EVOLUTIONARY CHANGES IN THE EUGLENOID CHLOROPLAST

By

Krystle Elaine Wiegert

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

MASTER OF SCIENCE

Plant Biology

2011

ABSTRACT

EVOLUTIONARY CHANGES IN THE EUGLENOID CHLOROPLAST

By

Krystle Elaine Wiegert

The chloroplasts of three exemplar taxa of photosynthetic euglenoids, have been sequenced and compared, providing an overview of chloroplast evolution in this highly diverse group of organisms which acquired their chloroplast through secondary endosymbiosis. The use of synteny mapping has shown that the chloroplast genomes of euglenoids are not fully collinear, and significant gene rearrangements have occurred across the lineage. The gene rearrangements however, have occurred largely through the repositioning of conserved gene clusters. Gene content among the taxa was similar with only a few variations noted including rpl32, psaI, psaM, rrn5, rpoA and roaA genes. Comparisons between the euglenoid chloroplast genomes also highlighted significant differences in genome sizes, due mainly to expansion/shrinkage of intergenic space and an overall increase in intron number. Further comparisons with the chloroplast genome of *Euglena gracilis* and green algal taxa provided insight into chloroplast genome evolution within the Euglenophyta. Phylogenies based upon 55 chloroplast genes are nearly identical to those based on nuclear encoded genes, indicating similar evolutionary pathways. Sequence comparisons and phylogenetic analyses have also provided evidence that the putative green algal chloroplast donor lineage was a scaly green algal flagellate belonging to the Prasinophyceae and that the chloroplast was acquired through a single endosymbiotic event.

ACKNOWLEDGMENTS

The completion of this research project would not have been possible without the decision of Dr. Richard Triemer to take me on as a Masters' student in a department largely focused on doctoral education. I am indebted to him for facilitating me the past two and a half years and demonstrating to me the highest level of professionalism in science. This thesis would not be what it is were it not for his patience and encouragement, and his willingness to take the time to work through challenges. I harbor immense respect for his ability to challenge me to think critically about research problems while simultaneously treating me as an equal in discussions surrounding my research.

I owe a huge thank you to Matthew Bennett for putting up with me every day. The work we completed developing protocols and testing different genomics programs and tools was invaluable to our understanding of the world of chloroplast genomics. I will miss the many discussions brainstorming new ideas and ways to tackle challenges, thank you. You have been a great example to me of how you can run a successful research lab and still manage to have fun!

To my committee members, Dr. Robin Buell and Dr. Shinhan Shiu, thank you for taking the time to counsel me in your areas of expertise as well as my future in science. Your knowledge in the areas of genomics and bioinformatics have been invaluable to my understanding and greatly improved the quality this project.

I also owe a big thank you to the hard working undergraduates of the Triemer lab, Donovan Watza and Lauren Pagett. Thank you for dealing with some of the more tedious of tasks, and taking them on without complaint. I wish you success in your future endeavors, and again, thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

BLAST – Basic Local Alignment Sequence Tool

bp – base pairs

bt – bootstrap support in Maximum Likelihood analysis

CCAP – Culture Collection of Algae and Protozoa

CDS – DNA Coding sequence or region

cp – chloroplast

CsCl – Cesium Chloride

DNA – DeoxyriboNucleic Acid

DOGMA – Dual Organellar GenoMe Annotator

e.g. – example

EMBL – European Molecular Biology Laboratory nucleotide sequence database

Gb – Giga base pairs

GDE – Genetic Data Environment

IR – inverted repeat region

Kb – Kilobase

ML – Maximum Likelihood

MSU RTSF – Michigan State University Research Technology Support Facility

NCBI – National Center Biotechnology Information

NNI – Nearest-Neighbor-Interchange

ORF (orf) – Open Reading Frame

PCR – Polymerase Chain Reaction

Pg – picogram

PVP – Polyvinylpyrrolidone

rDNA – Ribosomal DeoxyriboNucleic Acid

RNA – RiboNucleic Acid

rpl – ribosomal protein from large subunit

rps – ribosomal protein from small subunit

rRNA – Ribosomal RiboNucleic Acid

SAG – Sammlung von Algenkulturen Gottingen culture collection

TE buffer – Tris and EDTA containing buffer

tRNA – Transfer RiboNucleic Acid

VNTR – Variable Number Tandem Repeat

ycf – conserved gene with product of unknown function

Chapter 1

Methods Utilized in Determining the Chloroplast Genomes of *Colacium vesiculosum*, *Eutreptia viridis* and *Strombomonas acuminata*

Cultures

Cells were harvested from 3 unialgal cultures ranging from 10ml-1L in volume depending on the common cell density in culture observed for each species (*Colacium vesiculosum* CCAP 1211/3, *Eutreptia viridis* SAG 1226-1c, *Strombomonas acuminata* S716). Cultures were grown in AF-6 medium (Watanabe and Hiroki, 1997) at 20 – 22° C under 10:14 light:dark cycle with approximately 30 µmol photons • m^{-2} • s^{-1} provided from cool white fluorescent tubes. Prior to havesting, microbial contamination levels were assessed and a final confirmation of species was made using a Zeiss Axioskop 2 Plus microscope (Carl Zeiss Inc., Hallbergmoos, Germany).

Cell Isolation

Cells were concentrated by centrifugation at 2500rpm for 5min using an Eppendorf 5804R (15amp version) centrifuge (Eppendorf North America, Hauppauge, NY, USA). Resulting supernatant was decanted and cell pellets combined with fresh AF-6 medium and resuspended. This was repeated at least 3 times to reduce bacterial contamination. Cells suspended in fresh AF-6 medium were then layered over a gradient of colloidal polyvinylpyrrolidone (PVP) coated silica (Centricoll®, Sigma-Aldrich Chemical Co., St. Louis, MO, USA; Cat #C0580) for further removal of bacteria. Gradients were established in 15ml glass centrifuge tubes with the bottom layer of the gradient consisting of 1.5ml 100% Centricoll ®, followed by 4ml 50% Centricoll®, and about 2ml

of concentrated cells suspended in AF-6 medium. This was spun at 4000rpm for 5 minutes with ramp down so as to not cause mixing. Individual euglenoid cells concentrated on the 50:100% interface with some small clumps of cells sedimenting within the 100% gradient layer. Bacterial cells remained suspended on top of the gradient. If the expected separation did not occur, gradients were expanded to include a greater number of incremental layers (100%, 80%, 60%, 40%, 20% Centricoll®, and culture). All cells below the 50:100 % interface were removed and brought to a volume of 15ml with AF-6 medium to begin washing of the cells and removal of the Centricoll® from the sample. This sample was spun for 4 minutes at 4000rpm to pellet the cells and the supernatant was removed. Fresh AF-6 medium was added and the pellet resuspended and centrifuged again. This was repeated two additional times to ensure removal of the silica. After the final centrifugation, cells were resuspended in 1-3ml AF-6 depending on pellet size to provide a cleaned and concentrated culture for DNA extraction.

DNA Isolation

Whole genomic DNA was extracted from cultures using a Qiagen DNeasy Blood and Tissue Extraction Kit (Qiagen Co., Valencia, CA, USA; Cat #69504) or a 50:50 phenol/chloroform solution followed by precipitation in cold ethanol and 3M sodium acetate (pH 5.4) with rehydration in TE buffer pH 8.0.

Phenol/chloroform extractions were utilized in preparation for the separation of nuclear and chloroplast DNA fractions on cesium chloride (CsCl) gradients. CsCl gradients were prepared by grinding CsCl with a mortar and pestle to a very fine consistency. About

3.5g CsCl, 3ml TE, 10mg/ml Hoechst dye 33342 (Ana Spec Inc., San Jose, CA, USA; Cat #83218), and the DNA eluted in TE were combined as a starting point and adjusted to achieve a refractive index of 1.3850 (ABBE Refractometer, American Optical Co., Buffalo, NY, USA) which was determined to provide optimum separation of the organellar DNA fraction within the gradient. Gradients were centrifuged for no less than 16 hours at 50,000rpm and 15° C using a vertical rotor. DNA was drawn off the gradient using a 16G needle while venting the top with an additional needle. This extract was then purified using a 5:1 isopropanol:TE solution, reprecipitated with cold ethanol, and eluted again in TE with 50mM NaCl.

Quality and quantity of resulting DNA (whole genomic and/or chloroplast) was inferred based on A260/A280 and A260/A230 ratios collected from both an Eppendorf BioPhotometer plus spectrophotometer (Eppendorf North America, Hauppauge, NY, USA) as well as a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific Inc., Wilmington, DE, USA).

Sequence data collected from both *Eutreptia viridis* and *Strombomonas acuminata* were from whole genomic DNA isolated with the Qiagen DNeasy Blood and Tissue Extraction Kit. Data for *Colacium vesiculosum* were acquired through both whole genomic DNA sequencing as with *Eutreptia viridis* and *Strombomonas acuminata,* in addition to the sequencing of isolated chloroplast DNA using the CsCl gradients.

Sequencing

Sequencing was completed using a Roche 454 GS FLX/Titanium Genome Analyzer (single reads - 454 Life Sciences, a Roche Co., Branford, CT, USA) and Illumina

Genome Analyzer II (54bp paired-end reads - Illumina Inc., San Diego, CA, USA) at the Michigan State University Research Technology Support Facility (RTSF). Nucleic Acids Research Facilities at Virginia Commonwealth University were also utilized in sequencing using a Roche 454 GS FLX/Titanium Genome Analyzer (single reads - 454 Life Sciences, a Roche Co.).

In the sequencing of *C. vesiculosum* ¼ plates of Roche 454 were run for both whole genomic DNA and isolated chloroplast DNA respectively. Additionally, whole genomic DNA for *Colacium vesiculosum* and *Eutreptia viridis* were included as 2 of 4 organisms barcoded within a single Illumina lane. For *Eutreptia viridis* and *Strombomonas acuminata* ¼ plates were also run with whole genomic DNA.

Sequence Assembly - Roche 454

Raw sequencing reads were assembled using the Roche GS De Novo Assembler (Newbler) with default parameters. A local BLAST database(http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/pc_setup.html) including the known chloroplast genome of *Euglena gracilis, Euglena longa*, and other completed chloroplast genomes of green algal species (*Chlamydomonas reinhardtii*, *Chlorella vulgaris*, *Chlorokybus atmophyticus*, *Mesostigma virde*, *Monomastix sp.*, *Nephroselmis olivacea*, *Ostreococcus tauri*, *Parachlorella kessleri*, *Pedinomonas minor*, *Pycnococcus provasolii*, *Pyramimonas parkea*, *Scenedesmus obliquus*, *Zygnema circumcarinatum*) was used to identify contigs containing chloroplast sequences from the Roche 454 de novo assembler outputs (454AllContigs.fa and/or 454LargeContigs.fa files).

Sequence Assembly - Illumina

Raw read data were received as .FASTQ files from the sequencing facility. These data were assembled using Velvet version 1.0 (Zerbino and Birney 2008) by the MSU RTSF under default parameters to obtain contigs from the short reads. Resulting contigs were searched as with the Roche data against a local BLAST database to identify contigs containing chloroplast sequence.

In addition to assembly using the Roche 454 de novo assembler and Velvet (Illumina) designed for their respective sequencing data, assemblies were also generated using the CLC Genomics Workbench (CLC bio, Cambridge, MA, USA) high throughput de novo assembly option to compare and ensure consistency.

Sequence Alignment and Building of Consensus Sequences

Contigs in the assemblies that were identified as chloroplast sequences by local BLAST were run through the NCBI BLAST site (http://blast.ncbi.nlm.nih.gov/Blast.cgi) to ensure they were not from bacterial species (high similarity between ribosomal gene sequences often provided high identity hits in the local BLAST database). In addition, the web-based annotation software DOGMA (Wyman et al. 2004; http://dogma.ccbb.utexas.edu/) was utilized to provide a preliminary annotation of contig gene content. Based on this information these contigs were imported into the Genetic Data Environment (GDE, Smith et al. 1994). Here an attempt was made to align contigs using the built in CAP3 alignment function. If this was not successful, alignment of contigs was completed manually aided by the DOGMA annotation information which identified the 3’ and 5’ most genes on each of the contigs. All putative chloroplast sequences were identified and aligned within GDE creating a consensus

Table 1.1: Fill-in PCR primers used in the sequencing of chloroplast genomes**.**

Genome Position	Sequence (5'- 3')
Eutreptia viridis	
65003..65022	GGA AAC ATA ACC CGT CTT GC
65132..65153	GAA AAT CCA ACT GCA AGT AAA A
Colacium vesiculosum	
Complement(4133..4154)	GAT CAG CCT GTT ATC CCT AGA G
Complement(5666..5685);	
Complement(320..339)	CAC GCG GCA TTG CTC CGT CA
Complement(5320..5344)	AGC GTT CAT CCT GAG CCA GGA TCA A
128442..128459	GGG CTA TAT GCT TCA GGT
Complement(128442..128459)	ACC TGA AGC ATA TAG CCC
5129..5149	CTA ACT CTA CCA TTC GTG TTC
4716..4736	GTC TGA GGG TCA CCT CTT ATG
Complement(1081..1098)	GCA TGA TGA CTT GAC CTC
Strombomonas acuminata	
Complement(140015..140032)	KCA GTT TTA CTG GGG CGG
138947..138964	GGC GTA GCC AAG TGG TAA
Complement(139518..139538)	GAG TGG ATA ACT GCT GAA AGC
143913..143930	CAG CGT TCA TCC TGA GCA
397..419	TAG GTA TCA GGT GTA AGG CTG GT
486..506	CTC TGG CTT GTG GTA ACA CCT
Complement(83584..83602)	GGA GAA GGT GTC TGA GTG G
Complement(775..793)	GCT TAA TGC TTG CCG GTT C

sequence of the given species' chloroplast genome. Fill-in polymerase chain reaction (PCR) was necessary for *Colacium vesiculosum* and *Strombomonas acuminata* in order to bridge sequencing gaps and extend sequences.

Fill-in PCR

Primers were designed within the flanking regions of sequencing gaps with Primer3Plus (Untergasser and Nijveen 2007, http://www.bioinformatics.nl/cgi-bin/primer 3plus/primer3plus.cgi). Reactions were completed with GoTaq® PCR (Promega, Madison, WI, USA) using whole genomic DNA and/or isolated chloroplast DNA extracted as above. Reactions were carried out on either a Bio-Rad C1000 thermal cycler (Bio-Rad Laboratories, Hercules, CA, USA) or MJ Research PTC-200 Peltier thermal cycler (MJ Research, St. Bruno, Quebec, Canada) with denaturation at 96C, extension at 72C and varying annealing temperatures ranging from 50-60C dependent upon the primers in use (Table 1.1). PCR products were sized on 1% agarose gels and then purified using a Qiagen MiniElute Gel Extraction Kit (Cat #28606). The resulting purified DNA was tested for quality and quantity with an Eppendorf BioPhotometer Plus spectrophotometer before sequencing at the MSU RTSF using an ABI 3730 Genetic Analyzer (Applied Biosystems, Foster City, CA, USA). Sequences were aligned by eye in GDE 2.2 to consensus chloroplast sequences.

Annotation

Multiple forms of evidence were compiled to verifiy both protein coding and RNA genes (Figure 1.1). Coding genes were identified through the use of ORF-Finder available from NCBI using standard parameters for bacterial and plastid sequence (Table 11;

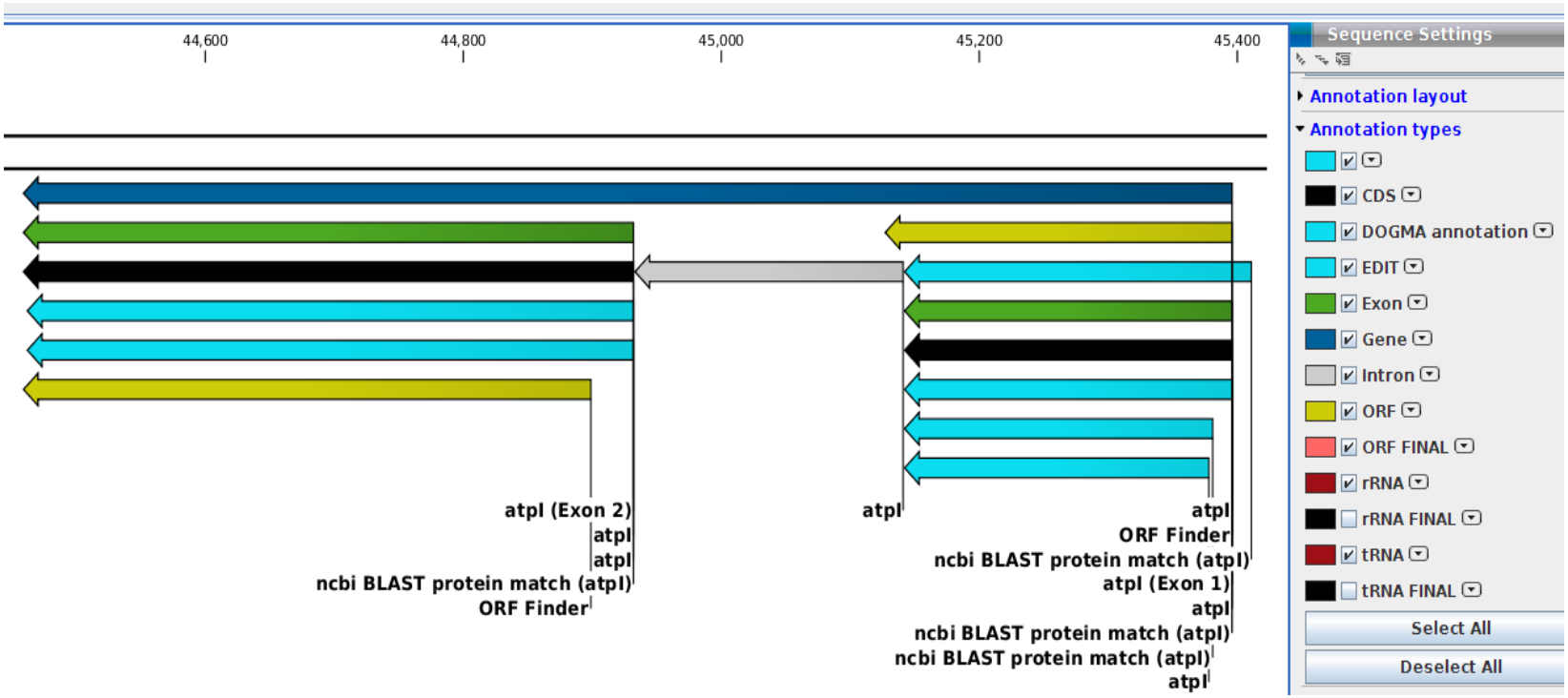

Figure 1.1: Example of gene annotation completed using the CLC Genomics Workbench. Shown is the atpI gene of *Eutreptia viridis*. Multiple colored tracks indicate different forms of evidence utilized in determining final gene characters which can be removed from view by selecting/deselecting of the given annotation type on the right (e.g. Exon, intron, CDS, gene). For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this thesis.

http://www.ncbi.nlm.ni h .gov/gorf/gorf .html), the Dual Organellar GenoMe Annotator - DOGMA (http://dogma.c cbb.utexas.edu/; Wyman et al., 2003), and various BLAST searches including nucleotide, protein, and the whole genome shotgun submission (wgs) databases (*Chlamydomonas reinhardii*; when no other hits were identified). Ribosomal RNA genes were also identified using DOGMA and BLAST. tRNAscan-SE (http://lowelab.ucsc.edu/tRNAscan-SE/; Lowe and Eddy 1997; Schattner et al., 2005) was utilized in the annotation of tRNAs. This information was compiled with the consensus sequence in the CLC Genomics Workbench allowing graphical representation for final annotation decisions in combination with the six-frame translation.

Annotation – Open Reading Frames (ORFs)

In determining the final annotation of both ribosomal and protein coding genes standards/thresholds were established. This included the decision that all orfs less than 300 nucleotides (100 amino acids) in length and lacking protein evidence were disregarded. All final orfs were labeled ‘orf’ followed by the length of the open reading frame in amino acid codons.

Annotation - Common Homopolymer Runs

Due to the known issues with homopolymeric runs in sequencing with Roche 454 pyrosequencing technology instances of primarily A/T insertions and deletions were identified. These became obvious in the return of split or overlapping BLAST hits and frame shifts within the sequence. Any predicted instances of indels based on the six-

frame translation of the consensus were often found to be locations of ambiguous bases within the sequence coverage.

Annotation - Alternative Start Codons

A number of genes were found to contain alternative start codons (Table 1.2). These were identified by either a lack of an ATG start codon or better correlation based on blast hits with a given alternative start codon.

Table 1.2: Standard start and stop codons found in bacterial and plant plastid genomes (allowed under Genetic Code Table 11). Shaded codons are those found within the sequenced chloroplast genomes.

Start	Stop
ATG (Methionine)	TAA
ATA (Isoleucine)	TAG
ATC (Isoleucine)	TGA
ATT (Isoleucine)	
CTG (Leucine)	
GTG (Valine)	
TTG (Leucine)	

Phylogenetic Analysis

Gene sequences were aligned with GDE (Genetic Data Environment) and/or MEGA 5 (Tamura et al. 2000). Maximum Likelihood phylogeny reconstruction with a general time reversible model was completed along with the heuristic method Nearest-Neighbor-Interchange (NNI) in MEGA5. A Gamma distribution was used to model evolutionary

rate differences among sites (5 categories (+*G*, parameter = 0.6260)). One hundred bootstrap replicates were completed. A total of 1863 informative sites were analyzed from the 22 tRNAs and 25858 informative sites were used in the 55 gene dataset.

Bayesian analysis was performed on the same datasets with Mr. Bayes 3.1.2 (Huelsenbeck and Ronquist 2001) using the Bayesian Information Criterion from MODELTEST 3.7, which recommended a (GTR+I+Γ) model for the 55 cp gene dataset, and (TVMef+I+G) for the 22 tRNA dataset. Four Markov chains were used in the analysis (2,000,000 generations per chain), with trees saved every 100 generations. The first 8,000 trees were discarded and convergence among the remaining trees was confirmed via the sump command. The latter trees were used to generate a majority-rule consensus tree.

Both ML and Bayesian phylogenies were midpointrooted. Trees were drawn to scale, with branch lengths measured in the number of substitutions per site.

Chapter 2

Evolution of the Chloroplast Genome in Photosynthetic Euglenoids: A Comparison of *Eutreptia viridis* and *Euglena gracilis* (Euglenophyta)

Abstract

The chloroplast genome of *Eutreptia viridis* Perty, a basal taxon in the euglenoid lineage, was sequenced and compared with that of *Euglena gracilis* Ehrenberg, a crown species. Several common gene clusters were identified and gene order, conservation, and sequence similarity was assessed through comparisons with *Euglena gracilis*. Significant gene rearrangements are present in the *Eutreptia viridis* and *Euglena gracilis* chloroplast genomes. In addition, major expansion has occurred in the *Euglena gracilis* chloroplast accounting for its larger size. However, the key chloroplast genes are present and differ only in the absence of rpl32, psaM, and roaA in *Eutreptia viridis*, and psaI in *Euglena gracilis*, suggesting a high level of gene conservation within the euglenoid lineage. Further comparisons with the plastid genomes of closely related green algal taxa have provided additional support for the hypothesis that a *Pyramimonas*-like alga was the euglenoid chloroplast donor via secondary endosymbiosis.

Introduction

The euglenoids represent a diverse, ancient eukaryotic lineage. Huber-Pestalozzi (1955) describes over 800 species belonging to about 40 genera, of which about two thirds are non-photosynthetic. Molecular phylogenies place euglenoids near the base of

the eukaryotic tree of life (Adl et al. 2005, Cavalier-Smith 1998, Lane and Archibald 2008, Moreira et al. 2007, Parfrey et al. 2010, Simpson 2004) with the phagotrophic forms positioned basal to the photosynthetic genera. This study will focus on the photosynthetic euglenoids and specifically on the acquisition and evolution of the chloroplast.

Euglenoids are believed to have gained the ability to photosynthesize through the acquisition of a chloroplast via secondary endosymbiosis of a green alga. This hypothesis originally was based on the existence of closely related non-photosynthetic phagotrophic euglenoids and the presence of three membranes surrounding the resulting chloroplast rather than the two membranes typical of green algae and higher plants (Gibbs 1978, Gibbs 1981). In recent years the sequencing of chloroplast genomes has been used to infer deep evolutionary relationships among photosynthetic lineages (Burger et al. 2007, Nock et al. 2011, Parks et al. 2009, Qiu et al. 2006). This is due in part to the small size and comparative simplicity of these genomes relative to their nuclear counterparts and to the ability to analyze multigene datasets without delving into the more complex eukaryotic nuclear genome. Sequencing of the chloroplast genome is also facilitated by the typically high chloroplast genome copy number present within cells relative to the nuclear genome and the ability to use multiplexed, massively parallel sequencing (Parks et al. 2009). At last count a total of 229 chloroplasts have been sequenced to completion, most of which are representative or economically important land plant species (Benson et al. 2011). Few are algal species and most of these are green algae with meager representation of other algal

lineages. Within the photosynthetic euglenoids, the only chloroplast genome that has been sequenced is from *Euglena gracilis* (Hallick et al. 1993).

Euglena gracilis is the model organism for photosynthetic euglenoids due to its ease in culturing and the ability to achieve high cell densities. However, *Euglena gracilis* is not a typical representative photosynthetic euglenoid and is not closely related to the phagotrophic forms believed to have been the host for the endosymbiont. The diversity of the euglenoid lineage warrants further exploration into their chloroplast evolution, which could elucidate understanding of this key basal eukaryotic lineage.

To date, all inferences regarding the chloroplast donor taxon have relied on the chloroplast sequence of *Euglena gracilis*. Based on established phylogenetic assessment of the photosynthetic euglenoids, *Euglena gracilis* is consistently a crown species of the lineage (Figure 2.1; Linton et al. 2010, Marin et al. 2003, Triemer et al. 2006).

Although euglenoids are thought to have obtained their chloroplast from a green alga, the chloroplast of *Euglena gracilis* shows some significant changes. Most notably, the *Euglena gracilis* chloroplast genome is not divided into large and small single copy regions separated by inverted repeats containing the rRNA genes (as well as a few other genes). Not only do euglenoids lack the inverted repeats, the ribosomal operon is organized in at least three and a half tandemly arranged copies (5S/23S/16S: 5S/23S/16S:5S/23S/16S:16S; Hallick et al. 1993). The *Euglena gracilis* gene content appears consistent with other sequenced green algal chloroplast genomes, but the arrangement is not (Turmel et al. 2009). This raises several questions with regard to

chloroplast evolution within the photosynthetic euglenoids. Is the chloroplast genome of *Euglena gracilis* representative of all photosynthetic euglenoids? Will the chloroplast genome of more basal taxa look more like that of green algae? What changes have occurred during the evolution of the chloroplast in the euglenoid lineage?

To address these questions the chloroplast genome of *Eutreptia viridis* was sequenced. *Eutreptia viridis* is the type species of the genus *Eutreptia* which consistently falls basal to photosynthetic euglenoids in phylogenetic assessments of both nuclear and chloroplast ribosomal genes as well as nuclear encoded protein coding genes (Figure 2.1; Linton et al. 2010, Marin et al. 2003, Triemer et al. 2006). A diagrammatic representation of euglenoid phylogeny based on analysis by Marin et al. in 2003 demonstrates the position of *Eutreptia viridis* relative to *Euglena gracilis* and the non-photosynthetic euglenoids (Figure 2.1).

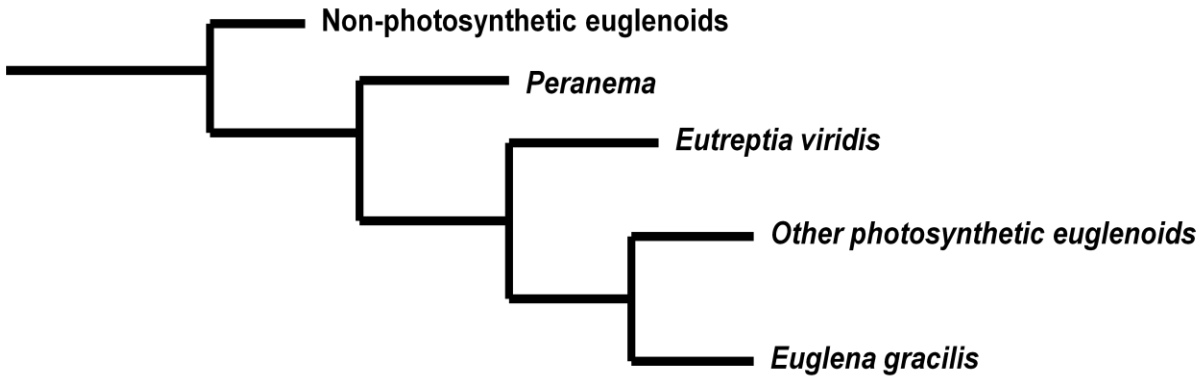

Figure 2.1: Diagrammatic representation of the current state of euglenoid phlogeny adapted from Marin et al. 2003. *Peranema* is a non-photosynthetic phagotrophic euglenoid found basal to the photosynthetic euglenoids including *Eutreptia viridis* and *Euglena gracilis*.

Furthermore, *Eutreptia viridis* differs morphologically from *Euglena gracilis* by the presence of two emergent flagella versus one. This is notable in that the closest extant

phagotrophic euglenoids such as *Peranema* also have two emergent flagella. This study will assess and compare gene content, arrangement, size, and sequence similarity between these two photosynthetic euglenoids.

Results and Discussion

The *Eutreptia viridis* chloroplast genome was assembled from reads obtained by sequencing total genomic DNA using Roche 454 sequencing and de novo assembly. Because the ribosomal genes form a tandem array rather than inverted repeats as in the green algae, it was difficult to determine the exact number of ribosomal operons and the 3' connection to the rest of the genome. Based on read coverage of the ribosomal operon components (5S, 23S, 16S) compared to single copy protein coding genes, it appears that at least two copies of the operon are present. Although a lack of overlap in sequence following the ribosomal genes to that before psaC does not allow the completion of the circularized genome, it is believed that all genes are present based on comparisons with *Euglena gracilis*. The chloroplast genome was assembled into a single contig of 65,513 bp (Figure 2.2) and is represented in circular form to facilitate comparison with other chloroplast genomes. Numbering of the sequence is not relative to the origin of replication however, as there is no evidence of a VNTR (variable number tandem repeat) sequence (Hallick et al. 1993) as found in *Euglena gracilis*. All genes excluding those of the ribosomal operon are found in single copy.

Overall, the chloroplast genome of *Eutreptia viridis* has an A+T content of 72.3% (percentages exclude the ribosomal operon) and exhibits a high gene density. A

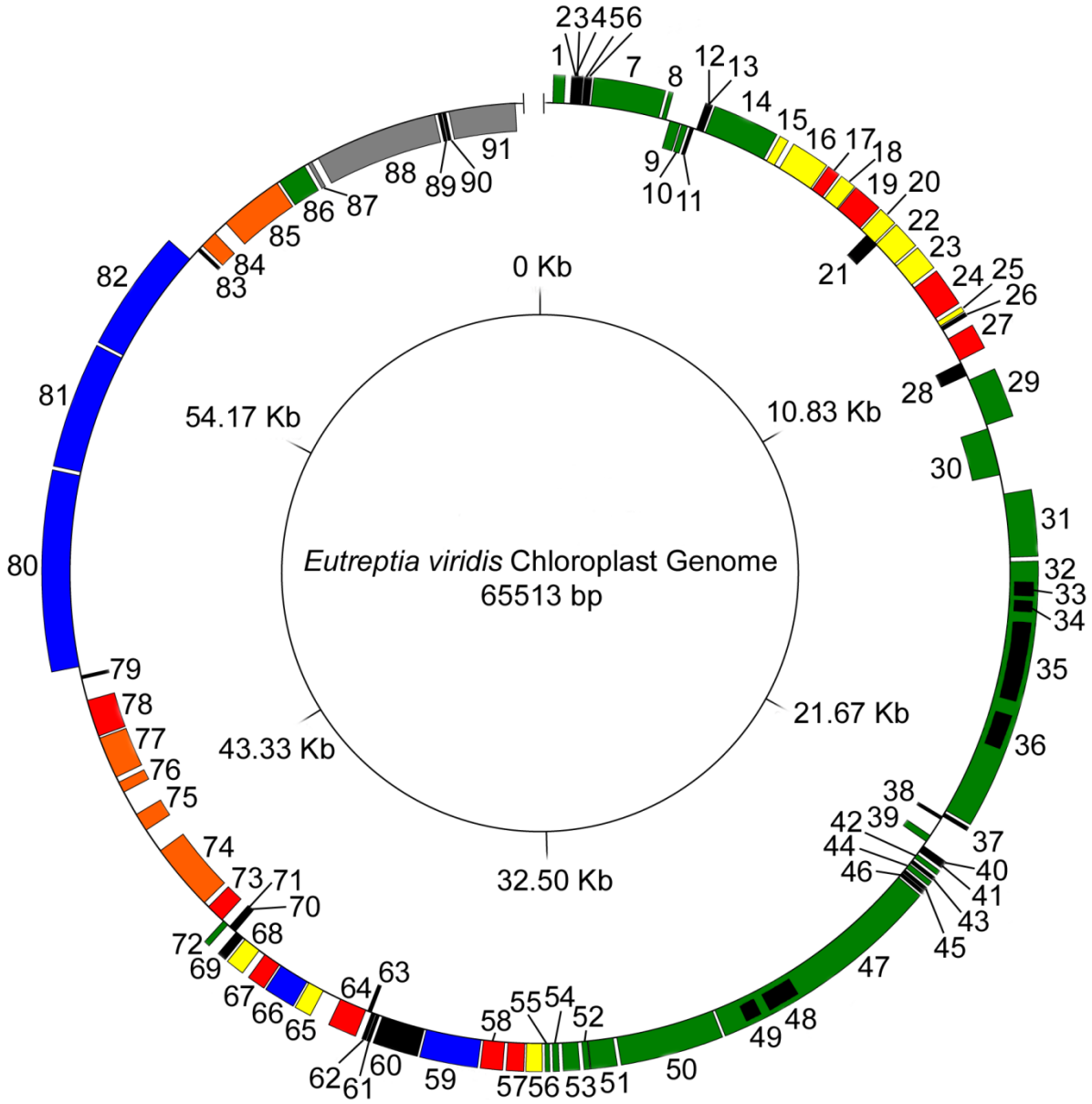

Figure 2.2: Circular map of the *Eutreptia viridis* chloroplast genome. Filled boxes of different colors represent genes of various functional groups (Green: photosystems/ photosynthesis; Yellow: large ribosomal proteins, rpl genes; Red: small ribosomal proteins, rps genes; Blue: rpo genes; Gray: ribosomal rRNAs; Orange: atp genes; Black: miscellaneous, ycf, orfs, tRNAs). Boxes are proportional to their sequence length including any introns present and those positioned on the outside of the circle are considered on the positive strand while those inside the circle are on the negative strand. Genes were annotated based on the best available nomenclature and tRNAs specifically by their single letter code with associated anticodon in parentheses. 1: psaC, 2: H(GUG), 3: M(CAU), 4: W(CCA), 5: E(UUC), 6: G(UCC), 7: psbB, 8: psbT,

Figure 2.2 con't: 9: psbH, 10: psbN, 11: L(UGA), 12: R(ACG), 13: V(UAC), 14: rbcL, 15: rpl23, 16: rpl2, 17: rps19, 18: rpl22, 19: rps3, 20: rpl16, 21: orf103, 22: rpl14, 23: rpl5, 24: rps8, 25: rpl36, 26: M(CAU), 27: rps14, 28: ycf65, 29: psbA, 30: ccsA/chlI, 31: psbD, 32:psbC, 33: orf98, 34: orf78, 35: orf564, 36: orf246, 37: L(UAA), 38: R(UCU), 39: psbK, 40: T(UGU), 41: M(CAU), 42: psbI, 43: D(GUC), 44: petG, 45: K(UUU), 46: F(GAA), 47: psaA, 48: orf229, 49: orf113, 50: psaB, 51: psbE, 52: psbF, 53: psbL, 54: psbJ, 55: psaI, 56: rpl20, 57: rps12, 58: rps7, 59: tufA, 60: ycf4, 61: Q(UUG), 62: S(GCU), 63: Y(GUA), 64: rps4, 65: rps11, 66: rpoA, 67: rps9, 68: rpl12, 69: ycf9, 70: S(UGA), 71: P(UGG), 72: psaJ, 73: rps18, 74: atpA, 75: atpF, 76: atpH, 77: atpI, 78: rps2, 79: C(GCA), 80: rpoB, 81: rpoC1, 82: rpoC2, 83: N(GUU), 84: atpE, 85: atpB, 86: petB, 87: rrn5, 88: rrn23, 89: A(UGC), 90: I(GAU), 91: rrn16.

total of 84 genes were identified including 56 protein coding, 25 tRNAs, 3rRNAs and 1 pseudogene rps2 (included in total/protein coding, Table 3). The rps2 gene was annotated as a pseudogene due to rampant internal frame shifts that could not be resolved with multiple tracks of evidence and for which there was no indication of sequencing error. The final annotation of the gene includes 3 exons. These exons were likely shortened in the process of annotation in order to eliminate frame shifts within the coding regions. Not included in the total of 84 genes were 7 open reading frames that were annotated (orfs 78, 98, 103,113, 229, 246, and 564, Figure 2.2). Taking into consideration only the protein coding genes, the average gene length is 587 bp (gene sizes represented in Figure 2.2 include introns). These genes were found to contain a total of 23 identified introns. In addition, alternative start codons were observed including that of ATT for the genes psaI, rps11, and atpE. The alternative start codon GTG was identified in petB.

The sequenced chloroplast genome of *Eutreptia viridis* is ~65,513 bp and lacks any evidence of an inverted repeat region (IR; Figure 2.2). In comparison, the published *Euglena gracilis* chloroplast genome is 143,170 bp in length and includes 3 identical copies of the ribosomal operon (5S, 23S, 16S) plus an additional copy of the 16S

(Hallick et al. 1993). If the size of the *Euglena gracilis* chloroplast is taken into consideration including only one copy of the repeated ribosomal operon, the genome totals about 120,500 bp. With the *Eutreptia viridis* chloroplast genome roughly 55,000 bp smaller, a huge difference in gene density and size is present. Thus when compared, the gene density of *Eutreptia viridis* is not only higher, but the average gene length of 587 bp increases to 925 bp in *Euglena gracilis*. In combination with this gene expansion, the chloroplast of *Euglena gracilis* also demonstrates a slightly higher A+T base composition of 76.4% (72.3% in *Eutreptia viridis*, excludes the ribosomal operons for comparative purposes; Hallick et al. 1993). Furthermore, in *Eutreptia viridis* nearly three-quarters of the genes are located on the plus strand (Figure 2.2), while in *Euglena gracils* only about half are located on the plus strand.

Introns

The genome expansion present in the *Euglena gracilis* chloroplast is most likely correlated with the larger intergenic regions and introns. The number of introns in *Euglena gracilis* (196; Benson et al. 2011) is nearly nine times that found in *Eutreptia viridis*. However, the introns of *Euglena gracilis* have been well studied and are unique in that they include the presence of group I, II, and III introns as well as twintrons (introns within introns) which have all been identified and annotated based on structural modeling (Christopher and Hallick 1989). This analysis has not been completed for *Eutreptia viridis,* therefore the number of introns will likely rise, especially with the possibility of having introns within other introns which defines a twintron (Copertino and Hallick 1993) and possible horizontal intron transfer (Sheveleva and Hallick 2004). Our sequencing and annotation of the *Eutreptia viridis* chloroplast genome has focused

primarily on the genes, thus further inquiry into differences in intron size, composition, and arrangement may add further to our understanding of the evolutionary divergence between the two species. It has been hypothesized that the chloroplast of ancestral photosynthetic euglenoids had no introns and that their origin was through the intergenomic proliferation of mobile introns (Copertino et al. 1991, Doetsch et al. 1998, Lambowitz and Belfort 1993, Thompson et al. 1995). Based solely on a comparison of the number of introns present between these two taxa, this hypothesis would be consistent with the genome expansion seen between the *Eutreptia viridis* and *Euglena gracilis* chloroplasts.

Gene Content

The gene content between the two chloroplasts is the same with regard to the essential ribosomal RNA and transcription/translation genes (Table 2.1; Hallick et al. 1993). However, the *Eutreptia viridis* chloroplast lacks the ribosomal protein L32 gene (rpl32), and photosynthetic protein psaM, the photosystem I M-polypeptide. The chloroplast of *Eutreptia viridis* also contains psaI, a photosystem I gene not found in the *Euglena gracilis* chloroplast. In addition, a more recently annotated gene in the *Euglena gracilis* chloroplast, roaA, once identified as orf516, was not found in the *Eutreptia viridis* chloroplast (Table 2.1, Figure 2.2). Further indication of the absence of roaA is that the distance between the rps3 and rpl16 genes, between which the roaA is found in *Euglena gracilis*, is only 53 bp in *Eutreptia viridis* (Christopher and Hallick 1990, Hallick et al. 1993, Jenkins et al. 1995). Pseudo-tRNAs were not identified in the *Eutreptia viridis* chloroplast in contrast to the 9 pseudo-tRNAs found in *Euglena gracilis* (Hallick et al. 1993). However, like *Euglena gracilis*, intron containing tRNAs were not found

Table 2.1: List of all genes identified in the *Eutreptia viridis* chloroplast genome. Gene abbreviations follow most recent usage and include orf: open reading frame; ycf: conserved open reading frame with unknown function between multiple genomes; rpl: ribosomal protein of the large subunit; rps: ribosomal protein of the small subunit.

Ribosomal RNAs and Proteins	
Gene	Description
23S rRNA	23S ribosomal RNA
16S rRNA	16S ribosomal RNA
5S rRNA	5S ribosomal RNA
rpl2	ribosomal protein L2
rpl5	ribosomal protein L5
rpl12	ribosomal protein L12
rpl14	ribosomal protein L14
rpl16	ribosomal protein L16
rpl20	ribosomal protein L20
rpl22	ribosomal protein L22
rpl23	ribosomal protein L23
rpl36	ribosomal protein L36
rps2	ribosomal protein S2
rps3	ribosomal protein S3
rps4	ribosomal protein S4
rps7	ribosomal protein S7
rps8	ribosomal protein S8
rps9	ribosomal protein S9
rps11	ribosomal protein S11
rps12	ribosomal protein S12
rps14	ribosomal protein S14
rps18	ribosomal protein S18
rps19	ribosomal protein S19
Transfer RNAs	
trnA	tRNA: Ala(UGC)
trnC	tRNA: Cys(GUC)
trnD	tRNA: Asp(GUC)
trnE	tRNA: Glu(UUC)
trnF	tRNA: Phe(GAA)
trnG	tRNA: Gly(UCC)
trnH	tRNA: His(GUG)
trnI	tRNA: Ile(GAU)
trnK	tRNA: Lys(UUU)
trnL	tRNA: Leu(UAG)

Table 2.1 (cont'd)

trnM	tRNA: Met(CAU)
trnN	tRNA: Asn(GUU)
trnP	tRNA: Pro(UGG)
trnQ	tRNA: Gln(UUG)
trnR	tRNA: Arg(UCU)
trnR	tRNA: Arg(ACG)
trnS	tRNA: Ser(GCU)
trnS	tRNA: Ser(UGA)
trnT	tRNA: Thr(UGU)
trnV	tRNA: Val(UAC)
trnW	tRNA: Trp(CCA)
trnY	tRNA: Tyr(GUA)
Transcription/Translation	
rpoA	RNA polymerase α subunit
rpoB	RNA polymerase β subunit
rpoC1	RNA polymerase β' subunit
rpoC2	RNA polymerase β '' subunit
tufA	translation elongation factor EF-Tu
Photosynthetic Proteins	
psaA	photosystem I P700 apoprotein A1
psaB	photosystem I P700 apoprotein A2
psaC	photosystem I subunit VII
psaI	photosystem I 4 kDa hydrophobic subunit
psaJ	photosystem I 5 kDa protein
psbA	photosystem II core 32 kDaprotein
psbB	photosystem II CP47 chlorophyll apoprotein
psbC	photosystem II CP43 chlorophyll apoprotein
psbD	photosystem II core 34 kDa protein
psbE	photosystem II cytochrome b559 α subunit
psbF	photosystem II cytochrome b559 β subunit
psbH	photosystem II 10 kDa protein
psbI	photosystem II I polypeptide
psbJ	photosystem II J protein
psbK	photosystem II 3.9 kDa protein
psbL	photosystem II L protein
psbN	photosystem II N protein
psbT	photosystem II T protein
petB	cytochrome b6

Table 2.1 (cont'd)

petG	cytochrome b6/f complex subunit V
rbcL	RuBisC/O large subunit
chlI/ccsA	chlorophyll biosynthesis
ATPases	
atpA	ATPase α subunit
atpB	ATPase β subunit
atpE	ATPase ϵ subunit
atpF	ATPase subunit I
atpH	ATPase subunit III
atpI	ATPase subunit IV
ORFs with identity to other chloroplast orfs	
ycf4	polar transcribed with tufA
ycf9	hydrophobic and in land plants
ORFs with no similarity/unknown function	
orf78	encoded in psbC
orf98	encoded in psbC
orf103	encoded on opposite strand spanning rpl14-rpl16
orf113	encoded in psaA
orf229	encoded in psaA
orf246	encoded in psbC
orf564	encoded in psbC

(Hallick et al. 1993, Lowe and Eddy 1997, Manhart and Palmer 1990, Schattner et al. 2005). Similarly, the conserved open reading frames ycf4 and ycf9 found in *Euglena gracilis* were identified in *Eutreptia viridis*, but ycf8, ycf12, and ycf13 were not present (Table 2.1; Hallick et al. 1993). These slight differences in gene content could be due to the inability to identify the genes based on the currently available homologous sequences. Gene identification would likely be improved with the addition of euglenoid chloroplast sequences. The genes not found in the *Eutreptia viridis* chloroplast are relatively small which also could have negatively influenced their chances of proper annotation.

The rpl32 gene is found in land plants, green algae, and *Euglena gracilis* (Lemieux et al. 2007, Turmel et al. 2009, Xiong et al. 2009). The absence of this gene in *Eutreptia viridis* suggests that it either has not been annotated or more likely was lost in the ancestor to the *Eutreptia* lineage.

In comparison with the land plants in general, there are several features common to the euglenoids chloroplasts. Both euglenoids lack the rpl33, infA, clpP, frxB, ndhA-K, petA, petD, psbM, rpl32, rps15, and rps16 genes which are common in land plant chloroplasts and contain rpl5, chlI (ccsA) and tufA not found in the land plant chloroplasts (Hallick et al. 1993).

Ribosomal Operons

The chloroplast genome of *Euglena gracilis* includes three identical copies of the ribosomal operon (5S, 23S, 16S) plus an additional copy of the 16S (Hallick et al. 1993). Studies have shown, however, that the arrangement and number of ribosomal operons can be chimeric in the *Euglena gracilis* chloroplast, therefore comparisons across species boundaries could be difficult (Koller and Delius 1982, Hallick et al. 1993). Only one copy of the operon has been included in the sequence of the *Eutreptia viridis* chloroplast. Sequencing coverage of the rRNA genes was more than twice that of known single copy protein coding genes indicating that at least two copies are present. Additional sequencing of the rRNA genes achieved through PCR and long range PCR all produced identical sequence indicating that all copies were equivalent. Analysis by Southern blot would have provided a more accurate indication of the number of copies, but was not completed due to the microgram amount of purified chloroplast DNA that would have been necessary for a single experimental replicate. Microgram quantities of

chloroplast genomic DNA are easily obtained with *Euglena gracilis,* but are difficult to obtain with the slower growing and less dense cultures of *Eutreptia viridis*.

Open Reading Frames

The method of naming orfs utilized in the annotation of the *Euglena gracilis* chloroplast genome (Hallick et al. 1993) was employed with *Eutreptia viridis*. The *Eutreptia viridis* chloroplast genome has seven named orfs (Table 2.1). These were named based on the length of the coding sequence in amino acid residues. A number of these orfs may be homologous to annotated genes within the *Euglena gracilis* chloroplast as well as other green algal chloroplast genomes but resulting BLAST hits were very weak (alighnment scores <40). The open reading frames orf98, orf78, orf564, and orf246 are within the psbC gene and have weak sequence similarity to the maturase-like proteins annotated in the *Euglena gracilis* chloroplast (Copertino et al. 1994, Doetsch et al. 1998, Mohr et al. 1993, Zhang et al. 1995). Doetsch et al. (1998) sequenced this region in a number of euglenoid species and were successful in obtaining transcripts of these mat-like proteins from *Euglena gracilis* but not for *Eutreptia sp*. In addition, psaA was also found to contain two orfs (orf229 and orf113), while orf103 was identified overlapping rpl16 and rpl14 on the opposite strand (negative strand; Figure 2.2).

Alternative Start Codons

The identification of alternative start codons (other than ATG) in the *Eutreptia viridis* chloroplast genome was unexpected due to the lack of alternative starts in *Euglena gracilis* (Hallick et al. 1993). In the *Euglena gracilis* chloroplast genome only one gene lacks a methionine (ATG) start codon and that is psbN for which no start codon has been identified. The two alternative start codons recognized in the annotation of the

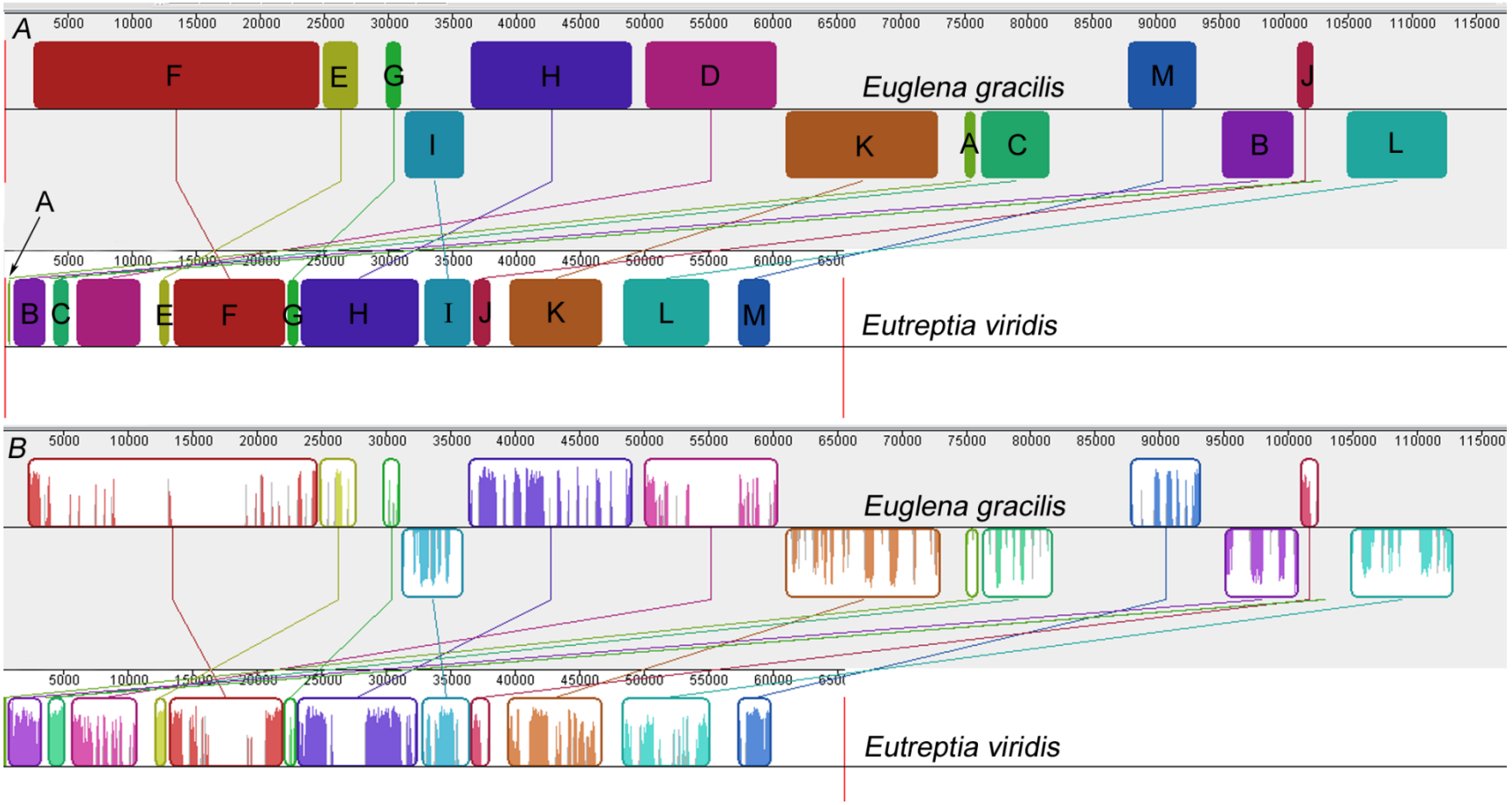

Figure 2.3: A. Relative rearrangements of gene clusters found between *Eutreptia viridis* and *Euglena gracilis* chloroplast genomes. *Eutreptia viridis* has been designated as the reference genome and as such all gene blocks are shown on the same strand. Strand orientation of the gene blocks cannot be distinguished from this representation, only comparison with regard to that of *Eutreptia viridis*. B. The same Mauve analysis showing the gene density within the homologous gene clusters.

Eutreptia viridis chloroplast genome, including ATT (psaI, rps11, and atpE) and GTG (petB) could result from introns near the 5' start of the gene leaving short coding sequences with the canonical start codon upstream. These are not likely to be identified through homology searches.

Gene Arrangement

Gene content and rearrangement was analyzed using Mauve (http://gel.ahabs.wisc.edu/mauve; Darling et al. 2004). Figure 2.3 illustrates the gene rearrangements between the *Eutreptia viridis* and *Euglena gracilis* chloroplast genomes. Each genome is displayed as a linear sequence with blocks representing homologous gene clusters. This alignment is rooted using the gene arrangement of *Eutreptia viridis* owing to its basal phylogenetic position relative to *Euglena gracilis*. Blocks lying above the center line are in the same orientation relative to the *Eutreptia viridis* genome sequence, while those below the center line are in the reverse orientation relative to the *Eutreptia viridis* genome sequence (Darling et al. 2004).

Although gene content is similar in the two euglenoids studied, there have been significant gene rearrangements as shown in Figure 2.3 panel A. Because of the large sequence divergence between the genes of the ribosomal operons (19.7% difference in the 16S gene alone) Mauve did not detect sequence similarity in this region of the genome (>115,000 bp in *Euglena gracilis*, >60,000 bp in *Eutreptia viridis*). Similarly, any regions for which homology could not be established are not shown in the analysis and likely contain elements specific to a particular genome (observed as empty space between blocks, Figure 2.3).

The 13 primary gene clusters (A-M; Table 2.2) identified show significant rearrangement in position and strand orientation between *Eutreptia viridis* and *Euglena gracilis* (Figure 2.3). For example, block D directly follows block H in *Euglena gracilis.* In *Eutreptia viridis* the two blocks are separated by ~ 15Kb and block D is positioned in front of block H. Block E has not only shifted position, but has switched strands and was inserted between blocks F and G in *Euglena gracilis*.

The rpo genes have shifted relative to the ribosomal RNA operon (Figure 2.2, Figure 2.3). In the *Euglena gracilis* chloroplast genome the rpoB gene flanks the ribosomal operon with the same polarity (same strand; Yepiz-Plascencia et al. 1990). In *Eutreptia viridis,* the order of the rpo genes has been a) inverted, b) shifted to the opposite strand and c) the gene block M (atpE-petB) moved between the rpo genes (rpoC2) and the ribosomal operon (Figure 2.3). Despite the numerous rearrangements, the gene order within clusters is conserved. A dot plot was also generated to aid in inferring the amount of rearrangement between the two chloroplast genomes (Figure 2.4, YASS,http://bioinfo . lifl.fr/yass/yass.php, Noe and Kucherov 2005). The three complete copies of the ribosomal operon can be observed in the bottom right corner while the rearrangements between the two genomes in the rest of the genome leave the remaining plot space a scattered matrix of hits with little correlation (Figure 2.4). However, identification of the homologous gene clusters found by Mauve were present and can be referenced by their block letter (Figure 2.3, Figure 2.4).

Gene Density and Homology

Figure 2.3 panel B shows the gene density and level of sequence homology between *Eutreptia viridis* and *Euglena gracilis*. Block F contains the genes chlI/ccsA, psbD,

Table 2.2: Gene clusters resulting from Mauve analysis. Blocks were labeled with letters for clarity (Figure 2.3). Genes flanking the clusters or the single genes present are listed. Additional genes present within the clusters are listed for each of the chloroplast genomes compared.

Block	Gene Clusters	*Euglena gracilis* additions	*Eutreptia viridis* additions
A	psaC		
B	tRNA-His(GUG) – psbN		His(GUG)
C	rbcL		
D	Rpl2 – rps8	rpl36	
E	psbA		
F	ccsA/chlI – psbC	Ycf13	
G	tRNA-Arg(UCU) – tRNA-Met(CAU)	ycf12, psaM, Gly(GCC)	
H	psbI – psbJ		Phe(GAA)
I	rpl20 – tRNA-Gln(UUG)	Ser(GCU)	
J	tRNA-Tyr(GUA) – rps11		
K	rpl12 – rps2	Phe(GAA), Cys(GCA)	
L	rpoB – rpoC2		
M	atpE - petB		

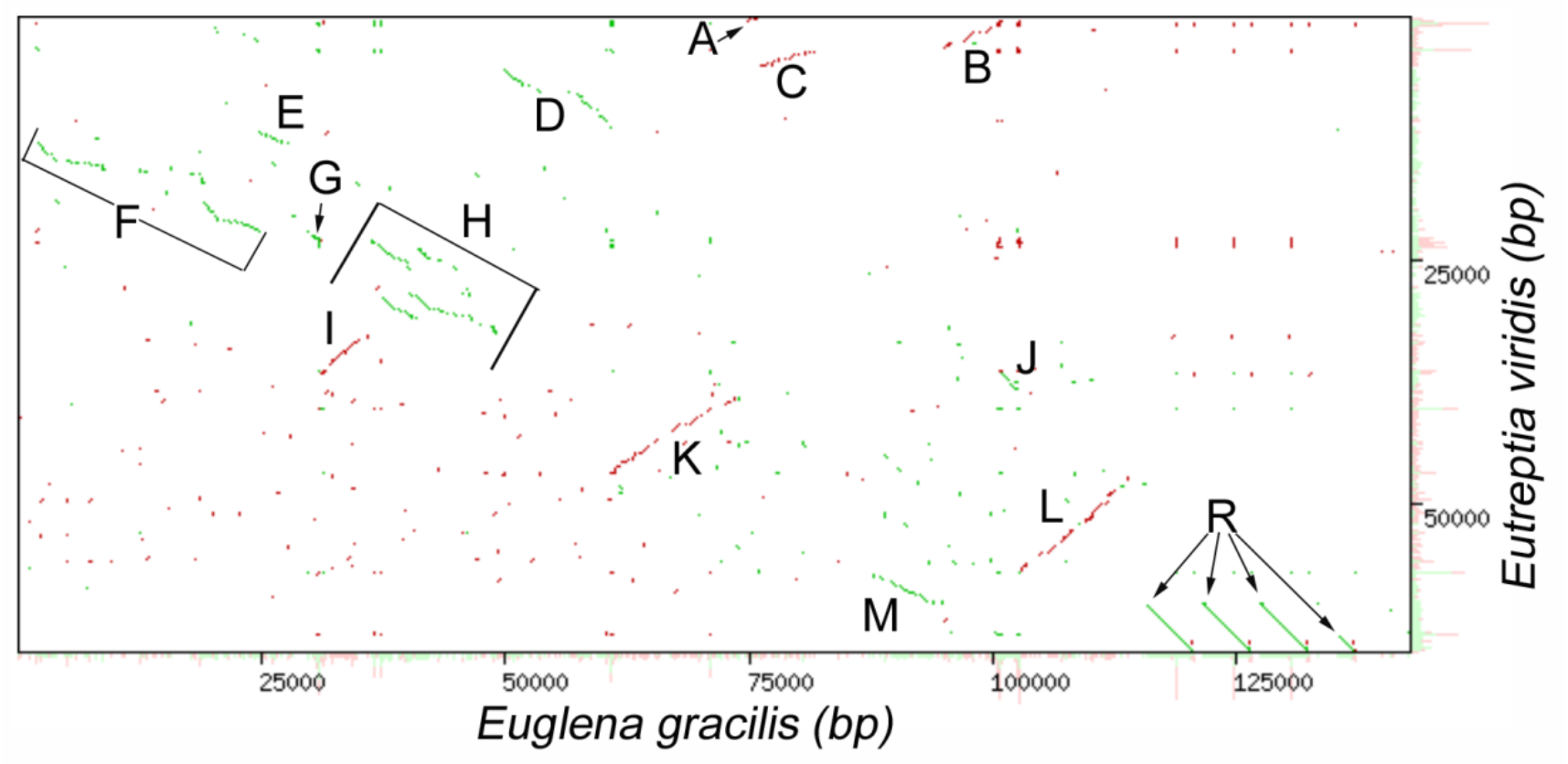

Figure 2.4: Dot plot demonstrating the overall synteny between the two chloroplast genomes of *Eutreptia viridis* and *Euglena gracilis* as calculated by YASS. Calculation was completed under standard parameters including an E value of 10 and considered both the forward (green) and reverse (red) sequence comparisons to identify synteny between gene clusters located on both strands. Letters A-M reference the gene clusters identified by Mauve in Figure 2.3 (Table 2.2). R identifies the ribosomal operon, which can be seen in 3 copies in the *Euglena gracilis* chloroplast genome.

psbC, and additionally ycf13 in *Euglena gracilis* and clearly illustrates the differences in gene density between the 2 chloroplast genomes (Table 2.2). The expansion of this gene cluster from *Eutreptia viridis* at ~5Kb to *Euglena gracilis* at ~25Kb is evidence of both an increase in introns and intergenic space (white space). The genes psbD and psbC each consist of 2 exons in the *Eutreptia viridis* chloroplast genome but have expanded to eleven exons each in *Euglena gracilis*. Each of the gene clusters (A-M) displays this expansion resulting in the large difference in chloroplast genome size between *Eutreptia viridis* and *Euglena gracilis*. Additional expansion can be seen between the blocks in *Euglena gracilis*, while the blocks in *Eutreptia viridis* are much more compact.

Evolutionary Implications

The chloroplast genome of the photosynthetic euglenoid *Eutreptia viridis* has provided insight into the evolution of this diverse lineage. The overall similarity in gene content between the chloroplast genomes of *Eutreptia viridis* and *Euglena gracilis* suggest that there was likely only a single secondary endosymbiotic event that led to the establishment of the euglenoid chloroplast. Alternatively, multiple endosymbiotic events could have occurred prior to the divergence of these two taxa. The small differences in gene content that were found are likely due to evolutionary changes or to limitations in current annotation.

The question now arises as to which taxon was the chloroplast donor. Turmel et al. (2009) using a chloroplast multigene phylogeny, showed that the chloroplast genes of *Euglena gracilis* (the only photosynthetic euglenoid chloroplast genome sequenced;

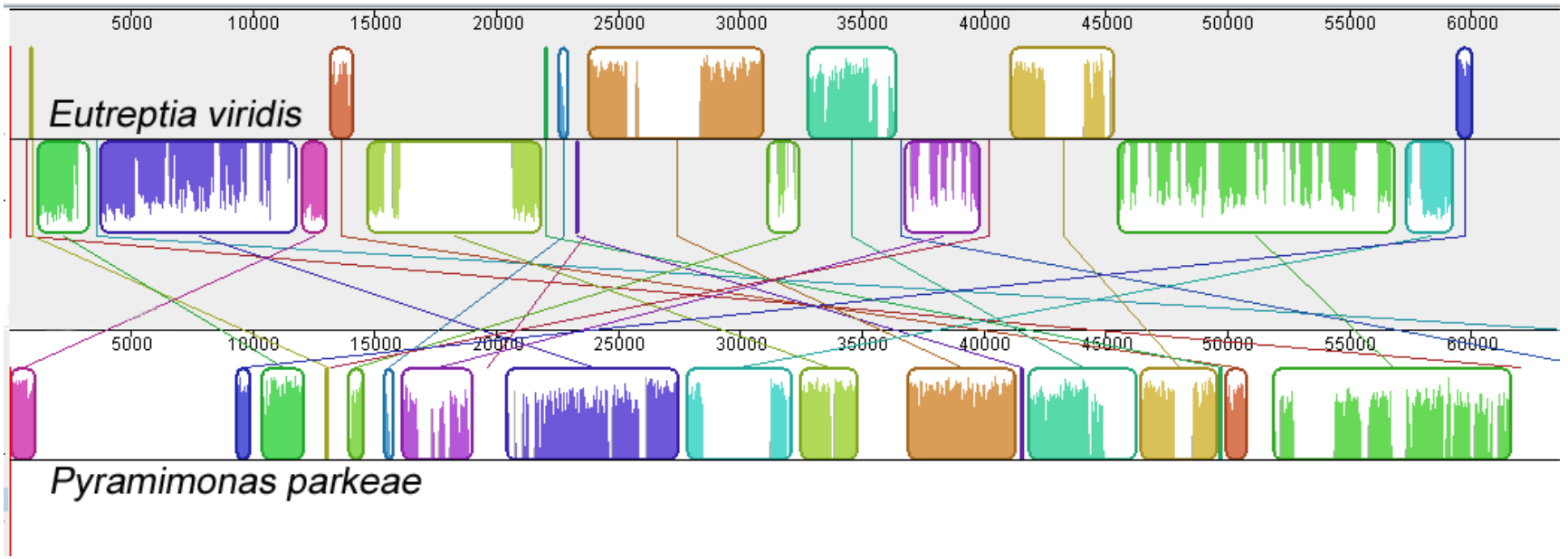

Figure 2.5: Rearrangements of gene clusters and levels of gene density and homology identified between the *Eutreptia viridis* and *Pyramimonas parkeae* chloroplast genomes. *Pyramimonas parkeae* has been designated as the reference genome. Strand orientation of the gene blocks cannot be distinguished from this representation, only comparison to that of *Pyramimonas parkeae*.

Pyramimonas parkae. Therefore, based on previous phylogenetic studies of photosynthetic euglenoids which established the basal position of *Eutreptia viridis* (Triemer et al. 2006), we hypothesized that the chloroplast genome of *Eutreptia viridis* would be more similar to that of the ancestral euglenoid chloroplast donated by the green alga than to the chloroplast genome of *Euglena gracilis*. This was found to be true. The individual chloroplast genes from *Eutreptia viridis* showed greater homology to those of *Pyramimonas parkeae* than to those of *Euglena gracilis* further substantiating an ancestral *Pyramimonas*-like species as the likely euglenoid chloroplast donor (Figure 2.5). The level of gene homology within blocks is indicated by the height of the lines which reach nearly to the top of the enclosed blocks (Figure 2.5). Compare these line heights to those shown in Figure 2.3 panel B with *Euglena gracilis*. The topology of the gene homology can be readily followed in corresponding blocks with small differences arising from expansion in the *Eutreptia viridis* chloroplast genome relative to *Pyramimonas parkeae* (additional white space). In comparing *Eutreptia viridis* and *Pyramimonas parkeae* it is clear that homology is greater between individual genes and that minimal gene expansion was found in *Eutreptia viridis* relative to that seen in *Euglena gracilis* (Figure 2.3, Figure 2.5).

Conclusions

This study has shown evolutionary changes that have occurred in the photosynthetic euglenoid chloroplast. While gene content is very similar in the chloroplasts of the two photosynthetic euglenoids examined and several common gene clusters have been identified, there have been significant rearrangements in the genomes and a major expansion of the genome has occurred in *Euglena gracilis*. Comparison of the *Eutreptia*

viridis chloroplast genome with *Pyramimonas parkae* provides additional support for the hypothesis that a *Pyramimonas*-like ancestor was the chloroplast donor to the euglenoid lineage.

Chapter 3

Tracing Patterns of Chloroplast Evolution in Euglenoids: Contributions from *Colacium vesiculosum* and *Strombomonas acuminata* (Euglenophyta)

Abstract

The chloroplast genomes of two photosynthetic euglenoids, *Colacium vesiculosum* Ehrenberg and *Strombomonas acuminata* (Schmarda) Deflandre have been sequenced. These chloroplast genomes in combination with those of *Euglena gracilis* and the recently sequenced *Eutreptia viridis* provide a snapshot of euglenoid chloroplast evolution allowing comparisons of gene content, arrangement, and expansion. *C. vesiculosum* and *S. acuminata* chloroplast genomes have been found to harbor similar gene content and although large rearrangements have occurred, the genomes were more like that of *Euglena gracilis* than the basal *Eutreptia viridis* chloroplast genome. These two new euglenoid chloroplast genomes also show evidence of further intergenic expansion, even more so than that in *Euglena gracilis,* contributing to slightly larger genome sizes. Therefore, no clear transition from the *Eutreptia viridis* chloroplast to that of *Euglena gracilis* was observed, especially in regard to intergenic expansion and genome size.

Introduction

The landscape of available sequence data for euglenoid chloroplasts had, until recently, been limited to *Euglena gracilis* (Hallick et al. 1993). This chloroplast (cp) genome was considered typical of the photosynthetic euglenoids based on the status of *Euglena gracilis* as a model experimental organism. In comparison with other sequenced algal and higher plant chloroplasts, the *Euglena* cp genome had unique gene content,

tandemly arrayed ribosomal genes (unlike the inverted repeat found in most algae and higher plants) and most notably, a large number of introns and repeated sequence. However, *Euglena gracilis* is not a typical euglenoid and considerable morphological and sequence diversity exists within the genus *Euglena* (Linton et al. 2010, Triemer et al. 2006). The recent sequencing of the cp genome of *Eutreptia viridis*, a basal photosynthetic euglenoid, provided evidence of significant changes in gene content and arrangement when compared to *Euglena gracilis*, as well as a much smaller genome size. The significant differences found between these two euglenoid chloroplast genomes, one from a basal species and one from a crown species, raises several questions about how the chloroplast has evolved within the euglenoid lineage. Can the differences found between the *Euglena* and *Eutreptia* cp genomes be tracked along the phylogenetic tree? Are the changes in genome size, intron number, gene content and gene re-arrangments transitional or random? To address these questions, the chloroplasts of two additional photosynthetic euglenoids, *Colacium vesiculosum* and *Strombomonas acuminata* were sequenced. These two species represent morphologically diverse genera within the photosynthetic euglenoids and are monophyletic as shown by a decade of nuclear ribosomal and protein coding gene sequencing (Linton et al. 2010, Triemer et al. 2006).

Colacium species are *Euglena*-like cells, which form mucilaginous stalks, and are commonly found growing in association with filamentous algae, and aquatic plants and animals (Leedale 1967). *Colacium* chloroplasts are discoid and parietal and may or may not contain a pyrenoid (Ciugulea and Triemer 2010). *Strombomonas* species are best recognized by their protective shell, or lorica. The chloroplasts are typically large

parietal disks with a pyrenoid that projects toward the cell center and is commonly covered by a cup-shaped paramylon cap (haplopyrenoid). Alternatively, some species may contain a pyrenoid covered on either side by a paramylon cap (diplopyrenoid) similar to that found in several *Euglena* species (Ciugulea and Triemer 2010). This study describes the chloroplast genome sequences of *Colacium* and *Strombomonas* and compares these with the known sequences from *Euglena* and *Eutreptia* to track the evolutionary changes occurring within the cp genome through the euglenoid lineage.

Results and Discussion

The chloroplast genomes of two photosynthetic euglenoids, *C. vesiculosum* (Figure 3.1) and *S. acuminata* (Figure 3.2), have been sequenced. Their chloroplast genomes are estimated at 128,900 bp and 144,166 bp in length, respectively. Genomes are presented in circular format for comparison with other available chloroplast sequences, however complete circularization was not achieved due to an unknown organization of the ribosomal genes. The 16S and 23S genes of several species of *Colacium* and *Strombomonas* have been sequenced recently and used in phylogenetic studies (Kim et al. 2010, Milanowski et al. 2001), thus we have focused primarily on the remaining genes of the cp genome and will consider the rrn genes only briefly. Two copies of the 16S and one copy of the 23S were identified in *C. vesiculosum*. Only one copy of the 16S matched the previously published sequence (Milanowski et al. 2001). This is unlike the cp genomes of *Euglena gracilis* and *Eutreptia viridis*, in which the rRNA genes (excluding pseudogenes) have identical sequence. Additional 23S sequence could not be obtained through multiple sequencing runs and fill-in PCR (Table 1.1), therefore the

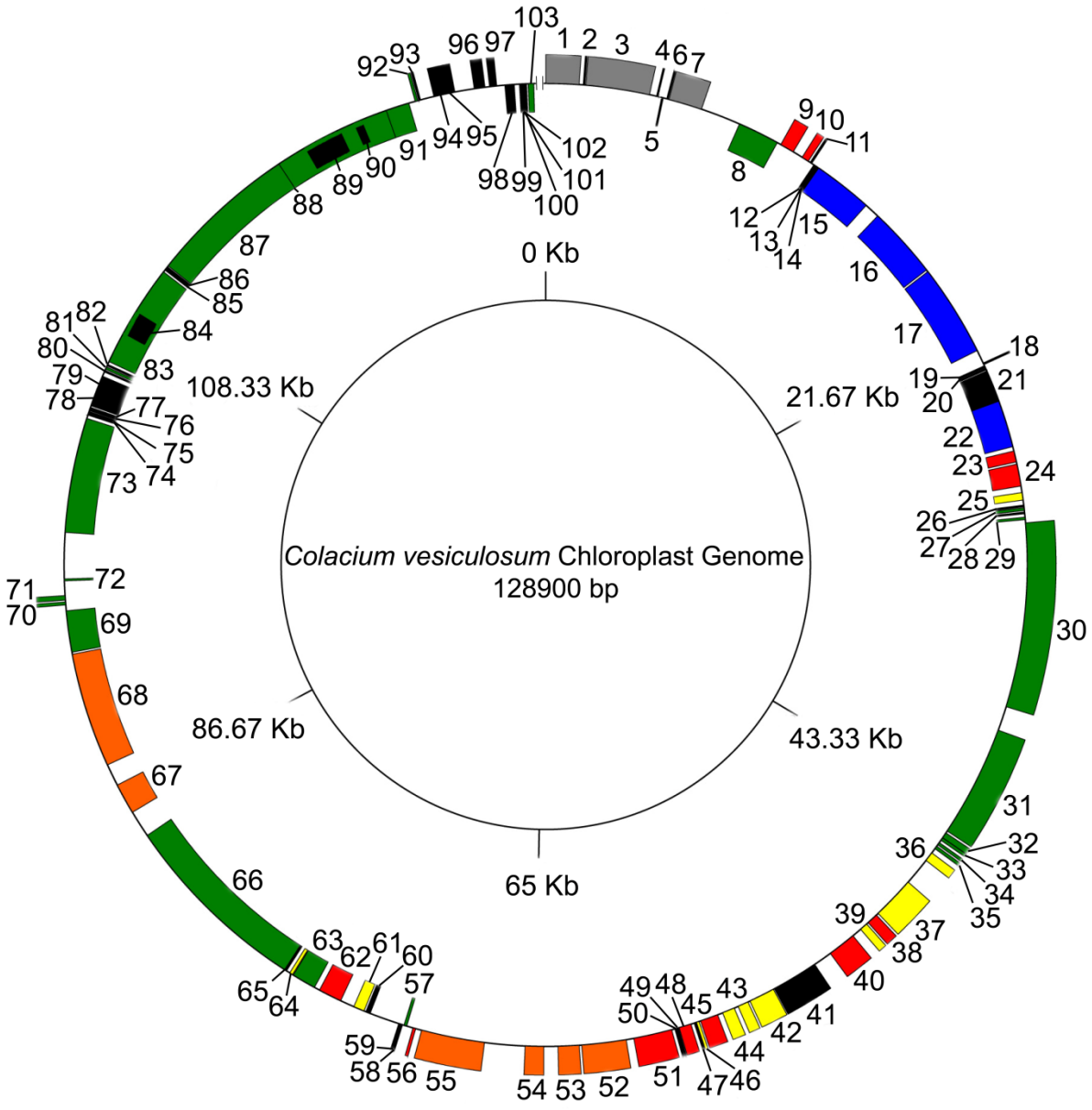

Figure 3.1: Chloroplast map of *Colacium vesiculosum*. Box colors represent genes of designated functional groups (Green: photosystems/photosynthesis; Yellow: large ribosomal proteins, rpl genes; Red: small ribosomal proteins, rps genes; Blue: rpo genes; Gray: ribosomal rRNAs; Orange: atp genes; Black: miscellaneous, ycf, orfs, tRNAs). Position of blocks on the inside or outside of the circle is representative of genes found on the positive and negative strands and proportional to their sequence length. 1: rrn16, 2: A(UGC), 3: rrn23, 4: tRNA(undetermined), 5: Gap feature, 6: L(UAA), 7: rrn16, 8: ccsA/chlI, 9: rps4, 10: rps11, 11: L(UAG), 12: R(ACG), 13: N(GUU),

Figure 3.1 Cont'd: 14: V(UAC), 15: rpoC2, 16: rpoC1, 17: rpoB, 18: H(GUG), 19: S(GCU), 20: Q(UUG), 21: ycf4, 22: tufA, 23: rps7, 24: rps12, 25: rpl20, 26: K(UUU), 27: petG, 28: D(GUC), 29: psbI, 30: psaA, 31: psaB, 32: psbE, 33: psbF, 34: psbL, 35: psbJ, 36: rpl23, 37: rpl2, 38: rps19, 39: rpl22, 40: rps3, 41: roaA, 42: rpl16, 43: rpl14, 44: rpl5, 45: rps8, 46: rpl36, 47: M(CAU), 48: rps14, 49: F(GAA), 50: C(GCA), 51: rps2, 52: atpI, 53: atpH, 54: atpF, 55: atpA, 56: rps18, 57: psaJ, 58: P(UGG), 59: S(UGA), 60: ycf9, 61: rpl12, 62: rps9, 63: psaC, 64: rpl32, 65: K(UUU), 66: rbcL, 67: atpE, 68: atpB, 69: petB, 70: psbN, 71: psbH, 72: psbT, 73: psbB, 74: G(UCC), 75: E(UUC), 76: W(CCA), 77: M(CAU), 78: orf348, 79: ycf65, 80: R(UCU), 81: psaM, 82: A(GAU), 83: psbA, 84: orf350, 85: L(CAA), 86: L(UAA), 87: psbC, 88: orf215, 89: orf550, 90: orf111, 91: psbD, 92: psaI, 93: Y(GUA), 94: orf171, 95: orf125, 96: orf153, 97: orf103, 98: orf133, 99: L(UAA), 100: T(UGU), 101: M(CAU), 102: G(GCC), 103: psbK.

total genome size of *C. vesiculosum* reported at 128,900 bp contains two copies of the 16S and one 23S (Figure 3.1). Based on read coverage it was determined that at least two copies of the ribosomal genes (16S and 23S) should be present. Fill-in PCR was completed in an attempt to link these ribosomal regions with no success (Table 1.1). The larger chloroplast genome of *S. acuminata*, like that of *C. vesiculosum*, also lacks complete circularization due to the presence of at least two copies of the ribosomal genes (16S, 23S) based on sequencing coverage (Figure 3.2). However, one full copy of the ribosomal region containing the genes 23S, A(UGC), I(GAU), and 16S is included, with the addition of a partial 23S on the opposite strand and with an inverted orientation (Figure 3.2) identified through fill-in PCR (Table 1.1). This makes determining the presence of an inverted repeat difficult especially in the absence of a small single copy region to distinguish them if they exist. More likely the *S. acuminata* ribosomal region is similar to those of *Euglena gracilis* and *Eutreptia viridis* with simple repeats and possible pseudogenes (Hallick et al. 1993). Regardless of the number of ribosomal gene copies, the chloroplast of *S. acuminata* is still much larger than that of *Euglena gracilis*, with much of the expansion due to additional intergenic regions (Figure

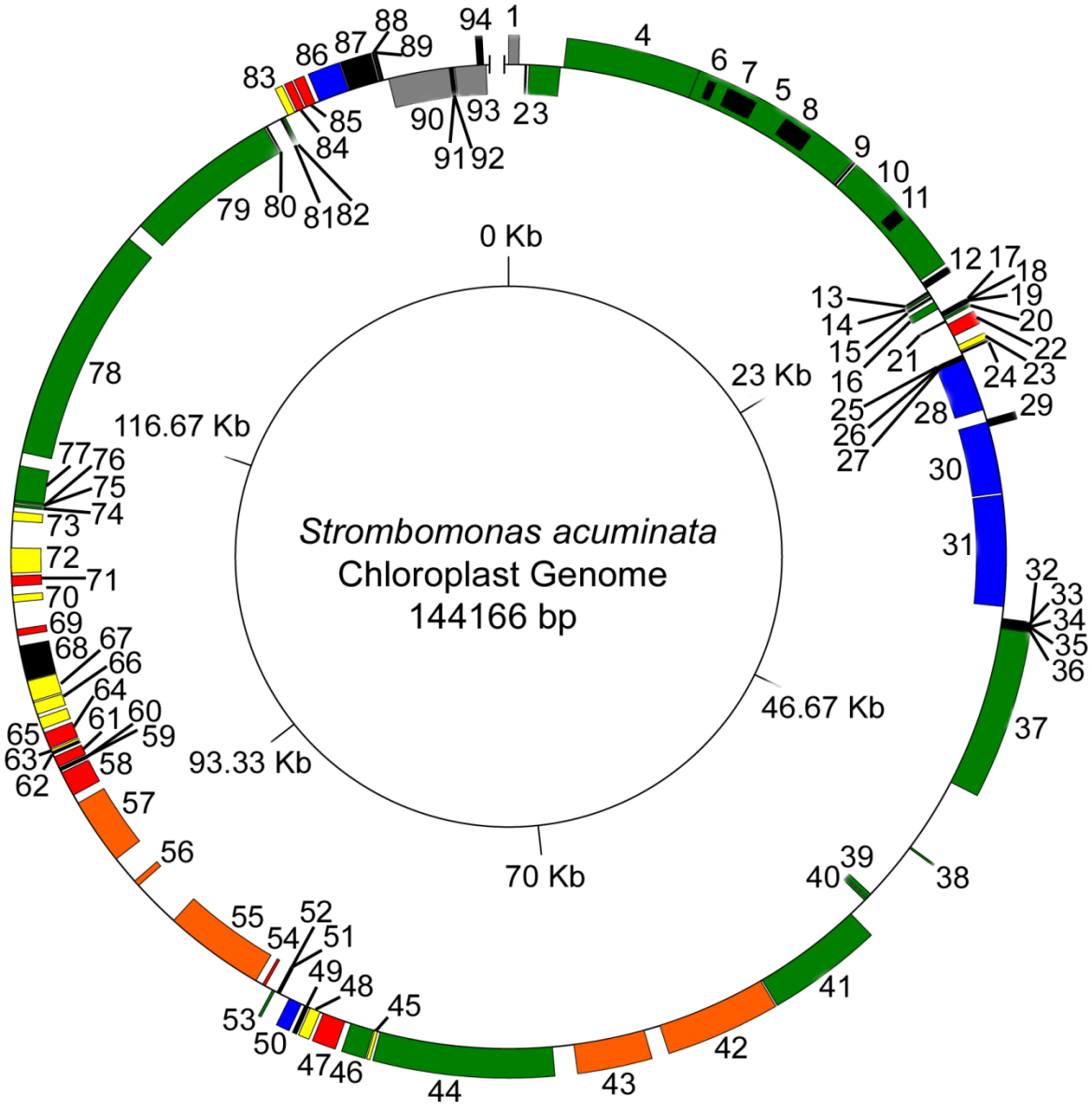

Figure 3.2: Circularized map of the *Strombomonas acuminata* chloroplast genome. Box colors represent genes of designated functional groups (Green: photosystems/ photosynthesis; Yellow: large ribosomal proteins, rpl genes; Red: small ribosomal proteins, rps genes; Blue: rpo genes; Gray: ribosomal rRNAs; Orange: atp genes; Black: miscellaneous, ycf, orfs, tRNAs). Position of blocks on the inside or outside of the circle is representative of genes found on the positive and negative strands and proportional to their sequence length. 1: rrn23, 2: L(CAA), 3: ccsA/chlI, 4: psbD, 5: psbC, 6: orf116, 7: orf463, 8: orf457, 9: L(UAA), 10: psbA, 11: orf190, 12: ycf65, 13: R(UCU), 14: psaM, 15: ycf12, 16: psbK, 17: T(UGU), 18: G(GCC), 19: M(CAU), 20: psaI, 21: Y(GUA), 22: rps4, 23: rps11, 24: L(UAG), 25: R(ACG), 26: N(GUU), 27: V(UAC), 28: rpoC2, 29: orf114, 30: rpoC1, 31: rpoB, 32: H(GUG), 33: M(CAU),

Figure 3.2 Cont'd: 34: W(CCA), 35: E(UUC), 36: G(UCC), 37: psbB, 38: psbT, 39: psbH, 40: psbN, 41: petB, 42: atpB, 43: atpE, 44: rbcL, 45: rpl32, 46: psaC, 47: rps9, 48: rpl12, 49: ycf9, 50: rpoA, 51: S(UGA), 52: P(UGG), 53: psaJ, 54: rps18, 55: atpA, 56: atpH, 57: atpI, 58: rps2, 59: C(GCA), 60: F(GAA), 61: rps14, 62: M(CAU), 63: rpl36, 64: rps8, 65: rpl5, 66: rpl14, 67: rpl16, 68: roaA, 69: rps3, 70: rpl22, 71: rps19, 72: rpl2, 73: rpl23, 74: psbJ, 75: psbL, 76: psbF, 77: psbE, 78: psbB, 79: psaA, 80: K(UUU), 81: D(GUC), 82: psbI, 83: rpl20, 84: rps12, 85: rps7, 86: tufA, 87: ycf4, 88: Q(UUG), 89: S(GCU), 90: rrn23, 91: A(UGC), 92: I(GAU), 93: rrn16, 94: orf100.

3.2). A variable number tandem repeat (VTNR) like that identified by Hallick et al. in the chloroplast of *Euglena gracilis* (1993) was not identified in either *C. vesiculosum* or *S. acuminata*. In addition, a 'start' was not established in the chloroplast maps (Figure 3.1, Figure 3.2) due to differences in arrangement and orientation as well as the lack of the VNTR (Hallick et al. 1993).

The A+T content of the *C. vesiculosum* chloroplast was measured at 73.8% while A+T content for *S. acuminata* was found to be 73.4%. This A+T percentage was measured over the entire sequence for each genome and therefore does not represent the gene or intragenic space specifically. However, it is known that coding regions are much lower in A+T content and this increases as the number of introns increases or the size of introns increases thereby raising the A+T percentage. Therefore, the drastic difference in gene density observed between *Eutreptia viridis* and the two sequenced chloroplast genomes here are due mainly to the increased number of introns and possibly intron size. Where *Eutreptia viridis* was found to contain at least 23 introns, *C. vesiculosum* has at least 92 and *S. acuminata* 103. The chloroplast genome of *Euglena gracilis* has 196 introns based on current GenBank data (Benson et al. 2011), however this includes identified group II and III introns as well as twintrons (introns within introns). This analysis has not yet been completed for *Eutreptia viridis*, *C. vesiculosum*, or *S.*

acuminata, and thus the intron number reported here is a minimum and likely to be higher.

Gene Content

Total gene content was fairly consistent between the two sequenced euglenoid chloroplasts (Figure 3.1, Figure 3.2). In *C. vesiculosum* a total of 91 genes were identified including 58 protein coding genes, 31 tRNAs, and the 2 rRNA (not including the 16S copy). This excludes 10 orfs that were annotated for *C. vesiculosum* (Figure 3.1). However, the total does include 3 tRNAs that are likely pseudogenes (2 Leucine tRNAs and 1 undetermined) and rpoC1 annotated as a pseudogene due to the presence of an internal stop codon.

The *S. acuminata* genome contains 86 total genes including 57 protein coding genes, 27 tRNAs, and 2 rRNA genes (not including the partial copy of 23S; Figure 3.2). None of the 6 orfs were included in this total gene count for *S. acuminata* and it again includes the genes annotated as pseudogenes including atpE, atpI, and rpl2 (Figure 3.2). Those annotated as pseudogenes were found to house internal stops or lacked a start codon, as was the case for atpI. A number of genes were also found to have alternative start codons. In *C. vesiculosum* 8 genes, rpl12, rps9, psbT, psbB, psbK, rps11, rpl23, and rps8, were annotated with alternative starts. Likewise 7 genes, rps8, ycf12, psbK, rps18, roaA, rpl2, and rpl23, had alternative start codons in *S. acuminata* (Table 3.1). The genes annotated with alternative starts for all four of the sequenced euglenoid chloroplasts can be seen in Table 3.1. No correlations between a gene and a given alternative start codon could be made, however psbK and rps9, each annotated with alternative start codons in *C. vesiculosum* and *S. acuminata* share an ATA start codon

Table 3.1: Alternative start codons identified in the euglenoid chloroplast genes of *C. vesiculosum*, *Euglena gracilis*, *Eutreptia viridis*, and *S. acuminata*. Total number of genes containing alternative starts is listed at the bottom. No ORFs or pseudogenes were included in this analysis. The absence of a gene from a given chloroplast genome is indicated by a dashed entry.

Gene Species	*Eutreptia viridis*	*S. acuminata*	*C. vesiculosum*	*Euglena gracilis*
atpE	ATT (Ile)			
roaA	------	TTG (Leu)		
petB	GTG (Val)			
psaI	ATT (Ile)			-----
psbB			ATC (Ile)	
psbD				undetermined
psbK		ATA (Ile)	ATA (Ile)	
psbN				undetermined
psbT			ATT (Ile)	
rpl2		ATT (Ile)		
rpl12			ATA (Ile)	
rpl23		ATA (Ile)	ATT (Ile)	
rps8			ATT (Ile)	
rps9		ATA (Ile)	ATA (Ile)	
rps11	ATT (Ile)		ATA (Ile)	
rps18		ATC (Ile)		
ycf12	------	ATA (Ile)	------	
Total Number	4	7	8	0

(Table 3.1). Overall, the chloroplast genome of *C. vesiculosum* has an average gene length of about 1803 bp. *S. acuminata* on the other hand has an average gene length of about 1323 bp, providing evidence for the expansion and the much larger chloroplast genome of *S. acuminata* (Figure 3.2).

Table 3.2 summarizes the changes in gene content when compared in all four of the sequenced euglenoid chloroplast genomes along with that of *Pyramimonas parkeae*, the hypothesized extant representative of the euglenoid chloroplast donor lineage (Turmel et al. 2009). *Eutreptia viridis*, located at the base of the photosynthetic lineage, lacks the psaM, roaA, rpl32, L(Leu)-CAA, and G(Gly)-GCC genes found in the other

euglenoids. roaA, an intron associated gene more recently annotated in *Euglena gracilis* (Jenkins et al. 1995) has been identified in *C. vesiculosum* and *S. acuminata*, but was not found in the basal *Eutreptia viridis*. *Eutreptia viridis* does however contain the rrn5 (5S) gene not present in the donor lineage represented by *Pyramimonas parkeae* (Turmel et al. 2009). rrn5 has been found in all euglenoid chloroplast genomes except for *S. acuminata*. There is low sequence homology seen between rrn5 in even very closely related species and therefore it may not have been recognized on the basis of the annotation procedures applied here. Although *S. acuminata* has the largest euglenoid chloroplast genome, it lacks atpF and petG genes (Figure 3.2, Table 3.2). The psaI gene was found in all euglenoid chloroplast genomes except *Euglena gracilis* (Hallick et al. 1993). Most notably, the recently identified rpoA gene (Sheveleva et al. 2002) that eluded annotation in the *Euglena gracilis* chloroplast genome long after it was first published (Hallick et al. 1993) has not been identified in the chloroplast genome of *C. vesiculosum* (Figure 3.1, Table 3.2).

Open reading frames (orfs) were identified in both the *C. vesiculosum* and *S. acuminata* chloroplast genomes. Orfs were annotated when open reading frames greater than or equal to 100 amino acids (300 nucleotides) occurred but lacked sufficient homology to known genes as described in the annotation protocols. Ten orfs were identified in *C. vesiculosum* (orf171, orf125, orf153, orf103, orf348, orf350, orf215, orf550, orf111, orf133) and six orfs were identified in *S. acuminata* (orf116, orf463, orf457, orf190, orf114, orf100). Orfs were not consistent in arrangement or size between *C. vesiculosum* and *S. acuminata*, or when compared to the chloroplast genomes of *Euglena gracilis* and *Eutreptia viridis* (Hallick et al. 1993).

Table 3.2: Differences in gene content between the chloroplast genomes of *C. vesiculosum*, *S. acuminata* and the previously sequenced *Euglena gracilis*, *Eutreptia viridis*, and *Pyramimonas parkeae*.

	rrn5 (5S)	rpl32	rpoA	atpF	petG	psaI	psaM	roaA	L(Leu) - CAA	G(Gly) - GCC
Euglena gracilis	+	+	+	+	+	-	+	+	+	+
S. acuminata	-	+	+	-	-	+	+	+	+	+
C. vesiculosum	+	+	-	+	+	+	+	+	+	+
Eutreptia viridis	+	-	+	+	+	+	-	-	-	-
Pyramimonas parkeae	-	+	+	+	+	+	+	-	+	+

Gene Arrangement

The extent of gene rearrangement was assessed with the use of Mauve (Figure 3.3; http://gel.ahabs. wisc.edu/mauve; Darling et al. 2004). This provided visualization and identification of conserved blocks or gene clusters and allowed us to trace them from genome to genome (Figure 3.3). Arrangements are based on the chloroplast genome of *Eutreptia viridis* which was used to root the alignment because of its basal position in phylogenetic assessments (Linton et al. 2010, Triemer et al. 2006). Genes consistently found in the labeled blocks (gene clusters, A-O) for each chloroplast genome can be found in Table 3.3. Figure 3.3 shows the dramatic differences in gene order among the four chloroplast genomes. The gene clusters are separated into 15 blocks labeled from A to O. Despite the rearrangements of the labeled blocks among the genomes there are patterns that can be seen among taxa. One collinear arrangement spans from blocks J-B in *Euglena gracilis* and *C. vesiculosum*. This same arrangement is also present in *S. acuminata*, but is inverted (Figure 3.3). Another example contains Blocks F, G, E and H seen spanning the region from 2.5-30Kb in *Euglena gracilis* and 0-26Kb in *S. acuminata*, but interrupted by the ribosomal genes of block O in *C. vesiculosum* (Figure 3.3). The genes atpF, petG, psaI, psaM, psbI, psbK, roaA, rpl23, rpl32, rpl36, rpoA, rps9, rps14, rrn5 (5S), ycf12, and ycf65 do not have consistent orientation within one of the gene blocks and/or were absent from one of the four analyzed genomes (Table 3.3). This analysis makes it clear that although the genes in the four sequenced euglenoid chloroplasts tend to be found in conserved gene clusters, the arrangement of those gene clusters is much more consistent with that found in *Euglena gracilis* compared to the gene arrangement of *Eutreptia viridis* (Figure 3.3). In addition, the Mauve analysis

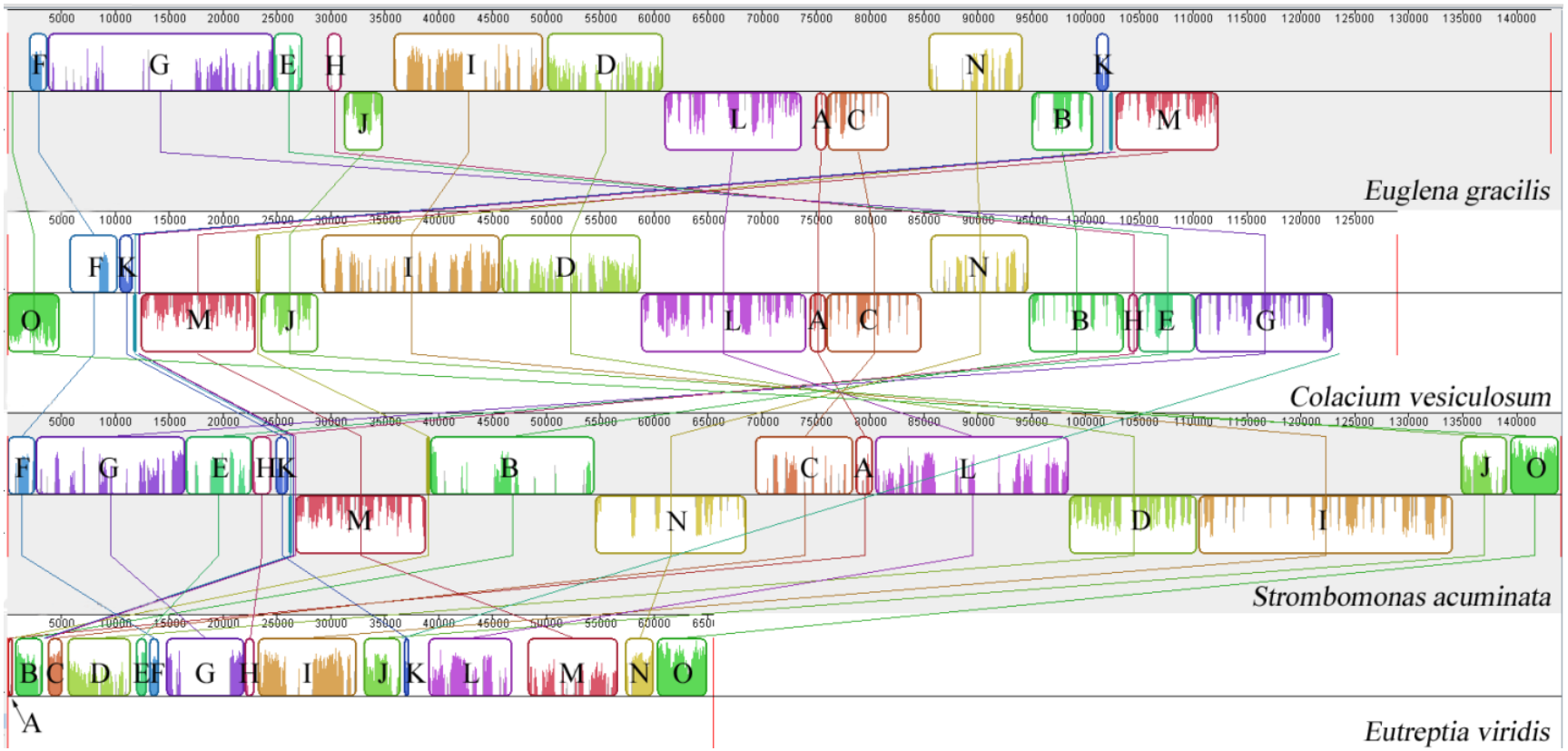

Figure 3.3: Mauve homology and synteny analysis of the four sequenced euglenoid chloroplast genomes. Like blocks are recognized by letters A-O throughout the four chloroplast genomes. Arrangement of blocks in regard to strand is relative only to the genome of *Eutreptia viridis*.

Table 3.3: Gene clusters resulting from Mauve analysis of the four sequenced euglenoid chloroplast genomes. Gene clusters (blocks) were identified with letters for more clarity and the genes contained within them are listed. The genes atpF, petG, psaI, psaM, psbI, psbK, roaA, rpl23, rpl32, rpl36, rpoA, rps9, rps14, rrn5, ycf12, and ycf65 do not have consistent orientation with one of the below mentioned gene blocks or are absent from one of the four analyzed genomes.

Block	Conserved genes present
A	psaC
B	psbB, psbT, psbH, psbN
C	rbcL
D	rpl2, rps19, rpl22, rps3, rpl16, rpl14, rpl5, rps8, rpl36, rps14
E	psbA
F	ccsA/chlI
G	psbD, psbC
H	No consistent gene content between the 4 genomes
I	psaA, psaB, psbE, psbF, psbL, psbJ
J	rpl20, rps12, rps7, tufA, ycf4
K	rps4, rps11
L	rps9, rpl12, ycf9, psaJ, rps18, atpA, atpH, atpI, rps2
M	rpoB, rpoC1, rpoC2
N	atpE, atpB, petB
O	23S, 5S, 16S

shows the gene density, homology, and relative size of the four euglenoid chloroplast genomes (Figure 3.3). The relative compactness of lines within homologous blocks

represents the gene density while the height of the lines reflects the level of homology present among the genes (Figure 3.3). Block G represents a smooth transition in increasing size from *Eutreptia viridis* to *Euglena gracilis*. Block B, however, is more representative of the majority of blocks in that they tend to be larger in the genomes of C. *vesiculosum* and *S. acuminata* and smaller in *Euglena gracilis* (Figure 3.3). Meanwhile blocks in *Eutreptia viridis* are consistently compact and smallest in size. These differences in density have contributed greatly to the larger genome sizes of *C. vesiculosum* and *S. acuminata*, and are easily seen in the amount of white space within each block representing intergenic space (Figure 3.3). Overall the Mauve analysis has emphasized the lack of a clear transition from basal to crown euglenoid species in chloroplast evolution.

The sequencing of the *C. vesiculosum* and *S. acuminata* chloroplast genomes has added greatly to our understanding of the evolutionary state of the euglenoid chloroplast. These species, which are consistently found intermediate to the basal *Eutreptia viridis* and crown species *Euglena gracilis*, have provided an evolutionary framework characterized by the lack of clear evolutionary transitions in gene content, arrangement, and expansion (Table 3.2, Figure 3.3). Table 3.2 shows the similar gene content of *Euglena gracilis*, *C. vesiculosum*, and *S. acuminata* with discrepancies likely based on limitations in current identification due to a lack of sequence conservation. However, the absence of key genes in *C. vesiculosum* (rpoA) and *S. acuminata* (rrn5 (5S), atpE, petG) as well as the presence of genes found throughout the euglenoid chloroplast genomes except for the basal *Eutreptia viridis* do not provide a clear transition of gene gain and/or loss (Figure 3.1, Figure 3.2, Table 3.2). However, the

most striking divergence from a clear evolutionary transition is evident in the expansion of intergenic regions in *C. vesiculosum* and *S. acuminata* which is much greater than that seen in *Euglena gracilis* (Figure 3.1, Figure 3.2, Figure 3.3). This is surprising especially with the chloroplast of *Euglena gracilis* previously considered the most intron diverse (twintrons) and intron rich chloroplast genome sequenced to date.

In summary, this study has provided the chloroplast sequences of the photosynthetic euglenoids *C. vesiculosum* and *S. acuminata*. These two sequences in combination with the previously sequenced chloroplast genomes of *Eutreptia viridis* and *Euglena gracilis* have shown that there are no clear evolutionary transitions from basal to crown species within the lineage. Genes are found consistently in homologous gene clusters, however the arrangement of these clusters is highly variable. In addition, differences in both the number of introns, as well as their size, have resulted in increase in genome size not observed in any other green algal chloroplast lineage.

Chapter 4

Phylogenetic Assessment of the Euglenoid Chloroplast

Abstract

Comparisons of the *Colacium vesiculosum*, *Euglena gracilis*, *Eutreptia viridis*, and *Strombomonas acuminata* chloroplast genomes have been made through the use of synteny mapping and homology searches. However, the sequence data from these chloroplast genomes has not been phylogenetically assessed. Current understanding of phylogenetics within the photosynthetic euglenoids is largely based on the 16S and 23S rRNA genes. Therefore, in combination with additional sequence data from the photosynthetic euglenoid *Discoplastis spathirhyncha* and a subset of green algal representatives, both Maximum Likelihood and Bayesian phylogenetic analyses were carried out by assessing two datasets, one of 22 tRNA genes, and another totaling 55 (the 22 tRNAs, 16S, and 33 protein coding) chloroplast genes. tRNA trees were used to establish the basal position of *Eutreptia* among the euglenoids and placed the green algae *Pyramimonas* and *Ostreococcus* at the base of the euglenoid clade, but could not be used to resolve relationships among the remaining photosynthetic euglenoids. The 55 gene dataset resulted in greater support within the euglenoids, placing *Eutreptia* at the base of the euglenoid lineage followed by the divergence of *Discoplastis* prior to the two *Euglena* species, *Colacium*, and *Strombomonas*. Both topology and support within the trees were consistent with previous studies of the 16S or 23S genes, as well as nuclear encoded genes. The phylogenies generated from nuclear encoded and chloroplast encoded genes are congruent suggesting that both genomes have followed similar evolutionary pathways. The basal position of *Pyramimonas* relative to the 6

euglenoids used in the chloroplast phylogenies strongly supports this taxon as an extant representative of the euglenoid green algal chloroplast donor and implies that acquisition of the chloroplast was the result of a single endosymbiotic event.

Introduction

It is generally accepted that the euglenoid chloroplast was acquired through the secondary endosymbiosis of a green alga (Gibbs 1979, 1981). At the time however, the donor lineage of the euglenoid chloroplast within the green algae was unknown. Sequencing of the *Euglena gracilis* chloroplast genome in 1993 provided the first glimpse into the gene content and arrangement of a euglenoid chloroplast genome (Hallick et al. 1993). The chloroplast genome of *Euglena gracilis* challenged characters thought to be common among "green" chloroplasts, with the lack of an inverted repeat and the largest amount of intergenic space recorded (Hallick et al. 1993). Turmel et al. (2009) compared the euglenoid chloroplast with several green algal species and identified a putative extant representative of the euglenoid chloroplast donor lineage, *Pyramimionas parkeae*. However, since only a single euglenoid chloroplast genome was available, the authors could not determine if acquisition of the plastid was the result of a single or multiple endosymbiotic events.

The recent addition of several euglenoid chloroplasts (*Euglena viridis* (Bennett et al. 2011), *Eutreptia viridis*, *Colacium vesiculosum*, and *Strombomonas acuminata*) to the dataset provided further evidence supporting *Pyramimionas parkeae* as the extant representative of the chloroplast donor lineage (Turmel et al. 2009). The close

homology of each of the sequenced euglenoid chloroplast genomes with the chloroplast of the prasinophyte *Pyramimonas parkeae*, was consistently greater than with any other green algal chloroplast, providing solid evidence for the hypothesis that the euglenoid chloroplast evolved from a single endosymbiotic event. Therefore, an ancestral non-photosynthetic phagotrophic euglenoid (of which there are several extant closely related species; Busse et al 2003, Busse and Preisfeld 2002, 2003, Leander et al. 2001, Linton et al. 1999, 2000, Marin et al. 2003, Preisfeld et al. 2000, 2001, Von der Heyden et al. 2004) engulfed a small *Pyramimonas*-like prasinophyte and retained its chloroplasts adopting a photosynthetic lifestyle.

However, the phylogenetic relationships among the chloroplast encoded genes of green algae and multiple euglenoids has not been assessed. Two chloroplast genes have been sequenced en mass including the 16S and 23S (rrn16, rrn23, Milanowski et al. 2001, Kim and Shin 2008). The analysis of these sequences from roughly 200 photosynthetic species has provided an outline of the phylogenetic relationships present. The 16S and 23S genes, however, are fairly conserved. This data has also been combined with the nuclear encoded SSU and LSU (Brosnan et al. 2003, Ciuglea et al. 2008, Kim et al. 2010, Linton et al. 2000, Linton et al. 2010, Marin et al. 2003, Milanowski et al. 2006, Preisfeld et al. 2000, Triemer et al. 2006) and other protein coding genes (5 gene paper), and together this analysis has provided a solid and consistent picture of phylogeny among the photosynthetic euglenoids. What is not known is if evolutionary relationships assessed from largely nuclear data concur with those from chloroplast encoded genes. This study analyzes the phylogenetic relationships and evolutionary trends present among the sequenced euglenoid

chloroplast genomes in combination with those of green algae, thereby assessing the history of these unique chloroplasts obtained from a green algal donor.

Results

Phylogenies based on concatenations of chloroplast genes from five euglenoid chloroplast genomes (*Colacium vesiculosum*, *Euglena gracilis*, *Eutreptia viridis*, and *Strombomonas acuminata*) were constructed using Maximum Likelihood (ML) and Bayesian analyses. In addition, sequence data from the chloroplasts of *Discoplastis spathirhyncha* and *Euglena viridis* (unpublished) were included along with data from a subset of green algal species.

In these analyses we tested the phylogenetic utility of two sets of chloroplast genes, a dataset of 22 tRNA genes and a dataset of 55 ribosomal and protein coding genes. The tRNA dataset produced phylogenies which placed all of the euglenoids into a single, well-supported clade along with two green algal species, *Ostreococcus tauri* and *Pyramimonas parkeae,* which diverged at the base of the euglenoid lineage (clade A, Figure 4.1, bt 97, Figure 4.2, pp=1.00). The remaining green algal taxa formed a sister clade (B) to the euglenoid containing clade. However, the relationships among the green algal taxa in clade B were unresolved. Similarly the relationships among the *Colacium*, *Strombomonas* and *Discoplastis* species in clade A were not resolved in the ML tree (Figure 4.1).

When the dataset was increased to include a total of 55 cp genes (22 tRNAs, 16SrRNA, and 32 protein coding genes, 25858 informative sites; Table 4.1) the resolution of the

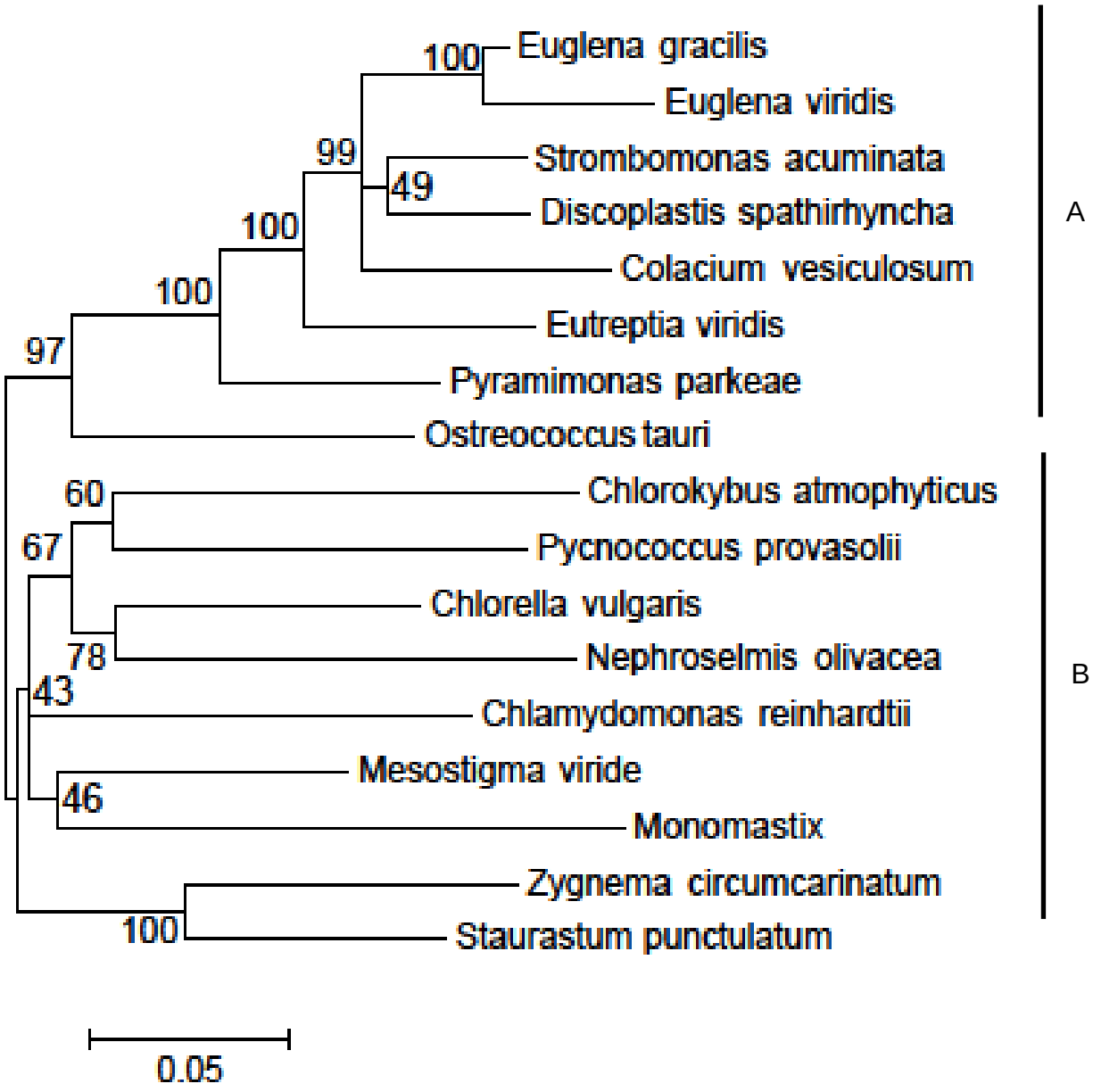

Figure 4.1: Maximum Likelihood analysis of 22 chloroplast tRNA genes from a total of 17 species of green algae and euglenoids. Tree constructed from 1856 informative sites and midpoint rooted. Clade A represents the euglenoid species along with the green algae *Pyramimonas parkeae* and *Ostreococcus tauri*. Clade B contains the remaining green algal species.

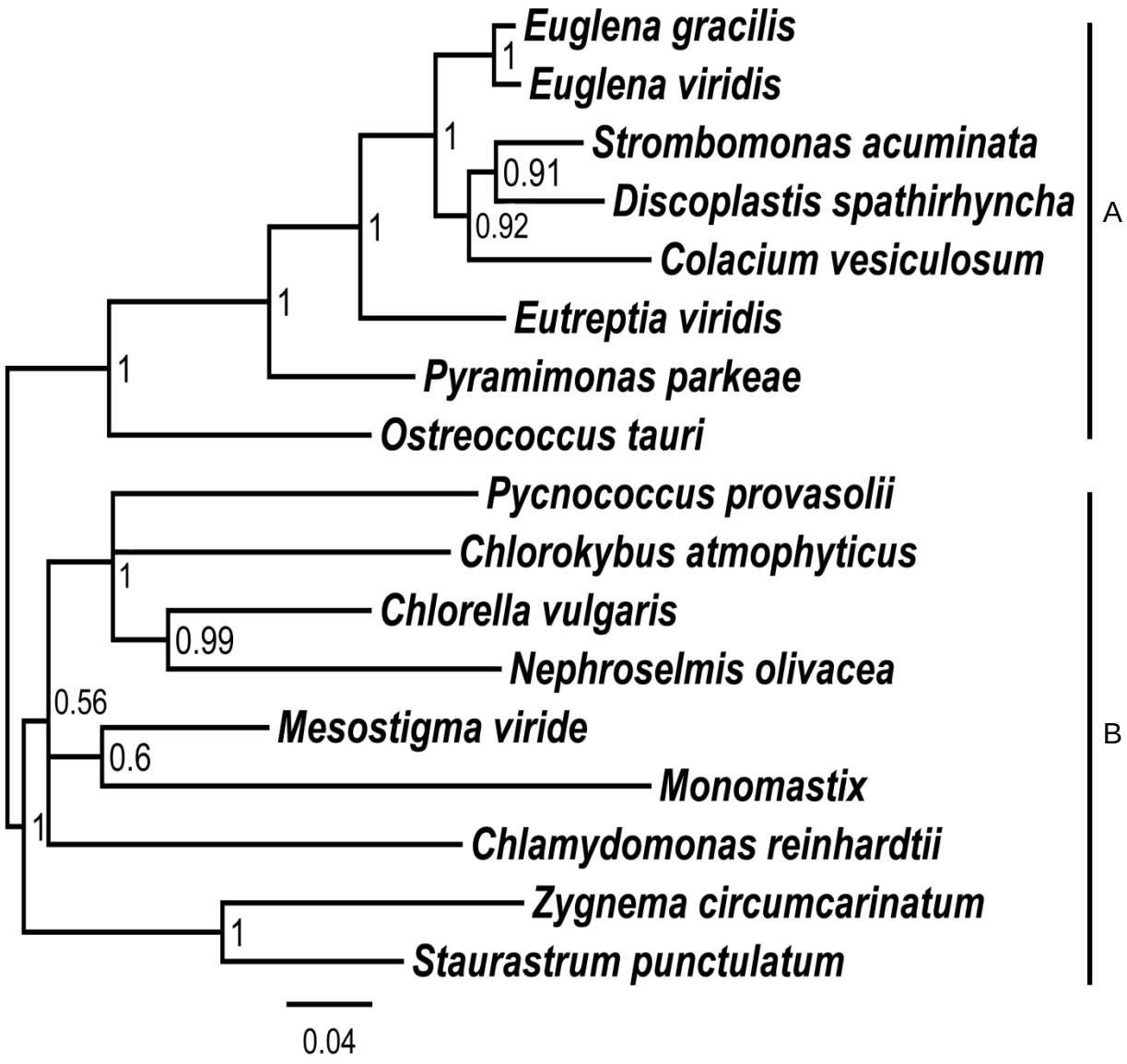

Figure 4.2: Bayesian analysis of 22 chloroplast tRNA genes from a total of 17 species of green algae and euglenoids. Tree constructed from 1856 informative sites and midpoint rooted. Clade A represents the euglenoid species along with the green algae *Pyramimonas parkeae* and *Ostreococcus tauri*. Clade B contains the remaining green algal species.

entire tree was greatly improved (Figure 4.3, Figure 4.4). The relationships among the green algae were fully resolved in the ML analysis and well resolved in the Bayesain analysis with the exception of one node (indicated by *) representing the position of *Nephroselmis* (Figure 4.4). In the 55 gene analyses *Ostreococcus tauri* was positioned among the green algae such that it joined the Chlorophyte lineage (*Nephroselmis*, *Chlorella*, *Chlamydomonas*, *Monomastix*, *Pycnococcus*, and *Ostreococcus*) while the Streptophyte (*Mesostigma*, *Chlorokybus*, *Staurastrum*, and *Zygnema*) lineages formed a sister clade (Figure 4.3, Figure 4.4).

Table 4.1: Chloroplast genes included in the 55 gene dataset for both Maximum Likelihood and Bayesian analysis. Genes are listed by functional category.

Functional Category	**Genes**
Ribosomal rRNA	16s
ATPases	atpA, atpB, atpE, atpF, atpH, atpI
Cytochrome subunits	petB, petG
Photosystem I	psaA, psaB, psaC, psaJ, psaM
Photosystem II	psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbK, psbL, psbT
Large ribosomal proteins	rpl14, rpl16, rpl23, rpl36, rpl5
Small ribosomal proteins	rps2, rps3, rps4
Transfer RNAs (tRNAs)	A(TGC), C(GCA), D(GTC), E(TTC), F(GAA), G(TCC), H(GTG), I(GAT), K(TTT), L(TAG), M(CAT), N(GTT), P(TGG), O(TTG), R(ACG), R(TCT), S(GCT), S(TGA), T(TGT), V(TAC), W(CCA), Y(GTA)

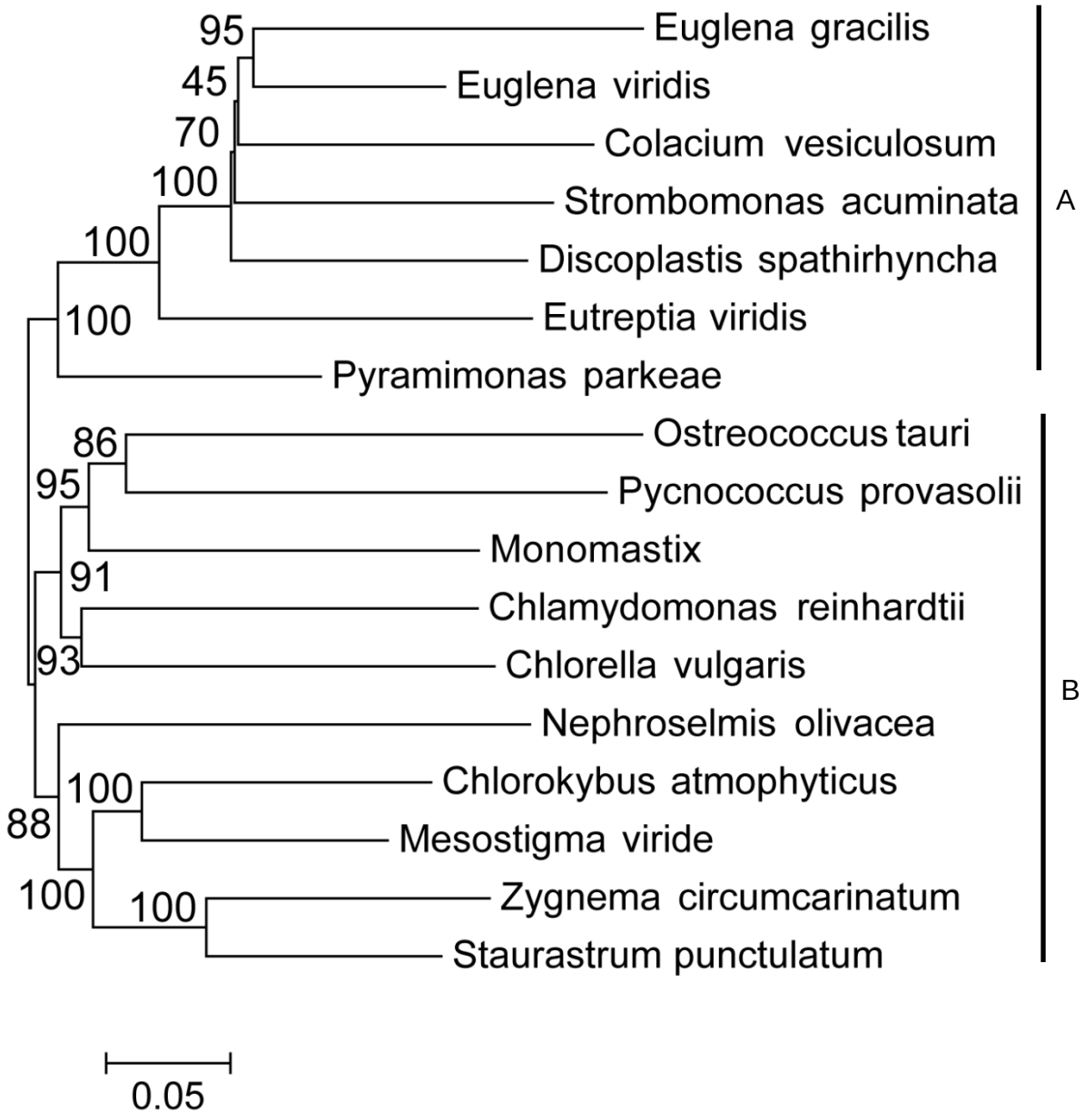

Figure 4.3: Maximum Likelihood analysis of 55 chloroplast genes including ribosomal and protein coding (22 tRNAs, 16SrRNA, and 32 protein coding genes). Tree constructed from 25858 informative sites and midpoint rooted. Clade A contains all euglenoid species with *Pyramimonas parkeae* positioned basally.

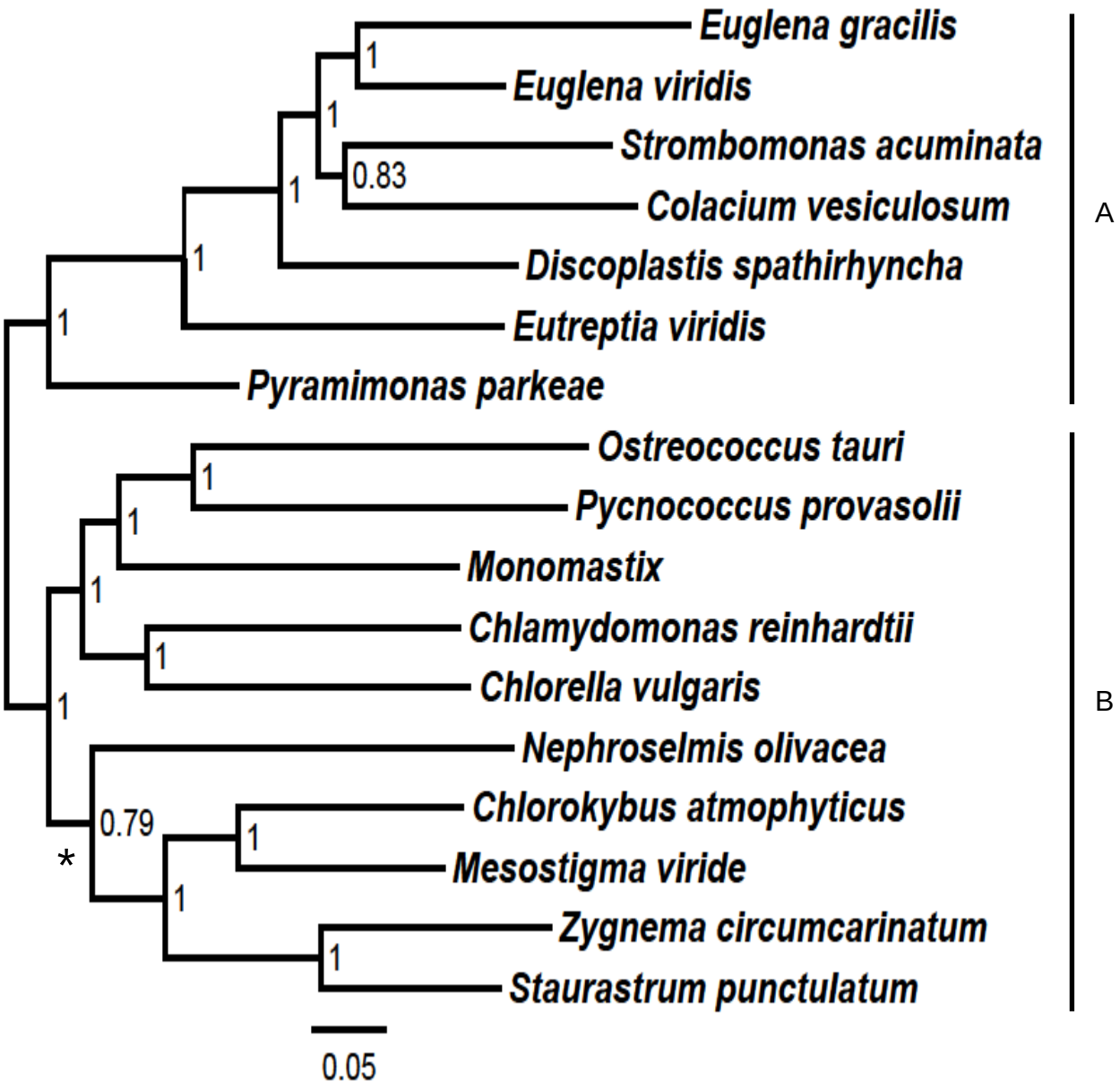

Figure 4.4: Bayesian analysis of 55 chloroplast genes including ribosomal and protein coding (22 tRNAs, 16SrRNA, and 32 protein coding genes). Tree contructed from 25858 informative sites and midpoint rooted. Clade A contains all euglenoid species with *Pyramimonas parkeae* positioned basally.

Pyramimonas parkeae was the single green algal species firmly positioned at the base of the euglenoid lineage in clade A (Figure 4.3, Figure 4.4). The position of *Eutreptia viridis* was also supported basal to the rest of the photosynthetic euglenoids (Figure 4.3, Figure 4.4). However, the two datasets differed in the positioning of *Discoplastis spathirhyncha*, which was well supported and positioned basal to *Colacium*, *Strombomonas* and the two *Euglena* species on in the 55 gene dataset (Figure 4.3, Figure 4.4). Both ML and Bayesian analysis failed to resolve the relationship between *Strombomonas* and *Colacium* which collapsed into a trichotomy with the two *Euglena* species in the Bayesian tree (Figure 4.4) and a polytomy in the ML tree which also included *Discoplastis* (Figure 4.3). Although the topologies varied somewhat between the trees, the position of *Pyramimonas parkeae* at the base of the euglenoid lineage diverging prior to *Eutreptia viridis* was consistent and well supported in all phylogenies.

Discussion

Phylogenetic analysis of euglenoid chloroplast sequence has focused largely on the 16S and 23S (rrn16 and rrn23) ribosomal genes. Our purpose in this study was to assess the phylogenetic utility of other chloroplast genes and compare the phylogenies with those based on chloroplast ribosomal genes and nuclear genes. tRNAs were first tested since they were easily identified in sequence data based on high levels of homology and represented a second source of non-protein coding genes. Generating phylogenies based upon tRNAs proved to be a very quick and useful method for generating trees in the early stages of sequence assembly, allowing preliminary

chloroplast phylogenetic analysis. This is due in part to the arrangement of many tRNAs in gene clusters. For example, in *Euglena gracilis*, 10 tRNAs are encoded within a single 2380 bp region of the chloroplast genome (100430..102810; Hallick et al. 1993). When tRNA phylogenies were compared with previous studies based on the chloroplast ribosomal genes alone, which contained many more taxa and informative sites, the ribosomal trees had greater support and resolution of the internal nodes (Kim and Shin 2008, Milanowski et al., 2001) but did not conflict with the overall topology of the tRNA trees. Therefore, chloroplast tRNAs do provide a quick and easy source of phylogenetic data and with increased taxon sampling (e.g. *Trachelomonas*, *Lepocinclis*, and *Phacus* are not represented) should be able to infer phylogeneies as well as ribosomal genes.

As expected, the 55 gene data set did improve the phylogenies. The Bayesian and ML analysis of the 55 chloroplast genes resulted in increased support, firmly positioning *Discoplastis spathirhyncha* basal to the two *Euglena* species, *Strombomonas*, and *Colacium* (Figure 4.4). The position of the genus *Discoplastis* basal to the remaining Euglenales agrees with the previous phylogenies based on nuclear ribosomal genes (Triemer et al. 2006). Overall the phylogenetic trees resulting from the analysis of the 55 chloroplast genes showed a similar topology to previous analyses of the 16S or 23S genes (Kim and Shin 2008, Milanowski et al., 2006). The 55 gene Bayesian and ML analysis failed to resolve the positions of *Strombomonas* and *Colacium* which was not surprising due to the small number of photosynthetic genera included in the study. The addition of chloroplast sequence data from a representative of the genus *Trachelomonas* may have resolved the relationship between *Strombomonas* and *Colacium*. These three genera together consistently form a clade with *Colacium* sister to

Strombomonas and *Trachelomonas*. However, the support for the position of *Colacium* is very weak (pp 0.77, bt 46; Kim et al. 2010, Triemer et al. 2006). The well supported basal position of *Eutreptia viridis* established in all four trees corresponds with previous phylogenetic analyses of the photosynthetic euglenoids (Leander et al. 2001, Marin et al. 2003, Preisfeld et al. 2001).

The analysis of the 22 chloroplast tRNAs, as well as the combined data set of 55 genes show phylogenetic relationships similar to those based upon chloroplast 16S and 23S genes, as well as to those based on nuclear encoded genes. Studies of the chloroplast 16S (Milanowski et al. 2001) and 23S (Kim et al. 2008) agree with the phylogenetic assessment of the nuclear LSU and SSU genes (Brosnan et al. 2003, Ciuglea et al. 2008, Linton et al. 2000, Marin et al. 2003, Preisfeld et al. 2000, Triemer et al. 2006). This consistency in support and position indicate that the chloroplast and nuclear genome phylogenies are congruent suggesting these genomes have followed similar evolutionary pathways. One characteristic to consider if this is the case is the number and increasing prevalence of intergenic regions within the euglenoid chloroplast (Hallick et al. 1993). The nuclear genome of photosynthetic euglenoids has been estimated to exceed 4 Gb in size (1 picogram (pg) of DNA is estimated at 978Mb, 4.4 pg DNA/nucleus in *Euglena gracilis* X 978Mb = 4303.2 Mb; Buetow 1978) and is hypothesized to contain large amounts of intergenic and repeated sequence. Could a similar evolutionary trend be driving the large amount of intergenic sequence in euglenoid chloroplasts (opposed to green algae with few to no annotated introns in the chloroplast)? Furthermore, these similarities in phylogenies from chloroplast and

nuclear genes provide additional support for the hypothesis that photosynthetic euglenoids acquired their chloroplast from a single endosymbiotic event.

This study has provided further evidence that the putative green algal chloroplast donor lineage was a scaly green algal flagellate belonging to the Prasinophyceae, with *Pyramimonas parkeae* the closest extant species based upon available chloroplast sequence data. In addition, the phylogenetic trees based upon the 55 chloroplast genes were found to be nearly identical to phylogenies based on nuclear encoded genes (Brosnan et al. 2003, Ciuglea et al. 2008, Linton et al. 2000, Marin et al. 2003, Preisfeld et al. 2000, Triemer et al. 2006). The close phylogenetic relationship of euglenoid chloroplast genomes to each other, more than to any other available green algal chloroplast genome, also indicates that a single endosymbiotic event led to the establishment of the euglenoid chloroplast.

REFERENCES

REFERENCES

Adl SM, Simpson AGB, Farmer MA, Andersen RA, Anderson OR, Barta JR, Bowser SS, Brugerolle G, Fensome RA, Fredericq S, James RY, Karpov S, Kugrens P, Krug J, Lane CE, Lewis LA, Lodge J, Lynn DH, Mann DG, McCourt RM, Mendoza L, Moestrup O, Mozley-Stadridge SE, Nerad RA, Shearer CA, Smirnov AV, Spiegel FW, Taylor MFJR (2005) The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J Eukaryot Microbiol 52(5):399-451

Bennett MS, Triemer RE (2011) The chloroplast genome of *Euglena viridis*. J Phycol 47:S1-S98

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2011) GenBank. Nucleic Acids Res 39:D32-37

Brosnan S, Shin W, Kjer KM, Triemer RE (2003) Phylogeny of the photosynthetic euglenophytes inferred from the nuclear SSU and partial LSU rDNA. Int J Syst Evol Microbiol 53(4):1175-1186

Buetow DE (1978) Nuclei and Chromatin from Euglena gracilis. In Handbook of Phycological Methods: Culture Methods and Growth Measurements (eds. Hellebust JA, Craigie JS) Cambridge University Press, 448 pp

Burger G, Lavrov DV, Forget L, Lang BF (2007) Sequencing complete mitochondrial and plastid genomes. Nature Protocols 2(3):603-614

Busse I, Patterson DJ, Preisfeld A (2003) Phylogeny of phagotrophic euglenids ("Euglenozoa"): A molecular approach based on culture material and environmental samples. J Phycol 39:828-836

Busse I, Preisfeld A (2002) Phylogenetic position of Rhynchopus sp and Diplonema ambulatory as indicated by analyses of euglenozoan small subunit ribosomal DNA. Gene 284:83-91

Busse I, Preisfeld A (2003) Systematics of primary osmotrophic euglenids: a molecular approach to the phylogeny of Distigma and Astasia (Euglenozoa). Int J Syst Evol Microbiol 53:617-624

Cavalier-Smith T (1998) A revised six-kingdom system of life. Biol Rev 73(3):203-266

Christopher DA, Hallick RB (1989) *Euglena gracilis* chloroplast ribosomal protein operon: A new chloroplast gene for ribosomal protein L5 and description of a novel organelle intron category designated group III. Nucleic Acids Res 17(19):7591-7608

Christopher DA, Hallick RB (1990) Complex RNA maturation pathway for a chloroplast ribosomal protein operon with an internal tRNA cistron. Plant Cell 2(7):659-671

Ciugulea I, Nudelman MA, Brosnan S, Triemer RE (2008) Phylogeny of the euglenoid loricate genera Trachelomonas and Strombomonas (Euglenophyta) inferred from nuclear SSU and LSU rDNA. J Phycol 44(2):406-418

Ciugulea I, Triemer RE (2010) A Color Atlas of Photosynthetic Euglenoids. Michigan State University Press, East Lansing, MI, USA, 204 pp.

Copertino DW, Hallick RB (1991) Group II twintron: An intron within an intron in a chloroplast cytochrome b-559 gene. EMBO (Eur Mol Biol Organ) 10(2):433-442

Coperino DW, Hallick RB (1993) Group II and group III introns of twintrons: Potential relationships with nuclear pre-mRNA introns. Trends in Biochemical Sciences 18(12):467-471

Copertino DW, Hall ET, Van Hook FW, Jenkins KP, Hallick RB (1994) A group III twintron encoding a maturase-like gene excises through lariat intermediates. Nucleic Acids Res 22 (6):1029-1036

Darling ACE, Mau B, Blattner FR, Perna NT (2004) Mauve: Multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14:1394-1403

Doetsch NA, Thompson MD, Hallick RB (1998) A maturase-ecoding group III twintron is conserved in deeply rooted euglenoid species: Are group III introns the chicken or the egg? Mol Biol Evol 15(1):76-86

Gibbs SP (1978) The chloroplasts of *Euglena* may have evolved from symbiotic green algae. C J Bot 56(22):2883-2889

Gibbs SP (1981) The chloroplasts of some algal groups may have evolved from endosymbiotic eukaryotic algae. Ann NY Acad Sci 361:193-208

Gockel G, Hachtel W (2000) Complete gene map of the plastid genome of the nonphotosynthetic euglenoid flagellate Astasia longa. Protist 151(4):347-351

Hallick RB, Hong L, Drager RG, Favreau MR, Monfort A, Orsat B, Spielmann A, Stutz E (1993) Complete sequence of *Euglena gracilis* chloroplast DNA. Nucleic Acids Res 21(15):3537-3544

Huber-Pestalozzi G (1955) Das Phytoplankton des Süsswassers, Stytematik und Biologie,4. Teil: Euglenophyceen. In Die Binnengewasser. Band 16, 4. Teil, ed. A. Thienemann. Stuttgart: Schweizerbart'sche Verlagsbuchhandlung, 606 pp

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754-5

Jenkins KP, Hong L, Hallick RB (1995) Alternative splicing of the *Euglena gracilis* chloroplast roaA transcript. RNA (NY) 1:624-633

Karnkowska-Ishikawa A, Watza D, Bennett MS, Triemer RE (in prep) Phylogenetic analysis of the photosynthetic euglenoids based upon 5 ribosomal and protein coding genes. J Phycol

Kim JI, Shin W (2008) Phylogeny of the euglenales inferred from plastid LSU rDNA sequences. J Phycol 44(4):994-1000

Kim JI, Shin W, Triemer RE (2010) Multigene analysis of photosynthetic euglenoids and new family, Phacaceae (Euglenales). J Phyc 46(6):1278-1287

Koller B, Delius H (1982) A chloroplast DNA of *Euglena gracilis* with five complete rRNA operons and two extra 16S rRNA genes. Mol Gen Genet 188(2):305-308

Lambowitz AM, Belfort M (1993) Introns as mobile genetic elements. Annu Rev Biochem 62:587-622

Lane CE, Archibald JM (2008) The eukaryotic tree of life: endosymbiosis takes its TOL. Trends Ecol Evol 23(5):268-275

Leander BS, Triemer RE, Farmer MA (2001) Character evolution in heterotrophic euglenoids. Eur J Protistol 37:337-356

Leedale GF (1967) Euglenida-Euglenophyta. Annu Rev Microbiol 21:31-48

Lemieux C, Otis C, Turmel M (2007) A clade uniting the green algae *Mesostigma viride* and *Chlorokybus atmophyticus* represents the deepest branch of the Streptophyta in chloroplast genome-based phylogenies. BMC Biol 5:2

Linton EW, Hittner D, Lewandowski C, Auld T, Triemer RE (1999) A molecular study of euglenoid phylogeny using small subunit rDNA. J Eukaryot Microbiol 46:217-223

Linton EW, Karnkowska-Ishikawa A, Kim JI, Shin W, Bennett MS, Kwiatowski J, Zakryś B, Triemer RE (2010) Reconstructing euglenoid evolutionary relationships using three genes: Nuclear SSU and LSU, and chloroplast SSU rDNA sequences and the description of *Euglenaria* gen. nov. (Euglenophyta). Protist 161(4):603-619

Linton EW, Nudelman MA, Conforti V, Triemer RE (2000) A molecular analysis of the euglenophytes using SSU rDNA. J Phycol 36(4): 740-746

Linton E, Shin W, Nudelman A, Monfils A, Bennett M, Brosnan S, Triemer RE (2006) Phylogeny of the Euglenales based upon combined SSU and LSU rDNA sequence

comparisons and description of *Discoplastis* Gen Nov (Euglenophyta). J Phycol 42:731-740

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

Manhart JR, Palmer JD (1990) The gain of two chloroplast tRNA introns marks the green algal ancestors of land plants. Nature 345:268-270

Manning JE, Wolstenholme DR, Ryan RS, Hunter JA, Richards OC (1971) Circular chloroplast DNA from Euglena gracilis. PNAS 68:1169-1173

Marin B, Palm A, Klingberg MAX, Melkonian M (2003) Phylogeny and taxonomic revision of plastid-containing euglenophytes based on SSU rDNA sequence comparisons and synapomorphic signatures in the SSU rRNA secondary structure. Protist 154(1):99-145

Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D (2002) Evolutionary analysis of Arabidopsis, cyanobacterial and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. PNAS 99(19):12246-12251

Martin W, Stoebe B, Goremykin V, Hansmann S, Hasegawa M, Kowallik KV (1998) Gene transfer to the nucleus and the evolution of chloroplasts. Nature 393:162-165

Milanowski R, Zakryś B, Kwiatowski J (2001) Phylogenetic analysis of chloroplast small-subunit rRNA genes of the genus Euglena Ehrenberg. Int J Syst Evol Micobiol 51(3):773-781

Milanowski R, Kosmala S, Zakrys B, Kwiatowski J (2006) Phylogeny of photosynthetic euglenophytes based on combined chloroplast and cytoplasmic SSU rDNA sequence analysis. J Phycol 42(3):721-730

Mohr G, Perlman PS, Lambowitz AM (1993) Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. Nucleic Acids Res 22(22):4991-4997

Moreira D, von der Heyden S, Bass D, López-García P, Chao E, Cavalier-Smith T (2007) Global eukaryote phylogeny: Combined small- and large-subunit ribosomal DNA trees support monophyly of Rhizaria, Retaria and Excavata. Mol Phylogenet Evol 44(1):255-266

Nock CJ, Waters DLE, Edwards MA, Bowen SG, Rice N, Cordeiro GM, Henry RJ (2011) Chloroplast genome sequences from total DNA for plant identification. Plant Biotechnol J 9(3):328-333

Noe L, Kucherov G (2005) YASS: Enhancing the sensitivity of DNA similarity search. Nucleic Acids Res 33(2):W540-543

Parfrey LW, Grant J, Tekle YI, Lasek-Nesselquist E, Morriosn HG, Sogin ML, Patterson DJ, Katz LA (2010) Broadly sampled multigene analyses yield a well-resolved eukaryotic tree of life. Syst Biol 59(5):518-533

Parks M, Cronn R, Liston A (2009) Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. BMC Biol 7(1):84

Preisfeld A, Berger S, Busse I, Liller S, Ruppel HG (2000) Phylogenetic analyses of various euglenoid taxa (Euglenozoa) based on 18S rdna sequence data. J Phycol 36(1):220-226

Preisfeld A, Busse I, Klingberg M, Talke S, Ruppel HG (2000) Phylogenetic analyses of various euglenoid taxa (Euglenozoa) based on 18S rDNA sequence data. J Phycol 36:220-226

Qiu Y-L, Li L, Wang B, Chen Z, Knoop V, Groth-Malonek M, Dombrovska O, Lee J, Kent L, Rest J, Estabrook GF, Hendry TA, Taylor DW, Testa CM, Ambros M, Crandall-Stotler B, Duff RJ, Stech M, Frey W, Quandt D, Davis CC (2006) The deepest divergences in land plants inferred from phylogenomic evidence. PNAS 103(42):15511-15516

Raubeson LA, Peery R, Chumley TW, Dziubek C, Fourcade HM, Boore JL, Jansen RK (2007) Comparative chloroplast genomics: analyses including new sequences from the angiosperms Nuphar *advena* and *Ranunculus macranthus*. BMC Genomics 8:174

Schattner P, Brooks AN, Lowe TM (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res 33:W686-689

Sheveleva EV, Giordani NV, Hallick RB (2002) Identifiaction and comparative analysis of the chloroplast α-subunit gene of DNA-dependent RNA polymerase from seven Euglena species. Nucl Acids Res 30(5):1247-1254

Sheveleva EV, Hallick RB (2004) Recent horizontal intron transfer to a chloroplast genome. Nucleic Acids Res 32(2):803-810

Simpson AGB, Roger AJ (2004) Protein phylogenies robustly resolve the deep-level relationships within Euglenozoa. Mol Phylogenet Evol 30(1):201-212

Smith SW, Overbeek R, Woese CR, Gilbert W, Gillevet PM (1994) The genetic data environment an expandable GUI for multiple sequence analysis. Bioinformatics 10(6):671-675

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, and Kumar S (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol Biol Evol (Epub ahead of print)

Triemer RE, Linton E, Shin W, Nudelman A, Monfils A, Bennett M, Brosnan S (2006) Phylogeny of the euglenales based upon combined SSU and LSU rDNA sequence comparisons and description of Discoplastis gen. nov. (Euglenophyta). J Phycol 42(3):731-740

Thompson MD, Copertino DW, Thompson E, Favreau MR, Hallick RB (1995) Evidence for the late origin of introns in chloroplast genes from an evolutionary analysis of the genus Euglena. Nucleic Acids Res 23(23):4745-4752

Turmel M, Gagnon M.-C, O'Kelly CJ, Otis C, Lemieux C (2009) The chloroplast genomes of the green algae *Pyramimonas*, *Monomastix*, and *Pycnococcus* shed new light on the evolutionary history of Prasinophytes and the origin of the secondary chloroplasts of euglenids. Mol Biol and Evol 26(3):631-648

Turmel M, Otis C, Lemieux C (1999) The complete chloroplast DNA sequence of the green alga Nephroselmis olivacea: Insights into the architecture of ancestral chloroplast genomes. PNAS USA 96:10248-10253

Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JAM (2007) Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res 35(2): W71-74

Von der Heyden S, Chao EE, Vickerman K, and Cavalier-Smith T (2004) Ribosomal RNA phylogeny of Bodonid and Diplonemid flagellates and the evolution of Euglenozoa. J Eukaryot Microbiol 51:402-416

Watanabe MM, Hiroki M (1997) In NIES - Collection List of Strains, 5th ed. National Institute for Environmental Studies, Tsukuba, 127 pp

Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20(17):3252-3255

Xiong A-S, Peng R-H, Zhuang J, Gao F, Zhu B, Fu X-Y, Xue Y, Jin X-F, Tian Y-S, Zhao W, Yao Q-H (2009) Gene duplication, transfer, and evolution in the chloroplast genome. Biotechnol Adv 27(4):340-347

Yepiz-Plascencia GM, Rodebaugh CA, Hallick RB (1990) The *Euglena gracilis* chloroplast rpoB gene: Novel gene organization and transcription of the RNA polymerase subunit operon. Nucleic Acids Res 18(7):1869-1878

Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821-829

Zhang L, Jenkins KP, Stutz E, Hallick R (1995) The *Euglena gracilis* intron-encoded mat2 locus is interrupted by three additional group II introns. RNA (NY) 1:1079-1088