* lineage containing five of the seven strains tested. Here, we ran a maximum-likelihood phylogeny using the concatenated DNA sequences of seven mtDNA-encoded proteins from the three *Polytomella* species plus six other chlorophycean algae for which complete or almost-complete mitochondrial-genome sequences are available. Our results support those of Mallet and Lee (2006): we find that *P. piriformis*, *P. parva*, and *P. capuana* represent separate distinct lineages within the "*Polytomella* clade," and that *P. capuana* is the deepest branching of the three *Polytomella* species (Figure 3.5). This analysis as well as that of Mallet and Lee (2006) reveals high rates of mtDNA evolution

for all three *Polytomella* species relative to other chlamydomonadalean algae (Figure 3.5). Previous phylogenic analyses (Pröschold et al. 2001; Nakada et al. 2008) show high rates of nucleotide substitution in the *P. parva* nuclear-encoded SSU rRNA gene relative to that of other chlamydomonadalean algae, therefore, suggesting that rapid evolutionary rate is a feature of the *Polytomella* lineage and not any one genetic compartment. Such a lineage feature could result from an enhanced number of mutations and/or an elevated level of mutation fixation in the *Polytomella* lineage relative to that of other chlamydomonadalean algae as discussed earlier for *P. parva* (Fan and Lee 2002).

P. piriformis is the seventh chlamydomonadalean alga (fourth from the *Reinhardtinia* clade) to have its mtDNA completely sequenced; moreover, almost complete mtDNA sequences are available for *V. carteri* and *C. incerta*, and mtDNA gel-electrophoresis results exist for a variety of other chlamydomonadalean species (Table 3.1). Altogether, these data allow for an intricate picture of the evolution of mitochondrial genome architecture in the Chlamydomonadales. We have summarized on a phylogenetic tree some of the diverse features of mitochondrial genome architecture in this group and proposed the events leading to their origin (Figure 3.6). It remains to be determined if this diversity in mitochondrial genome architecture in the Chlamydomonadales is greater than that of other comparable green algal groups or is the result of unequal sampling.

Table 3.1. Available Organelle-genome Data for Chlamydomonadalean Green Algae (With Some Examples of Non-green-algal Linear MtDNAs)

Genus and species	Clade[a] (or group)	Mitochondrial Genome			Mitochondrial-DNA Telomeres				Reference (Thesis chapter)
		Conformation	Size (kb)	%GC	Conformation of ends	Size (kb)	Telomere Type	%GC	
Chlamydomonadales[a]									
Chlamydomonas asymmetrica	*Reinhardtinia*	linear	~14	—	—	—	—	—	Laflamme & Lee 2003
Chlamydomonas incerta	*Reinhardtinia*	linear	~17.5[c]	~44[c]	—	—	—	—	Popescu & Lee 2007
Chlamydomonas reinhardtii	*Reinhardtinia*	linear	15.8⇔18.9[b]	~45	3′ extension	~0.6	TIR	~51	Michaelis et al. 1990, Smith & Lee 2008b (Chapter 5)
Chlamydomonas sphaeroides	*Reinhardtinia*	linear	~25f	—	—	—	—	—	Laflamme & Lee 2003
Pandorina morum	*Reinhardtinia*	linear	19⇔38	—	—	~1.8-3.3	TIR	—	Moore & Coleman 1989
Polytomella capuana	*Reinhardtinia*	linear	13.0	57.2	Hairpin loop or nicked loop	~0.9	TIR	58.7	Smith & Lee 2008a (Chapter 2)
Polytomella parva	*Reinhardtinia*	linear bipartite	~13 & 3	41.0[d]	Hairpin loop or nicked loop	~1.3	TIR	44.4	Fan & Lee 2002
Polytomella piriformis	*Reinhardtinia*	linear bipartite	~13 & 3	42.0[d]	Hairpin loop or nicked loop	~1.3	TIR	45.7	Smith et al. 2010a (Chapter 3)
Sphaerellopsis aulata	*Reinhardtinia*	linear	~25[f]	—	—	—	—	—	Laflamme & Lee 2003
Volvox carteri f. *nagariensis*	*Reinhardtinia*	circular[e]	~35[c]	~34[c]	N/A	N/A	N/A	N/A	Smith & Lee 2009b, 2010 (Chapter 7; Appendix E)
Chlamydomonas eugametos	*Moewusinia*	circular	22.9	34.6	N/A	N/A	N/A	N/A	Denovan-Wright et al. 1998
Chlamydomonas moewusii	*Moewusinia*	circular	~21	—	N/A	N/A	N/A	N/A	Boudreau et al. 1994
Chlamydomonas pitschmannii	*Moewusinia*	circular	~16.5	—	N/A	N/A	N/A	N/A	Boudreau & Turmel 1995
Lobochlamys culleus	*Oogamo-chlamydinia*	linear fragmented	—	~60	—	—	—	—	Borza et al. 2009
Lobochlamys segnis	*Oogamo-chlamydinia*	fragmented[g]	3⇔20	~55	—	—	—	—	Borza et al. 2009
Chlorogonium elongatum	*Chlorogonia*	circular	22.7	37.8	N/A	N/A	N/A	N/A	Kroymann & Zetsche 1998
Dunaliella salina	*Dunaliellinia*	circular	28.3	34.4	N/A	N/A	N/A	N/A	Smith et al. 2010b (Chapter 8)

Genus and species	Clade[a] (or group)	Mitochondrial Genome			Mitochondrial-DNA Telomeres				Reference
		Conformation	Size (kb)	%GC	Conformation of ends	Size (kb)	Telomere Type	%GC	
Other examples of linear mtDNAs									
Aurelia aurita	Cnidaria	linear	16.9	33.3	—	~0.4	TIR	~30	Shao et al. 2006
Hydra magnipapillata	Cnidaria	linear bipartite	8.2 & 7.7	23.7[d]	—	~0.2	TIR	~52	Voigt et al. 2008
Hydra oligactis	Cnidaria	linear	16.3	23.7	—	~1.5	TIR	~35	Kayal & Lavrov 2008
Candida parapsilosis	Yeast	linear	32.7	23.8	5′ extension	~1.4	TIR (TR)	15.2	Nosek et al. 2004b
Ochromonas danica	Chrysophyta	linear	41.0	~26	N/A	N/A	N/A	N/A	Coleman et al. 1991
Tetrahymena pyriformis	Ciliata	linear	42.1	~21	—	2.7	TIR (TR)	~27	Burger et al. 2000
Theileria parva	Apicomplexa	linear	6.6	21.3	—	~0.45	TIR	~29	Shukla & Nene 1998

Note: N/A, not applicable; (—), data not available; TIR, terminal inverted repeat; TIR (TR), terminal inverted repeat containing internal tandem repeats.

[a] Our definition of Chlamydomonadales follows that of Lewis and McCourt (2004); clades are defined by Nakada et al. (2008).

[b] These data vary because of the presence/absence of optional introns (Smith and Lee 2008b; Chapter 5).

[c] These data are based on almost-complete genome sequences.

[d] MtDNA consists of two fragments; data are based on the concatenation of these fragments.

[e] The circular conformation of the *V. carteri* mtDNA is based on genome-assembly data (Smith and Lee 2010; Chapter 7) and needs to be confirmed by gel electrophoresis results.

[f] MtDNAs were at the limit of the linear size range for the gels.

[g] It is unknown, at present, if these fragments have circular or linear conformations (Borza et al. 2009).

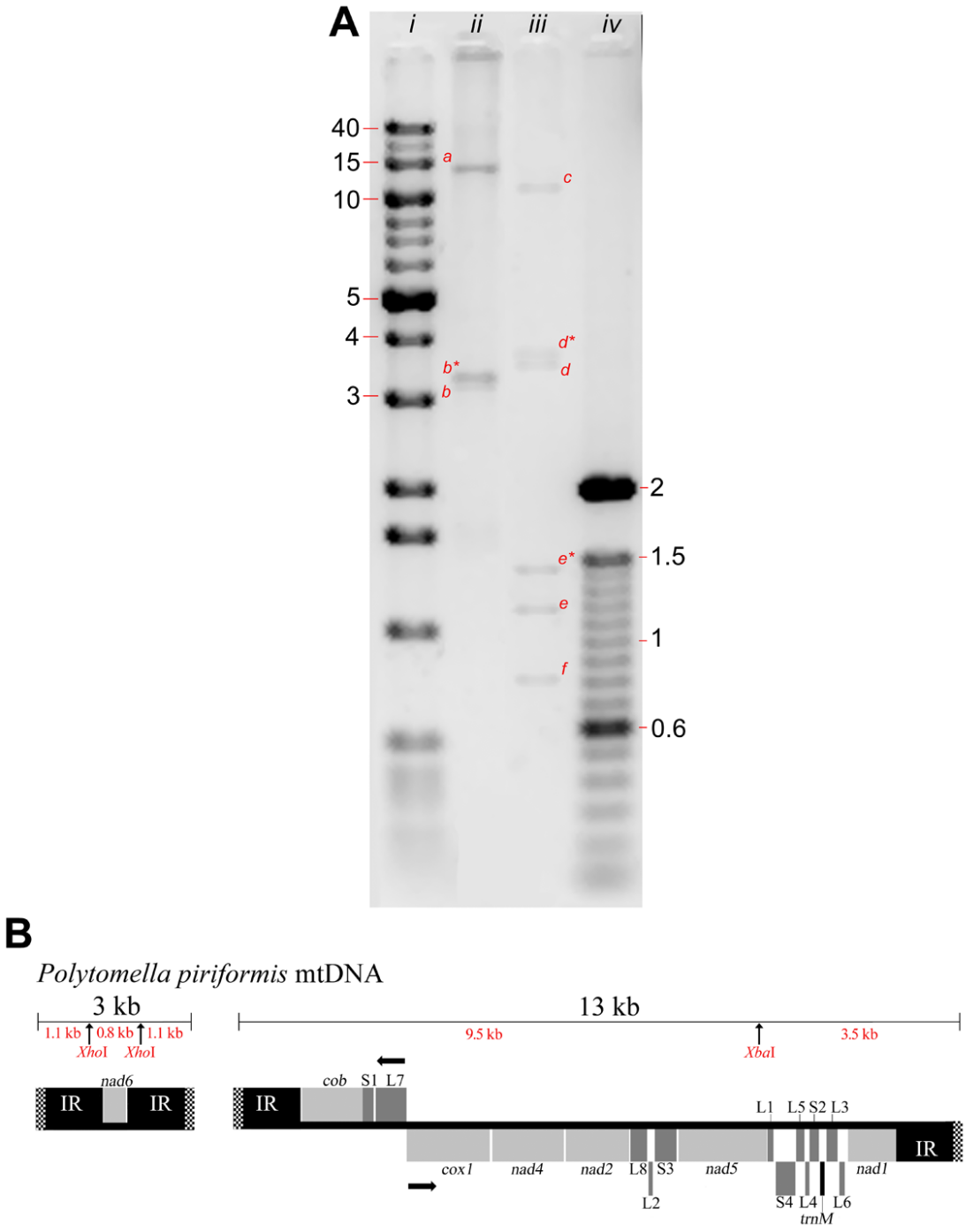

Figure 3.1. Gel Electrophoresis Analyses of *P. piriformis* MtDNA

(A) Agarose (0.8%) gel electrophoresis analyses of *P. piriformis* mtDNA. Lanes are as follows: i) Invitrogen 1 kb Extension Ladder (2 µg); ii) *P. piriformis* DNA isolated from a mitochondria-enriched fraction; iii) restriction digest of *P. piriformis* mtDNA with *Xho*I

and *Xba*I; iv) Invitrogen 100 bp Ladder (2 µg). Size of DNA markers are shown in kilobases. Bands marked on gel are as follows: *a*, *P. piriformis* mtDNA 13 kb piece; *b*, *P. piriformis* mtDNA 3 kb piece; *b**, *P. piriformis* mtDNA 3 kb piece shadow band; *c*, 9.5 kb *Xba*I fragment; *d*, 3.5 kb *Xba*I fragment; *d**, 3.5 kb *Xba*I fragment shadow band; *e*, 1.1 kb *Xho*I fragment; *e**, 1.1 kb *Xho*I fragment shadow band; *f*, 0.8 kb *Xho*I fragment. (B) Physical and genetic map of the *P. piriformis* mtDNA. *Xho*I and *Xba*I restriction sites as well as their fragment sizes are shown in red. Checkered regions on the terminal inverted repeats (IR) represent portions of the telomeres that can either have a closed (hairpin-loop) conformation or an open (nicked-loop) conformation (Dinouël et al. 1993; Smith and Lee 2008a; Chapter 2). Note, when the *P. piriformis* mtDNA is digested with *Xba*I and *Xho*I and run on an agarose gel each of the terminal fragments as defined by the restriction map appear as two bands: a main band (with a closed telomeric conformation), which migrates as expected and a shadow band (with an open telomeric conformation), which migrates more slowly.

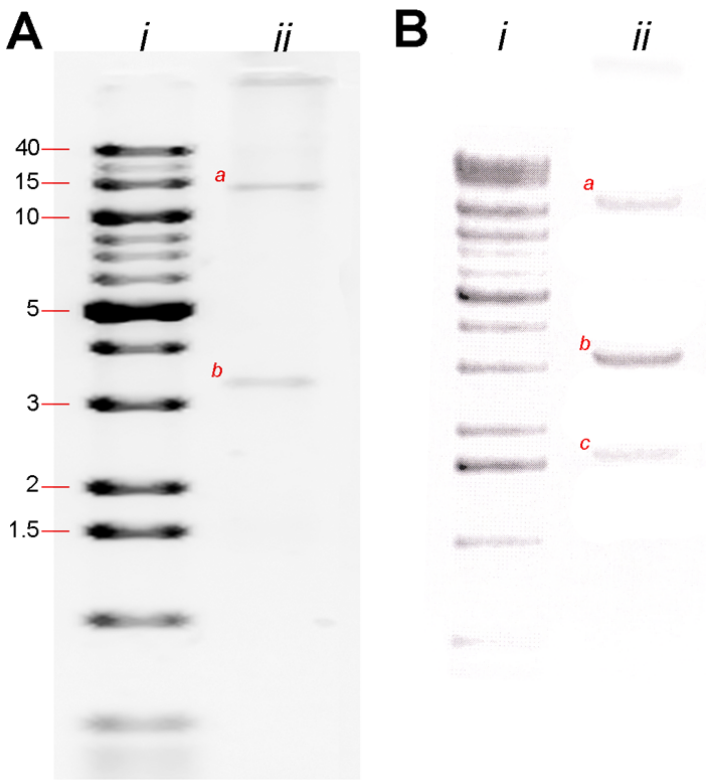

Figure 3.2. Southern Blot Analyses of *P. piriformis* MtDNA

(A) Agarose (0.8%) gel electrophoresis analysis of *P. piriformis* mtDNA. (B) Southern-blot analysis of *P. piriformis* mtDNA using a 600 nt portion of the mitochondrial telomeres as a probe. Invitrogen 1 kb DNA Extension Ladder (2 µg) (lane i); *P. piriformis* DNA isolated from a mitochondria-enriched fraction (lane ii). Sizes of DNA markers are shown in kilobases. Bands marked on gel and Southern blot are as follows: *a*, *P. piriformis* mtDNA 13 kb piece; *b*, *P. piriformis* mtDNA 3 kb piece; *c*, *P. piriformis* mtDNA 1.8 kb piece.

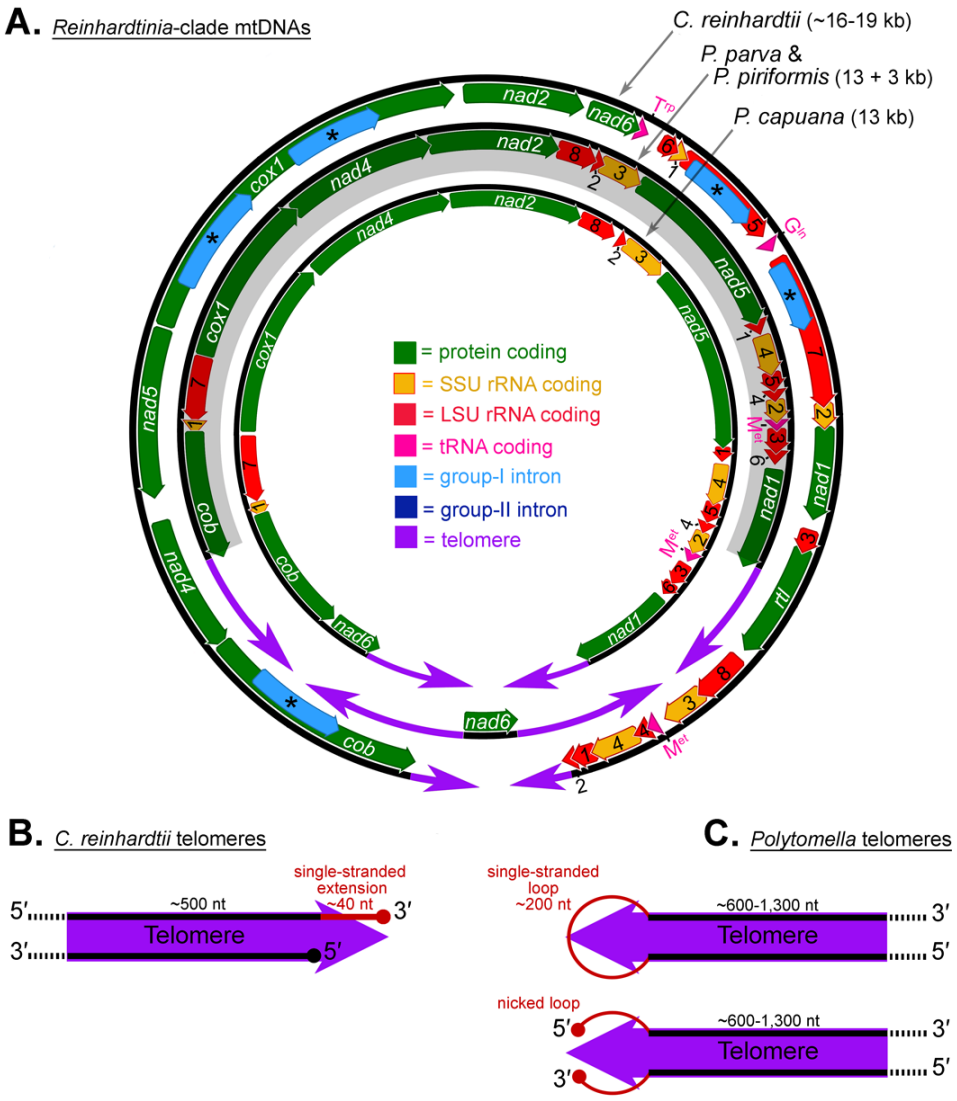

Figure 3.3. Genome and Telomere Maps of *Polytomella* and *C. reinhardtii* MtDNAs

(A) Arrows within the coding regions indicate transcriptional polarities. The small subunit and large subunit rRNA-coding modules are numbered based on their order in the gene. Portions of the *P. parva* and *P. piriformis* mitochondrial genome maps that are shaded gray represent regions of gene colinearity with the *P. capuana* mtDNA. Introns annotated with an asterisk are optional. (B) and (C) *In vitro* mtDNA telomere conformation for *C. reinhardtii* and *Polytomella* taxa. Sequencing and agarose gel electrophoresis analyses suggest that the *Polytomella* telomeres can exist in either a closed (hairpin loop) or an open (nicked loop) conformation (see Figure 3.1).

Figure 3.4. Multiple Alignment of the *P. piriformis* and *P. parva* MtDNA Telomeric Sequences

P. pir, *P. piriformis*; P. par, *P. parva*; L and R, mtDNA sequences from the left and right telomeres (as defined by the genetic map in Figure 3.1B), respectively. Percent-similarity values for shaded nucleotides are as follows: green, 100%; blue, 80-99%; yellow, 60-79%. Note, in the alignment position 1 corresponds to the extreme 5′ end of the telomeres. [Figure continues on next page.]


P. pir 3 kb L
P. pir 3 kb R
P. pir 13 kb L
P. pir 13 kb R
P. par 3 kb L
P. par 3 kb R
P. par 13 kb L
P. par 13 kb R
nad6 N-terminus
nad6 C-terminus
cob C-terminus
nad1 N-terminus
nad6 N-terminus
nad6 C-terminus
cob C-terminus
nad1 N-terminus

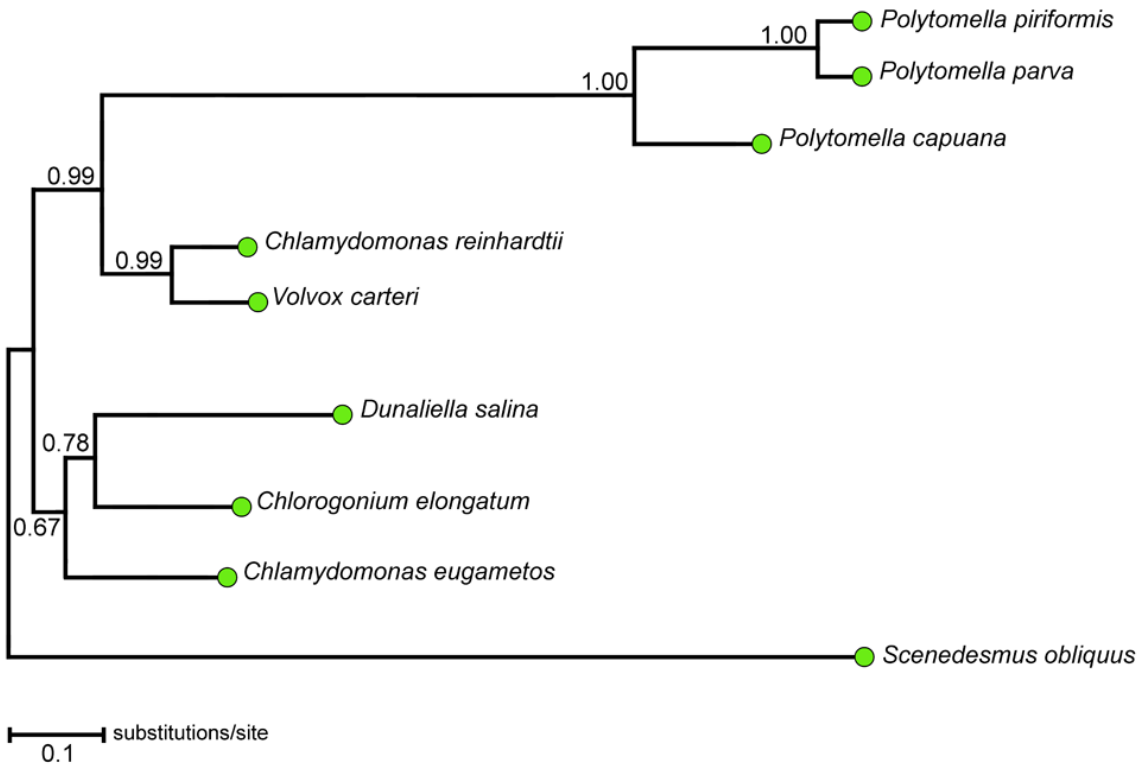

Figure 3.5. Maximum Likelihood Tree Inferred From the DNA Sequences of the Seven MtDNA-encoded Proteins that are Shared Among All Completely Sequenced Chlorophycean MtDNAs

GenBank accession numbers are as follows: *P. piriformis* (GU108480, GU108481), *P. capuana* (EF645804), *P. parva* (AY062933, AY062934), *C. reinhardtii* (EU306622), *V. carteri* (GU084821), *D. salina* (GQ250045), *C. eugametos* (AF008237), *C. elongatum* (Y13643, Y13644, Y07814), and *S. obliquus* (X17375). All taxa are members of the Chlorophyceae. The node support values were assessed by bootstrap resampling calculated using 500 replicates.

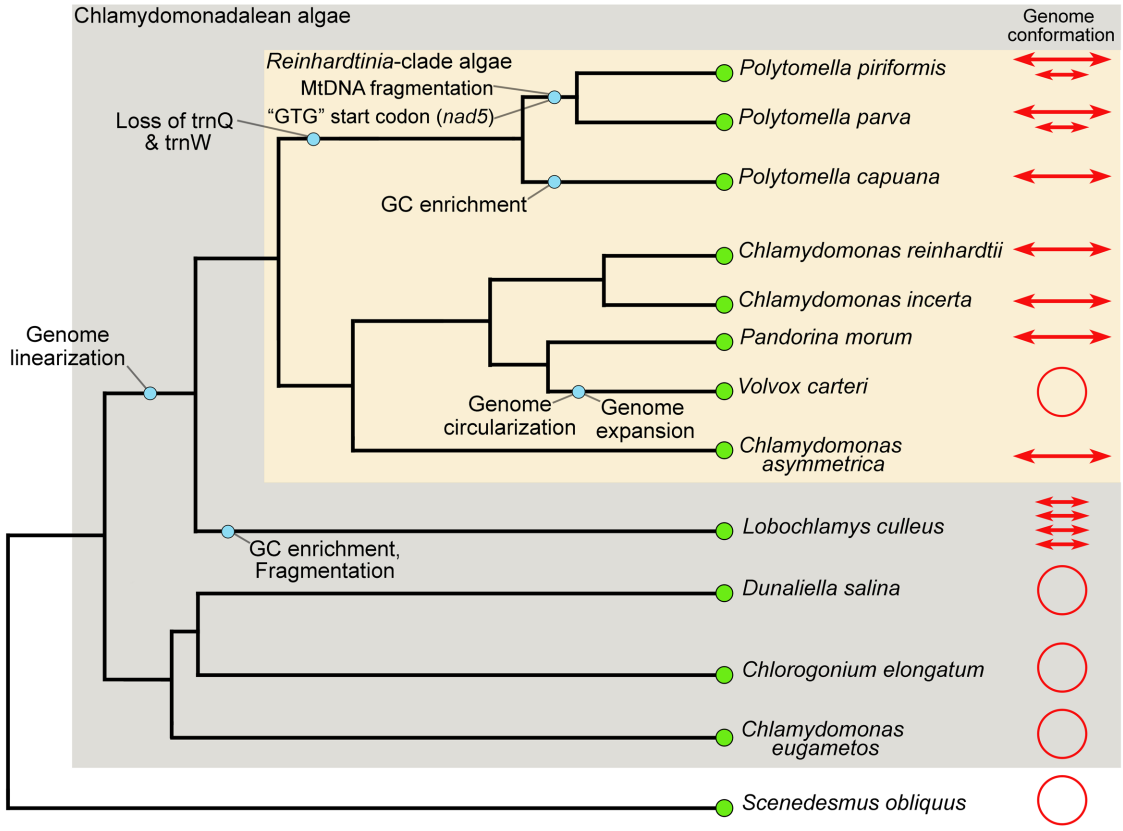

Figure 3.6. Hypotheses on the Evolution of Chlamydomonadalean Mitochondrial Genome Architecture

Phylogenetic relationships are based on the *18S*-rDNA gene tree of Nakada et al. (2008). Blue circles represent hypotheses on chlamydomonadalean mitochondrial genome evolution. Refer to Table 3.1 for a complete list of references and GenBank accession numbers.

CHAPTER 4: UNPARALLELED GC CONTENT IN THE PLASTID DNA OF *SELAGINELLA*

Published as:

Smith DR (2009) Unparalleled GC content in the plastid DNA of *Selaginella*. Plant Mol Biol 71:627-639.

Abstract

One of the more conspicuous features of plastid DNA (ptDNA) is its low guanine and cytosine (GC) content. As of February 2009, all completely sequenced plastid genomes have a GC content below 43% except for the ptDNA of the lycophyte *Selaginella uncinata*, which is 55% GC. The forces driving the *S. uncinata* ptDNA towards G and C are undetermined, and it is unknown if other *Selaginella* species have GC-biased plastid genomes. This study presents the complete ptDNA sequence of *Selaginella moellendorffii* and compares it with the previously reported *S. uncinata* plastid genome. Partial ptDNA sequences from 103 different *Selaginella* species are also described as well as a significant proportion of the *S. moellendorffii* mitochondrial genome. Moreover, *S. moellendorffii* expressed sequence tags are data-mined to estimate levels of plastid and mitochondrial RNA editing. Overall, these data are used to show that: i) there is a genus-wide GC bias in *Selaginella* ptDNA, which is most pronounced in South American articulate species; ii) within the Lycopsida class (and among plants in general), GC-biased ptDNA is restricted to the *Selaginella* genus; iii) the cause of this GC bias is arguably a combination of reduced AT-mutation pressure relative to other plastid genomes and a large number of C-to-U RNA editing sites; and iv) the mitochondrial DNA (mtDNA) of *S. moellendorffii* is also GC biased (even more so than the ptDNA) and is arguably the most GC-rich organelle genome observed to date — the high GC content of the mtDNA also appears to be influenced by RNA editing. Ultimately, these findings provide convincing support for the earlier proposed theory that the GC content of land plant organelle DNA is positively correlated and directly connected to levels of organelle RNA editing.

Introduction

A prominent feature of plastid DNA (ptDNA) is its low guanine and cytosine (GC) content. Indeed, all of the 150 completely sequenced plastid genomes available at the National Center for Biotechnology Information (NCBI) as of February 2009 have a GC content between 19.5% and 42.1% (average = 36.2%; SD = 4.6%) with the exception of the *Selaginella uncinata* ptDNA, which is 54.8% GC — a complete compilation is shown in Appendix B. The evolutionary forces shaping ptDNA nucleotide landscape are

unknown; however, several hypotheses have been proposed. For instance, some argue that a neutral process such as AT-mutation pressure or AT-biased gene conversion caused the low GC content of ptDNA (Howe et al. 2003; Kusumi and Tachida 2005; Khakhlova and Bock 2006). Others invoke selection for translational efficiency to explain the lack of G and C observed in plastid genomes (Morton 1993; Morton 1998). There is also the possibility that plastid genomes descend from an AT-rich bacterial genome, but it is generally thought that ptDNA has become GC poor since endosymbiosis (Howe 2003). Interestingly, convergent evolution towards a reduced GC content is seen in mitochondrial DNA (mtDNA), nucleomorph DNA, and in the genomes of symbionts, parasites, and pathogenic bacteria (Dybvig and Voelker 1996; Ogata et al. 2001; Lane et al. 2007; Smith and Lee 2008a; Chapter 2); if the forces biasing these genomes against G and C are similar to those acting on ptDNA, then understanding nucleotide composition from a plastidial framework could have wide-reaching implications.

The plastid genome of the lycophyte *S. uncinata* is exceptional in that it has a GC content above 50% (Tsuji et al. 2007). In addition to being GC biased, the *S. uncinata* ptDNA has several other distinguishing characteristics: i) it encodes only 12 distinct tRNAs, which is currently one of the most reduced tRNA-coding repertoires of any completely sequenced plastid genome (land plant plastid genomes typically contain more than 30 tRNA-coding genes); ii) it contains a unique ptDNA gene order, unlike *Huperzia lucidula* (the only other lycophyte with a completely sequenced ptDNA), which has a ptDNA gene arrangement similar to bryophytes (Wolf et al. 2005); and iii) it experiences extremely high levels of RNA editing, potentially higher than any other ptDNA sequence examined to date. It is predicted that RNA editing in *S. uncinata* restores the 79 non-standard start and stop codons found in the ptDNA protein-coding regions to their canonical state (Tsuji et al. 2007).

Although the ptDNA of *S. uncinata* has been described in detail, the plastid genomes from other *Selaginella* species remain unexplored and it is unknown if GC-biased ptDNA is a trait common to all members of the *Selaginella* genus or if it is restricted to only *S. uncinata*. One of the only reported cases of a GC-rich mitochondrial genome is that of the green algae *Polytomella capuana* (Smith and Lee 2008a; Chapter 2). Interestingly, the mtDNA from other *Polytomella* species are AT rich (Mallet and Lee

2006; Smith et al. 2010a; Chapter 3); thus, it would be fascinating to see if the same trend is apparent for *Selaginella* ptDNA. It would also be intriguing to explore the mitochondrial and nuclear genomes from *Selaginella* taxa — if they too are GC rich then their sequences may help pinpoint the processes that are biasing *Selaginella* nucleotide composition.

The *Selaginella* genus belongs to the class Lycopsida, the members of which are called lycophytes. Fossil records and phylogenetic analyses indicate that lycophytes are an ancient, monophyletic group of vascular plants (~400-million years old [Kenrick and Crane 1997]) comprised of three known families: the Isoetaceae (quillworts), the Lycopodiaceae (club mosses), and the Selaginellaceae. *Selaginella*, the only recognized genus within the Selaginellaceae, is an eclectic genus, containing around 700 species, which are spread throughout the world and cover an impressive range of habitats, including desert, tropical-rain-forest, alpine, and arctic habitats (Mabberley 1997). Significant efforts have been directed at resolving the evolutionary relationships among *Selaginella* species (Korall et al. 1999). Phylogenetic analyses using the plastid-encoded *rbcL* gene and nuclear-encoded rRNA genes suggest that the ancestries of *Selaginella* species are complex, with many subgroups existing (Korall and Kenrick 2002, 2004); these analyses have also shown that *rbcL* substitution rates among *Selaginella* species are high relative to those observed within other land plant families (Korall and Kenrick 2002); though, to a less dramatic extent when compared to other spore-producing vascular plants (Pryer et al. 2004).

One *Selaginella* species that has recently gained widespread attention is *Selaginella moellendorffii*; this is because in 2007 its nuclear genome was completely sequenced by the United States Department of Energy Joint Genome Institute (DOE JGI). *S. moellendorffii* was chosen as a candidate for sequencing because its nuclear genome is especially small (~110 mega bases [Mb]) relative to the nuclear DNA (nucDNA) of other land plants, and also because lycophytes represents an important evolutionary link between vascular plants and the nonvascular mosses, liverworts, and hornworts (Banks 2009). Now that the *S. moellendorffii* nuclear genome is sequenced, *Selaginella* is emerging as a model genus for comparative plant genomics.

The study presented here takes advantage of publicly-available DNA and RNA sequence data to investigate the evolution of nucleotide landscape in the plastid and mitochondrial genomes of *S. moellendorffii* and other *Selaginella* species. It is concluded that there is a genus-wide GC bias in *Selaginella* ptDNA and potentially one in the mtDNA as well, and that both of these nucleotide biases are affiliated with high levels of RNA editing. The overall implications of this nucleotide bias are discussed.

Materials and Methods

Assembly and Verification of the S. moellendorffii *Organelle Genome Sequences*

The complete plastid-genome sequence of *S. moellendorffii* was generated by collecting and assembling ptDNA trace files produced by the DOE JGI *S. moellendorffii* nuclear-genome sequencing project (http://genome.jgi-psf.org/Selmo1/Selmo1.home.html). Trace files were data-mined from the NCBI Trace Archive (http://www.ncbi.nlm.nih.gov/blast/mmtrace.shtml) using the *S. uncinata* ptDNA sequence as a BLAST (blastn 2.2.21+) query — similar approaches for assembling organelle genomes have been used in previous studies (e.g., Smith and Lee 2008b; Smith and Lee 2009a; Chapters 5 and 6). The BLAST parameters were as follows: an expectation value (E-value) of 10; a word size of 11; match and mismatch scores of 2 and -3, respectively; and gap-cost values of 5 (existence) and 2 (extension). Trace files showing >90% sequence identity to the *S. uncinata* ptDNA in BLAST alignments were downloaded and assembled using CodonCode Aligner Version 2.0.6 (CodonCode Corporation, Dedham, MA, USA), which employs the Phred, Cross-match, and Phrap algorithms for base calling, sequence comparison, and sequence assembly, respectively. Assemblies were performed with a minimum-percent-identity score of 98, a minimum-overlap length of 500 nucleotides (nt), a match score of 1, a mismatch penalty of -2, a gap penalty of -2, and an additional first-gap penalty of -3. Assembly of the *S. moellendorffii* ptDNA trace files ultimately gave a complete plastid-genome sequence with >50-fold coverage.

To verify that no nuclear-genome-located ptDNA-like sequences (NUPTs) were collected, the entire *S. moellendorffii* nucDNA sequence was scanned for regions that show similarity to ptDNA. This was performed by blasting (blastn version 2.2.21+) the *S.*

moellendorffii ptDNA sequence against the *S. moellendorffii* nuclear genome sequence (v1.0) using the same parameters that are listed above. Only the first 150 scaffolds of the nuclear genome assembly were analyzed: approximately 93.5% of the diploid nuclear genome is contained in these 150 scaffolds and their cumulative length is 198.93 Mb. PtDNA sequences that mapped to the nucDNA with >80% sequence identity and at least 30 nt of aligned length (in BLAST alignments) were counted as hits.

The same general approach as that described for the ptDNA was used to collect, assemble, and validate the 56 kilobases (kb) of *S. moellendorffii* mtDNA sequence data presented in this study. The *Physcomitrella patens* and *Marchantia polymorpha* mitochondrial genomes (GenBank accession numbers NC_007945 and NC_001660, respectively) were used as BLAST queries to data-mine *S. moellendorffii* mtDNA trace files from the NCBI Trace Archive.

Scanning the S. moellendorffii *Nuclear Genome for Plastid-targeted Sequences*

The *S. moellendorffii* nuclear genome was scanned for plastid-targeted sequences by constructing a custom BLAST databank of the first 150 nucDNA scaffolds and then blasting this databank with ptDNA queries using an E-value of 5, a word size of 7, a match score of 2, a mismatch penalty of -3, a gap open score of 5, and a extend value of 2. All of the queries came from the *H. lucidula* ptDNA — specifically, the pool of genes that are located in the *H. lucidula* ptDNA but absent from the *S. moellendorffii* plastid genome. The TargetP server was employed for the prediction of plastid transit peptide sequences (Emanuelsson et al. 2007).

S. moellendorffii *Expressed Sequence Tags*

Expressed sequence tag (EST) data for *S. moellendorffii* were obtained from the DOE JGI *S. moellendorffii* Genome Portal (v1.0) (http://genome.jgi-psf.org/Selmo1/Selmo1.download.html) on 1 January 2009. Plastid and mitochondrial RNA-derived ESTs were collected by blasting (employing the same BLAST parameters that were used for finding plastid-targeted sequences) this EST databank using *S. moellendorffii* ptDNA and mtDNA sequences as queries. All hits were subsequently checked against the *S. moellendorffii* nucDNA sequence to insure that they were not

derived from nuclear-genome-located ptDNA-like or mtDNA-like sequences (NUPTs or NUMTs). The *S. moellendorffii* ESTs that map to the plastid and mitochondrial genomes are shown in Table 4.1.

RbcL *Sequence Data*

The *Selaginella rbcL* sequences employed in this study come from either Korall and Kenrick (2002, 2004) or are unpublished data deposited in GenBank. A list of the *Selaginella* species from which *rbcL* sequences were data-mined (including GenBank accession numbers) is described in Appendix C — note, the GC content of these sequences has neither been presented nor discussed elsewhere.

The other non-*Selaginella rbcL* sequences described in this study were collected by downloading from the NCBI nucleotide sequence repository all of the entries that have an *rbcL* annotation and any of the following taxonomic identifications: Charophyceae, Marchantiophyta, Bryophyta, Lycopodiophyta, Moniliformopses, Coniferophyta, Cycadophyta, Ginkgophyta, Gnetophyta, and Magnoliophyta. Partial *rbcL* sequences were accepted as long as they were >900 nt in length.

XLSTAT-Pro, an add-in software package for Microsoft Excel, was employed for all statistical analyses of the *rbcL* dataset, including Tukey's Honestly Significant Difference (HSD) test.

Nucleotide Composition Analyses

Nucleotide-composition analyses, including the GC content of first-, second-, and third-position codons sites, were determined with DAMBE (Xia and Xie 2001). The GC content of fourfold-degenerate sites (i.e., synonymous sites) was calculated with INCA (Supek and Vlahovicek 2004) by measuring the proportion of G or C at third-position codon sites that can tolerate any of the four nucleotides without altering the amino acid specified.

Nucleotide Sequence Accession Numbers

The GenBank accession numbers of the *S. moellendorffii* organelle genome sequence data described in this study are FJ755183 (ptDNA) and GQ246802-GQ246808 (mtDNA).

Results and Discussion

General Features of the S. moellendorffii *Plastid Genome*

The entire ptDNA sequence of *S. moellendorffii* was produced by data mining and assembling publicly-available sequences generated by the DOE JGI *S. moellendorffii* nuclear genome sequencing project. To ensure that no nuclear-genome-located ptDNA-like sequences (NUPTs) were collected, the complete *S. moellendorffii* nuclear genome was analyzed for regions that show similarity to ptDNA. The results of this analysis, described in Table 4.2, demonstrate that there are very few ptDNA-like sequences embedded in the nuclear genome: ~21.5 kb distributed over 307 sites in the diploid nucDNA sequence (this is at the lower end of what is observed for other land plants [Richly and Leister 2004]). These findings are a strong indication that the sequences used to assemble the *S. moellendorffii* plastid genome are derived from ptDNA and are not NUPTs.

The *S. moellendorffii* plastid genome is 143.8 kb in length and assembles as a circular molecule (Figure 4.1). Fifty-four percent (78 kb) of the genome codes for proteins and structural RNAs; the remaining 45.8% (65.8 kb) represents noncoding DNA, which can be subdivided into intergenic regions (57.8 kb) and introns (8 kb). A pair of inverted repeats, each with a length of 12.1 kb, divide the genome into a large- (83.7 kb) and a small-single-copy region (35.9 kb), referred to as the LSC and SSC regions. These statistics are similar to those of the *S. uncinata* ptDNA, with the exception that the *S. uncinata* plastid genome is 390 nt longer and its LSC and SSC regions have lengths of 77.7 kb and 40.9 kb, respectively — these size discrepancies are primarily due to the fact that the *S. uncinata* ptDNA harbours four pseudogenes and four gene duplicates that are absent from the *S. moellendorffii* ptDNA, and also because three genes in the SSC region of the *S. uncinata* ptDNA are found in the LSC region of the *S. moellendorffii* ptDNA

(see Figure 4.1 for details). The only other complete ptDNA sequence from a lycophyte, that of the club moss *H. lucidula* (Wolf et al. 2005), is ~10 kb longer than its *Selaginella* counterparts (because of a larger gene repertoire) and has a significantly larger LSC region (104.1 kb) and a much smaller SSC region (19.5 kb).

Annotation of the *S. moellendorffii* ptDNA sequence revealed 99 genes, 7 of which are duplicates found in the inverted repeats (Figure 4.1); when ignoring these duplicates, there are 75 protein-, 4 rRNA-, and 13 tRNA-coding genes (including tRNA *fMet*), which is among the most reduced ptDNA gene contents from any photosynthetic land plant examined to date. Pseudogenes of *accD*, *rpl33*, and *infa* were identified; the presumed functional copies of these loci were discovered in the nucDNA (see Methods for details). Eleven group-II introns, all within protein-coding genes, were also discerned from the ptDNA sequence (Figure 4.1). The ptDNA gene complement of *S. moellendorffii*, including introns and pseudogenes, mirrors that of *S. uncinata*, with some exceptions: i) the *S. uncinata* ptDNA contains duplicate copies of *psbK*, *trnQ*, *rpl23*, and the 5′-end of *rpl2,* whereas in the *S. moellendorffii* plastid genome these genes are present only once; ii) the *S. uncinata* ptDNA harbours pseudogenes for *chlL*, *psaM*, *rps12*, and *rpl21* (the latter three loci exist in the ptDNA only as pseudogenes), whereas the *S. moellendorffii* plastid genome contains only a functional *chlL* and has neither functional nor pseudogene copies of *psaM, rps12 or rpl21*. A scan of the *S. moellendorffii* nucDNA did not expose functional copies of these loci. They most likely exist in the nucDNA but were not uncovered because of their small size and relatively nonconserved sequence; and iii) the *S. moellendorffii* ptDNA encodes *trnL*, a gene that is absent from the *S. uncinata* ptDNA. Compared to the *H. lucidula* ptDNA, the *S. moellendorffii* plastid genome has 16 fewer tRNA-coding genes and 10 fewer protein-coding genes.

The relatively reduced ptDNA gene repertoires of *S. moellendorffii* and *S. uncinata* are reflections of the surprisingly small number of tRNAs encoded in these genomes (13 and 12, respectively, not including duplicates); their nearest rivals, in this respect, are the plastid genomes of the alveolates *Babesia bovis* and *Theileria parva*, which each encode 24 tRNAs, and the ptDNA of the parasitic angiosperm *Epifagus virginiana*, which encodes 23 tRNAs. It is unknown how *S. moellendorffii* and *S.*

uncinata compensate for the tRNA-coding genes that appear to be absent from their plastid genomes. One hypothesis is that they are encoded in the nuclear genome and imported to the plastid from the cytosol — a similar process is known to occur for plant mitochondria (Glover et al. 2001). A scan of the *S. moellendorffii* nucDNA for the missing plastidial tRNAs (using plastid-encoded tRNAs from a close relative as search queries) revealed only one putative plastid-bound tRNA: *trnP*-CGG. An alternative hypothesis is that the missing tRNAs are imported to the plastid from the mitochondria — a process also proposed for *E. virginiana* (Modern et al. 1991). However, analysis of a 56 kb portion of the *S. moellendorffii* mitochondrial genome uncovered no tRNA-coding genes, suggesting that the mtDNA of *S. moellendorffii,* like those from other land plants, has a reduced tRNA-coding suite. A final hypothesis is that novel tRNAs are generated from those encoded in the ptDNA through RNA editing, a topic discussed in more detail below.

The ptDNA gene order for *S. moellendorffii* is similar to that of *S. uncinata*, with one significant difference: the *S. moellendorffii* plastid genome lacks a 20 kb inversion (from *trnC* to *psbI*) found in the *S. uncinata* ptDNA. This inversion, which is also absent from the *H. lucidula* ptDNA and available plastid genome sequences from bryophytes, is commonly found in the ptDNA of higher ferns and seed plants (Palmer and Stein 1986; Raubeson and Jansen 1992). In addition, *rpl23* and the 5′ end of *rpl2*, which are a part of the inverted repeat in the *S. uncinata* ptDNA, are in the LSC region of the *S. moellendorffii* plastid genome; and the position of one protein-coding and four tRNA-coding genes in the *S. moellendorffii* ptDNA (*petN, trnD*, *trnE*, *trnF,* and *trnY*) differ from that in the *S. uncinata* ptDNA. In a general sense, the *S. moellendorffii* ptDNA gene order is intermediary to that of *S. uncinata* and *H. lucidula*, and shares more similarities with bryophyte ptDNA than with those of other vascular plants. The discrepancies in gene order and gene content between the *S. moellendorffii* and *S. uncinata* plastid genomes are outlined with red blocks, arrows, and symbols on Figure 4.1.

Nucleotide Landscape of the S. moellendorffii *Plastid Genome*

The overall GC content of the *S. moellendorffii* ptDNA is 51%, which is less than that of *S. uncinata* (54.8%) but still the second most GC-rich plastid genome observed to

date. A schematic compilation comparing the ptDNA GC content of *S. moellendorffii* to that of completely sequenced plastid genomes is shown in Figure 4.2. From this plot it is apparent that the nucleotide composition of ptDNA forms a continuum from approximately 20% to 40% GC, to the exclusion of the *S. moellendorffii* and *S. uncinata* plastid genomes, which are positioned outside of this continuum, well above all other available ptDNA sequences in terms of GC content. Note that the lycophyte *H. lucidula* has a more typical ptDNA GC content of 36.2% (Figure 4.2).

Among the different portions of the *S. moellendorffii* plastid genome, RNA-coding regions have the highest GC content (57.8%), followed by protein-coding regions (55.5%), introns (50.3%), and intergenic spacers (49.9%). The inverted repeats are more GC-rich (55.7%) than the SSC region (50.5%) and the LSC region (49.9%). The allocation of G versus C (GC skew) on the main sense strand (the strand depicted in Figure 4.1) is negligible with a value of only 0.0003. These trends parallel that of the *S. uncinata* ptDNA. It is noteworthy that the RNA-coding regions from other plastid genomes also tend to be GC biased, having an average GC content of 52.9% (SD = 4.9%) among completely sequenced ptDNAs; however, the intergenic-spacer, intronic and protein-coding regions from plastid genomes are generally skewed towards A and T.

Within the protein-coding ptDNA of *S. moellendorffii*, the average GC content of first-position codon sites (55.9%) exceeds that of second- (50.8%) and third-positions (44.8%). A comparison of these data with those from *S. uncinata* and other available plastid genome sequences is presented in Figure 4.3 (the raw data from which this figure was derived are shown in Appendix B). It is evident from this figure that the overall GC content of the *S. moellendorffii* (and *S. uncinata*) protein-coding regions is the result of a relatively inflated GC content at all three codon-site positions; although, among codon sites, third-position synonymous sites in the *S. moellendorffii* and *S. uncinata* ptDNAs (46.5% and 51% GC, respectively) depart most significantly in GC content from those of other available plastid genome sequences, which on average are 25.3% GC (SD 7.7%). GC-rich codons (those that code for the amino acids alanine, glycine, proline, and in some cases arginine) represent 30% of the codons found in the *S. moellendorffii* plastid genome. The proportion of GC-rich codons in the *S. uncinata* ptDNA is even greater at

34%, whereas the GC-rich codon composition from other completely sequenced ptDNAs is on average only 17%.

The PtDNA GC Content of Other Selaginella *Species*

The observation of GC-rich ptDNA in *S. moellendorffii* and *S. uncinata* raises questions regarding the phylogenetic distribution of GC content within the Selaginellaceae, such as: is GC-biased ptDNA a trait common to many (or all) members of the *Selaginella* genus? And if yes, is *Selaginella* truly an outlier in terms of ptDNA nucleotide composition, or are there other plant lineages with similarly high GC contents? To address these questions, *rbcL* ptDNA sequences from a series of diverse plant taxa, including over 100 *Selaginella* species (representing most of the species diversity within the genus), were data-mined from NCBI and assessed for their GC-content. The *rbcL* gene was chosen as a ruler for assessing the overall plastid genome nucleotide composition because it is one of the only ptDNA genes whose sequence is readily available for many plant species (due to the fact that it is often used for phylogenetic analyses) and because its GC content scales reasonably well with the overall plastid-genome GC content: for complete ptDNA sequences, the Pearson correlation coefficient between the *rbcL* GC content and the whole-genome GC content is 0.82 (r^2 = 0.67). Altogether, *rbcL* sequences were collected for 167 charophytes, 911 liverworts, 811 mosses, 62 hornworts, 103 *Selaginella* species, 87 "non-*Selaginella*" lycophytes, 2,848 monilophytes, 855 gymnosperms, and 2,100 angiosperms. Summary statistics of the *rbcL* GC contents for these different plant lineages are shown in Figure 4.4.

The mean *rbcL* GC content for *Selaginella* species (52%; SD = 1.7%) is significantly higher than that from other plant lineages (Figure 4.4), including other lycophytes, which have an average *rbcL* GC composition of only 42.7% (SD = 0.9%). The monilophytes and gymnosperms are the closest to *Selaginella* with respect to *rbcL* GC content with values of 46.2% (SD = 2.3%) and 44.3% (SD = 1.3%), respectively. The charophytes and liverworts have the lowest observed mean *rbcL* GC contents at 39.6%, with standard deviations of 3.2% (charophytes) and 2.0% (liverworts). Overall, these findings suggest that *Selaginella* ptDNA has become GC biased since the Selaginellaceae

diverged from their common ancestor with quillworts and club mosses, which is believed to have occurred at least 400 million years ago (Kenrick and Crane 1997; Banks 2009).

All 103 *Selaginella* species that were analyzed have an *rbcL* GC content above 50%, except for *Selaginella sinensis,* which has an *rbcL* GC content of 44.8% (Appendix C). The most extreme *rbcL* GC content was observed for *Selaginella fragilis* (57.0%), which is greater than that of *S. moellendorffii* (50.6%) and *S. uncinata* (53.2%), and suggests that the ptDNA of *S. fragilis* may have a higher overall GC content than *S. uncinata* (i.e., >55%). The seven highest *rbcL* GC contents (ranging from 55-57% GC) come from *Selaginella* species that belong to the South American articulate subclade. Support for this subclade come from parsimony and Bayesian analyses using *rbcL* (Korall and Kenrick 2002, 2004) and from the observation that the *Selaginella* taxa that form this subclade possess a unique morphological marker: the rhizosphore develops from the upper surface of the stem and loops over the branch to grow downwards whereas in other *Selaginella* species it develops on the lower surface of the stem (Korall and Kenrick 2002). It should be mentioned that when maximum-likelihood analyses were performed on the *rbcL* dataset used by Korall and Kenrick (2002, 2004) the South American articulate subclade is still observed (data not shown). The phylogenetic affiliation of *S. sinensis*, the only *Selaginella* species shown to have an *rbcL* GC content below 50%, remains problematic. Parsimony analyses using *rbcL* place it (with low bootstrap support) as a sister to a clade containing all other species in the genus (Korall and Kenrick 2004). This could be an indication that the occurrence of GC-rich ptDNA in *Selaginella* taxa evolved after the split between the lineage that gave rise to *S. sinensis* and that leading to the other *Selaginella* species investigated in this study. That being said, parsimony inferred phylogenies are particularly sensitive to nucleotide composition biases (Eyre-Walker 1998), meaning they can cause distantly related organisms with similar GC contents to look more closely related than they actually are. Bayesian analyses with the same *rbcL* dataset (Korall and Kenrick 2004) place the GC-poor *S. sinensis* in a well-supported subclade with GC-rich *Selaginella* species — this position of *S. sinensis* is also supported by parsimony and Bayesian inferred phylogenies of 26S rDNA sequence data from *Selaginella* species (Korall and Kenrick 2004).

Evolution of Nucleotide Composition in Selaginella *PtDNA*

Why is *Selaginella* ptDNA GC biased? Or rather, why is *Selaginella* ptDNA not enriched in A and T like other available plastid-genome sequences? For virtually all completely sequenced plastid genomes the AT content is highest at what are considered to more neutrally evolving positions, such as fourfold-degenerate sites and noncoding regions (collectively defined as silent sites), and it is lowest at the more functionally constrained sites (first- and second-position codons sites and RNA-coding regions). Thus, it is generally believed that a neutral process, such as AT-mutation pressure or AT-biased gene conversion, is driving the nucleotide composition of most plastid genomes towards A and T (Howe et al. 2003; Kusumi and Tachida 2005; Khakhlova and Bock 2006). Could the observed GC content of *Selaginella* ptDNA be caused by the absence of either AT-mutation pressure or AT-biased gene conversion, or both? In the plastid genomes of *S. moellendorffii* and *S. uncinata* the GC content of silent sites is on average 48.9% and 52.5%, respectively. Similar values are also seen for the *rbcL* data from the different *Selaginella* species where the average GC content of fourfold-degenerate sites is 50.5% (Appendix C). Taken as a whole, these findings on the silent-site nucleotide composition of *Selaginella* ptDNA could (because %GC ≈ %AT) be a reflection of an unbiased mutation/gene-conversion process. There is also the possibility that two opposing neutral forces, such as AT-mutation pressure coupled with a GC-biased gene conversion mechanism (or GC-mutation pressure coupled with an AT-conversion bias) are balancing the silent-site nucleotide composition of *Selaginella* ptDNA resulting in a GC content of ~50%. The fact that *rbcL* synonymous substitution rates among *Selaginella* species are exceptionally high relative to those observed within most other land plant families (Korall and Kenrick 2002) may be an indication of an elevated mutation rate in *Selaginella* ptDNA; if true, this may imply a scenario where biased mutation pressure (AT or GC) is offset by a biased gene conversion mechanism.

There is also the possibility that natural selection is influencing the nucleotide composition of *Selaginella* ptDNA, and this may explain why, in addition to an elevated silent-site GC content, the more functionally constrained positions in *Selaginella* ptDNA, such as first- and second-position codon sites, are also skewed towards G and C relative to other plastid genomes (Figure 4.3). GC richness could be interpreted as an adaptation

for thermostability or UV-light tolerance. The thermostability hypothesis seems unlikely when considering that *Selaginella* species from northern climates, like Alaska, Canada, and Siberia have equally high *rbcL* GC contents as those from tropical and desert habitats. That being said, both hot- and cold-climate *Selaginella* species tend to grow in environments with reasonably high levels of UV radiation (Appendix C). Another adaptive hypothesis could be that there is selection for translational efficiency. Approximately one-third of the codons in the *S. moellendorffii* and *S. uncinata* plastid genomes are GC rich, which may correlate with the specific pool of tRNA anticodons that are available for plastid-gene translation. This topic is difficult to address because both the *S. moellendorffii* and *S. uncinata* plastid genomes encode a limited number of tRNAs. Nevertheless, of the 13 and 12 unique tRNAs that are respectively encoded in the *S. moellendorffii* and *S. uncinata* plastid genomes, all except one (*trnR*-ACG) are cognate to AT-rich codons. There is reason to believe, however, that many of the GC-rich codons in both the *S. moellendorffii* and *S. uncinata* plastid genomes are changed into AT-rich codons through RNA editing. If true, RNA editing may be influencing the GC content of *Selaginella* ptDNA.

RNA Editing in Selaginella *PtDNA and its Impact on Nucleotide Composition*

A few observations indicate that RNA editing is an important, widespread, and frequently occurring phenomenon in the plastids of *Selaginella* species. For instance, of the 75 protein-coding genes encoded in the *S. moellendorffii* plastid genome, 41 contain non-canonical start codons (A<u>C</u>G instead of A<u>T</u>G) and 23 have non-canonical stop codons (<u>C</u>GA, <u>C</u>AA, or <u>C</u>AG instead of <u>T</u>GA, <u>T</u>AA or <u>T</u>AG); the incidence of irregular start/stop codons in the *S. uncinata* ptDNA is even more prevalent with 50 and 29 non-canonical start and stop codons, respectively (Tsuji et al. 2007). If these codons were left unedited in mature transcripts, it would imply that 66% of the protein-coding genes in the *S. moellendorffii* ptDNA and 83% of those in the *S. uncinata* ptDNA are non-functional. However, preliminary investigations of plastid complementary DNA (cDNA) sequence data from *S. moellendorffii* (this study) and *S. uncinata* (Tsuji et al. 2007) indicate that RNA editing restores these irregular start/stop codons to their canonical states and induces C→U conversions in other plastid-RNA regions as well.

Analyses of 8,291 nt of cDNA sequence data from the *S. moellendorffii* plastid genome (4,501 nt from coding regions, 901 nt from intronic regions, and 2,889 nt from intergenic regions) reveals 104 edited sites, all of them corresponding to C→U changes (Table 4.3). Fifty-eight of these edited sites map to protein-coding ptDNA, 16 to intronic portions of the genome, and 30 to intergenic regions (Table 4.3). Of the cDNA data that covers nonstandard stop/start codons, all were restored to their canonical states. For *S. uncinata*, cDNA studies of the *rbcL* and *atpB* genes uncovered 54 and 111 C→U edited sites, respectively (Tsuji et al. 2007); this is one of the more massive examples of plastid RNA editing observed to date. These data, albeit providing only a small window into the degree of plastid RNA editing in *S. moellendorffii* and *S. uncinata*, indicate that RNA editing is a critical and prevalent process in these two taxa and operates at a greater level than that currently observed in other land plants. Studies suggest that in the plastid mRNAs of seed plants there are approximately 15-44 C→U editing sites (see Tillich et al. [2006] for a review). A significantly larger number of plastid RNA editing sites are observed for the fern *Adiantum capillus-veneris*, which has 315 C→U and 35 U→C editing sties (Wolf et al. 2004), and the hornwort *Anthoceros formosae*, which has 509 C→U and 433 U→C editing sites (Kugita et al. 2003). Organelle RNA editing in land plants is believed to be of monophyletic origin (Tillich et al. 2006). Although, among different species of land plant, the levels of RNA editing and the sites that get edited appear to be highly lineage specific (Jobson and Qiu 2008), leaving open the possibility that RNA editing has arisen multiple times in land plant evolution.

Considering that plastid RNA editing mostly involves C→U changes (with the exception of *A. formosae*), and that first- and second-position codon sites are generally the most edited positions in plastid genomes (Tillich et al. 2006; Jobson and Qiu 2008), then the elevated GC content of first- and second-position codon sites in the *S. moellendorffii* and *S. uncinata* ptDNAs may be a reflection of a large number of RNA editing sites in these genomes — an idea also suggested for *S. uncinata* by Tsuji et al. (2007). If true, this would imply a positive correlation between the ptDNA GC content and the number of C→U RNA editing sites. A cursory scan of complete land plant ptDNA sequences reveals that those with a large number of C→U RNA editing sites, such as *A. capillus-veneris*, are more GC-rich than those with only a few RNA editing

sites (Figure 4.2). Indeed, next to *S. moellendorffii* and *S. uncinata*, *A. capillus-veneris* has the most GC-biased land plant plastid genome observed to date (42% GC), whereas *Marchantia polymorpha*, which appears to lack plastid RNA editing (Freyer et al. 1997), and *Physcomitrella patens*, which is believed to have less than five plastid RNA editing sites (Miyata and Sugita 2004; Rüdinger et al. 2009), are the two most AT-rich land plant plastid genomes sampled thus far (Figure 4.2). The correlation between genomic GC content and levels of RNA editing has been highlighted in other studies (Malek et al. 1996; Jobson and Qiu 2008). Given these observations, one could suggest that RNA editing is acting as a genomic buffer against GC-biased mutation/conversion pressure by neutralizing T→C mutations, specifically those at functionally important first- and second-position codon sites. That being said, the evolution of RNA editing is a complicated topic and many sophisticated (and well articulated) models for its origins exist (Covello and Gray 1993; Lynch et al. 2006; Tillich et al. 2006; Jobson and Qiu 2008). I favor the model of Covello and Gray (1993), which posits that both the origin of RNA editing activity and the fixation of mutations at editable sites evolved primarily through random genetic drift but that the maintenance of RNA editing activity at specific sites is the result of natural selection. A salient point for any debate on RNA editing is that the genomic and nucleotide-composition contexts under which RNA editing evolved are unknown. The RNA editing machinery may have originated in a species with a moderately AT-rich organelle genome, but whose descendants were exposed to increasing GC pressure (see Reviewers' comments in Jobson and Qiu [2008]). If RNA editing is linked to the high GC content of *Selaginella* ptDNA, studies on the different *Selaginella* species, especially those at either extremes of the nucleotide-composition spectrum, may give insight into the link between RNA editing and GC content.

The GC Content in Other Genetic Compartments of Selaginella *Species*

The observation that *Selaginella* species have relatively GC-rich plastid genomes raises questions regarding their mtDNA — is it also GC-rich? An attempt at answering this question was made by collecting 56 kb of *S. moellendorffii* mtDNA sequence data. These data were generated by assembling mtDNA trace files produced by the DOE JGI *S. moellendorffii* nuclear-genome sequencing project. All of the mtDNA sequences were

subsequently blasted against the *S. moellendorffii* nucDNA assembly to ensure that no nuclear-genome-located mtDNA-like sequences (NUMTs) were collected. The results of these analyses suggest that there are a relatively small number of NUMTs in the nucDNA, < 20 kb distributed over ~50 sites (based on the 56 kb of mtDNA that were collected), and the NUMTs that are present are highly degenerate. Several attempts were made to complete the *S. moellendorffii* mitochondrial-genome sequence; however, the mtDNA of *S. moellendorffii*, like that from other land plant taxa, contains an abundance of repeats, which have spread throughout most of the intergenic and intronic regions. This feature of the *S. moellendorffii* mitochondrial genome means that the mtDNA trace files corresponding to intergenic and intronic regions collapse on top of one another upon assembly, resulting in a network of spurious repetitive motifs. Nonetheless, enough mtDNA sequence data were characterized to confidently describe the nucleotide landscape of this genome. In total, 8 protein-coding genes, 1 rRNA-coding gene, and 11 intergenic spacers were collected; the lengths and nucleotide compositions of these regions are summarized in Table 4.4. The overall GC content of the 56 kb of mtDNA sequence data is 67.8%, ranging from 62% for protein-coding genes, 65.1% for rRNA-coding genes, 69.1% for intronic regions, and 68.9% for intergenic spacers (Table 4.4). These nucleotide composition statistics suggests that the *S. moellendorffii* mtDNA is the most GC-rich mitochondrial genome observed to date, exceeding that of the green alga *Polytomella capuana*, which is 57% GC (Smith and Lee 2008a; Chapter 2). Could similar processes be driving the nucleotide composition of the *S. moellendorffii* mitochondrial and plastid genomes towards G and C? Analyses of cDNA sequences for *atp1* and *nad5* revealed 78 and 74 C→U RNA editing sites, respectively (Table 4.4). This implies that the *S. moellendorffii* mitochondrial genome is, like its plastid counterpart, experiencing exceptionally high levels of RNA editing; moreover, this provides support for the notion that RNA editing is connected to the high GC content of the *S. moellendorffii* organelle DNA. Interestingly, it is believed that in land plants the same machineries are responsible for editing both the mitochondrial and plastid derived transcripts (Freyer et al. 1997; Steinhauser et al. 1999).

On a final note, the GC content of the *S. moellendorffii* nuclear genome is ~45% — based on analyses of the first 150 scaffolds of the diploid genome assembly (~93.5%

of the complete nuclear-genome sequence), which is unremarkable in comparison to the nuclear genomes from other land plants. The discordance between the nucleotide composition of the *S. moellendorffii* nucDNA and organelle DNA could be a reflection of different mutation/conversion biases in these genomes. One crucial point, however, is that land plant nucDNA, unlike its organelle counterparts, is believed to experience very little (if any) RNA editing, thus, reinforcing the notion that RNA editing is connected to the high GC content of the *S. moellendorffii* organelle genomes.

Conclusions

There is a genus-wide GC bias in *Selaginella* ptDNA, which is most pronounced in South American articulate species. GC-rich ptDNA appears to be something unique to *Selaginella* species and is absent from lycophytes outside of the Selaginellaceae. It is argued that the cause of this GC bias is a combination of reduced AT-mutation pressure relative to other plastid genomes and a large number of C→U RNA editing sites. Partial-genome analysis of the *S. moellendorffii* mtDNA indicates that it is also GC biased (even more so than the ptDNA) and is arguably the most GC-rich organelle genome observed to date — the high GC content of the mtDNA also appears to be influenced by RNA editing. These findings provide convincing support for the earlier proposed theory that the GC content of land plant organelle DNA is positively correlated (and directly connected) to the levels of organelle RNA editing.

Acknowledgments

Many thanks to Jo Ann Banks for giving me permission to use the *S. moellendorffii* nuclear genome project trace file data.

Table 4.1. Expressed Sequence Tags (ESTs) From *S. moellendorffii* that Map to Organelle DNA

EST name[1]	Region in Plastid Genome (ptDNA coordinates)[2]
CAOU10168.fwd	*chlB* (29,807-30,499)
CAOU10168.rev	*chlB* (29,327-29,993)
CAOS8867.fwd	*psbT, psbN, psbH* (70,792-71,593)
CAOS8867.rev	*psbH* (71,209-71,853)
CAOS16609.fwd	*psbH-petB* intergenic region (71,663-72,228)
CAOS16609.rev	*psbH-petB* intergenic region (71,663-72,228)
CAOU13505.fwd	*psbH-petB* intergenic region (71,682-72,228)
CAOU13505.rev	*psbH-petB* intergenic region (71,682-72,228)
CAOT5212.fwd	*rrn16* (85,746-86,592)
CAOS7390.fwd	*rrn16* (86,101-86,741)
SmGB1262.fwd	*rrn23* (87,769-87,959)
SmGB324.fwd	*rrn23* (87,769-87,959)
SmGB611.fwd	*rrn23* (87,769-87,963)
SmGB1488.fwd	*rrn23* (87,769-87,973)
SmGB1504.fwd	*rrn23* (87,769-87,971)
CAOS1811.fwd	*rrn23* (88,238-88,942)
CAOS7317.fwd	*rrn23* (88,062-88,829)
CAOP540.fwd	*rrn23* (88,450-89,133)
CAOS7317.rev	*rrn23* (88,781-89,427)
CAOP540.rev	*rrn23* (88,728-89,367)
CAOT7732.fwd	*rrn23* (88,911-89,165)
CAOT7732.rev	*rrn23* (88,911-89,165)
CAOU12204.rev	*ndhA* (105,269-105,722)
CAOP8441.fwd	*ndhA* (105,467-106,165)
CAOP8441.rev	*ndhA* (105,469-106,165)
CAOS8728.fwd	*ndhA* (105,469-106,216)
CAOS8728.rev	*ndhA* (105,469-106,216)
CAOU12204.fwd	*ndhA* (106,347-107,013)
CAOP5639.fwd	*ycf1* (108,468-109,211)
CAOS13368.fwd	*psbC* (121,213-121,921)
CAOS13368.rev	*psbC* (121,693-122,255)

EST name[1]	Region in Mitochondrial Genome (mtDNA coordinates)[2]
CAOX650.fwd	*atp1* (23,644-24,321)
CAOX650.rev	*atp1* (24,686-24,069)
CAOU502.fwd	*nad7* (35,397-35,961)
CAOS3463.fwd	*cox1* (6,274-6,746 / 7,890-8,221)
CAOS3463.rev	*cox1* (10,910-10,394)

Note: All of the *Selaginella* EST sequences were downloaded from the DOE JGI *Selaginella moellendorffii* Genome Portal (v1.0) (http://genomeportal.jgi-psf.org/Selmo1/Selmo1.download.html) on 1 January 2009.

[1] EST names are searchable in GenBank; i.e., they are equivalent to a GenBank accession number.

[2] Plastid- and mitochondrial-genome coordinates correspond to the sequences (and their GenBank entries) presented in this study.

Table 4.2. Amount of DNA in the *S. moellendorffii* Diploid Nuclear Genome that Shares Homology with PtDNA

		# of similarity regions[a]	Mean length of similarity region (nt)	Longest similarity length (nt)	Cumulative length of similarity regions (nt)	Fraction of nuclear genome
Number of nucleotides in the *S. moellendorffii* diploid nuclear genome that share similarity with the plastid genome (by ptDNA subcategory[b])	Protein-coding genes[c]	100	72	229	7,207	3.62×10^{-5}
	Structural-RNA genes[d]	142	55	91	7,746	3.89×10^{-5}
	Introns[e]	19	91	301	1,726	0.86×10^{-5}
	Intergenic spacers[f]	93	52	138	4,823	2.42×10^{-5}
	Complete plastid genome[g]	307	67	530	21,502	10.79×10^{-5}

Note: Nuclear DNA analyses are based on the *S. moellendorffii* draft nuclear genome sequence (v1.0). Only the first 150 scaffolds of the nuclear-genome assembly were analyzed; ~93.5% of the *S. moellendorffii* nucDNA is contained in these 150 scaffolds and their cumulative length is 198.93 mega bases (note: this length represents two haplotypes).

[a] The number of distinct regions in the *S. moellendorffii* nucDNA that show >80% sequence identity and at least 30 nt of aligned length to the plastid DNA.

[b] Refers to the region of the plastid genome to which the nucDNA maps.

[c] Includes all of the identified protein-coding genes.

[d] Includes all of the identified tRNA- and rRNA-coding genes.

[e] Includes all of the identified introns.

[f] Includes all of the identified intergenic regions, including pseudogenes.

[g] Some of the "ptDNA similarity regions" in the *S. moellendorffii* nucDNA consist of coding and noncoding ptDNA-like sequences; thus, the overall sum of similarity regions (307) is less than the sum of the similarity regions based on ptDNA subcategory (354).

Table 4.3. RNA-editing Sites in the *S. moellendorffii* Plastid Genome

		Length (nt)[a]	# of RNA editing sites[b]
Protein-coding (by gene)	*ndhA*	411	25
	ndhH	128	6
	psbH	234	21
	psbN	132	1
	psbT	102	5
Intronic (by gene)	*ndhA*-intron	750	16
Intergenic (by region)	*ndhA/ndhH*	127	4
	psbC/psbZ	451	1
	psbN/psbH	86	3
	psbH/petB	800	22

Note: RNA-editing data could only be collected for regions described above; these data were derived from EST sequences produced by the DOE JGI *S. moellendorffii* nuclear-genome sequencing project.

[a] Refers to the length of the EST sequence covering the given region.

[b] All observed RNA editing sites involve C→U changes.

Table 4.4. General Features of the *S. moellendorffii* MtDNA Sequences

		Length (nt)	%GC	GC1	GC2	GC3	# of RNA editing sites[c]
All sites		55,780	67.8	---	---	---	---
rRNA-coding (*26S*)		4,150	65.1	---	---	---	---
Protein-coding (overall)		7,742	62.0	64.2	60.2	61.5	---
Protein-coding (by gene)	*atp1*	1,515	64.2	69.3	64.4	58.8	78
	atp9[a]	74	62.2	33.3	70.8	79.2	---
	cox1[a]	864	63.9	61.5	66.7	63.5	74
	nad2	1,055	63.2	68.1	60.4	61.3	---
	nad4[a]	879	61.5	58.0	65.2	61.5	---
	nad5	1,704	60.1	59.5	54.2	66.5	---
	nad7	1,096	59.8	68.5	54.0	56.7	---
	nad9	555	62.2	67.5	60.0	58.9	---
Intronic (overall)		27,490	69.1	---	---	---	---
Intronic (by gene)	*atp9*	3,547	69.7	---	---	---	---
	cox1	6,825	69.4	---	---	---	---
	nad2	4,798	68.6	---	---	---	---
	nad4	2,013	66.9	---	---	---	---
	nad5	5,580	69.5	---	---	---	---
	nad7	4,727	69.0	---	---	---	---
Intergenic (overall)		16,398	68.9	---	---	---	---
Intergenic (by region)[b]	?/*atp1*	930	67.1	---	---	---	---
	atp1/*nad5*	461	65.7	---	---	---	---
	nad5/?	2,916	70.6	---	---	---	---
	?/*atp9*	356	71.3	---	---	---	---
	?/*nad2*	769	66.8	---	---	---	---
	nad2/?	652	67.8	---	---	---	---
	?/*nad4*	283	67.8	---	---	---	---
	nad4/?	175	58.3	---	---	---	---
	?/*nad7*	540	65.7	---	---	---	---
	nad7/?	317	69.1	---	---	---	---
	nad9/*26S*	8,999	69.3	---	---	---	---

Note: GC1, GC2, and GC3 are the GC contents at first-, second-, and third-position codon sites, respectively.

[a] Partial sequence.

[b] A question mark (?) is used when the adjacent gene is undetermined.

[c] RNA-editing data could only be collected for *cox1* and *atp9*; these data were derived from EST sequences produced by the DOE JGI *S. moellendorffii* nuclear-genome sequencing project.

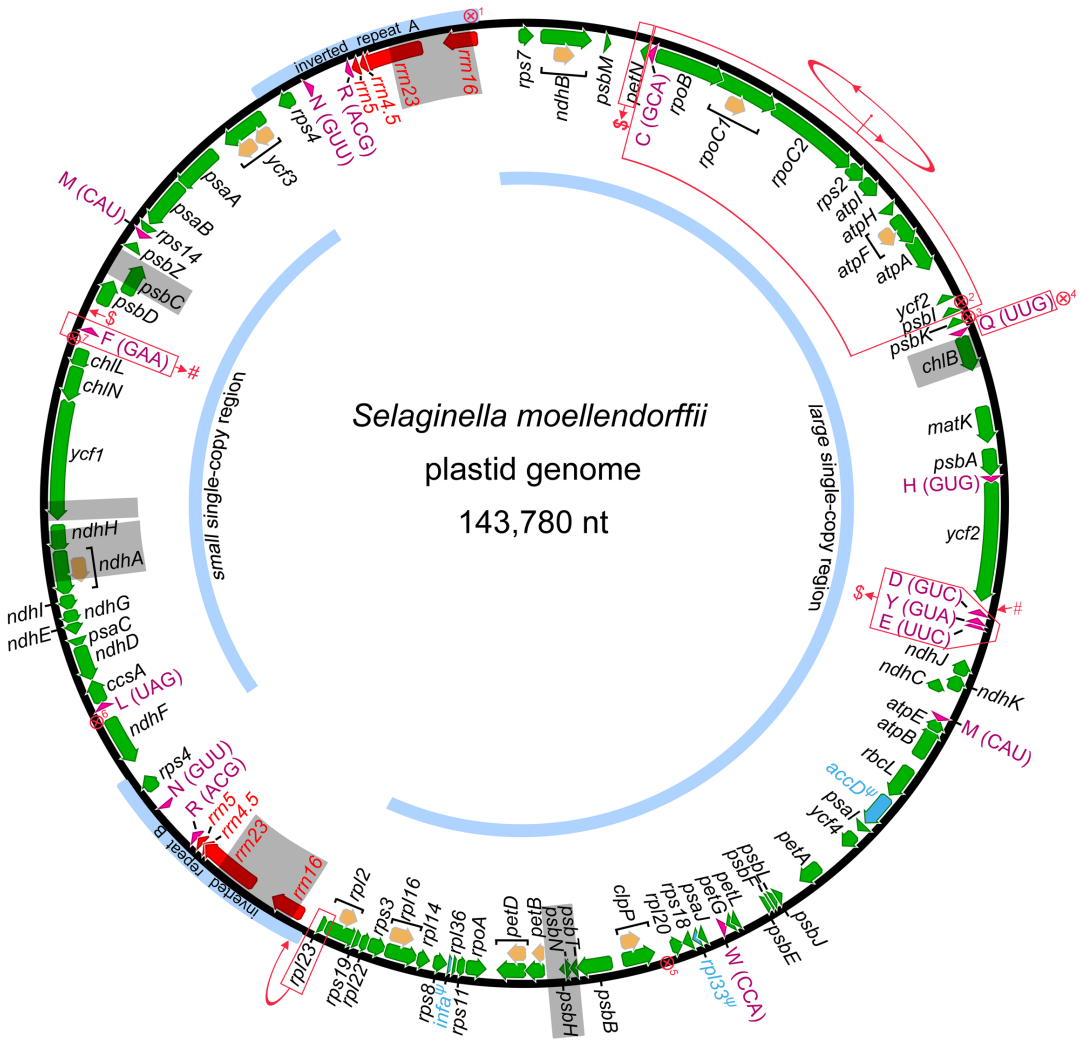

Figure 4.1. Genetic Map of the *S. moellendorffii* Plastid Genome

Ribosomal RNA-coding regions are red. Transfer RNA-coding regions are pink (their anticodon sequence is shown in brackets). Protein-coding regions are green and their introns are orange (all introns are of group-II affiliation). Pseudogenes are blue and are labeled with a ψ. Gray blocks highlight regions of the genome for which cDNA sequences were data mind. Differences in the ptDNA gene order between *S. moellendorffii* and *S. uncinata* are shown with red blocks, arrows, and symbols ($ and #). The genes and pseudogenes in the *S. uncinata* plastid genome that are absent from the *S. moellendorffii* ptDNA are signified as follows: $\otimes^1$ = *rpl23* and the 5′ end of *rpl2*; $\otimes^2$ = *psaM* pseudogene; $\otimes^3$ = *chlL* pseudogene; $\otimes^4$ = first copy of *trnQ*; $\otimes^5$ = *rps12* pseudogene; $\otimes^6$ = *rpl21* pseudogene; $\otimes^7$ = duplicate copy of *psbK* and *trnQ*.

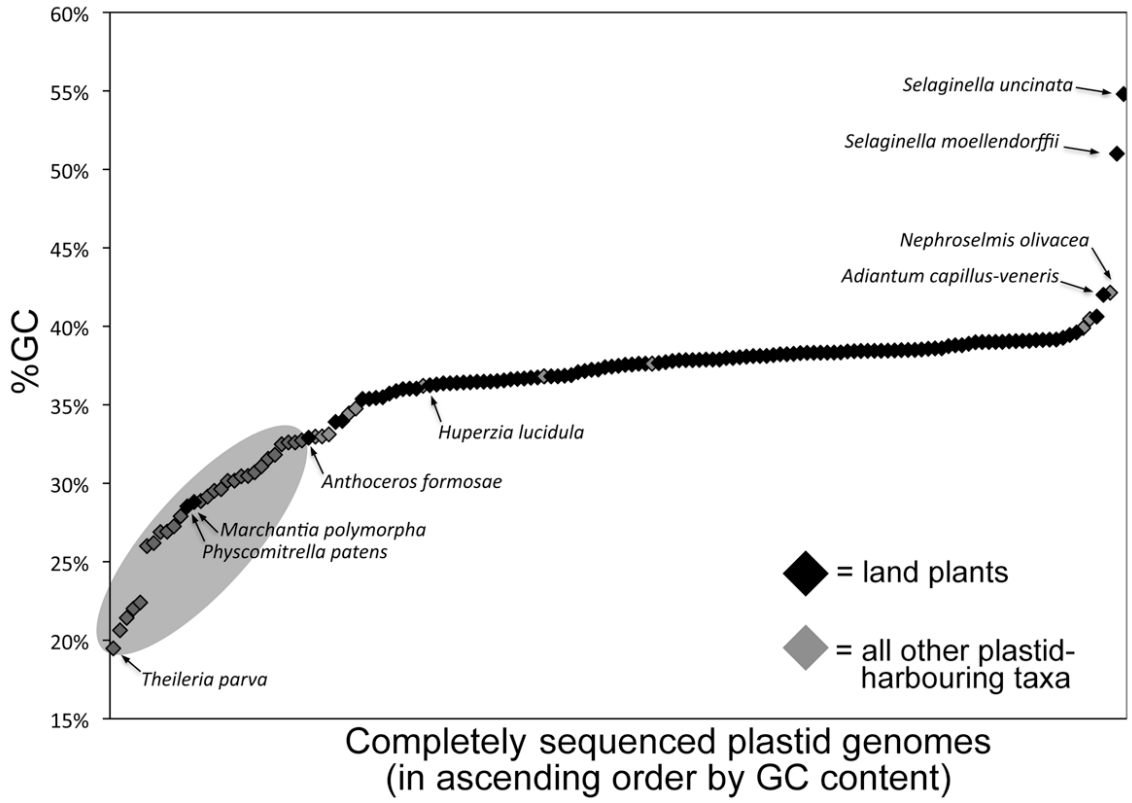

Figure 4.2. The GC Content of Completely Sequenced Plastid Genomes

Genomes are organized in ascending order by their GC content. The data points corresponding to species of interest are labeled. The shaded oval highlights some of the species for which plastid RNA-editing is believed to be either absent or restricted to less than five edited sites. The ptDNA GC-content data from which this graph was plotted are listed in Appendix B.

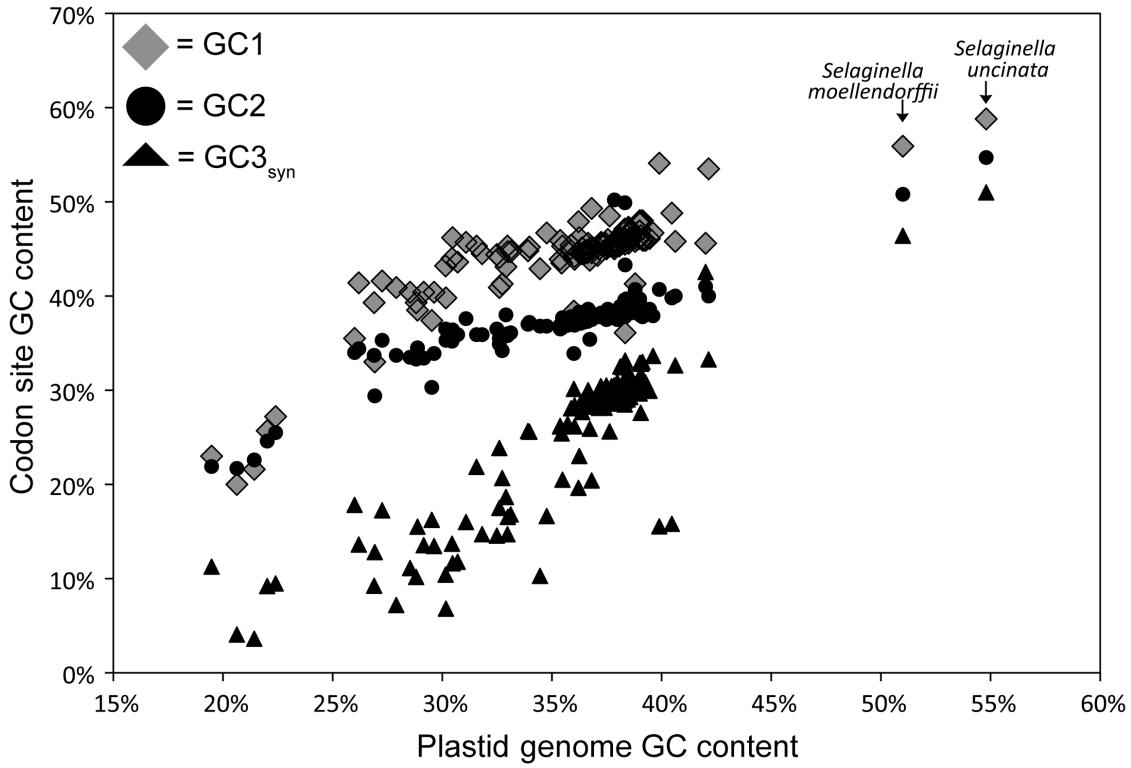

Figure 4.3. Scaling of Plastid Genome GC Content with GC Content at Different Codon-site Positions

GC1, GC2, and $GC3_{syn}$ represent the GC content at first-position, second-position, and third-position-synonymous codon sites, respectively. The ptDNA GC-content data from which this graph was plotted are listed in Appendix B.

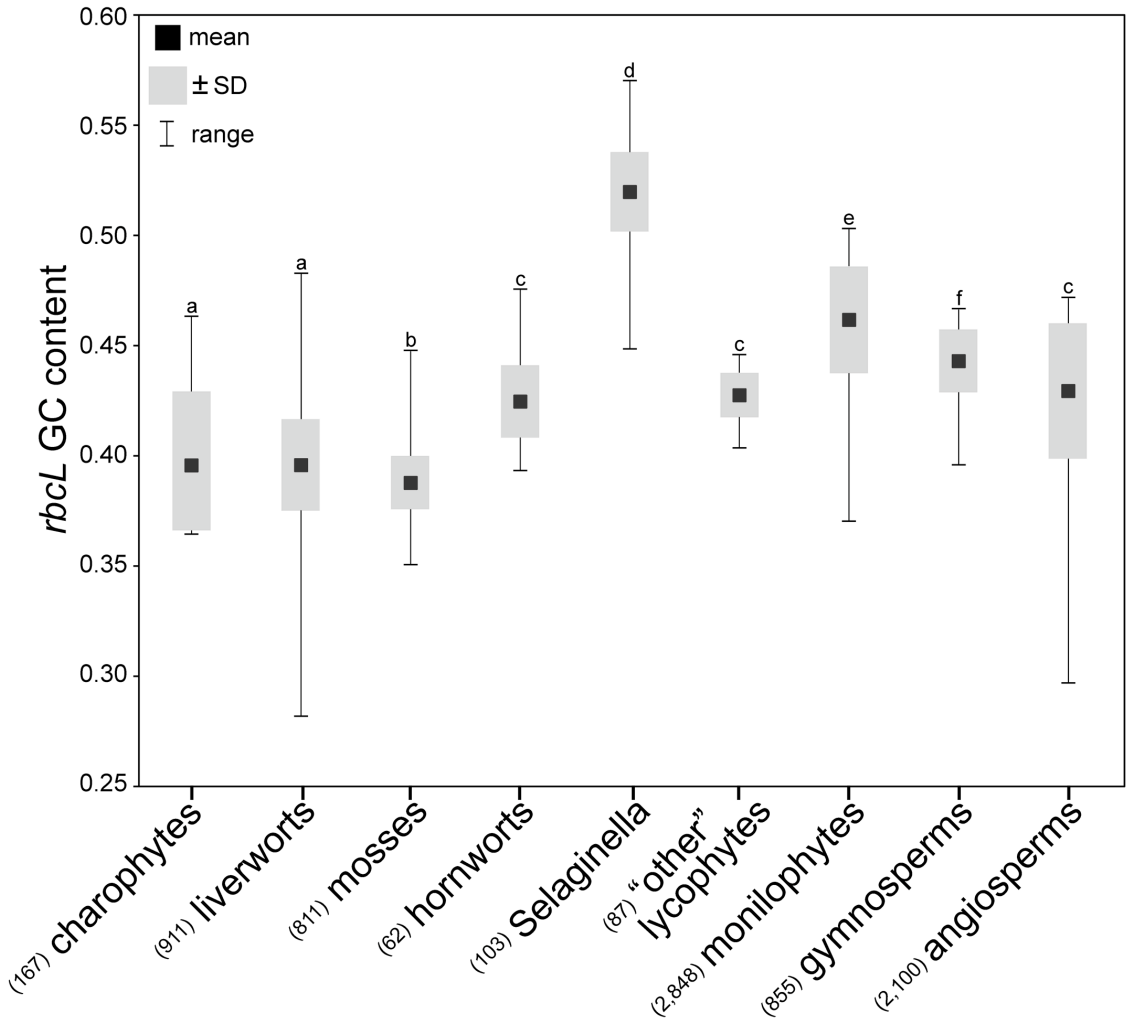

Figure 4.4. The *RbcL* GC Content for Major Plant Lineages

The number of species sampled from each lineage is shown in brackets. Plots with the same letter are not significantly different from one another (under the Tukey-Kramer method).

CHAPTER 5: NUCLEOTIDE DIVERSITY IN THE MITOCHONDRIAL AND NUCLEAR COMPARTMENTS OF *CHLAMYDOMONAS REINHARDTII*: INVESTIGATING THE ORIGINS OF GENOME ARCHITECTURE

Published as:

Smith DR, Lee RW (2008) Nucleotide diversity in the mitochondrial and nuclear compartments of *Chlamydomonas reinhardtii*: investigating the origins of genome architecture. BMC Evol Biol 8:156.

Abstract

Background: The magnitude of intronic and intergenic DNA can vary substantially both within and among evolutionary lineages; however, the forces responsible for this disparity in genome compactness are conjectural. One explanation, termed the mutational-hazard hypothesis, posits that genome compactness is primarily driven by two non-adaptive processes: mutation and random genetic drift — the effects of which can be discerned by measuring the nucleotide diversity at silent sites (π_{silent}), defined as noncoding sites and the synonymous sites of protein-coding regions. The mutational-hazard hypothesis holds that π_{silent} is positively correlated to genome compactness. We used the model organism *Chlamydomonas reinhardtii,* which has a streamlined, coding-dense mitochondrial genome and a non-compact, intron-rich nuclear genome, to investigate the mutational-hazard hypothesis. For measuring π_{silent} we sequenced the complete mitochondrial genome and portions of seven nuclear genes from seven geographical isolates of *C. reinhardtii*. Results: We found significantly more nucleotide diversity in the nuclear compartment of *C. reinhardtii* than in the mitochondrial compartment: net values of π_{silent} for the nuclear and mitochondrial genomes were 32×10^{-3} and 8.5×10^{-3}, respectively; and when insertions and deletions (indels) are factored in, these values become 49×10^{-3} for the nuclear DNA and 11×10^{-3} for the mitochondrial DNA (mtDNA). Furthermore, our investigations of *C. reinhardtii* revealed four previously undiscovered mitochondrial introns, one of which contains a fragment of the large-subunit (LSU) rRNA gene and another of which is found in a region of the LSU-rRNA gene not previously reported (for any taxon) to contain introns. Conclusions: At first glance our results are in opposition to the mutational-hazard hypothesis: π_{silent} was approximately 4 times greater in the nuclear compartment of *C. reinhardtii* relative to the mitochondrial compartment. However, when we consider the encumbrance of noncoding DNA in each of these *C. reinhardtii* compartments, we conclude that introns in the mtDNA impose a greater burden than those in the nuclear DNA and suggest that the same may be true for the intergenic regions. Overall, we cannot reject the mutational-hazard hypothesis and feel that more data on nucleotide diversity from green algae and other protists are needed.

Introduction

Genomic sequence data from the three domains of life have revealed a prodigious range of genome compactnesses; however, our knowledge of the processes responsible for this gamut of genomic architectures is contentious. One explanation is the mutational-hazard hypothesis (Lynch and Conery 2003; Lynch 2006; Lynch et al. 2006; Lynch 2007), which posits that genome compactness (defined by the proportion of intronic and intergenic DNA) is primarily driven by non-adaptive processes: namely, mutation and random genetic drift. The mutational-hazard hypothesis asserts that noncoding DNA (i.e., intronic and intergenic DNA) is a genetic liability because it is a target for deleterious and potentially lethal mutations, such as mutations affecting sequences involved with intron splicing and gene regulation. The hypothesis maintains that species with large effective population sizes (N_e) are more efficient at purging, or preventing the proliferation of, hazardous noncoding DNA because they experience less random genetic drift and thereby increase the efficacy of natural selection. The mutational-hazard hypothesis holds that the product of N_e and the mutation rate (μ) drives genome compactness; consequently, species whose genomes are coding rich should have a higher $N_e\mu$ than those whose genomes carry a surfeit of intronic and intergenic DNA.

Insights into N_e and μ can be acquired by measuring the nucleotide diversity at silent sites (π_{silent}), which are defined as noncoding sites and synonymous sites within protein-coding DNA. Since, for a diploid population at mutation-drift equilibrium, the rate at which new variation is introduced to a neutral nucleotide site in two randomly compared alleles is equivalent to 2μ (twice the mutation rate), and the rate at which variation is lost from a neutral site is $1/2N_e$, then the average number of nucleotide differences per neutral site is equivalent to the ratio of these two rates: $4N_e\mu$. Because silent sites are typically regarded as among the most neutrally evolving positions in a genome, measures of π_{silent} can provide an estimate of $4N_e\mu$. This formula can be simplified by substituting N_g, the effective number of genes per locus in a population, for N_e, giving a final equation of $\pi_{silent} = 2N_g\mu$, where N_g is equal to N_e for nuclear genes of haploid species and about one-half N_e for uniparentally-transmitted organelle genes (Birky et al. 1989). Uniparentally-transmitted organelle genomes (mitochondrial or plastid) are generally considered haploid, despite being present in multiple copies per

cell, because heteroplasmy (the existence of more than one organelle genome haplotype in the same individual) is rare. Moreover, in this instance, N_g is reduced further by the fact that during sexual reproduction only one of the parental sexual types transmits organelle genes to the next generation.

Large-scale studies have found a positive correlation between π_{silent} and genome compactness: in the compact genomes of prokaryotes π_{silent} tends to be > 50 x 10^{-3}; in the more bloated nuclear genomes of invertebrates and land plants it is in the range of 3 x 10^{-3} to 15 x 10^{-3}; and in the nuclear genomes of vertebrates, where noncoding DNA predominates, it appears to lie between 2 x 10^{-3} and 4 x 10^{-3} (Lynch and Conery 2003). There is evidence that this trend is also found in organelle genomes: comparative studies of nucleotide diversity in mitochondrial DNA (mtDNA) indicate that in the diminutive, coding-dense mitochondrial genomes of mammals π_{silent} is around 40 x 10^{-3}, whereas in the expanded mitochondrial genomes of land plants it is estimated to be <0.4 x 10^{-3} (Lynch et al. 2006). This disparity in π_{silent} between land plant and mammalian mitochondrial genomes is believed to be a reflection of the high mutation rates in mammalian mtDNA and the low mutation rates typically found in land plant mtDNA. Mutation rates have also been invoked to explain why, despite similar proposed values of N_g, the mitochondrial and nuclear genomes of mammals have opposite coding densities — in mammals estimates of μ for mtDNA are roughly 30 times those for nuclear DNA (Lynch et al. 2006). Although the relationship between π_{silent} and genome architecture is intriguing, the empirical data from which these correlations were derived are limited to a relatively small number of taxa and are generally skewed towards multicellular animals, with an overall lack of data for unicellular eukaryotes, especially green algae.

The model organism *Chlamydomonas reinhardtii*, a unicellular green alga of the chlorophycean class, is an excellent system for studying the evolution of genome compactness because it has a large, intron-rich nuclear genome and a small, compact mitochondrial genome, yet both genomes appear to have a similar mutation rate (Popescu et al. 2006; Popescu and Lee 2007). The nuclear genome of *C. reinhardtii*, which has been sequenced to 95% completion, is approximately 121 megabases (Mb), with about 17% of the nucleotides coding for proteins and structural RNAs (Merchant et al. 2007). Furthermore, the genome has an abundance of introns (~7 per protein-coding gene), and

the average intron length is longer than that of many eukaryotes and is more similar to multicellular organisms than to protists. In contrast, the mitochondrial genome of the standard laboratory strains of *C. reinhardtii* (derived from the Ebersold-Levine line) is streamlined, having a size of 15.8 kb and containing only 13 genes (Gray and Boer 1988; Michaelis et al. 1990; Vahrenholz et al. 1993). Moreover, at 82% coding it is one of the most compact mitochondrial genomes available from green algae (for a compilation see Pombert et al. [2006]), and although the mtDNA of one geographical isolate of *C. reinhardtii* (CC-1373) has an optional intron in *cob* (Colleaux et al. 1990) it is still >75% coding; this strain is often referred to as *Chlamydomonas smithii* but is in fact a member of the *C. reinhardtii* species (Boynton et al. 1987).

According to the mutational-hazard hypothesis, we would expect *C. reinhardtii* to have a high degree of silent-site nucleotide diversity in its mitochondrial genome (reflecting a large $2N_g\mu$) and a low degree of silent-site nucleotide diversity in its nuclear genome (reflecting a small $2N_g\mu$). To test this hypothesis and to investigate the correlation of $2N_g\mu$ with genome compactness, we measured π_{silent} in the mitochondrial and nuclear compartments from various geographical isolates of *C. reinhardtii*.

Materials and Methods

Strains, Culture Conditions, and DNA and RNA Extractions

All of the *C. reinhardtii* strains employed in this study were obtained from the Chlamydomonas Center at Duke University in July of 2006, with the exception to *C. reinhardtii* CC-277, which was obtained from the same source in 1991. Clonal cultures of each strain were prepared from a single vegetative colony recovered on agar medium (Harris 2009). For each of the seven strains, total genomic DNA was extracted using the DNeasy Plant Mini Kit (Qiagen, Germantown, MD), and total RNA was extracted using the RNeasy Plant Mini Kit (Qiagen).

Strain Confirmation

To verify that the seven *C. reinhardtii* isolates had been assigned the correct strain numbers we compared our sequences of the *ypt4*-VI, *ypt4*-VII, and *act*-VIII introns for

each isolate to those obtained in other studies (Dietmaier et al. 1995; Sugase et al. 1995; Liss et al. 1997) from the corresponding isolates. The results confirmed that the seven *C. reinhardtii* isolates employed here are the same as those used in previous reports.

Amplification and Sequencing of Genetic Loci

A PCR-based approach was used to amplify the mtDNA and the nuclear loci examined in this study. PCR experiments were performed in High Fidelity Platinum SuperMix (Invitrogen, Carlsbad, CA) using total genomic DNA as the template. Reverse transcriptase (RT) PCR reactions were performed with the SuperScript III One-Step RT-PCR System (Invitrogen) following the manufacturer's protocol. PCR and RT-PCR products were purified using the QIAquick PCR Purification Kit (Qiagen). The purified products were sequenced on both strands at the Macrogen sequencing facility, Rockville, MD, USA.

C. reinhardtii *Strain CC-503*

The complete mitochondrial genome of *C. reinhardtii* strain CC-503 (cw92 mt^{+}) was obtained by collecting and assembling mtDNA sequences generated from the *C. reinhardtii* nuclear genome sequencing project (http://genome.jgi-psf.org/Chlre3/Chlre3.home.html). These sequences were acquired by blasting the complete *C. reinhardtii* mitochondrial genome against the following databases at the United States Department of Energy Joint Genome Institute (DOE JGI): *C. reinhardtii* v3.1 unplaced reads and *C. reinhardtii* v3.1 bonus scaffolds. Blast hits showing >99% similarity to *C. reinhardtii* mtDNA were downloaded and assembled. All mitochondrial hits were subsequently checked against the draft nuclear genome of *C. reinhardtii* (v3.0 unmasked assembly) to insure that no nuclear-genome-located mtDNA-like sequences (NUMTs) were collected. Our assembly of the downloaded mtDNA sequences contains over 500 chromatogram reads and gives a complete *C. reinhardtii* CC-503 mitochondrial genome with 15-fold coverage.

Sequence Analyses

To ensure that the seven nuclear-encoded genes employed in this study are present only once in the *C. reinhardtii* nuclear genome we blasted each of the seven sequences against the *C. reinhardtii* draft nuclear genome. All seven genes returned a single hit, which is consistent with the hypothesis that these genes are present in single copies in the *C. reinhardtii* nuclear genome. The mitochondrial sequences obtained from each of the seven strains were also blasted against the *C. reinhardtii* draft nuclear genome to confirm that they are not NUMTs: the blast results suggest that there are very few copies of mitochondrial sequences in the nuclear compartment and the few that are present are highly degenerate; therefore, we are confident that none of the mtDNA sequences presented in this study are NUMTs. The mitochondrial introns of *cox1* and their secondary structures were identified using RNAweasel (Lang et al. 2007). DnaSP 4.0 (Rozas et al. 2003) was used for calculating all measures of genetic diversity. Theta was calculated using equation 3 of Tajima (1993). The McDonald-Kreitman test (McDonald and Kreitman 1991) and Tajima's D test (Tajima 1989) were performed with DnaSP.

Nucleotide Sequence Accession Numbers

The GenBank sequence accession numbers for the nucleotide sequences employed in this study are shown in Table 5.1.

Results

Strains and Their Genetic Loci

Seven geographical isolates of *C. reinhardtii* were employed in this study; their strain numbers, mating types, origins of isolation, and strain abbreviations are presented in Table 5.2. To assess levels of genetic diversity we sequenced the complete mitochondrial genome and portions of seven single-copy nuclear genes from each of the seven isolates. A genetic map of the *C. reinhardtii* mitochondrial genome is shown in Figure 5.1, and partial genetic maps of the seven nuclear genes are shown in Figure 5.2. We sequenced the entire mtDNA in order to employ both intergenic regions and

synonymous sites in our calculations of π_{silent} — previous studies on genetic diversity in mitochondrial genomes, due to a paucity of intraspecific sequence data, have tended to use only synonymous sites for estimating π_{silent}. Moreover, whole mtDNA sequences from *C. reinhardtii* allow for the comparison of synonymous-site nucleotide diversity (π_{syn}) in the standard mitochondrial protein-coding genes to that of *rtl*, a mitochondrial open reading frame (ORF) in the *C. reinhardtii* mtDNA coding for a putative reverse transcriptase-like protein (Boer and Gray 1988). It has been suggested that synonymous sites in *rtl* are under less selective constraints than those of the standard mtDNA protein-coding genes and that they may be more appropriate for estimating the neutral mutation rate in the mitochondrial compartment (Popescu and Lee 2007). For the nuclear loci, we sequenced mostly introns rather than exons because it is believed that in the *C. reinhardtii* nuclear genome intronic sites are more neutrally evolving than synonymous sites and may give more reliable estimates of the neutral mutation rate (Popescu et al. 2006). Sequences for two of the nuclear loci from the seven isolates have been previously reported (Dietmaier et al. 1995; Sugase et al. 1995; Liss et al. 1997) allowing us to confirm both our strain assignations and sequencing methods.

We were able to obtain the complete mtDNA sequence from an eigth strain of *C. reinhardtii* (CC-503) by collecting and assembling mtDNA sequences that were generated from the *C. reinhardtii* nuclear genome sequencing project (Merchant et al. 2007). Both *C. reinhardtii* CC-503 and *C. reinhardtii* CC-277 (one of the seven isolates described in Table 5.2) are cell-wall-less mutants recovered from the same "Ebersold-Levine" wild-type background of *C. reinhardtii,* but they have been separated for at least 35 years (Harris 2009). The mtDNA sequence of *C. reinhardtii* CC-503 is identical to that of *C. reinhardtii* CC-277; and when we downloaded the sequences of the seven nuclear loci for *C. reinhardtii* CC-503, they too were identical to those of *C. reinhardtii* CC-277. Therefore, for the purpose of this study we will be considering *C. reinhardtii* MA-1 as synonymous with *C. reinhardtii* CC-277 and CC-503.

Prior to this study a complete mtDNA sequence for *C. reinhardtii* was already available (GenBank accession number NC_001638); this sequence, which resulted from the accumulated efforts of multiple parties, mostly came from *C. reinhardtii* CC-277, or in some cases from strains having the same genetic background as *C. reinhardtii* CC-277.

The mtDNA sequence of *C. reinhardtii* CC-277 presented here differs at 46 positions relative to NC_001638; because 44 of these 46 differences are also present in the mtDNA of the other six *C. reinhardtii* isolates described here and because the *C. reinhardtii* CC-277 mtDNA sequence from this study was shown to be identical to the *C. reinhardtii* CC-503 mitochondrial genome, we feel that our version of the *C. reinhardtii* CC-277 mtDNA is currently the most accurate and that the discrepancies between our sequence and NC_001638 are the result of sequencing errors in the latter.

It is important to note that our annotation of the *C. reinhardtii* mitochondrial genome (Figure 5.1) does not contain the so-called rRNA-coding modules L2b and L3a. In previous studies each these modules was presumed to code for a non-core region of the large subunit (LSU) rRNA (Boer and Gray 1988). However, because sequence homologs of L2b and L3a have not been identified in the mtDNA of close relatives to *C. reinhardtii* (Popescu and Lee 2007; Denovan-Wright et al. 1994; Fan et al. 2003), or in any other genome, we have classified these regions as intergenic DNA, and have treated them as such for all genetic analyses.

Nucleotide Diversity

Summary statistics of the nucleotide diversity in *C. reinhardtii* are shown in Table 5.3. Two measures of nucleotide diversity were used to calculate variation within the *C. reinhardtii* mitochondrial and nuclear genomes: π, which is the average number of pair-wise nucleotide differences per site between sequences in a sample (Tajima 1989), and θ_W, which is based on the number of polymorphic sites in a sample of sequences but is independent of their frequency (Watterson 1975). With respect to both measures, the nuclear compartment shows significantly more silent-site nucleotide diversity than the mitochondrial compartment: net values of π_{silent} for the nuclear DNA and mtDNA were 31.96×10^{-3} and 8.54×10^{-3}, respectively. Net values of θ_W at silent sites are slightly higher at 33.02×10^{-3} for the nuclear compartment and $9.18\ 10^{-3}$ for the mitochondrial compartment. In all cases, silent sites in the various nuclear loci show more diversity than the silent sites in the mitochondrial compartment. The only exception to this is the nuclear gene *CBLP*, which has less synonymous-site diversity ($\pi_{syn} = 2.77 \times 10^{-3}$) than that of the mitochondrial protein-coding regions. Within the mitochondrial compartment,

diversity at intergenic and synonymous sites is similar (8.92×10^{-3} vs. 8.52×10^{-3}), as is the diversity of protein-coding regions and regions coding for structural RNAs (2.06×10^{-3} vs. 2.42×10^{-3}). The mitochondrial gene *rtl*, which encodes a putative reverse transcriptase, shows more diversity than the other mitochondrial protein-coding genes when all three codon sites are considered (3.07×10^{-3} vs. 2.06×10^{-3}) and slightly less diversity when looking only at synonymous sites (7.88×10^{-3} vs. 8.52×10^{-3}); however, it is unlikely that these observations are statistically significant.

Insertions and Deletions

For both the nuclear and mitochondrial compartments, insertions and deletions (indels) represent a large proportion of the observed polymorphisms (Table 5.3). In our alignments of the nuclear loci from the seven different strains of *C. reinhardtii*, 36% of mismatched nucleotides result from indels. The nuclear indels range from 1-31 nucleotides (nt) in length and have an average size of 4.5 nt. In the mitochondrial compartment indels represent 20% of the mismatched nucleotides. The mitochondrial indels range from 1-6 nt in length and have an average size of 2.5 nt. It is important to note that our estimates of nucleotide diversity shown in Table 5.3 are derived from sites in the alignment where all seven strains of *C. reinhardtii* have a nucleotide; therefore, sites corresponding to indels were removed from the alignment. If our methods for calculating π are modified to include indels (by counting each gap in the alignment as a nucleotide change) the overall values of π_{silent} in the nuclear and mitochondrial compartments become 49.27×10^{-3} ($\pm 4.89 \times 10^{-3}$) and 10.93×10^{-3} ($\pm 1.96 \times 10^{-3}$), respectively.

Testing for Neutrality

Two statistical tests were performed on the mitochondrial and nuclear datasets to examine for traces of selection: Tajima's *D*-test, which compares the average number of nucleotide differences between pairs of sequences to the total number of segregating sites (24), and the McDonald-Kreitman test, which compares the ratio of nonsynonymous to synonymous differences observed within a species to that observed between species (McDonald and Kreitman 1991). Tajima's *D* is slightly negative in all cases pertaining to

the mitochondrial compartment and in most cases pertaining to the nuclear compartment, but it is slightly positive for a few of the nuclear loci (the exons of *MAT3* and *PDK*, and the introns of *CBLP, PETC, PDK,* and *ACTIN*) (Table 5.3). In no case is Tajima's *D*-test statistically significant. The McDonald-Kreitman test was performed by comparing the ratio of nonsynonymous to synonymous polymorphisms within *C. reinhardtii* to the ratio of nonsynonymous to synonymous fixed differences between *C. reinhardtii* and *Chlamydomonas incerta* (one of the closest known non-interfertile relatives of *C. reinhardtii* (Pröschold et al. 2005) (Table 5.4) — this was done for all of the protein-coding regions surveyed in this study. Overall, no significant departures from neutral expectations were detected for any of the mitochondrial or nuclear loci, and in no case is the McDonald-Kreitman test statistically significant.

Mitochondrial Introns

Three of the *C. reinhardtii* strains (PA-1, MA-2, and FL) have introns in their mtDNA (Figure 5.1). *C. reinhardtii* MA-2 has a single intron, inserted into *cob*; *C. reinhardtii* FL has two introns, one in the L5-rRNA-coding module (the L5-intron) and one in the L7-rRNA-coding module (the L7-intron); and *C. reinhardtii* PA-2 has three introns, two in *cox1* and the L5-intron (note: the DNA sequence of the L5-intron in *C. reinhardtii* PA-2 is identical to that of *C. reinhardtii* FL). Of these introns only that of *cob* in *C. reinhardtii* MA-2 has been previously described (Colleaux et al. 1980). Like the intron of *cob* in *C. reinhardtii* MA-2, each of the four introns presented here has an ORF for which the deduced amino acid sequence shows similarity to a LAGLIDADG-type endonuclease. RT-PCR experiments confirm that all five introns, including their ORFs, are spliced-out in mature transcripts. Secondary-structure modeling suggests that the two introns in *cox1* are group-I introns belonging to subgroup D. Our analyses of the L5 and L7 introns suggest that they lack the core sequence and potential secondary structure necessary to be classified as either group-I or group-II introns; thus, at the present time they are considered highly-degenerate "unclassified" introns. A 35 nt duplicated portion of the L5 rRNA-coding module is found within the 5′ end of the L5 intron; RT-PCR experiments validate that this segment is in fact a component of the intron. The insertion sites of the L5 and L7 introns within the *C. reinhardtii* mtDNA and the nature of the

repeat found within the L5 intron are described in Figures 5.3A, 5.3B, and 5.3C, respectively. The insertion sites of the L5 and L7 introns in context to the *C. reinhardtii* LSU rRNA sequence are shown in Figure 5.3D.

Discussion

Accounting for the Differences in Π_{silent}

Before we interpret our data on nucleotide diversity in relation to the mutational-hazard hypothesis, let us first try to account for the values of π_{silent} that we observed. Overall, in *C. reinhardtii* we found 3.7-fold more nucleotide diversity at silent sites in the nuclear compartment than in the mitochondrial compartment; and when indels are taken into consideration, π_{silent} appears to be 4.5 times greater in the nuclear DNA compared to the mtDNA. Assuming that π_{silent} is equal to $2N_g\mu$, we can discuss our findings on nucleotide diversity in relation to μ and N_g.

In a recent study that compared silent-site substitution rates in the mitochondrial and nuclear genomes between *C. reinhardtii* and *C. incerta*, it was concluded that the mutation rate in the nuclear compartment of these taxa is approximately the same as that in the mitochondrial compartment (Popescu et al. 2006; Popescu and Lee 2007). If μ is similar in both the nuclear DNA and mtDNA of *C. reinhardtii*, then it appears that differences in N_g would have to explain the disparity in nucleotide diversity that we observe between these genomes.

In order to arrive at the values of π_{silent} observed in this study, N_g would have to be higher for the nuclear genome than for the mitochondrial genome (again, assuming equal mutation rates); there are a few reasons why this might be the case. For the haploid alga *C. reinhardtii*, nuclear genes are inherited biparentally and mitochondrial genes are inherited uniparentally (Harris 2001). As mentioned earlier, uniparental inheritance is thought to reduce N_g in the mitochondrial compartment by approximately one-half relative to that in the nucleus (Birky et al. 1989). Thus, for *C. reinhardtii* we might expect a value of π_{silent} in the mitochondrial compartment to be around one-half of what is observed in the nuclear compartment. Uniparental inheritance also implies that the mtDNA of *C. reinhardtii* has less opportunity for recombination than the nuclear DNA, which may make the mitochondrial genome as a whole more susceptible to the effects of

selective sweeps and purifying selection, both of which can reduce N_g, resulting in an even smaller than expected value of π_{silent} in the mitochondrial compartment (Bazin et al. 2006; Meiklejohn et al. 2007). One way to detect the influences of selection is with the McDonald-Kreitman test, where positive selection is inferred if the test returns a value for the neutrality index (NI) < 1, and purifying selection is indicated by NI > 1 (Rand and Kann 1998; Bazin et al. 2006; Meiklejohn et al. 2007). Although the results of the McDonald-Kreitman test for *C. reinhardtii* versus *C. incerta* showed no significant departure from neutral expectations, values of NI were < 1 for *rtl* and the concatenated sequence of the standard mitochondrial protein-coding regions (Table 5.4), which might be an indication of positive selection. Moreover, Tajima's *D* test returned negative values for all of the mtDNA regions that were examined (Table 5.3) — negative values of Tajima's *D* test can be an indication of a recent selective sweep of a linked mutation (Tajima 1993) — but again these findings were not statistically significant.

A further consideration is that the mutation rate in the *C. reinhardtii* mitochondrial compartment may be slightly lower than that in the nuclear compartment. When Popescu and Lee (2007) estimated μ to be similar for both the nuclear and mitochondrial genomes of *C. reinhardtii*, they used the synonymous substitution rate in *rtl* (which was about double that of the standard mitochondrial protein-coding regions) as an estimate of the neutral mutation rate in the mitochondrial compartment. In contrast, we found π_{syn} in *rtl* to be similar to that of the standard mitochondrial protein-coding regions (7.88×10^{-3} vs. 8.52×10^{-3}), and although this could be an artifact of a small sample size, it might suggest that Popescu and Lee overestimated μ in the *C. reinhardtii* mtDNA.

Π_{silent} in Relation to Previous Studies on C. reinhardtii *and Other Unicellular Eukaryotes*

Our estimates of π_{silent} in the nuclear and mitochondrial genomes of *C. reinhardtii* are approximately 32×10^{-3} and 8.5×10^{-3}, respectively; and with indels factored in, these values become ~ 50×10^{-3} for the nuclear DNA and about 11×10^{-3} for the mtDNA. The only other estimates of nucleotide diversity in the *C. reinhardtii* nuclear and mitochondrial genomes that we could find were that of Lynch and Connery (2003), who, using 11 kb of mostly noncoding nuclear DNA from *C. reinhardtii* strains MA-1 and MN, found π_{silent} in the nuclear compartment to be ~ 40×10^{-3} (they estimated $N_e\mu$ to be ~

20 x 10^{-3}), which is in agreement with our observations. With respect to other green algae, we are unaware of any reported estimates of within-population silent-site variation in either nuclear or mitochondrial genomes.

There is an overall lack of data on π_{silent} for unicellular species; however, the data that are available are relatively consistent with our results for *C. reinhardtii*. Among unicellular fungi, average values of π_{silent} in the nuclear and mitochondrial compartments (based on data for three different species) were estimated to be 50 x 10^{-3} and 12 x 10^{-3}, respectively (Lynch and Conery 2003; Lynch et al. 2006), which is comparable to what we observed in *C. reinhardtii*. Furthermore, the genera *Paramecium* and *Trypanosoma* appear to have ratios of $\pi_{silent(mitochondrion)}/\pi_{silent(nucleus)}$ of approximately 0.4 (Lynch et al. 2006), which is within the range of our estimates for *C. reinhardtii* (0.2-0.25). All studies deriving measures of π_{silent} from the mitochondrial genome of unicellular eukaryotes have used three or fewer loci for their estimates and have tended to focus only on synonymous sites, which makes our analysis one of the most comprehensive for any mitochondrial genome from a unicellular eukaryote to date.

It is also of interest to compare the nucleotide diversity of *C. reinhardtii* to that of multicellular species. Our estimation of nucleotide diversity in the nuclear genome of *C. reinhardtii* is much greater than that observed for animals: ~ 9 times larger than the average estimate for mammals (3.6 x 10^{-3}), and 2 times greater than the average for invertebrates, which is believed to be < 14.8 x 10^{-3} (Lynch and Conery 2003). Our approximation of π_{silent} for the mitochondrial genome of *C. reinhardtii* is much smaller than that of animals: 0.2 times the average for mammals (40 x 10^{-3}) and 0.1-0.6 times the average for invertebrates (11 x 10^{-3}– 67 x 10^{-3}) (Lynch 2006). In comparison to land plants, *C. reinhardtii* has twice the amount of silent-site nucleotide variation in its nuclear genome (π_{silent} in land plants is estimated to be ~ 15 x 10^{-3}), and although we are unaware of any reliable assessments of π_{silent} in land plant mtDNA, it is purported to be < 0.4 x 10^{-3} (Lynch et al. 2006), which is less than 0.05 times what we observed for the *C. reinhardtii* mtDNA.

Testing the Mutational-hazard Hypothesis

At first glance, our estimates of nucleotide diversity in the mitochondrial and nuclear genomes of *C. reinhardtii* appear contrary to what would be expected under the mutational-hazard hypothesis. We found π_{silent} to be ~ four times greater in the nuclear compartment than in the mitochondrial compartment, but based on the streamlined nature of the *C. reinhardtii* mitochondrial genome in relation to its noncompact nuclear genome, one might expect the mutational-hazard hypothesis to predict a greater value of π_{silent} for the mtDNA. However, before we conclude that our findings are in opposition to the mutational-hazard hypothesis, we must first consider what the actual "encumbrance" of noncoding DNA is for the mitochondrial and nuclear genomes in *C. reinhardtii*.

As described earlier, the basic premise of the mutational-hazard hypothesis is that noncoding DNA magnifies the target site for deleterious mutations, thereby, increasing the susceptibility of a genome to degenerative changes. The mutational disadvantage of noncoding DNA, therefore, is critically dependent on: 1) the number of nucleotides that are associated with gene function (n), and 2) the per-nucleotide mutation rate (μ). These two terms can be combined to define the overall mutational disadvantage (s), where $s = n\mu$ (Lynch et al. 2006; Lynch 2002). It is predicted that the threshold in a genome below which noncoding DNA can proliferate is $2N_g s < 1$, or alternatively $2N_g\mu < 1/n$ (Lynch et al. 2006; Lynch 2002). Although it is difficult to estimate n for intergenic regions, values of n for intronic regions can be predicted with reasonable confidence. For the spliceosomal introns of eukaryotic genomes, n is believed to be around 25 (Lynch 2002), which gives a threshold for intron colonization in the nuclear compartment of $2N_g\mu < 0.04$. Because mitochondrial introns are self-splicing and do not rely on a spliceosome for excision they have more nucleotides that are critical for proper splicing; thus, we can conservatively say that for mitochondrial introns n is between 75-100 (see Lang et al. [2007] for a review on mitochondrial-intron folding), giving a threshold for intron proliferation in the mitochondrial compartment of $\sim 2N_g\mu < 0.01$. Based on the mutational-hazard hypothesis, our estimates of $2N_g\mu$ (i.e. π_{silent}) in the nuclear and mitochondrial compartments of *C. reinhardtii*, whether including or excluding the influence of indels, lie too close to the predicted thresholds for intron proliferation in these genomes to accurately forecast intron abundance.

We do not know what the encumbrance of intergenic regions is for either the nuclear or the mitochondrial genome of *C. reinhardtii*, but if it is (or was at some point in the past) substantially higher for the mtDNA than the nuclear DNA, it would indicate that the mitochondrial compartment is a less permissive environment for the proliferation of intergenic DNA. And although this is highly speculative, there is one reason why this might be the case. In the *C. reinhardtii* mitochondrial genome, mature-RNA transcripts are generated by precise endonucleolytic cleavage of long polycistronic precursor-messenger RNAs; it is believed that these immature transcripts are cleaved in regions of the RNA corresponding to intergenic sites in the mitochondrial genome, and that processing is critically dependent on the primary sequence and the secondary structure of these regions (Gray and Boer 1988). This dependence implies that the intergenic sites in the *C. reinhardtii* mitochondrial genome may have a large mutational burden associated with them, perhaps large enough to impose a barrier on the amplitude of intergenic mtDNA. Moreover, the polycistronic nature of the *C. reinhardtii* mitochondrial transcripts suggest that the regulatory elements within the intergenic mtDNA carry the increased burden of being responsible for many genes — a burden not typically associated with the monocistronic gene regulation of nuclear DNA (Gurley et al. 2006).

If the burden of intronic and intergenic DNA is higher in the mtDNA of green algae and other protists, then we might expect to find very low values of π_{silent} in the mitochondrial genomes of species from these groups that have an abundance of intronic and intergenic sequences — for examples see references (Pombert et al. 2006; Turmel et al. 2007; Forget et al. 2002).

It is worth noting that the *C. reinhardtii* chloroplast genome when compared to its mitochondrial counterpart has a similarly low density of introns but a substantially greater proportion of intergenic DNA (Maul et al. 2002). Using 1500 nt of chloroplast DNA (composed of the *pet*A gene and a single intergenic region) we found no nucleotide polymorphisms in any of the seven geographical isolates of *C. reinhardtii* employed in this study (Smith DR and Lee RW, unpublished data). For a genome rich in intergenic DNA this value is consistent with what one might expect under the mutational-hazard hypothesis.

Novel Mitochondrial Introns

An unforeseen consequence of this study is the discovery of four previously unreported *C. reinhardtii* mitochondrial introns, two of which (the L5 and L7 introns) are unusual. Although a detailed description of these introns is beyond the scope of this paper, they each contain a characteristic that is notable: 1) the L5 intron carries a 35 nt portion of the L5 rRNA-coding module within the 5′ end of its DNA sequence — this is an unprecedented feature for a mitochondrial intron. And 2) the insertion site of the L7 intron in relation to the *C. reinhardtii* LSU-rRNA secondary-structure model of Boer and Gray (1988) corresponds to domain III (Figure 5.3); we believe this to be the first example (for any taxon) of an intron found in domain III of the LSU rRNA.

The discovery of four new optional introns in the *C. reinhardtii* mtDNA does not alter the notion that in *C. reinhardtii* the mitochondrial compartment is significantly more compact than the nuclear compartment: the mitochondrial genome of *C. reinhardtii* PA-1, the isolate with the most introns, is still ~ 67% coding compared to < 20% for the nuclear genome.

Conclusions

The main objective of this study was to investigate genome compactness from a population-genetic perspective and, in doing so, test a contemporary hypothesis regarding the origins of genome architecture — the mutational-hazard hypothesis. Our findings may not appear to be in full agreement with the mutational-hazard hypothesis: we found approximately four times more nucleotide diversity in the nuclear compartment of *C. reinhardtii* relative to the mitochondrial compartment. However, when the $2Ng\mu$-threshold for the proliferation of intronic and intergenic DNA is considered, we conclude that introns impose a greater burden on *C. reinhardtii* mtDNA and suggest that the intergenic regions of this genome do so as well. Overall, we cannot reject the mutational-hazard hypothesis.

Table 5.1. Genbank Accession Numbers of the *C. reinhardtii* MtDNA and NucDNA Sequences Employed for Measuring Nucleotide Diversity

Strain	mtDNA[a]	*CBLP*[a]	*PETC*[a]	*PDK*[a]	*MAT3*[a]	*SFA*[a]	*ACTIN*-VIII[b]	*YPT4*-VI[b]	*YPT4*-VII[b]
CC-277[c]	EU306622	EU306630	EU306651	EU306644	EU306632	EU306658	D50838	U13167	U13167
CC-1373	EU306617	EU306625	EU306646	EU306639	EU306633	EU306653	U70571	U55911	U55912
CC-1952	EU306621	EU306626	EU306647	EU306640	EU306634	EU306654	U70563	U55893	U55894
CC-2342	EU306620	EU306627	EU306648	EU306641	EU306635	EU306655	U70569	U55905	U55906
CC-2343	EU306623	EU306628	EU306649	EU306642	EU306636	EU306656	U70561	U55889	U55890
CC-2344	EU306619	EU306629	EU306650	EU306643	EU306637	EU306657	U70562	U55891	U55892
CC-2931	EU306618	EU306624	EU306645	EU306638	EU306631	EU306652	U70568	U55901	U55902

[a] Present study.

[b] Sequences from (Liss et al. 1997) except for D50838, which is from (Sugase et al. 1995), and U13167, which is from (Dietmaier et al. 1995).

[c] The mitochondrial genome from this strain was shown to be identical to that of *C. reinhardtii* CC-503.

Table 5.2. *C. reinhardtii* Strains Used for Measuring MtDNA and NucDNA Diversity

Strain	Mating Type	Strain Synonym	Geographical Origin (USA)	Abbreviation [a]	Reference
CC-277	mt^+	*cw15*	Amherst, Massachusetts	MA-1	Harris (2009)
CC-1373	mt^+	*C. smithii*	South Deerfield, Massachusetts	MA-2	Hoshaw and Ettl (1966)
CC-1952	mt^-	*C. grossii*	Plymouth, Minnesota	MN	Gross et al. (1988)
CC-2342	mt^-	Jarvik 6	Pittsburgh, Pennsylvania	PA-1	Spanier et al. (1992)
CC-2344	mt^+	Jarvik 356	Malvern, Pennsylvania	PA-2	Spanier et al. (1992)
CC-2931	mt^-	Harris 6	Durham, North Carolina	NC	Harris (2009)
CC-2343	mt^+	Jarvik 124	Melbourne, Florida	FL	Spanier et al. (1992)

[a] Strain abbreviations are based on the USA state from which the strains were isolated.

Table 5.3. Nucleotide Diversity in the Mitochondrial and Nuclear Compartments of *C. reinhardtii*

			# of sites	*S*	# of Indels[g] (length nt)	π x 10^{-3} (SD x 10^{-3})	θ_W x 10^{-3} (SD x 10^{-3})	π_{syn} x 10^{-3}	π_{nsyn} x 10^{-3}	Tajima's D test (P value)
mtDNA	Complete genome[a]		15280	134	15 (36)	3.35 (0.68)	3.66 (0.31)	---	---	-0.49 (>0.10)
	Protein-coding[b]		8154	44	1 (6)	2.06 (0.43)	2.20 (0.33)	8.52	0	-0.38 (>0.10)
	structural RNA genes[c]		3502	23	4 (4)	2.42 (0.42)	2.68 (0.56)	---	---	-0.55 (>0.10)
	rtl		1118	9	1 (3)	3.07 (0.87)	3.29 (1.10)	7.88	1.89	-0.35 (>0.10)
	Intergenic[d]		2434	58	9 (23)	8.92 (1.88)	9.73 (1.28)	---	---	-0.48 (>0.10)
	Silent sites[e]		5152	99	12 (26)	8.54 (1.03)	9.18 (0.65)	---	---	---
Nuclear DNA	Intronic (overall)		4294	359	47 (216)	33.50 (3.15)	33.81 (1.79)	---	---	-0.16 (>0.10)
	Silent sites[f]		4824	381	50 (219)	31.96 (3.03)	33.02 (1.88)	---	---	---
	Exonic (overall)		1614	25	1 (9)	6.02 (0.99)	6.58 (1.29)	19.57	1.42	-0.48 (>0.10)
	Exonic (by gene)	*CBLP*	420	1	0	0.68 (0.47)	0.97 (0.97)	2.77	0	-1.00 (>0.10)
		PETC	300	2	0	2.54 (0.78)	2.72 (1.92)	9.69	0	-0.27 (>0.10)
		PDK	222	3	0	6.86 (1.88)	5.52 (3.18)	26.35	0	1.10 (>0.10)
		MAT3	408	10	1 (9)	10.50 (2.70)	10.00 (3.16)	24.37	4.77	0.27 (>0.10)
		SFA	264	9	0	9.74 (2.80)	13.91 (4.64)	41.14	0	-1.59 (>0.10)
	Intronic (by gene)	*CBLP*	578	77	8 (75)	58.30 (7.41)	54.30 (6.20)	---	---	0.26 (>0.10)
		PETC	621	43	3 (5)	30.21 (6.07)	28.26 (4.31)	---	---	0.39 (>0.10)
		PDK	691	37	8 (42)	23.22 (5.58)	21.86 (3.59)	---	---	0.20 (>0.10)
		MAT3	560	30	4 (25)	21.60 (3.03)	21.87 (3.99)	---	---	-0.07 (>0.10)
		SFA	847	98	12 (15)	43.57 (4.53)	48.67 (4.77)	---	---	-0.61 (>0.10)
		YPT4	790	47	9 (39)	22.66 (4.50)	24.28 (3.54)	---	---	-0.38 (>0.10)
		ACTIN	207	27	3 (15)	53.37 (8.18)	53.37 (10.2)	---	---	0.01 (>0.10)

Note: *S*, number of segregating sites; Indels, insertion and deletions; π, nucleotide diversity; θ_W, Theta (per-site) from Watterson estimator; π_{syn}, nucleotide diversity at synonymous sites; π_{nsyn}, nucleotide diversity at nonsynonymous sites; SD, standard deviation.

[a] Includes only one telomere.

[b] Includes all protein-coding genes excluding *rtl*.

[c] Includes rRNA- and tRNA-coding regions.

[d] Includes intergenic regions and one telomere.

[e] Includes synonymous sites, intergenic regions, and one telomere.

[f] Includes synonymous sites and intronic regions.

[g] Nucleotide diversity does not include indels. Consecutive indel sites are counted as single event. Indel length refers to the sum of all indels.

Table 5.4. McDonald-Kreitman Test Comparing the Ratio of Nonsynonymous to Synonymous Differences within *C. reinhardtii* to that Found Between *C. reinhardtii* and *Chlamydomonas incerta*

			Polymorphisms within *C. reinhardtii*	Substitutions between *C. reinhardtii* and *C. incerta*	*NI*	*G*	*P*
mtDNA	Protein-coding[a]	Nonsynonymous	3	61	0.653	0.533	0.465
		Synonymous	36	478			
	rtl	Nonsynonymous	5	111	0.746	0.185	0.667
		Synonymous	4	119			
Nuclear DNA	Exonic (overall)[b]	Nonsynonymous	1	18	0.159	4.837	0.027
		Synonymous	21	60			
	CBLP	Nonsynonymous	0	0	undef	undef	undef
		Synonymous	1	1			
	PETC	Nonsynonymous	0	3	0.000	undef	undef
		Synonymous	2	4			
	PDK	Nonsynonymous	0	2	0.000	undef	undef
		Synonymous	3	18			
	MAT3	Nonsynonymous	1	9	0.677	1.081	0.298
		Synonymous	6	18			
	SFA	Nonsynonymous	0	4	0.000	undef	undef
		Synonymous	9	19			

Note: *NI*, neutrality index (ratio of nonsynonymous to synonymous polymorphisms within *C. reinhardtii* compared to the ratio of nonsynonymous to synonymous fixed differences between *C. reinhardtii* and *C. incerta*); *G,* G-test of independence (determines if the proportion of nonsynonymous substitutions is independent of whether the substitutions are fixed or polymorphic); *P,* probability of G-test; undef, undefined. *C. incerta* data came from (6) and (7). Note: in no case was the McDonald-Kreitman test statistically significant.

[a] Includes all protein coding genes excluding *rtl.*

[b] Concatenated exons from all 5 nuclear genes.

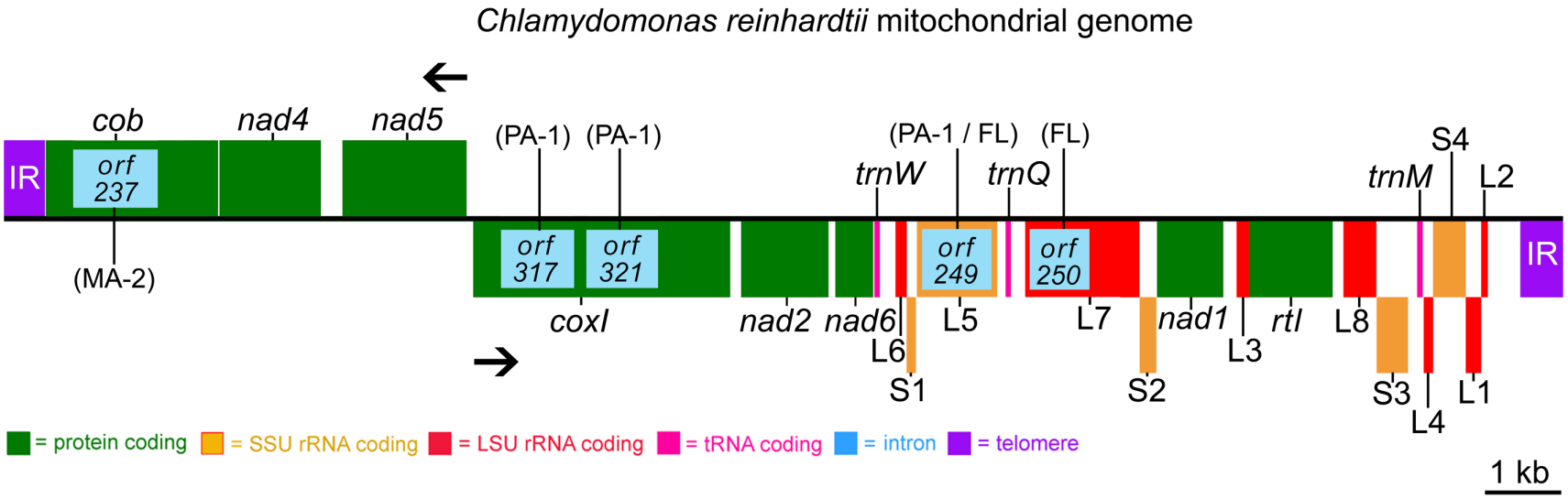

Figure 5.1. Genetic Map of the *C. reinhardtii* Mitochondrial Genome, Including All Currently Identified Optional Introns

Intronic regions and their open reading frames are boxed in blue inside their associated genes. The *C. reinhardtii* strains (Table 5.2) in which the different introns occur are labeled in parentheses. Solid arrows denote the transcriptional polarities. Note: due to the presence/absence of introns among the different strains, the size of the *C. reinhardtii* mitochondrial genome can vary from 15,782 nt to 18,990 nt.

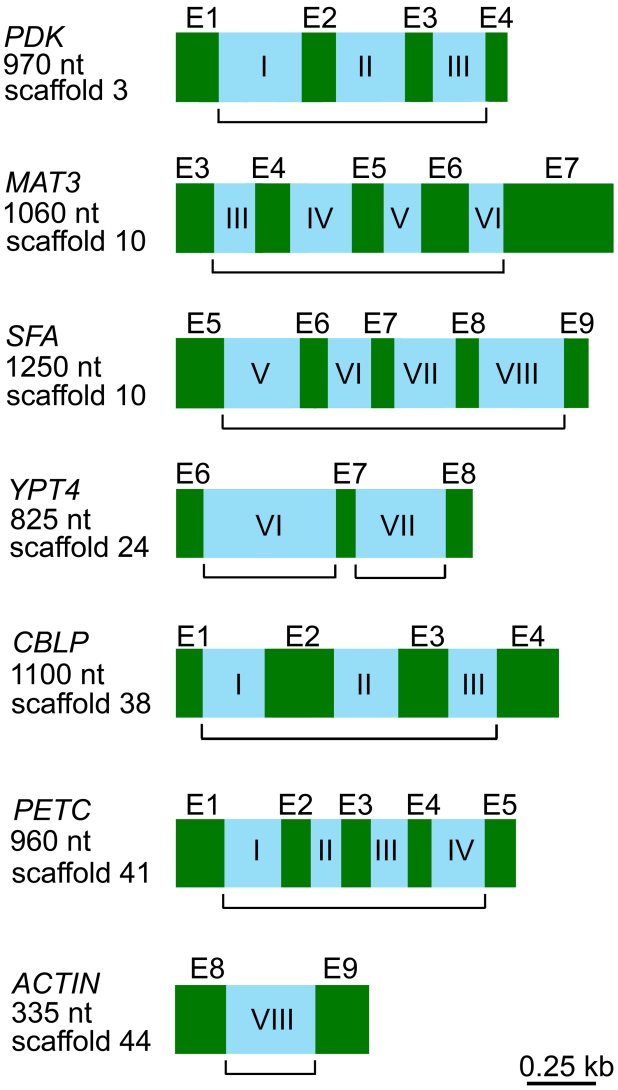

Figure 5.2. Partial Genetic Maps of the Seven *C. reinhardtii* Nuclear-encoded Genes Used for Measuring Nucleotide Diversity

The bracketed segment beneath each map represents the region that was PCR amplified. Left of each map is the name of the gene, the approximate size of the region that was PCR amplified, and the location of the gene within the *C. reinhardtii* nuclear genome — locations are based on the *C. reinhardtii* draft nuclear genome sequence version 3.0. Exons are green; they are labeled with an "E" and a number denoting their position within the gene. Introns are blue and are labeled with a roman numeral denoting their location within the gene. Note: each of these genes is present only once in the *C. reinhardtii* nuclear genome.

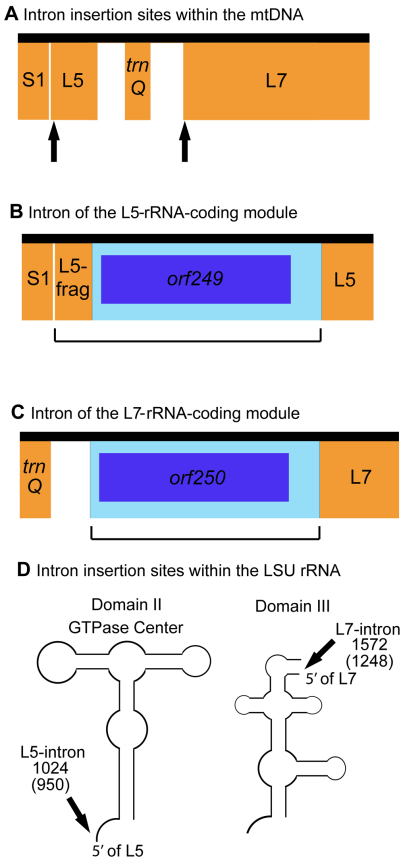

Figure 5.3. Schema of the Introns in the L5- and L7-rRNA-coding Modules

The vertical arrows in (A) show the intron insertion sites within the *C. reinhardtii* mtDNA. (B) and (C) depict the introns in the L5- and L7-rRNA-coding modules, respectively; rRNA- and tRNA-coding regions are orange; introns are light blue; intronic open reading frames are boxed in dark blue within their respective introns; L5-frag refers to a duplicated segment of the L5-rRNA-coding module (the first 35 nt of the module are duplicated); bracketed portions of the map represent regions that were shown to be spliced-out in mature transcripts. (D) depicts the intron insertion sites in the context of the large subunit (LSU) ribosomal RNA sequence of *C. reinhardtii*; arrows point to the region where the introns are inserted; numbers above the arrows denote the position of the residue that immediately precedes the insertion site: un-bracketed numbers correspond to the residue in the 23S rRNA gene of *Escherichia coli* (Cannone et al. 2002) and bracketed numbers correspond to the residue in the LSU-rRNA secondary-structure model of Boer and Gray (1988). Note: the *C. reinhardtii* strains in which these introns occur are shown in Figure 5.1.

CHAPTER 6: NUCLEOTIDE DIVERSITY IN THE *CHLAMYDOMONAS REINHARDTII* PLASTID GENOME: ADDRESSING THE MUTATIONAL-HAZARD HYPOTHESIS

Published as:

Smith DR, Lee RW (2009) Nucleotide diversity in the *Chlamydomonas reinhardtii* plastid genome: addressing the mutational-hazard hypothesis. BMC Evol Biol 9:120.

Abstract

Background: The mutational-hazard hypothesis argues that the noncoding-DNA content of a genome is a consequence of the mutation rate (μ) and the effective number of genes per locus in the population (N_g). The hypothesis predicts that genomes with a high $N_g\mu$ will have more compact genomes than those with a small $N_g\mu$. Approximations of $N_g\mu$ can be gained by measuring the nucleotide diversity at silent sites (π_{silent}). We addressed the mutation-hazard hypothesis apropos plastid genome evolution by measuring π_{silent} of the *Chlamydomonas reinhardtii* plastid DNA (ptDNA), the most noncoding-DNA-dense plastid genome observed to date. The data presented here in conjunction with previously published values of π_{silent} for the *C. reinhardtii* mitochondrial and nuclear genomes, which are respectively compact and bloated, allow for a complete analysis of nucleotide diversity and genome compactness in all three genetic compartments of this model organism. Results: In *C. reinhardtii*, the mean estimate of π_{silent} for the ptDNA (14.5×10^{-3}) is less than that of the nuclear DNA (32×10^{-3}) and greater than that of the mitochondrial DNA (8.5×10^{-3}). On average, *C. reinhardtii* has approximately four times more silent-site ptDNA diversity than the mean value reported for land plants, which have more compact plastid genomes. The silent-site nucleotide diversity of the different ptDNA loci that were studied varied significantly: from 0 to 71×10^{-3} for synonymous sites and from 0 to 42×10^{-3} for intergenic regions. Conclusion: Our findings on silent-site ptDNA diversity are inconsistent with what would be expected under the mutational-hazard hypothesis and go against the documented trend in other systems of π_{silent} positively correlating with genome compactness. Overall, we highlight the lack of reliable nucleotide diversity measurements for ptDNA and hope that the values presented here will act as sound data for future research concerning the mutational-hazard hypothesis and plastid evolution in general.

Introduction

The magnitude of noncoding DNA in genomes can differ dramatically both among and within evolutionary lineages. This statement holds true for prokaryotic genomes and for the nuclear, mitochondrial, and plastid genomes of eukaryotes. The

mutational-hazard hypothesis (Lynch 2007) asserts that much of this observed variation in genome compactness can be explained by the product of the effective genetic population size (represented in this study as the effective number of gene copies at a locus [N_g], not individuals) and the mutation rate (μ). The hypothesis maintains that an allele with more noncoding nucleotides than an alternative allele will be selectively disadvantageous because the excess noncoding DNA can accumulate hazardous mutations that may negatively impact gene function; the burden (or selective disadvantage) of the allele containing the surplus of noncoding DNA is determined by μ and the number of additional noncoding nucleotides in the larger allele that can affect gene function. The hypothesis proposes that natural selection is more efficient at perceiving the burden of the expanded allele when N_g is large; thus, genomes with a high $N_g\mu$ are predicted to be more compact than those with a small $N_g\mu$.

Population-genetic theory tells us that at mutation-drift equilibrium the nucleotide diversity at neutral sites ($\pi_{neutral}$) is equal to $2N_g\mu$ (where N_g of uniparentally inherited organelle genes is thought to be about half that of haploid nuclear genes [Birky et al. 1989]). Estimates of $\pi_{neutral}$ can be acquired by measuring the nucleotide diversity at silent sites (π_{silent}), which include noncoding sites and the synonymous sites of protein-coding DNA. Because there are many factors that can cause N_g to deviate from these neutral expectations, such as the influence of natural selection on linked variation, the only way to gain insight into $2N_g\mu$ is through empirical observation, i.e., by measuring π_{silent}.

As predicted by the mutational-hazard hypothesis, studies have found a positive correlation between π_{silent} and genome compactness: for the coding-rich DNA of prokaryotes π_{silent} is generally $> 50 \times 10^{-3}$; for the more noncoding-dense nuclear DNA (nucDNA) of land plants π_{silent} is in the range of 3×10^{-3} to 15×10^{-3}; and for the nuclear genomes of vertebrates, which abound with noncoding DNA, π_{silent} tends to be $\sim 3 \times 10^{-3}$ (Lynch and Conery 2003). Similar trends are also observed for mitochondrial genomes: in the streamlined mitochondrial DNA (mtDNA) of mammals π_{silent} is $\sim 40 \times 10^{-3}$, whereas that for land plant mtDNA, which is predominantly noncoding, is predicted to be $< 0.4 \times 10^{-3}$ (Lynch et al. 2006). The contrast in π_{silent} between mammalian and land plant mtDNA is thought to be a consequence of the high mutation rate in the former and the low mutation rate in the latter. Mutation rate has also been invoked to explain why,

despite similar proposed values of N_g, the mitochondrial and nuclear genomes of mammals have opposite coding densities — in mammals estimates of μ for mtDNA are roughly 30 times those for nucDNA (Lynch et al. 2006).

It is speculated that π_{silent} for plastid DNA (ptDNA) also correlates positively with genome compactness (Lynch et al. 2006; Lynch et al. 2007); however, this issue has not been formally addressed because there are very few ptDNA sequences for which both π_{silent} and genome-compactness data are available — we are aware of only two plastid genomes for which these two statistics are published: those of *Arabidopsis thaliana* and *Cycas taitungensis*; moreover, the silent-site diversities for these two genomes were derived in each case from only a single locus and, therefore, may have been unrepresentative because of a low sampling bias (see the supplementary material of Lynch et al. [2006]).

Of the 146 complete plastid-genome sequences available at the National Center for Biotechnology Information (NCBI) as of November 2008, the noncoding-DNA content ranges from 5%, in the apicomplexan *Eimeria tenella*, to 56%, in the unicellular green alga *Chlamydomonas reinhardtii* — a complete compilation is shown in Appendix D. Intriguingly, four of the five most bloated ptDNA sequences come from the Chlorophyta (a phylum containing most of the green-algal diversification), suggesting that this lineage is ideal for evaluating the mutational-hazard hypothesis vis-à-vis ptDNA. However, no studies as of yet have measured silent-site ptDNA diversity from the Chlorophyta. *C. reinhardtii*, a unicellular haploid alga, is a good candidate for investigating ptDNA diversity because it has a large (204 kilobases [kb]) and expanded plastid genome, and it is also a model organism for studying plastids and their photosynthetic processes (Harris 2001). From the viewpoint of the mutational-hazard hypothesis, π_{silent} in the *C. reinhardtii* ptDNA should be less than that of more compact organelle genomes.

A previous study on *C. reinhardtii* (Smith and Lee 2008b; Chapter 5) measured nucleotide diversity in its mitochondrial and nuclear genomes, which are respectively streamlined (~16-19 kb and ~20-30% noncoding, depending on the presence of optional introns) and bloated (~121 Megabases and ~83% noncoding). The mutational-hazard hypothesis would have forecasted π_{silent} for the mitochondrial genome to be greater than

that of the nuclear genome, but instead π_{silent} for the mtDNA was found to be four times smaller than that of the nucDNA (8.5×10^{-3} vs. 32×10^{-3}). Although these findings were in opposition to the mutational-hazard hypothesis, it was suggested that introns in the mtDNA impose a greater burden than those in the nuclear DNA and predicted that the same may be true for the mitochondrial intergenic regions (Smith and Lee 2008b; Chapter 5).

It would be interesting to see for *C. reinhardtii* how values of π_{silent} for the plastid genome compare to those of the mitochondrial and nuclear genomes. When considering the fraction of noncoding DNA in each of these genomes, the mutational-hazard hypothesis would predict π_{silent} for the ptDNA to be smaller than that of the mtDNA and larger than that of the nucDNA. But it is already known, as discussed above, that this is not the case: in *C. reinhardtii* the mtDNA has less silent-site diversity than the nucDNA. If the noncoding regions in the plastid genome carry an inflated burden, as suggested for those in the mtDNA, then we would expect a very low value of π_{silent} for the ptDNA, much smaller than that of the mtDNA (i.e., $<< 8.5 \times 10^{-3}$). However, if π_{silent} for the ptDNA is significantly larger than that of the mtDNA but still smaller than that of the nucDNA, it will be difficult to find any support in our data for the mutational-hazard hypothesis. In addition, silent-site ptDNA diversity data from *C. reinhardtii* will allow for a comparison of the π_{silent} values for the three genetic compartments of this species with those of *Arabidopsis lyrata*, the only other species for which reliable π_{silent} estimates from ptDNA, mtDNA, and nucDNA are published (Wright et al. 2008). Thus, to directly confront these issues, we measured π_{silent} from the ptDNA of various geographical isolates of *C. reinhardtii*.

Materials and Methods

The *C. reinhardtii* strains used in this study were obtained from the Chlamydomonas Center at Duke University. DNA was extracted from the same clonal isolate of each strain as used previously by Smith and Lee (2008b; Chapter 5) for studies on the nucleotide diversity of the *C. reinhardtii* mitochondrial and nuclear genomes. PtDNA was amplified by PCR using total genomic DNA as the template; the purified PCR products were sequenced on both strands. All of the ptDNA-sequence data

presented here were blasted against the *C. reinhardtii* draft nuclear genome sequence (v3.0) to insure that they are not nuclear-genome-located ptDNA-like sequences (NUPTs). Our blast results suggest that very few NUPTs are in the nuclear genome (<3 kb), and the few copies that are present are highly degenerate. Nucleotide diversity and its standard deviation were calculated with DnaSP 4.5 (Rozas et al. 2003) using the equations of Nei (1987), respectively. Two different methods for calculating silent-site nucleotide diversity were employed: one that excludes indels (indels-out), which was employed for calculating π_{silent}, and another that considers indels as polymorphic sites (indels-in), which was used for measuring $\pi_{silent+}$. For our estimates of $\pi_{silent+}$, indels involving more than one nucleotide were considered to be a single polymorphic site.

We acquired the complete plastid-genome sequence of *C. reinhardtii* strain CC-503 by assembling ptDNA sequences collected from the *C. reinhardtii* Whole Genome Shotgun Reads Trace Archive Database at GenBank. Blast hits showing >99% similarity to *C. reinhardtii* ptDNA were downloaded and assembled; all of the downloaded ptDNA sequences were subsequently blasted against the *C. reinhardtii* draft nuclear genome sequence (v3.0) to insure that no NUPTs were collected. Our assembly of the ptDNA data gave a complete CC-503 plastid genome with >50-fold coverage.

GenBank accession numbers for the ptDNA sequences produced in this study are FJ436944-FJ436977, FJ458164-FJ458275, and FJ423446; the latter number represents the CC-503 plastid-genome sequence.

Results

Strains and Their Genetic Loci

For our analysis we employed seven geographical isolates of *C. reinhardtii*, which are listed in Table 5.2 (Chapter 5). These are the same isolates that were previously used for calculating π_{silent} of the mtDNA and nucDNA. From each isolate, 14 distinct ptDNA regions were sequenced, amounting to 9.5 kb, 7.2 kb, and 2.7 kb of intergenic, protein-coding, and rRNA-coding ptDNA, respectively. A genetic map of the *C. reinhardtii* plastid genome highlighting these regions is shown in Figure 7.3 (Chapter 7).

We also produced a complete plastid-genome sequence for *C. reinhardtii* strain CC-503 by assembling ptDNA trace files generated by the *C. reinhardtii* nuclear-genome sequencing project (Merchant et al. 2007). The complete *C. reinhardtii* ptDNA sequence deposited at GenBank (accession number NC_005353) is a mosaic derived by linking the sequence data of various laboratory strains, most of which came from the "Ebersold-Levine" wild-type background of *C. reinhardtii* (Maul et al. 2002) — it is ideal to avoid using NC_005353 when calculating π_{silent} because sequence differences have been found between the ptDNA of some laboratory strains (Maul et al. 2002). A comparison of our CC-503 ptDNA sequence with NC_00533 reveals 471 single-nucleotide differences and 955 single-site indels; moreover, when the 14 ptDNA regions sequenced from the geographical isolates were sequenced from two additional laboratory strains belonging to the "Ebersold-Levine" wild-type background (CC-277 and CC-2454) the resulting data were identical to our CC-503-generated sequence but showed differences with NC_00533, suggesting that at least some of the discrepancies between CC-503 and NC_00533 are the result of sequencing errors in the latter. Thus, at present, the *C. reinhardtii* plastid genome sequence presented here appears to be the most accurate.

Twenty kilobases of intergenic ptDNA sequence data from an additional geographical isolate of *C. reinhardtii* (CC-2290) was obtained by data mining plastid sequences from GenBank (Table 6.1); because very little of these sequence data overlap with the 14 regions described above they were only compared to the ptDNA of CC-503.

Nucleotide Diversity

Nucleotide diversity measurements for the three genetic compartments of *C. reinhardtii* are summarized in Table 6.2. Net values of π_{silent} for the plastid genome are 14.5×10^{-3} when indels are removed from the alignment and 18.4×10^{-3} when indels are included and counted as polymorphisms ($\pi_{silent+}$); note, indels involving more than one nucleotide are considered to be a single polymorphic site. These values of π_{silent} and $\pi_{silent+}$ for the ptDNA are, respectively, 1.7 and 2 times those of the mtDNA, and 0.45 and 0.5 times those of the nucDNA. The nucleotide diversity values for the individual intergenic regions that were analyzed (outlined in Table 6.3) range from 0 to 41.6×10^{-3}

(average $\pi_{intergenic} = 11.3 \times 10^{-3}$), and the $\pi_{intergenic+}$ measurements for these same regions span from 0 to 53.2×10^{-3} (average $\pi_{intergenic+} = 14.4 \times 10^{-3}$). The synonymous-site nucleotide diversity of the different protein-coding genes that were sequenced varies from 0 to 71.1×10^{-3} (average $\pi_{syn} = 7.8 \times 10^{-3}$; Table 6.3). Relative to the mitochondrial and nuclear genomes, the ptDNA shows more variance in nucleotide diversity among different regions: $\pi_{intergenic}$ and π_{syn} of the various mtDNA loci range from 0 to 17.3×10^{-3} (average = 11.4×10^{-3}) and from 1.6×10^{-3} to 15.3×10^{-3} (average = 8.1×10^{-3}), respectively; and for the nucDNA, $\pi_{intergenic}$ varies from 21.6×10^{-3} to 58.3×10^{-3} (average = 36.1×10^{-3}) and π_{syn} extends from 2.8×10^{-3} to 41.1×10^{-3} (average = 20.9×10^{-3}). The ptDNA diversity of the rRNA-coding regions that were analyzed is 1.8×10^{-3}, which is slightly smaller than that of the mtDNA rRNA-coding regions (2.4×10^{-3}) — at present there are no nucleotide diversity data for rRNA-coding nucDNA.

The silent-site ptDNA diversity between CC-2290 (the strain from which ptDNA sequences were data mined) and CC-503 is 6.49×10^{-3} and $\pi_{silent+}$ is 18.78×10^{-3}; these values indicate that in the regions compared between CC-2290 and CC-503, single-site substitution differences are less frequent and indels are more frequent per site than in the regions compared in the group including CC-503 and the other six geographical isolates.

The various plastid-DNA loci were examined for traces of selection using Tajima's *D*-test (Table 6.3), which compares the average number of nucleotide differences between pairs of sequences (i.e., π) to the total number of segregating sites (*S*) (Tajima 1989). Tajima's *D* is positive for the protein-coding genes *atpA*, *cemA*, *psbA*, *rpoC2*, and *rpl2* and negative for *atpI*, *orf1995*, *rps9*, and *ycf3*. All of the analyzed intergenic regions show positive values for Tajima's *D*, with the exception of the *atpF-rps11* intergenic spacer, which has a negative *D* value. The only cases where Tajima's *D*-test is statistically significant are for the protein-coding gene *rpoC2* (Tajima's $D = 2.03$, P value < 0.05) and the region between the rRNA-coding genes *23S-1* and *23-2* (Tajima's $D = 2.10$, P value < 0.05).

Discussion

Accounting for the Observed Values of Π

At mutation-drift equilibrium, the nucleotide diversity at neutral sites should approximate $2N_g\mu$ (Lynch 2007); thus, an essential question of this study is: are the sites that we used to measure π_{silent} for the *C. reinhardtii* ptDNA neutrally evolving? We employed both noncoding sites and synonymous sites in our calculations of π_{silent}; these are generally considered to be among the more neutrally evolving positions in a genome. Indeed, the nucleotide diversity at these sites within the *C. reinhardtii* ptDNA exceeds that of the more functionally constrained positions, such as first and second codon positions and rRNA-coding sites. Among the different types of silent-sites, intergenic regions have ~1.8 times more nucleotide diversity than synonymous sites. Given that synonymous sites can be subject to selection for specific tRNA anticodons, one might expect them to be under more selective constraints than intergenic regions; therefore, it is not surprising that nucleotide diversity for the intergenic regions is greater than π_{syn}. Even so, because we sequenced more intergenic sites than synonymous sites, there is not a significant downward bias to our *C. reinhardtii* ptDNA-diversity measurements by including synonymous sites.

Another issue is the discrepancy in nucleotide diversity among the ptDNA loci that were studied. Factors that can result in inter-loci nucleotide diversity discrepancy include selection (e.g., balancing-, purifying-, or positive-selection) and inconsistencies in the mutation rate across the plastid genome; however, without interspecific ptDNA-divergence data, it would be overly speculative to focus on any one of these factors. Tajima's *D*-test did yield statistically significantly positive values for two of the loci that were studied, which could be an indication of balancing selection. It is noteworthy that the magnitude of variation among the *C. reinhardtii* ptDNA loci is significantly more pronounced than what is typically observed for ptDNA: the nucleotide diversity of most plastid genomes appears to be relatively homogeneous across loci (Wu et al. 2006; Wright et al. 2008). On the other hand, studies indicate that ptDNA substitution rates at

both synonymous and intergenic sites can vary considerably among loci within a genome (Wolfe et al. 1987; Shaw et al. 2005; Guisinger 2008).

It would be ideal if we could interpret our ptDNA nucleotide diversity measurements in relation to μ and N_g, but this is difficult because the mutation rate for the *C. reinhardtii* plastid genome is unknown. There is evidence that μ for the mtDNA and nucDNA of *C. reinhardtii* are approximately the same (Popescu and Lee 2007), and consequently the disparity of π_{silent} between these genomes can be explained by differences in N_g (see Popescu and Lee [2007] for a more detailed discussion). Other things being equal, in *C. reinhardtii* we would expect N_g of the uniparentally inherited plastid genome to be about the same as that of the mitochondrial genome, which is also uniparentally inherited, and about half that of the nuclear genome. Uniparental inheritance also implies that the organelle DNA has less opportunity for recombination as a result of sexual reproduction as compared with the nucDNA (Birky et al. 1989), meaning that it may be more prone to the influences of natural selection on linked variation (i.e., genetic hitch-hiking), which can cause $N_{g(organelle)}$ to deviate from neutral expectations (e.g., Bazin et al. [2006]). Nevertheless, the only study to seriously investigate this issue with respect to the ptDNA, mtDNA, and nucDNA from a single species, *A. lyrata*, found that N_g of the organelle DNA and nucDNA did not depart significantly from what was expected under neutrality (Wright et al. 2008). Thus, the fact that silent-site nucleotide diversity in *C. reinhardtii* ptDNA is only within a factor of two from that of the mtDNA and nucDNA can easily be accounted for by slight differences in μ and/or N_g.

Plastid DNA Diversity for C. reinhardtii *Relative to that of Other Taxa*

There is a paucity of nucleotide diversity data from ptDNA, and the estimates that are published are limited to a small number of model land plant species. Most of these available estimates are listed in the supplementary material of Lynch et al. (2006) who compiled a summary of silent-site ptDNA diversity values from 17 land plant species and found that on average π_{silent} is 3.7×10^{-3}, with a standard error of 1.1×10^{-3} — most of these diversity data were calculated using an indels-out approach but some were generated with the indels-in method (e.g., Huang et al. [2001]). More recently published

π_{silent} estimates from the ptDNA of land plants are concordant with these values: 0-1.2 x 10^{-3} (*Rhododendron* spp.) (Chung et al. 2006), ~4 x 10^{-3} (*Machilus* spp.) (Wu et al. 2006), and ~2 x 10^{-3} (*Silene* spp.) (Muir and Filatov 2007). In comparison, the silent-site ptDNA diversity of *C. reinhardtii* is four times the mean estimate for land plants (14.5 x 10^{-3} vs. 3.7 x 10^{-3}). The average π_{silent} estimates from the mtDNA and nucDNA of land plants are, respectively, 0.4 x 10^{-3} and 15.2 x 10^{-3} (Lynch and Conery 2003; Lynch et al. 2006). Thus, when considering all three genetic compartments, the π_{silent} values from *C. reinhardtii* match the general trend observed in land plants, with silent-site nucleotide diversity being intermediate for the plastid genome, lowest for the mitochondrial genome, and highest for the nuclear genome; however, there is an overall increase of silent-site diversity for *C. reinhardtii*, in all three of its genomes, relative to that of land plants.

To the best of our knowledge, the only species, heretofore, for which nucleotide diversity data are available from all three genetic compartments is *Arabidopsis lyrata* (Wright et al. 2008): values of π_{silent} for the ptDNA, mtDNA, and nucDNA are 1.0 x 10^{-3}, 0.35 x 10^{-3}, and 20 x 10^{-3}, respectively. Therefore, silent-site diversity in the *A. lyrata* ptDNA is 3 times that of the mtDNA and 0.05 times that of the nucDNA. Again, the same general trend is observed for *C. reinhardtii* but with a less dramatic difference between the silent-site diversity of the organelle DNA versus that of the nucDNA.

Addressing the Mutational-hazard Hypothesis

Contrary to what the mutational-hazard hypothesis forecasted, the π_{silent} data for the three genetic compartments of *C. reinhardtii* do not positively correlate with genome compactness. In fact, the opposite trend is observed, with silent-site diversity being lowest for the compact mitochondrial genome (8.5 x 10^{-3}), greatest for the bloated nucDNA (32.3 x 10^{-3}), and intermediary for the plastid genome (14.5 x 10^{-3}), which has a noncoding-DNA density that is halfway between the mtDNA and nucDNA.

Due to a lack of available data, it is difficult for us to compare π_{silent} and genome-compactness values of the *C. reinhardtii* ptDNA with those of other plastid genomes; we are aware of only two ptDNA sequences for which both these data are published: those of *Arabidopsis thaliana* ($\pi_{silent(ptDNA)}$ = 1.4 x 10^{-3}; 41% noncoding) and *Cycas taitungensis* ($\pi_{silent(ptDNA)}$ = 12.8 x 10^{-3}; 37% noncoding) (Lynch et al. 2006). Based on their relative

fractions of noncoding ptDNA, the mutational-hazard hypothesis would forecast *A. thaliana* and *C. taitungensis* to have more silent-site ptDNA diversity than *C. reinhardtii,* but instead they have less. However, it is important to stress that the π_{silent} values for the *A. thaliana* and *C. taitungensis* ptDNA are derived, in each case, from only a single locus (one protein-coding gene and one intergenic region), and, therefore, may be biased because of insufficient sampling.

If we assume that the mean π_{silent} estimate of land plant ptDNA (3.7×10^{-3}), derived by Lynch et al. (2006), is representative of the silent-site ptDNA diversity in land plants for which plastid-genome-compactness values are available (i.e., those with completely sequenced plastid genomes), then, based on the noncoding-DNA densities (Appendix D) the mutational-hazard hypothesis would predict less silent-site diversity for the *C. reinhardtii* ptDNA relative to the more coding-rich plastid genomes of land plants; however, *C. reinhardtii* appears to have four times more silent-site ptDNA diversity than the mean estimate for land plants.

Let us now compare the π_{silent} and genome-compactness measurements of the *C. reinhardtii* ptDNA to those of animal mtDNA — the only organelle genomes for which these data are readily available. As highlighted earlier, the size and noncoding-DNA density of the *C. reinhardtii* plastid genome is significantly larger than that of animal mitochondrial genomes, but contrary to what would be predicted under the mutational-hazard hypothesis, the silent-site diversity of animal mtDNA is not dramatically greater than that of the *C. reinhardtii* ptDNA. Although reported π_{silent} values for animal mitochondrial genomes can be as high as $\sim 67 \times 10^{-3}$ (nematodes), those for arthropods ($\sim 27 \times 10^{-3}$), birds ($\sim 17 \times 10^{-3}$), echinoderms ($\sim 11.7 \times 10^{-3}$), and mollusks ($\sim 13.5 \times 10^{-3}$) are between 0.8-1.9 times π_{silent} reported here for the *C. reinhardtii* ptDNA, which is reasonably close considering the stark contrast in genome architectures.

Of the 114 kb of noncoding nucleotides in the *C. reinhardtii* plastid genome, <2 kb represent intronic DNA — the remainder are intergenic DNA. Why have intergenic nucleotides proliferated in the *C. reinhardtii* plastid genome when intronic DNA has been kept at bay? One hypothesis could be that in the plastid genome of this taxon the encumbrance of intergenic DNA is significantly less than that of intronic DNA. Recall, that under the mutational-hazard hypothesis, the disadvantage of harboring noncoding

DNA is dependent on the: 1) number of noncoding nucleotides associated with gene function (n); 2) per-nucleotide mutation rate (μ); and 3) effective number of genes per locus in the population (N_g) — where the overall encumbrance of noncoding DNA is a product of $N_g\mu n$. By measuring nucleotide diversity we were able to approximate $2N_g\mu$; however, n is more difficult to estimate. For organelle introns n is believed to be relatively large, perhaps as high as 100 per intron (Lang et al. 2007), but n for organelle intergenic regions is generally unknown. One might ask, is there any reason to believe that intergenic DNA in the *C. reinhardtii* plastid genome carries a reduced burden (i.e., has fewer sites that are crucial for gene function relative to other plastid genomes)? In regards to this question, two observations are worth noting. In land plants, chloroplast genes are organized into operons, which are first transcribed into polycistronic primary transcripts and then subsequently processed into mature monocistronic units via endo- and exonucleolytic cleavage (Hudson et al. 1987; Barkan 1988; Haley and Bogorad 1990). In *C. reinhardtii*, however, most chloroplast genes appear to be transcribed into monocistronic (or in some cases dicistronic) transcripts (Sakamoto et al. 1994; Bruik and Mayfield 1998; Jiao et al. 2004). Although speculative, it is possible that the intergenic DNA in the *C. reinhardtii* plastid genome carries a reduced burden (small n) relative to that of land plant ptDNA — a mutation in the intergenic DNA of land plant ptDNA could affect the expression of many genes by interfering with transcriptional or posttranscriptional steps, an outcome that seems less likely for the *C. reinhardtii* ptDNA, which has a preponderance of monocistronically expressed genes. A final comment is that in *C. reinhardtii*, genes in the mtDNA, unlike those in the ptDNA, show extensive transcriptional linkage (Gray and Boer 1988) and although our estimates of $2Ng\mu$ for the mitochondrial genome are low, the intergenic regions are reduced in size, which may imply that n for mitochondrial intergenic DNA is relatively large.

Conclusion

The primary goal of this study was to measure nucleotide diversity for the ptDNA of *C. reinhardtii* and by doing so investigate a novel theory regarding genome evolution — the mutational-hazard hypothesis. Ultimately, the results presented in this study go

against the documented trend of π_{silent} positively correlating with genome compactness, and thus challenge the central premise of the mutational hazard hypothesis.

Table 6.1. GenBank Accession Numbers for the PtDNA Sequences Data-mined from *C. reinhardtii* strain CC-2290

GenBank Accession Number					
CACW10301.b1	CACW10301.g1	CACW10641.b1	CACW10641.g1	CACW11480.g1	CACW14215.b1
CACW14215.b2	CACW14215.g1	CACW14215.g2	CACW1919.b1	CACW1919.g1	CACW22128.b1
CACW23575.b1	CACW23575.g1	CACW23602.b1	CACW23882.b1	CACW24643.b1	CACW24643.g1
CACW24803.b1	CACW25840.b1	CACW25840.g1	CACW25856.b1	CACW25856.g1	CACW2634.b1
CACW2634.g1	CACW2890.b1	CACW2890.g1	CACW6507.b1	CACW6507.b2	CACW6507.g1
CACW6507.g2	CACW7106.g1	CCW7106.b1	CACW7225.b1	CACW7225.g1	CACW8237.b1
CACW8237.g1	CACW8272.b1	CACW8272.g1	CACW8423.b1	CACW8423.g1	CACW11480.b1
CACW22128.b1	CACW22128.g1	CACW23882.b1	CACW23882.g1		

Table 6.2. Nucleotide Diversity for the Plastid, Mitochondrial, and Nuclear Genomes of *C. reinhardtii*

	Protein-coding regions			Intronic/intergenic regions[e]			Silent sites[f]		
	ptDNA	mtDNA	nucDNA	ptDNA	mtDNA	nucDNA	ptDNA	mtDNA	nucDNA
# of sites[a]	7272	8160	1623	9438	2457	4510	16710	5550	5051
S	45	44	26	276	58	355	321	104	377
# of Indels[b]	1	1	1	85	9	47	86	11	48
(length nt)	(21)	(6)	(9)	(1672)	(23)	(216)	(1679)	(31)	(222)
π^c x 10^{-3}	2.82	2.06	6.02	15.17	8.92	33.50	14.53	8.51	32.29
(SD x 10^{-3})	(0.34)	(0.43)	(0.99)	(1.70)	(1.88)	(3.15)	(1.18)	(1.03)	(3.01)
π_+^d x 10^{-3}	---	---	---	19.54	10.29	38.63	18.36	9.23	36.00
(SD x 10^{-3})				(2.03)	(2.02)	(3.70)	(1.40)	(1.96)	(3.51)
π_{syn} x 10^{-3}	8.46	8.52	19.57	---	---	---	---	---	---
π_{nsyn} x 10^{-3}	1.14	0	1.42	---	---	---	---	---	---

Note: *S,* number of segregating (i.e., polymorphic) sites; Indels, insertion-deletion events; π, nucleotide diversity; π_+, nucleotide diversity including both polymorphic sites and insertion-deletion events; π_{syn}, nucleotide diversity at synonymous sites; π_{nsyn}, nucleotide diversity at nonsynonymous sites; SD, standard deviation. Mitochondrial- and nuclear-genome data come from (Smith and Lee 2008b; Chapter 5).

[a] Comprises all sites in the nucleotide alignment, including those with indels.

[b] Indels involving more than 1 nucleotide are counted as a single event. Indel length includes the sum of all indels and includes consecutive indel events.

[c] Only includes sites in the alignment without indels.

[d] Considers all sites, including those with indels. Consecutive indels are counted as a single polymorphic event. Indel states were measured using a multiallelic approach.

Table 6.3. Nucleotide Diversity (by Region) in the *C. reinhardtii* Plastid Genome

	# of sites[a]	*S*	# of Indels[b] (length nt)	π^c x 10^{-3} (SD x 10^{-3})	$\pi_+{}^d$ x 10^{-3} (SD x 10^{-3})	π_{syn} x 10^{-3}	π_{nsyn} x 10^{-3}	Tajima's *D*-Test (P value)
PROTEIN-CODING (by gene)								
atpA	501	18	0	17.30 (2.67)	---	71.08	0	0.06 (>0.1)
atpF	213	0	0	0	---	0	0	---
atpI	366	3	0	3.51 (1.43)	---	13.85	0	-1.01 (>0.1)
cemA	462	2	0	2.27 (4.40)	---	5.34	1.34	1.17 (>0.1)
orf1995	1896	8	0	1.41 (0.56)	---	2.57	1.09	-0.96 (>0.1)
petA	954	0	0	0	---	0	0	---
petG	45	0	0	0	---	0	0	---
psaJ	126	0	0	0	---	0	0	---
psbA	435	8	0	9.61 (1.46)	---	36.5	1.7	1.48 (>0.1)
psbK	72	0	0	0	---	0	0	---
rpoC2	252	6	0	13.61 (2.85)	---	10.85	14.33	2.03 (<0.05)
rpl2	321	2	0	3.56 (0.74)	---	7.28	2.36	1.64 (>0.1)
rps2	87	0	0	0	---	0	0	---
rps3	126	0	0	0	---	0	0	---
rps9	462	4	1 (21)	3.50 (1.01)	---	9.78	1.61	-0.04 (>0.1)
rps11	105	0	0	0	---	0	0	---
rps12	321	0	0	0	---	0	0	---
rps19	183	0	0	0	---	0	0	---
tufA	213	0	0	0	---	0	0	---
ycf3	315	2	0	1.81 (1.25)	---	3.84	1.19	-1.28 (>0.1)
ycf4	399	0	0	0	---	0	0	---

	# of sites[a]	*S*	# of Indels[b] (length nt)	π^c x 10^{-3} (SD x 10^{-3})	π_+^d x 10^{-3} (SD x 10^{-3})	π_{syn} x 10^{-3}	π_{nsyn} x 10^{-3}	Tajima's *D*-Test (P value)
INTERGENIC								
(by region)								
atpA/psbI	343	1	0	1.75 (0.51)	1.75 (0.51)	---	---	1.22 (>0.1)
atpF/rps11	1556	86	16 (368)	25.61 (9.77)	30.10 (11.15)	---	---	-1.08 (>0.1)
atpI/psaJ	328	9	0	12.49 (2.42)	12.49 (2.42)	---	---	0.61 (>0.1)
petG/rps3	805	5	3 (3)	2.97 (0.44)	4.50 (0.76)	---	---	0.57 (>0.1)
psaJ/rps12	310	2	0	3.38 (0.65)	3.38 (0.65)	---	---	1.17 (>0.1)
psbK/tufA	828	4	2 (117)	1.18 (0.37)	3.47 (0.53)	---	---	0.06 (>0.1)
psbI/cemA	271	3	2 (2)	5.20 (1.18)	8.86 (2.00)	---	---	0.00 (>0.1)
rpl2/rps19	808	36	9 (164)	27.36 (5.14)	33.69 (5.70)	---	---	0.99 (>0.1)
rps3/rpoC2	1474	88	32 (462)	41.60 (6.65)	53.19 (9.01)	---	---	0.71 (>0.1)
rps9/ycf4	344	1	1 (21)	1.65 (0.53)	3.29 (0.69)	---	---	1.03 (>0.1)
rps18/rps2-1	1115	45	19 (418)	30.47 (9.57)	42.17 (11.33)	---	---	0.90 (>0.1)
ycf3/ycf4	179	0	0	0	0	---	---	---
23S-1/23S-2	909	6	2 (46)	3.97 (0.67)	5.28 (0.89)	---	---	2.10 (<0.05)
23S-2/5S	89	0	0	0	0	---	---	---

Note: *S,* number of segregating (i.e., polymorphic) sites; Indels, insertion-deletion events; π, nucleotide diversity; π_+, nucleotide diversity including both polymorphic sites and insertion-deletion events; π_{syn}, nucleotide diversity at synonymous sites; π_{nsyn}, nucleotide diversity at nonsynonymous sites; SD, standard deviation.

[a] Comprises all sites in the nucleotide alignment, including those with indels.

[b] Indels involving more than 1 nucleotide are counted as a single event. Indel length includes the sum of all indels and includes consecutive indel events.

[c] Only includes sites without indels.

[d] Considers all sites, including those with indels. Consecutive indels are counted as a single polymorphic event. Indel states were measured using a multiallelic approach.

CHAPTER 7: LOW NUCLEOTIDE DIVERSITY FOR THE EXPANDED ORGANELLE AND NUCLEAR GENOMES OF *VOLVOX CARTERI* SUPPORTS THE MUTATIONAL-HAZARD HYPOTHESIS

Published as:

Smith DR, Lee RW (2010) Low nucleotide diversity for the expanded organelle and nuclear genomes of *Volvox carteri* supports the mutational-hazard hypothesis. Mol Biol Evol 27:1-13.

Abstract

The noncoding-DNA content of organelle and nuclear genomes can vary immensely. Both adaptive and non-adaptive explanations for this variation have been proposed. This study addresses a non-adaptive explanation called the mutational-hazard hypothesis and applies it to the mitochondrial, plastid, and nuclear genomes of the multicellular green alga *Volvox carteri*. Given the expanded architecture of the *V. carteri* organelle and nuclear genomes (60-85% noncoding DNA), the mutational-hazard hypothesis would predict them to have less silent-site nucleotide diversity (π_{silent}) than their more compact counterparts from other eukaryotes — ultimately reflecting differences in $2N_g\mu$ (twice the effective number of genes per locus in the population times the mutation rate). The data presented here support this prediction: analyses of mitochondrial, plastid, and nuclear DNAs from seven *V. carteri* forma *nagariensis* geographical isolates reveal low values of π_{silent} (0.00038, 0.00065, and 0.00528, respectively), much lower values than those previously observed for the more compact organelle and nuclear DNAs of *Chlamydomonas reinhardtii* (a close relative of *V. carteri*). We conclude that the large noncoding-DNA content of the *V. carteri* genomes is best explained by the mutational-hazard hypothesis, and speculate that the shift from unicellular to multicellular life in the ancestor that gave rise to *V. carteri* contributed to a low *V. carteri* population size and thus a reduced $2N_g\mu$. Complete mitochondrial- and plastid-genome maps for *V. carteri* are also presented and compared with those of *C. reinhardtii*.

Introduction

A striking observation to come from the genomics era is that the amount of noncoding DNA (defined here as intronic and intergenic DNA) in eukaryotic nuclear genomes varies by five orders of magnitude (Lynch and Conery 2003; Gregory 2005a). Similar findings are also observed for organelle genomes (Palmer 1991; Burger et al. 2003; Lynch et al. 2006), for which the noncoding-DNA content can range from 1-95% for mitochondrial DNA (mtDNA) and from 5-80% for plastid DNA (ptDNA) (Figure 7.1). The reason(s) why some eukaryotic genomes abound with seemingly useless DNA while others are paragons of compactness has proven to be a difficult question to answer,

a question that has been dubbed the "C-value enigma" (Gregory 2001), replacing the earlier "C-value paradox", which was concerned with the discordance between genome size and organismal complexity (Mirsky and Ris 1951; Thomas 1971).

There are many hypotheses that confront the C-value enigma, for reviews see Gregory (2005b), Bennett and Leitch (2005) and Lynch (2007). Some suggest that noncoding DNA is the direct product of natural selection (e.g., Cavalier-Smith 1982), whereas others, such as the selfish-DNA hypothesis (Doolittle and Sapienza 1980; Orgel and Crick 1980), maintain that noncoding DNA arises primarily through neutral processes. This study addresses a contemporary, non-adaptive hypothesis for the evolution of noncoding DNA called the mutational-hazard hypothesis (Lynch and Conery 2003; Lynch et al. 2006; Lynch 2006) and applies it to the mitochondrial, plastid, and nuclear genomes of the multicellular green alga *Volvox carteri* forma *nagariensis* (Chlorophyceae, Chlorophyta).

The mutational-hazard hypothesis is based on the premise that noncoding DNA is a mutational liability (i.e., more noncoding DNA means more chances for harmful mutations, such as those causing gene-expression problems). The hypothesis argues that "the tendency for mutationally-hazardous DNA to accumulate depends on both the population size and the mutation rate: the latter defines the burden of excess DNA, while the former defines the ability of natural selection to eradicate it" (Lynch 2007, p. 40). Therefore, organisms with large effective population sizes and high mutation rates are predicted to have more compact genomes than those with small effective population sizes and low mutation rates.

Insights into effective population size (represented in this study as the effective number of gene copies at a locus [N_g], not individuals) and mutation rate (μ; defined as the number of mutations per nucleotide site per generation) can be acquired by measuring the nucleotide diversity of silent sites (π_{silent}), which include intergenic and intronic sites and the synonymous sites of protein-coding DNA. Population-genetic theory tells us that at mutation-drift equilibrium the nucleotide diversity at neutral sites ($\pi_{neutral}$) is equal to $2N_g\mu$ [where N_g of uniparentally inherited organelle genes is thought to be about half that of haploid nuclear genes (Birky et al. 1989)]. Because silent sites are typically regarded as among the most neutrally evolving positions in a genome, measures of π_{silent} can

provide an estimate of $2N_g\mu$.

As predicted by the mutational-hazard hypothesis, large-scale studies have found a positive correlation between π_{silent} and genome compactness, with data sets that include bacterial, organelle, and nuclear DNAs (Lynch and Conery 2003; Lynch et al. 2006; Lynch 2006). Opponents of the mutational-hazard hypothesis have argued that the data from which these correlations were derived are weak (Daubin and Moran 2004), and some studies have found that the connection between π_{silent} and genome compactness does not hold up for some organisms or for some genetic compartments (Gregory and Witt 2008; Smith and Lee 2008b, 2009a; Chapters 5 and 6). However, a general lack of available π_{silent} data has prevented a thorough investigation of the mutational-hazard hypothesis, especially in relation to non-metazoan eukaryotic species. Moreover, the π_{silent} data that are available are often derived from a single locus, and only a few studies have measured π_{silent} for the organelle and nuclear DNAs of the same species.

Already a model for studying the origins of multicellularity and cellular differentiation (Kirk 1998), *V. carteri* is also an ideal species for addressing the mutational-hazard hypothesis. This sexually active alga has three distinct genomes (a haploid nuclear genome and uniparentally inherited mitochondrial and plastid genomes [Adams et al. 1990]) all of which contain an abundance of noncoding DNA. Previously, we presented partial mtDNA and ptDNA sequences for *V. carteri* and showed that both of these genomes abound with noncoding DNA (~60-80%) and are the most inflated organelle DNAs available from the Chlorophyta (Smith and Lee 2009b; Appendix E). Similarly, the *V. carteri* nuclear genome, sequenced in 2007 by the United States Department of Energy Joint Genome Institute (DOE JGI), is ~140 megabases (Mb), ~85% noncoding, and has an average intron-to-gene density of 7.8, placing it among the most bloated nuclear DNAs (nucDNA) observed from a photosynthetic protist. Indeed, when nucDNA-encoded genes from *V. carteri* are compared to their homologs from land plants and animals, *V. carteri* generally has a greater intron/gene ratio (Kirk 1998).

Given the expanded architecture of the *V. carteri* organelle and nuclear genomes, the mutational-hazard hypothesis would predict them to have less silent-site nucleotide diversity than their more compact counterparts from other eukaryotes — ultimately reflecting differences in $2N_g\mu$. Although data on π_{silent} are limited, especially for green

algae, we recently measured π_{silent} from the mitochondrial, plastid, and nuclear compartments of the unicellular green alga *Chlamydomonas reinhardtii* (Smith and Lee 2008b, 2009a; Chapters 5 and 6), a close relative of *V. carteri.* Our π_{silent} data for *C. reinhardtii* were difficult to interpret with respect to the mutational-hazard hypothesis because no other reliable π_{silent} estimates from green algae were available for comparison. Nucleotide-diversity data from *V. carteri* would allow for a more thorough evaluation of the forces affecting genome architecture in these two model species. *V. carteri* and *C. reinhardtii* make a good pair for investigating the evolution of noncoding DNA because although both species have similar organelle- and nuclear-DNA gene numbers, the genomes of *V. carteri* have more noncoding DNA than those of *C. reinhardtii* (Merchant et al. 2007; Smith and Lee 2009b; Appendix E; and see the *V. carteri* DOE JGI nuclear-genome portal: http://genome.jgi-psf.org/Volca1/Volca1.info.html). This is especially true for the mitochondrial and plastid genomes of *V. carteri*, which are more than twice the size of their *C. reinhardtii* counterparts.

Here we investigate the correlation between $2N_g\mu$ and genome compactness by presenting π_{silent} estimates for the mitochondrial, plastid, and nuclear genomes from seven geographical isolates of *V. carteri* f. *nagariensis*. These data are compared with our previously-published π_{silent} values for *C. reinhardtii* and then placed in a broad phylogenetic context. Moreover, by building upon our earlier work (Smith and Lee 2009b; Appendix E), we report complete mitochondrial- and plastid-genome maps for *V. carteri* (UTEX 2908).

Materials and Methods

Strains and Culture Conditions

The *V. carteri* strains employed in this study (Table 7.1) were obtained from either the Culture Collection of Algae at the University of Texas at Austin (UTEX) or the Microbial Culture Collection at the National Institute for Environmental Studies (NIES) in Ibaraki, Japan, with the exception of the *V. carteri* Isanuma male and female strains, which were graciously provided to us by Stephen Miller of the University Maryland, Baltimore County. The *V. carteri* strains were grown in Volvox medium (Provasoli and

Pintner 1960) at 28°C under a 16-h light/8-h dark cycle; the illumination at 35 µmol photons•m^{-2}•s^{-1} photosynthetically-active radiation (PAR) was provided by "cool-white" fluorescent bulbs.

DNA Extraction, Amplification, and Sequencing

For each *V. carteri* strain, total genomic DNA was extracted using the DNeasy Plant Mini Kit (Qiagen, Germantown, MD) following the manufacturer's protocol. The mtDNA, ptDNA, and nucDNA sequences presented in this study were amplified using a PCR-based approach. PCR reactions were performed with High Fidelity Platinum SuperMix (Invitrogen, Carlsbad, CA) using total genomic DNA as the template. PCR products corresponding to regions with extensive DNA repeats were cloned using the TOPO TA Cloning Kit (Invitrogen). Purified PCR products and isolated plasmids were sequenced on both strands using a 3730xl DNA Analyzer (Applied Biosystems, Foster City, CA) at the Macrogen Sequencing Facility in Rockville, MD.

Sequencing and Assembling the V. carteri *(UTEX 2908) Mitochondrial and Plastid Genomes*

The complete mitochondrial- and plastid-genome maps for *V. carteri* strain UTEX 2908 were generated using a two-pronged approach. First, mtDNA and ptDNA trace files generated by the DOE JGI *V. carteri* nuclear-genome sequencing project (http://genome.jgi-psf.org/Volca1/Volca1.home.html) were data mined from the National Center for Biotechnology Information (NCBI) *V. carteri* Trace Archive (http://www.ncbi.nlm.nih.gov/Traces/) using *V. carteri* organelle-DNA sequences (Smith and Lee 2009b; Appendix E) as BLAST (blastn 2.2.21+) queries. The BLAST parameters were as follows: an expectation value (E-value) of 10; a word size of 11; match and mismatch scores of 2 and -3, respectively; and gap-cost values of 5 (existence) and 2 (extension). Trace files showing >90% sequence identity to *V. carteri* organelle DNA in BLAST alignments were downloaded and then scanned for trace files originating from inserts (i.e., shotgun-sequencing fragments) with lengths greater than 7 kilobases (kb). Trace files from large inserts were pooled and used to arrange in the proper orientation the mtDNA and ptDNA contigs generated from our previous work (Smith and Lee

2009b; Appendix E). Long-range PCR was then used to fill the gaps between the organelle-DNA contigs. PCR reactions were performed with the LongRange PCR Kit (Qiagen) using total genomic DNA from *V. carteri* UTEX 2908 as the template. Sequences were assembled using CodonCode Aligner Version 3.0.1 (CodonCode Corporation, Dedham, MA), which employs the Phred, Cross-match, and Phrap algorithms for base calling, sequence comparison, and sequence assembly, respectively. Assemblies were performed with a minimum-percent-identity score of 98, a minimum-overlap length of 500 nucleotides (nt), a match score of 1, a mismatch penalty of -2, a gap penalty of -2, and an additional first-gap penalty of -3.

DNA Sequence Analyses

Nucleotide diversity and its standard deviation were calculated with DnaSP 4.5 (Rozas et al. 2003). Two different methods for calculating silent-site nucleotide diversity were employed: one that excludes indels (indels out), which was used for calculating π_{silent}, and another that considers indels as polymorphic sites (indels in), which was used for measuring $\pi_{silent+}$. For our estimates of $\pi_{silent+}$, indels involving more than one nucleotide were considered to be a single polymorphic site. To ensure that the six nucDNA-encoded genes employed in this study are each present only once in the *V. carteri* nuclear genome, we blasted (blastn 2.2.21+) the six sequences against the DOE JGI *V. carteri* draft nuclear genome sequence (v.1.0. [http://genome.jgi-psf.org/Volca1/Volca1.home.html]). All six genes returned a single hit, which is consistent with the hypothesis that these genes are in single copies in the *V. carteri* nuclear genome. The mtDNA and ptDNA sequences obtained from each of the seven strains were also blasted against the *V. carteri* draft nuclear genome sequence to confirm that they were not generated from nuclear-genome-located organelle-DNA-like sequences (NUMTs or NUPTs).

The Fraction of Noncoding DNA in Completely Sequenced Organelle Genomes

Completely sequenced organelle genomes were downloaded from the NCBI Reference Sequence (RefSeq) collection (http://www.ncbi.nlm.nih.gov/projects/RefSeq/) on 1 June 2009. The coding- and noncoding-DNA contents of these sequences were

calculated using the following methods and definitions: i) the number of coding nucleotides in the genome is equal to the collective length of all annotated protein-, rRNA-, and tRNA-coding regions, not including the portions of these regions that are also annotated as introns; ii) the amount of noncoding DNA is the genome length minus the number of coding nucleotides; iii) the number of intergenic nucleotides is equal to the genome length minus the collective length of regions annotated as genes (including their introns and intronic ORFs); and iv) the amount of intronic DNA is equivalent to the number of noncoding nucleotides minus the number of intergenic nucleotides. The above methods and definitions are contingent on the authors of the GenBank records having properly annotated their entry. If coding regions or introns have been ignored or inaccurately annotated this will result in incorrect coding- and noncoding-DNA content values. All records were quickly scanned for major errors, but due to the large number of organelle genomes deposited in GenBank, it was unfeasible to thoroughly review every record.

V. carteri *Nuclear-genome Statistics*

Length, noncoding-DNA-content, and average-intron-density estimates for the *V. carteri* nucDNA came from the DOE JGI *V. carteri* genome portal information page (accessed on 1 December 2009): http://genome.jgi-psf.org/Volca1/Volca1.home.html.

Nucleotide Sequence Accession Numbers

The mtDNA-, ptDNA- and nucDNA-sequence data that were used to measure nucleotide diversity are deposited in GenBank under the following accession numbers: GU169471-GU169512 (mtDNA), GU169513-GU169643 (ptDNA), and GU169644-GU169689 (nucDNA). The GenBank accession numbers for the updated mitochondrial- and plastid-genome sequences of *V. carteri* are GU048821 (mtDNA) and GU084820 (ptDNA).

Results and Discussion

The V. carteri *Organelle Genomes*

Complete genetic maps of the *V. carteri* UTEX 2908 organelle genomes are shown in Figures 7.2 (mtDNA) and 7.3 (ptDNA). For comparison, these figures also contain the corresponding genetic maps from *C. reinhardtii* (Gray and Boer 1988; Michaelis et al. 1990; Maul et al. 2002). Previously, we (Smith and Lee 2009b; Appendix E) described partial, fragmented genome-assemblies of the *V. carteri* mtDNA (seven fragments) and ptDNA (34 fragments). In the current work, we were able to overcome earlier difficulties (Smith and Lee 2009b; Appendix E) and bridge the different organelle-genome fragments (i.e., contigs) by data mining the *V. carteri* Trace Archive for organelle-DNA trace files generated from large, 8-38 kb shotgun-sequencing fragments. These trace files allowed us to position the mtDNA and ptDNA fragments in the correct orientation and from this perform long-range PCR reactions, linking the various organelle-DNA contigs. Although the *V. carteri* organelle-genome maps are complete, there still exist several irresolvable gaps within the intergenic and intronic regions of the mtDNA- and ptDNA-sequence data. These gaps, which are annotated using strings of n's on the *V. carteri* mtDNA and ptDNA GenBank entries, are due to complications with sequencing and assembling the repetitive elements found throughout the *V. carteri* organelle DNA — discussed in Smith and Lee (2009b; Appendix E) and Aono et al. (2002).

The mitochondrial and plastid genome maps of *V. carteri* assemble as circular molecules of ~35 and ~525 kb, respectively (Figures 7.2 and 7.3). Whether or not these maps (which are supported by PCR analyses) reflect *in vivo* unit-genome-sized circular structures needs to be confirmed by gel electrophoresis, especially for the mitochondrial genome as all other characterized mtDNAs from *Reinhardtinia*-clade algae [*sensu* Nakada et al. (2008)] appear to be linear genome-sized or subgenomic-sized structures (Laflamme and Lee 2003, Fan and Lee 2002, Smith and Lee 2008a; Smith et al. 2010a Chapters 2 and 3). Both organelle DNAs are dense with noncoding nucleotides, ~60% (mtDNA) and ~80% (ptDNA), and are among the most bloated organelle genomes observed from photosynthetic eukaryotes (Figure 7.1). In fact, *V. carteri* has the most expanded mitochondrial genome examined from the Chlorophyta, and its ptDNA has

more noncoding DNA, both fractionally and in sheer amount than any other plastid genome sequenced to date — its closest rivals, in this regard, are the plastid genomes of the chlamydomonadalean green algae *Dunaliella salina* (~65.5% noncoding; Smith et al. 2010b; Chapter 8) and *C. reinhardtii* (~57% noncoding), and the legume *Trifolium subterraneum* (~58% noncoding) (Maul et al. 2002; Cai et al. 2008). That said, available data suggest that the plastid genomes of some *Acetabularia* species may be upwards of 2 Mb [reviewed by Palmer (1985)], which implies that they have a noncoding DNA composition greater than 90%. It should be noted that the *V. carteri cox1* has an optional group-I intron (Figure 7.2); this intron is present in UTEX 1885 but absent from UTEX 2908 (Smith and Lee 2009b; Appendix E). Finally, the *V. carteri* mtDNA sequence structure appears to be more dynamic than initially thought (Smith and Lee 2009b; Appendix E). Long-range PCR analyses and mtDNA trace-file data, gleaned from the *V. carteri* Trace Archive, revealed two mitochondrial genome isomers (A and B), which differ from one another in gene arrangement. Isomer A, depicted in Figure 7.2, was readily amplified by PCR and was the most abundant assembly product from the mtDNA trace-file data. Isomer B, shown in Figure 7.4, was also attainable through PCR but was much less abundant in the mtDNA trace-file data; we were not able to confirm circularity or linearity of the isomer-B map. Illegitimate recombination between repetitive DNA elements has been suggested as a mechanism for the rearrangement of green algal mitochondrial genomes (Nedelcu 1998; Nedelcu and Lee 1998; Smith and Lee 2008a; Chapter 2); such a mechanism might explain the origin of the different isomers of *V. carteri* mtDNA.

V. carteri *Strains and Their Genetic Loci*

We used seven geographical isolates of *V. carteri* f. *nagariensis* to measure nucleotide diversity (Table 7.1). Six of the seven isolates can be divided into three sets of male/female interfertile pairs, which were respectively collected close to the Japanese cities of Kobe, Ichinomiya, and Kawagoe; the remaining isolate is a female from Poona, India (Table 7.1; Figure 7.5). We could find no published reports about fertility between the different Japanese geographical isolates of *V. carteri*; however, there is evidence of reduced fertility between Indian and Japanese isolates; laboratory-generated hybrid

zygotes of the Poona female (UTEX 2903) x Kobe male (UTEX 2864), and the reciprocal cross, had less than 10% of the post-zygotic survival as compared to zygotes from Kobe x Kobe or Poona x Poona crosses (Adams et al. 1990). These later results suggest that UTEX 2903 is a subspecies within the *V. carteri* f. *nagariensis* complex. As far as we are aware, the seven strains employed here represent the only distinct geographical isolates of *V. carteri* f. *nagariensis* that are currently maintained in public or private culture collections. Although eight strains of *V. carteri* are listed in Table 7.1, UTEX 2908 is derived from UTEX 1885 (Sessoms and Huskey 1973) and is therefore not a distinct geographical isolate. For this reason, the DNA sequences that we generated from UTEX 2908 were used only to verify our UTEX 1885 DNA-sequence data and were not included in the nucleotide-diversity analyses.

The regions of the *V. carteri* mitochondrial, plastid, and nuclear genomes that were sequenced and used to calculate nucleotide diversity are listed in Table 7.2 and highlighted in pink on the mitochondrial (Figure 7.2) and plastid (Figure 7.3) genome maps. Altogether, we sequenced 6 kb of mtDNA, 17.6 kb of ptDNA, and 8 kb of nucDNA from each of the seven *V. carteri* isolates. We made an effort to use intergenic-, intronic-, and synonymous-sites in our calculations of π_{silent}. Other studies on genetic diversity in organelle and nuclear genomes, because of limited intraspecific sequence data, have used mostly synonymous sites for estimating π_{silent} — because silent sites can be under selective constraints (Hershberg and Petrov 2008), it is best to use more than one type of silent site to approximate $2N_g\mu$. We also tried to measure the silent-site nucleotide diversity of *V. carteri* using the same genetic loci that were employed previously for estimating π_{silent} from *C. reinhardtii* (Smith and Lee 2008b, 2009a; Chapters 5 and 6), thereby allowing for a more meaningful comparison of genetic diversity between these two species. To this end, we were able to sequence the same nuclear-DNA loci from *V. carteri* that were sequenced from *C. reinhardtii*. However, because of differences in gene order between the *V. carteri* and *C. reinhardtii* organelle genomes, and because of the very large intergenic regions found in the *V. carteri* organelle DNAs, we were only able to sequence some of the same organelle-DNA loci.

Nucleotide Diversity of V. carteri

Nucleotide-diversity measurements for the three genetic compartments of *V. carteri* are summarized in Tables 7.2 and 7.3. When indels are removed from the aligned DNA sequences, net values of π_{silent} are 0.38 x 10^{-3} (mtDNA), 0.65 x 10^{-3} (ptDNA), and 5.28 x 10^{-3} (nucDNA). Thus, π_{silent} of the nuclear genome is approximately 14 and 8 times that of the mitochondrial and plastid genomes, respectively, and the silent-site diversity of the ptDNA is 1.7 times that of the mtDNA. When our methods for calculating π_{silent} are modified to incorporate indels ($\pi_{silent+}$), by counting each insertion and deletion in the alignment as a nucleotide change (with indels longer than one nucleotide considered as a single polymorphic site), the overall silent-site nucleotide diversity of the mitochondrial compartment remains unchanged (0.38 x 10^{-3}) and that of nuclear compartment is only slightly higher (5.83 x 10^{-3} vs. 5.28 x 10^{-3}); this is because the mtDNA dataset contains no indels and the nucDNA-sequence data have only 12 indels. For the ptDNA-sequence data, $\pi_{silent+}$ was almost twice that of π_{silent} (1.23 x 10^{-3} vs. 0.65 x 10^{-3}), reflecting the large number of indels in the sequenced ptDNA regions (Tables 7.2 and 7.3).

Within each genetic compartment, π_{silent} was relatively constant among the various loci and silent sites that were examined (Table 7.2). Alignments of the sequenced mtDNA loci revealed only four polymorphic positions, all of which are silent: two and one polymorphic synonymous site in *cox1* and *nad6*, respectively, and one polymorphic site in the group-I intron of *cob*. For the ptDNA, polymorphisms are observed in four of the nine intergenic regions and in the group-I intron of *atpA* (Table 7.2). No polymorphic sites were found in the protein-coding ptDNA. And for the nuclear-DNA data, polymorphic sites are distributed almost equally both among the various loci and between synonymous sites and intronic regions. We examined the organelle- and nuclear-DNA loci for traces of selection using Tajima's *D* test (Tajima 1989), which compares the average number of nucleotide differences between pairs of sequences (i.e., π) to the total number of segregating sites (*S*). Tajima's *D* was not statistically significant when considering the individual loci (Table 7.2) but was significant for the concatenated intron sequences of the nuclear genes *PETC*, *SFA*, and *YPT4* (*D* values of -1.69, -1.71, and -1.7, respectively), which could be a sign of purifying selection. That said, our sample size is most likely too small to critically test for neutrality.

When the Poona strain of *V. carteri* (UTEX 2903) is removed from our nucleotide-diversity analyses there is a substantial decrease in π_{silent} and $\pi_{silent+}$ for both the organelle- and nuclear-DNA datasets: for the mtDNA, π_{silent} and $\pi_{silent+}$ drop to 0.27 x 10^{-3}; the silent-site nucleotide diversity of ptDNA becomes 0.12 x 10^{-3} ($\pi_{silent+}$ = 0.29 x 10^{-3}); and for the nucDNA, π_{silent} and $\pi_{silent+}$ drop to 0.10 x 10^{-3}. This reduction in nucleotide diversity upon the removal of UTEX 2903 is not surprising because gene flow could have been restricted between this strain and that of the other isolates for the following reasons. UTEX 2903 was collected in India, at a location, which is more than 6,000 km away from the collection sites of the other six isolates, all of which are located in Japan within ~400 km of each other (Figure 7.5). More importantly, the reduced fertility between Poona and Kobe isolates (Adams et al. 1990) suggests that there is limited opportunity for gene flow between Indian and Japanese strains. Keep in mind, therefore, that in the following section when we compare the π_{silent} data for *V. carteri* to those from other eukaryotes, the *V. carteri* nucleotide-diversity estimates include UTEX 2903 and, thus, may be somewhat inflated as compared to a situation where Japanese and Indian strains were treated as being from separate populations.

Π_{silent} *of* V. carteri *Versus that of* C. reinhardtii *and Other Eukaryotic Species*

How do the estimated levels of silent-site nucleotide diversity for *V. carteri* compare with those from other photosynthetic eukaryotes? The only photosynthetic eukaryotes for which we were able to find published π_{silent} data from all three genetic compartments are *C. reinhardtii* and *Arabidopsis lyrata* (Wright et al. 2008; Smith and Lee 2008b, 2009a; Chapters 5 and 6). The silent-site nucleotide diversities of the *V. carteri* mitochondrial and plastid genomes are both 0.044 times the corresponding values from *C. reinhardtii*, and the silent-site nucDNA diversity of *V. carteri* is approximately 0.16 times that of *C. reinhardtii* — similar proportional relationships are obtained when indels are included in the nucleotide-diversity calculations (Table 7.3). Therefore, based on our analyses, the overall π_{silent} estimates for the mitochondrial, plastid, and nuclear genomes of *Volvox carteri* are much smaller than those of *C. reinhardtii*. Interestingly, π_{silent} for *V. carteri,* follows the same general trends that were observed for *C. reinhardtii*: $\pi_{silent(mtDNA)} < \pi_{silent(ptDNA)} < \pi_{silent(nucDNA)}$. Moreover, for both *V. carteri* and *C. reinhardtii,*

π_{silent} of mtDNA is half that of the ptDNA. Although for *V. carteri*, π_{silent} of the nucDNA is approximately eight times greater than that of the ptDNA, whereas for *C. reinhardtii*, π_{silent} of the nucDNA is only about twice that of the ptDNA.

Surprisingly, the π_{silent} estimates of *V. carteri* are more similar to those of the land plant *A. lyrata* than they are to those of its close relative *C. reinhardtii.* For example, *V. carteri* and *A. lyrata* have comparable levels of silent-site mtDNA (0.38×10^{-3} vs. 0.35×10^{-3}) and ptDNA (0.65×10^{-3}-1.23×10^{-3} vs. 1.0×10^{-3}) diversities. And the silent-site nucDNA diversity of *V. carteri* is about 0.3 times that of *A. lyrata* (5.28×10^{-3}-5.83×10^{-3} vs. 20×10^{-3}) (Wright et al. 2008); however, most of this difference can be explained by the nuclear genome of *A. lyrata* being diploid and that of *V. carteri* being haploid (N_g of diploid loci is thought to be about twice that of haploid loci). Furthermore, as observed for *V. carteri* and *C. reinhardtii*, the silent-site nucleotide diversity of *A. lyrata* follows the trend: $\pi_{silent(mtDNA)} < \pi_{silent(ptDNA)} < \pi_{silent(nucDNA)}$. Note that no indels were observed in the *A. lyrata* organelle- and nuclear-DNA sequences (Wright et al. 2008).

Silent-site nucleotide-diversity values for only the mtDNA, ptDNA, or nucDNA of eukaryotic taxa, unaccompanied by data from the neighboring genetic compartments, are more readily available. Many of these π_{silent} estimates are listed in the supplementary materials of Lynch and Conery (2003) and Lynch et al. (2006) who compiled a summary of average π_{silent} values for some major eukaryotic groups and found the following. For the mitochondrial and nuclear genomes of protists, mean values of π_{silent} were 17.5×10^{-3} and 57.3×10^{-3}, respectively. And for land plants, average nucleotide diversity estimates were 0.4×10^{-3} (mtDNA), 3.7×10^{-3} (ptDNA), and 15.2×10^{-3} (nucDNA), which are more or less in accordance with the findings of Wright et al. (2008) for *A. lyrata*, and in a general sense reflect the π_{silent} values (and their overall trends) that we observe for *V. carteri*. More recently published π_{silent} data from the ptDNA and nucDNA of land plants (Chen et al. 2008; Breen et al. 2009; Quang et al. 2009) are concordant with earlier approximations. Although reliable silent-site nucleotide diversity data are lacking for land plant mtDNA, it appears to be $<0.4 \times 10^{-3}$ (Lynch et al. 2006), which is in stark contrast to the $\pi_{silent(mtDNA)}$ estimates from animals, which are upwards of 35×10^{-3}. That said, both animals and land plants appear to have similar mean $\pi_{silent(nucDNA)}$ estimates (13.4×10^{-3} vs. 15.2×10^{-3}) (Lynch and Conery 2003; Lynch 2006). If we take the above

comparisons as a whole, the following conclusions can be made regarding the silent-site nucleotide-diversity measurements for the organelle and nuclear DNAs of *V. carteri*: i) they are generally lower than the corresponding π_{silent} values from other eukaryotes; ii) they are more similar to those of land plants than they are to those of *C. reinhardtii* and other protists; and iii) they follow the same trend that has been observed for other photosynthetic eukaryotes: $\pi_{silent(mtDNA)} < \pi_{silent(ptDNA)} < \pi_{silent(nucDNA)}$.

Interpreting Π_{silent} for V. carteri

Do the relative π_{silent} values that we observed across the genetic compartments of *V. carteri* make sense from a population-genetic standpoint? When a population is at mutation-drift equilibrium, the nucleotide diversity at silent sites should reflect $2N_g\mu$. In order to interpret our π_{silent} estimates for *V. carteri* we must therefore consider, for each genetic compartment, the mutation rate and the effective number of genes per locus in the population. Recall that the relative differences of π_{silent} for the mitochondrial, plastid, and nuclear compartments of *V. carteri* are about 1:2:14 when indels are ignored, and approximately 1:3:15 when indels are factored in. We can account for at least half of the difference between $\pi_{silent(organelle\ DNA)}$ and $\pi_{silent(nucDNA)}$ if the mitochondrial and plastid genomes of *V. carteri* are transmitted maternally as indicated by laboratory experiments (Adams et al. 1990), whereas the nuclear genome shows biparental inheritance (Starr 1969). Thus, we would expect N_g of the uniparentally inherited organelle DNAs to be about half that of the nucDNA. Even if we artificially reduce $\pi_{silent(nucDNA)}$ for *V. carteri* by one half to account for this difference in N_g, we are still left with relative silent-site mtDNA, ptDNA, and nucDNA diversities of about 1:2:7 (indels out) and 1:3:8 (indels in). This remaining variation in π_{silent} among the three genetic compartments of *V. carteri* could be due to a further reduction in N_g as a result of natural selection on linked variation in the organelle genomes, which, because of uniparental inheritance, have less opportunity for recombination during sexual reproduction as compared to the nucDNA (Birky et al., 1989). However, the only study to seriously investigate this issue in a photosynthetic eukaryote, namely *A. lyrata*, concluded that N_g of the organelle DNA and nucDNA did not depart significantly from neutral expectations (Wright et al. 2008). Differences in the mutation rates among the three genetic compartments could also

contribute to the range of π_{silent} values that are observed for *V. carteri*. Unfortunately, there are no available data for this taxon that bear on this possibility. In Chlamydomonas, however, which also shows lower π_{silent} values in the organelle DNAs, silent-site substitution-rate analyses for mtDNA and nucDNA suggest equal rates in these two genetic compartments (Popescu et al. 2006; Popescu and Lee 2007). Alternatively, a recent broad study of synonymous-substitution rates in the mitochondrial, plastid, and nuclear genes of seed plants suggest that the average relative mutation rates of these compartments is 1:3:10 (Drouin et al. 2008), which is almost exactly what we would predict for *V. carteri* if trying to account for the relative values of π_{silent} (after adjusting for differences in N_g).

Finally, why are the π_{silent} values for *V. carteri* significantly lower than those of *C. reinhardtii* (and other protists)? One possibility is that we have underestimated π_{silent} of the *V. carteri* f. *nagariensis* population, potentially because most of the strains that we employed were isolated in Japan within 400 km of each other and the true range of this population is much larger. Indeed, the one strain included in this study that came from outside Japan, the Indian one, showed greater genetic distance to the Japanese isolates than the Japanese isolates did to each other; but this Indian isolate, as mentioned previously, is also known to have reduced sexual compatibility with at least some of the Japanese strains and may not be a legitimate member of the Japanese population. Moreover, in the case of *C. reinhardtii*, which has a population of interfertile members that extend across eastern North America (Harris 2009, pg. 16), we found no simple and obvious relationship between genetic distance and the proximity of sites where isolates were recovered (Smith DR and Lee RW, unpublished data), and the same seems to be true for the volvocacean *Pandorina morum* (Kirk 1998). A second possibility is that *V. carteri* has a lower rate of mutation per generation in all three genetic compartments relative to *C. reinhardtii*. A third explanation is that the effective size of the *V. carteri* f. *nagariensis* population is considerably smaller than that of its unicellular cousin *C. reinhardtii* because of a smaller geographical range and/or a lower population density; the latter would be expected because *V. carteri* individuals are more than a thousand times larger than those of *C. reinhardtii* and both algae can be found in similar habitats (Kirk 1998).

V. carteri *Genome Architecture and the Mutational-hazard Hypothesis*

As predicted by the mutational-hazard hypothesis, the *V. carteri* mitochondrial, plastid, and nuclear genomes have less silent-site nucleotide diversity than their more compact counterparts from *C. reinhardtii* and other eukaryotes, and similar levels of diversity as those from eukaryotic species with comparable genome architectures, such as land plants (with some exceptions). In other words, our approximations of $2N_g\mu$ for the *V. carteri* organelle and nuclear genomes reflect their noncoding-DNA contents, in that $2N_g\mu$ positively correlates with genome compactness. For instance, the noncoding-DNA densities of the *V. carteri* mitochondrial, plastid, and nuclear genomes are approximately 60-65%, 80% and 85%, respectively, and the corresponding values for *C. reinhardtii* are about 20-30%, 57%, and 80%. Given these noncoding-DNA contents, the mutational-hazard hypothesis would predict the difference in π_{silent} between these two algae to be greatest for their organelle genomes. This is indeed the case, the relatively compact mitochondrial and plastid genomes of *C. reinhardtii* each have 22 times more silent-site nucleotide diversity than their more expanded *V. carteri* homologs, whereas the *C. reinhardtii* nuclear genome, which is only slightly more compact than that of *V. carteri*, has only around six times more silent-site nucleotide diversity than the *V. carteri* nucDNA. Similarly, the highly compact mtDNAs of animals (~10% noncoding) have, on average, 30-90 times more silent-site nucleotide diversity than the organelle DNAs of *V. carteri* (Lynch et al. 2006). Yet π_{silent} for the bloated nuclear genomes of animals is only around 2-3 times that of the *V. carteri* nucDNA (Lynch et al. 2006).

Now let us consider the genome architectures and nucleotide diversities of *V. carteri* in relation to those of land plants. Based on data from completely sequenced land plant genomes, average noncoding-DNA contents are around 84% (mtDNA), 42% (ptDNA), and >80% (nucDNA). Thus, *V. carteri* and land plants have similar mtDNA and nucDNA genome architectures, but the *V. carteri* plastid genome is more expanded than land plant ptDNA (~80% vs. ~42% noncoding). Now recall that π_{silent} measurements for land plants are in the range of 0.4×10^{-3} (mtDNA), $1\text{-}10 \times 10^{-3}$ (ptDNA), and $0\text{-}20 \times 10^{-3}$ (nucDNA), with means of 3.7×10^{-3} and 15.2×10^{-3} for the ptDNA and nucDNA, respectively (Lynch et al. 2006). As forecasted by the mutational-hazard hypothesis,

$\pi_{silent(mtDNA)}$ for *V. carteri* and land plants are essentially the same (~0.4 x 10^{-3}), $\pi_{silent(ptDNA)}$ for *V. carteri* is around 0.02-0.35 times the mean value for land plants, and $\pi_{silent(nucDNA)}$ for *V. carteri* falls within the range of that for land plants (but is below the mean). For both *V. carteri* and land plants, $\pi_{silent(mtDNA)}$ and $\pi_{silent(ptDNA)}$ are between 0.02-0.5 times smaller than $\pi_{silent(nucDNA)}$, despite the fact that the organelle DNAs of *V. carteri* and land plants are equally or more compact than their nuclear counterparts. This suggests that noncoding organelle DNA carries a greater burden than noncoding nuclear DNA, meaning that the $2N_g\mu$ threshold above which noncoding nucleotides can be perceived and eliminated by natural selection may be lower for organelle genomes than for nuclear genomes. This makes sense if we consider that organelle genomes are typically several orders of magnitude smaller than nuclear genomes. So even though organelle and nuclear genomes can have comparable fractional noncoding-DNA compositions, the sheer volume of noncoding DNA in nuclear genomes will, generally speaking, always be much greater (10-10^6 times) than that of organelle genomes. This implies that the number of noncoding nucleotides that are crucial to gene function relative to the number that are inert will be higher for organelle DNA than for nucDNA. This is pertinent to our study if we remember that the central premise of the mutation-hazard hypothesis is that noncoding DNA increases the susceptibility of a genome to degenerative changes, and that the mutational disadvantage of noncoding DNA is determined by the number of noncoding nucleotides that are associated with gene function (n), and the mutation rate (μ), where the overall mutational disadvantage (s) of an intron or intergenic region is $n\mu$ (Lynch 2002; Lynch et al. 2006). It is predicted that the threshold in a genome below which stretches of noncoding DNA can expand is $2N_gs < 1$, or alternatively $2N_g\mu < 1/n$ (Lynch 2002; Lynch et al. 2006). Although difficult to estimate, it has been argued that n for the intergenic and intronic regions of organelle DNA is larger than that of nuclear DNA (Lynch 2007; Smith and Lee 2008b, 2009a; Chapters 5 and 6). Based on this and earlier studies it appears that the threshold for noncoding-DNA proliferation may lie somewhere near $2N_g\mu < 0.002$ for organelle DNA and $2N_g\mu < 0.05$ for nucDNA. If the above suppositions are true, it would mean that when addressing the mutational burden of noncoding DNA, it is best to compare like with like: organelle DNA vs. organelle DNA and nucDNA vs. nucDNA, etc. Again, it is

important to re-emphasize that for all of the above π_{silent}/genome-compactness comparisons, the silent-site nucleotide diversity values for *V. carteri* include the Poona strain (UTEX 2903). As previously mentioned, if we remove UTEX 2903 from our π_{silent} calculations it would only make stronger the positive correlations between π_{silent} and genome compactness that we observe.

To summarize, it appears that the large quantities of noncoding DNA in the *V. carteri* organelle and nuclear genomes can be explained by a reduced ability to eradicate excess DNA, caused by a low $2N_g\mu$. If in fact a reduced $2N_g\mu$ is responsible for the expanded architecture of the *V. carteri* organelle and nuclear genomes, could it mean that in an indirect way the evolution of multicellularity in the lineage that gave rise to *V. carteri* was a major contributor to genome expansion? As we have already discussed, the evolution of multicellularity could cause a reduction in N_g. Lynch (2006) proposes that "A central challenge for evolutionary genomics is to determine the extent to which the expansion of eukaryotic gene and genomic complexity was a necessary prerequisite or an indirect consequence of the evolution of complex morphologies." It will be interesting to investigate other volvocine algae with ranging levels of complexity to see if there is a continuum from relatively compact genomes in unicellular species, moderately expanded genomes in the smaller multicelluar forms, to highly bloated genomes in macroscopic multicellular species.

Soon the DOE JGI will be releasing the nuclear-genome sequence from a male isolate of *V. carteri* f. *nagariensis* (presumably UTEX 2864) to complement the nucDNA sequence of the female isolate UTEX 1885 (the plans to sequence this genome are listed under the DOE JGI Community Sequencing Program FY2009). The data generated by this endeavor will allow for a complete view of π for the *V. carteri* organelle and nuclear genomes. It will be interesting to see if the data presented here will be consistent with whole-genome analyses.

Acknowledgements

We thank Stephen Miller for graciously supplying the Isanuma male and female strains of *V. carteri*.

Table 7.1. *V. carteri* f. *nagariensis* Strains Employed in Study

Strain	Mating Type	Geographical Origin of Isolation[c]
UTEX 2908[a]	female	Kobe, Hyogo, Japan
UTEX 1885	female	Kobe, Hyogo, Japan
UTEX 2864	male	Kobe, Hyogo, Japan
UTEX 2903	female	Poona, India
NIES 397	female	Ichinomiya, Aichi, Japan
NIES 398	male	Ichinomiya, Aichi, Japan
Isanuma F[b]	female	Kawagoe, Saitama, Japan
Isanuma M[b]	male	Kawagoe, Saitama, Japan

[a] 72-52 dissociator mutant, derived from UTEX 1885; this strain was used only to verify our sequences from UTEX 1885 — i.e., it was not included in our nucleotide-diversity analyses.

[b] These strains were kindly provided to us by Stephen Miller of the University Maryland, Baltimore County. They were originally collected from Isanuma Pond in the summer of 2005 by Atsushi Nakazawa of Ichiro Nishii's laboratory during an excursion with a group from Hisayoshi Nozaki's laboratory.

[c] See Figure 4 for geographical maps of Japan and India highlighting the origins of isolation for the different *V. carteri* strains.

Table 7.2. Nucleotide-diversity Estimates by Region for the *V. carteri* Mitochondrial, Plastid, and Nuclear Genomes

Loci	# of sites[a]	*S*	# of indels[b] (length nt)	π^c x 10^{-3} (SD x 10^{-3})	π_{syn} x 10^{-3}	π_{nsyn} x 10^{-3}	Tajima's *D* test (P value)
Mitochondrial DNA							
cox1	915	2	0	2.19 (1.09)	9.13	0	-1.23 (> 0.1)
cob	225	0	0	0	0	0	---
cob (group-I intron)	1,138	1	0	0.42 (0.15)	---	---	0.55 (> 0.1)
nad2/*nad6*	430	0	0	0	---	---	---
nad6	39	1	0	3.6 (1.4)	10.9	0	---
nad6/*nad5*	21	0	0	0	---	---	---
nad5	1,341	0	0	0	0	0	---
L7	412	0	0	0	---	---	---
L7/S2	113	0	0	0	---	---	---
S2	220	0	0	0	---	---	---
S2/*nad1*	15	0	0	0	---	---	---
nad1	882	0	0	0	0	0	---
nad1/L3	145	0	0	0	---	---	---
L3	133	0	0	0	---	---	---
RNA-coding	765	0	0	0	---	---	---
Protein-coding	**3,420**	**3**	**0**	**0.14 (0.08)**	**0.48**	**0**	---
Silent sites	**2,984**	**4**	**0**	**0.38 (0.17)**	---	---	---
Plastid DNA							
atpA	42	0	0	0	0	0	---
atpA (group-I intron)	1,816	8	5 (55)	1.41 (0.65)	---	---	-1.26 (> 0.1)
atpE	138	0	0	0	0	0	---
atpE/*ycf12*	776	0	0	0	---	---	---
clpP	1,140	0	0	0	0	0	---

Loci	# of sites[a]	*S*	# of indels[b] (length nt)	π^c x 10^{-3} (SD x 10^{-3})	π_{syn} x 10^{-3}	π_{nsyn} x 10^{-3}	Tajima's *D* test (P value)
petA	657	0	0	0	0	0	---
psaB	1,374	0	0	0	0	0	---
psaC	81	0	0	0	0	0	---
psaC/trnF	776	9	1 (1)	3.32 (2.28)	---	---	-1.59 (> 0.05)
trnF	54	0	0	0	---	---	---
psbC	1,230	0	0	0	0	0	---
psbF	54	0	0	0	0	0	---
psbF/psbL	1,161	6	1 (16)	1.05 (0.31)	---	---	-1.40 (> 0.1)
psbL	63	0	0	0	0	0	---
rpl16	168	0	0	0	0	0	---
rpl16/rpl14	1,138	0	0	0	---	---	---
rpl14	51	0	0	0	0	0	---
rpl23	33	0	0	0	0	0	---
rpl23/rpl2	1,055	0	2 (364)	0	---	---	---
rpl2	102	0	0	0	0	0	---
tufA/trnV	1,294	1	2 (145)	4.1 (1.5)	---	---	0.55 (> 0.1)
trnR/petD	964	6	3 (11)	1.80 (1.24)	---	---	-1.52 (> 0.05)
petD	75	0	0	0	0	0	---
psbK	18	0	0	0	0	0	---
psbK/trnE	838	0	1 (3)	0	---	---	---
rps14	132	0	0	0	0	0	---
rps14/psbM	1,239	0	1 (1)	0	---	---	---
rbcL	1,185	0	0	0	0	0	---
Protein-coding	**6,543**	**0**	**0**	**0**	**0**	**0**	---
Silent sites	**13,238**	**30**	**16 (596)**	**0.65 (0.32)**	---	---	---

Loci	# of sites[a]	*S*	# of indels[b] (length nt)	π^c x 10^{-3} (SD x 10^{-3})	π_{syn} x 10^{-3}	π_{nsyn} x 10^{-3}	Tajima's *D* test (P value)
Nuclear DNA							
ACTIN e8-e10	411	3	0	2.09 (1.43)	8.90	0	-1.32 (> 0.1)
ACTIN i8-i9	419	8	0	5.91 (3.22)	---	---	-1.26 (> 0.1)
PETC e2-e4	276	0	0	0	0	0	---
PETC i1-i4	901	22	4 (13)	7.08 (4.86)	---	---	-1.69 (< 0.05)
SFA e5-e9	450	3	0	1.90 (1.31)	7.87	0	-1.35 (> 0.1)
SFA i5-i9	2,363	30	0	3.63 (2.49)	---	---	-1.71 (< 0.05)
YPT4 e5-e7	273	1	0	1.05 (0.72)	4.72	0	-1.00 (> 0.1)
YPT4 i4-i7	993	26	2 (6)	7.53 (5.17)	---	---	-1.70 (< 0.05)
PDK e3	51	0	0	0	0	0	---
PDK i3	359	5	5 (18)	4.19 (2.88)	---	---	-1.48 (> 0.1)
MAT3 e4-e8	953	1	0	0.52 (0.28)	0	0.69	-0.61 (> 0.1)
MAT3 i4-i7	679	5	1 (2)	3.69 (1.96)	---	---	-0.79 (> 0.1)
Protein-coding	**2,415**	**8**	**0**	**1.37 (0.94)**	**5.82**	**0**	---
Silent sites	**6,519**	**96**	**12 (39)**	**5.28 (3.24)**	---	---	---

Note: *S,* number of segregating sites (i.e., polymorphic sites); Indels, insertion-deletion events; π, nucleotide diversity; π_{syn}, nucleotide diversity at synonymous sites; π_{nsyn}, nucleotide diversity at nonsynonymous sites; SD, standard deviation. Intergenic regions are labeled using their bordering genes and a dash. For the nuclear loci, exons and introns are respectively labeled with an "e" and "i" followed by a number denoting their position within the gene.

[a] Comprises all sites in the nucleotide alignment, including those with indels.

[b] Indels involving more than one nucleotide are counted as a single event. Indel length includes the sum of all indels and includes consecutive indel events.

[c] Only includes sites without indels.

Table 7.3. Silent-site Nucleotide Diversity for *V. carteri* Compared with that of *C. reinhardtii*

	Mitochondrial DNA		Plastid DNA		Nuclear DNA	
	Vc	*Cr*	*Vc*	*Cr*	*Vc*	*Cr*
# of sites[a]	2,984	5,550	13,238	16,710	6,519	5,051
S	4	104	30	321	96	377
# of Indels[b] **(length nt)**	0	11 (31)	16 (596)	86 (1,679)	12 (39)	48 (222)
π_{silent}[c] **x 10^{-3}** **(SD x 10^{-3})**	0.38 (0.17)	8.51 (1.03)	0.65 (0.32)	14.53 (1.18)	5.28 (3.24)	32.29 (3.01)
$\pi_{silent+}$[d] **x 10^{-3}** **(SD x 10^{-3})**	0.38 (0.17)	9.23 (1.96)	1.23 (0.75)	18.36 (1.40)	5.83 (3.56)	36.00 (3.51)

Note: *Vc*, *Volvox carteri*; *Cr*, *Chlamydomonas reinhardtii*; *S,* number of segregating sites (i.e., polymorphic sites); Indels, insertion-deletion events; π_{silent}, nucleotide diversity at silent sites; $\pi_{silent+}$, nucleotide diversity at silent sites including both polymorphic sites and insertion-deletion events; SD, standard deviation. The *C. reinhardtii* data come from Smith and Lee (2008b, 2009a; Chapters 5 and 6).

[a] Comprises all sites in the nucleotide alignment, including those with indels.

[b] Indels involving more than one nucleotide are counted as a single event. Indel length includes the sum of all indels and includes consecutive indel events.

[c] Only includes sites in the alignment without indels.

[d] Considers all sites, including those with indels. Consecutive indels are counted as a single polymorphic event. Indel states were measured using a multiallelic approach.

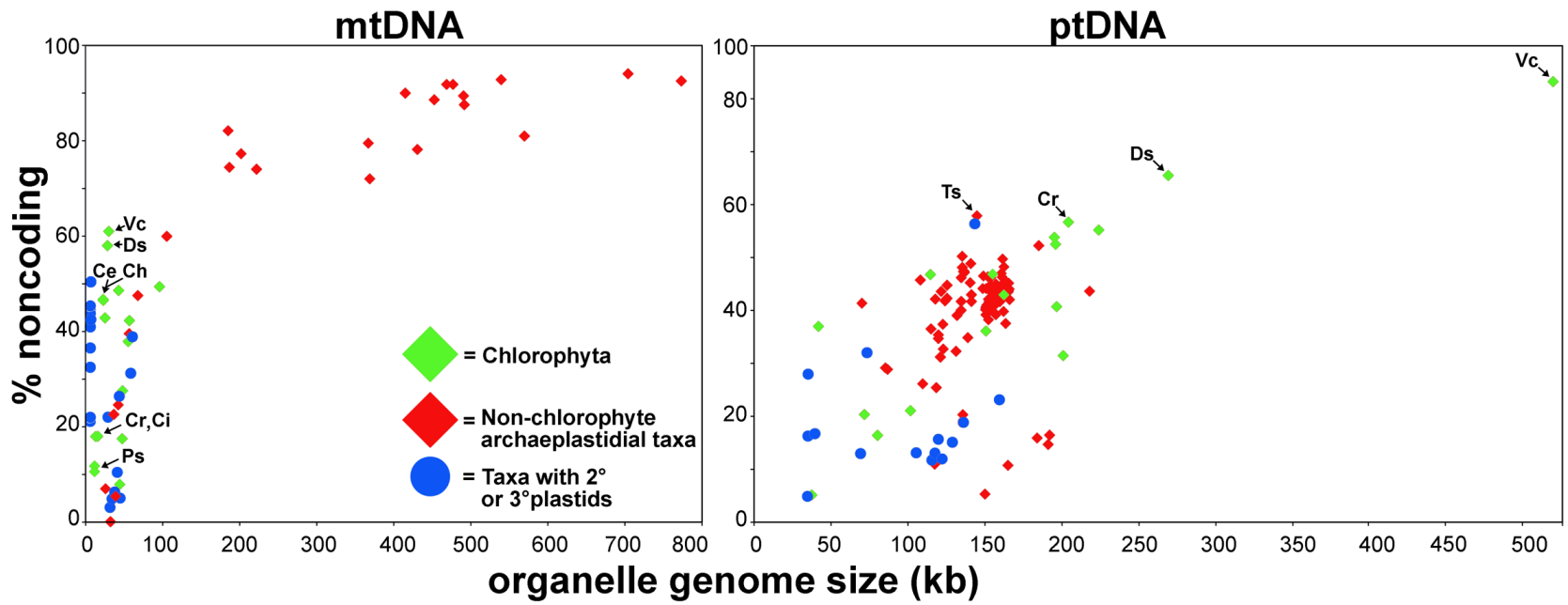

Figure 7.1. Scaling of Noncoding-DNA Content with Genome Size in Completely Sequenced Organelle DNAs

Chlamydomonadalean algae are labeled as follows: Ce = *Chlamydomonas eugametos*; Ch = *Chlorogonium elongatum*; Ci = *Chlamydomonas incerta*; Cr = *Chlamydomonas reinhardtii*; Ds = *Dunaliella salina*; Ps = *Polytomella capuana*, *Polytomella parva*, and *Polytomella piriformis* (strain SAG 63-10); Ts = *Trifolium subterraneum*; Vc = *Volvox carteri*.

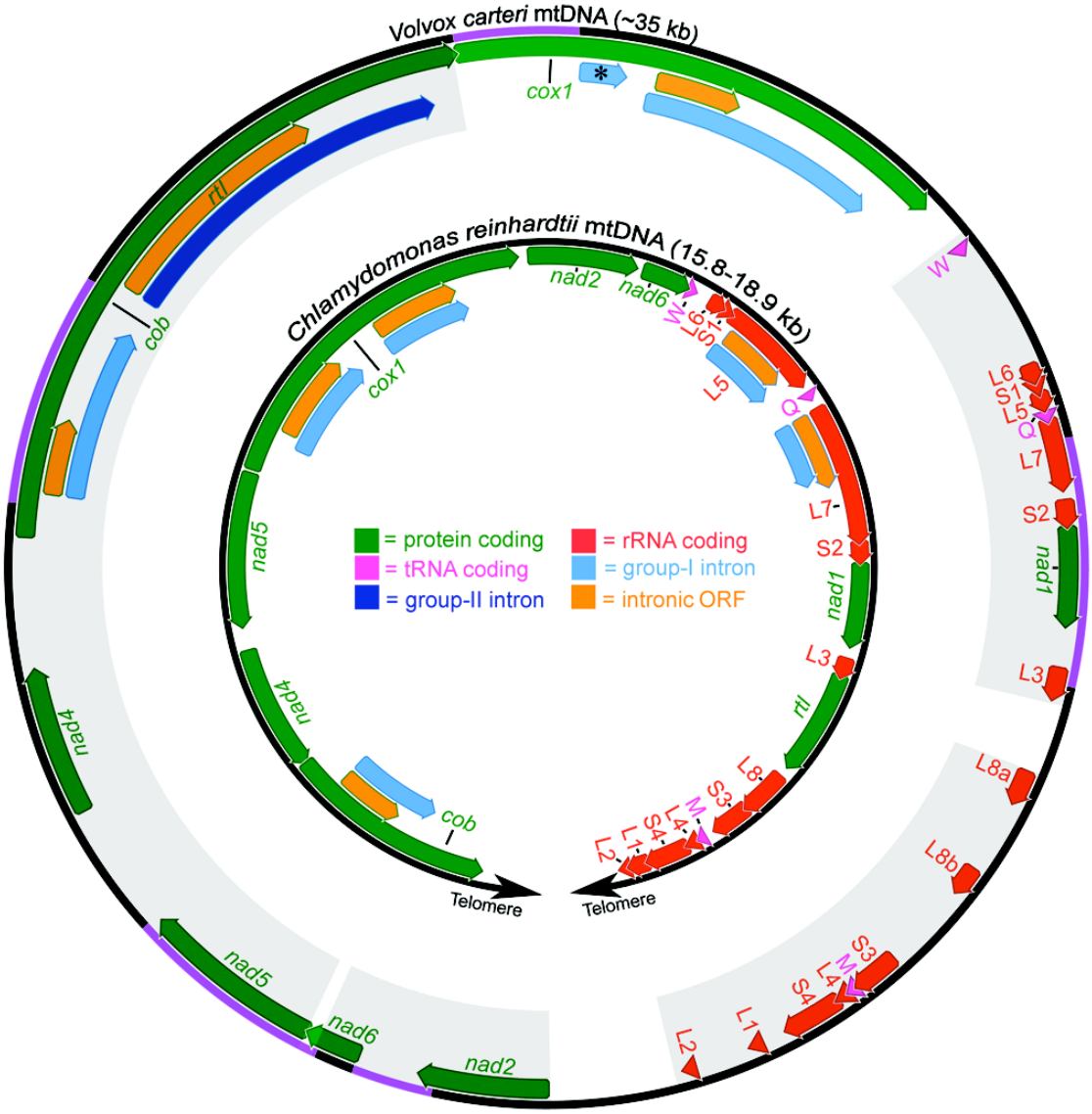

Figure 7.2. Complete Mitochondrial Genome Maps for *V. carteri* (Outer) and *C. reinhardtii* (Inner)

S1-S4 and L1-L8 represent the small-subunit and large-subunit rRNA–coding modules, respectively. Portions of the *V. carteri* mitochondrial genome map that are shaded gray represent regions of gene colinearity (not including introns) with the *C. reinhardtii* mtDNA, and portions that are pink correspond to the regions that were used for measuring nucleotide diversity. The *V. carteri cox1* group-I intron with an asterisk is optional. The GenBank accession number for the *V. carteri* (UTEX 2908) mitochondrial genome is GU048821 and those for the *C. reinhardtii* mtDNA are EU306617-EU306623.

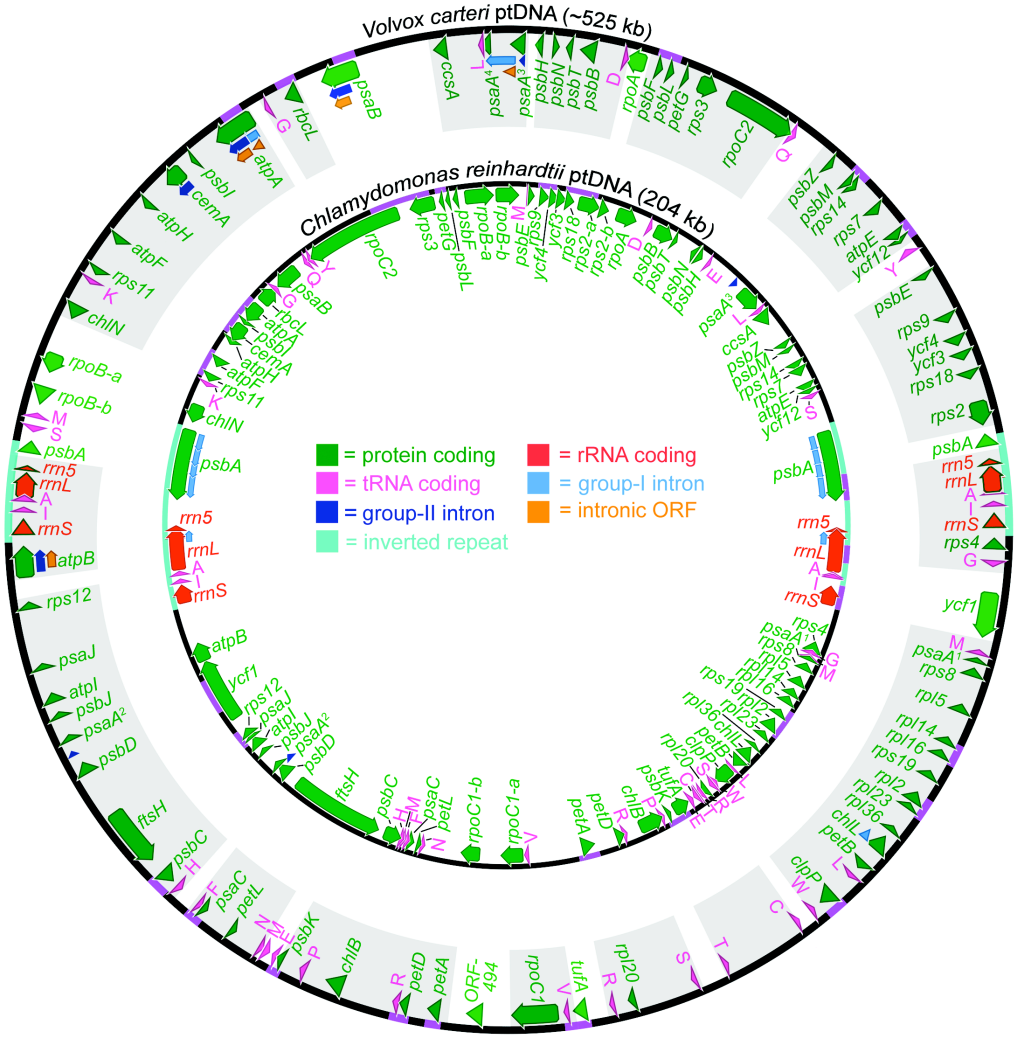

Figure 7.3. Complete Plastid Genome Maps for *V. carteri* (Outer) and *C. reinhardtii* (Inner)

Portions of the *V. carteri* plastid genome map that are shaded gray represent regions of gene colinearity (not including introns) with the *C. reinhardtii* plastid genome, and portions that are pink correspond to the regions that were used for measuring nucleotide diversity. GenBank accession numbers for the *V. carteri* and *C. reinhardtii* plastid genomes are GU084820 and FJ423446, respectively. For both genomes, the *psaA* gene is fragmented — the translational order of these fragments is signified with superscript numbers.

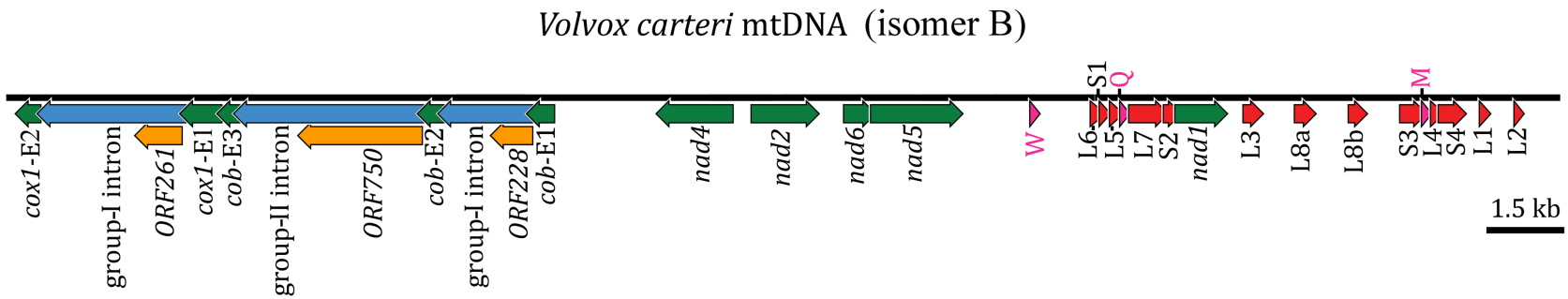

Figure 7.4. Mitochondrial Genome Map for *V. carteri* (Isomer B)

Protein-coding regions are green and their exons are labelled with an "E" followed by a number denoting their position within the gene. Introns and their associated open reading frames are blue and orange, respectively. Transfer RNA-coding regions are pink; they are designated by the single-letter abbreviation of the amino acid they specify. The large- and small-subunit rRNA-coding modules are red.

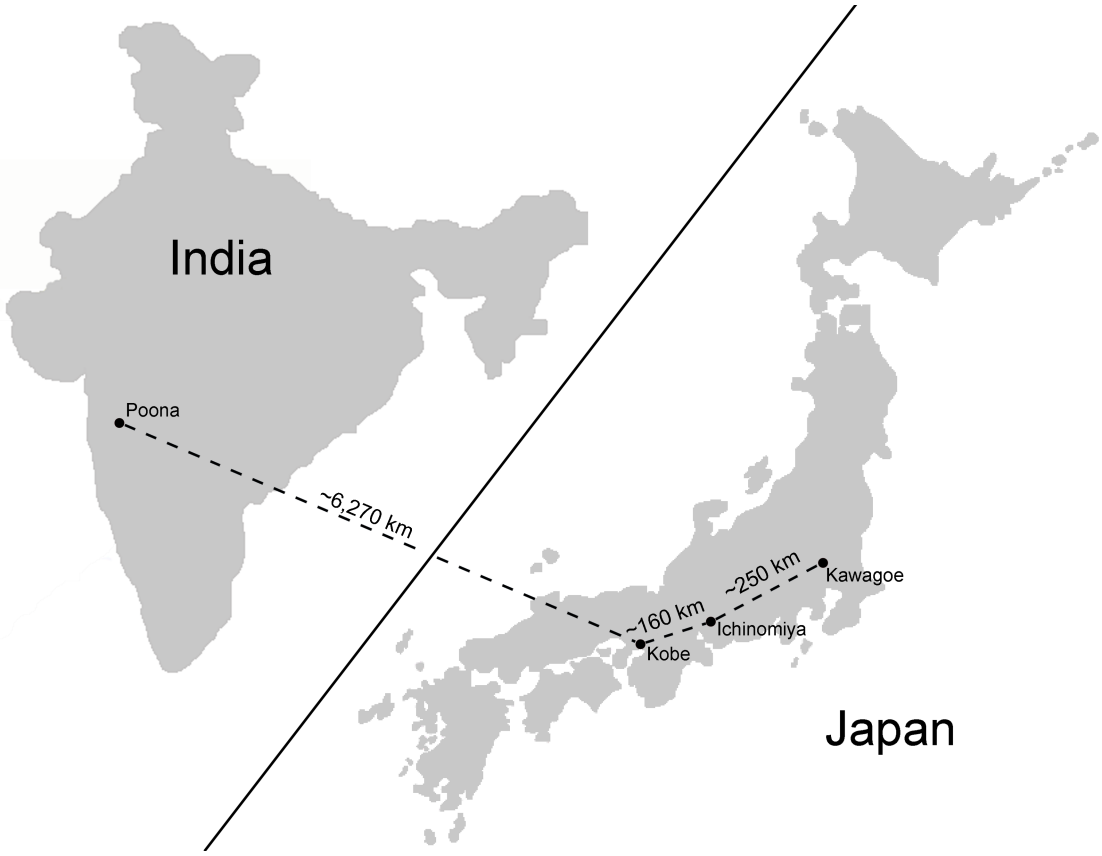

Figure 7.5. Geographical Maps of Japan and India Highlighting the Origins of Isolation of the *V. carteri* Strains Used in this Study

The distances in kilometers between origins of isolation are shown on the map. Refer to Table 7.1 for a list of the *V. carteri* strains.

CHAPTER 8: THE *DUNALIELLA SALINA* ORGANELLE GENOMES: LARGE SEQUENCES, INFLATED WITH INTRONIC AND INTERGENIC DNA

Published as:

Smith DR, Lee RW, Cushman JC, Magnuson JK, Tran D, Polle JEW (2010) The *Dunaliella salina* organelle genomes: large sequences, inflated with intronic and intergenic DNA. BMC Plant Biol 10:83.

Abstract

Background: *Dunaliella salina* Teodoresco, a unicellular, halophilic green alga belonging to the Chlorophyceae, is among the most industrially important microalgae. This is because *D. salina* can produce massive amounts of *β*-carotene, which can be collected for commercial purposes, and because of its potential as a feedstock for biofuels production. Although the biochemistry and physiology of *D. salina* have been studied in great detail, virtually nothing is known about the genomes it carries, especially those within its mitochondrion and plastid. This study presents the complete mitochondrial and plastid genome sequences of *D. salina* and compares them with those of the model green algae *Chlamydomonas reinhardtii* and *Volvox carteri*. Results: The *D. salina* organelle genomes are large, circular-mapping molecules with ~60% noncoding DNA, placing them among the most inflated organelle DNAs sampled from the Chlorophyta. In fact, the *D. salina* plastid genome, at 269 kb, is the largest complete plastid DNA (ptDNA) sequence currently deposited in GenBank, and both the mitochondrial and plastid genomes have unprecedentedly high intron densities for organelle DNA: ~1.5 and ~0.4 introns per gene, respectively. Moreover, what appear to be the relics of genes, introns, and intronic open reading frames are found scattered throughout the intergenic ptDNA regions — a trait without parallel in other characterized organelle genomes and one that gives insight into the mechanisms and modes of expansion of the *D. salina* ptDNA. Conclusions: These findings confirm the notion that chlamydomonadalean algae have some of the most extreme organelle genomes of all eukaryotes. They also suggest that the events giving rise to the expanded ptDNA architecture of *D. salina* and other Chlamydomonadales may have occurred early in the evolution of this lineage. Although interesting from a genome evolution standpoint, the *D. salina* organelle DNA sequences will aid in the development of a viable plastid transformation system for this model alga, and they will complement the forthcoming *D. salina* nuclear genome sequence, placing *D. salina* in a group of a select few photosynthetic eukaryotes for which complete genome sequences from all three genetic compartments are available.

Introduction

Dunaliella salina Teodoresco (1905) is one of the best-studied unicellular green algae (Oren 1005; Polle et al. 2009; Tafresh 2009). This is not only because *D. salina* is halotolerant, thriving in extreme saline environments (Polle et al. 2009), but also because it can produce large quantities of β-carotene (up to 10% of the cell's dry weight) in lipid globules located within the chloroplast (Ben-Amotz 1982; Katz et al. 1995). These traits make *D. salina* a model organism for investigating the evolution of salt adaptation (Oren 2005) and an attractive "cell factory" for the commercial production of β-carotene (Moulton et al. 1987; Ben-Amotz 2003). Although a great deal is known about the physiology and biochemistry of *D. salina* (Ben-Amotz et al. 2009; Avron and Ben-Amotz 1992), very little is known about the genomes it carries, especially those within its organelles. Until now, nothing was known about the size, conformation, or gene complement of either the mitochondrial or plastid genomes of *D. salina* (or those of any other *Dunaliella* species) even though the sequences of these genomes are essential to the development of new *D. salina* technologies, such as a viable plastid transformation system (Bock and Khan 2004; Walker et al. 2005; Fletcher et al. 2007; Verma and Daniell 2007; Purton 2007; Rosenberg et al. 2008).

Research on green-algal organelle genomes has led to significant advancements in genetic engineering. The first stable transformation of a plastid genome was achieved in 1988 using the unicellular green alga *Chlamydomonas reinhardtii* (Boynton et al. 1988) and, soon after, the first example of recombinant protein expression in a plastid was also achieved using *C. reinhardtii* (Goldschmidt-Clermont 1991). Since then, many techniques for plastid engineering have been first developed for green algae and then adapted for use in land plants (Maliga 2004). Given the relatively close evolutionary proximity of *C. reinhardtii* and *D. salina* (Nakada et al. 2008), it is reasonable to assume that many of the technologies for *C. reinhardtii* plastid transformation might be transferable to *D. salina*. Various groups have attempted to transform *D. salina* (Polle and Qin 2009); however, a lack of plastid-genome sequence data has prevented successful plastid transformation. Therefore, the first step in developing an efficient and reliable plastid transformation system for *D. salina* is to sequence its organelle genomes.

D. salina is an attractive alga for organelle genome research and plastome engineering for a variety of reasons: i) various strains and geographical isolates of *D. salina* are readily available from algal culture collections around the world; ii) *D. salina* is relatively easy to grow and maintain — it is one of the few microalgae that are being cultivated currently on a large scale; iii) *D. salina* lacks a rigid cell wall, facilitating organelle DNA extraction; iv) *D. salina* is unicellular, with only a single plastid, making it easier, as compared with multicellular species, to develop homoplasmic lines of plastid transformants; and v) being a close relative of the model green algae *C. reinhardtii* and *Volvox carteri*, means *D. salina* is an ideal species for comparative plant studies, especially comparative genomics, because the United States Department of Energy Joint Genome Institute (DOE JGI) is sequencing, or has sequenced, the *C. reinhardtii*, *V. carteri*, and *D. salina* nuclear genomes.

In 2006, the DOE JGI began sequencing the *D. salina* strain CCAP (Culture Collection of Algae and Protozoa) 19/18 nuclear genome, which is approximately 300 megabases (Mb) in length (DOE JGI, personal communication). *D. salina* was selected for genome sequencing because of its potential as a feedstock producer for biofuels production (Gouveia and Oliveira 2009) and its model status for studying saline adaptation. All of the *D. salina* whole genome shotgun sequencing (WGS) trace files that the DOE JGI produced are publicly available at the GenBank Trace Archive (http://www.ncbi.nlm.nih.gov/Traces/home/); soon a complete assembly of the *D. salina* nuclear DNA (nucDNA) will be made public. The fact that *D. salina* is the third chlamydomonadalean alga for which there is a genome sequencing project, reaffirms that the Chlamydomonadales are emerging as the one of foremost lineages for comparative genomics.

The taxonomic position of the *Dunaliella* genus is still under debate (González et al. 2009); however, it is often placed within the Chlamydomonadales (Chlorophyceae, Chlorophyta). For the purpose of this study, our definition of the Chlamydomonadales follows that of Lewis and McCourt (2004), which includes the *Dunaliella* genus, and is equivalent to both the basal-bodies-clockwise group (CW group) (Lewis and McCourt 2004) and the Volvocales *sensu* Nakada et al. (2008). Notably, some strains of *D. salina* were incorrectly identified in the past, which resulted in the deposition of inaccurately

labeled DNA sequence data in public databases. Moreover, there exists a debate regarding the delineation of the species *D. salina* Teod. and *Dunaliella bardawil* Avron et Ben-Amotz (Polle et al. 2009; González et al. 2009). According to Borowitzka and Borowitzka (1988) and Borowitzka and Siva (2007), the species *D. bardawil* is a *nomen nudum*; however, the name *D. bardawil* is still in use. Given the above issues, one should exercise great caution when using DNA sequence data in public databases that are said to have originated from *D. salina*.

As of November 1, 2009, complete and almost complete organelle DNA sequences are available for eight chlamydomonadalean algae, amounting to two plastid and eight mitochondrial genome sequences; moreover, comprehensive genetic maps and limited sequence data are available for the plastid genomes of an additional four taxa. The general features of these organelle genomes, as well as the species (and references) from which they come, are summarized in Table 8.1. Many of the available chlamydomonadalean organelle genome sequences are atypical in one way or another, having extreme sizes (e.g., large and expanded or highly compact [Michaelis et al. 1990; Smith and Lee 2009b; Appendix E]), unusual conformations (e.g., linear or linear fragmented [Vahrenholz et al. 1993; Fan and Lee 2002; Smith and Lee 2008a; Chapter 2]), and/or severely biased nucleotide composition (e.g., GC- or AT-rich [Borza et al. 2009). Furthermore, there can be extensive size, conformational and/or compositional differences among the organelle genomes of closely related chlamydomonadalean species (Boudreau and Turmel 1996; Mallet and Lee 2006; Borza et al. 2009; Smith et al. 2010a; Chapter 3). This wide assortment of genome architectures makes the Chlamydomonadales an ideal lineage for studying genome evolution (Nedelcu 1998; Popescu and Lee 2007; Smith and Lee 2010; Chapter 7).

Most of our knowledge of chlamydomonadalean organelle genomes comes from species within the *Reinhardtinia* clade (defined by Nakada et al. [2008]). Indeed, six of the eight chlamydomonadalean algae for which significant organelle DNA sequence data are available come from this clade (Table 8.1), including *C. reinhardtii* and *V. carteri.* Given that the *Reinhardtinia* clade contains only a small fraction of the species diversity within the Chlamydomonadales, it would be intriguing to explore the organelle genomes of algae from other chlamydomonadalean clades. It would be particularly interesting to

see if chlamydomonadalean algae from outside the *Reinhardtinia* clade have large, bloated plastid genomes. Both the *C. reinhardtii* and *V. carteri* plastid genomes, the only complete (or nearly complete) plastid DNA sequences that are available from the Chlamydomonadales, are among the largest and most noncoding-DNA dense plastid genomes observed to date, with sizes of 204 and ~525 kilobases (kb), respectively (Maul et al. 2002; Smith and Lee 2010; Chapter 7). The forces driving these genomes towards distention are unknown, but they may be connected to the combined effects of a low mutation rate and a low effective population size (Smith and Lee 2010; Chapter 7).

Here we present the complete mitochondrial DNA (mtDNA) and plastid DNA (ptDNA) sequences of *D. salina* strain CCAP 19/18 — a member of the *Dunaliellinia* clade (Nakada et al. 2008). The salient features of these genomes are described and compared with other chlamydomonadalean organelle genomes, particularly those of *C. reinhardtii* and *V. carteri*. The evolutionary and biotechnological implications of these sequences are discussed. The overreaching goals of this study are to use the *D. salina* organelle DNA data to test contemporary theories on genome evolution and to lay the foundation for a *D. salina* plastid transformation system.

Materials and Methods

D. salina *Strain Information*

The organelle DNA sequence data presented in this study come from *D. salina* strain CCAP 19/18, which is maintained at the Culture Collection of Algae and Protozoa (CCAP) in Argyll, Scotland. *D. salina* CCAP 19/18 originates from the hypersaline Hutt Lagoon in Western Australia.

Assembly of the D. salina *Organelle-genome Sequences*

The complete mitochondrial and plastid genome sequences of *D. salina* were generated by collecting and assembling the publicly available mtDNA and ptDNA trace files that the DOE JGI *D. salina* nuclear genome sequencing project produced. Trace files were data-mined from the National Center for Biotechnology Information (NCBI) *D. salina* Trace Archive (http://www.ncbi.nlm.nih.gov/Traces/home/) using the following

complete organelle genome sequences as trace BLAST (blastn 2.2.21+) queries: the *C. elongatum* and *C. eugametos* mitochondrial genomes, and the *C. reinhardtii* and *V. carteri* mtDNAs and ptDNAs — similar approaches to assembling organelle genomes have been used in other studies (e.g., Smith and Lee [2010; Chapter 7]; Voigt et al. [2008]). The BLAST parameters were as follows: an expectation value (E-value) of 10; a word size of 11; match and mismatch scores of 2 and -3, respectively; and gap-cost values of 5 (existence) and 2 (extension). Trace files showing >80% sequence similarity to the BLAST queries were downloaded and then assembled with CodonCode Aligner Version 2.0.6 (CodonCode Corporation, Dedham, MA, USA), which employs the Phred, Cross-match, and Phrap algorithms for base calling, sequence comparison, and sequence assembly, respectively. Assemblies were performed with a minimum percent identity score of 98, a minimum overlap length of 500 nt, a match score of 1, a mismatch penalty of -2, a gap penalty of -2, and an additional first gap penalty of -3. Gaps in the assemblies were filled by trace file walking, which was carried out by using *D. salina* mtDNA and ptDNA trace files as BLAST queries against the *D. salina* trace archive — a process that allows one to "walk" slowly in both directions along the contigs, thereby, filling in any gaps. The final assemblies of the *D. salina* mtDNA and ptDNA trace files gave complete mitochondrial and plastid genome sequences with greater than 50-fold coverage.

Analyses of Introns and Repetitive DNA in the D. salina *Organelle Genomes*

Introns in the *D. salina* organelle genomes were detected, classified, and folded into secondary structures using RNAweasel (Lang et al. 2007) and Rfam (Griffiths-Jones et al. 2003). Introns that were not detected by these programs were identified by their ability to be folded into suitable secondary structures. Intron/gene ratios were calculated by dividing the number of introns in the genome by the gene number; for the *D. salina* ptDNA, intergenic introns were included in this calculation. Dotplot similarity matrices were generated with JDotter (version 12.2.0) using a sliding window size of 50 (Brodie et al. 2004). Mfold (Zuker 2003) was employed for all secondary structure analyses.

The *D. salina* organelle-DNA sequences were initially scanned for repeats with REPuter (Kurtz and Schleiermacher 1999) using the Hamming distance option and a minimal repeat size setting of 12 nt. Forward, reverse, complement, and reverse

complement repeats were all considered under REPuter. More detailed analyses of the *D. salina* organelle genomes for repeats were performed by building a custom BLAST databank of the mtDNA and ptDNA sequences and then comparing (blastn version 2.2.21+) this databank with specific regions from the mitochondrial and plastid genomes using an E-value of 5, a word size of 7, a match score of 2, a mismatch penalty of -3, a gap open score of 5, and a extend value of 2.

Nucleotide Sequence Accession Numbers

The GenBank accession numbers of the *D. salina* organelle-genome sequences are GQ250045 (mtDNA) and GQ250046 (ptDNA).

Results and Discussion

Overview of the D. salina *Organelle Genomes*

The organelle genome sequences of *D. salina* were assembled using publicly available trace files that the DOE JGI *D. salina* nuclear genome sequencing project produced (see the Methods section for a detailed description of how the genome assembly was performed). Genetic maps of the *D. salina* organelle genomes are shown in Figures 8.1 (mtDNA) and 8.2 (ptDNA); for comparison, these two figures also include the corresponding genetic maps from *C. reinhardtii* and *Volvox carteri*. Table 8.1 outlines the general features of the *D. salina* organelle genomes, including their length, coding and noncoding DNA contents and nucleotide compositions, and compares these statistics to those from other chlamydomonadalean organelle DNAs. A Venn diagram highlighting the differences in gene content among available mtDNA sequences from the Chlamydomonadales is presented in Figure 8.3. A schematic compilation comparing the amounts of noncoding DNA in the *D. salina* organelle genomes with those from other completely sequenced (and almost complete) organelle genomes is shown in Figure 7.1 (Chapter 7), and analyses of the repetitive elements within the *D. salina* organelle DNA are summarized in Figure 8.4 and Figures 8.5 and 8.6.

Size, Conformation, and Nucleotide Composition

The *D. salina* mitochondrial and plastid genomes are 28.3 and 269 kb, respectively, and assemble as circular molecules — an observation that adds further support to the hypothesis that linear mitochondrial genomes in chlamydomonadalean algae are restricted to species within the *Reinhardtinia* clade (Table 8.1). The mitochondrial genome of *D. salina* is small relative to those of non-chlamydomonadalean green algae, which are, on average, 51.5 kb; however, it is still 5.4-15.3 kb larger than all other available chlamydomonadalean mtDNAs, except for that of *V. carteri*, which is ~35 kb (Smith and Lee 2009b, 2010; Chapter 7; Appendix E). The size of the *D. salina* plastid genome is more pronounced than its mitochondrial counterpart, being the largest ptDNA sequenced thus far. Its nearest rivals are the 223.9 kb ptDNA of the chlorophycean green alga *Stigeoclonium helveticum* and the 217.9 kb plastid genome of the geranium *Pelargonium* X *hortorum* (Bélanger et al. 2006; Chumley et al. 2006). Large plastid genomes are a common theme among chlamydomonadalean algae: the *C. reinhardtii* plastid genome, at 204.2 kb (Maul et al. 2002), is the fourth largest completely sequenced ptDNA, partial sequence data indicate that the *V. carteri* ptDNA is ~525 kb in length (Smith and Lee 2009b, 2010; Chapter 7; Appendix E), and gel electrophoresis results place the plastid genomes of *Chlamydomonas gelatinosa* (of the *Reinhardtinia* clade) and *Chlamydomonas moewusii* (of the *Moewusinia* clade [Nakada et al. 2008]) at ~285 kb and ~292 kb in length, respectively (Boudreau et al. 1994; Boudreau and Turmel 1996). The impressive size of the *D. salina* plastid genome (and those from other chlamydomonadalean algae) is a reflection of a prodigious noncoding DNA content rather than an unusually large gene repertoire. Within the *D. salina* ptDNA, a pair of inverted repeats, each with a length of 14.4 kb, divides the genome into a large (127.3 kb) and a small single-copy region (112.9 kb), referred to as the LSC and SSC regions. The *D. salina* inverted repeats are 6.2 kb shorter than their *C. reinhardtii* counterparts. This size discrepancy occurs because the *C. reinhardtii* inverted repeats contain *psbA*, a gene that is located in the SSC region of the *D. salina* ptDNA. Unlike in *D. salina*, the SSC and LSC regions of the *C. reinhardtii* ptDNA are virtually indistinguishable with sizes of ~80 kb. The precise lengths of the LSC, SSC, and inverted repeat regions for the *V. carteri* plastid genome are unknown; however, preliminary size

estimates place them at >25 kb (Smith and Lee 2010; Chapter 7). Southern blot analyses and partial sequence data indicate that the inverted repeats of the *C. moewusii* ptDNA may be upwards of 40 kb (Boudreau et al. 1994).

The GC content of the *D. salina* organelle DNAs is 34.4% (mtDNA) and 32.1% (ptDNA), which is unremarkable in relation to other archaeplastidial organelle genomes (i.e., those from eukaryotes with primary plastids). However, they are still the most GC-poor (or AT-rich) organelle DNAs observed within the Chlamydomonadales, which is significant because the Chlamydomonadales are one of the few lineages out of all eukaryotes known to contain species with GC-rich mitochondrial genomes (Table 8.1) (36,43). The GC content among the different regions of the *D. salina* mitochondrial and plastid genomes is relatively constant: $33\%_{(mtDNA)}$ and $34\%_{(ptDNA)}$ for coding DNA; $34\%_{(mtDNA)}$ and $32\%_{(ptDNA)}$ for introns and intronic open reading frames (ORFs); and $37\%_{(mtDNA)}$ and $31\%_{(ptDNA)}$ for intergenic regions. For the different codon-site positions of the mtDNA and ptDNA protein-coding regions, the GC content is approximately $38\%_{(mtDNA)}$ and $42\%_{(ptDNA)}$ (1^{st} position); $38\%_{(mtDNA)}$ and $52\%_{(ptDNA)}$ (2^{nd} position); and $19\%_{(mtDNA)}$ and $13\%_{(ptDNA)}$ (3^{rd} position). Cumulative GC-skew analyses, (often used to pinpoint origins of replication [Grigoriev 1998]) of the *D. salina* mtDNA show a strong positive correlation with the transcriptional orientation (data not shown), reflecting the slightly higher GC content of the coding and intronic regions relative to the noncoding mtDNA. The same analysis of the ptDNA gives a more disordered plot, but one typical of ptDNA, because of the many shifts in transcriptional polarity throughout the genome.

Coding Content

Like other chlamydomonadalean species, *D. salina* has a severely diminished mtDNA gene content of only 12 genes, which represent seven proteins, two rRNAs, and three tRNAs. Outside of the Chlamydomonadales, the only species known to have more reduced mtDNA gene contents are found in the phyla Apicomplexa and Dinoflagellata and arguably some species within the supergroup Excavata (Kairo et al. 1994; Nash et al. 2007). Various studies have tried to explain why chlamydomonadalean algae have such reduced mtDNA gene contents (Lynch et al. 2006; Smith and Lee 2008b; Chapter 5), but at present no straightforward answer to this question exists. The *D. salina* mtDNA gene

inventory mirrors those of *Chlorogonium elongatum* and *Chlamydomonas eugametos*, but it shows some differences to those of *Reinhardtinia*-clade algae. These differences, which can be visualized on the Venn diagram in Figure 8.3, involve changes in tRNA-coding content and in the number of rRNA-coding fragments found on the genome. For example, the *D. salina* mtDNA encodes three tRNAs, whereas *Polytomella* mtDNA contains only *trnM*, and the mitochondrial *rrns* and *rrnl* genes of *D. salina* are divided into three (S1-S3) and six (L1-L6) coding modules, whereas in available *Reinhardtinia*-clade mtDNAs the *rrns* and *rrnl* genes are fragmented into at least four and eight coding modules, respectively. Given the similarities among the *D. salina*, *C. elongatum*, and *C. eugametos* mitochondrial genomes, these findings add further appreciation for the stability of mtDNA gene content among chlamydomonadalean species outside of the *Reinhardtinia* clade and underscore the instability of mtDNA gene content among *Reinhardtinia*-clade taxa (Figure 8.3).

The *D. salina* plastid genome is much more gene rich than its mitochondrial counterpart, with 102 genes — five of which are duplicates found in the inverted repeats. When ignoring these duplicates, there are a total of 66 protein-, 3 rRNA-, and 28 tRNA-coding genes. This gene content is reduced from those of green-plant species outside the Chlamydomonadales, which on average have 123 ptDNA-encoded genes, representing 85 proteins, 3 rRNAs, and 35 tRNAs. It appears that chlamydomonadalean algae, at some point during their evolution, went through a major reduction in ptDNA (and mtDNA) coding content relative to most other photosynthetic eukaryotes. The *D. salina* ptDNA gene repertoire is identical to those of *C. reinhardtii* and *V. carteri* with the following minor exceptions: i) the *D. salina* plastid genome encodes three copies of *trnI* — one more than the *C. reinhardtii* and *V. carteri* ptDNAs; ii) *D. salina*, like *C. reinhardtii*, has two ptDNA copies of *trnE*, whereas *V. carteri* has only one; iii) for *D. salina* and *V. carteri*, the *rps2* gene is represented by a single open reading frame, whereas for *C. reinhardtii*, *rps2* is fragmented into two adjacent open reading frames (*rps2-a* and *rps2-b*); and iv) the *D. salina* ptDNA does not contain the *RoaA*-like gene (*orf494*), which is present in the *V. carteri* ptDNA. Preliminary investigations of the plastids from *Moewusinia*-clade algae (Broudreau et al. 1994; Turmel et al. 1987) indicate that their ptDNA gene contents are similar to those of *D. salina*, *C. reinhardtii*, and *V. carteri.*

Altogether, these findings suggest that ptDNA gene content is uniform throughout the Chlamydomonadales, save for some minor differences in the number of tRNA-coding genes.

Gene Order

All 12 genes on the *D. salina* mitochondrial genome are encoded on the same strand (i.e., have identical transcriptional polarities), a characteristic shared by the three other circular-mapping chlamydomonadalean mtDNAs sequenced thus far (Table 8.1). Sequence data from chlamydomonadalean algae whose mtDNAs map as linear molecules, such as *C. reinhardtii* and *Polytomella* spp., reveal genomes that have two unequally sized gene clusters (i.e., a group of two or more genes that are situated close to one another) with opposing transcriptional polarities, which proceed outwards toward the ends of the genome (Figure 8.1). Based on these and the above observations, it is reasonable to assume that the ancestral chlamydomonadalean mtDNA mapped as a circular molecule with ~12 genes, all of which were encoded on the same strand, and that the events giving rise to linear mtDNA were connected with, or resulted in, a shift in transcriptional orientation of approximately one-third of the genes. The mtDNA gene order of *D. salina* is unique and differs from those of other chlamydomonadalean algae. Very few conserved mtDNA gene clusters are shared among *D. salina* and other chlamydomonadalean species (Figure 8.1), but this is not surprising considering that mtDNA gene arrangements can vary significantly even among closely related species in this group (Denovan-Wright et al. 1998; Kroymann and Zetsche 1998; Mallet and Lee 2006; Smith and Lee 2010; Chapter 7). Previous reports suggest that homologous or illegitimate recombination between mtDNA repeats is causing mitochondrial genome rearrangements in chlamydomonadalean algae (Nedelcu 1997; Nedelcu and Lee 1998; Smith and Lee 2008a; Chapter 2). The *D. salina* mitochondrial genome does contain minor amounts of repetitive DNA (discussed below); these repeats may be catalysts for genome rearrangements. In *C. reinhardtii*, mitochondrial genes are organized into operons, which are first transcribed into polycistronic primary transcripts and then subsequently processed into mature monocistronic units via endo- and exonucleolytic cleavage (Gray and Boer 1988). A scan of the *D. salina* mitochondrial genetic map

reveals clusters of tightly packed genes separated by large stretches of noncoding DNA (Figure 8.1). These gene clusters may reflect the layout of operons in the genome, a theory supported by the fact that they are punctuated by regions of noncoding DNA that can be folded into secondary structures.

In contrast to the mtDNA, genes in the *D. salina* plastid genome are found on both strands and occur in small groups of two to four genes, which are distributed among the LSC and SSC regions and the inverted repeats (Figure 8.2). The former two regions contain approximately 50 and 45 genes, respectively, whereas the inverted repeats contain only five genes — fewer than any chlamydomonadalean inverted repeat explored heretofore. The following additional genes are observed in the inverted repeats of other chlamydomonadalean algae: *psbA* (*C. reinhardtii*, *V. carteri, C. moewusii, C. eugametos, C. gelatinosa*, and *Chlamydomonas pitschmannii*), *rbcL* (*C. moewusii* and *C. eugametos*), and *atpB* (*C. gelatinosa*). Regions of gene synteny between the ptDNA of *D. salina* and those of *C. reinhardtii* and *V. carteri* (the only other chlamydomonadalean algae for which complete ptDNA maps are available) are highlighted in gray on Figure 8.2. The allocation of *D. salina* genes into small clusters is consistent with what is known for the *C. reinhardtii* plastid genome, where genes appear to be transcribed into monocistronic and dicistronic transcripts (Sakamoto et al. 1994; Jiao et al. 2004) rather than the larger polycistronic transcripts that are observed for the mtDNA. Thus, regions of gene colinearity between *D. salina* and *C. reinhardtii* (or *V. carteri*) may represent conserved transcriptional units.

Introns and Intergenic Regions

One of the more salient features of the *D. salina* organelle genomes is their noncoding DNA content: 58% of the mtDNA and 65.5% of the ptDNA consist of either intergenic or intronic DNA. These values approach those of the *V. carteri* mitochondrial (>60% noncoding) and plastid (>80% noncoding) genomes, which are currently the most inflated organelle DNA sequences from the Chlorophyta (a phylum containing most of the identified classes of green algae [Lewis and McCourt 2004]) (Figure 7.1; Chapter 7). In fact, next to *V. carteri*, the *D. salina* ptDNA has a greater noncoding DNA composition than any other plastid genome sequenced to date, exceeding that of the

legume *Trifolium subterraneum* (57.9%) and *C. reinhardtii* (56.7%) (Figure 7.1; Chapter 7) (Maul et al. 2002; Cai et al. 2008). However, one would expect the unsequenced plastid genomes of *C. gelatinosa* and *C. moewusii*, based on their estimated sizes (Boudreau et al. 1994; Boudreau and Turmel 1996), to have more noncoding DNA than *D. salina* but less than that of *V. carteri* (i.e., between 65-80% noncoding). Interestingly, both the mitochondrial and plastid genomes of *D. salina* have equally large noncoding DNA densities (58% vs. 65.5%). This observation goes against what is seen in *C. reinhardtii* where the mtDNA and ptDNA have opposing architectures (~20% vs. ~57% noncoding), but it is consistent with the *V. carteri* organelle genomes, which are both distended with noncoding DNA (>60%).

The noncoding DNA in the *D. salina* organelle genomes can be subdivided into two categories: intergenic regions, which make up 8.37 kb (29.5%) of the mtDNA and 139.65 kb (52%) of the ptDNA, and introns and intronic ORFs, which together represent 8.05 kb (28.5%) and 36.49 kb (13.5%) of mitochondrial and plastid genomes, respectively. For the *D. salina* ptDNA, it is sometimes difficult to distinguish between intergenic DNA and intronic DNA because intron-like sequences (including intronic ORFs) are found in many of the intergenic regions (Figure 8.2). Altogether, 18 putative group-I introns were found in the mtDNA (two of which contain intronic ORFs) and 43 putative introns were discerned in the ptDNA: 36 within genes (35 of group-I and 1 of group-II affiliation) and 7 within intergenic regions (all of group-II affiliation). See Figures 8.1 and 8.2 as well as Appendix F for a comparison of the organelle genome intron content of *D. salina* with those of *C. reinhardtii* and *V. carteri*. Note, because of the inverted repeats, 11 of the 43 introns in the ptDNA are duplicates (the single-copy-intron count for the ptDNA is 32). Seventeen of the gene-located ptDNA introns contain ORFs (their families are shown on Figure 8.2), whereas no ORFs were found within the ptDNA intergenic introns. The remnants of eight intronic ORFs (pseudo ORFs) were found in the intergenic ptDNA regions; these pseudo ORFs, which are often located adjacent to intergenic introns (Figure 8.2), appear to be nonfunctional because they contain frameshifts in their coding regions. All eight of the pseudo ORFs show sequence similarity to genes that are typically found in either group-I or group-II introns (Figure 8.2), such as genes coding for integrase-, maturase-, reverse-transcriptase- and

endonuclease-like proteins. Intergenic introns have been identified in other genomes (Simon et al. 2008; Tourasse and Kolstø 2008), but until now they had never been observed in chlamydomonadalean organelle DNA, or, to the best of our knowledge, any other green-algal organelle genomes. Most of the intergenic introns are highly derived and could only be identified using domain V, which is the most conserved secondary structure element of group-II introns (Lang et al. 2007). Further experiments will need to be performed to confirm that the intergenic introns are functional (i.e., removed from mature transcripts) rather than inert sequences. If they are functional, then it would imply that many of the intergenic regions of the *D. salina* plastid genome are transcribed. There is also the possibility that the individual intergenic introns represent the fragments of larger introns that assemble after translation (i.e., the RNA fragments come together via base-pairing to form larger RNA species that are capable of splicing). However, secondary structure modeling of the intergenic introns gave no obvious indications that this was the case.

The intron/gene ratios for the *D. salina* mitochondrial and plastid genomes are 1.5 and 0.42, respectively. These values are much larger than those of other chlamydomonadalean organelle genomes, which range from 0 to 0.75 for available mitochondrial genomes, and are less than 0.07 for the two available ptDNA sequences. Notably, the *D. salina* ptDNA intron/gene ratio exceeds the average value for land plant mitochondrial genomes (~0.6) (Lynch et al. 2007), which are considered to be among the most intron-dense organelle DNAs.

The *D. salina* organelle DNAs contain significantly more introns than their *C. reinhardtii* and *V. carteri* counterparts (Figures 8.1 and 8.2). In stark contrast to the 18 introns found in the *D. salina* mtDNA, *C. reinhardtii* and *V. carteri* have five (all group I) and four (two group I and one group II) mitochondrial introns, respectively. Moreover, for *C. reinhardtii* all five introns are "optional" and, as of yet, the maximum number found in a single strain is three (Smith and Lee 2008b; Chapter 5). A similar trend is observed for the ptDNA: when counting duplicate genes only once, the plastid genome of *C. reinhardtii* has six introns (five group I and one group II) and that of *V. carteri* has eight (three group I and five group II); both these values are significantly less than the 32 unique putative introns found in the *D. salina* plastid genome. Interestingly, for all of the

genes that contain introns in the *C. reinhardtii* and *V. carteri* organelle DNAs, their homologues in *D. salina* also contain introns, with the exception of *cemA*, which is intronless in *D. salina* but contains a group-II intron in *V. carteri*. The number of introns per gene and the intron insertion sites can differ among these three algae.

Other notable features of the *D. salina* noncoding organelle DNA include three pseudogenes (*rps7*$^{\psi}$, *atpB*$^{\psi}$, and *chlL*$^{\psi}$) in the plastid genome. These pseudogenes, whose functional copies are also present in the ptDNA, were classified as such, not because they contain frameshifts in their coding sequence or because they appear highly degenerate relative to their functional counterparts, but because they are missing the first half or first two-thirds of their coding sequences. Furthermore, *atpB*$^{\psi}$ and *chlL*$^{\psi}$ are located immediately downstream of their functional copies, and in both cases a group-I intron is sandwiched between the functional gene and the pseudogene (Figure 8.2).

Repeats

Unlike the mtDNA, which is relatively devoid of repeats, the *D. salina* plastid genome abounds with repetitive elements. The difference in repeat content between the mitochondrial and plastid genomes can be visualized by comparing their respective dotplot similarity matrices, which are shown in Figures 8.5 (mtDNA) and 8.6 (ptDNA). Looking at the ptDNA dotplot, it is apparent that the ptDNA repeats are found in intergenic regions, introns, and in some of the longer protein-coding genes, such as *ftsH*, *rpoC2*, and *ycf1*. Nucleotide BLAST analyses of the *D. salina* plastid genome indicate that there are upwards of 5,000 repeats in the ptDNA, forming approximately 100 repeat subclasses. With some exceptions, these repeats range from 30-60 nt in length and are 70-90% AT. Some shorter (10-20 nt) GC-rich repeat elements (>50% GC) were also identified. The high degree of sequence similarity among the different ptDNA repeats is attributable to homopolymer runs rather than a recurring sequence motif. The *C. reinhardtii* and *V. carteri* plastid genomes are also rich in repetitive DNA and Southern blot analyses suggest that the *C. gelatinosa* ptDNA is as well (Boudreau and Turmel 1996; Maul et al. 2002; Smith and Lee 2009b; Appendix E). Presumably, *C. pitschmannii* plastid genome, the smallest ptDNA observed from the Chlorophyceae (~187 kb), contains fewer repeats than other chlamydomonadalean ptDNAs (Boudreau and Turmel

1995), implying a connection between repeat content and plastid genome size. In a general sense, the *D. salina* ptDNA repeats are analogous to the short dispersed repeats described for the *C. reinhardtii* ptDNA (Maul et al. 2002), but they lack the consistent motif of the palindromic elements of the *V. carteri* plastid genome (Smith and Lee 2009b; Appendix E).

The *D. salina* mtDNA dotplot reveals a mostly blank matrix, with the exception of some small diagonal lines, which correspond to palindromic repeats (i.e., repeats that can be folded into hairpin structures). Clusters of palindromic repeats are found in 18 different noncoding regions of the mitochondrial genome. The genomic breadth of these clusters can be seen in Figure 8.1, where purple asterisks pinpoint the precise location of these repeat clusters. A consensus sequence of one of the more frequently occurring clusters is depicted in Figure 8.4. The mean length of the palindromic repeat clusters is 110 nt and, on average, they have a GC content of 50%. Each cluster contains approximately three palindromic elements (i.e., three putative hairpin structures), and the individual palindromes within each cluster range from 23-38 nt in length, and can be either AT- or GC-rich (Figure 8.4). Ten of the eighteen intergenic regions that are found in the mtDNA contain either one or two palindromic clusters (Figure 8.1), an arrangement that suggests that the palindromes may play a role in gene processing. Palindromic repeats have been described in other chlamydomonadalean mitochondrial genomes, including those of *C. reinhardtii*, *V. carteri*, and *Polytomella capuana* (Gray and Boer 1988; Nedelcu and Lee 1998; Smith and Lee 2008a, 2009b; Chapter 2; Appendix E), and in many cases chlamydomonadalean mtDNA palindromic repeats have been implicated in RNA processing (Boer and Gray 1991; Smith and Lee 2008a, 2009b; Chapter 2; Appendix E). A notable observation is that the mitochondrial palindromic repeats of both *D. salina* and *V. carteri* often contain the sequence 5'-TTTA-3' (or 5'-TTT-3') in the loops of their hairpin structures (Figure 8.4).

Organelle DNA from Other D. salina *Strains*

Prior to this study, organelle DNA sequence data for *D. salina* was limited to 14 GenBank entries amounting to 6.7 kb of ptDNA, divided over six protein-coding loci (no mtDNA sequence data were available). Most of these 14 entries do not list the strains of

D. salina that were used for sequencing and only a few are associated with published articles. Alignments of the CCAP 19/18 ptDNA data generated here with the *D. salina* ptDNA data available at GenBank reveal a significant amount of sequence divergence: pairwise nucleotide diversity values varied from 2.4% to 17.6%, which suggests that none of the 14 *D. salina* entries are derived from strain CCAP 19/18 and that at least some of the sequences at GenBank labeled *D. salina* are not the same species as CCAP 19/18. In comparison, the average silent-site ptDNA diversity among seven different North American isolates of *C. reinhardtii* was estimated to be 1.4% (Smith and Lee 2009a; Chapter 6), and a recent study on seven *V. carteri* geographical isolates found π_{silent} of the ptDNA to be ~0.065% (Smith and Lee 2010; Chapter 7). The high degree of sequence diversity among *D. salina* GenBank entries appears to be a reflection of the known issue of misidentification of *Dunaliella salina* isolates. *Dunaliella* researchers should be mindful of the high levels of sequence divergence between CCAP 19/18 and the available *D. salina* GenBank entries if using these data for sequence-based studies, such as designing PCR primers.

Paving the Way Towards Plastid Transformation

The development of a reliable, high-efficiency genetic transformation system for *D. salina* is an important objective for the *Dunaliella* research community, especially when considering the many industrial applications that this technology could provide. Genetic transformation of the *D. salina* nuclear genome has been successful (Polle and Qin 2009; Feng et al. 2009); however, a lack of ptDNA sequence data has prevented successful attempts at transforming the *D. salina* plastid genome — although an unsuccessful attempt was made to transform the plastid of *Dunaliella tertiolecta* (Walker et al. 2005). For *Dunaliella* species, there are significant advantages to plastid transformation over nuclear transformation (see Verma and Daniell [2007] for a review), some of which relate to the fact that ptDNA is polyploid and experiences high levels of gene expression. Now that the *D. salina* ptDNA sequence is available, scientists will be able to use these data to develop plastid transformation vectors targeting specific regions of the *D. salina* ptDNA. Plastid transformation occurs via homologous recombination between an engineered vector and a selected region of the plastid genome. Moreover,

promoter and 3´ UTR regions of genes (Fletcher et al. 2007) can now be used to design vectors for *D. salina* with improved expression levels. In principle, transgenes can be integrated into any site of the plastid genome, but transcriptionally-active intergenic regions are ideal (Verma and Daniell 2007). One of the most frequently used integration sites is the *trnI-trnA* intergenic spacer, which is found in the inverted repeat of most plastid genomes, including that of *D. salina*. Given its popularity, the *trnI-trnA* intergenic region would be an ideal site for early attempts at transforming the *D. salina* ptDNA. However, the discovery of intergenic introns in the *D. salina* ptDNA is an indication that many of the intergenic regions are transcriptionally active, which, if true, should allow for a diverse range of transformation targeting sites.

Conclusion

The *D. salina* organelle genomes are large, intron-dense molecules comprised predominantly of noncoding nucleotides. Repetitive elements punctuate the noncoding regions of the mitochondrial and plastid genomes, but are much more prevalent in the ptDNA. Overall, the discovery of putative intergenic introns in the *D. salina* ptDNA adds a new layer of complexity to the diverse repertoire of organelle genome architectures found in the Chlamydomonadales. Already a model organism for synthesizing β-carotene and studying salt adaptation, *D. salina* is now an ideal species for investigating organelle genome expansion. The high level of repetitive elements found in the plastid genome of *D. salina* may mirror the high level of repetitive sequences that is expected to be present in the nucDNA (DOE JGI, unpublished data). Publication of the plastid genome of *D. salina* is expected to result in major advances of plastid engineering with generation and use of transgenic *D. salina* strains for a number of new applications in the fields of biofuels as well as in vaccine antigene and biopharmaceutical production. The complete sequence of the *D. salina* nuclear genome will be made available soon, placing *D. salina* in a group of a select few photosynthetic eukaryotes for which complete genome sequences from all three genetic compartments are available.

Acknowledgements

This work was supported by a SunGrant Initiative grant awarded to JCC. We thank Leyla Hathwaik for technical assistance and the DOE JGI for sequencing the *D. salina* genome and providing the data for this study. Support for the Nevada Genomics, Proteomics, and Bioinformatics Centers was made possible by NIH Grant Number P20 RR-016464 from the INBRE-BRIN Program of the National Center for Research Resources and the NIH IDeA Network of Biomedical Research Excellence (INBRE, RR-03-008). The work conducted by the DOE JGI is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Table 8.1. Available Organelle Genome Data for Chlamydomonadalean Algae

Genus and species	Clade	Mapping conformation	Size (kb)	% coding[a]	%GC	GenBank Accession	Reference (Chapter)
MITOCHONDRIAL GENOMES							
Chlamydomonas incerta	*Reinhardtinia*	linear	~17.5[c]	~75[c]	~44[c]	DQ373068	Popescu & Lee 2007
Chlamydomonas reinhardtii	*Reinhardtinia*	linear	15.8 –18.9[b]	67-82	~45	EU306617-23	Michaelis et al. 1990 Chapter 5
Polytomella capuana	*Reinhardtinia*	linear	13.0	82.0	57.2	EF645804	Smith & Lee 2008a Chapter 2
Polytomella parva	*Reinhardtinia*	linear-fragmented	16.2[d]	65.5[d]	41.0[d]	AY062933-4	Fan & Lee 2002
Polytomella piriformis[g]	*Reinhardtinia*	linear-fragmented	16.1[d]	65.8[d]	42.0[d]	GU108480-1	Smith et al. 2010a Chapter 3
Volvox carteri f. *nagariensis*	*Reinhardtinia*	circular[f]	~35[c]	<40[c]	~34[c]	EU760701, GU084821	Smith & Lee 2009b, 2010 Chapter 7, Appendix E
Chlamydomonas eugametos	*Moewusinia*	circular	22.9	53.4	34.6	AF008237	Denovan-Wright et al. 1998
Chlamydomonas moewusii	*Moewusinia*	circular[e]	~21[e]	—	—	—	Boudreau et al. 1994
Chlamydomonas pitschmannii	*Moewusinia*	circular[e]	~16.5[e]	—	—	—	Boudreau & Turmel 1995
Chlorogonium elongatum	*Chlorogonia*	circular	22.7	53.3	37.8	Y13643-4, Y07814	Kroymann & Zetsche 1998
Dunaliella salina	*Dunaliellinia*	circular	28.3	42.0	34.4	GQ250045	Smith et al. 2010b Chapter 8
PLASTID GENOMES							
Chlamydomonas gelatinosa	*Reinhardtinia*	circular[e]	~285[e]	—	—	—	Boudreau & Turmel 1996
Chlamydomonas reinhardtii	*Reinhardtinia*	circular	204.2	43.3	34.5	FJ423446	Maul et al. 2002 Chapter 6
Volvox carteri f. *nagariensis*	*Reinhardtinia*	circular	~525[c]	<20[c]	~43[c]	GU084820	Smith & Lee 2009b, 2010 Chapter 7, Appendix E
Chlamydomonas	*Moewusinia*	circular[e]	~243[e]	—	—	—	Boudreau et al. 1994

Genus and species	Clade	Mapping conformation	Size (kb)	% coding[a]	%GC	GenBank Accession	Reference (Chapter)
eugametos							
Chlamydomonas moewusii	*Moewusinia*	circular[e]	~292[e]	—	—	—	Boudreau et al. 1994
Chlamydomonas pitschmannii	*Moewusinia*	circular[e]	~187[e]	—	—	—	Boudreau & Turmel 1995
Dunaliella salina	*Dunaliellinia*	circular	269.0	34.5	32.1	GQ250046	Smith et al. 2010b Chapter 8

Note: "—", data not available. Clades are defined by Nakada et al. (2008).

[a] Intronic open reading frames were not considered as coding DNA.

[b] These data vary because of the presence/absence of optional introns.

[c] These data are based on almost-complete genome sequences.

[d] MtDNA consists of two fragments; data are based on the concatenation of these fragments.

[e] These data are based on gel-electrophoresis and Southern-blot analyses.

[f] The circular conformation of the *V. carteri* mtDNA is based on genome-assembly data (Smith and Lee 2010) and needs to be confirmed by gel electrophoresis experiments.

[g] This strain is formally known as *Polytomella* SAG 63-10.

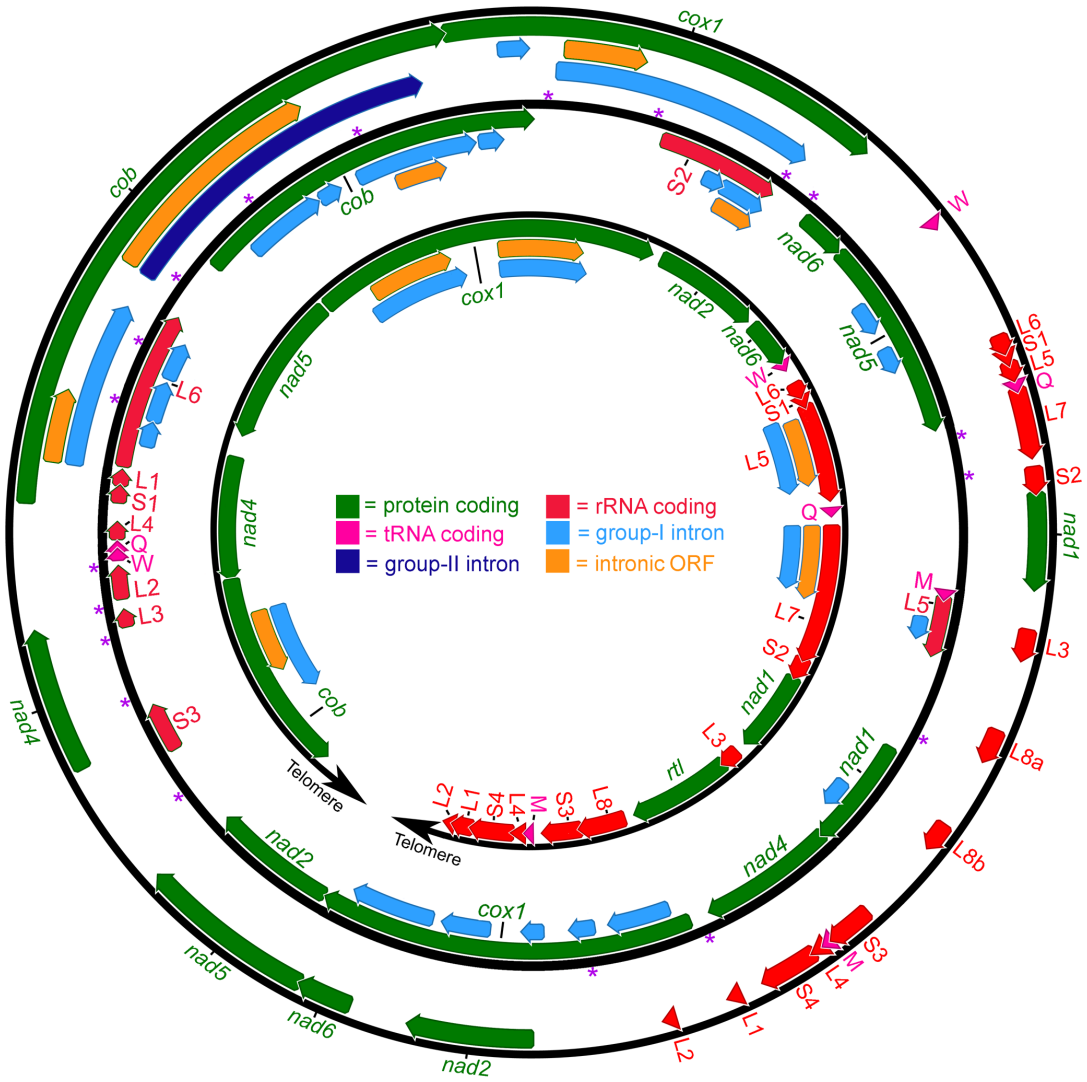

Figure 8.1. Complete Mitochondrial Genome Maps for *D. salina* (Middle), *C. reinhardtii* (Inner), and *V. carteri* (Outer)

The mitochondrial genome of *D. salina* (this study) is 28.3 kb, that of *C. reinhardtii* (GenBank accession numbers EU306617-23) ranges from 15.8-18.9 kb, depending on the presence of optional introns, and that of *V. carteri* (GenBank accession numbers EU760701 and GU084821) is ~35 kb. Note that the *C. reinhardtii* mtDNA is a linear molecule. Arrows within the coding regions denote transcriptional polarities. The small subunit and large subunit rRNA-coding regions are fragmented into modules. Transfer RNA-coding regions are designated by the single-letter abbreviation of the amino acid they specify. Purple asterisks denote the sites of palindromic repeat clusters (see Figure 8.4 for more details). *Rtl* codes for a putative reverse-transcriptase-like protein.

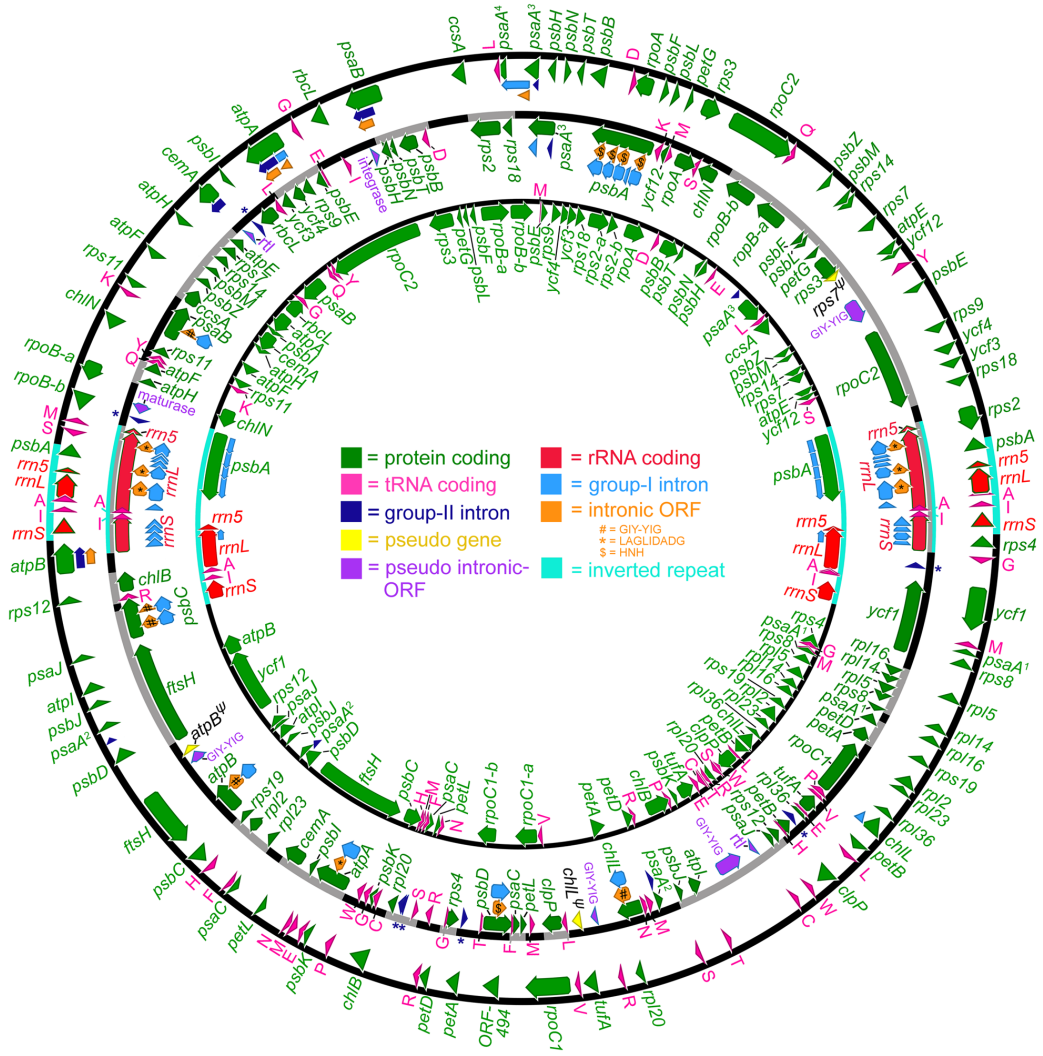

Figure 8.2. Complete Plastid Genome Maps for *D. salina* (Middle), *C. reinhardtii* (Inner), and *V. carteri* (Outer)

The *D. salina* plastid genome (this study) is 269 kb. The *C. reinhardtii* and *V. carteri* plastid genomes (GenBank accession numbers FJ423446 and GU084820) are 204.2 kb and ~525 kb, respectively. Arrows within the coding regions denote transcriptional polarities. Transfer RNA-coding regions are designated by the single-letter abbreviation of the amino acid they specify. Introns within intergenic regions are labeled with blue asterisks. Pseudogenes are labeled with a ψ. For all three genomes, the *psaA* gene is fragmented; the translational order of these fragments is set out using superscript numbers. The portions of the *D. salina* genome map that are gray (as opposed to black) highlight gene colinearity (not including introns) with either the *C. reinhardtii* or *V. carteri* plastid genomes.

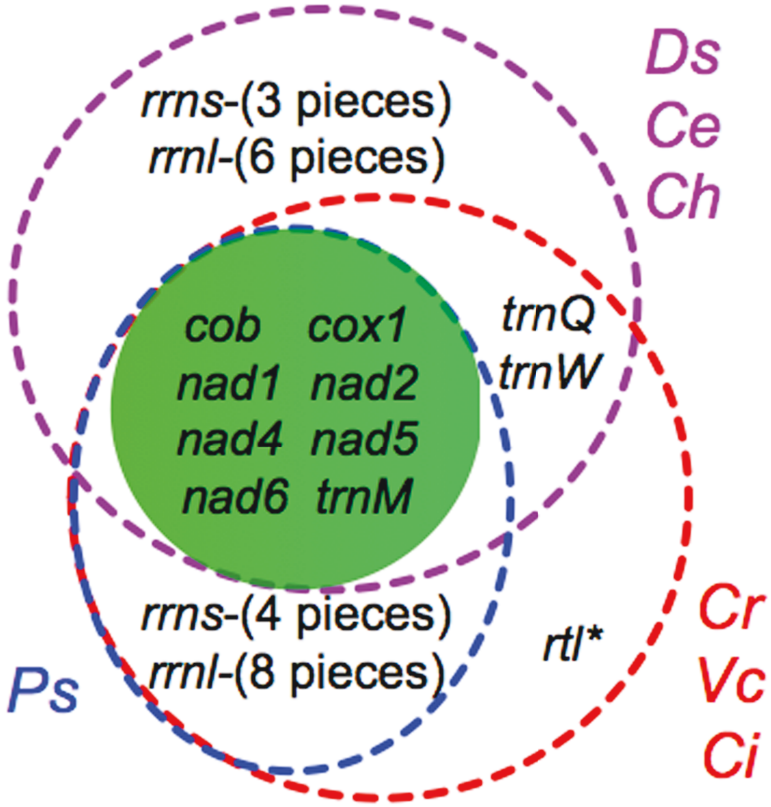

Figure 8.3. Venn Diagram Comparing the Gene Repertoires of Chlamydomonadalean Mitochondrial Genomes

Chlamydomonadalean algae are labeled as follows: Ce = *Chlamydomonas eugametos*; Ch = *Chlorogonium elongatum*; Ci = *Chlamydomonas incerta*; Cr = *Chlamydomonas reinhardtii*; Ds = *Dunaliella salina*; Ps = *Polytomella capuana*, *Polytomella parva*, and *Polytomella piriformis* (strain SAG 63-10); Vc = *Volvox carteri*. **Rtl* codes for a putative reverse-transcriptase-like protein: in *C. reinhardtii* and *C. incerta* this gene is independent of an intron, whereas in *V. carteri* it is within a group-II intron. Note, the *C. eugametos* mtDNA contains a duplicate copy of *trnM*.

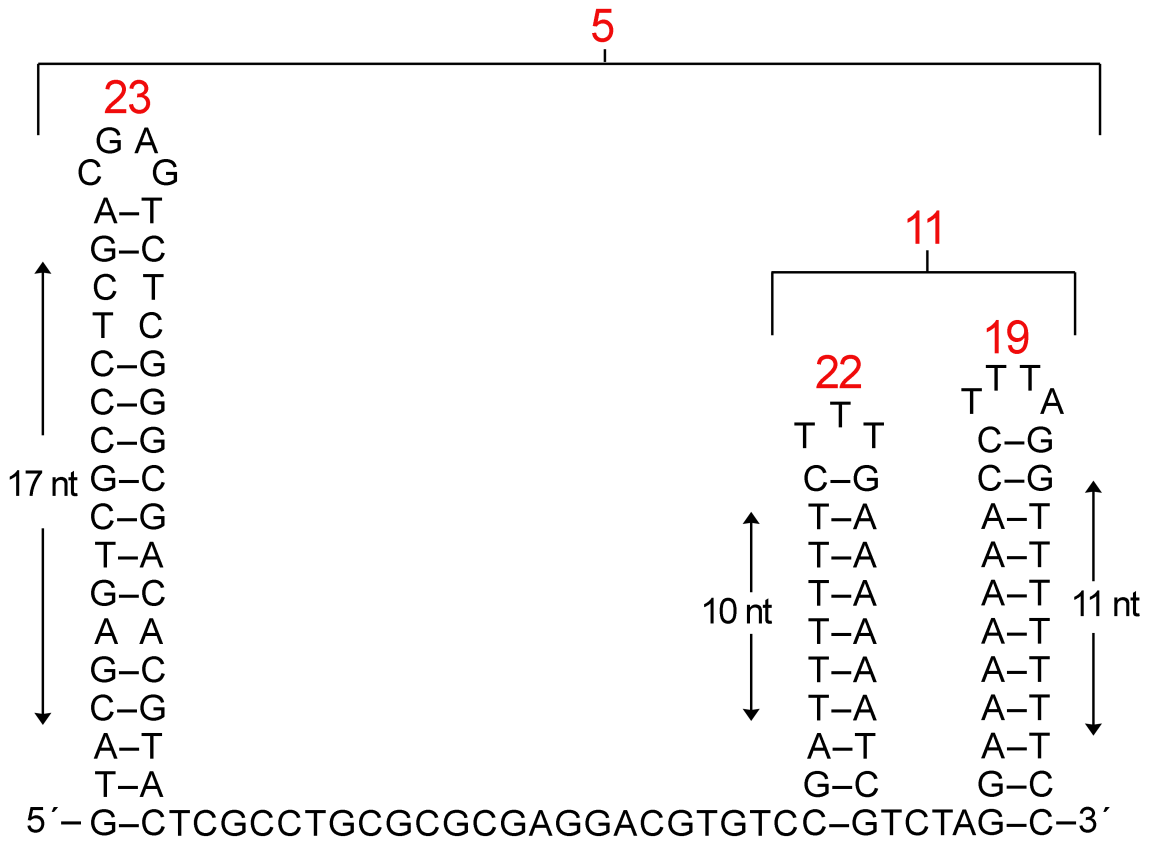

Figure 8.4. Consensus Sequences and Secondary Structures of the *D. salina* Mitochondrial Palindromic Repeat Elements

The number of times each element appears in the *D. salina* mitochondrial genome is shown in red numbers. The locations of these palindromic elements within the mtDNA are depicted on Figure 7.1 using purple asterisks.

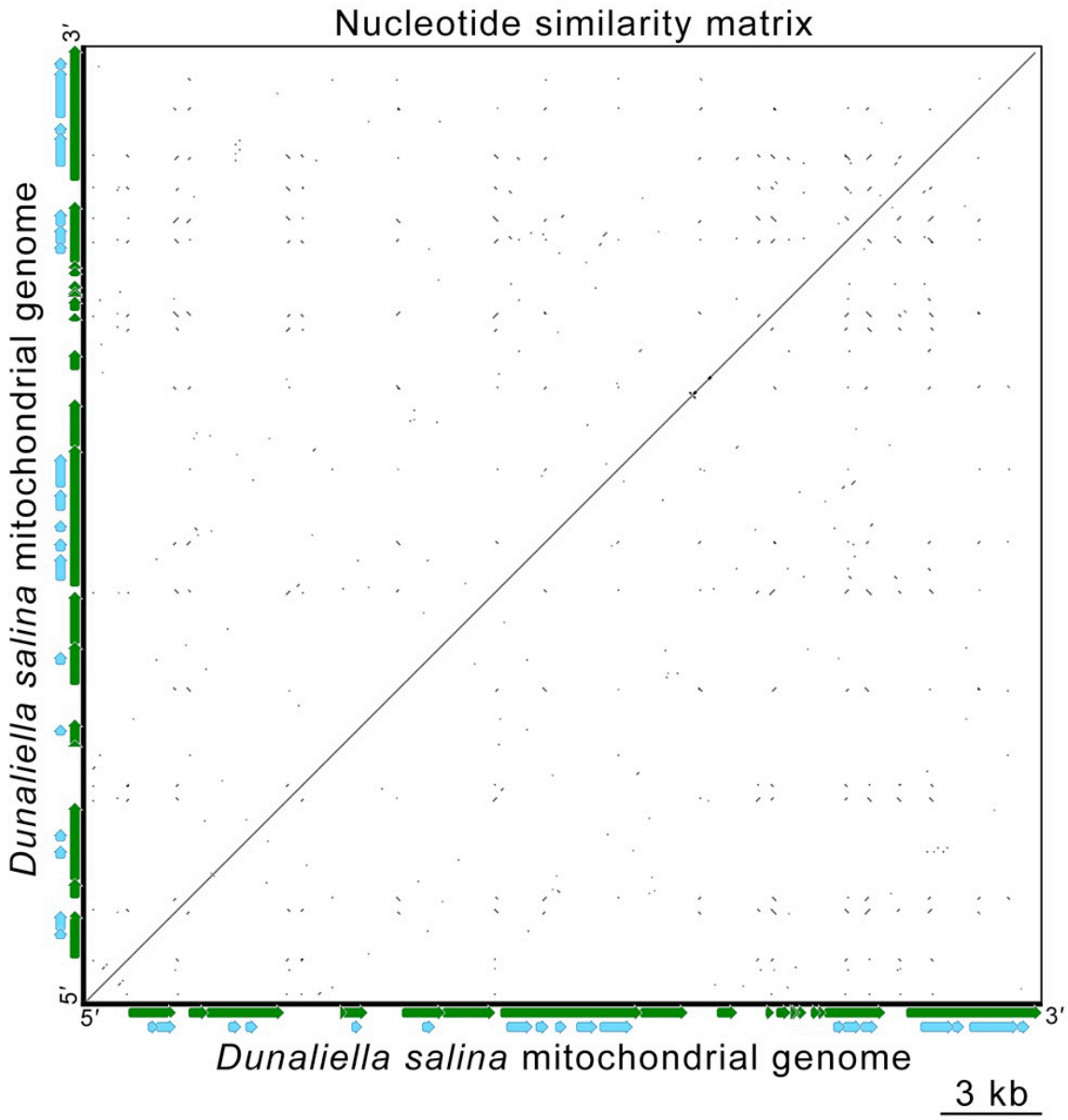

Figure 8.5. Dotplot Similarity Matrix of the *D. salina* Mitochondrial Genome

The X- and Y-axes each represent the *D. salina* mitochondrial genome (28.3 kb). For clarity, genetic maps of the *D. salina* mtDNA are placed below and beside these axes — on these maps coding regions are green and introns are blue (refer to Figure 8.1 for the complete annotation). Dots in the nucleotide similarity matrix represent regions of sequence similarity. The matrix was generated using a sliding-window size of 50.

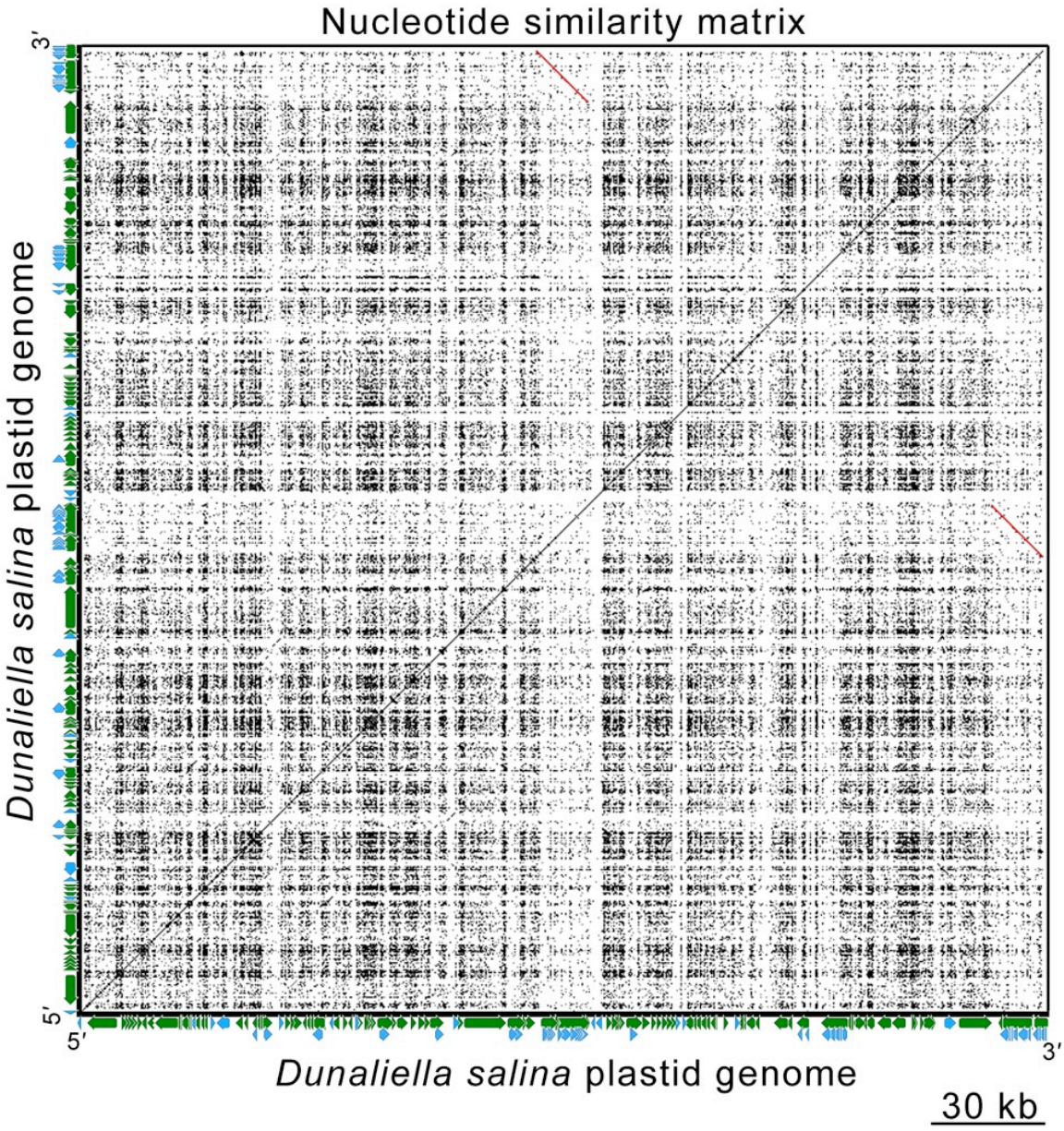

Figure 8.6. Dotplot Similarity Matrix of the *D. salina* Plastid Genome

The X- and Y-axes each represent the *D. salina* plastid genome (260 kb). For clarity, genetic maps of the *D. salina* ptDNA are placed below and beside these axes — on these maps coding regions are green and introns are blue (refer to Figure 8.2 for the complete annotation). Dots in the nucleotide similarity matrix represent regions of sequence similarity. The matrix was generated using a sliding-window size of 50. The inverted repeats are highlighted in red in the matrix.

CHAPTER 9: CONCLUSION

This thesis explored the architectural and nucleotide-sequence diversity of chloroplastidial genomes and attempted to pinpoint the fundamental evolutionary processes responsible for this diversity. The data presented throughout the seven primary chapters (Chapters 2 to 8) centered on the mitochondrial and plastid (and in some cases nuclear) DNAs from the green algal species *Polytomella capuana*, *Polytomella piriformis*, *Chlamydomonas reinhardtii*, *Volvox carteri*, and *Dunaliella salina*, and members of the land plant genus *Selaginella*. Analyses of these data revealed an impressive range of organelle genome sizes, compactnesses, conformations, and nucleotide compositions within the Chloroplastida, and underscored the role that non-adaptive forces have had in shaping this diversity.

Much of this dissertation documented extraordinary and unprecedented departures from what was considered "typical" of organelle genome architecture. It was shown, using the mitochondrial DNA (mtDNA) of *P. capuana* and the plastid DNAs (ptDNAs) of *Selaginella* species, that organelle genomes are not *all* AT-rich molecules, as some had believed, but that they can be driven to GC excess. Interestingly, we found that this GC richness is most prominent at silent nucleotide sites as well as those that experience RNA editing, suggesting a "neutral" underpinning for this nucleotide composition bias. Further data from *P. capuana* in conjunction with data on *P. piriformis* revealed new insights into mitochondrial genome linearization and fragmentation, and telomere evolution; these data indicate that the variation in genomic fragmentation among *Polytomella* taxa could be the result of recombination among GC-rich repeats and random genetic drift. Analyses of the *V. carteri* and *D. salina* mitochondrial and plastid DNAs exposed the most expanded organelle genomes observed from green algae, and eukaryotes in general, and demonstrated that this expansion can be manifested in either intron or repeat content. Finally, population genetic studies showed that for *C. reinhardtii* and *V. carteri* the product of the mutation rate and the effective genetic population size is positively correlated to genome compactness, and is, in my opinion, responsible for shaping the architectural characteristics of the organelle and nuclear DNAs of these two algae — particularly genome size. Thus, taken as whole, this dissertation can be condensed into two main points: i) that chlamydomonadalean algae boast some of the most varied and

unusual organelle genomes from all eukaryotes, and ii) that much of this variation can be framed in non-adaptive terms.

If, indeed, non-adaptive processes have tailored the architecture of genes and genomes in the Chloroplastida, and in life as a whole, then we must ask ourselves: what impact have these factors had on the higher-level organization of organisms, such as metabolism, multicellularity, and cellular differentiation? I would venture to say that non-adaptive factors have shaped or have had a cascading effect on all levels of cellular and organismal structure, both at the genomic and molecular level, as described in this thesis, as well as at the whole-organism level. If true, one may expect to find — if focusing on chlamydomonadalean algae for example — a positive scaling of genome size (and genomic ornamentations in general) with organism size, which tends to have an inverse relationship to the effective genetic population size. Thus, I hypothesize, that as more chlamydomonadalean algae are sampled, particularly members of the volvocine line (i.e., the lineage which contains *Volvox* and many other multicellular/colonial species) we will discover that levels of multicellularity and cellular differentiation (which should reflect an organism's size and therefore can be used, in a very general sense, as a proxy for population size) are positively correlated to the degree of genomic embellishment (e.g., the amount of noncoding, intron, and repeat DNA). Similarly, I would predict unicellular chlamydomonadalean algae, like *C. reinhardtii* and the *Vitreochlamys*, to have more compact and less exaggerated genomes than their multicellular/colonial relatives. It cannot be emphasized enough what an excellent group the chlamydomonadales are for addressing these types of evolutionary questions, especially the Volvocaceae, which contains species ranging in complexity from 16 cells up to 50,000 cells (Herron et al. 2009).

Studying the relationship between multicellularity and genome architecture within the chlamydomonadales may give insights into how the effective genetic population size can influence genome evolution, but it will not reveal the role that mutation rate plays in this process. One way to gain an understanding of mutation rate (aside from measuring nucleotide diversity) and how it is governing the genomic architecture of green algae would be to look at silent-site substitution rates between closely related species within a lineage. More specifically, one could measure the relative silent-site substitution rates

among the mitochondrial, plastid, and nuclear genomes. At present, there is lack of these types of data for green algae. However, the United States Department of Energy Joint Genome Institute has sequenced the genomes (mitochondrial, plastid, and nuclear) of two closely related *Micromonas* species, three *Ostreococcus* species, and two species of *Chlorella*; thus, the sequence data for these types of analyses are readily available. Knowledge of the relative silent-site substitution rates for various green algal lineages may help us understand why the genomes within some green algal species have opposing architectures (e.g., *C. reinhardtii*) whereas within other species (e.g., *Ostreococcus* spp.) all three genetic compartments have a similar architecture. Moreover, they may provide an explanation for why I found that when looking at silent-site nucleotide diversity values it is best to compare like with like, i.e., mtDNA versus mtDNA, ptDNA versus ptDNA, etc.

A final thought on which to conclude this thesis is the fact that most biologists, and scientists as a whole, and certainly many in the general public, have, as Michael Lynch put it, "an uncritical acceptance of natural selection as an explanatory force for all aspects of biodiversity" (Lynch 2007). In other words, when scientists observe diverse biological traits at the micro or macro level they often conclude that natural selection surely produced the end products. And although this may be true in many instances, I would argue that the adaptive hypotheses should be considered *only after* dismissing the non-adaptive "null" hypothesis. Even Darwin had a concept of how evolution could proceed under non-adaptive mechanisms. So let us leave the final word to him: "Variations neither useful nor injurious would not be affected by natural selection, and would be left either a fluctuating element, as perhaps we see in certain polymorphic species, or would ultimately become fixed, owing to the nature of the organism and the nature of conditions" (1859).

REFERENCES

Adams CR, Stamer KA, Miller JK, McNally JG, Kirk MM, Kirk DL (1990) Patterns of organellar and nuclear inheritance among progeny of two geographically isolated strains of *Volvox carteri*. Curr Genet 18:141-153.

Adl SM, Simpson AG, Farmer MA, et al. 28 co-authors) (2005) The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J Eukaryot Microbiol 52:399-451.

Almasan A, Mishra NC (1988) Molecular characterization of the mitochondrial DNA of *Schizosaccharomyces pombe* mutator strains. J Mol Biol. 202:725-734.

Aono N, Shimizu T, Inoue T, Shiraishi H (2002) Palindromic repetitive elements in the mitochondrial genome of *Volvox*. FEBS Lett 521:95-99.

Avron M, Ben-Amotz A (1992) *Dunaliella*: physiology, biochemistry, and biotechnology. Boca Raton, FL: CRC Press Inc.

Bah A, Bachand F, Clair E, Autexier C, Wellinger RJ (2004) Humanized telomeres and an attempt to express a functional human telomerase in yeast. Nucleic Acids Res 32:1917-1927.

Banks JA (2009) Selaginella and 400 million years of separation. Annu Rev Plant Biol 60:223-238.

Barkan A (1988) Proteins encoded by a complex chloroplast transcription unit are each translated from both monocistronic and polycistronic RNAs. EMBO J 7:2637-2644.

Baroudy BM, Venkatesan S, Moss B (1982) Incompletely base-paired flip-flop terminal loops link the two DNA strands of the vaccinia virus genome into one uninterrupted polynucleotide chain. Cell 28:315-324.

Baroudy BM, Venkatesan S, Moss B (1983) Structure and replication of vaccinia virus telomeres. Cold Spring Harb Symp Quant Biol 47:723-729.

Bateman AJ (1975) Simplification of palindromic telomere theory. Nature 253:379-380.

Bazin E, Glémin S, Galtier N (2006) Population size does not influence mitochondrial genetic diversity in animals. Science 312:570-572.

Bélanger AS, Brouard JS, Charlesbois P, Otis C, Lemieux C, Turmel M (2006) Distinctive architecture of the chloroplast genome in the chlorophycean green alga *Stigeoclonium helveticum*. Mol Genet Genomics 276:464-477.

Ben-Amotz A (2003) Industrial production of microalgal cell-mass and secondary products – major industrial species. In: Richmond A, editor. Handbook of Microalgal Culture: Biotechnology and Applied Phycology. Oxford, UK: Blackwell Publishing. p. 273-280.

Ben-Amotz A, Katz A, Avron M (1982) Accumulation of *ß*-carotene in halotolerant algae: Purification and characterization of *ß*-carotene-rich globules from *Dunaliella bardawil* (Chlorophyceae). J Phycol 18:529-537.

Ben-Amotz A, Polle JEW, Rao DVS (2009) The alga Dunaliella: biodiversity, physiology, genomics, and biotechnology. Enfield, NH: Science Publishers.

Bennett MD, Leitch IJ (2005) Genome size evolution in plants. In: Gregory TR, editor. The evolution of the genome. San Diego, CA: Elsevier. p. 89-162.

Birky CW, Fuerst P, Maruyama T (1989) Organelle gene diversity under migration, mutation, and drift: equilibrium expectations, approach to equilibrium, effects of heteroplasmic cells, and comparison to nuclear genes. Genetics 121:613-627.

Birky CW, Walsh JB (1992) Biased gene conversion, copy number, and apparent mutation rate differences within chloroplast and bacterial genomes. Genetics 130:677-683.

Bock R, Khan MS (2004) Taming plastids for a green future. Trends Biotechnol 22:311-318.

Boer PH, Gray MW (1986) The URF 5 gene of *Chlamydomonas reinhardtii* mitochondria: DNA sequence and mode of transcription. EMBO J 5:21-28.

Boer PH, Gray MW (1988) Genes encoding a subunit of respiratory NADH dehydrogenase (ND1) and a reverse transcriptase-like protein (RTL) are linked to ribosomal RNA pieces in *Chlamydomonas reinhardtii* mitochondrial DNA. EMBO J 7:3501-3508.

Boer PH, Gray MW (1988) Scrambled ribosomal RNA gene pieces in Chlamydomonas reinhardtii mitochondrial DNA. Cell 55:399-411.

Boer PH, Gray MW (1988) Transfer RNA genes and the genetic code in *Chlamydomonas reinhardtii* mitochondria. Curr Genet 14:583-590.

Boer PH, Gray MW (1991) Short dispersed repeats localized in spacer regions of *Chlamydomonas reinhardtii* mitochondrial DNA. Curr Genet 19:309-312.

Borowitzka MA, Borowitzka LJ (1988) *Dunaliella*. In: Borowitzka MA, Borowitzka LJ, editors. Microalgal Biotechnology. Cambridge, UK: Cambridge University Press. p. 27-58.

Borowitzka MA, Siva C (2007) The taxonomy of the genus *Dunaliella* (Chlorophyta, Dunaliellales) with emphasis on the marine and halophilic species. J Appl Phycol 19:567-590.

Borza T, Redmond EK, Laflamme M, Lee RW (2009) Mitochondrial DNA in the *Oogamochlamys* clade (Chlorophyceae): high GC content and unique genome architecture for green algae. J Phycol 45:1323-1334.

Boudreau E, Otis C, Turmel M (1994) Conserved gene clusters in the highly rearranged chloroplast genomes of *Chlamydomonas moewusii* and *Chlamydomonas reinhardtii*. Plant Mol Biol 24:585-602.

Boudreau E, Turmel M (1995) Gene rearrangements in *Chlamydomonas* chloroplast DNAs are accounted for by inversions and by the expansion/contraction of the inverted repeat. Plant Mol Biol 27:351-364.

Boudreau E, Turmel M (1996) Extensive gene rearrangements in the chloroplast DNAs of *Chlamydomonas* species featuring multiple dispersed repeats. Mol Biol Evol 13:233-243.

Boynton JE, Gillham NW, Harris EH, Hosler JP, Johnson AM, Jones AR, Randolph-Anderson BL, Robertson D, Klein TM, Shark KB (1988) Chloroplast transformation in *Chlamydomonas* with high velocity microprojectiles. Science 240:1534-1538.

Boynton JE, Harris EH, Burkhart BD, Lamerson PM, Gillham NW (1987) Transmission of mitochondrial and chloroplast genomes in crosses of *Chlamydomonas*. Proc Natl Acad Sci USA 84:2391-2395.

Breen AL, Glenn E, Yeager A, Olson MS (2009) Nucleotide diversity among natural populations of a North American poplar (*Populus balsamifera*, Salicaceae). New Phytol 182:763-773.

Bridge D, Cunningham CW, Schierwater B, DeSalle R, Buss LW (1992) Class-level relationships in the phylum Cnidaria: evidence from mitochondrial genome structure. Proc Natl Acad Sci USA 89:8750-8753.

Brodie R, Roper RL, Upton C (2004) JDotter: a java interface to multiple dotplots generated by dotter. Bioinformatics 20:279-281.

Bruick RK, Mayfield SP (1998) Processing of the *psbA* 5′ untranslated region in *Chlamydomonas reinhardtii* depends upon factors mediating ribosome association. J Cell Biol 143:1145-1153.

Burger G, Forget L, Zhu Y, Gray MW, Lang BF (2003) Unique mitochondrial genome architecture in unicellular relatives of animals. Proc Natl Acad Sci USA 100:892-897.

Burger G, Gray MW, Lang BF (2003) Mitochondrial genomes: anything goes. Trends Genet 19:709-716.

Burger G, Lang BF (2003) Parallels in genome evolution in mitochondria and bacterial symbionts. IUBMB Life 55:205-212.

Burger G, Zhu Y, Littlejohn TG, Greenwood SJ, Schnare MN, Lang BF, Gray MW (2000) Complete sequence of the mitochondrial genome of *Tetrahymena pyriformis* and comparison with *Paramecium aurelia* mitochondrial DNA. J Mol Biol 297:365-380.

Cai Z, Guisinger M, Kim HG, Ruck E, Blazier JC, McMurtry V, Kuehl JV, Boore J, Jansen RK (2008) Extensive reorganization of the plastid genome of *Trifolium subterraneum* (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions. J Mol Evol 67: 696-704.

Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, Du Y, Feng B, Lin N, Madabusi LV, Müller KM, Pande N, Shang Z, Yu N, Gutell RR (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics 3:15.

Cavalier-Smith T (1974) Palindromic base sequences and replication of eukaryote chromosome ends. Nature 250:467-470.

Cavalier-Smith T (1982) Skeletal DNA and the evolution of genome size. Annu Rev Biophys Bioeng 11:273-302.

Chen H, Morrell PL, de la Cruz M, Clegg MT (2008) Nucleotide diversity and linkage disequilibrium in wild avocado (*Persea americana* Mill.). J Hered 99:382-389.

Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, Boore JL, Jansen RK (2006) The complete chloroplast genome sequence of *Pelargonium* X *hortorum*: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. Mol Biol Evol 23:2175-2190.

Chung JD, Lin TP, Chen YL, Cheng YP, Hwang SY (2006) Phylogeographic study reveals the origin and evolutionary history of a *Rhododendron* species complex in Taiwan. Mol Phylogenet Evol 42:14-24.

Clark-Walker GD (1989) In vivo rearrangement of mitochondrial DNA in *Saccharomyces cerevisiae*. Proc Natl Acad Sci USA 86:8847-8851.

Clayton DA (1984) Transcription of the mammalian mitochondrial genome. Annu Rev Biochem 53:573-594.

Coleman AW, Thompson WF, Coff LJ (1991) Identification of the mitochondrial genome in the chrysophyte alga *Ochromonas danica*. J Protozool 38:129-135.

Colleaux L, Michel-Wolwertz MR, Matagne RF, Dujon B (1990) The apocytochrome *b* gene of *Chlamydomonas smithii* contains a mobile intron related to both *Saccharomyces* and *Neurospora* introns. Mol Gen Genet 223:288-296.

Covello PS, Gray MW (1993) On the evolution of RNA editing. Trends Genet 9:265-268.

Darwin C (1856) On the origin of species by means of natural selection, or the perservation of favoured races in the struggle for life. London, John Murray.

Daubin V, Moran NA (2004) Comment on "The origins of genome complexity". Science 306:978.

de Zamaroczy M, Bernardi G (1986) The GC clusters of the mitochondrial genome of yeasts and their evolutionary origin. Gene 41:1-22.

Denovan-Wright EM, Lee RW (1994) Comparative structure and genomic organization of the discontinuous mitochondrial ribosomal RNA genes of *Chlamydomonas eugametos* and *Chlamydomonas reinhardtii*. J Mol Biol 241:298-311.

Denovan-Wright EM, Nedelcu AM, Lee RW (1998) Complete sequence of the mitochondrial DNA of *Chlamydomonas eugametos*. Plant Mol Biol 36:285-295.

Dieckman CL, Gandy B (1987) Preferential recombination between GC clusters in yeast mitochondrial DNA. EMBO J 6:4197-4203.

Dietmaier W, Fabry S, Huber H, Schmitt R (1995) Analysis of a family of *ypt* genes and their products from *Chlamydomonas reinhardtii*. Gene 158:41-50.

Dinouël N, Drissi R, Miyakawa I, Sor F, Rousset S, Fukuhara H (1993) Linear mitochondrial DNAs of yeasts: closed-loop structure of the termini and possible linear-circular conversion mechanisms. Mol Cell Biol 13:2315-2323.

Doolittle WF, Sapienza C (1980) Selfish genes, the phenotype paradigm and genome evolution. Nature 284: 601-603.

Drouin G, Daoud H, Xia J (2008) Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. Mol Phylogenet Evol 49: 827-831.

Duret L, Eyre-Walker A, Galtier N (2006) A new perspective on isochore evolution. Gene 385:71-74.

Dybvig K, Voelker LL (1996) Molecular biology of mycoplasmas. Annu Rev Microbiol 50:25-57.

Duby F, Cardol P, Matagne RF, Remacle C (2001) Structure of the telomeric ends of mt DNA, transcriptional analysis and complex I assembly in the dum24 mitochondrial mutant of *Chlamydomonas reinhardtii.* Mol Genet Genomics 266: 109-114.

Ender A, Shierwater B (2003) Placozoa are not derived cnidarians: evidence from molecular morphology. Mol Biol Evol 20:130-134.

Eichler EE, Sankoff D. 2003. Structural dynamics of eukaryotic chromosome evolution. Science 301:793-797.

Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP, and related tools. Nat Protoc 2:953-971.

Eyre-Walker A (1993) Recombination and mammalian genome evolution. P Roy Soc B-Biol Sci 252:237-243.

Eyre-Walker A (1998) Problems with parsimony in sequences of biased base composition. J Mol Evol 47:686-690.

Fan J, Lee RW (2002) Mitochondrial genome of the colorless green alga *Polytomella parva*: two linear DNA molecules with homologous inverted repeat termini. Mol Biol Evol 19:999-1007.

Fan J, Schnare MN, Lee RW (2003) Characterization of fragmented mitochondrial ribosomal RNAs of the colorless green alga *Polytomella parva*. Nucleic Acids Res 31:769-778.

Feng S, Xue L, Liu H, Lu P (2009) Improvement of efficiency of genetic transformation for *Dunaliella salina* by glass beads method. Mol Biol Rep 36:1433-1439.

Fletcher SP, Muto, M., Mayfield SP (2007) Optimization of recombinant protein expression in the chloroplasts of green algae. Adv Exp Med Biol 616:90-98.

Forget L, Ustinova J, Wang Z, Huss VA, Lang BF (2002) *Hyaloraphidium curvatum*: a linear mitochondrial genome, tRNA editing, and an evolutionary link to lower fungi. Mol Biol Evol 19:310-319.

Förstemann K, Höss M, Lingner J (2000) Telomerase-dependent repeat divergence at the 3´ ends of yeast telomeres. Nucleic Acids Res 28:2690-2694.

Freyer R, Kiefer-Meyer MC, Kössel H (1997) Occurrence of plastid RNA editing in all major lineages of land plants. Proc Natl Acad Sci USA 94:6285-6290.

Galtier N (2004) Recombination, GC-content and the human pseudoautosomal boundary paradox. Trends Genet 20:347-349.

Galtier N, Duret L (2007) Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends Genet 23:273-277.

Galtier N, Piganeau G, Mouchiroud D, Duret L (2001) GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics 159:907-911.

Gerloff-Elias A, Spijkerman E, Pröschold T (2005) Effect of external pH on the growth, photosynthesis and photosynthetic electron transport of *Chlamydomonas acidophila* Negoro, isolated from an extremely acidic lake (pH 2.6). Plant Cell Environ 28:1218-1229.

Glass JI, Lefkowitz EJ, Glass JS, Heiner CR, Chen EY, Cassell GH (2000) The complete sequence of the mucosal pathogen *Ureaplasma urealyticum*. Nature 407: 757-762.

Glover KE, Spencer DF, Gray MW (2001) Identification and structural characterization of nucleus-encoded transfer RNAs imported into wheat mitochondria. J Biol Chem 276:639-648.

Goldschmidt-Clermont M (1991) Transgenic expression of aminoglycoside adenine transferase in the chloroplast: a selectable marker for site-directed transformation of *Chlamydomonas*. Nucleic Acids Res 19:4083-4089.

González MA, Gómez PI, JEW Polle (2009) Taxonomy and phylogeny of the genus *Dunaliella*. In: Ben-Amotz A, Polle JEW, Rao DVS, edtiors. The alga *Dunaliella*: biodiversity, physiology, genomics, and biotechnology. Enfield, NH: Science Publishers. p. 15-44.

González A, Talavera A, Almendral JM, Viñuela E (1986) Hairpin loop structure of African swine fever virus DNA. Nucleic Acids Res 14:6835-6844.

Gouveia L, Oliveira AC (2009) Microalgae as a raw material for biofuels production. J Ind Microbiol Biot 36:269-274.

Gray MW, Boer PH (1988) Organization and expression of algal (*Chlamydomonas reinhardti*i) mitochondrial DNA. Philos Trans R Soc Lond, B, Biol Sci 319:135-147.

Gregory TR (2001) Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol Rev Camb Philos Soc 76:65-101.

Gregory TR (2005a) Synergy between sequence and size in large-scale genomics. Nat Rev Genet. 6:699-708.

Gregory TR (2005b) Genome size evolution in animals. In: Gregory TR, editor. The evolution of the genome. San Diego, CA: Elsevier. p. 89-162.

Gregory TR, Witt JD (2008) Population size and genome size in fishes: a closer look. Genome 51:309-313.

Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR (2003) Rfam: an RNA family database. Nucleic Acids Res 31:439–441.

Grigoriev A (1998) Analyzing genomes with cumulative skew diagrams. Nucleic Acids Res 26:2286-2290.

Gross CH, Ranum LPW, Lefebvre PA (1988) Extensive restriction fragment length polymorphisms in a new isolate of *Chlamydomonas reinhardtii*. Curr Genet 13:503-508.

Gurley WB, O'Grady K, Czarnecka-Verner E, Lawit SI (2006) General transcription factors and the core promoter: ancient roots. In: Grasser KD, editor. Regulation of transcription in plants. Oxford, UK: Blackwell Publishing. p. 1-21.

Guisinger MM, Kuehl JV, Boore J, Jansen RK (2008) Genome-wide analyses of Geraniaceae plastid DNA reveal unprecedented patterns of increased nucleotide substitutions. Proc Natl Acad Sci USA 105:18424-18429.

Haley J, Bogorad L (1990) Alternative promoters are used for genes within maize chloroplast polycistronic transcription units. Plant Cell 2:323-333.

Handa H (2007) Linear plasmids in plant mitochondria: peaceful coexistences or malicious invasions? Mitochondrion 8:15-25.

Harris EH (2001) *Chlamydomonas* as a model organism. Annu Rev Plant Physiol Plant Mol Biol 52:160-174.

Harris EH (2009) The Chlamydomonas sourcebook second edition. San Diego, CA: Elsevier.

Herron MD, Hackett JD, Aylward FO, Michod RE (2009) Triassic origin and early radiation of multicellular volvocine algae. Proc Natl Acad Sci USA 106:3254-3258.

Hershberg R, Petrov DA (2008) Selection on codon bias. Annu Rev Genet 42:287-299.

Hinnebusch J, Barbour AG (1991) Linear plasmids of *Borrelia burgdorferi* have a telomeric structure and sequence similar to those of a eukaryotic virus. J Bacteriol 173:7233-7239.

Holmquist GP (1992) Chromosome bands, their chromatin flavors, and their functional features. Am J Hum Genet 51:17-37.

Hoshaw RW, Ettl H (1966) *Chlamydomonas smithii* sp. nov.: a chlamydomonad interfertile with *Chlamydomonas reinhardtii*. J Phycol 2:93-96.

Howe CJ, Barbrook AC, Koumandou VL, Nisbet RE, Symington HA, Wightman TF (2003) Evolution of the chloroplast genome. Philos Trans R Soc Lond, B, Biol Sci 358:99-106.

Huang S, Chiang YC, Schaal BA, Chou CH, Chiang TY (2001) Organelle DNA phylogeography of *Cycas taitungensis*, a relict species in Taiwan. Mol Ecol 10:2669-2681.

Hudson GS, Mason JG, Holton TA, Koller B, Cox GB, Whitfeld PR, Bottomley W (1987) A gene cluster in the spinach and pea chloroplast genomes encoding one CF_1 and three CF_0 subuntis of the H^+-ATP synthase complex and ribosomal protein S2. J Mol Biol 196:283-298.

Jabbari K, Bernardi G (2004) Body temperature and evolutionary genomics of vertebrates: a lesson from the genomes of *Takifugu rubripes* and *Tetraodon nigroviridis*. Gene 333:179-181.

Jiao HS, Hicks A, Simpson C, Stern DB (2004) Short dispersed repeats in the *Chlamydomonas* chloroplast genome are collocated with sites for mRNA 3′ end formation. Curr Genet 45:311-322.

Jobson RW, Qiu YL (2008) Did RNA editing in plant organellar genomes originate under natural selection or through genetic drift? Biol Direct 3:43.

Kairo A, Fairlamb AH, Gobright E, Nene V (1994) A 7.1 kb linear DNA molecule of *Theileria parva* has scrambled rDNA sequences and open reading frames for mitochondrially encoded proteins. EMBO J 13:898-905.

Katz A, Jiménez C, Pick U (1995) Isolation and characterization of a protein associated with carotene globules in the alga *Dunaliella bardawil.* Plant Physiol 108:1657-1664.

Katz LA, Curtis EA, Pfunder M, Landweber LF (2000) Characterization of novel sequences from distantly related taxa by walking PCR. Mol Phylogenet Evol 14:318-321.

Kayal E, Lavrov DV (2008) The mitochondrial genome of *Hydra oligactis* (Cnidaria, Hydrozoa) sheds new light on animal mtDNA evolution and cnidarian phylogeny. Gene 410:177-86.

Kellogg CA, Paul JH (2002) Degree of ultraviolet radiation damage and repair capabilities are related to G+C content in marine vibriophages. Aquat Microb Ecol 27:13-20.

Kenrick P, Crane PR (1997) The origin and early evolution of plants on land. Nature 389:33-39.

Khakhlova O, Bock R (2006) Elimination of deleterious mutations in plastid genomes by gene conversion. Plant J 46:85-94.

Kimura M (1983) The neutral theory of molecular evolution. Cambridge, UK: Cambridge University Press.

Kirk DL (1998) *Volvox*: molecular-genetic origins of multicellularity and cellular differentiation. New York, NY: Cambridge University Press.

Korall P, Kenrick P (2002) Phylogenetic relationships in Selaginellaceae based on *rbcL* sequences. Am J Bot 89:506-517.

Korall P, Kenrick P (2004) The phylogenetic history of Selaginellaceae based on DNA sequences from the plastid and nucleus: extreme substitution rates and rate heterogeneity. Mol Phylogenet Evol 31:852-864.

Korall P, Kenrick P, Therrien JP (1999) Phylogeny of Selaginellaceae: evaluation of generic/subgeneric relationships based on *rbcL* gene sequences. Int J Plant Sci 160:585-594.

Kroymann J, Zetsche K (1998) The mitochondrial genome of *Chlorogonium elongatum* inferred from the complete sequence. J Mol Evol 47:431-440.

Kugita M, Yamamoto Y, Fujikawa T, Matsumoto T, Yoshinaga K (2003) RNA editing in hornwort chloroplasts makes more than half the genes functional. Nucleic Acids Res 31:2417-2423.

Kurtz S, Schleiermacher C (1999) REPuter: fast computation of maximal repeats in complete genomes. Bioinformatics 15:426-427.

Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res 29:4633-4642.

Kusumi J, Tachida H (2005) Compositional properties of green-plant plastid genomes. J Mol Evol 60:417-425.

Laflamme M, Lee RW (2003) Mitochondrial genome conformation among CW-group chlorophycean algae. J Phycol 39:213-220.

Lane CE, van den Heuvel K, Kozera C, Curtis BA, Parsons B, Bowman S, Archibald JM (2007) Nucleomorph genome of *Hemiselmis andersenii* reveals complete intron loss and compaction as a driver of protein structure and function. Proc Natl Acad Sci USA 104:19908-19913.

Lang BF, Laforest MJ, Burger G (2007) Mitochondrial introns: a critical view. Trends Genet 23:119-125.

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

Liss M, Kirk DL, Beyser K, Fabry S (1997) Intron sequences provide a tool for high-resolution phylogenetic analysis of volvocine algae. Curr Genet 31:214-227.

Lynch M (2002) Intron evolution as a population-genetic process. Proc Natl Acad Sci USA 99:6118-6123.

Lynch M (2006) Streamlining and simplification of microbial genome architecture. Annu Rev Microbiol 60:327-349.

Lynch M (2007) The origins of genome architecture. Sunderland, MA: Sinauer Associates.

Lynch M, Conery JS (2003) The origins of genome complexity. Science 302:1401-1404.

Lynch M, Koskella B, Schaack S (2006) Mutation pressure and the evolution of organelle genomic architecture. Science 311:1727-1730.

Mabberley DJ (1997) The Plant-book: a Portable Dictionary of the Vascular Plants. New York, NY: Cambridge University Press.

Malek O, Lättig K, Hiesel R, Brennicke A, Knoop V (1996) RNA editing in bryophytes and a molecular phylogeny of land plants. EMBO J 15:1403-1411.

Maliga P (2004) Plastid transformation in higher plants. Annu Rev Plant Biol 55:289-313.

Mallet MA, Lee RW (2006) Identification of three distinct *Polytomella* lineages based on mitochondrial DNA features. J Eukaryot Microbiol 53:79-84.

Maul JE, Lilly JW, Cui L, dePamphilis CW, Miller W, Harris EH, Stern DB (2002) The *Chlamydomonas reinhardtii* plastid chromosome: islands of genes in a sea of repeats. Plant Cell 14:2659-2679.

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 351:652-654.

Meiklejohn CD, Montooth KL, Rand DM (2007) Positive and negative selection on the mitochondrial genome. Trends Genet 23:259-263.

Merchant SS, Prochnik SE, Vallon O, et al. 117 co-authors (2007) The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science 318:245-250.

Michaelis G, Vahrenholz C, Pratje E (1990) Mitochondrial DNA of *Chlamydomonas reinhardtii*: the gene for apocytochrome b and the complete functional map of the 15.8 kb DNA. Mol Gen Genet 223:211-216.

Mirsky AE, Ris H (1951) The deoxyribonucleic acid content of animal cells and its evolutionary significance. J Gen Physiol 34:451–462.

Miyashita S, Hirochika H, Ikeda JE, Hashiba T (1990) Linear plasmid DNAs of the plant pathogenic fungus *Rhizoctonia solani* with unique terminal structures. Mol Gen Genet 220:165-171.

Miyata Y, Sugita M (2004) Tissue- and stage-specific RNA editing of *rps14* transcripts in moss (*Physcomitrella patens*) chloroplasts. J Plant Physiol 161:113-115.

Modern CW, Wofe K, dePamphilis CW, Palmer JD (1991) Plastid translation and transcription genes in a non-photosynthetic plant: intact, missing and pseudo genes. EMBO J 10:3281-3288.

Moore LJ, Coleman AW (1989) The linear 20 kb mitochondrial genome of *Pandorina morum* (Volvocaceae, Chlorophyta). Plant Mol Biol 13:459-465.

Moran NA (2002) Microbial minimalism: genome reduction in bacterial pathogens. Cell 108:583-586.

Morton BR (1993) Chloroplast DNA codon use: evidence for selection at the psb A locus based on tRNA availability. J Mol Evol 37:273-280.

Morton BR (1998) Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. J Mol Evol 46:449-459.

Moulton TP, Borowitzka LJ, Vincent DJ (1987) The mass culture of *Dunaliella salina* for *ß*-carotene: from pilot plant to production plant. *Hydrobiologia* 151/152:99-105.

Muir G, Filatov D (2007) A selective sweep in the chloroplast DNA of dioecious *Silene* (section *Élisanthe*). Genetics 177:1239-1247.

Nakada T, Misawa K, Nozaki H (2008) Molecular systematics of Volvocales (Chlorophyceae, Chlorophyta) based on exhaustive 18S rRNA phylogenetic analyses. Mol Phylogenet Evol 48:281-291.

Nakayama T, Watanabe S, Mitsui K, Uchida H, Inouye I (1996) The phylogenetic relationship between the Chlamydomonadales and Chlorococcales inferred from 18S rDNA sequence data. Phycol Res 44:47-55.

Nakazono M, Tsutsumi N, Sugiura M, Hirai A (1995) A small repeated sequence contains the transcription initiation sites for both *trnfM* and *rrn26* in rice mitochondria. Plant Mol Biol 28:343-346.

Nash EA, Barbrook AC, Edwards-Stuart RK, Bernhardt K, Howe CJ, Nisbet RE (2007) Organization of the mitochondrial genome in the dinoflagellate *Amphidinium carterae*. Mol Biol Evol 24:1528-1536.

Nedelcu AM (1997) Fragmented and scrambled mitochondrial ribosomal RNA coding regions among green algae: a model for their origin and evolution. Mol Biol Evol 14:506-517.

Nedelcu AM (1998) Contrasting mitochondrial genome organizations and sequence affiliations among green algae: potential factors, mechanisms, and evolutionary scenarios. J Phycol 34:16-28.

Nedelcu AM, Lee RW (1998) Short repetitive sequences in green algal mitochondrial genomes: potential roles in mitochondrial genome evolution. Mol Biol Evol 15:690-701.

Nedelcu AM, Lee RW, Lemieux C, Gray MW, Burger G (2000) The complete mitochondrial DNA sequence of *Scenedesmus obliquus* reflects an intermediate stage in the evolution of the green algal mitochondrial genome. Genome Res 10:819-831.

Nei M (1987) Molecular evolutionary genetics. New York, NY: Columbia University Press.

Nosek J, Tomáska L (2002) Mitochondrial telomeres: alternative solutions to the end-replication problem. In: Krupp G, Parwaresch R, editors. Telomerases, telomeres and cancer. New York, NY: Kluwer Academic/Plenum Publishers. p. 396–417.

Nosek J, Tomáska L (2003) Mitochondrial genome diversity: evolution of the molecular architecture and replication strategy. Curr Genet 44:73-84.

Nosek J, Tomáska L, Kucejová B (2004a) The chromosome end replication: lessons from mitochondrial genetics. J Appl Biomed 2:71-79.

Nosek J, Novotna M, Hlavatovicova Z, Ussery DW, Fajkus J, Tomáska L (2004b) Complete DNA sequence of the linear mitochondrial genome of the pathogenic yeast *Candida parapsilosis*. Mol Genet Genomics 272:173-180.

Ogata H, Audic S, Renesto-Audiffren P, et al. 11 co-authors (2001) Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. Science 293:2093-2098.

Olovnikov AM (1971) Principle of marginotomy in template synthesis of polynucleotides. Dokl Akad Nauk SSSR 201:1496–1499.

Oren A (2005) A hundred years of *Dunaliella* research: 1905-2005. Saline Syst 1:2.

Orgel LE, Crick FH (1980) Selfish DNA: the ultimate parasite. Nature 284:604-607.

Palmer JD (1985) Comparative organization of chloroplast genomes. Annu Rev Genet 19:325-354.

Palmer JD (1991) Plastid chromosomes: structure and evolution. In: Bogorad LK, Vasil I, editors. Cell culture and somatic cell genetics of plants, Volume 7. San Diego, CA: Elsevier. p. 5-53.

Palmer JD, Stein DB (1986) Conservation of chloroplast genome structure among vascular plants. Curr Genet 10:823-833.

Polle JEW, Tran D, Ben-Amotz A (2009) History, distribution, and habitats of algae of the genus *Dunaliella* Teodoresco (Chlorophyceae). In: Ben-Amotz A, Polle JEW, Rao DVS, editors. The alga *Dunaliella*: biodiversity, physiology, genomics, and biotechnology. Enfield, NH: Science Publishers. p. 1-14.

Polle JEW, Qin S (2009) Development of genetics and molecular tool kits for species of the unicellular green alga *Dunaliella* (Chlorophyta). In: Ben-Amotz A, Polle JEW, Rao DVS, editors. The alga *Dunaliella*: biodiversity, physiology, genomics, and biotechnology. Enfield, NH: Science Publishers. p. 403-422.

Pombert JF, Beauchamp P, Otis C, Lemieux C, Turmel M (2006) The complete mitochondrial DNA sequence of the green alga *Oltmannsiellopsis viridis*: evolutionary trends of the mitochondrial genome in the Ulvophyceae. Curr Genet 50:137-147.

Popescu CE, Lee RW (2007) Mitochondrial genome sequence evolution in Chlamydomonas. Genetics 175:819-826.

Popescu CE, Borza T, Bielawski JP, Lee RW (2006) Evolutionary rates and expression level in Chlamydomonas. Genetics 172:1567-1576.

Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817-818.

Pringsheim EG (1955) The Genus *Polytomella*. J Protozool 2:137–145.

Pringsheim EG (1963) Farblose Algen. Ein Beitrag zur Evolutionsforschung. Gustav Fischer, Stuttgart.

Pritchard AE, Cummings DJ (1981) Replication of linear mitochondrial DNA from *Paramecium*: sequence and structure of the initiation-end crosslink. Proc Natl Acad Sci USA 78:7341-7345.

Pröschold T, Harris EH, Coleman A (2005) Portrait of a species: *Chlamydomonas reinhardtii*. Genetics 170:1601-1610.

Pröschold T, Marin B, Schlösser UG, Melkonian M (2001) Molecular phylogeny and taxonomic revision of *Chlamydomonas* (Chlorophyta). I. Emendation of *Chlamydomonas* Ehrenberg and *Chloromonas* Gobi, and description of *Oogamochlamys* gen. nov. and *Lobochlamys* gen. nov. Protist 152:265-300.

Provasoli L, Pintner IJ (1960) Artificial media for fresh-water algae: problems and suggestions. In: Tyron CA, Hartman RT, editors. The ecology of algae. Pittsburgh, PA: University of Pittsburgh Press. p. 84-96.

Pryer KM, Schuettpelz E, Wolf PG, Schneider H, Smith AR, Cranfill R (2004) Phylogeny and evolution of ferns (monilophytes) with a focus on the early leptosporangiate divergences. Am J Bot 91:1582-1598.

Purton S (2007) Tools and techniques for chloroplast transformation of *Chlamydomonas*. Adv Exp Med Biol 616:34-45.

Quang ND, Ikeda S, Harada K (2009) Patterns of nucleotide diversity at the methionine synthase locus in fragmented and continuous populations of a wind-pollinated tree, *Quercus mongolica* var. *crispula*. J Hered 100:762-770.

Rand DM, Kann LM (1998) Mutation and selection at silent and replacement sites in the evolution of animal mitochondrial DNA. Genetica 103:393-407.

Raubeson LA, Jansen RK (1992) Chloroplast DNA evidence on the ancient evolutionary split in vascular land plants. Science 255:1697-1699.

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

Rocha EP, Danchin A (2002) Base composition bias might result from competition for metabolic resources. Trends Genet 18:291-294.

Rosenberg JN, Oyler GA, Wilkinson L, Betenbaugh MJ (2008) A green light for engineered algae: redirecting metabolism to fuel a biotechnology revolution. Curr Opin in Biotech 19:430-436.

Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496-2497.

Rohozinski J, Girton LE, Van Etten JL (1989) *Chlorella* viruses contain linear nonpermuted double-stranded DNA genomes with covalently closed hairpin ends. Virology 168:363-369.

Rüdinger M, Funk HT, Rensing SA, Maier UG, Knoop V (2009) RNA editing: only eleven sites are present in the *Physcomitrella patens* mitochondrial transcriptome and a universal nomenclature proposal. Mol Genet Genomics 281:473-481.

Ryan R, Grant D, Chiang KS, Swift H (1978) Isolation and characterization of mitochondrial DNA from *Chlamydomonas reinhardtii*. Proc Natl Acad Sci USA 75:3268-3272.

Sakamoto W, Sturm NR, Kindle KL, Stern DB (1994) *petD* mRNA maturation in *Chlamydomonas reinhardtii* chloroplasts: role of 5′ endonucleolytic processing. Mol Cell Biol 14:6180-6186.

See DR, Brooks S, Nelson JC, Brown-Guedira G, Friebe B, Gill BS (2006) Gene evolution at the ends of wheat chromosomes. Proc Natl Acad Sci USA 103:4162-4167.

Sessoms AH, Huskey RJ (1973) Genetic control of development in Volvox: isolation and characterization of morphogenetic mutants. Proc Natl Acad Sci USA 70:1335-58.

Shao Z, Graf S, Chaga OY, Lavrov DV (2006) Mitochondrial genome of the moon jelly *Aurelia aurita* (Cnidaria, Scyphozoa): A linear DNA molecule encoding a putative DNA-dependent DNA polymerase. Gene 381:92-101.

Shaw J, Lickey EB, Beck JT, Farmer SB, Liu W, Miller J, Siripun KC, Winder CT, Schilling EE, Small RL (2005) The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. Am J Bot 9:142-166.

Shukla GC, Nene V (1998) Telomeric features of *Theileria parva* mitochondrial DNA derived from cycle sequence data of total genomic DNA. Mol Biochem Parasitol 95:159-63.

Simon DM, Clark N, McNeil BA, Johnson I, Pantuso D, Dai L, Chai D, Zimmerly S (2008) Group II introns in Eubacteria and Archaea: ORF-less introns and new varieties. *RNA* 14:1704-1713.

Singer CE, Ames BN (1970) Sunlight ultraviolet and bacterial DNA base ratios. Science 170:822-825.

Smith (2009) Unparalleled GC content in the plastid DNA of *Selaginella*. Plant Mol Biol 71:627-639.

Smith DR, Lee RW (2008a) Mitochondrial genome of the colorless green alga *Polytomella capuana*: a linear molecule with an unprecedented GC content. Mol Biol Evol 25:487-496.

Smith DR, Lee RW (2008b) Nucleotide diversity in the mitochondrial and nuclear compartments of *Chlamydomonas reinhardtii*: investigating the origins of genome architecture. BMC Evol Biol 8:156.

Smith DR, Lee RW (2009a) Nucleotide diversity in the *Chlamydomonas reinhardtii* plastid genome: addressing the mutational-hazard hypothesis. BMC Evol Biol 9:120.

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

Smith DR, Lee RW (2010) Low nucleotide diversity for the expanded organelle and nuclear genomes of *Volvox carteri* supports the mutational-hazard hypothesis. Mol Biol Evol *In press.*

Smith DR, Hua J, Lee RW (2010a) Evolution of linear mitochondrial DNA in three known lineages of *Polytomella*. Curr Genet *in press*.

Smith DR, Lee RW, Cushman JC, Magnuson JK, Tran D, Polle JEW (2010b) The *Dunaliella salina* organelle genomes: large sequences, inflated with intronic and intergenic DNA. BMC Plant Biol 10:83.

Spanier JG, Graham JE, Jarvik JW (1992) Isolation and preliminary characterization of three *Chlaymydomonas* strains interfertile with *Chlamydomonas reinhardtii* (Chlorophyta). J Phycol 28:822-828.

Starr RC. 1969. Structure, reproduction, and differentiation in *Volvox carteri* f. *nagariensis* Iyengar, strains HK 9 and 10. Arch Protistenkd. 111:204-222.

Steinhauser S, Beckert S, Capesius I, Malek O, Knoop V (1999) Plant mitochondrial RNA editing. J Mol Evol 48:303-312.

Sugase Y, Hirono M, Kindle KL, Kamiya R (1995) Cloning and characterization of the actin-encoding gene of *Chlamydomonas reinhardtii*. Gene 168:117-121.

Supek F, Vlahovicek (2004) INCA: synonymous codon usage analysis and clustering by means of self-organizing map. Bioinformatics 20:2329-2330.

Swofford DL (2003) PAUP*: Phylogenetic analysis using parsimony (and other methods, 4.0 Beta. Sunderland, MA: Sinauer Associates.

Tafresh AH, Shariati M (2009) *Dunaliella* biotechnology: methods and applications. J Appl Microbiol 107:14-35.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595.

Tajima F (1993) Measurement of DNA polymorphisms. In: Takahat N, Clark AG, editors. *Mechanisms of molecular evolution*. Sunderland MA: Sinauer Associates. p. 37-59.

Tatarenkov A, Avise JC (2007) Rapid concerted evolution in animal mitochondrial DNA. Philos Trans R Soc Lond, B, Biol Sci 274:1795-1798.

Teodoresco EC (1905) Organisation et développement du *Dunaliella*, nouveau genre de Volvocacée-Polyblépharidée. Beih Bot Zentralblatt Bd 18 Abt 1:215-232.

Thomas FC (1971) The genetic organization of chromosomes. Annu Rev Genet 5:237-256.

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673-4680.

Tillich M, Lehwark P, Morton BR, Maier UG (2006) The evolution of chloroplast RNA editing. Mol Biol Evol 23:1912-1921.

Tourasse NJ, Kolstø AB (2008) Survey of group I and group II introns in 29 sequenced genomes of the *Bacillus cereus* group: insights into their spread and evolution. Nucleic Acids Res 14:4529-4548.

Traktman P (1996) Poxvirus DNA replication. In: DePamphilis ML, editor. DNA replication in eukaryotic cells. Cold Spring Harbor, NY: Cold Spring Harbour Laboratory Press. p. 775-798.

Tsuji S, Ueda K, Nishiyama T, Hasebe M, Yoshikawa S, Konagaya A, Nishiuchi T, Yamaguchi K (2007) The chloroplast genome from a lycophyte (microphyllophyte), *Selaginella uncinata*, has a unique inversion, transposition and many gene losses. J Plant Res 120:281-290.

Turmel M, Bellemare G, Lemieux C (1987) Physical mapping of differences between the chloroplast DNAs of the interfertile algae *Chlamydomonas eugametos* and *Chlamydomonas moewusii*. Curr Genet 11:543-552.

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

Turmel M, Otis C, Lemieux C (2010) A Deviant Genetic Code in the Reduced Mitochondrial Genome of the Picoplanktonic Green Alga *Pycnococcus provasolii*. J Mol Evol 70:203-214.

Vahrenholz C, Riemen G, Pratje E, Dujon B, Michaelis G (1993) Mitochondrial DNA of *Chlamydomonas reinhardtii*: the structure of the ends of the linear 15.8-kb genome suggests mechanisms for DNA replication. Curr Genet 24:241-247.

Verma D, Daniell H (2007) Chloroplast vector systems for biotechnology applications. Plant Physiol 145:1129-1143.

Voigt O, Erpenbeck D, Wörheide G (2008) A fragmented metazoan organellar genome: the two mitochondrial chromosomes of *Hydra magnipapillata*. BMC Genomics 9:350.

Walker TL, Black D, Becker DK, Dale JL, Collet C (2005) Isolation and characterization of components of the *Dunaliella tertiolecta* chloroplast genome. J Appl Phycol 17:495-508.

Walsh JB (1992) Intracellular selection, conversion bias, and the expected substituion rate of organelle genes. Genetics 130:939-946.

Warrior R, Gall J (1985) The mitochondrial DNA of *Hydra attenuata* and *Hydra littoralis* consists of two linear molecules. Arch Sci Geneva 38:439-445.

Watson JD (1972) Origin of concatemeric T7 DNA. Nature New Biol 239:197-201.

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7:256-276

Weiller G, Schueller CME, Schweyen RJ (1989) Putative target sites for mobile G + C rich clusters in yeast mitochondrial DNA: single elements and tandem arrays. Mol Gen Genet 218:272-283.

Wolfe KH, Li W-H, Sharp PM (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci USA 84:9054-9058.

Wolfe PG, Rowe CA, Hasebe M (2004) High levels of RNA editing in a vascular plant chloroplast genome: analysis of transcripts from the fern *Adiantum capillus-veneris*. Gene 339:89-87.

Wolf PG, Karol KG, Mandoli DF, Kuehl J, Arumuganathan K, Ellis MW, Mishler BD, Kelch DG, Olmstead RG, Boore JL (2005) The first complete chloroplast genome sequence of a lycophyte *Huperzia lucidula* (Lycopodiaceae). Gene 350:117-128.

Wright SI, Nano N, Foxe JP, Dar VU (2008) Effective population size and tests of neutrality at cytoplasmic genes in *Arabidopsis*. Genet Res 90:119-128.

Wu SH, Hwang CY, Lin TP, Chung JD, Cheng YP, Hwang SY (2006) Contrasting phylogeographic patterns of two closely related species, *Machilus thunbergii* and *Machilus kusanoi* (Lauraceae), in Taiwan. J Biogeogr 33:936-947.

Xia X, Xie Z (2001) DAMBE: Data analysis in molecular biology and evolution. J Hered 92:371-373.

Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31:3406-3415.

APPENDICES

Copyright Permission Letters

Copyright Permission for Chapter 2

OXFORD UNIVERSITY PRESS LICENSE
TERMS AND CONDITIONS

Mar 15, 2010

This is a License Agreement between David R Smith ("You") and Oxford University Press ("Oxford University Press") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Oxford University Press, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	2390320739717
License date	Mar 15, 2010
Licensed content publisher	Oxford University Press
Licensed content publication	Molecular Biology and Evolution
Licensed content title	Mitochondrial Genome of the Colorless Green Alga Polytomella capuana: A Linear Molecule with an Unprecedented GC Content
Licensed content author	David R. Smith, et. al.
Licensed content date	March 2008
Volume number	25
Issue number	3
Type of Use	Thesis/Dissertation
Requestor type	Academic/Educational institute
Format	Print
Portion	Full article
Will you be translating?	No
Author of this OUP article	Yes
Order reference number	
Title of your thesis / dissertation	The evolution of organelle genome architecture
Expected completion date	Aug 2010
Estimated size(pages)	300
Terms and Conditions	

Copyright Permission for Chapter 3

SPRINGER LICENSE
TERMS AND CONDITIONS

Jun 29, 2010

This is a License Agreement between David R Smith ("You") and Springer ("Springer") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Springer, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	2458180639757
License date	Jun 29, 2010
Licensed content publisher	Springer
Licensed content publication	Current Genetics
Licensed content title	Evolution of linear mitochondrial DNA in three known lineages of *Polytomella*
Licensed content author	David Roy Smith
Licensed content date	Jan 1, 2010
Type of Use	Thesis/Dissertation
Portion	Full text
Number of copies	4
Author of this Springer article	Yes and you are the sole author of the new work
Order reference number	
Title of your thesis / dissertation	The evolution of organelle genome architecture
Expected completion date	Aug 2010
Estimated size(pages)	300
Total	0.00 USD
Terms and Conditions	

Copyright Permission for Chapter 4

SPRINGER LICENSE
TERMS AND CONDITIONS

Mar 15, 2010

This is a License Agreement between David R Smith ("You") and Springer ("Springer") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Springer, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	2390311477409
License date	Mar 15, 2010
Licensed content publisher	Springer
Licensed content publication	Plant Molecular Biology
Licensed content title	Unparalleled GC content in the plastid DNA of *Selaginella*
Licensed content author	David Roy Smith
Licensed content date	Jan 1, 2009
Volume number	71
Issue number	6
Type of Use	Thesis/Dissertation
Portion	Full text
Number of copies	1
Author of this Springer article	Yes and you are the sole author of the new work
Order reference number	
Title of your thesis / dissertation	The evolution of organelle genome architecture
Expected completion date	Aug 2010
Estimated size(pages)	300
Total	0.00 CAD
Terms and Conditions	

Copyright Permission for Chapter 7

OXFORD UNIVERSITY PRESS LICENSE
TERMS AND CONDITIONS

Jun 15, 2010

This is a License Agreement between David R Smith ("You") and Oxford University Press ("Oxford University Press") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Oxford University Press, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	2450300109865
License date	Jun 15, 2010
Licensed content publisher	Oxford University Press
Licensed content publication	Molecular Biology and Evolution
Licensed content title	Low nucleotide diversity for the expanded organelle and nuclear genomes of Volvox carteri supports the mutational-hazard hypothesis
Licensed content author	David Roy Smith, et. al.
Licensed content date	April 29, 2010
Type of Use	Thesis/Dissertation
Institution name	
Title of your work	The evolution of organelle genome architecture
Publisher of your work	n/a
Expected publication date	Aug 2010
Permissions cost	0.00 USD
Value added tax	0.00 USD
Total	0.00 USD
Total	0.00 USD
Terms and Conditions	

The GC Content of Completely Sequenced Plastid Genomes

Taxon	Genome length (nt)	Fraction GC	GC1	GC2	GC3	GC3syn	GenBank accession #
Acorus americanus	153,819	0.386	0.461	0.382	0.322	0.301	NC_010093
Acorus calamus	153,821	0.386	0.461	0.382	0.322	0.301	NC_007407
Adiantum capillus-veneris	150,568	0.420	0.456	0.410	0.383	0.425	NC_004766
Aethionema cordifolium	154,168	0.366	0.446	0.373	0.294	0.285	NC_009265
Aethionema grandiflorum	154,243	0.368	0.446	0.375	0.297	0.290	NC_009266
Agrostis stolonifera	136,584	0.385	0.472	0.395	0.303	0.290	NC_008591
Amborella trichopoda	162,686	0.383	0.361	0.433	0.376	0.332	NC_005086
Aneura mirabilis	108,007	0.406	0.458	0.400	0.315	0.326	NC_010359
Angiopteris evecta	153,901	0.355	0.453	0.377	0.244	0.205	NC_008829
Anthoceros formosae	161,162	0.329	0.431	0.380	0.228	0.187	NC_004543
Arabidopsis thaliana	154,478	0.363	0.445	0.371	0.294	0.286	NC_000932
Arabis hirsuta	153,689	0.364	0.445	0.373	0.294	0.289	NC_009268
Atropa belladonna	156,687	0.376	0.452	0.376	0.311	0.297	NC_004561
Babesia bovis T2Bo	35,107	0.220	0.257	0.246	0.150	0.092	NC_011395
Barbarea verna	154,532	0.364	0.445	0.372	0.294	0.283	NC_009269
Bigelowiella natans	69,166	0.302	0.398	0.353	0.133	0.068	NC_008408
Brachypodium distachyon	135,199	0.386	0.468	0.394	0.312	0.293	NC_011032
Buxus microphylla	159,010	0.381	0.453	0.380	0.318	0.297	NC_009599
Calycanthus floridus var. *glaucus*	153,337	0.393	0.459	0.385	0.328	0.308	NC_004993
Capsella bursa-pastoris	154,490	0.366	0.448	0.373	0.297	0.287	NC_009270
Carica papaya	160,100	0.369	0.451	0.378	0.305	0.290	NC_010323
Ceratophyllum demersum	156,252	0.382	0.456	0.382	0.316	0.294	NC_009962
Chaetosphaeridium globosum	131,183	0.296	0.404	0.339	0.166	0.135	NC_004115
Chara vulgaris	184,933	0.262	0.414	0.344	0.192	0.136	NC_008097
Chlamydomonas reinhardtii	203,828	0.345	0.429	0.368	0.170	0.103	NC_005353
Chloranthus spicatus	157,772	0.389	0.460	0.387	0.317	0.298	NC_009598

Taxon	Genome length (nt)	Fraction GC	GC1	GC2	GC3	GC3syn	GenBank accession #
Chlorella vulgaris	150,613	0.316	0.453	0.359	0.227	0.219	NC_001865
Chlorokybus atmophyticus	152,254	0.362	0.479	0.380	0.243	0.196	NC_008822
Cicer arietinum	125,319	0.339	0.448	0.370	0.270	0.257	NC_011163
Citrus sinensis	160,129	0.385	0.460	0.384	0.322	0.319	NC_008334
Coffea arabica	155,189	0.374	0.456	0.381	0.300	0.290	NC_008535
Crucihimalaya wallichii	155,199	0.364	0.446	0.374	0.295	0.288	NC_009271
Cryptomeria japonica	131,810	0.354	0.459	0.365	0.279	0.262	NC_010548
Cucumis sativus	155,293	0.371	0.443	0.380	0.310	0.292	NC_007144
Cuscuta exaltata	125,373	0.381	0.462	0.389	0.312	0.325	NC_009963
Cuscuta gronovii	86,744	0.377	0.453	0.377	0.297	0.304	NC_009765
Cuscuta obtusiflora	85,286	0.378	0.453	0.379	0.297	0.305	NC_009949
Cuscuta reflexa	121,521	0.382	0.456	0.383	0.314	0.324	NC_009766
Cyanidioschyzon merolae strain 10D	149,987	0.376	0.485	0.377	0.257	0.256	NC_004799
Cyanidium caldarium	164,921	0.327	0.413	0.342	0.242	0.207	NC_001840
Cyanophora paradoxa	135,599	0.305	0.462	0.364	0.139	0.116	NC_001675
Cycas taitungensis	163,403	0.394	0.461	0.386	0.321	0.299	NC_009618
Daucus carota	155,911	0.377	0.453	0.384	0.301	0.291	NC_008325
Dioscorea elephantipes	152,609	0.372	0.450	0.378	0.297	0.281	NC_009601
Draba nemorosa	153,289	0.365	0.445	0.374	0.294	0.286	NC_009272
Drimys granadensis	160,604	0.388	0.458	0.384	0.324	0.302	NC_008456
Eimeria tenella strain Penn State	34,750	0.206	0.200	0.217	0.040	0.041	NC_004823
Emiliania huxleyi	105,309	0.368	0.493	0.380	0.244	0.204	NC_007288
Epifagus virginiana	70,028	0.360	0.384	0.339	0.280	0.301	NC_001568
Eucalyptus globulus subsp. *globulus*	160,286	0.369	0.450	0.381	0.302	0.283	NC_008115
Euglena gracilis	143,171	0.260	0.355	0.340	0.214	0.178	NC_001603
Euglena longa	73,345	0.224	0.272	0.255	0.126	0.095	NC_002652
Fagopyrum esculentum subsp. *ancestrale*	159,599	0.380	0.454	0.375	0.315	0.308	NC_010776

Taxon	Genome length (nt)	Fraction GC	GC1	GC2	GC3	GC3syn	GenBank accession #
Festuca arundinacea	136,040	0.384	0.471	0.397	0.304	0.293	NC_011713
Glycine max	152,218	0.354	0.439	0.369	0.275	0.262	NC_007942
Gossypium barbadense	160,317	0.372	0.455	0.378	0.315	0.304	NC_008641
Gossypium hirsutum	160,301	0.372	0.457	0.382	0.309	0.301	NC_007944
Gracilaria tenuistipitata var. *liui*	183,883	0.292	0.404	0.334	0.175	0.135	NC_006137
Guillardia theta	121,524	0.330	0.453	0.358	0.185	0.147	NC_000926
Guizotia abyssinica	151,762	0.376	0.453	0.380	0.303	0.286	NC_010601
Helianthus annuus	151,104	0.376	0.453	0.379	0.306	0.286	NC_007977
Helicosporidium sp. ex *Simulium jonesii*	37,454	0.269	0.330	0.294	0.126	0.128	NC_008100
Heterosigma akashiwo	159,370	0.304	0.440	0.352	0.170	0.137	NC_010772
Hordeum vulgare subsp. *vulgare*	136,462	0.383	0.471	0.396	0.300	0.285	NC_008590
Huperzia lucidula	154,373	0.362	0.462	0.383	0.258	0.230	NC_006861
Illicium oligandrum	148,553	0.390	0.461	0.388	0.325	0.300	NC_009600
Ipomoea purpurea	162,046	0.375	0.450	0.375	0.312	0.305	NC_009808
Jasminum nudiflorum	165,121	0.380	0.457	0.378	0.313	0.297	NC_008407
Lactuca sativa	152,765	0.375	0.460	0.386	0.308	0.289	NC_007578
Lemna minor	165,955	0.357	0.447	0.369	0.293	0.264	NC_010109
Lepidium virginicum	154,743	0.365	0.447	0.375	0.294	0.284	NC_009273
Leptosira terrestris	195,081	0.273	0.416	0.353	0.199	0.172	NC_009681
Liriodendron tulipifera	159,886	0.392	0.461	0.385	0.325	0.301	NC_008326
Lobularia maritima	152,659	0.365	0.447	0.374	0.296	0.286	NC_009274
Lolium perenne	135,282	0.382	0.468	0.390	0.307	0.288	NC_009950
Lotus japonicus	150,519	0.360	0.442	0.369	0.293	0.284	NC_002694
Manihot esculenta	161,453	0.359	0.454	0.378	0.287	0.281	NC_010433
Marchantia polymorpha	121,024	0.288	0.393	0.333	0.129	0.102	NC_001319
Medicago truncatula	124,033	0.340	0.452	0.372	0.270	0.256	NC_003119
Mesostigma viride	118,360	0.302	0.432	0.365	0.140	0.104	NC_002186

Taxon	Genome length (nt)	Fraction GC	GC1	GC2	GC3	GC3syn	GenBank accession #
Morus indica	158,484	0.364	0.451	0.375	0.294	0.277	NC_008359
Nandina domestica	156,599	0.383	0.455	0.499	0.543	0.297	NC_008336
Nasturtium officinale	155,105	0.364	0.445	0.372	0.295	0.284	NC_009275
Nephroselmis olivacea	200,799	0.421	0.535	0.400	0.357	0.333	NC_000927
Nicotiana sylvestris	155,941	0.378	0.450	0.502	0.536	0.301	NC_007500
Nicotiana tabacum	155,943	0.378	0.458	0.383	0.319	0.300	NC_001879
Nicotiana tomentosiformis	155,745	0.378	0.452	0.377	0.316	0.302	NC_007602
Nuphar advena	160,866	0.391	0.460	0.386	0.332	0.309	NC_008788
Nymphaea alba	159,930	0.392	0.459	0.385	0.334	0.312	NC_006050
Odontella sinensis	119,704	0.318	0.445	0.359	0.174	0.147	NC_001713
Oedogonium cardiacum	196,547	0.295	0.374	0.303	0.178	0.162	NC_011031
Oenothera argillicola	165,055	0.391	0.480	0.378	0.337	0.330	NC_010358
Oenothera biennis	164,807	0.391	0.479	0.380	0.336	0.329	NC_010361
Oenothera elata subsp. hookeri	165,728	0.391	0.479	0.379	0.336	0.328	NC_002693
Oenothera glazioviana	165,225	0.390	0.480	0.380	0.336	0.329	NC_010360
Oenothera parviflora	163,365	0.391	0.478	0.381	0.336	0.329	NC_010362
Olimarabidopsis pumila	154,737	0.365	0.446	0.373	0.296	0.288	NC_009267
Oltmannsiellopsis viridis	151,933	0.405	0.488	0.398	0.243	0.158	NC_008099
Oryza nivara	134,494	0.390	0.469	0.392	0.326	0.315	NC_005973
Oryza sativa Indica Group	134,496	0.390	0.477	0.397	0.308	0.297	NC_008155
Oryza sativa Japonica Group	134,525	0.390	0.469	0.392	0.326	0.309	NC_001320
Ostreococcus tauri	71,666	0.399	0.541	0.407	0.270	0.155	NC_008289
Panax ginseng	156,318	0.381	0.455	0.380	0.315	0.299	NC_006290
Pelargonium x hortorum	217,942	0.396	0.467	0.379	0.350	0.336	NC_008454
Phaeodactylum tricornutum	117,369	0.326	0.441	0.355	0.188	0.175	NC_008588
Phalaenopsis aphrodite subsp. *formosana*	148,964	0.367	0.456	0.386	0.320	0.300	NC_007499
Phaseolus vulgaris	150,285	0.354	0.435	0.366	0.271	0.254	NC_009259

Taxon	Genome length (nt)	Fraction GC	GC1	GC2	GC3	GC3syn	GenBank accession #
Physcomitrella patens subsp. *patens*	122,890	0.285	0.404	0.335	0.142	0.111	NC_005087
Picea sitchensis	109,798	0.387	*not available*	*not available*	*not available*	*not available*	NC_011152
Pinus contorta	115,615	0.381	*not available*	*not available*	*not available*	*not available*	NC_011153
Pinus gerardiana	116,997	0.385	*not available*	*not available*	*not available*	*not available*	NC_011154
Pinus koraiensis	117,190	0.388	0.413	0.407	0.364	0.405	NC_004677
Pinus krempfii	115,555	0.384	*not available*	*not available*	*not available*	*not available*	NC_011155
Pinus lambertiana	112,521	0.382	*not available*	*not available*	*not available*	*not available*	NC_011156
Pinus longaeva	115,896	0.384	*not available*	*not available*	*not available*	*not available*	NC_011157
Pinus monophylla	114,607	0.384	*not available*	*not available*	*not available*	*not available*	NC_011158
Pinus nelsonii	110,575	0.383	*not available*	*not available*	*not available*	*not available*	NC_011159
Pinus thunbergii	119,707	0.385	0.456	0.383	0.320	0.310	NC_001631
Piper cenocladum	160,624	0.383	0.461	0.389	0.328	0.311	NC_008457
Platanus occidentalis	161,791	0.380	0.457	0.382	0.321	0.299	NC_008335
Populus alba	156,505	0.367	0.451	0.378	0.298	0.286	NC_008235
Populus trichocarpa	157,033	0.367	0.449	0.377	0.301	0.289	NC_009143
Porphyra purpurea	191,028	0.330	0.447	0.360	0.205	0.165	NC_000925
Porphyra yezoensis	191,952	0.331	0.447	0.361	0.210	0.168	NC_007932
Pseudendoclonium akinetum	195,867	0.326	0.409	0.349	0.225	0.239	NC_008114
Psilotum nudum	138,829	0.360	0.439	0.372	0.293	0.262	NC_003386
Ranunculus macranthus	155,129	0.379	0.454	0.381	0.310	0.299	NC_008796
Rhodomonas salina	135,854	0.348	0.467	0.368	0.234	0.166	NC_009573
Saccharum hybrid cultivar SP-80-3280	141,182	0.384	0.471	0.396	0.320	0.308	NC_005878
Saccharum officinarum	141,182	0.384	0.464	0.392	0.327	0.316	NC_006084
Scenedesmus obliquus	161,452	0.269	0.393	0.337	0.143	0.092	NC_008101
Selaginella moellendorffii	173,780	0.510	0.559	0.508	0.448	0.464	FJ755183
Selaginella uncinata	144,170	0.548	0.588	0.547	0.493	0.510	AB197035
Solanum bulbocastanum	155,371	0.379	0.456	0.383	0.309	0.293	NC_007943

Taxon	Genome length (nt)	Fraction GC	GC1	GC2	GC3	GC3syn	GenBank accession #
Solanum lycopersicum	155,461	0.379	0.454	0.379	0.313	0.295	NC_007898
Solanum tuberosum	155,298	0.379	0.453	0.379	0.314	0.297	NC_008096
Sorghum bicolor	140,754	0.385	0.473	0.394	0.306	0.295	NC_008602
Spinacia oleracea	150,725	0.368	0.449	0.378	0.294	0.282	NC_002202
Staurastrum punctulatum	157,089	0.325	0.444	0.365	0.185	0.146	NC_008116
Stigeoclonium helveticum	223,902	0.289	0.385	0.345	0.181	0.155	NC_008372
Thalassiosira pseudonana	128,814	0.307	0.436	0.359	0.149	0.118	NC_008589
Theileria parva strain *Muguga*	39,579	0.195	0.230	0.219	0.131	0.113	NC_007758
Toxoplasma gondii RH	34,996	0.214	0.216	0.226	0.050	0.036	NC_001799
Trachelium caeruleum	162,321	0.383	0.460	0.378	0.316	0.303	NC_010442
Trifolium subterraneum	144,763	0.390	0.464	0.383	0.287	0.276	NC_011828
Triticum aestivum	134,545	0.383	0.470	0.391	0.307	0.289	NC_002762
Vaucheria litorea	115,341	0.279	0.409	0.337	0.086	0.072	NC_011600
Vitis vinifera	160,928	0.374	0.454	0.382	0.301	0.281	NC_007957
Welwitschia mirabilis	119,726	0.367	0.438	0.354	0.282	0.259	NC_010654
Zea mays	140,384	0.385	0.467	0.394	0.323	0.311	NC_001666
Zygnema circumcarinatum	165,372	0.311	0.457	0.376	0.222	0.160	NC_008117

Note: GC1, GC2, and GC3 are the fractional GC contents at first-, second-, and third-position codon sites, respectively. $GC3_{syn}$ is the proportion of G or C nucleotides at fourfold-degenerate sites, which are third-position codon sites that can tolerate any of the four nucleotides without altering the amino acid specified.

Available *RbcL* ptDNA Sequences from *Selaginella* Species

Taxon	Subgenus	Geographical distribution	Sequence length (nt)	G+C Fraction	GC1	GC2	GC3	$GC3_{syn}$	GenBank accession number
Selaginella acanthostachys	Stachygynandrum[1]	S. America[1]	1299	0.50	0.61	0.47	0.42	0.42	AJ295884
Selaginella alopecuroides	Stachygynandrum[1]	Borneo[1]	1287	0.50	0.61	0.47	0.42	0.43	AJ295875
Selaginella apoda	Stachygynandrum[1]	N. America[1]	1428	0.51	0.59	0.43	0.52	0.48	SLGCPRBCL
Selaginella arenicola	Tetragonostachys[2]	N. America[3]	480	0.52	0.55	0.45	0.55	0.60	AF419084
Selaginella arizonica	Tetragonostachys[1]	N. America[1]	1323	0.52	0.59	0.43	0.53	0.56	AF419078
Selaginella arsenei	Tetragonostachys[2]	Mexico[3]	1350	0.52	0.60	0.43	0.52	0.53	AF419056
Selaginella articulata	Stachygynandrum[1]	S. & C. America[1]	1272	0.56	0.61	0.44	0.62	0.63	AJ295894
Selaginella asprella	Tetragonostachys[2]	N. America, Mexico[3]	1350	0.52	0.60	0.44	0.52	0.54	AF419064
Selaginella australiensis	Stachygynandrum[1]	Australia[1]	1296	0.51	0.61	0.46	0.47	0.39	AJ295890
Selaginella balansae	Tetragonostachys[2]	Morocco[3]	1350	0.52	0.60	0.43	0.53	0.56	AF419080
Selaginella bigelovii	Tetragonostachys[2]	N. America[3]	1320	0.51	0.59	0.43	0.50	0.52	AF419082
Selaginella bombycina	Stachygynandrum[1]	S. & C. America[1]	1341	0.50	0.61	0.47	0.42	0.43	AJ010848
Selaginella brooksii	Stachygynandrum[2]	Borneo[3]	1284	0.50	0.61	0.47	0.41	0.41	AJ295876
Selaginella caffrorum	Tetragonostachys[2]	E. Africa & Arabia[3]	1165	0.50	0.59	0.41	0.51	0.53	AF419070
Selaginella ciliaris	*incertae sedis*	SE Asia, Australia[3]	1380	0.51	0.60	0.48	0.46	0.41	EU126658
Selaginella cinerascens	Tetragonostachys[2]	N. America, Mexico[3]	1350	0.52	0.60	0.44	0.52	0.56	AF419063
Selaginella deflexa	Selaginella[1]	Hawaii[1]	1338	0.51	0.60	0.45	0.48	0.47	AF093253
Selaginella densa	Tetragonostachys[2]	N. America[3]	1350	0.52	0.60	0.44	0.51	0.53	AF419069
Selaginella denticulata	Stachygynandrum[1]	Mediterrranean[1]	1293	0.53	0.62	0.51	0.45	0.49	AJ010853
Selaginella diffusa	Stachygynandrum[1]	S. & C. America[1]	1329	0.55	0.60	0.44	0.62	0.63	AJ010852
Selaginella digitata	Stachygynandrum[1]	Madagascar[1]	1284	0.51	0.60	0.45	0.48	0.48	AJ295895
Selaginella douglasii	Stachygynandrum[2]	N. America[3]	1335	0.54	0.61	0.53	0.47	0.51	AF419049
Selaginella dregei	Tetragonostachys[2]	Africa[3]	1350	0.52	0.60	0.44	0.53	0.54	AF419055
Selaginella echinata	Tetragonostachys[2]	Madagascar[3]	1350	0.51	0.59	0.44	0.51	0.51	AF419071
Selaginella eremophila	Tetragonostachys[2]	N. America, Mexico[3]	1332	0.52	0.59	0.44	0.53	0.56	AF419079
Selaginella erythropus	Stachygynandrum[1]	S. America[1]	1272	0.50	0.62	0.47	0.41	0.42	AJ295877
Selaginella exaltata	Stachygynandrum[1]	S. & C. America[1]	1314	0.51	0.59	0.44	0.50	0.47	AJ010849
Selaginella extensa	Tetragonostachys[2]	Mexico[3]	1026	0.52	0.60	0.46	0.51	0.53	AF419085

Taxon	Subgenus	Geographical distribution	Sequence length (nt)	G+C Fraction	GC1	GC2	GC3	$GC3_{syn}$	GenBank accession number
Selaginella firmuloides	Stachygynandrum[1]	New Caledonia[1]	1287	0.50	0.62	0.47	0.42	0.42	AJ295870
Selaginella flabellate	Stachygynandrum[1]	Grenada[1]	1296	0.50	0.61	0.47	0.42	0.43	AJ295885
Selaginella flagellata	Tetragonostachys[1]	Mexico to S. America[1]	1293	0.50	0.61	0.47	0.42	0.42	AJ295866
Selaginella fragilis	Stachygynandrum[1]	S. America[1]	1272	0.57	0.61	0.44	0.64	0.66	AJ295872
Selaginella frondosa	Stachygynandrum[1]	SE Asia[1]	1299	0.50	0.61	0.48	0.42	0.42	AJ295874
Selaginella gracillima	Ericetorum[1]	Australia, Tasmania[1]	1275	0.52	0.59	0.43	0.55	0.50	AJ010844
Selaginella grisea	Tetragonostachys[2]	*Stat not available*	1164	0.50	0.58	0.43	0.51	0.53	AF419072
Selaginella haematodes	Stachygynandrum[1]	S. & C. America[1]	1281	0.50	0.62	0.47	0.42	0.44	AJ010846
Selaginella hansenii	Tetragonostachys[2]	N. America[3]	1350	0.51	0.59	0.44	0.51	0.53	AF419057
Selaginella helioclada	Stachygynandrum[1]	Madagascar[1]	1272	0.51	0.59	0.45	0.48	0.49	AJ295896
Selaginella helvetica	Stachygynandrum[1]	E. Europe & N. Asia[1]	678	0.53	0.58	0.51	0.49	0.54	AJ295891
Selaginella imbricata	Stachygynandrum[1]	E. Africa & Arabia[1]	1266	0.50	0.59	0.45	0.47	0.49	AJ295897
Selaginella indica	Tetragonostachys[2]	*Stat not available*	1353	0.51	0.60	0.44	0.48	0.50	AF419052
Selaginella intermedia	*incertae sedis*	SE Asia[3]	1380	0.50	0.61	0.47	0.42	0.50	EU086853
Selaginella kerstingii	Stachygynandrum[1]	New Guinea[1]	1296	0.50	0.62	0.47	0.41	0.41	AJ295881
Selaginella kraussiana	Stachygynandrum[1]	S. & E. Africa and widely introduced[1]	1293	0.52	0.60	0.44	0.54	0.52	AJ010845
Selaginella landii	Tetragonostachys[2]	Mexico[3]	1323	0.52	0.59	0.43	0.53	0.56	AF419086
Selaginella lepidophylla	Stachygynandrum[1]	S. USA, C. America[1]	1353	0.52	0.61	0.43	0.50	0.50	AF419051
Selaginella leucobryoides	Tetragonostachys[2]	N. America[3]	1350	0.51	0.60	0.44	0.50	0.53	AF419068
Selaginella lingulata	Stachygynandrum[1]	S. & C. America[1]	1278	0.56	0.59	0.43	0.50	0.63	AJ295882
Selaginella longiaristata	Stachygynandrum[1]	SE Asia[1]	1299	0.50	0.61	0.47	0.42	0.41	AJ295873
Selaginella longipinna	Stachygynandrum[1]	Australia[1]	1293	0.50	0.61	0.47	0.42	0.41	AJ295860
Selaginella lyallii	Stachygynandrum[1]	Madagascar[1]	1248	0.53	0.59	0.43	0.56	0.52	AJ295898
Selaginella martensii	Stachygynandrum	Mexico & C. America[3]	1287	0.50	0.61	0.47	0.42	0.42	AJ295878
Selaginella mayeri	*incertae sedis*	SE Asia[3]	936	0.53	0.65	0.50	0.44	0.42	EU197125
Selaginella moellendorffii	Stachygynandrum[2]	SE Asia[3]	1428	0.50	0.59	0.47	0.42	0.44	FJ755183
Selaginella moratii	Stachygynandrum[1]	Madagascar[1]	1248	0.53	0.60	0.43	0.56	0.52	AJ295899
Selaginella moritziana	Heterostachys[1]	S. America[1]	1278	0.50	0.61	0.47	0.42	0.44	AJ010856

Taxon	Subgenus	Geographical distribution	Sequence length (nt)	G+C Fraction	GC1	GC2	GC3	$GC3_{syn}$	GenBank accession number
Selaginella mutica	Tetragonostachys[2]	N. America[3]	1338	0.51	0.60	0.43	0.52	0.53	AF419058
Selaginella myosurus	Stachygynandrum[1]	W. tropical Africa[1]	818	0.51	0.60	0.43	0.50	0.50	AJ295863
Selaginella nivea	Tetragonostachys[2]	E. Africa[3]	1164	0.50	0.58	0.42	0.51	0.53	AF419073
Selaginella njamnjamensis	Tetragonostachys[2]	Africa[3]	1353	0.52	0.60	0.44	0.51	0.51	AF419074
Selaginella novae-hollandiae	Heterostachys[1]	S. & C. America[1]	1290	0.50	0.60	0.47	0.43	0.43	AJ295865
Selaginella novae-hollandiae	Heterostachys[1]	S. & C. America[1]	1293	0.50	0.61	0.47	0.42	0.43	AJ295883
Selaginella oregana	Tetragonostachys[2]	N. America[3]	1341	0.51	0.60	0.44	0.51	0.51	AF419066
Selaginella pallescens	Stachygynandrum[1]	S. & C. America[1]	1341	0.50	0.60	0.46	0.43	0.44	AF419050
Selaginella peruviana	Tetragonostachys[2]	S. & C. America[3]	1323	0.52	0.59	0.43	0.53	0.56	AF419087
Selaginella pervillei	Stachygynandrum[1]	Madagascar[1]	1110	0.53	0.64	0.49	0.46	0.47	AJ295901
Selaginella phillipsiana	Tetragonostachys[2]	E. Africa[3]	1341	0.52	0.60	0.44	0.52	0.54	AF419061
Selaginella pilifera	Stachygynandrum[1]	S. USA & N. Mexico[1]	1284	0.51	0.60	0.45	0.50	0.50	AJ295862
Selaginella plana	Stachygynandrum[1]	SE Asia, introduced in the New World[1]	984	0.53	0.66	0.50	0.44	0.45	AJ295880
Selaginella polymorpha	Stachygynandrum[1]	Madagascar[1]	1245	0.53	0.60	0.43	0.56	0.53	AJ295900
Selaginella pulcherrima	Stachygynandrum[1]	Mexico[1]	1350	0.50	0.61	0.46	0.44	0.45	AJ010847
Selaginella pygmaea	Ericetorum[1]	South Africa, Australia[1]	1293	0.51	0.59	0.43	0.49	0.47	AJ295892
Selaginella radiata	Heterostachys[1]	S. America[1]	1228	0.50	0.61	0.46	0.43	0.45	AJ295867
Selaginella remotifolia	Stachygynandrum[1]	E. & SE Asia[1]	1098	0.52	0.60	0.41	0.54	0.54	AJ295864
Selaginella roxburghii	*incertae sedis*	SE Asia[3]	1245	0.50	0.59	0.47	0.42	0.47	EU140945
Selaginella rupestris	Tetragonostachys[1]	N. America[1]	1356	0.51	0.59	0.44	0.51	0.53	AF093255
Selaginella rupincola	Tetragonostachys[1]	S. USA[1]	1326	0.52	0.60	0.43	0.52	0.53	AF419083
Selaginella sanguinolenta	*Stat not available*	E. Asia[3]	1347	0.52	0.62	0.52	0.43	0.43	EU197124
Selaginella sartorii	Tetragonostachys[2]	S. & C. America[3]	1353	0.51	0.59	0.43	0.52	0.54	AF419054
Selaginella selaginoides	Selaginella[1]	Circumboreal[1]	1332	0.51	0.59	0.44	0.49	0.47	AF419048
Selaginella sericea	Stachygynandrum[1]	S. America[1]	1278	0.56	0.61	0.44	0.64	0.65	AJ295871
Selaginella shakotanensis	Tetragonostachys[2]	E. Asia[3]	1341	0.51	0.59	0.43	0.51	0.53	AF419059
Selaginella sibirica	Tetragonostachys[2]	N. America (Alaska), E. Asia[3]	1266	0.52	0.59	0.43	0.52	0.55	AF419076
Selaginella simplex	Heterostachys[1]	S. America[1]	1287	0.50	0.61	0.47	0.42	0.42	AJ295888

Taxon	Subgenus	Geographical distribution	Sequence length (nt)	G+C Fraction	GC1	GC2	GC3	$GC3_{syn}$	GenBank accession number
Selaginella sinensis	Stachygynandrum[1]	China[1]	1287	0.44	0.58	0.44	0.31	0.27	AJ295868
Selaginella stauntoniana	Stachygynandrum[1]	E. Asia[1]	1281	0.51	0.59	0.45	0.49	0.50	AJ295869
Selaginella steyermarkii	Tetragonostachys[2]	S. & C. America[3]	1347	0.52	0.60	0.45	0.51	0.53	AF419088
Selaginella suavis	Stachygynandrum[1]	S. America[1]	1296	0.56	0.61	0.45	0.63	0.63	AJ295886
Selaginella sulcata	Stachygynandrum[1]	S. America[1]	1299	0.56	0.60	0.44	0.62	0.63	AJ295887
Selaginella tamariscina	Stachygynandrum[1]	E. Asia, northern parts of SE Asia[1]	1296	0.51	0.59	0.46	0.47	0.48	AJ295861
Selaginella tortipila	Tetragonostachys[2]	N. America[3]	1350	0.52	0.60	0.43	0.53	0.55	AF419081
Selaginella uliginosa	Ericetorum[1]	Australia, Tasmania[1]	1284	0.52	0.59	0.43	0.54	0.55	AJ010843
Selaginella umbrosa	Stachygynandrum[1]	S. America[1]	1284	0.50	0.61	0.47	0.42	0.43	AJ295879
Selaginella uncinata	Stachygynandrum[2]	China, E. Asia[3]	1428	0.53	0.62	0.50	0.46	0.48	AB197035
Selaginella underwoodii	Tetragonostachys[2]	N. America[3]	849	0.51	0.57	0.43	0.53	0.56	AF419077
Selaginella utahensis	Tetragonostachys[2]	N. America[3]	1350	0.51	0.60	0.44	0.50	0.53	AF419067
Selaginella vardei	Tetragonostachys[2]	China, E. Asia[3]	1341	0.51	0.59	0.43	0.51	0.53	AF419060
Selaginella wallacei	Tetragonostachys[2]	N. America[3]	1350	0.51	0.60	0.43	0.52	0.53	AF419065
Selaginella watsonii	Tetragonostachys[2]	N. America[3]	1350	0.51	0.60	0.43	0.51	0.54	AF419090
Selaginella weatherbiana	Tetragonostachys[2]	N. America[3]	1335	0.51	0.60	0.43	0.51	0.52	AF419075
Selaginella wightii	Tetragonostachys[2]	SE Asia[3]	1341	0.52	0.60	0.44	0.52	0.54	AF419062
Selaginella wildenowii	Stachygynandrum[1]	SE Asia[1]	756	0.54	0.66	0.50	0.45	0.46	AJ295893
Selaginella wrightii	Tetragonostachys[2]	C. & N. America[3]	1038	0.52	0.59	0.44	0.54	0.60	AF419089

Note: GC1, GC2, and GC3 are the GC contents at first-, second-, and third-position codon sites, respectively. $GC3_{syn}$ is the proportion of G or C nucleotides at fourfold-degenerate sites, which are third-position codon sites that can tolerate any of the four nucleotides without altering the amino acid specified.

[1] This statistic came from the publication affiliated with the GenBank accession number.

[2] In instances where the GenBank accession number is not affiliated with a publication, information listed in the GenBank entry was used to determine subgenus.

[3] In instances where the GenBank accession number is not affiliated with a publication, Michael Hassier and Brian Swale's Checklist of ferns and fern allies (http://homepages.caverock.net.nz/~bj/fern/list.htm) was used to determine geographical distribution.

The Fraction of Noncoding DNA in Completely Sequenced Plastid Genomes

Taxon	Genome length (nt)	% Noncoding	Genbank accession number
PLASTID GENOMES			
Acorus americanus	153819	42	NC_010093
Acorus calamus	153821	42	NC_007407
Adiantum capillus-veneris	150568	39	NC_004766
Aethionema cordifolium	154168	41	NC_009265
Aethionema grandiflorum	154243	41	NC_009266
Agrostis stolonifera	136584	47	NC_008591
Amborella trichopoda	162686	44	NC_005086
Aneura mirabilis	108007	46	NC_010359
Angiopteris evecta	153901	46	NC_008829
Anthoceros formosae	161162	46	NC_004543
Arabidopsis thaliana	154478	41	NC_000932
Arabis hirsuta	153689	41	NC_009268
Atropa belladonna	156687	43	NC_004561
Barbarea verna	154532	41	NC_009269
Bigelowiella natans	69166	13	NC_008408
Buxus microphylla	159010	42	NC_009599
Calycanthus floridus var. *glaucus*	153337	41	NC_004993
Capsella bursa-pastoris	154490	41	NC_009270
Carica papaya	160100	43	NC_010323
Ceratophyllum demersum	156252	42	NC_009962
Chaetosphaeridium globosum	131183	31	NC_004115
Chara vulgaris	184933	51	NC_008097
Chlamydomonas reinhardtii	204159	56	FJ423446
Chloranthus spicatus	157772	41	NC_009598
Chlorella vulgaris	150613	36	NC_001865
Chlorokybus atmophyticus	152254	38	NC_008822
Cicer arietinum	125319	39	NC_011163
Citrus sinensis	160129	43	NC_008334
Coffea arabica	155189	41	NC_008535
Crucihimalaya wallichii	155199	41	NC_009271
Cryptomeria japonica	131810	39	NC_010548
Cucumis sativus	155293	42	NC_007144
Cuscuta exaltata	125373	45	NC_009963
Cuscuta gronovii	86744	29	NC_009765
Cuscuta obtusiflora	85286	29	NC_009949
Cuscuta reflexa	121521	44	NC_009766
Cyanidioschyzon merolae strain 10D	149887	5	NC_004799
Cyanidium caldarium	164921	11	NC_001840
Cyanophora paradoxa	135599	20	NC_001675
Cycas taitungensis	163403	37	NC_009618
Daucus carota	155911	43	NC_008325
Dioscorea elephantipes	152609	41	NC_009601
Draba nemorosa	153289	41	NC_009272
Drimys granadensis	160604	43	NC_008456

Taxon	Genome length (nt)	% Noncoding	Genbank accession number
Eimeria tenella strain Penn State	34750	5	NC_004823
Emiliania huxleyi	105309	13	NC_007288
Epifagus virginiana	70028	41	NC_001568
Eucalyptus globulus subsp. *globulus*	160286	43	NC_008115
Euglena gracilis	143171	52	NC_001603
Euglena longa	73345	32	NC_002652
Glycine max	152218	41	NC_007942
Gossypium barbadense	160317	43	NC_008641
Gossypium hirsutum	160301	43	NC_007944
Gracilaria tenuistipitata var. liui	183883	16	NC_006137
Guillardia theta	121524	11	NC_000926
Guizotia abyssinica	151762	41	NC_010601
Helianthus annuus	151104	41	NC_007977
Helicosporidium sp. ex *Simulium jonesii*	37454	5	NC_008100
Heterosigma akashiwo	159370	23	NC_010772
Hordeum vulgare subsp. *vulgare*	136462	47	NC_008590
Huperzia lucidula	154373	45	NC_006861
Illicium oligandrum	148553	44	NC_009600
Ipomoea purpurea	162046	40	NC_009808
Jasminum nudiflorum	165121	43	NC_008407
Lactuca sativa	152765	45	NC_007578
Lemna minor	165955	42	NC_010109
Lepidium virginicum	154743	41	NC_009273
Leptosira terrestris	195081	53	NC_009681
Liriodendron tulipifera	159886	44	NC_008326
Lobularia maritima	152659	40	NC_009274
Lolium perenne	135282	46	NC_009950
Lotus japonicus	150519	45	NC_002694
Manihot esculenta	161453	50	NC_010433
Marchantia polymorpha	121024	31	NC_001319
Medicago truncatula	124033	42	NC_003119
Mesostigma viride	118360	26	NC_002186
Morus indica	158484	43	NC_008359
Nandina domestica	156599	42	NC_008336
Nasturtium officinale	155105	41	NC_009275
Nephroselmis olivacea	200799	32	NC_000927
Nicotiana sylvestris	155941	38	NC_007500
Nicotiana tabacum	155943	38	NC_001879
Nicotiana tomentosiformis	155745	38	NC_007602
Nuphar advena	160866	44	NC_008788
Nymphaea alba	159930	43	NC_006050
Odontella sinensis	119704	16	NC_001713
Oedogonium cardiacum	196547	41	NC_011031
Oenothera argillicola	165055	44	NC_010358
Oenothera biennis	164807	44	NC_010361
Oenothera elata	165728	44	NC_002693
Oenothera glazioviana	165225	44	NC_010360
Oenothera parviflora	163365	44	NC_010362

Taxon	Genome length (nt)	% Noncoding	Genbank accession number
Olimarabidopsis pumila	154737	41	NC_009267
Oltmannsiellopsis viridis	151933	42	NC_008099
Oryza nivara	134494	47	NC_005973
Oryza sativa Japonica Group	134525	42	NC_001320
Ostreococcus tauri	71666	20	NC_008289
Panax ginseng	156318	42	NC_006290
Pelargonium x *hortorum*	217942	43	NC_008454
Phaeodactylum tricornutum	117369	12	NC_008588
Phalaenopsis aphrodite	148964	46	NC_007499
Phaseolus vulgaris	150285	40	NC_009259
Physcomitrella patens subsp. *patens*	122890	33	NC_005087
Pinus thunbergii	119707	31	NC_001631
Piper cenocladum	160624	46	NC_008457
Platanus occidentalis	161791	44	NC_008335
Populus alba	156505	42	NC_008235
Populus trichocarpa	157033	39	NC_009143
Porphyra purpurea	191028	15	NC_000925
Porphyra yezoensis	191952	16	NC_007932
Pseudendoclonium akinetum	195867	46	NC_008114
Psilotum nudum	138829	35	NC_003386
Ranunculus macranthus	155129	42	NC_008796
Rhodomonas salina	135854	19	NC_009573
Saccharum officinarum	141182	43	NC_006084
Saccharum hybrid cultivar SP-80-3280	141182	43	NC_005878
Scenedesmus obliquus	161452	39	NC_008101
Solanum bulbocastanum	155371	41	NC_007943
Solanum lycopersicum	155461	41	NC_007898
Solanum tuberosum	155298	42	NC_008096
Sorghum bicolor	140754	49	NC_008602
Spinacia oleracea	150725	39	NC_002202
Staurastrum punctulatum	157089	43	NC_008116
Stigeoclonium helveticum	223902	51	NC_008372
Thalassiosira pseudonana	128814	15	NC_008589
Theileria parva strain Muguga	39579	17	NC_007758
Toxoplasma gondii strain RH	34996	16	NC_001799
Trachelium caeruleum	162321	48	NC_010442
Triticum aestivum	134545	46	NC_002762
Vitis vinifera	160928	44	NC_007957
Welwitschia mirabilis	119726	34	NC_010654
Zea mays	140384	43	NC_001666
Zygnema circumcarinatum	165372	45	NC_008117

The mitochondrial and plastid genomes of *Volvox carteri*: bloated molecules rich in repetitive DNA

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

BMC Genomics

BioMed Central

Research article

Open Access

The mitochondrial and plastid genomes of *Volvox carteri*: bloated molecules rich in repetitive DNA

David Roy Smith* and Robert W Lee

Address: Department of Biology, Dalhousie University, Halifax, Nova Scotia, Canada

Email: David Roy Smith* - smithdr@dal.ca; Robert W Lee - robert.lee@dal.ca

* Corresponding author

Published: 26 March 2009

BMC Genomics 2009, **10**:132 doi:10.1186/1471-2164-10-132

Received: 27 September 2008

Accepted: 26 March 2009

This article is available from: http://www.biomedcentral.com/1471-2164/10/132

Abstract

Background: The magnitude of noncoding DNA in organelle genomes can vary significantly; it is argued that much of this variation is attributable to the dissemination of selfish DNA. The results of a previous study indicate that the mitochondrial DNA (mtDNA) of the green alga *Volvox carteri* abounds with palindromic repeats, which appear to be selfish elements. We became interested in the evolution and distribution of these repeats when, during a cursory exploration of the *V. carteri* nuclear DNA (nucDNA) and plastid DNA (ptDNA) sequences, we found palindromic repeats with similar structural features to those of the mtDNA. Upon this discovery, we decided to investigate the diversity and evolutionary implications of these palindromic elements by sequencing and characterizing large portions of mtDNA and ptDNA and then comparing these data to the *V. carteri* draft nuclear genome sequence.

Results: We sequenced 30 and 420 kilobases (kb) of the mitochondrial and plastid genomes of *V. carteri*, respectively – resulting in partial assemblies of these genomes. The mitochondrial genome is the most bloated green-algal mtDNA observed to date: ~61% of the sequence is noncoding, most of which is comprised of short palindromic repeats spread throughout the intergenic and intronic regions. The plastid genome is the largest (>420 kb) and most expanded (>80% noncoding) ptDNA sequence yet discovered, with a myriad of palindromic repeats in the noncoding regions, which have a similar size and secondary structure to those of the mtDNA. We found that 15 kb (~0.01%) of the nuclear genome are homologous to the palindromic elements of the mtDNA, and 50 kb (~0.05%) are homologous to those of the ptDNA.

Conclusion: Selfish elements in the form of short palindromic repeats have propagated in the *V. carteri* mtDNA and ptDNA, resulting in the distension of these genomes. Copies of these same repeats are also found in a small fraction of the nucDNA, but appear to be inert in this compartment. We conclude that the palindromic repeats in *V. carteri* represent a single class of selfish DNA and speculate that the derivation of this element involved the lateral gene transfer of an organelle intron that first appeared in the mitochondrial genome, spreading to the ptDNA through mitochondrion-to-plastid DNA migrations, and eventually arrived in the nucDNA through organelle-to-nucleus DNA transfer events. The overall implications of palindromic repeats on the evolution of chlorophyte organelle genomes are discussed.

Background

The amount of noncoding DNA (i.e., intronic and intergenic DNA) in organelle genomes varies significantly. One evolutionary lineage where the gamut of organelle-genome compactness is particularly pronounced is the Chlorophyta (a phylum containing most known classes of green algae [1]) for which the noncoding-DNA contents span from 10% (*Ostreococcus tauri*) to 53% (*Pseudendoclonium akinetum*) for mitochondrial DNA (mtDNA) and from 5% (*Helicosporidium* sp. ex *Simulium jonesii*) to 56% (*Chlamydomonas reinhardtii*) for plastid DNA (ptDNA). Although the processes influencing genome compactness are poorly understood, it is suggested that they are associated with the proliferation of selfish DNA [2-4], which we will define for the purpose of this study as any noncoding DNA element with the ability to spread its sequence to new genomic locations. Selfish DNA is ubiquitous in eukaryotic nuclear genomes and is also present, though less pervasive, in organelle genomes [5-7], including those of chlorophytes [8] – and see Hurst and Werren [9] for a review.

One proposed type of selfish DNA that is often observed in the mitochondrial and plastid genomes of chlorophytes are short palindromic repeats; these repetitive elements are marked by their short length [10–100 nucleotides (nt)] and their ability to be folded into hairpin (i.e., stem-loop) structures. Short palindromic repeats have been identified in the mtDNA and ptDNA from a variety of chlorophyte taxa (for compilations see [6,10-13]), and there are indications that the organelle genomes of some chlorophyte species are more prone to the dissemination of palindromic repeats than those of other chlorophyte species. For example, the mtDNA of *P. akinetum*, the largest chlorophyte mitochondrial genome sequenced to date [95 kilobases (kb)], contains a series of short palindromic repeats that have proliferated throughout its intergenic and intronic regions [11]; moreover, these same palindromic elements are found in the noncoding regions of the *P. akinetum* ptDNA [12], indicating that the lateral transfer of repetitive DNA between organelle compartments is taking place – it is speculated that this process involves the hitchhiking of repeats on introns that migrate between the mtDNA and ptDNA [12]. Short palindromic repeats in chlorophyte organelle DNA, as well as adding to the amount of noncoding DNA in a genome, have been credited for: 1) genome rearrangements [10], including the fragmentation and scrambling of ribosomal-RNA (rRNA) coding modules [14]; 2) changes in genome conformation and chromosome number [13,15]; and 3) having regulatory functions [13,16].

A previous study found that the mitochondrial genome of the multicellular, chlorophyte green alga *Volvox carteri*, a close relative to *C. reinhardtii*, is profuse with short palindromic repeats [17]. By sequencing ~8 kb of mtDNA, corresponding to portions of *cob*, *cox1*, and their group-I introns, and an intergenic region between *nad2* and *nad6*, these authors showed that at least two group-I introns and one intergenic region of the *V. carteri* mitochondrial genome contain extensive palindromic sequences. Ten classes of short palindromic repeats, all of which share sequence identity with one another, were discerned. Aono et al. [17] concluded that these repetitive elements are selfish DNA, that they are expanding the intronic and intergenic regions of the mitochondrial genome, and that they may have played a role in the fragmentation of rRNA-coding modules.

In June 2007, the United States Department of Energy Joint Genome Institute (DOE JGI) released the draft nuclear genome sequence of *V. carteri* [18]. This genome is ~140 megabases (Mb) in length [18-20], which is the largest chlorophyte nuclear genome sequence currently available (see [21] for a compilation), and the preliminary annotation of this sequence suggests that it is rich in noncoding DNA, much of which appears to be selfish [18].

Our interest regarding the evolution of selfish DNA in *V. carteri* was sparked when we located short palindromic repeats in the *V. carteri* draft nuclear genome sequence that are similar to those in the mtDNA. To see if palindromic repeats are also present in the *V. carteri* ptDNA, we PCR amplified and sequenced an intergenic region from the plastid genome; our analysis of this sequence confirmed that the ptDNA contains palindromic repeats that have a similar size and potential secondary structure to those in the mitochondrial compartment. We, therefore, set out to do a detailed investigation of the palindromic repeats in *V. carteri* by sequencing and characterizing large portions of mtDNA and ptDNA and then comparing these data to the nuclear DNA (nucDNA). Our motives for this study are to gain insights into the role selfish DNA plays in organelle-genome expansion and its ability to move between genetic compartments.

Results

General features of the mitochondrial- and plastid-DNA sequences from V. carteri

Using a long-range PCR approach in conjunction with cloning, we sequenced from *V. carteri* 29,961 nt of the mitochondrial genome and 420,650 nt of the plastid genome; partial genetic maps of these genomes describing the coding and noncoding regions that were sequenced are respectively shown in Figures 1 and 2. Regions of the *V. carteri* organelle DNA that were previously characterized (~8 kb of mtDNA and ~5 kb of ptDNA) are highlighted in pink on these maps. Although we attempted to completely sequence the mitochondrial and plastid

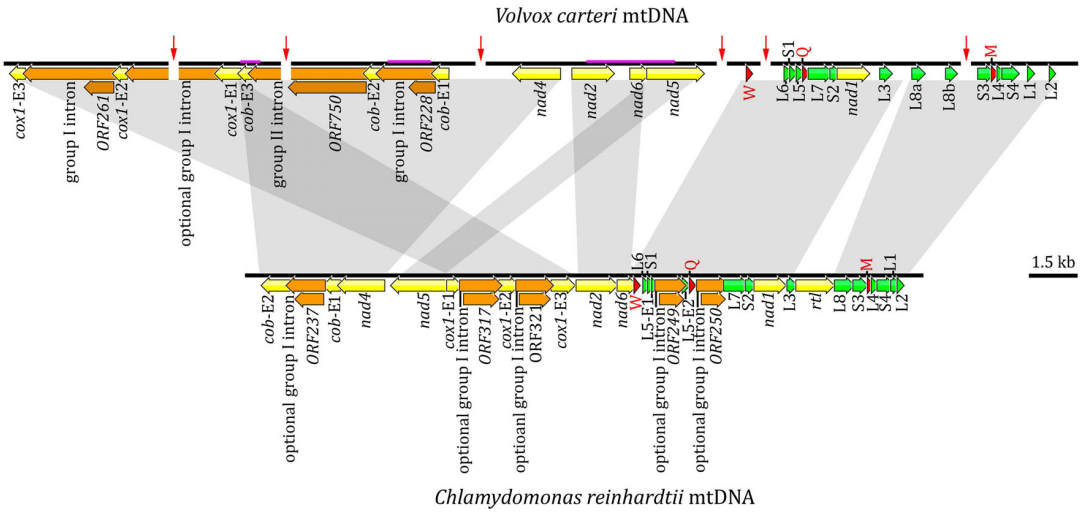

Figure 1

Partial genetic map of the *Volvox carteri* mitochondrial genome compared to the complete mtDNA genetic map of *Chlamydomonas reinhardtii*. Protein-coding regions are yellow and their exons are labelled with an "E" followed by a number denoting their position within the gene. Introns and their associated open reading frames are orange. Transfer RNA-coding regions are red; they are designated by the single-letter abbreviation of the amino acid they specify. The large-subunit and small-subunit rRNA-coding modules are green. Arrows within the coding regions denote their transcriptional polarities. Solid red arrows perpendicular to the genome map indicate regions of the genome assembly where sequence data is either unreadable or lacking. The mtDNA regions that were previously sequenced and described by Aono et al. [17] are underlined in pink on the genome map. Gray blocks highlight regions of synteny between the *V. carteri* and *C. reinhardtii* mitochondrial genomes. Note: the optional group-I intron in *cox1* is found in the mtDNA of *V. carteri* strain HK10 (UTEX 1885); this intron is absent from *V. carteri* strain 72-52 (UTEX 2908) – the *C. reinhardtii* strains in which the different introns occur are listed in [50].

genomes, presumed secondary structures in the mtDNA and ptDNA templates likely caused many of the sequencing reactions to suddenly stop – even when using protocols designed to alleviate this problem [22]. Furthermore, the repetitive nature of the organelle genomes means that much of the mtDNA and ptDNA sequence data are irresolvable using the currently available genome-assembly software programs: many of the organelle intergenic and intronic DNA sequences collapse into networks of spurious repetitive motifs upon assembly. Moreover, the fact that most of the mtDNA and ptDNA intergenic regions are much longer than a typical sequencing read (some intergenic regions exceed 15 kb) means that these collapsed repeats are irresolvable. At present, the most sophisticated assembly programs use the paired-end sequencing data from whole-genome shotgun reads to resolve complex repeat regions. Because there is a *V. carteri* nuclear genome sequencing project [18], we have access to paired-end sequencing reads for the mitochondrial and plastid genomes (see Methods for details); but even with these data, neither the assembly programs nor our own manual, by-eye assembly methods can untangle these repeats. Because of these difficulties, our *V. carteri* mitochondrial-genome assembly, although contained in a single contig, contains six regions where the mtDNA sequence is either unreadable or unavailable (Figure 1), and the assembly of the ptDNA is divided into 34 contigs (Figure 2). Nevertheless, we did sequence and characterize enough mtDNA and ptDNA to confidently describe the abundance and various types of noncoding DNA in each of these organelle genomes.

The organelle-DNA sequences presented in this study were validated by collecting and assembling mtDNA and ptDNA sequence data that were generated by the DOE JGI *V. carteri* nuclear genome sequencing project [18]. This was performed by: 1) downloading DNA-sequence trace files corresponding to the *V. carteri* mitochondrial and plastid genomes; 2) assembling these trace files into contigs; and 3) mapping the trace-file contigs to the *V. carteri* mtDNA and ptDNA sequences produced in this study. Ultimately, the mtDNA and ptDNA sequences coming from the DOE JGI covered all of our self-generated *V. carteri* sequence data with >50-fold redundancy. It is important to note that the DOE JGI data that we used to confirm our mtDNA and ptDNA sequences came from *V. carteri*

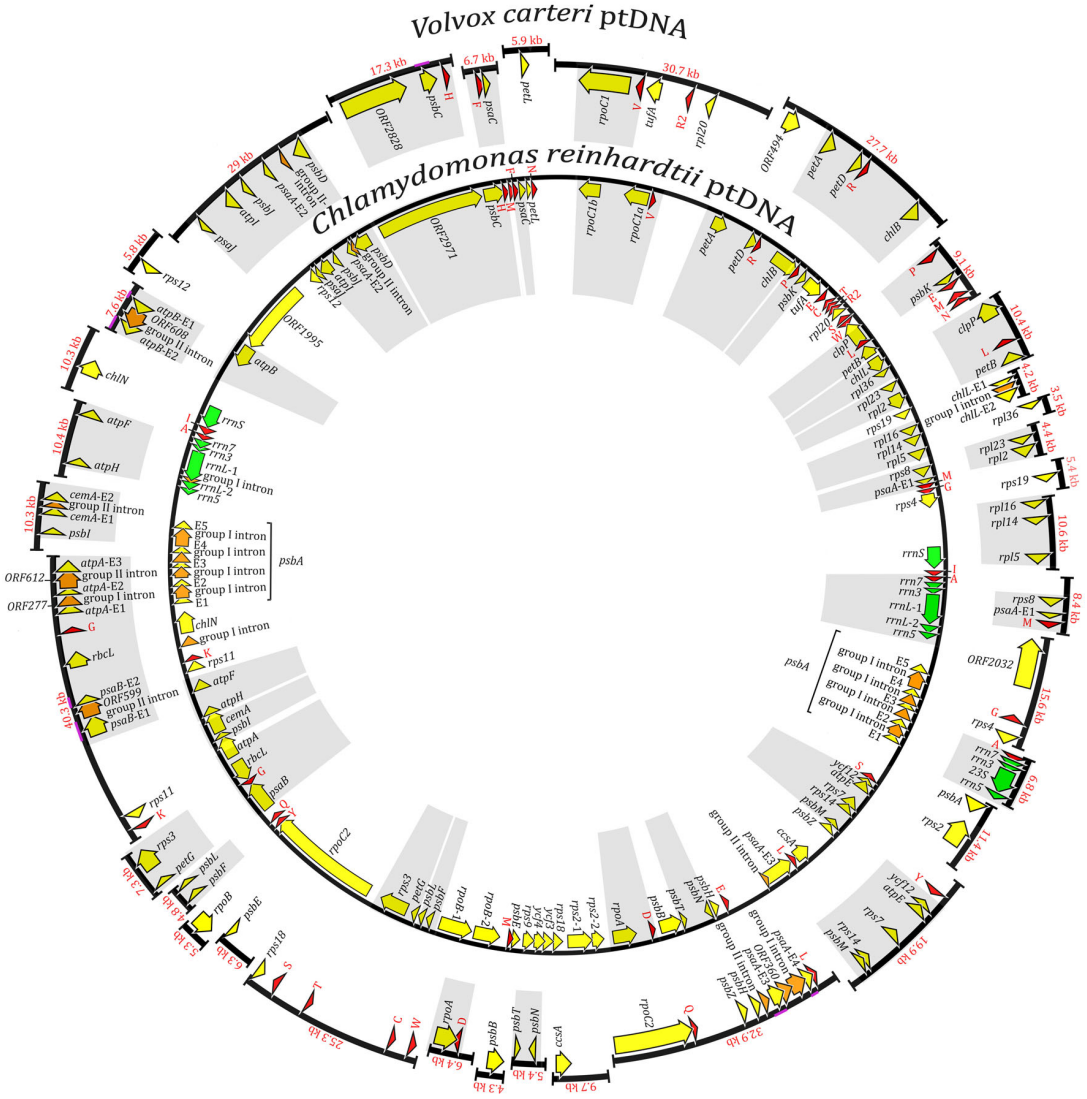

Figure 2

Partial genetic map of the *Volvox carteri* plastid genome compared to the complete ptDNA genetic map of *Chlamydomonas reinhardtii*. Regions encoding proteins are yellow and their exons are labelled with an "E" followed by a number signifying their order within the gene. Introns and their associated open reading frames are orange. Transfer RNA-coding regions are red; they are designated by the single-letter abbreviation of the amino acid they specify. Ribosomal RNA-coding regions are green. The coding regions are shaped into arrows that denote their transcriptional polarities. Regions previously described by Aono et al. [17] are underlined in pink. Gray blocks highlight regions of synteny between the *V. carteri* and *C. reinhardtii* plastid genomes. Note, the true order of the *V. carteri* contigs are unknown.

strain HK10 (UTEX 1885), whereas the *V. carteri* mtDNA and ptDNA sequences that we generated came from strain 72-52 (UTEX 2908), which is a dissociator mutant derived from HK10 [23,24]. In all instances, the organelle DNA sequence data coming from strain HK10 were identical to those of strain 72-52 (i.e., no ambiguities between the DOE JGI trace-file contigs and our sequences were observed), with the exception of a group-I intron that is present in the mtDNA of HK10 but absent in that of 72-52 (see below for details).

Of the 29,961 nt of *V. carteri* mtDNA sequence data presented here, 18,355 nt (61%) are noncoding, which include 7,870 nt (26%) of intronic DNA and 10,485 nt (35%) of intergenic DNA; the remaining 11,606 nt (39%) are comprised of 8,166 nt (27%) coding for proteins and 3,440 nt (12%) coding for structural RNAs. The intergenic regions range from 0 to >1,400 nt in length, and, on average, are 455 nt long. The AT content of the 29,961 nt mtDNA sequence is 66%. Our annotation of the *V. carteri* mtDNA includes 7 protein-coding genes; the full suite of rRNA-coding modules required for the formation of the large-subunit and small-subunit rRNAs; 3 tRNA-coding genes; and 3 introns, 2 of group-I affiliation, located in *cox1* and *cob*, and 1 of group-II affiliation, located in *cob* (Figure 1). Both group-I introns contain an open reading frame (ORF) encoding a putative LAGLIDADG endonuclease. The sole group-II intron has an ORF for which the deduced amino-acid sequence shows similarity to a reverse transcriptase (Figure 1). The DOE JGI *V. carteri* mtDNA sequences that we assembled (derived from *V. carteri* strain HK10) have, as mentioned above, an additional group-I intron in *cox1* that is not present in *V. carteri* strain 72-52 (Figure 1). The coding suite that we acquired for the *V. carteri* mtDNA is identical to that of the *C. reinhardtii* mitochondrial genome [16,25,26] as is the gene order save for two rearrangements, which are outlined on Figure 1. There are two interesting features of the *V. carteri* mtDNA relative to its *C. reinhardtii* counterpart. First, the *V. carteri* L8 rRNA-coding module harbours a 725 nt insertion composed of short palindromic repeats, whereas that of *C. reinhardtii* contains no repeats (Figure 3a). When the *V. carteri* L8 module is folded into a putative secondary-structure model within the context of the LSU rRNA it contains two structural constituents: L8a and L8b (corresponding to the LSU rRNA domains V and VI, respectively), where the 3' end of L8a and the 5' end of L8b border the 725 nt insertion (Figure 3b). At present, we do not know if this insertion is removed from the primary transcript so that separate L8a and L8b mature transcripts are produced or if a single mature L8 transcript is generated with the insertion. The second point of interest is that although a putative reverse transcriptase gene is found in the mtDNA of both of *V. carteri* (*ORF750*) and *C. reinhardtii* (*rtl*), that of *V. carteri* appears to be part of a group-

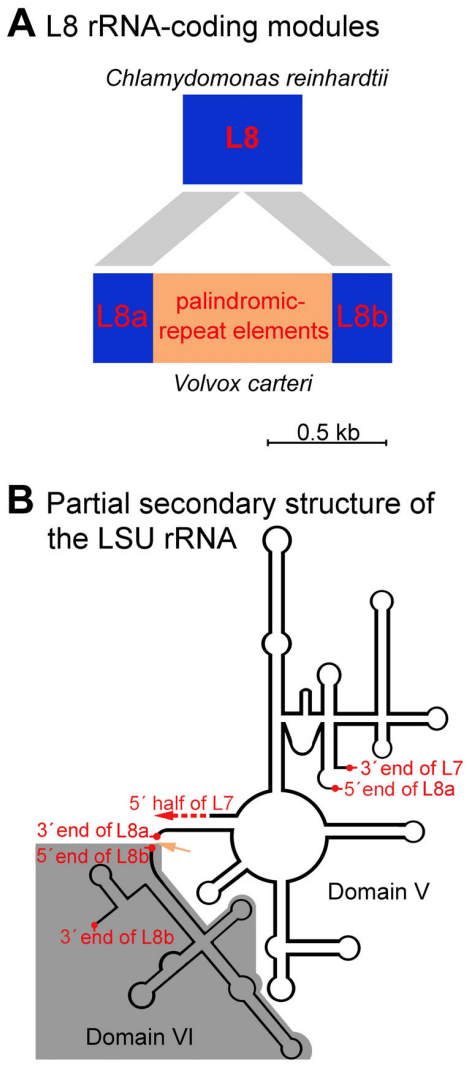

Figure 3
Schema of the L8 rRNA-coding module in the *Volvox carteri* mitochondrial genome and its relationship to that in the *Chlamydomonas reinhardtii* mtDNA. The grey bars in **A** denote regions of sequence identity between the L8 rRNA-coding modules of the *V. carteri* and *C. reinhardtii* mtDNA. **B** depicts the *V. carteri* L8 coding module in the context of the large-subunit (LSU) rRNA secondary-structure model; the orange arrow points to the repetitive region separating the L8a and L8b components of the L8 module. The LSU rRNA secondary-structure model is based on that of Boer and Gray [49].

II intron located in *cob*, whereas in *C. reinhardtii*, *rtl* is a free standing gene that is lacking an intron but speculated to have originated from one [27,28] – see Popescu and Lee [29] for further discussion. The deduced amino-acid sequences of both *ORF750* and *rtl* have a conserved domain that resembles that of a reverse transcriptase with group-II intron affiliation; however, the amino-acid sequence of *ORF750* also has a conserved domain with similarity to a type-II intron maturase, while that of *rtl* does not. This extra domain encoded in *ORF750* also explains why this ORF is twice the size of *rtl* (2,250 nt versus 1,119 nt).

In regard to the 420,650 nt of *V. carteri* ptDNA sequence data that were generated, 338,557 nt (80%) are noncoding, of which 16,005 nt are intronic DNA and 322,552 nt are intergenic DNA; 77,335 nt (19%) code for proteins and 4,758 nt (1%) code for structural RNAs. The intergenic regions that were sequenced range from 87 nt to >12,444 nt in length and have an average size of 5,103 nt. The 420 kb of ptDNA are 57% AT. Our annotation of the ptDNA sequences includes 91 genes: 60 coding for standard plastid proteins, 27 coding for structural RNAs (23 tRNAs and 4 rRNAs), and 4 corresponding to ORFs (*ORF494*, *ORF2032*, *ycf12*, *ORF2828*) that have been previously found in plastid genomes (Figure 2). Three group-I introns were observed, located in *chlL*, *psaA*, and *atpA*; those in the later two genes contain an ORF encoding a putative LAGLIDAD endonuclease. Five group-II introns were discerned, situated in *psaA*, *cemA*, *psaB*, *atpA*, and *atpB*; the introns of the latter two genes have an ORF for which the inferred amino-acid sequence resembles that of a reverse transcriptase. The group-II intron of *psaA* is fragmented into two separate modules, which is also the case for *C. reinhardtii* (Figure 2). The 91 *V. carteri* ptDNA genes presented here are all found in the *C. reinhardtii* plastid genome with the exception of *ORF494*. The only apparent homolog of *ORF494* is the ribosomal operon-associated gene (*roaA*) found in the *Euglena gracilis* plastid genome. Note, the *C. reinhardtii* ptDNA encodes a further 4 tRNA-coding and one rRNA-coding regions that we were unable to amplify from *V. carteri*.

A graph comparing both the estimated sizes and the fraction of noncoding nucleotides in the mitochondrial and plastid genomes of *V. carteri* relative to those of the currently available complete organelle-genome sequences from chlorophyte-, streptophyte- and other plastid-harbouring-taxa is shown in Figure 4 [and see Additional file 1]. Values of 30 kb and 420 kb, respectively, were chosen, based on our sequence data, as minimum-estimate genome sizes for the *V. carteri* mtDNA and ptDNA.

Short palindromic repeats in the mitochondrial and plastid genomes of V. carteri

Scanning of the *V. carteri* mitochondrial- and plastid-DNA sequences for repetitive elements lead to the identification of a series of short palindromic repeats in both of the organelle genomes; the consensus sequences, complementary bases, and copy numbers of the mtDNA and

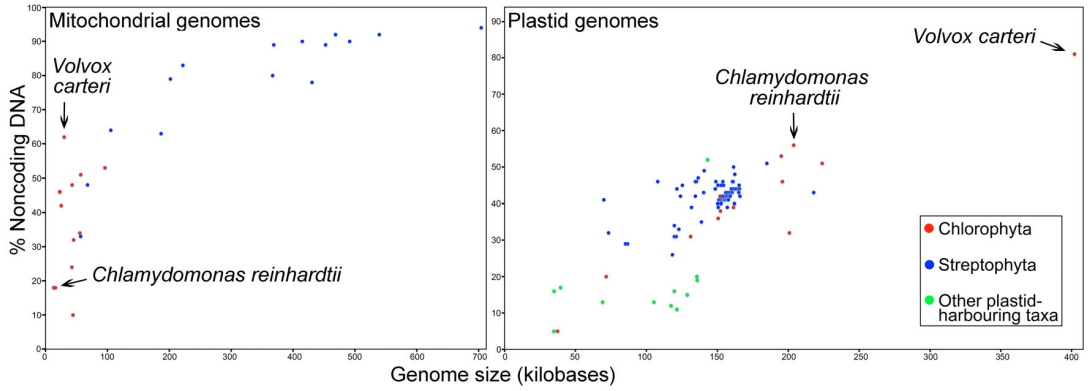

Figure 4

Fraction of noncoding DNA plotted against genome size for the available organelle genomes from streptophytes, chlorophytes, and other plastid-harbouring taxa. The data points corresponding to the mtDNA and the ptDNA of *V. carteri* and those of its close relative *C. reinhardtii* are labelled and marked with arrows on the appropriate graph. The noncoding-DNA contents and genome sizes from which these two graphs were plotted are listed in Supplementary Table S1 [see Additional file 1]. Values of 30 kb and 420 kb, respectively, were chosen, based on our sequence data, as minimum-size-estimates of the mitochondrial and plastid genomes.

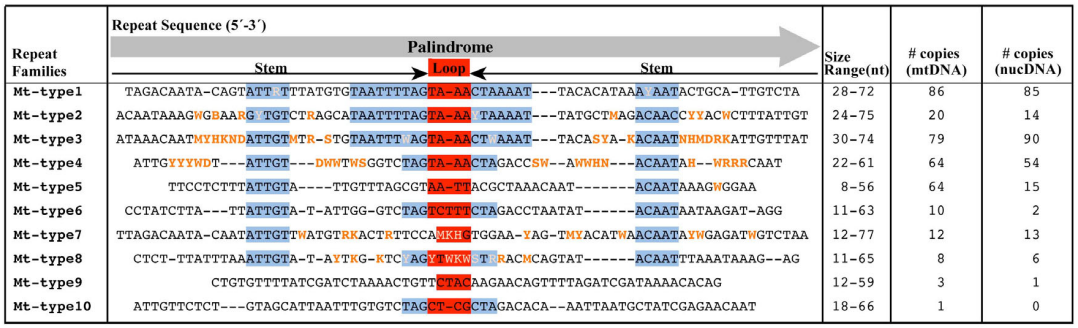

Figure 5

Abundance and classification of the *Volvox carteri* mitochondrial-DNA palindromic repeat elements. Regions of high sequence identity among the different repeat families are shaded in blue; variable sites are orange; the loop portions of the putative hairpin structures are shaded in red, and the stems (i.e., complementary bases) of these structures are located beneath the black arrows. Nuclear DNA analyses were performed using the first 75 scaffolds of the *V. carteri* draft nuclear genome sequence (version 1) at the DOE JGI [18].

ptDNA palindromic elements are outlined in Figures 5 and 6, respectively. Although the short palindromic repeats of the mtDNA share many of the same structural traits as those of the ptDNA (discussed below), they differ by >50% in sequence identity and, therefore, must be considered as distinct repeats relative to those of the plastid genome.

The short palindromic repeats in the *V. carteri* mtDNA are restricted to intergenic and intronic regions, with the exception of the palindromic elements in the L8 rRNA-coding module. All of the intergenic regions that measure >50 nt in length consist predominantly of palindromic repeats; the few intergenic regions with lengths <50 nt are composed of non-repetitive DNA. Within the intronic regions, the palindromic repeats are confined to the non-ORF portions of the group-I and group-II introns. All four of the identified mitochondrial introns contain short palindromic repeats in their non-ORF regions, including the optional group-I intron of *cox1*, which was found in the mtDNA of *V. carteri* strain HK10. Approximately 14,600 nt (~80%) of the 18,355 nt of noncoding mtDNA that were sequenced are composed of short palindromic repeats. A dotplot similarity matrix of the *V. carteri* mtDNA plotted against itself, shown in Supplementary Figure S1 [see Additional file 2], emphasizes the magnitude of repetitive DNA in this genome and draws attention to the high degree of sequence identity between the different palindromic elements within and among the various intergenic and intronic regions.

The short palindromic repeat elements identified in the *V. carteri* mtDNA show >50% sequence identity with one another and share similar structural and compositional traits (Figure 5). The individual palindromes range from 11–77 nt in length (the average size is 50 nt) and from 71–84% in their AT content. When the palindromes are folded into hairpin structures, the stem component of the hairpin varies from 4–37 nt in length, and the loop portion is usually 3–5 nt long and frequently has the sequence 5'-TAAA-3' or 5'-TTTA-3' (Figure 5). In many instances, a short palindromic repeat is found inserted within another palindromic repeat, resulting in larger, more elaborate repetitive elements; these larger repeats have a maximum length of 633 nt, and, in a few cases, are found at multiple locations in the mitochondrial genome. For example, a 550 nt repeat sequence composed of complete and incomplete short palindromic units is found in the group-I introns of *cob* and *cox1*, in the intergenic regions between *cob* and *nad4*, and in the group-II intron of *cob*. Some of these more complex repeats can also be folded into tRNA-like structures – as shown in Supplementary Figure S2 [see Additional file 3].

Like the mitochondrial genome, the noncoding regions of the *V. carteri* plastid genome abound with short palindromic sequences. These palindromic elements are observed in all of the sequenced intergenic regions that have lengths >100 nt and in the non-ORF portions of the *psaA* and *atpA* group-I introns. No palindromic elements are located in the *chlL* group-I intron or in any of the identified group-II introns. Overall, the short palindromic repeats constitute ~80% (~270 kb) of the 338,557 noncoding nucleotides in the *V. carteri* ptDNA. A dotplot similarity matrix of the *V. carteri* plastid genome plotted against itself (Supplementary Figure S3 [see Additional file 4]) shows the high level of sequence identity among the various palindromic repeats.

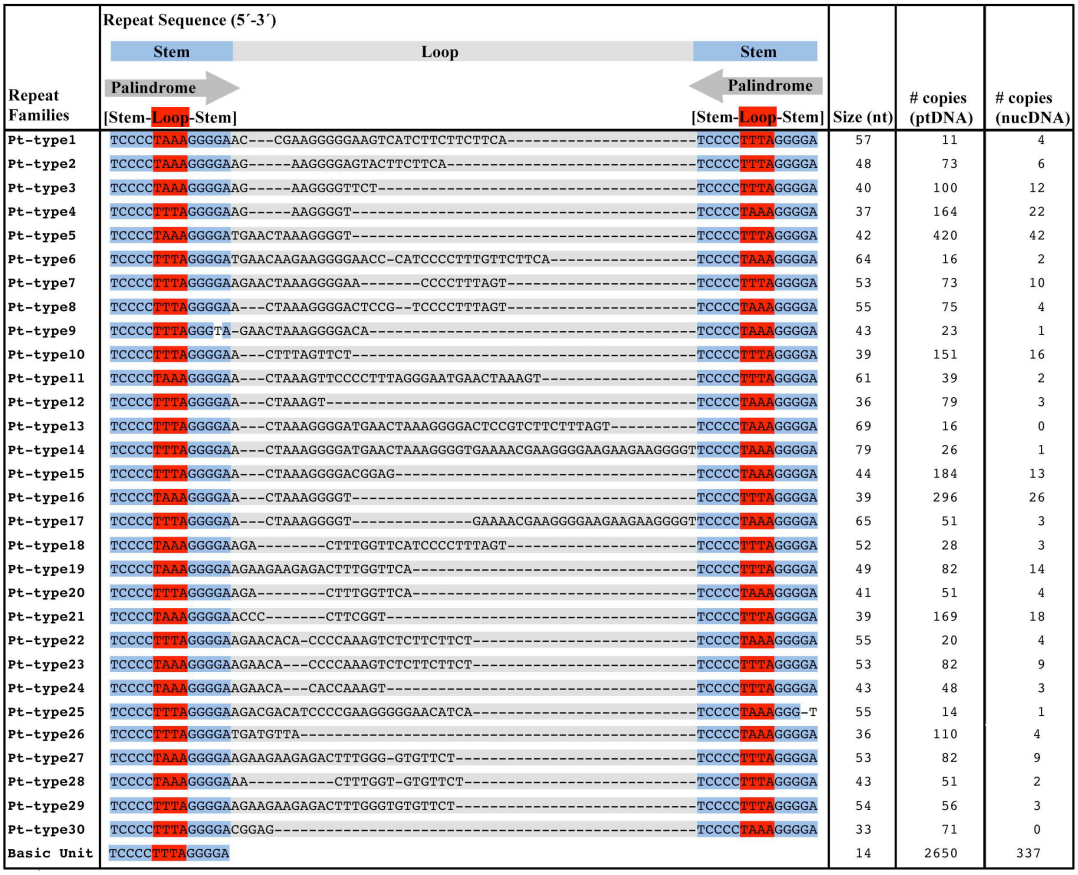

Figure 6

Abundance and classification of the *Volvox carteri* plastid-DNA palindromic repeat elements. Regions of high sequence identity among the different repeat families are shaded in blue; the loop portions of the putative hairpin structures are shaded in either red or grey, and the stems (i.e., complementary bases) of these structures are located beneath the grey arrows. Nuclear DNA analyses were performed using the first 75 scaffolds of the *V. carteri* draft nuclear genome sequence (version 1) at the DOE JGI [18].

In the *V. carteri* ptDNA, most of the short palindromic repeats contain the sequence motif 5'-TCCCCTTTAG-GGA-3' (Figure 6). The palindromes have a size range of 14–79 nt, with an average length of 50 nt, and, when folded into hairpin structures, their stems and loops vary in length from 5–29 nt and 3–5 nt, respectively. In most cases, the loops of the hairpin structures contain the sequence 5'-TAAA-3' or 5'-TTTA-3' (Figure 6). The AT content of the ptDNA palindromes varies from 39–55%. As observed for the mtDNA, the ptDNA palindromic elements are often found inserted into one another, the consequence of which is a series of multifarious repetitive sequences.

The short palindromic repeats of the mitochondrial and plastid compartments have similar structural attributes: 1) they have proliferated in intergenic regions and non-ORF segments of introns; 2) they have an average size of 50 nt and a maximum length of ~78 nt; and 3) the loops of their hairpin structures are generally 3–5 nt long with the sequence 5'-TTTA-3' or 5'-TTTA-3'.

The nuclear genome of V. carteri *shares sequence identity with organelle DNA*

To investigate if the short palindromic repeats in the mtDNA and ptDNA of *V. carteri* are present in the nucDNA, we analyzed the draft nuclear genome sequence

Table 1: Amount of nuclear DNA in *Volvox carteri* that maps to the mitochondrial and plastid genomes.

		# of similarity regions[a]	Avg. length of similarity region (nt)	Max similarity length (nt)	Cumulative lengths of similarity regions (nt)	Fraction of nuclear genome
Amount of nucDNA mapping to the mitochondrial genome (by mtDNA subcategory[b])	Protein-coding genes[c]	114	64	291	7335	6.70×10^{-5}
	Sructural-RNA genes[d]	73	40	207	2969	2.70×10^{-5}
	Intronic ORFs[e]	278	30	233	278	0.25×10^{-5}
	Intergenic and non-ORF intronic regions[f]	337	38	933	14782	13.53×10^{-5}
	Subtotal	802	39	933	33452	30.63×10^{-5}
Amount of nucDNA mapping to the plastid genome (by ptDNA subcategory[b])	Protein-coding genes[c]	365	48	462	17440	15.90×10^{-5}
	Structural-RNA genes[d]	127	31	170	4008	3.60×10^{-5}
	Intronic ORFs[e]	31	34	154	1075	0.98×10^{-5}
	Intergenic and non-ORF intronic regions[f]	927	22	3430	50631	46.36×10^{-5}
	Subtotal	1450	29	3430	73154	66.99×10^{-5}
Total nucDNA mapping to organelle DNA		2252	33	3430	106606	97.62×10^{-5}

Note: Nuclear DNA analyses are based on the *V. carteri* draft nuclear genome sequence (version 1) at the DOE JGI [18]. Only the first 75 scaffolds of the nuclear-genome assembly were analyzed; approximately 78% of the *V. carteri* nucDNA is contained in these 75 scaffolds and their cumulative length is 109.2 Mb.

[a] The number of distinct regions in the *V. carteri* nucDNA that show >90% sequence identity and at least 25 nt of aligned length to organelle DNA.

[b] Refers to the region of the organelle genome to which the nucDNA maps.

[c] Includes all of the identified protein-coding genes.

[d] Includes all of the identified tRNA- and rRNA-coding genes.

[e] Includes all of the identified group I and II intronic-ORFs.

[f] Includes all of the identified intergenic and non-ORF intronic regions.

of *V. carteri* at the DOE JGI [18]. Manual curation of the *V. carteri* nuclear genome is still underway; therefore, only the first 75 scaffolds of the nuclear-genome assembly were analyzed. Approximately 78% of the *V. carteri* nucDNA is contained in these 75 scaffolds, their cumulative length is 109.2 Mb, and each scaffold is at least 0.5 Mb long. The amount of nucDNA in these 75 scaffolds that map to mtDNA and ptDNA is described in Table 1; the approximate number of nucDNA-located organelle-like repeats is outlined in Figures 5 and 6.

Thirty-three kilobases of nucDNA (~0.03% of the nuclear genome) share >90% identity with mtDNA; 14.7 kb (44%) of this shared sequence are homologous to the

short palindromic repeats in the intergenic and non-ORF intronic regions of the mitochondrial genome; the remaining 10.6 kb (56%) map to the coding and intronic-ORF portions of the mtDNA (Table 1). In the nucDNA, 802 distinct regions show homology to mtDNA; the average mapping length of these regions is 39 nt. Of the 75 nuclear scaffolds that were analyzed, all but two (scaffold 66 and 75) have at least one region that shows homology to mtDNA.

Seventy-three kilobases of the *V. carteri* nucDNA (~0.07% of the nuclear genome) share >90% identity with ptDNA; 50.6 kb (69%) of this shared sequence are homologous to the short palindromic repeats in the intergenic and non-ORF-intronic portions of the ptDNA, and 22.5 kb (31%) are homologous to the coding regions and intronic ORFs of the plastid genome (Table 1). In the nuclear genome, 1,450 different regions show homology to ptDNA, and the average similarity length is 29 nt. All of the 75 nuclear scaffolds that were examined have at least one region that shows homology to ptDNA.

In total, 65.4 kb (~0.06%) of the *V. carteri* nuclear-genome-sequence data that were analyzed share sequence identity to the short palindromic repeats of the organelle genomes. The secondary structure and general characteristics of the nuclear-palindromic repeat elements are the same as those described for the organelle palindromes in Figures 5 and 6.

In order to place the *V. carteri* data described above in a broader context, we analyzed the *C. reinhardtii* nuclear genome for regions that show sequence identity to its mtDNA and ptDNA – these results are summarized in Supplementary Table S2 [see Additional file 5]. Only 0.007% of the *C. reinhardtii* nuclear genome maps to organelle DNA (0.0035% to the mtDNA and 0.0035% to the ptDNA), which is 10-times less than what is observed for *V. carteri* (0.01%).

Discussion

Expanded architectures and the proliferation of selfish DNA in the organelle genomes of V. carteri

The *V. carteri* organelle genomes are profuse with noncoding DNA. Our analyses indicate that the mitochondrial genome is >61% noncoding, which is the most expanded chlorophyte-mtDNA sequence currently deposited in Genbank [30] – other complete mitochondrial genome sequences from chlorophyte are larger only because they contain more genes. The plastid genome is the largest (>420 kb) and most bloated (>81% noncoding) ptDNA sequence from any taxon observed to date (Figure 4). Heretofore, the largest documented plastid genome was 223.9 kb, belonging to the chlorophyte *Stigeoclonium helveticum* [31]; and the plastid genome with the highest fraction of noncoding DNA was that of *C. reinhardtii* (56% noncoding) [6]. At present, due to a lack of available mtDNA and ptDNA sequence data, we do not know if other colonial or multicellular algae in the volvocine line of the Chlorophyceae have bloated organelle genomes as a result of excessive noncoding DNA. Preliminary investigations of the *Volvox aureus* mitochondrial genome indicate that it harbours palindromic repeats; these repeats do not share sequence similarity with those of the *V. carteri* mtDNA [17].

The fact that short palindromic repeats respectively constitute ~80% and ~75% of the mtDNA and ptDNA noncoding regions suggests that these elements precipitated the expansion of the organelle genomes. We must, therefore, ask: how did the palindromic repeats disseminate their sequence throughout the noncoding regions of the organelle DNA? The processes by which this could have occurred include transposition, retrotransposition via an RNA intermediate, and recombination-based mechanisms, such as gene conversion (for a review on the mobility of selfish DNA see Austin and Trivers [32]); at present, we are not partial to any one of these mechanisms. It is worth noting, however, that both the mitochondrial and plastid genomes of *V. carteri* encode a putative reverse transcriptase and a putative endonuclease; the mutual association of these enzymes has been invoked for mediating the retrotransposition of selfish DNA. Koll et al. [33] propose that the mobility of an ultrashort invasive element in the mtDNA of the filamentous fungus *Podospora anserina* is instigated by a group-II-intron-encoded reverse transcriptase; they also suggest that an intron-encoded endonuclease generates the 3'-hydroxyl required for reverse transcription. Of all the intronic ORFs deposited in Genbank, the deduced amino-acid sequence of *ORF261* in the *Volvox* mtDNA shows the greatest identity (46%; expectation values = 6×10^{-48}) to that of the *cox1*-intronic ORF of *P. anserina*. Another intriguing observation is that the short palindromic repeats in the *V. carteri* organelle and nuclear genomes are found in both orientations on the same strand, i.e., the same sequence can occur in the 3' to 5' and 5' to 3' directions – this is suggestive of a transposition-mediated mechanism of mobility rather than one based on recombination.

Short palindromic repeats in the three genetic compartments of V. carteri

Short palindromic repeats are found in the three genetic compartments of *V. carteri*. The palindromic elements of the mtDNA are structurally similar but different in nucleotide sequence to those of the ptDNA, whereas the nucDNA palindromes are of two types: those that map to the mtDNA and those that map to ptDNA. These observations evoke several questions, such as are the palindromic

repeats in the mtDNA and ptDNA related? And if so, in what genetic compartment did they first appear? Moreover, how did the nucDNA acquire palindromes that share sequence identity to those of the mitochondrial and plastid genomes, and why are they not as abundant as those of the organelle DNA? And finally, are the palindromic elements indigenous to *V. carteri* or are they the products of lateral gene transfer?

Although the mtDNA palindromic repeats differ in nucleotide sequence to the palindromes of the ptDNA, these two groups of repeats share enough structural features to suggest that they have descended from a common repetitive element and, thus, represent a single class of selfish DNA. For instance, in both the mitochondrial and plastid compartments the palindromic repeats reside in identical genomic landscapes (intergenic regions and non-ORF portions of introns) and they also inhabit these noncoding regions in similar abundances, representing ~80% of the noncoding nucleotides in both the mitochondrial and plastid genomes. Moreover, the mtDNA and ptDNA palindromes fold into hairpin structures where the loop portion is consistently 5'-TAAA-3' or 5'-ATTT-3'.

The same palindromic elements that have propagated in the *V. carteri* organelle DNA are also found in a small fraction (~0.06%) of the nucDNA. Although these nuclear-palindromic repeats can share up to 100% sequence identity with those of the mtDNA or ptDNA, many are degenerate with mismatches in the stem component of their hairpin structures. This observation, coupled with the relatively low abundance of these repeats in the nuclear genome, leads us to suggest that the nucDNA palindromes are inert, accumulating in the nucDNA through both mitochondrial-to-nucleus and plastid-to-nucleus DNA-transfer events – the movement of organelle DNA to nuclear genomes is well documented [34-37]. Further evidence to support this hypothesis is the fact that the proposed organelle-derived palindromes are present within the nuclear genome in the same proportions as other genetic regions from the mitochondrial and plastid genomes, such as coding regions (Table 1).

Two observations regarding the derivation of the short palindromic repeats in *V. carteri* are worth noting. First, the palindromic elements appear to have a strong affinity for the non-ORF regions of organelle introns, especially those in the mitochondrial genome. Because organelle introns are, themselves, a type of selfish element, which can migrate between organelle compartments, both within a species and between unrelated species ([38,39], and see [40] for a discussion on volvocalean ptDNA group-I introns), it is not unreasonable to surmise that the origin of the short palindromic repeats in *V. carteri* is linked to lateral intron transfer. The second observation is that the short palindromic repeats have propagated in only two of the nine introns in the *V. carteri* plastid genome: the group-I introns of *psaA* and *atpA*. Why have the other seven plastid introns (one group-I and six group-II introns) remained inviolate from palindromic elements? The two group-I plastid introns that contain short palindromic repeats belong to subgroup IB. Considering that the other group-I intron in the ptDNA (that of *chlL*), which is devoid of palindromes, is also of subgroup IB, it seems unlikely that the palindromic elements are favouring a certain class of intron. It is noteworthy that the *chlL* group-I intron lacks a LAGLIDADG homing endonuclease, whereas the group-I introns of *psaA* and *atpA*, as well as those of the mtDNA, harbour LAGLIDADG ORFs – perhaps, as suggested above, the mobility of the palindromic elements is dependent on intronic endonuclease proteins. Another possibility is that the plastid genome was seeded with short palindromic repeats after the inception of palindromic elements in the mtDNA and, therefore, the plastid palindromic repeats have not had sufficient time to spread to all of the intronic regions in the ptDNA; moreover, some of ptDNA introns may have arrived more recently in evolutionary time than those of *psaA* and *atpA*, and, thus, have not yet been seeded with repeats.

Short palindromic repeats and the evolution of fragmented ribosomal-RNA-coding modules

When Aono et al. [17] first identified short palindromic repeats in the *V. carteri* mitochondrial genome they predicted that these repeats would be associated with the fragmentation of rRNA-transcripts. The mtDNA data presented here may support these predications. Our annotation of the *V. carteri* mitochondrial genome contains eight modules encoding the LSU rRNA, and their arrangement within the mtDNA is identical to that of *C. reinhardtii*, with the exception that the L8 module of *V. carteri* contains a large block of palindromic repeats, whereas that of *C. reinhardtii* harbours no repetitive mtDNA. At the present time we do not know if this insertion represents a fragmentation point within the L8 coding module that is removed from the primary transcript or if it is maintained as a variable region within an intact L8 transcript. We favour the former possibility because in the latter scenario the insertion is three-fold larger than any previously reported variable region identified in ribosomal RNA – based on the Comparative RNA Website [41].

Conclusion

The goal of this study was to investigate the genomic breadth and the evolutionary implications of short palindromic repeats in the organelle and nuclear genomes of *V. carteri*. Our findings indicate that selfish DNA, in the form of palindromic elements, have proliferated in the *V. carteri* mtDNA and ptDNA; and although copies of this element exist in the nuclear compartment, we suggest that they are

inert and arrived in the nuclear genome via rare organelle-to-nucleus DNA transfer events. We speculate that the palindromic repeats in *V. carteri* descended from a single invasive element, perhaps first seeded in *V. carteri* through the lateral gene transfer of a mitochondrial intron, eventually spreading to the ptDNA through mitochondrial-to-plastid DNA migration. Overall, the palindromic repeats appear to be involved with the expansion of the *V. carteri* organelle DNA and have potentially precipitated a gene fragmentation event in the mitochondrial genome.

Methods

Strain and DNA extractions

The sequence data generated in this study were obtained from the 72-52 dissociator mutant of *V. carteri* (UTEX 2908), which is derived from *V. carteri* strain HK10 (UTEX 885) [23,24]. Total genomic DNA was extracted using the DNeasy Plant Mini Kit (Qiagen, Germantown, MD, USA) following manufacturer's protocol.

Amplification, cloning, and sequencing of DNA fragments

The organelle loci examined in this study were amplified using a PCR-based approach. PCR reactions were performed with the LongRange PCR Kit (Qiagen) using total genomic DNA as the template. PCR products were cloned using the TOPO TA Cloning Kit (Invitrogen, Carlsbad, CA, USA). Purified PCR products and isolated plasmids were sequenced on both strands at the Macrogen Sequencing Facility, Rockville, MD.

Assembly of the mitochondrial and plastid DNA sequences

Sequences were edited and assembled using CodonCode Aligner Version 2.0.6 (CodonCode Corporation, Dedham, MA, USA), which employs the Phred, Cross-match, and Phrap algorithms for base calling, sequence comparison, and sequence assembly, respectively. Assemblies were performed with a minimum-percent-identity score of 98, a minimum-overlap length of 500 nt, a match score of 1, a mismatch penalty of -2, a gap penalty of -2, and an additional first gap penalty of -3.

Sequence confirmation

The mtDNA and ptDNA data presented in this study were validated by collecting and assembling mitochondrial- and plastid-genome sequences that were generated by the *V. carteri* nuclear genome sequencing project [18]. These sequences were obtained by blasting our mtDNA and ptDNA data against the *V. carteri* f. *nagariensis* Whole Genome Shotgun Reads Trace Archive Database at the National Center for Biotechnological Information (NCBI) [42]. Blast hits showing >99% similarity to our *V. carteri* mtDNA and ptDNA sequences were downloaded and assembled (using the assembly program and parameters described above); the downloaded mitochondrial and plastid sequences were subsequently blasted against the *V. carteri* draft nuclear genome sequence (v1.0 Repeatmasked) [18] to verify that no nuclear-genome-located mtDNA-like or ptDNA-like sequences were collected.

Analyses of repeats, introns, and intergenic regions

An initial scan for repetitive elements in the *V. carteri* mtDNA and ptDNA sequence data was performed with REPuter [43,44] using the Hamming distance option and a minimal-repeat-size setting of 12 nt – note, forward, reverse, complement, and reverse complement repeats were all considered under REPuter. Further analyses of the *V. carteri* organelle DNA repeats were performed in Geneious (Biomatters LtD, Auckland, NZ) by building a Blast databank of the mtDNA and ptDNA sequences and then blasting these databanks with specific regions from the mitochondrial and plastid genomes. Mfold [45] was employed for secondary-structure analyses. The mtDNA and ptDNA introns were detected, classified, and folded into secondary structures using RNAweasel [46,47]. The noncoding-DNA estimates presented for the *V. carteri* organelle genomes were inferred from the average-intergenic-spacer sizes and intron-densities of the mtDNA and ptDNA data that were collected. Dotplot similarity matrices were plotted with JDotter [48]. The LSU rRNA secondary-structure model depicted in Figure 3 is based on that of Boer and Gray [49]. The *C. reinhardtii* mtDNA-introns shown in Figure 1 are described elsewhere [50].

Inspection of **V. carteri** *nuclear DNA for organelle-like sequences*

The *V. carteri* nucDNA was scanned for regions of identity to organelle DNA by blasting [51] the mtDNA and ptDNA sequences produced in this study against the *V. carteri* draft nuclear genome sequence (v1.0 Repeatmasked) [18] using an expectation value of 1×10^{-5} and a word size of 11. Organelle-DNA sequences that mapped to the nuclear genome with >90% identity and at least 25 nt of aligned length were counted as hits. The same protocol employed for *V. carteri* was used to scan the *C. reinhardtii* nucDNA for regions that show identity to organelle sequences. The *C. reinhardtii* nucDNA scaffolds (version 3.1) were downloaded from the DOE JGI [52].

Accession numbers

The mtDNA and ptDNA sequences generated in this study are found in Genbank under the accession numbers EU760701 and EU755264–EU755299, respectively.

Authors' contributions

DRS carried out the molecular studies, data analyses, and wrote the manuscript. RWL helped in interpreting the data and revising the manuscript. Both DRS and RWL have read and approved the final version of this manuscript.

Additional material

Additional File 1

Supplementary Table S1. *The fraction of noncoding DNA in completely-sequenced mitochondrial and plastid genomes from streptophytes, chlorophytes, and other plastid-harbouring taxa.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-132-S1.pdf]

Additional File 2

Supplementary Figure S1. *Dotplot similarity matrix of the* Volvox carteri *mitochondrial DNA plotted against itself.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-132-S2.pdf]

Additional File 3

Supplementary Figure S2. *Putative secondary-structure diagrams of the tRNA pseudogenes identified in the mitochondrial genome of* Volvox carteri.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-132-S3.pdf]

Additional File 4

Supplementary Figure S3. *Dotplot similarity matrix of the* Volvox carteri *plastid DNA plotted against itself.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-132-S4.pdf]

Additional File 5

Supplementary Table S2. *Amount of nuclear DNA in* Chlamydomonas reinhardtii *that maps to its mitochondrial and plastid genomes.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-132-S5.pdf]

Acknowledgements

This work was supported by a grant to R.W.L. from the Natural Sciences and Engineering Research Council (NSERC) of Canada. D.R.S. is an Izaak Walton Killam Memorial Scholar and holds a Canada Graduate Scholarship from NSERC.

References

1. Lewis LA, McCourt RM: **Green algae and the origin of land plants.** *Am J Bot* 2004, **91:**1535-1556.
2. Doolittle WF, Sapienza C: **Selfish genes, the phenotype paradigm and genome evolution.** *Nature* 1980, **284:**601-603.
3. Orgel LE, Crick FH: **Selfish DNA: the ultimate parasite.** *Nature* 1980, **284:**604-607.
4. Lynch M: *The Origins of Genome Architecture* Sunderland: Sinauer Associates, Inc; 2007.
5. Paquin B, Laforest MJ, Lang BF: **Double-hairpin elements in the mitochondrial DNA of Allomyces: evidence for mobility.** *Mol Biol Evol* 2000, **17:**1760-1768.
6. Maul JE, Lilly JW, Cui L, dePamphilis CW, Miller W, Harris EH, Stern DB: **The *Chlamydomonas reinhardtii* plastid chromosome: islands of genes in a sea of repeats.** *Plant Cell* 2002, **14:**2659-2679.
7. Smith DR, Snyder M: **Complete mitochondrial DNA sequence of the scallop *Placopecten magellanicus*: evidence of transposition leadng to an uncharacteristically large mitochondrial genome.** *J Mol Evol* 2007, **65:**380-391.
8. Fan WH, Woelfe MA, Mosig G: **Two copies of a DNA element 'Wendy', in the chloroplast chromosome of *Chlamydomonas reinhardtii* between rearranged gene clusters.** *Plant Mol Biol* 1995, **29:**63-80.
9. Hurst GDD, Werren JH: **The role of selfish genetic elements in eukaryotic evolution.** *Nat Rev Genet* 2001, **2:**597-606.
10. Nedelcu AM, Lee RW: **Short repetitive sequences in green algal mitochondrial genomes: potential roles in mitochondrial genome evolution.** *Mol Biol Evol* 1998, **15:**690-701.
11. Pombert JF, Beauchamp P, Otis C, Lemieux C, Turmel M: **The complete mitochondrial DNA sequence of the green alga *Pseudendoclonium akinetum* (Ulvophyceae) highlights distinctive evolutionary trends in the Chlorophyta and suggests a sister-group relationship between the Ulvophyceae and Chlorophyceae.** *Mol Biol Evol* 2004, **21:**922-935.
12. Pombert JF, Otis C, Lemieux C, Turmel M: **The chloroplast genome sequence of the green alga *Pseudendoclonium akinetum* (Ulvophyceae) reveals unusual structural features and new insights into the branching order of chlorophyte lineages.** *Mol Biol Evol* 2005, **22:**1903-1918.
13. Smith DR, Lee RW: **Mitochondrial genome of the colorless green alga *Polytomella capuana*: a linear molecule with an unprecedented GC content.** *Mol Biol Evol* 2008, **25:**487-496.
14. Nedelcu AM: **Fragmented and scrambled mitochondrial ribosomal RNA coding regions among green algae: a model for their origin and evolution.** *Mol Biol Evol* 1997, **14:**506-517.
15. Nedelcu AM: **Contrasting mitochondrial genome organizations and sequence affiliations among green algae: potential factors, mechanisms, and evolutionary scenarios.** *J Phycol* 1998, **34:**16-28.
16. Gray MW, Boer PH: **Organization and expression of algal (*Chlamydomonas reinhardtii*) mitochondrial DNA.** *Philos Trans R Soc Lond B Biol Sci* 1988, **319:**135-147.
17. Aono N, Shimizu T, Inoue T, Shiraishi H: **Palindromic repetitive elements in the mitochondrial genome of *Volvox*.** *FEBS Lett* 2002, **521:**95-99.
18. **United States Department of Energy Joint Genome Institute (JGI) *Volvox carteri* nuclear genome assembly v1.0** [http://genome.jgi-psf.org/Volca1/Volca1.home.html]
19. Kirk DL, Harper JF: **Genetic, biochemical and molecular approaches to *Volvox* development and evolution.** *Int Rev Cytol* 1986, **99:**217-93.
20. Kirk DL: *Volvox: Molecular-genetic Origins of Multicellularity and Cellular Differentiation* Cambridge: Cambridge University Press; 1998.
21. Merchant SS, Prochnik SE, Vallon O, 117 co-authors, *et al.:* **The *Chlamydomonas* genome reveals the evolution of key animal and plant functions.** *Science* 2007, **318:**245-250.
22. **Macrogen Sequencing Facility: sequencing of difficult templates** [http://www.macrogen.com/eng/sequencing/dna.jsp]
23. Sessoms AH, Huskey RJ: **Genetic control of development in *Volvox*: isolation and characterization of morphogenetic mutants.** *P Natl Acad Sci USA* 1973, **70:**1335-1338.
24. Dauwalder M, Whaley WG, Starr RC: **Differentiation and secretion in *Volvox*.** *J Ultrastruct Res* 1980, **70:**318-335.
25. Michaelis G, Vahrenholz C, Pratje E: **Mitochondrial DNA of *Chlamydomonas reinhardtii*: the gene for apocytochrome *b* and the complete functional map of the 15.8 kb DNA.** *Mol Gen Genet* 1990, **223:**211-216.
26. Vahrenholz C, Riemen G, Pratje E, Dujon B, Michaelis G: **Mitochondrial DNA of *Chlamydomonas reinhardtii*: the structure of the ends of the linear 15.8-kb genome suggests mechanisms for DNA replication.** *Curr Genet* 1993, **24:**241-247.
27. Boer PH, Gray MW: **Genes encoding a subunit of respiratory NADH dehydrogenase (ND1) and a reverse transcriptase-like protein (RTL) are linked to ribosomal RNA gene pieces in *Chlamydomonas reinhardtii* mitochondrial DNA.** *EMBO J* 1988, **7:**3501-3508.
28. Nedelcu AM, Lee RW: **A degenerate group II intron in the intronless mitochondrial genome of *Chlamydomonas reinhardtii*: evolutionary implications.** *Mol Biol Evol* 1988, **7:**918-922.
29. Popescu CE, Lee RW: **Mitochondrial genome sequence evolution in Chlamydomonas.** *Genetics* 2007, **175:**819-826.

30. **National center for biotechnology information entrez organelle-genome database** [http://www.ncbi.nlm.nih.gov/genomes/genlist.cgi?taxid=2759&type=4&name=Eukaryotae%20Organelles]
31. Bélanger AS, Brouard JS, Charlebois P, Otis C, Lemieux C, Turmel M: **Distinctive architecture of the chloroplast genome in the chlorophycean green alga *Stigeoclonium helveticum*.** *Mol Genet Genomics* 2006, **276:**464-477.
32. Austin B, Trivers R: *Genes in Conflict: the Biology of Selfish Genetic Elements* Cambridge: Harvard University Press; 2006.
33. Koll F, Boulay J, Belcour L, d'Aubenton-Carafa Y: **Contribution of ultra-short invasive elements to the evolution of the mitochondrial genome in the genus *Podospora*.** *Nucleic Acids Res* 1996, **24:**1734-1741.
34. Adams KL, Palmer JD: **Evolution of mitochondrial gene content: gene loss and transfer to the nucleus.** *Mol Phylogenet Evol* 2003, **29:**380-395.
35. Timmis JN, Ayliffe MA, Huang CY, Martin W: **Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes.** *Nat Rev Genet* 2004, **5:**123-135.
36. Richly E, Leister D: **NUMTs in sequenced eukaryotic genomes.** *Mol Biol Evol* 2004, **21:**1081-1084.
37. Richly E, Leister D: **NUPTs in sequenced eukaryotes and their genomic organization in relation to NUMTs.** *Mol Biol Evol* 2004, **21:**1972-1980.
38. Turmel M, Côté V, Otis C, Mercier JP, Gray MW, Lonergan KM, Lemieux C: **Evolutionary transfer of ORF-containing group I introns between different subcellular compartments (chloroplast and mitochondrion).** *Mol Biol Evol* 1995, **12:**533-545.
39. Turmel M, Lemieux C, Burger G, Lang BF, Otis C, Plante I, Gray MW: **The complete mitochondrial DNA sequences of *Nephroselmis olivacea* and *Pedinomonas minor*. Two radically different evolutionary patterns within green algae.** *Plant Cell* 1999, **11:**1717-1729.
40. Nozaki H, Takahara M, Nakazawa A, Kita Y, Yamada T, Takano H, Kawano S, Kato M: **Evolution of *rbcL* group IA introns and intron open reading frames within the colonial Volvocales (Chlorophyceae).** *Mol Phylogenet Evol* 2002, **23:**326-338.
41. **Comparative RNA web site and project** [http://www.rna.ccbb.utexas.edu/]
42. **Trace archive database mega BLAST search** [http://www.ncbi.nlm.nih.gov/blast/mmtrace.shtml]
43. Kurtz S, Schleiermacher C: **REPuter: fast computation of maximal repeats in complete genomes.** *Bioinformatics* 1999, **15:**426-427.
44. **REPuter** [http://www.genomes.de/]
45. Zuker M: **Mfold web server for nucleic acid folding and hybridization prediction.** *Nucleic Acids Res* 2003, **31:**3406-3415.
46. Lang BF, Laforest MJ, Burger G: **Mitochondrial introns: a critical view.** *Trends Genet* 2007, **23:**119-125.
47. **RNAweasel** [http://megasun.bch.umontreal.ca/RNAweasel/]
48. Brodie R, Roper RL, Upton C: **JDotter: a Java interface to multiple dotplots generated by dotter.** *Bioinformatics* 2004, **20:**279-281.
49. Boer PH, Gray MW: **Scrambled ribosomal RNA gene pieces in *Chlamydomonas reinhardtii* mitochondrial DNA.** *Cell* 1988, **55:**399-411.
50. Smith DR, Lee RW: **Nucleotide diversity in the mitochondrial and nuclear compartments of *Chlamydomonas reinhardtii*: investigating the origins of genome architecture.** *BMC Evol Biol* 2008, **8:**156.
51. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
52. **United States Department of Energy Joint Genome Institute (JGI) *Chlamydomonas reinhardtii* nuclear genome assembly v3.0** [http://genome.jgi-psf.org/Chlre3/Chlre3.home.html]

Intron Content of the *D. salina*, *C. reinhardtii*, and *V. carteri* Organelle Genomes

Numbers of introns per gene are shown in brackets; insertion sites relative to the *C. reinhardtii* gene sequence are shown in red.

Gene	*D. salina*		*C. reinhardtii*		*V. carteri*	
MtDNA	Group-I intron	Group-II intron	Group-I intron	Group-II intron	Group-I intron	Group-II intron
cob	(4) 429, 504, 756, 828	—	(1) 411[b]	—	(1) 411	(1) 822
cox1	(5) 189, 384, 711, 1110, 1281	—	(2) 384[b], 849[b]	—	(2) 700[b], 1089	—
nad1	(1) 600	—	—	—	—	—
nad2	—	—	—	—	—	—
nad4	—	—	—	—	—	—
nad5	(2) 663, 252	—	—	—	—	—
nad6	—	—	—	—	—	—
rrnL1	—	—	—	—	—	—
rrnL2	—	—	—	—	—	—
rrnL3	—	—	—	—	—	—
rrnL4	—	—	—	—	—	—
rrnL5	(1) 264	—	(1) 50[b]	—	—	—
rrnL6	(3) 259, 311, 398	—	—	—	—	—
rrnL7[a]	N/A	N/A	(1[b]) ~5[b]	—	—	—
rrnL8[a]	N/A	N/A	—	—	—	—
rrnS1	—	—	—	—	—	—
rrnS2	(2) 585, 622	—	—	—	—	—
rrnS3	—	—	—	—	—	—
rrnS4	—	—	—	—	—	—
PtDNA	Group-I intron	Group-II intron	Group-I intron	Group-II intron	Group-I intron	Group-II intron
atpA	(1) 492	—	—	—	(1) 492	(1) 756
atpB	(1) 1443	—	—	—	—	(1) 717
atpE	—	—	—	—	—	—
atpF	—	—	—	—	—	—
atpH	—	—	—	—	—	—
atpI	—	—	—	—	—	—
ccsA	—	—	—	—	—	—
cemA	—	—	—	—	(1) 351	—

Gene	*D. salina*		*C. reinhardtii*		*V. carteri*	
chlB	—	—	—	—	—	—
chlL	(1) 876	—	—	—	(1) 201	—
chlN	—	—	—	—	—	—
clpP	—	—	—	—	—	—
ftsH	—	—	—	—	—	—
petA	—	—	—	—	—	—
petB	—	—	—	—	—	—
petD	—	—	—	—	—	—
petG	—	—	—	—	—	—
petL	—	—	—	—	—	—
psaA	(1) 2040	(2) 90, 270	—	(2) 90, 270	(1) 1605	(2) 90, 270
psaB	(1) 939	—	—	—	—	(1) 1920
psaC	—	—	—	—	—	—
psaJ	—	—	—	—	—	—
psbA	(5) 276, 384, 414, 570, 900	—	(4) 184, 204, 525, 726	—	—	—
psbB	—	—	—	—	—	—
psbC	(2) 543, 882	—	—	—	—	—
psbD	(1) 567	—	—	—	—	—
psbE	—	—	—	—	—	—
psbF	—	—	—	—	—	—
psbH	—	—	—	—	—	—
psbI	—	—	—	—	—	—
psbJ	—	—	—	—	—	—
psbK	—	—	—	—	—	—
psbL	—	—	—	—	—	—
psbM	—	—	—	—	—	—
psbN	—	—	—	—	—	—
psbT	—	—	—	—	—	—
psbZ	—	—	—	—	—	—
rbcL	—	—	—	—	—	—
rpl14	—	—	—	—	—	—
rpl16	—	—	—	—	—	—
rpl2	—	—	—	—	—	—
rpl5	—	—	—	—	—	—
rpl20	—	—	—	—	—	—
rpl23	—	—	—	—	—	—
rpl36	—	—	—	—	—	—
rpoA	—	—	—	—	—	—
rpoBa	—	—	—	—	—	—
rpoBb	—	—	—	—	—	—
rpoC1	—	—	—	—	—	—
rpoC2	—	—	—	—	—	—
rps2	—	—	—	—	—	—
rps3	—	—	—	—	—	—
rps4	—	—	—	—	—	—

Gene	*D. salina*		*C. reinhardtii*		*V. carteri*	
rps7	—	—	—	—	—	—
rps8	—	—	—	—	—	—
rps9	—	—	—	—	—	—
rps11	—	—	—	—	—	—
rps12	—	—	—	—	—	—
rps14	—	—	—	—	—	—
rps18	—	—	—	—	—	—
rps19	—	—	—	—	—	—
rrn5	—	—	—	—	—	—
rrnL	(7) 276, 1902, 1994, 2322, 2509, 2561, 2659	—	(1) 2222	—	—	—
rrnS	(4) 415, 476, 740, 881	—	—	—	—	—
tufA	—	—	—	—	—	—
ycf1	—	—	—	—	—	—
ycf3	—	—	—	—	—	—
ycf4	—	—	—	—	—	—
ycf12	—	—	—	—	—	—
PtDNA intergenic [c]	Intron-like sequence		Intron-like sequence		Intron-like sequence	
rps3/rpoC2	(1)		—		—	
rrnS/ycf1	(1)		—		—	
rpl36/petB	(1)		—		—	
psaJ/atpI	(2)		—		—	
chlL/clpP	(1)		—		—	
psbD/rps4	(1)		—		—	
trnS/rpl20	(2)		—		—	
atpB/ftsH	(1)		—		—	
rrn5/atpH	(2)		—		—	
atpE/rbcL	(2)		—		—	
trnI/psbH	(1)		—		—	

Note: Intron insertion sites are shown in red (*C. reinhardtii* was used as the reference genome for determining the nucleotide position of the intron insertion); dash (i.e., —) means no introns. As there are no introns in the organelle-DNA encoded tRNAs of *D. salina*, *C. reinhardtii*, and *V. carteri* only protein- and rRNA-coding genes are shown in the table.

[a] The LSU rRNA-coding regions of the *D. salina* mitochondrial genome are fragmented into six modules whereas those of *C. reinhardtii* and *V. carteri* are fragmented into eight coding modules (see Figures 8.1 and 8.3 [Chapter 8] for more details).

[b] Denotes optional intron.

[c] Refers to intron-like sequences found in the intergenic regions of the *D. salina* ptDNA (see Figure 8.2 [Chapter 8] for more details).