SHEDDING LIGHT ON EUGLENOID EVOLUTION AND SYSTEMATICS THROUGH THE CHLOROPLAST GENOMES OF *EUGLENA VIRIDIS* AND *EUGLENAFORMIS* [*EUGLENA*] *PROXIMA*

By

Matthew Scott Bennett

A THESIS

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

Plant Biology - Master of Science

2013

ABSTRACT

SHEDDING LIGHT ON EUGLENOID EVOLUTION AND SYSTEMATICS THROUGH THE CHLOROPLAST GENOMES OF *EUGLENA VIRIDIS* AND *EUGLENAFORMIS* [*EUGLENA*] *PROXIMA*

By

Matthew Scott Bennett

The chloroplast genomes of *Euglena viridis* and *Euglenaformis* [*Euglena*] *proxima* were sequenced and analyzed against the chloroplast genomes of other previously sequenced algal taxa. The chloroplast genome of *E. viridis* was sequenced in order to explore intrageneric chloroplast evolution, and our results revealed that while the chloroplast genome of *E. viridis* closely resembled that of *Euglena gracilis*, it did show significant differences. The chloroplast genome of *E. viridis* was far more compact, had a gene cluster that was reversed in both gene order and strand orientation, had a region that was comprised almost entirely of open reading frames, and had substantially fewer introns than *E. gracilis*. However, despite these differences, it was clear that the majority of chloroplast evolution in the genus *Euglena* probably occurred before its divergence from the rest of the photosynthetic euglenoids. The chloroplast genome of *E. proxima* was sequenced in an attempt to clarify its relationship to the rest of the photosynthetic euglenoids. Genetic data obtained from the chloroplast genome sequence were used in phylogenomic analyses to compare 78 chloroplast-encoded genes from *E. proxima* with those found in six photosynthetic euglenoids and three prasinophytes. The results of these analyses were consistent with the results of previous phylogenetic analyses using a small number of small subunit and large subunit rDNA genes and supported the position of *E. proxima* as sister to all of the Euglenaceae. Based on these data, *E. proxima* was removed from the genus *Euglena*, and a new genus, *Euglenaformis*, was erected for this taxon.

ACKNOWLEDGEMENTS

I must thank, first and foremost, my advisor Dr. Richard Triemer for encouraging me to pursue my Master's Degree through his lab. I have had the opportunity to work for Dr. Triemer since 2002 as his lab manager, and since that time, I have been in search of a field of research that I found interesting enough in which to pursue a graduate degree. In 2010, his lab began to explore a new field of study in Chloroplast Genomics, and Dr. Triemer quickly recognized my interest and abilities in this research and encouraged me to pursue my Master's degree. Dr. Triemer has been the best advisor and boss that anyone could ask for, and I know I would not be the scientist, and person, that I am today without his wisdom and guidance for all these years.

Secondly, I would like to thank Krystle Wiegert for all of her assistance on these projects, and for dealing with me on a day-to-day basis. When I think about the struggles we went through to develop protocols, to determine which programs will do what we need them to do, and to determine the best way to conduct our research projects, I'm amazed that we've gotten to the point we are at now.

I would like to thank my committee members, Dr. Robin Buell and Dr. Shinhan Shiu, for agreeing to be on my committee. I know you are both extremely busy with your own research programs, and yet you found the time to meet with me and answer questions when I had them. Your wisdom and advice was invaluable throughout the completion of these projects.

Finally, I would like to thank my wife, Melody, and my children, Anna and Isaac. You have had to put up with an extremely busy and stressed husband and father for the last few years. I only hope that all the time and effort put into this degree will make our lives better in the end.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

BLAST – Basic Local Alignment Search Tool

CDS – DNA Coding sequence or region

DNA – DeoxyriboNucleic Acid

DOGMA – Dual Organellar GenoMe Annotator

GDE – Genetic Data Environment

Kb – Kilobase

MEGA – Molecular Evolutionary Genetics Analysis

MUSCLE - MUltiple Sequence Comparison by Log- Expectation

NCBI – National Center Biotechnology Information

PCR – Polymerase Chain Reaction

rRNA – Ribosomal RiboNucleic Acid

SAG – Sammlung von Algenkulturen Gottingen culture collection

tRNA – Transfer RiboNucleic Acid

Chapter 1

Comparative chloroplast genomics between *Euglena viridis* and *Euglena gracilis* (Euglenophyta)

Introduction

It has been accepted for some time that the origin of the euglenoid chloroplast was through a secondary endosymbiotic event wherein a phagotrophic euglenoid engulfed a green algal cell (Gibbs 1978, 1981). However, until recently, little research had been conducted to explore the validity of these claims or to explore chloroplast evolution within the photosynthetic euglenoids.

The first published chloroplast genome (cpGenome) of a photosynthetic euglenoid was that of *Euglena gracilis* Klebs (Rawson *et al*. 1978), which described its gene content, intron content and gene order. This study also verified previous work that showed *E. gracilis* contains a region in which the ribosomal RNA operon is tandemly repeated as opposed to the inverted repeat that is typically found in green algae and higher plants (Hallick 1982). A subsequent comparative study of cpGenomes also recognized that *E. gracilis* has a highly reduced number of genes within the genome as compared to all other sequenced chloroplasts (Turmel *et al*. 1999).

The *E. gracilis* cpGenome was followed by the cpGenome of *Euglena* (*Astasia*) *longa* (Pringsheim) Marin & Melkonian, a colorless euglenoid that had secondarily lost its ability to photosynthesize due to the loss of many of the genes associated with photosynthesis (Gockel & Hachtel 2000). The genes remaining in the *E. longa* cpGenome showed both a high degree of gene conservation and a nearly identical transcriptional and translational gene content with *E. gracilis*, missing only one of the ribosomal proteins (*rps*18; Gockel & Hachtel 2000).

Recently, the chloroplast genomes of *Eutreptia viridis* Perty (Wiegert *et al*. 2012), *Eutreptiella gymnastica* Throndsen (Hrdá *et al*. 2012), *Colacium vesiculosum* Ehrenberg and *Strombomonas acuminata* (Schmarda) Deflandre (Wiegert *et al*. 2013) have been published. All four of these genomes showed that there is a large amount of gene content conservation within the photosynthetic euglenoids; however, the studies also showed that there have been substantial gene rearrangements and large disparities in genome sizes. While previous studies have concentrated on the similarities and differences in cpGenomes between genera, it is still not known if the extensive gene rearrangements and differences in genome sizes have occurred among genera within a genus.

The genus *Euglena* was chosen as a strategic group in which to explore intrageneric chloroplast evolution because this genus has the greatest morphological chloroplast diversity found in the photosynthetic euglenoids (Linton *et al*. 2010). Chloroplasts can be plate-like, lobed, disc shaped, lenticular, incised, ribbon-like, fimbriate or stellate (Ciugulea & Triemer 2010). In addition, the chloroplasts may be with or without pyrenoids, and the pyrenoids may be either with or without a paramylon cap (Ciugulea & Triemer 2010). In an effort to gain a better understanding of this genus and to explore intrageneric cpGenome evolution, the cpGenome of *Euglena viridis* Ehrenberg was sequenced.

Euglena viridis was chosen for several reasons. First, *E. viridis* is the type species, and the culture of *E. viridis* used in this study is the same culture that was used to establish the lyophilized epitype (Shin & Triemer 2004). Second, the cells are morphologically different from *E. gracilis* in body shape, position of the chloroplast, type of chloroplast and organization of the paramylon grains surrounding the pyrenoid. *Euglena gracilis* cells are cylindrically shaped with multiple disc-shaped chloroplasts and diplopyrenoids, while *E. viridis* cells are spindle shaped

with a single stellate chloroplast surrounded by short paramylon rods to form a 'paramylon center' (Table 1.1). Third, in phylogenetic trees, *E. viridis* and *E. gracilis* occupy positions in separate subclades within the greater *Euglena* clade, indicating that while these two taxa are related, they are both highly diverged from their last common ancestor (Figure 1; Triemer *et al*. 2006; Linton *et al*. 2010). These differences between the two photosynthetic *Euglena* species have allowed us to explore intrageneric chloroplast evolution, and to discover what, if any, changes have occurred within cpGenomes among divergent but related taxa.

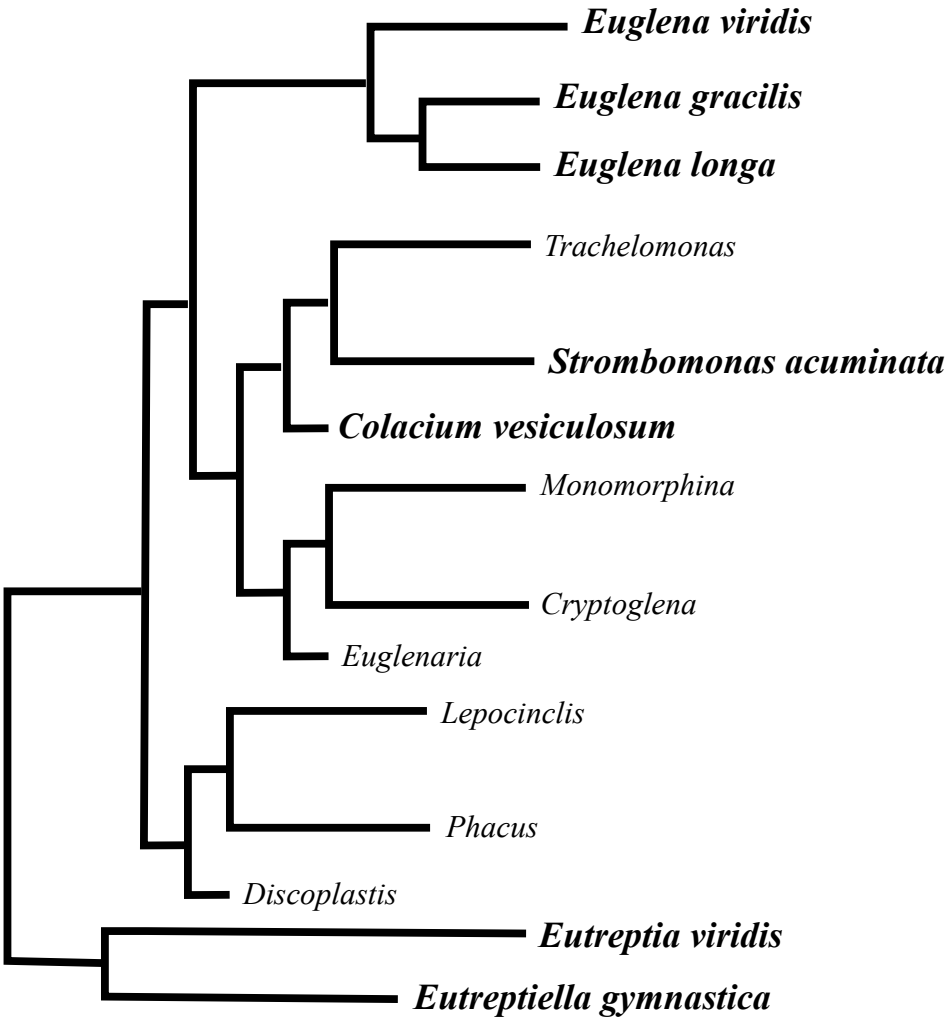

Figure 1.1: Diagrammatic phylogeny showing the relative positions of all seven photosynthetic euglenoid taxa with sequenced cpGenomes. Figure was redrawn based on Linton *et al*. (2010, figure 2).

Table 1.1: Comparison of chloroplast morphology and genome size for the six sequenced photosynthetic euglenoids. Genome size data: *Colacium vesiculosum* and *Strombomonas acuminata* (Wiegert *et al*. 2013), *Euglena gracilis* (Hallick *et al*. 1993), *Euglena longa* (Gockel & Hachtel 2000), *Eutreptia viridis* (Wiegert *et al*. 2012), *Eutreptiella gymnastica* (Hrdá *et al*. 2012).

Taxon	Culture strain	Chloroplast morphology		Genome Size (bp)
		Shape	Pyrenoid Type	
Colacium vesiculosum	CCAP 1211/3	Disc-shaped	Haplopyrenoids	128,900
Euglena gracilis	Pringsheim, strain Z	Disc-shaped	Diplopyrenoids	143,170
Euglena longa	SAG 1204-17a	Colorless Plastid	Not Applicable	73,345
Euglena viridis	ATCC PRA-110	Stellate with deeply lobed extensions	Surrounded by short paramylon grains forming a paramylon center	91,606
Eutreptia viridis	SAG 1226-1c	Elongated and arranged in a stellate-like pattern	Surrounded by short paramylon grains forming a paramylon center	65,513
Eutreptiella gymnastica	SCCAP K-0333	Reticulate	Diplopyrenoid	67,622
Strombomonas acuminata	S716	Disc-shaped	Haplopyrenoids	144,166

Materials and Methods

Euglena viridis cells (strain ATCC PRA-110; American Type Culture Collection, Manassas, VA, USA) were grown in AF-6 medium (Watanabe & Hiroki 1997) under the following growth conditions: 20–22ºC; 10hr:14hr light:dark cycle under cool white fluorescent tubes which provided approximately 30 µmol photons m^{-2} s^{-1} of light. After 14 days of growth, cells from all culture tubes were combined and washed three times in order to minimize bacterial content. Washing was completed by centrifuging cells for 4 min at 1800 x *g*, removing the resulting supernatant as well as any bacterial layer (which appeared white) that was above the green layer of cells and resuspending cells in fresh AF-6 medium. Cells were further cleaned using a Centricoll (Sigma Inc., St. Louis, MO, USA; catalog no. C0580) gradient (1 ml, 100%; 3 ml, 60%; 2 ml, 40%) that was centrifuged for 10 min at 4000 x *g*. Cells at the 60% / 40% interface were collected and washed three times, following the washing protocol above, to remove the Centricoll. The cleaned cells were run through Qiagen DNeasy Blood & Tissue Kit columns (Qiagen, Germantown, MD, USA; catalog no. 69504), following the protocol for purification of total DNA from animal tissues to extract the genomic DNA (gDNA). Extracted gDNA was then sequenced using single reads on a half plate of Roche 454 GS FLX/Titanium Genome Analyzer (454 Life Sciences, Branford, CT, USA) at the Virginia Commonwealth University Nucleic Acids Research Facility.

Raw sequencing data was assembled using the Roche GS De novo Assembler (454 Life Sciences), with default settings, to create contigs. The *E. gracilis* chloroplast genome (GenBank accession no. NC_001603) was then used as a query sequence to BLAST against the contigs with BLAST+ (Camacho *et al*. 2009;

ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST) to identify chloroplast sequences. The resulting contigs were viewed using Tablet (Milne *et al*. 2010; http://bioinf.hutton.ac.uk/tablet), with the 'Trim poor quality reads using QA tags (ACE only)' option deselected under the Preferences/Importing menu in order to view the low-quality read ends of the contigs that had been clipped off during assembly. The previously clipped sequences were then extracted from each end of the contigs and imported into MacGDE (http://macgde.bio.cmich.edu), where they were manually realigned and consensus sequences produced. These consensus sequences were used as query sequences to BLAST against all of the assembled contigs using BLAST+. As a result, additional contigs were found that brought the consensus sequence to near completion. PCR primers were manually created (Table 1.2) based on the nucleotide sequences near the ends of the contigs and checked with the Primer_Check function in Primer3Plus (Untergasser *et al*. 2007; http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi), with the following General Settings differing from default: Primer Tm (Min: 40, Opt: 50, Max: 65) and Primer GC% (Opt: 50). Fill-in PCR was then employed to bridge the gaps between the contigs. The PCR reactions were performed using a Bio-Rad C1000 Thermal Cycler (Bio-Rad Laboratories, Hercules, CA, USA), and the program used for all PCR reactions was as follows: 96ºC 2 min, 35 cycles (95ºC 30 s, 40–65ºC 30 s, 72ºC 1 min), 72ºC 6 min, 15ºC hold. Following amplification, products were run on 1% agarose gels in order to identify, size and purify the DNA bands. The DNA bands were then excised from the gel and extracted using a Qiagen MinElute Gel Extraction Kit (catalog no. 28606) following the manufacturer's protocol. The extracted DNA was sequenced on an ABI 3730XL DNA Analyzer (Applied Biosystems Inc., Foster City, CA, USA) at the Michigan State University Research Technology Support Facility to at least double coverage. All sequences (contigs and PCR

products) were viewed and manually aligned in MEGA 5 (Tamura *et al*. 2011; http://megasoftware.net) to create the completed chloroplast genome sequence.

Table 1.2: PCR primers created for the chloroplast genome of *Euglena viridis.*

Location in the Genome	**Sequence**
Complement(259..278)	ACG GAT CCC TTA TCC TAA CG
12515..12534	CTG AAG TTA TAA ATG ACT GG
Complement(12996..13013)	GAA GCA TTA TCC ATG CAA
16923..16940	GAA GGC CTA GGC GTG AAC
Complement(17304..17324)	CTA ATT CCA TTT CAA GAT CAG
26081..26100	AGC GTC ACA GAT AGG AAT CG
Complement(26611..26630)	ATT TAT CGA GGT AAG TAC GC
90491..90510	CAC GCG GCA TTG CTC CGT CA
90835..90859	AGC GTT CAT CCT GAG CCA GGA TCA A[1]

[1]Primer originally published in Wiegert *et al*. 2012.

Once the chloroplast genome sequence was created, it was submitted to DOGMA (Wyman *et al*. 2004; http://dogma.ccbb.utexas.edu) in order to generate a 'backbone' upon which to build the annotation. To identify all open reading frames, the cpGenome was submitted to NCBI's ORF (open reading frame) Finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html), using the option for standard genetic codes. Also, in order to identify tRNAs, the cpGenome was submitted to tRNAscan-SE (Schattner *et al*. 2005; http://lowelab.ucsc.edu/tRNAscan-SE) using default

parameters and the source chosen as Mito/Chloroplast. Only those tRNAs with a cove score above 50 were accepted from tRNAscan-SE and included in the final annotation. The chloroplast genome sequence as well as resulting DOGMA annotations, NCBI ORF Finder results and tRNAscan-SE results were inputted into CLC Genomics Workbench (CLC Bio, Cambridge, MA, USA) for final annotation. The option to display translations in all six frames was turned on as well as having the Bacterial and Plant Plastid table chosen and the 'only AUG start codons' option deselected in order to aid in annotation. Annotation was performed using a total-evidence approach, and the following rules were applied to the process of manual annotation.

Protein coding genes were identified using both DOGMA annotation and BLASTX searches. When possible, the protein sequences from individual genes were then extracted from the genome and manually aligned in MEGA 5 against the gene sequences from other photosynthetic euglenoids, as well as green algal representatives, to better determine intron/exon boundaries. In all cases an annotation with a methionine start codon was preferred; however, there were a few cases where alternative start codons were accepted due to the lack of traditional start codons.

The plastid-encoded 16S and 23S rRNA genes were identified using both DOGMA annotation and BLASTN searches. The plastid-encoded 5S rRNA gene was not identified by either of these methods, so an alternative approach was utilized: The nucleotide sequence between *psa*I and the 23S rRNA gene was extracted from the genome and imported into MEGA 5, as well as the annotated 5S genes from both the *E. gracilis* (GenBank accession no. NC_001603) and *E. longa* (GenBank accession no. NC_002652) cpGenomes. A MUSCLE (Edgar 2004) alignment was performed, and the portion of the extracted chloroplast genome that aligned with the other 5S genes was identified as the 5S.

With one exception, all open reading frames that were at least 300 nucleotides long, lacked BLAST evidence for being an identified protein-coding gene and did not overlap with an identified gene were included in the final annotation. The lone exception was an instance where two ORFs were identified in the same region of the genome but on opposite strands. In this case the larger ORF that was located on the same strand as the other annotated features in that area was retained. In all cases the ORFs were named according to the number of amino acids in the open reading frame.

To help determine the number of ribosomal operons present in the genome, long-range PCR was employed using the Qiagen LongRange PCR kit (catalog no. 206401) following the manufacturer's protocol for 0.1–10 Kb. Primers utilized for long-range PCR were 90491..90510 and Complement(259..278) (Table 1.2).

Synteny between the cpGenomes of *E. viridis* and *E. gracilis* was determined by using Mauve v. 2.3.1 (Darling *et al*. 2004; http://asap.ahabs.wisc.edu/mauve). This program allows for syntenic comparison between multiple genomes and displays these regions graphically. The cpGenome sequence, along with annotations, has been accessioned into GenBank (accession no. JQ237893).

Results

Following assembly with the Roche GS De novo Assembler, four contigs containing chloroplast sequences were returned (Figure 1.2). A total of 5,385 reads were used to create the four contigs, with an average coverage depth of 41 reads. These contigs, along with the utilization of fill-in PCR to bridge gaps between the contigs, allowed us to close the circular cpGenome (Figure 1.2). The *E. viridis* cpGenome was found to be 91,606 base pairs (bp) in length, with position 1 of the

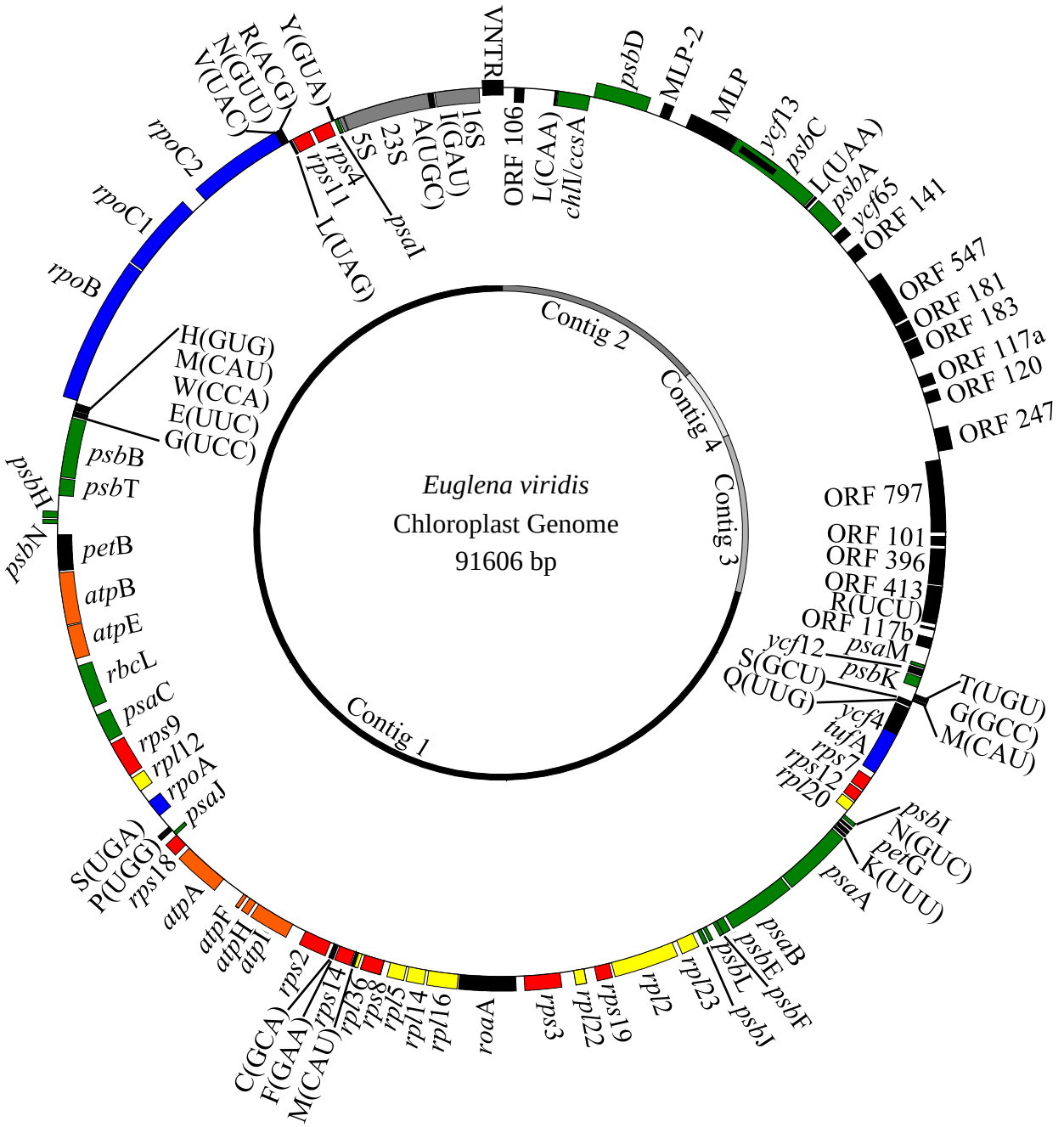

Fig. 1.2: Gene map of the *Euglena viridis* chloroplast genome. Outer ring: The box colors are common for genes of similar functional groups. Green: photosystems/photosynthesis genes; yellow: large ribosomal proteins, *rpl* genes; red: small ribosomal proteins, *rps* genes; blue: transcription/translation related genes, *rpo* genes, *tuf*A; orange: *atp* genes; black: miscellaneous, conserved hypothetical proteins (*ycf*), open reading frames (ORF), tRNAs, maturase-like proteins (MLP), Variable Number Tandem Repeat (VNTR); grey: ribosomal rRNAs. The positions of the genes on the outside or inside of the outer ring are representative of the positive and negative strands, respectively. The block size for each gene is proportional to its sequence length. Inner ring: Positions of the four contigs that make up the assembled chloroplast genome. Contigs are numbered in decreasing size order: Contig 1 is the largest. To interpret references to color in this and all other figures, the reader is referred to the electronic version of this thesis.

sequence immediately following the Variable Number Tandem Repeat (VNTR) sequence as was established by Hallick *et al*. (1993). Bases were numbered clockwise through the genome, and two complete copies of the VNTR sequence plus the partial repeat were included (Figure 1.2). The overall nucleotide content of the cpGenome was 26.4% G+C and 73.6% A+T, which was remarkably similar to the base composition reported for *E. gracilis*: 26.1% G+C and 73.9% A+T (Hallick *et al*. 1993).

Overall, 91 genes were identified and annotated in the cpGenome of *E. viridis*. This includes 61 protein-coding genes, 27 tRNAs and 3 rRNAs (Figure 1.2). For the protein-coding genes, the gene size ranged from 4,740 bp in *rpo*B to 96 bp in *psa*M, with an overall average gene length of approximately 947 bp (including introns).

The cpGenome of *E. viridis* contained more ORFs than had been previously seen in other photosynthetic euglenoids, with 13 annotated overall (Figure 1.2). In addition to having the most ORFs, a large region of 13,773 bp, extending from approximately the 2:00 (*ycf*65) to the 4:00 (*psa*M) position of the cpGenome (Figures 1.2, 1.3), was identified, containing 12 of the 13 annotated ORFs and was punctuated only by a single tRNA. This region differed significantly from the rest of the cpGenome in that it contained no known chloroplast genes; whereas, the rest of the cpGenome was tightly packed with recognized chloroplast genes. Also of note is that this region continued the strandedness trends seen in the genes annotated both before and after the region.

A BLASTP analysis was conducted against the NCBI nonredundant protein sequences (nr) database to determine if any of the ORFs had functional similarity to previously sequenced genes (Table 1.3). Only one ORF, ORF 183, had a significant similarity match – to ORF 295 in the

Table 1.3: BLASTP analysis of the 13 ORFs annotated in the *E. viridis* cpGenome against the NCBI non-redundant protein sequences (nr) database. For each ORF the best match is reported. [1]dash (-) = no significant similarity

ORF	Accession #	BLASTP		E-value[1]
		Organism	Product	
106	YP_005713705.1	*Sinorhizobium meliloti* BL225C	unnamed protein product	3.5
141	YP_005444535.1	*Phycisphaera mikurensis* NBRC 102666	unnamed protein product	8
547	EFW41724	*Capsaspora owczarzaki* ATCC 30864	hypothetical protein CAOG_06856	0.75
181	-	-	-	-
183	NP_041938.1	*Euglena gracilis*	hypothetical protein EugrCp050 (ORF 295)	2E-09
117a	XP_003396255.1	*Bombus terrestris*	PREDICTED: hypothetical protein LOC100646143	5.6
120	XP_973932.1	*Tribolium castaneum*	PREDICTED: similar to adenosine deaminase	0.24
247	AEX62269.1	*Moumouvirus Monve*	hypothetical protein mv_R64	0.014
797	EGW09712.1	*Cricetulus griseus*	Tubulin alpha chain	1.7
101	-	-	-	-
396	YP_001373361.1	*Bacillus cytotoxicus* NVH 391-98	replication initiation factor	1.3
413	YP_002607027.1	*Nautilia profundicola* AmH	hypothetical protein NAMH_0610	4.9
117b	YP_001885467.1	*Clostridium botulinum* B str. Eklund 17B	hypothetical protein CLL_A1269	1.5

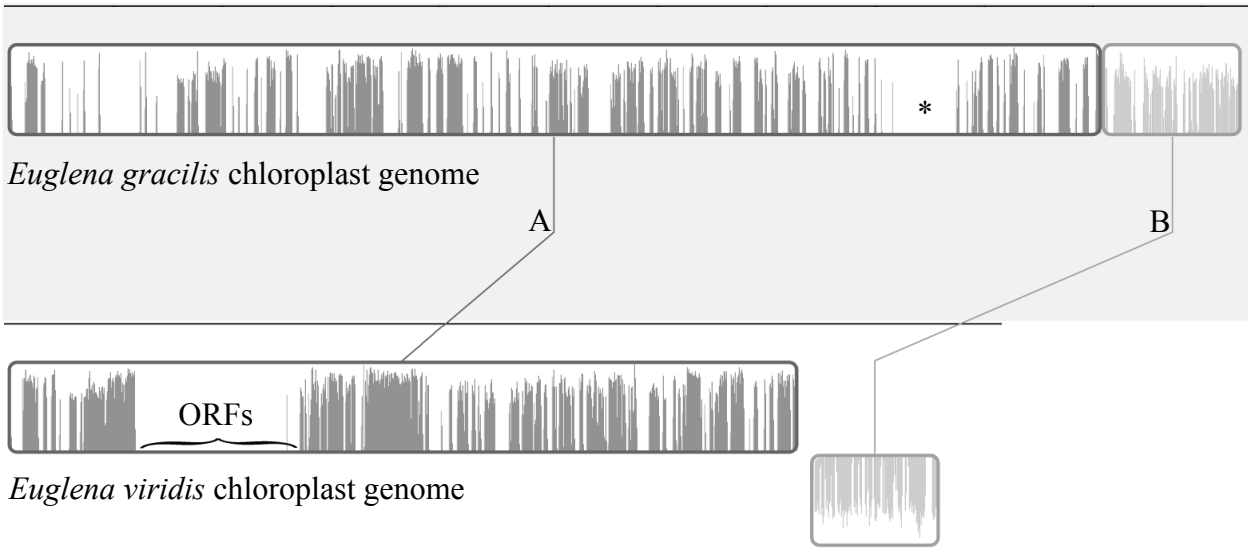

Figure 1.3: A Mauve analysis comparing the chloroplast genomes of *Euglena gracilis* and *Euglena viridis*. Linearized cpGenomes are shown with boxes that represent homologous gene clusters, and vertical lines within the boxes represent the extent that the genome sequence is conserved in that region. The positions of the boxes are relative to the *E. gracilis* cpGenome; consequently, boxes that lie above the horizontal line represent gene clusters that are oriented in the same direction as *E. gracilis*, while boxes that lie below the horizontal line represent gene clusters that are oriented in the opposite direction of *E. gracilis*. Box A: Base 1 – H(GUG). Box B: Y(GUA) – *rpo*B in *E. gracilis*; *rpo*B – Y(GUA) in *E. viridis*. ORFs = region of 12 ORFs in *E. viridis*. * = Region between genes *rbc*L and *atp*E in *E. gracilis*.

cpGenome of *E. gracilis* (Table 1.3). However, this match was only to a small section of *E. gracilis* ORF 295, and no other significant matches were found to any other organism. Interestingly, similar to ORF 183, *E. gracilis* ORF 295 occurs in a region of the *E. gracilis* cpGenome that contains only ORFs (Figure 1.3, indicated by *), although the other two ORFs annotated in that region share no similarity matches to any part of the *E. viridis* cpGenome. In addition to the previous functional analysis, the 13 ORFs were subjected to a BLASTP analysis against the NCBI Whole Gene Shotgun database to determine if similarities could be found to any previously reported sequence. This analysis returned no significant similarity matches for any of the 13 ORFs (Table 1.4).

Table 1.4: BLASTP analysis of the 13 ORFs annotated in the *E. viridis* cpGenome against the NCBI Whole Genome Shotgun (WGS) database. For each ORF the best match is reported. [1]dash (-) = no significant similarity

ORF	WGS Database		
	Accession #	Organism	E-value[1]
106	AAGW02074114.1	*Oryctolagus cuniculus*	0.003
141	-	-	-
547	-	-	-
181	-	-	-
183	-	-	-
117a	AAWZ02016552.1	*Anolis carolinensis*	3.7
120	ABJB010004481.1	*Ixodes scapularis*	0.27
247	AACT01019649.1	*Ciona savignyi*	5.1
797	-	-	-
101	-	-	-
396	AFSB01150306.1	*Heterocephalus glaber*	6.6
413	-	-	-
117b	CAAP03012806.1	*Vitis vinifera*	2.3

Discussion

A substantial size difference was found between the cpGenomes of *E. viridis* (91,606 bp) and *E. gracilis* (143,170 bp). One of the reasons for this difference is the number of ribosomal operons. The *E. gracilis* cpGenome reported by Hallick *et al*. (1993) contained three tandemly repeated complete copies of the ribosomal operon plus an additional 16S rRNA, while the cpGenome of *E. viridis* contained only one complete ribosomal operon. However, previous studies have shown that the number of ribosomal operons contained within the euglenoid chloroplast were not consistent, with different strains of *E. gracilis* containing one (Wurtz & Buetow 1981), two (Ravel-Chapuis *et al*. 1984), three (Wurtz & Buetow 1981) or five (Koller & Delius 1982) complete ribosomal operons as well as one or two (Koller & Delius 1982) additional partial operons.

Because of the nature of sequence reassembly following Next-Generation sequencing, a tandem repeat orientation of the ribosomal operon would make it difficult to determine the actual number of ribosomal operons contained within the chloroplast. In an attempt to overcome this, recent studies (Wiegert *et al*. 2012; Wiegert *et al*. 2013) compared the number of sequence reads over the ribosomal operon (read coverage) to the read coverage over single copy protein-coding genes. This method proved to be useful for estimating the potential number of tandem ribosomal repeats in other euglenoid cpGenomes. However, in *E. viridis*, the read coverage over the ribosomal operon was consistent with that of all other annotated genes contained in the same contig (Contig 1, Figure 1.2).

We further investigated whether a misassembled tandem repeat was present by visualizing Contig 1 (Figure 1.2) in Tablet, with the option to not trim the low-quality reads selected as before (see Materials and Methods). If multiple copies of the ribosomal operon were present, a minority of the reads at the beginning and the end of the assembled operon would contain the sequences of a preceding or following ribosomal operon (or both). However, when the contig was viewed in this manner, it was found that all reads making up the assembled ribosomal operon agreed with the reported alignment, indicating that no additional copies of the ribosomal operon were present.

A final test to determine the number of ribosomal operons present was performed using long-range PCR. Primers were created to amplify a section of the genome stretching from 376 bp into the 16S rRNA to bp 278 at the beginning of the genome (a size of 1,374 bp if only one copy of the ribosomal operon was present). The ribosomal operon within the *E. viridis* cpGenome totaled 4,712 bp in length; consequently, if multiple copies of the ribosomal operon were present, PCR products would be seen minimally at 1,374 bp and 6,086 bp. Following the long-range PCR

reaction, one major PCR product was present at ~ 1,400 bp with two smaller PCR products also present due a repetition of the primer Complement (259..278) sequence three times at the beginning of the genome. No other PCR products were amplified using long-range PCR.

All three methods used to test for the number of ribosomal operons gave the same result, indicating that the *E. viridis* cpGenome contains only one copy of the ribosomal operon. The single copy is consistent with most other sequenced photosynthetic euglenoid chloroplasts because it does not contain an inverted repeat, including the ribosomal operon. To date, only *Eutreptiella gymnastica* has an inverted repeat (Hrdá *et al*. 2012), a taxon located at the base of the photosynthetic euglenoid lineage (Figure 1.1). The lack of a ribosomal operon inverted repeat in *E. viridis* provides further evidence that this loss occurred near the base of the photosynthetic euglenoid lineage, with the division of the *Eutreptiella* and *Eutreptia* genera.

In addition to the number of ribosomal operon repeats, a substantial size difference between the two cpGenomes can be attributed to the variance in intergenic sequence length. This is noticeably apparent when the genomes are analyzed using Mauve (Figure 1.3), where the cpGenome of *E. viridis* shows markedly smaller intervening gene sequences. One example of this difference is the distance between the genes *rbc*L and *atp*E (Figure 1.3, indicated by *). In the *E. gracilis* cpGenome there is an intergenic sequence length of 5,816 bp, with enough distance to encode three ORFs. However, in *E. viridis*, that same intergenic sequence length is only 687 bp.

Gene Content and Synteny

The 91 cpGenome genes of *E. viridis* were similar to that of *E. gracilis*, which contained 88 annotated genes (Hallick *et al*. 1993). One difference between the two cpGenomes was the

presence of conserved chloroplast genes *ycf*65 and *psa*I in *E. viridis* (Figure 1.2) but not in *E. gracilis* (Hallick *et al*. 1993). We further investigated this difference by performing a tBLASTx search of the *E. gracilis* genome using the gene sequences from *E. viridis* as the query. We were unable to find any matches to these genes, even if e-value thresholds were lowered far below accepted limits (to 1E-1). With *ycf*65 this was surprising because it has been found in all other photosynthetic euglenoid sequences (Hrdá *et al*. 2012; Wiegert *et al*. 2012; Wiegert *et al*. 2013), so it was presumed by the authors that the gene either had been missed during the initial *E. gracilis* annotation or had been completely lost in the *Euglena* lineage. It now appears that this gene loss is restricted to *E. gracilis* and possibly its close relatives.

Another difference in the reported number of genes is that the maturase-like proteins in *E. gracilis* occur within *psb*C (Hallick *et al*. 1993); whereas, the maturase-like proteins in *E. viridis* occur as separate protein-coding genes (Figure 1.2, MLPs). Additionally, when compared to the cpGenome of *E. gracilis*, the cpGenome of *E. viridis* is missing only one gene, *rpl*32.

Other than the genes previously discussed, the overall gene content and synteny was extremely conserved. The only difference in synteny occurs in a section extending from the *rpo*B gene to tRNA Y(GUA), approximately the 9:00 to the 11:00 position in Figure 1.2, where this section is inverted and on the opposite strand as compared to *E. gracilis*. This movement is clearly evident in the Mauve analysis (Figure 1.3, Box B).

Alternative Start Codons

A total of five genes were annotated with alternative start codons, and all but one of these genes used an Isoleucine as the start codon, with the exception being *rpo*A, which used a Leucine (Table 1.5). This number of alternative-start-codon genes is not substantially different than what

Table 1.5: Alternative start codon usage in the chloroplast genome of *Euglena viridis* and a comparison of those genes to the alternative start codons used in other photosynthetic euglenoids for the same genes.

Gene	Start Codon	Amino Acid	Shared Taxa (Start Codon)(Amino Acid)
*atp*F	I	ATT	
*atp*I	I	ATA	
*psb*K	I	ATA	*C. vesiculosum*(ATA)(I)
*roa*A	I	ATT	*S. accuminata*(TTG)(L)
*rpo*A	L	TTG	

has been reported in all other cpGenomes of photosynthetic euglenoids, with four genes reported in *Eutreptia viridis* (Wiegert *et al*. 2012), and one gene reported in both *C. vesiculosum*, and *S. accuminata* (Wiegert *et al*. 2013). The cpGenome of *E. gracilis* only has two genes that do not contain a traditional methionine start codon, *psb*D and *psb*N, with both of them having an undetermined start codon.

Introns and VNTR Sequence

Seventy-one introns were identified in the cpGenome of *E. viridis*. This number differs significantly from the number of introns in *E. gracilis*, where over 150 have currently been found. However, the number of introns reported in *E. gracilis* identified all types of introns, including twintrons (introns within introns) (Hallick *et al*. 1993). An extensive analysis of intron types has not been performed on the cpGenome of *E. viridis*, and, consequently, the intron total reported here should be considered a minimum.

As found in the cpGenome of *E. gracilis* (Hallick *et al*. 1993), a VNTR sequence was identified in the cpGenome of *E. viridis* (Figure 1.2). The VNTR sequence in *E. viridis* was much larger than that found in *E. gracilis* (54 bp; Hallick *et al*. 1993) and consists of identical 284 bp repeats, followed by a 95 bp partial repeat. When this section of the cpGenome was sequenced via PCR, the products contained either one or two copies of the VNTR sequence and was always followed by the partial repeat.

Overall, the cpGenomes of *E. viridis* and *E. gracilis* were similar. Their G+C / A+T content, gene content, presence of a VNTR sequence, lack of an inverted repeat, and overall synteny was very comparable. The main differences between the cpGenomes were the sizes of the genomes (including the number of repeats of the ribosomal operon and the intergenic sequence lengths), a region of the genome consisting mostly of ORFs in *E. viridis*, and the number of introns present. These similarities and differences between the two cpGenomes indicate that while some intrageneric evolution has occurred, most of the major evolutionary changes in the genus *Euglena* occurred prior to the separation of the genus *Euglena* from the rest of the photosynthetic euglenoids. This finding is somewhat surprising given the extreme differences in chloroplast morphology between the two taxa. In view of this, we conclude that the chloroplast shape bears no significant consequence on the chloroplast genome content or synteny.

Chapter 2

Characterization of new genus *Euglenaformis* and the chloroplast genome of *Euglenaformis* [*Euglena*] *proxima*

Introduction

Euglena proxima is a widely distributed, pan-global, photosynthetic euglenoid species that was first described by P.A. Dangeard (1901) from water in a "muddy pit" near Poitiers, France. Since that time, this taxon has been identified in field collections on most continents outside of Europe, including Asia (Pham *et al.* 2011), Australia (Grimes 1988), North America (Smith 2010), and South America (Alves-da-Silva *et* Menezes 2010). Due to the fact that this taxon was commonly found in field collections and that its description fits well into the generic description of *Euglena*, little research had been conducted on this taxon since its initial description.

Beginning in 2006, a series of phylogenetic analyses using multiple genes revealed that despite its phenotypic association with the genus *Euglena*, *E. proxima* does not group with other *Euglena* taxa. Milanowski *et al*. (2006) showed in an analysis combining nuclear SSU rDNA (18S) and chloroplast SSU rDNA (16S) sequences that *E. proxima* has a well-supported (0.97 posterior probability (pp)) sister relationship not only to other *Euglena* taxa, but to a larger clade that also includes *Euglenaria*, *Monomorphina*, *Cryptoglena*, *Colacium*, *Strombomonas*, and *Trachelomonas* taxa. This sister relationship was later confirmed in 2010 by two independent phylogenetic analyses that used different combined datasets: 18S, nuclear LSU rDNA (28S) and 16S with a 0.99 pp (Linton *et al.* 2010); 18S, 16S and chloroplast LSU rDNA (23S) with a 1.0 pp (Kim *et al*. 2010). Both authors indicated that while the resulting paraphyly of the genus *Euglena* was unfortunate, they were unwilling to reassign *E. proxima* to a new genus until more data, or additional taxa pairing with *E. proxima*, had been obtained.

In an effort to acquire additional data to help elucidate the relationship of *E. proxima* with the rest of the photosynthetic euglenoid taxa, the chloroplast genome (cpGenome) was sequenced. This genome allowed us to perform more comprehensive analyses of these relationships and to discover if the cpGenome of *E. proxima* can help determine/confirm its phylogenetic relationships.

Materials and Methods

Euglena proxima strain SAG 1224-11a (Culture Collection of Algae at the University of Goettingen, Germany) cells were grown in AF-6 medium (Watanabe & Hiroki 1997) for 36 days at 20-22ºC with a 10hr:14hr light:dark cycle using cool white fluorescent bulbs that delivered approximately 30µmol photons m^{-2} s^{-1} of light. Cells were then washed and had their genomic DNA extracted as described in Bennett *et al*. (2012). Total genomic DNA was sequenced with Illumina 2x100 paired end reads using version 3 reagents at the Virginia Commonwealth University Nucleic Acids Research Facility. Due to the large number of reads generated in the sequencing process, the sequencing facility divided the paired-end read sequences into 31 separate pairs of files, and those file pairs were individually used for assembly. The paired-end reads were assembled into contigs with the 'De Novo Assembly' program in CLC Genomics Workbench (CLC Bio, Cambridge, MA, USA; http://www.clcbio.com), with the default settings. The resulting contigs were searched for chloroplast sequences using BLAST+ (Camacho et al. 2009; ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast?/LATEST) and the *Euglena gracilis* cpGenome (GenBank accession# NC_001603) as the query sequence. All contigs that contained chloroplast sequence were then manually aligned in MacGDE (http://macgde.bio.cmich.edu),

and a consensus sequence was produced. In order to connect the two ends of the circular genome, primers were created using Primer 3 (Rozen & Skaletsky 2000), accessed through the 'Create New Primers' function in Geneious Pro (Biomatters Ltd, Auckland, New Zealand; http://www.geneious.com/), based on the nucleotide sequences located near the ends of the consensus sequence (Table 2.1) and fill-in PCR was employed. Fill-in PCR, gel purification, gel extraction, and sequencing of PCR products were performed as described in Bennett *et al*. (2012). The resulting PCR sequences were manually aligned with the assembled consensus sequence in MacGDE and a final cpGenome sequence was produced. This cpGenome sequence was then submitted to DOGMA (Wyman et al. 2004; http://dogma.ccbb.utexas.edu) in order to generate a basic annotation of the genome and to aid in the final annotation process. The final annotation file was created in Geneious Pro, with the option to translate the nucleotide sequence in all frames selected, and the genetic code identified as bacterial.

Table 2.1: PCR primers created for the chloroplast genome of *Euglenaformis proxima.*

Location in the Genome	**Sequence**
complement(11..30)	TTA ATT ATC AAG TGC ACA CC
complement(372..393)	ACA CCC AGG AAA ACG TTG CAT T
77858..77877	CAC GCG GCA TTG CTC CGT CA[1]
complement(78322..78343)	AGC GCG TTG CTA CGA ACT ACG A
93030..93049	AGC ATG TTC CGC CCA ACC CG
93521..93274	CAT AGC TTC TAC CAC TAC GAG ACA

[1]Primer originally published in Bennett *et al*. 2012.

Protein coding genes and intron/exon boundaries were identified by aligning extracted portions of the cpGenome against GenBank sequences from both photosynthetic euglenoids and selected green algal representatives. A traditional Methionine start codon was always preferred for annotations; however, in one case (*rpl*20) a Methionine start codon could not be determined, so an alternative start codon was accepted.

The chloroplast-encoded 16S & 23S rRNA genes were determined by performing BLASTN searches. The 5S rRNA could not be identified through a BLASTN search, so the following procedure was employed in order to identify the 5S: the region of the cpGenome between psbH and 23S was extracted and imported into MEGA5 (Tamura *et al*. 2011) along with the 5S sequences from the cpGenomes of *Euglena gracilis*, *Euglena longa* (GenBank accession no. NC_002652), *Euglena viridis* (GenBank accession no. JQ237893), and *Eutreptia viridis* (GenBank accession no. JN643723). MUSCLE (Edgar 2004) was then used to align the 5S sequences and the extracted portion of the cpGenome. The portion of the cpGenome that aligned with the 5S sequences was then identified as the 5S gene. In order to determine the total number of ribosomal operons present in the cpGenome, long-range PCR was utilized with the Qiagen LongRange PCR kit (Qiagen, Germantown, MD, USA; catalog no. 206401), following the manufacturer's protocol for 0.1–10 Kb, and primers 77858..77877 and complement(78322..78343) (Table 2.1).

In order to determine the number of repeats present in the VNTR sequence, Primers 93521..93274 and complement(11..30) (Table 2.1), which directly flank the VNTR region, were used with the Qiagen LongRange PCR kit.

All tRNAs were identified with tRNAscan-SE (Schattner et al. 2005; http://lowelab.ucsc.edu/tRNAscan-SE) using default parameters and the source identified as 'Mito/Chloroplast'. Only tRNAs with a cove score above 49 were accepted for the final annotation.

Open Reading Frames (ORFs) were determined using NCBI's ORF Finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html), with the 'standard genetic codes' option chosen. Only those ORFs that were at least 300 nucleotides long, did not overlap with an identified gene, and lacked BLAST evidence for being a previously identified protein-coding gene were included in the final annotation. All ORFs were named 'ORF' followed by the number of amino acids in the open reading frame.

Synteny between the cpGenomes of *Euglena gracilis*, *Eutreptia viridis*, and *Euglenaformis* [*Euglena*] *proxima* was determined and visualized with progressive Mauve (Darling et al. 2004) accessed through the 'Align Whole Genomes…' function in Geneious Pro, using default parameters. Mauve performs syntenic comparisons between multiple genomes and displays these syntenic regions graphically.A phylogenomic analysis was conducted using six photosynthetic euglenoid cpGenome sequences and three cpGenome sequences from prasinophyte algae, the putative chloroplast donor (Turmel *et al*. 2009, Wiegert *et al*. 2012), available from GenBank. The following 78 genes were used in the phylogenomic analysis: 2 rRNA genes (16S, 23S), 22 tRNAs (V(UAC), W(CCA), F(GAA), C(GCA), Y(GUA), Q(UUG), N(GUU), H(GUG), E(UUC), K(UUU), N(GUC), G(UCC), I(GAU), L(UAG), M(CAU), P(UGG), R(ACG), R(UCU), S(GCU), S(UGA), T(UGU), A(UGC)), and 54 protein coding genes (*atp*A, *atp*B, *atp*E, *atp*F, *atp*H, *atp*I, *chl*I/*ccs*A, *pet*B, *pet*G, *psa*A, *psa*B, *psa*C, *psa*D, *psa*I, *psa*M, *psb*A, *psb*B, *psb*C, *psb*D, *psb*E, *psb*F, *psb*H, *psb*I, *psb*J, *psb*K, *psb*L, *psb*M, *psb*N, *psb*T, *rbc*L, *rpl*2, *rpl*5, *rpl*12,

*rpl*14, *rpl*16, *rpl*20, *rpl*22, *rpl*23, *rpl*32, *rpl*36, *rps*2, *rps*3, *rps*4, *rps*7, *rps*8, *rps*9, *rps*11, *rps*12, *rps1*4, *rps*18, *rps*19, *tuf*A, *ycf*4, *ycf*9). Individual genes were manually aligned in MEGA5 and MacGDE and only

homologous sites were used in the analysis. For protein-coding genes, the nucleotide CDSs were extracted from the GenBank files and aligned according to their amino acid sequences in order to determine homologous sites. Two Bayesian analyses were performed using Mr. Bayes 3.2.1 (Ronquist *et al*. 2012), with the protein-coding genes analyzed as nucleotide sequences, or as amino acid sequences. The dataset was partitioned into 3 groupings (rRNA genes, tRNAs, and protein-coding genes) so that an independent model would represent the evolutionary history of each gene category. The model used for nucleotide dataset partitions was determined by jModeltest 2.1.1 (Darriba *et al*. 2012), and the model used for the protein-coding genes amino acid dataset partition was determined by the 'Find Best DNA/Protein Models' function in MEGA5. The following models were used in the analyses: rRNA – GTR+I+G, tRNA – GTR+G, protein-coding genes (nucleotide) – GTR+I+G, protein-coding genes (amino acid) – cpREV+I+G. The analyses utilized four Markov chains (2,000,000 generations per chain), with trees saved every 100 generations, and the first 4,000 trees discarded. A majority-rule consensus tree was created from the remaining trees and convergence among the trees was confirmed by using the 'sump' command.

Results

Following contig assembly, 23 of the 31 paired-end read files returned a near complete cpGenome. The average read coverage per assembly ranged from approximately 109 to 178 reads, with an overall average read coverage of approximately 135 reads. Fill-in PCR using

nucleotide sequences at each end of the assembled genome allowed us to complete the circular genome, which was found to be at least 94,185 base pairs (bp) in length (Figure 2.1). Abiding by the conventions established by Hallick *et al*. (1993), base 1 of the sequence was assigned to the first base immediately following the Variable Number Tandem Repeat (VNTR) sequence, and the cpGenome was oriented such that the rRNA genes were on the reverse stand. The bases were numbered clockwise through the sequence, and the reported cpGenome includes 4 copies of the VNTR sequence plus a partial repeat. The base composition was 26.9% G+C and 73.1% A+T, which was comparable to the nucleotide content reported in previously sequenced photosynthetic euglenoid cpGenomes (*Euglena gracilis,* 73.9% A+T, Hallick *et al*. 1993; *Euglena viridis,* 73.6% A+T, Bennett *et al*. 2012; *Eutreptia viridis,* 72.3% A+T, Wiegert *et al*. 2012; *Eutreptiella gymnastica,* 65.68% A+T, Hrda *et al*. 2012; *Colacium vesiculosum,* 73.8% A+T, and *Strombomonas acuminata,* 73.4% A+T, Wiegert *et al*. 2013).

Ninety-one genes were identified in the cpGenome, including: 27 tRNAs, three rRNAs, and 61 protein coding genes (Figure 2.1). For one gene, *rpl*20, a Methionine start codon could not be identified, so a Valine (GTG) was accepted as the start codon. The protein coding gene sizes ranged from at least 6,648 bp in *psb*C to 96 bp in *psa*M, with an overall average gene size of 1,330 bp including introns. 68.9% of the protein coding genes contained at least one intron with a total of 113 introns annotated in the cpGenome and an average of 1.85 introns per gene. It should be noted that an extensive analysis on the types of introns has not been conducted on this cpGenome, and the number of introns reported here is considered a minimum. In addition to the features already mentioned, one ORF was identified in the cpGenome, ORF144, which occurred between the genes *psb*Z and tRNA S(UGA) (Figure 2.1). The cpGenome sequence with annotations (Figure 2.1) has been accessioned into GenBank (accession no. KC684276).

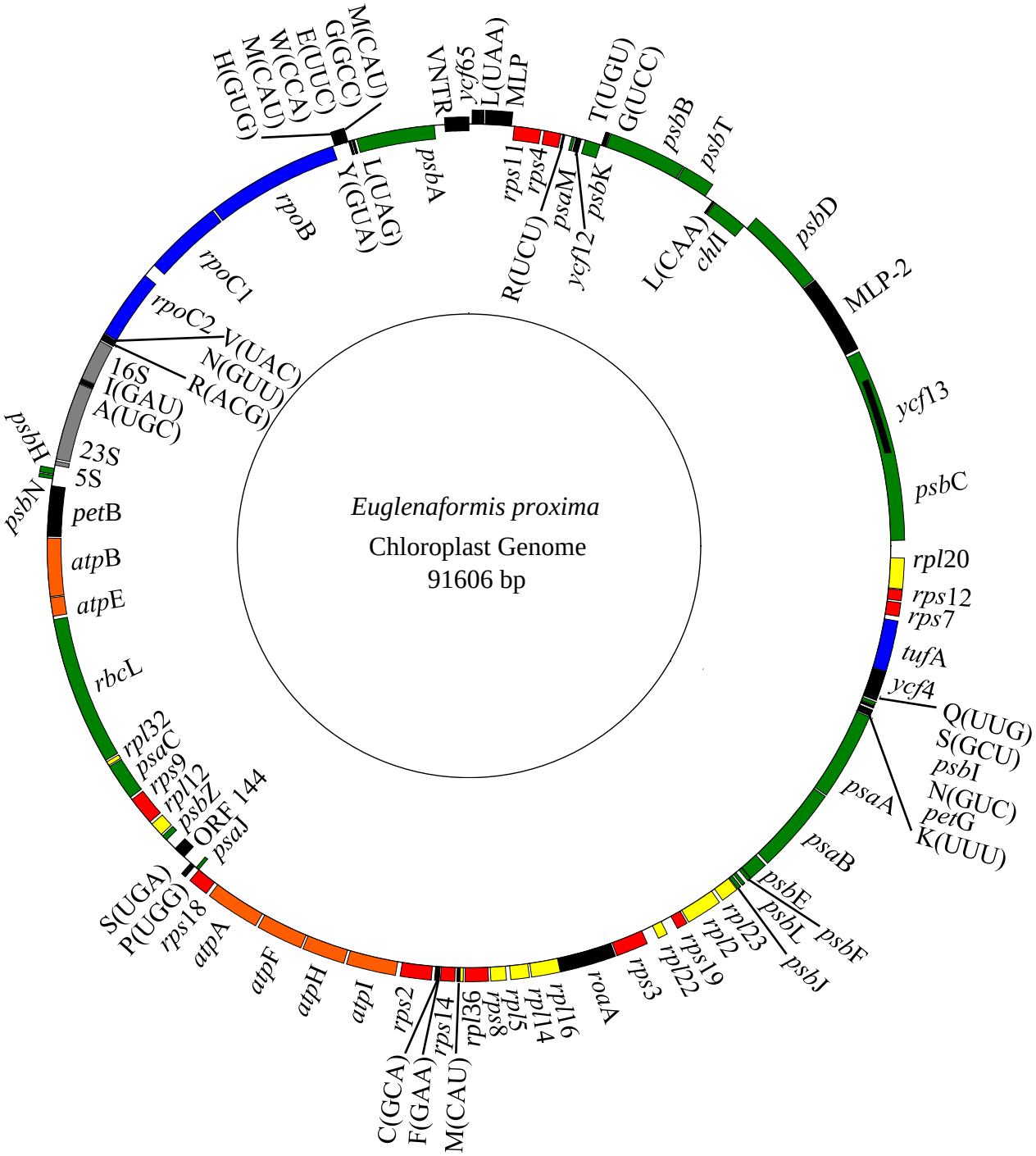

Figure 2.1. Gene map of the *Euglenaformis proxima* chloroplast genome. The box colors are common for genes of similar functional groups. Green: photosystems/photosynthesis genes; yellow: large ribosomal proteins, rpl genes; red: small ribosomal proteins, rps genes; blue: transcription/translation related genes, rpo genes, tufA; orange: atp genes; black: miscellaneous, conserved hypothetical proteins (ycf), open reading frames (ORF), tRNAs, maturase-like proteins (MLP), Variable Number Tandem Repeat (VNTR); grey: ribosomal rRNAs. Positions of the genes on the outside or inside of the outer ring are representative of the positive and negative strands, respectively. Block size for each gene is proportional to its sequence length.

The phylogenomic analyses (Figure 2.2, Figure 2.3) resulted in trees that were consistent with previously reported tree topologies (Milanowski *et al*. 2006, Kim *et al*. 2010, Linton *et al.* 2010). The position of *Euglenaformis* [*Euglena*] *proxima*, relative to the rest of the "crown" photosynthetic euglenoid taxa, was consistent and well supported (1.0 posterior probability (pp), Figure 2.2; 0.93 pp, Figure 2.3). The preponderance of evidence from the phylogenomic analyses, as well as evidence from past phylogenetic analyses, has shown that *E. proxima* should not be considered as a member of the genus *Euglena* and should be transferred to its own genus.

Taxonomic Revision:

Euglenaformis M. S. Bennett *et* Triemer, *gen. nov*.

Diagnosis:

Cells free-living, solitary, with one emergent flagellum when swimming; spindle-shaped, narrowing to the posterior and tapering into a pointed tail-piece; cells metabolic with flexible pellicle, displaying euglenoid movement; discoid chloroplasts without pyrenoids.

Type species:

Euglenaformis proxima (Dangeard) M. S. Bennett *et* Triemer, *comb. nov*.

Basionym:

Euglena proxima 1901. Dangeard, P.A. Recherches sur Les Eugléniens. *Le Botaniste* 8:154-157, Fig. 6 A-F.

Lectotype:

Figure 6A, Dangeard, P.A. 1901. Recherches sur Les Eugléniens. *Le Botaniste* 8:154-157.

Cultures representing the Lectotype:

SAG 1224-11a, SAG 1224-11b.

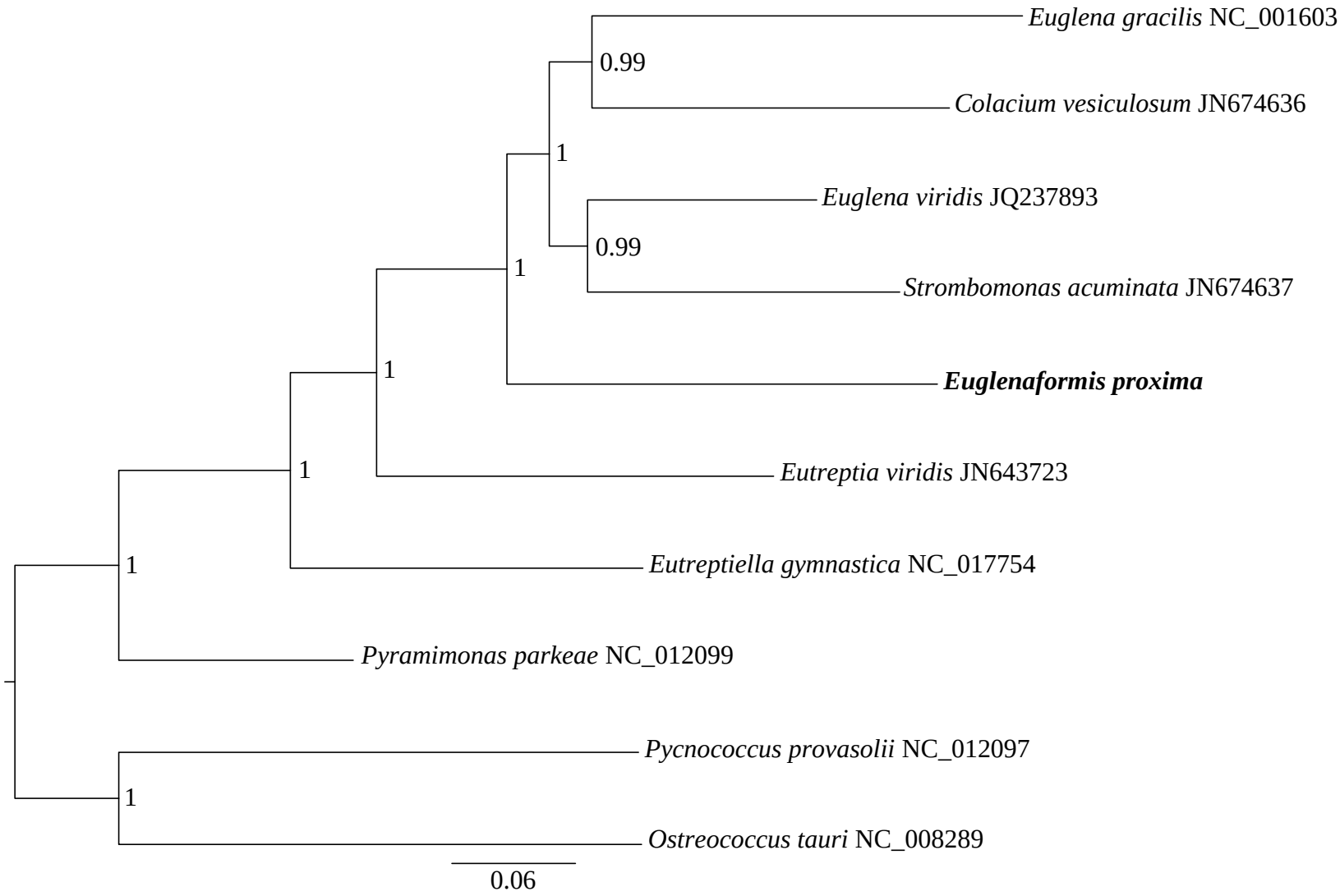

Figure 2.2. Nucleotide phylogenomic tree based on 34,230 sites, which were partitioned into 3 datasets (rRNA genes, tRNAs, and protein-coding genes), and all datasets were analyzed as nucleotide sequences. The numbers on the nodes are the Bayesian posterior probability (pp) values and the GenBank accession number for the chloroplast genome follows each taxon name. The tree was posteriorly rooted with *Pycnococcus provasolii* and *Ostreococcus tauri* and the scale bar represents the number of substitutions/site.

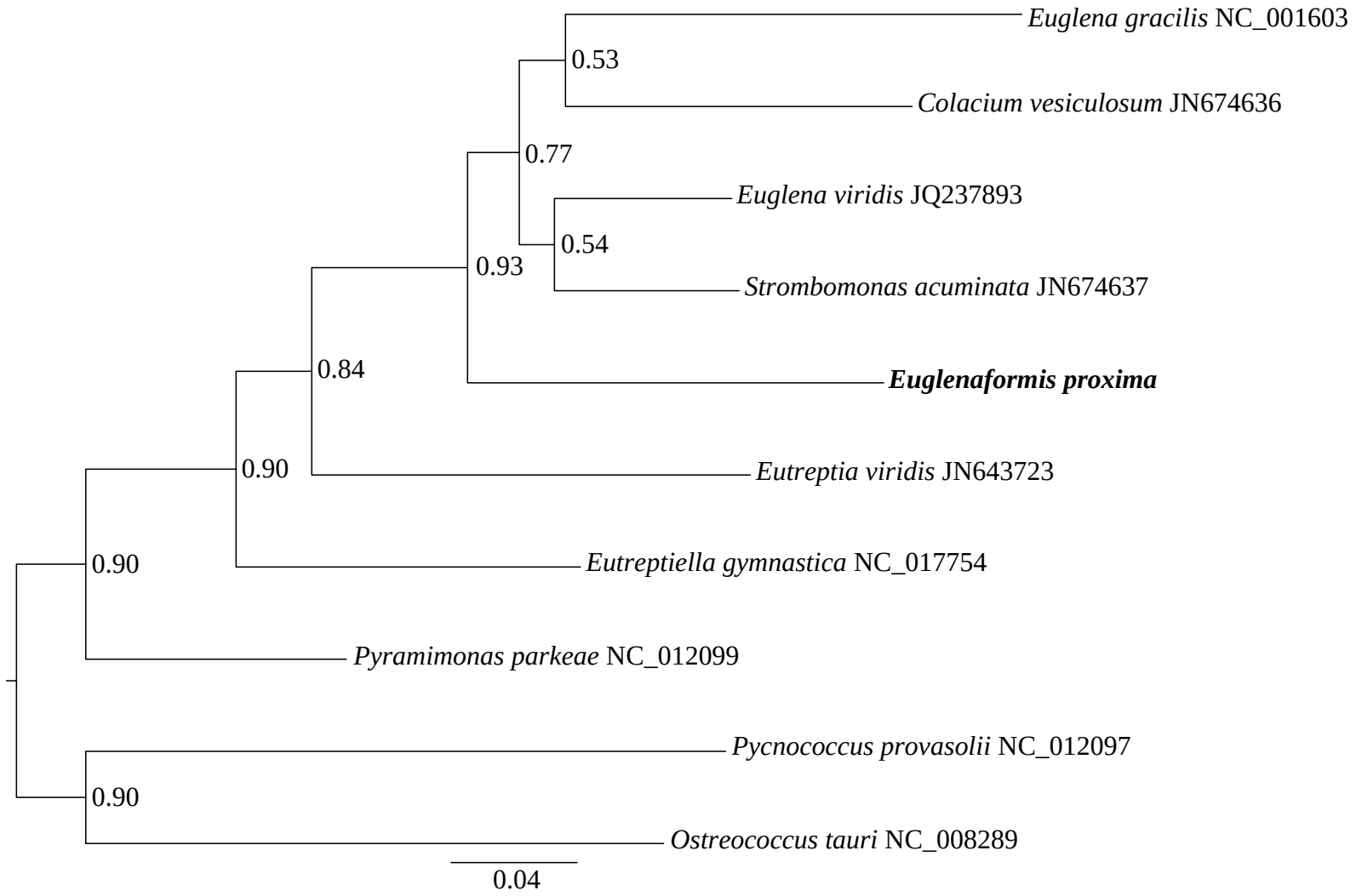

Figure 2.3. Mixed-character phylogenomic tree based on 14,589 sites, which were partitioned into 3 datasets (rRNA genes, tRNAs, and protein-coding genes). The protein-coding genes were analyzed as amino acid sequences and the rRNA genes and tRNAs were analyzed as nucleotide sequences. The numbers on the nodes are the Bayesian posterior probability (pp) values and the GenBank accession number for the chloroplast genome follows each taxon name. The tree was posteriorly rooted with *Pycnococcus provasolii* and *Ostreococcus tauri* and the scale bar represents the number of substitutions/site.

The name *Euglenaformis* was derived from the generic name *Euglena* and the Latin suffix ‘‘-formis’’, meaning “form, likeness, shape.” The name was a reference to the fact that *Euglenaformis proxima* was morphologically indistinguishable from taxa in either the genus *Euglena* or the genus *Discoplastis*, which was also comprised of former *Euglena* taxa. However, *Efs. proxima* was genetically distinct from other photosynthetic euglenoid taxa based on both phylogenetic and phylogenomic analyses. In addition, three molecular markers in the 16S rRNA gene allowed *Efs. proxima* to be easily distinguished from all other photosynthetic euglenoid taxa. The first marker was located in bases 4-6 of helix 25’, where *Efs. proxima* had a sequence of ‘TTT.’ Almost all other taxa had the sequence ‘GTA’, no other sequenced taxon contained a ‘T’ in the fourth position, and only one other taxon, *Phacus salina*, contained a ‘T’ in the sixth position. The second marker was located in a series of 7 bases between helix 36 and 37, where *Efs. proxima* had a sequence of ‘GAGATAT.’ Most other taxa had the sequence ‘TTGACAT,’ and no other sequence was similar to that of *Efs. proxima* in that region. The third marker was located in a series of 4 bases between helix 38’ and 36’, where *Efs. proxima* had a sequence of ‘TTCG.’ Most all other taxa had the sequence ‘TTAT,’ and no other sequenced taxa had either a ‘C’ in the third position or a ‘G’ in the fourth position.

A syntenic comparison between *Efs. proxima*, *Eutreptia virids*, and *Euglena gracilis* is shown in Figure 2.4. This combination of photosynthetic euglenoid taxa showed the gene rearrangements that have taken place when “basal” and “crown” taxa were compared to *Efs. proxima* (as informed by the topology of the phylogenomic trees in Figures 2.2 & 2.3). The comparison resulted in 21 blocks of genes that are shared between the 3 taxa (A-U, Figure 2.4; Table 2.2).

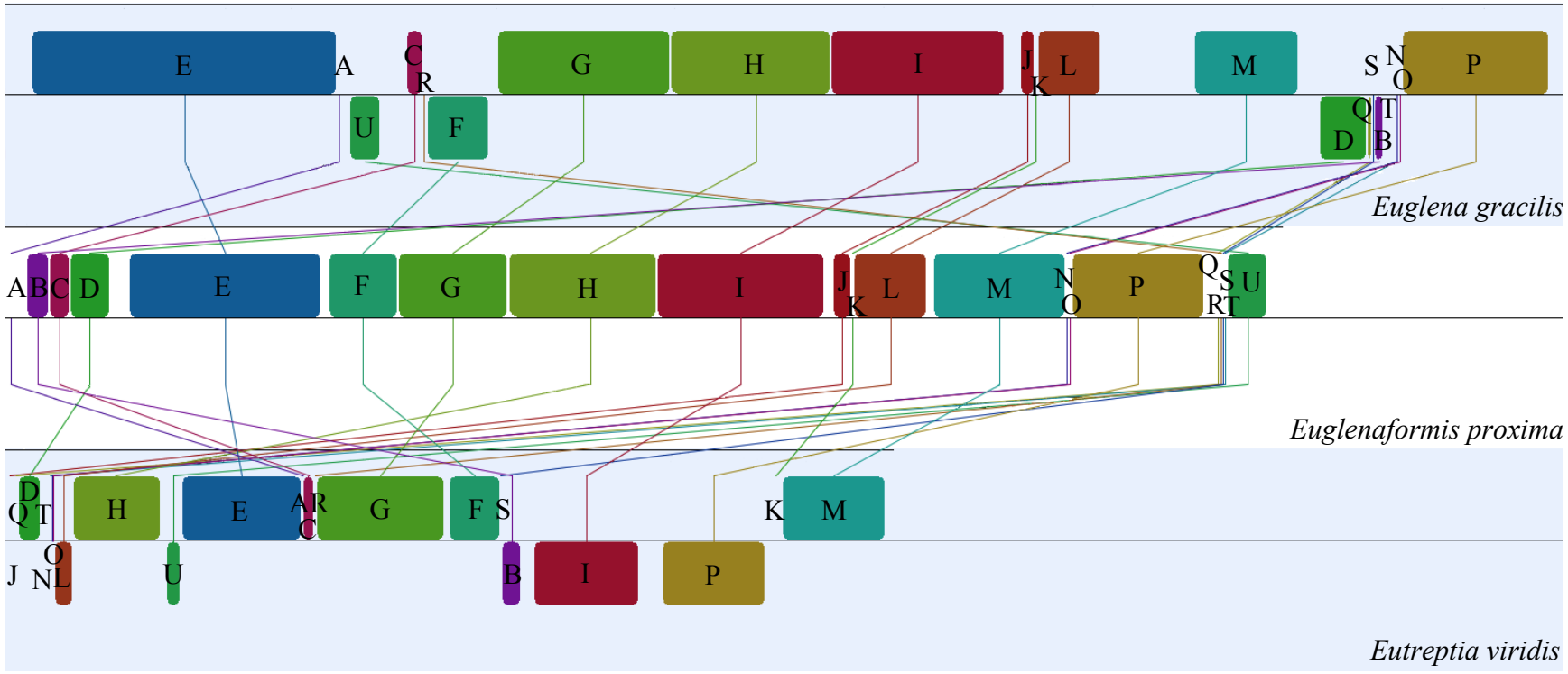

Figure 2.4. A Mauve analysis comparing the chloroplast genomes of *Euglena gracilis*, *Euglenaformis proxima*, and *Eutreptia viridis*. The linearized cpGenomes are shown with colored boxes that represent homologous gene clusters, and these boxes are lettered A-U to help identify the boxes between the genomes. The positions of the boxes are relative to the *Efs. proxima* cpGenome; consequently, boxes that lie above the horizontal line represent gene clusters that are oriented in the same direction as *Efs. proxima*, while boxes that lie below the horizontal line represent gene clusters that are oriented in the opposite direction of *Efs. proxima*. A list of the genes contained in each gene cluster is outlined in Table 2.2.

Table 2.2: Gene clusters identified in Mauve analysis of three sequenced euglenoid cpGenomes. Gene clusters were identified with letters for more clarity, and the genes contained within them are listed in the order they appear in *Euglenaformis proxima*. Only those genes that exist in at least two of the genomes are listed.

Cluster	Conserved genes present in *Euglenaformis proxima*
A	L(UAA)
B	*rps*11, *rps*4
C	R(UCU), *psa*M, *ycf*12, *psb*K, T(UGU)
D	G(UCC), *psb*B, *psb*T
E	L(CAA), *chl*I, *psb*D, MLP-2, *psb*C, *ycf*13
F	*rpl*20, *rps*12, *rps*7, *tuf*A, *ycf*4, Q(UUG), S(GCU)
G	*psb*I, N(GUC), *pet*G, K(UUU), *psa*A, *psa*B, *psb*E, *psb*F, *psb*L, *psb*J
H	*rpl*23, *rpl*2, *rps*19, *rpl*22, *rps*3, *roa*A, *rpl*16, *rpl*14, *rpl*5, *rps*8, *rpl*36, M(CAU), *rps*14
I	F(GAA), C(GCA), *rps*2, *atp*I, *atp*H, *atp*F, *atp*A, *rps*18, *psa*J, P(UGG), S(UGA), *psb*Z, *rpl*12, *rps*9
J	*psa*C
K	*rpl*32
L	*rbc*L
M*	*atp*E, *atp*B, *pet*B, *psb*N, *psb*H, 5S, 23S, A(UGC), I(GAU), 16S
N	R(ACG)
O	N(GUU), V(UAC)
P	*rpo*C2, *rpo*C1, *rpo*B
Q	H(GUG), M(CAU), W(CCA), E(UUC)

Table 2.2 (cont'd)

R	G(GCC), M(CAU)
S	Y(GUA)
T	L(UAG)
U	*psb*A

* = Progressive Mauve had difficulty correctly parsing the genes present in cluster M: *psb*N and *psb*H occur between clusters D and T in *Eutreptia viridis*; 5S, 23S, A(UGC), I(GAU), and 16S occur after cluster P in *Euglena gracilis*.

Discussion

The size of the *Efs. proxima* cpGenome (at least 94,185 bp) is comparable to the cpGenomes of previously sequenced photosynthetic euglenoids, which range in size from 65,513 bp in *Eutreptia viridis* (Wiegert *et al*. 2012) to 144,167 bp in *Strombomonas acuminata* (Wiegert *et al*. 2013). Unfortunately, we are unable to determine the full length of the cpGenome sequence in *Efs. proxima* due to the size of it's VNTR region. The existence of a VNTR sequence is not uncommon in photosynthetic euglenoids, and is present in the cpGenomes of *E. gracilis* (54 bp, Hallick *et al*. 1993), *Euglena viridis* (284 bp, Bennett *et al*. 2012), and *Eutreptiella gymnastica* (3x11 bp and 3.4x33 bp, Hrda *et al*. 2012). Based on PCR sequencing of the VNTR region, *Efs. proxima* has at least four 200 bp repeats, plus a partial repeat. However, due to the limitations of PCR, the total number of repeats can not be determined because 800 - 900 bp is the maximum for Sanger sequencing to double coverage with a single set of primers. Long-range PCR over the region shows at least four copies of the repeat region, although the gel bands become more faint as the size increases making it difficult to determine if additional copies exist.

The presence, or number, of repeats of the ribosomal operon varies greatly between the cpGenomes of photosynthetic euglenoids. However, previous research shows that this character

is not consistent, even within a single taxon (see Bennett *et al*. 2012 for a discussion on this topic). Following sequencing and assembly, only one copy of the ribosomal operon appears to be present in the cpGenome of *Efs. proxima*. Because the assembly of reads created by Next-Generation sequencing can mask the presence of a tandem repeat sequence, we performed a Long-Range PCR to determine the number of ribosomal operons in the cpGenome. The PCR amplified region stretches from 376 bp into the 16S rRNA to the next annotated gene, tRNA R(ACG) (Figure 2.1), a region of 485 bp if a single copy of the ribosomal operon is present. The ribosomal operon of *Efs. proxima* totals 4,707 bp, so if multiple copies of the ribosomal operon are present, gel bands should be seen (minimally) at 485 bp and 5,192 bp. The results from this Long-Range PCR show that only one gel band is present, at a position slightly less than 500 bp. The number of genes present, and the gene content of *Efs. proxima*, is similar to that seen in other photosynthetic euglenoid cpGenomes. In fact, the cpGenome is only missing two genes that are annotated in other photosynthetic euglenoid cpGenomes, *psa*I and *rpo*A. The absence of these two genes is not unprecedented, with *Euglena gracilis* also missing both genes (Hallick *et al*. 1993), *C. vesiculosum* missing *rpo*A (Wiegert *et al*. 2013), and *Eutreptiella gymnastica* missing *psa*I (Hrda *et al*. 2012). Interestingly, ORF 144 (Figure 2.1) occurs in a position that is occupied by *rpo*A in *Euglena viridis* (Bennett *et al*. 2012) and *S. acuminata* (Wiegert *et al*. 2013). However, when a BLASTx analysis is performed on the ORF, it returns only a single non-significant match (e-value 4.1) to a putative protein in human body louse (*Pediculus humanus corporis*).

The longest gene in the cpGenome of *Efs. proxima*, *psb*C, is annotated with the first portion of the gene missing. When this gene sequence is aligned against the gene sequences of other photosynthetic euglenoids and green algae (see Materials and Methods), the portion of the gene

sequence that we have labeled “exon 2” is the first portion of the gene that can be readily aligned. Despite an extensive search from the beginning of gene *psb*D to the beginning of *psb*C “exon 2”, we are unable to locate the fairly-conserved 5’ portion of this gene that is present in all other taxa. Due to the fact that this gene is essential for photosynthesis to occur, we are not comfortable identifying the gene as a pseudogene, and further analysis will need to be performed in order to determine the 5’ portion of the gene.

Phylogenomic Analyses

The two phylogenomic analyses (Figure 2.2, Figure 2.3) result in trees with the exact same topologies, and differ only in the support given to each node. There are two major reasons that can account for the differences between the nodal support numbers. The first is that the Bayesian analysis with the protein-coding genes analyzed as amino acid sequences (amino acid analysis) (Figure 2.3) has only 42.6% of the sites available for analysis that the Bayesian analysis with the protein-coding genes analyzed as nucleotides (nucleotide analysis) (Figure 2.2) has (14,589 sites vs. 34,230 sites). The second, and more significant reason, is that the synonymous substitutions in the amino acid sequence mask the underlying genetic variability present in the nucleotide sequence, which further reduces the number of phylogenetically informative sites. For example, there are cases where a single Leucine site in the amino acid sequence actually represents all 6 of the nucleotide codon variants, depending on the specific taxon (personal observation). Despite the inability of the amino acid analysis (Figure 2.3) to resolve relationships among the “crown” photosynthetic euglenoids, it still separates those taxa from *Efs. proxima*, with good support (0.93 pp). In fact, the separation of *Efs. proxima* from the “crown” photosynthetic taxa is the best-supported node in the amino acid analysis. This analysis, as well as the strong support from the nucleotide analysis (1.0 pp, Figure 2.2) and strong support from previous research that shows

the same relationship (Milanowski *et al*. 2006, Kim *et al*. 2010, Linton *et al.* 2010), leads us to the removal of *Euglena proxima* from the genus *Euglena* and to the formation of the new genus *Euglenaformis* for the taxon.

Synteny

As can be seen in the Mauve analysis (Figure 2.4, Table 2.2), there is little consistency in the arrangement of the 21 gene clusters, and it appears that the cpGenomes have been rearranged indiscriminately over evolutionary time. As more euglenoid cpGenomes are published (Hallick *et al*. 1993, Bennett *et al*. 2012, Hrda *et al*. 2012, Wiegert *et al*. 2012, Wiegert *et al.* 2013), it is becoming increasingly apparent that massive gene rearrangements have occurred in the cpGenomes of these taxa. It will be interesting to determine if, as even more cpGenomes are published, a pattern of gene rearrangement emerges within genera or phylogenetic clades, or if gene clusters ("operons") will consistently move together, even if the overall arrangement of these clusters varies between taxa.

Taxonomy

Dangeard (1901) discusses the phenotypic similarities of *Euglena proxima* (= *Efs. proxima*) and *Euglena variabilis* G.A. Klebs in the manuscript first describing the species *Euglena proxima*. Based on that description we believe it is possible that the taxon *Euglena variablis* also belongs in the newly erected genus *Euglenaformis*. However, cultures of *Euglena variablis* are not available, and we have never seen this taxon in field samples, so we are hesitant to formally move this species without being able to confirm the presence of the genetic synapomorphies for *Euglenaformis*.

In his initial description of *Euglena proxima*, Dangeard (1901) intimates that he is not completely confident in assigning the newly described taxon to the genus *Euglena* due to the presence of discoid chloroplasts and the lack of pyrenoids. Over 110 years later, through modern phylogenetic and phylogenomic techniques, we have now demonstrated that his inklings regarding the generic assignment of this taxon were correct, and that this species should be considered as a completely separate genus.

REFERENCES

REFERENCES

ALVES-DA-SILVA S.M. & MENEZES M. 2010. Eugleophyceae. In: Catálogo de plantas e fungos do Brasil. Vol. 1. (Forzza, R.C. Eds), pp. 383-404. Rio de Janeiro: Andrea Jakobsson Estúdio; Instituto de Pesquisas Jardim Botânico do Rio de Janeiro.

BENNETT M.S., WIEGERT K.E. & TRIEMER R.E. 2012. Comparative chloroplast genomics between *Euglena viridis* and *Euglena gracilis* (Euglenophyta). *Phycologia* 51: 711–718.

CAMACHO C., COULOURIS G., AVAGYAN V., MA N., PAPADOPOULOS J., BEALER K. & MADDEN T.L. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10: 421.

CIUGULEA I. & TRIEMER R.E. 2010. *A color atlas of photosynthetic euglenoids*. Michigan State University Press, East Lansing. 204 pp.

DANGEARD P.A. 1901. Recherches sur Les Eugléniens. *Le Botaniste* 8: 97-370.

DARLING A.C.E., MAU B., BLATTER F.R. & PERNA N.T. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research* 14: 1394–1403.

DARRIBA D., TABOADA G.L., DOALLO R. & POSADA D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* 9: 772.

EDGAR R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32:1732–1797.

GIBBS S. P. 1978. The chloroplasts of Euglena may have evolved from symbiotic green algae. *Canadian Journal of Botany* 56: 2883–2889.

GIBBS S. P. 1981. The chloroplasts of some algal groups may have evolved from endosymbiotic eukaryotic algae. *Annals of the New York Academy of Sciences* 361: 193–208.

GOCKEL G. & HATCHEL W. 2000. Complete gene map of the plastid genome of the nonphotosynthetic euglenoid flagellate *Astasia longa*. *Protist* 151: 347–351.

GRIMES J.A. 1988. The Algae, in G.Scott (ed.), Lake Broadwater. The Natural History of an Inland Lake and its Environs, Darling Downs Institute Press (in Association with Lake Broadwater Natural History Association), pp. 105-133.

HALLICK R. B., HONG L., DRAGER R. G., FAVREAU M. R., MINFORT A., ORSAT B., SPIELMANN A. & STUTZ E. 1993. Complete sequence of *Euglena gracilis* chloroplast DNA. *Nucleic Acids Research* 21: 3537–3544.

HRDÁ Š, FOUSEK J., SZABOVÁ J., HAMPL V. & VLČEK Č. 2012. The plastid genome of *Eutreptiella* provides a window into the process of secondary endosymbiosis of plastid in euglenids. *PLoS ONE* 7: e33746. DOI:10.1371/journal.pone.0033746.

KIM J.I., SHIN W. & TRIEMER R.E. 2010. Multigene analyses of photosynthetic euglenoids and new family, Phacaceae (Euglenales). *Journal of Phycology* 46: 1278-1287.

KOLLER B. & DELIUS H. 1982. A chloroplast DNA of *Euglena gracilis* with five complete rRNA operons and two extra 16S rRNA genes. *Molecular and General Genetics* 188: 305–308.

LINTON E.W., KARNKOWSKA-ISHIKAWA A., KIM J.I., SHIN W., BENNETT M.S., KWIATOWSKI J., ZAKRYŚ B. & TRIEMER R.E. 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: 603-619.

MILANOWSKI R., KOSMALA S., ZAKRYS B. & KWIATOWSKI J. 2006. Phylogeny of photosynthetic euglenophytes based on combined chloroplast and cytoplasmic SSU rDNA sequence analysis. *Journal of Phycology* 42: 721-730.

MILNE I., BAYER M., CARDLE L., SHAW P., STEPHEN G.,WRIGHT F. & MARSHALL D. 2010. Tablet – next generation sequence assembly visualization. *Bioinformatics* 26: 401–402.

PHAM M.N., TAN H.T.W., MITROVIC S. & YEO H.H.T. 2011. A checklist of the algae of Singapore. pp. 1-100. Singpore: Raffles Museum of Biodiversity Research, National University of Singapore.

RAVEL-CHAPIUS P., FLAMANT F., NICOLAS P., HEIZMANN P. & NIGON V. 1984. Diversity of the ribosomal structures in the *Euglena gracilis* chloroplast genome: description of a mutant with two ribosomal operons and possible mechanism for its production. *Nucleic Acids Research* 12: 1039–1048.

RAWSON J.R.Y., KUSHNER S.R., VAPNEK D., KIRBY N., BOERMA A. & BOERMA C.L. 1978. Chloroplast ribosomal RNA genes in *Euglena gracilis* exist as three clustered tandem repeats. *Gene* 3: 191–209.

RONQUIST F., TESLENKO M., VAN DER MARK P., AYRES D.L., DARLING A., SEVASTIAN H., LARGET B., LIU L., SUCHARD M.A. & HUELSENBECK J.P. 2012. MRBAYES 3.2: Efficient Bayesian phylogenetic inference and model selection across a large model space. *Systematic Biology* 61: 539-542.

ROZEN S. & SKALETSKY H.J. 2000. Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ, PP 365-386.

SCHATTNER P., BROOKS A.N. & LOWE T.M. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Research* 33: W686–W689.

SHIN W. & TRIEMER R.E. 2004. Phylogenetic analysis of the genus *Euglena* (Euglenophyceae) with particular reference to the type species *Euglena viridis*. *Journal of Phycology* 40: 759–771.

SMITH T.E. 2010. Revised list of algae from Arkansas, U.S.A. and new additions. *International Journal on Algae* 12(3): 230-256.

TAMURA K., PETERSON D., PETERSON N., STECHER G., NEI M. & KUMAR S. 2011. MEGA5: Molecular Evolutionary Genetics Analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* 28: 2731–2739.

TRIEMER R.E., 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*). Journal of Phycology* 42: 731–740.

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. *Proceedings of the National Academy of Sciences of the United States of America* 96: 10248–10253.

TURMEL M., GAGNON M.-C., O'KELLY C.J., 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. *Molecular Biology and Evolution* 26: 631–648.

UNTERGASSER A., NIJVEEN H., RAO X., BISSELING T., GEURTS R. & LEUNISSEN J.A.M. 2007. Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Research* 35: W71–W74. DOI:10.1093/nar/gkm306.

WATANABE M. M. & HIROKI M. 1997. NIES – collection list of strains, ed. 5. National Institute for Environmental Studies, Tsukuba. 127 pp.

WIEGERT K.E., BENNETT M.S. & TRIEMER R.E. 2012. Evolution of the chloroplast genome in photosynthetic euglenoids: a comparison of Eutreptia viridis and Euglena gracilis (Euglenophyta). *Protist* 163: 832-843.

WIEGERT K.E., BENNETT M.S. & TRIEMER R.E. 2013. Tracing patterns of chloroplast evolution in euglenoids: Contributions from *Colacium vesiculosum* and *Strombomonas acuminata* (Euglenophyta). *Journal of Eukaryotic Microbiology* (in press). doi:10.1111/jeu.12025

WURTZ E. A. & BUETOW D. E. 1981. Intraspecific variation in the structural organization and redundancy of chloroplast ribosomal DNA cistrons in *Euglena gracilis*. Current Genetics 3: 181–187.

WYMAN S.K., JANSEN R.K. & BOORE J.L. 2004. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20: 3252–3255.