# The genome of *Euglena gracilis*: Annotation, function and expression

ThankGod Echezona Ebenezer

Churchill College

Department of Biochemistry

University of Cambridge

This dissertation is submitted for the degree of Doctor of Philosophy

May 2018

**Full name:** ThankGod Echezona Ebenezer

**Title:** The genome of *Euglena gracilis*: Annotation, Function and Expression

## ABSTRACT

*Euglena gracilis* is a species of unicellular photosynthetic flagellate that inhibits aquatic ecosystems. *E. gracilis* belongs to the supergroup Excavata, and are an important component of the global biosphere, have biotechnological potential and is useful biological model due to their evolutionary history and complex biology. Whilst the evolutionary position of *E. gracilis* is now clear, their relationship with other protists such as *Naegleria*, *Giardia*, and Kinetoplastids, remains to be investigated in detail. Investigating and understanding the biology of this complex organism is a promising way to approach many evolutionary puzzles, including secondary endosymbiotic events and the evolution of parasitism, due to their relationship with Kinetoplastids. Here, I report a draft genome for *E. gracilis*, together with a high quality transcriptome and proteomic analysis. The estimated genome size is ~ 2 Gbp, with a GC content of ~ 50 % and a protein coding potential predicted at 36,526 Open Reading Frames (ORFs). Less than 25% of the genome is single copy sequence, indicating extensive repeat structure. There are evidences for large number of paralogs amongst specific gene families, indicating expansions and possible polyploidy as well as extensive sharing of genes with other non photosynthetic and photosynthetic eukaryotes: red and green algael genes, together with trypanosomes and other members of the excavates. Functional resolution into several of the biological systems indicates multiple similarities with the trypanosomatids in terms of orthology, paralogy, relatedness and complexity. Several biological systems such as nuclear architecture (e.g. chromosome segregation, nuclear pore complex, nuclear lamins), protein trafficking, translation, surface, consist of conserved and divergent components. For instance, several gene families likely associated with the cell surface and signal transduction possess very large numbers of lineage-specific paralogs, suggesting great flexibility in environmental monitoring and, together with divergent mechanisms for metabolic control, novel solutions to adaptation to extreme environments. I also demonstrate that the majority of control of protein expression levels is post-transcriptional and absence of transcriptional regulation, despite the presence of conventional introns. These data are a major advance in the understanding of the nuclear genome of Euglenids and provide a platform for investigation of the contributions of *E. gracilis* and relatives to the biosphere.

# DECLARATION

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

It does not exceed the prescribed word limit for the relevant Degree Committee – 'except as required post examination'.

ThankGod E. Ebenezer

May, 2018

# ACKNOWLEDGEMENT

I would like to thank the funding agencies whose funding made this piece of work possible. First, I would like to thank the Yousef Jameel Academic Program for sponsoring, in full, my PhD research program through the Yousef Jameel PhD Scholarship and whose funding made this piece of work possible. Second, I would like to thank the Cambridge Commonwealth, European and International Trust for administering these funds and for providing additional financial scholarship required for this work. Third, I would also like to thank the Cambridge Philosophical Society (CPS), the Cambridge University Student Registry, and the Medical Research Council (MRC Grant #: P009018/1 to MCF) for providing additional financial support for this work.

I would like to thank Prof. Mark C. Field at the University of Dundee, UK, Dr. Steven Kelly at the University of Oxford, UK, and my supervisor Prof. Mark Carrington for providing the supervision for this work. I would like to thank members of the research collaboration and annotation team (Principal Investigators and respective research students) for participating in the community annotation section of this work. This include: Prof. Julius Lukes (Czech Academy of Sciences, Czech Republic), Dr. Samuel Obado (The Rockefeller University, USA), Dr. Andrew Jackson (University of Liverpool, UK), Dr. Damien Devos (Universidad Pablo de Olavide, Spain), Dr. Joel B. Dacks (University of Alberta, Canada), Prof. Michael Lebert (University of Erlangen-Nuremberg, Germany) Prof. Peter Myler (Center for Infectious Disease Research, Seattle, USA), Dr. Susan Vaughan (Oxford Brookes University, UK), Dr. Vladimír Hampl (Charles University in Prague, Czech Republic), Dr. Ellis O'Neill (University of Oxford, UK), and Prof. Michael Ginger (University of Huddersfield, UK).

I would like to thank Prof. William Martin (Heinrich Heine University, Düsseldorf) for the kind gift of the *E. gracilis* strain (Z). I would like to thank Purification Gomez-Lopes and David Moriara for the donation of PacBio data and Prof. Rob Field for permission to reutilize transcriptome data. I would like to thank Dr. Jenny Barna who was formerly in the Department of Biochemistry (now with the University Information Services, University of Cambridge) for administering the servers (whale-data, elephant and

molerate) where several of the ‘heavy duty’ computational tasks were carried out. I would like to thank the Microscopy unit of the Department of Pathology, University of Cambridge, UK, for assisting in producing the first trial electron microscopy live cell imaging for this work which is now being used on the welcome page of the EuglenaDB website. I would like to thank Dr. Douglas Lamont and the Fingerprints proteomics facility at the University of Dundee, UK, for the label-free proteomics analysis. I would like to thank Dr. Martin Zoltner in the Field Laboratory, University of Dundee, UK, who I worked with in the light and dark analysis of the transcriptomics and proteomics experiment. I would like to thank members of the Field lab, Kelly lab, and Carrington lab for the support during the time this work was undertaken. Finally, I would like to thank my family for their moral support.

# ABSTRACT

*Euglena gracilis* is a species of unicellular photosynthetic flagellate that inhibits aquatic ecosystems. *E. gracilis* belongs to the supergroup Excavata, and are an important component of the global biosphere, have biotechnological potential and is useful biological model due to their evolutionary history and complex biology. Whilst the evolutionary position of *E. gracilis* is now clear, their relationship with other protists such as *Naegleria*, *Giardia*, and Kinetoplastids, remains to be investigated in detail. Investigating and understanding the biology of this complex organism is a promising way to approach many evolutionary puzzles, including secondary endosymbiotic events and the evolution of parasitism, due to their relationship with Kinetoplastids. Here, I report a draft genome for *E. gracilis*, together with a high quality transcriptome and proteomic analysis. The estimated genome size is ~ 2 Gbp, with a GC content of ~ 50 % and a protein coding potential predicted at 36,526 Open Reading Frames (ORFs). Less than 25% of the genome is single copy sequence, indicating extensive repeat structure. There are evidences for large number of paralogs amongst specific gene families, indicating expansions and possible polyploidy as well as extensive sharing of genes with other non photosynthetic and photosynthetic eukaryotes: red and green algael genes, together with trypanosomes and other members of the excavates. Functional resolution into several of the biological systems indicates multiple similarities with the trypanosomatids in terms of orthology, paralogy, relatedness and complexity. Several biological systems such as nuclear architecture (e.g. chromosome segregation, nuclear pore complex, nuclear lamins), protein trafficking, translation, surface, consist of conserved and divergent components. For instance, several gene families likely associated with the cell surface and signal transduction possess very large numbers of lineage-specific paralogs, suggesting great flexibility in environmental monitoring and, together with divergent mechanisms for metabolic control, novel solutions to adaptation to extreme environments. I also demonstrate that the majority of control of protein expression levels is post-transcriptional and absence of transcriptional regulation, despite the presence of conventional introns. These data are a major advance in the understanding of the nuclear genome of Euglenids and provide a platform for investigation of the contributions of *E. gracilis* and relatives to the biosphere.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF APPENDICES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| AP | Adaptin Proteins |
| AT | Adenine-Thiamine |
| BLAST | Basic Local Alignment Search Tool |
| CDD | Conserved Domain Database |
| cDNA | complementary DNA |
| CEG | Core Eukaryotic Genes |
| CEGMA | Core Eukaryotic Gene Mapping Approach |
| cpDNA | chloroplast DNA |
| DNA | Deoxyribonucelic acid |
| EBI-EMBL | European Bioinformatics Institute European Molecular Biology Laboratory |
| EC | Enzyme Code |
| EGT | Endosymbiotic Gene Transfer |
| ER | Endoplasmic Recticulum |
| EST | Expressed Sequence Tag |
| EuPath | Eukaryotic Pathogen Database Resources Center |
| FDR | False Discovery Rate |
| GC | Guanine-Cytosine |
| gDNA | genomic DNA |
| GMST | GeneMarkS-T |
| GO | Gene Ontology |
| gRNA | guide RNA |
| HMM | Hidden Markov Model |
| INDELS | Insertions/deletions |
| JGI | Joint Genome Institute |
| kDNA | kinetoplast DNA |
| KEGG | Kyoto Encyclopedia of Genes and Genomes |
| KKIP | KKT-interacting protein |
| KKT | Kinetoplastid Kinetochore |
| LECA | Last Eukaryotic Common Ancestor |

| | |
|---|---|
| LGT | Lateral Gene Transfer |
| DO | Dubious ORF |
| LHC | Light Harvesting Complexes |
| LSU | Large Sub Unit |
| mRNA | Messenger RNA |
| MSA | Multiple Sequence Alignment |
| miDNA | mitochondrial DNA |
| NCBI | National Center for Biotechnology Information |
| NGS | Next Generation Sequencing |
| NPC | Nuclear Pore Complex |
| ORF | Open Reading Frame |
| PIC | Pre-initiation complex |
| QC | Quality Control |
| QUAST | Quality Assessment Tool for Genome Assemblies |
| RBP | RNA binding protein |
| rDNA | Ribosomal DNA |
| RNA | Ribonucleic acid |
| RNAseq | RNA sequencing |
| RNP | Ribonucleoprotein |
| RPS-BLAST | Reverse Position Specific BLAST |
| RRM | RNA recognition motif |
| rRNA | Ribosomal ribonucleic acid |
| SDA-PAGE | Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis |
| SGA | String Graph Assembler |
| SL | Spliced Leader |
| snoRNA | Small nucelolar RNA |
| SNPs | Single Nucleotide Polymorphism |
| SSPACE | SAKE-based Scaffolding of Pre-Assembled Contigs after Extension |
| SSU | Small Sub Unit |
| TEM | Transmission Electron Microscope |
| TGN | Trans-Golgi Network |

| | |
|---|---|
| tRNA | transfer RNA |
| UTR | Untranslated region |

# CHAPTER ONE

## 1.0 INTRODUCTION

In this chapter, the biology of *Euglena gracilis* is discussed from a molecular, cellular, biochemical, genetic, and evolutionary point of view. The chapter begins with a brief on the background of the research, and concludes with specific aims and objectives. I have attempted to capture the generic and specific biology of *E. gracilis,* putting this in context and in comparison to other members of the Excavates, Plantae, and other supergroups within the eukryotic domain as well as relating literatures with this current work.

### 1.1 BACKGROUND AND CLASSIFICATION

*Euglena gracilis* is a species of unicellular photosynthetic flagellate that inhabits aquatic ecosystems (Buetow, 1982) (Figure 1.1). The Eugleniods were first discovered by Antony van Leeuwenhoek in 1674 who referred to them as animalcules because he observed them being of diverse colours, and exhibiting movements now known as metaboly (Dobell, 1932). They were further detailed by Ehrenberg in 1830 who established the genus to accommodate those euglenoid organisms that have eyespots. *E. gracilis*, and members of its genus, may represent one of the earliest derived eukaryotic cells. Figure 1.2 presents the consensus eukaryotic and Euglena classifications respectively (for a detailed review see Keeling, *et al.,* 2005, and Adl, *et al.,* 2012, pages 479 - 488). This classification shows the emergence of the kingdom Excavata, superphylum Discoba, and subphylum Euglenozoa – which contains *E. gracilis* and other members of the *Euglena* genus (Euglenids), Kinetoplastids (e.g. Trypanosoma and Leishmania), and Heterolobosea (Naegleria) (Cavalier-Smith, 2002; Simpson, 2003; Keeling, *et al.,* 2005; Adl, *et al.,* 2012).

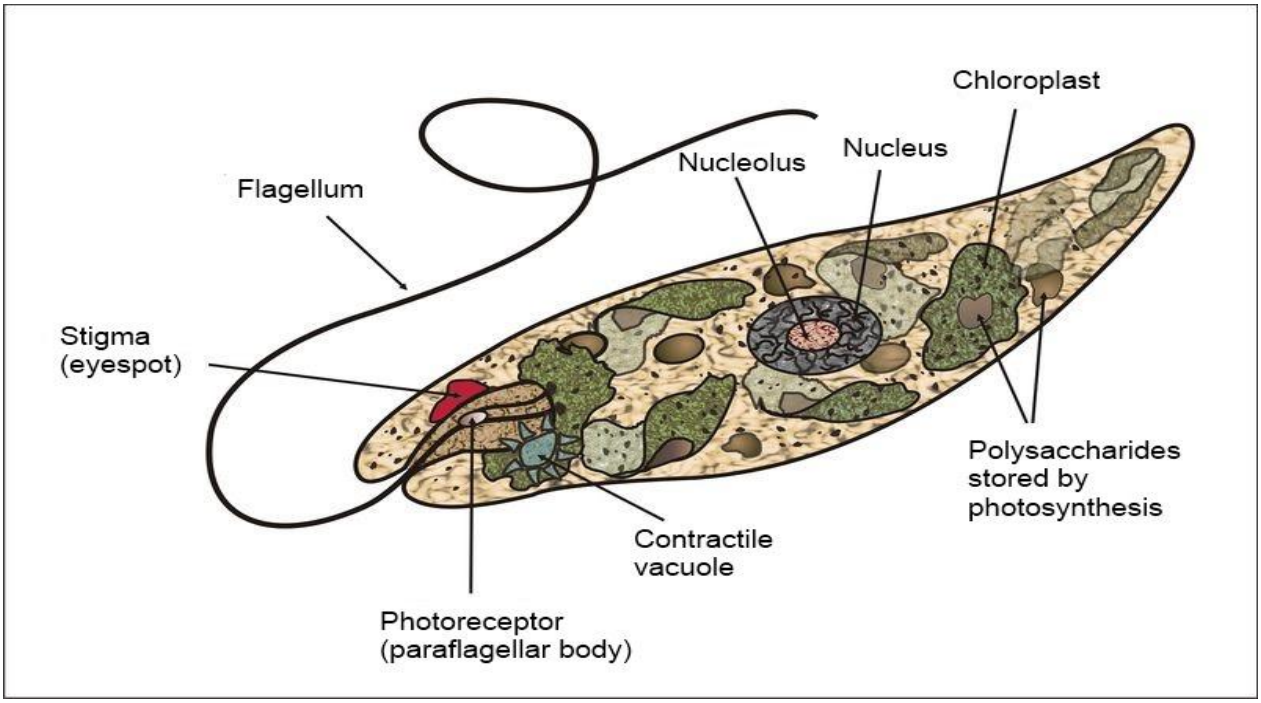



**Figure 1.1: *E. gracilis* is a unicellular secondary endosymbiotic green flagellate.** The diagram shows an *E. gracilis* cell with the organelles: Nucleus, chloroplast, polysaccharides granules, photoreceptor, contractile vacuole, stigma or eye spot, nucleolus, pellicle, and a flagellum. The chloroplast and mitochondria arose as a result of secondary endosymbiosis with a cyanobacteria and a proteobacteria respectively (http://en.wikipedia.org/wiki/Euglena).

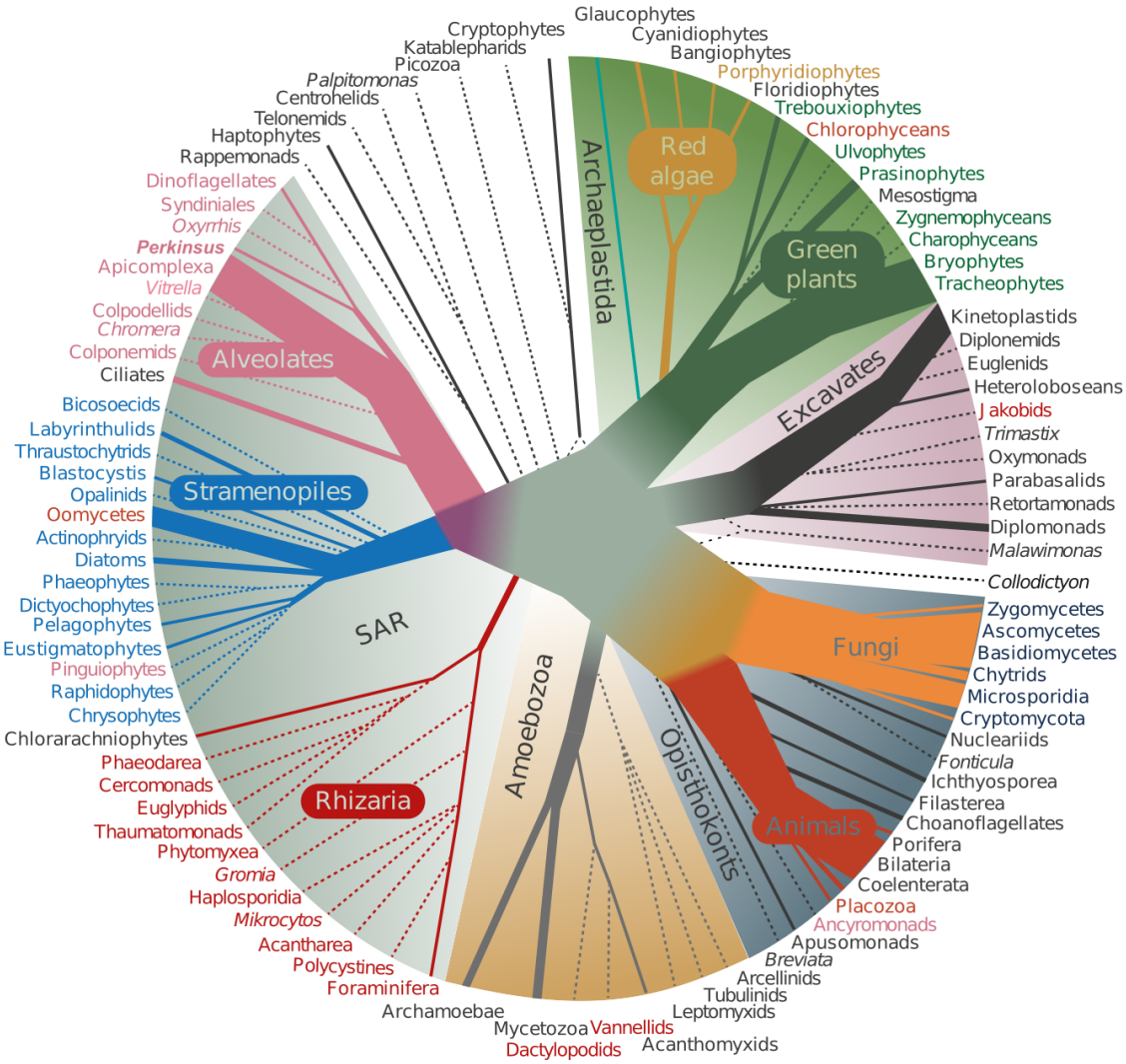



**Figure 1.2: Euglenids are early diverged members of the Euglenozoans and distant relatives to the Kinetoplastids.** A view of the evolutionary relationships among eukaryotes reflecting the classification presented herein. *E. gracilis* is grouped within the Kingdom, Excavata – in light purple colouration, in the Subphylum, Discoba, and distantly related to the members of the Kinetoplastids and Heterobolosean (adapted from Burki and Keeling, 2014). Other eukaryotic groups contained herein include: Archaeplastidae, SAR, Amoebozoa, and Opisthokonts. For extensive diagrammatic annotation see Burki and Keeling, 2014.

There are well over 250 species of the genus *Euglena*, with 20 taxa being predominantly cosmopolitan, i.e. present in many habitats (Kim, *et al.,* 1998; Gojdics, 1953; Zakrys, 1986; Zakrys and Walne, 1994). The cosmopolitan groups includes, amongst others, *E. viridis, E. caudata, E. oblonga, E. obtusa, E. deses, and E. gracilis* (Kim, *et al*., 1998). Of these, the most widely studied species is *E. gracilis,* which has been a useful model organism for biologists interested in understanding the biochemistry and molecular biology of living organisms due to their evolutionary history or complex biology (dos Santos Ferreira, 2007). The biological complexity of *E. gracilis* arose from secondary endosymbiosis (see sub section 1.2.2 and 1.2.3) which resulted in diverse mode of nutrition, possessing the ability to produce energy via photosynthetic and heterotrophic processes (dos Santos Ferreira, 2007). Endosymbiosis explains the origin of mitochondria and chloroplast as well as their physiological processes and functions. More interesting, are the clues they could provide to further understand the evolution of parasitism; delineating how a once free-living common ancestor of *E. gracilis* and kinetoplastids evolved into parasitic trypanosomes (see sub section 1.5).

The evolutionary position of *E. gracilis* has recently become clear, with the incorporation of new molecular and biochemical data (see Adl, *et al.,* 2012, for an extensive review). However, there are many aspects that remain puzzling, especially in their relationship with other protists such as *Naegleria, Giardia*, and Kinetoplastids (Tessier, 1997). According to Linton, *et al.,* 2010 and Adl, *et al.,* 2012, studies of euglenoid phylogenies have used only nuclear encoded genes. Most often this was the nuclear SSU rDNA (Busse and Preisfeld, 2002, 2003; Busse, *et al.*, 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) with one study (Talke and Preisfeld, 2002) using the nuclear encoded flagellar PAR1 and PAR2 genes, the results from which agreed with early nuclear SSU rDNA studies. Early molecular evidence noted that the genus, *Euglena*, is not monophyletic (Linton, *et al.,* 1999, 2000; Milanowski, *et al.,* 2001; Mullner, *et al.*, 2001), which led to several taxonomic revisions (Adl, *et al.,* 2012). An extensive recent study on the phylogeny of the genus, *Euglena*, suggests they are paraphyletic in comparison with *Naegleria* and Trypanosomes (see Linton, *et al.,* 2010; Adl, *et al.,* 2012 for an extensive review).

According to Gray, 1989, the *E. gracilis* photosynthetic machinery did not come directly from the primary (prokaryotic/eukaryotic) endosymbiotic event that gave rise to the chloroplasts of higher plants, as previously proposed. The host cell of the secondary endosymbiont (in which a eukaryote engulfs another eukaryote) (Palmer and Delwiche, 1996) seems to have been a protozoan which is believed to have evolved in parallel with trypanosomatids. The evolutionary relationship between *E. gracilis* and *Trypanosoma* is supported by morphological considerations (Kivic and Walne 1984), by nuclear rRNA alignments (Sogin, *et al.,* 1989), and by the addition of a leader sequence (spliced leader sequence SL) at the 5' end of the nuclear pre-messenger RNAs through a *trans*-splicing mechanism (Tessier, *et al.*, 1991) as well as the orthologous clustering analysis which is reported in this work. Considering their relative position on the eukaryotic evolutionary tree, both organisms (E. gracilis and Trypanosoma) are considered to belong to one of the earliest mitochondrion-containing branches (Sogin, *et al.*, 1986; Schnare, *et al.,* 1990). Current evidence suggests that all mitochondria have evolved from prokaryotic endosymbiont(s) that originated from within the α-proteobacteria (Gray, 1992) (see section 1.2.3 for details).

Investigating and understanding the biology of this complex organism, *E. gracilis,* is a promising way to resolve several evolutionary puzzles, secondary endosymbiotic events, biological processes, as well as their theoretical and practical application in fully understanding the evolution of parasitism. The Next Generation Sequencing (NGS) era for genomes and transcriptomes provides a unique opportunity to address these issues directly; provide the information to understand genome architecture and full coding potential for this organism and to infer functions such as RNA processing, plastid evolution, cytoskeletal mechanics, mitochondria, surface proteins, membrane protein targeting, and metabolome, as will be discussed in the subsequent chapters.

The interplay between genes that are specific to *E. gracilis*, genes more general to the kinetoplastids, genes more general to other eukaryotes, and those genes acquired during secondary endosymbiosis (which are plastid associated but nuclear encoded) are unknown. Trypanosomatids are as good a model group for tracing the evolution of parasitism as is available (Lukes, *et al.,* 2014), and by analyzing the genomic information from free-living bodonid and euglenid relatives and early-

branching trypanosomatids, e.g. *P. confusum* (Flegontov, *et al.,* 2013), can provide insight into the genetic changes associated with accession of parasitism as well as a comparison to free-living but non-photosynthetic bodonids (Lukes, *et al.,* 2014).

The genome of *E. gracilis* is comprised of three elements: nuclear, chloroplast and mitochondria genomes (Figure 1.3). The *E. gracilis* chloroplast and mitochondria genomes have been sequenced (Hallick, *et al.,* 1993; Perez, *et al.,* 2014; Dobakova, *et al.,* 2015; Faktorova, *et al.,* 2016), however, little is known about the nuclear genome (Pavel and Lukes, 2012). The only available nuclear DNA sequence organization information is that described by Rawson, *et al.,* 1979, using a combination of techniques, as well as isolation and sequencing of an ORF coding for the elongation factor protein EF-1 alpha by Montandon and Stutz, 1990, and alpha- and beta-tubulin cDNA *Euglena* lambda gt11 expression library sequencing and/or hybridization to *Euglena* RNA and DNA (alpha- and beta-tubulin cDNA) by Levvasseur, et al., 1994. To highlight many of these questions in more detail and to place in context with current understanding, this introductory chapter is divided into five subsections, each covering a distinct aspect of Euglena biology. A final section details the specific questions being addressed.

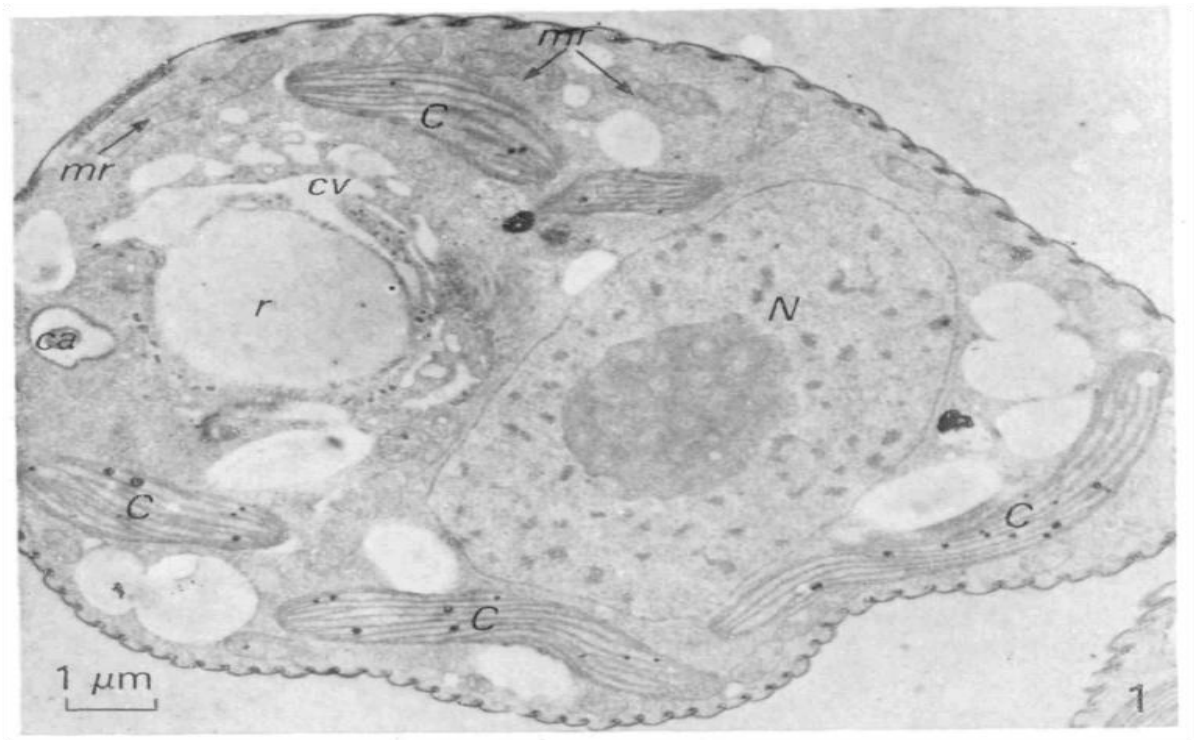



**Figure 1.3: *E. gracilis* is a hybrid of three genomes: Nuclear, chloroplast, and mitochondria.** The diagram shows the median sections of *E. gracilis* cells with spindle-shaped chloroplasts, a complex lobed one (C) and numerous mitochondrial profiles (mr). *ca,* canal; *cv,* contractile vacuole; *N,* nucleus; *r,* reservoir. Thiéry test,x 9000 (Pellegrini, 1980).

## 1.2 GENOMIC STRUCTURES, EVOLUTION, AND FUNCTION

### *1.2.1 Nucleus*

#### *1.2.1.1 Chromatin*

*E. gracilis* chromosomes appear heterochromatic, i.e. uniformly condensed, throughout the cell cycle (Falchuk, *et al.,* 1975), while an uncondensed state (Ebenezer, *et al.,* manuscript in preparation) has recently been reported as discussed in chapter 5 and appendices. Other Euglenidae and dinoflagellates also possess continuously condensed chromosomes. Interphase chromatin of higher eukaryotes has been fractionated into dispersed (euchromatin) and condensed (heterochromatin) fractions (Sirlin, 1972; Johmann, *et al.,* 1973). *E. gracilis* chromosomes always appear condensed under microscopy, with a portion of chromatin in an uncondensed state, which is active in RNA synthesis and likely the nucleolus (Falchuk, *et al.,* 1975). The euchromatin fraction contains only 14 % of the nuclear DNA, but shows a ten-fold increase in RNA polymerase activity compared to the heterochromatin fraction. It remains to be determined what advantage might accrue to *E. gracilis* in maintaining over 80 % of its chromatin in a condensed state, even in the period of interphase of the cell cycle, as well as what factors are responsible for condensation of chromatin (Lynch, *et al.,* 1974). Histones were suggested to be such factors (Mirskv, *et al.,* 1968), but little difference has been noted in histone content between condensed metaphase chromosomes and diffuse interphase chromatin (Comings, 1967). The modification of specific histone fractions may be instrumental in inducing chromosome condensation. The treatment of heterochromatin with trypsin causes an increased template activity for the chromatin (Lynch, *et al.,* 1974). Certain proteins do appear to mediate the degree of condensation of chromatin, but the precise mechanism involved is still open to debate, but the condensation of its chromatin could be the final result of a change in ionic content or composition of the nucleus (Lynch, *et al.,* 1974). Recent work in trypanosomes however suggests that the modification of histones and the likely mechanisms of heterochromatinisation are broadly similar to higher eukaryotes (Jack and Hake, 2014).

#### *1.2.1.2 Nuclear genome structure*

Preliminary investigations into nuclear genomic DNA of *E. gracilis* have been described by James, *et al.,* 1979. They studied the sequence organization of nuclear DNA using a combination of techniques involving: the comparison of the reassociation kinetics of DNA fragments 300 bp, 2000 bp and 8100 bp nucleotides long, the reassociation of $^{32}$P-labeled DNA fragments of various lengths with driver fragments 300 nucleotides long, the hyperchromicity of DNA structures formed by the reassociation of repetitive sequences, and the direct measurement of the size of the duplex regions of reassociated repetitive DNA resistant to S1 nuclease. Initial studies in characterizing the genome of *E. gracilis* by reassociation kinetics suggested that there were two kinetic components, and that only 36 % of the genome consisted of single copy DNA (Rawson, 1975) (this is similar to the 25 % reported in chapters 3 and 5 in this work). These single copy DNA sequences are approximately 1500 nucleotide pairs long and interspersed with repetitive DNA sequences. The repetitive DNA, consisting of both highly repetitive and middle repetitive sequences, comprises one fraction of nucleotide sequences (0.67) with an average size of 4900 nucleotide pairs and a second fraction (0.33) with an average size of 1000 nucleotide pairs. 34 % of the DNA consists of foldback sequences which are present on 45 % of the DNA 4000 nucleotides long. Following their investigation, other independent nuclear genome related investigation has also been carried out and includes: isolation and sequencing of an ORF coding for the elongation factor protein EF-1 alpha (Montandon and Stutz, 1990), and alpha- and beta-tubulin cDNA *Euglena* lambda gt11 expression library sequencing and/or hybridization to *Euglena* RNA and DNA (alpha- and beta-tubulin cDNA) by Levasseur, *et al.,* 1994. Overall this suggests a highly complex genome with extensive low complexity sequence.

*Primary transcripts and splice leaders*

A common euglenozoan feature is the processing of primary nuclear transcripts by spliced leader (SL) RNA-mediated trans-splicing (Frantz, *et al.,* 2000). This process includes replacement of the 5'-end of pre-mRNA by the 5'-end of SL-RNA, donating identical 5'-termini to the mRNA molecules. Similar to nuclear cis-splicing, trans-splicing process is also mediated by spliceosomes, but a Y-branch intron structure is formed instead of a lariat (Bonen, 1993). The only currently known nuclear mRNA

lacking the SL sequence in *E. gracilis* is that of the fibrillarin gene (Russel, *et al.,* 2005). Since SL-trans-splicing does not occur in organelles (mitochondria and plastids), the presence (or absence) of an SL sequence at the 5'-end of a euglenozoan mRNA is diagnostic for its synthesis in the nucleus.

*Microsatellite DNA*

Microsatellite DNA markers in a large range of eukaryotic organisms have proven to be a powerful tool in studies of population genetics, high-resolution genotyping, gene mapping, linkage analysis, diagnosis of parasites, phylogenetic relationships, and evolution, as well as the degree and pattern of genetic variability within and between populations. This is due to their co-dominant inheritance, easy genotyping, ubiquity, genetic neutrality, high mutation rates, and high degree of polymorphism (De Luca *et al.,* 2002; Dobrowolski *et al.,* 2002; Lai and Sun 2004; Scott, Rodney, and Ira 2005; Weissenbach, 1993). Microsatellites in *E. gracilis* have been investigated by means of randomly amplified polymorphic DNA (RAPD) fingerprinting. Fragments containing microsatellite loci have been detected by Southern hybridization and sequenced, revealing eight microsatellite loci in *E. gracilis* (Wuthisuthimethavee, *et al.,* 2003), however, the present *E. gracilis* genome assembly cannot support an extensive analysis for microsatellite DNA.

*rRNA*

In *E. gracilis,* the cytoplasmic large subunit (LSU) rRNA is composed of 14 discrete small RNA species that must interact in the functional ribosome (Donald, *et al.,* 1999) (see sub-sections on Translation, RNA processing and modification, and Transcription in this section as well as chapter 4 and 5). The native complexes of *E. gracilis* rRNA has been isolated and the largest of these complexes contains eight of the 14 LSU rRNA species. Several of these small rRNA species are able to associate in vitro to reform an isolated domain of LSU rRNA structure. Spliceosomal GT-AG cis-splicing occurs in *E. gracilis* (see chapter 5 and appendices), in addition to the nonconventional cis-splicing and spliced leader trans-splicing previously recognized in this early diverging unicellular eukaryote (see chapter 5 on tubulins and Donald, *et al.,* 1996).

*1.2.1.3 Evolution of the nuclear genome*

*E. gracilis* possesses some distinctive characteristics in the nucleus such as their permanently condensed chromosomes, intact nuclear envelope (Haapala and Soyer, 1975), and a nuclear matrix (see chapter 5 of this work). The permanently condensed state of the chromosome ensures that the giant DNA structures are tightly fitted into the *E. gracilis* nucleus. The nuclear matrix is relevant because it contains residual elements of the pore-complex and lamina (which is absent in *E. gracilis*, see chapter 5 of this work), the nucleolus, and an intra-nuclear fibrous network that provides the basic shape and structure of the nucleus (Plaumann and Pelzer-Reith, 1997). The 18S rRNA tree suggests that *E. gracilis* diverged very early from other eukaryotes (Cavalier-Smith, 1995). The fact that all lower unicellular and/or primitive eukaryotes possesses nuclear matrix suggests that the nuclear matrix must have originated very early in the progress from prokaryotic cells to eukaryotic cells, and its origin must be an important foundation of the origin of the eukaryotic nucleus (see Nigg, 1992; Yang, *et al,* 1997; Stick and Hausen, 1985; Lehner, *et al.,* 1987; Steward and Burke, 1987; Stick, 1992; Wen, 2000 for detailed review on nuclear matrix and lamina).

*Common and uncommon evolutionary patterns with Kinetoplastids, and other supergroups: β-D-glucosyl-hydroxymethyluracil or J*

*E. gracilis* also possess common and distinct evolutionary patterns with Trypanosomatids, Naegleria, and Giardia. For instance, β-d-glucosyl-hydroxymethyluracil or J is a DNA modification discovered in the DNA of the African trypanosome, *Trypanosoma brucei*, where it replaces 0.5 – 1.0 % of all thymines (Gommers-Ampt, *et al.,* 1993; Borst and Van Leeuwen, 1997). J has only been found in non-transcribed, and transcribed, repetitive sequences of *T. brucei* (Van Leeuwen, *et al.,* 1997) and a substantial fraction of J is present in both strands of the telomeric hexamer (GGGTTA) repeats of this organism (Van Leeuwen, *et al.,* 1996). Dooijes, *et al.,* 2000, speculated that J may have an analogous function in trypanosomes as 5-methylcytosine (5-MeC) in plants and animals, and that it is involved in transcriptional repression and/or suppression of recombination between homeologous sequences, and regions of transcriptional initiation (Liu, et al., 2014; Borst and Van Leeuwen, 1997, Van Leeuwen, *et al.,* 1997, 1998).

The DNA of *E. gracilis* have been analysed for the presence of the unusual minor base J, thus far only found in kinetoplastid flagellates and in *Diplonema*. Using antibodies specific for J and post-labeling of DNA digestion followed by two-dimensional thin-layer chromatography of labeled nucleotides, Dooijes, *et al.,* 2000, demonstrated that approximately 0.2 mole percent of *Euglena* DNA consists of J, an amount similar to that found in DNA of *T. brucei*. By staining permeabilized *E. gracilis* cells with anti-J antibodies, they showed that J is rather uniformly distributed in the *E. gracilis* nucleus, and does not co-localize to a substantial extent with (GGGTTA) repeats, the putative telomeric repeats of *E. gracilis*. Hence, most of J in *E. gracilis* appears to be non-telomeric. It seems also likely that *E. gracilis* has much longer chromosomes than *T. brucei* (the reason behind this is not yet known), and that most of the J are intra-chromosomal in this organism. Their result adds to the existing evidence for a close phylogenetic relation between kinetoplastids and euglenids as well as suggesting that base J might have evolved or been present early in eukaryogenesis.

*The impact of endosymbiosis on the nuclear genome*

The impact of endosymbiosis on the euglenophyte nuclear genome is not fully understood due to its complex nature as a 'hybrid' of a non-photosynthetic host cell and a secondary endosymbiont (Maruyama, *et al.,* 2011). Morphological, biochemical, and phylogenetic analyses suggest that only the last common ancestor of the extant plastid-harboring euglenophytes experienced the secondary endosymbiosis, but not the common ancestor of Euglenozoa as a whole (Euglenida, Diplonemea and Kinetoplastea) (Leader, 2004). However, the discovery of algal-type genes and the specific features of a mitochondrion-targeted protein in Kinetoplastea showing similarity to those of euglenophytes led to the hypothesis that a plastid was present in the common ancestor of Kinetoplastida, or Euglenozoa (Bodyt, *et al.,* 2010; Krnacova, *et al.,* 2012). A preliminary expressed sequence tag (EST) analysis of *E. gracilis* showed a complex history of nuclear genes in this organism (Ahmadinejad, *et al.,* 2007), but many aspects of how the *E. gracilis* nuclear genome integrated genes from the green algal endosymbiont via secondary endosymbiosis are unclear. Moreover, recent molecular phylogenies suggested the presence of 'red lineage' genes in the nuclear genome of *E. gracilis* (as well as evidences from this work; see chapter 5), but their origins and evolutionary histories have not been

explored in detail (Petersen, *et al.,* 2006; Teich, *et al.,* 200; Frommolt, *et al.,* 2008; Sanches-Perez, *et al.,* 2008).

Genome mosaicism has also been reported in *E. gracilis* and *P. trichophorum*. *E. gracilis* acquired at least 14 genes via eukaryote-to-eukaryote lateral gene transfer from algal sources other than the green algal endosymbiont that gave rise to its current plastid (this work). Some or all of them were independently acquired by lateral gene transfer and contributed to the successful integration and functioning of the green algal endosymbiont as a secondary plastid. Alternative hypotheses include the presence of a phagocytosed alga as the single source of those genes, or a cryptic tertiary endosymbiont harboring secondary plastid of red algal origin, which the eukaryovorous ancestor of euglenophytes had acquired prior to the secondary endosymbiosis of a green alga (Maruyama, *et al.,* 2011).

*1.2.1.3 Nuclear genome functions*

*Life history, cell division and growth, cell control, chromosome partitioning, defense mechanisms, and apoptosis*

There are three different stages in the life cycle of *E. gracilis*: The free swimming flagellated form, palmella stage, and the cyst stage (Tannreuther, 1923, Jahn, 1946). The free living flagellated forms (encysted state) refers to as the adult developmental stage (or mother cell) where the organelles and flagellar are more conspicuous with pronounced cellular movement (Tannreuther, 1923, Jahn, 1946). The palmella stage (cell division also occur here) are the non motile cells of Euglena that is covered with mucilage, and have the possibility of becoming flagellated again (Tannreuther, 1923, Jahn, 1946). The cyst stage consist of three forms: The protective form (thick-walled closed cyst), the reproductive form (thin-walled cyst), and the transitory or resting form where the cyst wall is thick but not completely closed – the cell is usually flagellated with a pronounced sense of motility and response to stimuli such as sunlight (Tannreuther, 1923, Jahn, 1946).

In *E. gracilis*, there are no confirmed reports of sexual reproduction (except a possible cytological evidence by Leedale, 1962, in the Euglenoid *Hyalophacus ocellatus*) (Jahn, 1946, Leedale, 1962), and reproduction is usually by asexual reproduction (Tannreuther, 1923, Ratcliffe, 1927, Jahn, 1946). Similarly, the question

of whether the Euglenoid cell is normally haploid or diploid has been discussed (Chadefaud, 1940, Leedale, 1962, and Rawson, 1975), however, in the absence of proof of the existence of sexual phenomena this question seems somewhat far-fetched. There are three forms of asexual reproduction propagated by *E. gracilis*: 1) Reproduction by longitudinal binary fission of the flagellated stage, 2) Division of non flagellated cells contained within reproductive or division cysts, 3) Division of cells in palmelloid stage which is common in *E. viridis*, *E. stellata*, *E. schmitzii*, and *E. pisciformis* (Tannreuther, 1923, Ratcliffe, 1927). *E. gracilis* propagates mainly by a process known as longitudinal binary fission which is a form of asexual reproduction (Tannreuther, 1923, Ratcliffe, 1927). This involves the division of the nucleus (containing the genetic material) through the process of mitosis. Mitosis involves five stages in *E. gracilis*: Interphase, prophase, metaphase, anaphase, and telophase (Ratcliffe, 1927). These stages ensure that the nuclei is duplicated and temporarily resident in the same cell. The rest of the cell is duplicated and separated through cytokinesis resulting in two identical daugther Euglenoid cells which are smaller than the mother cells. These daughter cells contain the two nuclei and equal proportion of the organelles. Under favourable conditions, in the presence of nutrients and optimum growth conditions, the growth of the daughter cell results in the development and achievement of the optimum number of organelles, and the cell division cycle is repeated (see Jahn, 1946 for an extensive review on the life history of Euglenoids).

The ploidy nature of the Euglenids have been contentious for over five decades. While there has been no recent attempt to uncover this ploidy nature, some extensive reviews and investigations have been carried out by Chadefaud, 1940, Leedale, 1962, and Rawson, 1975. Using experimental findings from Dobell, 1908, Berliner, 1909, and Biecheler, 1937, Chadehaud concluded that the Euglenids are diploids rather than haploids as seen in *Phacus spp* (Chadefaud, 1940). Leedale, 1962, suggested that the Euglenids are possibly haploid. This is because since there is a possible true meiosis in the Euglenid (*Hyalophacus spp*) and the chromosome counts in this Euglenid suggest that it is following autogamy or sexual fusion of a gamete nuclei rather that preceding the formation of a gamete implying a haploid state. Several literatures have reported high chromosome numbers, large DNA per cell, and cell morphology in most modern Euglenids which are indicative of the

presence of fairly high degrees of polyploidy (Leedale, 1958a, 1959b, and 1961, Rawson, 1975). Following from these findings, Leedale, 1962, proposed that if meiosis does occur in the Euglenids, it does this as a rare phenomenon and will possibly be a perculiar form of meiosis specific to the Euglenids as is the mitosis (Leedale, 1962). Similarly, Rawson, 1975, suggested that *E. gracilis* is not polyploid by carrying out a reassociation kinetics of *E. gracilis* DNA. He determined the $C_ot_{1/2}$ of the unique DNA of *E. gracilis* which is 2000, and the amount of DNA per cell (3 pg) to propose this suggestion. Three kinetically definable classes of DNA were observed in *E. gracilis*: a highly repetitive fraction, a middle repetitive fraction, and a non-repetitive fraction, and each fraction characterised by its melting properties. The presence of a non repetitive DNA fraction with observed $C_ot_{1/2}$ C similar to that expected for diploid organism containing 3 pg of DNA, indicates that *E. gracilis* is not a polyploid, with a possible genetic recombination (see Rawson, 1975 for an extensive discussion). These studies were indicative of the diploid or polyploid nature of *E. gracilis* without strong evidences for a haploid state. The data from this work also suggests a non haploid state (see Chapter 5), and a complete *E. gracilis* genome sequence will provide further evidence to the ploidy level of *E. gracilis*.

Two principle factors actively control the cell division cycle in *E. gracilis*, these include: cell growth (stationary and logarithmic phases) and light/darkness (circadian rhythm) (see Goto and Beneragama, 2010). In autotrophic cultures, chlorophyll synthesis and growth rate is dependent on the concentration of nitrogen and carbon (Regnault, *et al.,* 1990), and cobalamin (Ross, 1952; Varma *et al.,* 1961; Watanabe, *et al.,* 1992; Watanabe *et al.,* 1988a; 1992). Dark grown cultures lose their chlorophyll, and their chloroplasts regress to form proplastids. Upon exposure to light in inorganic medium, they are able to re-differentiate chloroplasts, resuming photosynthetic functions (Buetow, 1968).

In *E. gracilis,* a number of striking chemical derangements accompany growth arrest which includes: severe repression of RNA and protein synthesis, doubling of the cellular volume, DNA contents, peptides, amino acids, nucleotides, pyrophosphates, and proteins of unusual composition accumulates, with a marked increase in the intracellular content of $Ca^{2+}$, $Mn^{2+}$ and $Fe^{2+}$ (see Wacker, 1962; Wacker, Kornicker & Pothier, 1965; Falchuk, *et al.,* 1975 for extensive reviews and data). The specific

biochemical events underlying these derangements are yet unknown which will be an interesting area of research, but the profound metabolic disturbances resulting from zinc deficiency point to events critical to cellular division and nucleic acid metabolism (Wacker, 1962; Wacker & Vallee, 1959; Shin & Eichhorn, 1968a, *b;* Prask& Plocke, 1971; Slater, Mildvan & Loeb, 1971; Scrutton, Wu & Goldthwait, 1971; Auld, Kawaguchi, Livingston & Vallee, 1974a, *b*) which may provide some explanations. In *E. gracilis,* the chromosomes appear to remain condensed at interphase (as previously mentioned above), resembling the metaphase chromosomes of other organisms (Leedale 1968; Moyne et al. 1975; Pellegrini, 1980; Hayashi-Isimaru, *et al.* 1993; Ueda and Hayashi-Ishimaru, 1996). The structures responsible for motive force development in cell aperiodic contraction and re-extension, cytokinesis and chromosome separation during mitosis in the unicellular euglenoid, *E. gracilis,* have not yet been identified (Vanni and Poli, 1983). However, contractile proteins, primarily actin, have been proposed to be involved in these activities considering the role of actin cytoskeleton in spindle assembly and positioning in eukaryotes (Hofmann and Bouck 1976, Lefort-Tran, *et al.,* 1980; Bre, *et al.,* 1981; Murray, 1981; Carter, 1967; Copeland, 1974; Brown and Spudich, 1979; Kunda and Baum, 2009). Similarly, kinetochores which are macromolecular protein complexes have been evidenced to play specific role in chromosome segregation, and the presence of the conventional and non conventional types in *E. gracilis* raises the question on the origin of the kinetochore proteins (see chapter 5 of this work for discussion).

In *E. gracilis*, an apoptosis-like process (Gordeeva, *et al.,* 2004; Deponte, 2008) similar to the processes in *Leishmania* and *Trypanosoma* (Nakhasi, 2003; Deponte, 2008; Ridgley, *et al.,* 1999; Mukberiee, *et al.,* 2002; Simpson, et al., 2004) has not been reported. It is has been hypothesized that this apoptotic-like process (Scheuerlein, *et al.,* 1995) is a derived feature in kinetoplastids rather than ancestral for all euglenozoans (Bumbulis and Balog, 2013). If it is an ancestral trait, it is expected that a similar death process in the more distantly related *E. gracilis* would have been evidenced (see Bumbulis and Balog, 2013). In *E. gracilis,* defense mechanism involves avoidance that is specific to circadian rhythm (Goto and Beneragama, 2010; Goto, *et al.,* 1985), homeostasis (see Greer and Brunet, 2008; Kirkwood and Shanley, 2005; Goto and Beneragama, 2010), optimum strategy

adopted in the absence of increased cells mortality and reproductive overhead (Shanley and Kirkwood, 2000; Kirkwood and Shanley, 2005; Bolige, *et al.,* 2005b), the production of biofilms (Morales-Calderón, *et al.,* 2012) which is considered a metal protection mechanism, and finally, the production of algae toxins (Zimba, *et al.,* 2010) such as Euglenophycin which has been reported in *E. sanguine* and plays significant role in cellular defense (see Rodri´Guez-Zavala, 2007 for the biochemical and molecular mechanisms underpinning this processes). It is not yet fully understood which part of the *E. gracilis* genome controls cellular defence mechanism. But it's been proposed to be in the nucleus considering that the nucleus controls a majority of the cellular processes in eukaryotes, including cell division and apoptosis. This means that a susceptibility, or resistant, to external stress would most certainly be controlled by the nucleus (see Ebenezer, *et al.,* 2018, and Chapter 5, and Appendix I-a of this work).

*Transcription, replication, recombination and repair, chromatin structure and dynamics*

Comparative genomics of parasitic protists (such as trypanosomatids) and their free-living relatives (such as *N. gruberi, B. saltans*, and *E. gracilis*) are profoundly impacting our understanding of regulatory systems involved in transcription and chromatin dynamics (Iyer, *et al.,* 2008). While some parts of these systems are highly conserved, other parts are rapidly evolving, thereby providing the molecular basis for the variety in the regulatory adaptations of eukaryotes (Iyer, *et al.,* 2008). The gross number of specific transcription factors (promoters and repressors) and chromatin proteins are positively correlated with proteome size in eukaryotes, and the individual types of specific transcription factors and polymerases show huge variety within and across eukaryotic lineages (Iyer, *et al.,* 2008).

In Euglenids, RNA polymerases have been reported which are substantially influenced by external conditions. For instance, in *E. gracilis*, the existence of RNA Pol I and II have since been reported using zinc-sufficient cells which substantially differs from RNA polymerases species in zinc-deficient cells (Falchuk, *et al.,* 1976). This suggests that regulatory mechanisms which serve to synthesize or preferentially activate different classes of RNA polymerases appear to be called into play in zinc-deficiency, and which may potentially result to derangements in RNA metabolism,

and ultimately producing defects in protein synthesis of zinc deficient cells either during initiation, elongation, or termination (see Falchuk, *et al.,* 1976, for discussions on similar control processes in sea urchin, liver, *Helianthus tuberosus*, amoeba, and *E. coli,* with respect to the activities of RNA polymerase I and II variation as a function of the eukaryotic species stage of development).

In eukaryotes, except yeast $tRNA^{Ser}$ -$tRNA^{Met}$ gene locus, most tRNA genes so far examined are present as single transcription units (DeRobertis and Olson, 1979, Garber and Gage, 1979). In *E. gracilis* chloroplast, the processing of the chloroplast polycistronic tRNA transcript resembles those of prokaryotes (Gruissem and Prescott, 1982), and the tRNA gene cluster ($tRNA^{Val}$ –$tRNA^{Asn}$ –$tRNA^{Arg}$) are faithfully transcribed into polycistronic precursor tRNAs (~ 268 nucleotides) by RNA polymerase III, by recognizing internal sequences of the prokaryotic tRNA genes for transcription initiation rather than a 5' flanking prokaryotic promoter sequence (Gruissem and Prescott, 1982). These precursor RNAs possess a consensus sequence, and are subsequently processed into the mature or nearly mature tRNAs by a defined pathway which involves trimming of the 5' leaders and 3'trailers of the primary transcript as well as the intermediates (Gruissem and Prescott, 1982).

In trypanosomes, as in *E. gracilis*, the mechanism for the control of gene expression is unusual, and transcription by RNA Polymerase II is polycistronic with only a few transcriptional units having been identified so far (Ivens, *et al.,* 2005). The sequences present in the polycistronic units code for proteins having unrelated functions or similar metabolic pathways (De Gaudenzi, *et al.,* 2011). This enhances gene expression by post-transcriptional events similar to the phenomenon in *E. gracilis* (see Chapter 6 of this work). In *T. brucei*, RNA Polymerase I transcribes the rRNA gene cluster, and this enzyme also mediates the transcription of two life cycle stage-specific mRNAs, bloodstream-form VSGs and procyclic-form procyclins (Ruan, *et al.,* 2004). During mRNA maturation in trypanosomes, polycistronic pre-mRNA units are processed into monocistrons by two coupled cleavage reactions *Trans*-splicing attaches a capped 39-nt SL (spliced leader) sequence at the 5' end of virtually all mRNAs, and the 3' end is polyadenylated (De Gaudenzi, *et al.,* 2011). The polyadenylation machinery in *E. gracilis* is not yet fully understood compared to those in trypanosomes. In *T. brucei*, a high throughput RNA sequencing has

demonstrated that most genes have between one and three alternative splice-acceptor sites, but contain several polyadenylated sites (Siegel, *et al.,* 2010).

During rRNA processing and maturation, pre-rRNA undergoes post-transcriptional modification including 2'-O-methylation and pseudouridylation events (Kiss 2001; Lafontaine and Tollervey 2001; Maden, 1990). In prokaryotes, they are primarily specified by ribonucleoprotein (RNP) complexes containing small guide RNAs associated with well-defined protein components (Bachellerie, *et al.,* 2002; Decatur and Fournier, 2003; Kiss 2002; Omer, *et al.,* 2003). In eukaryotes, these small guide RNAs localize to the nucleolus and are thus termed small nucleolar RNAs (snoRNAs), (Moore and Russell, 2012). There are two main types of snoRNAs, box C/D and box H/ACA, each classified by unique conserved sequence elements (box elements) and secondary structures (Kiss, 2002). *E. gracilis* snoRNAs display specific characteristics compared to some other eukaryotes (Moore and Russell, 2012), with significantly more $O^{2'}$ -methyl than uridine (ψ) sites (Schnare and Gray, 2011), and frequent gene duplication - a common mechanism driving snoRNA emergence and evolution in *Euglena* spp that has resulted in both the large number and clustered patterning of rRNA modification sites (Charette and Gray, 2009). Comparison of *E. gracilis* modified rRNA positions to those of other eukaryotes, including *Trypanosomes* (Dunbar, *et al.,* 2000; Liang, *et al.,* 2005; Uliel, *et al.,* 2004; Xu, *et al.*, 2001) indicates significant differences, including many *Euglena*-specific sites (Schnare and Gray, 2011; Russell, *et al.,* 2006) which also indicates that the large subunit (LSU) rRNA of *E. gracilis* is extensively fragmented, and that *Euglena spp* rRNA is heavily modified, containing significantly more $O^{2'}$-methylated nucleotides than the rRNAs of yeast or vertebrates. However, little is currently known about the mechanism producing the highly modified rRNA pattern (Moore and Russell, 2012), and the present genome data in this work does not support an extensive analysis to uncover this (see Chapter 3 and 4 of this work).

Ribonucleotide reductases (RNRs) are enzymes that provide all cells with the deoxyribonucleoside triphosphates (dNTPs) required for DNA replication and repair (Eklund, *et al.,* 2001; Jordan and Reichard, 1998; Lawrence and Stubbe, 1998, Sjoberg, 1978, 1997; Torrents, *et al.,* 2005; Reichard, 1993; Torrent, *et al.,* 2002; Stubbe, 2003; Eklund, *et al.,* 2001; Jordan and Reichard, 1998; Lawrence and

Stubbe, 1998). *E. gracilis* is an eukaryotic unicellular organism with a class II ribonucleotide reductase (Gleason and Hogenkamp, 1970; Hamilton, 1974) and represents an exception from the general rule that eukaryotes contain class I enzymes (Torrents, *et al.,* 2005). *E. gracilis* RNR belongs to the group of monomeric RNRs. Ribonucleoside triphosphates serve as the substrate and enzyme activity is absolutely dependent on adenosyl cobalamin. The structure and function of the *E. gracilis* enzyme is of interest with regard to the evolution of ribonucleotide reduction, but it also bears upon the evolution of vitamin $B_{12}$ dependence among eukaryotes in general (Torrents, *et al.,* 2005). Only prokaryotes synthesize $B_{12}$ (Roth, *et al.,* 1996), but numerous eukaryotes have a $B_{12}$ requirement for growth. $B_{12}$ dependence in *E. gracilis* relates to its $B_{12}$-dependent ribonucleotide reductase (Gleason and Hogenkamp, 1970; Hamilton, 1974) and methylmalonyl-CoA mutase (Watababe, et al., 1996), which likely forms the basis of the classical *E. gracilis* clinical assay for diagnosing human $B_{12}$ deficiency and for quantifying human serum $B_{12}$ levels (Mollin, *et al.,* 1976). In vitamin $B_{12}$-deficient *E. gracilis*, ribonucleotide reductases are inducible enzymes and are synthesized just before and during the S-phase (Follmann, 1974). There are two possible reasons why *E. gracilis* contain a gene for a $B_{12}$-dependent class II RNR (and methylmalonyl-CoA mutase) different from other eukaryotes (Watanabe, *et al.,* 1996; Torrent, *et al.,* 2005): lateral gene transfer or inheritance from the eukaryote common ancestors. The function of the nucleus also involves maintaining chromatin structure and dynamics. However, there is no recent information on the chromatin structure and dynamics of *E. gracilis* except the heterochromatin organization as reported in this work, and existing literature suggests that there are a number of descriptive responses *of E. gracilis* to deficiencies of the essential metals, for instance, zinc, iron, manganese, or magnesium or to cold shock (Falchuk, *et al.,* 1986). The effects of Zn deficiency on chromatin composition and structure are characteristic and specific. This usually results to the absence of histones in acid extracts from chromatin and an increased resistance of the latter to nuclease digestion which are unique and specific to Zn deficiency in *E. gracilis* (see Falchuk, *et al.,* 1986, for details).

*Translation, ribosomal structure and biogenesis, RNA processing, and modification*

The conversion of genetic information stored in DNA into a protein product proceeds through the obligatory intermediate of messenger RNA (Monde, *et al.,* 2000). The

steady-state level of an mRNA is determined by its relative synthesis and degradation rates, i.e., an interplay between transcriptional regulation and control of RNA stability. When the biological status of an organism requires that a gene product's abundance varies as a function of developmental stage, environmental factors or intracellular signals, increased or decreased RNA stability can be the determining factor. RNA stability and processing have long been known as important regulatory points in chloroplast gene expression (Monde, *et al.,* 2000). Several interesting features of rRNA biosynthesis in *E. gracilis* combine to make this organism an attractive model system for studies of ribosome biogenesis. This include few integrated chromosomal copies of rDNA (Ravel-Chapuis, 1988), 14 individual coding RNA pieces contained in the nucleocytoplasmic LSU rRNA (Schnare and Gray, 1990; Schnare, *et al.,* 1990) and processing of the SSU and LSU rRNAs from a large precursor of approximately 10.3 kb (Brown and Haselkorn, 1971).

Initial characterization of *E. gracilis* rRNA (involving chemical and/or enzymatic sequencing of 5'-end-labeled and 3'-end-labeled RNA species) (Schnare and Gray, 1990) suggested that *E. gracilis* rRNAs are highly modified (Decatur and Fournier, 2002; Schnare and Gray, 1990; Murray, *et al.,* 2011). *E. gracilis* cytoplasmic rRNAs (particularly the LSU rRNA) are the most highly modified rRNAs yet described. Comparison with previously published modification patterns for other eukaryotic rRNAs identifies conserved sites of modification (presumably ones involved in general eukaryotic ribosome function) and species-specific sites, the functional significance of which remains to be determined. In *E. gracilis*, the additional rRNA modifications are correlated with aspects of potential secondary structure, suggesting that supernumerary modification may play a role in stabilizing higher-order structure in the case of a highly fragmented rRNA.

In *E. gracilis* rRNAs, biogenesis requires a considerably larger number of snoRNAs than is the case in other eukaryotes (Murray, *et al.,* 2011). The *E. gracilis* modification maps could also aid in the prediction of modified sites in other eukaryotic rRNAs where mapping data are not yet available. The lower overall number of modified nucleosides in the cytoplasmic rRNA of other eukaryotes compared to *E. gracilis* raises the question of whether there has been an

evolutionary loss of modifications in other eukaryotes, or a gain, in *Euglena*. Some sites that are highly conserved among eukaryotes are not modified in *E. gracilis*, suggesting a specific loss in the evolutionary lineage leading to the genus *Euglena*. Modifications that are located in variable regions must represent gain; for instance, the unusual cluster of five 1-methyladenosine ($m^1A$) residues in LSU species 6 is located in a region that is absent in most other LSU rRNAs and is not even conserved among *Euglena* species. It's been proposed that elevated levels of snoRNA gene duplication and divergence in the *E. gracilis* genome have resulted in an unusually wide array of potential snoRNAs having novel target sites within rRNA. (see Maden, 1990; Hury, *et al.,* 2006; Bokov and Steinberg, 2009; Kawai, 1992; Arnez and Steitz, 1994; Davis, 1995, 1998; Auffinger and Westhof, 1998; Davis*, et al.,* 1998; Yarian, *et al.,* 1999; Durant and Davis, 1999; Charette and Gray, 2000; Murray, *et al.,* 2011 for the most highly modified SSU and LSU rRNAs, most ancient parts of *E. gracilis* rRNA, the role of supernumerary modification in the stabilization of RNA structure, RNA modification maps and snoRNA genes in *E. gracilis*).

*Post-translational modification*

ADP-ribosylation is a post-translational modification of proteins, and both poly- and mono-ADP-ribosylations have been reported in *E. gracilis* (Ueda and Hayaishi, 1985) which are cell cycle dependent (Shigeo, *et al.,* 1995). In *E.gracilis*, poly ADP-ribose (see Edmunds, 1984a & b; Matsuyama and Tsuyama, 1991; Scaife, *et al.,* 1992Shigeo, *et al.,* 1995 for details on physiological role of mono and poly ADP-ribosylation) is synthesized only in S phase and not at the transition from the G2 to the M phase. How poly ADP-ribose is synthesized in the S phase and cleaved, as well as the function of mono ADP-ribosylation in the S phase are yet unknown (Shigeo, *et al.,* 1995). ADP-ribosylation functions specifically during the S phase in the chloroplast and mitochondrial fractions organelles, where it might be related to genome replication and/or repair. ADP-ribosyltransferase or acceptor proteins are sensitive to trypsin; thus, ADP-ribosylation is never found in organelles obtained using partial trypsin digestion (Shigeo, *et al.,* 1995).

### *1.2.2 Chloroplast*

#### *1.2.2.1 The photosynthetic apparatus*

The eukaryotic chloroplast genome encodes over 30 proteins that are involved in photosynthesis, including components of photosystems I and II, cytochrome *bf* complex, ATP synthase, and one of the subunits of ribulose bisphosphate carboxylase (RUBISCO). Although, chloroplasts encode more of their own proteins than mitochondria, about 90% of chloroplast proteins are still encoded by nuclear genes, suggesting that many constituents required for chloroplast development may come from outside the plastid in *E. gracilis* (Cooper, 2000). This also suggests that certain chloroplast molecules might be supplied through synthetic activities and informational contributions from outside the chloroplast. *E. gracilis* contains at least three different genomes localized in the nucleus, mitochondrion, and chloroplast respectively (Bovarnick, 1974; Schiff, 1971; Nass and Ben-Shaul, 1972), and at least two and perhaps three sets of ribosomes (Bovarnick, 1974), the 68S of the plastid, the 87S of the cytoplasm and perhaps a mitochondrial species (Bovarnick, 1974; Avadhani & Buetow, 1972a, b). The mitochondrial DNA is small (~ 19 kbp) (Shori, Ben-Shaul & Edelman, 1970; Bovarnick, 1974); it therefore possesses a very limited amount of information. Recent studies suggests there are only 7 protein-coding genes within the mitochondrial of *E. gracilis* (Dobakova, *et al.,* 2015). The nucleus and chloroplast contain relatively large amounts of DNA, and are the sources of information for plastid biosynthesis.

*E. gracilis* cells are characterized by thylakoid dismantling and chloroplast dedifferentiation (Scheer & Parthier 1982; Reinbothe, *et al.,* 1991). Thylakoid dismantling is a process whereby chloroplasts are converted to non-photosynthetically active plastids called proplastids (Ferroni, *et al.,* 2009) (see Chapter 6 of this work for current understanding on the transition from light to dark and vice versa). It characterizes processes such as the chloroplast-to-gerontoplast transition in senescing photosynthetic tissues and the chloroplast-to-chromoplast transition in reproductive structures such as petals and fruits (Cheung, *et al.* 1993; Bonora, *et al.* 2000; Thomas, *et al.,* 2003; Keskitalo, *et al.*, 2005). The removal of thylakoid components, primarily chlorophyll (chl)–protein complexes, leads to progressive changes in organisation of the inner membrane system of the organelle (Ferroni, *et al.,* 2009). The degradation of key complexes, i.e. photosystem II (PSII), photosystem I (PSI) and their light-harvesting complexes (LHC), has been mainly characterized in higher plant leaves, where senescence processes are induced

(Humbeck et al. 1996; Miersch et al. 2000; Humbeck & Krupinska 2003). Although thylakoid dismantling can vary considerably in different experimental systems, degradation of the photosynthetic complexes presents common characteristics (Ferroni, *et al.,* 2009). Chl–protein complexes are usually degraded in a multi-step fashion; in particular, LHCs normally undergo slower degradation than the photosynthetic reaction centres (RCs) (Kura-Hotta, *et al.,* 1987; Noode´n, *et al.,* 1997; Miersch, *et al.*, 2000; Humbeck & Krupinska, 2003; Tang, *et al.,* 2005).

Chloroplast differentiation occurs only in dividing cells and generally leads to the formation of plastids with nearly absent thylakoid system (Ferroni, *et al.,* 2009), which may contain internal membranes (Schift and Epstein, 1961). Typical chloroplasts in *E. gracilis* cells (Figure 1.1, 1.3), contains long lamellae, each consisting of 2 – 3 thylakoids (Tomoko Ehara, *et al*., 1975). *E. gracilis* acquired its chlorophyll a-b containing plastid by secondary endosymbiosis as previously mentioned; the organelle is bounded by three membrane envelope, attesting to the multiple endosymbiotic events that led to the present state (Keeling, 2004). It has been suggested that the *E. gracilis* photosystems use a common antenna system composed of both LHCI and LHCII (Doege, *et al.,* 2000); however, the LHCI of *E. gracilis* seems to be non-homologous to LHCI of green algae (Koziol, *et al.,* 2007). When grown in the dark in the presence of organic carbon sources, *E. gracilis* is etiolated and contains about 10 small, polyp-shaped plastids, commonly called ‘proplastids’ (Schwartzbach & Schiff, 1983). Upon exposure to light, proplastids differentiate into photosynthetically active chloroplasts with elongated three-thylakoid lamellae. The sequence of events leading to formation of the proplastids still remains obscure in many aspects; including thylakoid re-arrangement and degradation (see Ferroni, *et al.,* 2009)

#### *1.2.2.2 Chloroplast genome structure*

The only completely sequenced *E. gracilis* non nuclear DNA (to date), are the plastid and the mitochondrial genome (Hallick, *et al.,* 1993, Dobakova, *et al.,* 2015). The nuclear genomic DNA which is a focus of this work (see chapter 2 – 5) has been sequenced using Next Generation Sequencing (NGS) technology. In *E. gracilis*, the plastid genome is highly asymmetrical in its composition. One asymmetrical feature that has been noted previously is that the circular chromosome is strongly biased in

terms of which strand is the coding strand (Hallick, *et al.,* 1993), and its entire plastid genome consists of 143,170 bp, a total of 97 proteins and gene loci - 46 of these are protein-coding genes, and 51 RNA-coding genes (see Hallick, *et al.,* 1993 for details), including group II and III introns, and twintrons (i.e. introns inserted within introns) (Thomson, *et al.,* 1996). The two strands of the *E. gracilis* chloroplast genome also differ in compositional properties, and codon usage bias (Morton, 1999), and it has been proposed that this bias is due to selection to coordinate transcription and DNA replication (see Morton, 1998, 1999; Hallick, *et al.,* 1993).

*Primary transcripts, intergenic spacers transcripts, introns and twintrons*

An *E. gracilis* cell possesses approximately eight secondary plastids bounded by three membranes (Chem, *et al.,* 1977; Rawso and Boerma, 1976), as previously mentioned. The plastid genome of *E. gracilis* is circular, and the evolutionary transition from an endosymbiont to the plastid organelle was accompanied by a loss of many genes and gene transfer from the endosymbiont genome (s) (see sub section 1.2.3 for a review on endosymbiosis and evolution of the genus, *Euglena*) to the host genome (Martin and Herrman, 1998). Gene transfer from plastids and mitochondria is an ongoing process, as it has been demonstrated in animals, plants, fungi as well as other protists (Hazkani-Covo, *et al.,* 2010; Smith, *et al.,* 2011). The polyadenylation of plastid transcripts in *E. gracilis* has been previously studied by hybridization experiments of labeled plastid RNA to plastid DNA, and it was concluded that plastid rRNAs were polyadenylated to some extent, while plastid mRNAs were not (Milner, *et al.,* 1979). A small portion of plastid mRNAs has been demonstrated to be polyadenylated as well, and similar mechanism exists in secondary plastids (Zahonova, *et al.,* 2014). One of the most remarkable features of the *E. gracilis* plastid genome (cpDNA) is the presence of over 150 introns (160 total) that interrupt almost all of its protein-coding genes (Hallick, *et al.,* 1993; and GenBank X70810.2). Self-splicing catalytic ribozymes, such as these, are common in organellar genomes (Cech, 1990; Lambowtiz and Zimmerly, 2011), but the introns found in *E. gracilis* cpDNA are highly unusual. The majority of its introns are so-called group II introns, with several other group III introns (Lambowtiz and Zimmerly, 2004, 2011; Copertino and Hallick, 1993) (see Chapter 5 & 6 of this work for indications of novel splicing mechanisms in *E. gracilis*).

The *E. gracilis* plastid genome also contains about 15 twintrons in total. These intricate nested introns must be removed sequentially to result in accurate splicing. The unorthodox *E. gracilis* introns have been found in other *Euglena species* (Doetsch, *et al.,* 1998, 2000; Sheveleva, *et al.,* 2002) including *Euglena* (formerly *Astasia*) *longa* (Gockel and Hachtel, 2000), a close relative that has retained a plastid genome despite the loss of photosynthesis, but until recently it was not known if these peculiar genetic elements were representative of the euglenids as a whole. Early hypotheses postulated that these introns were acquired late in the evolution of photosynthetic euglenids (Thompson, *et al.,* 1995) and the subsequent finding that the plastid genome of the prasinophyte, *P. parkeae,* did not contain any highly-reduced group II introns (Turmel, *et al.,* 2009) gave further credence to the notion that these introns were gained after the secondary endosymbiotic plastid uptake. In *E. gracilis*, analysis of intergenic spacer transcripts suggests the possibility of 'read around' transcription, i.e transcription that proceeds multiple times around rDNA circle without termination (Spencer, 2001; see Yasuko, *et al.,* 1993 for studies on the structure of the nucleoids in *E. gracilis*).

*1.2.2.3 Evolution of the chloroplast genome*

Within the photosynthetic euglenoids, the only chloroplast genome that has been sequenced is from *E. gracilis* (Hallick, *et al.,* 1993). *E. gracilis* is not a typical representative photosynthetic euglenoid due to the diversity within the lineage, and it is not closely related to the phagotrophic forms believed to have been the host for the endosymbiont (Weigert, *et al.,* 2012). 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 *E. gracilis by* Hallick, *et al.,* 1993. Based on established phylogenetic assessment of the photosynthetic euglenoids, *E. gracilis* is consistently a crown species of the lineage (Linton, *et al.,* 2010; Marin, *et al.*, 2003; Triemer, *et al.,* 2006).

There are some significant changes that have been demonstrated by the *E. gracilis* chloroplast in comparison with other photosynthetic algae. Firstly, the *E.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).

Secondly, not only does *E. gracilis* lack the inverted repeats, but 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). Thirdly, the *E. gracilis* gene content appears consistent with other sequenced green algal chloroplast genomes, but the arrangement is not (Turmel, *et al.,* 2009). *E. gracilis* chloroplast DNA is the richest known source of introns, and an ideal system for studying the evolution and proliferation of group II, group III introns, and twintrons. During the evolution of the *Euglena* plastid DNA, introns were added to intronless progenitors and twintrons were formed by the insertion of one or more introns into an existing intron (Thompson, *et al.,* 1995). *E. gracilis* chloroplast introns arose late in the evolution of this lineage. The group II introns are the proposed progenitors of nuclear spliceosomal introns and are found in ancient genes from modern organisms.

*Endosymbiosis: explanation, potential use in understanding endosymbiosis mechanisms*

Several independent eukaryotic lineages have acquired their photosynthetic lifestyle from a secondary endosymbiont (Gibbs, 1978; Stoebe and Maier, 2002). Both primary and secondary endosymbiosis were accompanied by endosymbiotic gene transfer (EGT) - the relocation of genes from the organelle to the chromosomes of the host (Martin, *et al.,* 1993; Archibald, *et al.,* 2003; Timmis, *et al.,* 2004). Estimations for the frequency of EGT during the primary endosymbiosis range between 18 % of the *Arabidopsis thaliana* genome (Martin, *et al.,* 2002) and 11% of the Cyanophora genome (Reyes-Prieto, *et al.,* 2006). *E. gracilis* is well suited for the study of endosymbiosis and endosymbiotic gene transfer because its plastid was acquired by a secondary endosymbiosis, but it includes no remains of the endosymbiotic nucleus (McFadden, 2001). Moreover, *E. gracilis* shares a common ancestor with the Kinetoplastida (Adl, *et al.,* 2005, 2012), none of which seem to have experienced a secondary endosymbiosis (Rogers, *et al.,* 2006). Hence, by the sequence similarity criterion, the genome of *E. gracilis* is expected to be a hybrid composed of four main gene classes: (i) *Euglena*-specific genes, (ii) Kinetoplastida-specific genes, (iii) eukaryotic genes that are spread in other eukaryotes, and (iv) genes acquired during the secondary endosymbiosis (Ahmadinejad, *et al.,* 2007) (see Chapter 4, 5, and Appendices of this work for an extensive discussion). Clear

candidates for the latter group are genes that have homologues only in photoautotrophic eukaryotes. In some cases the acquired gene replaced an orthologous gene within the host (Henze, *et al.,* 1995). Such genes have homologues in both photoautotrophic and heterotrophic eukaryotes, but are closely related to the former. Because of these distinct gene origins, a single bifurcating tree cannot accurately describe the complex gene collection of *E. gracilis* because genome evolution by endosymbiotic gene transfer is a non tree-like process; thus buttressing that genome of *E. gracilis* is a hybrid of photoautotrophic and heterotrophic genomes (Ahmadinejad, *et al.,* 2007).

Although the majority of euglenids live in freshwater, the basal lineage of the autotrophic clade contains the marine species *Eutreptia* and *Eutreptiella*, corroborating the hypothesis of a marine origin of photosynthetic euglenids (Marin, *et al.,* 2003; Marin, 2004). Comparative studies on the plastid genomes of *E. gracilis, Eutreptia viridis perty*, and closely related green algal taxa have provided support for the hypothesis that a *Pyramimonas*-like alga was the euglenoid chloroplast donor via secondary endosymbiosis. Significant gene rearrangements exists between *E. gracilis* plastid genomes and that of other members of the group such as *Eutreptia viridis*; the key chloroplast genes are present and differ only in the absence of *psaM* and *roaA* in *Eutreptia viridis*, and *psaI* in *E. gracilis*, suggesting a high level of gene conservation within the euglenoid lineage (Weigert, *et al.,* 2012). The comparative analysis of the gene content between the plastid genome of *Pyramimonas parkeae*, which encodes 110 conserved genes (81 protein and 29 RNA species), (Turmel, *et al.,* 2009), and *E. gracilis*, which comprises 97 conserved genes (46 protein and 51 RNA species) (Hallick, *et al.,* 1993), has revealed a substantial loss of genes (for example all genes of NADH plastoquinone oxidoreductase of the plastidial respiratory chain) happening from the common ancestor of *P. parkeae* to extant *E. gracilis*. This reduction of gene repertoire is explained as a consequence of secondary endosymbiosis, although comparable gene losses took place in the prasinophyte lineages leading to *Pycnococcus* and to the coccoid microalgae *Ostreococcus* and *Monomastix* (Turmel, *et al.,* 2009).

Further gene loss in euglenids accompanying the loss of photosynthetic activity has been observed in the closely related but non-photosynthetic *E. longa*, which has

maintained 56 conserved genes (26 protein and 30 RNA species) (Gockel and Hachtel, 2000). Despite the reduction of coding capacity of the *E. gracilis* plastid in comparison to that of *P. parkeae*, the size of the *E. gracilis* plastid genome increased (143.2 vs. 101.6 kb in *P. parkeae*). The increase in the genome size should mainly be ascribed to the expansion of self splicing introns (see Ebenezer, et al., 2017, and Chapter 3 and 4 of this work for discussion on genome sizes). While *P. parkeae* features a single group II intron, the genome of the *E. gracilis* plastid contains ~160 group II and group III introns (15 of which formed twintrons), which is by far the most of all known organellar genomes (Thompson, et al., 1996; Doetsch, *et al.,* 1998). There are indications that the expansion of introns may be a feature specific to *E. gracilis* and its relatives (Thompson, *et al.,* 1996; Doetsch, *et al.,* 1998); however, no other plastid genome of euglenids has been completely sequenced, which would be necessary to enable comprehensive comparisons (Hrda, *et al.,* 2012). This phenomenon is also true in comparison with other euglenid relatives. For instance, the plastid genome of *E. gracilis* and *Eutreptiella* also provides a window into the process of secondary endosymbiosis of plastid euglenids (Hrda, *et al.,* 2012).

#### *1.2.2.4 Chloroplast genome functions*

Insights into the overall functional organization of the *E. gracilis* plastid chromosome and the transcriptional regulation of its genes (Geimer, *et al.,* 2009) has now been understood through a microarray-based transcriptome analysis using the complete sequence of the plastid chromosome (Hallick, *et al.*, 1993). The plastid chromosome appears to be constitutively expressed under all chosen external conditions including stresses such as UV light, temperature, antiplastidial agents, herbicide and heavy metal exposure. The euglenoid organelle transcriptome is qualitatively relatively insensitive to the environment, but exhibits marked overall quantitative changes. Global changes in patterns of gene expressions demonstrates that RNA turnover, translational, proteolytic and/or metabolic control regulate organelle gene expression in this flagellate (Geimer, *et al.,* 2009).

*Intracellular trafficking and targeting, secretion and vesicular transport, posttranslational modification and protein turnover*

As previously mentioned, the plastid of *E. gracilis* was acquired secondarily through an endosymbiotic event with a eukaryotic green alga, and as a result, it is

surrounded by three membranes (see Bachvaroff, *et al.,* 2005; Yoon, *et al.,* 2002). Nevertheless, the similar morphological architecture of their plastids may also require analogous transport mechanisms (Nassoury, *et al.,* 2003). The majority of the plastid proteome of these secondary plastids, as in primary plastids, is encoded in the host's nuclear genome (see Armbrust*, et al.*, 2004; Douglas, *et al.,* 2001; Gilson, *et al.,* 2006; Rogers, *et al.*, 2006, 2007). Thus, hundreds of proteins needed for essential plastid functions must cross either three or four membranes to reach their final destination in the plastid stroma (Bolte, *et al.,* 2009). The thylakoid membrane can be considered an additional membrane barrier, if the protein is destined for the thylakoid lumen. Although protein import into complex plastids is much more complicated than import into primary plastids. Recent investigations into the cellular mechanisms involved have nonetheless resulted in a preliminary albeit incomplete picture of the evolution of these import machineries (Bolte, *et al.,* 2009).

This membrane complexity raises the question of how the plastid proteins are targeted to and imported into the organelle, and the complex routes of plastid targeting that must exist in this organism (Dion, *et al.,* 2006). Many proteins synthesized in the cytoplasm are post-translationally directed to their final destination by crossing membranes, which surround specific compartments (Bolte, *et al.,* 2009). Most have to be synthesized as preproteins with one or more topogenic signals, which act like addresses for identification and targeting by specific receptors. The complexes formed in this process are directed to the target compartment and then transported by way of compartment-specific translocons. During symbiogenesis, new compartments originated, and as in mitochondria and plastids, symbiont nucleic acids were mixed intracellularly. Transport of nucleus-encoded proteins into the symbiont is a prerequisite for the maintenance of the partnership.

In *E. gracilis*, as in many other Euglenoids, pre-protein classification is generally divided into three classes (Durnford and Gray 2006; Patron et al. 2005). Class I pre-proteins comprises of tripartite targeting sequences, an N-terminal topogenic signal consisting of a signal peptide at the N-terminus, followed by a transit peptide domain and a predicted stop transfer sequence (hydrophobic region) in front of the coding sequence for the mature protein (Bolte, *et al.,* 2009). Class II pre-proteins; however, contain a bipartite signal sequence without the putative stop-transfer sequence

(hydrophobic region). They also possess only a signal sequence at the N terminus, followed by what, in amino acid composition, resembles a plastid transit peptide. A few unrelated plastid-targeted proteins also exhibit highly similar transit sequences, implying either a recent swapping of these domains or a conserved function (Dion, *et al.,* 2006). Despite the similarities in their sequences, pre-proteins crossing the thylakoid membrane, in addition to the three plastid-surrounding membranes, seemingly belonging to a third group, are nonetheless classified as Class Ib proteins (Durnford and Gray, 2006; Patron, *et al.,* 2005). Proteins related to these latter classes possess an additional region, which seems to be required for import into thylakoids (Inagaki, *et al.,* 2000). The transport across the plastid envelop as well as the intra-plastidal transport may be directly comparable to the situation in primary plastids (see Inagaki, *et al.,* 2000; Nassoury, *et al.,* 2003; Slavikova, *et al.,* 2005; Sulli, *et al.*, 1999; Chaal and Green, 2005; Dooren, *et al.,* 2001; Bolte, *et al.*, 2009; Osafune, *et al.,* 1991 for details and import experiments of Class I and Class II proteins).

Light-harvesting complexes (LHCs) are a superfamily of chlorophyll- and carotenoid-binding proteins that are responsible for the capture of light energy and its transfer to the photosynthetic reaction centers in *E. gracilis* (Koziol and Durnford, 2007). Unlike those of most eukaryotes, the LHCs of *E. gracilis* are translated from large mRNAs, producing polyprotein precursors consisting of multiple concatenated LHC subunits that are separated by conserved decapeptide linkers. These precursors are posttranslationally targeted to the chloroplast and cleaved into individual proteins. Several of the individual LHC-coding subunits both within and between transcriptional units evolve in concert, suggesting that gene conversion has been a significant mechanism for LHC evolution in *E. gracilis* (Koziol and Durnford, 2007). These polyproteins have previously been classified as LHCI or LHCII based upon whether their LHC subunits are predicted to associate exclusively with photosystem I (Houlne and Schantz, 1988) or with photosystem II (Houlne and Schantz, 1987; Muchhal and Schwartzbach, 1992). Additionally, both *E. gracilis* and dinoflagellates encode multiple nuclear-encoded plastid-targeted proteins translated as polyprotein precursors. These polyproteins are routed through the endomembrane system (Osafune, *et al.,* 1990) and directed to the chloroplasts via complex N-terminal targeting sequences (Durnford and Gray, 2006). Once in the chloroplast, the

individual proteins are liberated through proteolytic cleavage of conserved decapeptide linkers (polylinkers) (Muchhal and Schwartzbach, 1992; Hiller, *et al.,* 1995) as previously mentioned.

Two types of RNA polymerases are known to exist in chloroplasts (Greenberg, 1984): one is a single subunit, nuclear encoded enzyme related to various phage RNA polymerases, and multi-subunit eubacterial RNA polymerases. Genes for four subunits, *rpoA*, *rpoB*, *rpoC1* and *rpoC2*, are normally encoded by the plastid genome (Hess and Borner, 1999). When the sequence of the *E. gracilis* chloroplast genome was reported in 1993 the α-subunit gene (rpoA) of RNA polymerase appeared to be missing, based on a comparison of all putative reading frames to the then known rpoA loci. The rpoA gene products in *E. gracilis* appear to be the most variable in this gene family when compared to the rpoA gene in other species of bacteria, algae and plants. Additionally, *E. gracilis* rpoA proteins lack a C-terminal domain required for interaction with some regulatory proteins, a feature shared only with some chlorophyte green algae. The *E. gracilis* rpoA gene is the distal cistron of a multigene cluster that includes genes for carbohydrate biosynthesis, photosynthetic electron transport, an antenna complex and ribosomal proteins (Sheveleva, *et al.,* 2002).

The light-induced development of proplastids into chloroplasts is a classic example of subcellular differentiation (Pathier, 1982) and protein turnover (Cushman and Price, 1986). Extremely low rates of protein synthesis by isolated proplastids from *E. gracilis* have also been reported (Miller, *et al.,* 1983). ATP combined with equimolar $Mg^{2+}$ are capable of stimulating translation rates in pea chloroplasts in the dark to levels equal to or greater than, those observed in the light (Fish, *et al.,* 1983; Navison, 1984). In *E. gracilis* proplastids supplemented with ATP and $Mg^{2+}$ and other external factors, can synthesize proteins (large subunit of ribulose-1,5-bisphosphate carboxylase (LS)) at rates about half those of light-driven chloroplasts and most translation products of proplastids are distinct from those of mature chloroplasts (Cushman and Price, 1986). The primary limiting factor in proplastid protein synthesis is energy (light, and other supplementary and inorganic elements such as amino acid, $Mg^{2+}$, $K^{2+}$) and not in the availability of transcripts (Dehesh and Apel, 1983; Liu and Jagendorf, 1984; Malek, *et al.,* 1984; Cushman and Price, 1986).

*Translation, ribosomal structure and biogenesis, RNA processing, and modification*

A comparison by two-dimensional gel electrophoresis of the most abundant in vivo and in vitro *Euglena* translation products indicates that translation rather than transcription is the major site of photoregulation (Monroy, *et al.,* 1987; McCarthy and Schwartzbach, 1984; Bonham-Smith and Bourque, 1989; Koo and Spremull, 1994; Kozak, 1989). Light Harvesting Chlorophyll a/b-binding protein of photosystem II(LHCPII) is one of the light-induced *E. gracilis* proteins whose synthesis is controlled at the translational rather than the transcriptional level – this phenomenon is also evident in this work (see Chapter 6 of this work). LHCPII mRNA is translated on membrane-bound polysomes in *E. gracilis*, and light regulates pLHCPII (LHCPII polysomes) synthesis at the level of polypeptide chain elongation (Kishore and Schwartzbach, 1992). Similarly, little is known of the interdependence and coordinated regulation of plastid and nuclear gene expression during light-induced chloroplast development (Parthier, 1982). Detailed analysis at the protein and RNA levels (Monroy and Schwartzbach, 1984; Monroy, *et al.,* 1986; 1987; McCarthy and Schwartzbach, 1984; Krauspe, *et al.,* 1987) merely revealed that nuclear and plastid genes for chloroplast proteins are differentially expressed during the transformation of proplastids to chloroplasts (as previously mentioned; see Chapter 6 of this work). Changes in other plastid constituents like galacto- or sulfo- lipids (Pohl, 1973), chlorophylls (Kindman, *et al.,* 1978; Gomez-Silva, *et al.,* 1985) and carotenoids (Dolphin, 1970; Cunningham, 1986) emphasized the complexity of the biochemical processes. They only reflect but do not explain how the expression of plastid and nuclear genes is coordinated (Reinbothe and Parthier, 1990). Scholarly evidences suggests that, in *E. gracilis*, translational controls and posttranslational events may also play a role in regulating the synthesis and accumulation of chloroplast localized proteins (Devic and Schantz, 1983; Duberttret and Lefort-Tran, 1982; Mccarthy and Schwartbach, 1984; Miller, *et al.*, 1983; Reardon and Price, 1983) as also evident in this work.

### *1.2.3 Mitochondria*

#### *1.2.3.1 The mitochondrial machinery*

The mitochondria, much like the chloroplasts, contain their own genetic system, and grow in a coordinated process that requires the contribution of two separate genetic systems - one in the organelle and one in the cell nucleus. Most of the proteins in

these organelles are encoded by nuclear DNA, synthesized in the cytosol, and then imported individually into the organelle, and are inherited by non Mendelian-mechanism (Alberts, *et al.,* 2002). The biosynthesis of new mitochondria, much like chloroplasts, requires lipids in addition to nucleic acids and proteins. Mitochondria, in contrast to chloroplasts, import most of their lipids. In animal cells, the phospholipids phosphatidylcholine and phosphatidylserine are synthesized in the endoplasmic reticulum and then transferred to the outer membrane of mitochondria (Alberts, *et al.,* 2002)

*1.2.3.2 Mitochondrial genome structure*

Mitochondrial genomes are unique and diverse across the various eukaryotic lineages. In most animals, fungi and plants, mitochondrial DNAs (mtDNAs) consist of a single chromosome whose physical structure is monomeric circular, or multimeric linear (but circular mapping; Nosek and Tomaska, 2003). Electron microscopic observations have suggested that virtually all of the *E. gracilis* mitochondrial genome is in the form of small linear pieces of average length 0.9 – 1 µm (Spencer and Gray, 2012). Sizes of *E. gracilis* mtDNA vary from ~15 - 20 kbp to several hundred kbp. The mitochondrial genomes (mtDNA) of the kinetoplastids are referred to as the kinetoplast DNA (kDNA), consisting of a maxicircle (present in a dozen copies) and numerous minicircles (present in several thousand copies) (Lukes, *et al.,* 2005). The physical and spatial structure of kinetoplastid mtDNAs is diverse; designated as eu-kDNA in *Trypanosoma* and *Leishmania*, pro-kDNA in *Bodo saltans*, Pan-kDNA in *Bodo caudatus*, poly-kDNA in *Dimastigella mimosa*, *D. trypaniformis* and *Cruzella marina*, and finally, mega-kDNA in *Trypanoplasma* and *Jarellia* (Lukes, *et al.* 2002; Joannie, *et al.,* 2007). In a number of characteristics, *E. gracilis* mtDNA more closely resembles the mitochondrial genome found in dinoflagellates, an apparent evolutionary convergence. Nevertheless, the presence of dispersed gene fragments and their organization in *E. gracilis* mtDNA do suggest a model for the evolutionary emergence of coding sequences for the guide-type RNAs that function in mitochondrial U insertion/ deletion editing in the kinetoplastids (Spencer and Gray, 2012).

RNA species of 14S and 11S have been identified in *E. gracilis* mitochondria (Krawiec and Eisenstadt, 1970b) and shown to hybridize to the mtDNA of this

organism (Crouse, *et al.,* 1974); in contrast, 21S and 16S rRNA species having a very low G + C content (30%) have been isolated from 71S mitochondrial ribosomes composed of 50S and 32S subunits (Avadhani and Buetow, 1972). Apart from these observations, nothing is known about mitochondrial rRNA gene organization or expression in euglenid flagellates (Spencer and Gray, 2012). The *Euglena* mitochondrial genome does not appear as a compact and homogeneous structure and it possess high A+T content (about 76%) whereas this value is less than 50% in nuclear DNA, but Rawson, *et al.,* 1979, reported an overall base composition GC content of 48 %.

In the past, the mitochondrial genomes of euglenids are poorly characterized, due to several reasons, however, this has now fully been characterized (Dobakova, *et al.,* 2015). In the model organism, *E. gracilis*, mtDNA structure and size had remained elusive despite intensive studies (Buetow, 1989; Gray, *et al.*, 2004; Manning, *et al.,* 1971; Nass *et al.*, 1974; Talen, *et al.,* 1974; Yasuhira and Simpson, 1997; Lukes, *et al.,* 2005), and for the other euglenids, there is lack of established culture conditions, axenic strains, or specialized organelle isolation protocols (Joannie, *et al.,* 2007) to propel sufficient investigations. Recent observations suggest that the unusual mitochondrial genome organization seen in Kinetoplastids might occur in other Euglenozoa as well (Joannie, *et al.,* 2007), including *E. gracilis*. Earlier studies described a single (‘giant’) mitochondrion in *E. gracilis* (Pellegrini, 1980b). Previously, the only mitochondrial genes identified in *E. gracilis* encode fragmented LSU and SSU rRNAs (Gray, *et al.,* 2004; Buetow, 1989) and subunit 1 of cytochrome oxidase (Tessier, *et al.,* 1997; Yasuhira and Simpson, 1997). The architecture of *E. gracilis* mtDNA has now been described (Perez, *et al.,* 2014, Dobakova, et al., 2015).

### *1.2.3.3 Evolution of the mitochondrial genome*

In recent years, alternatives to the exclusively adaptationist evolutionary approach favored by many molecular biologists have become popular (Koonin, 2009). There is a growing understanding that mutation bias, drift and selectively neutral ‘ratchets’, rather than positive selection, may underpin the main evolutionary mechanisms for genomes, multi-molecular machines, and regulatory networks (Covello and Gray, 1993; Gray, *et al.,* 2010; Lukes, *et al.,* 2011; Lynch, 2007; Lynch and Conery, 2003;

Stoltzfus, 1999). It's been proposed that the model of constructive neutral evolution (Covello and Gray, 1993; Stoltzfus, 1999), recently expanded as an explanation for the seemingly 'irremediable complexity' of the molecular systems and machines characteristic of eukaryotes (Gray, *et al.,* 2010; Lukes *et al.,* 2011) is particularly well suited to explain the explosion of forms and complexities of mitochondrial genomes and their expression in euglenozoans.

The Phylum Euglenozoa comprises three groups of eukaryotic organisms (kinetoplastids, diplonemids, and euglenids; see sub-section 1.1), the mitochondrial (mt) genomes of which exhibit radically different modes of organization and expression. Gene fragmentation is a striking features of both euglenid and diplonemid mtDNAs (Flegontov, *et al.,* 2011). To rationalize the emergence of these highly divergent mtDNA types and the existence of insertion/deletion RNA editing (in kinetoplastids) and trans-splicing (in diplonemids), Flegontov, *et al.,* 2011, proposed that in the mitochondrion of the common evolutionary ancestor of Euglenozoa, small expressed gene fragments promoted a rampant neutral evolutionary pathway. And that it is safe to predict that the biochemical mechanisms operating on mtDNA and mtRNA in diplonemids and euglenids may even surpass in complexity the biochemical mechanisms underlying kinetoplastid RNA editing and kDNA replication (see Flegontov, *et al.,* 2011 for an extensive review on the adaptationist evolutionary approach and constructive neutral evolution).

*Common and uncommon patterns: Trypanosomatids, Naegleria, Giardia, and other super groups*

*A. Respiratory complexes*

Numerous additional subunits of the respiratory-chain complexes have been described in *Trypanosoma brucei* and *Trypanosoma cruzi*, and had been thought to have no counterparts in other organisms. They were interpreted as potentially associated with the parasitic trypanosome lifestyle. However, recent investigation by Perez, *et al.,* 2014, using two complementary approaches to characterize the subunit composition of respiratory complexes in *E. gracilis* revealed that at least half of the subunits recently reported in *T. brucei* and *T. cruzi* are actually not specific to Trypanosomatidae, but extend at least to other excavates (*Naegleria gruberi* and *E. gracilis*), and that their origin and function are thus not specifically associated with

the parasitic lifestyle. *E. gracilis* mitochondria contain a minimum of 40 subunits previously identified only in parasitic Trypanosomatidae (kinetoplastids). However, many of these subunits are not found beyond Euglenozoa (or Discoba), suggesting the biochemical peculiarities observed for the respiratory-chain complexes of kinetoplastids and euglenids (Perez, *et al.,* 2014).

The sequence of the *E. gracilis* cox1 gene, encoding subunit 1 of cytochrome oxidase, has been reported (Tessier, *et al.,* 1997; Yasuhira and Simpson, 1997), and the sequence containing cox2 (cytochrome oxidase subunit 2). The sequence information indicates the likely absence of mitochondrial U insertion/deletion editing in this organism and apparent use of the standard genetic code in mitochondrial translation. There is radically different pattern of mitochondrial genome organization in the euglenid, *E. gracilis*, compared with its kinetoplastid relatives (Spencer and Gray, 2012). COX 1 is one of the mitochondrion genes encoding subunit 1 of cytochrome c oxidase in *E. gracilis*; the nucleotide sequence does not exhibit any intron, and an amino-acid alignment of the *E. graclis* COX1 with homologous proteins shows that the universal genetic code is used (Tessier, et al., 1997). Comparisons of the genomic and cDNA sequences of *Euglena cox1* indicate that the transcript does not undergo RNA editing as found in trypanosomes and in higher plants. The phylogeny obtained using COX1 protein sequences shows similarity with that obtained with nuclear rRNA sequences and places *Euglena* and *Trypanosoma* far apart from other eukaryotes (Tessier*, et al.,* 1997). The size of the *E. gracilis* mitochondrial genome has been predicted to be between 5 kbp – 8 kbp (Dobáková, et al., 2015). The presence of a single mitochondrial reticulum has been assessed by electron microscopy studies, and the structure of this reticulum has been proven to vary noticeably with growth conditions (Buetow, 1989).

### *B. Glycolytic pathway*

The Embden-Meyerhof-Parnas pathway of glycolysis is the backbone of energy metabolism (ATP synthesis) in eukaryotes. Enolase (2-phospho-D-glycerate hydrolase; EC 4.2.1.11), a highly conserved enzyme across eukaryotes and prokaryotes, and catalyzes the forward and reverse reaction during gluconeogenesis (Hannaert, *et al.,* 2000). In all eukaryotes and many prokaryotes, enolase is biologically active as a dimer. Catalytically active dimers may be generated with

products of different enolase genes, for example, in *Saccharomyces cerevisiae* (McAlister and Holland, 1982). In the vast majority of eukaryotes, such as *Giardia lamblia,* studied to date, glycolysis is a cytosolic pathway and compartmentalized (Muller, 1998; Chavez, Balamuth and Gong, 1986; Brugerolle, 1993).

However, among other protists, there are some notable exceptions to be found with regard to the compartmentation of glycolysis. In *Chlamydomonas reinhardtii,* for example, all glycolytic enzyme activites studied were found to be localized in the chloroplast, rather than in the cytosol (Schnarrenberger, *et al.,* 1990). A better-known and well-studied exception is the Kinetoplastids (trypanosomes and sister organisms), in which most of the glycolytic pathway is compartmentalized in specialized microbodies, glycosomes (Opperdoes and Borst, 1977; Hannaert and Michels, 1994; Clayton and Michels, 1996). In these organisms, only the last three enzymes of the pathway leading to pyruvate production, including enolase, are found in the cytosol. In relatives of the Kinetoplastids, the Euglenids, which possess plastids but not glycosomes, some of the enzymes that are common to glycolysis and the Calvin cycle occur as distinct chloroplast-cytosol isoenzyme pairs (Kitaoka, *et al.,* 1989; Henze, *et al.*, 1995; Plaumann, *et al.,* 1997), as is also found in many higher plants (Martin and Schnarrenberger, 1997).

Several key glycolytic enzymes from trypanosomes and amitochondriate protists differ in their regulatory properties from those of other eukaryotes, e.g., yeast or vertebrates, and these differences appear to correlate with the lifestyle of these organisms (Mertens, *et al.,* 1992; Mertens 1993; Hannaert and Michels, 1994; Michels and Hannaert, 1994; Clayton and Michels, 1996; Bakker, *et al.*, 1997; Park, *et al.,* 1997). In *E. gracilis,* the glycolytic pathways differ from Kinetoplastids not only with respect to compartmentation, but also with respect to regulation of the enzymes involved (Kitaoka, *et al.*, 1989; Bakker, *et al.,* 1997). Molecular analysis of cytosolic enolase from the amitochondriate protist *M. balamuthi*, and *T. brucei* and two isoenzymes of enolase from *E. gracilis* have been reported, with *E. gracilis* possessing a putative N-terminal plastid-targeting peptide. Enolase is an exception among eukaryotic glycolytic enzymes in that it does not show markedly more similarity to eubacterial homologs than it does to archaebacterial homologs, except the enolases from *E. gracilis* (Hannaert, *et al.,* 2000).

*C. Predicted mitochondria targeting presequences*

The *E. gracilis* mitochondrial genome is represented by a heterodisperse collection of short molecules (approximately 4 kb) encoding gene fragments flanked by repeats (Spencer and Gray, 2011). The nucleus-encoded mitochondrial precursor proteins in *E. gracilis* and trypanosomatids apparently possess presequences sharing common features. *E. gracilis* possess homologs of mitochondria-targeted proteins from other organisms (*T. brucei*, *T. cruzi* and *L. major*), and other trypanosomatid mitochondrial protein precursor (e.g., those involved in RNA editing). Mitochondrial presequences of *E. gracilis* and these trypanosomatids are highly variable in sequence length (5–118 aa), but share statistically significant similarities and this is probably responsible for recognition via import apparatus of mitochondrial outer membrane (Krnacova, *et al.,* 2012). In Kinetoplastids, mitochondrial transcripts encoded by maxicircles are edited by guide RNAs (gRNAs) encoded by minicircles (Hajduk, *et al.* 1993), although in some trypanosomatids a small proportion of gRNAs is also encoded by maxicircles (Simpson, *et al.,* 2000; Stuart and Panigrahi, 2002). The RNA editing in mitochondria of Kinetoplastids includes uridine insertions and deletions (Krnacova, *et al.,* 2012).

The engulfment of α-proteobacterial ancestor of mitochondria by a host entity was probably a key moment in eukaryogenesis (Martin and Muller, 1998; Vesteg and Krajcovic, 2008, 2011). The acquisition of α-proteobacterium and its evolution to a primitive mitochondrion was accompanied by the transfer of endosymbiont genes to the host genome and the evolution of a mechanism for import of proteins to mitochondria including the evolution of mitochondria-targeting presequences. Most of the proteins necessary for mitochondrial function were probably nucleus-encoded in LECA (Desmond, *et al.*, 2011). The potentially primitive mitochondrial import apparatus in a potentially early branching eukaryotic group (either euglenozoans or excavates) could be a good model to trace the evolution of mitochondrial import mechanism of the first eukaryote. However, the data about mitochondrial targeting presequences of Euglenozoa are fragmentary (Krnacova, *et al.,* 2012), and hopefully the Euglena genome data will help to address many of these gaps.

### *1.2.3.4 Mitochondrial genome function*

*Enzymology and pathways*

The glyoxylate cycle, amongst 17 other experimentally determined pathways present within *E. gracilis* (see Caspi, *et al.,* 2010 for the remaining 16 pathways), is a modified form of the tricarboxylic acid cycle, which enables organisms to synthesize carbohydrates from C2 compounds (Nakazawa, *et al.,* 2011). In *E. gracilis*, the key enzyme activities of the glyoxylate cycle, isocitrate lyase (ICL) and malate synthase (MS), are conferred by a single bifunctional protein named glyoxylate cycle enzyme (*Euglena gracilis* glyoxylate cycle enzyme [EgGCE]) with a 62-kDa N-terminal and 67-kDa C-terminal domains respectively. The N-terminal residues of EgGCE are critical for both the ICL and MS activities. The mitochondrion of *E. gracilis* is a facultatively anaerobic organelle that produces ATP in the presence and absence of $O_2$. Unlike the mitochondria of most eukaryotes, which have a pyruvate dehydrogenase multienzyme complex, the mitochondrion of *E. gracilis* has an unusual $O_2$-sensitive enzyme for the oxidative decarboxylation of pyruvate (i.e. pyruvate: NADP1 oxidoreductase) to produce acetyl-CoA (Nakazawa, *et al.,* 2000). Acetyl-CoA regulates the ICL reaction by binding to a site other than the catalytic center of the MS reaction (Caspi, *et al.,* 2010).

*E. gracilis,* grown aerobically*,* expresses pyruvate dehydrogenase (PDH) in mitochondria (Hoffmeister M, *et al.* 2004) and respires $O_2$ but by using a slightly modified Krebs cycle that is also found among some α-proteobacteria (Green, *et al.,* 2000) as a mechanism for energy production and conversion. The process involves the replacement of α-ketoglutarate dehydrogenase by α-ketoglutarate decarboxylase and succinate semialdehyde dehydrogenase (Buetow, 1989). When grown under anaerobic conditions *E. gracilis* uses acetyl-CoA as the terminal electron acceptor, produces wax esters as end products of metabolism (Green, *et al.,* 2000; Inui, *et al.,* 1982, 1983, 1984; Tucci, *et al.,* 2010; Teerawanichpan, 2010) and expresses pyruvate:NADP+ oxidoreductase (PNO) (Inui, *et al.,* 1985, 1987; 1991), which performs the oxidative decarboxylation of pyruvate (Nakazawa, *et al.,* 2003; Rotte, *et al.,* 2001; Rotte, *et al.,* 2001; Hug, *et al.,* 2010; Horner, *et al.,* 1999; Rotter, *et al.,* 2001). Upon a return to an oxic environment, the stored waxes are degraded via aerobic dissimilation in the mitochondrion (Inui, *et al.,* 1982). Similar to the situation for anaerobic mitochondria of metazoa, wax ester fermentation of *E. gracilis* involves mitochondrial fumarate reduction and thus utilizes rhodoquinone (RQ) (Hoffmeister, *et al.*, 2004) for the synthesis of propionyl-CoA, although reduced RQ can also

donate electrons to other components of the *E. gracilis* mitochondrial respiratory chain (Castro-Guerrero, *et al.,* 2005), including the cytochrome *bc*1 complex and the alternative oxidase (Castro-Guerrero, *et al.,* 2004).

In *E. gracilis*, sulphate metabolism is linked to energy metabolism. The mitochondria contain the enzymes capable of sulphate activation (ATP sulphurylase, EC 2.7.7.4; adenylylsulphate kinase, EC 2.7.1.25, and inorganic pyrophosphatase, EC 3.6.1.1) and reduction (Brunold and Schiff, 1976; Saidha, *et al.,* 1985, 1988;). Sulphate metabolizing centre in *E. gracilis* is located outside of the mitochondrial inner membrane. This centre appears to supply the mitochondrion and the rest of the cell with the products of sulphate activation as well as with reduced sulphur in the form of cystein (Saidha, *et al.,* 1985, 1998). This amino acid is incorporated into mitochondrial proteins (Saidha and Schiff, 1986), and sulphate reduction is caused by inorganic phosphates such as $KH_2PO_4$ (Saidha, *et al.,* 1988).

## 1.3 OTHER INTRA- AND EXTRA- CELLULAR STRUCTURES AND FUNCTIONS

### *1.3.1 Motor apparatus and photoreceptors*

In *E. gracilis*, the structure of the motor apparatus consists of the canal (gullet) and reservoir, contractile vacuole, flagellar apparatus, basal bodies and related structures, transition zone, axonemes, paraflagellar rod, paraflagellar membrane, and Mastigonemes (Flagellar Hairs, Hairlike Appendages), while the structure of the photoreceptors consists of the stigma, and the paraflagellar swelling (PFS) or body (PFB) (Figure, 1.1, 1.3). The flagellar swelling plays significant role as a primary photic organelle (Lenci and Ghetti, 1989), however, two theories on the transduction mechanisms seem to have consistency, and have been refered to as the synapsis theory and the piezoelectric theory. While the former entails the role of a synapsis between the two *E. gracilis* flagella (emergent and non emergent) or between the membranes of PFS and the stigma, the latter deals with the alteration of cation flux from segment to segment of the quasi-crystalline microtubules of the emergent flagellum, driven by piezoelectric currents generated in the bending and straightening of flagellar segments in sequence which can constitute a possible mechanism of impulse transfer. (see Giovanna, *et al.,* 1991 for a detailed review of the *E. gracilis* motor apparatus and photoreceptors).

The configuration of the *E. gracilis* eyespot, photoreceptor and flagellum represents a simple but complete visual system. The photoreceptor is a three-dimensional natural crystal of about 1 x 0.7 x 0.7 µm in dimension. This crystal is made up of a single protein and can be interpreted as a three-dimensional crystal of type I (Michel, 1990), i.e. a stack of two-dimensional crystal protein layers characterized by in-plane hydrophobic interactions and held together by hydrophilic interactions as confirmed on the basis of cryo-fractured images (Walne, *et al.,* 1998). The purification of this protein that forms the photoreceptor of *E. gracilis* has also been reported. It is a 27 kDa protein (Erh) with a photocycle resembling that of sensory rhodopsin (rhodopsin-like protein), but with at least one stable intermediate, and a simple photocycle. This protein shows optical bistability, without thermal deactivation (Walne, *et al.,* 1998).

*E. gracilis* shows pronounced movement behavior towards light and gravity (phototaxis and gravitaxis respectively). Both reactions are based on physiological mechanisms and are triggered by blue photo-activated adenylyl cyclases (PAC *α* and PAC β), which synthesizes 3'-5'-cyclic adenosine monophosphate (cAMP) from ATP upon blue light absorption (see Daiker, *et al.,* 2011) and calcium stimulation (Richter, *et al.,* 2001b; Sakato, *et al.*, 2007). Calmodulin (CaM.2), and other related potential calcium sensors (Ebenezer, *et al.,* 2017), also play a crucial role in the signal transduction chain of gravitaxis in *E. gracilis* (see Toda, *et al.,* 1992; Lonergan, 1984, 1985, 1990; Fontana and Devreotes, 1984; Van Haastert, 1985; Pasquale and Goodenough, 1988, for detailed mechanisms and additional investigations on effects of calmodulin on *E. gracilis*). The proteins consist of two PACa ($M_r$ 105 kDa) and two PACb (*Mr* 90 kDa) subunits. The proteins are believed to be located in the paraxonemal body (PAB), along the total length of the flagellum. The photoactivated adenylyl cyclase (PAC) family genes responsible for the PAC proteins in the PAB and flagella are identical but divergent, as full sequence analysis using PCR and 3’ and 5’ RACE had indicated a substantial divergence between strains with a homology between 45 and 100%. It’s been hypothesized that in *E. gracilis* strains, there is a family of very dissimilar PAC proteins located in the PAB and the flagellum where they serve different functions in phototaxis and step-up photophobic reactions, however, their physiological roles have not yet been revealed (Ntefidou and Hader, 2005).

*1.3.2 The pellicular complex and cytoskeleton*

Eukaryotic cells are distinctive in possessing a nucleus, an endomembrane system, a complex cytoskeleton and mitochondria (or hydrogenosomes), which are modern-day descendants of proteobacterial endosymbionts (Embly and Martin, 2006). The cytoskeleton of eukaryotic cells serves a variety of functions (see Huttenlauch and Stick, 2003), although some of these functions are absent in unicellular organisms, many Excavates are among those that show the most elaborated cytoarchitecture (Huttenlauch and Stick, 2003). Microfilaments, microtubules, and intermediate filaments (IF) are the principal cytoskeletal elements in most eukaryotic cells (Huttenlauch and Stick, 2003; Honts and Williams, 1990; Levy, *et al.,* 1996) (see Grain, 1986; Peck, 1977, 1986; Bouck and Ngo, 1996; Dubreuil and Bouck, 1985; Grain, 1986; Bricheux and Brugerolle, 1986; Dubreuil and Bouck, 1985; Huttenlauch and Peck, 1991; Peck, *et al.,* 1991; Vigue, *et al.,* 1984; Huttenlauch and Stick, 2003, for detailed investigations and reviews on the cortical cytoplasm and skeletal nature of the epiplasm). It has been proposed that endosymbiosis played significant role in the evolution of *E. gracilis* cytoskeleton. Once photosynthesis was established in a previously phagotrophic cell, the evolutionary pressures on the cytoskeletal systems involved in locomotion and feeding changed, and gave rise to fundamental modifications of cell structures found in the descendants of these complex cells (see Leander, *et al.,* 2007 for the review on the current understanding of the Euglenid cytoskeleton and it's evolution, and references to their sister relatives – the Trypanosomes).

One cytoskeletal feature distinguishes the euglenids from all other protists-like excavates and apicomplexa: a pellicle composed of longitudinal articulated strips that arise in the reservoir and extend the length of the cell immediately beneath the membrane (Arnott & Walne, 1967; Conforti & Tell, 1989). Deep-etching technique has revealed that the ultrastructure of the pellicle of *E.gracilis* are of four different structural levels: the cell membrane, the electron opaque layer organized in ridges and grooves or proteinaceous strips, the microtubular system, and the subpellicular tubular cisternae of the endoplasmic reticulum (see Leader, *et al.,* 2007; Leander and Farmer, 2000; 2000; Dubreuil, *et al.,* 1992; Vismara, *et al.,* 2000 for detailed ultrastructural analysis of the euglenoid pellicular complex). The euglenid pellicle proteinaceous strips are helically twisted. Helically arranged strips are associated

with cell plasticity (metaboly) (Pringsheim, 1948; Arnott & Walne, 1966 and 1967) (see Chapter 5 and Appendices on subsection, Surfaceome/Surface, in this work), which is facilitated by relative translational movements between adjacent strips (Angeler, *et al.,* 1999; Suzaki and Williamson, 1985, 1986; Galo and Schrevel, 1982). The underlying mechanism for euglenoid movement, however, is not well understood. Although experiments have shown that the pellicle microtubules are controlled by calcium stored in the ER and play an important role in changing cell shape. One of the major gaps in knowledge has to do with the precise distribution and morphogenesis of the pellicle microtubules during cytokinesis (Leader, *et al.,* 2007) as well as the nature, structure, function, and organization of the euglenoids pellicle despite employing diverse techniques (Kirk & Juniper, 1964; Hofmann & Bouck, 1976; Silverman & Hikida, 1976; Miller & Miller, 1978; Lefort-Tran, *et al.,* 1980; Hilenski & Walne, 1983; Dubreuil & Bouck, 1985; Suzaki & Williamson, 1985, 1986; Dawson, *et al.,* 1988; Dawson & Walne, 1991; Zakrys & Walne, 1998).

Euglenoid flagellates have striped surface structures comprising pellicles, which allow the cell shape to vary from rigid to flexible during the characteristic movement of the flagellates (as previously mentioned). In *E. gracilis*, the pellicular strip membranes are covered with paracrystalline arrays of a major integral membrane protein, IP39, a putative four-membrane-spanning protein with the conserved sequence motif of the PMP-22/EMP/MP20/Claudin superfamily. The three-dimensional structure of *Euglena* IP39 determined by electron crystallography have been reported by Suzuki, *et al.,* 2013. Two dimensional crystals of IP39 appear to form a striated pattern of antiparallel double-rows in which trimeric IP39 units are longitudinally polymerised, resulting in continuously extending zigzag-shaped lines. Structural analysis revealed an asymmetric molecular arrangement in the trimer, and suggested that at least four different interactions between neighbouring protomers are involved. A combination of such multiple interactions would be important for linear strand formation of membrane proteins in a lipid bilayer. They concluded that the four-transmembrane protein IP39 of *E. gracilis* forms strands by a trimeric unit repeat.

The cortex of ciliates, dinoflagellates and euglenoids comprises a unique structure called the epiplasm implicated in pattern processes of the cell cortex and in

maintaining cell shape (see Bouck and Ngô, 1996, for an extensive review). While there are differences in the structural organization of the epiplasm and cortex of these organisms, articulin is the principal constituent of the epiplasm in the euglenoid, *Euglena*, and the ciliate, *Pseudomicrothorax* (Huttenlauch, *et al.,* 1998). In paramecium, epiplasmins, a group of polypeptides with common biochemical properties are the major constituents of the epiplasm. These two proteins contribute to the organization of the membrane skeleton in these protists (see Huttenlauch, *et al.,* 1998, for an extensive discussion).

*1.3.3 Microtubules and the Golgi Apparatus*

Extra-cisternal substances, known as microtubules, exist within the Golgi apparatus and have been reported in various literatures (Morre, *et al.,* 1971; Mollenhauer, 1974). These substances occupy at least 20 - 40 % of the golgi apparatus volume and take the form of zones of exclusion (Mollenhauer & Morre, 1972, 1973; Morre, *et al.,* 1971; Mollenhauer, 1974), intercisternal elements (Cunningham, *et al.,* 1966; Mollenhauer, 1965; Mollenhauer and Morre, 1966, 1972, 1973; Turner and Whaley, 1965; Mollenhauer, 1974), bonding substances (Mollenhauer and Morre, 1966; Mollenhauer, *et al.,* 1971; Mollenhauer and Morre, 1972, 1973; Mollenhauer, 1974) and other elements or structures which have been reported (Amos and Grimstone, 1968; Kartenbeck and Franke, 1971; Mollenhauer, 1974). The existence of these various constituents cannot be doubted, but what each does functionally and the mechinisms that drives them are not yet fully understood. Since they surround the Golgi apparatus (Mollenhauer and Morre, 1972, 1973; Morre, *et al.* 1971), hold dictyosomes together (Mollenhauer & Morre, 1966; Mollenhauer *et al.* 1971; Mollenhauer & Morre, 1972, 1973), and alter or position dictyosome structure (Mollenhauer, 1974; Mollenhauer & Morre, 1972), they are believed to play a significant role in functional processes of the Golgi apparatus function (Mollenhauer, 1974). Microtubules could add an element of structural rigidity or anisotropy to the Golgi apparatus and act to position dictyosomes within the cell or organize partially its internal structure (Mollenhauer, 1974). Microtubules might also serve to guide or to transport soluble precursors into or out of the Golgi apparatus in a manner similar to that in neurons and/or other cells (Burton and Fernandez, 1973; Fernandez, Burton and Samson, 1971; Lane and Treherne, 1970; Yamada, *et al.,* 1971;

Mollenhauer, 1974) or to direct the movement of secretion vesicles from the Golgi apparatus (Mollenhauer, 1974).

## 1.4 METABOLISM

*E. gracilis* transports and metabolizes amino acids, primarily since they are not capable of utilizing nitrate, nitrite, and urea as nitrogen source. Therefore, ammonium is supplied as an N-source in the lab (provided as diammonium-hydrogenphosphate $(NH_4)_2HPO_4$) to cell cultures. While nitrate exerts low toxicity to organisms, ammonium is harmful for many aquatic organisms especially, at high pH-values, which causes the ionic $NH_4^+$ (low toxicity) to be partially transformed into the highly toxic ammonia, $NH_3$. In other reports, *E. gracilis* has been described to grow with various amino acids as sole N-source. Glycine, glutamine, glutamic acid, leucine, and threonine, and their uptake kinetic have been reported, and found to have the potential of being an alternative N-source for *E. gracilis* cultivation. Hence, these amino acids can be used as a non-toxic surrogate for $(NH_4)_2HPO_4$ (Richter, *et al.,* 2015).

*E. gracilis* represents an evolutionary intermediate in a metabolic transformation (coenzyme transport and metabolism) of a primary heterotroph to a photoautotroph through secondary endosymbiosis as mentioned previously. Koreny and Obornik, 2011, noted that *E. gracilis* possess a tetrapyrrole pathway, which serves as a highly informative marker for the evolution of plastids and plays a crucial role in the loss or gain of plastids since it is involved in chlorophyll formation in photoauthotrophs. Phylogenetic analyses and protein localization predictions have supported the presence of two separated tetrapyrrole pathways in *E. gracilis*. One of these pathways resembles the heme synthesis in primarily heterotrophic eukaryotes and was presumably present in the host cell prior to secondary endosymbiosis with a green alga. The second pathway is similar to the plastid-localized tetrapyrrole syntheses in plants and photosynthetic algae. It appears to be localized to the secondary plastid, presumably derived from an algal endosymbiont and probably serves only for the production of plastidial heme and chlorophyll (see Koreny and Obornik, 2011).

Inorganic ion transport and metabolism has also been reported in *E. gracilis*. It is well documented that most organisms are susceptible to heavy metals exposure,

affecting their growth, development, and morphology (Cervantes, *et al.,* 2001). However, some plants, bacteria and microalgae species are able to survive in heavy metal polluted environments by means of internal and/or external detoxification mechanisms such as: (1) diminished uptake; (2) internal binding; (3) biotransformation; (4) compartmentalization; and (5) external chelation (Yang, *et al.,* 2005). The effective onset of several resistance mechanisms of heavy metals such as chromium in *E. gracilis*, including synthesis of heavy metal-chelating molecules with thiol and carboxylate-groups, sub-cellular compartmentalization, metal chemical reduction and secretion of chelating molecules have also been reported (see Rodríguez-Zavala, *et al.,* 2007).

*E. gracilis* grows autotrophically, heterotrophically or myxotrophically (Dos Santos Ferreira, *et al.,* 2007; Matsuda, *et al.,* 2011) mainly via primary metabolites production (Bersanti, *et al.,* 2000; 2001; Rodriquez, *et al.,* 2010), but little is known about secondary metabolites biosynthesis (Tolivia, *et al.,* 2013). *E. gracilis* has a wide range of nutritional requirements, suggesting the existence of diverse physiological patterns, generating different metabolites and/or variation in the proportion they are biosynthesised. Two such biosynthetic metabolites are flavonoids and tannins, which are generally regarded to be bioactive and having free radical scavenging properties (Heim, 2002). When nutrients become scarce, *E. gracilis* cells enter stationary phase and develop a multiple-stress resistance response. The presence of flavonoids in the stationary phase may be associated to that response. In some *E. gracilis* strains (e.g. UTEX), tannins are produced in the exponential phase (Tolivia, 2013).

## 1.5 THE EVOLUTION OF PARASITISM

The most recent reviews on the evolution of parasitism is that as described by Lukes, *et al.,* 2014 and Jackson, *et al.,* 2015, 2016. Kinetoplastids are evolutionarily possibly more ancestral within the eukaryotic lineage compared to the majority of other groups of parasitic protists as well as very widespread and adaptable (Lukes, *et al.,* 2014). The proposal is that parasitism must have emerged separately in the kinetoplastid lineage, and excitingly challenging to identify genetic changes and/or inventions underlying this dramatic switch to a parasitic life style, and the Trypanosomatids might serve as a good model group for tracing the evolution of

parasitism. It is unclear how a once free living organism (Trypanosomatids) adopted a parasitic lifestyle (Lukes, *et al.,* 2014). However, this will be clear when the whole genomes for these sister clades of kinetoplastids are available (Lukes, *et al.,* 2014). Species-rich and morphologically diverse euglenids and diplonemids are almost exclusively free-living, so this life style was likely the ancestral state of early kinetoplastids (Lukes, *et al.,* 2014). Morphological and molecular evidence strongly support the origin of obligatory parasitic trypanosomatids from the clade of free-living *Bodo saltans* (Simpson, *et al.,* 2006; Jackson, *et al.,* 2008; Maslov, *et al.,* 2001; Stevens, 2014). Until recently, the most basal branch of this clade was the genus *Trypanosoma*, supporting the scenario in which ancestral trypanosomatids established themselves in vertebrates already at an early stage, and only subsequently invaded insects via their blood-sucking fellows (Minchin, 1908).

Monoxenous trypanosomatids of the genera *Crithidia*, *Leptomonas* and *Herpetomonas* can survive and multiply in their hosts anal scent glands. Monoxenous *species* could pre-adapt in this regions to the dixenous lifecycle because within the scent glands parasites are protected from the host immune system and the body temperature is lower (Deane and Jansen, 1988). Recent description of the likely monoxenous *Paratrypanosoma confusum*, which constitutes a well-supported branch between the free-living *B. saltans* on one side and *Trypanosoma* plus all other trypanosomatids on the other side (Votypka, *et al.,* 2012; Steven, 2014) supports a scenario, in which the ancestral flagellate first invaded insects or other invertebrate hosts and only subsequently, probably by blood feeding, entered vertebrates – a theory proposed by Léger in 1904. Comparative analysis of the genomes of *B. saltans* and *P. confusum* may shed key light on this dramatic change in life strategy (Lukes, *et al.,* 2014).

Within the Kinetoplastids, the Trypanosomatids might serve as a good model group for tracing the evolution of parasitism (Lukes, *et al.,* 2014). However, the solution comes in using modern methods of genome analysis which allows direct comparison of the gene content between different groups of Trypanosomatida, especially the dixenous group. By analyzing the genomic information from their free-living bodonid and euglenid relatives and early-branching trypanosomatid *P. confusum* (Flegontov, *et al.,* 2013), Parasitologist will be able to identify features shared between these

organisms as well as those representing key differences, especially in terms of gene loss and/or gain (Lukes, *et al.,* 2014, Jackson, *et al.,* 2016) (see Chapter 5 and Appendices on Orthologous Groups Clustering analysis). From the currently available data in major genome databases, most differences are associated with genes encoding metabolic and cell surface proteins (Lukes, *et al.,* 2014). The underlying assumption of this approach is that there must be some genes responsible for adaptation to the dixenous lifestyle. Those genes should exclusively be either present or absent in *Trypanosoma, Leishmania, and Phytomonas* as compared to their monoxenous kins (see Lukes, *et al.,* 2002, 2009, 2014, Waller and Jackson, 2009; Flegontov, *et al.,* 2011; Lui, *et al.,* 2006; Jensen and Englund, 2012; Noyes, 1998; Noyes, *et al.,* 1997; Croan, *et al.,* 1997 for recent reviews).

## 1.6 THE SURFACE OF KINETOPLASTIDS

In *E. gracilis*, the cell surface plays huge role in response to external factors. Similary, in kinetoplastids, it is crucial for mediating host-parasite interactions and is instrumental to the initiation, maintenance and severity of infection (Jackson, *et al.,* 2013, 2016). Within members of the kinetoplastids physiological similarities associated with shared ancestry are common, the cell-surface architectures are highly divergent, reflecting the evolution of specific mechanisms for immune evasion and survival by each parasite, and majority of the cell surface phylome are taxon or species-specific (see Jackson, *et al.,* 2013, 2016 for an extensive discussion).

## 1.7 PROJECT AIMS AND OBJECTIVES

To understand the biology of *E. gracilis*, next generation sequencing (NGS) and multi-omics approaches are being employed, focusing on four main aims and objectives:

i. To determine the genome and transcriptome architecture and content of *E. gracilis,* including ORF complexity, paralogy and orthology distribution, exon and intron size/length, tRNAs, rRNAs, and splice leaders locus.
ii. To characterize the protein-coding genes, and the biological interplay of this network in *E. gracilis*; catalogue genes, gene products, and putative surface proteomes - signal peptides and transmembrane proteins, assign functions, classify enzymes based on chemical reactions they catalyze,

delineate metabolisms and pathways (biological processes), and molecular functions.

iii. To comparatively investigate the *E. gracilis* sequenced genome with reference genomes, transcriptomes, and proteomes of other members of selected excavates (*T. brucei and L. major*) and chlorophytes (*C. reindhardtii* and *V. carteri*); elucidate possible insights into endosymbiosis and eukaryogenesis using core eukaryotic genes, conserved Trypanosomatids specific genes, and Chlorophytes specific genes.

iv. To determine *E. gracilis* responses to light and dark environmental conditions; investigate gene expressions in the transition from light to dark and dark to light; delineate posttranscriptional events and metabolic capacity.

# CHAPTER TWO

## 2.0 METHODS AND MATERIALS

### 2.1 CULTIVATION OF *Euglena gracilis*

The *E. gracilis* strain (Z) utilized in this project was the kind gift of William Martin (Heinrich Heine University, Düsseldorf). Cells for DNA and RNA sequencing projects (genome and transcriptome) were cultivated at room temperature under continous illumination in photosynthetic Hutner‘s media (Table 2.1, Hutner, 1966), and collected at exponential phase of growth between 7 x $10^5$ cells/ml and 9 x $10^5$ cells/ml, with cell density measured using a haemocytometer counting chamber. For light/dark experiments, cells were cultivated in heterotrophic Hutner’s media (Table 2.2, Hutner, 1966). To generate dark adapted cells for light/dark transcriptomics and proteomics experiments, cultures were grown in the absence of light for 16 days to produce bleached cells.

**Table 2.1:** Ingredients for phototrophic Hutner's media preparation for genome sequencing.

| S/N | Compound | Concentration (grams/litre) |
|---|---|---|
| *Essential ingredients* | | |
| 1. | $MgCO_3$ | 0.3 |
| 2. | $MgSO_4.3H_2O$ | 0.2 |
| 3. | $CaCO_3$ | 0.03 |
| 4. | HEDTA | 0.2 |
| 5. | $K_3$Citrate | 0.4 |
| 6. | Citric acid.H2O | 4 |
| 7. | $KH_2PO_4$ | 0.15 |
| 8. | L-Histidine HCl.$H_2O$ | 1 |
| 9. | $NH_4HCO_3$ | 0.5 |
| 10. | Thiamine HCl | 0.001 |
| 11. | Vitamin $B_{12}$ | 0.002 |
| 12. | pH | 3.2 – 3.5 |
| *Trace metals* | | |
| 13. | $Fe(NH_4)_2(SO_4)_2.H_2O$ | 0.042 |
| 14. | $MnSO_4.H_2O$ | 0.0155 |
| 15. | $ZnSO_4.7H_2O$ | 0.022 |
| 16. | $(NH_4)6MoO_{24}.4H_2O$ | 0.0036 |
| 17. | CuSO4 (anhydrous) | 0.001 |
| 18. | $Na_3VO_4$ | 0.0037 |
| 19. | $CoSO_4.7H_2O$ | 0.00048 |
| 20. | $H_3BO_3$ | 0.00057 |
| 21. | Pentaerythritol | 0.0888 |

**Note:** The table above describes the ingredients utilized for the phototrophic Hutner's media preparation (Hutner, 1966). The first column describes the serial number, the second column describes the compound, while the third column indicates the concentrations which are in grams/litre except for pH without a unit.

**Table 2.2:** Ingredients for heterotrophic Hutner's media preparation for light/dark experiment.

| S/N | Compound | Concentration (grams/litre) |
|---|---|---|
| *Essential ingredients* | | |
| 1. | $KH_2PO_4$ | 0.4 |
| 2. | $MgSO_4.3H_2O$ | 0.1 |
| 3. | $MgCO_3$ | 0.4 |
| 4. | $CaCO_3$ | 0.1 |
| 5. | DL-Malic acid | 5 |
| 6. | L-Glutamic acid | 5 |
| 7. | Glucose (anhydrous) | 10 |
| 8. | Urea | 0.4 |
| 9. | $Na_2$ Succinate.$6H_2O$ | 0.1 |
| 10. | Glycine | 2.5 |
| 11. | DL-Aspartic acid | 2 |
| 12. | Thiamine HCl | 0.0006 |
| 13. | Vitamin B12 | 0.0005 |
| 14. | pH | 3.1 – 3.4 |
| *Trace metals* | | |
| 15. | $Fe(NH_4)_2(SO_4)_2.H_2O$ | 0.021 |
| 16. | $MnSO_4.H_2O$ | 0.00775 |
| 17. | $ZnSO_4.7H_2O$ | 0.011 |
| 18. | $(NH_4)6MoO_{24}.4H_2O$ | 0.0018 |
| 19. | CuSO4 (anhydrous) | 0.0005 |
| 20. | $Na_3VO_4$ | 0.00185 |
| 21. | $CoSO_4.7H_2O$ | 0.00024 |
| 22. | $H_3BO_3$ | 0.000285 |
| 23. | Pentaerythritol | 0.0444 |

**Note:** The table above describes the ingredients utilized for the heterotrophic Hutner's media preparation (Hutner, 1966). The first column describes the serial number, the second column describes the compound, while the third column indicates the concentrations which are in grams/litre except for pH without a unit.

## 2.2 NUCLEIC ACID ISOLATION AND PURIFICATION

The *E. gracilis* gDNA isolation (version 0.7) was previously performed by Lebert, *et al.,* and Field, *et al.,* (unpublished data) using the methods of Medina-Acosta and Cross (Cross, Wieland *et al*., 1993) and TELT (Tris, LiCL, EDTA, and Triton X). *E. gracilis* genomic DNA (versions 0.8 and 0.9) was isolated using the Qiagen purification system to obtain low and high molecular weight DNA for generating Illumina paired-end and mate-pair read libraries (100bp paired-end libraries with insert sizes of 170bp, 500bp, 800bp, and mate-pair libraries with insert sizes of 2kbp, 5kbp and 40kbp) as well as PacBio libraries. The protocol for obtaining DNA fragments for the shorter-length libraries differed from that used for the 40 kb insert library. For the shorter length libraries (≤ 5kbp), cells were harvested by centrifugation for 10 min at 1000g, and DNA was extracted using the Qiagen DNAeasy extraction kit for total genomic DNA - the DNAeasy blood and tissue kit (Catalogue No. 69504) was utilized. The procedure designed for cultured animal cells was modified and involved firstly, using 1 x $10^7$ cells and secondly, prior to adding Buffer AL, 200 µl of RNase A was added to eliminate RNA contamination. Immediately after the wash step with Buffer AW2, centrifugation was performed for 1 min at 20,000 g to eliminate traces of ethanol. To obtain high molecular weight DNA fragments for the ≥ 40kb insert size library, the Qiagen Genomic-DNA isolation kit (Blood and Cell Culture DNA kit - Maxi, Catalogue No. 13362) was used. In this case 1 x $10^8$ cells were harvested. Prior to adding Buffer C1, samples were ground in liquid nitrogen using a ball mill at 300 rpm for 3 mins. Four wash steps were performed to remove contaminants including traces of RNA. To determine/estimate molecular weight, 400 ng/µl of DNA was loaded onto a 0.45 % agarose gel in TAE buffer/Ethidium bromide system, stained with Thermo Scientific 6X Orange Loading Dye, and electrophoresed at 80 V for 2 hours. A NanoDrop spectrophotometer was used to determine concentration and purity. Total RNA was isolated using the Qiagen RNeasy Mini kit (Cat. No. 74104). The protocol for the purification of total RNA from animal cells using spin technology was employed with minor modifications: 40 µl of DNase I stock solution was added to the mixture to sufficiently remove any traces of DNA. RNA yield, purity, and concentration were determined as previously described above for the DNA.

## 2.3 LIBRARY PREPARATION AND SEQUENCING

Genome and transcriptome library preparation and sequencing data were performed at the Beijing Genomic Institute (BGI), Hong Kong respectively, using Illumina Genome Analyzer HiSeq 2000 and MiSeq. Paired-end genomic sequence of multiple read lengths (49 bp to 300 bp) corresponding to five insert size libraries (170 bp, 250 bp, 500 bp, 540 bp, 800 bp, 2kbp, 5 kbp, and 40 kbp) were generated with a combined length of ~ 44 Gbp. Additional PacBio libraries were generated by Peter Myler (University of Washington) and Purificación López-García and David Moreira, while the Roche454 read libraries were generated by Michael Ginger. A combined total of 914,547 PacBio sequenced reads were generated with estimated average length of 9560.59 and an estimated coverage of ~ 4.4X (see Table 3.1 and 3.2, for details on genome and transcriptome sequence data).

## 2.4 GENOME AND TRANSCRIPTOME ASSEMBLY

### *2.4.1 Quality controls and assembly*

#### Genome

This analysis was carried out in collaboration with Dr. Steve Kelly at the Department of Plant Sciences, University of Oxford, UK, were I also spent 5 months (May – September 2014) assembling the initial draft of the genome. Several attempts to generate a genome assembly, using different datasets and informatics routes were explored (Ebenezer, *et al.,* 2017) (Figure 2.1, 2.2). The most successful strategy, as assessed by the Core Eukaryotic Gene Mapping Approach (CEGMA) and the percentage of the RNAseq reads from the transcriptome that mapped to the assembly (Hornett and Wheat, 2012; O'Neil and Emrich, 2013), utilised Platanus, SSPACE, and SGA (Figure 2.2). Assembling the quality controlled error corrected reads with Platanus followed by subsequent rounds of scaffolding with SSPACE and gapfilling using SGA resulted in a *de novo* asembly with a scaffold $N_{50}$ of 955 bp, comprising 2,066,288 scaffolds. Of note with this assembly was that ~ 87 % of the RNAseq reads from the transcriptome successfully mapped to the genome.

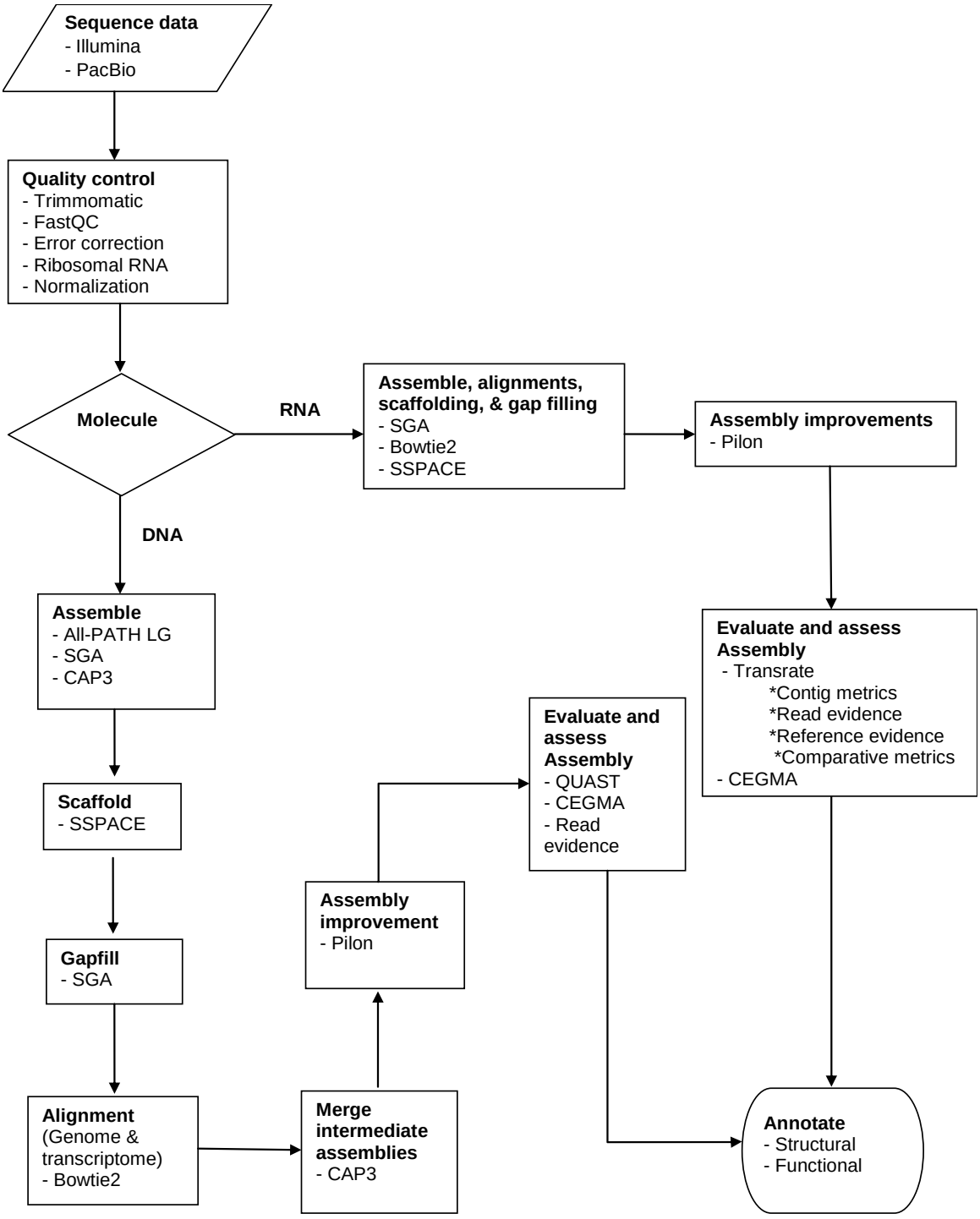



**Figure 2.1: The standard genome and transcriptome assembly pipeline involved several bioinformatics algorithms.** Overview of the actual genome and transcriptome assembly and improvements: The workflow involves the supply of sequence data, quality control, actual assembly, alignment and re-assembly, scaffolding, gap filling, assembly improvements, assembly evaluation and assessment, and production of data for structural and functional annotation.

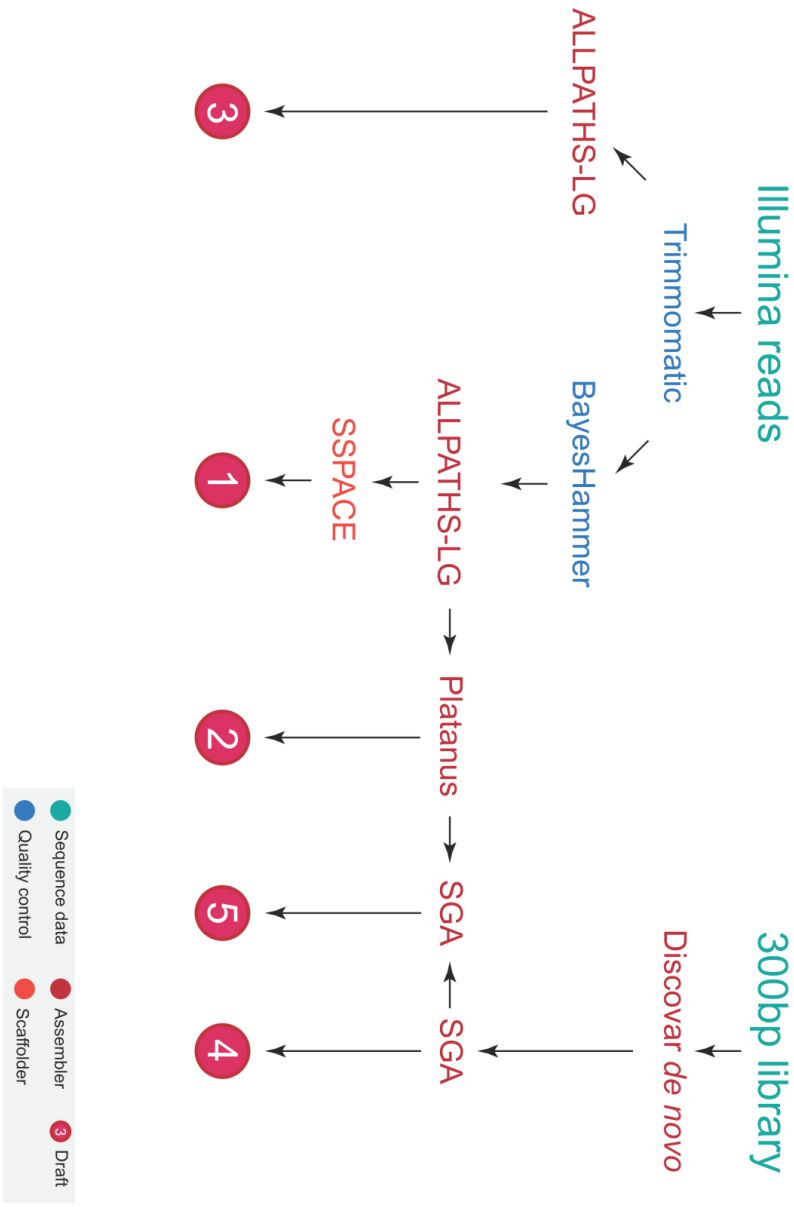



**Figure 2.2: The optimized genome assembly approach involved multiple assembly pipelines.** Overview of the optimized genome assembly pipelines attempted. The several genome assembly pipelines included the standard assembly pipelines (Figiure 2.1) as well as the optimized pipelines. Colours highlighted represent sequence data, quality control, assembler, scaffolder, and draft genomes (5) respectively. The numbers (1 – 5) in coloured circles represent the draft genomes generated through the respective assembly pipelines. The final draft is draft genome number 5.

Transcriptome

This analysis was carried out in collaboration with Dr. Steve Kelly at the Department of Plant Sciences, University of Oxford, UK. A transcriptome was assembled using a combination of data generously provided by Rob Field (O'Neill, *et al.,* 2015) and additional data generated in house (Ebenezer, *et al.,* 2017). Quality control checks on raw sequence data from high throughput sequencing pipelines were performed using FastQC (Barbraham bioinformatics software). According to the developers, it provides a modular set of analyses which gives a quick impression about the quality of the sequenced data. Read statistics and nucleotide distribution were also determined using this toolkit (data not shown). Reads were quality filtered using Trimmomatic and searched for Illumina adaptors (the default option) with settings LEADING:10 TRAILING:10 SLIDINGWINDOW:5:15 MINLEN:50. Ribosomal RNA was removed using SortMeRNA using the default settings, before read error correction using BayesHammer with the default settings. Reads were normalized using khmer with settings –C 20 -k 21 -M 8e9 and overlapping paired-end reads joined using ALLPATHS-LG and all reads subject to *de novo* assembly using SGA, minimum overlap size of 80 nucleotides and no mismatches. These assembled contigs were reassembled using SGA, with each round reducing overlap size by 10 nucleotides. The filtered, normalized joined reads were then mapped to this assembly using Bowtie2. Reads that were absent from the assembly were identified and combined to the assembly file. The combined un-assembled reads and assembled contigs were subject to assembly using SGA with an overlap size of 70. This process of identifying unmapped reads and reassembling with SGA was repeated each time decreasing the overlap size by 10 nucleotides. Contigs were then subject to scaffolding using SSPACE and the full set of non-ribosomal, corrected, normalized paired-end reads using the settings -k 10 -a 0.7 -n 50 -o 20. Scaffolds were subject to gap filling using the SGA gap filling function. Finally, the assembled contigs were subject to base-error correction using Pilon with the default settings (see sub section 2.4.2 and Figure 2.1).

### *2.4.2 Assembly improvements and variants detections*

To make improvements to the quality of the draft genome and transcriptome assembly by recognizing and correcting errors involving: single bases and small

insertion/deletion events (indels) or gap filling, an integrated assembly improvement of all error types was performed using the Pilon software tool (Walker, *et al.,* 2014) (Figure 2.1). Pilon treats assembly improvement and variant detection as a single pipeline (see Walker, *et al.,* 2014 for details on Pilon's workflow). To greatly reduce run time, normalized sequence data (up to 20x) were used for assembly improvements.

### *2.4.3 Assembly evaluation and assessment*

#### *2.4.3.1 Genome*

To evaluate the performance of the assembly pipeline and the quality of the assembly, a variety of metrics were used (Figure 2.1). These include: statistical description of assembly, presence of core eukaryotic conserved genes, and calculating average predicted gene length, and transcriptome read mapping evidence (Bradnam, *et al.,* 2013). Statistical metrics were generated using the Quality Assessment Tool for Genome Assemblies (QUAST) (Gurevich, 2013). A QC pipeline using the Core Eukaryotic Gene Mapping Approach (CEGMA) (Parra and Bradnam, 2007) for eukaryotic genomes and transcriptome assemblies was utilized whereby the number of 258 core eukaryotic conserved genes mapped suggests the number of genes present in the assembly - the higher the proportion, the better (Bradnam, *et al.,* 2013). In summary, the algorithm allows assessment of the completeness of a genome or transcriptome assembly based on the percentage of core eukaryotic genes (CEG) which align to the assembled contigs or scaffolds. A cumulative contig distribution function as described by Bradnam, *et al.,* 2013, to evaluate when the maximum genome size was attained by the fewest contigs was used, where the slope of the line is proportional to fragmentation of the assembly (this approach was also used to assess the transcriptome). Finally, assessment quality was determined by mapping RNA-seq evidences to the draft genome to determine percentage of expressed transcripts in the genome assembly.

#### *2.3.3.2 Transcriptome*

The criteria to assess genome assemblies are now well developed, but standards for systematically assessing the quality of transcriptome assemblies have not fully been established (Martin and Wang, 2011) (Figure 2.1). Five matrices (Martin and Wang,

2011; O'Neil and Emrich, 2013) were taken into account in evaluating the assembly quality: accuracy, completeness, contiguity, chimerism, and variant resolution (see Martin and Wang, 2011, for extensive definitions of these terms). To achieve this aim, the Transrate algorithm (Smith-Unna, *et al.,* 2014) was used which weaves these metrices into three super metrix: contigs metrics (based entirely on analyzing the set of contigs), read mapping metrics (based on aligning the reads used in the assembly to the assembled contigs), and comparative metrices (involves comparing the assembly to a related reference species) (Smith-Unna, *et al.,* 2014). Another method of assessment was performed involving mapping cDNA or ESTs obtained from NCBI database (~ 26,131 ESTs for *E. gracilis* available at the NCBI: http://www.ncbi.nlm.nih.gov/nucest/?term=euglena (as at January 4, 2015)) and 454 Roche generated transcriptome reads (Field, *et al,* unpublished data) to the assembled transcriptome. The set of 248 CEGMA CEGs were also used to interrogate the transcriptome assembly (Hornett and Wheat, 2012).

## 2.5 GENOME AND TRANSCRIPTOME STRUCTURAL AND FUNCTIONAL AUTOMATIC ANNOTATION

### *2.5.1 Transcriptome: Gene predictions and automatic functional analysis*

Insilico analysis such as open reading frame (ORF) determination and gene predictions were carried out using GMST and TransDecoder, while Gene Ontology (GO), Enzyme Code, KEGG maps (biological pathways), and taxa distribution were performed as part of the automatic functional annotation previously described by Kaneshisa, *et al.,*2004, 2010, with minor modifications (Figure 2.3). Six frame translation, ORF determination, and gene prediction of assembled transcriptome sequences were predicted using TransDecoder prediction tool (https://transdecoder.github.io/) and GMST (Tang, *et al.,* 2015), and the longest ORF with coding characteristics, BLAST homology, and PFAM domain were extracted (see Stein, *et al.,* 2002). The predicted ORF was queried against the NCBI non-redundant protein database using BLASTP homology searches, and the top hit for each protein with an E-value cutoff of 10 retained. Using the Blast2GO automatic functional annotation tool (Conesa, *et al.,* 2005), the GO annotations of the best BLAST results with an E-value cutoff of 10 were generated from the GO database. The protein domain, biological pathway analyses, and top species distributions were

determined using InterPro, BLAST, Enzyme code and KEGG tool - as part of the Blast2GO functions using default settings. To greatly reduce run times, BLASTP and InterProScan were processed locally prior to uploading to Blast2GO in .xml file formats.

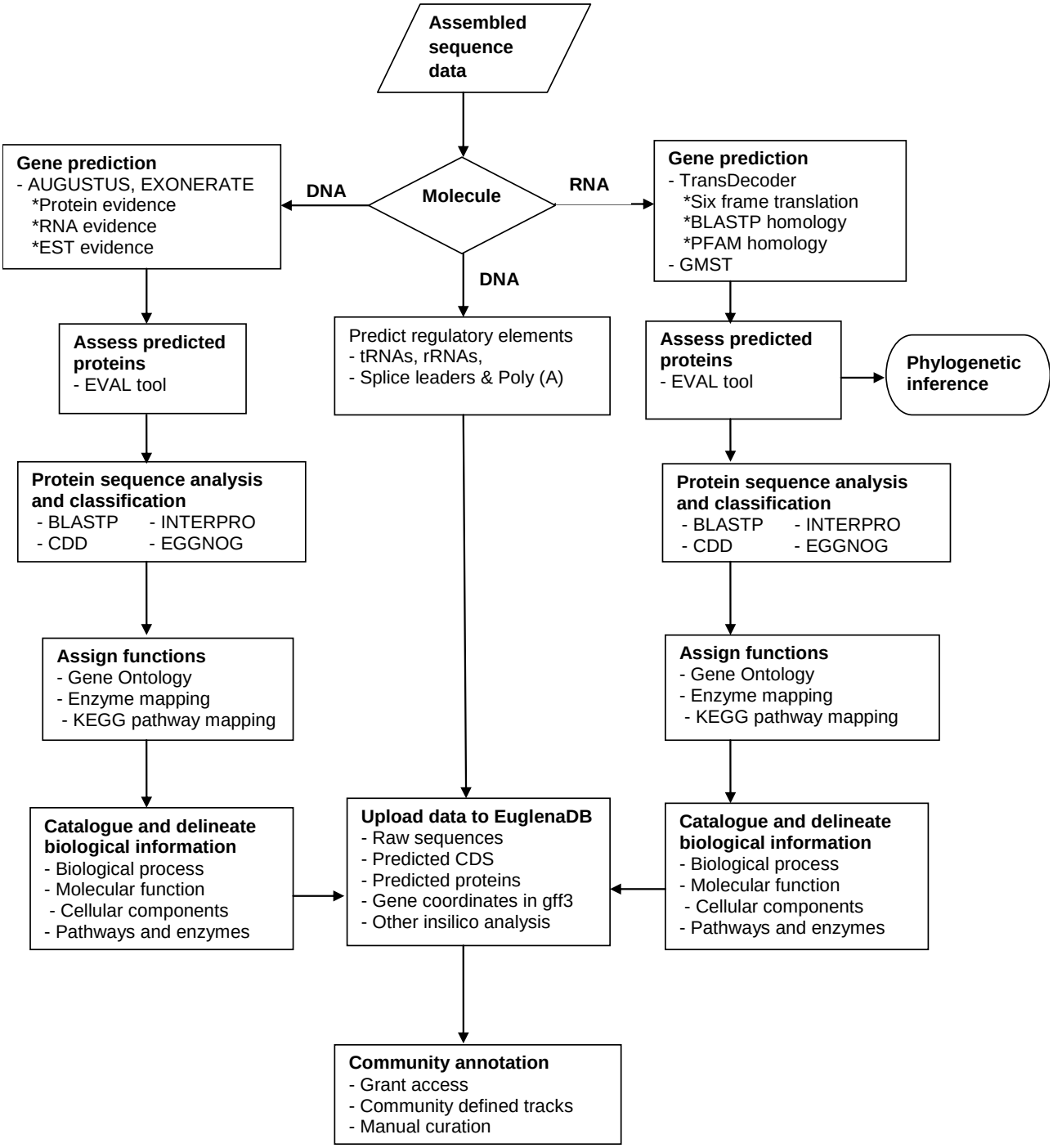



**Figure 2.3: The standard genome and transcriptome structural and functional annotation involved multiple bioinformatics approaches and biological categories.** Overview of actual genome and transcriptome structural and functional annotation: The workflow involves the production of an assembled sequenced data, gene predictions, protein sequence analysis and classifications, assignment of functions, cataloguing and delineation of biological information, upload to EuglenaDB, and community-based annotation.

*2.5.2 Assembling sequence data, data mining and phylogenetic inference*

Orthologs of reference proteins were inferred from the OrthoFinder algorithm (Emms and Kelly, 2015) predicted orthogroups, or via custom searches as highlighted: Homology searches for orthologs and paralogs were performed against the predicted proteome for *E. gracilis* using BLASTp (Figure 2.4). Clustering to 100 % identity was performed for the predicted *E. gracilis* proteins using the Cluster Database at High Identity (CD-HIT) algorithm (Li and Godzik, 2006) to remove gapped/incomplete and redundant sequences. Sequences with significant BLASTP top hit search (evalue = 1e-3) were subjected to both Reversed Position Specific Blast *RPS-BLAST* (NCBI CDD Search) and InterProScan (Chapter 5, and Appendix I-a). The annotated sequences with domain and/or protein signatures matches were extracted using a combination of UNIX command and BioPerl script (available on request), and clustered to 99 % identity using CD-HIT algorithm (Li and Godzik, 2006). The CD-HIT program outputs a set of 'non-redundant' (nr) protein representative sequences. The clustered representatives were aligned to known eukaryotic protein reference sequences using the Multiple Sequence Alignment (MSA) tool, MAFFT (Katoh and Standley, 2013) and/or ClustalX2 (Larkin, *et al.,* 2007). Poorly aligned positions or gaps were removed using the gap deletion menu/command, trimmed manually in Jalview and with trimAl (Capella-Gutierrez and Silla-Martinez, 2009) prior to alignment. The final alignments were processed locally for phylogenetic inference with the PhyML Command Line Interface (CLI) using the default settings (Guindon and Gascuel, 2003; Le and Gascuel, 2008; Yang, 1993; Guindon *et al.,* 2010), RaxML (Stamatakis, 2006),FastTree (Price, et al., 2010), MrBayes (Huelsenbeck and Ronquist, 2001) and/or FastTree (Price,et al., 2010) as appropriate. Annotations of the trees were performed using TreeGraph2 (Stöverand Müller, 2010), InkScape (https://inkscape.org/en/), and Adobe Illustrator. For analytical methods of specific aspects of the biology of *E. gracilis* which include, amongst others, surfaceome, mitochondria proteome, plastid proteome see Ebenezer, *et al.,* 2017, and corresponding sub subsections in Chapters 5.

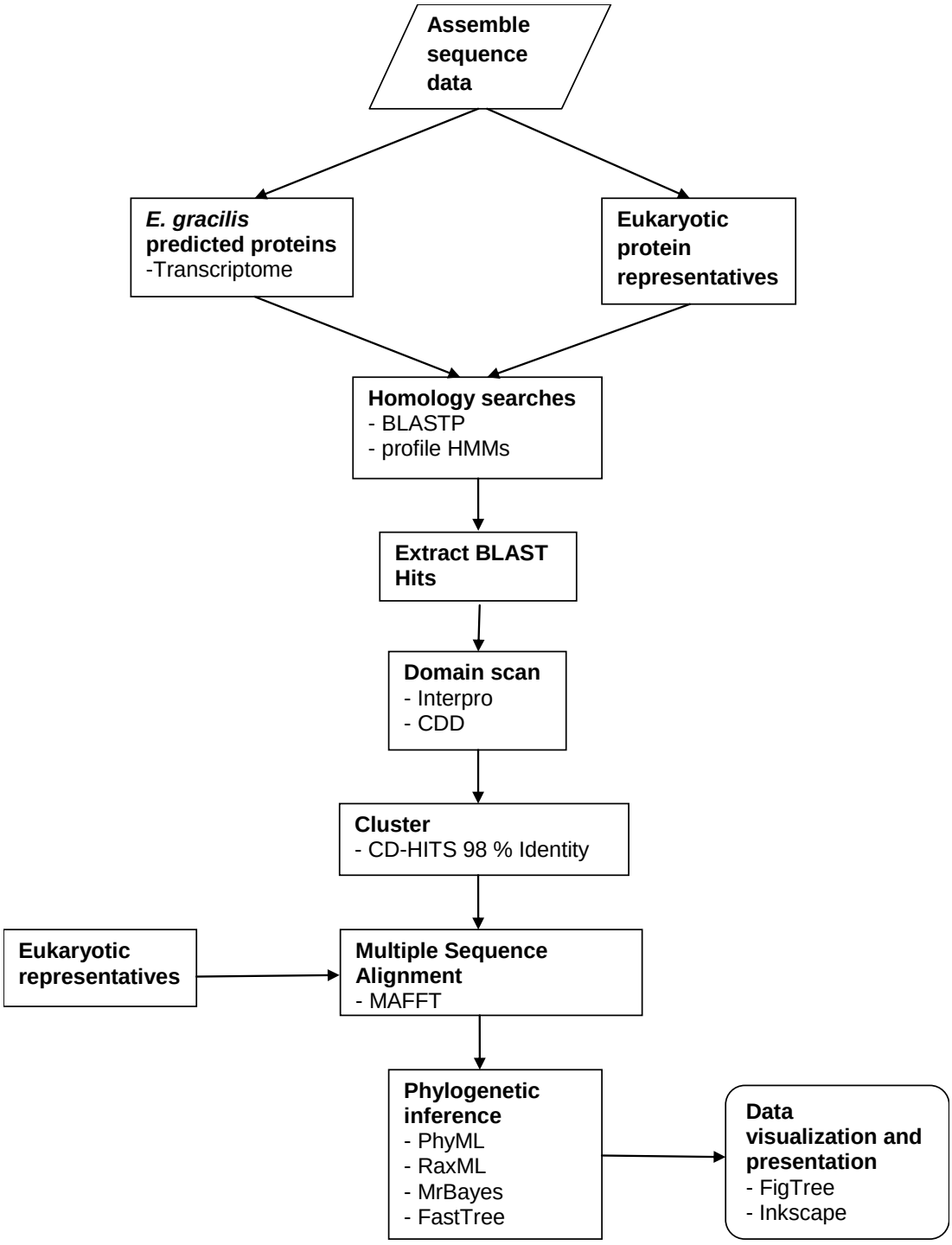



**Figure 2.4: Phylogenetic inference relationship is a multi-step process.** Overview of *E. gracilis* phylogenetic inference protein annotation: *E.gracilis* predicted proteins and their respective reference proteins are assembled. Homology predictions and domain scans are carried out using BLASTP, HMMs, and InterProScan/CDD, redundancies, alignments, and phylogenetic trees are constructed using CD-HITS, MAFFT, and tree building algorithms.

### *2.5.3 Large contigs (>10kbp) in the E. gracilis genome*

For an initial resolution into the architecture of the genome, contigs >10kbp were analysed. These contigs were interrogated using tblastn with the *E. gracilis* predicted proteome from the transcriptome. Sequences with hits were further interrogated using the Exonerate algorithm (with --protein2genome option) for a resolution into splicing mechanisms and coding regions. Pre-processing of Exonerate outputs suitable for further analysis and visualizations were performed using performed using a combination of custom scripts, Unix/Linux commands, and the Genome Tools algorithms (Gremme, et al., 2013). Sequences, and their respective splicing coordinates in gff3, were uploaded to the IGV genome viewer. Exon and intron general statistics was performed using EVAL assessment tool (Keibler and Brent, 2003). Functional genome annotation involves BLASTP homology searches (evalue = 10) against the National Center for Biotechnology Information (NCBI) non redundant (nr) protein database (http://blast.ncbi.nlm.nih.gov/Blast.cgi), protein domain scan using InterProScan (Jones, *et al.,* 2014), and their mappings to Gene Ontology (GO) (Ashburner, *et al.,* 2000), enzyme code and KEGG metabolic pathways (biological inference) (Kanehisa, *et al.*, 2006; Aoki and Kanehisa, 2005; Kaas, 2007) using Blast2GO (Gotz, et al., 2008). Transmembrane helices and signal peptides were predicted using TMHMM (Krogh, *et al.,* 2001) and SignalP (Petersen, *et al.,* 2011) respectively, which are components of the InterPro Scan utility. Prediction of tRNAs and rRNAs were preliminary performed using tRNA-Scan (Lowe and Eddy, 1997) and rRNAmmer (Lagesen, 2007) respectively.

## 2.6 ORTHOLOGOUS GROUP CLUSTERING

This orthologous groups clustering analysis was carried out in collaboration with Dr. Steve Kelly at the Department of Plant Sciences, University of Oxford, UK. To identify orthologous genes in *E. gracilis*, across eukaryotic taxa, the *E. gracilis* proteome was clustered with thirty eukaryotic taxa using Orthofinder (Emms and Kelly, 2015) – thanks to Dr. Steve Kelly at the University of Oxford, UK. Eukaryotic groups captured in the clustering include members of the Euglenozoa, Unikonta, secondary hosts, red algae, green algae, land plants (primary hosts), and members of the Excavata (see Chapter 4, and Appendix I-a). Clustering heat maps were

produced using a custom Perl script (available on request) and annotated using Adobe Illustrator.

## 2.7 GENE EXPRESSION STUDIES

### *2.7.1 Transcriptome analysis for gene expression studies*

This transcriptome analysis for gene expression studies was carried out in collaboration with Dr. Steve Kelly at the Department of Plant Sciences, University of Oxford, UK. RNA were extracted, purified, and sequenced at the Beijing Genomics Institute (http://www.genomics.cn/en/index). Analysis and comparisons of the data were performed using standard bioinformatics pipelines developed by Dr Steve Kelly at the University of Oxford (scripts available on request). An estimated 62M clean reads were generated which were subject to quality filtering using trimmomatic (Bolger, et al., 2014), to remove low quality bases and read-pairs as well as contaminating adaptor sequences prior to assembly. Sequences were searched for all common Illumina adaptors and the settings used for read processing by trimmomatic were LEADING:10 TRAILING:10 SLIDING WINDOW:5:15 MIN LEN:50. The trimmed filtered reads were then used to quantify the *de novo* assembled transcriptome using Salmon (Patro, et al., 2016) with the bias Correct option. Expected counts were integerised before being subjected to differential expression testing using DESeq2 (Love, et al., 2014) using default parameters. In the transcriptomics analysis 66542 distinct sequence classes were detected and the data was reduced to 41045 applying the same rejection criteria (minimum three replicates). Evolutionary selections (dN/dS ratios) of transcripts which encode known proteins functional information (biological pathway-directed) were carried out using the Codon Alignment v2.1.0 (Codon Align tool on https://www.hiv.lanl.gov/) and Synonymous Non-synonymous Analysis Program (SNAP) (Korber, 2000) to investigate if regulatory response is enriched in ratio types or to provide an indication of selective pressure on protein coding genes in the light and dark experiment.

### *2.7.2 Proteomics analysis for gene expression studies*

This proteomic analysis for gene expression studies was carried out in collaboration with Dr. Martin Zoltner in Prof. Mark Fields Laboratory, University of Dundee, UK. Liquid chromatography tandem mass spectrometry ($LCMS^2$) was performed at the University of Dundee, UK, and thanks to the Proteomic Facility team of the School of

Life Sciences for the analysis. Samples were analysed on an UltiMate 3000 RSLCnano System (Thermo Scientific) coupled to a Q-exactive mass spectrometer (Thermo Scientific) and mass spectra were analysed using MaxQuant version 1.5 (Cox and Mann, 2008) searching the predicted *E. gracilis* proteome (*de novo* transcriptome assembly) reported here, Dr. Martin Zoltner performed this analysis. Minimum peptide length was set at six amino acids, isoleucine and leucine were considered indistinguishable and false discovery rates (FDR) of 0.01 were calculated at the levels of peptides, proteins and modification sites based on the number of hits against the reversed sequence database. Ratios were calculated from label free quantification intensities using only peptides that could be uniquely mapped to a given protein. If the identified peptide sequence set of one protein contained the peptide set of another protein, these two proteins were assigned to the same protein group. P values were calculated applying t-test based statistics using Perseus (Tyanova et *al*., 2016). 8661 distinct protein groups were identified in the MaxQuant analysis. For further analyses data were reduced to 4297 protein groups by rejecting those groups not identified at the peptide level in each of the three replicates for one state. Additionally, a cohort of 384 protein groups was extracted that were observed in only one state (232 light and 152 dark in all three respective replicates).

### *2.8.3 Electron microscopy of E. gracilis cells*

This section was carried out in collaboration with Dr. Sue Vaughan and Alana Burrell at the Department of Biological and Medical Sciences, Oxford Brookes University, UK. 50 ml of light and dark adapted cells were grown to mid logarithmic phase and primarily fixed. These primary fixatives (glutaraldehyde) and fixing protocol were a kind gift from Dr. Sue Vaughan's laboratory. 25 % glutaraldehyde solution was added directly to the Euglena light and dark 50 ml cultures to give a final concentration of 2.5 %. The flasks were swirled gently and ensured that the cells have stopped moving. The content of the flask were then transferred to 50 ml Falcons tubes (25 ml in each) and centrifuged into pellets. The supernatant were pipette and discarded. 3 ml of the primary fixatives (2.5% glut and 2% PFA in 0.1M cacodylate buffer pH 7.2) were added and the pellet re-suspended by gentle pipetting. The light and dark primarily fixed samples were then shipped in parafilmed falcon tubes and sealed box to Dr. Vaughan's laboratory for Transmission Electron Microscopy (TEM) analysis.

# CHAPTER THREE

## 3.0 GENOME ORGANISATION AND FUNCTION

### 3.1 INTRODUCTION

In this chapter, the genome organization and function of *E. gracilis* was investigated. There are three genetic systems in *E. gracilis*: chloroplast, mitochondria, and nuclear, with the first two already in the public domain – the focus of this genome project is on the nuclear genome. Over the course of this project, the *E. gracilis* genome was sequenced using multiple read libraries across several sequencing platforms: Roche 454, Illumina, and PacBio. Similarly, several assembly pipelines were attempted until an optimum assembly was generated. However, this assembly is ~ 20 % complete in terms of Core Eukaryotic Genes (CEGs), highly repetitive and fragmentary with over half of the assembly being < 1 kb. To understand the genome function and structure with the draft genome assembly, the > 10 kb sequences were analysed and discussed here. Part of this chapter was done in collaboration with Steve Kelly at the University of Oxford, UK, and has been published in preliminary form in Ebenezer, *et al.,* 2017.

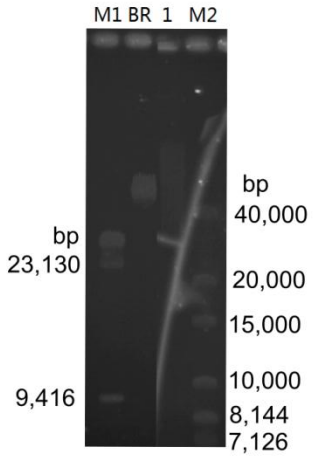
### 3.2 GENOME SEQUENCING AND ASSEMBLY

#### *3.2.1 DNA isolation and purification*

*E. gracilis* Z strain gDNA was extracted as previously described (see materials and methods) for Illumina and PacBio samples respectively. The extraction method employed produced high quality gDNAs with good and clear bands, ranging in sizes between 23 kbp and > 40 kbp for the low and high molecular weight protocols respectively (Figure 3.1). In both extraction methods, the purity of *E. gracilis* gDNA preparations were first established by the ratios of the extinction coefficients at 230, 260 and 280 nm. Concentrations and yield ranges from 47.0 – 71.6 ng/µl and 10.34 – 14.39 µg for low molecular weight isolation protocol, and 450 ng/µl and 132.75 µg for high molecular weight isolation methods respectively. DNA dissolved in Qiagen buffer AE and TE buffer shows OD260/230 ratio ranging from 2.15 – 2.81, and OD260/280 ratio ranging from 1.94 – 2.29 respectively, indicating no evident

contamination with proteins or RNA in the extraction procedures respectively (Table 3.1).

**A**


M1 BR 1 M2
bp
23,130
9,416
bp
40,000
20,000
15,000
10,000
8,144
7,126

**B**

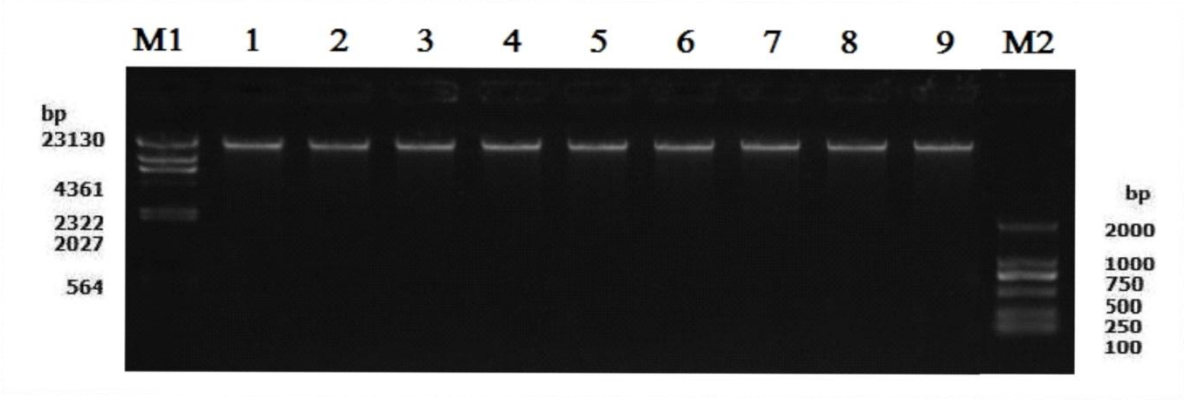

M1 1 2 3 4 5 6 7 8 9 M2
bp
23130
4361
2322
2027
564
bp
2000
1000
750
500
250
100

**Figure 3.1: Nucleic acid Pulse-Field Gel Electrophoresis for *E. gracilis* DNA preparation.** Panel A shows the 40 Kbp mate pair insert size. M1, M2, and BR are the various markers utilized for the analysis with their respective ladders in base pairs (bp). The *E. gracilis* DNA sample is labeled number 1. Panel B shows the ≤800 bp insert size. M1 and M2 are the various ladders utilized for the analysis with their respective ladders. Samples labeled 1-8, & 9, are represent the different DNA preparations carried out to maximize volume, yield, and quality. These are not replicates.

**Table 3.1:** Pre-DNA sequence quality control test for 40 Kb Mate pair and ≤800 bp insert size

| S/No. | Sample name | Sample number | Library type | Concentration (ng/µl) | Volume (µl) | Total mass (µg) | OD260/280 | OD260/230 | Test result |
|---|---|---|---|---|---|---|---|---|---|
| 1. | Euglena gDNA (T1) | 0150208895 | 40K Mate pair | 450 | 295 | 132.75 | 1.99 | 2.15 | Level A |
| 2. | Euglena gDNA (R1) | 85215022084 | ≤800 bp Insert size | 71.6 | 220 | 15.75 | 2.11 | 2.73 | Level A |
| 3. | Euglena gDNA (R2) | 85215022085 | ≤800 bp Insert size | 54.4 | 220 | 11.97 | 1.94 | 2.26 | Level A |
| 4. | Euglena gDNA (R3) | 85215022086 | ≤800 bp Insert size | 47 | 220 | 10.34 | 2.29 | 2.81 | Level A |
| 5. | Euglena gDNA (R4) | 85215022087 | ≤800 bp Insert size | 65.4 | 220 | 14.39 | 2.12 | 2.61 | Level A |
| 6. | Euglena gDNA (R5) | 85215022088 | ≤800 bp Insert size | 54.8 | 220 | 12.06 | 2.13 | 2.65 | Level A |
| 7. | Euglena gDNA (R6) | 85215022089 | ≤800 bp Insert size | 49.8 | 220 | 10.96 | 2.05 | 2.57 | Level A |
| 8. | Euglena gDNA (R7) | 85215022090 | ≤800 bp Insert size | 58.2 | 220 | 12.8 | 2.08 | 2.7 | Level A |
| 10. | Euglena gDNA (R9) | 85215022091 | ≤800 bp Insert size | 64.2 | 220 | 14.12 | 2.03 | 2.7 | Level A |
| 11. | Euglena gDNA (R10) | 85215022092 | ≤800 bp Insert size | 56.2 | 220 | 12.36 | 2.14 | 2.27 | Level A |

**Note:** The table represent the pre-sequencing result of *E. gracilis* genome for both 40K mate pair and ≤800 bp insert size libraries. *Parameter descriptions:* **Sample name** refers to the alphabetical identifier of the sample. **Sample number** refers to the numerical identifier of the sample. **Library type** refers to the specific insert size sequenced. **Concentration**, **Volume**, and **Total Mass** refers to these respective parameters for each sample. OD260/280 and OD260/230 refers to the **Optical Density** at wavelengths of 260/280 and 260/230. **Test Result** refers to the classification of the pre-sequencing quality control; Level A means that the sample is qualified for DNA sequencing, and the amount of sample is sufficient for two or more libraries.

*3.2.2 Library preparation, Illumina, Roche 454, and PacBio sequence data*

The genome sequence consists of three versions: version 0.7, 0.8, and 0.9 (Table 3.1). For the genome sequence version 0.7, five paired-end reads illumina Hiseq 2000 sequencing runs using five fragment libraries with insert sizes of 170 bp, 500 bp, 800 bp, 2 kbp, and 5 kbp and one Roche 454 single read respectively were carried out. Approximately 26,905 M total clean reads were generated with read lengths ranging from 49 – 200 bp (~ 19 Gb in total bases) and an estimated coverage of 0.1X, 21.2X, and 57.59X for Roche 454 and Illumina read libraries respectively. For the genome sequence version 0.8, three paired-end reads illumina Hiseq 2000 sequencing runs using three fragment libraries with insert sizes of 250 bp, 540 bp and 40 kbp respectively were carried out. More than 442.52 M clean reads were generated with read lengths ranging from 100 – 300 bp (~ 67 Gb in total bases) and a combined estimated physical coverage of 229X. For the genome sequence version 0.9, 0.9 M clean reads was generated using the PacBio platform (Table 3.2) to produce a merged assembly using read data from the version 0.8. Overall, the qualities of these reads were satisfactory, showing phred quality values (Q20) ≥ 82 % and moderate GC content between 47.80 – 51.55 % (Table 3.2) (the FastQC sequence visualization .html files for all read sequences in Table 3.2 are available on request).

.

**Table 3.2:** Data summary of *E. gracilis* whole genome and transcriptome sequencing

| Molecule | Version | Sequence technology | Read library (bp) | Orientation | Length | Clean reads (M) | Clean bases (Mbp) | Total clean bases (Mbp) | Q20 (%) | GC (%) | **Phys Cov. (X) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Genome | 0.9 (Merged) | PacBio | Myler | Single | 9914* | 561233** | 5571.29 | 8827.11 | 82.8* | 49.89 | 4.4 |
| | | | Puri/Dav | | 9207* | 353314** | 3255.82 | | - | 50.10 | |
| | 0.8 | Illumina HiSeq 2000 | 250 | Paired-overlap | 150 | 337.844 | 50675.25 | 67303.97 | 95.05 | 51.07 | 229 |
| | | | 540 | | 300 | 31.307 | 9392.17 | | 82.93 | 51.55 | |
| | | | 40000 | Paired-jump | 100 | 72.366 | 7236.55 | | 88.65 | 51.09 | |
| | 0.7 | Roche454 | 454 | Single | 200 | 493166 | 174 | 18661.67 | - | 54.15 | 0.1 |
| | | Illumina HiSeq 2000 | 170a | Paired-overlap | 100 | 45.926 | 4592.55 | | 89.85 | 47.8 | 21.2 |
| | | | 170b | | | | | | | | |
| | | | 500 | | | 51.372 | 5137.22 | | 85.10 | 48.3 | |
| | | | 800 | | | 45.389 | 4538.90 | | 83.80 | 48.65 | |
| | | | 2000a | Paired-jump | 49 | 31.682 | 3104.85 | | 90 | 50.5 | 57.59 |
| | | | 2000b | | | | | | | | |
| | | | 2000c | | | | | | | | |
| | | | 5000 | | | 26.289 | 1288.15 | | 90.30 | 50.2 | |
| Transcriptome | 1.0 | Illumina HiSeq 2000 | 189 | Paired-overlap | 100 | 14.549 | 2438.72 | 2822.14 | 97.09 | 58.93 | 21.5 |
| | | | 125 | | | 450.90 | 383.42 | | 93.93 | 55.5 | 7.15 |

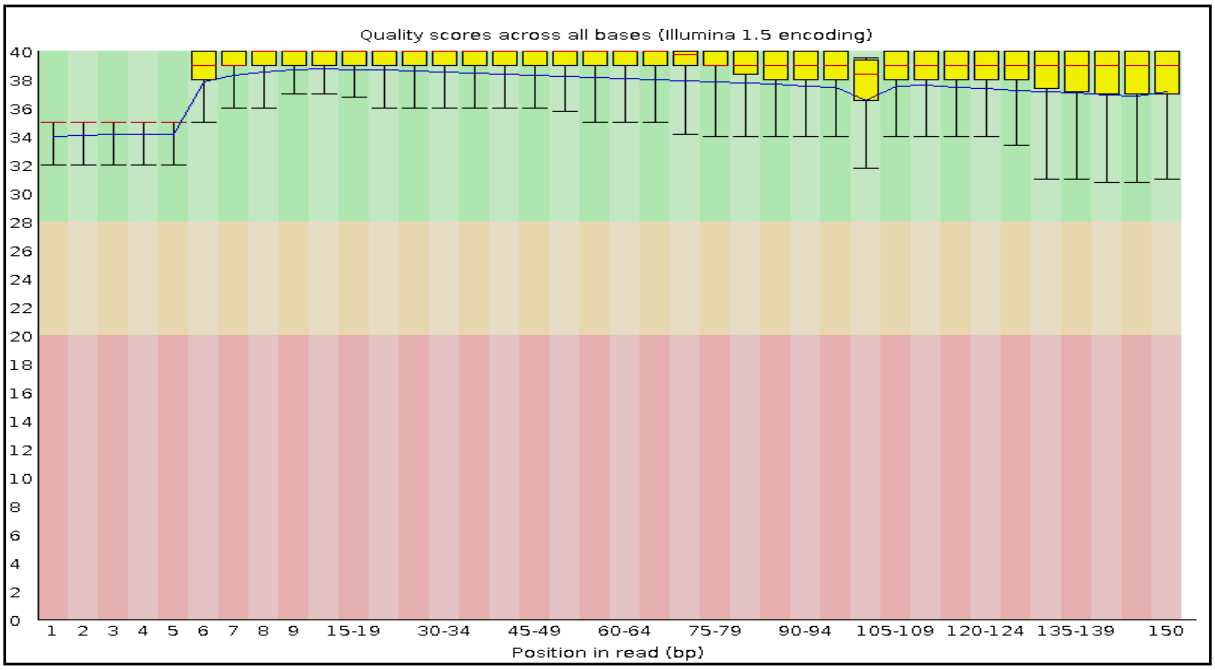
**Column description: Molecule** refers to either DNA (genome) or RNA (transcriptome). **Version** refers to the genome or transcriptome version produced. Orientation refers to the read orientation. **Sequence technology** refers to the sequence platform employed. **Read library** refers to the library, sequence identification number that identify the sample, or an insert size which is the distance between the forward and reverse Mate pairs (read1 and read2); **Length** is the length of the total reads (in bp); **Clean reads** refers to total reads after post sequencing quality control; **Clean bases** refers to total nucleotides (in base pair) after post sequencing quality control (sample filtration) for individual libraries; **Total clean bases** refers to total nucleotides (in base pair) after post sequencing quality control (sample filtration) for a libraries combined; **Q20 (%)** is the number of nucleotide with quality higher than 20/nucleotide (i.e. clean forward and reverse reads, R1 & R2), and **Q20** is the probability that a base is called incorrectly – probability of incorrect base is 1 in 100 with a call accuracy of 99 %. **GC (%)** is the percentage of total nucleotides with guanine-cytocine bases. **Phys. Cov. (X)** = L * N / G, where L is read length, N is number of reads, and G is genome length. M = millions. Mbp = megabases. † Average read length. ***** = Mean value. ****** Read lengths are not in Megabases. (see http://www.illumina.com/science/education/sequencing-quality-scores.html)

*3.2.3 Genome assembly, assembly improvements and variant detections, assembly evaluation and assessments*

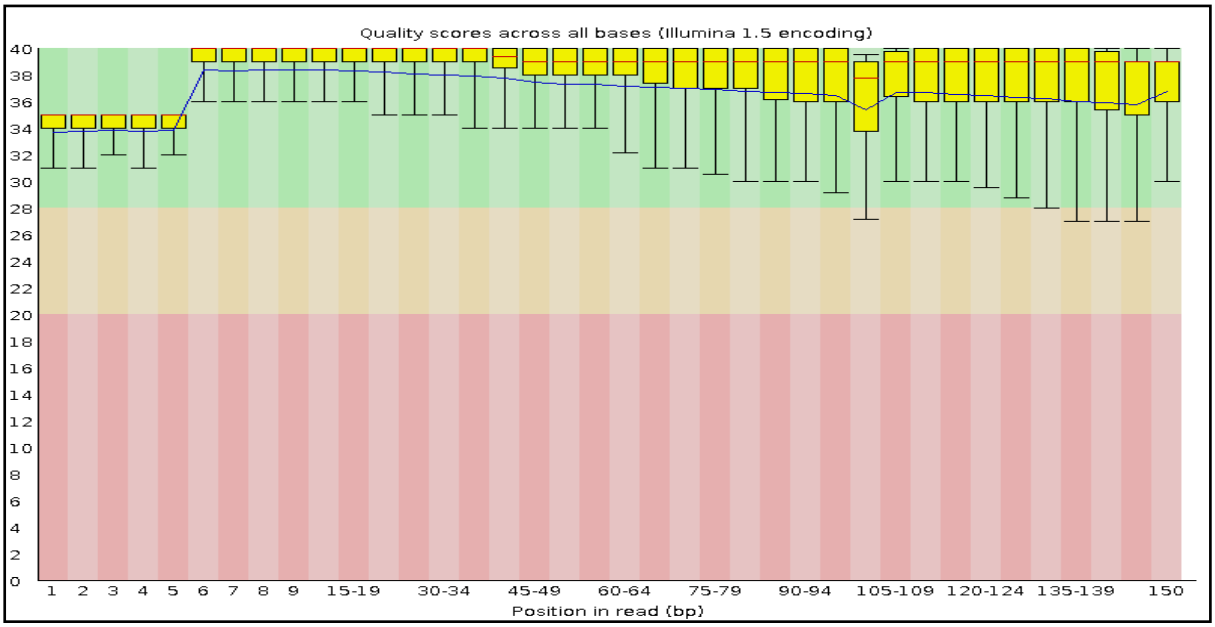
Quality Control

The genome was assembled as described in methods (see chapter 2). Quality control was performed using the Trimmomatic algorithm, and visualized using FastQC quality tool. There are at least eleven (11) quality parameters to take into account when performing QC for NGS data analysis. These include per base sequence quality, per tile sequence quality, per sequence quality scores, per base sequence content, per sequence GC content, per base N content, sequence length distribution, Sequence duplication levels, Overrepresented sequences, and Adapter content, and Error Correct Reads (Patel and Jain, 2012; Gnerre, *et al.,* 2011). Out of these 11, the per base sequence quality is the most important parameter because it defines the proportion of both correctly and incorrectly sequenced bases. For conciseness, the result of per base sequence quality‘ parameter for 250 bp library 150PE and 540 bp library 300PE sequencing will only be presented (Figure 3.2). This is because the results for the other libraries (see Table 3.2) are primarily similar after QC, and generally entails ensuring that all reads and bases (100 %) are within the acceptable threshold prior to input into the assembly pipeline (other QC files are available on request). These libraries have been chosen due to two reasons: the first figure gives a general indication of QC results for ≤ 150PE sequences, while the second figure gives an indication of ≤ 300PE sequences, and perhaps the role of increasing read lengths on sequence quality (as will be highlighted below).

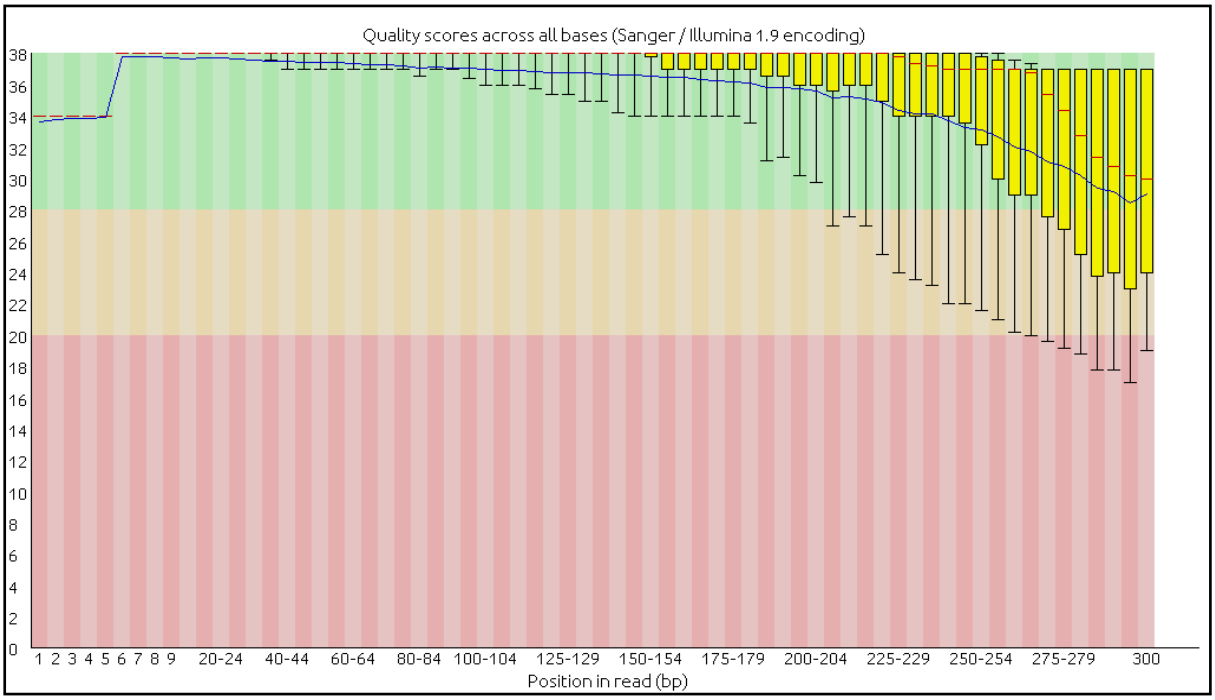
## A: 250 bp library, 150PE read (Forward, Read 1)




## B: 250 bp library, 150PE (Reverse, Read 2)



## C: 540 bp library, 300PE (Forward, Read 1)




## D: 540 bp library, 300PE (Reverse, Read 2)

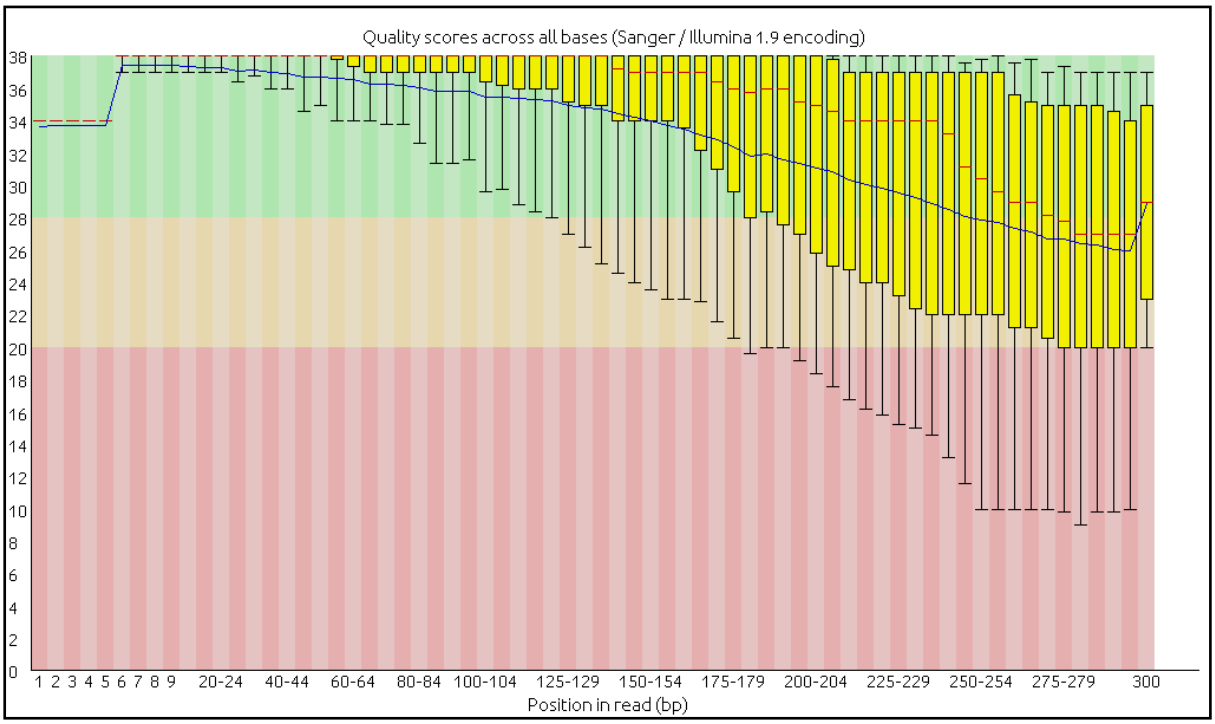


**Figure 3.2: Quality trimmed reads are within quality control thresholds.** The diagram shows the per base sequence quality for 250 bp and 540 bp read libraries. Panel A: Per base sequence quality for 250 bp library 150PE trimmed sequence (Forward, Read 1). Panel B: Per base sequence quality for 250 bp library 150PE trimmed sequence (Reverse, Read 2). Panel C: Per base sequence quality for 540 bp library 300PE trimmed sequence (Forward, Read 1). Panel D: Per base sequence quality for 540 bp library 300PE trimmed (Reverse, Read 2). Green zones show reads of very good quality; orange zones show read of moderate quality; red zones show reads of poor quality.

Actual assembly

The assembly process includes a workflow that employed several assembly pipelines and intermediates to produce an optimal assembly. According to Armstrong, *et al.,* (unpublished manuscript & poster presentation), an assembly would be considered optimal if it contained the fewest contigs, the total length of the contigs or scaffolds closely matched the predicted genome or transcriptome size and the contig/scaffold N50 was larger than other assembly intermediates. The draft genome has a summary statistics as presented in Table 3.3. The assembly resulted in a set of "scaffolds", with each scaffold made up of a set of contiguous sequences (contigs), ordered and oriented with respect to one another by mate-pairs such that the gaps between adjacent contigs are of known size or distance (using estimates from the long insert sizes of 800 bp, 2 kbp, 5 kbp, and 40 kb) and are spanned by clones with end-sequences flanking the gap (see Figure 2.1 for assembly workflow) as previously described by Myers, *et al.,* 2000. Present in the assembly are gaps of unknown base (N). The types of gaps observed were both sequence and physical gaps. According to Myers, *et al.,* 2000, "gaps within scaffolds are called sequence gaps; gaps between scaffolds are called 'physical gaps' because there are no clones identified spanning the gap". The draft genome assembly has a total sequence of 2066288 sequences with an N50 of 955.

Assembly improvements and variant detection

The Pilon algorithm was run for multiple iterations on the last Illumina assembly intermediate prior to merging with the PacBio data, and recorded changes include insertions (in), deletions (dels), gap filling, SNPs, number of sequences, insertion length, deletion length, and gap filled length. In the assembly version 0.7, Pilon made significant improvements to the contiguity of the draft assembly, increasing the genome coverage, and contig N50 size by 437 kb (data not shown). According to Walker, *et al.,* 2014, observed gains in genome coverage and contig N50s when using Pilon are principally due to Pilon's ability to recognize and fill (or partially fill), by local assembly, "captured gaps", i.e., missing sequence between contigs and within a scaffold (Walker, *et al.,* 2014). In the same version 0.7, Pilon completely and accurately filled 69.65 kb of the > 100,000 captured gaps (69.64 % closure rate). Pilon was able to improve the assembly and added 18.952 kb of sequence. Pilon

improved the quality of the draft assembly, resulting in fewer and longer contigs and an improved gene set as previously report by Walker, *et al.,* 2014 for other model organisms. Insertions/deletions (INDELS) and single nucleotide polymorphisms (SNPs) were identified resulting in 451.78 kb, 458.26 kb, and 3029.44 kb events respectively (data not shown).

**Table 3.3:** QUAST genome assembly statistics

| Statistics without reference | *Euglena_gracilis*_genome_V1 |
|---|---|
| # contigs | 2066288 |
| # contigs (>= 0 bp) | 2066288 |
| # contigs (>= 1000 bp) | 373610 |
| # contigs (>= 10000 bp) | 1459 |
| # contigs (>= 100000 bp) | 2 |
| Median sequence length | 457 |
| Mean sequence length | 694 |
| Minimum sequence length | 106 |
| Largest contig | 166587 |
| Total length | 1435499417 |
| Total length (>= 0 bp) | 1435499417 |
| Total length (>= 1000 bp) | 691997200 |
| Total length (>= 10000 bp) | 22999641 |
| N50 | 955 |
| N75 | 504 |
| L50 | 399971 |
| L75 | 924484 |
| GC (%) | 50.60 |
| No. of gaps | 0 |
| No. bases in gaps | 0 |
| No. of gaps per 100 kbp | 0 |

**Key:** Sequence here refers to contigs (nucleotides). Number of sequences (#) is the total number of contigs in the assembly; no. of sequence > x kbp is total number of contigs of length ≥ x kbp - where x is 1, 10, or 100 kbp; combined sequence length is the total number of bases in the assembly; N50 is the length for which the collection of all contigs of that length or longer covers at least half an assembly; Gaps is the total number of uncalled bases (N's) in the assembly; Min and Max sequence length are the lengths of the shortest and largest contigs/proteins in the assembly respectively; Mean and Median sequence length are the average length of the sequences and the sequence length separating the minimum and maximum sequence lengths in the assembly respectively. All statistics are based on contigs of size >= 1 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs. L50 and L75 is the number of contigs equal to or longer than N50 or N75 respectively. For example, L50 is the minimal number of contigs that cover half the size of the assembly. While N50 has to do with contig length, L50 has to do with contig numbers (for an extensive description of terminologies, see Gurevich, *et al.,* 2013).

Assembly evaluation and assessment

The genome assembly workflow employed several QC analysis pipelines for validating assembly completeness (see Chapter 2), and these include: QUAST assessment metrices, CEGMA statistical metrices, RNA-seq reads and EST sequence mapping evidences – which are all discussed below. As a gene discovery project, the most important evaluation is the CEGMA statistical metrics since it tells us the proportion of core eukaryotic genes (CEG) present in the respective assemblies. The CEG datasets were aligned to the genome and transcriptome assembly in order to determine the percentage CEG‘s contained in the assembly. According to Armstrong, *et al.,* (unpublished manuscript and poster presentation), it is assumed that an assembly which is more complete will have a higher percentage of complete and partial CEG alignments than a less complete assembly (see Parra, *et al.,* 2007; and CEGMA metrices below for details).

I. Quality Assessment Tool for Genome Assemblies (QUAST) statistical metrices

Several statistical parameters were calculated for the genome assembly (see Table 3.3). This includes number of contigs, mean sequence length, N25, N50, and N75 (see Gurevich, *et al.,* 2013 for detailed definitions of these vocabularies, and many more that are highlighted below). The most important is the N50, and according to Gurevich, *et al.,* 2013, is calculated by summing all sequence lengths, starting with the longest, and observing the length that takes the sum length past 50 % of the total assembly length. The N25, N50, and N75 metrices are based on using 25 %, 50 %, and 75 % thresholds (see Table 3.3). A recently emerging metric was also adopted in version 0.7 draft assembly intermediates (data not shown), and is called the NG50 length (Earl, *et al.,* 2011; Bradnam, *et al.,* 2013). According to Bradnam, *et al.,* 2013, this normalizes for differences in the sizes of the genome assemblies being compared. It is calculated in the same way as the N50, except that the total assembly size is replaced with the estimated genome size when making the calculation (Bradnam, *et al.,* 2013). Other parameters that were taken into account in this particular draft version 0.7 assembly (data not shown) include the NG values, NAx, NGAx, and cumulative alignment lengths (see Bradnam, *et al.,* 2013 for details of these vocabularies and their utility).

The *E. gracilis* assembly showed some significant differences in its total assembly length, contig and scaffold lengths or basic statistical matrices (Table 3.3). In the version 0.7 draft assembly, using the NGx statistical criteria, which involves assessing the size of the genome assembly in relation to the estimated genome size (~ 750 Mbp), it was evident that the draft assembly is a less competitive assembly (containing ~ 650 Mbp which is ~ 87 % of the ~ 750 Mbp estimated amount of sequence) when compared against the sequenced genomes of *T. brucei* and *L. major* as a references – even though these species may not be suitable as a reference genome since they are distantly related. The graph (data not shown) shows that *E. gracilis* draft assembly is not yet saturated when compared to the estimated genome size. This suggests that assessing and evaluating assemblies by their basic statistics may be misleading (Armstrong, *et al.,* unpublished manuscript and poster presentation), and Yandell and Ence, 2012 further implied that NG50 and N50 can be poor predictors of the suitability of an assembly for gene-finding purposes, suggesting that someone who is looking to use a genome assembly for gene finding may not need to be overly concerned by low N50 or NG50 values (Yandell and Ence, 2012).

II. CEGMA statistical metrices

A set of 248 CEGs (Parra, *et al.,* 2007) were used to interrogate the presence of core eukaryotic genes by testing for > 70 % alignment of these genes against the contigs in the assembly. The presence, or absence, of these genes is determined by employing the statistical model called the Hidden Markov Model (HMM) approach (Eddy, 2004). This analysis was carried out using CEGMA (Parra, *et al.,* 2007; see Materials and Methods), and in comparison with other completely sequenced genomes of *T. brucei* and *L. major*. According to Bradnam, *et al.,* 2013, the CEGMA analysis can assess presence, but not accuracy of the given genes within the assemblies. *E. gracilis* draft assembly contains a total of 50 of the 248 ultra conserved eukaryotic proteins which is 20.16 %. Of these, 8.87 % are complete while 11.29 % are partial (see Table 3.4). When compared to *T. brucei* and *L. major*, the overall percentage CEG presence for these organisms are higher with 82.66 % and 82.26 % respectively. Out of these, 79.03 % and 78.23 % are complete for *T. brucei* and *L. major* respectively.

**Table 3.4:** Comparative statistics of the completeness of the genome and transcriptome based on 248 CEGs

| Assembly | Organism | Gene status | #Prots | %Completeness | #Total | Average | %Ortho |
|---|---|---|---|---|---|---|---|
| Genome | *E. gracilis* | Complete | 22 | 8.87 | 37 | 1.68 | 54.55 |
| | | Partial | 50 | 20.16 | 89 | 1.78 | 56.00 |
| | *T. brucei* | Complete | 196 | 79.03 | 259 | 1.32 | 24.49 |
| | | Partial | 205 | 82.66 | 282 | 1.38 | 28.29 |
| | *L. major* | Complete | 194 | 78.23 | 220 | 1.13 | 11.34 |
| | | Partial | 204 | 82.26 | 245 | 1.20 | 15.69 |
| Transcriptome | *E. gracilis* | Complete | 187 | 75.40 | 390 | 2.09 | 65.78 |
| | | Partial | 218 | 87.90 | 506 | 2.32 | 69.72 |
| | *T. brucei* | Complete | 190 | 76.61 | 393 | 2.07 | 60.00 |
| | | Partial | 205 | 82.66 | 448 | 2.19 | 63.41 |
| | *L. major* | Complete | 133 | 53.63 | 275 | 2.07 | 64.66 |
| | | Partial | 194 | 78.23 | 405 | 2.10 | 64.43 |

**Key:** Prots = number of 248 ultra-conserved CEGs present in genome; %Completeness = percentage of 248 ultra-conserved CEGs present; Total = total number of CEGs present including putative orthologs; Average = average number of orthologs per CEG; %Ortho = percentage of detected CEGs that have more than 1 ortholog; Complete = those predicted proteins in the set of 248 CEGs that when aligned to the HMM for the KOG for that protein-family, give an alignment length that is 70% of the protein length. i.e. if CEGMA produces a 100 amino acid protein, and the alignment length to the HMM to which that protein should belong is 110, then we would say that the protein is 'complete' (91% aligned); Partial = those predicted proteins in the 248 sets that are incomplete, but still exceeds a pre-computed minimum alignment score. Keys are as described by Parra, *et al.,* 2007.

III. RNA-seq and EST mapping evidences

The completeness and correctness of protein-coding genes in assemblies is of paramount importance for diverse applications in gene discovery (Yandell and Ence, 2012). In this project, I acquired genome and transcriptome sequences in parallel, and to enable alignment of the transcriptome sequence data to the genome assembly. This directly assesses the presence of genes in terms of completeness and correctness. While ~ 23k EST sequences are available for *E. gracilis* in the NCBI database, there is no available full length transcript RNA-seq data in any of the database; hence, the production and assembling of the initial transcriptome data into draft to produce full length transcripts. To achieve this aim, RNA-seq reads and EST sequences were mapped to the genome to determine the proportion of expressed genes using the RNA-seq reads presented in Table 3.2, and NCBI EST evidences for *E. gracilis.* Mapping evidenced an overall alignment rate of ≥ 88.05 for all read libraries using Bowtie2.

## 3.3 GENOME STRUCTURAL AND FUNCTIONAL AUTOMATIC ANNOTATION

### *3.3.1 Gene predictions and features*

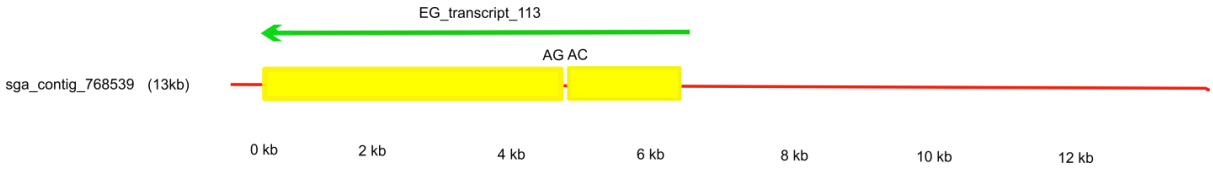
Whole genome sequencing technology was used to determine ~ 70 % of the estimated 2 Gbp nucleotide sequence of the *E. gracilis* genome. This is an estimated 57 – 63 times the size of *T. brucei* and *L. major* genome respectively (Figure 3.3). The genome assembly is only ~ 20.16 % complete (in terms of CEGs), and can support an initial analysis of genome structure and preliminary gene annotation and interpretation (especially the > 10 kbp contigs), but not an extended analysis (for access to data see http://euglenadb.org/ and BioProject #: PRJNA298469). The gene prediction analysis was focused on the > 10 kbp contigs (~ 1.6 % of the total assembled genome size) since these are more likely to encode genes than > 1 kbp which I already anticipate to be highly fragmentary.

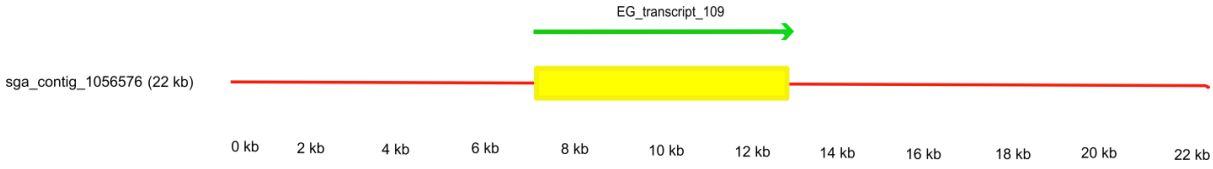
Using a combination of BLAST and Exonerate alignment algorithm (with --protein2genome option), the transcriptome was aligned to the > 10 kbp contigs (1490 sequences). Of these 1490 contigs, 267 sequences have Open Reading Frame (ORF) based on Exonerate predictions (Figure 3.4, 3.5, 3.6, and Table 3.5). These 267 sequences were then screened for chloroplast and mitochondria

genomes (since these organellar genomes have already previously been sequenced), reducing the number to 214 contigs. Out of these 214 contigs, 135 protein-coding genes were identified with 144 transcripts (Table 3.5, Figure 3.5) and 214 translated proteins (Figure 3.5, 3.6, Table 3.5).

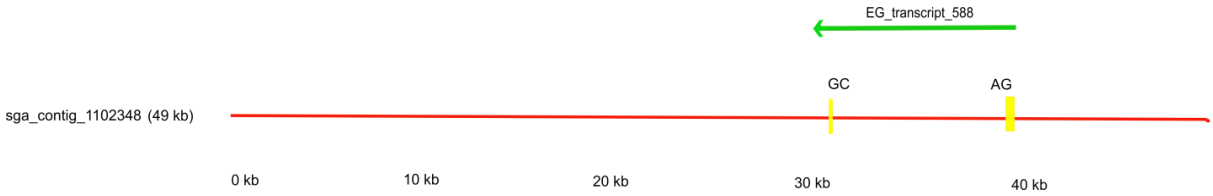
13 S. cerevisiae
13 O. lucimarinus
17 C. merolae
35 T. brucei
41 N. gruberi
100 C. elegans
105 C. crispus
120 C. reinhardtii
123 D. melanogaster
135 A. thaliana
160 T. vaginalis
3000 H. sapiens
E. gracilis
1000 (Dooijes et al., 2000)
2000 (Ebenezer et al., 2016)
3000 (Cook and Robert 1985)
4400 (Zakrys 1988)
9000 (Stankeiwicz et al., 1983, Marzus et al., 1984)

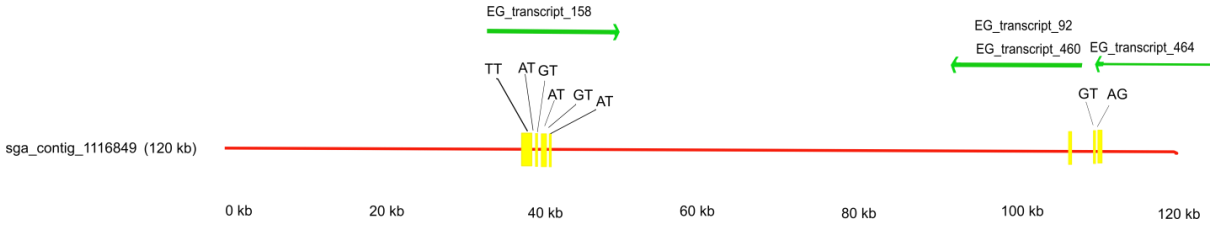
**Figure 3.3: Predicted *E. gracilis* genome sizes are variable.** The figure shows the comparison of genome sizes across selected eukaryotes. The area (×3 magnifications) of the circles represents the sizes of the selected eukaryotic genomes in megabases which are highlighted by numbers corresponding to respective circle. Color codes: *Green* photosynthetic green organisms, *red* photosynthetic red organisms, *purple* kinetoplastids, *grey* other members of the eukaryotic species. Circles embed into each other is *E. gracilis* genome sizes which are calculated/predicted estimates from Zakrys (1988), Cook and Roxby (1985), Dooijes et al. (2000), Mazus et al. (1984) and Stankiewicz et al. (1983), and Ebenezer, *et al.* (2017).


EG_transcript_113
AG AC
sga_contig_768539 (13kb)
0 kb
2 kb
4 kb
6 kb
8 kb
10 kb
12 kb


EG_transcript_109
sga_contig_1056576 (22 kb)
0 kb
2 kb
4 kb
6 kb
8 kb
10 kb
12 kb
14 kb
16 kb
18 kb
20 kb
22 kb


EG_transcript_159
sga_contig_1079396 (21 kb)
0 kb
10 kb
20 kb
18 kb
20 kb


EG_transcript_588
GC
AG
sga_contig_1102348 (49 kb)
0 kb
10 kb
20 kb
30 kb
40 kb


EG_transcript_158
EG_transcript_92
EG_transcript_460
EG_transcript_464
TT
AT
GT
AT
GT
AT
GT
AG
sga_contig_1116849 (120 kb)
0 kb
20 kb
40 kb
60 kb
80 kb
100 kb
120 kb

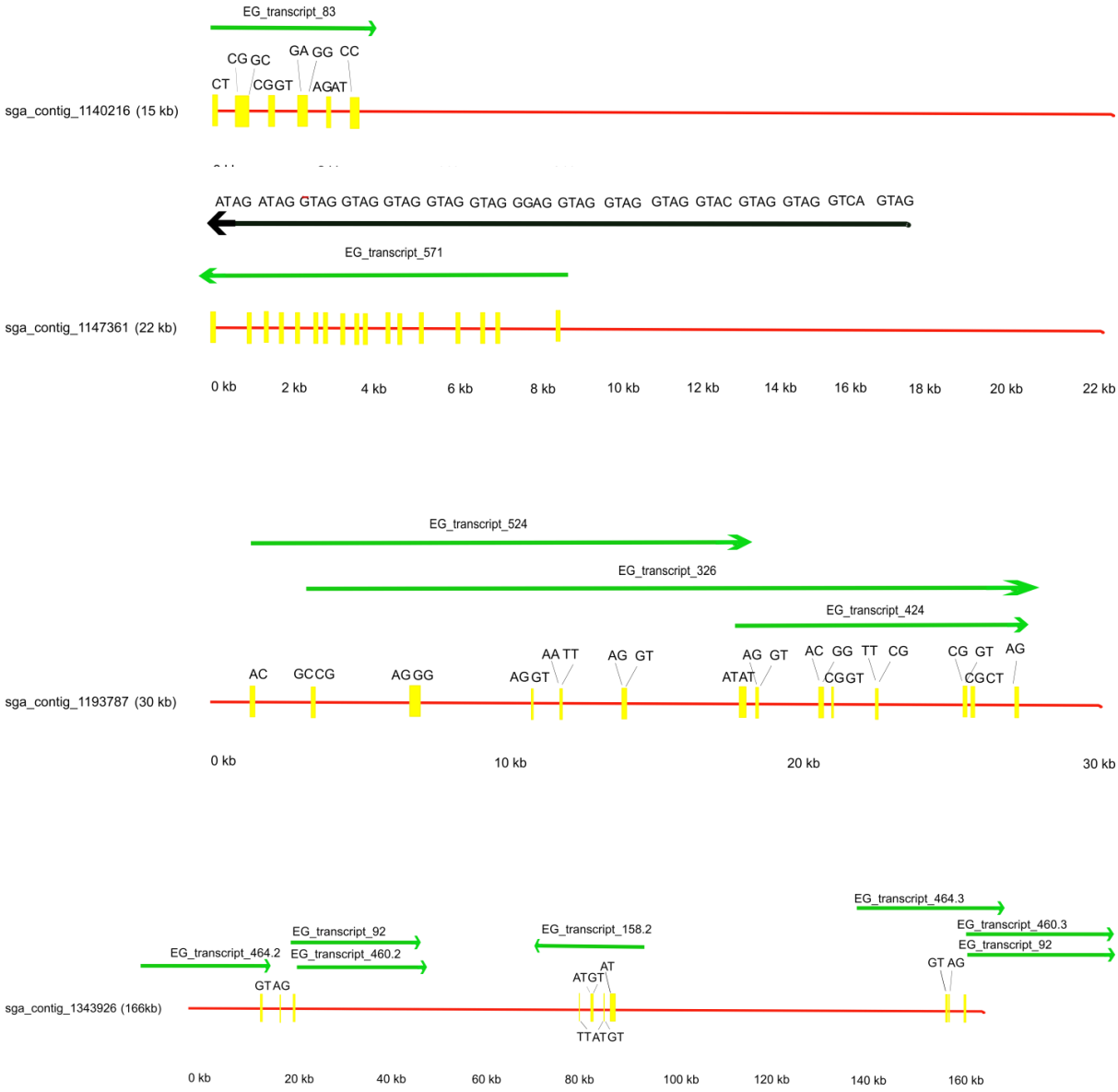



**Figure 3.4: Alternation between high and low coding sequence and gene density regions**. Structural representations of *Euglena gracilis* selected genes: Individual contig figures show the structural representations of selected *E. gracilis* genes. Each contig/gene structure is defined by a sequence identifier prefixed by sga_contig_. Numbers in bracket represent the length of the contig in kilobase pair. Lines coloured in red defines whole contigs spanned with exons/CDS (coloured yellow). Arrows (coloured green) shows orientation (+/-) of the gene. Black arrows or lines shows splicing or splice junctions. The corresponding transcriptome transcript for each contig/gene structure is prefixed EG_transcript_.

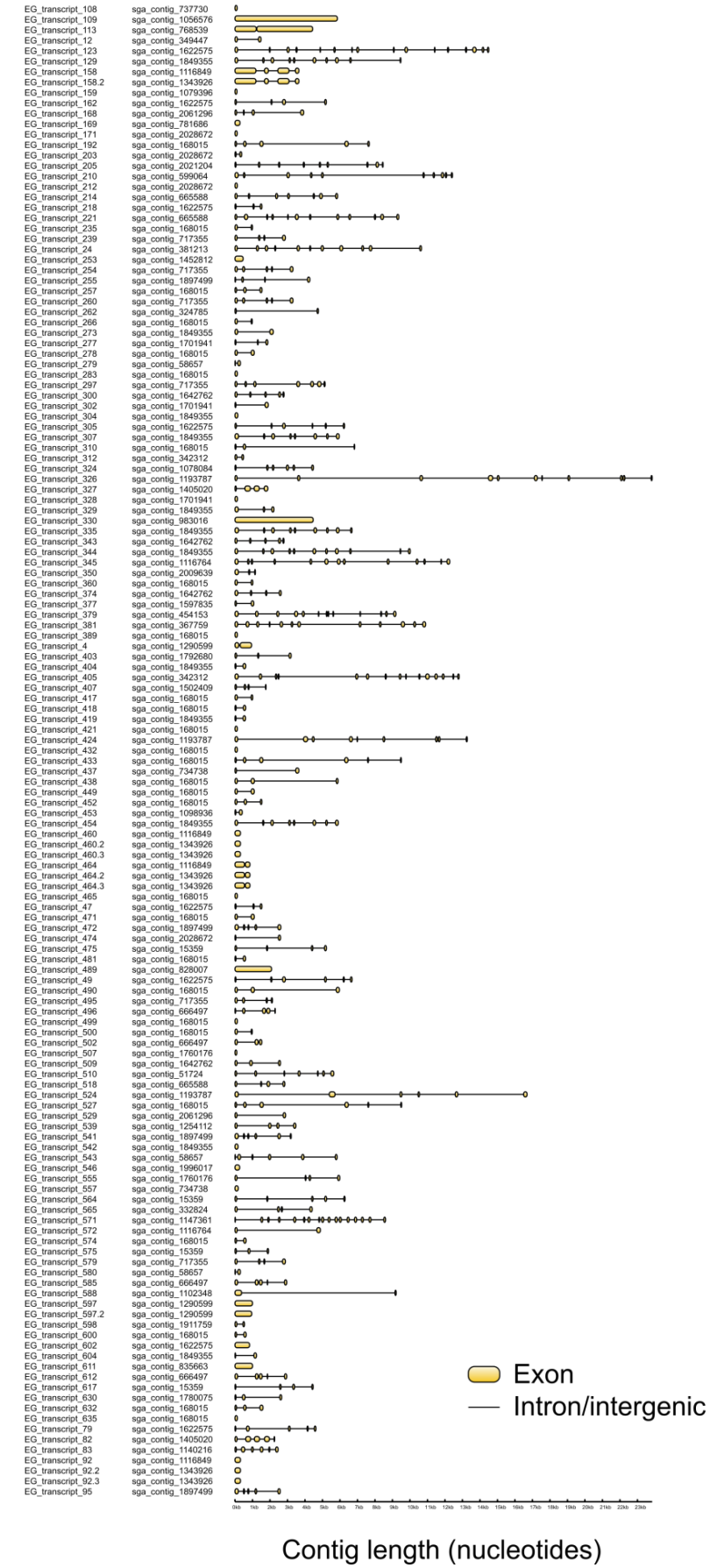

Transcript Contig | Organisation
EG_transcript_108 sga_contig_737730
EG_transcript_109 sga_contig_1056576
EG_transcript_113 sga_contig_768539
EG_transcript_12 sga_contig_349447
EG_transcript_123 sga_contig_1622575
EG_transcript_129 sga_contig_1849355
EG_transcript_158 sga_contig_1116849
EG_transcript_158.2 sga_contig_1343926
EG_transcript_159 sga_contig_1079396
EG_transcript_162 sga_contig_1622575
EG_transcript_168 sga_contig_2061296
EG_transcript_169 sga_contig_781686
EG_transcript_171 sga_contig_2028672
EG_transcript_192 sga_contig_168015
EG_transcript_203 sga_contig_2028672
EG_transcript_205 sga_contig_2021204
EG_transcript_210 sga_contig_599064
EG_transcript_212 sga_contig_2028672
EG_transcript_214 sga_contig_665588
EG_transcript_218 sga_contig_1622575
EG_transcript_221 sga_contig_665588
EG_transcript_235 sga_contig_168015
EG_transcript_239 sga_contig_717355
EG_transcript_24 sga_contig_381213
EG_transcript_253 sga_contig_1452812
EG_transcript_254 sga_contig_717355
EG_transcript_255 sga_contig_1897499
EG_transcript_257 sga_contig_168015
EG_transcript_260 sga_contig_717355
EG_transcript_262 sga_contig_324785
EG_transcript_266 sga_contig_168015
EG_transcript_273 sga_contig_1849355
EG_transcript_277 sga_contig_1701941
EG_transcript_278 sga_contig_168015
EG_transcript_279 sga_contig_58657
EG_transcript_283 sga_contig_168015
EG_transcript_297 sga_contig_717355
EG_transcript_300 sga_contig_1642762
EG_transcript_302 sga_contig_1701941
EG_transcript_304 sga_contig_1849355
EG_transcript_305 sga_contig_1622575
EG_transcript_307 sga_contig_1849355
EG_transcript_310 sga_contig_168015
EG_transcript_312 sga_contig_342312
EG_transcript_324 sga_contig_1078084
EG_transcript_326 sga_contig_1193787
EG_transcript_327 sga_contig_1405020
EG_transcript_328 sga_contig_1701941
EG_transcript_329 sga_contig_1849355
EG_transcript_330 sga_contig_983016
EG_transcript_335 sga_contig_1849355
EG_transcript_343 sga_contig_1642762
EG_transcript_344 sga_contig_1849355
EG_transcript_345 sga_contig_1116764
EG_transcript_350 sga_contig_2009639
EG_transcript_360 sga_contig_168015
EG_transcript_374 sga_contig_1642762
EG_transcript_377 sga_contig_1597835
EG_transcript_379 sga_contig_454153
EG_transcript_381 sga_contig_367759
EG_transcript_389 sga_contig_168015
EG_transcript_4 sga_contig_1290599
EG_transcript_403 sga_contig_1792680
EG_transcript_404 sga_contig_1849355
EG_transcript_405 sga_contig_342312
EG_transcript_407 sga_contig_1502409
EG_transcript_417 sga_contig_168015
EG_transcript_418 sga_contig_168015
EG_transcript_419 sga_contig_1849355
EG_transcript_421 sga_contig_168015
EG_transcript_424 sga_contig_1193787
EG_transcript_432 sga_contig_168015
EG_transcript_433 sga_contig_168015
EG_transcript_437 sga_contig_734738
EG_transcript_438 sga_contig_168015
EG_transcript_449 sga_contig_168015
EG_transcript_452 sga_contig_168015
EG_transcript_453 sga_contig_1098936
EG_transcript_454 sga_contig_1849355
EG_transcript_460 sga_contig_1116849
EG_transcript_460.2 sga_contig_1343926
EG_transcript_460.3 sga_contig_1343926
EG_transcript_464 sga_contig_1116849
EG_transcript_464.2 sga_contig_1343926
EG_transcript_464.3 sga_contig_1343926
EG_transcript_465 sga_contig_168015
EG_transcript_47 sga_contig_1622575
EG_transcript_471 sga_contig_168015
EG_transcript_472 sga_contig_1897499
EG_transcript_474 sga_contig_2028672
EG_transcript_475 sga_contig_15359
EG_transcript_481 sga_contig_168015
EG_transcript_489 sga_contig_828007
EG_transcript_49 sga_contig_1622575
EG_transcript_490 sga_contig_168015
EG_transcript_495 sga_contig_717355
EG_transcript_496 sga_contig_666497
EG_transcript_499 sga_contig_168015
EG_transcript_500 sga_contig_168015
EG_transcript_502 sga_contig_666497
EG_transcript_507 sga_contig_1760176
EG_transcript_509 sga_contig_1642762
EG_transcript_510 sga_contig_51724
EG_transcript_518 sga_contig_665588
EG_transcript_524 sga_contig_1193787
EG_transcript_527 sga_contig_168015
EG_transcript_529 sga_contig_2061296
EG_transcript_539 sga_contig_1254112
EG_transcript_541 sga_contig_1897499
EG_transcript_542 sga_contig_1849355
EG_transcript_543 sga_contig_58657
EG_transcript_546 sga_contig_1996017
EG_transcript_555 sga_contig_1760176
EG_transcript_557 sga_contig_734738
EG_transcript_564 sga_contig_15359
EG_transcript_565 sga_contig_332824
EG_transcript_571 sga_contig_1147361
EG_transcript_572 sga_contig_1116764
EG_transcript_574 sga_contig_168015
EG_transcript_575 sga_contig_15359
EG_transcript_579 sga_contig_717355
EG_transcript_580 sga_contig_58657
EG_transcript_585 sga_contig_666497
EG_transcript_588 sga_contig_1102348
EG_transcript_597 sga_contig_1290599
EG_transcript_597.2 sga_contig_1290599
EG_transcript_598 sga_contig_1911759
EG_transcript_600 sga_contig_168015
EG_transcript_602 sga_contig_1622575
EG_transcript_604 sga_contig_1849355
EG_transcript_611 sga_contig_835663
EG_transcript_612 sga_contig_666497
EG_transcript_617 sga_contig_15359
EG_transcript_630 sga_contig_1780075
EG_transcript_632 sga_contig_168015
EG_transcript_635 sga_contig_168015
EG_transcript_79 sga_contig_1622575
EG_transcript_82 sga_contig_1405020
EG_transcript_83 sga_contig_1140216
EG_transcript_92 sga_contig_1116849
EG_transcript_92.2 sga_contig_1343926
EG_transcript_92.3 sga_contig_1343926
EG_transcript_95 sga_contig_1897499
0kb 1kb 2kb 3kb 4kb 5kb 6kb 7kb 8kb 9kb 10kb 11kb 12kb 13kb 14kb 15kb 16kb 17kb 18kb 19kb 20kb 21kb 22kb 23kb
Exon
Intron/intergenic
Contig length (nucleotides)

**Figure 3.5: Gene clustering organization in *E. gracilis* are arranged in a functionally related architecture.** A comprehensive representation of the Open Reading Frame Organisation of the 53 contigs possessing coding capacity. This compliments the contigs shown in Figure 3.3. The first column represent the sequence ID from the *E. gracilis* transcriptome, the second column represent the corresponding sequence ID from the *E. gracilis* genome arising from the alignment of the transcriptome to the genome. The third column represent the organization of the predicted genes showing high and low gene density regions, presence of higly variable structures, multiple transcripts corresponding respectively to different positions of the genome on a single contig, and the tandem array arrangement of genes on these contigs. Contig lengths are indicated by a scale bar (kb).

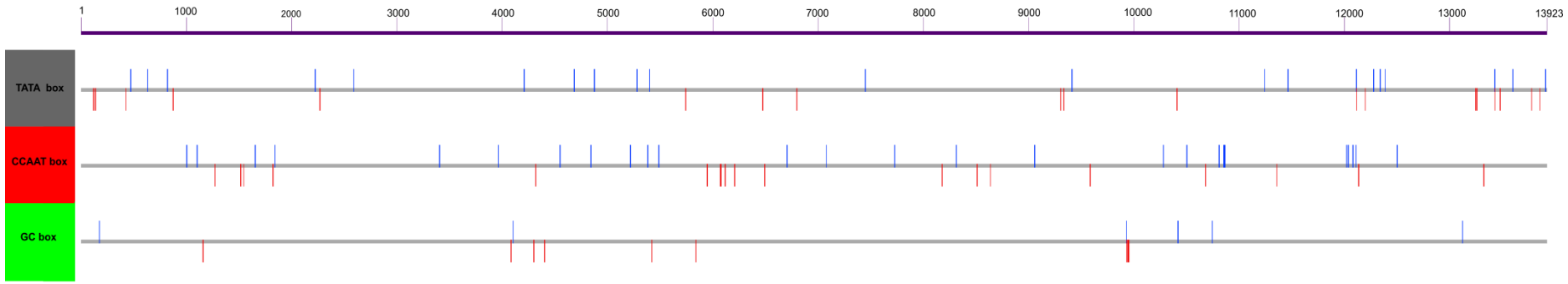
**Table 3.5:** Quantitative and functional description of gene clusters in the *E. gracilis* genome

| S/N | Genome Contig ID | Total number of transcripts mapped | Representative transcript ID* | Common Protein Family | Common Biological Process | Common Molecular Function | Common Cellular Component |
|---|---|---|---|---|---|---|---|
| 1. | sga_contig_1622575 | 8 | EG_transcript_123 | NAD:arginine ADP-ribosyltransferase, ART (IPR000768) | GO:0006471 protein ADP-ribosylation | GO:0005515 protein binding | NP |
| 2. | sga_contig_1849355 | 12 | EG_transcript_129 | NP | NP | GO:0005524 ATP binding | GO:0016020 membrane |
| 3. | sga_contig_1116849 | 3 | EG_transcript_460 | NP | GO:0006351 transcription, DNA-templated | GO:0003677 DNA binding | NP |
| 4. | sga_contig_1343926 | 4 | EG_transcript_464 | NP | GO:0006351 transcription, DNA-templated | GO:0003677 DNA binding | NP |
| 5. | sga_contig_2061296 | 2 | EG_transcript_168 | NAD:arginine ADP-ribosyltransferase, ART (IPR000768) | GO:0006471 protein ADP-ribosylation | GO:0003956 NAD(P)+-protein-arginine ADP-ribosyltransferase activity | NP |
| 6. | sga_contig_2028672 | 4 | EG_transcript_171 | NP | GO:0006629 lipid metabolic process | NP | NP |
| 7. | sga_contig_168015 | 28 | EG_transcript_235 | NP | GO:0009190 cyclic nucleotide biosynthetic process | GO:0016849 phosphorus-oxygen lyase activity | NP |
| 8. | sga_contig_665588 | 4 | EG_transcript_171 | NP | GO:0006629 lipid metabolic process | NP | NP |
| 9. | sga_contig_717355 | 10 | EG_transcript_239 | NP | GO:0055085 transmembrane transport | GO:0005524 ATP binding | GO:0016021 integral component of membrane |
| 10. | sga_contig_1897499 | 4 | EG_transcript_255 | Kinesin-like protein (IPR027640) | GO:0007018 microtubule-based movement | GO:0003777 microtubule motor activity | NP |

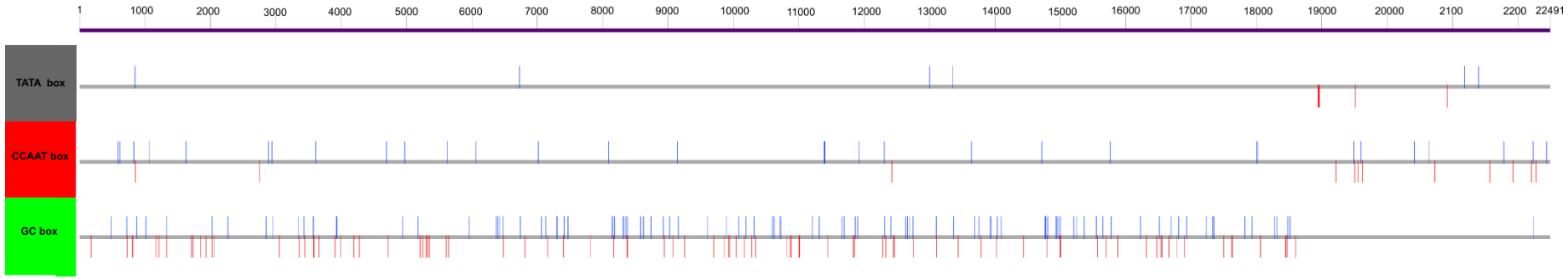
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 11. | sga_contig_1701941 | 3 | EG_transcript_277 | NP | NP | NP | NP |
| 12. | sga_contig_58657 | 3 | EG_transcript_279 | NP | GO:0009190 cyclic nucleotide biosynthetic process | GO:0016849 phosphorus-oxygen lyase activity | NP |
| 13. | sga_contig_1642762 | 4 | EG_transcript_300 | NP | NP | NP | NP |
| 14. | sga_contig_342312 | 2 | EG_transcript_312 | NP | GO:0006629 lipid metabolic process | NP | NP |
| 15. | sga_contig_1193787 | 3 | EG_transcript_524 | NAD:arginine ADP-ribosyltransferase, ART (IPR000768) | GO:0006471 protein ADP-ribosylation | GO:0003956 NAD(P)+-protein-arginine ADP-ribosyltransferase activity | NP |
| 16. | sga_contig_1405020 | 2 | EG_transcript_327 | NP | GO:0008152 metabolic process | GO:0003824 catalytic activity | NP |
| 17. | sga_contig_1116764 | 2 | EG_transcript_345 | NCFP | NCFP | NCFP | NCFP |
| 18. | sga_contig_1290599 | 2 | EG_transcript_4 | NCFP | NCFP | NCFP | NCFP |
| 19. | sga_contig_734738 | 2 | EG_transcript_437 | NP | NP | NP | NP |
| 20. | sga_contig_15359 | 4 | EG_transcript_575 | NP | GO:0009190 cyclic nucleotide biosynthetic process | GO:0016849 phosphorus-oxygen lyase activity | NP |
| 21. | sga_contig_666497 | 4 | EG_transcript_496 | NP | GO:0009190 cyclic nucleotide biosynthetic process | GO:0016849 phosphorus-oxygen lyase activity | GO:0016849 phosphorus-oxygen lyase activity |
| 22. | sga_contig_1760176 | 2 | EG_transcript_507 | NP | GO:0009190 cyclic nucleotide biosynthetic process | GO:0016849 phosphorus-oxygen lyase activity | GO:0016849 phosphorus-oxygen lyase activity |

**Note:** The table describes the clustering properties of the genome contigs predicted to possess coding characteristics. **Keys:** S/N = serial number. Genome contig ID = Sequence IDs of contigs with clustering properties. Total number of transcripts or gene mapped represents the overall number of clustering for respective contigs. Representative transcript ID or gene ID describes the sequence ID that represent the clustered transcript or gene. *For additional members see Figure 3.4 above. Common Biological Process, Molecular Function, and Cellular Component describes the top 1% functional process common amongst the clustered transcripts/genes and genome contigs respectively. Function information were inferred using Interpro database. NP = None predicted. NCFP = No Common Property.

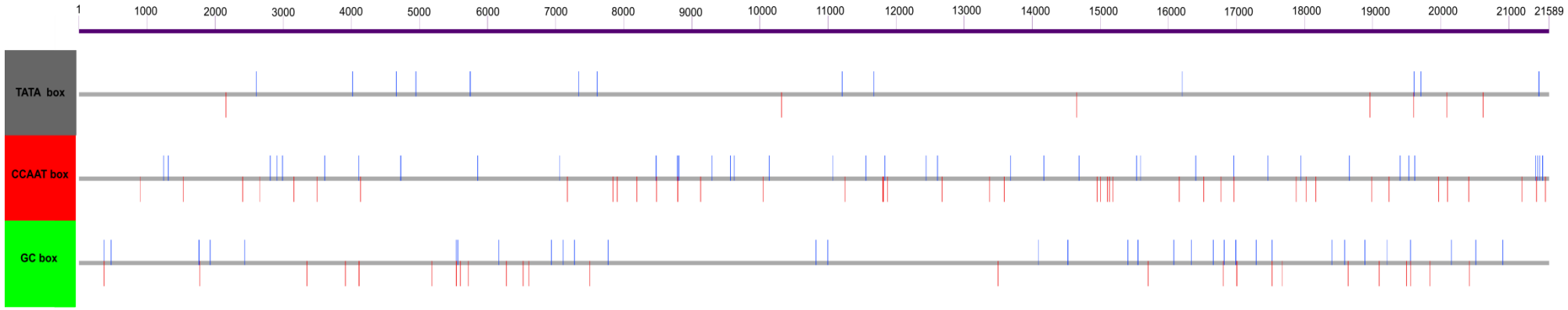
sga_contig_768539



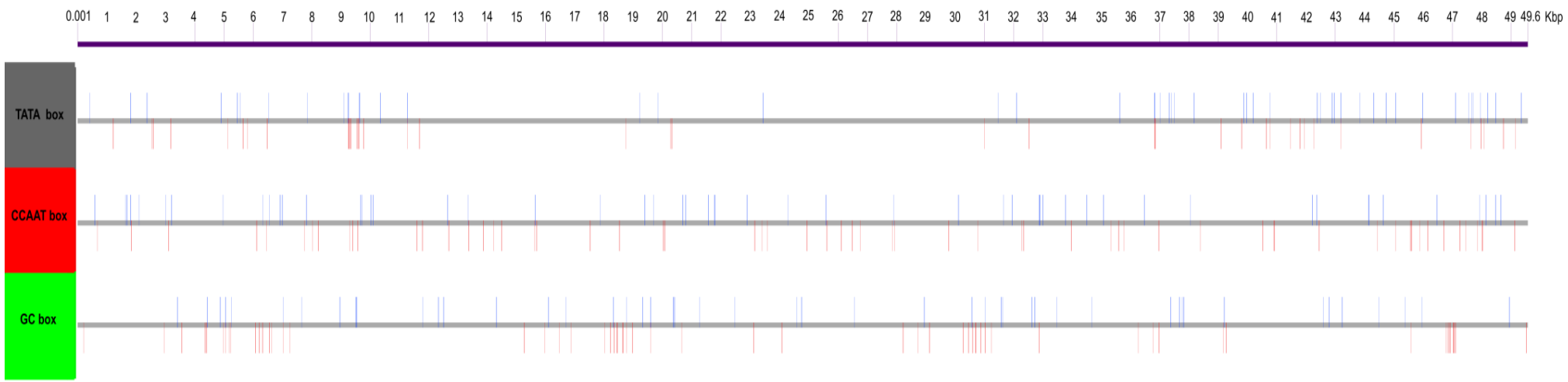

sga_contig_1056576



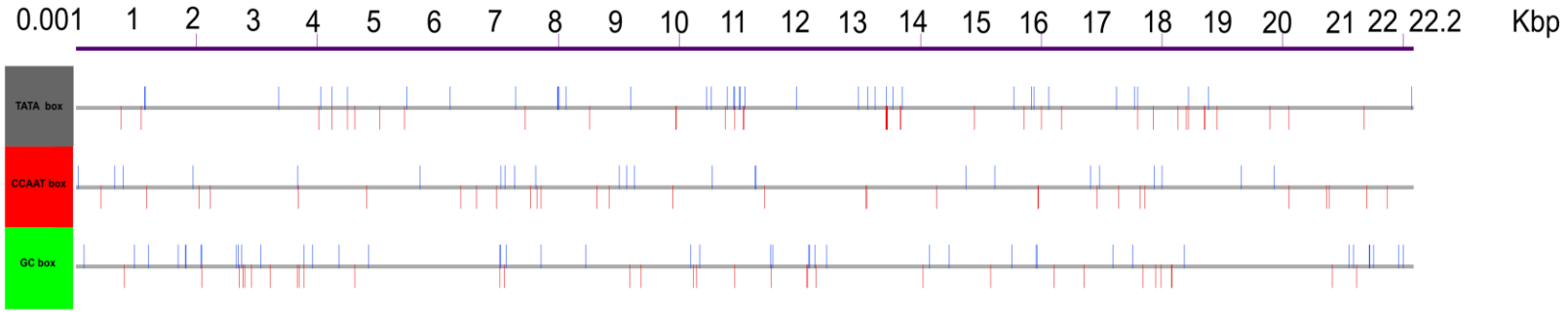
sga_contig_1079396



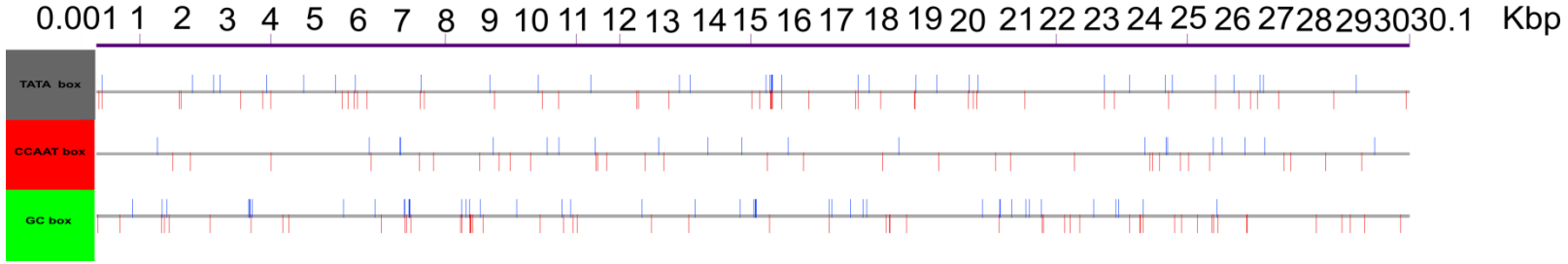

sga_contig_1102348



**sga_contig_1147361**




sga_contig_1193787



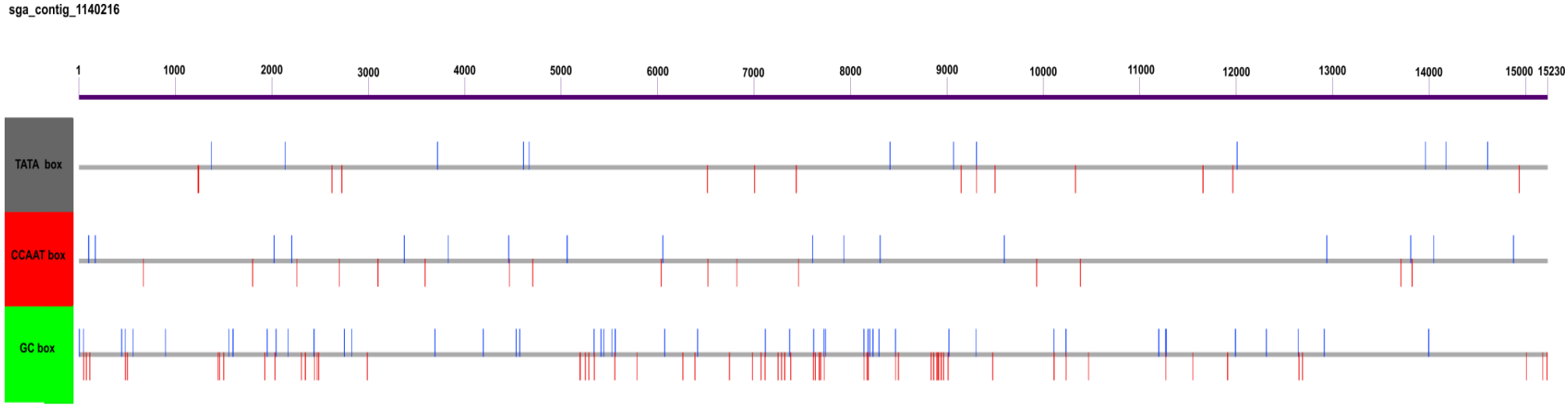



**Figure 3.6: *E. gracilis* genome is replete with promoter islands.** The diagrams describe the core promoter types and positions in select *E. gracilis* contigs: sga_contig_768539, sga_contig_1056576, sga_contig_1079396, sga_contig_1102348, sga_contig_1140216, sga_contig_1147361, sga_contig_1193787. Each respective contig number is presented at the top of each diagram. Coloured boxes represent the core promoters identified. Gray = TATA box, Red = CCAAT box, Green = GC box. Purple lines describe the length of the contig, with corresponding numbers representing scaling in base pairs (except where otherwise noted). Blue and red lines represent promoter positions in the forward and reverse strands respectively.

In these contigs, the genes are not evenly distributed across contigs - there is an alternation between high gene density regions and low gene density regions across contigs in both forward (+) and reverse (-) orientations (Figure 3.4, 3.5, 3.6, and Table 3.5). Nine (9) contigs (Figure 3.4) represent the global architecture of the *E. gracilis* genome. This include contigs with 1 or 2 large clustered and unclustered exon islands between 1 kb and 5kb in length (sga_contig_768539, sga_contig_1056576), contigs with 1 or 2 small clustered and unclustered exon islands of up to 2 kb in length (sga_contig_1079396, sga_contig_1102348), and contigs with multiple small-sized exon islands of up to 3 kb in length, with moderate and huge clustering properties (sga_contig_1116849, sga_contig_1140216, sga_contig_1147361, sga_contig_1193787, sga_contig_1343926). Introns occupy a significant proportion of the contigs, and spliced junctions are both conventional and unconventional as has previously been reported in this system (Figure 3.4, 3.5, 3.6, Milanowski, *et al.,* 2014). Gene arrangements in the genome are organized in long, forward- and reverse- stranded tandem arrays (Figure 3.4, 3.5, 3.6, and Table 3.5) as seen in *T. brucei* where the primary purpose is to increase gene dosage in an environment where individual gene promoters are otherwise absent. Out of the 53 contigs analysed, 22 possess gene clusters with functionally related properties while 31 possess no gene clustering properties (Figure 3.5, 3.6, and Table 3.5). For instance, 8 transcripts (EG_transcript_123, EG_transcript_162, EG_transcript_218, EG_transcript_305, EG_transcript_47, EG_transcript_49, EG_transcript_602, and EG_transcript_79) aligned to the contig sga_contig_1622575 with common protein family (NAD:arginine ADP-ribosyltransferase, ART), biological processes (protein ADP-ribosylation), molecular functions (protein binding), and cellular component properties (NP) (Figure 3.4 and Table 3.5). Analysis of promoters using GPMiner (Lee, *et al.,* 2012) in select contigs show the presence of the core promoter types of TATA box, CCAAT, and GC box (Figure 3.6) across the entire length of the contigs.

In the *T. brucei* genome sequence (TriTrypDB May, 2017), there are 131 contigs/sequences with 129 being greater than 10 kbp, and having ~ 12094, 9068 and 11202 transcripts, protein-coding genes and proteins respectively. This suggests that the *E. gracilis* gene contents are lesser in terms of numbers and quality when compared to *T. brucei* (Table 3.3 and 3.4). This also suggests an

average gene density of 5.87 genes/Mb in *E. gracilis*, and an overall GC composition is 51.19 % (see Table 3.3). Out of the estimated 2 Gbp of the *E. gracilis* genome, it is suggested that < 25 % of these are likely to be single copy genes – with the majority being repeats. In the > 10 kbp (22999641 sequences) contig analysis, 2.16 % are protein-coding genes. Introns occupy 57.83 % of these genes, and there are 271 introns, occupying 278354 of the sequence, with an average length of 1027.14 bases. Exons occupy 14.79 % of the genes, and there are 554 exons, occupying 73482 of the sequence, with an average length of 175 bp. The remaining parts of the contigs (outside the protein coding genes) are intergenic regions, and this is 98.84 %. So, from this analysis, if 98.84 % of the ~ 0.02 Gbp (> 10 kbp) of the genome is non-coding, this means that the actual coding sequence in the estimated 2 Gbp genome size will be < 2 %. The mean length of *E. gracilis* genes (exons or coding sequences only) is ~ 174.54 bases, substantially smaller than in the other sequenced excavates such as *T. brucei* and *L. major* where the average gene length range from 1.3 kb to 1.9 kb. This low numbers in mean length may be due to the fragmentary nature of the *E. gracilis* genome.

### 3.3.2 Non coding RNA genes

Many transcripts function at the RNA level, including rRNAs, tRNAs, snRNAs, and snoRNAs (Hillier, *et al.,* 2005). Using tRNAscan-SE and RNAmmer, I did not find tRNA and rRNA genes. I did not also progress to investigate for all other non coding RNAs since I could not find the essential ones in the > 10 kbp contigs. The present draft genome does not have enough data for investigating non coding RNAs (rRNA, tRNA, etc) as well as comparison and assessment with other sequenced species.

### 3.3.3 BLAST homology, InterProScan, gene ontology, kegg maps, and enzyme codes

Functional analysis of conserved protein signatures using InterProScan further revealed biological functions for the predicted genes in *E. gracilis.* Since the genome is only about 20.01 % complete, there was not enough protein signatures represented in the dataset (Figure 3.7), which also explain the incompleteness of the data (Table 3.6 and Table 3.7). However, the gene model is regarded as preliminary, and primarily due to the fragmented nature of the genome, insufficient sequence reads and repeatitive regions. The low statistical nature of the gene models cannot

be attributed to the ESTs/RNA-seq sequences and/or protein evidences used for gene prediction with Exonerate, but an intrinsic characteristic of the present assembly.

Of the 214 predicted proteins, 33.65 % (> 72 proteins) did not have sufficient BLASTP similarity to proteins in other organisms to delineate functional information (Table 3.7 and Figure 3.7). It is not yet clear if these proteins (33.65 %) appear to be unique to *E. gracilis*. However, if this is the case, this is a moderate proportion as observed in other eukaryotes (Gardner, *et al.,* 2002). However, it will be interesting to know what proportion of the entire genome would have similarity with other known eukaryotic databases as this will reflect a greater evolutionary divergence between *E. gracilis* and other eukaryotic sequenced genomes. The *E. gracilis* complete genome sequence will provide more insight into this. About 142 proteins (~ 66.36 %) had BLASTP significant similarity to hypothetical or putative proteins in other organisms. 8.41 % (18) of the predicted proteins had InterPro domain hits, with 61.1 % (11) and 11.11 % (2) of the InterProScan annotated proteins having one or more transmembrane domains and putative signal peptides respectively (Table 3.7 and Figure 3.7). The Gene Ontology (GO) database is a controlled vocabulary that describes the roles of genes and gene products in organisms (Ashburner, *et al.,* 2000). GO terms were assigned automatically to the gene products. 33.65 % (72) of the proteins have GO terms, and revealed a total of 192 functional descriptions distributed among biological processes, molecular functions, and cellular structures (Table 3.5, 3.7, and Figure 3.7).

**Table 3.6:** Statistics of structural genome annotation

| Molecule | Feature | *E. gracilis* | |
|---|---|---|---|
| | | Counts | % |
| DNA | **Nucleotide stats** | | |
| | Contigs | | |
| | - Total contigs analysed (>10bp) | 1459 | |
| | - Length | 22999641 | - |
| | - Base composition (GC)* | | |
| | Overall | 11773516 | 51.19 |
| | Coding†† | 36837 | 0.31 |
| | Non coding†† | 11736679 | 99.69 |
| | - Contigs with CDS | 53 | |
| | Genes | | |
| | - Number of genes | 135 | - |
| | - Gene density (genes per Mb) | 5.87 | - |
| | - Total gene length* | 497008 | 2.16 |
| | - Average gene length (bp) | 3475 | - |
| | - Number of transcripts | 143 | - |
| | - Transcripts per gene | 1.06 | - |
| | Transcript | | |
| | - Average Length | 3475.58 | |
| | - Total Length | 497008 | |
| | - Average Coding Length | 620.91 | |
| | - Ave. exons per transcript | 3.87 | |
| | - Total Coding Length | 88790 | |
| | | | |
| | Exons | | |
| | - Number of exons | 554 | - |
| | - Total length (bp)† | 73482 | 14.79 |
| | - Average exon per transcript | 3.87 | - |
| | - Average size/length (bp) | 174.54 | - |
| | Introns | | |
| | - Number of introns | 271 | - |
| | - Total length (bp) † | 278354 | 56.01 |
| | - Average introns per transcript | 1.90 | |
| | - Average size/length (bp) | 1027.14 | - |
| | - Introns per CDS | 0.64 | |
| | - Conventional introns | 218 | |
| | - Intermediate non conventionals | 30 | |
| | - Non conventional introns | 23 | |
| | UTRs | | |
| | - Number of 5’ UTR | 264 | |
| | - Number of 3’ UTR | 266 | |

| | | | |
|---|---|---|---|
| | - Total length (bp) † | 1060 | |
| | - Average size/length (bp) | 2 | |
| | Intergenic regions | 22502633 | 98.84 |
| | tRNAs | | |
| | - Total tRNAs | 0 | - |
| | - tRNAs with introns | 0 | - |
| | rRNAs | | |
| | - Total rRNAs | 0 | - |
| | - Number of 8S rRNAs | 0 | - |
| | - Number of 28S rRNAs | 0 | - |
| | - Number of 18S rRNAs | 0 | - |
| | Splice Leaders (SL) and PolyA††† | - | - |

**Summary gene statistics of *E. gracilis* draft genome (nt):** Analysis is based on > 10 kbp contigs. Counts = finite numbers; % = percentage. nt = nucleotides. *In percentage of actual assembly size; †In percentage of total gene size; †† In percentage of overall GC length. †††Analysis not done due to incomplete genome.

**Table 3.7:** Statistics of functional genome annotation

| Molecule | Feature | *E. gracilis* | |
|---|---|---|---|
| | | **Counts** | **%** |
| Proteome | **Protein Stats** | | |
| | *Without chloroplast and mitochondria screening* | | |
| | Total proteins | 267 | - |
| | Average length | 160 | - |
| | Total length | 42737 | - |
| | *Screened for chloroplast and mitochondria* | | |
| | Total proteins | 214 | |
| | Average length | 155 | |
| | Total length | 33290 | |
| | | | |
| | **BLAST** | | |
| | **With BLAST hits*** | **142** | **66.34** |
| | Top 7 species distribution† | | |
| | **Genlisea aurea* | 16 | 11.27 |
| | **Albugo laibachii* | 8 | 5.63 |
| | **Acanthamoeba castellanii* | 5 | 3.52 |
| | **Aureococcus anophagefferens* | 4 | 2.82 |
| | **Trypanosoma cruzi* | 3 | 2.11 |
| | **Guillardia theta* | 3 | 2.11 |
| | **Euglena gracilis* | 3 | 2.11 |
| | *Other species | 100 | 70.42 |
| | | | |
| | **InterProScan** | | |
| | **With Interpro Domain*** | **18** | **8.41** |
| | With atleast one TM domain† | 11 | 61.11 |
| | With putative signal peptides† | 2 | 11.11 |
| | | | |
| | **Functional Classification** | | |
| | **Total annotated sequences based on GO Mapping** | **72** | **33.65** |
| | - Biological processes†† | 82 BP | - |
| | - Molecular functions †† | 57 MF | - |
| | - Cellular components†† | 53 CC | - |
| | Kegg maps | | |
| | - Pathways | 3 | - |

**Summary comparative statistics of *E. gracilis* draft genome (aa) Keys:** signal peptides have > 0.95 % specificity; aa = amino acid; *percentages are with respect to total protein sequences (267). †percentages are with respect to total BLAST hits (142), total InterProScan hits (85). ††Number of identified processes, functions, and components. BP – Biological Process; MF = Molecular Function; CC = Cellular Component.

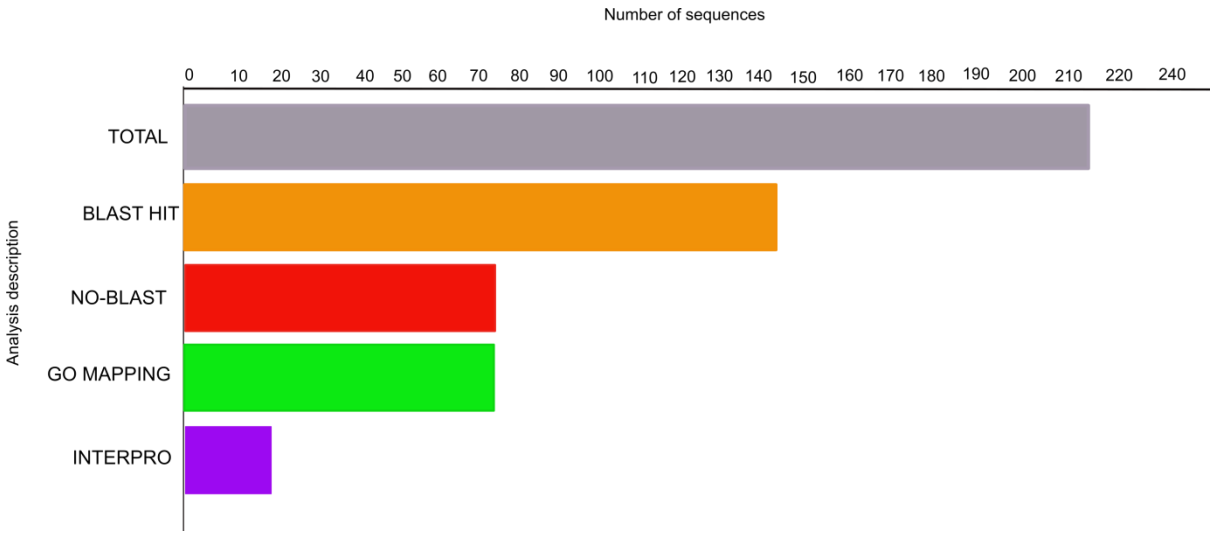



**Figure 3.7: The *E. gracilis* predicted proteome (genome) is characteristic of moderate sequence homology.** The figure shows the data distribution of functional genome annotation**.** TOTAL (gray) = total number of protein sequences for analysis; BLAST HIT (orange) = proportion of protein sequences with BLAST hits against the NCBI NR database; NO-BLAST HIT (red) = proportion of sequences without a BLAST hit against the NCBI NR database; GO MAPPING (green) = proportion of BLASTed sequences with Gene Ontology (GO) mapping; INTERPRO (purple) = proportion of protein sequences with InterProScan hits.

## 3.4 DISCUSSION AND CONCLUSION

### *3.4.1 DNA isolation, sequence data quality, Assembly evaluation and assessment*

The *E. gracilis* genome has been sequenced; with multiple illumina, Roche 454, and Pacbio libraries, and an assembly of the data into draft. The sequenced data is of moderate quality. The assembly improvement and evaluation for the genome shows that it‘s only ~ 20.01 % complete. The genome could only support an initial or preliminary resolution into gene structure and products, and could not support an extensive annotation analysis as well as comparison with other eukaryotic/kinetoplastid taxa such as *T. brucei*, and *L. major.* The GC content of *E. gracilis* genome is 50.60 % (Table 3.6) which is also at par with the GC content of *T. brucei* and *L. major,* and suggesting that the complete *E. gracilis* genome may be more GC rich than AT rich. This also suggests that most Excavates are likely to have an average GC content of ~ 50 %. In terms of the contig metrics assessment, *T. brucei* and *L. major* had larger N50, and > 36 % of their genome coding for proteins (data not shown). It is not yet known what proportion of the *E. gracilis* genome codes for proteins until the genome is fully sequenced. However, this proportion is < 2 % (< 40 Mb) of the estimated 2 Gbp. In the present data of 1.4 Gbp actual genome size, this will be ~ 28 Mbp that are coding while the remaining are non coding sequences.

### *3.4.2 Structural and functional genome annotation*

<u>In silico analysis</u>

The predicted genome sizes of *E. gracilis* are variable in comparison with other members of the eukaryotes (see Figure 3.3 and Ebenezer, *et al.,* 2017 for details). The number of chromosomes in *E. gracilis* has also been somewhat contentious for over half a century, with estimates varying between four (O’Donnell 1965) and fourty-five (Leedale, 1958a, b). In addition to the uncertainty concerning chromosome number, the size of the nuclear genome remains unknown, with estimates varying by an order of magnitude in *E. gracilis* (Zakrys, 1988, Ebenezer, *et al.,* 2017), variations and surprisingly large genome sizes within members of the Euglena genus (Zakrys, 1988), extensive repeat elements (Rozby, 1985), and changes in genome sizes within a single species with respect to environmental perturbations (see Ebenezer, *et al.,* 2017 for details). While *E. gracilis* is distantly related to the kinetoplastids, it is

also not yet known if *E. gracilis* also retains some level of higher order gene organisation that is similar to that in the kinetoplastids.

The most basic metrics for genome assemblies involves the size of the assembled sequence in relation to the inputed reads. This includes intrinsic statistics such as the assembly size, percentage of reads assembled into contigs, and counts of contigs and/or scaffold. As the *E. gracilis* genome reads coverage increased, the percentage of reads incorporated into the assembly increased proportionately in the assembly (Table 3.2, 3.3). Other statistical parameters includes the inherent properties of the assembled sequences including length, coverage, and N50 which are all satisfactory (except for the N50) considering the amount of sequence data and coverage. The comparison of distantly related eukaryotic genomes (*T. brucei* and *L. major*, and other super groups) provided a useful opportunity for further interpretation – suggesting overall that the genomes of *T. brucei* and *L. major* are much better (data not shown) when compared with the present genome data.

Generally, the assembled draft genome provides a slight overview of gene content, functions, and scaffold organization in the *E. gracilis* 2066288 supercontigs, spanning ~ 1.4 Gbp of DNA. The subjection of the sequenced genome assemblies to multiple insilico analysis also revealed several protein-coding genes (but not non coding RNAs such as tRNA and rRNA) genes, similarities to ESTs and other proteins in protein databases. The ~ 1.4 Gbp total sequences showed that 22 Mbp are > 10 kbp, which contain 135 predicted protein-coding genes. All the genes have been evaluated, and revealed an average density of 5.87 predicted gene per Megabases. Each gene has an average of 2.01 introns, and 0.32 **%** (73482) of the genome (> 10 kbp) resides in predicted exons. The number of genes is less than those found in *L. major* and *T. brucei* (the > 10 kbp sequences). However, it is predicted that the number of genes present in the actual Euglena genome sequence will most likely be more than that observed in *T. brucei, L. major,* and the human genomes (IHGC, 2004). This prediction stems from the analysis of the version 0.7 genome assembly which produced 44,162 protein-coding genes as the then actual and estimated genome sizes of 650 Mb and 750 Mb respectively. If this is the case, then the predicted protein-coding genes for Euglena will be about that observed in the human genome (IHGC, 2004). This high number may be due to its huge genome

size, but this preliminary in silico analysis is not linear considering that the human genome is ~ 3 Gb in size.

Similarities to known proteins provide an understanding into the possible functions of the predicted genes. Top BLASTP hits shows protein products have close matches within the Excavates and related supergoups (Table 3.7); most of these matches contain functional information. In addition to the protein-coding genes, the draft genome (> 10 kbp) does not contain any gene for non-coding RNAs such as tRNAs and rRNAs. The structural and functional preliminary genome annotation revealed the presence of several genes and pathways. Among these, and most represented, are the genes involved in purine metabolism, pyrimidine metabolism, and thiamine metabolism (see Table 3.7). The *E. gracilis* genome sequences provide insights not only into the unique aspects of the biology of this complex organism, but also eukaryote evolution, given their early divergence. The availability of the entire or complete gene content of *E. gracilis* will provide the platform for the identification and wholistic functional analysis of eukaryotic evolution due to their position in the eukaryotic phylogenetic tree. Sequence comparisons to sister supergroups within Excavates and Chlorophyte, may provide useful information on the evolution of parasitism and endosymbiosis. To this end, the genome is ~ 20.01 % complete; and there are several reasons for completely sequencing a genome, however, the most important over time and as described by Hillier, *et al.,* 2005 is that the whole is an archive for the future, containing all the genetic information required.

It is not fully clear why the *E. gracilis* genome (version 0.9) is about 20.01 % complete in terms of core eukaryotic genes. However, there are three possible reasons for this, and none of these is as a result of assembly protocols, rather it may be that the *E. gracilis* genome may require a novel bioinformatics assembly algorithm. First, the estimated coverage of the *E. gracilis* genome version 0.9 is likely to be realistically low, or perhaps undercovered by the version 0.9 sequenced reads, when compared to the actual genome size (this value is in the order of 2 Gb) of *E. gracilis.* Though 1 Gbp – 9 Gbp have been reported in many scientific literatures (Ebenezer, *et al.,* 2017); thus missing key genes. Secondly, the fragmented nature of the genome and the unavailability of long insert size libraries (>= 100 kb) means that genes spanning fragmented scaffold lengths tend not to be contiguous, with many

scaffolds not containing either of the genetic bases (A, C, T, or G) or containing missing bases represented by Ns. This instance is over half of the genome, meaning that half of the genome is gapped or have unknown base representation and are replaced by gaps (called N‘s). This long insert reads may be obtained from either sufficient PacBio or Oxford Nanopore sequencing – but this will be very expensive, with the PacBio having a high error rate. Thirdly, the insufficient huge data of short insert sizes (250 bp – 540 bp) means that gapped regions (or regions with unknown bases) cannot be completely filled. Even with huge illumina short reads data, I still expect that this will be highly fragmentary; hence the solution lies with longer reads. Fourth, are the potential issues of repeats and Base J that has also been reported in scholarly articles. While the repeat challenge are evident from the genome assembly (whereby some genes do have multiple repeats) and will require development of novel bioinformatics assembly pipelines to deal with polyploidy in *E. gracilis*, I do not yet have clear evidence yet for the later. However, I sent some DNA samples to Peter Myler at the University of Washington for PacBio sequencing and Base J investigation. He could not find any evidence for problems arising from Base J, partly due to low coverage – but this will be worth investigating extensively. These points need to be factored in the provision of a high quality genome for *E. gracilis* in the future.

*3.4.3 Significance of findings – adaptations, functions, and applications*

The genomic organization and function of the *E. gracilis* genome was investigated to understand its structure and content. This is useful in understanding the organization of it genes, splicing mechanisms, intron and exon architecture and sizes, genome size, size of intergenic regions, repeat elements, as well as the proportion of GC:AT content. The understanding of this utility play specific roles in understanding how the *E. gracilis* genome structure and content influence biological functions in this model organism. Over the next several paragraphs, I will discuss the significance of the findings in the *E. gracilis* genome with reference to adaptions, functions (biology), and applications.

*Large genome size, repetitive elements, large and unusual introns correlate with increased biological complexity and function:* There are several factors that could affect biological complexity in an organism: Genome size, number of genes, large

introns (in terms of size and numbers), and repetitive/selfish DNA (see Latorre and Silva, 2013, Pray, 2008, Gregory, 2005, for extensive reviews). In prokaryotes (Archae and Bacteria), there is a linear correlation between genome size, biological complexity, number of genes and lifestyle. For instance, bacteria with smaller genomes are specialists, such as obligate parasites and endosymbionts, and bacteria with larger genomes are generalists, and may even have a certain degree of development, such as sporulation in *Bacillus* (Latorre and Silva, 2013). In eukaryotes, there is no linear correlation between genome sizes and biological complexity (known as the C-value paradox) (Latorre and Silva, 2013). For instance, the marbled lungfish, *Protopterus aethiopicus*, has more than 40 times the amount of DNA per cell than humans. One haploid copy of this fish's genome is composed of a whopping 132.8 billion base pairs, while one copy of a human haploid genome has only 3.5 billion base pairs. It is also clear that the biology of the marbled lungfish cannot be more complex than humans, otherwise, humans would have at least as much DNA as the marbled lungfish, although probably much more. Therefore, genome size is clearly not an indicator of the genomic or biological complexity of an organism (Latorre and Silva, 2013). On the contrary, in eukaryotes, number of genes correlates with increased biological complexity (Latorre and Silva, 2013), though contrary opinions had previously been reported (Pray, 2008). For instance, the nematode worm *C. elegans* has 18,000 genes, about 5,000 more than Drosophila, a more complex organism. *H. sapiens* has ~2000 genes more than *C. elegans*, with the biology of *H. sapiens* being more complex. In *E. gracilis*, while the large genome size may be attributed to repetitive sequences and hybrid genome phenomenon, there are more than 35,000 predicted genes suggesting that the biology of *E. gracilis* may be more complex than the kinetoplastids and *H. sapiens* with smaller and moderately large genomes respectively, following this assumption or theory. *E. gracilis* may be one of the eukaryotic members to demonstrate the linear relationship between genome size and biological complexity. While I have also reported varying predicted genome sizes in *E. gracilis* (Figure 3.3), mutational mechanisms leading to whole genome duplications (or loss and gain of nucleotides) (Latorre and Silva, 2013) and varying analytical technologies may be attributed to this variance in predicted genome sizes. The size of intergenic/non genic DNA also correlates with the genome sizes. This is because intergenic/non genic DNA are not only accumulated but lost during evolution, since the continous accumulation would

create a metabolic and structural burden on the organism (Floreano and Mattiussi, 2008). It has also been argued that spontenous deletion of intergenic regions/non genic DNA (mutational and deletion bias, and recombination phenomenon) may be a major factor to account for different genome sizes (Hartl, 2000, Ometto, *et al.,* 2005) which play roles in several evolutionary adaptations (Floreano and Mattiussi, 2008). This could explain the variations in genome sizes between *E. gracilis* (2 Gbp) and trypanosomatids (e.g. *T. brucei* = 35 Mbp). Similarly, the presence of both the conventional and non conventional introns in *E. gracilis* shows the presence and utility of several regulatory mechanisms within this system. The alternation between high and low gene density regions in the E. gracilis genome (Figure 3.4, 3.5, and 3.6) has been suggested to be common in plants with large genomes such as *Glycine max* (1,115 Mb) and *Zea mays* (2,300 Mb) (Pingault, *et al.,* 2015). This predicts a non-random distribution of genes along chromosomes in a complete *E. gracilis* genome, resulting in clusters of genes sharing the same expression profile, the same function or involved in the same metabolic pathway (Lee and Sonnhammer, 2003, Liu and Han, 2009, Ren, *et al.,* 2005, 2007, Williams and Bowles, 2004, Xu, *et al.,* 2008, Zhan, *et al.,* 2006).

*Moderate sequence homology and low database annotation information reaffirms Euglena biological uniqueness:* As previously mentioned, *E. gracilis* genome analysis showed low to moderate protein database (NCBI, InterPro, GO) hits. The fact that some of the proteins have moderate BLAST (66.34 %), low InterPro (8.41) and Gene Ontology hits show that there is a low sequence homology and protein domain annotations representation for Euglena on the protein databases. Many of these proteins are classified as hypotheticals, with about 33.66 % not showing sequence similarity. While these sequences do have predicted ORFs/proteins, it remains to be investigated if these are real proteins or artifacts. The orthologous group clustering (Chapter 4) analysis suggest that an estimated 150 orthologous groups are Euglena-specific. If this is the case, the predicted ORFs/proteins in the genome data without protein database hits might be some of the Euglena-specific proteins. It is fair to anticipate the absence of these Euglena-specific proteins considering that there are no comprehensive sequence data for *E. gracilis* on any of the protein databases.

*Nucleotide and protein statistics, introns, exons, and scaffold structure predicts chromosome architecture and function:* Of the several statistical parameters, GC content is one of the most important. GC content has been predicted to affect genome functioning and species ecology (Smarda, *et al.,* 2014), with several hypothesis linking GC content variation with several biological fundamental processes. In prokaryotes, several lines of evidences exists to suggest that GC content variation is governed by genome replication and DNA repair mechanisms (see Wu, *et al.,* 2012 for details). In eukaryotes, and most prokaryotes, the variations in GC content is correlated with genome size, ecological and evolutionary adaptations, and natural selection (gene mutation). They could either play subsidiary roles or rely indirectly on different mutator genes to fine-tune the GC content (Wu, *et al.,* 2012). Very high GC content may mean large genome size, while very low GC content may mean small genome size. The moderate GC content of *E. gracilis* genome (though higher GC content in the transcriptome) may provide explanations for it's large genome size, ecological adaptability and resilience (and survivability through natural selection). In higher plants, such as monocots, GC content has been associated with holocentric chromosomal structure (entire length of the chromosome acts as the centromere) (Smarda, *et al.,* 2014).

*Introns, exons, and transcripts play huge functional benefits in gene expression.* In Table 3.6, transcript arrangement shows an average of 1.06 transcript per gene. One explanation for this is the presence of alternative splicing events in the genome (which is not evident in the transcriptome, see Chapter 4). This is important considering the role of alternative splicing in genetic control, and a crucial step in gene expression (see Chapter 6). Introns are predicted to occupy 56.01 % of the *E. gracilis* genome (Table 3.6). This means that the *E. gracilis* cell will require an extensive amount of energy to splice out these introns during transcription, considering that the density of introns (i.e., the genic regions consuming large amounts of energy for nothing in terms of protein synthesis) is greater than that of exons in genomes. However, introns are of huge utility in eukaryotic organisms where they perform putative functional roles in splicing, mRNA transport, nonsense mediated decay, expression regulation as well as mutational buffers (see Jo, *et al.,* 2015, for details). This means that there is an extensive biological benefit of *E. gracilis* intron:exon ratio that has a huge energy cost, implying that high intron:exon

load results in specific and efficient gene expression and a greater transcriptional complexity (Pingault, *et al.,* 2015). How *E. gracilis* generate this massive amount of energy will be an interesting area of investigation.

*Genome coverage may be an imperfect 'numerator' for E. gracilis genome sequencing:* Genome coverage is one of the factors considered when carrying out a Next Generation Sequencing (NGS) project. As a rule of thumb, it is believed that genome coverages of up to 20X provides strong evidences for the presence of conserved genes (Sims, *et al.,* 2014). However, this is not the case with *E. gracilis* genome where > 200X coverage still lacks core eukaryotic genes (as evident from the CEGMA score and read mapping evidence) – this may be due to other factors such as read length and sequencing platforms, suggesting that genome coverages may be an imperfect singular criteria for genome sequencing. This also means that genome coverage required to capture more than 90 % of genes in an organism may be different across taxa. Other indirect numerators would include a combination of the contig matrix, CEGMA, and read mapping evidence which have proven to present a reasonably precise test for assembly quality, though these are less likely prior to sequencing.

# CHAPTER FOUR

## 4.0 TRANSCRIPTOME ARCHITECTURE, FUNCTION AND EVOLUTION

### 4.1 INTRODUCTION

In this chapter, the *E. gracilis* transcriptome organization, function and evolution were investigated. The *E. gracilis* transcriptome was sequenced using the Illumina sequencing platform as well as the aquisition of multiple sources from sequence databases and kind donations. The final transcriptome is ~ 90 % complete in terms of Core Eukaryotic Genes. Approximately 90 % of all the transcriptome reads (including 23k ESTs on NCBI) map back to the transcriptome assembly. Part of this chapter was done in collaboration with Steve Kelly at the Department of Plant Sciences, University of Oxford, UK. In this chapter, the focus will be on the transcriptome assembly and automatic structural and functional annotation. The community-based manual functional curation will be discussed in more details in chapter 5.

### 4.2 TRANSCRIPTOME SEQUENCING AND ASSEMBLY

#### *4.2.1 RNA isolation and purification*

RNA isolation and purification for K1E (sample name) reads were previously performed by Carrington, *et al.,* in 2012 and sequenced at the Beijing Genomic Institute (BGI), Hong Kong.

#### 4.2.2 *Library preparation and Illumina sequence data*

For the transcriptome sequence, two paired-ends reads illumina Hiseq 2000 sequencing runs were carried out using two fragment libraries with insert sizes of ~ 125 bp and 189 bp respectively (Table 3.1). Approximately 466 M reads were generated with read lengths of 100 bp (> 3 Gb in total clean bases). In general, the qualities of these reads were satisfactory, showing phred quality values (Q20) ≥ 90 % and moderate GC content between 55 – 60 % (see Table 3.1). To validate the quality of the transcriptome assembly, one Roche/454 GS FLX runs single-end read was acquired, and ~23k EST sequences were obtained from the NCBI database for transcriptome assembly probing and read/sequence mapping or alignment. The

Roche/454 GS FLX dataset consists of 582, 319 reads, and 214 Mbp in total length, with an average read length of 367 bp (data not shown) (the FastQC sequence visualization .html files for all read sequences in Table 3.1 are available on request).

4.2.3 *Transcriptome assembly, assembly improvements and variant detections, and assembly evaluation and assessments*

Quality control and actual assembly

To completely assemble the transcriptome, quality trimmed and error corrected reads of the KIE_l1 illumina and DOHN9ACXX (sample name) illumina read libraries (kind donation by Rob Field) (see Table 3.1) were assembled into final/complete transcriptome (Figure 2.2). The final assembly consisted of a total of 72509 sequences, with a total length and N50 of 63050794 and 1242 respectively (see Table 4.1, 4.2 and 4.3 and Chapter 2, sub section 2.4).

Assembly evaluation and assessment

More like the genome, the transcriptome assembly workflow employed several QC analysis pipelines for validating assembly completeness, and these include: Transrate statistical metrices, CEGMA statistical metrices, RNA-seq reads and EST sequence mapping evidences. As a gene discovery project, the most important evaluation is the CEGMA statistical metrics since it tells us the proportion of core eukaryotic genes (CEG) present in the respective assemblies. The CEG datasets were aligned to the transcriptome assembly in order to determine the percentage CEG‘s contained in the assembly. According to Bradnam, *et al.*, 2013, it is assumed that an assembly which is more complete will have a higher percentage of complete and partial CEG alignments than a less complete assembly (see Parra, *et al.,* 2007; and CEGMA metrices below for details).

**Table 4.1:** Comparative transrate statistical description of transcriptome assembly (transcripts)

| **Metrics** | **Parameter** | ***E. gracilis*** | ***T. brucei*** | ***L. major*** |
|---|---|---|---|---|
| Contigs/Transcripts | n seqs | 72509 | 12094 | 9378 |
| | smallest | 202 | 3 | 54 |
| | largest | 25763 | 66816 | 52179 |
| | n bases | 63050794 | 22255660 | 16084351 |
| | mean_len | 869.55818 | 1834.86 | 1706.75 |
| | n under 200 | 0 | 628 | 878 |
| | n over 1k | 19740 | 8585 | 5679 |
| | n over 10k | 25 | 66 | 69 |
| | n with orf | 30467 | 9855 | 7764 |
| | mean orf percent (%) | 78.77 | 73.47 | 99.13 |
| | N50 | 1242 | 2425 | 2511 |
| | gc (%) | 60.9 | 48.32 | 62.44 |
| | bases n | 0 | 1054 | 2 |

**Keys: n** = numbers; **p** = proportion; **bases_n** = number of bases that are N or number of bases with gap; Smallest = minimum contig length; Largest = maximum contig length; Mean_len = mean sequence length; Mean ORF percent (%) = percentage sequences with open reading frame; N50 = proportion of sequences that makes up half the size of the genome; gc (%) = percent proportion of Guanine and Cytosine. Contig metrics are measures based entirely on analysing the set of contigs or the intrinsic statistical characteristics of a contig. Read mapping metrics are based on aligning the reads used in the assembly to the assembled contigs. Comparative metrics involve comparing the assembly to a related reference species. See Smith-Unna, *et al.,* 2015 for detail definition of the vocabularies and statistical parameters in this table. This table contains the most basic parameters; additional statistical parameters are available on request. Keys are as described by Smith-Unna, *et al.,* 2016.

**Table 4.2:** Transcriptome assembly statistics (CDS)

| Parameter | CDS | Proteins |
|---|---|---|
| Number of sequences | 36526 | 36526 |
| Median sequence length | 765 | 254 |
| Mean sequence length | 1041 | 346 |
| Max. sequence length | 25218 | 8406 |
| Min. sequence length | 297 | 98 |
| No. sequence > 1kbp | 13991 | 1290 |
| No. sequence > 10kbp | 24 | - |
| N50 | 1413 | 471 |
| Combined sequence length | 38030668 | - |

**Keys:** Sequences refer to either nucleotide or protein sequence. Max. and Min. = Maximum and Minimum respectively; No. sequence > X kbp = number of sequences greater than length X; N50 = proportion of sequences that makes up half the size of the genome.

**Table 4.3:** Statistical description of transcriptome read mapping evidences

| Sequencing technology | Read library ID* | Total number of reads aligned | % paired/unpaired | Overall alignment rate (%) |
|---|---|---|---|---|
| Illumina HiSeq 2000 | KIE_l1 | 13412134 | 100 | 88.05 |
| Illumina HiSeq 2000 | DOHN9ACXX | 9924418 | 100 | 92.13 |
| Roche/454 | A107065 | 587000 | 100 | 91.45 |
| NCBI ESTs | ESTs | 23394 | 100 | 91.99 |

**Keys:** Sequencing technology means platform where sequencing was carried out. Read library ID designates the unique identifier for the sequencing experiment. Total number of aligned reads = number of reads from the corresponding read library that maps to the transcriptome assembly using Bowtie2; % paired/unpaired = proportion of sequences that are either paired or unpaired; Overall alignment rate (%) = proportion of corresponding read libraries that maps to the transcriptome assembly. See Table 3.1 for a description of these read libraries.

Transrate

The final transcriptome assembly was assessed comparing three standard metrics: the contig matrics, read mapping metrics, and the comparative metrics (Smith-Unna, *et al.,* 2015) commonly used in ascertaining the quality of an assembly (see Table 4.1, 4.2, 4.3 and Chapter 2, sub section 2.4). In general, statistical parameters include: total number of contigs; longest contig length; mean and median contig length; N50 (the median contig size, length weighted); and the summed contig lengths, or the raw size of the draft transcriptme assembly (see Smith Smith-Unna, *et al.,* 2015 for detail definition of these vocabularies and statistical parameters). According to Hornett and Wheat, 2012, the contrasting insights provided by the basic metrics illustrate their limited utility. They argued that metrics based upon contig lengths (e.g. mean, median, N50) do not provide quantitative insights into how much of the target species transcriptome is represented in the de novo transcriptome assembly.

Another metric adopted from Hornett and Wheat, 2012, is the calculation of additional metric to gauge the quality of the transcriptome assembly by taking advantage of the genomic information of reference species such as predicted proteomes. Bradnam, *et al.,* 2013 proposed that an optimal assembly will have near full length contigs similar to that expected from the actual transcriptome of the target species. In this instance, annotated transcripts from *T. brucei* and *L. major* were used as references (see Table 4.1). For the contig metrices, the numbers of sequences are 72509, 12094, and 9378 for *E. gracilis, T. brucei,* and *L. major* respectively, the N50 is 1242, 2425, and 2511 for *E. gracilis, T. brucei,* and *L. major* respectively (see Table 4.1 for additional details), meaning that 50 % of the assembled sequences are longer than these respective values. For the read mapping metrics, the number of read pairs mapping in a way indicative of good assembly (good mappings), as defined by Smith-Unna, *et al.,* 2015, for *E. gracilis, T. brucei,* and *L. major* respectively (data not shown) was satisfactory for *E. gracilis* but not for the Kinetoplastids – hence suggesting possible evolutionary divergence. The GC content of predicted open reading frames (see Materials and Methods) of the final assembly is 60.90 %, which is relatively at par value compared to the coding sequence GC content of *T. brucei* (48.32 %) and *L. major* (62.44 %) transcripts. Why *E. gracilis* and other Excavates maintain a low GC content is still not yet understood.

Generally, it has been shown that coding regions are more GC-rich compared to surrounding genomic regions (Eyre-Walker and Hurst, 2001; Lander, *et al.,* 2001). To estimate the sequencing depth in the assembled transcriptome (see Materials and Methods), all read libraries and ESTs were mapped to the final assembly and assessed. Out of 40780.72 Megabases of inputted reads, 63 Megabases (2.23 %) of these have matching transcriptome assembly sequences (Table 3.1 and Table 4.1), while direct RNA-seq read and EST mapping with Bowtie2 suggests that > 88 % are represented in the assembled transcriptome (Table 4.3).

Core Eukaryotic Gene Mapping Approach (CEGMA)

Table 3.2 presents the proportion of CEGs contained in the *E. gracilis* transcriptome assembly (87.90 %), as well as when compared with *T. brucei* (82.66 %) and *L. major* (78.23 %) respectively. Out of these, 75.40 %, 76.61 %, and 53.63 % are complete respectively. Bradnam, *et al.,* 2013, posited that the performance of CEGMA in determining complete or partial genes differ, and could be attributable to several reasons, including fracturing of a given genic region across multiple scaffolds within an assembly, or exons lying in gaps within a single scaffold. They further suggested that it is also possible that for some, highly paralogous genes, CEGMA is detecting a paralog and not the true ortholog - this means that the assemblies likely contain additional core genes that are too fragmented to be detected by the original analysis as proposed for other eukaryotic transcriptomes.

## 4.3 TRANSCRIPTOME STRUCTURAL AND FUNCTIONAL AUTOMATIC ANNOTATION

### *4.3.1 Gene predictions, features, ORF distribution and statistics*

Gene prediction was carried out using GMST and TransDecoder. The transcriptome assembly recorded a GC base composition of 60.87 % (38379018) out of the 63 Mbp transcript sequences, with 26.26 % (25096438) coding. ORF distribution and statistics are presented in Table 4.1 and 4.2. The number of transcript with open reading frames and mean ORF percent in the transcriptome assembly is 36526 and 73.38 % respectively. These values are a little and much lower in *T. brucei* and *L. major* 9855 (73.47 %) and 7764 (99.13 %) respectively, implying that *E. gracilis* is likely to have more protein-coding genes than *T. brucei* and *L. major,* and other sister groups. There were 36526 predicted genes with a corresponding 36526 transcripts – 1 transcript per gene. The gene density, average gene length and total gene length are 1.36 per Mbp, 1266 bp, and 46269013 bp respectively. There were 36526 coding sequences with an average length and total length of 1072.08 and 27790548 respectively. ORFs and exons make up 73.38 % and 60.06 % of the transcriptome assembly and genes respectively. There were 5280 and 13013 5'UTR and 3'UTR respectively (Table 4.4). tRNA and rRNA non conding candidate genes were not found.

### *4.3.2 Blast homology search*

The assembly was annotated using Blast2GO automatic functional annotation tool (see Materials and Methods). The results of each step of Blast2GO annotation of the transcriptome assembly are summarized in Table 4.4, 4.5, and Figure 4.1, 4.2. All 36,526 predicted proteome (translated transcripts) of the final assembly were interrogated against the NCBI non-redundant protein database (nr) using BLASTP. 82.22 % (30033) of all translated transcripts (Table 4.4 and 4.5) showed similarity to at least one protein in the nr database (Figure 4.1), with an evalue cutoff of 10. A reduction in the number of genes predicted in the transcriptome is anticipated; this is taking into account the < 2 % of the genome size which is coding as well as the ~ 25 % of the total genes that are single copies. I anticipate that the production of a complete *E gracilis* genome will reveal the actual number of genes present in this

species. Top taxa distribution or best BLAST hit was also populated. The top or highest proportion of best BLAST hit include: *Bodo saltans, Guilardia theta, Naegleria gruberi, Chrysochromulina spp., and Euglena gracilis*, suggesting a high level of relatedness to other sister groups or other photosynthetic groups (Table 4.5).

**Table 4.4:** Statistics of transcriptome structural annotation

| Molecule | Feature | *E. gracilis* | |
|---|---|---|---|
| | | Counts | % |
| | **Nucleotide stats** | | |
| | Transcriptome assembly | | |
| DNA/RNA | - Length | 63050794 | - |
| | - Base compostion (GC)* | | |
| | Overall | 38379018 | 60.87 |
| | Coding | 25096438 | 26.26 |
| | Non coding | 13282580 | 34.61 |
| | Genes | | |
| | - Number of genes | 36526 | |
| | - Gene density (genes per Mb) | 1.36 | |
| | - Total gene length* | 46269013 | 73.38 |
| | - Average gene length (bp) | 1266 | |
| | - Number of transcripts | 36526 | |
| | - Transcripts per gene | 1.00 | |
| | Transcripts (based on ORF) | | |
| | - Average Length | 1266.74 | - |
| | - Total Length | 46269013 | - |
| | - Average Coding Length | 1072.08 | - |
| | - Average exons per transcript | 3.82 | |
| | - Total Coding Length† | 39158820 | 77.75 |
| | Exons | | |
| | - Number of exons | 36526 | - |
| | - Total length (bp) † | 27790548 | 60.06 |
| | - Average exon per transcript | 1 | - |
| | - Average size/length (bp) | 1181.32 | - |
| | UTRs | | |
| | - Number of 5’ UTR | 5280 | |
| | - Number of 3’ UTR | 13013 | - |
| | - Total length (bp) † | 4721467 | 10.20 |
| | - Average size/length (bp) | 229 | - |
| | tRNAs | | |
| | - Total tRNAs | 0 | 0 |
| | - tRNAs with introns | 0 | |
| | rRNAs | | |
| | - Total rRNAs (8s_rRNA) | 0 | 0 |

**Summary gene statistics of *E. gracilis* transcriptome assembly (nt). Keys:** nt = nucleotides; *In percentage of actual assembly size; †In percentage of total gene or transcript size; ORF = Open Reading Frame; UTR = UnTranslated Region.

**Table 4.5:** Statistics of transcriptome functional annotation

| Molecule | Feature | *E. gracilis* | |
|---|---|---|---|
| | | Counts | % |
| Proteome | **Protein stats** | | - |
| | Total proteins | 36526 | - |
| | Average length | 346 | - |
| | Total length | 12655111 | - |
| | | | |
| | **BLAST** | | |
| | **With BLAST hits*** | 30033 | 82.22 |
| | Top 5 species distribution† | | |
| | **Bodo saltans* | 1220 | 4.06 |
| | **Guillardia theta* | 591 | 1.97 |
| | **Naegleria gruberi* | 574 | 1.91 |
| | **Chrysochromulina sp.* | 564 | 1.88 |
| | **Euglena gracilis* | 471 | 1.57 |
| | **InterProScan** | | |
| | With INTERPRO DOMAIN* | 23866 | 65.34 |
| | With atleast one TM domain† | 6442 | 26.98 |
| | With putative signal peptides† | 2551 | 6.98 |
| | | | |
| | **Functional Classification** | | |
| | **Total annotated sequences based on GOs Mapping** | 14112 | 38.64 |
| | - Biological processes | 40 BP | - |
| | - Molecular functions | 12 MF | - |
| | - Cellular components | 22 CC | - |
| | Kegg maps | | |
| | - Pathways | 89 | - |

**Summary comparative statistics of *E. gracilis* transcriptome assembly (aa).** Key: aa = amino acid; *percentages are with respect to total protein sequences. †percentages are with respect to total BLAST hits, total INTERPRO hits, or total annotated sequences. TBD = to be determined. BP = Biological Processes, MF = Molecular Function, CC = Cellular Component.

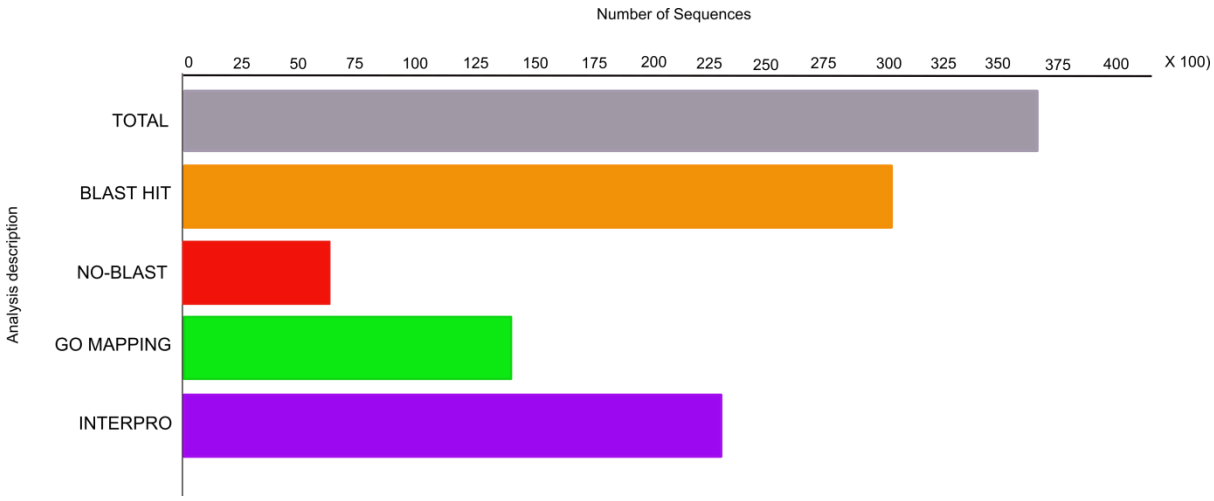
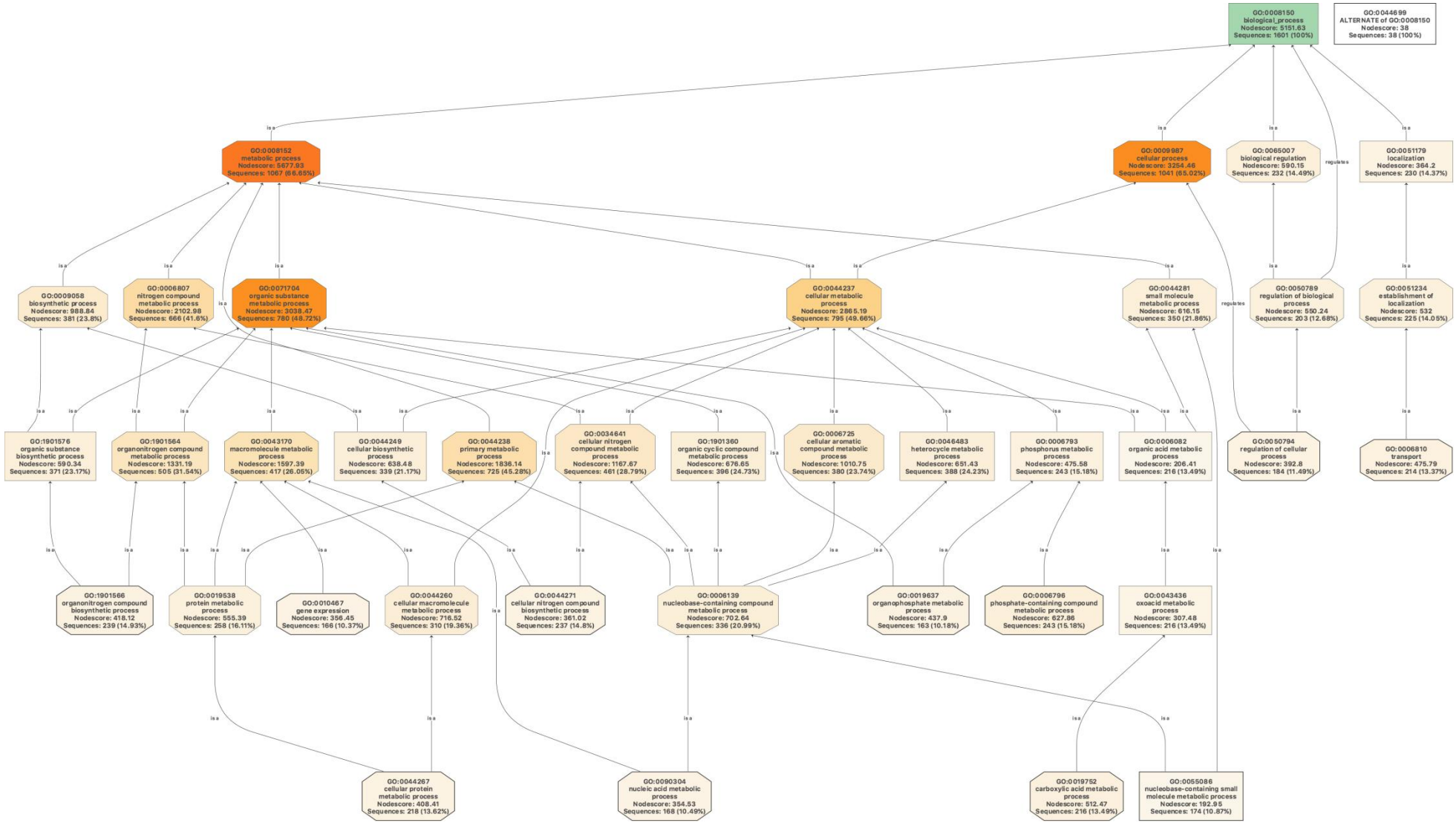
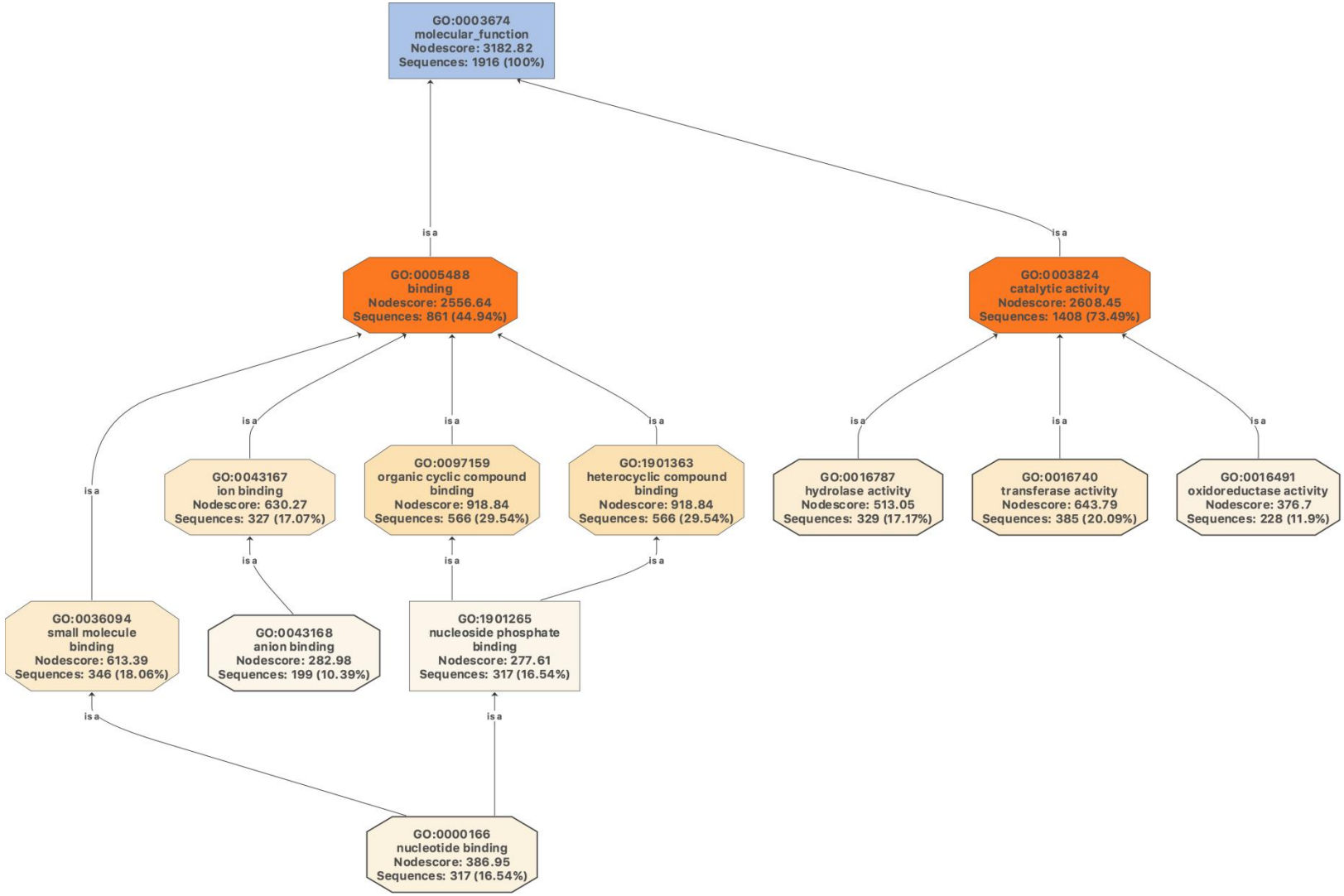



**Figure 4.1: The *E. gracilis* predicted proteome (transcriptiome) is characteristic of high sequence homology.** Data Distribution of functional transcriptome annotation. TOTAL (gray) = total number of protein sequences for analysis; BLAST HIT (orange) = proportion of protein sequences with BLAST hits against the NCBI NR database; NO-BLAST HIT (red) = proportion of sequences without a BLAST hit against the NCBI NR database; GO MAPPING (green) = proportion of BLASTed sequences with Gene Ontology (GO) mapping; INTERPRO (purple) = proportion of protein sequences with InterProScan hits.

## A: Biological process



## B: Molecular function



## C: Cellular component

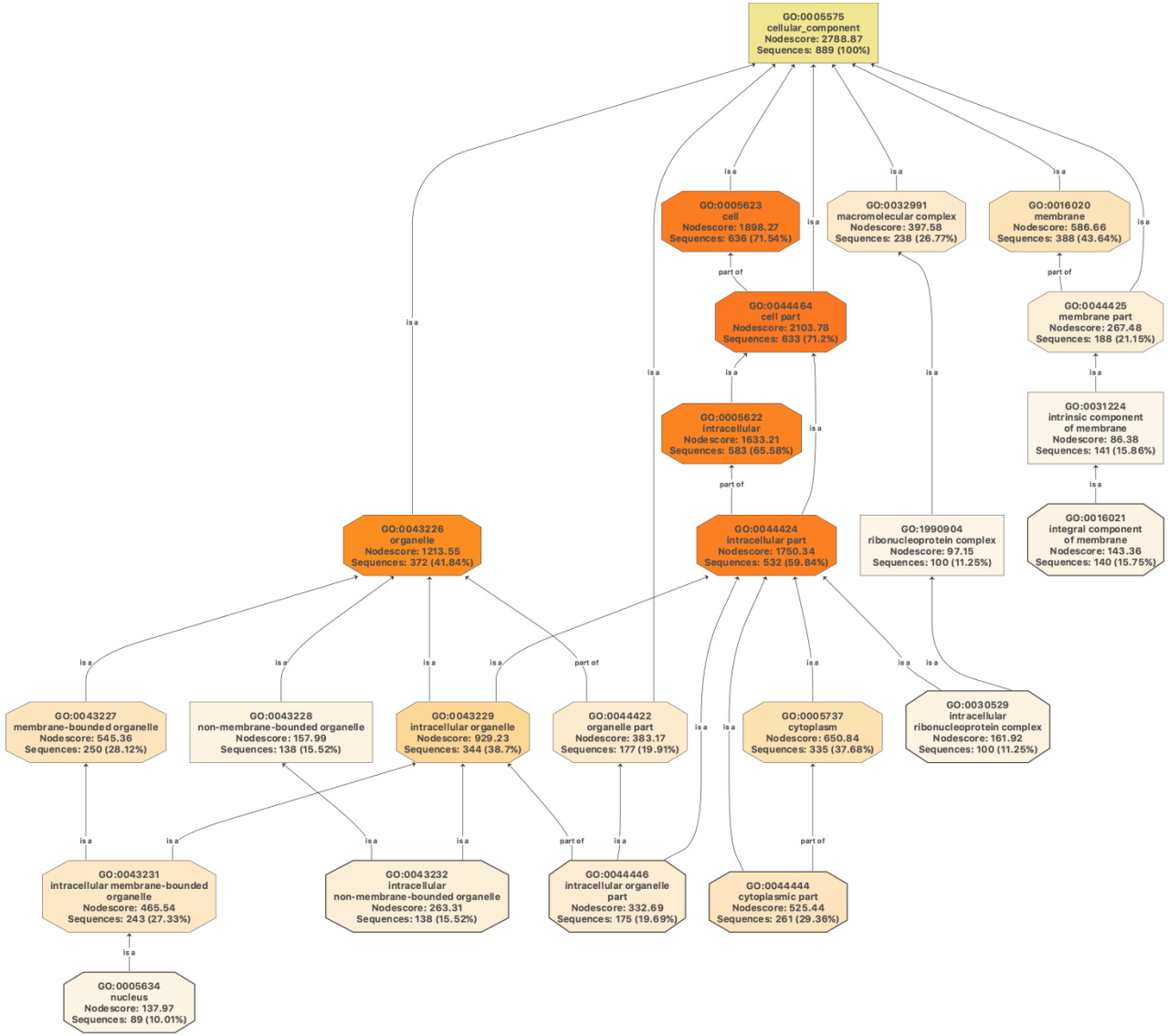


**Figure 4.2: *E. gracilis* Gene Ontology (GO) classifications are distributed across conserved eukaryotic functional categories.** The chart describes the functional categorization of *E. gracilis* gene ontology. Panel A: Biological Processes (BP) showing the weighted GO number, nodescore, number of sequences, associated hierarchy, and respective biological process. Panel B: Molecular Functions (MF) showing the weighted GO number, nodescore, number of sequences, associated hierarchy, and respective molecular function. Panel C: Cellular Components (CC) showing the weighted GO number, nodescore, number of sequences, associated hierarchy, and respective cellular component. Colour codes are indicative of the positional hierarchies. Green (BP), Blue (MF), and Yellow (CC) represent top level or higher hierarchy. Orange (BP, MF, CC) represent middle level hierarchies. Light and dark brown (BP, MF, CC) represent lower level hierarchies.

*4.3.3 InterProScan and Gene Ontology*

The mapping (NCBI NR BLAST) and InterPro analysis of Blast2GO identified 14112 (38.64) sequences with GO terms (Figure 4.1, 4.2 and Table 4.5). The proteins were assigned terms with specific GO term weighted average per protein (Figure 4.2 and Table 4.5). Significant amounts of mapping data (sequences with mapping information) were derived from multiple databases (Figure 4.1). A query with InterProScan increased the number of annotated proteins by ~ 20 % (data not shown). Characterization of cellular components, molecular functions, and biological processes revealed the presence of 40 biological process (BP), 12 molecular functions (MF), and 22 cellular components (CC) (see Figure 4.2 and Table 4.5). Out of the 14112 sequences with InterProScan hits, 6442 (26.98 %) had transmembrane domain while 2551 (6.98 %) had signal peptides. This suggests that ~ 27 % of the predicted proteins are membrane or organelle associated. The reasons for low frequency of signal peptides are unclear. However, since signal petides are present at the N-terminus of proteins destined torwards the secretory pathways (Blobel and Dobberstein, 1975), two factors could explain this low frequency: 1) This may be due to truncation (at the N-terminus) of the potential signal peptides sequences in the transcriptome assembly, thereby reducing the signal peptide candidacy of a sequence. 2) Signal peptides function to prompt cells to translocate synthesized proteins to cell membranes. In eukaryotes, signal peptides directs newly synthesized proteins to sec61 channel (translocons) in the endoplastic recticulum (ER) in a process known as translocation. So, it may mean that there is no active translocation acitivity, no active protein secretion, or perhaps, less protein secretion going on within this system. If this is the case, then this may suggest evidence of the utility of uncovnetional secretion mechanism (Interleukin, Galectin) in *E. gracilis* in a process known as unconventional protein secretion (UPS) (Nickel and Seedorf, 2008, Agrawal, *et al.,* 2010).

*4.3.4 Enzyme Code ad Kegg map*

To delineate information on the possible biological pathways, and to populate these pathways and their involvement in biological processes and functions, the KEGG database (through BLAST2GO) was searched without restriction to prior knowledge of biological pathways specific to *E. gracilis*. About 89 pathway components were

found in the *E. gracilis* transcriptome. This is similar but substantially lower when compared to the pathways found in the kinetoplastid genomes (Berriman, *et al.,* 2005, Ivens, *et al.,* 2005, El-Sayed, *et al.,* 2005), and distribution across metabolism and transport, carbohydrate metabolism, electron transport and oxidative phosphorylation, glycosylphosphatidylinositol anchor biosynthesis, amino acid metabolism, amongst others. A comparison of metabolic pathways encoded by the genomes of *T. brucei*, *T. cruzi*, and *Leishmania major* reveals the least overall metabolic capability in *T. brucei* and the greatest in *L. major* (Berriman, *et al.,* 2005).

### *4.3.5 Orthologous groups clustering and evolution*

Orthologous groups clustering was carried out as previously described in Chapter 2. The distribution of the *E. gracilis* orthogroups across selected eukaryotes suggest the hybrid and pan eukaryotic nature of the Euglena genome. The conserved nature of the eukaryotic core genes is consistent with other members of the eukaryotes. While there are conserved eukaryotic genes as well as genes shared with select eukaryotic taxa, there are also genes unique to *E. gracilis*. It is not yet clear the functions of these genes in Euglena – partly due to insufficient Euglena information on the protein databases. A third cohort represents nuclear transfer of endosymbiotic genes as a result of acquisition of the plastid or mitochondrion (see Figure 4.3 for genes shared specifically with plants/algae). Significantly, I was able to confirm earlier work by Ahmadinejad, *et al.,* 2007, and Maruyama, *et al.,* 2011, that suggests a complex acquisition in terms of gene complement, with contributions from green, red and brown plastids as shown in Figure 4.3.

More than 20 different versions of endosymbiotic theory have been presented in literatures to explain the origin of eukaryotes and their mitochondria (Martin, *et al.,* 2015). Given the complex history of the secondary endosymbiotic event leading to the current configuration in *E. gracilis* (Ahmadinejad, *et al.,* 2007) it is not yet clear which of these surveys of the endosymbiotic theories for the origin of eukaryotes and mitochondria, plastid, and for the origin of eukaryotic nucleus (Martin, *et al.,* 2015) *E. gracilis* subscribes to? To understand the origin of these complexities in *E. gracilis*, whether these were all present in the endosymbiont, or have been picked up over multiple events, will require the incorporation of lateral gene transfer among prokaryotes, endosymbiosis and gene transfer from the organelles to the nucleus

(Martin, *et al.,* 2015). However, individual genes have individual and multiple origins meaning that this may be challenging, thus, there is need to integrate all individual gene trees taking the evolutionary affinities of the plastid (a cyanobacterium), the mitochondrion (a proteobacterium) and the host (an archaeon) into account (Martin, *et al.,* 2015). This is particularly important since genome evolution by endosymbiotic gene transfer is a non tree-like process (Ahmadinejad, *et al.,* 2007).

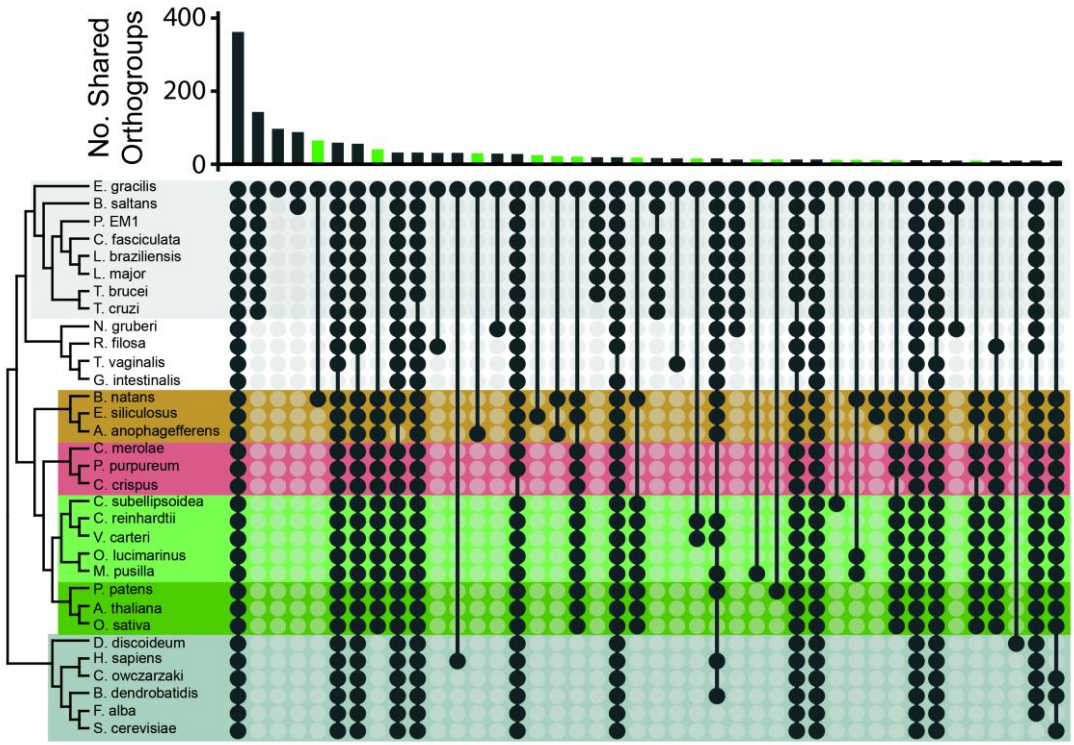



**Figure 4.3: *E. gracilis* posseses genes shared across several eukaryotic taxa.** The diagram depicts the Orthogroup clusters in *E. gracilis*. Orthogroups shared between *E. gracilis* and additional taxa are shown, together with numerical group descriptors for comparison to data. The top diagram (histogram) shows the proportion of shared orthogroups while the bottom diagram (phylogenetic tree and dot plot) shows the presence or absence of these orthogroups in respective eukaryotic taxa. X-axis = proportion of shared orthogroups, y-axis = orthogroup type. Gray = Kinetoplastids. White = Other members of the Excavates. Brown = secondary host. Pink = red algae. Light green = green algae. Darker green = land plants. Light blue = Unikonts. Black circle and connecting lines = presence of corresponding orthogroup. Gray circle without connecting lines = absence of corresponding orthogroup.

### *4.3.6 The Model Organism Database* (MOD)*, EuglenaDB*

To make these information available to the community, I set up the Model Organism Database (MOD), EuglenaDB (Ebenezer, *et al.,* unpublished data; http://euglenadb.org/), which is a web resource for the genome, and transcriptome of *E. gracilis*. Employing the methods as previously described by Hillier, *et al.,* 2005, EuglenaDB was set up, or developed, as part of the *E. gracilis* genome project to integrate the genome and transcriptome assembled sequence with manually expert curated biological information by providing public access to the data. The database is BLASTable, downloadable, and currently contains information about the genome and transcriptome of *E. gracilis.* However, it is expected that a wide range of information will be contained within the Euglena database as the complete *E. gracilis* genome sequence becomes available (which may eventually be donated to Eukaryotic Pathogen (EuPath) for further curation after publishing the genome paper). Presently, EuglenaDB acts as the repository for all preliminary structural genome and transcriptome annotation for *E. gracilis*, and provides a platform to reconcile automatic annotations with community manual curation. EuglenaDB supports four different methods of user access through its graphical web interface, which are also access restricted, and these methods are as previously applied in similar community directed genome projects (Hillier, *et al.,* 2005). Currently, EuglenaDB contains 2066288 DNA Sequences, 36,526 Proteins, *E. gracilis* genome raw assembled sequences, and final transcriptome assembled sequences as BLASTable resource, as well as predicted proteins (amino acids), the General Feature Files (GFF) delineating coordinates and corresponding coding sequences.

## 4.4 DISCUSSION AND CONCLUSION

### *4.4.1 RNA isolation, sequence data quality, assembly evaluation and assessment*

The *E. gracilis* transcriptome was sequenced with multiple illumina libraries, and an assembly of the data into a final/complete assembly. The sequenced data is of good quality. The assembly improvement and evaluation for the transcriptome is ~ 90 % complete in terms of the presence of core eukaryotic genes. When compared to *T. brucei* and *L. major* curated transcripts, the *E. gracilis* transcriptome appears to be at par. The GC content of *E. gracilis* transcriptome is 60.90 % respectively, which is also at par with the GC content of *T. brucei* and *L. major* which are 48.32 % and

62.44 % respectively. This suggests that most Excavates are likely to have an average GC content of ~ 55 %. In terms of the contig metrics assessment, *T. brucei* and *L. major* had larger N50, and > 36 % of their transcripts coding for proteins (data not shown).

*4.4.2 Structural and functional transcriptome annotation*

In silico analysis

Among euglenoids, many studies have focused on the genus *Euglena*, but none of them involved large-scale genomic and transcriptomic sequencing (Jonathan and Badger, 2013). The insufficient genomic and transcriptomic data for this unique organism, might explain why a moderate proportion (20 - 25 %) of the total sequences had no matches to proteins in the database, BLAST and InterproScan respectively. This might define an extensive catalogue of novel or putative protein-coding genes in *E. gracilis*. According to Jonathan and Badger, 2013, many genes are known to encode poorly conserved and short polypeptides, with no significant BLAST hits on the database, and this is likely the case with the present data. Many genes without a BLAST hit tend to also have InterProScan domains (and vice versa), but with no GO terms. It is likely that these are all novel genes that are not yet available on these databases, since there are ~ 39k GO terms.

*4.4.3 Significance of findings – adaptations, functions and applications*

In order to validate the quality and utility of the genome, a transcriptome assembly was produced simultenously alongside the genome assembly. Since the transcriptome contains information for expressed transcript, it provides a great deal of information for validating the assembly – not just in terms of expressed transcripts, but in terms of structure and content – complementary to the genome data. It also provides an indication into genome organization and function and overall biology drawing from transcriptome structure and function which are discussed below.

*Transcriptome sizes correlates with genome sizes:* A positive correlation exists between whole transcriptome size and whole genome size. This means that transcriptome sizes, contents and structures are indicative of genome architectures

(taking expression/transcripts profiling into account). So, in future genome projects (intra- and inter- species), it will be relevant to carry out an initial transcriptome studies in order to understand the magnitude, depth, and forecast of genome architectures. This is useful to reduce time and cost. For instance, while the estimated genome size for *E. gracilis* is 2 Gbp, the genome size for *T. brucei* is approximately 35 Mbp - this is 57 times less than the *E. gracilis* genome. In the present genome studies, the transcriptome size is larger than the transcriptome size of other members of the Kinetoplastids (Table 4.1), being more than twice the size of *T. brucei* transcripts. This further provides insights into the chromosome architecture of *E. gracilis* in comparison to kinetoplastids, suggesting that majority of the *E. gracilis* genome are either repetitive or contains more introns/intergenic regions than expressed genes. Increase in transcriptome size or output may also be correlated with the capacity to deal with inherently complex biological processes. This suggests that the *E. gracilis* transcriptome data (in the present studies) may be mined for future comparative study, and as an annotation tool for a complete *E. gracilis* whole genome sequencing project (Jensen, *et al.,* 2015).

*Mean sequence length is not a direct function of transcriptome size:* Comparatively, across species, the *E. gracilis* mean sequence length is much shorter than those of other members of the kinetoplastids (Table 4.3 and 4.4). Since the size of *E. gracilis* transcriptome is larger than the other members of the kinetoplastids (Table 4.3 and 4.4), I anticipate that the mean transcript/gene length will be larger (Hou and Lin, 2009) - however, this is not the case. Two possible explanations may be available for this: Truncations arising from assembly processes or non transcript length bias inherent in RNA-seq data (Oshlack and Wakefield, 2009). To provide evidence for truncations, about a third of the transcriptome sequences are missing an N-terminal, C-terminal, or both. Similarly, a recent study showed that a technical feature of using high-throughput sequencing to interrogate full length transcripts is that longer transcripts produce more reads relative to short transcripts of similar expression (Oshlack and Wakefield, 2009). This higher sampling means that there is more statistical power to detect differential expression for long transcripts compared to short ones. Since the later may now be ruled out, it is evident that the short mean transcript length compared with kinetoplastids may be due to truncations and may not suggest any biological function.

*Nucleotide and protein statistics are pointers to gene functions and evolutionary charactristics:* Nucleotide and protein statistics such as GC content, gene density, average gene length and protein length play significant role in delineating biological functions in eukaryotes and prokaryotes. For instance, GC content is correlated with various genomic features, including repeat element distribution, methylation pattern, and, most remarkably, gene density (Galtier, *et al.,* 2001). GC-rich regions include many genes with short introns while GC-poor regions are essentially deserts of genes (Galtier, *et al.,* 2001). In higher eukaryotes, such as mammals, the distribution of GC content could have some protein-coding functional relevance, raising the issue of origin and evolution (Galtier, *et al.,* 2001). In Bacteria and Archaea, several environmental factors (e.g. oxygen, nitrogen) and mutational bias (GC-biased gene conversion) potentially affecting genomic GC-content have been proposed (Lassalle, *et al.,* 2015). For instance, it has been proposed that GC content plays an adaptive role to temperatures. *E. gracilis* possesses an overall GC content of 60.87 % (Table 4.4), suggesting a strong coding potential for the transcriptome and one of the explanations for it's adaption to extreme conditions. Similarly, comparative mean ORF and GC percentage may be an evolutionary characteristics within the Euglenozoa. For instance, similar mean ORF and GC percentage (Table 4.1 and 4.4) exists among *E. gracilis*, *T. brucei* and *L. major.* This is similar to the evidence in virtually all bacteria where two independent analyses have shown that, independently of their genomic GC-content, there is an excess of G/C→A/T mutations (Lassalle, *et al.,* 2015).

*Top species BLAST hits moderately correlate with orthologous groups clustering analysis:* In the orthologous groups clustering analysis (see Chapter 5 and Appendix I-a), the highest number of shared orthogroups are conserved across eukaryotes, the second largest are shared with the Kinetoplastids while the third largest are shared with *Bodo saltans*. In the top species distribution (Table 4.5), *B. saltans* shows the highest distribution, and twice the value recorded for the next top hits (*Guillardia theta, Naegleria gruberi, Chrysochromulina sp., and Euglena gracilis*). In *E. gracilis* and *B. saltans*, these orthology and homology may suggest some evolutionary relationship in the genes playing some specific roles in adaptive responses to freeliving lifestyles.

# CHAPTER FIVE

## 5.0 COMMUNITY-BASED TRANSCRIPTOME ANNOTATION

### 5.1 INTRODUCTION

This chapter discusses the community-based annotation of the *E. gracilis* transcriptome taking into account the various aspects of its biology such as cellular processes and signaling, information storage and processing, metabolism, and evolution (see paragraphs below). The community-based annotation was carried out in collaboration with 13 researchers. The researchers and their respective analysis are listed therein: ThankGod Ebenezer (meiosis, protein trafficking, bilobes, nuclear cohorts, tubulins, transporters, histones, rhodopsin, translational apparatus, pre-initiation complex, kinetochores, ribosome, mRNA metabolism, exosomes, spliceosomes and related proteins, TAC/CAP proteins, RNAi pathway, GalF, PPG, heterochromatin, and orthologous groups clustering), Andrew Jackson (surfaceome), Vladimir Hampl (plastid proteome), Julius Lukes (mitochondrial genes), Sam Obado and Carlos Santana and Mark Field (nuclear cohorts), Michael Lebert (signal transduction), Mukund Thattai (dynamins), Ellis O'Neil (glycosylation machinery, carbohydrate active enzymes), Michael Ginger (glycosomes, metabolic control), Steve Kelly and Carlos Santana (orthologous groups clustering), Joel Dacks (protein trafficking), Steve Kelly (Horizontal Gene Transfer) and Sue Vaughan (heterochromatin electron microscopy). The biological aspects that I specifically investigated in this work were performed under the supervision of Prof. Mark Field and Professor Mark Carrington (see above). The other biological aspects listed above which I did not specifically investigate were coordinated by me with supervision from Prof. Mark Field. In this Chapter, I will discuss the community based annotation that I carried out (or contributed directly in the analysis) with supervision from Prof. Mark Field and Prof. Mark Carrington, and except where otherwise stated, standard analytical approach involved the generation of an eukaryotic reference set, homology and orthology searches using BLAST and OrthoFinder, and quality functional delineation and control using protein and/or conserved domain databases such as InterPro, CDD, and Uniprot. I have also dedicated a sub section (see sub section 5.6) where I will discuss briefly the analysis

I only coordinated but was not involved in the direct analysis or investigation. For access to sequence data files, visit: http://euglenadb.org/ using the following log-in details - Username: guest; Password: frid@y80. For an extensive discussion of these analysis (including additional statistical parameters employed), see the pre-print version of the paper, Ebenezer, *et al.,* 2018, doi: https://doi.org/10.1101/228015 and Appendix-Ia).

## 5.2 CELLULAR PROCESSES AND SIGNALING

All eukaryotic organisms contain a nucleus from whence most cellular processes are directed. These cellular processes are essential as well as form fundamental system that involves multiple biochemical reactions and signaling pathways, and include, protein trafficking, meiosis, bilobes, nuclear cohorts, tubulins, transporters, histones, and calmodulins and rhodopsins. In this section, and over the next paragraphs, these cellular processes and signaling pathways in *E. gracilis* will be discussed.

### *5.2.1 Protein trafficking*

Proteins associated with the endomembrane systems were investigated to understand the function of the organelles (e.g. Golgi apparatus, endoplasmic recticulum, etc), compartmentalization and transport/trafficking system in *E. gracilis*. More importantly, it provides specific insight into the nuclear envelope with demonstrated understanding in the nuclear matrix. To delineate the role of the Ras superfamily, Adaptor complexes, Coatomer complexes, TSETS, TSPOON, SNARE, and SNAP superfamilies in protein trafficking in my model organism, the *E. gracilis* transcriptome was analyzed using homology, orthology-based searches and phylogenetic inference (Dacks and Field, 2007, Emms and Kelly, 2015; see Chapter 2) to achieve this purpose. A combined BLAST homology search and InterPro domain/Conserved Domain scan was performed to preliminarily identify the homologs in the transcriptome assembly. Protein families investigated include those as highlighted in section 5.1 above, and similar methodology and result analysis output exists for these protein families. With respect to the trafficking proteins, specific orthologs were not able to be identified in some cases.

An eukaryotic reference dataset of manually curated protein trafficking genes (e.g. small GTPases, adaptor complexes and coats; Elias, *et al.,* 2012; Rojas, *et al.,* 2012,

Abbasi, *et al.,* 2011, Manna, *et al.,* 2013), Perez, *et al.,* 2014 were assembled and used to interrogate the transcriptome assembly with functional phylogenetic inference carried out using maximum likelihood (PhyML, RaxML, FastTree) and posterior probability approaches (MrBayes), as previously reported. Query taxa distribution was widely spread across selected eukaryotic kingdoms (*T. brucei*, *L. major*, *T. cruzi*, *N. gruberi, H. sapiens, D. melanogaster, and S. cerevisiae*), except for the Rab small GTPases where a narrow distribution specific to the trypanosomatids (*T. brucei* and *T. cruzi*) was utilized since these reference sets have experimentally been tested (Figure 5.1, Table 5.1; Elias, *et al.,* 2012) as well as being closely related to *E. gracilis*. *E. gracilis* possesses a near complete set of membrane trafficking machinery as that of the query organisms, but there are instances where an entire complex is missing such as in APs, TSET, and COG (Figure 5.1, Table 5.1). This may represent lineage-specific losses, or failure of our methods to detect the protein sequence due to divergence in sequence or length of sequence.

Phylogenetic inference show statistical supports ranging from moderate to strong values (Figure 5.2 – 5.6). Analyses of the small GTPases provided some resolutions into 16 Rab subfamilies which are also present in trypanosomatids. The Rab subfamilies in *E. gracilis* constructed possesses Rabs 1, 2, 18, 14, 4, 11, 5, 21, 23, 6, 32, 28, 7, X1, X2, and X3 (Figure 5.1, 5.2). This is between 48 - 58 % (19 – 23) of the minimum distinct Rab and Rab-related proteins proposed by Elias, *et al.,* 2012 to be represented in the Last Eukaryotic Common Ancestor (LECA). Rab 11 and 32 which are conserved across all eukaryotes were also found in the *E. gracilis* transcriptome (Figure 5.1, 5.2). Two unclassified Rabs were found in *E. gracilis* which are not present in trypanosomatids (Figure 5.1, 5.2). Interrogation of these two Rab sequences against a database of Rabs (Rabifier) and NCBI nr database suggested that it is a different kind of Rab X. Rab clade was defined as containing sequences from at least one or two supergroups (Bright, *et al.,* 2010; Pereira-Leal, 2008; Pereira-Leal and Seabra, 2001), supported by at least an average of 0.50 posterior probabilities (PP) and/or 50 % bootstrap (BP) support by the Maximum Likelihood method, and possessing a GTP binding domain.

Within the ARFs protein family, only two ARFs (Arf1, and Arf 7&8) were found (Figure 5.1). Nine (9) Arf-like (Arl5, Arl1, Arl7&8, Arl2, Arl3, Arl6, Arl8, Arl X1 & X2) subfamilies were found with one Sar1 subfamily. Of all the ARFs found in *E. gracilis*, only Arf1, Arl1, Arl2, Arl3, and Sar1 are present in kinetoplastids. No Rho protein was found in *E. gracilis*; this is consistent with other kinetoplastids where a Rho protein is missing, except for the mitochondria Rho (Miro). Only one Ran protein was found in *E. gracilis* which is also consistent with other eukaryotes. The Ran present shows similarity to kinetoplastids than to any other eukaryotes.

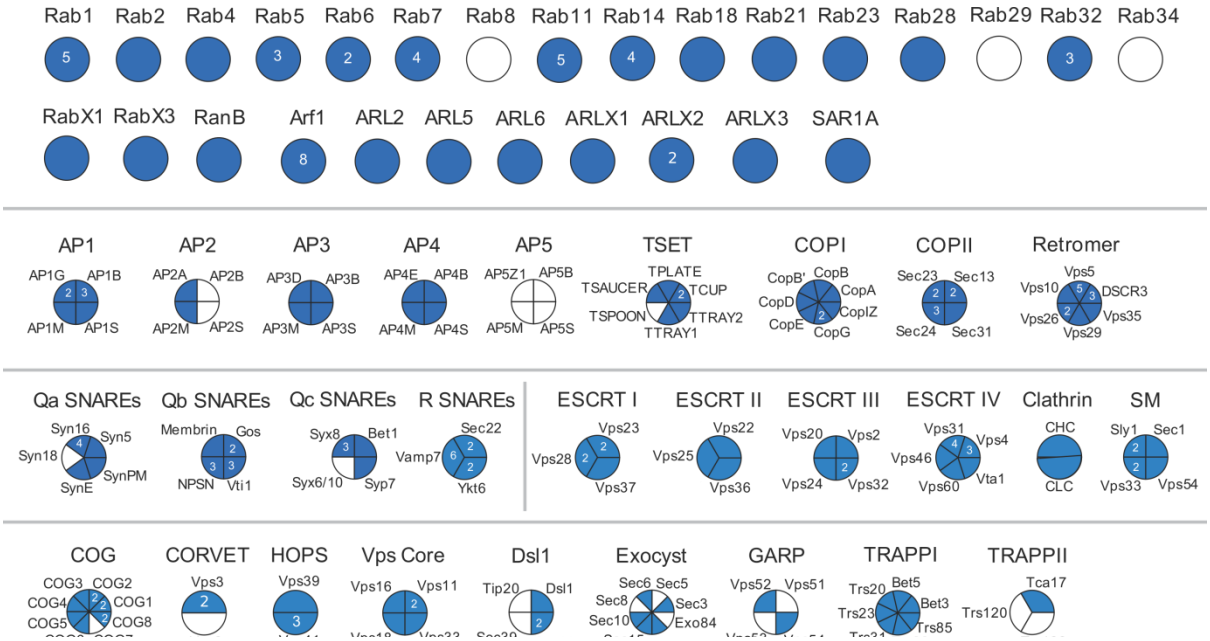



**Figure 5.1: The endomembrane system of *Euglena gracilis* is complex.** Blue circles represent the presence of the protein while white represents proteins that were not retrieved by search methods. Paralog numbers are shown in white within circles for, from the top Rab, Arf, ARL, and Sar proteins families, adaptin and coatomer protein complexes, SNARE proteins, endosomal protein complexes and multi-subunit tethering complexes.

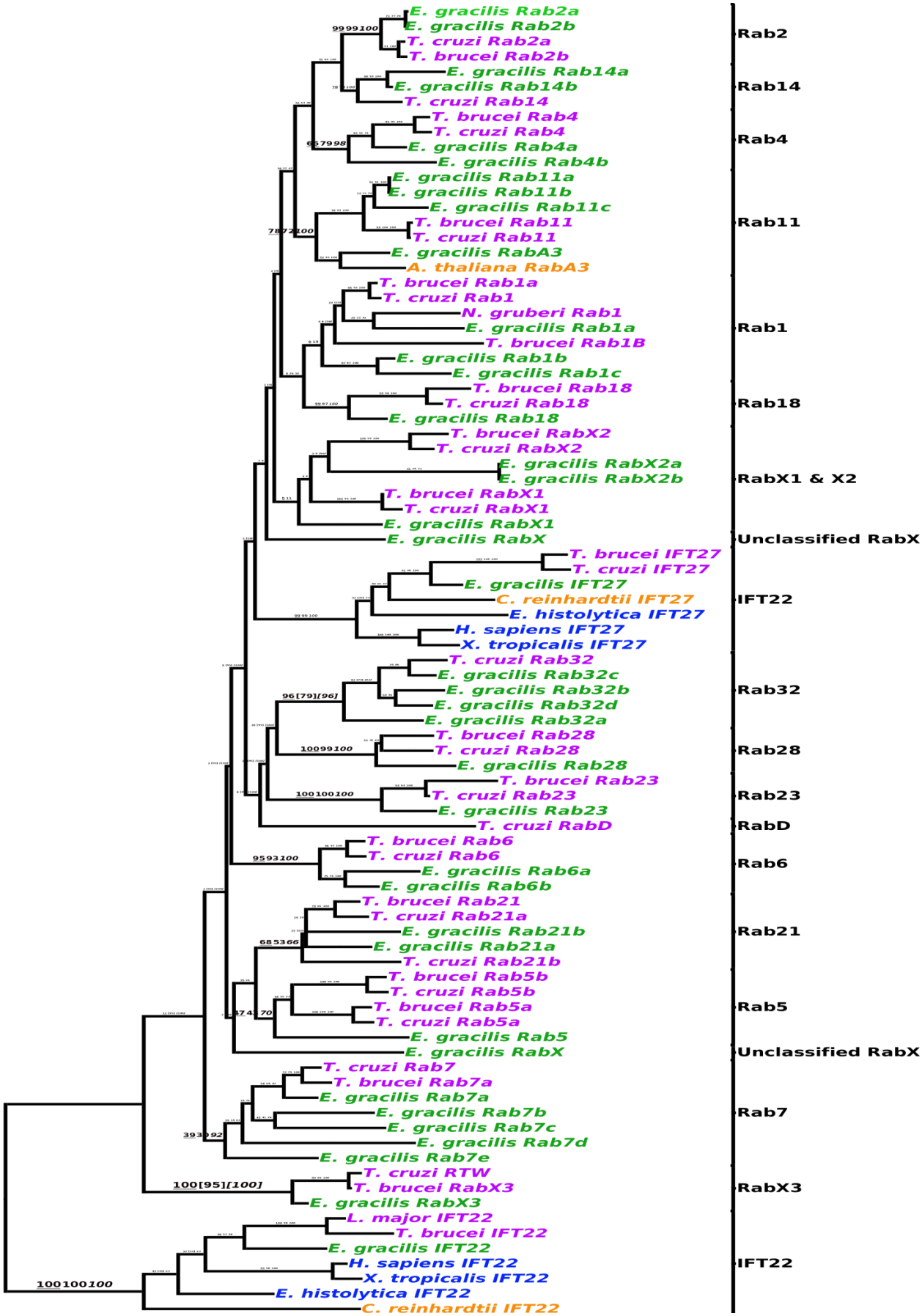

E. gracilis Rab2a
E. gracilis Rab2b
T. cruzi Rab2a
T. brucei Rab2b
E. gracilis Rab14a
E. gracilis Rab14b
T. cruzi Rab14
T. brucei Rab4
T. cruzi Rab4
E. gracilis Rab4a
E. gracilis Rab4b
E. gracilis Rab11a
E. gracilis Rab11b
E. gracilis Rab11c
T. brucei Rab11
T. cruzi Rab11
E. gracilis RabA3
A. thaliana RabA3
T. brucei Rab1a
T. cruzi Rab1
N. gruberi Rab1
E. gracilis Rab1a
T. brucei Rab1B
E. gracilis Rab1b
E. gracilis Rab1c
T. brucei Rab18
T. cruzi Rab18
E. gracilis Rab18
T. brucei RabX2
T. cruzi RabX2
E. gracilis RabX2a
E. gracilis RabX2b
T. brucei RabX1
T. cruzi RabX1
E. gracilis RabX1
E. gracilis RabX
T. brucei IFT27
T. cruzi IFT27
E. gracilis IFT27
C. reinhardtii IFT27
E. histolytica IFT27
H. sapiens IFT27
X. tropicalis IFT27
T. cruzi Rab32
E. gracilis Rab32c
E. gracilis Rab32b
E. gracilis Rab32d
E. gracilis Rab32a
T. brucei Rab28
T. cruzi Rab28
E. gracilis Rab28
T. brucei Rab23
T. cruzi Rab23
E. gracilis Rab23
T. cruzi RabD
T. brucei Rab6
T. cruzi Rab6
E. gracilis Rab6a
E. gracilis Rab6b
T. brucei Rab21
T. cruzi Rab21a
E. gracilis Rab21b
E. gracilis Rab21a
T. cruzi Rab21b
T. brucei Rab5b
T. cruzi Rab5b
T. brucei Rab5a
T. cruzi Rab5a
E. gracilis Rab5
E. gracilis RabX
T. cruzi Rab7
T. brucei Rab7a
E. gracilis Rab7a
E. gracilis Rab7b
E. gracilis Rab7c
E. gracilis Rab7d
E. gracilis Rab7e
T. cruzi RTW
T. brucei RabX3
E. gracilis RabX3
L. major IFT22
T. brucei IFT22
E. gracilis IFT22
H. sapiens IFT22
X. tropicalis IFT22
E. histolytica IFT22
C. reinhardtii IFT22
Rab2
Rab14
Rab4
Rab11
Rab1
Rab18
RabX1 & X2
Unclassified RabX
IFT22
Rab32
Rab28
Rab23
RabD
Rab6
Rab21
Rab5
Unclassified RabX
Rab7
RabX3
IFT22

**Figure 5.2: *E. gracilis* possess proportionate number of trypanosome conserved Rab protein families.** The figure describes the *E. gracilis* Rab proteins identified utilizing mainly trypanosomes as reference sets. Statistical supports are in the order PhyML, RaxML, and MrBayes (in %), and visible for major nodes**.** [#] = conflicting topology; where # are values. IFT22 was used as the outgroup. Where available, taxa color codes is as described: green = *E. gracilis*; purple = Excavates; orange = Plants/Algae; blue = other eukaryotes. Colours do not have specific connotation except for green which shows *E. gracilis* is a photosynthetic green organism. Other Chlorophytes/Plants are in orange colour to differentiate from the species of interest – *E. gracilis.*

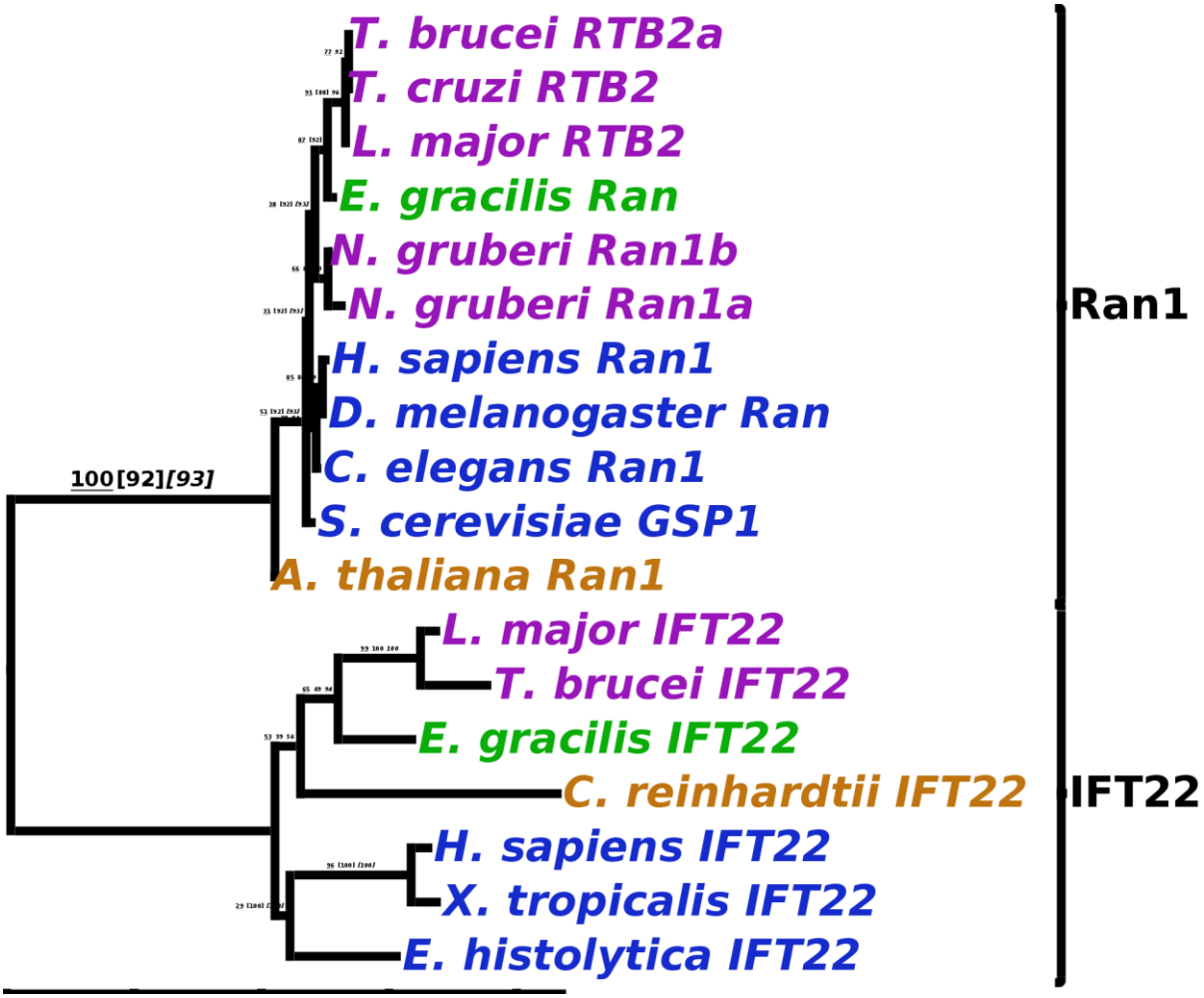



**Figure 5.3: *E. gracilis* possess a Ran protein with similarities to trypanosomes**. The figure describes the *E. gracilis* Ran proteins identified utilizing mainly eukaryotic reference sets. Statistical supports are in the order PhyML, RaxML, and MrBayes (in %), and visible for major nodes**.** [#] = conflicting topology; where # are values. IFT22 was used as the outgroup. Where available, taxa color codes is as described: green = *E. gracilis*; purple = Excavates; orange = Plants/Algae; blue = other eukaryotes. Colours do not have specific connotation except for green which shows *E. gracilis* is a photosynthetic green organism. Other Chlorophytes/Plants are in orange colour to differentiate from the species of interest – *E. gracilis.*

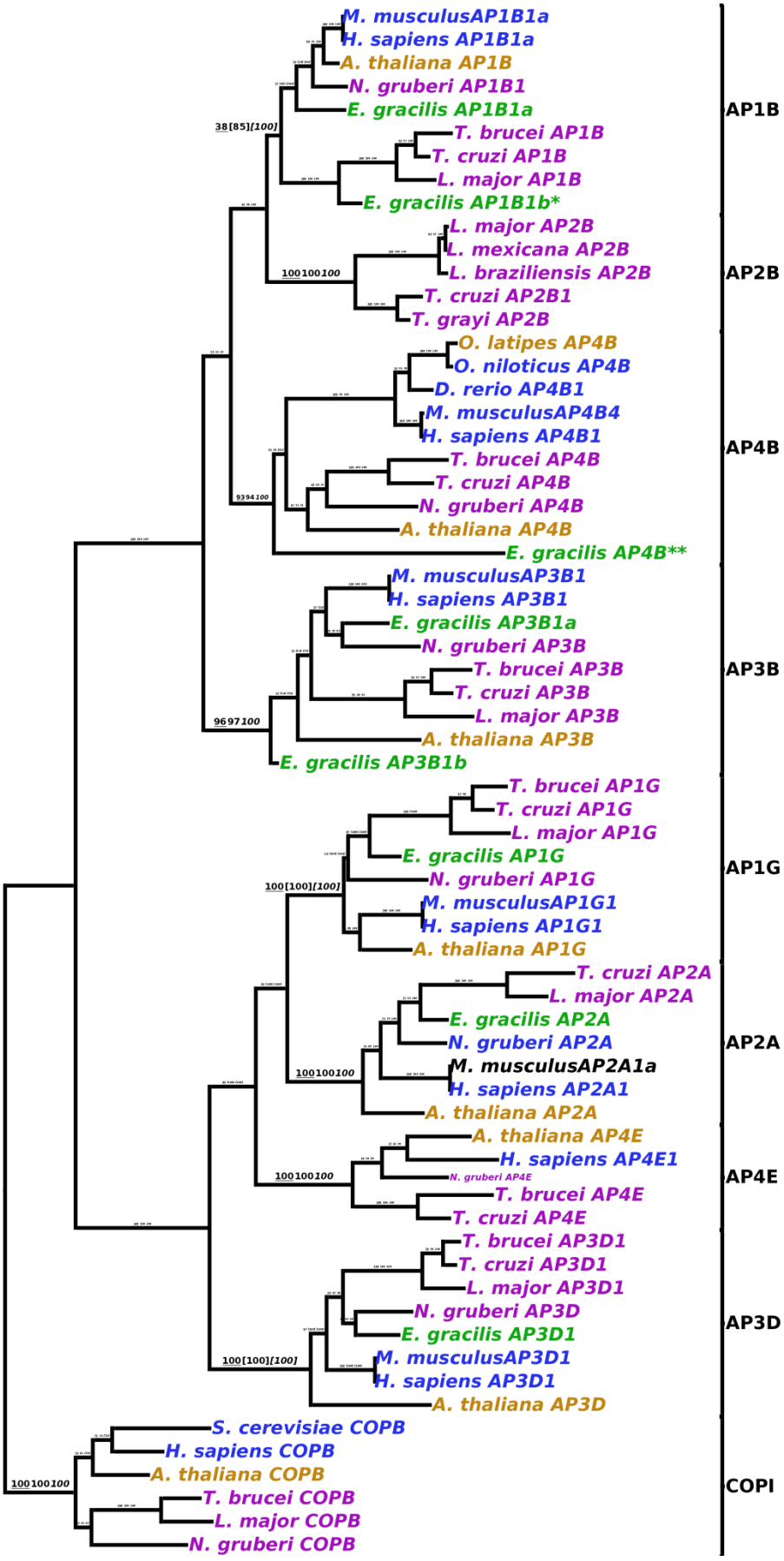

M. musculusAP1B1a
H. sapiens AP1B1a
A. thaliana AP1B
N. gruberi AP1B1
E. gracilis AP1B1a
38[85][100]
T. brucei AP1B
T. cruzi AP1B
L. major AP1B
E. gracilis AP1B1b*
L. major AP2B
L. mexicana AP2B
L. braziliensis AP2B
100100100
T. cruzi AP2B1
T. grayi AP2B
O. latipes AP4B
O. niloticus AP4B
D. rerio AP4B1
M. musculusAP4B4
H. sapiens AP4B1
T. brucei AP4B
T. cruzi AP4B
N. gruberi AP4B
9394100
A. thaliana AP4B
E. gracilis AP4B**
M. musculusAP3B1
H. sapiens AP3B1
E. gracilis AP3B1a
N. gruberi AP3B
T. brucei AP3B
T. cruzi AP3B
L. major AP3B
9697100
A. thaliana AP3B
E. gracilis AP3B1b
T. brucei AP1G
T. cruzi AP1G
L. major AP1G
E. gracilis AP1G
N. gruberi AP1G
100[100][100]
M. musculusAP1G1
H. sapiens AP1G1
A. thaliana AP1G
T. cruzi AP2A
L. major AP2A
E. gracilis AP2A
N. gruberi AP2A
100100100
M. musculusAP2A1a
H. sapiens AP2A1
A. thaliana AP2A
A. thaliana AP4E
H. sapiens AP4E1
100100100
N. gruberi AP4E
T. brucei AP4E
T. cruzi AP4E
T. brucei AP3D1
T. cruzi AP3D1
L. major AP3D1
N. gruberi AP3D
E. gracilis AP3D1
100[100][100]
M. musculusAP3D1
H. sapiens AP3D1
A. thaliana AP3D
S. cerevisiae COPB
H. sapiens COPB
A. thaliana COPB
100100100
T. brucei COPB
L. major COPB
N. gruberi COPB
AP1B
AP2B
AP4B
AP3B
AP1G
AP2A
AP4E
AP3D
COPI

**Figure 5.4: *E. gracilis* large subunit adaptor complex possess a hybrid origin.** The figure describes the *E. gracilis* large adaptor proteins identified utilizing select eukaryotes as reference sets, showing relatedness of some adaptor large subunits to either plant and/or animal. Statistical supports are in the order PhyML, RaxML, and MrBayes (in %), and visible for major nodes**.** [#] = conflicting topology; where # are values. COPI was used as the outgroup. Where available, taxa color codes is as described: green = *E. gracilis*; purple = Excavates; orange = Plants/Algae; blue = other eukaryotes. Colours do not have specific connotation except for green which shows *E. gracilis* is a photosynthetic green organism. Other Chlorophytes/Plants are in orange colour to differentiate from the species of interest – *E. gracilis.*

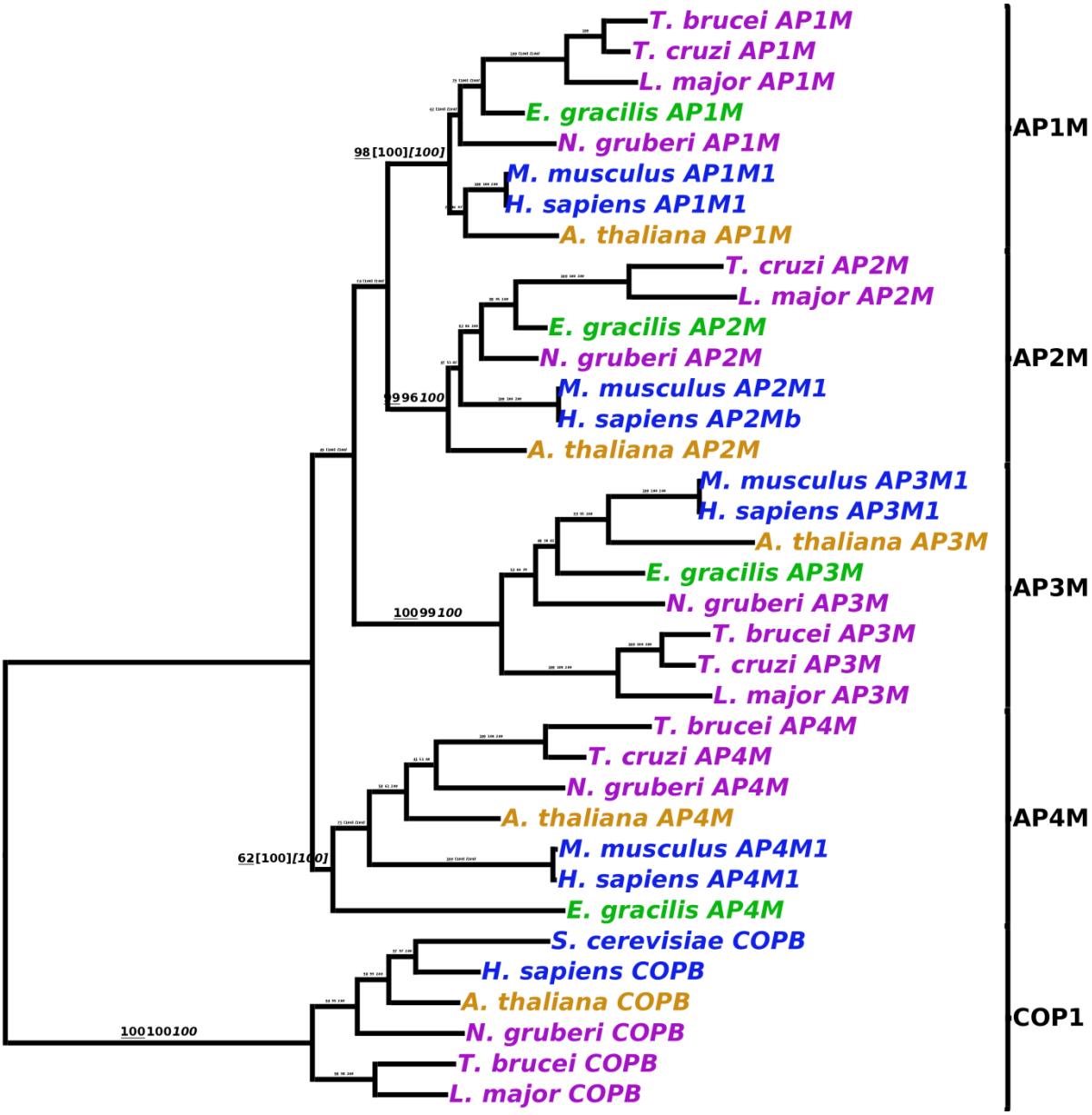



**Figure 5.5: *E. gracilis* medium subunit adaptor complex possess a hybrid origin.** The figure describes the *E. gracilis* medium adaptor proteins identified utilizing select eukaryotes as reference sets, showing relatedness of some adaptor medium subunits to either plant and/or animal. Statistical supports are in the order PhyML, RaxML, and MrBayes (in %), and visible for major nodes**.** [#] = conflicting topology; where # are values. COPI was used as the outgroup. Where available, taxa color codes is as described: green = *E. gracilis*; purple = Excavates; orange = Plants/Algae; blue = other eukaryotes. Colours do not have specific connotation except for green which shows *E. gracilis* is a photosynthetic green organism. Other Chlorophytes/Plants are in orange colour to differentiate from the species of interest – *E. gracilis.*

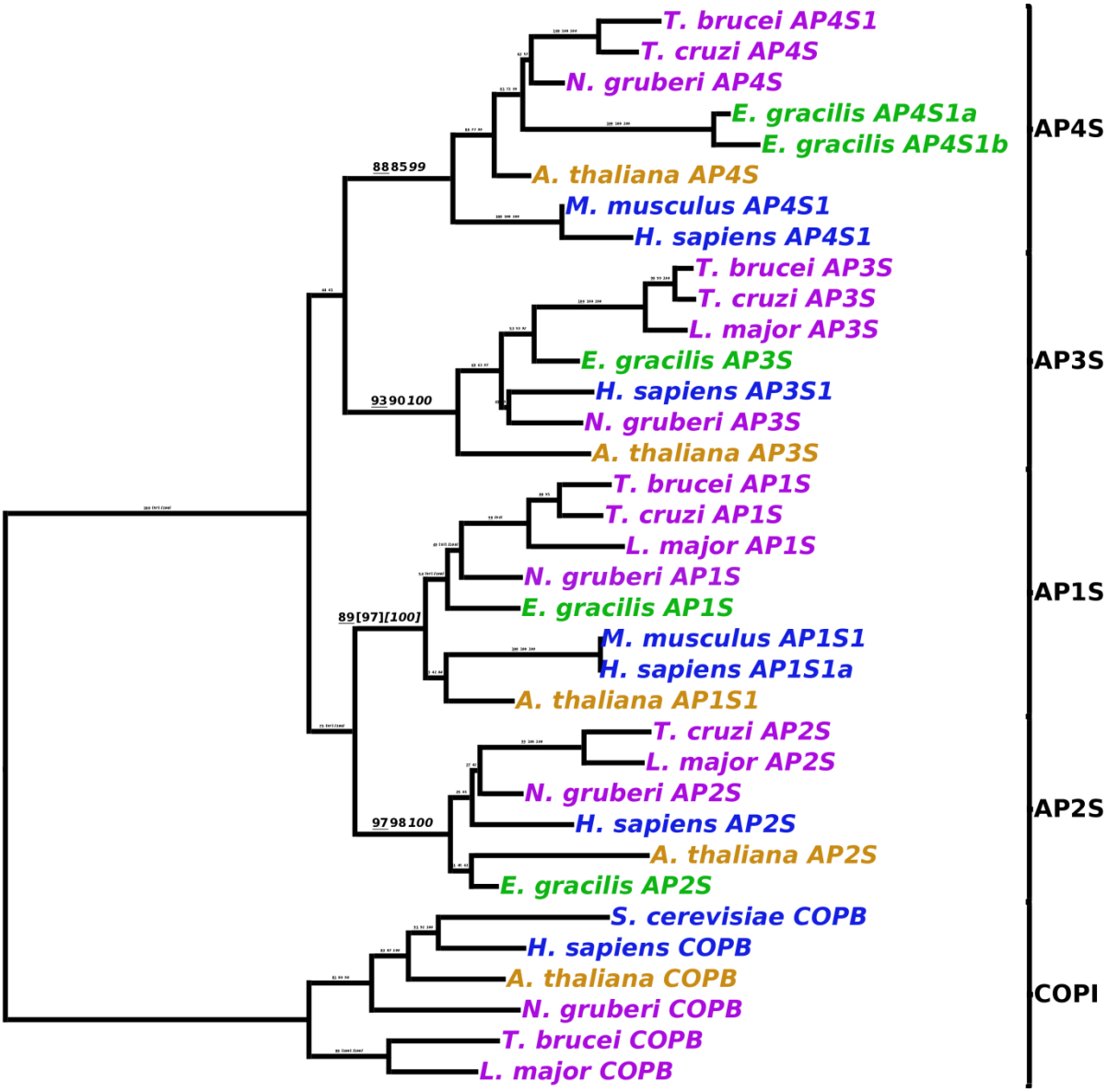



**Figure 5.6: *E. gracilis* small subunit adaptor complex possess a hybrid origin.** The figure describes the *E. gracilis* small adaptor proteins identified utilizing select eukaryotes as reference sets, showing relatedness of some adaptor small subunits to either plant and/or animal. Statistical supports are in the order PhyML, RaxML, and MrBayes (in %), and visible for major nodes**.** [#] = conflicting topology; where # are values. COPI was used as the outgroup. Where available, taxa color codes is as described: green = *E. gracilis*; purple = Excavates; orange = Plants/Algae; blue = other eukaryotes. Colours do not have specific connotation except for green which shows *E. gracilis* is a photosynthetic green organism. Other Chlorophytes/Plants are in orange colour to differentiate from the species of interest – *E. gracilis.*

**Table 5.1:** The endomembrane system of *E. gracilis*

| Reference sequence ID | Paralogous sequence ID | Final classification |
|---|---|---|
| EG_transcript_22226 | EG_transcript_23097, EG_transcript_24244, EG_transcript_39897, EG_transcript_40714, EG_transcript_51388, EG_transcript_68851, EG_transcript_28124 | Rab1 |
| EG_transcript_20574 | | Rab2 |
| EG_transcript_13352 | | Rab4 |
| EG_transcript_19330 | EG_transcript_23555 | Rab5 |
| EG_transcript_16137 | EG_transcript_19234, EG_transcript_27572, EG_transcript_33556, EG_transcript_23166 | Rab6 |
| EG_transcript_19917 | EG_transcript_28802, EG_transcript_33428, EG_transcript_40039, EG_transcript_30017 | Rab7 |
| EG_transcript_12480 | EG_transcript_12776, EG_transcript_22338, EG_transcript_23728, EG_transcript_25212, EG_transcript_58838, EG_transcript_23963 | Rab11 |
| EG_transcript_20767 | EG_transcript_21965, EG_transcript_21965, EG_transcript_23506 | Rab14 |

| | | |
|---|---|---|
| EG_transcript_25311 | | Rab18 |
| EG_transcript_29860 | | Rab21 |
| EG_transcript_19532 | | Rab23 |
| EG_transcript_19476 | | Rab28 |
| EG_transcript_21240 | EG_transcript_23963, EG_transcript_29393 | Rab32 |
| EG_transcript_30796 | | RabX1 |
| EG_transcript_7088 | EG_transcript_20384, EG_transcript_26047, EG_transcript_26290, EG_transcript_31729, EG_transcript_33105, EG_transcript_40926 | Arf1 |
| EG_transcript_10906 | EG_transcript_17530 | Arf4 |
| EG_transcript_45648 | | Arf8 |
| EG_transcript_1688 | | ARLX1 |
| EG_transcript_26745 | | ARL2 |
| EG_transcript_18062 | | ARL6 |
| EG_transcript_22239 | | Sar1 |
| EG_transcript_8695 | EG_transcript_14956, EG_transcript_10202, EG_transcript_63537 | AP1B |
| EG_transcript_5122 | | AP1G |
| EG_transcript_8953 | | AP1M |
| EG_transcript_35562 | | AP1S |
| EG_transcript_1519 | EG_transcript_52977 | AP2A |
| EG_transcript_10768 | EG_transcript_62169 | AP2M |
| EG_transcript_3438 | EG_transcript_51116 | AP3B |
| EG_transcript_1077 | | AP3D |
| EG_transcript_10486 | | AP3M |

| | | |
|---|---|---|
| EG_transcript_36248 | | AP3S |
| EG_transcript_35588 | EG_transcript_46098 | AP4B |
| EG_transcript_10404 | | AP4M |
| EG_transcript_22484 | | AP4S |
| EG_transcript_12790 | EG_transcript_24023 | TCUP |
| EG_transcript_1367 | | TPLATE1 |
| EG_transcript_5169 | | TSAUCER1 |
| EG_transcript_517 | | TTRAY1 |
| EG_transcript_4777 | | TTRAY2 |
| EG_transcript_12005 | | Syn5 |
| EG_transcript_13070 | EG_transcript_22398 | Syn16.1B |
| EG_transcript_11115 | EG_transcript_29436 | Syn16.2B |
| EG_transcript_17067 | EG_transcript_41040 | SynPM.1 |
| EG_transcript_10123 | EG_transcript_25281, EG_transcript_25867 | NPSN1 |
| EG_transcript_22629 | EG_transcript_23411, EG_transcript_29398 | Vti1.1 |
| EG_transcript_18194 | EG_transcript_19152 | Gos |
| EG_transcript_60540 | | Membrin |
| EG_transcript_63709 | | Bet1 |
| EG_transcript_13273 | | Syp7 |
| EG_transcript_19260 | EG_transcript_23261, EG_transcript_20573 | Syx8 |
| EG_transcript_23133 | EG_transcript_37021 | Sec22 |
| EG_transcript_36195 | EG_transcript_21972, EG_transcript_15955, EG_transcript_17762, EG_transcript_22047, EG_transcript_22965 | VAMP7 |
| EG_transcript_47648 | EG_transcript_15564, | Ykt6 |

| | | |
|---|---|---|
| | EG_transcript_22181, EG_transcript_72155 | |

**Note:** The table shows the trafficking proteins inventoried in *E. gracilis*. Column 1 = Reference sequence ID. Column 2 = Paralogous sequence ID. Column 3 = Final classifications. Green coloured cell and corresponding text means the presence of the corresponding protein and the protein name.

Discerning the membrane trafficking proteins possessed by *E. gracilis* provides insight into the evolutionary loss of certain protein families within the Excavate supergroup. Phylogenetic inference yielded well conserved adaptin subunit homologues in the transcriptome assembly, and using wide sampling dataset representing a broad range of eukaryotic diversity as described by Dacks, *et al.,* 2008 and Manna, *et al.,* 2013, all members of the Adaptor complexes (large-, medium-, and small- subunits) were found except for AP1B2, AP4E, and AP5 (Figure 5.1, 5.4, 5.5, and 5.6). The analysis of the adaptin subunits produced well supported clades, which in all cases corresponded to four AP complexes I - IV subunits totaling 16 AP subunits in *E. gracilis* (see Fugure 5.1, 5.4, 5.5, and 5.6). There are five separate complexes, each composed of four subunits (small- Σ, medium- μ, and the large δ-, α-, β-and ɣ-subunits) (Dacks, *et al.,* 2007). Proportion of AP complexes identified in *E. gracilis* include seven large subunits (AP1B1, AP2B1, AP3B1/2, AP4B, AP3D1, AP1G1, AP2A1/2), three medium subunits (AP2M, AP1M, and AP3M), and five small subunits (APS1, AP4S2, AP3S, AP2S, and AP1S) (Figure 5.1, 5.4, 5.5, and 5.6). Complexes which were not found in the present assembly include AP2B and AP5 (Figure 5.1, 5.4, 5.5, and 5.6). The presence of AP I - IV complexes in the present assembly, suggests that overall, all the AP complexes appear to have been present very early, and likely in the LECA (Field, *et al.,* 2007; Nevin and Dacks, 2009). In the resulting dataset, the AP complexes each formed separate clades with PhyML, RaxML, and/or MrBayes bootstrap and posterior probability statistical support values of > 50 %, and distributed amongst the various subunits of small- Σ, medium- μ, and the large δ-, α-, β-and ɣ-subunits. The AP5 complex is lost in Kinetoplastid organisms such as *T.brucei* and *Leishmania major (*Hirst, *et al.,* 2011). However, some subunits (μ and β) are still expressed in the Heterolobosea; *N. gruberi*. The loss of all AP5 subunits in *E. gracilis,* suggests that the loss of the AP5 μ and β subunits occurred in the common ancestor of the Euglenozoa, following their divergence from their common ancestor with the Heterolobosea.

Similar to the distribution of AP5, TSET is present in the Heterolobosea but lost in the Kinetoplastids (Hirst, *et al.,* 2014). However, *E. gracilis* contains the complete set of TSET subunits, except for TSPOON. This infers that the loss of TSET occurred

after divergence of the euglenids, in the common ancestor of the Kinetoplastids. The same conclusion can be reached for the TRAPPII component Tca17 which is present in *N. gruberi* (Klinger, *et al.,* 2013) and *E. gracilis* but not present in any of the Kinetoplastid organisms.

All members of the coatomer complex were found having the same copy numbers as that of their distant relatives, the kinetoplastids (Figure 5.1). At least one copy of subfamilies were found and include: COP1 (COPG, COPB1, COPE, COPA and COPB2), COPII (SEC13, SEC31, SEC23 and SEC24) and Clathrin (CHC and CLC) (Schlacht and Dacks, 2015, Adung'a, *et al.,* 2013; van Dam, *et al.,* 2013; Hirst, *et al.,* 2011; McMahon, *et al.,* 2011), suggesting that these subfamilies were present in LECA (Schlacht and Dacks, 2015). The medium and small subunits of the COP1 subfamily were missing. I was unable to identify SEC12 and SEC16 which is consistent with the results across several eukaryotic genomes (Schlacht and Dacks, 2015). Overall, there are strong evidences of kinetoplastid sister relatedness, Endosymbiotic Gene Transfer and paralogous gene expansions within the *E. gracilis* protein sub family clades across all the trafficking genes investigated. For instance, 9 Rab subfamilies show paralogous expansions, *E. gracilis* AP1G is more related to trypanosome AP1G, while *E. gracilis* AP4B is more related to *A. thaliana* AP4B (Figure 5.4). The paralogous gene expansion result is particularly intriguing given the fact that protein targeting to the secondary plastid in euglena has been shown to involve trafficking via the Golgi body. Therefore, this implies that at least two novel membrane-trafficking pathways should exist in Euglena, one anterograde TGN to plastid and a retrograde Plastid to TGN pathway (Ebenezer, *et al.,* 2017). We predict that the relevant machinery for such pathways would be produced via two potential mechanisms, either EGT from the green algal host or via gene duplications of host membrane-trafficking machinery.

### *5.2.1 Meiosis*

The inventory of ciliate meiotic genes from Chi, *et al.,* 2013 and *T. brucei* genomes (Berriman, *et al.,* 2005) in the Tritryp database (Aslett, *et al.,* 2010) were used to interrogate the *E. gracilis* transcriptome using a combination of BLASTp, PSI-BLAST, and the OrthoFinder algorithm (Emms and Kelly, 2015). Chi, *et al.,* 2013, inventoried 11 meiosis-specific genes. Seven (7) of these genes are conserved in at

least one of the ciliates. Of these 7 conserved genes, 6 were found and classified into functions in *E. gracilis* and include: SPO11/REC12, DMC1, HOP2/TBP1, MND1, MSH4 and MSH5 (Table 5.2), suggesting the presence of conserved meiotic activities in *E. gracilis* as present in most eukaryotes. REC8 which is present in only *Tetrahymena thermophila* is absent in the *E. gracilis* transcriptome. REC8 is the only conserved meiotic specific gene I could not find. The absence of this gene may be due to the incomplete dataset or that *E. gracilis* may have a different way of regulating sister chromatid cohesion and recombination between its homologous chromosomes (Parisi, *et al.,* 1999).

Funtional analysis suggests that SPO11 product initiates recombination by forming double-strand breaks in DNA (Chi, *et al.,* 2013); DMC1 is essential for the homolog (nonsister) bias in meiotic recombination (Bugreev, *et al.,* 2011); HOP2 and MND1 protein products form a complex that stabilizes the association of DMC1 with DNA (Chen,*et al.,* 2004), while MSH4 and MSH5 products act as a heterodimer (Snowden,*et al.,* 2004) and are believed to stabilize recombination intermediates (Nishant,*et al.,* 2010). Additional meiosis-related genes were found within our dataset, and include genes involved in bouquet formation, DNA damage sensing/response, double-strand break repair (nonhomology end join), recombinational repair and meiotic entry (Chi, *et al.,* 2013; Hanson, *et al.,* 2013; Akiyoshi, *et al.,* 2013; Peacock, *et al.,* 2011; and Passoss-Silva, *et al.,* 2010). Overall, multiple copies of HOP2 (meiosis-specific) and meiosis related (RAD23, RAD50, KU70, LIG4/DNL, EXO1, and MSH2) genes were identified in the *E. gracilis* transcriptome suggesting high degrees of expansions in copy numbers. Multiple copies of meiotic genes have also recently been associated with parthenogenesis, asexual reproduction, which is present in major lineages (Schurko, *et al.,* 2015). This suggests that reproduction in *E. gracilis* may be a combination of sexual and asexual.

**Table 5.2:** Meiotic genes inventoried in *E. gracilis*

| Category/complex | Gene/convention | nrBLAST evalue | Ref. Seq. ID | Paralogous Seq. ID | Final classification |
|---|---|---|---|---|---|
| Double-strand break formation | SPO11/REC12 | | | | |
| Crossover regulation | DMC1 | 2E-95 | EG_transcript_23829 | | RAD51 |
| | HOP2/TBP1 | 0.00005 | EG_transcript_36872 | EG_transcript_16379 | HOP2 |
| | HOP1 | 6E-43 | EG_transcript_10398 | | HOP1 |
| | MND1 | 2E-57 | EG_transcript_22650 | | MND1 |
| Double-strand break repair and meiotic divisions | REC8 | | | | |
| Bouquet formation | MPS3/SUN-1(c)/SAD1(m) | 5E-37 | EG_transcript_24787 | | MPS3/SUN-1(c)/SAD1(m) |
| DNA damage sensing/response | MEC1/ATR | 0 | EG_transcript_350 | | MEC1/ATF |
| | MRE11 | 2E-131 | EG_transcript_8155 | | MRE11 |
| | RAD17 | | | | |
| | RAD23 | 2E-32 | EG_transcript_24159 | EG_transcript_55709 | RAD23 |
| | RAD24 | 2E-122 | EG_transcript_19426 | | |
| | RAD50 | 3.00E-093 | EG_transcript_10538 | EG_transcript_2370 | RAD50 |
| Double-strand break repair (nonhomology end join) | KU70 | 3E-42 | EG_transcript_4457 | EG_transcript_3154 | |
| | KU80 | 4E-137 | EG_transcript_2345 | | |
| | LIG4/DNL1 | 6E-12 | EG_transcript_14285 | EG_transcript_14864, EG_transcript_1394 | |
| Recombinational repair | BRCA2 | 8E-44 | EG_transcript_13110 | | BRCA2 |
| | DNA2 | 9.00E-111 | EG_transcript_2563 | | |
| | MMS4/EME1 | | | | |
| | EXO1 | | EG_transcript_8622 | EG_transcript_11040 | EXO1 |

| | | | | | |
|---|---|---|---|---|---|
| | MSH2 | 0 | EG_transcript_2281 | EG_transcript_29157 | MSH2 |
| | MSH3 | 0 | EG_transcript_1023 | | MSH3 |
| | MSH4 | 3.00E-161 | EG_transcript_5741 | | MSH2 |
| | MSH5 | 3.00E-172 | EG_transcript_5382 | | MSH2 |
| Cell cycle control | CDC2 | | | | |
| Meiotic entry | | | | | |

**Note:** The table shows the meiotic genes inventoried in *E. gracilis*. Column 1 = category or complex. Column 2 = gene/convention. Column 3 = nrBLAST evalue for the reference sequence. Column 4 = Reference sequence ID. Column 5 = Paralogous sequence ID. Column 6 = Final classifications. Green coloured cell and corresponding text means the presence of the corresponding protein and the protein name. Empty white cells in columns means that the corresponding protein /process was not found or that a final classification could not be made due to insufficient statistical confidence.

### *5.2.3 Bilobes and associated proteins*

The bilobe is a discrete cytoskeletal structure and tightly associated with flagellar skeleton (Esson, *et al.,* 2012; Morriswood, *et al.,* 2013; Wang, *et al.,* 2012; Zhou, *et al.,* 2010; McAllaster, *et al.,* 2015; Aslett, *et al.,* 2010; Bugreev, *et al.,* 2011). In kinetoplastids, the bilobe plays a central role in Golgi, flagellar pocket collar and flagellum attachment zone biogenesis (Zhou, *et al.*, 2010). Five (MORN1, RRP1, BILBO1, and Centrin-4) of the six proteins which localizes within the bilobe structure were found in *E. gracilis* in multiple copies*,* with PLK proteins missing in the dataset (Table 5.3). Centrins are integral to the discovery of the bilobe, as it was first observed using the pan-centrin monoclonal antibody 20H5 (Esson, *et al.,* 2012). The bilobe architecture in *E. gracilis* is similar to that found in trypanosomatids and it's proteins possibly performs similar functions which include precise positioning, biogenesis, and inheritance of single-copied structures of the Golgi apparatus during cell cycle (Gheiratmand and He, 2014). This is significant because the bilobe structure and it's associated proteins play a role in facilitating protein entry into the cell, thus providing a link between the cytoskeleton and the endomembrane system (Morriswood, 2015), bilobe and FAZ biogenesis, consequently affecting flagellum-driven cell motility and division (Brasseur, *et al.,* 2014).

**Table 5.3:** The Bilobe proteins of *E. gracilis*

| Category/complex | Gene/convention | nrBLAST evalue | Ref. Seq. ID | Paralogous Seq. ID | Final classification |
|---|---|---|---|---|---|
| MORN1 | MORN1 | 5.00E-164 | EG_transcript_12856 | EG_transcript_1091, EG_transcript_23152, EG_transcript_27434, EG_transcript_32257, EG_transcript_5232, EG_transcript_5370, EG_transcript_55440, EG_transcript_6309, EG_transcript_6990, EG_transcript_8913 | MORN1 |
| LRRP1 | LRRP1 | 1E-64 | EG_transcript_16676 | | LRRP putative |
| Centrin2 | Centrin2 | | | | |
| Centrin4 | Centrin4 | 3E-121 | EG_transcript_27131 | EG_transcript_21599, EG_transcript_30159, EG_transcript_30159, EG_transcript_31683, EG_transcript_41051 | Centrin4/Centrin/Centrin2 |
| BILBO1 | BILBO1 | 4E-25 | EG_transcript_29908 | | |
| PLK1 | PLK1 | | | | |
| Plectin | Plectin | | | | |
| Tbccd1 | ccd1 | 1.00E-040 | EG_transcript_7280 | | |

**Note:** The table shows the meiotic genes inventoried in *E. gracilis*. Column 1 = category or complex. Column 2 = gene/convention. Column 3 = nrBLAST evalue for the reference sequence. Column 4 = Reference sequence ID. Column 5 = Paralogous sequence ID. Column 6 = Final classifications. Green coloured cell and corresponding text means the presence of the corresponding protein and the protein name. Empty white cells in column 3 and 4 means that the corresponding protein /process was not found or that a final classification could not be made due to insufficient statistical confidence.

*5.2.4 Nuclear cohorts, architecture, and chromatin organization*

*T. brucei, A. thaliana, S. cerevisiae* and *H. sapiens* nucleoporin and lamina proteins were used to interrogate (BLAST and HMMER) the *E. gracilis* transcriptome dataset for the presence of the lamina proteins (NUPs/Nuclear Pore Complex) and nucleoporins (Nups). All the instances of the Nucleoporins/NPC subcomplexes were found except the lamina proteins (NUP-1 and NUP-2) (Figure 5.7). Nucleoporins not found in *E. gracilis* include Mlp2, ScNup133, ScNup159, ScNup42, HsNup43, HsNup37, ScNup2, TbNup109, TbNup37, ScPom34, ScPom121, TbNup110, and ScNDC1 (see Figure 5.7 for additional Nucleoporins). In some cases where a Nuceloporin not listed here may be absent in *E. gracilis*, this data may not have been shown. All the nucleoporins found in *E. gracilis* are also conserved across *T. brucei*, *S. cerevisiae*, and *H. sapiens*, except TbNup48, TbNup109, TbNup41, ScPom152, TbNup140, FGNup153, Mex67a, Mex67b, Gle2a, Gle2b, and ScDBP6 (see Figure 5.7 for additional Nucleoporins) (Obado, *et al.,* 2016; Kabachinski, *et al.,* 2015; Neumann, *et al.,* 2010; Holden, *et al.,* 2014; Koreny and Field, 2016; Dubois, *et al.,*2012; Wilson and Dawson, 2011; Curtis, *et al.,* 2012; and Ibarra, *et al.,* 2015). Protein structure of the Nucleoporin found include α-solenoid, β-propellar, transmembrane, and coiled coil (see Figure 5.7 for details). DBP5/DD19 which is present in *H. sapiens* and *S. cerevisiae* but absent in trypanosomes was also found in the dataset (Figure 5.8). While the nuclear cohorts found in *E. gracilis* are conventional with other reference eukaryotes, the nuclear lamins (NUP-1 and NUP-2) were not found. Similarly, analysis of the heterochromatin organization (Figure 5.9) shows a less condensed heterochromatin and the absence of peripheral heterochromatin around the nuclear envelope.

| Protein structure | No. | Tb | At | Hs | Sc | Eg | NPC subcomplex |
|---|---|---|---|---|---|---|---|
| α-solenoid | 1 | TbNup225 | Nup205 | Nup205 | Nup192 | EG_transcript_112 | Inner ring |
| | 2 | TbNup181 | Nup188 | Nup188 | Nup188 | EG_transcript_119 | |
| | 3 | TbNup96 | Nup93 | Nup93 | Nic96 | EG_transcript_2443 | |
| | 4* | TbNup158 | Nup98 | Nup98 | Nup145N | EG_transcript_2261 | Outer ring |
| | 5 | TbNup158 | Nup96 | Nup96 | Nup145C | EG_transcript_4195 | |
| | 6 | TbNup89 | Nup75 | Nup75 | Nup85 | EG_transcript_5081 | |
| | 7 | TbNup82 | Nup107 | Nup107 | Nup84 | EG_transcript_2952 | |
| β-propeller α-solenoid (β-α) | 8 | TbNup144 | Nup155 | Nup155 | Nup170 | EG_transcript_351 | Inner ring |
| | 9 | TbNup119 | Nup155 | Nup155 | Nup157 | EG_transcript_351 | |
| | 10 | TbNup152 | Nup160 | Nup160 | Nup120 | EG_transcript_636 | Outer ring |
| | 11 | TbNup132 | Nup133 | Nup133 | Nup133 | EG_transcript_1761 | |
| | 12 | TbNup109 | | | | | |
| β-propeller | 13 | TbSec13 | Sec13 | Sec13 | Sec13 | EG_transcript_21910 | |
| | 14 | | Seh1 | Seh1 | Seh1 | EG_transcript_15911 | |
| | 15 | TbNup41 | Nup43 | Nup43 | | EG_transcript_14251 | |
| | 16 | | | Nup37 | | | |
| β-propeller, coiled coil | 17 | TbNup48 | ALADIN | ALADIN | | EG_transcript_10094 | |
| RRM, ALPS motif, TM (Trypanosomes) | 18 | TbNup76 | Nup88 | Nup88 | Nup82 | EG_transcript_3164 | Cytoplasmic (opisthokonts), Symmetric (Trypanosomes) |
| *Trans*-membrane | 19 | TbNup65 | Nup35 | Nup35 | Nup53/59 | EG_transcript_16344 | Inner ring |
| | 20 | x | Gp210 | Gp210 | Pom152 | EG_transcript_157 | Pore membrane ring |
| | 21 | x | NDC1 | NDC1 | NDC1 | | |
| | 22 | x | x | x | Pom34 | | |
| | 23 | x | x | Pom121 | x | | |
| Coiled Coils | 24 | TbNup110 | x | x | x | | Nuclear basket |
| | 25 | TbNup92 | x | x | x | EG_transcript_2749 | |
| | 26 | x | NUA | TPR | Mlp1 | EG_transcript_9861 | |
| | 27 | x | x | x | Mlp2 | | |

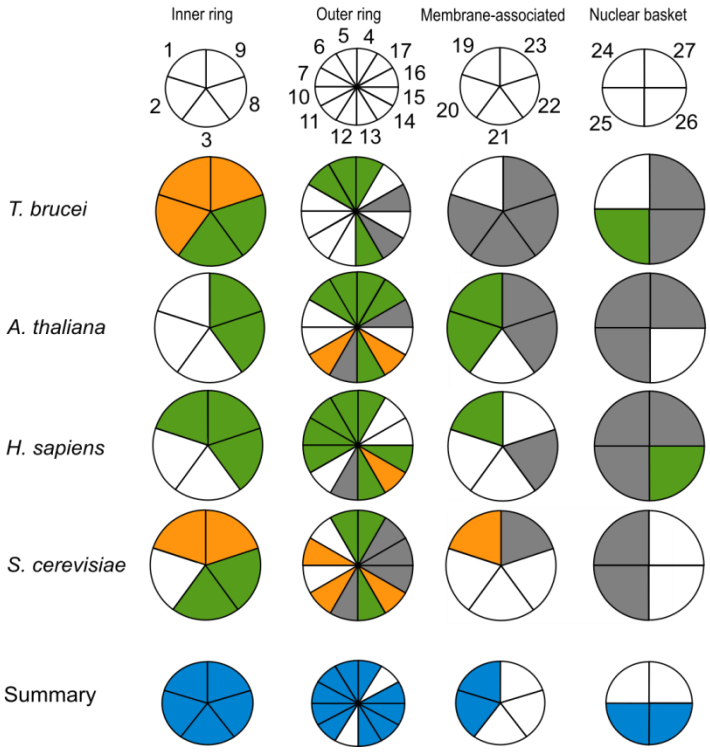


**Figure 5.7: The Nuclear Pore Complex of *E. gracilis* is conventional.** Predicted orthology relationships for *E. gracilis* nucleoporins compared to human, yeast, plant and African trypanosomes. Nuclear pore proteins are grouped according to secondary structural architecture (left) and subcomplex (right) and color coded for clarity. Orthologs for each reference species together with the predicted *E. gracilis*accession are shown. Open (white) cells indicate absence of an *E. gracilis* candidate and empty colored cells where a reference species lacks a known candidate. The lower panel is a Coulson plot for subcomplexes using the numerical scheme in the table. Colors indicate the confidence with which Euglena orthologs are ascribed against each reference species, with white, orange, green, and gray (for blastp= 1e-04 and psi-blast = 1e-05) respectively. Gray = absence of gene in reference proteome. White = presence of gene in reference proteome but no detection of homologs through blastp nor psi-blast. Green = homolog found through reciprocal blastp (most confident). Yellow = homologg found through psi-blast (less confident). The summary (blue; found) is the most parsimonious interpretation of the presence or absence of the NPC subunits based on the above.

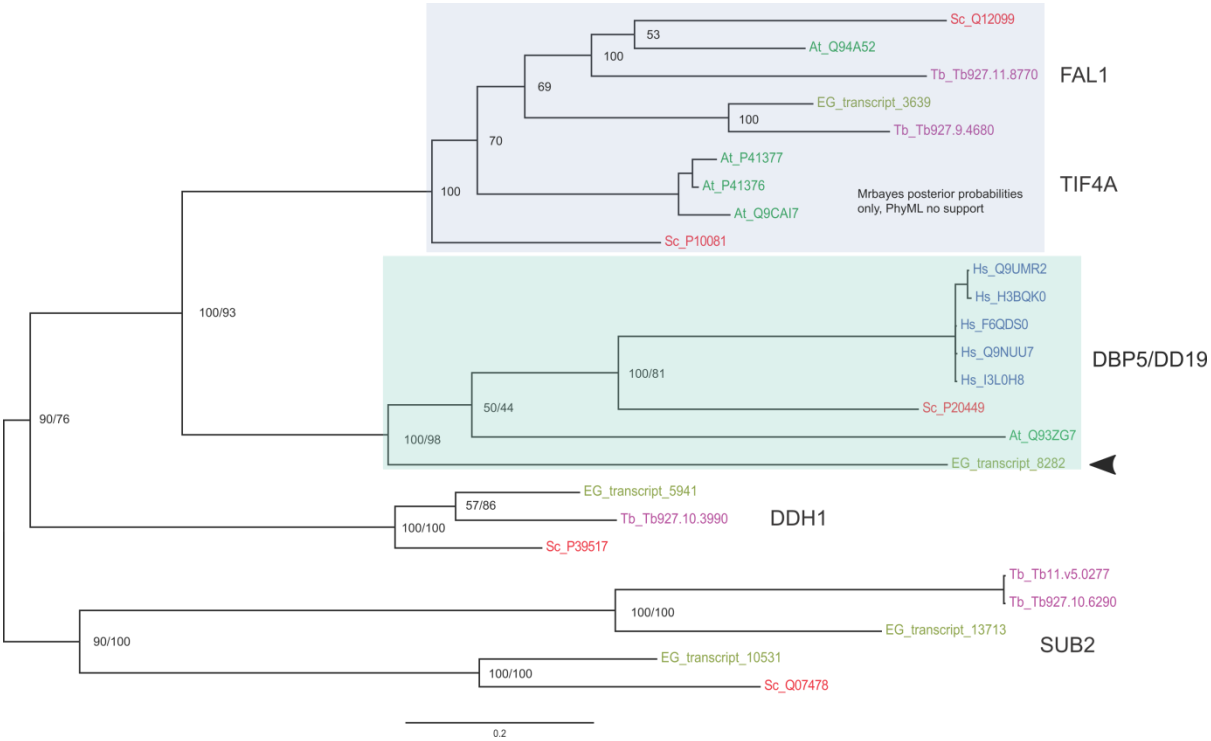



**Figure 5.8: DBP5 is a characteristic feature of *E. gracilis* mRNA export mechanism.** The tree shows the phylogenetic evidence for the presence of a Euglenid Dbp5 RNA helicase. A maximum-likelihood/MrBayes phylogenetic reconstruction of selected taxa, containing Dpb5 orthologs and their closest relatives is shown. Contrary to African trypanosomes, *E. gracilis* has a clear Dbp5 ortholog (arrowhead), consistent with themore complex mRNA splicing pathways present.*S. cerevisiae*; red, *A. thaliana*; green, *E. gracilis*; Khaki, *T. brucei*; purple, *H. sapiens*: blue.

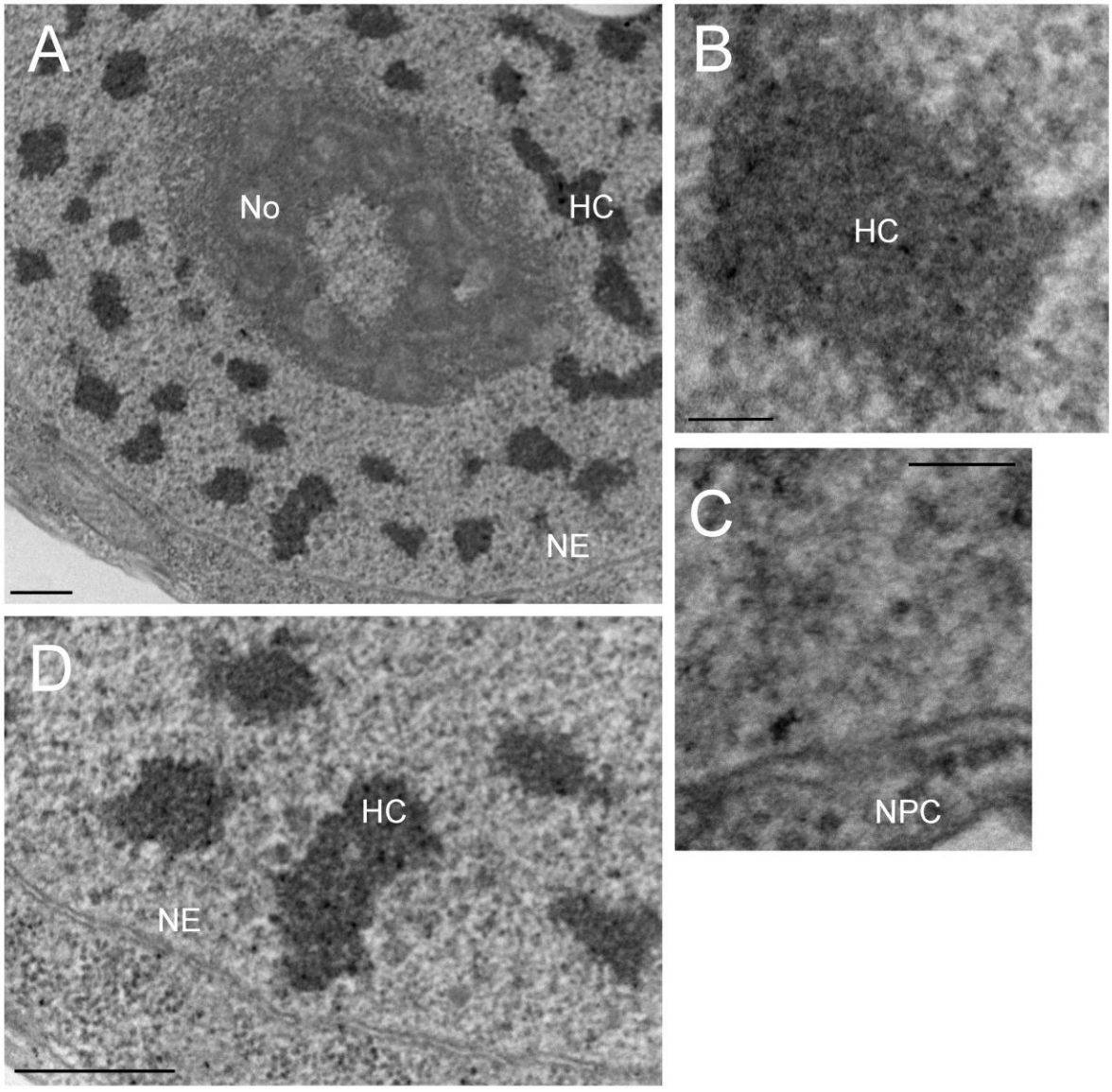



**Figure 5.9: The *E. gracilis* Nuclear Pore Complex are lacking in peripheral heterochromatin organisation and nuclear lamins.** Gallery of electron micrographs illustrating the unusual 'currant bun' arrangement of heterochromatin in *E. gracilis*. NE; nuclear envelope, HC; heterochromatin, No; nucleolus and NPC; nuclear pore complex.

*5.2.5 Tubulins*

Reference tubulin genes from eight (8) eukaryotic taxa (*L. major, P. EM1*, *T. brucei, T. cruzi, B. saltans, H. sapiens, S. cerevisiae,* and *A. thaliana*) were utilized to interrogate the *E. gracilis* transcriptome and genome for the instances of tubulin gene families (β, α, ε, ɣ, δ, ζ, and FtsZ). The beta-, alpha-, epsilon- (2 copies), gamma-, delta-, zeta-, and FtsZ-tubulin gene families were all found within the transcriptome assembly with strong statistical supports (Figure 5.10, Table 5.4). The *E. gracilis* tubulin genes showed similarity to orthologs from the kinetoplastids than they do to *A. thaliana*, *H. sapiens*, and *S. cerevisiae*, except for beta- and gamma-tubulins which are closely related to *A. thaliana* than to the kinetoplastids. The FtsZ tubulin genes (outgroup in Figure 5.10) that are mostly found in prokaryotes and other lower eukaryotes is also found in *E. gracilis*, suggesting that the prokaryotic and eukaryotic homologs of the cell division machinaries may well be present in *E. gracilis*. Finally, there are 18 introns in the *E. gracilis* tubulin genes which are conventional, intermediate non conventional, or non conventional introns (see Ebenezer, *et al.,* 2017, for an extensive discussion). Hence, demonstrating some novel splicing mechanism within this biological system (Ebenezer, *et al.,* 2017, Ebenezer, *et al.,* 2018).

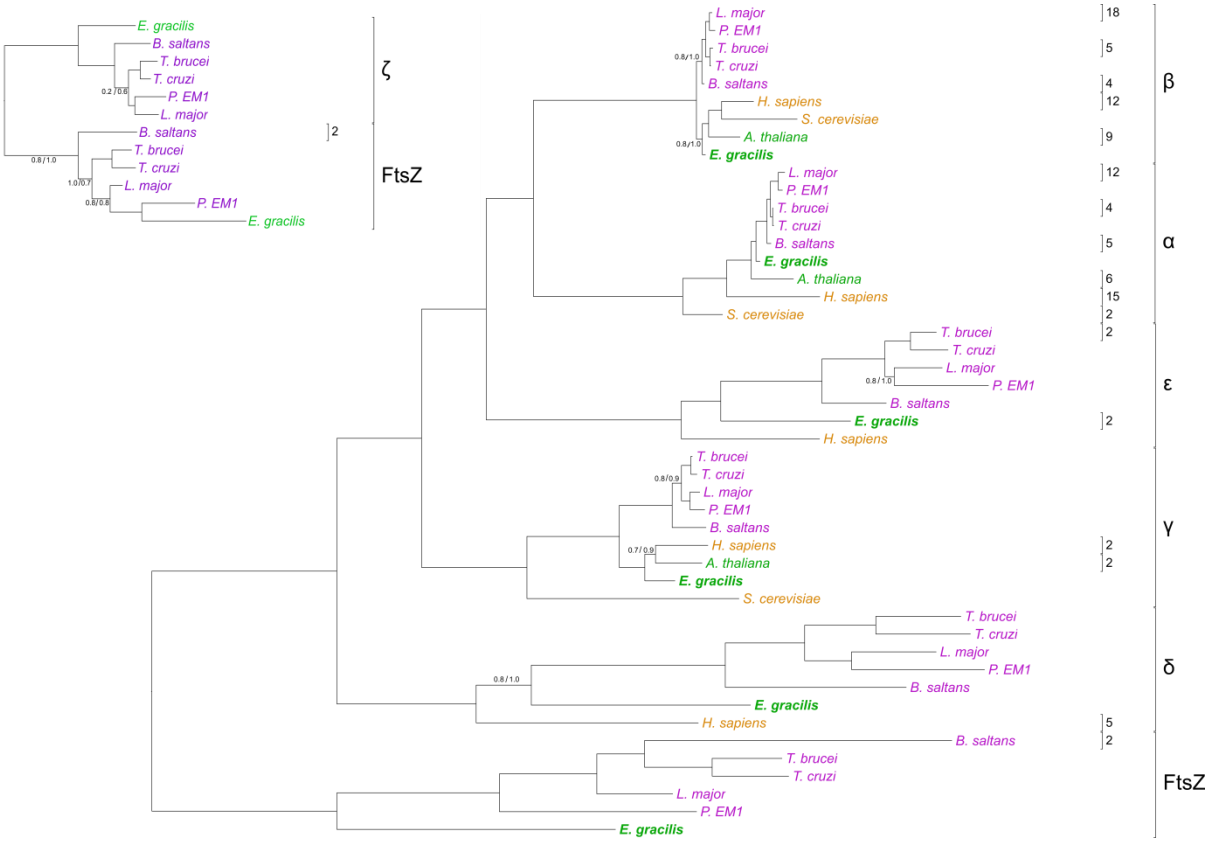



**Figure 5.10: The tubulin genes of *E. gracilis* are conventional and ancient.** The phylogenetic tree shows α-, β-, ɣ-, ε-, δ-, ζ-, and FtsZ tubulin genes in selected eukaryotes. The tree at the top left corner of the diagram depicts ζ-tubulin genes while larger tree depicts α-, β-, ɣ-, ε-, and δ-tubulin genes. In both instances, FtsZ tubulin family was used as the outgroup. Statistical support are in the order FastTree/MrBayes, and generated using Maximum Likelihood based on 1000 bootstrappings and Posterior Probability respectively. Nodes without statistical support implies that this is above the threshold of ≥99/99. The numbers on the right of the tree, corresponding to specific taxa name, represent the copies of the tubulin sub family found in that taxa. Each clade correspond to each tubulin sub family and are classified as either α-, β-, ɣ-, ε-, δ-, ζ-, or FtsZ tubulin. Color codes: *Green* photosynthetic organisms, *purple* kinetoplastids, *orange* other members of the eukaryotes. I decided to show the phylogenetic tree for the transcriptome since all tubulin subfamilies were found in this dataset, and it is a global representatives of the tubulin genes in *E. gracilis.*

**Table 5.4:** Tubulin gene families of *E. gracilis*

| Reference Seq ID | Final classification |
|---|---|
| EG_transcript_14010 | Beta-, β- |
| EG_transcript_10562 | Alpha- , α- |
| EG_transcript_10714 | Gamma-, ɣ- |
| EG_transcript_18972 | Epsilon-, ε- |
| EG_transcript_17300 | Epsilon-, ε- |
| EG_transcript_7888 | Delta-, δ- |
| EG_transcript_10899 | Zeta-, ζ- |
| EG_transcript_23781 | FtsZ tubulin family |

**Note:** The table shows the tubulin gene families inventoried in *E. gracilis*. Column 1 = Reference sequence ID. Column 2 = Final classifications. Green coloured cell and corresponding text means the presence of the corresponding protein and the protein name respectively.

### *5.2.6 Transporters*

Members of the transporter superfamilies (ATP binding, SLC, ion channels, and phosphotransferase system (PTS)) (Elboune, *et al,* 2017) were investigated using *T. cruzi*, *T. brucei*, *L. major*, *B. saltans*, *N. gruberi*, and *H. sapiens* orthologs as a reference set. Members of the ATP superfamilies were all found (Table 5.5, and Appendix I-a) as evident in other eukaryotic taxa, suggesting a conserved transport activities in *E. gracilis*. However, some families within these superfamilies were not present, but consistent with the reference taxa.

**Table 5.5:** The transporter proteins of *E. gracilis*

| **Reference sub family** | ***E. gracilis* Ref. Seq. ID** | **Final Euglena classification** |
|---|---|---|
| ABCA2 | EG_transcript_559 | |
| ABCB11 | EG_transcript_634 | |
| ABCC10 (MRP7) | EG_transcript_260 | |
| ABCD2 | EG_transcript_2824 | |
| ABCG2 | EG_transcript_3327 | |
| | EG_transcript_2841 | |
| F-type ATPase α subunit | x | |
| V-type ATPase V1 motor E2 subunit | EG_transcript_21665 | V-type ATPase V1 motor E2 subunit |
| sodium/potassium-transporting ATPase subunit β-2 | x | |
| PMCA1 | EG_transcript_950 | PMCA1 |
| ATP12A | EG_transcript_2393 | ATP12A |
| ATP7B | EG_transcript_901 | ATP7B |
| ATP8B3 (Phospholipid-transporting ATPase IK) | EG_transcript_907 | ATP8B3 (Phospholipid-transporting ATPase IK) |
| synaptic vesicle glycoprotein 2A | EG_transcript_9149 | synaptic vesicle glycoprotein 2A |
| EAAT5 (Excitatory amino acid transporter 5 / SLC1A7) | x | |
| ASCT2 (Alanine/serine/cysteine transporter 2 / SLC1A5) | x | |
| GLUT2 (Glucose transporter 2 / SLC2A2) | EG_transcript_9149 | GLUT2 (Glucose transporter 2 / SLC2A2) |
| GLUT8 (Glucose transporter 8 / SLC2A8) | EG_transcript_9149 | - |
| HMIT (Proton myo-inositol cotransporter / SLC2A13) | EG_transcript_9149 | - |
| rBAT / SLC3A1 | x | |
| xCT (Cystine/glutamate transporter / SLC7A11) | EG_transcript_6332 | xCT (Cystine/glutamate transporter / SLC7A11) |

| | | |
|---|---|---|
| AE2 (Anion exchange protein 2 / SLC4A2) | x | |
| BTR1 (NaBC1 / SLC4A11) | x | |
| SGLT1 (Sodium/glucose cotransporter 1 / SLC5A1) | EG_transcript_495 | |
| CHT / SLC5A7 | x | |
| SMCT1 / SLC5A8 | x | |
| SMIT1 (SMIT / SLC5A3) | x | |
| DAT / SLC6A3 | x | |
| GAT3 / SLC6A11 | x | |
| PROT / SLC6A7 | EG_transcript_5892 | |
| NTT5 / SLC6A16 | x | |
| NCX1 (Sodium/calcium exchanger 1 / SLC8A1) | x | |
| NHE9 (Sodium/hydrogen exchanger 9 / SLC9A9) | EG_transcript_4181 | NHE9 (Sodium/hydrogen exchanger 9 / SLC9A9) |
| P5 (Sodium/bile acid and sulphated solute cotransporter 5 / SLC10A5) | EG_transcript_21665 | |
| DMT1 / SLC11A2 | x | |
| NKCC1 (Basolateral Na-K-Cl symporter / SLC12A2) | EG_transcript_6332 | |
| NaC3 (Na+/dicarboxylate cotransporter 3 / SLC13A3) | EG_transcript_4581 | NaC3 (Na+/dicarboxylate cotransporter 3 / SLC13A3) |
| UT-A (Kidney urea transporter / SLC14A2) | x | |
| PHT2 (Peptide transporter 3 / SLC15A3) | EG_transcript_5713 | PHT2 (Peptide transporter 3 / SLC15A3) |
| MCT8 (Monocarboxylate transporter 8 / SLC16A2) | x | |
| NPT1 (Sodium/phosphate cotransporter 1 / SLC17A1 | EG_transcript_33365 | |
| AST (Sialin / SLC17A5) | EG_transcript_11531 | AST (Sialin / SLC17A5) |
| VGLUT3 (Vesicular glutamate transporter 3 / SLC17A8) | EG_transcript_11531 | VGLUT3 (Vesicular glutamate transporter 3 / SLC17A8) |
| VNUT (Vesicular nucleotide transporter / SLC17A9) | EG_transcript_11531 | VNUT (Vesicular nucleotide transporter / SLC17A9) |

| | | |
|---|---|---|
| VAChT (Vesicular acetylcholine transporter / SLC18A3) | EG_transcript_8497 | |
| FOLT (Reduced folate transporter 1 / SLC19A1) | EG_transcript_6540 | |
| PiT1 (Sodium-dependent phosphate transporter 1 / SLC20A1) | EG_transcript_9187 | PiT1 (Sodium-dependent phosphate transporter 1 / SLC20A1) |
| OCT3 (Organic cation transporter 3 / SLC22A3) | EG_transcript_9149 | OCT3 (Organic cation transporter 3 / SLC22A3) |
| OCTN2 (Organic cation/carnitine transporter 2 / SLC22A5) | EG_transcript_5405 | OCTN2 (Organic cation/carnitine transporter 2 / SLC22A5) |
| OAT5 (Organic anion transporter 5 / SLC22A10) | EG_transcript_5393 | OAT5 (Organic anion transporter 5 / SLC22A10) |
| URAT1 (Urate anion exchanger 1 / SLC22A12) | EG_transcript_9149 | |
| SLC22A24 | EG_transcript_5393 | |
| SVCT2 (Sodium-dependent vitamin C transporter 2 / SLC23A2) | x | |
| NKCX4 (Sodium/potassium/calcium exchanger 4 / SLC24A4 | x | |
| OGC (Mitochondrial oxoglutarate carrier / SLC25A11) | EG_transcript_15350 | |
| CAC (Carnitine/acylcarnitine carrier / SLC25A20 | EG_transcript_17132 | CAC (Carnitine/acylcarnitine carrier / SLC25A20 |
| PHC (Mitochondrial phosphate carrier / SLC25A3) | EG_transcript_11544 | PHC (Mitochondrial phosphate carrier / SLC25A3) |
| APC2 (Mitochondrial phosphate carrier 2 / SLC25A23) | EG_transcript_11597 | APC2 (Mitochondrial phosphate carrier 2 / SLC25A23) |
| UCP1 (Uncoupling protein 1 / SLC25A7) | EG_transcript_14195 | UCP1 (Uncoupling protein 1 / SLC25A7) |
| SLC25A46 | EG_transcript_14195 | SLC25A46 |
| Sat-1 / SLC26A1 | EG_transcript_6540 | |
| Pendrin / SLC26A4 | EG_transcript_6540 | Pendrin / SLC26A4 |
| SLC26A9 | EG_transcript_6540 | SLC26A9 |

| | | |
|---|---|---|
| Tat1 / SLC26A8 | EG_transcript_6540 | Tat1 / SLC26A8 |
| FATP2 (Fatty acid transport protein 2 / SLC27A2) | EG_transcript_6570 | FATP2 (Fatty acid transport protein 2 / SLC27A2) |
| CNT2 (Sodium/nucleoside cotransporter 2 / SLC28A2) | x | |
| ENT3 (Equilibrative nucleoside transporter 3 / SLC29A3 | EG_transcript_7890 | ENT3 (Equilibrative nucleoside transporter 3 / SLC29A3 |
| ZnT4 (Zinc transporter 4 / SLC30A4) | EG_transcript_6135 | ZnT4 (Zinc transporter 4 / SLC30A4) |
| CTR2 (Copper transporter 2 / SLC31A2) | x | |
| VIAAT (Vesicular inhibitory amino acid transporter / SLC32A1) | EG_transcript_8425 | VIAAT (Vesicular inhibitory amino acid transporter / SLC32A1) |
| ACATN1 (AcetylCoA transporter / SLC33A1) | x | |
| NaPi-IIc (Sodium phosphate 3 / SLC34A3) | x | |
| YEA / SLC35B4 | EG_transcript_17625 | YEA / SLC35B4 |
| PAT4 (Proton-coupled Amino acid Transporter 4 / SLC36A4) | EG_transcript_8073 | PAT4 (Proton-coupled Amino acid Transporter 4 / SLC36A4) |
| SPX4 (Glucose-6-phosphate transporter / SLC37A4) | EG_transcript_11531 | SPX4 (Glucose-6-phosphate transporter / SLC37A4) |
| SNAT1 (sodium-coupled neutral amino acid transporter 1 / SLC38A1) | EG_transcript_8425 | SNAT1 (sodium-coupled neutral amino acid transporter 1 / SLC38A1) |
| SNAT3 (Sodium-coupled neutral amino acid transporter 3 / SLC38A3) | EG_transcript_8425 | SNAT3 (Sodium-coupled neutral amino acid transporter 3 / SLC38A3) |
| PP1744 (Putative sodium-coupled neutral amino acid transporter 10 / SLC38A10) | EG_transcript_8425 | PP1744 (Putative sodium-coupled neutral amino acid transporter 10 / SLC38A10) |
| ZIP6 (Zinc transporter 6 / SLC39A6) | x | |
| IREG1 (Ferroportin / SLC40A1) | x | |

| | | |
|---|---|---|
| MgtE (Solute carrier family 41 member 1 / SLC41A1) | x | |
| RhCG (Ammonium transporter Rh type C / SLC42A3) | EG_transcript_5298 | RhCG (Ammonium transporter Rh type C / SLC42A3) |
| EEG1 (Solute carrier family 43 member 3 / SLC43A3) | x | |
| CTL3 (Choline transporter-like 3 / SLC44A3) | EG_transcript_4904 | CTL3 (Choline transporter-like 3 / SLC44A3) |
| Membrane-associated transporter protein / SLC45A2 | x | |
| SLC46A3 | x | |
| MATE1 (Multidrug and toxin extrusion / SLC47A1) | EG_transcript_7180 | |
| HRG1 (Heme transporter / SLC48A1) | EG_transcript_239 | |
| DIRC2 (Disrupted in renal carcinoma 2 / SLC49A4) | EG_transcript_9532 | DIRC2 (Disrupted in renal carcinoma 2 / SLC49A4) |
| RAG1AP1 (SLC50 sugar exporter / SLC50A1) | EG_transcript_24253 | RAG1AP1 (SLC50 sugar exporter / SLC50A1) |
| OSTβ (Organic solute transporter subunit β / SLC51A1BP) | x | |
| RFVT2 (solute carrier family 52 member 2 / SLC52A2) | x | |
| OATP4C1 / SLCO4C1 | EG_transcript_5393 | |
| ACC | x | |
| Amt | EG_transcript_5298 | Amt |
| annexin | x | |
| Bcl-2 | x | |
| | x | |
| CD20 | x | |
| connexin | x | |
| Calcium release-activated calcium channel protein 1 | x | |
| E-CIC | x | |
| ENaC | x | |
| GIC | x | |

| | | |
|---|---|---|
| Hsp70 | EG_transcript_8430 | Hsp70 |
| ICC | x | |
| innexin | x | |
| IRK-C | x | |
| Acetylcholine receptor subunit alpha | x | |
| Mid1 | x | |
| aquaporin 9 | EG_transcript_19390 | aquaporin 9 |
| CorA | EG_transcript_8861 | |
| MscL | x | |
| MscS | EG_transcript_6444 | MscS |
| NSCC2 | x | |
| PCC | x | |
| PLM | x | |
| RIR-CaC | EG_transcript_43 | RIR-CaC |
| TRP-CC | x | |
| UAC | x | |
| UT | x | |
| VIC | EG_transcript_18356 | VIC |

**Note:** The table shows the transporter genes inventoried in *E. gracilis*. Column 1 = Reference sub family. Column 2 = *E. gracilis* Reference sequence ID. Column 3 = Final classifications. Green coloured cell and corresponding text means the presence of the corresponding protein and the protein name respectively. Empty white cells or “x” means that the corresponding protein /process was not found or that a final classification could not be made due to insufficient statistical confidence.

### *5.2.7 Histones*

Five major families of histones exist: H1/H5, H2A, H2B, H3, and H4 (Cox, *et al.,* 2005; Bhasin, *et al.,* 2006; Hartl, *et al.,* 1988). Histones H2A, H2B, H3 and H4 are known as the core histones, while histones H1 and H5 are known as the linker histones (Tilber, *et al.,* 2012). In the *E. gracilis* transcriptome, I found all of the core histones (Table 5.6), suggesting eukaryotic conservation activity (Tilber, *et al.,* 2012). The analysis also revealed other histone variants (macroH2, H3.2, and H3.3), with strong statistical confidence for H3.2 and H3.3, and low statistical confidence for macroH2 (Table 5.6, and Appendix-Ia). The number of histone variants in *E. gracilis* is quite low compared to those found in other taxa (Tilber, *et al,* 2012).

**Table 5.6:** List of histones identified in *E. gracilis*

| Category/complex | Convention | nrBLAST evalue | Ref. Seq. ID | Paralogous Seq. ID | Final classification |
|---|---|---|---|---|---|
| *Canonical/replication coupled histones* | | | | | |
| H1 | H1 | 2E-10 | EG_transcript_37574 | EG_transcript_53416 | H1 |
| H2A | H2A | 1E-88 | EG_transcript_34406 | EG_transcript_25106 | H2A |
| H2B | H2B | 9.00E-047 | EG_transcript_48180 | EG_transcript_67478 | H2B |
| H3 | H3 | 5.00E-34 | EG_transcript_63430 | EG_transcript_27533, EG_transcript_30647, EG_transcript_33619 | H3 |
| H4 | H4 | 7.00E-48 | EG_transcript_37574 | EG_transcript_53416 | H4 |
| H5 | H5 | | | | |
| | | | | | |
| *Variant histones* | | | | | |
| H1 | H1.0 | | | | |
| | H1.10 | | | | |
| | OO H1.8 | | | | |
| | scH1 | | | | |
| | TS H1.6 | | | | |
| | TS H1.7 | | | | |
| | TS H1.9 | | | | |
| H2A | H2A.X | | | | |
| | H2A.Z.1 | | | | |
| | H2A.Z.2.1 | | | | |
| | H2A.Z.2.2 | | | | |
| | H2A.Bbd/H2A.B | | | | |
| | macroH2A1.1 | | | | |
| | macroH2A1.2 | | | | |
| | macroH2A2 | | | | |
| | H2A.Z | | | | |

| | | | | | |
|---|---|---|---|---|---|
| | H2A.1 | | | | |
| | H2A.L | | | | |
| | H2A.P | | | | |
| | H2A.W | | | | |
| | macroH2A | 6E-43 | EG_transcript_8713 | | |
| H2B | H2B.W (H2B type WT) | | | | |
| | H2B type 1A (TSH2B) | | | | |
| | H2B.1 | | | | |
| | H2B.W | | | | |
| | H2B.Z | | | | |
| | Sperm H2B | | | | |
| | subH2B | | | | |
| H3 | H3.3 | | EG_transcript_64233 | EG_transcript_21219 | H3.3 |
| | H3.2 | | EG_transcript_29591 | | H3.2 |
| | CENP-A/cenH3 | | | | |
| | H3.1T | | | | |
| | H3.5 | | | | |
| | H3.X/H3.Y.2 | | | | |
| | H3.Y/H3.Y.1 | | | | |
| | H3.V | | | | |
| | TS H3.4 | | | | |
| | H2B.V | | | | |
| H4 | H4.V | | | | |

**Note:** The table shows the meiotic genes inventoried in *E. gracilis*. Column 1 = category or complex. Column 2 = gene/convention. Column 3 = nrBLAST evalue for the reference sequence. Column 4 = Reference sequence ID. Column 5 = Paralogous sequence ID. Column 6 = Final classifications. Green coloured cell and corresponding text means the presence of the corresponding protein and the protein name. Empty white cells in column 3 and 4 means that the corresponding protein /process was not found or that a final classification could not be made due to insufficient statistical confidence.

### 5.2.8 Calmodulins and Rhodopsins

Analysis of the genome and transcriptome suggests the presence of calmodulin genes and the absence of rhodopsin genes. Interrogation of the data suggests the absence of camodulin 1 – 7 in the genome, with a single copy present in the transcriptome. Forty four CaM-related potential calcium sensors (McCormack and Braam, 2003) were found in the *E. gracilis* transcriptome, while only 23 were found in the genome (Ebenezer, *et al.,* 2017, 2018), suggesting a highly complex mechanism for the control of signal transduction and that CaM/CaM-related proteins likely have very important roles in the responses of *E. gracilis* to environmental and other cues (see Ebenezer, *et al.,* 2017, for extensive discussion, data, and publication of full result). Analysis for the presence of rhodopsin did not reveal any strong rhodopsin candidate (Table 5.7. Genes found include the Adenylyl/Guanylyl cyclase and the Rhodopsin-guanylyl cyclase (RhGC) (see Table 5.7, and Appendix I-a).

**Table 5.7:** Rhodopsin proteins in *E. gracilis*

| Category/complex | Gene | | nrBLAST evalue | Ref. Seq. ID | Final classification |
|---|---|---|---|---|---|
| | **Microbial or animal convention** | **Reference seqid** | | | |
| Adenylyl/Guanylyl cyclase | HsGC1 | Q02846 | 2.00E-015 | EG_transcript_2162 | GC1 |
| | HsGC2 | P51841 | 2.00E-015 | EG_transcript_9517 | GC2 |
| | EgPACA | Q8S9F2 | 0 | EG_transcript_1073 | PACA |
| | EgPACB | Q8S9F1 | 0 | EG_transcript_2020 | PACB |
| Histidine kinase (CheA) | DmCheA | Q4AB33 | | | |
| | EcCheA | P07363 | | | |
| | HsCheA | B0R4J9 | | | |
| Histidine Kinase rhodopsin (HKR) | CrHKR | Q6WRU3 | 6.00E-060 | EG_transcript_1624 | |
| Rhodopsin-guanylyl cyclase (RhGC) | SrGC | F2TZN0 | 2.00E-117 | EG_transcript_12299 | GC |
| Visual pigment rhodopsin (G protein coupled receptor) | Homo sapiens rhodopsin | P08100 | | | |
| | Danio | P35359 | | | |
| | Gallus | P22328 | | | |
| Retinochrome (G protein coupled receptor) | Homo sapiens rhodopsin | P47804 | | | |
| G protein arrestin (GPCR) | ARRB2 | P32121 | 7.00E-014 | EG_transcript_11761 | ARRB1 |
| | ARRB1 | P49407 | 1.00E+000 | EG_transcript_11761 | |
| Sensory rhodopsin (SR) | HsSRI | P0DMH8 | | | |
| | HvSRI | M0IXQ5 | | | |
| | SrSRI | D5HC99 | | | |
| | HsSRII | P71411 | | | |
| | NpSRII | P42196 | | | |
| | HvSRII | P42199 | | | |
| Bacteriorhodopsin (BR): H+ | GR | Q7NP59 | | | |

| | | | | | |
|---|---|---|---|---|---|
| Pump | XR | Q2S2F8 | | | |
| | GPR1 | EAQ40507 | | | |
| | GPR2 | ZP_02194911 | | | |
| | GPR3 | Q4FMZ3 | | | |
| | BPR4 | BAL68143 | | | |
| | BPR5 | AAK30179 | | | |
| | GPR6 | AAG10475 | | | |
| | GPR7 | AAK30176 | | | |
| | LR | Q9HGT7 | | | |
| | Ph2 | Q0V5A7 | | | |
| | NR | Q9UW81 | | | |
| | Ph1 | Q0V6M3 | | | |
| | Ace1 | AEF12206 | | | |
| | Ace2 | AEF12207 | | | |
| | MR | WP_011570314.1 | | | |
| | BR | P02945 | | | |
| | AR2 | P29563 | | | |
| | AR1 | P69051 | | | |
| | AR3 | P96787 | | | |
| Halorrhodopsin (HR): Cl- Pumps | NpHR | P15647 | | | |
| | SrHR | Q5UNE6 | | | |
| | HsHR | B0R2U4 | | | |
| Krokinobacterhodopsin (KR): Na+ pumps | Krodi_2220 | F4AZU8 | | | |
| | Krodi_0510 | F4AYM8 | | | |
| Cation Conduction Channel Rhodopsin (ChR) | VcChR2 | ABZ90902 | | | |
| | CrChR2 | XP_001701725 | | | |
| | VcChR1 | ABZ90900 | | | |
| | CrChR1 | AAL08946 | | | |
| | MvChR1 | AEI83869 | | | |
| Anion Conduction Channel Rhodopsin (ACR) | GtACR1 | L1J207 | | | |
| | GtACR2 | L1IFZ3 | | | |

**Note:** The table shows the Rhodopsin genes inventoried in *E. gracilis*. Column 1 = Category/complex. Column 2 = Gene name. Column 3 = nr BLAST. Column 4 = *E. gracilis* reference sequence ID. Column 5 = Final classifications. Green coloured cell and corresponding text means the presence of the corresponding protein and the protein name respectively. Empty white cells in column 3 and 4 means that the corresponding protein /process was not found or that a final classification could not be made due to insufficient statistical confidence. Colour codes in column 1: Orange = present in both microbes, fungi and algae. Purple = present in only animals.

## 5.3 INFORMATION STORAGE AND PROCESSING

The reception, storage and processing of information by cellular organisms (within and outside their surroundings) is an essential cellular activity. A biological cell can be viewed as a dynamic information-processing system that responds to and interacts with a varied and changing environment (D'Onofrio and An, 2010). Cellular actions rely on a set of operations between the genetic information encoded in the cell's DNA and its intracellular information-processing infrastructure (RNA and proteins) (D'Onofrio and An, 2010). The structure and function of this information processing complex are of great interest in the study of both normal cellular functions (such as differentiation and metabolism) and pathological conditions (such as oncogenesis and dysregulation) (D'Onofrio and An, 2010). In eukaryotes, many of these information storage and processings include, transcription, translation, splicing, RNA editing, and RNAi pathway. In this section, these information storage and processes in *E. gracilis* will be discussed.

### *5.3.1 Translational apparatus*

The characterized *T. brucei* translational apparatus reference sets (Aslett, *et al.,* 2010) were used to predict the minimal translational apparatus in *E. gracilis* using BLAST, OthorFinder and phylogenetic inference (Table 5.8 and Appendix-Ia). Overall, 158 translational genes were found in *E. gracilis* and this include: Ribosomal proteins (70), Aminoacyl-tRNA (50), translational initiation (18), elongation (6), and termination (3) (Grosjean, *et al.,* 2014; Hernández, 2012; Hernández, *et al.,* 2012). RPL37, RPL37a, RPL38, RPL39, RPP0b, RPP0c, RPS21, RPS27, RPS30, RPS33, glycyl-tRNA synthetase, the recycling factor (RF) were not found (Table 5.8). These are also consistent with the translational apparatus evident in other members of the kinetoplastid and eukaryotes.

**Table 5.8:** The set of proteins for functional translational apparatus in *E. gracilis.*

| Category/complex | Gene/convention | Ref. Seq. ID | Paralogous Seq. ID | Final classification |
|---|---|---|---|---|
| LSU (60S) | Nop53 | | | |
| | RPL10 | | | |
| | RPL10a | EG_transcript_29418 | | RPL10a |
| | RPL11 | EG_transcript_31984 | EG_transcript_26203, EG_transcript_4233 | RPL11 |
| | RPL12 | EG_transcript_26978 | | RPL12 |
| | RPL13 | EG_transcript_43409 | EG_transcript_43409, EG_transcript_18309, EG_transcript_31449 | RPL13 |
| | RPL13a | EG_transcript_33082 | | RPL13a |
| | RPL14 | EG_transcript_29685 | | RPL14 |
| | RPL15 | EG_transcript_49234 | EG_transcript_54111 | RPL15 |
| | RPL16 | EG_transcript_26376 | | RPL16 |
| | RPL17 | EG_transcript_45600 | EG_transcript_10319, EG_transcript_26134 | RPL17 |
| | RPL18 | EG_transcript_45000 | | RPL18 |
| | RPL18a | EG_transcript_41078 | | RPL18a |
| | RPL18b | | | |
| | RPL19 | EG_transcript_23570 | | RPL19 |
| | RPL2 | EG_transcript_25307 | | RPL2 |
| | RPL21e | EG_transcript_36505 | | RPL21e |
| | RPL22 | EG_transcript_35834 | | RPL22 |
| | RPL23 | EG_transcript_56253 | | RPL23 |
| | RPL23a | EG_transcript_42001 | | RPL23a |
| | RPL24 | EG_transcript_30802 | EG_transcript_63500 | RPL24 |
| | RPL25 | EG_transcript_42001 | | RPL25 |
| | RPL26 | EG_transcript_48931 | | RPL26 |
| | RPL27 | | | |
| | RPL27a | EG_transcript_34772 | EG_transcript_31697, EG_transcript_71017 | RPL27a |
| | RPL28 | EG_transcript_34772 | EG_transcript_31697, EG_transcript_71017 | RPL28 |
| | RPL29 | EG_transcript_34772 | EG_transcript_31697, EG_transcript_71017 | RPL29 |
| | RPL3 | EG_transcript_14730 | EG_transcript_18974 | RPL3 |

| | | | | |
|---|---|---|---|---|
| | RPL30 | | | RPL30 |
| | RPL32 | | | |
| | RPL34 | | | |
| | RPL35 | EG_transcript_52389 | | RPL35 |
| | RPL35a | | | |
| | RPL36 | | | |
| | RPL37 | EG_transcript_32885 | | RPL37 |
| | RPL37a | | | |
| | RPL38 | | | |
| | RPL39 | | | |
| | RPL4 | EG_transcript_15379 | EG_transcript_15615 | RPL4 |
| | RPL40 | EG_transcript_12396 | EG_transcript_35776, EG_transcript_45651, EG_transcript_55709 | RPL40 |
| | RPL44 | EG_transcript_56296 | | RPL44 |
| | RPL5 | EG_transcript_22286 | | RPL5 |
| | RPL51 | EG_transcript_30305 | | RPL51 |
| | RPL6 | EG_transcript_34325 | | RPL6 |
| | RPL7 | EG_transcript_30542 | | RPL7 |
| | RPL7a | EG_transcript_26369 | | RPL7a |
| | RPL7Ae | EG_transcript_23482 | | RPL7Ae |
| | RPL9 | EG_transcript_35793 | | RPL9 |
| | RPL-like | EG_transcript_42748 | | RPL-like |
| | RPP0 | EG_transcript_54363 | EG_transcript_17739 | RPP0 |
| | RPP0b | | | |
| | RPP0c | | | |
| | RPP2 | EG_transcript_61646 | EG_transcript_50867 | RPP2 |
| SSU (40S) | RPS10 | EG_transcript_38985 | EG_transcript_38627, EG_transcript_17391 | RPS10 |
| | RPS11 | EG_transcript_40456 | EG_transcript_68512 | RPS11 |
| | RPS12 | | | RPS12 |
| | RPS13 | EG_transcript_33850 | | RPS13 |
| | RPS14 | | | RPS14 |

| | | | | |
|---|---|---|---|---|
| | RPS15 | EG_transcript_43430 | EG_transcript_37274 | RPS15 |
| | RPS15a | EG_transcript_52704 | | RPS15a |
| | RPS16 | EG_transcript_43548 | | RPS16 |
| | RPS17 | | | |
| | RPS18 | | | |
| | RPS19 | EG_transcript_41111 | | RPS19 |
| | RPS2 | EG_transcript_23458 | EG_transcript_10203 | RPS2 |
| | RPS21 | EG_transcript_31840 | | RPS21 |
| | RPS23 | EG_transcript_41429 | | RPS23 |
| | RPS24E | EG_transcript_59714 | | RPS24E |
| | RPS25 | EG_transcript_53666 | | RPS25 |
| | RPS26 | | | |
| | RPS27 | EG_transcript_52992 | | RPS27 |
| | RPS27a | EG_transcript_12396 | | RPS27a |
| | RPS3 | EG_transcript_34746 | | RPS3 |
| | RPS30 | | | |
| | RPS33 | EG_transcript_26209 | | RPS33 |
| | RPS3a | EG_transcript_26209 | | RPS3a |
| | RPS4 | EG_transcript_25958 | | RPS4 |
| | RPS5 | EG_transcript_50540 | | RPS5 |
| | RPS6 | EG_transcript_34339 | EG_transcript_33134 | RPS6 |
| | RPS7 | EG_transcript_30512 | EG_transcript_50050 | RPS7 |
| | RPS8 | EG_transcript_19137 | EG_transcript_24975, EG_transcript_27678 | RPS8 |
| | RPS9 | EG_transcript_51504 | | RPS9 |
| | RPSa | EG_transcript_32698 | | RPSa |
| Class I (tRNA synthetase) | alanyl-tRNA synthetase | EG_transcript_3246 | EG_transcript_1083 | alanyl-tRNA synthetase |
| | antigenic protein | EG_transcript_2790 | | antigenic protein |
| | arginyl-tRNA synthetase | EG_transcript_2535 | EG_transcript_3489 | arginyl-tRNA synthetase |
| | asparagine | | | |

| | | | | |
|---|---|---|---|---|
| | synthetase a | | | |
| | asparaginyl-tRNA synthetase | EG_transcript_2696 | | asparaginyl-tRNA synthetase |
| | aspartyl-tRNA synthetase | EG_transcript_3395 | EG_transcript_7672 | aspartyl-tRNA synthetase |
| | bifunctional aminoacyl-tRNA synthetase | EG_transcript_4109 | EG_transcript_9052 | bifunctional aminoacyl-tRNA synthetase |
| | Biotin--acetyl-CoA-carboxylase ligase | EG_transcript_5040 | | Biotin--acetyl-CoA-carboxylase ligase |
| | cysteinyl-tRNA synthetase | EG_transcript_11617 | EG_transcript_24982, EG_transcript_3436 | cysteinyl-tRNA synthetase |
| | D-tyrosyl-tRNA deacylase | EG_transcript_26133 | EG_transcript_28837 | D-tyrosyl-tRNA deacylase |
| | glutaminyl-tRNA synthetase | | | |
| | glutamyl-tRNA synthetase | EG_transcript_5201 | | glutamyl-tRNA synthetase |
| | glycyl-tRNA synthetase | | | glycyl-tRNA synthetase |
| | histidyl-tRNA synthetase | EG_transcript_10079 | EG_transcript_7879 | histidyl-tRNA synthetase |
| | isoleucyl-tRNA synthetase, putative (IleRS) | EG_transcript_971 | | isoleucyl-tRNA synthetase, putative (IleRS) |
| | leucyl-tRNA synthetase | EG_transcript_1177 | | leucyl-tRNA synthetase |
| | lipoate-protein ligase | EG_transcript_29972 | | lipoate-protein ligase |
| | lysyl-tRNA synthetase | EG_transcript_14094 | EG_transcript_3471 | lysyl-tRNA synthetase |
| | methionyl-tRNA | EG_transcript_22934 | EG_transcript_3873 | methionyl-tRNA |

| | | | | |
|---|---|---|---|---|
| | synthetase, putative (MetRS) | | | synthetase, putative (MetRS) |
| | Octanoyltransferase | EG_transcript_19552 | | Octanoyltransferase |
| | peptidyl-tRNA hydrolase | EG_transcript_21075 | EG_transcript_31253, EG_transcript_33263, EG_transcript_34638 | peptidyl-tRNA hydrolase |
| | phenylalanyl-tRNA synthetase (beta subunit) | EG_transcript_10635 | | phenylalanyl-tRNA synthetase (beta subunit) |
| | phenylalanyl-tRNA synthetase alpha chain | EG_transcript_7994 | | phenylalanyl-tRNA synthetase alpha chain |
| | RF-1 domain containing protein | EG_transcript_29168 | | RF-1 domain containing protein |
| | selenophosphate synthetase, putative (SPS2) | EG_transcript_17462 | | selenophosphate synthetase, putative (SPS2) |
| | seryl-tRNA synthetase | EG_transcript_26096 | | seryl-tRNA synthetase |
| | Threonyl and Alanyl-tRNA synthetase | | | |
| | threonyl-tRNA synthetase | EG_transcript_2180 | EG_transcript_4373 | threonyl-tRNA synthetase |
| | tryptophanyl-tRNA synthetase | EG_transcript_17328 | EG_transcript_7320, EG_transcript_7650 | tryptophanyl-tRNA synthetase |
| | tyrosyl/methionyl-tRNA synthetase | EG_transcript_22934 | | tyrosyl/methionyl-tRNA synthetase |
| | tyrosyl-tRNA synthetase | EG_transcript_11394 | EG_transcript_26133, EG_transcript_27695, EG_transcript_28837 | tyrosyl-tRNA synthetase |
| | valyl-tRNA synthetase, | EG_transcript_1453 | EG_transcript_1687 | valyl-tRNA synthetase, putative (ValRS) |

| | | | | |
|---|---|---|---|---|
| | putative (ValRS) | | | |
| Class II (tRNA-ligases) | Alanine--tRNA ligase | | | |
| | Asparagine--tRNA ligase, cytoplasmic | | | |
| | asparaginyl-tRNA synthetase 2 | | | |
| | Bifunctional glutamate/proline--tRNA ligase | | | |
| | Glycine--tRNA ligase | EG_transcript_3792 | EG_transcript_4976 | Glycine--tRNA ligase |
| | Lysine--tRNA ligase | EG_transcript_21843 | | Lysine--tRNA ligase |
| | Lysine--tRNA ligase, cytoplasmic | | | |
| | Lysine--tRNA ligase, mitochondrial | | | |
| | Phenylalanine--tRNA ligase alpha subunit | | | |
| | Phenylalanine--tRNA ligase beta subunit | EG_transcript_31175 | | Phenylalanine--tRNA ligase beta subunit |
| | Tyrosine--tRNA ligase, cytoplasmic | EG_transcript_17808 | | Tyrosine--tRNA ligase, cytoplasmic |
| | Phenylalanine--tRNA ligase, mitochondrial | EG_transcript_8464 | EG_transcript_66085 | Phenylalanine--tRNA ligase, mitochondrial |
| | Serine--tRNA ligase | EG_transcript_7877 | | Serine--tRNA ligase |

| | | | | |
|---|---|---|---|---|
| | Serine--tRNA ligase, mitochondrial | | | |
| | Cysteine--tRNA ligase | EG_transcript_8663 | | Cysteine--tRNA ligase |
| | Glutamate--tRNA ligase | EG_transcript_4918 | | Glutamate--tRNA ligase |
| | Glutamine--tRNA ligase | EG_transcript_4663 | EG_transcript_6722 | Glutamine--tRNA ligase |
| | Leucine--tRNA ligase | EG_transcript_1718 | | Leucine--tRNA ligase |
| | Isoleucine--tRNA ligase, cytoplasmic | EG_transcript_1330 | | Isoleucine--tRNA ligase, cytoplasmic |
| | Methionine--tRNA ligase | EG_transcript_1145 | | Methionine--tRNA ligase |
| | Threonine--tRNA ligase | | | |
| Initiation Factors | eIF2a | EG_transcript_9985 | | eIF2a |
| | eIF2B | EG_transcript_9484 | | eIF2B |
| | eIF2B-alpha | EG_transcript_6440 | | eIF2B-alpha |
| | eIF2B-beta | EG_transcript_24882 | | eIF2B-beta |
| | eIF2B-delta | EG_transcript_9484 | | eIF2B-delta |
| | eIF2B-epsilon | EG_transcript_5123 | | eIF2B-epsilon |
| | eIF2G | EG_transcript_10211 | | eIF2G |
| | eIF4a | EG_transcript_10211 | EG_transcript_6824, EG_transcript_8282, EG_transcript_9065, EG_transcript_23012, EG_transcript_52954 | eIF4a |
| | eIF4e (1 - 6) | EG_transcript_23493 | EG_transcript_5504, EG_transcript_9983, EG_transcript_12176, EG_transcript_12824, EG_transcript_13797, EG_transcript_14547, EG_transcript_20040, EG_transcript_20943, EG_transcript_23493, EG_transcript_27095 | eIF4e (1 - 6) |

| | | | | |
|---|---|---|---|---|
| | eIF4G1 | EG_transcript_11223 | | eIF4G1 |
| | eIF4G2 | EG_transcript_10211 | | eIF4G2 |
| | eIF4G3 | EG_transcript_4306 | | eIF4G3 |
| | eIF4G4 | EG_transcript_4306 | EG_transcript_735 | eIF4G4 |
| | eIF4G5 | EG_transcript_4306 | | eIF4G5 |
| | K1 | EG_transcript_388 | | K1 |
| | K2 | EG_transcript_1834 | EG_transcript_19804 | K2 |
| | K3 | EG_transcript_1834 | | K3 |
| | PAPB1 | EG_transcript_3151 | EG_transcript_12193, EG_transcript_21222, EG_transcript_21345, EG_transcript_21408, EG_transcript_2280, EG_transcript_23784, EG_transcript_43186, EG_transcript_45856, EG_transcript_5117, EG_transcript_5441, EG_transcript_9154 | PAPB1 |
| | PAPB2 | EG_transcript_19503 | EG_transcript_4302 | PAPB2 |
| Elongation Factors (EF) | EF-1-alpha (TEF1a) | EG_transcript_12357 | EG_transcript_18855, EG_transcript_45355, EG_transcript_5617 | EF-1-alpha (TEF1a) |
| | EF-1-beta (TEF1b) | EG_transcript_64722 | EG_transcript_20319 | EF-1-beta (TEF1b) |
| | EF-1-gamma (TEF1g) | EG_transcript_11223 | | EF-1-gamma (TEF1g) |
| | EF-2 (TEF-2) | EG_transcript_5039 | EG_transcript_51028 | EF-2 (TEF-2) |
| | GTP-binding EF-Tu family | EG_transcript_6874 | EG_transcript_8722, EG_transcript_1495 | GTP-binding EF-Tu family |
| | selenocysteine-tRNA-specific EF | EG_transcript_6473 | | selenocysteine-tRNA-specific EF |
| Termination Release Factors | eukaryotic peptide chain release factor subunit 1, putative (ERF1) | EG_transcript_5686 | | eukaryotic peptide chain release factor subunit 1, putative (ERF1) |
| | eukaryotic release factor 3, putative (ERF3) | EG_transcript_4245 | | eukaryotic release factor 3, putative (ERF3) |

| | | | | |
|---|---|---|---|---|
| | RF-1 domain containing protein | EG_transcript_6340 | | RF-1 domain containing protein |
| Recycling | RRF1 | | | |

**Note:** The table shows the sets of proteins for functional translational apparatus inventoried in *E. gracilis*. Column 1 = Reference sub family. Column 2 = Gene convention. Column 3 = *E. gracilis* Reference sequence ID. Column 4 = Paralogous sequence ID. Column 4 = Final classifications. Green coloured cell and corresponding text means the presence of the corresponding protein and the protein name respectively. Empty white cells in columns means that the corresponding protein /process was not found or that a final classification could not be made due to insufficient statistical confidence.

*5.3.2 Pre-initiation complex* (Transcription regulatory complexes)

Using BLAST, PSI-BLAST and OrthoFinder, I found the core transcriptional pre-initiation complex proteins (Table 5.9). Proteins found include the core promoters (TATA-binding protein), the general transcription factors (TFIIA - GTF2A1, GTF2A2; TFIIB - GTF2B; TFIID - TAF1-7, 10 – 12; and TFIIH - CDK7/cyclin H kinase complex), DNA helicase, RNA Plymerase (RPB1 and RPB2) and the Activators and Repressors (Ctk1, CDK7, CDK8, CCNC, MEDs, CCNT, and TLF). TFIIA, B –TFIID, TFIID (BTAF1, BTF3, BTF3L4, EDF1, 6L, 7L, 11L), TFIIE, TFIIF, TFIIJ, TFIIK, other members of the RNA Polymerase (RNAP), and the Activators and Repressors (CBP-1 and CTD kinases) (Maree and Patterson, 2014; Blackwell and Walker, 2006; Samorodnitsky and Pugh, 2010; Yin and Wang, 2014; Poss, *et al.,* 2013; Rhess and Pugh, 2012; Afek and Lukatsky, 2013; Lauberth, *et al.,* 2013; and Murakami, *et al.,* 2013). The absence of the proteins may be due to incomplete dataset and not due to incomplete transcriptional machinery in *E. gracilis*.

**Table 5.9:** Transcription regulatory complexes (RNA Pol II and general transcription factors) in *E. gracilis.*

| Category/complex | Gene/convention | nrBLAST evalue | Ref. Seq. ID | Paralogous Seq. ID | Final classification |
|---|---|---|---|---|---|
| | | | | | |
| Core Promoters | TATA-binding protein (TBP) | 3.00E-052 | EG_transcript_17190 | EG_transcript_16379 | TATA-binding protein (TBP) |
| TFIIA | GTF2A1 | | | | |
| | GTF2A2 | 2.00E-006 | EG_transcript_63671 | | GTF2A2 |
| TFIIB | GTF2B | 1.00E-062 | EG_transcript_21874 | | RNA polymerase III transcription factor IIIB |
| B-TFIID | BTAF1 | 0 | EG_transcript_588 | EG_transcript_11305, EG_transcript_1160, EG_transcript_15685, EG_transcript_1871, EG_transcript_2607, EG_transcript_6408, EG_transcript_744, EG_transcript_993 | |
| TFIID | BTAF1 | | | | |
| | BTF3 | 3.00E-043 | EG_transcript_32626 | | BTF3 |
| | BTF3L4 | 3.00E-043 | EG_transcript_32626 | | BTF3L4 |
| | EDF1 | 1.00E-041 | EG_transcript_25186 | | EDF1 |
| | TAF1 – 15 | 1.00E-018 | EG_transcript_31652 | EG_transcript_17472, EG_transcript_31652, EG_transcript_3972, EG_transcript_10703, EG_transcript_6811, EG_transcript_28897, EG_transcript_17038, EG_transcript_31437, EG_transcript_63637, EG_transcript_20588, EG_transcript_27352, EG_transcript_11294, EG_transcript_16079, EG_transcript_20105, EG_transcript_26855, EG_transcript_27352, EG_transcript_31447, EG_transcript_36755 | TAF1-15 (except for TAF3, TAF4, TAF10, TAF11, and TAF13) |
| | 6L | 6.00E-046 | EG_transcript_6811 | | 6L |
| | 7L | | | | |

| | | | | | |
|---|---|---|---|---|---|
| | 11L | | | | |
| TFIIE | α-subunit | 1.00E-008 | EG_transcript_18459 | | α-subunit |
| | β-subunit | 3.00E-043 | EG_transcript_32626 | | |
| TFIIF | RAP30 (large subunit) | | | | |
| | RAP74 (small subunit) | | | | |
| | | | | | |
| TFIIH | CDK7/cyclin H kinase complex | 1.00E-015 | EG_transcript_4913 | EG_transcript_24942, EG_transcript_12795, EG_transcript_17017, EG_transcript_15271, EG_transcript_9986, EG_transcript_18332, EG_transcript_27826 | Cyclin H kinase complex |
| DNA Helicase | DNA Helicase | 0 | EG_transcript_2707 | EG_transcript_1436, | RPB1 (POLR2A) |
| RNA Polymerase (RNAP) | RPB1 (POLR2A) | 0 | EG_transcript_92 | EG_transcript_26129, EG_transcript_30765, EG_transcript_40868, EG_transcript_54794 | RPB2 (POLR2B) |
| | RPB2 (POLR2B) | 0 | EG_transcript_464 | EG_transcript_1166, EG_transcript_15521, EG_transcript_20482 | RPB3 (POLR2C) |
| | RPB3 (POLR2C) | 4.00E-081 | EG_transcript_15850 | | RPB4 (POLR2D) |
| | RPB4 (POLR2D) | 2.00E-020 | EG_transcript_32084 | | RPB5 (POLR2E) |
| | RPB5 (POLR2E) | 7.00E-065 | EG_transcript_23275 | | RPB6 (POLR2F) |
| | RPB6 (POLR2F) | 4E-39 | EG_transcript_35731 | | RPABC2/RPB6 |
| | RPB7 (POLR2G) | 3.00E-050 | EG_transcript_24901 | | RPB7 (POLR2G) |

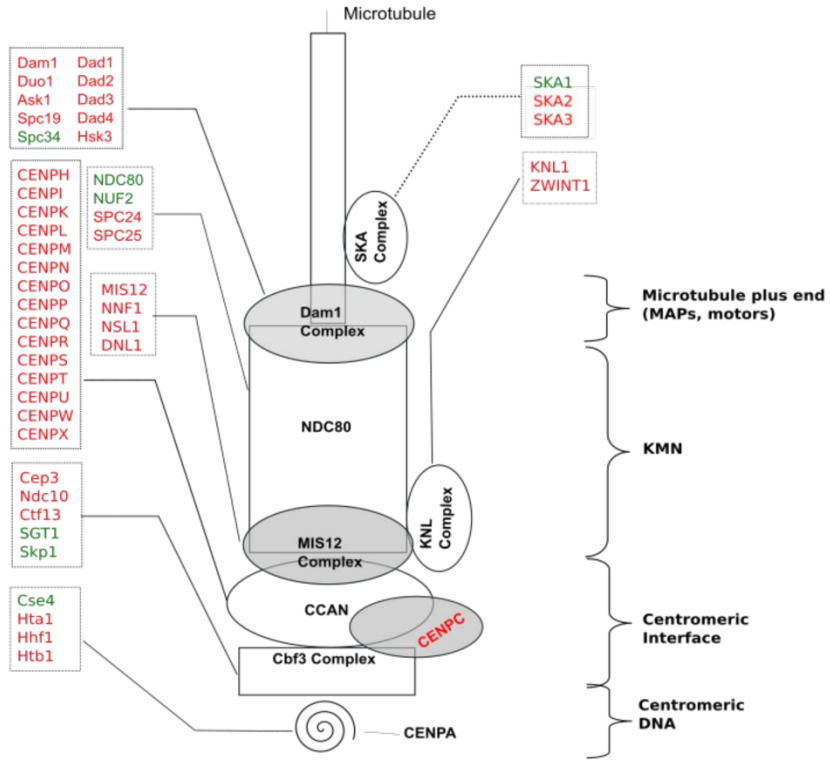
| | | | | | |
|---|---|---|---|---|---|
| | RPB8 (POLR2H) | 1.00E-033 | EG_transcript_26233 | | RPB8 (POLR2H) |
| | RPB9 (POLR2I) | 4.00E-034 | EG_transcript_26985 | | RPB9 (POLR2I) |
| | RPB10 (POLR2L) | | | | |
| | RPB11-a (POLR2J) | 3.00E-028 | EG_transcript_26129 | | RPB11-a (POLR2J) |
| | RPB11-b (POLR2J2) | | | | |
| | RPB11-c (POLR2J3) | | | | |
| | RPB12 (POLR2K) | | | | |
| | RNAPII0 | | | | |
| | RNAPIIA | | | | |
| | RNA POL I (Subunit 1) | 2.00E-101 | EG_transcript_2051 | EG_transcript_7627 | RNA POL I (Subunit 1) |
| | RNA POL III | 0 | EG_transcript_460 | | RNA POL III |
| Activators and Repressors | Ctk1 (CDK9) | 3.00E-116 | EG_transcript_20268 | | Ctk1 (CDK9) |
| | CDK7 | 6E-144 | EG_transcript_12795 | | CDK2 |
| | CDK8 | 6E-144 | EG_transcript_12795 | | |
| | cyclin C (CCNC) | 6E-144 | EG_transcript_10654 | EG_transcript_12795 | |
| | CBP-1 | | | | |
| | CTD kinases | 2.00E-097 | EG_transcript_15279 | EG_transcript_15687, EG_transcript_15863, EG_transcript_17202, EG_transcript_17838, EG_transcript_14503 | CTD kinases |
| | MED1 to MED 31 | 2.00E-011 | EG_transcript_41685 | EG_transcript_26163, EG_transcript_3741, EG_transcript_2126 | MED6, MED7, MED12 & MED14 (MED1-5, 8–11, 13, 15 - |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | 13 absent) |
| | cyclin T (CCNT) | | EG_transcript_31481 | | cyclin T/Cyclin L (CCNT) |
| | TLF | 3.00E-052 | EG_transcript_17190 | | TLF |
| | Cyclin 1 - 10 | 5.00E-144 | EG_transcript_12795 | EG_transcript_12795, EG_transcript_20268 | CDK1, CDK7, CDK8, CDK9 (CDK2,3,4,5 absent) |

**Note:** The table shows the sets of proteins for transcription regulatory complexes (RNA Pol II and general transcription factors) inventoried in *E. gracilis*. Column 1 = Reference sub family. Column 2 = Gene convention. Column 3 = nrBLAST evalue for the reference sequence. Column 4 = *E. gracilis* Reference sequence ID. Column 5 = Paralogous sequence IDs. Column 6 = Final classifications. Green coloured cell and corresponding text means the presence of the corresponding protein and the protein name respectively. Empty white cells in columns means that the corresponding protein /process was not found or that a final classification could not be made due to insufficient statistical confidence.

### *5.3.3 Kinetochores*

To identify proteins that are relevant for mitosis (conventional kinetochores), I used BLAST, PSI-BLAST and OrthoFinder clustering algorithms, using the 19 KKTs and 7 KKIPs reported in *T. brucei* (Akiyoshi and Gull, 2014; D'Archivio and Wickstead, 2016) and a selection of eukaryotic conventional kinetochores (Lampert and Westerman, 2011, Meraldi*et al.* 2006, Kitagawa and Heiter, 2001) as a reference set. Out of the 19 KKTs and 7 KKIPs present in *T. brucei*, I found three of them in the *E. gracilis* transcriptome (KKT10, KKT19, and KKIP7) with multiple copies for KKT10/19 (Figure 5.11, Table 5.10). The KKT proteins show clear homology to KKT10 than they did to KKT19. The KKTs found showed huge sequence similarity and conservation when aligned with MAFFT, however, additional analysis will be required to confirm the classification of *E. gracilis* KKT10 and KKT19. Protein families of members of the microtubule plus-end, KMN, and centromeric interface, and centromeric DNA were found. Some of the *E. gracilis* kinetochores do not show similarity to two of the conventional kinetochore complexes in other eukaryotes (Lampert and Westerman, 2011, Meraldi*et al.* 2006, Kitagawa and Heiter, 2001), however, I found members of the KMN and CCAN which are core multi-protein complexes of the kinetochores (Figure 11, Table 5.10, and Appendix-Ia). Three complexes, MIS12, CCAN, and the CENPC, were not found in our dataset, suggesting a rather incomplete dataset or truncations of sequences as no *E. gracilis* sequences showed significant similarities to these complexes. Other proteins found are those involved in microtubule attachment, regulatory and signaling checkpoints.



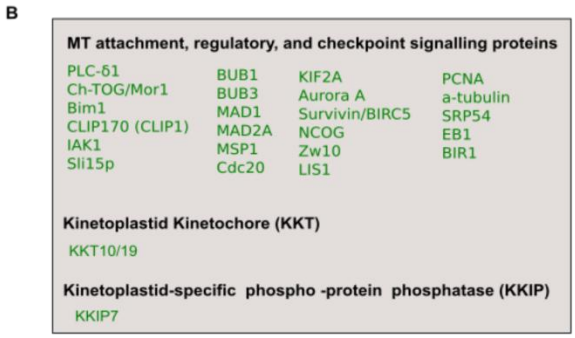




**Figure 5.11: Conventional and non conventional Kinetochores are characteristic of the *E. gracilis* Kinetochore machinery.** Architectural representation of *E. gracilis* Kinetochores. Panel A: Showing the general structure of the kinetochore highlighting the microtubule plus-end (Dam1 and SKA1), the KMN (NDC80, MIS12 complex, KNL1 complex), the centromeric interface (CCAN, CENPC, Cbf3 complex), and the centromeric DNA (CENPA). Boxes linked to these respective complexes shows the proteins associated with it. Text coloured green means presence in *E. gracilis* whilst those coloured red indicate absence in *E. gracilis*. Diagram is as previously described by Lampert and Westerman, 2011, and superimposed. Panel B: Showing the microtubule attachments, regulatory and checkpoint signalling, kinetoplastid kinetochores (KKT) and kinetoplastid-specific phosphor protein phosphatase proteins (KKIP) associated with the kinetochores. The KKTs and KKIPs are restricted to the kinetoplastids but some sub families (KKT10/19 and KKIP7 respectively) are also found in *E. gracilis*. Conventions and classifications are as designated by Lampert and Westerman, 2011, Meraldi*et al.* 2006, Kitagawa and Heiter, 2001, Akiyoshi and Gull, 2014 (Kinetoplastid Kinetochores, KKT), D'Archivio and Wickstead, 2016 (Kinetoplastid-specific phosphor protein phosphatase, KKIP).

**Table 5.10:** Kinetochore and it's protein composition in *E. gracilis*.

| Kinetochore complex | Baker's yeast proteins | H. sapiens proteins | nrBLAST evalue | Ref. Seq. ID | Paralogous Seq. ID | Final classification |
|---|---|---|---|---|---|---|
| **Ndc80 complex** | Ndc80p | NDC80 (Hec1) | 2.00E-016 | EG_transcript_7526 | N/A | NDC80 |
| | Spc24p | SPC24 | | | | |
| | Spc25p | SPC25 | 2.00E-016 | EG_transcript_7526 | N/A | |
| | Nuf2p | NUF2R | 2.00E-018 | EG_transcript_11202 | N/A | NUF2 |
| **MIS12 (Mtw12)** | Mtw1p | Mis12 | | | | |
| | Nnf1p | NNF1 | | | | |
| | Nsl1p | NSL1 | | | | |
| | Dsn1p | DSN1 | | | | |
| **Spc105 complex** | Spc105p | KNL1 (Blinkin) | | | | |
| | Ydr532p | ZWINT1 | | | | |
| **CENPA** | Cse4p | CENPA | 1.00E-085 | EG_transcript_21219 | EG_transcript_21219, EG_transcript_27533, EG_transcript_29591, EG_transcript_30647, EG_transcript_33619, EG_transcript_64233 | Histone H3 |
| | Cbf1p | *T*FE3 | 1.00E-074 | EG_transcript_3336 | N/A | |
| **CENPC** | Mif2p | CENPC | | | | |
| **CCAN (Human)** | | CENPK | | | | |
| | | CENPL | | | | |
| | | CENPM | | | | |
| | | CENPN | | | | |
| | | CENPO | | | | |
| | | CENPP | | | | |
| | | CENPQ | | | | |
| | | CENPR | | | | |
| | | CENPS | | | | |
| | | CENPT | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | CENPU | | | | |
| | | CENPW | | | | |
| | | CENPX | | | | |
| | | CENPB | | | | |
| | | CENPE | | | | |
| **CCAN (Yeast)** | Ctf19p | N/A | | | | |
| | Okp1p | CENPF | | | | |
| | Mcm21p | Mal2 | | | | |
| | Ame1p | N/A | | | | |
| | Ctf3p | CENPI | | | | |
| | Plc | PLC-δ1 | 2E-83 | EG_transcript_1338 | N/A | PLC-δ1 |
| | Mcm16p | | | | | |
| | Mcm22p | | | | | |
| | Chl4p | | | | | |
| | Iml3p | | | | | |
| | Nkp1p | | | | | |
| | Nkp2p | | | | | |
| | Cnn1p | | | | | |
| | Slk19p | CENPF | | | | |
| | Sim4 | | | | | |
| | Fta1 | | | | | |
| | Fta3 | CENPH | | | | |
| | Fta4 | | | | | |
| **DASH (Dam1) complex** | Ask1p | N/A | | | | |
| | Dam1p | N/A | | | | |
| | Dad1p | N/A | | | | |
| | Dad2p | N/A | | | | |
| | Dad3p | N/A | | | | |
| | Dad4p | N/A | | | | |
| | Duo1p | N/A | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Hsk3p | N/A | | | | |
| | Spc19p | N/A | | | | |
| | Spc34p | N/A | 8.00E-124 | EG_transcript_9981 | N/A | Spc34 |
| **SKA Complex** | N/A | SKA1 | 4.00E-036 | EG_transcript_17333 | EG_transcript_50775 | SKA1 |
| | N/A | SKA2 | | | | |
| | N/A | SKA3 | | | | |
| **CBF3** | Ndc10p | N/A | | | | |
| | Ctf13p | N/A | | | | |
| | Cep3p | N/A | | | | |
| | Sgt1p | SGT1 | 8.00E-061 | EG_transcript_13576 | EG_transcript_13576, EG_transcript_16816, EG_transcript_17578, EG_transcript_22504, EG_transcript_22585, EG_transcript_24612, EG_transcript_39776 | Protein SGT1 |
| | Skp1p | P19/SKP1 | 2.00E-085 | EG_transcript_13383 | N/A | Skp1 |
| **Regulatory proteins (Sli 15 complex)** | Ipl1p | IAK1 | 0 | EG_transcript_13984 | N/A | Ipl1 protien kinase |
| | Sli15p | | 5.00E-095 | EG_transcript_1470 | N/A | Sli15 |
| | Borealin | Borealin | | | | |
| | Bir1p | BIR1 | 6.00E-011 | EG_transcript_4348 | N/A | Bir1 |
| | Bub1p | BUB1 | 6.00E-029 | EG_transcript_4679 | EG_transcript_4679, EG_transcript_15337, | Bub1 |
| | Bub3p | BUB3 | 3.00E-091 | EG_transcript_15004 | N/A | Bub3 |
| | Mad1p | MAD1 | 2.00E-012 | EG_transcript_3089 | N/A | MAD1 |
| | Mad2p | MAD2 (MAD2L1) | 7.00E-094 | EG_transcript_22781 | N/A | MAD2A |
| | Mad3p | MAD3 (BUB1R) | 6.00E-029 | EG_transcript_15337 | N/A | Bub1 |
| | Mps1p | MPS1 (PYT/TTK1) | 1.00E-075 | EG_transcript_10874 | EG_transcript_10874, EG_transcript_24615 | Msp1 |
| | Cdc20 | Cdc20 | 3.00E-158 | EG_transcript_8342 | N/A | Cdc20 |
| **MT attachment,** | Plc | PLC-δ1 | 2E-83 | EG_transcript_1338 | N/A | PLC-δ1 |
| | Stu2p | Ch-TOG | 3.00E-148 | EG_transcript_458 | N/A | MOR1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **regulatory, and checkpoint signalling proteins** | Bim1p | | 2.00E-054 | EG_transcript_24591 | N/A | Bim1 |
| | Bik1p | CLIP170 (CLIP1) | 8.00E-063 | EG_transcript_1988 | N/A | Bik1 |
| | | KIF2A | 1E-62 | EG_transcript_2827 | | |
| | | KIF2B | | | | |
| | | KIF2C/MCAK | | | | |
| | | INCENP | | | | |
| | | Aurora B | | | | |
| | | Aurora A | 0 | EG_transcript_13984 | N/A | Aurora A |
| | | Survivin/BIRC5 | 3.00E-017 | EG_transcript_18602 | EG_transcript_22643 | BIRC5 |
| | | ICIS/MTUS1 | | | | |
| | | ICIS/MTUS2 | | | | |
| | Glc7 | PP1 (PPP1CC) | 0 | EG_transcript_6968 | N/A | PP1 (PPP1CC) |
| | | Zw10 | 2E-21 | EG_transcript_7524 | N/A | Zw10 |
| | | Zwint-1 | | | | |
| | | Rod (CCHCR1) | | | | |
| | | LIS1 (PAFAH1B1) | 2E-126 | EG_transcript_11974 | N/A | LIS1 (PAFAH1B1) |
| | | CLASP1 | 3E-18 | EG_transcript_1303 | | |
| | | CLASP2 | | | | |
| | | CLIP-170 (CLIP1) | 8E-63 | EG_transcript_1988 | | |
| | | EB1 (MAPRE1) | 2E-54 | EG_transcript_24591 | N/A | EB1 (MAPRE1) |
| | PCNA | N/A | 6.00E-124 | EG_transcript_24437 | N/A | PCNA |
| | a-tubulin | a-tubulin | 0 | EG_transcript_10562 | N/A | a-tubulin |
| | Srp54p | SRP54 | 0 | EG_transcript_8662 | N/A | SRP54 |
| | | | | | | |
| | | | | | | |
| | ***T. brucei*** | | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Kinetoplastid Kinetochore (KKT)** | KKT1 | | | | | |
| | KKT2 | | 3.00E-108 | EG_transcript_16061 | | Protein kinase |
| | KKT3 | | 2.00E-144 | EG_transcript_14276 | | polo-like protein kinase |
| | KKT4 | | 4.00E-083 | EG_transcript_6409 | | Ankyrin repeat protein |
| | KKT5 | | | | | |
| | KKT6 | | | | | |
| | KKT7 | | | | | |
| | KKT8 | | | | | |
| | KKT9 | | | | | |
| | KKT10 | | 7.00E-090 | EG_transcript_18328 | EG_transcript_18328, EG_transcript_13506, EG_transcript_18595 | KKT10, putative |
| | KKT11 | | | | | |
| | KKT12 | | | | | |
| | KKT13 | | 5.00E-012 | EG_transcript_11775 | | |
| | KKT14 | | | | | |
| | KKT15 | | | | | |
| | KKT16 | | | | | |
| | KKT17 | | | | | |
| | KKT18 | | | | | |
| | KKT19 | | 1.00E-090 | EG_transcript_18328 | EG_transcript_24137 | KKT10/KKT19, putative |
| **Kinetoplastid-specific phospho-protein phosphatase (KKIP)** | KKIP1 | | | | | |
| | KKIP2 | | | | | |
| | KKIP3 | | | | | |
| | KKIP4 | | | | | |
| | KKIP5 | | | | | |
| | KKIP6 | | | | | |
| | KKIP7 | | 8.00E-101 | EG_transcript_3659 | EG_transcript_5075, EG_transcript_9673 | KKIP7 putative |

**Note:** The table shows the protein composition of the *E. gracilis* kinetochore. Column 1 = Reference sub family. Column 2 = Gene convention. Column 3 = nrBLAST evalue for the reference sequence. Column 4 = *E. gracilis* Reference sequence ID. Column 5 = Paralogous sequence ID. Column 6 = Final classifications. Green coloured cell and corresponding text means the presence of the corresponding protein and the protein name respectively. Empty white cells in column 3 and 4 means that the corresponding protein /process was not found or that a final classification could not be made due to insufficient statistical confidence.

### *5.3.4 mRNA metabolism*

I analyzed the *E. gracilis* transcriptome for mRNA metabolism gene candidates: DHH1, SCD6, XRNA, XRNB, XRNC, XRND, NOT1, and UPF1. Using a BLAST, PSI-BLAST and OrthoFinder searches of reference *T. brucei* mRNA metabolism genes, I found the entire mRNA metabolism proteins (except XRNC) from the reference set with moderate sequence similarity and amino acid conservation (Table 5.11).

**Table 5.11:** RNA metabolism proteins in *E. gracilis*.

| Category/complex | Gene/convention | nrBLAST evalue | Ref. Seq. ID | Paralogous Seq. ID | Final classification |
|---|---|---|---|---|---|
| DHH1 | DHH1 | 0.0 | | EG_transcript_5941 | DHH1 |
| SCD6 | SCD6 | 2E-26 | EG_transcript_4567 | EG_transcript_7661, EG_transcript_16210, EG_transcript_16290, EG_transcript_27759 | SCD6 |
| XRNA | XRNA | 0 | EG_transcript_425 | EG_transcript_16941, EG_transcript_4369 | XRNA |
| XRNB | XRNB | 5.00E-115 | EG_transcript_3595 | EG_transcript_3555, EG_transcript_5287, EG_transcript_3611 | XRNB |
| XRNC | XRNC | | | | |
| XRND | XRND | 9.00E-104 | EG_transcript_2132 | EG_transcript_2268 | XRND |
| NOT1 | NOT1 | 4E-148 | EG_transcript_467 | | NOT1 |
| UPF | UPF | 0 | EG_transcript_1144 | | UPF1 |

**Note:** The table shows the protein composition of the *E. gracilis* kinetochore. Column 1 = Reference sub family. Column 2 = Gene convention. Column 3 = nrBLAST evalue for reference sequence. Column 4 = *E. gracilis* Reference sequence ID. Column 5 = Paralogous sequence ID. Column 6 = Final classifications. Green coloured cell and corresponding text means the presence of the corresponding protein and the protein name respectively. Empty white cells in column 3 and 4 means that the corresponding protein /process was not found or that a final classification could not be made due to insufficient statistical confidence.

### *5.3.5 Exosomes*

Exosome, which are 40-100 nm vesicles of endocytic origin that are formed within multivesicular bodies (MVBs) (Eliaz, *et al.,* 2017), were identified in the *E. gracilis* proteome using *T. brucei* reference sets. Exosomes found include: RRP41A, RRP41B, RRP45, EAP4, RRP4, RRP40, DIS3-L, CSL4/5, RRP44, and RRP6. EAP1 - 3 were not found (Table 5.12).

**Table 5.12**: Exosome proteins present in *E. gracilis*.

| Category/complex | Gene/convention | nrBLAST evalue | Ref. Seq. ID | Paralogous Seq. ID | Final classification |
|---|---|---|---|---|---|
| RRP41A | RRP41A | 7E-78 | EG_transcript_19722 | | RRP41A |
| RRP41B | RRP41B | 6.00E-033 | EG_transcript_29615 | | RRP46 |
| RRP45 | RRP45 | 3E-76 | EG_transcript_12341 | EG_transcript_22407 | RRP45 |
| EAP1 | EAP1 | | | | |
| EAP2 | EAP2 | 3E-76 | EG_transcript_12341 | | |
| EAP3 | EAP3 | | | | |
| EAP4 | EAP4 | 5E-48 | EG_transcript_29813 | | EAP2 |
| RRP4 | RRP4 | 3E-51 | EG_transcript_18024 | | RRP4 |
| RRP40 | RRP40 | 5E-49 | EG_transcript_23635 | | RRP40 |
| CSL4 | CSL4 | 1E-50 | EG_transcript_24025 | | CSL5 |
| DIS3-L | DIS3-L | 9E-113 | EG_transcript_1172 | | DIS3-L |
| RRP44 | RRP44 | 0 | EG_transcript_1783 | | RRP44 |
| RRP6 | RRP6 | 7E-115 | EG_transcript_4586 | | RRP6 |

**Note:** The table shows the Exosome proteins found in *E. gracilis*. Column 1 = Reference sub family. Column 2 = Gene convention. Column 3 = nrBLAST evalue for reference sequence. Column 4 = *E. gracilis* Reference sequence ID. Column 5 = Paralogous sequence ID. Column 6 = Final classifications. Green coloured cell and corresponding text means the presence of the corresponding protein and the protein name respectively. Empty white cells in column 3 and 4 means that the corresponding protein /process was not found or that a final classification could not be made due to insufficient statistical confidence.

### *5.3.6 Spliceosomes and related proteins*

Spliceosomal and related proteins were predicted using BLAST and the OrthoFinder algorithm. The reference protein sequences are as contained in Yu, *et al.,* 2011 and Preußer, *et al.,* 2012. All subfamilies of the spliceosomal and related proteins were found except for the hnRNPs which I could not find (Table 5.13). Some specific proteins within the subfamilies were not found. For example, proteins missing are Sm D2, Sm N, SF3b145 (U2) U11/U12, U4atac, U6atac, Prp19 (fSap33, PRP19, CRN, SYF1) sub families. Subfamilies found include: SNRNPs (core proteins, major U2 type snRNP, minor U2 type snRNP, Tri-snRNP), non-snRNP (SR, PRP19 complex proteins, catalytic step II and late acting proteins, EJC, others with known motifs, others without know motifs), hnRNP proteins associated with the spliceosomes, and additional splicing regulators (Table 5.13).

**Table 5.13**: Spliceosomal proteins (snRNP and Trans-spliceosomes) and related factors in *E. gracilis*.

| Category/complex | Gene/convention | Ref. Seq. ID | Paralogous Seq. ID | Final classification |
|---|---|---|---|---|
| Sm/Lsm | Sm B/B' | EG_transcript_22228 | EG_transcript_31945 | Sm B/B' |
| | Sm D1 | EG_transcript_36320 | | Sm D1 |
| | Sm D2 | | | |
| | Sm D3 | EG_transcript_19730 | | Sm D3 |
| | Sm F | EG_transcript_49700 | | Sm F |
| | Sm N | | | |
| | LSm2 | EG_transcript_32750 | | LSm2 |
| | LSm4 | EG_transcript_43397 | | LSm4 |
| | LSm5 | EG_transcript_34490 | | LSm5 |
| U1 | U1-70K | EG_transcript_11874 | | U1-70K |
| | CROP | EG_transcript_16849 | | CROP |
| | U1 A | EG_transcript_54080 | | U1 A |
| | U1 C | EG_transcript_32896 | EG_transcript_1536 | U1 C |
| | Fnbp3 (Prp40) | EG_transcript_27104 | EG_transcript_15196 | Fnbp3 (Prp40) |
| U2 | U2 A' | EG_transcript_22467 | EG_transcript_10571, EG_transcript_13248, EG_transcript_16069, EG_transcript_16828, EG_transcript_18804, EG_transcript_20962, EG_transcript_22467, EG_transcript_22868, EG_transcript_7776 | U2 A' |
| | U2 B" | EG_transcript_4302 | | U2 B" |
| | SF3a120 (Prp21) | EG_transcript_12396 | EG_transcript_35518, EG_transcript_35776, EG_transcript_45651 | SF3a120 (Prp21) |

| | | | | |
|---|---|---|---|---|
| | SF3a66 (Prp11p) | EG_transcript_8543 | | SF3a66 (Prp11p) |
| | SF3b155 (Hsh155p) | EG_transcript_34272 | | SF3b155 (Hsh155p) |
| | SF3b145 (Cus1p) | | | |
| | SF3b130 (Rse1p) | EG_transcript_1075 | | SF3b130 (Rse1p) |
| | SF3b49 (Hsh49p) | EG_transcript_40823 | EG_transcript_17860 | SF3b49 (Hsh49p) |
| | SF3b10 | EG_transcript_33121 | EG_transcript_5552 | SF3b10 |
| | SF3b14a | | | SF3b14a |
| | SF3b14b Rds3p) | EG_transcript_32818 | | SF3b14b Rds3p) |
| U5 | U5-220 kDa | EG_transcript_261 | EG_transcript_12019, EG_transcript_20164, EG_transcript_21910, EG_transcript_33457, EG_transcript_42349, EG_transcript_5130, EG_transcript_5786, | U5-220 kDa |

| | | | | |
|---|---|---|---|---|
| | | | EG_transcript_6032, EG_transcript_6256, EG_transcript_6622, EG_transcript_8280 | |
| | U5-200 kDa | EG_transcript_141 | EG_transcript_261 | U5-200 kDa |
| | U5-116 kDa | EG_transcript_1536 | EG_transcript_59471 | U5-116 kDa |
| | U5-102 kDa | EG_transcript_2027 | | U5-102 kDa |
| | U5-100 kDa | | EG_transcript_33457, EG_transcript_5130 | U5-100 kDa |
| | U5-40 kDa | EG_transcript_42349 | | U5-40 kDa |
| | U5-15 kDa | EG_transcript_32934 | | U5-15 kDa |
| | U5-52 kDa | EG_transcript_23482 | | U5-52 kDa |
| U4/U6 | U4/U6-90 kDa | EG_transcript_8282 | EG_transcript_12249, EG_transcript_13536, EG_transcript_47952, EG_transcript_7007 | U4/U6-90 kDa |
| | U4/U6-60 kDa (Prp4) | EG_transcript_6069 | | U4/U6-60 kDa (Prp4) |
| | U4/U6-61 kDa (Prp31) | EG_transcript_11058 | EG_transcript_8570 | U4/U6-61 kDa (Prp31) |
| | U4/U6-20 kDa | EG_transcript_21220 | | U4/U6-20 kDa |
| | U4/U6-15.5 kDa | EG_transcript_34339 | | U4/U6-15.5 kDa |
| U11/U12 | U11-25 | | | |

| | | | | |
|---|---|---|---|---|
| | U11-35 | | | |
| | U11-48 | | | |
| | U11-59 | | | |
| | U11/U12-20 | | | |
| | U11/U12-31 | | | |
| | U11/U12-65 | | | |
| | C114 | | | |
| | YB1 | | | |
| | Toe-1 | | | |
| U4atac | U4atac | | | |
| U6atac | U6atac | | | |
| U4/U6.U5 tri-snRNP-specific | Tri-snRNP 110 kDa | EG_transcript_7210 | | Tri-snRNP 110 kDa |
| | Tri-snRNP 65 kDa | EG_transcript_31193 | | Tri-snRNP 65 kDa |
| | Tri-snRNP 27 kDa | EG_transcript_29809 | | Tri-snRNP 27 kDa |
| SR | 9G8 | EG_transcript_20514 | | 9G8 |
| | Tra2-beta | EG_transcript_21345 | | Tra2-beta |
| | Tra2-alpha | EG_transcript_21345 | | Tra2-alpha |
| Prp19 complex | CDC5 | EG_transcript_13535 | | CDC5 |
| | PRP5 | EG_transcript_2790 | | PRP5 |

| | | | | |
|---|---|---|---|---|
| | PRP46 | EG_transcript_12337 | | PRP46 |
| | fSap33 | | | |
| | PRP19 | | | |
| | CRN | | | |
| | SYF1 | | | |
| Catalytic step II & late-acting | Prp22 (DHX8) | EG_transcript_5745 | | Prp22 (DHX8) |
| | Prp43 | EG_transcript_3381 | | Prp43 |
| | Prp16 | EG_transcript_2127 | | Prp16 |
| | MCG9280 | | | MCG9280 |
| | Prp17 | | | Prp17 |
| | Prp18 | EG_transcript_17464 | | Prp18 |
| Exon junction complex (EJC) | SRm160 | EG_transcript_19739 | EG_transcript_3865, EG_transcript_5276 | SRm160 |
| | UAP56 | EG_transcript_10531 | | UAP56 |
| | LDC2 | EG_transcript_29944 | | LDC2 |
| | Y14 | EG_transcript_35526 | | Y14 |
| | Magoh | EG_transcript_32711 | | Magoh |
| DexD/H | DDX3 | | | |
| | DBY | | | |
| | p68 | EG_transcript_6137 | | p68 |
| | p72 | | | |
| | Dbp5 | | | |
| | DICE1 | EG_transcript_3163 | | DICE1 |
| | FLJ41215 | | | |
| | DDXL | | | |
| | Abstrakt | | | |

| | | | | |
|---|---|---|---|---|
| | NMP265 | EG_transcript_23012 | EG_transcript_3639, EG_transcript_11281 | NMP265 |
| | RHA | EG_transcript_759 | | RHA |
| | Prp2 | | | |
| | DDX 35 | EG_transcript_4862 | | DDX 35 |
| | fSAP118 | EG_transcript_1831 | | fSAP118 |
| | FLJ21972 | EG_transcript_8077 | | FLJ21972 |
| | SF3b125 | | | |
| | SKI2W | EG_transcript_6545 | EG_transcript_8090 | SKI2W |
| Cyclophilins | CyPL1 | EG_transcript_27668 | | CyPL1 |
| | CyP60 | EG_transcript_27668 | | CyP60 |
| | CyPJ | EG_transcript_6327 | | CyPJ |
| | CyP64 | EG_transcript_14492 | | CyP64 |
| | NY-CO-10 | EG_transcript_8339 | | NY-CO-10 |
| WD40s | TEX1 | EG_transcript_9419 | | TEX1 |
| | fSAP57 | EG_transcript_10335 | EG_transcript_57994 | fSAP57 |
| | fSAP35 | EG_transcript_13815 | | fSAP35 |
| | MGC4238 | EG_transcript_17141 | | MGC4238 |
| Cap Binding | CBP80 | EG_transcript_7013 | EG_transcript_10490, EG_transcript_22536 | CBP80 |
| | CBP20 | EG_transcript_19699 | EG_transcript_22332 | CBP20 |
| Polyadenylated machinery | PAB1 | EG_transcript_2280 | EG_transcript_12193, EG_transcript_23784, EG_transcript_45856, EG_transcript_47270, EG_transcript_5101, EG_transcript_5441 | PAB1 |

| | | | | |
|---|---|---|---|---|
| | PAB2 | EG_transcript_27391 | | PAB2 |
| Znf motif | CPSF 160 | | | |
| | ZNF183 | EG_transcript_18213 | EG_transcript_36959 | ZNF183 |
| | fSAP47 | | | |
| | ZNF207 | | | |
| | FLJ31121 | EG_transcript_26007 | | FLJ31121 |
| Other motifs | U2 AF35 | EG_transcript_46732 | EG_transcript_20807, EG_transcript_20361, EG_transcript_20135 | U2 AF35 |
| | U2 AF65 | EG_transcript_2354 | | U2 AF65 |
| | fSAP164 | EG_transcript_514 | EG_transcript_64781 | fSAP164 |
| | fSAP152 | | | |
| | SPF45 | | | |
| | fSAP94 | | | |
| | ZFM1 | EG_transcript_12083 | | ZFM1 |
| | fSAP | EG_transcript_16024 | EG_transcript_25189, EG_transcript_4423, EG_transcript_71966 | fSAP |
| | TAT-SF1 | | | TAT-SF1 |
| | fSAP59 | EG_transcript_27544 | EG_transcript_39982, EG_transcript_44698, EG_transcript_46762 | fSAP59 |
| | SKIP | EG_transcript_11585 | | SKIP |
| | Hpr1 | EG_transcript_8521 | | Hpr1 |
| | fSAPb | | | fSAPb |
| | CRK7 | EG_transcript_44387 | EG_transcript_54791, EG_transcript_57218, EG_transcript_29712, EG_transcript_27768, EG_transcript_24615, EG_transcript_18904, | CRK7 |

| | | | | |
|---|---|---|---|---|
| | | | EG_transcript_15534, EG_transcript_14792, EG_transcript_13734 | |
| | TIP39 | | | |
| | Prp4 kinase | | | |
| | CPSF5 | | | |
| | fSAPc | | | |
| | FLJ10374 | EG_transcript_17862 | | FLJ10374 |
| | RRP6 | | | RRP6 |
| | SMC1 | EG_transcript_318 | | SMC1 |
| | CAPE | EG_transcript_864 | | CAPE |
| | HUB1 | | | |
| | Prp38 | | | |
| | Prp39 | | | |
| | fSAP17 | | | |
| | MRPS4 | EG_transcript_51504 | | MRPS4 |
| | fSAP113 | | | |
| | SPF30 | | | |
| | CGI-25 | | | |
| Other spliceosomes with unknown motifs | ASR2 | | | |
| | THO2 | EG_transcript_72417 | | THO2 |
| | fSAP24 | EG_transcript_28851 | | fSAP24 |
| | fSAP71 | | | |
| | MFAP1 | | | |
| | FEM-2 | | | |
| | fSAP105 | EG_transcript_8429 | | fSAP105 |
| | CCAP2 | | | |

| | | | | |
|---|---|---|---|---|
| | DGSI | EG_transcript_11864 | | DGSI |
| | NAP | EG_transcript_5609 | | NAP |
| | SPF27 | EG_transcript_14165 | | SPF27 |
| | LUCA15 | | | |
| | KIAA0122 | EG_transcript_62587 | | KIAA0122 |
| | Lupus La | | | |
| | Kin17 | | | |
| | G patch containing 1 | EG_transcript_53097 | | G patch containing 1 |
| | FLJ39430 | EG_transcript_17764 | | FLJ39430 |
| | fSAP23 | EG_transcript_14588 | | fSAP23 |
| | fSAP29 | | | |
| hnRNPs | hnRNP L | | | |
| | hnRNP R | EG_transcript_23877 | | hnRNP R |
| | hnRNP Q | | | |
| | Bub3 | EG_transcript_15004 | EG_transcript_20441, EG_transcript_12355 | Bub3 |
| | HSP70 | | | |
| | HSP71 | | | |
| | HSP90 | EG_transcript_5918 | EG_transcript_4403, EG_transcript_3387, EG_transcript_1060 | HSP90 |
| | GRP78 | EG_transcript_5458 | EG_transcript_5157, EG_transcript_3371 | GRP78 |
| | RNPL | EG_transcript_12387 | EG_transcript_18218 | RNPL |

| | | | | |
|---|---|---|---|---|
| | hnRNP M | EG_transcript_10425 | | hnRNP M |
| | hnRNP H2 | EG_transcript_14181 | | hnRNP H2 |
| | hnRNP A0 | EG_transcript_15874 | EG_transcript_16348, EG_transcript_18642, EG_transcript_20926 | hnRNP A0 |
| | hnRNP A1 | EG_transcript_40301 | EG_transcript_20570 | hnRNP A1 |
| | hnRNP R | EG_transcript_23877 | | hnRNP R |
| | hnRNP U | EG_transcript_26337 | | hnRNP U |
| | hnRNP I | | | |
| | hnRNP K | | | |
| | hnRNP C | | | |
| | HNRNPA2B1 | EG_transcript_9473 | | HNRNPA2 B1 |
| | PTBP2 (nPTB) | EG_transcript_9729 | EG_transcript_6757, EG_transcript_6546, EG_transcript_2834 | PTBP2 (nPTB) |
| | PTBP3 | EG_transcript_22826 | EG_transcript_46910, EG_transcript_8244 | PTBP3 |
| | HuR | | | |
| | HuB | | | |
| | HuC | | | |
| | HuD | | | |
| | Ku70 | | | |
| Additional splicing | CUG-BP | EG_transcript_13296 | EG_transcript_5117 | CUG-BP |

| | | | | |
|---|---|---|---|---|
| regulators | ETR3 | | | |
| | CAGH4 | EG_transcript_5985 | | CAGH4 |
| | Bruno-like 4 | | | |
| | Bruno-like 5 | EG_transcript_32788 | | Bruno-like 5 |
| | Bruno-like 6 | EG_transcript_4547 | | Bruno-like 6 |
| | SFRSK1 | EG_transcript_46064 | | SFRSK1 |
| | SFRSK2 | | | |
| | CDC-like kinase 1 | | | |
| | CDC-like kinase 2 | | | |
| | CDC-like kinase 3 | EG_transcript_6191 | EG_transcript_12549 | CDC-like kinase 3 |
| | CDC-like kinase 4 | EG_transcript_24137 | | CDC-like kinase 4 |

**Note:** The table shows the meiotic genes inventoried in *E. gracilis*. Column 1 = category or complex. Column 2 = gene/convention. Column 3 = nrBLAST evalue for the reference sequence. Column 4 = Reference sequence ID. Column 5 = Paralogous sequence ID. Column 6 = Final classifications. Green coloured cell and corresponding text means the presence of the corresponding protein and the protein name. Empty white cells in columns means that the corresponding protein /process was not found or that a final classification could not be made due to insufficient statistical confidence.

### *5.3.7 Editosomes and related proteins*

I predicted the repertoire of editosome proteins in *E. gracilis* using *T. Brucei* (Salavati, *et al.,* 2012, Deschamps, *et al.,* 2011, and Goringer, 2012) and *A. thaliana* reference sets (Takenaka, 2014). All editosome subfamilies and related proteins were found except for GAP2, MORF1 (recently discovered in *A. thaliana*), and some members of the 20S which do not have any *E. gracilis* ortholog (Table 5.14).

**Table 5.14**: Proteins involved in editing found in *E. gracilis* transcriptome

| Category/complex | Gene/convention | nrBLAST evalue | Ref. Seq. ID | Paralogous Seq. ID | Final classification |
|---|---|---|---|---|---|
| | | evalue | | | |
| Accessory protein | Cobalt ion binding | | | | |
| | DAG | | | | |
| | DAL1 | | | | |
| KPAP1 | EF-Tu | 0 | EG_transcript_8722 | EG_transcript_1495, EG_transcript_45355 | KPAP1 (EF-Tu) |
| MRB1 | GAP2 | | | | |
| Accessory protein | GR-RBP3 | 6E-24 | EG_transcript_39982 | EG_transcript_40823, EG_transcript_43186, EG_transcript_44698, EG_transcript_46762, EG_transcript_47270, EG_transcript_54308 | GR-RBP3 |
| KPAF | KPAF1 | 4E-74 | EG_transcript_3221 | EG_transcript_56042 | KPAF1 |
| | KPAF2 | 8E-44 | EG_transcript_12380 | | KPAF2 |
| KPAP1 | KPAP1 | | | | |
| 20S Editosomes | KREL1 | | | | |
| | KREN1 | | | | |
| | KREN2 | | | | |
| | KREN3 | | | | |
| | KREPA1 | | | | |
| | KREPA2 | 3E-25 | EG_transcript_1612 | | |
| | KREPA3 | | | | |
| | KREPA4 | | | | |
| | KREPA5 | | | | |
| | KREPB4 | | | | |
| | KREPB5 | | | | |
| | KREPB6 | | | | |

| | | | | | |
|---|---|---|---|---|---|
| | KREPB7 | | | | |
| | KREPB8 | | | | |
| | KRET1 | 2E-25 | EG_transcript_1661 | EG_transcript_10831 | KRET1 |
| | KRET2 | 3E-25 | EG_transcript_9521 | | KRET2 |
| | KREX1 | 6E-23 | EG_transcript_4534 | | KREX1 |
| | KREX2 | 6E-23 | EG_transcript_4534 | | KREX2 |
| | KRIPP1 | 6E-24 | EG_transcript_3096 | | KRIPP1 |
| | LETM1 | | | | |
| | MEAT1 | | | | |
| | MHEL61 | 0 | EG_transcript_2406 | EG_transcript_11281, EG_transcript_12010, EG_transcript_1328, EG_transcript_13713, EG_transcript_14317, EG_transcript_15110, EG_transcript_23012, EG_transcript_2748, EG_transcript_2752, EG_transcript_2769, EG_transcript_2790, EG_transcript_2941, EG_transcript_2951, EG_transcript_3119, EG_transcript_3639, EG_transcript_3822, EG_transcript_3940, EG_transcript_4123, EG_transcript_4811, EG_transcript_5115, EG_transcript_5184, EG_transcript_5242, EG_transcript_52954, EG_transcript_5308, EG_transcript_5350, EG_transcript_5511, EG_transcript_5632, EG_transcript_5792 EG_transcript_5941, EG_transcript_5967, EG_transcript_6094, EG_transcript_6137, EG_transcript_6232, EG_transcript_6255, EG_transcript_6824, EG_transcript_6868, EG_transcript_6971, EG_transcript_7161 EG_transcript_7389, EG_transcript_7490, EG_transcript_7585, EG_transcript_7645, | MHEL61 |

| | | | | | |
|---|---|---|---|---|---|
| | | | | EG_transcript_7679, EG_transcript_7901, EG_transcript_8174, EG_transcript_8744, EG_transcript_9013, EG_transcript_9065 | |
| MRB1 | MRB1 (MRB3010) | 1.00E-044 | EG_transcript_1813 | EG_transcript_19151 EG_transcript_39166 EG_transcript_56935 EG_transcript_59715 | MRB1 (MRB3010) |
| | MRB4150 | | | | |
| | MRB4160 | | | | |
| | MRB8170 | | | | |
| | MRB8620 | | | | |
| | MRG | 7.00E-013 | EG_transcript_9906 | | MRG/bromodomain putative |
| | MSSU | | | | |
| | REH2 | | EG_transcript_1067 | | REH2 |
| | RGG | | EG_transcript_763 | | RGG |
| | RGG2 | | EG_transcript_31056 | | RGG2 |
| | RRM/RBD/RNP | | EG_transcript_39982 | | RRM/RBD/RNP |
| MORF Family | MORF1 | | | | |
| | MORF2 | | | | |
| | MORF3 | | | | |
| | MORF4 | | | | |
| | MORF5 | | | | |
| | MORF6 | | | | |
| | MORF7 | | | | |
| | MORF8 | | | | |
| | MORF9 | | | | |
| | MORF10 | | | | |

**Note:** The table shows the meiotic genes inventoried in *E. gracilis*. Column 1 = category or complex. Column 2 = gene/convention. Column 3 = nrBLAST evalue for the reference sequence. Column 4 = Reference sequence ID. Column 5 = Paralogous sequence ID. Column 6 = Final classifications. Green coloured cell and corresponding text means the presence of the corresponding protein and the protein name. Empty white cells in columns means that the corresponding protein /process was not found or that a final classification could not be made due to insufficient statistical confidence.

### *5.3.8 TAC/CARP proteins*

Using eukaryotic TAC/CAP reference sets (Zhou, *et al.,* 2012 and Lopez-Bellido, *et al.,* 2015), I could not predict the Tachykinin proteins: TAC1, TAC3s, TAC4, TACR1, TACR2, TACR3, and TACR4 as well as the Tripartite Attachment Complex (Appendix-Ia). This also include additional Tachykinin related proteins: TACL, TACL-86, substance-P receptor, substance-K receptor, protachykinin-1 isoform X1, peptides receptor 99D like, peptides receptor 99D, neuromedin-K receptor, and (Tachykinin-like) G protein-coupled receptor. I only found two CARPs and include CARP1 and 4 (Appendix I-a). The absence of Tachykinin and Tripartite Attachment Complex in *E. gracilis* is understandable since these proteins are unique to mammals and kinetoplastids (due to the presence of a kinetoplast) respectively.

### *5.3.9 RNAi pathway*

I searched for all the members of the RNA-induced silencing complex (RISC) using eukaryotic reference sets (Owens and Malham, 2015; Dunoyer, *et al.,* 2015; and Dang, *et al.,* 2011). Present are all the subfamilies of the Dicer and Drosha - RNase III proteins, Argonaute proteins (except PIWI and Zwille), small ncRNAs (except ssRNA binding), and Eri-1 like nuclease. dsRBM, RdRP, sid-1-like, AGO2 and AGO3 were not found (Table 5.15).

**Table 5.15**: RNAi pathway proteins found in *E. gracilis* transcriptome

| Category/complex | Gene/convention | nrBLAST evalue | Ref. Seq. ID | Paralogous Seq. ID | Final classification |
|---|---|---|---|---|---|
| Dicer and Drosha - RNase III proteins | DCR1/DCL1 | 7E-38 | EG_transcript_14317 | | DCR1/DCL1 |
| | DCR2/DCL2 | 0 | EG_transcript_2748 | | DCR2/DCL2 |
| | DCR3/DCL3 | 4E-87 | EG_transcript_7679 | | DCR3/DCL3 |
| | DCR4/DCL4 | 1E-127 | EG_transcript_12010 | | DCR4/DCL4 |
| | Drosha | 4E-13 | EG_transcript_20514 | | |
| Argonaute proteins | AGO1 (PAZ, PIWI-PAZ domains) | 1E-113 | EG_transcript_1984 | | AGO1 (PAZ, PIWI-PAZ domains) |
| | PIWI | | | | |
| | Zwille | | | | |
| | PIWI-like 1 | 1E-113 | EG_transcript_1984 | | AGO1 (PAZ, PIWI-PAZ domains) |
| | PIWI-like 2 | 9E-18 | EG_transcript_7865 | | |
| | AGO3 | | | | |
| | AGO2 | | | | |
| small ncRNAs | dsRNA binding/Sid-1 | 2E-111 | EG_transcript_5115 | | RNA helicase |
| | ssRNA binding | | | | |
| dsRBM | dsRBM | | | | |
| RdRP | RdRP | | | | |
| Sid-1-like | Sid-1-like | | | | |

**Note:** The table shows the meiotic genes inventoried in *E. gracilis*. Column 1 = category or complex. Column 2 = gene/convention. Column 3 = nrBLAST evalue for the reference sequence. Column 4 = Reference sequence ID. Column 5 = Paralogous sequence ID. Column 6 = Final classifications. Green coloured cell and corresponding text means the presence of the corresponding protein and the protein name. Empty white cells in columns means that the corresponding protein /process was not found or that a final classification could not be made due to insufficient statistical confidence.

## 5.4 METABOLISM

Cellular metabolism, represented by complements of enzymatic and transport reactions, is a fundamental biological system required for sustaining life (Peregrín-Alvarez, et al., 2009), and involves glycolysis and gluconeogenesis. Glycolysis (the Embden–Meyerhof–Parnas pathway) and gluconeogenesis are common to all domains of life and play fundamental roles in essential cellular processes. Some specific metabolic pathways are common to members of the kinetoplastids, and possibly, Excavates and other members of the eukaryotic kingdom. For instance, the carbohydrates (Galactofuranose, Galf) present in glycoconjugates of the cell surface of trypanosomatids are essential to a variety of cell events, including infectivity and/or virulence in insect vectors and mammalian hosts (Stoco, *et al.,* 2012). Similarly, a common feature of *Leishmania*-infected sand flies is the presence of a mass of parasites embedded within a gel-like material called the promastigote secretory gel (PSG), which contains proteophosphoglycans (PPGs) and lipophosphoglycan (LPG) (Stierhof, *et al.,* 1999). In this section, these processes (Galf and PPGs) will be discussed in *E. gracilis* (see section 5.6 for an extensive community discussion of the additional processes).

### *5.4.1 GalF*

I found the Gipl galf transferase (GIPL galf), galactofuranosyltransferase lpg1-like protein (LPG1L), beta galactofuranosyl glycosyltransferase (B-GalF), and beta galactofuranosyl transferase (B-GalFT) (Stoco, *et al.,* 2012;Tefsen, *et al.,* 2012; Matsunaga, et al., 2015; Komachi, et al., 2013), which are conserved and consisted with the eukaryotic reference sets (Table 5.16).

**Table 5.16**: Galf biosynthesis genes in *E. gracilis*

| Category/complex | Gene/convention | nrBLAST evalue | Ref. Seq. ID | Paralogous Seq. ID | Final classification |
|---|---|---|---|---|---|
| Gipl galf transferase | GIPL galf | | | | |
| Galactofuranosyl transferase lpg1-like protein | LPG1L | 6E-14 | EG_transcript_14245 | EG_transcript_5711, EG_transcript_6863 | beta galactofuranosyl transferase |
| Beta galactofuranosyl glycosyltransferase | B-GalF | 2E-19 | EG_transcript_10895 | EG_transcript_15940, EG_transcript_16921 | glycosyltransferases |
| Beta galactofuranosyl transferase | B-GalT | 6E-14 | EG_transcript_14245 | | |
| Galactofuranosyl transferase A | glfA | 1E-150 | EG_transcript_1069 | EG_transcript_1167, EG_transcript_751 | Galactofuranosyl transferase A/udp-galactopyranose mutase |

**Note:** The table shows the meiotic genes inventoried in *E. gracilis*. Column 1 = category or complex. Column 2 = gene/convention. Column 3 = nrBLAST evalue for the reference sequence. Column 4 = Reference sequence ID. Column 5 = Paralogous sequence ID. Column 6 = Final classifications. Green coloured cell and corresponding text means the presence of the corresponding protein and the protein name. Empty white cells in columns means that the corresponding protein /process was not found or that a final classification could not be made due to insufficient statistical confidence.

### *5.4.2 PPG*

Multiple copies of Phosphoglycan beta (PPG) were found: PPG1, PPG3, PPG4, and PPG5, except for PPG2 which was not found (Table 5.17) (Satheesh, et al., 2014; Rogers, 2012; Secundino, *et al.,* 2010; Gabriela, et al., 2015). The presence of LPG and PPGs in *E. gracilis* may suggest a novel mechanism for resistance to external stressors as evident in *L. major*. They postulate that parasitism evolved at least four times in kinetoplastids, suggesting that obligate parasitic trypanosomatids are a relatively 'derived' group within kinetoplastids; their closest relative is likely to be the free-living *Bodo saltans*, and the ancestral trypanosomatids were probably parasites of insects.

**Table 5.17**: Proteophosphoglycan proteins in *E. gracilis*

| Category/complex | Gene/convention | nrBLAST | Ref. Seq. ID | Paralogous Seq. ID | Final classification |
|---|---|---|---|---|---|
| Phosphoglycan beta 1 | PPG1 | 0 | EG_transcript_9168 | EG_transcript_9330, EG_transcript_12167 | Protein kinase/LRRP |
| Phosphoglycan beta 2 | PPG2 | | | | |
| Phosphoglycan beta 3 | PPG3 | 7.00E-75 | EG_transcript_4652 | | Leucine Rich Repeat Protein |
| Phosphoglycan beta 4 | PPG4 | 2.00E-17 | EG_transcript_14557 | EG_transcript_34202 | Leucine Rich Repeat Protein |
| Phosphoglycan beta 5 | PPG5 | 2.00E-17 | EG_transcript_14557 | | Leucine Rich Repeat Protein |

**Note:** The table shows the meiotic genes inventoried in *E. gracilis*. Column 1 = category or complex. Column 2 = gene/convention. Column 3 = nrBLAST evalue for the reference sequence. Column 4 = Reference sequence ID. Column 5 = Paralogous sequence ID. Column 6 = Final classifications. Green coloured cell and corresponding text means the presence of the corresponding protein and the protein name. Empty white cells in columns means that the corresponding protein /process was not found or that a final classification could not be made due to insufficient statistical confidence.

## 5.5 EVOLUTION

Members of Euglenozoa have diverse modes of nutrition, including predation, parasitism, and photoautotrophy (Lukes, et al., 2009). Little is known about the evolution of parasitism in kinetoplastids: According to Simpson, *et al.,* 2006, recent improvements in the taxon sampling for nuclear rRNA genes and several protein markers have transformed this understanding. Similarly, the genetic basis for parasitism is not yet known, however, it is possible that the loss and gain in parasitic lifestyle within the Euglenozoans may be due to Horizontal Gene Transfer (HGT) (Soucy, *et al.,* 2015). Many apparent gene duplications events, for example, are now known to be the result of HGT, not autochthonous gene duplication, resulting in a 'web of life' rather than in a steadily bifurcating tree (Treangen, and Rocha, 2011; Swithers, *et al.,* 2012). In this section, the results of the orthologous groups clustering will be discussed putting in context the endosymbiotic phenomenon and the transfer of genes.

### *5.5.1 Orthologous groups clustering*

To depict orthology accross the eukaryotic kingdom such as families shared across eukaryotes and those restricted to specific taxonomic groupings, a set of 30 eukaryotic taxa was assembled which were then clustered with the *E. gracilis* transcriptome using the OrthoFinder algorithm (Emms and Kelly, 2015). This clustering analysis is important to understand the origin of the genes in *E. gracilis* as well as its hybrid nature. Overall, of the 36,526 *E. gracilis* predicted proteins, ~ 20,000 (55 % ; orthogroups n = 400) have orthology to one or more of the 30 eukaryotic reference sets, and are responsible for core metabolic, structural and informational processes - these are referred to as pan eukaryotic genes while about 45 % are shared accross selected eukaryotes (other euglena shared genes). Within this sets, euglena have genes specifically shared with trypanosomatids, broader sets of excavates, unikonts, land plants, red algae, secondary hosts, and green algae as well as genes unique to *E. gracilis* (see Chapter 5, Figure 4.3). For example, 82 orthogroups were found in common with *Bodo saltans,* 245 shared across all other sampled kinetoplastids and 36 across the kinetoplastids, *E. gracilis* and the heterolobosid *Naegleria gruberi* (or Excavates). These classes are broadly within the relative frequencies of what was expected from previous analysis of Excavate

genomes (Simpson, *et al.,* 2006) and consistent with the phylogenetic positions of the relevant taxa (see Ebenezer, *et al.,* 2018, for an extensive discussion).

## 5.6 OVERVIEW OF COMMUNITY-BASED SYSTEMS INVESTIGATED

In this section, the community based annotation which I coordinated but did not directly analyse, will be discussed briefly. For an extensive discussion, see Ebenezer et al., 2018, and Appendix I. Biological aspects and the name of the corresponding Annotator/PI that will be discussed briefly include : Surfaceome (Andrew Jackson), plastid proteome (Vladimir Hampl), mitochondrial genes (Julius Lukes), signal transduction (Michael Lebert), dynamins (Mekund Thattai), and metabolism (Ellis O'Neill).

Surface protein predictions were estimated for *E. gracilis* employing the method as described in *Jackson, et al.,* 2013 using *T. brucei, T. cruzi, L. major,* and *B. saltans* as a reference set. The main criteria to classify surface genes were the presence of signal peptides and corresponding orthologous clustering annotation criteria when available (see the supplimentary section of the main genome paper, Ebenezer, *et al.,* manuscript in preparation). Out of the 36,526 *E. gracilis* predicted proteins, 3259 (9 %) were putative surface proteins. Generally, protein families existing in most organisms were found in *Euglena*. These include, amongst others, ABC transporters, adenylyl cyclases, amino acid transporters and ATP-binding cassette proteins (Jackson, *et al.,* 2013; Jackson, 2016; Jackso, *et al.,* 2015; Wang, *et al.,* 2010; Jackson*, et al,* 2012; Depledge, *et al.,* 2010). In common with *B. saltans*, it appears that *E. gracilis* has a distinct class of amastin; a major surface protein common to the kinetoplastids, but that arose from a single ancestral form shared with the last common ancestor of euglenids (Figure 5.12). It is also of interest in this regard that gp63, a major surface protein from *Leishmania* and present in nearly all eukaryotes, is absent from the genome, and hence must represent a secondary loss following separation from the kinetoplastid lineage.

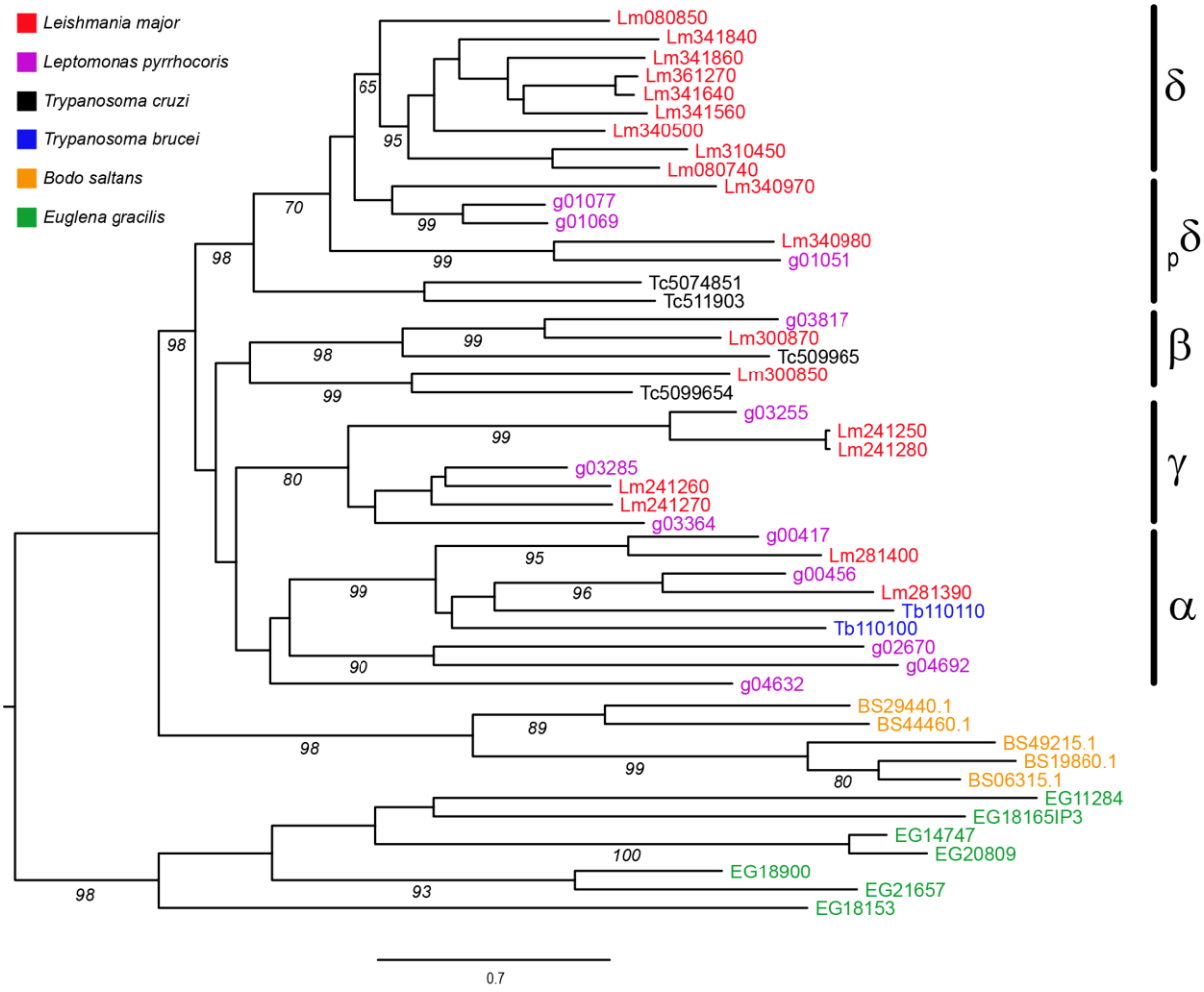



**Figure 5.12: *E. gracilis* possess an unusual surface.** The diagram shows the Neighbor Joining phylogeny of Amastin protein sequences from *Euglena* and diverse kinetoplastids. Colour codes: Taxa highlighted in red, purple, black, orange, and green represent *L. major*, *L. pyrrhocoris*, *T. cruzi*, *T. brucei*, *B. saltans*, and *E. gracilis* respectively. Bars in fron of taxa nodes represent the classifications of the Amastin protein identified: δ, ρδ, β, ɣ, and α.

Plastid proteome was predicted in *E. gracilis* using automatic functional pipelines (SignalP, PrediSI, ChloroP; Petersen, *et al.,* 2011; http://www.predisi.de/home.html; Emanuelsson, *et al.,* 1999) and a set of reference (*A. thaliana* chloroplast genes) (https://www.arabidopsis.org/). Out of the 36,526 *E. gracilis* predicted proteome, 1906 were predicted to be plastid associated (Figure 5.13, 5.14, 5.15). 1179 of these are hypothetical, 475 were annotated and sorted to categories, while 252 were annotated but could not be sorted to categories. To further ascertain the reliability of our annotated dataset, the data were analyzed using proteins generated through Mass Spectrometry. 472 sequences overlap between the In silico and Mass Spectrometry proteomes, and they show similar patterns in terms of proportion annotated and categorized – except that there are lower numbers of hypothetical proteins in the MS proteome as well as bias towards better annotated sequences. The plastid protein sequences were sorted into 24 categories and these include photosynthesis, carbohydrate metabolism, and ion transport (Figure 5.14).

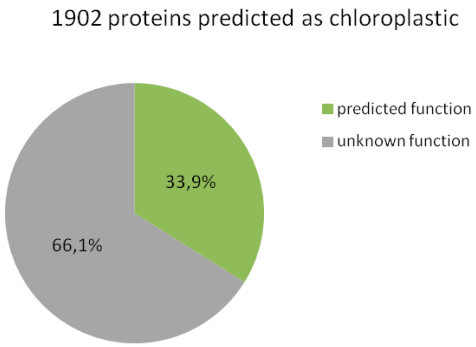



**Figure 5.13: Pie chart depicting functional distribution of *E. gracilis* functional predicted chloroplastic proteins.** The ratio between annotated proteins with a certain predicted function and proteins of unknown homology and/or function. In the set of 1,902 plastid candidates, 1,257 proteins (66.1 %) remained without a predicted function, 1,008 being completely unidentifiable with no homologs in bioinformatic databases and 249 having unclear or completely unknown function.

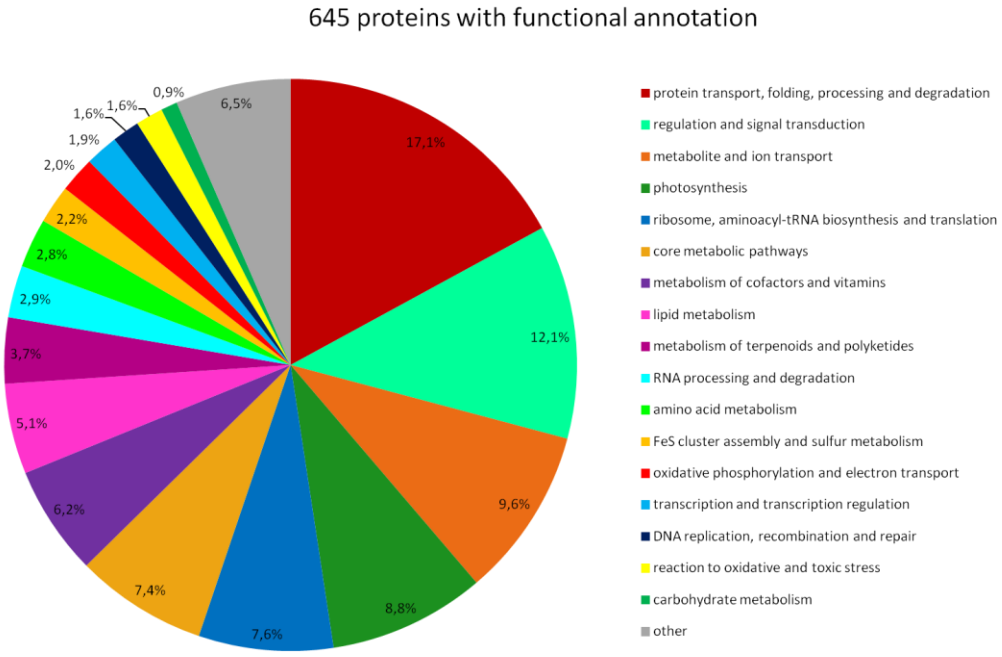



**Figure 5.14: Protein transport, folding, processing and degradation, are the major biological activity of the *E. gracilis* plastid.** Pie chart depicting functional annotation of *E. gracilis* predicted chloroplastic proteins. The remaining 645 (33.9 %) proteins with predicted function were used for metabolic and cellular pathways reconstruction and sorted into 18 functional categories. The largest portion of proteins were those involved in the synthesis, post-translational modification and folding of newly synthesized proteins and with regulatory and/or signalling functions. Other major categories include photosynthesis and chlorophyll biosynthesis,

ribosome biogenesis and protein synthesis and transport of ions and non-protein molecules.

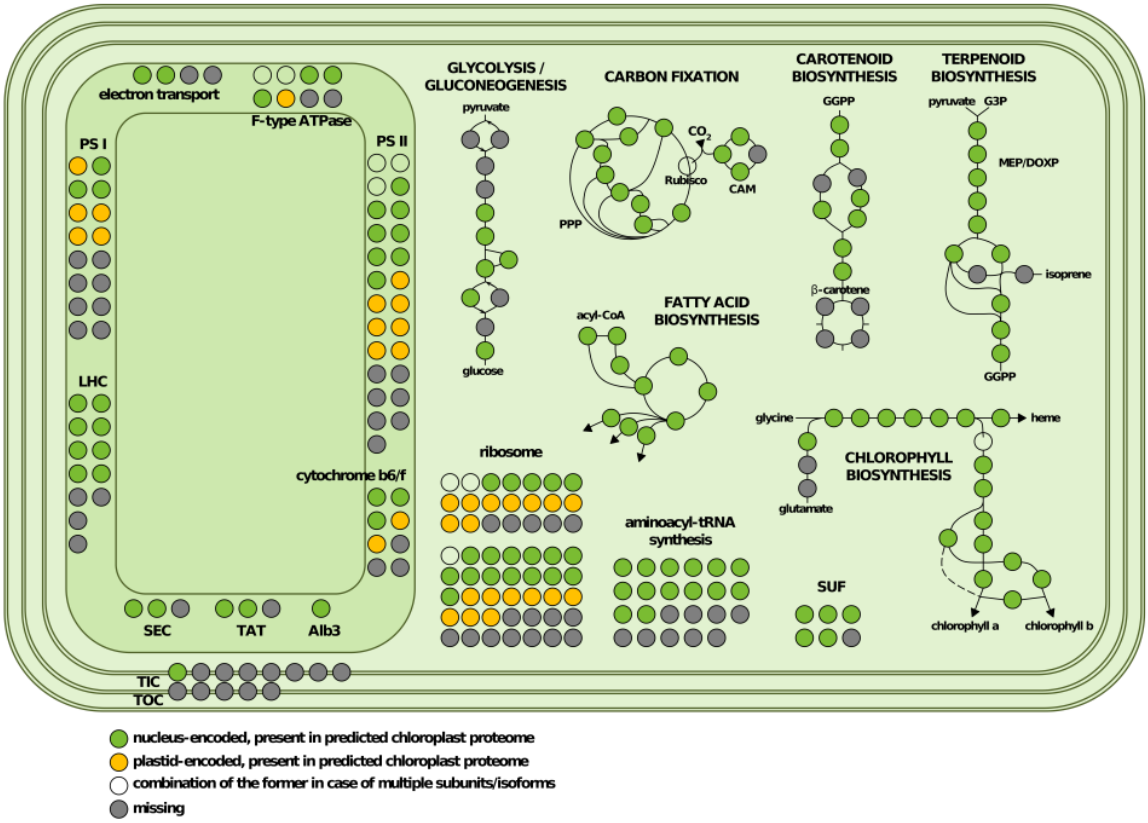



**Figure 5.15: The *Euglena gracilis* plastid possesses broad metabolic potential.** Proteins involved in core plastid metabolic pathways were identified and include: glycolysis/gluconeogenesis, carbon fixation, fatty acid biosynthesis, caretenoid biosynthesis, terpenoid biosynthesis, and chlorophyll biosynthesis. Colour codes: green, nucleus encoded, present in predicted chloroplast proteome; amber, plastid encoded, present in predicted chloroplast proteome; light green/white, combination of green and amber in case of multiple subunits/isoforms and grey, expected but not found.

There are three ways mitochondrial protein localization were predicted in the dataset: Blast2GO (https://www.blast2go.com/), TargetP (Emanuelsson, *et al.,* 2000), and BLASTP searches against the *T. brucei, H. sapiens*, and *S. cerevisiae* mitoproteomes (Appendix I). Out of the 36,526 predicted *E. gracilis* proteome, 900 proteins were finally predicted to be mitochondrial associated with 37 pathways annotated from KEGG (Figure 5.16 – 5.18). The list of proteins found associated with mitochondria localization also include the seven protein-coding genes recently published by Dobakova, *et al.,* 2015 (Faktorová, *et al.,* 2016; Buhrman, *et al.,* 2013; Zhang, *et al,* 2010; and Lukes, 2014).

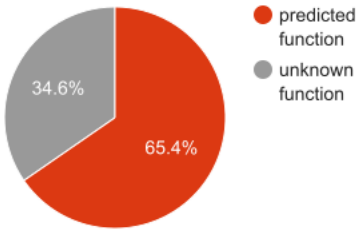



**Figure 5.16: *E. gracilis* mitochondria possess more known functional information.** The predicted mitochondrial proteome of *E. gracilis*. The ratio between annotated proteins with a certain predicted function and proteins of unknown homology and/or function.In the set of 1075 mitochondrial candidates, 703 (65.4 %) remained with a predicted function while 372 (34.6) do not possess a predicted function.

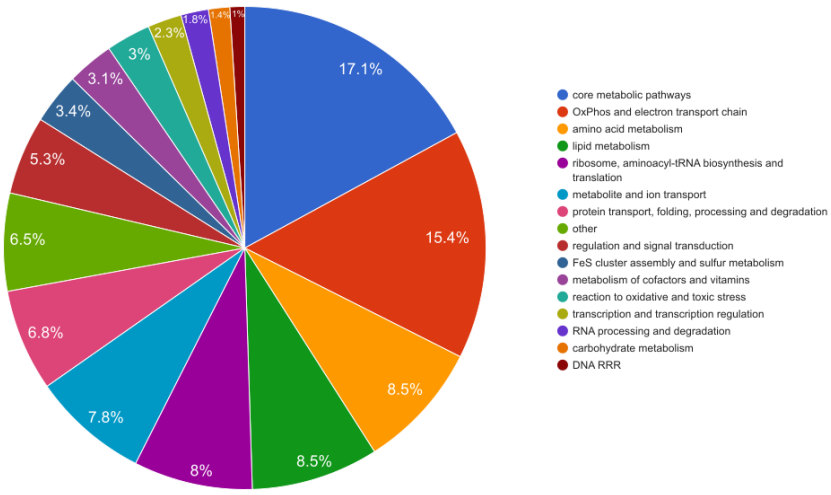



**Figure 5.17: The *E. gracilis* plastid possess core metabolic pathways.** The predicted mitochondrial proteome of *E. gracilis* with predicted functions were used for metabolic and cellular pathways reconstruction and sorted into 16 functional categories as shown in the diagram. These functional categories are highlighted in colour codes above.

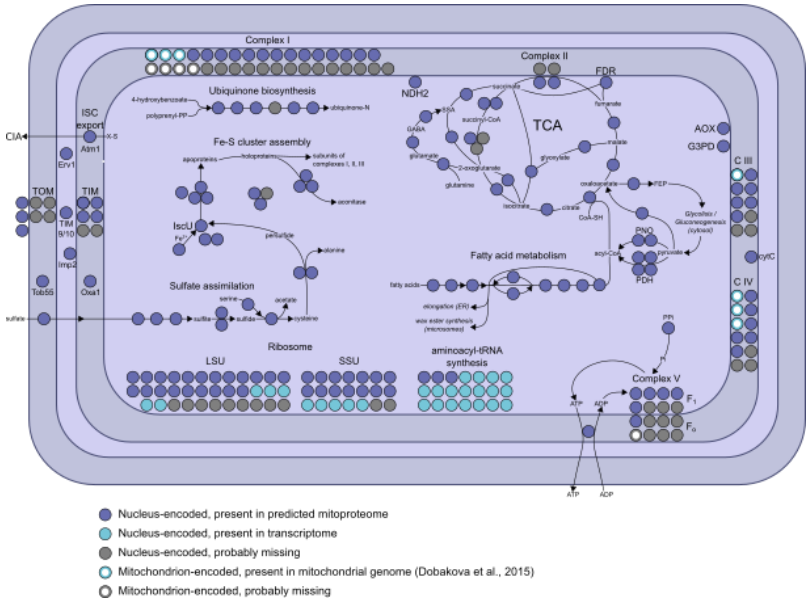



**Figure 5.18: A reconstruction of likely metabolic pathways present within the mitochondria organelle of *E. gracilis.*** The diagram shows the five mitochondrial complexes found in *E. gracilis*: Complex I, II, III, IV, and V, alongside LSU and SSU rRNAs, and accessory plastid proteins.

Signal transduction genes were identified in *E. gracilis* using BLAST and Blast2GO (https://www.blast2go.com). Out of the 36,526 protein sequences in *E. gracilis*, 1495 matched proteins with annotated signal transduction functions. The distribution of the predicted signal transduction genes include, amongst others, protein kinase (14%) and cyclase (13%) (Appendix I-b). The distribution of the top BLASTP hits shows a distribution across the kingdoms of life (data not shown), highlighting the diversity of possible origins of genetic material present in the *Euglena* genome, which have been integrated through its complex phenotypic and genotypic history (Ahmadinejad, *et al.*, 2007; O'Neill, *et al.,* 2015).

Members of the dynamin superfamily are characterized by three major domains: an N-terminal GTPase domain (ND), a stalk region formed by helices of the middle domain (MD), and a GTPase effector domain (GED) that folds back across the stalk to regulate GTPase activity (Purkanti and Thattai, 2015). A detailed annotation of dynamins encoded by the *E. gracilis* genome was carried out using Pfam Hidden Markov Models and HMMER (Finn,*et al.,* 2014) to find proteins containing the GTPase domain (Pfam PF00350), middle domain (Pfam 01031), and the GTPase effector domain (Pfam PF02212) (Table 5.18). It was found that *E. gracilis* has 20 potential dynamins (after resolving splice isoforms): 3 Class A, 6 Class B, 1 Class C, with the remaining 10 unclassified (Table 5.18). Only one member each from Class A (EG_transcript_8006) and Class B (EG_transcript_3945) have complete dynamin-like architectures with all three domains. I found one FtsZ homolog in the *E. gracilis* genome (using BLAST OrthoFinder algorithm and phylogenetic analysis) and all evidence suggest that this is mitochondrial associated as the *E. gracilis* FtsZ homolog shows similarities to kinetoplastids than to red or green photosynthetic algae/prokaryotes orthologs, suggesting a mitochondrial origin.

**Table 5.18**. Répertoire of Dynamin proteins in *E. gracilis*.

| Transcript ID | Splice isoforms | Domain architecture | Class | Signature |
|---|---|---|---|---|
| EG_transcript_8006 | | Dynamin_N--Dynamin_M--COMP--GED | Class A | :AAAAAAA |
| EG_transcript_3945 | EG_transcript_41670, EG_transcript_45081 | Dynamin_N--Dynamin_M--GED | Class B | B::B:::K |
| EG_transcript_7126 | | Dynamin_N | Class C | :M:----- |
| EG_transcript_6599 | | Dynamin_N--Dynamin_M | Class A | :F:::::- |
| EG_transcript_16277 | | Dynamin_N | Class A | :F:----- |
| EG_transcript_6469 | | Dynamin_N | Class B | :B:----- |
| EG_transcript_14051 | | Dynamin_N--Dynamin_M | Class B | B:BB:::- |
| EG_transcript_18246 | | Dynamin_M--GED | Class B | ---::::K |
| EG_transcript_29706 | | Dynamin_N | Class B | :B:----- |
| EG_transcript_6775 | EG_transcript_21609 | Dynamin_N--DUF4351--DUF2366 | Class B | BB:----- |
| EG_transcript_28821 | | Dynamin_N | Unclassified | :::----- |
| EG_transcript_14648 | | Dynamin_N | Unclassified | :::----- |
| EG_transcript_60550 | | GED | Unclassified | -------: |
| EG_transcript_4990 | | Dynamin_N--AAA_25 | Unclassified | :::----- |
| EG_transcript_561 | | Dynamin_N | Unclassified | :::----- |
| EG_transcript_3279 | | Dynamin_N--DUF2552 | Unclassified | :::----- |
| EG_transcript_1987 | | PRA-PH--Dynamin_N--DUF2205 | Unclassified | :::----- |
| EG_transcript_2596 | | SAM_1--Dynamin_N--PI3K_p85B | Unclassified | :::----- |
| EG_transcript_12528 | EG_transcript_25672 | Dynamin_N | Unclassified | :::----- |
| EG_transcript_8748 | | Dynamin_N | Unclassified | :::----- |

**Note:** Colume 1 = *E. gracilis* transcript ID. Column 2 = Spliced isoforms. Column 3 = Domain architecture. Column 4 = Class of dynamins. Column 5 = dynamin signature. Domains are according to Pfam. Signatures correspond to the notation used in Purkanti and Thattai, 2015.

Using BLAST searches and the CAZy database, three superfamilies of the Carbohydrate Active Enzymes: Glycoside Hydrolase (GH), Glycosyltransferases (GTs), and the distant relatives of the Glycosyltransferases (GT-related) were found (Yoshida, *et al.,* 2016; O'Neill and Field, 2015; Takeda, *et al.,* 2015; Grimma, *et al.,* 2015; O'Neill, *et al.,* 2015a&b; Lombard, *et al.,* 2013; Kuhaudomlarp, et al). 32, 36 and 10 families that are associated with the GHs, GTs, and GT-related were found respectively, with corresponding enzymatic activities ranging from 5'-AMP-activated protein kinase beta-subunit to Glycosyltransferase (Appendix I-a). Many of these has the N-, O- and C-linked glycosylated sites (Chauhan, *et al.,* 2013; Li, *et al.,* 2015; Caragea, *et al.,* 2007; Martínez-Duncker, *et al.,* 2014; Aebi, 2013; Ulvskov, *et al.,* 2013; Moremen, *et al.,* 2012).

## 5.7 DISCUSSIONS AND CONCLUSION

### *5.7.1 Cellular processess and signaling*

The analysis of the protein trafficking pathways in *E. gracilis* provides a detail insight into the evolution of genes involved in this pathway across eukaryotes (Elias, *et al.,* 2012; Rojas, *et al.,* 2012 ; van Dam, *et al.,* 2013; Abbasi, *et al.,* 2011; Manna, *et al.,* 2013; Hirst, *et al.,* 2011; Schlacht and Dacks, 2015; Adunga, *et al.,* 2013; McMahon, H. T. and Boucrot, 2011). In the Rab protein family, which provides a potent force for endomembrane and cellular evolution across the entire range of Eukaryota (Elias, *et al.,* 2012), there are multiple conserved paralogues of some specific subfamilies such as Rab32, Rab11 and Rab7. The presence of the core endocytic and exocytic pathways in *E. gracilis* suggests details of cellular complexity in LECA (Elias, *et al.,* 2012) as previously described in *N. gruberi* (Fritz-Laylin, *et al.,* 2010; Koonin, 2010). The multiple Rab paraloguos in *E. gracilis* indicate the presence of potentially multiple anterograde routes, and also implies the presence of active autophagic systems, while other trafficking pathways identified include complex endosomal network containing multiple sorting and recycling steps (Rab5, 21 and 22), late endosomal and/or lysosomal trafficking (Rab7, 2 and 32), retrograde transport through the Golgi complex (Rab2 and 6) and the endosomal recycling and exocytic system (Rab4, 11). This clearly indicate that bidirectional movement of molecules through the endomembrane system was firmly established in the LECA (Elias, *et al.,* 2012). The presence of IFT27, IFT22, Rab23, Rab8 and Rab11, all suggest multiple

transport pathways integrating the endomembrane system and the flagellum (Elias, *et al.,* 2012). The detection of several ancestral, widely distributed Rabs with no known function (e.g. RabX1, RabX2, and RabX3) suggests that there remain many fundamental aspects of Rab biology that are yet to be described (Elias, *et al.,* 2012).

The presence of IFT22 and IFT27 in this dataset also supports the role of these proteins in intraflagellar transport as well as the complex evolution of the IFT and its origin from a protocoatomer complex which corroborates the involvement of IFT components in vesicle transport (van Dam, *et al.,* 2013). It is not yet clear the significance of the absence of the Rho family in *E. gracilis*, however, while animals and fungi possess multiple Rho paralogous involved in signal transduction, cytoskeletal function and cellular proliferation, the trypanosomatids possess a much simpler Rho-signaling system (Abbasi, *et al.,* 2011). For instance, *T. brucei* genome contains only a single divergent Rho-related gene which has recently been found to play important role in spindle formation and mitosis (Abbasi, *et al.,* 2011).

The absence of AP4E and AP2B (and the presence of all members of the AP subunits) may be due to slight fragment of the transcriptome assembly but these proteins are also absent in some trypanosomatids such as *T. congolense* (Manna, *et al.,* 2013), suggesting that the pathway that mediate selective transport to the endosome rather than exocytose (Burgos, *et al.,* 2010) may have been lost in *E. gracilis* and reflecting a simplification in post-Golgi transport events (Manna, *et al.,* 2013). The recently discovered AP5 subunit which is absent in all kinetoplastids (Manna, *et al.,* 2013) but present in *N. gruberi* was also not found in *E. gracilis*. While Manna, *et al.,* 2013, proposed that the AP5 arose after the separation of the kinetoplastids from the main eukaryotic lineage, the absence of AP5 in *E. gracilis* have pushed the root of this protein further up the eukaryotic tree, suggesting an ancient ancestral divergence of AP5 after the separation of the Euglenozoans from the main eukaryotic lineage (Manna, *et al.,* 2013). The AP5 is involved in endosomal sorting (Hirst, *et al.,* 2011) and appears to mediate trafficking of the cation independent mannose-6-phosphate receptor, a specialised lysosomal pathway that is absent from the trypanosomes (Manna, *et al.,* 2013), suggesting that this pathway may also be absent in *E. gracilis*.

The distribution of the coat complex suggests that these proteins were present in LECA (Neumann, *et al.,* 2010). It is not yet clear the significance of the absence of Sec12 and Sec16, however, very few organisms are missing both subunits and could suggest that lower levels of ER to Golgi trafficking are necessary in these organisms or that, given the appropriate cellular conditions, these factors are not necessary for the formation of COPII-coated vesicles (Schlacht and Dacks, 2015). In this scenario, Sec12 and Sec16 would serve to increase the speed and efficiency of coat formation, rather than acting as integral steps in the process or the other GEFs/scaffold proteins may have functionally replaced Sec12 or Sec16 (see Schlacht and Dacks, 2015, and Ebenezer, *et al.,* 2018, for an extensive discussion).

The results of this investigation indicate that an expansion in the membrane trafficking proteins partially facilitates a plastid directed membrane trafficking pathway. Trafficking proteins from the ER to the Golgi is facilitated by an increase in COPII to allow for vesicle formation (Robinson, 2004), and Rab1; to target the vesicle to the Golgi and recruit the necessary tethers (Stenmark, 2009; Brocker, *et al.,* 2010). At the cis-Golgi the additional Sly1 homolog can function to regulate the SNARE-SNARE interactions (Dulubova, *et al.,* 2002; Yoshizawa, *et al.,* 2006) of the additional SNARE proteins Gos and Sec22.

Meiotic sex, the fusion of haploid meiotic products from different individuals, is thought to have originated in the common ancestor of all eukaryotes (Dacks and Roger, 1999), are particularly present in *E. gracilis*, except for REC8 which is involved in double-strand break repair and meiotic divisions. However, this is consistent with published studies which found that although meiosis-specific and meiosis-related genes are generally conserved across all the major clades of eukaryotes (Schurko and Logsdon, 2008), absence of one or several of them is common and may be present in paralogs (see Chi, *et al.,* 2013 for details). For example, DMC1, HOP2, MER3, and MND1 are missing in *Caenorhabditis elegans* and *Drosophila melanogaster* (Masson and West, 2001; Pezza, *et al.*, 2007); while REC8 which is also involved in mitosis in *Tetrahymena thermophila* (Howard-Till, *et al.*, 2013) but absent in *Paramecium tetraurelia, Ichthyophthirius multifiliis, and Oxytricha trifallax.* Almost all eukaryotes capable of meiosis form synaptonemal complex (SC) (see Wettstein, *et al.,* 1984 for details) which are protein complexes

that consist of two parallel axial elements that form along the axes of paired chromosomal partners (Chi, *et al.,* 2014). They are thought to provide a tight physical link between homologous chromosomes and are involved in crossover regulation. Three meiosis-specific proteins (Hop1, Red1, and Zip1) are major components of the canonical SC, and none of these SC-related genes are present in *E. gracilis*. This is not unusual as only a few eukaryotes are able to perform crossing over in their absence (Chi, *et al.,* 2014).

The Bilobe was first discovered in trypanosomes, however, the cryptic nature of this cytoskeletal structure and the fact that it evaded detection for decades raise the question of whether there are homologous structures in other taxa (Esson, *et al.,* 2012). For a clear understanding whether these bilobe associated proteins found in *E. gracilis* are true homologs to the trypanosomatids counterparts since bioinformatics analysis as well as existing methods of cytoskeletal protein identification such as generation of monoclonal antibody panels, proteomics, affinity purification, and yeast two-hybrid screens, all have drawbacks (Morriswood, *et al.,* 2013), Esson, *et al.,* 2012, have proposed that examining whole mounts of the flagella and associated structures, especially using Proximity-dependent biotin identification (BioID) (Morriswood, *et al.,* 2013) in euglenozoan taxa such as *Leishmania*, *Bodo*, *Diplonema*, and *Euglena* might be a way to resolve this. If indeed the centrin and bilobe associated proteins are present in *E. gracilis*, then it may not be playing a role in parasitic lifestyle as previously proposed by Esson, *et al.,* 2012. The presence of the flagellar pocket collar protein BILBO1 indirectly supports the notion that the flagellar pocket collar and the bilobe are non distinct structures (Esson, *et al.,* 2012). In yeast and mammals, centrins and Polo-like Kinase (PLK) are important for organelle duplication and/or subsequent cell division (Zhou, *et al.,* 2010). The absence of PLK in *E. gracilis* suggests these functions are possibly reduced.

In *T. brucei*, and other protistan parasites, the cell surface is crucial for mediating host-parasite interactions and is instrumental to the initiation, maintenance and severity of infection (Jackson, *et al.,* 2013). In *E. gracilis*, the cell surface may well play sensory and defense roles. Two previously identified *Bodo* sequences belong to 2 different lineages, only one of which has been inherited by trypanosomatids.

Disulphide isomerases appear to be reduced in trypanosomatids compared to *Bodo*. The introduction of the *Euglena* repertoire allows to test whether this is due to loss in trypanosomatids or gain in *Bodo*. This gene loss or gain is also evident in other families as well as in the ABC transporters, and the combined effects of these gene loss or gain is reflected in the topologies of the phylogenies (see Figure 5.12).

The complex evolutionary history of plastid acquisition in *Euglena* suggests a possible equally involved relationship between the photosynthetic organelle and the gene complement encoded in the nucleus. The categories distributions, pathways, and complexes present are also consistent with other plastid containing organisms: The metabolic pathways expected in the plastid, based on previous studies are present in the predicted set with a few proteins missing (such as glucose-6-phosphate isomerase in glycolysis/gluconeogenesis or 15-*cis*-phytoene desaturase in carotenoid biosynthesis pathway): these gaps are likely false negatives arising from incomplete sequence data, since both of these pathways function normally in *Euglena* (Figures, 5.13, 5.14, and 5.15) (Hallick, *et al.,* 1993).

Members of the euglenozoans carry a highly diverse array of mitochondrial genomes (Dobakova, *et al.,* 2015). Overall, 2629 protein groups were found and these also include the seven mitochondria resident protein-coding genes and single rRNA previously reported by Dobakova, *et al.,* 2015 (Figures 5.16, 5.17, and 5.18). The remaining are associated with the mitochondria and possibly imported from the nucleus (Dobakova, *et al.,* 2015, Perez, *et al.,* 2014). For example, the ATP synthase (complex V, atp6) which are missing in the mitochondria genome but found in the dataset.

*E. gracilis* is an exception amongst excavates, as evident from this work, both in terms of number of dynamins as well as their diversity, with dynamins from all three functional classes found (Table 5.18). The complete Class A dynamin might be a bi-functional protein involved in both vesicle scission as well as mitochondrial division. The single Class C dynamin is closely related to cytokinetic dynamins of Amoebozoans (Miyagishisma, *et al.,* 2008). Most excavates appear to have lost FtsZ at their root, however, an FtsZ-like tubulin family is present and/or found in kinetoplastids and *E. gracilis*, and lost Class B and Class C dynamins on multiple independent occasions.

The presence of many characterized members of the CAZy glycosyl hydrolase and transferase families suggest an active carbohydrate metabolism and paramylon synthesis (Cantarel, *et al.,* 2009) in *E. gracilis*, encoding a wide array of carbohydrate active enzymes. The enzymes likely to be involved in β-glucan synthesis and degradation can be predicted, based on similarity to those used in plants and fungi for cell wall metabolism (O'Neill, *et al.,* 2015). Many enzymes involved in the synthesis and degradation of complex plant cell wall type carbohydrates, including galactosidases, xylosidases and rhamnosidases, are encoded in the *Euglena* transcriptome, suggesting they are capable of digesting complex plant material.

The extreme size of the *E. gracilis* genome, together with a rather complex splicing machinery **(**Breckenridge, *et al.,* 1999, Milanowski, *et al.,* 2016), suggests unusual mechanisms may be present for organization of chromatin, mRNA processing and transcription. Regulatory mechanisms, for example, are expected to be rather more complex than those reported for kinetoplastids, as both *trans* and *cis* splicing are known to be present (Günzl, 2010, Michaeli, 2011). Furthermore, the ultrastructure of the *E. gracilis* nucleus suggests a rather unusual heterochromatin organization, where electron-dense regions of the nucleus appear as numerous foci throughout the nucleoplasm, rather than as predominantly peripheral, as is the case in trypanosomatids and most organisms (Figure 5.9), and inspection of the genome revealed multiple significant features.

It was found that the two known proteins (NUP1 and NUP2) of the nuclear lamina of trypanosomes are absent, as is the more widespread lamin system of metazoan and other lineages (Koreny and Field, 2016). By contrast, much of the nuclear pore complex appears to be well conserved with African trypanosomes (Figure 5.7), and *Euglena* possesses orthologs of trypanosomatid nuclear basket proteins Nup92 and Nup110. However, unlike African trypanosomes orthologs for DBP5 and Gle1 were identified, two proteins involved in mRNA export in mammalian, yeast and plant NPCs but absent from trypanosomes. This is consistent with the hypothesis that the absence of these proteins is connected to the loss of *cis*-splicing mechanisms in most kinetoplastids (Obado, *et al.,* 2016), but does indicate the presence of a hybrid NPC, with shared elements from trypanosomes and higher eukaryotes. Secondly, I

also observed that the *Euglena* kinetochore possesses both elements of the trypanosome-specific and more canonical structures (see Figure 5.11, Table 5.10, and subsequent paragraphs for discussions) (Akiyoshi and Gull, 2014, D'Archivio and Wickstead, 2017).

There are some contrast between the transcription initiation in trypanosomes and *E. gracilis*. For instance, the arrangement of functionally unrelated genes in units and the absence of classic Pol II promoters led to the belief that transcription initiation was not a key factor in regulating trypanosome mRNA production (Ekanayake and Sabatini, 2011) which is in contrast to *E. gracilis*. Significantly, control of protein expression level also appears to be predominantly post-transcriptional in *Euglena* (Van Assche, *et al.,* 2015, Araújo and Teixeira, 2011, and see below). Overall these observations suggest a highly unusual set of mechanisms are likely present in the *E. gracilis* nucleus, but that these also reflect the transition between the conventional kinetochores, lamins and nuclear pores into the more radical kinetoplastida state.

In the signal transduction analysis, the distribution of the top BLASTP hits shows a distribution across the eukaryotic kingdoms (Ebenezer, *et al.,* 2018, Appendix I), highlighting the diversity of possible origins of genetic material present in the *Euglena* genome, which have been integrated through its complex phenotypic and genotypic history (Ahmadinejad, *et al.*, 2007; O'Neilla, *et al.,* 2015). This data suggests that *E. gracilis* has a remarkable richness of channel regulation for signal transduction in response to diverse stimuli. The tubulin genes, which are eukaryotic cytoskeletal components, are present in *E. gracilis* and involved in cell division. *E. gracilis* possesses both the eukaryotic (tubulins) and prokaryotic (FtsZ-tubulin family) cell division machinery (Figure 5.10). The FtsZ tubulin genes showed homology to the non photosynthetic prokaryotic FtsZ tubulins as well as those found in other non photosynthetic lower eukaryotes suggesting that, in *E. gracilis*, the FtsZ tubulin may have originated from the proteobacteria (mitochondria organelle) during the endosymbiotic phenomenon (see Ebenezer, *et al.,* 2017 for an extensive review). The organization of the tubulins in *E. gracilis* provides a preliminary understanding into the cytoskeletal architecture of *E. gracilis*. The transporter analysis (Table 5.5) suggests an active expression of membrane transport proteins in *E.gracilis* via multiple routes that allow for nutrient uptake, establishment of ion gradients, efflux

metabolites, translocation of compounds from one intracellular compartment to another, and uptake and export metals (Scott, 2008).

### *5.7.2 Information storage and processing*

Most evidence for molecular and functional diversification among the translation components has been found in the eIF4 proteins, and all major eukaryotic lineages possess several paralog genes for members of the eIF4 families (Hernández, *et al.,* 2012). In contrast to the initiation step, the process of elongation is highly conserved among all forms of life (Hernández, *et al.,* 2012). eEF1A which is found in *E. gracilis* is lacking in a number of eukaryotic lineages. It is a canonical factor that delivers aa-tRNAs to the A-site of ribosomes during the elongation step (Hernández, *et al.,* 2012). Instead, they possess a related factor called elongation factor-like (EFL) protein that retains the residues critical for eEF1A (Keeling and Inagaki, 2004). It was later found that EFL-encoding species are scattered widely across eukaryotes and that *eEF1A* and *EFL* genes display mutually exclusive phylogenetic distributions. Thus, it is assumed that eEF1A and EFL are functionally equivalent (Keeling and Inagaki, 2004; Sakaguchi, *et al.,* 2009; Gile, *et al.,* 2009; Cocquyt, *et al.,* 2009; Noble, *et al.,* 2007; Kamikawa, *et al.,* 2008; Guo*, et al.,* 2010; Szymański*, et al.,* 2000). It is thought that *eEF1A* is ancestral to all extant eukaryotes and that a single duplication event in a specific lineage gave rise to *EFL*. eRF1 is universally present in eukaryotes and, with the exception of some vascular plants and some ciliates, eukaryotes contain only one *eRF1* gene (Kim, *et al.,* 2005; Moreira, *et al.,* 2002; Atkinson, *et al.,* 2008; Inagaki and Doolittle, 2001), however, this is not consistent with *E. gracilis* and other members of the excavates where eRF1 and eRF3 are present.

In eukaryotes, transcription initiation is a key regulatory point in controlling the level of gene expression, and there are some constrast between the transcription initiation in trypanosomes and *E. gracilis.* For instance, the arrangement of functionally unrelated genes in units and the absence of classic Pol II promoters led to the belief that transcription initiation was not a key factor in regulating trypanosome mRNA production (Ekanayake and Sabatini, 2011) which is in contrast to *E. gracilis*.

Representatives of the chromosome segregation machinery (kinetochores) found in *E. gracilis* are distinct from those present in the kinetoplastids – possessing a

conventional kinetochore as present in other members of the eukaryotes. However, KKT19 and KKT10 homologous were found in *E. gracilis* but will require additional experimental validation since the KKTs are unique to kinetoplastids. The kinetoplastids possess unusual unconventional kinetochore proteins known as the kinetoplastid kinetochores (KKT). The presence of the conventional type kinetochores in *E. gracilis* has further extended the root of these proteins to the base of the Euglenozoa (Akiyoshi and Gull, 2014). In trypanosomes, the exosomes perform a major nuclear function in rRNA, snoRNA and mRNA quality control; necessary for the degradation of incompletely spliced mRNAs, and the majority of the exosome is in the nucleus, rather than in the cytoplasm (Kramer, *et al.,* 2016). *E. gracilis* possess the full exosome compliment seen in trypanosomes, however, EAP1, EAP2, and EAP3 are missing from our dataset suggesting that these proteins may not be playing huge role in mRNA degradation. The presence of these exosome compliments in *E. gracilis* suggests it may be performing the same function as in trypanosomes. In *E. gracilis*, the snRNPs are highly conserved with multiple paralogous with high sequence similarity to their corresponding eukaryotic reference sets, suggesting complex RNA processing activities in *E. gracilis*. Similarly, some of the splicing factors (such as HuR, B, C & D (hnRNPs)) which are present in Opisthokonts are not all present in *E. gracilis* suggesting a loss in these factors in Euglenozoa.

While there is a strong KPAP1 candidate and the presence of KPAF in *E. gracilis*, MORF is absent, so are some members of the MRB1 proteins. The absence of the MORF family proteins in *E. gracilis* suggests is not associated with plastid containing organisms (Salavati, *et al.,* 2012). The absence of some core kinetoplastid editosome components in *E. gracilis* may suggest some variations in the RNA editing mechanism between the euglenids and the Kinetoplastids, or perhaps, that these core RNA editing machinery are too divergent to be easily detect by homology/orthology relationships. The Tripartite Attachment Complex which is unique to Kinetoplastids, and connects the Kinetoplast to the flagellar, is expectedly absent in *E. gracilis* since they do not have Kinetoplast.

*E. gracilis* has about 2 Argonaute proteins consisting of the AGO1 (PAZ-PIWI domain) and PIWI-like 1 domains. Each of these argonautes interacts with specific

classes of ncRNAs to potentially regulate distinct biological processes (Couvillion, *et al.,* 2009). *E. gracilis* also have identified members of the RNase III proteins and small ncRNAs (dsRNA binding/Sid-1), but lack the dsRBM, Eri-1 like nuclease, and Sid-1-like. In trypanosomatids, RNAi pathways have diverged and been lost throughout their evolution (Lye, *et al.,* 2010). This group of protists has two distinct types of argonaute proteins that are somewhat more similar to each other than to other eukaryotic proteins (Garcia, *et al.,* 2010) and have also been found in the *E. gracilis* genome. One type, referred to as AGO-tryp, is absent in several of the species of Trypanosomatids including *T. cruzi* (Garcia, *et al.,* 2010). A second type of argonaute protein in trypanosomatids (referred to as PIWI-tryp) seems to be present in all members of this group where a genome sequence is available (Garcia, *et al.,* 2010). PIWI-tryp is highly conserved and is expressed throughout the life cycle of *T. cruzi,* which suggests it plays an important biological role in trypanosomatids (Garcia, *et al.,* 2010). While the RNAi pathway mediated by AGO-tryp is likely associated with silencing of transposable elements, the function of PIWI-tryp is unclear, and it remains unclear what the impact of the loss of AGO-tryp had on the biology of trypanosomatids. *T. cruzi* and *T. brucei* are both parasitic trypanosomatids irrespective of the presence of AGO-tryp, and AGO-tryp possesses an ortholog in *E. gracilis*, which suggests the loss of this RNAi pathway in trypanosomatids had no impact on parasitism per se. I propose here that *E. gracilis* elaborates an active RNAi pathway machinery.

### *5.7.3 Metabolism and evolution*

The *E. gracilis* galactofuranosyl glycosyltransferase (beta GALFT) protein shows high similarity to the active LPG-specific GALFT encoded by LPG1 from *L. major* and other GALFTs found in other kinetoplastid species (Parodi, 1993; de Lederkremer and Colli, 1995). A highly abundant class of free GPI referred to as glycoinositolphospholipids (GIPLs) cover the surface of numerous trypanosomatids, and is also present in *E. gracilis*. In *T. cruzi*, the principal GIPL and major component at the surface of the insect stage are called lipopeptidophosphoglycan (LPPG) and contain Galf (de Lederkremer, *et al.,* 1980, 1991, 1993; Previato, *et al.,* 1990). The presence of Galf, Galft, and GIPLs in *E. gracilis* suggests that their biosynthesis play a crucial role in the elaboration of the *E. gracilis* surface coat (Tefsen, *et al.,* 2012).

In *Leishmania*, studies into the interaction between the parasite, vector, and host have uncovered an important missing ingredient during transmission. *Leishmania* actively adapt their sand fly hosts into efficient vectors by secreting Promastigote Secretory Gel (PSG), a proteophosphoglycan (PPG)-rich, mucin-like gel which accumulates in sand fly gut and mouthparts (Rogers, 2012). Orthologs of these PPGs (PPG1, PPG3, PPG4, PPG5) has been found in *E. gracilis*. It is not yet clear the utility of PPGs in *E. gracilis* even though it aids parasitism in *Leishmania*. For instance, PPG have been found to function similarly to mammalian mucins to protect the surface of developing promastigotes against proteolytic damage (Secundino, *et al.,* 2010), which may well be of utility in the response of *E. gracilis* to environmental stressors. Similarly, analysis of the orthogroup clusters using GO terms indicated the presence of differing predicted functionality between the clusters, with, for example increased regulatory function genes in green/secondary plastid orthogroups, whilst phosphorus metabolic processes predominated the orthogroups shared only with kinetoplastids.

*5.7.5 Significance of findings – adaptations, functions and applications*

The draft genome and a final transcriptome of *E. gracilis* was produced. In other to further understand the biology, a community based annotation (as described above) was embarked on which focused on the transcriptome data due to its quality, which is useful to understand the biology of this organism. Indications from this analysis suggests that the biology is hugely complex, propelling the question about the origin and significance of its genome; hybrid genome. In this section, I will discuss the significance of the findings from the *E. gracilis* genome and transcriptome community annotation with reference to adaptions, functions (biology), and applications.

*Complex endomembrane system is a function of complex cellular activities:* The presence of multiple copies of specific trafficking proteins suggest an increased organellar complexity and an active cellular transport via multiple routes and pathways. For instance, transport between the endoplasmic recticulum and the Golgi apparatus is mediated by several GTPases, particularly the Rab1 and Rab2, which function to co-ordinate cytoskeletal interactions and specificity in fusion with membranes at the *cis*-Golgi (Brighouse, *et al.,* 2010). This suggests that in *E. gracilis,* exit from the Golgi apparatus can take several routes (e.g. division of labour

or a common response to evolutionary pressure) into exocytic vesicles, plasma membrane and other membrane bound destinations. It is evident that understanding the endomembrane system also provide useful information about eukaryotic evolutionary history through the delineation of the origins of membrane trafficking specificity (see Brighouse, *et al.,* 2010 for details). The findings also suggest an increased level of internal compartmentalization following from increased cellular activities as seen in the trafficking genes. While the Rab proteins have been used as a reference, several of the other endomembrane proteins play significant roles in trafficking and the significance of their findings is similar to the Rabs.

*Evidence of meiotic cell division machinery:* Reproductive analysis of meiosis and mitosis suggests a co-occurrence of sexual (they have sex) and asexual reproduction in *E. gracilis*. The presence of Kinetochores (involved in chromosome segregation) and core meiotic genes (in this work, though will require experimental evidences) provide evidences for mitosis and meiosis in *E. gracilis* respectively. Mitosis has also been extensively reported in *E. gracilis* (see Chapter 1), however, reports on meiosis (this chapter) has not previously been reported. The co-occurrence of these two forms of reproduction in *E. gracilis* further supports the hypothesis that the nucleus may not have been derived from the endosymbiotic relationship between the cyanobacteria (chloroplast) and the proteobacteria (mitochondria). This is obvious since mitosis only occurs in eukaryotic cells, and that prokaryotes reproduce by a process called binary fission. It is also not yet clear why *E. gracilis* possesses these two forms of reproduction (though this is universal in eukaryotes) and mechanisms, or if the utility of these varying forms of reproduction are mutually exclusive, or perhaps, what propels the switch from one form of reproduction to the other. One possible explanation for this would be in the conservation and maximization of energy and time since sexual reproduction requires more energy and time consuming than asexual reproduction (Neiman, *et al.,* 2010). A combined utility of these two forms of reproduction could also be that while meiosis maintains chromosome numbers and ensures diversity with cell populations (of relevance in natural selection), mitosis would play roles in cell development and growth, cell replacement, and cell regeneration. A combined presence of mitosis and meiosis is uncommon among the Kinetoplastids but evident in the European orchid, *Epipogium aphyllum* (Krawczyk, *et al.,* 2016).

*Bilobe structure in Euglena may play similar role as in Trypanosomes:* A particularly enigmatic feature that has recently been discovered in *T. brucei* is the bilobe – it localizes near the flagellar pocket and has been proposed to mediate biogenesis of the Golgi apparatus (Esson, *et al.,* 2012), forming both stable and dynamic association with other cytoskeletal components (and other membrane-bound organelles) including the basal bodies (tripartite attachment complex, kinetoplast) that seed the flagellum and the flagellar pocket collar that is critical for flagellar pocket biogenesis (Gheiratmand, *et al.,* 2013). A major function of the bilobe structure in *T. brucei* is to ensure that single-copied organelles (e.g. Golgi apparatus) duplicate and segregate in a highly coordinated fashion during cell cycle (Gheiratmand, *et al.,* 2013). The presence of all proteins (TbCentrin4, TbMORN1, and TbLRRP1) associated with the bilobe in *E. gracilis* (except for TbCentrin2), suggests that the bilobe structure in *E. gracilis* performs the same function as in *T. brucei*. While this analysis may present strong indications for the presence of bilobe structure in *E.. gracilis*, experimental analysis (biochemical, ultrastructural, phylogenetics) of a whole mount of *E. gracilis* flagellar may be required to further support this discovery (Esson, *et al.,* 2012). It has also been predicted that the bilobe structure may well play an adaptive role to a parasitic lifestyle (Esson, *et al.,* 2012), however, the presence of the bilobe associated proteins (orthologs/homologs) in freeliving *E. gracilis* suggests this is not case. It also suggests that the bilobe structure may have been evolutionarily derived and traces back to LECA. This is because it has previously been thought that the bilobe structure is only associated with the kinetoplastids (Esson, *et al.,* 2012), the discovery of these proteins in *E. gracilis* extends the root of their origin to the base of the Euglenozoa.

*Conventional, unconventional, and unusual biological systems:* In this paragraph (and the next), I will discuss the significance of the annotation discoveries associated with the nuclear system, and a brief highlight of the other systems. The results of the NPC components suggests that the nuclear system in *E. gracilis* is conventional in nature as this is consistent with the respective NPC subcomplexes and protein structures present in the reference sets of *T. brucei*, *A. thaliana*, *S. cerevisiae*, and *H. sapiens*, except that the nuclear lamina is absent. One unconventional observation, with respect to kinetoplastids (*T. brucei*) in the nuclear system is the presence of Dbp5 in *E. gracilis*, which is absent in kinetoplastids. In most eukaryotes

where this is present, Dbp5 mediates mRNA/proteins (mRNPs) export through the NPCs (Hodge, *et al.,* 2011). This occurs by the systematic triggering of the removal of mRNP proteins in a spatially controlled manner (Hodge, *et al.,* 2011). It remains to be seen how Kinetoplastids achieve this aim, as the precise sequence of events within this mechanism has not been fully defined in higher eukaryotes (Hodge, *et al.,* 2011).

In eukaryotes, heterochromatin which are conserved organizational feature of eukaryotic nuclei, localises to the peripheral of the nucleus (peripheral heterochromatin compartmentalization), providing a protected area for epigenetically silent genes and gene-poor DNA (Poleshko and Katz, 2014). In metazoan cells, the peripheral heterochromatin compartment is associated with the nuclear lamina, the protein meshwork at the inner edge of the nucleus. Heterochromatin-nuclear lamina interactions promote epigenetic gene silencing, which may drive many normal and diseased biological processes (Poleshko and Katz, 2014). A previously unstudied human protein, PRR14, participates in the tethering of heterochromatin to the inner nuclear periphery alongside heterochromatin protein 1 (HP1) (Poleshko and Katz, 2014). The results in this chapter suggests that there is an absence of a peripheral heterochromatin in *E. gracilis*, and that the heterochromatin organization is unusual and not condensed as previously reported. Following from this, the absence of nuclear lamins might explain the absence of peripheral heterochromatin in *E. gracilis*. Further evidence to prove this will be the interrogation of the *E. gracilis* genome with PRR14 and HP1 proteins which are associated with heterochromatin tethering. The absence of these proteins will further open up the question: Why would *E. gracilis* possess a non peripheral heterochromatin which is quite unusual with other eukaryotes?

Another conventional features of the nuclear system are seen in the results of tubulin genes, dynamins, and histones, suggesting that these are consistent in terms of functions and evolution with those found in the Kinetoplastids and other eukaryotes. The results of the kinetochore analysis is quite unusual in that *E. gracilis* possesses both the conventional and non conventional kinetochores (Akiyoshi and Gull, 2014; D' Archivio and Wickstead, 2016). Though this may require some experimental validation, it might indeed mean that these non conventional kinetochores have

evolved in LECA, or that they simply perform similar function as in Kinetoplastids. Other conventional biological systems are seen in the information processing system (translational apparatus, pre-initiation complex, mRNA metabolism, Exosomes, Spliceosomes and related proteins, Editosomes, TAC/CARP proteins, and RNAi pathway), organellar (plastid and mitochondria), and metabolism (PPG, Galf, and Carbohydrate Active Enzymes), and they perform similar function as in Kinetoplastids and other reference eukaryotes (see Results section of this Chapter).

*Surface and signal transduction play an adaptive and evolutionary role:* In trypanosomatids, amastins, a glycoprotein are found and expressed in the cell surface or plasma membrane (Jackson, 2010). In *Leishmania*, there is a substantial expansion of amastin repertoire that is directly associated with it's origin suggesting that amastin genes evolved novel functions crucial to cell functions in this species (Jackson, 2010). In *E. gracilis*, it is not yet clear the complete components of the surface (pellicle). A biotinylation experiment carried out in this work (data not shown) did not provide any protein evidence. While Leishmania seem to demonstrate strong functional characterization through evolutionary dynamics of amastin variation (Jackson, 2010), the in silico results in this work further supports this amastin divergence and extends it's ancestry to the base of the Euglenozoa. This is evident since the amastin gene family predicted in *E. gracilis* is different from those in trypanosomatids, suggesting the uniqueness of these genes in *E. gracilis*. In trypanosomatids, and more commonly in Leishmania, amastin play a role in host-parasite interactions (de Paiva, *et al.,* 2015). In *E. gracilis*, alongside transporters (Moreno-Sánchez, *et al.,* 2017) and signal transduction genes (Hader, *et al.,* 2006), it is anticipated that the uniqueness of the *E. gracilis* surface (amastins) play a role in response to external or environmental conditions.

*Evidence of endosymbiosis and hybrid genome origins:* In this work, *E. gracilis* have been shown to possess genes shared between itself and the prokaryotic endosymbiont as well as with other members of the eukaryotes. However, why would this unique organism accumulate several genes across prokaryotic and eukaryotic taxa? One possible explanation would be an active Lateral Gene Transfer (LGT) events in *E. gracilis*, which could play both evolutionary and functional roles. In prokaryotes, LGT are seen as a driving evolutionary force (Sieber, *et al.,* 2017).

For instance, sexual reproduction is considered an evolutionary advantage because offspring have increased genetic diversity that arose from the male and female gametes. Given that bacteria reproduce asexually, bacterial offspring lack genetic diversity from the sexual reproduction of two parents. In the absence of sexual reproduction, the transfer of DNA between organisms independent of sexual reproduction via lateral gene transfer (LGT) enables bacteria to increase genetic diversity and therefore potentially increase evolutionary fitness (Sieber, *et al.,* 2017). In *E. gracilis*, LGT may explain the reason behind the occurrence of sexual and asexual reproduction (see results and discussions above in this Chapter). LGT from prokaryotes to eukaryotes are also common, and involves the acquisition of nuclear genes in mitochondria and chloroplast (Sieber, *et al.,* 2017). This phenomenon will allow the host eukaryote, such as *E. gracilis*, to continue to perform the once ancient biological functions of the prokaryotes (cyanobacteria and proteobacteria) - photosynthesis and respiration respectively. This also has several adaptive functions such as the capability to adapt to both the presence and absence of light (see Chapter 6 of this work).

# CHAPTER SIX

## 6.0 GENE EXPRESSION STUDIES

### 6.1 INTRODUCTION

In Euglena, considerable advances have now been made in understanding the integration of both the mitochondrion and chloroplast with the host cell, and how their metabolic activities interact and are co-ordinated (O'Neill, *et al.,* 2015). *E. gracilis* are extremely metabolically flexible having acquired, by secondary endosymbiosis, a chloroplast with complex secondary metabolism (O'Neill, *et al.,* 2015) and also being highly resistant to noxious conditions. The presence of the chloroplast provides a major source of energy, and may also be connected with diurnal changes within the cell (Ebenezer, *et al.,* manuscript in preparation).

In the transition from light to dark conditions, and vice versa, *E. gracilis* shows several physiological and morphological changes (Gibbs, 1960; Wolken, 1956; Seigesmund, 1962; Pellegrini, 1980; Rocchetta et. al, 2007; Kivic and Vesk, 1972; Ben-Shaul, et. al., 1963; Lyman, et. al., 1961; Mollenhauer, et. al., 1967). For instance, in the absence of light, *E. gracilis* losses its chloroplast and undergo differentiation to form what is now known as proplastids which could adopt heterotrophic lifestyle – accumulating numerous paramylon granules and lipid inclusions, showing variably shaped and sized plastids, apparently lacking in ribosomes and showing a deeply disorganized membrane system (Vannini, 1983). When returned back to light, the photosynthetic apparatus, prothykaloids and chloroplast are restored and the organellar structures normalized (Siegesmund, *et. al.,* 1962; Wolken, 1956; Kivic and Vesk, 1972; Ben-Shaul, *et. al.,* 1963; Lyman, *et. al.,* 1961), with an increased levels (when compared to dark cells) of the three enzymes of the reductive pentose phosphate cycle: fructose-1,6" dlphosphate aldolase (class I), NADP-dependent glyceraldehyde-3-phosphate dehydrogenase and 3-phosphoglycerate kinase (Russel and Draffan, 1978).

In eukaryotic cells, gene expression follows the universal mechanism of a gene being transcribed into an mRNA that, in turn, is translated into a protein. Of recent, an unusual extension of this mechanism was discovered in *E. gracilis* (Houlné and Schantz, 1993*).* In gene expression in *E. gracilis,* novel mechanism of transport and

processing are utilized, suggesting that light controls the gene expression at a post-transcriptional level, either by mobilization onto the polysomes or at the translational step, depending on the greening conditions (Houlné and Schantz, 1993).

In *E. gracilis*, treatments with xenobiotics affecting prokaryotic translation, transcription and DNA replication can result in the loss of chloroplast DNA reflecting the cyanobacterial ancestry of *E. gracilis* chloroplast genome (see review by Krajcovic, Ebringer, and Schwartzbach 2002). This process is accompanied by irreversible elimination of functional chloroplasts (bleaching phenomenon). Similarly, *E. gracilis* grown in the dark exhibits a similar bleaching phenomenon, but this is completely reversible. In higher plants, plastid-derived signals can affect the expression of chloroplast-encoded genes (Puthiyaveetil, *et al.*, 2008) and the expression of nuclear genes that encode chloroplast proteins (Nott, *et al.*, 2006). For instance, a functional plastid transcription apparatus was shown to be necessary for the transcription of nucleus-encoded genes Lhcb (formerly cab gene, encoding light-harvesting chlorophyll a/b-binding protein of photosystem II—LHCPII) and RbcS (encoding ribulose-1,5-bisphosphate carboxylase/oxygenase small subunit) in wheat and barley (Rapp and Mullet, 1991). In Euglena, however, the situation appears to differ somewhat with respect to the influence of plastids on nuclear gene expression (Vesteg, *et al.,* 2009). Quite recently, it has been proven that mRNA levels of nucleus-encoded genes for chloroplast proteins in *E. gracilis* do not depend on either light or plastid function. For instance, Northern hybridization experiments showed that mRNA levels of nuclear photosynthetic genes Lhcb, RbcS and Pbgd (porphobilinogen deaminase) are similar in wild-type cells and non-photosynthetic *E. gracilis* white mutants (Vacula, *et al.*, 2001), and a constant levels of Apx transcripts were reported in *E. gracilis* during light adaptation while Apx mRNAs in white mutants and in the wild-type were comparable (Madhusudhan, *et al.,* 2003).

*E. gracilis* genes are differentially expressed under different conditions such as environmental stress, aerobic and anaerobic conditions and nutrition deprivation. For instance, using cDNA sequencing and microarray analysis, dos Santos Ferreira, *et al.,* 2007, found that 90 out of 610 identified ESTs changed expression levels in response to different stress treatments of Chromium, Streptomycin or darkness. Yoshida, *et al.,* 2016, described that in the transitioning from aerobic to anaerobic

conditions, approximately 5 % of the *E. gracilis* transcriptome are differentially expressed with pathway enrichment analysis suggesting an emphasis on photosynthesis, nucleotide metabolism, oxidative phosphorylation and fatty acid metabolism with insight that paramylon and wax ester metabolic pathways are regulated at post-transcriptional rather than the transcriptional level in response to anaerobic conditions. Vitamin B12, or cobalamin (Cbl), is an organometallic cofactor that supports the activities of enzymes in organisms, ranging from bacteria to humans, and it is required by *E. gracilis* for its growth (Yabuta, *et al.,* 2013).

Despite a considerable number of gene expression studies in *E. gracilis* involving light and dark environmental conditions, none of these have taken into account Next Generation Sequence (RNA-seq) approaches and/or a combination of RNA-seq and label-free proteomics quantification. In this chapter, for an indication into gene functions and physiological control, the gene expression of *E. gracilis* under light and dark environmental conditions will be discussed.

## 6.2 LIGHT AND DARK RNA-seq DIFFERENTIAL EXPRESSION ANALYSIS AND LABEL-FREE PROTEOMIC QUANTIFICATION

### *6.2.1 Light and dark adaptation of E. gracilis cells and RNA-seq analysis,*

RNA sequencing for light and dark *E. gracilis* cells was performed as previously described in Chapter 2. The RNA preparation method employed produced high quality RNA (Table 6.1). Three replicates of sequencing runs were produced each for light and dark conditions respectively, with a length of 100 bp, total clean bases of 6153 Mbp, and minimum Q20 of 98.59 % respectively (Table 6.2, and BioProject #: PRJNA310762). To ensure that *E. gracilis* cell have sufficiently been adapted to light and dark conditions, photographs, microscopy, growth analysis, protein electrophoresis and spectrophotometry analysis were taken showing light (greenish) and dark (grayish/colourless) adapted cells respectively (Figure 6.1, 6.2). The greenish and grayish (or colourless) oval structures (and lines or bands) in Figure 6.2 correspond to the chloroplast and proplastids respectively. Growth analysis suggests a more increased growth rate for cultures grown in light conditions than in dark conditions. Protein electrophoresis demonstrates the presence and absence of chlorophyll in the light and dark cultures respectively. Spectrophotometry analysis

demonstrates the presence of two peaks/spectra corresponding to chlorophyll b and a respectively in the light adapted cultures, while this is flattened or absent in the dark adapted cultures. Comparative analyses of the light and dark cells suggest significant correlations (Figure 6.2).

**Table 6.1:** Pre-RNA sequence quality control test for light and dark regime RNA purification

| S/No. | Sample name | Sample number | Library type | Concentration (ng/µl) | Volume (µl) | Total mass (µg) | OD260/280 | OD260/230 | Test result |
|---|---|---|---|---|---|---|---|---|---|
| 1. | Euglena RNA L1 | 8521512002690 | HiSeq Eukaryotic Transcriptome | 876 | 59 | 51.68 | 2.18 | 1.96 | Level A |
| 2. | Euglena RNA L2 | 8521512002691 | HiSeq Eukaryotic Transcriptome | 528 | 64 | 33.79 | 2.21 | 1.5 | Level A |
| 3. | Euglena RNA L3 | 8521512002692 | HiSeq Eukaryotic Transcriptome | 680 | 61 | 41.48 | 2.18 | 2.21 | Level A |
| 4. | Euglena RNA D1 | 8521512002693 | HiSeq Eukaryotic Transcriptome | 480 | 62 | 29.76 | 2.2 | 1.75 | Level A |
| 5. | Euglena RNA D2 | 8521512002694 | HiSeq Eukaryotic Transcriptome | 332 | 60 | 19.92 | 2.15 | 1.09 | Level A |
| 6. | Euglena RNA D3 | 8521512002695 | HiSeq Eukaryotic Transcriptome | 320 | 38 | 12.16 | 2.16 | 1.16 | Level A |

**Note:** The table represent the pre-sequencing result of *E. gracilis* light and dark regime RNA purifications. *Parameter descriptions:* **Sample name** refers to the alphabetical identifier of the sample. **Sample number** refers to the numerical identifier of the sample. **Library type** refers to the specific insert size sequenced. **Concentration**, **Volume**, and **Total Mass** refers to these respective parameters for each sample. OD260/280 and OD260/230 refers to the **Optical Density** at wavelengths of 260/280 and 260/230. **Test Result** refers to the classification of the pre-sequencing quality control; Level A means that the sample is qualified for DNA sequencing, and the amount of sample is sufficient for two or more libraries.

**Table 6.2:** Data summary of *E. gracilis* light and dark transcriptome sequencing experiment

| Molecule | Sample name | Sequence technology | Read library (bp) | Orientation | Length (bp) | Clean reads (M) | Clean bases (Mbp) | Total clean bases (Mbp) | Q20 (%) | GC (%) | **Phys Cov. (X) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| RNA | EuglenaRNAD1 | Illumina HiSeq2000 | Light and dark | Paired-overlap | 100 | 10.262 | 1026 | 6153 | 98.63 | 59.54 | 16.276 |
| | EuglenaRNAD2 | | | | | 10.274 | 1027 | | 98.67 | 59.24 | 16.294 |
| | EuglenaRNAD3 | | | | | 10.344 | 1034 | | 98.68 | 59.17 | 16.405 |
| | EuglenaRNAL1 | | | | | 10.239 | 1023 | | 98.59 | 59.69 | 16.239 |
| | EuglenaRNAL2 | | | | | 10.209 | 1020 | | 98.61 | 59.48 | 16.192 |
| | EuglenaRNAL3 | | | | | 10.205 | 1020 | | 98.60 | 59.55 | 16.185 |

**Column description: Sample name (Library ID)** refers to library or sequence indetification number that identify each sample; **Insert size** is the distance between the forward and reverse Mate pairs (read1 and read2); **Read length** is the length of the total reads (in bp); **Clean reads** refers to total reads after post sequencing quality control; **Clean bases** refers to total nucleotides (in base pair) after post sequencing quality control (sample filtration) for individual libraries; **Total clean bases** refers to total nucleotides (in base pair) after post sequencing quality control (sample filtration) for a libraries combined; **Q20 (%)** is the number of nucleotide with quality higher than 20/nucleotide (i.e. clean forward and reverse reads, R1 & R2), and **Q20** is the probability that a base is called incorrectly – probability of incorrect base is 1 in 100 with a call accuracy of 99 % (see: http://www.illumina.com/science/education/sequencing-quality-scores.html). **GC (%)** is the percentage of total nucleoties with guanine-cytocine bases. † Average read length. ***** = Mean value. ****** Read lengths are not in Megabases. **Phy Cov.** Is the physical coverage.

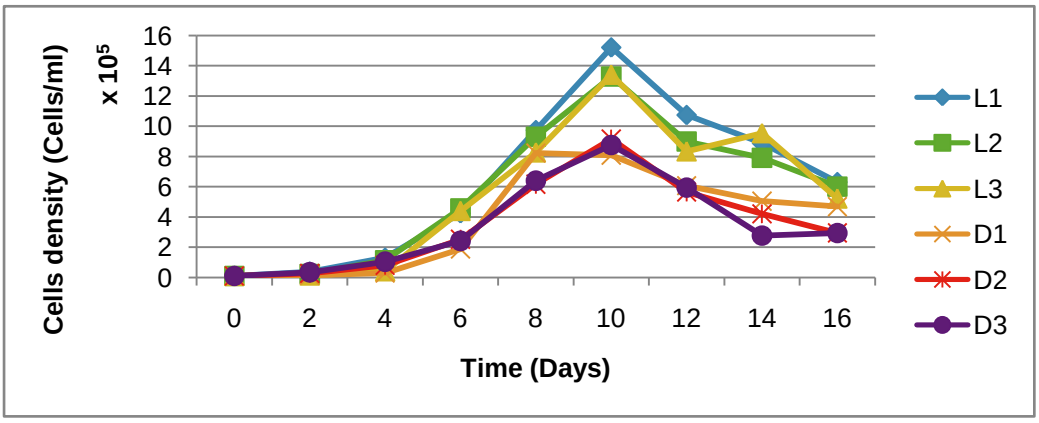



**Figure 6.1: Light is a catalyst for growth in *E. gracilis*.** The diagram shows the light and dark experiment growth curve. Growth adaptions or conditions were carried out for 16 days 'till cultures plateau. Each growth condition (light and dark) include 3 replicates represented by L1 (blue), L2 (green), L3 (yellow), D1 (amber), D2 (red), D3 (purple) respectively. Y axis = cell density in cells/ml (multiplied by a factor of $10^5$), and X axis = Time in days. Samples were analysed at optimum growth conditions.

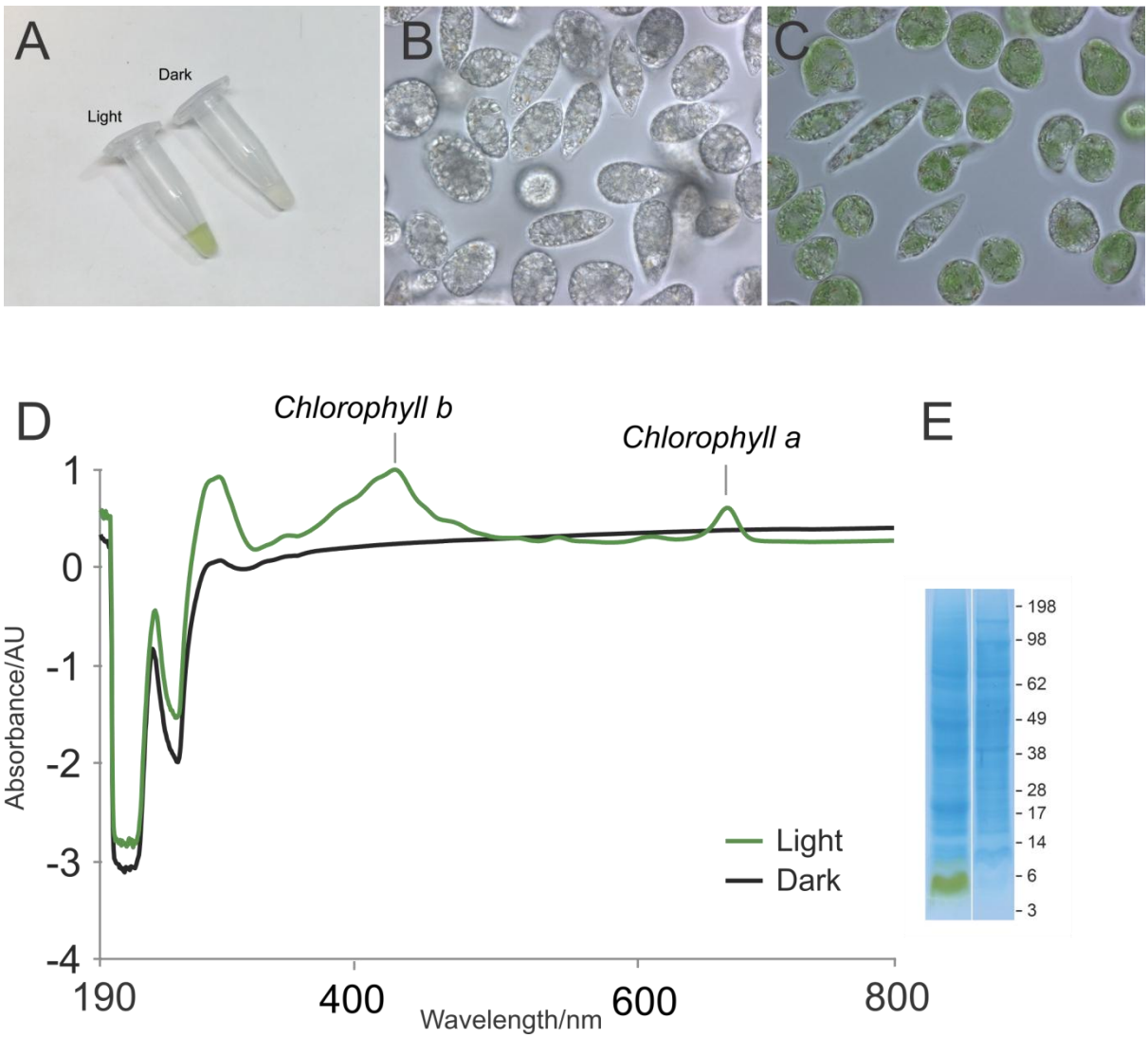



**Figure 6.2: Dark adapted cells lack chlorophyll.** All panels represent cells analysed following subculturing and then six days of continuous culture in the light or dark. Panel A: Samples of cultures demonstrating loss of pigments from the dark cultured cells. Panels B and C: Phase contrast images of cells after dark and light culture respectively. Panel D: UV-VIS absorbance spectra of cultures demonstrating the loss of characteristic absorbance peaks associated with chlorophyll and other pigments. Panel E: Coomassie-stained one dimension SDS-PAGE resolution of cells. Left lane are light and right dark cultured cells respectively, and molecular weight standards (kDa) are shown at right.

### 6.2.2 *Differential Expression analysis*

Differential Expression (DE) analysis was performed as previously described in Chapter 2, producing 66542 differentially expressed transcripts (Appendix II & III). Out of these 66542 differentially expressed transcripts, 64 were significantly expressed in the dark samples (this does not correlate to 61 upregulated proteins in the proteomics experiment) while 398 were significantly expressed in the light samples (this does not correlate to the 347 proteins upregulated in the proteomics experiment) using a 2-fold change criterion (-1.00 to -1.91 for dark regime, and 1.00 to 3.28 for light regime) (see Table 6.3, 6.4, Figure 6.6, and section 6.3; for an extensive details see Ebenezer, *et al.,* 2018, Appendix III, Table 1A-B). These 66542 differentially expressed transcripts were further reduced to 41045 by retaining those that are present in each of the 3 replicates in a specific state or condition as well as taking into account the infinite changes defined by the "$\log_2$ fold change" parameter (Appendix III), and the ratios for the 41045 transcripts were quantified (Figure 6.3, 6.4, and Appendix III, Table 1A-B).

There are significantly more transcripts than protein-coding genes or Open Reading Frame (ORF) within the data, as can be seen in the subsequent sub sections. A possible explanation is because not all manufactured transcripts were eventually translated. For instance, recent studies compared several yeast genomes and showed that many of the yeast ORFs believed to be protein-coding genes are actually not conserved even in closely related species. These studies suggested that the non-conserved ORFs do not actually code for proteins and were called Dubious ORF (DO) (Havilio, *et al.,* 2005). Another reason for more abundant transcript than protein-coding genes may be because more than one transcript are likely required to produce only one ORF. For instance, while it is expected that correctly spliced and reconstructed eukaryotic transcript should carry a single functional protein-coding gene, there are instances were two or more genes are produced by a single transcript (prokaryotes or some eukaryote with operon organization) (Tang, *et al.,* 2015) or more than one transcripts producing a single gene. When several protein-coding genes are predicted in a single transcript (or vice versa) one could think about either biological (e.g. presence of alternative isoforms) or technological reasons (e.g. erroneous sequencing or assembly) (Tang, *et al.,* 2015).

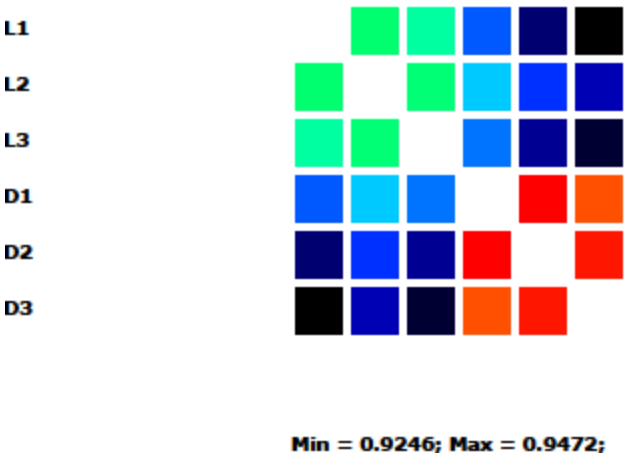



**Figure 6.3: Experimental triplicates are correlated**. Correlation heat map of light and dark RNA-seq Differential Expression (DE) analysis. L1, L2, and L3 represent light replicate experiments while D1, D2, and D3 represent dark replicate experiments respectively. Similar colourations of green, blue/black, or red, represent correlation significance. Minimum and maximum correlation values are 0.93 and 0.94 respectively. The diagram suggest that the three light replicate samples are more similar to each other than they are to the dark sample and the three dark replicates are more similar to each other than they are to the light.

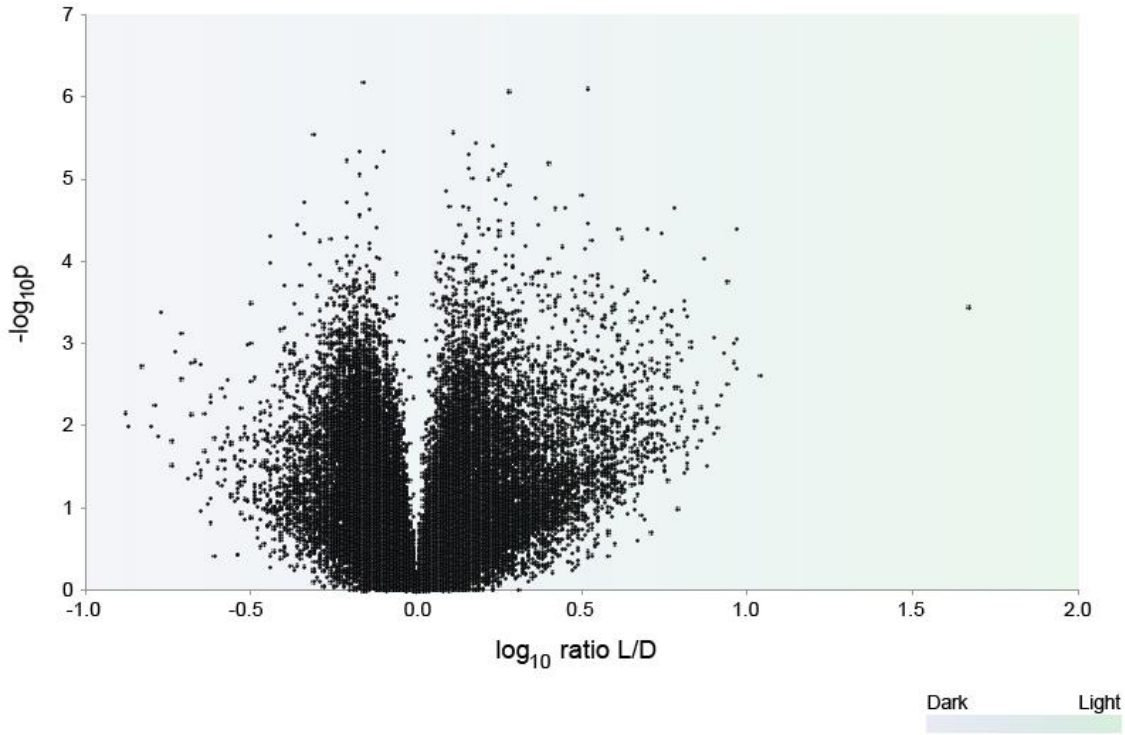



**Figure 6.4: Altered transcriptome of dark adapted cells.** Volcano plot of RNAseq data for cells maintained in the dark for six days against light grown cells. Data are plotted as $\log_{10}$ probability versus $\log_{10}$ of the ratio of intensities for dark/light transcript abundance. Data are the mean of triplicate RNA extractions. First half (left dotted) of the plot represent transcripts of dark adapted cells while the second half (right dotted) of the plot represent transcripts of light adapted cells. Thanks to Martin Zoltner for his kind help in generating this figure using Microsoft Excel and for Mark Field's help in reviewing it.

*6.2.3 Light and dark label-free proteomic quantification*

Label-free proteomic quantification was performed for light and dark adapted *E. gracilis* cells as previously described in Chapter 2, thanks to the Proteomics Facility at the University of Dundee, UK, for the Mass Spectrometry and Martin Zoltner for assistance with data analysis. Total protein (3 replicates for each environmental condition) were prepared from the *E. gracilis* light and dark adapted cells and separated by SDS-PAGE. Eight (8) fractions were obtained for each replicate in each condition respectively, and were subjected to Liquid chromatography–mass spectrometry/Mass Spectrometry (LC-MS/MS). This resulted in dataset of 48 samples which were analysed using MaxQuant (Cox and Mann, 2008) by searching the translated transcriptome. 80083 peptides were identified corresponding to 8661 distinct protein groups. Ratios for 4681 protein groups were quantified (Figure 6.5). A cohort of 384 protein groups was extracted belonging to only one state or condition (232 infinite changes specifically found in the light condition and 152 specifically found in the dark condition) (Figure 6.5). The protein groups not identified at the peptide level in each of the 3 replicates for one state were rejected from further analysis, and functional information corresponding to Biological Processes (BP), Cellular Component (CC), and Molecular Functions (MF) were assigned to all significant proteins identified (Table 6.3, 6.4, Figure 6.6, Appendix III). Further analysis of the Enzyme Code and KEGG pathways were performed (Figure 6.7, and Table 6.5). Combined, these analysis suggests the active presence of chloroplast and mitochondrial associated processes, with a higher proportion of amino acid metabolism and energy (Figure 6.6, 6.7 and Table 6.3, 6.4, 6.5)

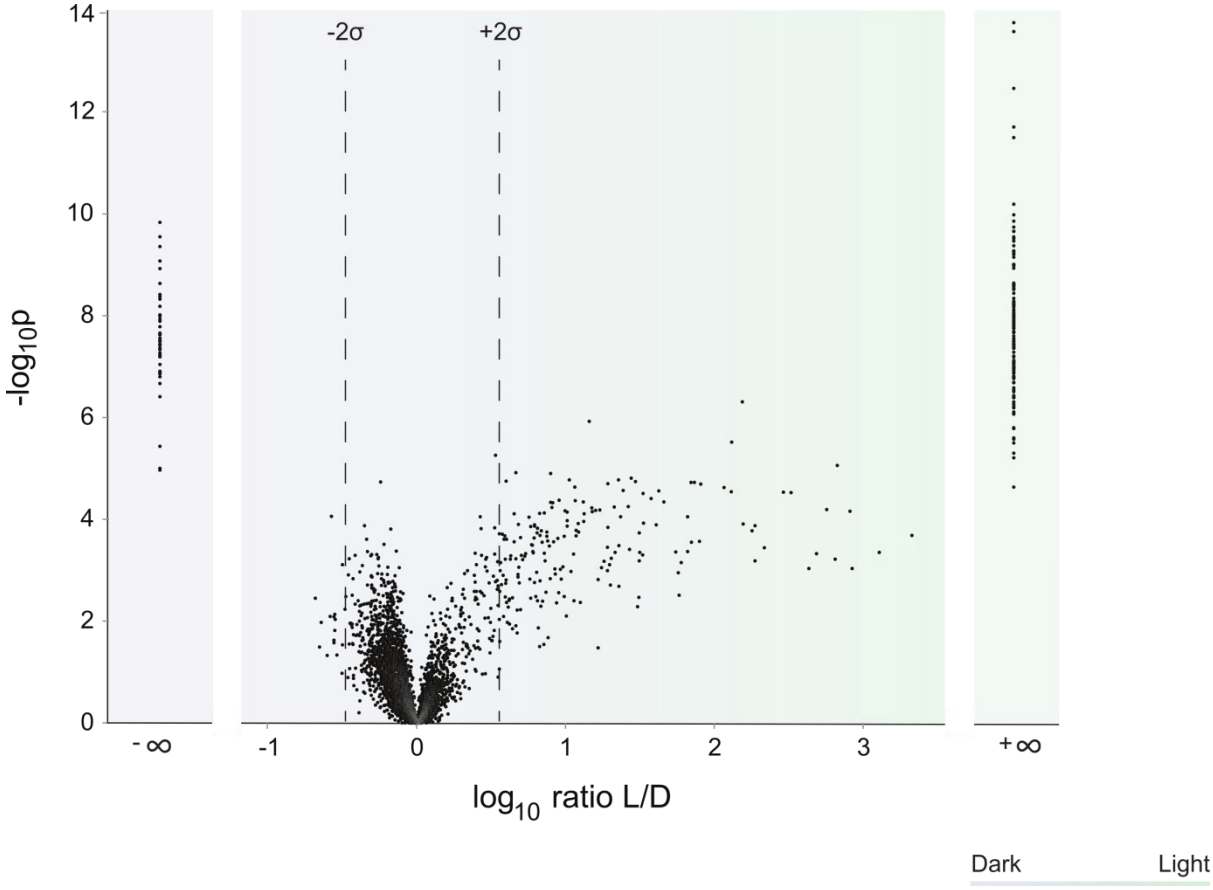



**Figure 6.5: Altered proteome of dark adapted cells**. Volcano plot of mass spectrometric proteomic analysis for cells maintained in the dark for six days against control cultures maintained in the light. Data are plotted as $\log_{10}$ probability versus $\log_{10}$ of the ratio of intensities for dark/light transcript abundance and are derived from triplicate analyses. The dotted lines indicate the significance threshold for altered abundance, and the groups at far left and far right are proteins only detected in one condition, i.e. infinite change. Thanks to Martin Zoltner for his kind help in generating this figure using Microsoft Excel and for Mark Field's help in reviewing it.

**Table 6.3:** Top upregulated proteins with functional information in the dark regime

| S/No. | Sequence ID | Gene Ontology number | BLAST Hit description |
|---|---|---|---|
| 1 | EG_transcript_10989 | GO:0008152, GO:0008270, GO:0046872, GO:0004386, GO:0003676 | gi\|361124091\|gb\|EHK96212.1\|putative ATP-dependent RNA helicase glh-4 [Glarea lozoyensis 74030] |
| 2 | EG_transcript_11489 | GO:0003676 | small RNA degrading nuclease 5 |
| 3 | EG_transcript_11553 | No annotation | |
| 4 | EG_transcript_13727 | GO:0006979, GO:0004601, GO:0016491, GO:0016020, GO:0016021, GO:0020037, GO:0055114, GO:0098869, GO:0016688 | gi\|229002753\|dbj\|BAC05484.2\|ascorbate peroxidase [Euglena gracilis] |
| 5 | EG_transcript_14442 | No annotation | |
| 6 | EG_transcript_14708 | GO:0016020, GO:0016021 | gi\|167535710\|ref\|XP_001749528.1\|hypothetical protein [Monosiga brevicollis MX1] |
| 7 | EG_transcript_14739 | No annotation | |
| 8 | EG_transcript_15222 | GO:0005515 | hypothetical protein RMCBS344292_15445 |
| 9 | EG_transcript_16204 | No annotation | |
| 10 | EG_transcript_16257 | GO:0008152, GO:0003824 | gi\|683461375\|gb\|KFZ55942.1\|hypothetical protein N321_07366, partial [Caprimulgus carolinensis] |
| 11 | EG_transcript_16757 | GO:0004190, GO:0006508, GO:0016020, GO:0016021 | gi\|675885768\|ref\|XP_009028386.1\|hypothetical protein HELRODRAFT_187510 [Helobdella robusta] |
| 12 | EG_transcript_16925 | GO:0008152, GO:0003824, GO:0003676 | gi\|567139595\|ref\|XP_006394828.1\|hypothetical protein EUTSA_v10003904mg [Eutrema salsugineum] |
| 13 | EG_transcript_17065 | GO:0008152, GO:0016491, GO:0055114 | gi\|496476580\|ref\|WP_009185341.1\|short-chain dehydrogenase [Cecembia lonarensis] |
| 14 | EG_transcript_18533 | GO:0003723, GO:0003676 | gi\|922868115\|gb\|KOO34586.1\|kh domain-containing protein [Chrysochromulina sp. CCMP291] |
| 15 | EG_transcript_19278 | GO:0003824, GO:0030246, GO:0005975 | gi\|496563941\|gb\|EON77991.1\|hypothetical protein ADIS_1530 [Lunatimonas lonarensis] |
| 16 | EG_transcript_19915 | No annotation | |
| 17 | EG_transcript_1993 | GO:0000166, GO:0046872, GO:0016020, GO:0016021, GO:0043231, GO:0016787, GO:0016887, GO:0015992, GO:1902600, GO:0008553, GO:0006810, GO:0006811, | gi\|159490822\|ref\|XP_001703372.1\|plasma membrane hydrogen ATPase [Chlamydomonas reinhardtii] |

| | | | |
|---|---|---|---|
| | | GO:0005524, GO:0005887, GO:0006754 | |
| 18 | EG_transcript_20465 | GO:0005515 | hypothetical protein |
| 19 | EG_transcript_21291 | No annotation | |
| 20 | EG_transcript_22379 | No annotation | |
| 21 | EG_transcript_22708 | GO:0008152, GO:0016874 | gi\|528220015\|gb\|EPY21792.1\|E3 ubiquitin-protein ligase CHFR [Strigomonas culicis] |
| 22 | EG_transcript_22899 | GO:0008171 | catechol O-methyltransferase |
| 23 | EG_transcript_24867 | GO:0003824 | peroxisomal bifunctional enzyme |
| 24 | EG_transcript_25896 | No annotation | |
| 25 | EG_transcript_2641 | GO:0008168, GO:0016740, GO:0032259 | gi\|818421176\|gb\|KKQ96760.1\|C-methyltransferase [Candidatus Levybacteria bacterium GW2011_GWA1_39_11] |
| 26 | EG_transcript_26745 | GO:0000166, GO:0005525, GO:0007264, GO:0005622 | gi\|302844586\|ref\|XP_002953833.1\|small Arf-related GTPase [Volvox carteri f. nagariensis] |
| 27 | EG_transcript_26940 | No annotation | |
| 28 | EG_transcript_27379 | GO:0004514, GO:0016757, GO:0004516, GO:0019363, GO:0009435, GO:0019358, GO:0016740, GO:0016874 | gi\|917634208\|ref\|WP_052203209.1\|nicotinate phosphoribosyltransferase [Corynebacterium riegelii] |
| 29 | EG_transcript_28262 | GO:0000166 | hypothetical protein |
| 30 | EG_transcript_3030 | No annotation | |
| 31 | EG_transcript_3171 | GO:0051087 | hypothetical protein SDRG_07330 |
| 32 | EG_transcript_32460 | No annotation | |
| 33 | EG_transcript_32537 | No annotation | |
| 34 | EG_transcript_33626 | GO:0005515 | hypothetical protein H257_03940 |
| 35 | EG_transcript_33977 | GO:0016020, GO:0016021 | gi\|910756073\|gb\|KNH03764.1\|hypothetical protein XU18_4840 [Perkinsela sp. CCAP 1560/4] |
| 36 | EG_transcript_35573 | No annotation | |
| 37 | EG_transcript_38082 | GO:0043231, GO:0051537 | gi\|924559594\|gb\|ALC44606.1\|CG1458 [Drosophila busckii] |
| 38 | EG_transcript_4196 | GO:0016758 | hypothetical protein PBRA_000926 |
| 39 | EG_transcript_42257 | GO:0006979, GO:0004601, GO:0046872, GO:0016491, GO:0020037, GO:0055114, GO:0098869, GO:0005576, GO:0042744 | gi\|224612181\|gb\|ACN60162.1\|peroxidase [Tamarix hispida] |
| 40 | EG_transcript_4227 | GO:0016567, GO:0004842 | gi\|118348690\|ref\|XP_001007820.1\|von willebrand factor type A |

| | | | |
|---|---|---|---|
| | | | domain protein, putative [Tetrahymena thermophila SB210] |
| 41 | EG_transcript_43582 | GO:0008233, GO:0045047, GO:0016020, GO:0016021, GO:0006465, GO:0030176, GO:0005787 | gi\|118784204\|ref\|XP_313570.3\|AGAP004296-PA [Anopheles gambiae str. PEST] |
| 42 | EG_transcript_45390 | No annotation | |
| 43 | EG_transcript_4580 | GO:0008152, GO:0003824, GO:0016020, GO:0016021 | gi\|488814079\|ref\|WP_002726485.1\|peptide synthetase [Phaeospirillum molischianum] |
| 44 | EG_transcript_4879 | GO:0008536, GO:0006886, GO:0005622 | gi\|290988107\|ref\|XP_002676763.1\|karyopherin beta [Naegleria gruberi] |
| 45 | EG_transcript_488 | GO:0016020, GO:0016021 | gi\|907095846\|gb\|KND01500.1\|hypothetical protein SPPG_03300 [Spizellomyces punctatus DAOM BR117] |
| 46 | EG_transcript_4887 | GO:0005515 | hypothetical protein PHYSODRAFT_247101 |
| 47 | EG_transcript_52895 | No annotation | |
| 48 | EG_transcript_57369 | GO:0005739, GO:0016020, GO:0016021, GO:0070469, GO:0055114, GO:0005743 | gi\|1174868\|sp\|P43266.1\|UCR9_EUGGRRecName: Full=Ubiquinol-cytochrome-C reductase complex subunit IX, mitochondrial |
| 49 | EG_transcript_57440 | No annotation | |
| 50 | EG_transcript_58502 | No annotation | |
| 51 | EG_transcript_59561 | GO:0016020, GO:0016021 | gi\|554937696\|gb\|ESL08680.1\|hypothetical protein TRSC58_03614 [Trypanosoma rangeli SC58] |
| 52 | EG_transcript_5961 | GO:0006397 | ---NA--- |
| 53 | EG_transcript_6094 | GO:0008152, GO:0000166, GO:0004386, GO:0005524, GO:0016787, GO:0003676 | gi\|761931188\|gb\|KIY52138.1\|DEAD-domain-containing protein [Fistulina hepatica ATCC 64428] |
| 54 | EG_transcript_61813 | GO:0000287, GO:0005737, GO:0004427, GO:0046872, GO:0006796, GO:0016787 | gi\|511100189\|ref\|WP_016330362.1\|inorganic pyrophosphatase [Thermus oshimai] |
| 55 | EG_transcript_63679 | No annotation | |
| 56 | EG_transcript_72309 | GO:0008152, GO:0046872, GO:0003824 | gi\|597570211\|ref\|XP_007288162.1\|peptidase M16 inactive domain-containing protein [Marssonina brunnea f. sp. 'multigermtubi' MB_m1] |
| 57 | EG_transcript_7608 | No annotation | |
| 58 | EG_transcript_7901 | GO:0008152, GO:0000166, GO:0004386, GO:0005524, GO:0016787, GO:0003676 | gi\|901802269\|gb\|KMZ62062.1\|ATP dependent RNA helicase [Zostera marina] |
| 59 | EG_transcript_8844 | No annotation | |
| 60 | EG_transcript_9170 | GO:0045892 | gi\|242080973\|ref\|XP_002445255.1\|hypothetical protein SORBIDRAFT_07g006910 [Sorghum bicolor] |

| | | | |
|---|---|---|---|
| 61 | EG_transcript_9625 | GO:1901838, GO:0000166, GO:0070062, GO:0003723, GO:0001525, GO:0042162, GO:0030529, GO:0016020, GO:0005938, GO:0036464, GO:0042802, GO:0008022, GO:0005654, GO:0044822, GO:0005730, GO:0005634, GO:0003676, GO:0045944 | gi\|874448046\|ref\|XP_005010398.2\|PREDICTED: nucleolin [Anas platyrhynchos] |

**Note:** The table describes the list of top upregulated proteins (61 proteins) in the dark regime corresponding to the proteomics experiment. Column descriptions are in the order: Sequence ID = *Euglena* sequence ID, Gene Ontology number = GO number for the corresponding protein. BLAST Hit description = description of BLAST top hit. Protein IDs with corresponding description “No annotation” under the GO number column and empty cells under BLAST Hit description column means that an annotation could not be assigned. Individual proteins are represented by multiple GO numbers (e.g. EG_trascript_9625), and single GO numbers are also shared by multiple proteins e.g. GO:0008152 (EG_transcript_10989 and EG_transcript_16257). In the case of multiple GOs against a specific proteins, the first GO on the series is the top hit or reference GO.

**Table 6.4:** Top upregulated proteins with functional information in the light regime

| S/No. | Sequence ID | Gene Ontology | BLAST Hit description |
|---|---|---|---|
| 1 | EG_transcript_10021 | GO:0045261, GO:0046961, GO:0046933, GO:0016787, GO:0015986 | gi\|299469805\|emb\|CBN76659.1\|ATP synthase gamma chain [Ectocarpus siliculosus] |
| 2 | EG_transcript_10056 | GO:0008152, GO:0003824, GO:0016740, GO:0004314 | gi\|255069969\|ref\|XP_002507066.1\|malonyl-CoA:ACP transacylase [Micromonas commoda] |
| 3 | EG_transcript_10193 | GO:0016021 | ATP-binding cassette sub-family C member 9 isoform X3 |
| 4 | EG_transcript_10194 | GO:0016021 | major facilitator superfamily |
| 5 | EG_transcript_10203 | GO:0015935 | 30S ribosomal S5, chloroplastic |
| 6 | EG_transcript_10277 | GO:0018193 | gi\|922866029\|gb\|KOO32969.1\|beta- aspartyl asparaginyl family [Chrysochromulina sp. CCMP291] |
| 7 | EG_transcript_10319 | GO:0003735, GO:0030529, GO:0005840, GO:0006412, GO:0005622 | gi\|302770833\|ref\|XP_002968835.1\|hypothetical protein SELMODRAFT_90107, partial [Selaginella moellendorffii] |
| 8 | EG_transcript_10323 | GO:0005524 | hypothetical protein |
| 9 | EG_transcript_10421 | GO:0016020, GO:0016021 | gi\|546310592\|ref\|XP_005714398.1\|unnamed protein product [Chondrus crispus] |
| 10 | EG_transcript_10484 | No annotation | |
| 11 | EG_transcript_10513 | No annotation | |
| 12 | EG_transcript_10575 | GO:0016491, GO:0055114 | gi\|308798893\|ref\|XP_003074226.1\|Glycine/D-amino acid oxidases-like (ISS) [Ostreococcus tauri] |
| 13 | EG_transcript_1060 | GO:0006950, GO:0051082, GO:0005524, GO:0006457 | gi\|528265585\|gb\|EPY39843.1\|TNF receptor-associated protein 1 [Angomonas deanei] |
| 14 | EG_transcript_10675 | GO:0000287, GO:0009507, GO:0015979, GO:0015977, GO:0016984, GO:0009536 | gi\|480512027\|gb\|AEW12959.2\|ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (chloroplast) [Colacium vesiculosum] |
| 15 | EG_transcript_10721 | No annotation | |
| 16 | EG_transcript_10778 | GO:0006096, GO:0004618, GO:0016310, GO:0016740, GO:0016301, GO:0005524 | gi\|551676214\|ref\|XP_005840861.1\|hypothetical protein GUITHDRAFT_91943 [Guillardia theta CCMP2712] |
| 17 | EG_transcript_10846 | No annotation | |
| 18 | EG_transcript_10877 | GO:0008152, GO:0016740, GO:0008080 | gi\|499899234\|ref\|WP_011579968.1\|N-acetyltransferase GCN5 [Chelativorans sp. BNC1] |
| 19 | EG_transcript_10888 | GO:0003824 | phosphatase 2C |

| | | | |
|---|---|---|---|
| 20 | EG_transcript_10984 | GO:0004853 | uroporphyrinogen decarboxylase, partial |
| 21 | EG_transcript_11021 | GO:0016020 | TPA_inf: chloroplast light-harvesting complex I precursor Lhca3 |
| 22 | EG_transcript_11176 | GO:0016701, GO:0051213, GO:0003868, GO:0016491, GO:0046872, GO:0055114, GO:0009072 | gi\|545704129\|ref\|XP_005704388.1\|4-hydroxyphenylpyruvate dioxygenase [Galdieria sulphuraria] |
| 23 | EG_transcript_11279 | No annotation | |
| 24 | EG_transcript_11323 | No annotation | |
| 25 | EG_transcript_11387 | GO:0000166, GO:0015995, GO:0006779, GO:0009507, GO:0015979, GO:0016851, GO:0016874, GO:0005524, GO:0009536 | gi\|830260407\|gb\|AKL82342.1\|chlorophyll biosynthesis (chloroplast) [Euglena gracilis var. bacillaris] |
| 26 | EG_transcript_11490 | GO:0004784 | unnamed protein product |
| 27 | EG_transcript_11539 | GO:0016491, GO:0016717, GO:0006629, GO:0016020, GO:0016021, GO:0055114, GO:0006633 | gi\|168023862\|ref\|XP_001764456.1\|predicted protein [Physcomitrella patens] |
| 28 | EG_transcript_11638 | GO:0016791 | phosphoglycolate phosphatase 1B, chloroplastic-like |
| 29 | EG_transcript_11739 | GO:0005524 | gi\|342360007\|gb\|AEL29575.1\|chloroplast rubisco activase [Chlorella pyrenoidosa] |
| 30 | EG_transcript_11779 | GO:0004316, GO:0008152, GO:0051287, GO:0016491, GO:0055114, GO:0006633 | gi\|522071708\|ref\|WP_020582917.1\|beta-ketoacyl-ACP reductase [Endozoicomonas elysicola] |
| 31 | EG_transcript_11790 | GO:0004550 | hypothetical protein Ctob_013751 |
| 32 | EG_transcript_11810 | No annotation | |
| 33 | EG_transcript_11885 | GO:0005515 | hypothetical protein SELMODRAFT_180340 |
| 34 | EG_transcript_12003 | No annotation | |
| 35 | EG_transcript_12015 | GO:0005515 | predicted protein |
| 36 | EG_transcript_12048 | No annotation | |
| 37 | EG_transcript_1207 | GO:0005506, GO:0046872, GO:0016705, GO:0020037, GO:0055114, GO:0004497 | gi\|219115858\|ref\|XP_002178724.1\|lutein deficient 1-like protein [Phaeodactylum tricornutum CCAP 1055/1] |
| 38 | EG_transcript_121 | GO:0030976 | unnamed protein product |
| 39 | EG_transcript_12119 | GO:0006096, GO:0004618, GO:0016310, GO:0016740, GO:0016301 | gi\|51860718\|gb\|AAU11483.1\|chloroplast phosphoglycerate kinase precursor [Euglena gracilis] |
| 40 | EG_transcript_12174 | GO:0005515 | photosystem II stability assembly factor HCF136 |
| 41 | EG_transcript_12275 | No annotation | |
| 42 | EG_transcript_12307 | No annotation | |

| | | | |
|---|---|---|---|
| 43 | EG_transcript_12327 | No annotation | |
| 44 | EG_transcript_12361 | GO:0005515 | hypothetical protein GUITHDRAFT_102686 |
| 45 | EG_transcript_12362 | No annotation | |
| 46 | EG_transcript_12440 | No annotation | |
| 47 | EG_transcript_12460 | No annotation | |
| 48 | EG_transcript_12610 | GO:0006779, GO:0004109, GO:0016020, GO:0016021, GO:0055114 | gi\|327493895\|gb\|AEA86534.1\|coproporphyrinogen oxidase [Euglena gracilis] |
| 49 | EG_transcript_12645 | No annotation | |
| 50 | EG_transcript_12751 | GO:0008152, GO:0003755, GO:0000413, GO:0016853, GO:0006457 | gi\|551653172\|ref\|XP_005829364.1\|hypothetical protein GUITHDRAFT_95747 [Guillardia theta CCMP2712] |
| 51 | EG_transcript_12753 | No annotation | |
| 52 | EG_transcript_12810 | GO:0008150, GO:0009507 | gi\|357114051\|ref\|XP_003558814.1\|PREDICTED: uncharacterized protein SYNPCC7002_A1590 [Brachypodium distachyon] |
| 53 | EG_transcript_12845 | GO:0009507, GO:0009536 | gi\|11467045\|ref\|NP_041952.1\|hypothetical protein EugrCp064 [Euglena gracilis] |
| 54 | EG_transcript_12982 | GO:0016020, GO:0016021 | gi\|224011820\|ref\|XP_002294563.1\|predicted protein [Thalassiosira pseudonana CCMP1335] |
| 55 | EG_transcript_13040 | GO:0016705, GO:0016491, GO:0055114, GO:0016117 | gi\|551676633\|ref\|XP_005841070.1\|lycopene beta cyclase [Guillardia theta CCMP2712] |
| 56 | EG_transcript_13066 | GO:0009507, GO:0046872, GO:0006098, GO:0019288, GO:0009055, GO:0009688, GO:0022900, GO:0051537, GO:0010103, GO:0051536 | gi\|159489964\|ref\|XP_001702961.1\|apoferredoxin [Chlamydomonas reinhardtii] |
| 57 | EG_transcript_13092 | No annotation | |
| 58 | EG_transcript_13115 | No annotation | |
| 59 | EG_transcript_13141 | GO:0008152, GO:0000166, GO:0006810, GO:0005524, GO:0016887 | gi\|302840511\|ref\|XP_002951811.1\|iron-sulfur cluster assembly protein [Volvox carteri f. nagariensis] |
| 60 | EG_transcript_1328 | GO:0008152, GO:0000166, GO:0004386, GO:0005524, GO:0016787, GO:0003676 | gi\|751841078\|emb\|CEL55341.1\|ATP-dependent RNA helicase DDX5/DBP2 [Rhizoctonia solani AG-1 IB] |
| 61 | EG_transcript_13343 | GO:0009654 | psbP family |
| 62 | EG_transcript_13435 | GO:0016491, GO:0055114, GO:0004324 | gi\|926794315\|ref\|XP_013905631.1\|ferredoxin--NADP+ reductase [Monoraphidium neglectum] |
| 63 | EG_transcript_1352 | GO:0008152, GO:0000166, GO:0005524, GO:0016887 | gi\|298707277\|emb\|CBJ25904.1\|ABC transporter [Ectocarpus |

| | | | |
|---|---|---|---|
| | | | siliculosus] |
| 64 | EG_transcript_13553 | GO:0015995, GO:0046872, GO:0016491, GO:0055114, GO:0015979, GO:0048529 | gi\|302833487\|ref\|XP_002948307.1\|copper target 1 protein [Volvox carteri f. nagariensis] |
| 65 | EG_transcript_13633 | GO:0016491, GO:0055114 | gi\|397615104\|gb\|EJK63221.1\|hypothetical protein THAOC_16135 [Thalassiosira oceanica] |
| 66 | EG_transcript_1376 | GO:0000166, GO:0005524, GO:0019538 | gi\|302841992\|ref\|XP_002952540.1\|hypothetical protein VOLCADRAFT_75431 [Volvox carteri f. nagariensis] |
| 67 | EG_transcript_13779 | GO:0008152, GO:0008270, GO:0016829, GO:0004089, GO:0015976 | gi\|58613427\|gb\|AAW79300.1\|chloroplast carbonic anhydrase [Heterocapsa triquetra] |
| 68 | EG_transcript_13834 | GO:0003824 | hypothetical protein |
| 69 | EG_transcript_13890 | GO:0018026, GO:0016279 | gi\|168002824\|ref\|XP_001754113.1\|predicted protein [Physcomitrella patens] |
| 70 | EG_transcript_13986 | GO:0005524 | adenylate kinase |
| 71 | EG_transcript_14000 | No annotation | |
| 72 | EG_transcript_14037 | GO:0031408, GO:0006636, GO:0044272, GO:0000096, GO:0050518, GO:0016740, GO:0009072, GO:0019216, GO:0016117, GO:0019748, GO:0016779, GO:0015995, GO:0015994, GO:0008299, GO:0009695, GO:0009117, GO:0009106, GO:0008652, GO:0006733, GO:0006546, GO:0006766, GO:0009416, GO:0009108 | gi\|159481752\|ref\|XP_001698942.1\|4-diphosphocytidyl-2C-methyl-D-erythritol synthase [Chlamydomonas reinhardtii] |
| 73 | EG_transcript_14068 | GO:0006508, GO:0006364, GO:0004222 | gi\|298711736\|emb\|CBJ32782.1\|conserved unknown protein [Ectocarpus siliculosus] |
| 74 | EG_transcript_14130 | GO:0016020, GO:0016021, GO:0043231, GO:0051537 | gi\|906458901\|gb\|KNC22034.1\|hypothetical protein FF38_14215 [Lucilia cuprina] |
| 75 | EG_transcript_14187 | GO:0005515 | predicted protein |
| 76 | EG_transcript_14312 | No annotation | |
| 77 | EG_transcript_14460 | GO:0016020, GO:0016021 | gi\|308805572\|ref\|XP_003080098.1\|unnamed protein product [Ostreococcus tauri] |
| 78 | EG_transcript_14464 | GO:0043085, GO:0031977, GO:0006098, GO:0005509, GO:0009523, GO:0010207, GO:0006364, GO:0009654, GO:0015979, GO:0019898, GO:0010103, GO:0009657 | gi\|226491834\|ref\|NP_001147590.1\|chloroplast oxygen-evolving complex/thylakoid lumenal 25.6kDa protein [Zea mays] |
| 79 | EG_transcript_14553 | No annotation | |
| 80 | EG_transcript_14654 | No annotation | |

| | | | |
|---|---|---|---|
| 81 | EG_transcript_14865 | GO:0016209, GO:0016491, GO:0016020, GO:0016021, GO:0055114, GO:0098869, GO:0051920 | gi\|594591713\|dbj\|BAO53977.1\|peroxiredoxin [Euglena gracilis] |
| 82 | EG_transcript_1495 | GO:0000166, GO:0005525, GO:0003746, GO:0009507, GO:0003924, GO:0006414, GO:0006412, GO:0009536, GO:0005622 | gi\|11466993\|ref\|NP_041900.1\|elongation factor Tu [Euglena gracilis] |
| 83 | EG_transcript_1542 | GO:0006950, GO:0051082, GO:0005524, GO:0006457 | gi\|452769133\|gb\|AGG11508.1\|hypothetical protein, partial [Trypanosomatidae sp. TS-2013] |
| 84 | EG_transcript_15491 | GO:0016851 | Magnesium chelatase, subunit H, N-terminal, partial |
| 85 | EG_transcript_15674 | GO:0000774 | GrpE protein homolog |
| 86 | EG_transcript_158 | GO:0000287, GO:0009507, GO:0046872, GO:0016491, GO:0018298, GO:0016020, GO:0016021, GO:0055114, GO:0051539, GO:0015979, GO:0016168, GO:0051536, GO:0009055, GO:0009579, GO:0009535, GO:0009522, GO:0009536 | gi\|830260425\|gb\|AKL82360.1\|photosystem I P700 apoprotein A1 (chloroplast) [Euglena gracilis var. bacillaris] |
| 87 | EG_transcript_15917 | GO:0005840 | 50S ribosomal L3, chloroplastic-like |
| 88 | EG_transcript_16010 | GO:0009507, GO:0016020, GO:0018298, GO:0016021, GO:0009579, GO:0009765, GO:0009523, GO:0015979, GO:0016168 | gi\|510388\|emb\|CAA43633.1\|light harvesting chlorophyll a /b binding protein of PSII [Euglena gracilis] |
| 89 | EG_transcript_16160 | GO:0005737, GO:0008233, GO:0006508, GO:0008236, GO:0004252, GO:0016787 | gi\|499390682\|ref\|WP_011078149.1\|ATP-dependent Clp protease proteolytic subunit [Vibrio vulnificus] |
| 90 | EG_transcript_16166 | GO:0019684, GO:0009535, GO:0009534, GO:0006364, GO:0010207, GO:0010206, GO:0003674, GO:0035304, GO:0009657 | gi\|159485408\|ref\|XP_001700736.1\|hypothetical protein CHLREDRAFT_187371 [Chlamydomonas reinhardtii] |
| 91 | EG_transcript_16296 | GO:0009507, GO:0016020, GO:0016021, GO:0009536 | gi\|830260409\|gb\|AKL82344.1\|hypothetical protein (chloroplast) [Euglena gracilis var. bacillaris] |
| 92 | EG_transcript_16400 | GO:0006779, GO:0016491, GO:0004729, GO:0055114 | gi\|327493899\|gb\|AEA86536.1\|protoporphyrinogen oxidase [Euglena gracilis] |
| 93 | EG_transcript_1656 | GO:0000166, GO:0006812, GO:0046872, GO:0070574, GO:0016020, GO:0098655, GO:0043231, GO:0016021, GO:0030001, GO:0071577, GO:0016787, GO:0008152, GO:0019829, GO:0015086, GO:0005385, GO:0005887 | gi\|156390845\|ref\|XP_001635480.1\|predicted protein [Nematostella vectensis] |
| 94 | EG_transcript_16567 | GO:0004392, GO:0055114, GO:0006788 | gi\|922864246\|gb\|KOO31539.1\|heme oxygenase 1 [Chrysochromulina sp. CCMP291] |
| 95 | EG_transcript_16570 | GO:0005506 | hypothetical protein GPECTOR_1g86 |

| | | | |
|---|---|---|---|
| 96 | EG_transcript_16722 | GO:0005515 | predicted protein |
| 97 | EG_transcript_1682 | GO:0008152, GO:0003824, GO:0046872 | gi\|585109176\|gb\|EWM27006.1\|Peptidase M16 [Nannochloropsis gaditana] |
| 98 | EG_transcript_16836 | GO:0004421 | hypothetical protein |
| 99 | EG_transcript_1707 | GO:0000166, GO:0005524, GO:0019538 | gi\|590670650\|ref\|XP_007038115.1\|CLPC [Theobroma cacao] |
| 100 | EG_transcript_17073 | GO:0031072, GO:0051082 | gi\|308808229\|ref\|XP_003081425.1\|unnamed protein product [Ostreococcus tauri] |
| 101 | EG_transcript_17203 | GO:0098599, GO:0008474, GO:0002084 | gi\|813102405\|ref\|XP_012193809.1\|hypothetical protein SPRG_00319 [Saprolegnia parasitica CBS 223.65] |
| 102 | EG_transcript_17297 | GO:0016020, GO:0016021 | gi\|495511008\|ref\|WP_008235653.1\|hypothetical protein [Richelia intracellularis] |
| 103 | EG_transcript_17391 | GO:0003723, GO:0003735, GO:0005739, GO:0005840, GO:0006412, GO:0005622 | gi\|168031515\|ref\|XP_001768266.1\|predicted protein [Physcomitrella patens] |
| 104 | EG_transcript_17545 | GO:0006417, GO:0045893, GO:0010319, GO:0019288, GO:0010207, GO:0015979, GO:0035304, GO:0010182, GO:0009773, GO:0007186, GO:0034660, GO:0009535, GO:0042793, GO:0009534, GO:0009941, GO:0045037, GO:0006364, GO:0045038, GO:0009532, GO:0042742, GO:0009902, GO:0009528, GO:0009657, GO:0006655 | gi\|357122407\|ref\|XP_003562907.1\|PREDICTED: protein THYLAKOID FORMATION1, chloroplastic [Brachypodium distachyon] |
| 105 | EG_transcript_17548 | No annotation | |
| 106 | EG_transcript_17571 | GO:0015995, GO:0016740, GO:0008168, GO:0032259, GO:0046406 | gi\|922865272\|gb\|KOO32364.1\|mg-protoporphyrin ix methyltransferase [Chrysochromulina sp. CCMP291] |
| 107 | EG_transcript_17576 | GO:0003779, GO:0005509, GO:0005938, GO:0005634, GO:0030479, GO:0006897, GO:0006886 | gi\|452837206\|gb\|EME39148.1\|hypothetical protein DOTSEDRAFT_38392 [Dothistroma septosporum NZE10] |
| 108 | EG_transcript_17636 | No annotation | |
| 109 | EG_transcript_17719 | GO:0031514 | gi\|159475433\|ref\|XP_001695823.1\|flagellar associated protein [Chlamydomonas reinhardtii] |
| 110 | EG_transcript_17723 | GO:0016705, GO:0016491, GO:0055114, GO:0016117 | gi\|551676633\|ref\|XP_005841070.1\|lycopene beta cyclase [Guillardia theta CCMP2712] |
| 111 | EG_transcript_17799 | GO:0005509, GO:0009523, GO:0009654, GO:0015979, GO:0019898 | gi\|356553956\|ref\|XP_003545316.1\|PREDICTED: oxygen-evolving enhancer protein 2, chloroplastic [Glycine max] |
| 112 | EG_transcript_17812 | GO:0031977, GO:0006636, GO:0019761, GO:0019288, GO:0019684, GO:0016020, GO:0016021, GO:0019344, GO:0010218, GO:0010207, GO:0016311, GO:0010206, | gi\|159489872\|ref\|XP_001702915.1\|predicted protein [Chlamydomonas reinhardtii] |

| | | | |
|---|---|---|---|
| | | GO:0015979, GO:0035304, GO:0010114, GO:0016117, GO:0015995, GO:0009773, GO:0009534, GO:0006364, GO:0009637, GO:0003993, GO:0009657 | |
| 113 | EG_transcript_17836 | GO:0046872, GO:0009055, GO:0051537, GO:0051536 | gi\|545375973\|ref\|XP_005652111.1\|ferredoxin, partial [Coccomyxa subellipsoidea C-169] |
| 114 | EG_transcript_1793 | GO:0003735, GO:0030529, GO:0015934, GO:0005840, GO:0006412 | gi\|223994385\|ref\|XP_002286876.1\|predicted protein [Thalassiosira pseudonana CCMP1335] |
| 115 | EG_transcript_18039 | GO:0016020, GO:0016021 | gi\|300244356\|gb\|ADJ93793.1\|chloroplast ferredoxin precursor [Euglena gracilis] |
| 116 | EG_transcript_18155 | No annotation | |
| 117 | EG_transcript_18210 | GO:0016020, GO:0016021 | gi\|698792322\|ref\|XP_009829193.1\|hypothetical protein H257_05882 [Aphanomyces astaci] |
| 118 | EG_transcript_18295 | No annotation | |
| 119 | EG_transcript_18309 | GO:0005840 | 50S ribosomal L13, chloroplastic |
| 120 | EG_transcript_18342 | GO:0016020, GO:0042132, GO:0016021, GO:0016311, GO:0042578, GO:0016787, GO:0005975 | gi\|99904205\|gb\|ABF68597.1\|chloroplast fructose-1,6-bisphosphatase [Euglena gracilis] |
| 121 | EG_transcript_18353 | No annotation | |
| 122 | EG_transcript_18719 | GO:0055114, GO:0008113 | gi\|545361125\|ref\|XP_005645876.1\|hypothetical protein COCSUDRAFT_37270 [Coccomyxa subellipsoidea C-169] |
| 123 | EG_transcript_1878 | GO:0000166, GO:0005737, GO:0042026, GO:0005524 | gi\|145348995\|ref\|XP_001418926.1\|chaperonin 60 beta chain, chloroplast [Ostreococcus lucimarinus CCE9901] |
| 124 | EG_transcript_18816 | GO:0005509, GO:0009523, GO:0009654, GO:0015979, GO:0019898 | gi\|612392145\|ref\|XP_007511593.1\|predicted protein [Bathycoccus prasinos] |
| 125 | EG_transcript_18865 | No annotation | |
| 126 | EG_transcript_19210 | No annotation | |
| 127 | EG_transcript_19233 | No annotation | |
| 128 | EG_transcript_19306 | GO:0009507, GO:0018298, GO:0016020, GO:0016021, GO:0009579, GO:0009765, GO:0009523, GO:0016168, GO:0015979 | gi\|157965835\|gb\|ABW06954.1\|chloroplast light-harvesting complex II protein precursor [Euglena gracilis] |
| 129 | EG_transcript_1939 | GO:0005506, GO:0008299, GO:0055114, GO:0016114, GO:0046429 | gi\|612390679\|ref\|XP_007510861.1\|predicted protein [Bathycoccus prasinos] |
| 130 | EG_transcript_19396 | No annotation | |
| 131 | EG_transcript_19863 | GO:0015035, GO:0046872, GO:0045454, GO:0009055, | gi\|504565180\|ref\|WP_014752282.1\|monothiol glutaredoxin, Grx4 |

| | | | |
|---|---|---|---|
| | | GO:0055114, GO:0051537, GO:0005623, GO:0051536 | family [Advenella kashmirensis] |
| 132 | EG_transcript_19892 | GO:0005524 | P-loop containing nucleoside triphosphate hydrolase |
| 133 | EG_transcript_19916 | GO:0005840 | 50S ribosomal L19 |
| 134 | EG_transcript_20060 | No annotation | |
| 135 | EG_transcript_2012 | GO:0016310, GO:0004672, GO:0016301, GO:0005524, GO:0006468 | gi\|545368684\|ref\|XP_005649120.1\|kinase-like protein [Coccomyxa subellipsoidea C-169] |
| 136 | EG_transcript_202 | GO:0015995, GO:0009058, GO:0016851, GO:0016874 | gi\|1011382130\|ref\|WP_062294880.1\|magnesium chelatase [Nostoc piscinale] |
| 137 | EG_transcript_20306 | GO:0008152, GO:0008270, GO:0016829, GO:0004089, GO:0015976 | gi\|58613427\|gb\|AAW79300.1\|chloroplast carbonic anhydrase [Heterocapsa triquetra] |
| 138 | EG_transcript_20407 | GO:0005840 | chloroplast ribosomal L4 precursor |
| 139 | EG_transcript_20472 | GO:0000023, GO:0043085, GO:0006098, GO:0019252, GO:0009535, GO:0009941, GO:0009534 | gi\|573957305\|ref\|XP_006660867.1\|PREDICTED: rhodanese-like domain-containing protein 14, chloroplastic isoform X1 [Oryza brachyantha] |
| 140 | EG_transcript_20496 | GO:0008233, GO:0006508, GO:0008236, GO:0016787 | gi\|504986470\|ref\|WP_015173572.1\|C-terminal processing peptidase-2 [Geitlerinema sp. PCC 7407] |
| 141 | EG_transcript_20679 | GO:0003723, GO:0003735, GO:0030529, GO:0019843, GO:0005840, GO:0006412, GO:0005622 | gi\|493575931\|ref\|WP_006529066.1\|50S ribosomal protein L9 [Gloeocapsa sp. PCC 73106] |
| 142 | EG_transcript_20725 | No annotation | |
| 143 | EG_transcript_20841 | GO:0004602, GO:0006979, GO:0004601, GO:0016491, GO:0055114, GO:0098869 | gi\|351727154\|ref\|NP_001236895.1\|uncharacterized protein LOC100306570 [Glycine max] |
| 144 | EG_transcript_20863 | GO:0008152, GO:0016740 | gi\|255081496\|ref\|XP_002507970.1\|glutathione s-transferase [Micromonas commoda] |
| 145 | EG_transcript_20958 | GO:0016020, GO:0016021 | gi\|551634553\|ref\|XP_005820119.1\|hypothetical protein GUITHDRAFT_98431 [Guillardia theta CCMP2712] |
| 146 | EG_transcript_21028 | No annotation | |
| 147 | EG_transcript_21122 | GO:0005509, GO:0009523, GO:0009654, GO:0015979, GO:0019898 | gi\|926784492\|ref\|XP_013900720.1\|PsbP domain-containing protein 6 [Monoraphidium neglectum] |
| 148 | EG_transcript_21198 | No annotation | |
| 149 | EG_transcript_21224 | GO:0035556 | PDZ domain-containing 8 isoform X2 |
| 150 | EG_transcript_21284 | GO:0003755, GO:0000413, GO:0016853, GO:0006457 | gi\|551568844\|ref\|XP_005770377.1\|peptidyl-prolyl isomerase [Emiliania huxleyi CCMP1516] |

| | | | |
|---|---|---|---|
| 151 | EG_transcript_21485 | GO:0016021 | unknown protein |
| 152 | EG_transcript_21961 | GO:0009507, GO:0016491, GO:0045156, GO:0016020, GO:0016021, GO:0055114, GO:0015979, GO:0008150, GO:0009055, GO:0009535, GO:0009579, GO:0042651, GO:0009512, GO:0003674, GO:0009536, GO:0009767 | gi\|38604764\|sp\|Q84TU6.1\|PETD_EUGGRRecName: Full=Cytochrome b6-f complex subunit 4, chloroplastic |
| 153 | EG_transcript_22249 | GO:0005840 | hypothetical protein GPECTOR_10g892 |
| 154 | EG_transcript_22491 | GO:0006979, GO:0004601, GO:0016491, GO:0016020, GO:0016021, GO:0020037, GO:0055114, GO:0098869, GO:0016688 | gi\|229002753\|dbj\|BAC05484.2\|ascorbate peroxidase [Euglena gracilis] |
| 155 | EG_transcript_22652 | GO:0003723, GO:0003735, GO:0030529, GO:0019843, GO:0005840, GO:0006412, GO:0005622 | gi\|818483467\|gb\|KKR38094.1\|30S ribosomal protein S17 [Parcubacteria group bacterium GW2011_GWF2_40_10] |
| 156 | EG_transcript_22691 | GO:0016491, GO:0016020, GO:0016021, GO:0055114, GO:0051537, GO:0042128, GO:0008942 | gi\|255081861\|ref\|XP_002508149.1\|rieske [2Fe-2S] domain protein protein [Micromonas commoda] |
| 157 | EG_transcript_22745 | No annotation | |
| 158 | EG_transcript_22790 | GO:0042651, GO:0009523, GO:0009654, GO:0015979 | gi\|255083855\|ref\|XP_002508502.1\|photosystem II PsbR protein, chloroplast precursor [Micromonas commoda] |
| 159 | EG_transcript_23265 | GO:0016020, GO:0016021, GO:0006810 | gi\|551672884\|ref\|XP_005839198.1\|hypothetical protein GUITHDRAFT_84771 [Guillardia theta CCMP2712] |
| 160 | EG_transcript_23566 | GO:0045263, GO:0015992, GO:0015078, GO:0016021, GO:0006810, GO:0006811, GO:0015986 | gi\|545366936\|ref\|XP_005648338.1\|chloroplast ATP synthase subunit II [Coccomyxa subellipsoidea C-169] |
| 161 | EG_transcript_23694 | GO:0005737, GO:0016836, GO:0006633 | gi\|585105186\|gb\|EWM23904.1\|beta-hydroxyacyl-acp dehydratase precursor [Nannochloropsis gaditana] |
| 162 | EG_transcript_23917 | GO:0003735, GO:0005840, GO:0006412, GO:0005622 | gi\|552826410\|ref\|XP_005847014.1\|hypothetical protein CHLNCDRAFT_35672 [Chlorella variabilis] |
| 163 | EG_transcript_23949 | GO:0016857 | ribulose-phosphate 3-epimerase-like |
| 164 | EG_transcript_24057 | GO:0016020, GO:0016021 | gi\|302835946\|ref\|XP_002949534.1\|hypothetical protein VOLCADRAFT_104324 [Volvox carteri f. nagariensis] |
| 165 | EG_transcript_24310 | No annotation | |
| 166 | EG_transcript_24363 | GO:0046872 | gi\|615478074\|ref\|XP_007603251.1\|MYND finger [Colletotrichum fioriniae PJ7] |
| 167 | EG_transcript_24497 | GO:0016020, GO:0016021 | gi\|552826630\|ref\|XP_005847073.1\|hypothetical protein CHLNCDRAFT_134749 [Chlorella variabilis] |
| 168 | EG_transcript_2508 | GO:0009507, GO:0016020, GO:0018298, GO:0016021, GO:0009579, GO:0009523, GO:0009765, GO:0016168, | gi\|157965821\|gb\|ABW06947.1\|chloroplast light-harvesting complex I protein precursor [Euglena gracilis] |

| | | | |
|---|---|---|---|
| | | GO:0015979 | |
| 169 | EG_transcript_25199 | GO:0009507, GO:0018298, GO:0016020, GO:0016021, GO:0009579, GO:0009765, GO:0009523, GO:0016168, GO:0015979 | gi\|157965833\|gb\|ABW06953.1\|chloroplast light-harvesting complex I protein precursor [Euglena gracilis] |
| 170 | EG_transcript_2535 | GO:0006418, GO:0000166, GO:0005737, GO:0004812, GO:0004814, GO:0006420, GO:0016874, GO:0005524, GO:0006412 | gi\|398017578\|ref\|XP_003861976.1\|arginyl-tRNA synthetase, putative [Leishmania donovani] |
| 171 | EG_transcript_2565 | GO:0000166, GO:0005524, GO:0019538 | gi\|302841992\|ref\|XP_002952540.1\|hypothetical protein VOLCADRAFT_75431 [Volvox carteri f. nagariensis] |
| 172 | EG_transcript_25832 | GO:0003824, GO:0016020, GO:0006096, GO:0016021, GO:0016829, GO:0004332 | gi\|1150392\|emb\|CAA61911.1\|fructose-1,6-bisphosphate aldolase [Euglena gracilis] |
| 173 | EG_transcript_25934 | GO:0009654 | hypothetical protein GPECTOR_79g122 |
| 174 | EG_transcript_26043 | No annotation | |
| 175 | EG_transcript_2646 | GO:0042132 | plastid sedoheptulose-1,7-bisphosphatase, partial |
| 176 | EG_transcript_26562 | GO:0003743, GO:0003723, GO:0009507, GO:0043022, GO:0019843, GO:0006412, GO:0009536, GO:0006413 | gi\|113170451\|ref\|YP_717243.1\|InfA (chloroplast) [Ostreococcus tauri] |
| 177 | EG_transcript_26637 | No annotation | |
| 178 | EG_transcript_2664 | GO:0000166, GO:0003676 | gi\|397643182\|gb\|EJK75699.1\|hypothetical protein THAOC_02569 [Thalassiosira oceanica] |
| 179 | EG_transcript_26703 | No annotation | |
| 180 | EG_transcript_26804 | No annotation | |
| 181 | EG_transcript_27332 | No annotation | |
| 182 | EG_transcript_2756 | GO:0005524 | unnamed protein product |
| 183 | EG_transcript_27835 | GO:0016020, GO:0016021 | gi\|545367249\|ref\|XP_005648469.1\|periplasmic binding protein-like II [Coccomyxa subellipsoidea C-169] |
| 184 | EG_transcript_28 | GO:0008152, GO:0016491, GO:0008299, GO:0003824, GO:0055114, GO:0031177, GO:0016740, GO:0004421 | gi\|818211082\|gb\|AKG25413.1\|putative polyketide synthase [Hematodinium sp. SG-2015] |
| 185 | EG_transcript_28382 | GO:0005515 | LOW PSII ACCUMULATION 1, chloroplastic-like isoform X1 |
| 186 | EG_transcript_28442 | GO:0016491, GO:0055114 | gi\|219115856\|ref\|XP_002178723.1\|predicted protein [Phaeodactylum tricornutum CCAP 1055/1] |
| 187 | EG_transcript_28662 | GO:0008289 | gi\|145354609\|ref\|XP_001421572.1\|predicted protein [Ostreococcus lucimarinus CCE9901] |

| | | | |
|---|---|---|---|
| 188 | EG_transcript_2884 | GO:0003924 | Translation elongation initiation factor Ribosomal, beta-barrel |
| 189 | EG_transcript_2894 | GO:0005524 | dnaJ homolog 1, mitochondrial-like |
| 190 | EG_transcript_2899 | GO:0006605, GO:0000166, GO:0016020, GO:0015031, GO:0006810, GO:0017038, GO:0005524, GO:0005622, GO:0006886 | gi\|545373296\|ref\|XP_005651027.1\|protein translocase subunit secA, chloroplastic [Coccomyxa subellipsoidea C-169] |
| 191 | EG_transcript_29305 | GO:0009507, GO:0018298, GO:0016020, GO:0016021, GO:0009579, GO:0009523, GO:0009765, GO:0016168, GO:0015979 | gi\|459020\|gb\|AAA16605.1\|light harvesting chlorophyll a/b binding protein of PSII, partial [Euglena gracilis] |
| 192 | EG_transcript_29382 | GO:0000166, GO:0005737, GO:0042026, GO:0005524 | gi\|145348995\|ref\|XP_001418926.1\|chaperonin 60 beta chain, chloroplast [Ostreococcus lucimarinus CCE9901] |
| 193 | EG_transcript_29936 | GO:0003735, GO:0070181, GO:0019843, GO:0005840, GO:0006412 | gi\|159477657\|ref\|XP_001696925.1\|plastid ribosomal protein S6 [Chlamydomonas reinhardtii] |
| 194 | EG_transcript_2997 | GO:0008152, GO:0004802, GO:0003824, GO:0016020, GO:0016021 | gi\|146335205\|gb\|ABQ23342.1\|plastid transketolase [Euglena gracilis] |
| 195 | EG_transcript_3018 | GO:0045261, GO:0046961, GO:0046034, GO:0016820, GO:0016787, GO:0015986, GO:0015992, GO:0033178, GO:0015991, GO:0006810, GO:0006811, GO:0046933, GO:0005524 | gi\|545363691\|ref\|XP_005646958.1\|P-loop containing nucleoside triphosphate hydrolase protein [Coccomyxa subellipsoidea C-169] |
| 196 | EG_transcript_3047 | GO:0008152, GO:0003755, GO:0005789, GO:0005528, GO:0061077, GO:0009535, GO:0009534, GO:0042631, GO:0000413, GO:0016853, GO:0006457 | gi\|159483091\|ref\|XP_001699596.1\|peptidyl-prolyl cis-trans isomerase, FKBP-type [Chlamydomonas reinhardtii] |
| 197 | EG_transcript_30510 | GO:0009507, GO:0016020, GO:0016021, GO:0009523, GO:0015979 | gi\|300244372\|gb\|ADJ93801.1\|chloroplast photosystem II reaction center W precursor [Euglena gracilis] |
| 198 | EG_transcript_30720 | GO:0005840 | mitochondrial or chloroplast ribosomal L31 precursor |
| 199 | EG_transcript_3104 | GO:0008152, GO:0000166, GO:0005524, GO:0016887 | gi\|298710208\|emb\|CBJ26283.1\|ABC transporter, ATP-binding protein [Ectocarpus siliculosus] |
| 200 | EG_transcript_31070 | GO:0009507, GO:0016020, GO:0009579, GO:0009535, GO:0009522, GO:0015979, GO:0009538, GO:0009536 | gi\|159489252\|ref\|XP_001702611.1\|photosystem I 8.1 kDa reaction center subunit IV [Chlamydomonas reinhardtii] |
| 201 | EG_transcript_31122 | GO:0003755, GO:0000413, GO:0016853, GO:0006457 | gi\|294893560\|ref\|XP_002774533.1\|Peptidyl-prolyl cis-trans isomerase, putative [Perkinsus marinus ATCC 50983] |
| 202 | EG_transcript_3119 | GO:0008152, GO:0000166, GO:0003723, GO:0004386, GO:0016787, GO:0005634, GO:0005524, GO:0003676 | gi\|299115354\|emb\|CBN74178.1\|DEAD box helicase [Ectocarpus siliculosus] |
| 203 | EG_transcript_3146 | No annotation | |
| 204 | EG_transcript_3159 | No annotation | |

| | | | |
|---|---|---|---|
| 205 | EG_transcript_3196 | GO:0000166, GO:0005525, GO:0003746, GO:0003924, GO:0006414, GO:0005622 | gi\|397584410\|gb\|EJK52971.1\|hypothetical protein THAOC_27681 [Thalassiosira oceanica] |
| 206 | EG_transcript_3254 | GO:0008152, GO:0016208, GO:0003987, GO:0003824, GO:0016874, GO:0019427 | gi\|159484368\|ref\|XP_001700230.1\|acetyl CoA synthetase [Chlamydomonas reinhardtii] |
| 207 | EG_transcript_32888 | GO:0008152, GO:0003755, GO:0000413, GO:0016853, GO:0006457 | gi\|674594452\|emb\|CDS26820.1\|fk506 binding protein [Hymenolepis microstoma] |
| 208 | EG_transcript_33051 | GO:0016020, GO:0016021, GO:0005509, GO:0009523, GO:0009654, GO:0015979, GO:0019898 | gi\|585101757\|gb\|EWM21524.1\|psbp domain-containing protein chloroplastic-like protein [Nannochloropsis gaditana] |
| 209 | EG_transcript_3371 | GO:0005524 | heat shock 70B |
| 210 | EG_transcript_34237 | No annotation | |
| 211 | EG_transcript_3435 | GO:0006779, GO:0016491, GO:0050661, GO:0055114, GO:0006782, GO:0008883, GO:0033014 | gi\|327493883\|gb\|AEA86528.1\|glutamyl-tRNA reductase [Euglena gracilis] |
| 212 | EG_transcript_3493 | GO:0005515 | hypothetical protein GPECTOR_3g291 |
| 213 | EG_transcript_3515 | GO:0008152, GO:0003824, GO:0016874 | gi\|488762439\|ref\|WP_002685639.1\|propionyl-CoA synthetase [Beggiatoa alba] |
| 214 | EG_transcript_3641 | GO:0006508, GO:0016020, GO:0004222, GO:0005524 | gi\|302794532\|ref\|XP_002979030.1\|hypothetical protein SELMODRAFT_152929 [Selaginella moellendorffii] |
| 215 | EG_transcript_3654 | GO:0016020, GO:0006508, GO:0016021, GO:0004222, GO:0005524 | gi\|302833547\|ref\|XP_002948337.1\|hypothetical protein VOLCADRAFT_80122 [Volvox carteri f. nagariensis] |
| 216 | EG_transcript_3663 | GO:0005515 | intracellular chloride channel family |
| 217 | EG_transcript_3677 | No annotation | |
| 218 | EG_transcript_3688 | GO:0009522, GO:0015979, GO:0009538 | gi\|504986591\|ref\|WP_015173693.1\|photosystem I protein PsaD [Geitlerinema sp. PCC 7407] |
| 219 | EG_transcript_3765 | GO:0003723, GO:0003735, GO:0005840, GO:0006412, GO:0005622 | gi\|83584321\|gb\|ABC24935.1\|plastid 30S ribosomal protein S20 [Prototheca wickerhamii] |
| 220 | EG_transcript_3792 | GO:0006418, GO:0000166, GO:0005737, GO:0004812, GO:0006426, GO:0004820, GO:0005524 | gi\|922887354\|gb\|KOO52976.1\|glycyl-tRNA synthetase [Chrysochromulina sp. CCMP291] |
| 221 | EG_transcript_38684 | No annotation | |
| 222 | EG_transcript_3877 | GO:0052886, GO:0052887, GO:0016491, GO:0016719, GO:0055114, GO:0052889, GO:0016117 | gi\|493556155\|ref\|WP_006509674.1\|9,9'-di-cis-zeta-carotene desaturase [Xenococcus sp. PCC 7305] |
| 223 | EG_transcript_39172 | GO:0016020, GO:0016021 | gi\|298715709\|emb\|CBJ28206.1\|expressed unknown protein [Ectocarpus siliculosus] |
| 224 | EG_transcript_3948 | GO:0005506, GO:0046872, GO:0016491, GO:0016705, | gi\|585112948\|gb\|EWM30289.1\|cytochrome p450 [Nannochloropsis |

| | | | |
|---|---|---|---|
| | | GO:0020037, GO:0055114, GO:0004497 | gaditana] |
| 225 | EG_transcript_3956 | No annotation | |
| 226 | EG_transcript_3991 | GO:0046872, GO:0016491, GO:0055114, GO:0051537, GO:0010277, GO:0051536 | gi\|545358292\|ref\|XP_005644725.1\|PaO-domain-containing protein, partial [Coccomyxa subellipsoidea C-169] |
| 227 | EG_transcript_40006 | GO:0009507, GO:0016020, GO:0018298, GO:0016021, GO:0009579, GO:0009523, GO:0009765, GO:0016168, GO:0015979 | gi\|157965841\|gb\|ABW06957.1\|chloroplast light-harvesting complex II protein precursor [Euglena gracilis] |
| 228 | EG_transcript_4083 | GO:0008152, GO:0009570, GO:0009507, GO:0046872, GO:0016787 | gi\|566198955\|ref\|XP_002319540.2\|hypothetical protein POPTR_0013s02230g [Populus trichocarpa] |
| 229 | EG_transcript_41111 | GO:0005840 | putative uncharacterized protein |
| 230 | EG_transcript_41739 | GO:0016020 | TPA_inf: chloroplast light-harvesting complex II precursor Lhcbm3 |
| 231 | EG_transcript_4197 | No annotation | |
| 232 | EG_transcript_4275 | GO:0000166, GO:0009507, GO:0045261, GO:0046034, GO:0016020, GO:0016787, GO:0016820, GO:0015986, GO:0015992, GO:0033178, GO:0015991, GO:0009535, GO:0009579, GO:0006810, GO:0006811, GO:0005524, GO:0046933, GO:0006754, GO:0009536 | gi\|11467034\|ref\|NP_041941.1\|ATP synthase CF1 beta subunit [Euglena gracilis] |
| 233 | EG_transcript_4353 | GO:0004784 | superoxide dismutase |
| 234 | EG_transcript_4370 | GO:0030604, GO:0046872, GO:0016491, GO:0008299, GO:0070402, GO:0055114, GO:0016853 | gi\|209402475\|gb\|ACI45960.1\|putative plastid 1-deoxy-D-xylulose 5-phosphate reductoisomerase precursor [Pyropia yezoensis] |
| 235 | EG_transcript_4487 | GO:0000166, GO:0005737, GO:0042026, GO:0005524, GO:0006457 | gi\|159491478\|ref\|XP_001703692.1\|chaperonin 60A [Chlamydomonas reinhardtii] |
| 236 | EG_transcript_4530 | GO:0016020, GO:0016021, GO:0051205 | gi\|676388263\|ref\|XP_009037410.1\|hypothetical protein AURANDRAFT_11603, partial [Aureococcus anophagefferens] |
| 237 | EG_transcript_4570 | GO:0042132 | chloroplastic fructose-1,6-bisphosphataseII |
| 238 | EG_transcript_4631 | GO:0005886 | Peptidase S49 |
| 239 | EG_transcript_4685 | GO:0008152, GO:0000166, GO:0004386, GO:0016787, GO:0005524, GO:0003676 | gi\|756169266\|ref\|WP_042617393.1\|RNA helicase [Agrobacterium tumefaciens] |
| 240 | EG_transcript_4693 | GO:1901838, GO:0070062, GO:0003723, GO:0097421, GO:0001525, GO:0042393, GO:0030529, GO:0016020, GO:0005509, GO:0043066, GO:2000778, GO:2000232, GO:0036464, GO:0007283, GO:0032760, GO:0009986, GO:0042134, GO:0005102, GO:0043565, GO:0045944, | gi\|21750187\|dbj\|BAC03738.1\|unnamed protein product [Homo sapiens] |

| | | | |
|---|---|---|---|
| | | GO:0006897, GO:0003677, GO:0005515, GO:0000166, GO:0005737, GO:1990631, GO:0042162, GO:0005938, GO:0071222, GO:0035368, GO:0042802, GO:0008022, GO:0043236, GO:0044822, GO:0005654, GO:0005730, GO:0001650, GO:0003697, GO:0001651, GO:0005634, GO:0003676 | |
| 241 | EG_transcript_4707 | GO:0004190, GO:0006508, GO:0016020, GO:0016021 | gi\|698780165\|ref\|XP_009823126.1\|hypothetical protein H257_01558 [Aphanomyces astaci] |
| 242 | EG_transcript_4720 | No annotation | |
| 243 | EG_transcript_4724 | GO:0004784 | unnamed protein product |
| 244 | EG_transcript_4741 | GO:0006779, GO:0004109, GO:0016020, GO:0016021, GO:0055114 | gi\|327493895\|gb\|AEA86534.1\|coproporphyrinogen oxidase [Euglena gracilis] |
| 245 | EG_transcript_4868 | No annotation | |
| 246 | EG_transcript_4888 | GO:0005737, GO:0016491, GO:0045454, GO:0016020, GO:0016021, GO:0019430, GO:0004791, GO:0055114 | gi\|594591723\|dbj\|BAO53982.1\|NADPH-dependent thioredoxin reductase [Euglena gracilis] |
| 247 | EG_transcript_4918 | GO:0006418, GO:0000166, GO:0005737, GO:0004812, GO:0050561, GO:0004818, GO:0016874, GO:0016876, GO:0043039, GO:0000049, GO:0005524, GO:0006412, GO:0006424 | gi\|298715644\|emb\|CBJ28170.1\|Glutamyl-tRNA Synthetase, chloroplast precursor [Ectocarpus siliculosus] |
| 248 | EG_transcript_4984 | GO:0030170, GO:0003824, GO:0008483, GO:0042286, GO:0033014 | gi\|193890955\|gb\|ACF28631.1\|glutamate semialdehyde synthase [Amphidinium carterae] |
| 249 | EG_transcript_5124 | GO:0008152, GO:0016747, GO:0003824, GO:0016740, GO:0006633 | gi\|545355322\|ref\|XP_005643461.1\|3-oxoacyl-synth [Coccomyxa subellipsoidea C-169] |
| 250 | EG_transcript_516 | GO:0016851 | magnesium-chelatase H subunit |
| 251 | EG_transcript_5175 | GO:0018193 | gi\|922866029\|gb\|KOO32969.1\|beta- aspartyl asparaginyl family [Chrysochromulina sp. CCMP291] |
| 252 | EG_transcript_52231 | GO:0008152, GO:0004792, GO:0016740 | gi\|748141586\|ref\|WP_039716800.1\|thiosulfate sulfurtransferase [Scytonema millei] |
| 253 | EG_transcript_5228 | GO:0042651 | cytochrome b6-f complex Fe-S subunit |
| 254 | EG_transcript_5279 | GO:0003824, GO:0016020, GO:0006096, GO:0016021, GO:0016829, GO:0004332 | gi\|1150392\|emb\|CAA61911.1\|fructose-1,6-bisphosphate aldolase [Euglena gracilis] |
| 255 | EG_transcript_5309 | GO:0016491, GO:0055114 | gi\|224006652\|ref\|XP_002292286.1\|predicted protein [Thalassiosira pseudonana CCMP1335] |

| | | | |
|---|---|---|---|
| 256 | EG_transcript_5335 | GO:0055085, GO:0006813, GO:0098655, GO:0016020, GO:0016021, GO:0008324 | gi\|546315769\|ref\|XP_005715956.1\|unnamed protein product [Chondrus crispus] |
| 257 | EG_transcript_5362 | GO:0016020, GO:0016021 | gi\|922887238\|gb\|KOO52893.1\|ABC1 protein [Chrysochromulina sp. CCMP291] |
| 258 | EG_transcript_5493 | GO:0006418, GO:0000166, GO:0004812, GO:0005524, GO:0003676 | gi\|223995725\|ref\|XP_002287536.1\|predicted protein [Thalassiosira pseudonana CCMP1335] |
| 259 | EG_transcript_5539 | GO:0008168 | bifunctional demethylmenaquinone methyltransferase 2-methoxy-6-polyprenyl-1,4-benzoquinol methylase |
| 260 | EG_transcript_5655 | No annotation | |
| 261 | EG_transcript_56611 | No annotation | |
| 262 | EG_transcript_5678 | No annotation | |
| 263 | EG_transcript_5684 | GO:0016020, GO:0006629, GO:0016021 | gi\|308813329\|ref\|XP_003083971.1\|delta 12 fatty acid desaturase (ISS) [Ostreococcus tauri] |
| 264 | EG_transcript_5740 | No annotation | |
| 265 | EG_transcript_5780 | GO:0006979, GO:0004601, GO:0016491, GO:0016020, GO:0016021, GO:0020037, GO:0055114, GO:0098869, GO:0016688 | gi\|229002753\|dbj\|BAC05484.2\|ascorbate peroxidase [Euglena gracilis] |
| 266 | EG_transcript_5798 | GO:0009570, GO:0010155, GO:0009507, GO:0046777, GO:0009941, GO:0006413 | gi\|145354801\|ref\|XP_001421664.1\|predicted protein, partial [Ostreococcus lucimarinus CCE9901] |
| 267 | EG_transcript_5873 | GO:0003746, GO:0006414, GO:0006412, GO:0003676, GO:0005622 | gi\|612395376\|ref\|XP_007513208.1\|elongation factor Ts [Bathycoccus prasinos] |
| 268 | EG_transcript_5886 | No annotation | |
| 269 | EG_transcript_59611 | GO:0003735, GO:0005840, GO:0006412 | gi\|902189586\|gb\|KNA11413.1\|hypothetical protein SOVF_135540 [Spinacia oleracea] |
| 270 | EG_transcript_6058 | GO:0008152, GO:0003824, GO:0008661, GO:0016114 | gi\|922852149\|gb\|KOO20858.1\|1-deoxy-d-xylulose-5-phosphate synthase [Chrysochromulina sp. CCMP291] |
| 271 | EG_transcript_6141 | GO:0016020 | predicted protein |
| 272 | EG_transcript_6340 | GO:0005737, GO:0003747, GO:0006415, GO:0016149 | gi\|255079106\|ref\|XP_002503133.1\|predicted protein [Micromonas commoda] |
| 273 | EG_transcript_6397 | GO:0016226 | SUF system cluster assembly, |
| 274 | EG_transcript_65 | GO:0000166, GO:0003824, GO:0003989, GO:0046872, GO:0004075, GO:0016874, GO:0005524, GO:0006633 | gi\|578896496\|gb\|AHI17198.1\|acetyl-CoA carboxylase [Nannochloropsis oculata] |
| 275 | EG_transcript_6500 | GO:0016020 | TPA_inf: chloroplast light-harvesting complex I precursor Lhca2 |

| | | | |
|---|---|---|---|
| 276 | EG_transcript_6636 | GO:0004853 | uroporphyrinogen decarboxylase |
| 277 | EG_transcript_6647 | GO:0008152, GO:0016491, GO:0055114, GO:0016630 | gi\|219117802\|ref\|XP_002179689.1\|predicted protein [Phaeodactylum tricornutum CCAP 1055/1] |
| 278 | EG_transcript_669 | GO:0000166 | ---NA--- |
| 279 | EG_transcript_6699 | GO:0003743, GO:0009507, GO:0006412, GO:0009536, GO:0006413 | gi\|547710\|sp\|P36177.1\|IF3C_EUGGRRecName: Full=Translation initiation factor IF-3, chloroplastic |
| 280 | EG_transcript_6713 | GO:0003824, GO:0030246, GO:0005975, GO:0016853 | gi\|676391527\|ref\|XP_009039041.1\|hypothetical protein AURANDRAFT_4661, partial [Aureococcus anophagefferens] |
| 281 | EG_transcript_6735 | GO:0005524 | gi\|168065377\|ref\|XP_001784629.1\|predicted protein [Physcomitrella patens] |
| 282 | EG_transcript_6853 | No annotation | |
| 283 | EG_transcript_6882 | GO:0000023, GO:0019252, GO:0019288, GO:0034660, GO:0009902, GO:0010027, GO:0005623 | gi\|145350419\|ref\|XP_001419603.1\|predicted protein [Ostreococcus lucimarinus CCE9901] |
| 284 | EG_transcript_6932 | GO:0005737, GO:0006508, GO:0008235, GO:0004177, GO:0030145, GO:0019538, GO:0005622 | gi\|695464043\|ref\|XP_009538858.1\|hypothetical protein PHYSODRAFT_353200 [Phytophthora sojae] |
| 285 | EG_transcript_6976 | GO:0016020, GO:0016021 | gi\|219109882\|ref\|XP_002176694.1\|triose phosphate/phosphate translocator [Phaeodactylum tricornutum CCAP 1055/1] |
| 286 | EG_transcript_7051 | GO:0018160, GO:0015995, GO:0006779, GO:0009507, GO:0004418, GO:0006782, GO:0016740, GO:0033014, GO:0009536 | gi\|122842\|sp\|P13446.1\|HEM3_EUGGRRecName: Full=Porphobilinogen deaminase, chloroplastic |
| 287 | EG_transcript_7071 | GO:0016020, GO:0016021 | gi\|397624597\|gb\|EJK67440.1\|hypothetical protein THAOC_11524 [Thalassiosira oceanica] |
| 288 | EG_transcript_7105 | GO:0008152, GO:0051741, GO:0008168, GO:0010189, GO:0032259 | gi\|302841643\|ref\|XP_002952366.1\|hypothetical protein VOLCADRAFT_105494 [Volvox carteri f. nagariensis] |
| 289 | EG_transcript_7123 | GO:0009507, GO:0019253, GO:0016491, GO:0016829, GO:0009853, GO:0055114, GO:0004497, GO:0015979, GO:0016984, GO:0015977, GO:0009536 | gi\|132143\|sp\|P16881.1\|RBS_EUGGRRecName: Full=Ribulose bisphosphate carboxylase small chains, chloroplastic |
| 290 | EG_transcript_7174 | GO:0016020, GO:0016021 | gi\|255086363\|ref\|XP_002509148.1\|predicted protein [Micromonas commoda] |
| 291 | EG_transcript_7192 | GO:0000166, GO:0009507, GO:0045261, GO:0046961, GO:0046034, GO:0016020, GO:0016787, GO:0016820, GO:0015986, GO:0015992, GO:0033178, GO:0015991, GO:0009535, GO:0009579, GO:0006810, GO:0006811, GO:0046933, GO:0005524, GO:0009536, GO:0006754 | gi\|11467021\|ref\|NP_041928.1\|ATP synthase CF1 alpha subunit [Euglena gracilis] |

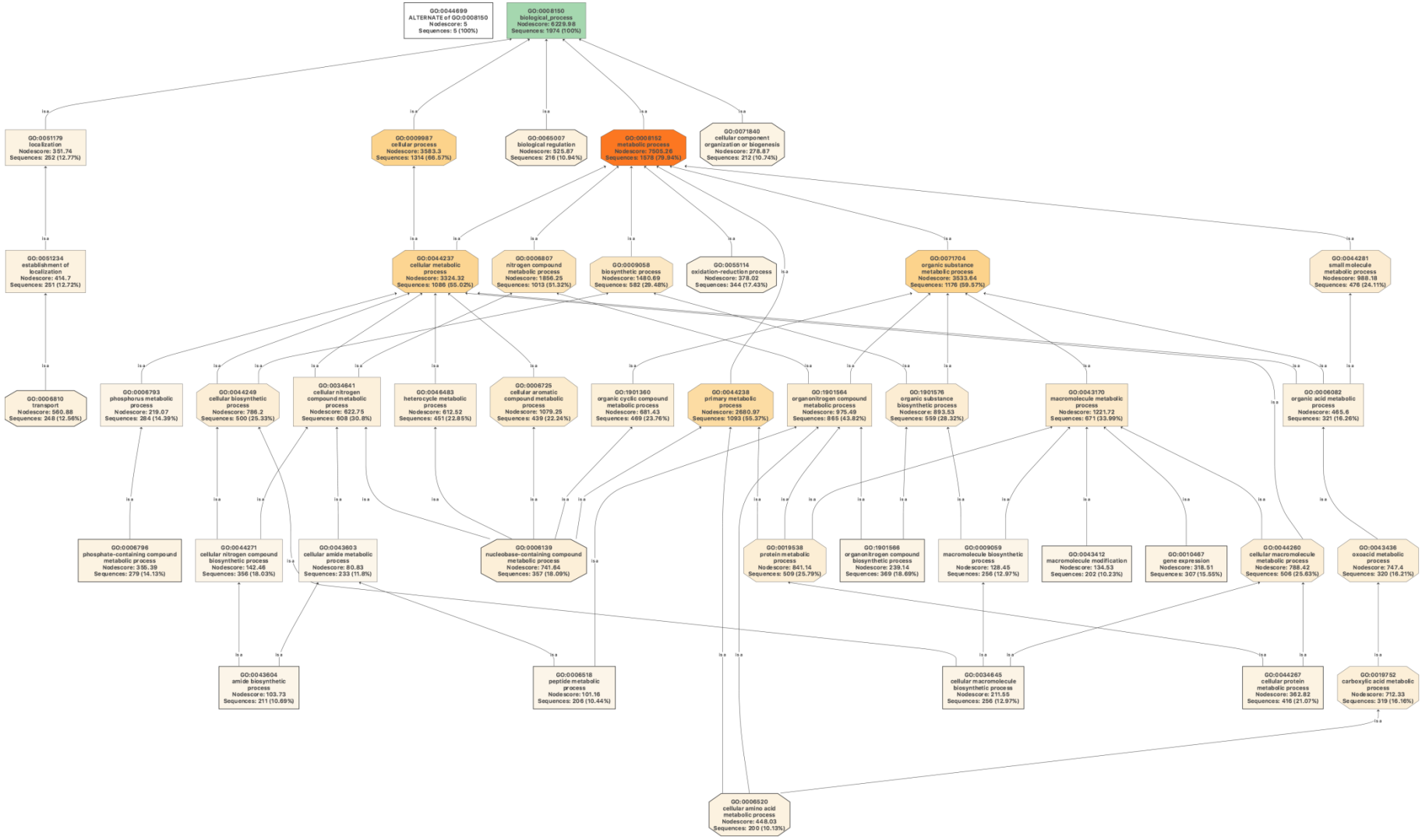
| | | | |
|---|---|---|---|
| 292 | EG_transcript_7239 | No annotation | |
| 293 | EG_transcript_7247 | GO:0050343, GO:0016491, GO:0005739, GO:0051287, GO:0050661, GO:0006629, GO:0055114, GO:0006631, GO:0006633 | gi\|62287512\|sp\|Q5EU90.1\|TER_EUGGRRecName: Full=Trans-2-enoyl-CoA reductase, mitochondrial |
| 294 | EG_transcript_7257 | GO:0008152, GO:0005829, GO:0016787, GO:0005886 | gi\|159466964\|ref\|XP_001691668.1\|predicted protein [Chlamydomonas reinhardtii] |
| 295 | EG_transcript_7295 | No annotation | |
| 296 | EG_transcript_7354 | GO:0016021 | hypothetical protein JH06_2455 |
| 297 | EG_transcript_7356 | GO:0008152, GO:0000166, GO:0005525, GO:0005737, GO:0090502, GO:0043023, GO:0043022, GO:0016787, GO:0004523, GO:0005524, GO:0003676, GO:0016887 | gi\|224003407\|ref\|XP_002291375.1\|GTP binding protein [Thalassiosira pseudonana CCMP1335] |
| 298 | EG_transcript_7358 | GO:0016491, GO:0045300, GO:0055114, GO:0006631 | gi\|612391857\|ref\|XP_007511449.1\|predicted protein [Bathycoccus prasinos] |
| 299 | EG_transcript_7385 | No annotation | |
| 300 | EG_transcript_7394 | GO:0016020, GO:0042132, GO:0016021, GO:0016311, GO:0042578, GO:0016787, GO:0005975 | gi\|99904205\|gb\|ABF68597.1\|chloroplast fructose-1,6-bisphosphatase [Euglena gracilis] |
| 301 | EG_transcript_7656 | GO:0009523 | chloroplast photosystem II M precursor |
| 302 | EG_transcript_7728 | No annotation | |
| 303 | EG_transcript_7821 | GO:0006508, GO:0008236 | gi\|545376288\|ref\|XP_005652247.1\|ClpP/crotonase [Coccomyxa subellipsoidea C-169] |
| 304 | EG_transcript_7827 | GO:0003824, GO:0006096, GO:0004332 | gi\|612391897\|ref\|XP_007511469.1\|predicted protein [Bathycoccus prasinos] |
| 305 | EG_transcript_7877 | GO:0006418, GO:0000166, GO:0005737, GO:0004812, GO:0004828, GO:0016874, GO:0006434, GO:0005524 | gi\|553195230\|ref\|XP_005856106.1\|seryl-tRNA synthetase [Nannochloropsis gaditana CCMP526] |
| 306 | EG_transcript_7991 | GO:0016491 | (2Fe-2S)-binding protein |
| 307 | EG_transcript_8042 | GO:0016620 | plastid glyceraldehyde 3-phosphate dehydrogenase, partial |
| 308 | EG_transcript_8143 | GO:0016491 | hypothetical protein GPECTOR_42g802 |
| 309 | EG_transcript_8253 | GO:0016717 | palmitoyl-monogalactosyldiacylglycerol delta-7 desaturase |
| 310 | EG_transcript_8312 | GO:0016020, GO:0016021 | gi\|219109882\|ref\|XP_002176694.1\|triose phosphate/phosphate translocator [Phaeodactylum tricornutum CCAP 1055/1] |
| 311 | EG_transcript_8319 | GO:0016020, GO:0006508, GO:0016021, GO:0004222, GO:0005524 | gi\|302833547\|ref\|XP_002948337.1\|hypothetical protein VOLCADRAFT_80122 [Volvox carteri f. nagariensis] |

| | | | |
|---|---|---|---|
| 312 | EG_transcript_8383 | GO:0008762 | hypothetical protein |
| 313 | EG_transcript_8435 | GO:0003677 | gi\|731701726\|ref\|XP_010700693.1\|high mobility group protein tdp-1, putative [Leishmania panamensis] |
| 314 | EG_transcript_8444 | GO:0008152, GO:0016747, GO:0016020, GO:0016021 | gi\|145352756\|ref\|XP_001420703.1\|predicted protein [Ostreococcus lucimarinus CCE9901] |
| 315 | EG_transcript_8456 | GO:0006779, GO:0004109, GO:0016020, GO:0016021, GO:0055114 | gi\|327493895\|gb\|AEA86534.1\|coproporphyrinogen oxidase [Euglena gracilis] |
| 316 | EG_transcript_8464 | GO:0006418, GO:0000287, GO:0000166, GO:0005737, GO:0004812, GO:0004826, GO:0008033, GO:0006432, GO:0000049, GO:0043039, GO:0016874, GO:0005524 | gi\|567961404\|gb\|ETK86946.1\|phenylalanine-tRNA ligase [Phytophthora parasitica] |
| 317 | EG_transcript_8528 | GO:0050992, GO:0046872, GO:0019288, GO:0055114, GO:0051745 | gi\|552833460\|ref\|XP_005848540.1\|hypothetical protein CHLNCDRAFT_59658 [Chlorella variabilis] |
| 318 | EG_transcript_854 | GO:0003676 | ATP-dependent helicase |
| 319 | EG_transcript_8610 | GO:0005840, GO:0003676 | gi\|612398819\|ref\|XP_007514929.1\|30S ribosomal protein S1 [Bathycoccus prasinos] |
| 320 | EG_transcript_8626 | No annotation | |
| 321 | EG_transcript_8688 | GO:0006779, GO:0003824, GO:0046872, GO:0016829, GO:0033014, GO:0004655 | gi\|397575727\|gb\|EJK49854.1\|hypothetical protein THAOC_31221 [Thalassiosira oceanica] |
| 322 | EG_transcript_8709 | GO:0042549, GO:0009507, GO:0016020, GO:0016021, GO:0005509, GO:0009535, GO:0009579, GO:0009523, GO:0009654, GO:0015979, GO:0019898, GO:0009536 | gi\|7008029\|dbj\|BAA03529.2\|oxygen-evolving enhancer protein 1 precursor [Euglena gracilis] |
| 323 | EG_transcript_8727 | GO:0042132 | plastid fructose 1,6-bisphosphatase isoform 2, partial |
| 324 | EG_transcript_8736 | GO:0005506, GO:0009507, GO:0046872, GO:0016020, GO:0016021, GO:0055114, GO:0015979, GO:0031361, GO:0009055, GO:0020037, GO:0009579, GO:0009535, GO:0009523, GO:0009522, GO:0009536 | gi\|38604661\|sp\|Q8GZR2.2\|CYF_EUGGRRecName: Full=Cytochrome f, chloroplastic |
| 325 | EG_transcript_8740 | No annotation | |
| 326 | EG_transcript_8798 | No annotation | |
| 327 | EG_transcript_8931 | GO:0046872, GO:0009055, GO:0051536 | gi\|299472521\|emb\|CBN77306.1\|Ferredoxin [Ectocarpus siliculosus] |
| 328 | EG_transcript_8951 | No annotation | |
| 329 | EG_transcript_8954 | No annotation | |
| 330 | EG_transcript_899 | GO:0009507, GO:0045156, GO:0019684, GO:0016020, GO:0018298, GO:0016021, GO:0016168, GO:0015979, | gi\|830260467\|gb\|AKL82402.1\|photosystem II CP47 chlorophyll apoprotein (chloroplast) [Euglena gracilis var. bacillaris] |

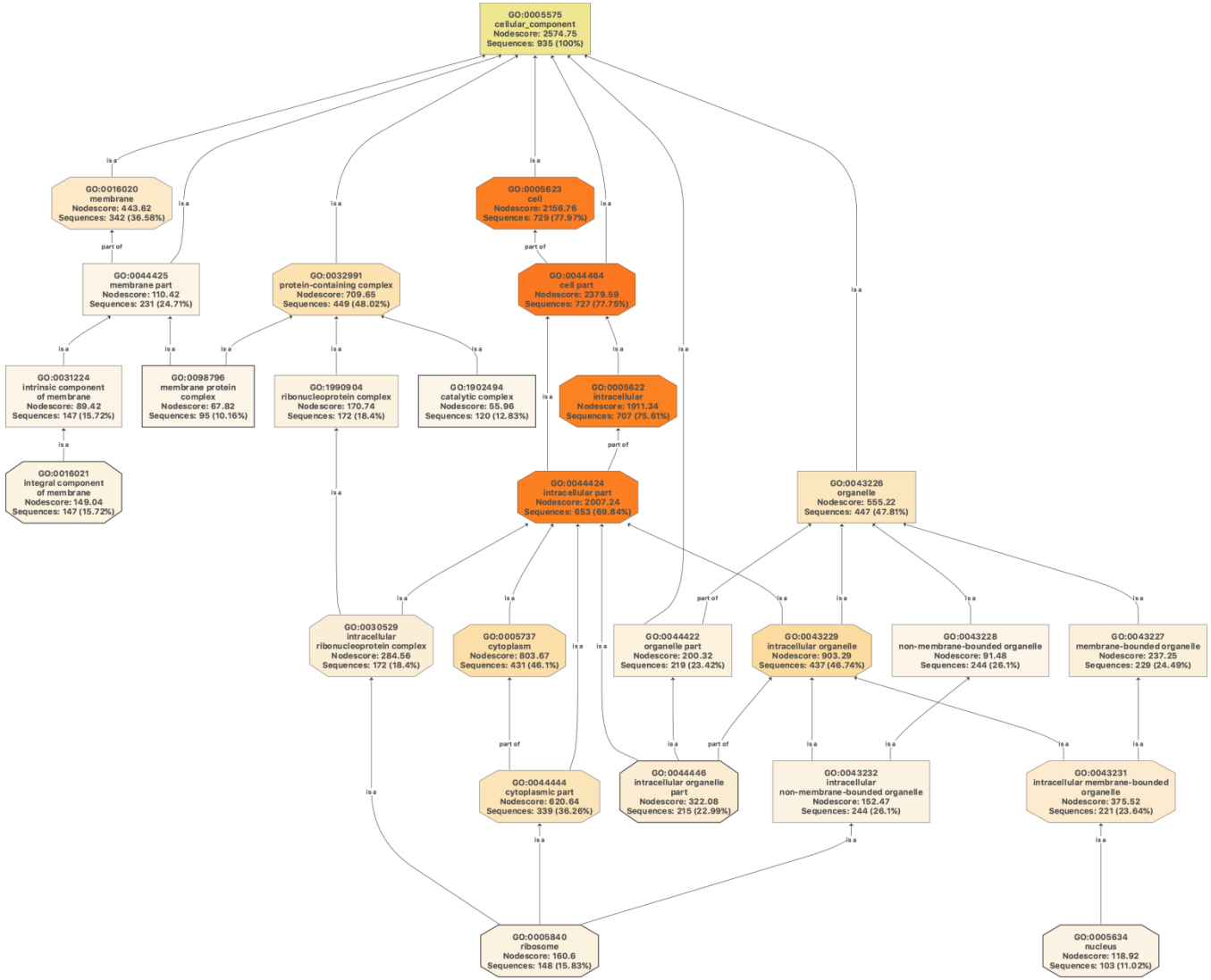
| | | | |
|---|---|---|---|
| | | GO:0009772, GO:0009535, GO:0009579, GO:0009523, GO:0009521, GO:0009536, GO:0009767 | |
| 331 | EG_transcript_9015 | GO:0008152, GO:0000166, GO:0016020, GO:0016021, GO:0008974, GO:0016310, GO:0016740, GO:0016301, GO:0005524, GO:0005975 | gi\|60101680\|gb\|AAX13964.1\|chloroplast phosphoribulokinase [Euglena gracilis] |
| 332 | EG_transcript_9052 | GO:0006418, GO:0000166, GO:0005737, GO:0004812, GO:0004827, GO:0006433, GO:0016874, GO:0005524 | gi\|545705603\|ref\|XP_005705121.1\|prolyl-tRNA synthetase [Galdieria sulphuraria] |
| 333 | EG_transcript_9148 | No annotation | |
| 334 | EG_transcript_9197 | No annotation | |
| 335 | EG_transcript_9217 | GO:0005515 | small glutamine-rich tetratricopeptide repeat |
| 336 | EG_transcript_9317 | GO:0016020, GO:0016021 | gi\|219109882\|ref\|XP_002176694.1\|triose phosphate/phosphate translocator [Phaeodactylum tricornutum CCAP 1055/1] |
| 337 | EG_transcript_9321 | GO:0031418, GO:0005506, GO:0016705, GO:0016491, GO:0055114 | gi\|545359684\|ref\|XP_005645287.1\|hypothetical protein COCSUDRAFT_57307 [Coccomyxa subellipsoidea C-169] |
| 338 | EG_transcript_9385 | GO:0015995, GO:0004659, GO:0016020, GO:0016021, GO:0046408 | gi\|298714003\|emb\|CBJ27235.1\|Chlorophyll synthase, putative chloroplast precursor [Ectocarpus siliculosus] |
| 339 | EG_transcript_9401 | No annotation | |
| 340 | EG_transcript_941 | No annotation | |
| 341 | EG_transcript_943 | No annotation | |
| 342 | EG_transcript_9509 | No annotation | |
| 343 | EG_transcript_9571 | GO:0050662, GO:0003954, GO:0003824, GO:0055114, GO:1901006 | gi\|224011014\|ref\|XP_002294464.1\|predicted protein [Thalassiosira pseudonana CCMP1335] |
| 344 | EG_transcript_9617 | GO:0015934 | 50S ribosomal L1 |
| 345 | EG_transcript_9691 | GO:0009507, GO:0016020, GO:0018298, GO:0016021, GO:0009579, GO:0009765, GO:0009523, GO:0016168, GO:0015979 | gi\|157965829\|gb\|ABW06951.1\|chloroplast light-harvesting complex I protein precursor [Euglena gracilis] |
| 346 | EG_transcript_9991 | GO:0006418, GO:0000166, GO:0004812, GO:0005524, GO:0003676 | gi\|223995725\|ref\|XP_002287536.1\|predicted protein [Thalassiosira pseudonana CCMP1335] |

**Note:** The table describes the list of top upregulated proteins (346 proteins) in the light regime corresponding to the proteomics experiment. Column descriptions are in the order: Sequence ID = *Euglena* sequence ID, Gene Ontology number = GO number for the corresponding protein. BLAST Hit description = description of BLAST top hit. Protein IDs with corresponding description “No annotation” under the GO number column and empty cells under BLAST Hit description column means that an annotation could not be assigned. Individual proteins are represented by multiple GO numbers (e.g. EG_transcript_10021), and single GO numbers are also shared by multiple proteins e.g. GO:0016021 (EG_transcript_10193 and EG_transcript_10194). In the case of multiple GOs against a specific proteins, the first GO on the series is the top hit or reference GO.

## A: Biological Process



## B: Cellular Component



## C: Molecular Function

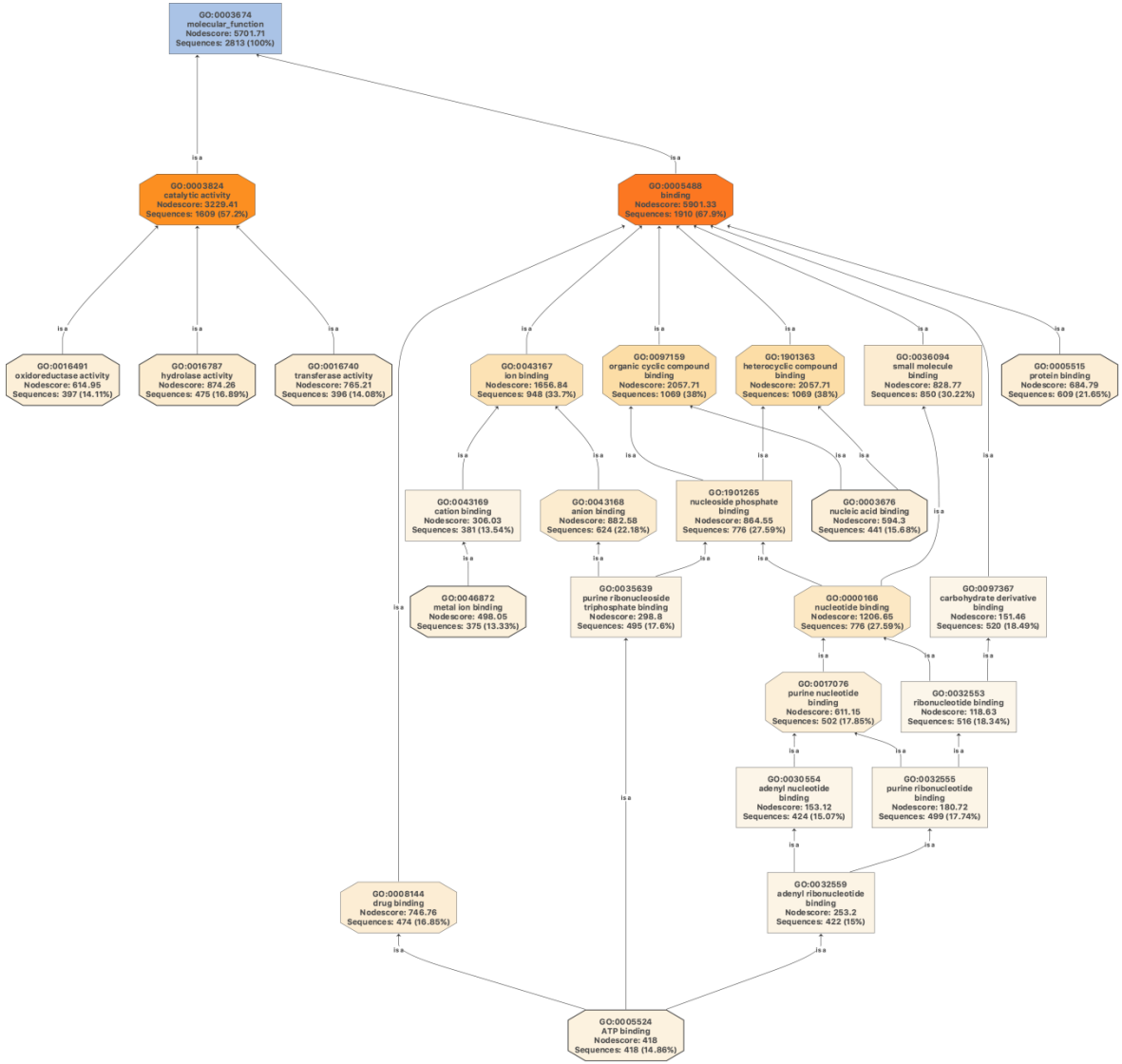
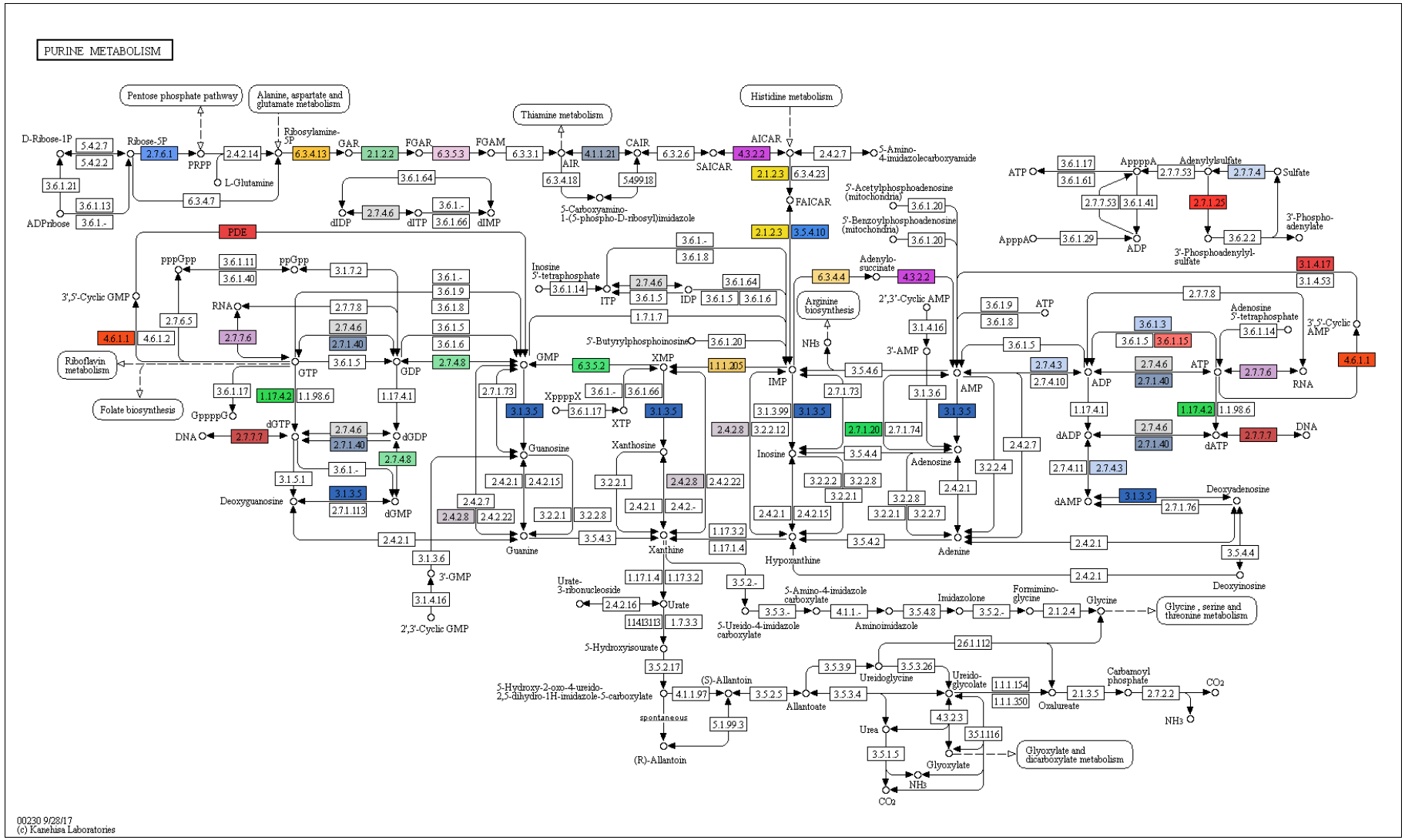
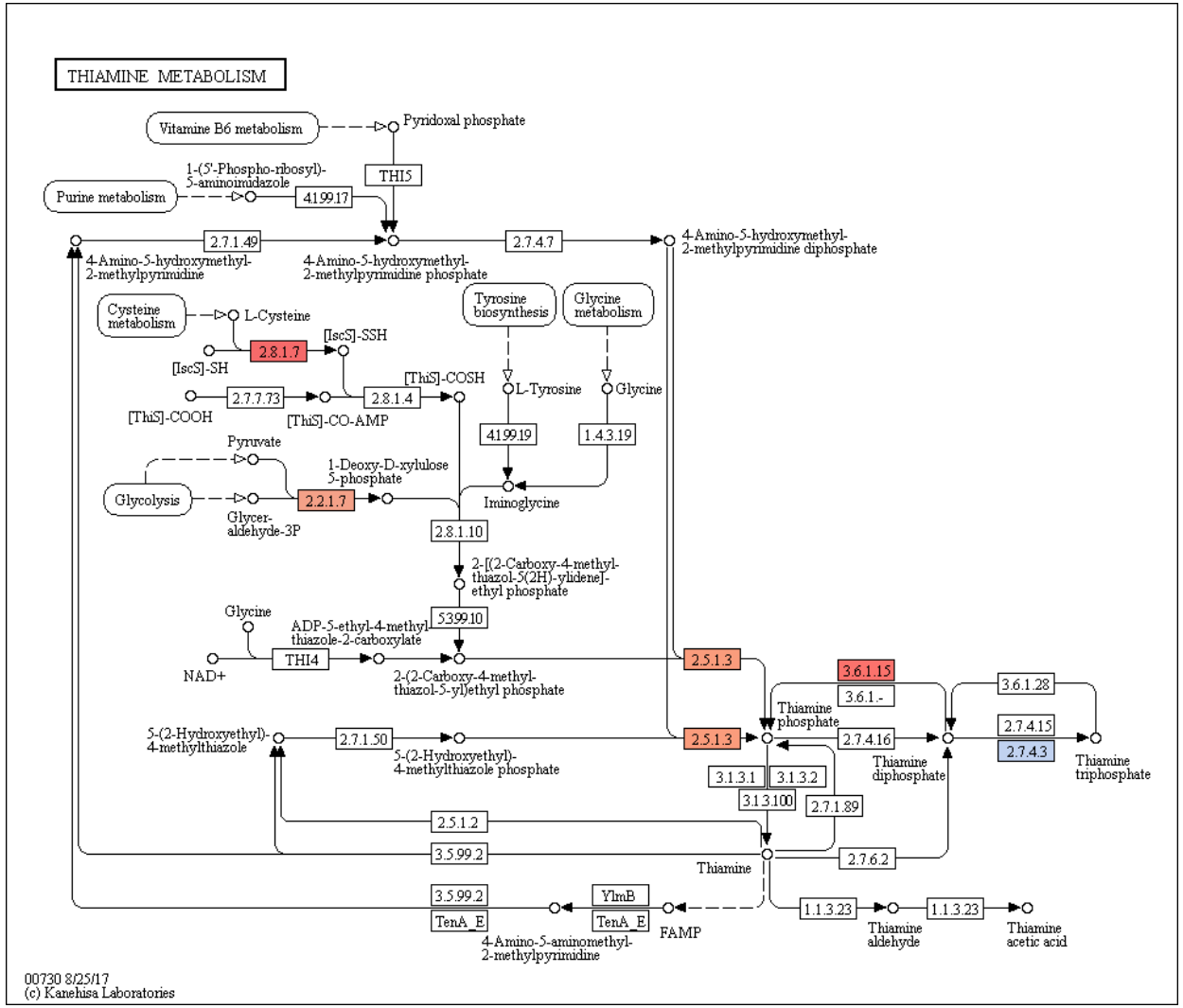


**Figure 6.6: *E. gracilis* cellular processes, structure, and functions are influenced by light or dark external conditions.** The diagrams show the GO combined graphs for all significant regulated proteins (4681 quantified proteins) in the *E. gracilis* light and dark proteomic experiment. Panel A: Biological Process (BP). Panel B: Cellular Component (CC). Panel C: Molecular Functions (MF). Boxes are in hierarchy and correspond to the downward flow of the graph. Green, blue, and orange, are of higher hierarchies in Biological Process, Cellular Component, and Molecular Functions. Each box contains the specific GO number for the functional information, corresponding GO description, Nodescore, and number of sequences that constitute that process expressed in percentage (%).

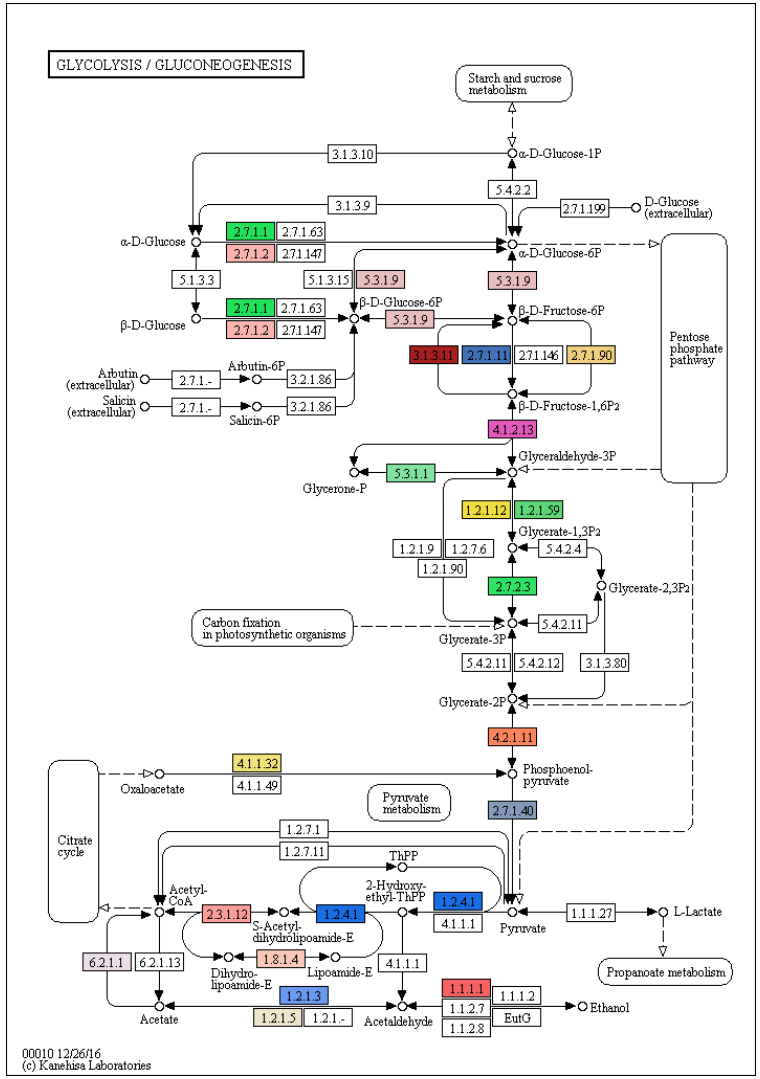
## Panel A: Purine Metabolism



## PANEL B: Thiamine Metabolism



## PANEL C: Glycolysis / Gluconeogenesis



## PANEL D: Carbon Fixation in Photosynthetic Organisms

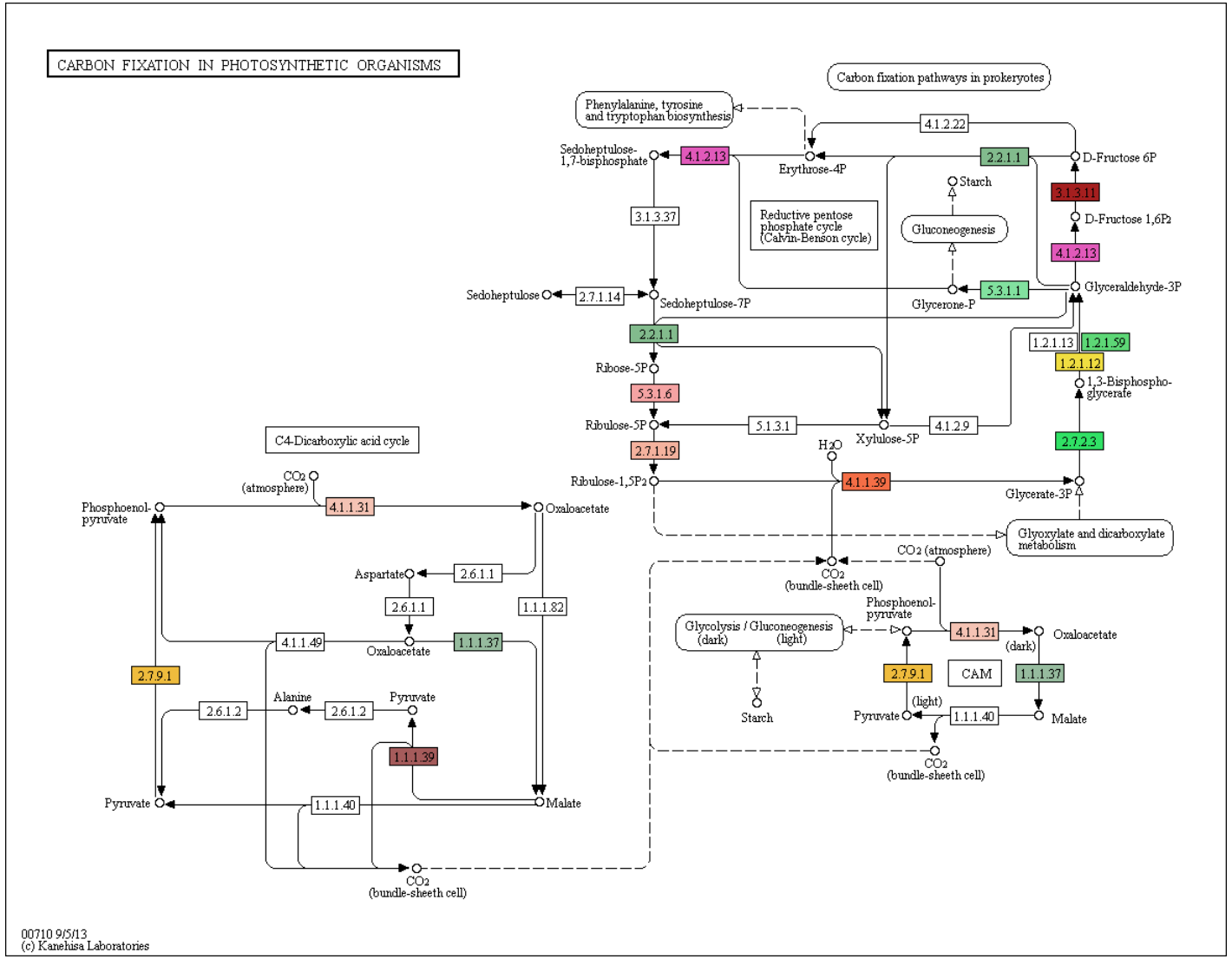


**Figure 6.7: *E. gracilis* possess core biological heterotrophic and photosynthetic pathways.** The diagram shows the core pathways represented in the significantly up- and down- regulated proteins in the light and dark experiment. Panel A: shows the Purine metabolism pathway with the highest number of corresponding number of sequences (Table 6.5). Panel B: shows the Thiamine metabolism pathway with the second highest number of corresponding sequences (Table 6.5). Panel C and D: shows the Glycolysis/Gluconeogenesis metabolism and Carbon Fixation in Photosynthetic Organisms pathways respectively. Colour codes represent the Enzyme number associated with specific pathways are described in Table 6.5. Text in flowcharts describes pathway steps, process, and products.

**Table 6.5:** Enzyme Code and KEGG Pathway Maps for *E. gracilis* light and dark protein regime

| Pathway | Pathway ID | #Enzs in Pathway | Enzyme | Seqs |
|---|---|---|---|---|
| Glyoxylate and dicarboxylate metabolism | map00630 | 9 | ec:2.1.2.1 - hydroxymethyltransferase, ec:4.1.1.39 - carboxylase, ec:4.2.1.3 - hydratase, ec:6.3.1.2 - synthetase, ec:4.1.3.1 - lyase, ec:1.1.1.37 - dehydrogenase, ec:2.3.3.9 - synthase, ec:2.3.3.1 - (Si)-synthase, ec:5.4.99.2 - mutase | EG_transcript_9546, EG_transcript_7542, EG_transcript_3442, EG_transcript_9405, EG_transcript_7123, EG_transcript_10675, EG_transcript_2009, EG_transcript_3524, EG_transcript_13142, EG_transcript_24268, EG_transcript_19219, EG_transcript_729, EG_transcript_12212, EG_transcript_15919, EG_transcript_24127, EG_transcript_2530, EG_transcript_729, EG_transcript_2898, EG_transcript_6849, EG_transcript_10302, EG_transcript_11595, EG_transcript_4204 |
| Glycolysis / Gluconeogenesis | map00010 | 21 | ec:1.8.1.4 - dehydrogenase, ec:2.7.1.90 - 1-phosphotransferase, ec:5.3.1.1 - isomerase, ec:4.1.2.13 - aldolase, ec:5.3.1.9 - isomerase, ec:1.1.1.1 - dehydrogenase, ec:4.1.1.32 - carboxykinase (GTP), ec:4.2.1.11 - hydratase, ec:2.7.2.3 - kinase, ec:2.7.1.40 - kinase, ec:2.7.1.2 - glucokinase (phosphorylating), ec:2.7.1.1 - hexokinase type IV glucokinase, ec:1.2.1.5 - dehydrogenase [NAD(P)+], ec:1.2.1.3 - dehydrogenase (NAD+), ec:1.2.1.59 - dehydrogenase (NAD(P)+) (phosphorylating), ec:2.7.1.11 - phosphohexokinase, ec:6.2.1.1 - ligase, ec:2.3.1.12 - acetyltransferase, ec:3.1.3.11 - hexose diphosphatase, ec:1.2.1.12 - dehydrogenase (phosphorylating), ec:1.2.4.1 - dehydrogenase (acetyl-transferring) | EG_transcript_8679, EG_transcript_1151, EG_transcript_14120, EG_transcript_30563, EG_transcript_18179, EG_transcript_30456, EG_transcript_44253, EG_transcript_7827, EG_transcript_21524, EG_transcript_43521, EG_transcript_16008, EG_transcript_25832, EG_transcript_5279, EG_transcript_15855, EG_transcript_5108, EG_transcript_21676, EG_transcript_19272, EG_transcript_61784, EG_transcript_12376, EG_transcript_6665, EG_transcript_12119, EG_transcript_10778, EG_transcript_12260, EG_transcript_2957, EG_transcript_6211, EG_transcript_8623, EG_transcript_8623, EG_transcript_5269, EG_transcript_6432, EG_transcript_15045, EG_transcript_16406, EG_transcript_1151, EG_transcript_7325, EG_transcript_6970, EG_transcript_6957, EG_transcript_4603, EG_transcript_3254, EG_transcript_9669, EG_transcript_5584, EG_transcript_7394, EG_transcript_2646, EG_transcript_8727, EG_transcript_7683, EG_transcript_4570, EG_transcript_15632, EG_transcript_18342, EG_transcript_13690, EG_transcript_16406, EG_transcript_14931, EG_transcript_9827 |
| Glutathione metabolism | map00480 | 10 | ec:1.8.1.7 - reductase, ec:1.11.1.9 - peroxidase, ec:6.3.2.3 - synthase, ec:6.3.2.2 - ligase, ec:1.1.1.44 - dehydrogenase (NADP+-dependent, decarboxylating), ec:1.1.1.42 - dehydrogenase (NADP+), ec:2.3.2.2 - glutamyl transpeptidase, ec:1.1.1.49 - dehydrogenase (NADP+), ec:1.11.1.11 - peroxidase, ec:1.11.1.15 - thioredoxin peroxidase | EG_transcript_8715, EG_transcript_29973, EG_transcript_21143, EG_transcript_20841, EG_transcript_27461, EG_transcript_8300, EG_transcript_1613, EG_transcript_13898, EG_transcript_7631, EG_transcript_16470, EG_transcript_15075, EG_transcript_9064, EG_transcript_10115, EG_transcript_4351, EG_transcript_6875, EG_transcript_5619, EG_transcript_2747, EG_transcript_5995, EG_transcript_9049, EG_transcript_22491, EG_transcript_5780, EG_transcript_17359, EG_transcript_29745, EG_transcript_14865, |

| | | | | |
|---|---|---|---|---|
| | | | | EG_transcript_51669, EG_transcript_11677, EG_transcript_22903, EG_transcript_42708 |
| Tryptophan metabolism | map00380 | 3 | ec:1.4.3.2 - oxidase, ec:1.2.1.3 - dehydrogenase (NAD+), ec:1.1.1.35 - dehydrogenase | EG_transcript_5672, EG_transcript_15045, EG_transcript_7580, EG_transcript_29059, EG_transcript_26319, EG_transcript_862, EG_transcript_21457, EG_transcript_3075 |
| Pyruvate metabolism | map00620 | 20 | ec:1.8.1.4 - dehydrogenase, ec:6.4.1.2 - carboxylase, ec:2.3.3.14 - synthase, ec:2.3.3.13 - synthase, ec:3.6.1.7 - acetylphosphatase, ec:4.1.1.31 - carboxylase, ec:4.1.1.32 - carboxykinase (GTP), ec:4.2.1.2 - hydratase, ec:3.1.2.6 - hydrolase, ec:2.7.1.40 - kinase, ec:2.7.9.1 - phosphate dikinase, ec:1.2.1.3 - dehydrogenase (NAD+), ec:6.2.1.1 - ligase, ec:1.1.1.37 - dehydrogenase, ec:4.4.1.5 - lyase, ec:1.1.1.39 - dehydrogenase (decarboxylating), ec:1.1.1.38 - dehydrogenase (oxaloacetate-decarboxylating), ec:2.3.1.12 - acetyltransferase, ec:2.3.3.9 - synthase, ec:1.2.4.1 - dehydrogenase (acetyl-transferring) | EG_transcript_8679, EG_transcript_65, EG_transcript_97, EG_transcript_6633, EG_transcript_4221, EG_transcript_7582, EG_transcript_5664, EG_transcript_52815, EG_transcript_1597, EG_transcript_61784, EG_transcript_4764, EG_transcript_14054, EG_transcript_2957, EG_transcript_6211, EG_transcript_1653, EG_transcript_15045, EG_transcript_4603, EG_transcript_3254, EG_transcript_9669, EG_transcript_12212, EG_transcript_15919, EG_transcript_24127, EG_transcript_18923, EG_transcript_27221, EG_transcript_5257, EG_transcript_8836, EG_transcript_16327, EG_transcript_14050, EG_transcript_4839, EG_transcript_27221, EG_transcript_5257, EG_transcript_8836, EG_transcript_16327, EG_transcript_14050, EG_transcript_4839, EG_transcript_5584, EG_transcript_2530, EG_transcript_729, EG_transcript_14931, EG_transcript_9827 |
| Cyanoamino acid metabolism | map00460 | 2 | ec:2.1.2.1 - hydroxymethyltransferase, ec:2.3.2.2 - glutamyl transpeptidase | EG_transcript_9546, EG_transcript_7542, EG_transcript_3442, EG_transcript_9405, EG_transcript_5619 |
| Photosynthesis | map00195 | 1 | ec:1.10.9.1 - reductase | EG_transcript_5228 |
| Biosynthesis of antibiotics | map01130 | 99 | ec:1.8.1.4 - dehydrogenase, ec:2.3.3.10 - synthase, ec:6.4.1.2 - carboxylase, ec:2.1.2.3 - formyltransferase, ec:2.3.3.14 - synthase, ec:2.6.1.16 - transaminase (isomerizing), ec:2.1.2.2 - formyltransferase, ec:2.1.2.1 - hydroxymethyltransferase, ec:2.7.1.90 - 1-phosphotransferase, ec:5.3.1.1 - isomerase, ec:4.1.2.13 - aldolase, ec:2.7.4.6 - kinase, ec:5.3.1.6 - isomerase, ec:2.7.4.3 - kinase, ec:4.1.1.48 - synthase, ec:5.3.1.9 - isomerase, ec:2.7.7.60 - 4-phosphate cytidylyltransferase, ec:2.7.7.64 - uridylyltransferase, ec:6.3.4.5 - synthase, ec:4.1.1.33 - decarboxylase, ec:4.1.1.39 - carboxylase, ec:1.3.5.1 - dehydrogenase, ec:1.1.1.1 - dehydrogenase, ec:4.1.1.32 - carboxykinase (GTP), ec:4.2.1.9 - dehydratase, ec:4.2.1.3 - hydratase, ec:4.2.1.2 - hydratase, ec:1.17.7.1 - synthase (ferredoxin), ec:4.2.1.20 - synthase, ec:2.6.1.52 - transaminase, ec:2.6.1.9 - | EG_transcript_8679, EG_transcript_16836, EG_transcript_28, EG_transcript_5258, EG_transcript_65, EG_transcript_97, EG_transcript_6633, EG_transcript_5356, EG_transcript_4221, EG_transcript_5074, EG_transcript_4891, EG_transcript_36522, EG_transcript_9546, EG_transcript_7542, EG_transcript_3442, EG_transcript_9405, EG_transcript_1151, EG_transcript_14120, EG_transcript_30563, EG_transcript_18179, EG_transcript_30456, EG_transcript_44253, EG_transcript_7827, EG_transcript_21524, EG_transcript_43521, EG_transcript_16008, EG_transcript_25832, EG_transcript_5279, EG_transcript_15855, EG_transcript_17910, EG_transcript_11790, EG_transcript_26205, EG_transcript_15619, EG_transcript_20009, EG_transcript_22143, EG_transcript_1490, EG_transcript_5108, EG_transcript_14037, EG_transcript_6731, EG_transcript_12921, EG_transcript_10767, EG_transcript_17330, EG_transcript_7123, EG_transcript_10675, EG_transcript_6467, EG_transcript_21676, EG_transcript_19272, EG_transcript_61784, |

| | | | | |
|---|---|---|---|---|
| | | | transaminase, ec:1.1.1.95 - dehydrogenase, ec:2.5.1.47 - synthase, ec:4.2.1.11 - hydratase, ec:2.2.1.2 - dihydroxyacetonetransferase, ec:4.2.1.10 - dehydratase, ec:2.2.1.1 - glycolaldehydetransferase, ec:2.2.1.6 - synthase, ec:5.3.3.2 - Delta-isomerase, ec:2.5.1.54 - synthase, ec:2.2.1.7 - synthase, ec:2.7.2.4 - kinase, ec:2.7.2.3 - kinase, ec:4.1.1.82 - decarboxylase, ec:4.3.2.1 - lyase, ec:4.3.2.2 - lyase, ec:1.2.1.38 - reductase, ec:2.7.1.40 - kinase, ec:2.7.2.11 - 5-kinase, ec:1.2.1.41 - dehydrogenase, ec:1.4.1.14 - synthase (NADH), ec:4.3.1.19 - ammonia-lyase, ec:1.3.1.13 - dehydrogenase (NADP+), ec:1.3.1.12 - dehydrogenase, ec:2.7.1.36 - kinase, ec:4.2.1.46 - 4,6-dehydratase, ec:2.7.1.2 - glucokinase (phosphorylating), ec:2.7.1.1 - hexokinase type IV glucokinase, ec:4.2.3.5 - synthase, ec:1.1.1.86 - reductoisomerase (NADP+), ec:4.2.3.4 - synthase, ec:2.6.1.42 - transaminase, ec:1.2.1.3 - dehydrogenase (NAD+), ec:1.2.1.59 - dehydrogenase (NAD(P)+) (phosphorylating), ec:3.5.1.14 - acid amidohydrolase, ec:4.2.1.36 - hydratase, ec:2.7.1.11 - phosphohexokinase, ec:4.2.1.51 - dehydratase, ec:5.5.1.4 - synthase, ec:2.5.1.19 - 1-carboxyvinyltransferase, ec:6.2.1.1 - ligase, ec:1.1.1.37 - dehydrogenase, ec:1.1.1.35 - dehydrogenase, ec:1.1.1.34 - reductase (NADPH), ec:2.3.1.12 - acetyltransferase, ec:3.1.3.11 - hexose diphosphatase, ec:1.1.1.44 - dehydrogenase (NADP+-dependent, decarboxylating), ec:1.1.1.42 - dehydrogenase (NADP+), ec:2.7.7.4 - adenylyltransferase, ec:1.1.1.49 - dehydrogenase (NADP+), ec:2.7.7.9 - uridylyltransferase, ec:5.4.2.8 - mannose phosphomutase, ec:5.4.2.9 - mutase, ec:1.2.1.11 - dehydrogenase, ec:1.2.1.12 - dehydrogenase (phosphorylating), ec:6.3.5.3 - synthase, ec:1.1.1.25 - dehydrogenase, ec:2.4.2.18 - phosphoribosyltransferase, ec:6.3.4.13 - ligase, ec:1.5.1.2 - reductase, ec:2.3.3.1 - (Si)-synthase, ec:2.7.6.1 - diphosphokinase, ec:1.1.1.267 - reductoisomerase, ec:1.2.4.1 - dehydrogenase (acetyl-transferring), ec:3.5.4.10 - cyclohydrolase, ec:1.1.1.282 - dehydrogenase, ec:3.1.1.31 - phosphogluconolactonase, | EG_transcript_6624, EG_transcript_2009, EG_transcript_4764, EG_transcript_1939, EG_transcript_8284, EG_transcript_7914, EG_transcript_18439, EG_transcript_1490, EG_transcript_6489, EG_transcript_11162, EG_transcript_13699, EG_transcript_5154, EG_transcript_16158, EG_transcript_10012, EG_transcript_12808, EG_transcript_13795, EG_transcript_11659, EG_transcript_12376, EG_transcript_6665, EG_transcript_13075, EG_transcript_1208, EG_transcript_2997, EG_transcript_4605, EG_transcript_7788, EG_transcript_15457, EG_transcript_12896, EG_transcript_47268, EG_transcript_13245, EG_transcript_13871, EG_transcript_6058, EG_transcript_2820, EG_transcript_12119, EG_transcript_10778, EG_transcript_12260, EG_transcript_15698, EG_transcript_7726, EG_transcript_16693, EG_transcript_14099, EG_transcript_2957, EG_transcript_6211, EG_transcript_12607, EG_transcript_6612, EG_transcript_107, EG_transcript_6007, EG_transcript_6066, EG_transcript_6066, EG_transcript_16291, EG_transcript_21799, EG_transcript_8623, EG_transcript_8623, EG_transcript_11718, EG_transcript_27401, EG_transcript_13481, EG_transcript_1208, EG_transcript_9159, EG_transcript_15045, EG_transcript_16406, EG_transcript_11815, EG_transcript_11950, EG_transcript_21446, EG_transcript_1151, EG_transcript_7325, EG_transcript_6970, EG_transcript_6957, EG_transcript_6066, EG_transcript_12179, EG_transcript_6431, EG_transcript_1208, EG_transcript_4603, EG_transcript_3254, EG_transcript_9669, EG_transcript_12212, EG_transcript_15919, EG_transcript_24127, EG_transcript_7580, EG_transcript_29059, EG_transcript_26319, EG_transcript_862, EG_transcript_21457, EG_transcript_3075, EG_transcript_10083, EG_transcript_5584, EG_transcript_7394, EG_transcript_2646, EG_transcript_8727, EG_transcript_7683, EG_transcript_4570, EG_transcript_15632, EG_transcript_18342, EG_transcript_13690, EG_transcript_13898, EG_transcript_7631, EG_transcript_16470, EG_transcript_15075, EG_transcript_9064, EG_transcript_10115, EG_transcript_4351, EG_transcript_6875, EG_transcript_13843, EG_transcript_12509, EG_transcript_2747, EG_transcript_5995, EG_transcript_9049, EG_transcript_6731, EG_transcript_12921, EG_transcript_19893, EG_transcript_13804, EG_transcript_14099, EG_transcript_16406, EG_transcript_583, EG_transcript_21328, EG_transcript_15415, EG_transcript_10994, EG_transcript_18673, EG_transcript_20866, EG_transcript_2898, EG_transcript_6849, |

| | | | | |
|---|---|---|---|---|
| | | | ec:5.4.99.5 - mutase, ec:5.3.1.24 - isomerase | EG_transcript_10302, EG_transcript_11595, EG_transcript_7528, EG_transcript_14984, EG_transcript_13095, EG_transcript_4370, EG_transcript_14931, EG_transcript_9827, EG_transcript_5356, EG_transcript_21328, EG_transcript_2747, EG_transcript_19914, EG_transcript_12179, EG_transcript_1490 |
| Neomycin, kanamycin and gentamicin biosynthesis | map00524 | 2 | ec:2.7.1.2 - glucokinase (phosphorylating), ec:2.7.1.1 - hexokinase type IV glucokinase | EG_transcript_8623, EG_transcript_8623 |
| Selenocompound metabolism | map00450 | 6 | ec:1.8.1.9 - reductase, ec:2.1.1.13 - synthase, ec:2.1.1.14 - S-methyltransferase, ec:4.4.1.8 - beta-lyase, ec:2.7.7.4 - adenylyltransferase, ec:6.1.1.10 - ligase | EG_transcript_4888, EG_transcript_14795, EG_transcript_6835, EG_transcript_3052, EG_transcript_1784, EG_transcript_2610, EG_transcript_11187, EG_transcript_13843, EG_transcript_12509, EG_transcript_3873, EG_transcript_1145 |
| Monobactam biosynthesis | map00261 | 3 | ec:2.7.2.4 - kinase, ec:2.7.7.4 - adenylyltransferase, ec:1.2.1.11 - dehydrogenase | EG_transcript_2820, EG_transcript_13843, EG_transcript_12509, EG_transcript_14099 |
| Synthesis and degradation of ketone bodies | map00072 | 1 | ec:2.3.3.10 - synthase | EG_transcript_16836, EG_transcript_28, EG_transcript_5258 |
| alpha-Linolenic acid metabolism | map00592 | 1 | ec:1.1.1.1 - dehydrogenase | EG_transcript_21676, EG_transcript_19272 |
| Steroid degradation | map00984 | 1 | ec:1.1.1.145 - dehydrogenase | EG_transcript_15039 |
| Carotenoid biosynthesis | map00906 | 1 | ec:1.3.5.6 - desaturase | EG_transcript_3877 |
| Valine, leucine and isoleucine biosynthesis | map00290 | 8 | ec:2.3.3.13 - synthase, ec:4.2.1.9 - dehydratase, ec:2.2.1.6 - synthase, ec:4.3.1.19 - ammonia-lyase, ec:4.2.1.33 - dehydratase, ec:1.1.1.86 - reductoisomerase (NADP+), ec:1.1.1.85 - dehydrogenase, ec:2.6.1.42 - transaminase | EG_transcript_7582, EG_transcript_5664, EG_transcript_6624, EG_transcript_7788, EG_transcript_6007, EG_transcript_4532, EG_transcript_27401, EG_transcript_13481, EG_transcript_12096, EG_transcript_9159 |
| Fatty acid elongation | map00062 | 2 | ec:3.1.2.22 - hydrolase, ec:1.1.1.35 - dehydrogenase | EG_transcript_17203, EG_transcript_7580, EG_transcript_29059, EG_transcript_26319, EG_transcript_862, EG_transcript_21457, EG_transcript_3075 |
| Styrene degradation | map00643 | 1 | ec:3.7.1.2 - beta-diketonase | EG_transcript_9792 |
| Valine, leucine and isoleucine degradation | map00280 | 10 | ec:1.8.1.4 - dehydrogenase, ec:2.3.3.10 - synthase, ec:1.4.3.2 - oxidase, ec:1.2.1.27 - dehydrogenase (CoA-acylating), ec:2.6.1.42 - transaminase, ec:1.2.1.3 - dehydrogenase (NAD+), ec:1.1.1.35 - dehydrogenase, ec:1.1.1.31 - dehydrogenase, ec:6.2.1.16 - ligase, ec:5.4.99.2 - mutase | EG_transcript_8679, EG_transcript_16836, EG_transcript_28, EG_transcript_5258, EG_transcript_5672, EG_transcript_15045, EG_transcript_29172, EG_transcript_9159, EG_transcript_15045, EG_transcript_7580, EG_transcript_29059, EG_transcript_26319, EG_transcript_862, EG_transcript_21457, EG_transcript_3075, EG_transcript_13898, EG_transcript_4805, EG_transcript_4204 |

| | | | | |
|---|---|---|---|---|
| Lipoic acid metabolism | map00785 | 1 | ec:2.8.1.8 - synthase | EG_transcript_8825 |
| Galactose metabolism | map00052 | 8 | ec:2.7.7.64 - uridylyltransferase, ec:2.7.1.2 - glucokinase (phosphorylating), ec:2.7.1.1 - hexokinase type IV glucokinase, ec:2.7.1.6 - galactokinase (phosphorylating), ec:2.7.1.11 - phosphohexokinase, ec:5.1.3.2 - 4-epimerase, ec:2.7.7.9 - uridylyltransferase, ec:5.4.99.9 - mutase | EG_transcript_6731, EG_transcript_12921, EG_transcript_8623, EG_transcript_8623, EG_transcript_12330, EG_transcript_1151, EG_transcript_7325, EG_transcript_6970, EG_transcript_6957, EG_transcript_10795, EG_transcript_6731, EG_transcript_12921, EG_transcript_8646 |
| Amino sugar and nucleotide sugar metabolism | map00520 | 13 | ec:2.6.1.16 - transaminase (isomerizing), ec:5.3.1.9 - isomerase, ec:2.7.7.64 - uridylyltransferase, ec:4.1.1.35 - decarboxylase, ec:4.2.1.47 - 4,6-dehydratase, ec:2.7.1.2 - glucokinase (phosphorylating), ec:2.7.1.1 - hexokinase type IV glucokinase, ec:2.7.1.6 - galactokinase (phosphorylating), ec:5.1.3.2 - 4-epimerase, ec:2.7.7.9 - uridylyltransferase, ec:5.4.2.8 - mannose phosphomutase, ec:1.1.1.22 - 6-dehydrogenase, ec:5.4.99.9 - mutase | EG_transcript_5074, EG_transcript_4891, EG_transcript_5108, EG_transcript_6731, EG_transcript_12921, EG_transcript_12235, EG_transcript_15475, EG_transcript_8623, EG_transcript_8623, EG_transcript_12330, EG_transcript_10795, EG_transcript_6731, EG_transcript_12921, EG_transcript_19893, EG_transcript_14324, EG_transcript_8646 |
| Toluene degradation | map00623 | 1 | ec:1.1.1.35 - dehydrogenase | EG_transcript_7580, EG_transcript_29059, EG_transcript_26319, EG_transcript_862, EG_transcript_21457, EG_transcript_3075 |
| Biosynthesis of unsaturated fatty acids | map01040 | 2 | ec:1.1.1.100 - reductase, ec:1.14.19.2 - 9-desaturase | EG_transcript_11779, EG_transcript_7358 |
| Metabolism of xenobiotics by cytochrome P450 | map00980 | 2 | ec:1.1.1.1 - dehydrogenase, ec:1.2.1.5 - dehydrogenase [NAD(P)+] | EG_transcript_21676, EG_transcript_19272, EG_transcript_5269, EG_transcript_6432 |
| Phenylalanine metabolism | map00360 | 5 | ec:1.4.3.2 - oxidase, ec:4.1.3.39 - aldolase, ec:2.6.1.9 - transaminase, ec:1.2.1.5 - dehydrogenase [NAD(P)+], ec:1.13.11.27 - dioxygenase | EG_transcript_5672, EG_transcript_3463, EG_transcript_11162, EG_transcript_5269, EG_transcript_6432, EG_transcript_11176, EG_transcript_10118 |
| mTOR signaling pathway | map04150 | 1 | ec:2.7.11.24 - protein kinase | EG_transcript_7330 |
| Tyrosine metabolism | map00350 | 6 | ec:1.4.3.2 - oxidase, ec:1.1.1.1 - dehydrogenase, ec:2.6.1.9 - transaminase, ec:3.7.1.2 - beta-diketonase, ec:1.2.1.5 - dehydrogenase [NAD(P)+], ec:1.13.11.27 - dioxygenase | EG_transcript_5672, EG_transcript_21676, EG_transcript_19272, EG_transcript_11162, EG_transcript_9792, EG_transcript_5269, EG_transcript_6432, EG_transcript_11176, EG_transcript_10118 |
| Aminoacyl-tRNA biosynthesis | map00970 | 20 | ec:6.1.1.18 - ligase, ec:6.1.1.17 - ligase, ec:6.1.1.19 - ligase, ec:6.1.1.14 - ligase, ec:6.1.1.16 - ligase, ec:6.1.1.15 - ligase, ec:6.1.1.10 - ligase, ec:6.1.1.12 - ligase, ec:6.1.1.11 - ligase, ec:6.1.1.6 - ligase, ec:6.1.1.7 - ligase, ec:6.1.1.9 - ligase, ec:6.1.1.2 - ligase, ec:6.1.1.3 - ligase, | EG_transcript_2696, EG_transcript_6722, EG_transcript_4918, EG_transcript_2535, EG_transcript_3489, EG_transcript_3792, EG_transcript_4976, EG_transcript_11617, EG_transcript_4109, EG_transcript_9052, EG_transcript_3873, EG_transcript_1145, EG_transcript_7672, EG_transcript_9682, EG_transcript_26096, |

| | | | | |
|---|---|---|---|---|
| | | | ec:6.1.1.4 - ligase, ec:6.1.1.5 - ligase, ec:6.1.1.1 - ligase, ec:6.1.1.21 - ligase, ec:6.1.1.20 - ligase, ec:6.1.1.22 - ligase | EG_transcript_7877, EG_transcript_3471, EG_transcript_3246, EG_transcript_1453, EG_transcript_7320, EG_transcript_4373, EG_transcript_1177, EG_transcript_1330, EG_transcript_1145, EG_transcript_7879, EG_transcript_10635, EG_transcript_7994, EG_transcript_7320, EG_transcript_8464, EG_transcript_2696 |
| Aminobenzoate degradation | map00627 | 2 | ec:3.6.1.7 - acetylphosphatase, ec:3.1.3.41 - nitrophenyl phosphatase | EG_transcript_52815, EG_transcript_8839, EG_transcript_8518, EG_transcript_26406, EG_transcript_18287, EG_transcript_18342, EG_transcript_680, EG_transcript_13690, EG_transcript_17863, EG_transcript_14035, EG_transcript_7683, EG_transcript_4570, EG_transcript_15687, EG_transcript_8577, EG_transcript_2212, EG_transcript_21163, EG_transcript_12518, EG_transcript_2107, EG_transcript_2646, EG_transcript_7616, EG_transcript_4802, EG_transcript_6968, EG_transcript_8727, EG_transcript_1615, EG_transcript_19003, EG_transcript_7394, EG_transcript_9851, EG_transcript_8387, EG_transcript_24788, EG_transcript_9672, EG_transcript_15632, EG_transcript_24665, EG_transcript_13679, EG_transcript_1851, EG_transcript_11638 |
| Novobiocin biosynthesis | map00401 | 2 | ec:2.6.1.9 - transaminase, ec:1.3.1.12 - dehydrogenase | EG_transcript_11162, EG_transcript_6066 |
| Phosphonate and phosphinate metabolism | map00440 | 4 | ec:4.1.1.82 - decarboxylase, ec:2.6.1.37 - transaminase, ec:5.4.2.9 - mutase, ec:3.11.1.2 - hydrolase | EG_transcript_15698, EG_transcript_14715, EG_transcript_13804, EG_transcript_8786 |
| Histidine metabolism | map00340 | 9 | ec:2.6.1.9 - transaminase, ec:4.2.1.19 - dehydratase, ec:1.2.1.5 - dehydrogenase [NAD(P)+], ec:1.2.1.3 - dehydrogenase (NAD+), ec:3.6.1.31 - diphosphatase, ec:1.1.1.23 - dehydrogenase, ec:2.4.2.17 - phosphoribosyltransferase, ec:5.3.1.16 - isomerase, ec:3.5.4.19 - cyclohydrolase | EG_transcript_11162, EG_transcript_24054, EG_transcript_5269, EG_transcript_6432, EG_transcript_15045, EG_transcript_52739, EG_transcript_6602, EG_transcript_10604, EG_transcript_26398, EG_transcript_19445, EG_transcript_31973 |
| Tropane, piperidine and pyridine alkaloid biosynthesis | map00960 | 1 | ec:2.6.1.9 - transaminase | EG_transcript_11162 |
| Oxidative phosphorylation | map00190 | 6 | ec:1.6.5.3 - reductase (H+-translocating), ec:3.6.1.1 - diphosphatase, ec:1.3.5.1 - dehydrogenase, ec:1.10.2.2 - reductase, ec:1.9.3.1 - oxidase, ec:1.6.99.3 - dehydrogenase | EG_transcript_19403, EG_transcript_32395, EG_transcript_19840, EG_transcript_2633, EG_transcript_19429, EG_transcript_61813, EG_transcript_1590, EG_transcript_6467, EG_transcript_23844, EG_transcript_5228, EG_transcript_36506, EG_transcript_19403, EG_transcript_32395, EG_transcript_19840 |
| Taurine and | map00430 | 3 | ec:2.3.2.2 - glutamyl transpeptidase, ec:1.4.1.2 - | EG_transcript_5619, EG_transcript_791, EG_transcript_8124, |

| | | | | |
|---|---|---|---|---|
| hypotaurine metabolism | | | dehydrogenase, ec:1.4.1.1 - dehydrogenase | EG_transcript_11425 |
| Phosphatidylinositol signaling system | map04070 | 2 | ec:2.7.1.107 - kinase (ATP), ec:2.7.1.127 - 3-kinase | EG_transcript_14838, EG_transcript_23163 |
| Cysteine and methionine metabolism | map00270 | 16 | ec:2.5.1.6 - adenosyltransferase, ec:4.1.1.50 - decarboxylase, ec:3.1.3.77 - synthase, ec:1.4.3.2 - oxidase, ec:2.1.1.13 - synthase, ec:2.1.1.14 - S-methyltransferase, ec:3.3.1.1 - S-adenosylhomocysteine synthase, ec:2.5.1.47 - synthase, ec:2.7.2.4 - kinase, ec:6.3.2.3 - synthase, ec:6.3.2.2 - ligase, ec:2.6.1.42 - transaminase, ec:1.13.11.54 - dioxygenase [iron(II)-requiring], ec:1.1.1.37 - dehydrogenase, ec:4.4.1.8 - beta-lyase, ec:1.2.1.11 - dehydrogenase | EG_transcript_10098, EG_transcript_6938, EG_transcript_17863, EG_transcript_5672, EG_transcript_3052, EG_transcript_1784, EG_transcript_2610, EG_transcript_19742, EG_transcript_16158, EG_transcript_10012, EG_transcript_12808, EG_transcript_13795, EG_transcript_11659, EG_transcript_2820, EG_transcript_8300, EG_transcript_1613, EG_transcript_9159, EG_transcript_25256, EG_transcript_12212, EG_transcript_15919, EG_transcript_24127, EG_transcript_11187, EG_transcript_14099 |
| Folate biosynthesis | map00790 | 4 | ec:6.3.2.17 - synthase, ec:1.5.1.3 - reductase, ec:3.5.4.25 - cyclohydrolase II, ec:3.5.4.16 - cyclohydrolase I | EG_transcript_13840, EG_transcript_8517, EG_transcript_3866, EG_transcript_6561 |
| Inositol phosphate metabolism | map00562 | 4 | ec:5.3.1.1 - isomerase, ec:2.7.1.127 - 3-kinase, ec:5.5.1.4 - synthase, ec:1.2.1.18 - dehydrogenase (acetylating) | EG_transcript_14120, EG_transcript_30563, EG_transcript_18179, EG_transcript_30456, EG_transcript_44253, EG_transcript_23163, EG_transcript_6431, EG_transcript_15045 |
| Polyketide sugar unit biosynthesis | map00523 | 1 | ec:4.2.1.46 - 4,6-dehydratase | EG_transcript_21799 |
| beta-Alanine metabolism | map00410 | 5 | ec:6.3.2.1 - ligase (AMP-forming), ec:1.2.1.5 - dehydrogenase [NAD(P)+], ec:1.2.1.3 - dehydrogenase (NAD+), ec:4.1.1.9 - decarboxylase, ec:1.2.1.18 - dehydrogenase (acetylating) | EG_transcript_16919, EG_transcript_5269, EG_transcript_6432, EG_transcript_15045, EG_transcript_9247, EG_transcript_15045 |
| Glycine, serine and threonine metabolism | map00260 | 11 | ec:1.8.1.4 - dehydrogenase, ec:2.1.2.1 - hydroxymethyltransferase, ec:2.1.2.10 - S-aminomethyldihydrolipoylprotein:(6S)-tetrahydrofolate aminomethyltransferase (ammonia-forming), ec:1.1.1.1 - dehydrogenase, ec:4.2.1.20 - synthase, ec:2.6.1.52 - transaminase, ec:1.4.4.2 - dehydrogenase (aminomethyl-transferring), ec:1.1.1.95 - dehydrogenase, ec:2.7.2.4 - kinase, ec:4.3.1.19 - ammonia-lyase, ec:1.2.1.11 - dehydrogenase | EG_transcript_8679, EG_transcript_9546, EG_transcript_7542, EG_transcript_3442, EG_transcript_9405, EG_transcript_9664, EG_transcript_21676, EG_transcript_19272, EG_transcript_8284, EG_transcript_7914, EG_transcript_18439, EG_transcript_1490, EG_transcript_6489, EG_transcript_1548, EG_transcript_13699, EG_transcript_5154, EG_transcript_2820, EG_transcript_6007, EG_transcript_14099 |
| Fatty acid degradation | map00071 | 3 | ec:1.1.1.1 - dehydrogenase, ec:1.2.1.3 - dehydrogenase (NAD+), ec:1.1.1.35 - dehydrogenase | EG_transcript_21676, EG_transcript_19272, EG_transcript_15045, EG_transcript_7580, EG_transcript_29059, EG_transcript_26319, EG_transcript_862, EG_transcript_21457, EG_transcript_3075 |
| Starch and sucrose metabolism | map00500 | 5 | ec:5.3.1.9 - isomerase, ec:2.7.1.2 - glucokinase (phosphorylating), ec:2.7.1.1 - hexokinase type IV | EG_transcript_5108, EG_transcript_8623, EG_transcript_8623, EG_transcript_6731, EG_transcript_12921, EG_transcript_11758, |

| | | | | |
|---|---|---|---|---|
| | | | glucokinase, ec:2.7.7.9 - uridylyltransferase, ec:2.4.1.34 - synthase | EG_transcript_73 |
| Drug metabolism - other enzymes | map00983 | 8 | ec:2.7.4.6 - kinase, ec:1.1.1.205 - dehydrogenase, ec:3.6.1.23 - diphosphatase, ec:6.3.5.2 - synthase (glutamine-hydrolysing), ec:3.1.1.1 - ali-esterase, ec:2.4.2.10 - phosphoribosyltransferase, ec:2.4.2.8 - phosphoribosyltransferase, ec:2.4.2.3 - phosphorylase | EG_transcript_17910, EG_transcript_11790, EG_transcript_26205, EG_transcript_9707, EG_transcript_8320, EG_transcript_13311, EG_transcript_24436, EG_transcript_4795, EG_transcript_4109, EG_transcript_17245, EG_transcript_2747, EG_transcript_1330, EG_transcript_11766, EG_transcript_19914, EG_transcript_1177, EG_transcript_1453, EG_transcript_22241, EG_transcript_9327, EG_transcript_20333, EG_transcript_25833 |
| Th1 and Th2 cell differentiation | map04658 | 1 | ec:3.1.3.16 - phosphatase | EG_transcript_2107, EG_transcript_4802, EG_transcript_6968, EG_transcript_1615, EG_transcript_26406, EG_transcript_19003, EG_transcript_18287, EG_transcript_14035, EG_transcript_9851, EG_transcript_9672, EG_transcript_8577, EG_transcript_13679, EG_transcript_2212, EG_transcript_1851, EG_transcript_12518 |
| Alanine, aspartate and glutamate metabolism | map00250 | 16 | ec:1.4.3.16 - oxidase, ec:2.6.1.16 - transaminase (isomerizing), ec:6.3.4.4 - synthase, ec:6.3.4.5 - synthase, ec:1.4.3.2 - oxidase, ec:6.3.1.2 - synthetase, ec:4.3.2.1 - lyase, ec:4.3.2.2 - lyase, ec:1.4.1.14 - synthase (NADH), ec:2.1.3.2 - carbamoyltransferase, ec:6.3.5.4 - synthase (glutamine-hydrolysing), ec:6.3.5.5 - synthase (glutamine-hydrolysing), ec:1.4.1.3 - dehydrogenase [NAD(P)+], ec:1.4.1.2 - dehydrogenase, ec:1.4.1.1 - dehydrogenase, ec:3.5.1.2 - glutaminase I | EG_transcript_5672, EG_transcript_5074, EG_transcript_4891, EG_transcript_929, EG_transcript_10767, EG_transcript_5672, EG_transcript_3524, EG_transcript_13142, EG_transcript_24268, EG_transcript_19219, EG_transcript_7726, EG_transcript_16693, EG_transcript_107, EG_transcript_7470, EG_transcript_14992, EG_transcript_5296, EG_transcript_224, EG_transcript_478, EG_transcript_791, EG_transcript_791, EG_transcript_8124, EG_transcript_11425, EG_transcript_2202 |
| Fatty acid biosynthesis | map00061 | 5 | ec:6.4.1.2 - carboxylase, ec:2.3.1.85 - synthase, ec:1.1.1.100 - reductase, ec:1.14.19.2 - 9-desaturase, ec:1.3.1.9 - reductase (NADH) | EG_transcript_65, EG_transcript_97, EG_transcript_6633, EG_transcript_9, EG_transcript_11779, EG_transcript_11779, EG_transcript_7358, EG_transcript_9 |
| T cell receptor signaling pathway | map04660 | 1 | ec:3.1.3.16 - phosphatase | EG_transcript_2107, EG_transcript_4802, EG_transcript_6968, EG_transcript_1615, EG_transcript_26406, EG_transcript_19003, EG_transcript_18287, EG_transcript_14035, EG_transcript_9851, EG_transcript_9672, EG_transcript_8577, EG_transcript_13679, EG_transcript_2212, EG_transcript_1851, EG_transcript_12518 |
| Porphyrin and chlorophyll metabolism | map00860 | 17 | ec:1.2.1.70 - reductase, ec:4.1.1.37 - decarboxylase, ec:2.7.8.26 - ribazoletransferase, ec:1.16.1.5 - reductase (NADPH), ec:2.1.1.11 - protoporphyrin IX methyltransferase, ec:1.3.1.33 - reductase, ec:4.2.1.24 - synthase, ec:2.5.1.62 - synthase, ec:2.5.1.61 - synthase, ec:1.14.13.122 - oxygenase, ec:5.4.3.8 - 2,1-aminomutase, ec:1.14.13.81 - IX monomethyl ester | EG_transcript_3435, EG_transcript_10984, EG_transcript_6636, EG_transcript_17708, EG_transcript_9735, EG_transcript_12772, EG_transcript_2112, EG_transcript_790, EG_transcript_17571, EG_transcript_6647, EG_transcript_8688, EG_transcript_16460, EG_transcript_9385, EG_transcript_7051, EG_transcript_3991, EG_transcript_4984, EG_transcript_13553, EG_transcript_13855, EG_transcript_4918, EG_transcript_16400, EG_transcript_9328, |

| | | | | |
|---|---|---|---|---|
| | | | (oxidative) cyclase, ec:4.99.1.1 - ferrochelatase, ec:6.1.1.17 - ligase, ec:1.3.3.4 - oxidase, ec:1.3.3.3 - oxidase, ec:6.6.1.1 - chelatase | EG_transcript_12610, EG_transcript_4741, EG_transcript_8456, EG_transcript_516, EG_transcript_2756, EG_transcript_202, EG_transcript_11387, EG_transcript_15491 |
| Pyrimidine metabolism | map00240 | 16 | ec:2.7.4.6 - kinase, ec:2.1.1.45 - synthase, ec:3.5.2.3 - carbamoylaspartic dehydrase, ec:3.1.3.5 - uridine 5'-nucleotidase, ec:6.3.4.2 - synthase (glutamine hydrolysing), ec:1.3.5.2 - dehydrogenase (quinone), ec:4.1.1.23 - decarboxylase, ec:1.17.4.2 - reductase (thioredoxin), ec:2.1.3.2 - carbamoyltransferase, ec:3.6.1.23 - diphosphatase, ec:2.7.7.6 - RNA polymerase, ec:2.7.7.7 - DNA polymerase, ec:6.3.5.5 - synthase (glutamine-hydrolysing), ec:2.4.2.10 - phosphoribosyltransferase, ec:2.4.2.6 - deoxyribosyltransferase, ec:2.4.2.3 - phosphorylase | EG_transcript_17910, EG_transcript_11790, EG_transcript_26205, EG_transcript_8517, EG_transcript_14846, EG_transcript_8387, EG_transcript_4156, EG_transcript_10912, EG_transcript_8554, EG_transcript_9327, EG_transcript_9422, EG_transcript_20718, EG_transcript_2191, EG_transcript_7470, EG_transcript_14992, EG_transcript_13311, EG_transcript_1236, EG_transcript_7627, EG_transcript_18378, EG_transcript_15458, EG_transcript_2051, EG_transcript_15521, EG_transcript_26985, EG_transcript_32593, EG_transcript_92, EG_transcript_464, EG_transcript_23275, EG_transcript_356, EG_transcript_13664, EG_transcript_1356, EG_transcript_224, EG_transcript_478, EG_transcript_9327, EG_transcript_23092, EG_transcript_25833 |
| Fructose and mannose metabolism | map00051 | 10 | ec:2.7.1.90 - 1-phosphotransferase, ec:2.7.1.105 - phosphofructokinase 2, ec:5.3.1.1 - isomerase, ec:4.1.2.13 - aldolase, ec:5.3.1.6 - isomerase, ec:4.2.1.47 - 4,6-dehydratase, ec:2.7.1.1 - hexokinase type IV glucokinase, ec:2.7.1.11 - phosphohexokinase, ec:3.1.3.11 - hexose diphosphatase, ec:5.4.2.8 - mannose phosphomutase | EG_transcript_1151, EG_transcript_10541, EG_transcript_8196, EG_transcript_10980, EG_transcript_4882, EG_transcript_7956, EG_transcript_14120, EG_transcript_30563, EG_transcript_18179, EG_transcript_30456, EG_transcript_44253, EG_transcript_7827, EG_transcript_21524, EG_transcript_43521, EG_transcript_16008, EG_transcript_25832, EG_transcript_5279, EG_transcript_15855, EG_transcript_15619, EG_transcript_20009, EG_transcript_15475, EG_transcript_8623, EG_transcript_1151, EG_transcript_7325, EG_transcript_6970, EG_transcript_6957, EG_transcript_7394, EG_transcript_2646, EG_transcript_8727, EG_transcript_7683, EG_transcript_4570, EG_transcript_15632, EG_transcript_18342, EG_transcript_13690, EG_transcript_19893 |
| Isoquinoline alkaloid biosynthesis | map00950 | 1 | ec:1.4.3.2 - oxidase | EG_transcript_5672 |
| Arginine and proline metabolism | map00330 | 5 | ec:4.1.1.50 - decarboxylase, ec:2.7.2.11 - 5-kinase, ec:1.2.1.41 - dehydrogenase, ec:1.2.1.3 - dehydrogenase (NAD+), ec:1.5.1.2 - reductase | EG_transcript_6938, EG_transcript_12607, EG_transcript_6612, EG_transcript_15045, EG_transcript_18673, EG_transcript_20866 |
| Aflatoxin biosynthesis | map00254 | 1 | ec:6.4.1.2 - carboxylase | EG_transcript_65, EG_transcript_97, EG_transcript_6633 |
| Xylene degradation | map00622 | 1 | ec:4.1.3.39 - aldolase | EG_transcript_3463 |
| Indole alkaloid biosynthesis | map00901 | 1 | ec:4.3.3.2 - synthase | EG_transcript_20082 |
| Phenylpropanoid | map00940 | 1 | ec:1.11.1.7 - lactoperoxidase | EG_transcript_29973, EG_transcript_25331, EG_transcript_20841, |

| | | | | |
|---|---|---|---|---|
| biosynthesis | | | | EG_transcript_5780, EG_transcript_21143, EG_transcript_22491, EG_transcript_13727, EG_transcript_27461 |
| Lysine degradation | map00310 | 2 | ec:1.2.1.3 - dehydrogenase (NAD+), ec:1.1.1.35 - dehydrogenase | EG_transcript_15045, EG_transcript_7580, EG_transcript_29059, EG_transcript_26319, EG_transcript_862, EG_transcript_21457, EG_transcript_3075 |
| Caprolactam degradation | map00930 | 1 | ec:1.1.1.35 - dehydrogenase | EG_transcript_7580, EG_transcript_29059, EG_transcript_26319, EG_transcript_862, EG_transcript_21457, EG_transcript_3075 |
| Naphthalene degradation | map00626 | 1 | ec:1.1.1.1 - dehydrogenase | EG_transcript_21676, EG_transcript_19272 |
| Biotin metabolism | map00780 | 3 | ec:6.3.3.3 - synthase, ec:2.8.1.6 - synthase, ec:1.1.1.100 - reductase | EG_transcript_38077, EG_transcript_14092, EG_transcript_11779 |
| Phenylalanine, tyrosine and tryptophan biosynthesis | map00400 | 17 | ec:4.1.1.48 - synthase, ec:1.4.3.2 - oxidase, ec:4.2.1.20 - synthase, ec:2.6.1.9 - transaminase, ec:4.2.1.10 - dehydratase, ec:2.5.1.54 - synthase, ec:1.3.1.13 - dehydrogenase (NADP+), ec:1.3.1.12 - dehydrogenase, ec:4.2.3.5 - synthase, ec:4.2.3.4 - synthase, ec:4.2.1.51 - dehydratase, ec:2.5.1.19 - 1-carboxyvinyltransferase, ec:1.1.1.25 - dehydrogenase, ec:2.4.2.18 - phosphoribosyltransferase, ec:1.1.1.282 - dehydrogenase, ec:5.4.99.5 - mutase, ec:5.3.1.24 - isomerase | EG_transcript_1490, EG_transcript_5672, EG_transcript_8284, EG_transcript_7914, EG_transcript_18439, EG_transcript_1490, EG_transcript_11162, EG_transcript_1208, EG_transcript_12896, EG_transcript_47268, EG_transcript_13245, EG_transcript_13871, EG_transcript_6066, EG_transcript_6066, EG_transcript_11718, EG_transcript_1208, EG_transcript_6066, EG_transcript_12179, EG_transcript_1208, EG_transcript_21328, EG_transcript_15415, EG_transcript_21328, EG_transcript_12179, EG_transcript_1490 |
| Lysine biosynthesis | map00300 | 4 | ec:2.3.3.14 - synthase, ec:2.7.2.4 - kinase, ec:4.2.1.36 - hydratase, ec:1.2.1.11 - dehydrogenase | EG_transcript_4221, EG_transcript_2820, EG_transcript_11950, EG_transcript_21446, EG_transcript_14099 |
| Sulfur metabolism | map00920 | 5 | ec:3.1.3.7 - nucleotidase, ec:2.5.1.47 - synthase, ec:2.7.1.25 - kinase, ec:1.8.4.8 - reductase (thioredoxin), ec:2.7.7.4 - adenylyltransferase | EG_transcript_8518, EG_transcript_16158, EG_transcript_10012, EG_transcript_12808, EG_transcript_13795, EG_transcript_11659, EG_transcript_68647, EG_transcript_13057, EG_transcript_13843, EG_transcript_12509 |
| Pantothenate and CoA biosynthesis | map00770 | 9 | ec:2.1.2.11 - hydroxymethyltransferase, ec:4.2.1.9 - dehydratase, ec:2.2.1.6 - synthase, ec:6.3.2.1 - ligase (AMP-forming), ec:1.1.1.86 - reductoisomerase (NADP+), ec:2.6.1.42 - transaminase, ec:2.7.1.24 - kinase, ec:2.7.8.7 - synthase, ec:1.1.1.169 - 2-reductase | EG_transcript_14237, EG_transcript_6624, EG_transcript_7788, EG_transcript_16919, EG_transcript_27401, EG_transcript_13481, EG_transcript_9159, EG_transcript_39815, EG_transcript_9, EG_transcript_24449 |
| Steroid hormone biosynthesis | map00140 | 1 | ec:1.1.1.145 - dehydrogenase | EG_transcript_15039 |
| Nicotinate and nicotinamide metabolism | map00760 | 5 | ec:1.4.3.16 - oxidase, ec:3.1.3.5 - uridine 5'-nucleotidase, ec:1.6.1.2 - transhydrogenase (Re/Si-specific), ec:6.3.5.1 - synthase (glutamine-hydrolysing), ec:2.4.2.19 - diphosphorylase (carboxylating) | EG_transcript_5672, EG_transcript_8387, EG_transcript_8446, EG_transcript_2419, EG_transcript_6238, EG_transcript_45301, EG_transcript_17502 |
| Purine metabolism | map00230 | 27 | ec:2.1.2.3 - formyltransferase, ec:2.1.2.2 - | EG_transcript_5356, EG_transcript_36522, EG_transcript_1073, |

| | | | | |
|---|---|---|---|---|
| | | | formyltransferase, ec:4.6.1.1 - cyclase, ec:2.7.4.8 - kinase, ec:2.7.4.6 - kinase, ec:2.7.4.3 - kinase, ec:3.1.3.5 - uridine 5'-nucleotidase, ec:6.3.4.4 - synthase, ec:3.6.1.3 - adenylpyrophosphatase, ec:4.1.1.21 - carboxylase, ec:3.1.4.17 - phosphodiesterase, ec:4.3.2.2 - lyase, ec:1.17.4.2 - reductase (thioredoxin), ec:2.7.1.40 - kinase, ec:1.1.1.205 - dehydrogenase, ec:2.7.1.25 - kinase, ec:2.7.1.20 - kinase, ec:3.6.1.15 - phosphatase, ec:2.7.7.4 - adenylyltransferase, ec:2.7.7.6 - RNA polymerase, ec:2.7.7.7 - DNA polymerase, ec:6.3.5.2 - synthase (glutamine-hydrolysing), ec:6.3.5.3 - synthase, ec:6.3.4.13 - ligase, ec:2.7.6.1 - diphosphokinase, ec:3.5.4.10 - cyclohydrolase, ec:2.4.2.8 - phosphoribosyltransferase | EG_transcript_2020, EG_transcript_3792, EG_transcript_24336, EG_transcript_20583, EG_transcript_17910, EG_transcript_11790, EG_transcript_26205, EG_transcript_22143, EG_transcript_8387, EG_transcript_929, EG_transcript_4118, EG_transcript_10317, EG_transcript_1888, EG_transcript_567, EG_transcript_7192, EG_transcript_6491, EG_transcript_13141, EG_transcript_8030, EG_transcript_3260, EG_transcript_5164, EG_transcript_11643, EG_transcript_1241, EG_transcript_1121, EG_transcript_1245, EG_transcript_3018, EG_transcript_4745, EG_transcript_159, EG_transcript_950, EG_transcript_74, EG_transcript_155, EG_transcript_795, EG_transcript_20888, EG_transcript_6088, EG_transcript_1350, EG_transcript_1352, EG_transcript_1993, EG_transcript_944, EG_transcript_10538, EG_transcript_5627, EG_transcript_36488, EG_transcript_260, EG_transcript_21665, EG_transcript_23566, EG_transcript_12798, EG_transcript_1663, EG_transcript_5227, EG_transcript_14, EG_transcript_1139, EG_transcript_1656, EG_transcript_32956, EG_transcript_253, EG_transcript_18273, EG_transcript_254, EG_transcript_10021, EG_transcript_251, EG_transcript_22409, EG_transcript_19235, EG_transcript_8289, EG_transcript_1372, EG_transcript_3154, EG_transcript_7996, EG_transcript_3310, EG_transcript_7356, EG_transcript_1016, EG_transcript_35708, EG_transcript_129, EG_transcript_4714, EG_transcript_2819, EG_transcript_10193, EG_transcript_6572, EG_transcript_5363, EG_transcript_11167, EG_transcript_4275, EG_transcript_29, EG_transcript_8912, EG_transcript_3104, EG_transcript_9603, EG_transcript_10923, EG_transcript_910, EG_transcript_34, EG_transcript_1953, EG_transcript_353, EG_transcript_10241, EG_transcript_7497, EG_transcript_18009, EG_transcript_16984, EG_transcript_46, EG_transcript_31678, EG_transcript_901, EG_transcript_1901, EG_transcript_744, EG_transcript_8893, EG_transcript_5143, EG_transcript_5387, EG_transcript_6513, EG_transcript_4327, EG_transcript_5, EG_transcript_52, EG_transcript_854, EG_transcript_2947, EG_transcript_579, EG_transcript_297, EG_transcript_331, EG_transcript_573, EG_transcript_25828, EG_transcript_49747, EG_transcript_23012, EG_transcript_2423, EG_transcript_6217, EG_transcript_7167, EG_transcript_22300, EG_transcript_13675, EG_transcript_16693, EG_transcript_9422, EG_transcript_20718, EG_transcript_2191, EG_transcript_2957, |

| | | | | |
|---|---|---|---|---|
| | | | | EG_transcript_6211, EG_transcript_9707, EG_transcript_8320, EG_transcript_68647, EG_transcript_18833, EG_transcript_1523, EG_transcript_4118, EG_transcript_10317, EG_transcript_1888, EG_transcript_567, EG_transcript_321, EG_transcript_7192, EG_transcript_13141, EG_transcript_8030, EG_transcript_3260, EG_transcript_3381, EG_transcript_8154, EG_transcript_11643, EG_transcript_3148, EG_transcript_3018, EG_transcript_1079, EG_transcript_6648, EG_transcript_1877, EG_transcript_2847, EG_transcript_74, EG_transcript_10562, EG_transcript_795, EG_transcript_12190, EG_transcript_6088, EG_transcript_13158, EG_transcript_6889, EG_transcript_1993, EG_transcript_10538, EG_transcript_30386, EG_transcript_664, EG_transcript_36488, EG_transcript_19234, EG_transcript_19476, EG_transcript_3280, EG_transcript_21665, EG_transcript_20574, EG_transcript_23963, EG_transcript_1663, EG_transcript_5227, EG_transcript_1535, EG_transcript_1656, EG_transcript_1536, EG_transcript_18273, EG_transcript_91, EG_transcript_22409, EG_transcript_19235, EG_transcript_8289, EG_transcript_8722, EG_transcript_3154, EG_transcript_7996, EG_transcript_4245, EG_transcript_16527, EG_transcript_5456, EG_transcript_35708, EG_transcript_526, EG_transcript_2819, EG_transcript_7027, EG_transcript_2494, EG_transcript_3584, EG_transcript_2134, EG_transcript_4312, EG_transcript_8912, EG_transcript_21240, EG_transcript_3104, EG_transcript_3337, EG_transcript_3458, EG_transcript_10923, EG_transcript_515, EG_transcript_1831, EG_transcript_1953, EG_transcript_2925, EG_transcript_1959, EG_transcript_10241, EG_transcript_15140, EG_transcript_22226, EG_transcript_22347, EG_transcript_7497, EG_transcript_23555, EG_transcript_505, EG_transcript_744, EG_transcript_10211, EG_transcript_191, EG_transcript_14010, EG_transcript_15106, EG_transcript_2274, EG_transcript_6513, EG_transcript_4327, EG_transcript_1056, EG_transcript_5, EG_transcript_854, EG_transcript_2947, EG_transcript_297, EG_transcript_25828, EG_transcript_8006, EG_transcript_2387, EG_transcript_1126, EG_transcript_28802, EG_transcript_6491, EG_transcript_5164, EG_transcript_8557, EG_transcript_1120, EG_transcript_1241, EG_transcript_1121, EG_transcript_1245, EG_transcript_5844, EG_transcript_4745, EG_transcript_159, EG_transcript_950, EG_transcript_155, EG_transcript_150, EG_transcript_5390, EG_transcript_19532, |

| | | | | |
|---|---|---|---|---|
| | | | | EG_transcript_20888, EG_transcript_8662, EG_transcript_1470, EG_transcript_1350, EG_transcript_1352, EG_transcript_5039, EG_transcript_944, EG_transcript_4416, EG_transcript_4537, EG_transcript_5627, EG_transcript_260, EG_transcript_23566, EG_transcript_12798, EG_transcript_1144, EG_transcript_2113, EG_transcript_14, EG_transcript_1139, EG_transcript_5617, EG_transcript_32956, EG_transcript_253, EG_transcript_254, EG_transcript_255, EG_transcript_10021, EG_transcript_251, EG_transcript_1372, EG_transcript_7234, EG_transcript_1495, EG_transcript_3310, EG_transcript_7356, EG_transcript_1016, EG_transcript_129, EG_transcript_3745, EG_transcript_4714, EG_transcript_367, EG_transcript_25212, EG_transcript_10193, EG_transcript_19330, EG_transcript_6572, EG_transcript_5363, EG_transcript_11167, EG_transcript_4275, EG_transcript_29, EG_transcript_9603, EG_transcript_910, EG_transcript_34, EG_transcript_2649, EG_transcript_353, EG_transcript_3170, EG_transcript_13352, EG_transcript_18009, EG_transcript_16984, EG_transcript_3176, EG_transcript_2884, EG_transcript_46, EG_transcript_31678, EG_transcript_901, EG_transcript_25311, EG_transcript_1901, EG_transcript_12480, EG_transcript_5020, EG_transcript_8893, EG_transcript_5143, EG_transcript_5387, EG_transcript_12357, EG_transcript_6599, EG_transcript_2550, EG_transcript_1100, EG_transcript_1221, EG_transcript_5941, EG_transcript_3521, EG_transcript_1102, EG_transcript_4853, EG_transcript_3639, EG_transcript_52, EG_transcript_579, EG_transcript_331, EG_transcript_573, EG_transcript_15990, EG_transcript_3196, EG_transcript_6461, EG_transcript_49747, EG_transcript_23012, EG_transcript_2423, EG_transcript_10506, EG_transcript_13843, EG_transcript_12509, EG_transcript_1236, EG_transcript_7627, EG_transcript_18378, EG_transcript_15458, EG_transcript_2051, EG_transcript_15521, EG_transcript_26985, EG_transcript_32593, EG_transcript_92, EG_transcript_464, EG_transcript_23275, EG_transcript_356, EG_transcript_13664, EG_transcript_1356, EG_transcript_24436, EG_transcript_4795, EG_transcript_583, EG_transcript_10994, EG_transcript_7528, EG_transcript_14984, EG_transcript_13095, EG_transcript_5356, EG_transcript_20333 |
| Vitamin B6 metabolism | map00750 | 2 | ec:2.6.1.52 - transaminase, ec:2.7.1.35 - kinase | EG_transcript_6489, EG_transcript_14152 |

| | | | | |
|---|---|---|---|---|
| Glycerolipid metabolism | map00561 | 4 | ec:2.7.1.107 - kinase (ATP), ec:2.7.1.30 - kinase, ec:1.2.1.3 - dehydrogenase (NAD+), ec:2.3.1.20 - O-acyltransferase | EG_transcript_14838, EG_transcript_9510, EG_transcript_15045, EG_transcript_27587, EG_transcript_6847, EG_transcript_5274 |
| Arginine biosynthesis | map00220 | 8 | ec:6.3.4.5 - synthase, ec:6.3.1.2 - synthetase, ec:4.3.2.1 - lyase, ec:1.2.1.38 - reductase, ec:3.5.1.14 - acid amidohydrolase, ec:1.4.1.3 - dehydrogenase [NAD(P)+], ec:1.4.1.2 - dehydrogenase, ec:3.5.1.2 - glutaminase I | EG_transcript_10767, EG_transcript_3524, EG_transcript_13142, EG_transcript_24268, EG_transcript_19219, EG_transcript_7726, EG_transcript_14099, EG_transcript_11815, EG_transcript_791, EG_transcript_791, EG_transcript_2202 |
| Arachidonic acid metabolism | map00590 | 1 | ec:1.11.1.9 - peroxidase | EG_transcript_29973, EG_transcript_21143, EG_transcript_20841, EG_transcript_27461 |
| Drug metabolism - cytochrome P450 | map00982 | 3 | ec:1.1.1.1 - dehydrogenase, ec:1.14.13.8 - monooxygenase, ec:1.2.1.5 - dehydrogenase [NAD(P)+] | EG_transcript_21676, EG_transcript_19272, EG_transcript_8408, EG_transcript_7495, EG_transcript_5269, EG_transcript_6432 |
| Benzoate degradation | map00362 | 2 | ec:4.1.3.39 - aldolase, ec:1.1.1.35 - dehydrogenase | EG_transcript_3463, EG_transcript_7580, EG_transcript_29059, EG_transcript_26319, EG_transcript_862, EG_transcript_21457, EG_transcript_3075 |
| Retinol metabolism | map00830 | 1 | ec:1.1.1.1 - dehydrogenase | EG_transcript_21676, EG_transcript_19272 |
| Ether lipid metabolism | map00565 | 2 | ec:2.5.1.26 - synthase, ec:3.1.1.47 - esterase | EG_transcript_5463, EG_transcript_11766 |
| Methane metabolism | map00680 | 12 | ec:2.1.2.1 - hydroxymethyltransferase, ec:3.1.3.71 - phosphatase, ec:4.1.2.13 - aldolase, ec:4.1.1.31 - carboxylase, ec:2.6.1.52 - transaminase, ec:1.1.1.95 - dehydrogenase, ec:4.2.1.11 - hydratase, ec:2.7.1.11 - phosphohexokinase, ec:6.2.1.1 - ligase, ec:1.1.1.37 - dehydrogenase, ec:3.1.3.11 - hexose diphosphatase, ec:3.1.2.12 - hydrolase | EG_transcript_9546, EG_transcript_7542, EG_transcript_3442, EG_transcript_9405, EG_transcript_21163, EG_transcript_7827, EG_transcript_21524, EG_transcript_43521, EG_transcript_16008, EG_transcript_25832, EG_transcript_5279, EG_transcript_15855, EG_transcript_1597, EG_transcript_6489, EG_transcript_13699, EG_transcript_5154, EG_transcript_12376, EG_transcript_6665, EG_transcript_1151, EG_transcript_7325, EG_transcript_6970, EG_transcript_6957, EG_transcript_4603, EG_transcript_3254, EG_transcript_9669, EG_transcript_12212, EG_transcript_15919, EG_transcript_24127, EG_transcript_7394, EG_transcript_2646, EG_transcript_8727, EG_transcript_7683, EG_transcript_4570, EG_transcript_15632, EG_transcript_18342, EG_transcript_13690, EG_transcript_17257 |
| One carbon pool by folate | map00670 | 11 | ec:2.1.2.3 - formyltransferase, ec:2.1.2.2 - formyltransferase, ec:2.1.2.1 - hydroxymethyltransferase, ec:2.1.2.10 - S-aminomethyldihydrolipoylprotein:(6S)-tetrahydrofolate aminomethyltransferase (ammonia-forming), ec:2.1.1.45 - synthase, ec:6.3.4.3 - ligase, ec:1.5.1.20 - reductase [NAD(P)H], ec:2.1.1.13 - synthase, ec:3.5.4.9 - cyclohydrolase, ec:1.5.1.3 - reductase, ec:1.5.1.5 - dehydrogenase (NADP+) | EG_transcript_5356, EG_transcript_36522, EG_transcript_9546, EG_transcript_7542, EG_transcript_3442, EG_transcript_9405, EG_transcript_9664, EG_transcript_8517, EG_transcript_17312, EG_transcript_15809, EG_transcript_12465, EG_transcript_3052, EG_transcript_1784, EG_transcript_17805, EG_transcript_8517, EG_transcript_17805, EG_transcript_17733 |

| | | | | |
|---|---|---|---|---|
| Nitrogen metabolism | map00910 | 7 | ec:4.2.1.1 - anhydrase, ec:6.3.1.2 - synthetase, ec:4.2.1.104 - cyanate lyase, ec:1.4.1.14 - synthase (NADH), ec:1.13.12.16 - monooxygenase, ec:1.4.1.3 - dehydrogenase [NAD(P)+], ec:1.4.1.2 - dehydrogenase | EG_transcript_20306, EG_transcript_13779, EG_transcript_11777, EG_transcript_12758, EG_transcript_3524, EG_transcript_13142, EG_transcript_24268, EG_transcript_19219, EG_transcript_14786, EG_transcript_107, EG_transcript_22015, EG_transcript_791, EG_transcript_791 |
| Dioxin degradation | map00621 | 1 | ec:4.1.3.39 - aldolase | EG_transcript_3463 |
| Pentose and glucuronate interconversions | map00040 | 3 | ec:2.7.7.64 - uridylyltransferase, ec:2.7.7.9 - uridylyltransferase, ec:1.1.1.22 - 6-dehydrogenase | EG_transcript_6731, EG_transcript_12921, EG_transcript_6731, EG_transcript_12921, EG_transcript_14324 |
| C5-Branched dibasic acid metabolism | map00660 | 2 | ec:2.2.1.6 - synthase, ec:1.1.1.85 - dehydrogenase | EG_transcript_7788, EG_transcript_12096 |
| D-Glutamine and D-glutamate metabolism | map00471 | 2 | ec:1.4.1.3 - dehydrogenase [NAD(P)+], ec:3.5.1.2 - glutaminase I | EG_transcript_791, EG_transcript_2202 |
| Terpenoid backbone biosynthesis | map00900 | 9 | ec:2.3.3.10 - synthase, ec:2.7.7.60 - 4-phosphate cytidylyltransferase, ec:4.1.1.33 - decarboxylase, ec:1.17.7.1 - synthase (ferredoxin), ec:5.3.3.2 - Delta-isomerase, ec:2.2.1.7 - synthase, ec:2.7.1.36 - kinase, ec:1.1.1.34 - reductase (NADPH), ec:1.1.1.267 - reductoisomerase | EG_transcript_16836, EG_transcript_28, EG_transcript_5258, EG_transcript_14037, EG_transcript_17330, EG_transcript_1939, EG_transcript_15457, EG_transcript_6058, EG_transcript_16291, EG_transcript_10083, EG_transcript_4370 |
| Ubiquinone and other terpenoid-quinone biosynthesis | map00130 | 2 | ec:2.2.1.9 - synthase, ec:1.13.11.27 - dioxygenase | EG_transcript_121, EG_transcript_11176, EG_transcript_10118 |
| Butanoate metabolism | map00650 | 6 | ec:2.3.3.10 - synthase, ec:1.3.5.1 - dehydrogenase, ec:2.2.1.6 - synthase, ec:1.3.1.44 - reductase (NAD+), ec:1.1.1.35 - dehydrogenase, ec:6.2.1.16 - ligase | EG_transcript_16836, EG_transcript_28, EG_transcript_5258, EG_transcript_6467, EG_transcript_7788, EG_transcript_7247, EG_transcript_7580, EG_transcript_29059, EG_transcript_26319, EG_transcript_862, EG_transcript_21457, EG_transcript_3075, EG_transcript_4805 |
| Glucosinolate biosynthesis | map00966 | 1 | ec:2.6.1.42 - transaminase | EG_transcript_9159 |
| Chloroalkane and chloroalkene degradation | map00625 | 2 | ec:1.1.1.1 - dehydrogenase, ec:1.2.1.3 - dehydrogenase (NAD+) | EG_transcript_21676, EG_transcript_19272, EG_transcript_15045 |
| Riboflavin metabolism | map00740 | 6 | ec:1.1.1.193 - reductase, ec:3.1.3.2 - phosphatase, ec:4.1.99.12 - synthase, ec:2.7.1.26 - kinase, ec:3.5.4.26 - deaminase, ec:3.5.4.25 - cyclohydrolase II | EG_transcript_6629, EG_transcript_8839, EG_transcript_7616, EG_transcript_680, EG_transcript_3866, EG_transcript_30923, EG_transcript_6629, EG_transcript_3866 |

| | | | | |
|---|---|---|---|---|
| Biosynthesis of vancomycin group antibiotics | map01055 | 1 | ec:4.2.1.46 - 4,6-dehydratase | EG_transcript_21799 |
| Primary bile acid biosynthesis | map00120 | 1 | ec:1.1.1.35 - dehydrogenase | EG_transcript_7580, EG_transcript_29059, EG_transcript_26319, EG_transcript_862, EG_transcript_21457, EG_transcript_3075 |
| Glycerophospholipid metabolism | map00564 | 5 | ec:3.1.4.46 - phosphodiesterase, ec:2.7.1.107 - kinase (ATP), ec:1.1.1.8 - dehydrogenase (NAD+), ec:1.1.1.94 - dehydrogenase [NAD(P)+], ec:1.1.5.3 - dehydrogenase | EG_transcript_13559, EG_transcript_14838, EG_transcript_8021, EG_transcript_18235, EG_transcript_8021, EG_transcript_18235, EG_transcript_9333, EG_transcript_13673 |
| Cutin, suberine and wax biosynthesis | map00073 | 1 | ec:2.3.1.20 - O-acyltransferase | EG_transcript_27587, EG_transcript_6847, EG_transcript_5274 |
| Thiamine metabolism | map00730 | 5 | ec:2.5.1.3 - phosphate synthase, ec:2.7.4.3 - kinase, ec:2.8.1.7 - desulfurase, ec:2.2.1.7 - synthase, ec:3.6.1.15 - phosphatase | EG_transcript_22875, EG_transcript_22143, EG_transcript_11568, EG_transcript_6058, EG_transcript_1523, EG_transcript_4118, EG_transcript_10317, EG_transcript_1888, EG_transcript_567, EG_transcript_321, EG_transcript_7192, EG_transcript_13141, EG_transcript_8030, EG_transcript_3260, EG_transcript_3381, EG_transcript_8154, EG_transcript_11643, EG_transcript_3148, EG_transcript_3018, EG_transcript_1079, EG_transcript_6648, EG_transcript_1877, EG_transcript_2847, EG_transcript_74, EG_transcript_10562, EG_transcript_795, EG_transcript_12190, EG_transcript_6088, EG_transcript_13158, EG_transcript_6889, EG_transcript_1993, EG_transcript_10538, EG_transcript_30386, EG_transcript_664, EG_transcript_36488, EG_transcript_19234, EG_transcript_19476, EG_transcript_3280, EG_transcript_21665, EG_transcript_20574, EG_transcript_23963, EG_transcript_1663, EG_transcript_5227, EG_transcript_1535, EG_transcript_1656, EG_transcript_1536, EG_transcript_18273, EG_transcript_91, EG_transcript_22409, EG_transcript_19235, EG_transcript_8289, EG_transcript_8722, EG_transcript_3154, EG_transcript_7996, EG_transcript_4245, EG_transcript_16527, EG_transcript_5456, EG_transcript_35708, EG_transcript_526, EG_transcript_2819, EG_transcript_7027, EG_transcript_2494, EG_transcript_3584, EG_transcript_2134, EG_transcript_4312, EG_transcript_8912, EG_transcript_21240, EG_transcript_3104, EG_transcript_3337, EG_transcript_3458, EG_transcript_10923, EG_transcript_515, EG_transcript_1831, EG_transcript_1953, EG_transcript_2925, EG_transcript_1959, EG_transcript_10241, EG_transcript_15140, EG_transcript_22226, EG_transcript_22347, EG_transcript_7497, EG_transcript_23555, EG_transcript_505, EG_transcript_744, EG_transcript_10211, EG_transcript_191, EG_transcript_14010, |

| | | | | |
|---|---|---|---|---|
| | | | | EG_transcript_15106, EG_transcript_2274, EG_transcript_6513, EG_transcript_4327, EG_transcript_1056, EG_transcript_5, EG_transcript_854, EG_transcript_2947, EG_transcript_297, EG_transcript_25828, EG_transcript_8006, EG_transcript_2387, EG_transcript_1126, EG_transcript_28802, EG_transcript_6491, EG_transcript_5164, EG_transcript_8557, EG_transcript_1120, EG_transcript_1241, EG_transcript_1121, EG_transcript_1245, EG_transcript_5844, EG_transcript_4745, EG_transcript_159, EG_transcript_950, EG_transcript_155, EG_transcript_150, EG_transcript_5390, EG_transcript_19532, EG_transcript_20888, EG_transcript_8662, EG_transcript_1470, EG_transcript_1350, EG_transcript_1352, EG_transcript_5039, EG_transcript_944, EG_transcript_4416, EG_transcript_4537, EG_transcript_5627, EG_transcript_260, EG_transcript_23566, EG_transcript_12798, EG_transcript_1144, EG_transcript_2113, EG_transcript_14, EG_transcript_1139, EG_transcript_5617, EG_transcript_32956, EG_transcript_253, EG_transcript_254, EG_transcript_255, EG_transcript_10021, EG_transcript_251, EG_transcript_1372, EG_transcript_7234, EG_transcript_1495, EG_transcript_3310, EG_transcript_7356, EG_transcript_1016, EG_transcript_129, EG_transcript_3745, EG_transcript_4714, EG_transcript_367, EG_transcript_25212, EG_transcript_10193, EG_transcript_19330, EG_transcript_6572, EG_transcript_5363, EG_transcript_11167, EG_transcript_4275, EG_transcript_29, EG_transcript_9603, EG_transcript_910, EG_transcript_34, EG_transcript_2649, EG_transcript_353, EG_transcript_3170, EG_transcript_13352, EG_transcript_18009, EG_transcript_16984, EG_transcript_3176, EG_transcript_2884, EG_transcript_46, EG_transcript_31678, EG_transcript_901, EG_transcript_25311, EG_transcript_1901, EG_transcript_12480, EG_transcript_5020, EG_transcript_8893, EG_transcript_5143, EG_transcript_5387, EG_transcript_12357, EG_transcript_6599, EG_transcript_2550, EG_transcript_1100, EG_transcript_1221, EG_transcript_5941, EG_transcript_3521, EG_transcript_1102, EG_transcript_4853, EG_transcript_3639, EG_transcript_52, EG_transcript_579, EG_transcript_331, EG_transcript_573, EG_transcript_15990, EG_transcript_3196, EG_transcript_6461, EG_transcript_49747, EG_transcript_23012, EG_transcript_2423, EG_transcript_10506 |

| | | | | |
|---|---|---|---|---|
| Carbon fixation pathways in prokaryotes | map00720 | 15 | ec:6.4.1.2 - carboxylase, ec:6.3.4.3 - ligase, ec:1.5.1.20 - reductase [NAD(P)H], ec:1.3.5.1 - dehydrogenase, ec:4.1.1.31 - carboxylase, ec:4.2.1.3 - hydratase, ec:4.2.1.2 - hydratase, ec:3.5.4.9 - cyclohydrolase, ec:2.7.9.1 - phosphate dikinase, ec:6.2.1.1 - ligase, ec:1.1.1.37 - dehydrogenase, ec:1.1.1.35 - dehydrogenase, ec:1.1.1.42 - dehydrogenase (NADP+), ec:1.5.1.5 - dehydrogenase (NADP+), ec:5.4.99.2 - mutase | EG_transcript_65, EG_transcript_97, EG_transcript_6633, EG_transcript_17312, EG_transcript_15809, EG_transcript_12465, EG_transcript_6467, EG_transcript_1597, EG_transcript_2009, EG_transcript_4764, EG_transcript_17805, EG_transcript_1653, EG_transcript_4603, EG_transcript_3254, EG_transcript_9669, EG_transcript_12212, EG_transcript_15919, EG_transcript_24127, EG_transcript_7580, EG_transcript_29059, EG_transcript_26319, EG_transcript_862, EG_transcript_21457, EG_transcript_3075, EG_transcript_9064, EG_transcript_10115, EG_transcript_4351, EG_transcript_6875, EG_transcript_17805, EG_transcript_17733, EG_transcript_4204 |
| Geraniol degradation | map00281 | 1 | ec:1.1.1.35 - dehydrogenase | EG_transcript_7580, EG_transcript_29059, EG_transcript_26319, EG_transcript_862, EG_transcript_21457, EG_transcript_3075 |
| Biosynthesis of ansamycins | map01051 | 1 | ec:2.2.1.1 - glycolaldehydetransferase | EG_transcript_2997, EG_transcript_4605 |
| Carbon fixation in photosynthetic organisms | map00710 | 14 | ec:5.3.1.1 - isomerase, ec:4.1.2.13 - aldolase, ec:5.3.1.6 - isomerase, ec:4.1.1.39 - carboxylase, ec:4.1.1.31 - carboxylase, ec:2.2.1.1 - glycolaldehydetransferase, ec:2.7.2.3 - kinase, ec:2.7.9.1 - phosphate dikinase, ec:1.2.1.59 - dehydrogenase (NAD(P)+) (phosphorylating), ec:2.7.1.19 - phosphopentokinase, ec:1.1.1.37 - dehydrogenase, ec:1.1.1.39 - dehydrogenase (decarboxylating), ec:3.1.3.11 - hexose diphosphatase, ec:1.2.1.12 - dehydrogenase (phosphorylating) | EG_transcript_14120, EG_transcript_30563, EG_transcript_18179, EG_transcript_30456, EG_transcript_44253, EG_transcript_7827, EG_transcript_21524, EG_transcript_43521, EG_transcript_16008, EG_transcript_25832, EG_transcript_5279, EG_transcript_15855, EG_transcript_15619, EG_transcript_20009, EG_transcript_7123, EG_transcript_10675, EG_transcript_1597, EG_transcript_2997, EG_transcript_4605, EG_transcript_12119, EG_transcript_10778, EG_transcript_12260, EG_transcript_1653, EG_transcript_16406, EG_transcript_9015, EG_transcript_12212, EG_transcript_15919, EG_transcript_24127, EG_transcript_27221, EG_transcript_5257, EG_transcript_8836, EG_transcript_16327, EG_transcript_14050, EG_transcript_4839, EG_transcript_7394, EG_transcript_2646, EG_transcript_8727, EG_transcript_7683, EG_transcript_4570, EG_transcript_15632, EG_transcript_18342, EG_transcript_13690, EG_transcript_16406 |
| Ascorbate and aldarate metabolism | map00053 | 5 | ec:2.7.7.64 - uridylyltransferase, ec:1.2.1.3 - dehydrogenase (NAD+), ec:1.11.1.11 - peroxidase, ec:1.3.2.3 - dehydrogenase, ec:1.1.1.22 - 6-dehydrogenase | EG_transcript_6731, EG_transcript_12921, EG_transcript_15045, EG_transcript_22491, EG_transcript_5780, EG_transcript_10254, EG_transcript_14324 |
| Streptomycin biosynthesis | map00521 | 4 | ec:4.2.1.46 - 4,6-dehydratase, ec:2.7.1.2 - glucokinase (phosphorylating), ec:2.7.1.1 - hexokinase type IV glucokinase, ec:5.5.1.4 - synthase | EG_transcript_21799, EG_transcript_8623, EG_transcript_8623, EG_transcript_6431 |
| Carbapenem | map00332 | 2 | ec:2.7.2.11 - 5-kinase, ec:1.2.1.41 - dehydrogenase | EG_transcript_12607, EG_transcript_6612 |

| | | | | |
|---|---|---|---|---|
| biosynthesis | | | | |
| Pentose phosphate pathway | map00030 | 13 | ec:2.7.1.90 - 1-phosphotransferase, ec:4.1.2.13 - aldolase, ec:5.3.1.6 - isomerase, ec:5.3.1.9 - isomerase, ec:2.2.1.2 - dihydroxyacetonetransferase, ec:2.2.1.1 - glycolaldehydetransferase, ec:2.7.1.11 - phosphohexokinase, ec:3.1.3.11 - hexose diphosphatase, ec:4.1.2.4 - aldolase, ec:1.1.1.44 - dehydrogenase (NADP+-dependent, decarboxylating), ec:1.1.1.49 - dehydrogenase (NADP+), ec:2.7.6.1 - diphosphokinase, ec:3.1.1.31 - phosphogluconolactonase | EG_transcript_1151, EG_transcript_7827, EG_transcript_21524, EG_transcript_43521, EG_transcript_16008, EG_transcript_25832, EG_transcript_5279, EG_transcript_15855, EG_transcript_15619, EG_transcript_20009, EG_transcript_5108, EG_transcript_13075, EG_transcript_2997, EG_transcript_4605, EG_transcript_1151, EG_transcript_7325, EG_transcript_6970, EG_transcript_6957, EG_transcript_7394, EG_transcript_2646, EG_transcript_8727, EG_transcript_7683, EG_transcript_4570, EG_transcript_15632, EG_transcript_18342, EG_transcript_13690, EG_transcript_20144, EG_transcript_13898, EG_transcript_7631, EG_transcript_16470, EG_transcript_15075, EG_transcript_2747, EG_transcript_5995, EG_transcript_9049, EG_transcript_7528, EG_transcript_14984, EG_transcript_13095, EG_transcript_2747, EG_transcript_19914 |
| Other glycan degradation | map00511 | 1 | ec:3.2.1.25 - mannanase | EG_transcript_4493 |
| Insect hormone biosynthesis | map00981 | 1 | ec:1.2.1.3 - dehydrogenase (NAD+) | EG_transcript_15045 |
| Limonene and pinene degradation | map00903 | 1 | ec:1.2.1.3 - dehydrogenase (NAD+) | EG_transcript_15045 |
| Citrate cycle (TCA cycle) | map00020 | 10 | ec:1.8.1.4 - dehydrogenase, ec:1.3.5.1 - dehydrogenase, ec:4.1.1.32 - carboxykinase (GTP), ec:4.2.1.3 - hydratase, ec:4.2.1.2 - hydratase, ec:1.1.1.37 - dehydrogenase, ec:2.3.1.12 - acetyltransferase, ec:1.1.1.42 - dehydrogenase (NADP+), ec:2.3.3.1 - (Si)-synthase, ec:1.2.4.1 - dehydrogenase (acetyl-transferring) | EG_transcript_8679, EG_transcript_6467, EG_transcript_61784, EG_transcript_2009, EG_transcript_4764, EG_transcript_12212, EG_transcript_15919, EG_transcript_24127, EG_transcript_5584, EG_transcript_9064, EG_transcript_10115, EG_transcript_4351, EG_transcript_6875, EG_transcript_2898, EG_transcript_6849, EG_transcript_10302, EG_transcript_11595, EG_transcript_14931, EG_transcript_9827 |
| Propanoate metabolism | map00640 | 7 | ec:1.8.1.4 - dehydrogenase, ec:6.4.1.2 - carboxylase, ec:1.2.1.27 - dehydrogenase (CoA-acylating), ec:4.1.1.9 - decarboxylase, ec:6.2.1.1 - ligase, ec:1.2.1.18 - dehydrogenase (acetylating), ec:5.4.99.2 - mutase | EG_transcript_8679, EG_transcript_65, EG_transcript_97, EG_transcript_6633, EG_transcript_15045, EG_transcript_29172, EG_transcript_9247, EG_transcript_4603, EG_transcript_3254, EG_transcript_9669, EG_transcript_15045, EG_transcript_4204 |
| Acarbose and validamycin biosynthesis | map00525 | 1 | ec:4.2.1.46 - 4,6-dehydratase | EG_transcript_21799 |

**Note:** The table describes the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Enzyme Code (EC) for all significant proteomic sequences.

### 6.2.4 Evolutionary pressures and selections - dN/dS ratios

Evolutionary selections or selective pressure analysis (dN/dS ratios) were performed for genes with known pathways/functions (Figure 6.6, 6.7, and Table 6.3, 6.4, 6.5). Transcripts (627) which correspond to these genes in the light and dark experiment were investigated for synonymous and non synonymous substitutions (dN dS). Overall, 17 codons were recorded with 196,251 comparisons (Figure 6.8, Table 6.6, 6.7, 6.8). Indels were not recorded except on the 17th codon, while the number of of stop codons per sequence is 0 (Figure 6.8, Table 6.6). The average dN/dS ratios recorded across all pairwise comparisons and first sequence compared to others are 0.87 and 0.99 respectively (dN/dS < 1). The cumulative behavior increased linearly for both the synonymous and non synonymous substitutions until the 17th codon where indel is present.

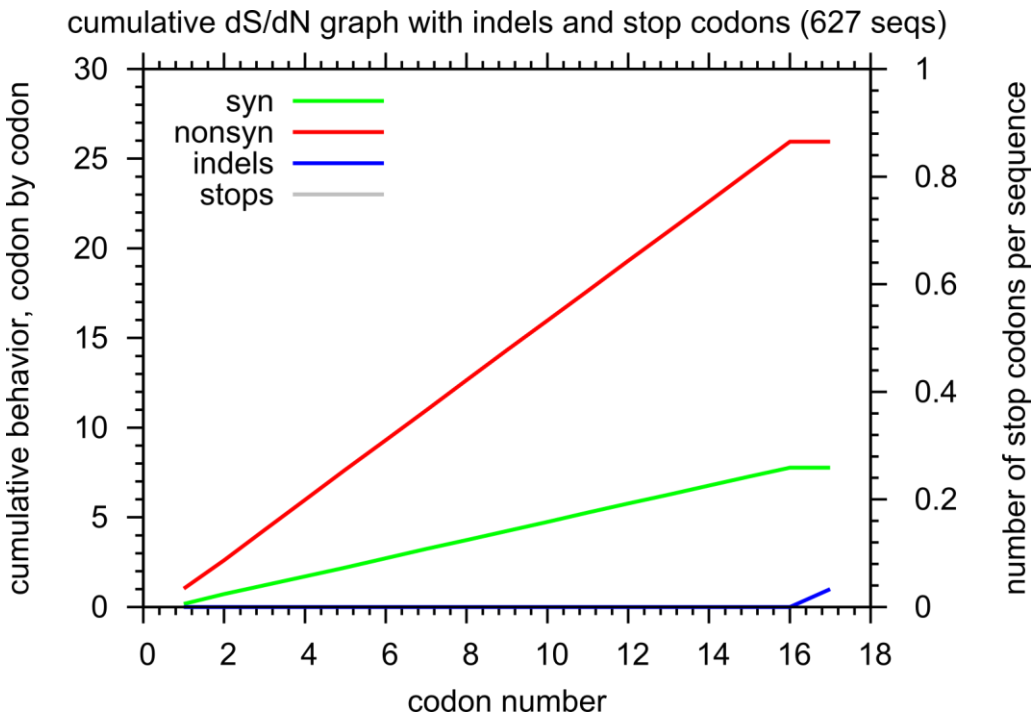



**Figure 6.8: Protein coding genes in *E. gracilis* are not under selective pressure.** The XY plot shows the cumulative dS/dN average for 196,251 comparisons in functional (biological pathway-directed) transcripts (627 sequences) predicted from the light and dark experiment. **X-axis** = codon number. **Y-axis 1** = cumulative behavior, codon by codon. **Y-axis 2** = number of stop codons per sequence. **Legends:** Green = synonymous substitution. Red = non synonymous substitution. Blue = indels. Grey = stod codons. dN = nonsynonymous. dS = synonymous. dS/dN = ratio of nonsynonymous to synonymous substitutions.

**Table 6.6:** Counts of codons, indel, synonymous, and non synonymous substitutions.

| | Cumulative | | | Per Codon | | | | |
|---|---|---|---|---|---|---|---|---|
| **Codon** | **indel** | **syn** | **nonsyn** | **indel** | **syn** | **nonsyn** | **stopc** | **aa** |
| 1 | 0 | 0.17 | 1.03 | 0 | 0.17 | 1.03 | 0 | M |
| 2 | 0 | 0.73 | 2.61 | 0 | 0.56 | 1.59 | 0 | E |
| 3 | 0 | 1.21 | 4.3 | 0 | 0.48 | 1.68 | 0 | D |
| 4 | 0 | 1.71 | 5.97 | 0 | 0.5 | 1.67 | 0 | E |
| 5 | 0 | 2.2 | 7.65 | 0 | 0.49 | 1.69 | 0 | K |
| 6 | 0 | 2.72 | 9.31 | 0 | 0.52 | 1.66 | 0 | G |
| 7 | 0 | 3.24 | 10.97 | 0 | 0.52 | 1.65 | 0 | Q |
| 8 | 0 | 3.73 | 12.65 | 0 | 0.49 | 1.68 | 0 | D |
| 9 | 0 | 4.24 | 14.32 | 0 | 0.51 | 1.67 | 0 | Q |
| 10 | 0 | 4.74 | 15.98 | 0 | 0.5 | 1.66 | 0 | K |
| 11 | 0 | 5.27 | 17.63 | 0 | 0.53 | 1.65 | 0 | L |
| 12 | 0 | 5.77 | 19.3 | 0 | 0.5 | 1.67 | 0 | L |
| 13 | 0 | 6.27 | 20.96 | 0 | 0.5 | 1.66 | 0 | A |
| 14 | 0 | 6.77 | 22.61 | 0 | 0.5 | 1.65 | 0 | V |
| 15 | 0 | 7.28 | 24.29 | 0 | 0.51 | 1.68 | 0 | V |
| 16 | 0 | 7.77 | 25.95 | 0 | 0.5 | 1.66 | 0 | H |
| 17 | 1 | 7.77 | 25.95 | 1 | 0 | 0 | 0 | - |

**Note:** For full description of codon, indel, syn, nonsyn, stopc, and aa, see Korber, 2000.

**Table 6.7:** Averages of all pairwise and first sequence comparisons

| | ds | dn | ds/dn | ps/pn |
|---|---|---|---|---|
| Averages of all pairwise comparisons: | 1.4674 | 1.9538 | 0.8708 | 0.8816 |
| Averages of the first sequence compared to others: | 1.6082 | 1.8416 | 0.9879 | 0.9506 |

**Note:** For full description of ds, dn, ds/dn, and ps/pn, see Korber, 2000.

**Table 6.8:** Top 100 comparisons from the 196,251 cumulative comparisons

| Compare | | Sequences_names | | Sd | Sn | S | N | ps | pn | ds | dn | ds/dn | ps/pn |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | EG_transcript_5 | EG_transcript_9 | 6.3333 | 26.6667 | 11.1667 | 36.8333 | 0.5672 | 0.724 | 1.0586 | 2.521 | 0.4199 | 0.7834 |
| 0 | 2 | EG_transcript_5 | EG_transcript_14 | 9.5 | 28.5 | 12.3333 | 35.6667 | 0.7703 | 0.7991 | nan | nan | nan | 0.964 |
| 0 | 3 | EG_transcript_5 | EG_transcript_28 | 12.167 | 28.8333 | 10.3333 | 37.6667 | 1.1774 | 0.7655 | nan | nan | nan | 1.5381 |
| 0 | 4 | EG_transcript_5 | EG_transcript_29 | 9.1667 | 18.8333 | 11.3333 | 36.6667 | 0.8088 | 0.5136 | nan | 0.866 | nan | 1.5747 |
| 0 | 5 | EG_transcript_5 | EG_transcript_34 | 6.5 | 24.5 | 10.3333 | 37.6667 | 0.629 | 0.6504 | 1.3684 | 1.5145 | 0.9035 | 0.9671 |
| 0 | 6 | EG_transcript_5 | EG_transcript_46 | 10.333 | 25.6667 | 11.1667 | 36.8333 | 0.9254 | 0.6968 | nan | 1.985 | nan | 1.328 |
| 0 | 7 | EG_transcript_5 | EG_transcript_52 | 6.1667 | 24.8333 | 9.5 | 38.5 | 0.6491 | 0.645 | 1.5046 | 1.4747 | 1.0203 | 1.0064 |
| 0 | 8 | EG_transcript_5 | EG_transcript_65 | 10.667 | 23.3333 | 13 | 35 | 0.8205 | 0.6667 | nan | 1.6479 | nan | 1.2308 |
| 0 | 9 | EG_transcript_5 | EG_transcript_74 | 6.5 | 26.5 | 10.6667 | 37.3333 | 0.6094 | 0.7098 | 1.2555 | 2.1951 | 0.572 | 0.8585 |
| 0 | 10 | EG_transcript_5 | EG_transcript_91 | 4.1667 | 21.8333 | 12 | 36 | 0.3472 | 0.6065 | 0.4663 | 1.2402 | 0.376 | 0.5725 |
| 0 | 11 | EG_transcript_5 | EG_transcript_92 | 6.6667 | 26.3333 | 10.6667 | 37.3333 | 0.625 | 0.7054 | 1.3438 | 2.116 | 0.6351 | 0.8861 |
| 0 | 12 | EG_transcript_5 | EG_transcript_97 | 8.6667 | 25.3333 | 11.8333 | 36.1667 | 0.7324 | 0.7005 | 2.8139 | 2.038 | 1.3807 | 1.0456 |
| 0 | 13 | EG_transcript_5 | EG_transcript_107 | 6.8333 | 26.1667 | 11.1667 | 36.8333 | 0.6119 | 0.7104 | 1.2693 | 2.2061 | 0.5754 | 0.8614 |
| 0 | 14 | EG_transcript_5 | EG_transcript_121 | 6.5 | 20.5 | 12.8333 | 35.1667 | 0.5065 | 0.5829 | 0.8437 | 1.1263 | 0.7491 | 0.8689 |
| 0 | 15 | EG_transcript_5 | EG_transcript_129 | 8.3333 | 26.6667 | 11.3333 | 36.6667 | 0.7353 | 0.7273 | 2.9489 | 2.6224 | 1.1245 | 1.011 |
| 0 | 16 | EG_transcript_5 | EG_transcript_150 | 6.5 | 23.5 | 10.6667 | 37.3333 | 0.6094 | 0.6295 | 1.2555 | 1.3711 | 0.9157 | 0.9681 |
| 0 | 17 | EG_transcript_5 | EG_transcript_155 | 6.8333 | 26.1667 | 12.5 | 35.5 | 0.5467 | 0.7371 | 0.9789 | 3.0465 | 0.3213 | 0.7417 |
| 0 | 18 | EG_transcript_5 | EG_transcript_159 | 7.3333 | 25.6667 | 12.6667 | 35.3333 | 0.5789 | 0.7264 | 1.1086 | 2.5946 | 0.4273 | 0.797 |
| 0 | 19 | EG_transcript_5 | EG_transcript_191 | 5.3333 | 27.6667 | 11.5 | 36.5 | 0.4638 | 0.758 | 0.7225 | nan | nan | 0.6118 |
| 0 | 20 | EG_transcript_5 | EG_transcript_202 | 9.3333 | 24.6667 | 12.1667 | 35.8333 | 0.7671 | 0.6884 | nan | 1.8742 | nan | 1.1144 |
| 0 | 21 | EG_transcript_5 | EG_transcript_224 | 6.3333 | 27.6667 | 10.1667 | 37.8333 | 0.623 | 0.7313 | 1.3316 | 2.7678 | 0.4811 | 0.8519 |
| 0 | 22 | EG_transcript_5 | EG_transcript_251 | 6.6667 | 24.3333 | 13.1667 | 34.8333 | 0.5063 | 0.6986 | 0.8432 | 2.0098 | 0.4195 | 0.7248 |
| 0 | 23 | EG_transcript_5 | EG_transcript_253 | 7 | 30 | 9 | 39 | 0.7778 | 0.7692 | nan | nan | nan | 1.0111 |
| 0 | 24 | EG_transcript_5 | EG_transcript_254 | 9 | 18 | 11.5 | 36.5 | 0.7826 | 0.4932 | nan | 0.8037 | nan | 1.587 |
| 0 | 25 | EG_transcript_5 | EG_transcript_255 | 6.5 | 26.5 | 9.5 | 38.5 | 0.6842 | 0.6883 | 1.8252 | 1.8735 | 0.9742 | 0.994 |
| 0 | 26 | EG_transcript_5 | EG_transcript_260 | 8.8333 | 24.1667 | 12.3333 | 35.6667 | 0.7162 | 0.6776 | 2.3251 | 1.7531 | 1.3263 | 1.057 |
| 0 | 27 | EG_transcript_5 | EG_transcript_297 | 9.8333 | 20.1667 | 12 | 36 | 0.8194 | 0.5602 | nan | 1.0305 | nan | 1.4628 |
| 0 | 28 | EG_transcript_5 | EG_transcript_321 | 8 | 25 | 10.5 | 37.5 | 0.7619 | 0.6667 | nan | 1.6479 | nan | 1.1429 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 29 | EG_transcript_5 | EG_transcript_331 | 8.5 | 22.5 | 11.6667 | 36.3333 | 0.7286 | 0.6193 | 2.6665 | 1.3102 | 2.0352 | 1.1765 |
| 0 | 30 | EG_transcript_5 | EG_transcript_353 | 7.5 | 29.5 | 12 | 36 | 0.625 | 0.8194 | 1.3438 | nan | nan | 0.7627 |
| 0 | 31 | EG_transcript_5 | EG_transcript_356 | 9 | 23 | 11 | 37 | 0.8182 | 0.6216 | nan | 1.3238 | nan | 1.3162 |
| 0 | 32 | EG_transcript_5 | EG_transcript_367 | 5.5 | 27.5 | 10 | 38 | 0.55 | 0.7237 | 0.9913 | 2.5124 | 0.3946 | 0.76 |
| 0 | 33 | EG_transcript_5 | EG_transcript_464 | 8.5 | 30.5 | 9.1667 | 38.8333 | 0.9273 | 0.7854 | nan | nan | nan | 1.1806 |
| 0 | 34 | EG_transcript_5 | EG_transcript_478 | 8 | 30 | 11.6667 | 36.3333 | 0.6857 | 0.8257 | 1.8426 | nan | nan | 0.8305 |
| 0 | 35 | EG_transcript_5 | EG_transcript_505 | 7 | 30 | 11.5 | 36.5 | 0.6087 | 0.8219 | 1.2519 | nan | nan | 0.7406 |
| 0 | 36 | EG_transcript_5 | EG_transcript_515 | 9.1667 | 24.8333 | 12.5 | 35.5 | 0.7333 | 0.6995 | 2.855 | 2.024 | 1.4106 | 1.0483 |
| 0 | 37 | EG_transcript_5 | EG_transcript_516 | 8.8333 | 24.1667 | 11.5 | 36.5 | 0.7681 | 0.6621 | nan | 1.6079 | nan | 1.1601 |
| 0 | 38 | EG_transcript_5 | EG_transcript_526 | 8.8333 | 23.1667 | 11.3333 | 36.6667 | 0.7794 | 0.6318 | nan | 1.3859 | nan | 1.2336 |
| 0 | 39 | EG_transcript_5 | EG_transcript_567 | 9 | 27 | 13 | 35 | 0.6923 | 0.7714 | 1.9237 | nan | nan | 0.8974 |
| 0 | 40 | EG_transcript_5 | EG_transcript_573 | 7.1667 | 30.8333 | 10.8333 | 37.1667 | 0.6615 | 0.8296 | 1.6031 | nan | nan | 0.7974 |
| 0 | 41 | EG_transcript_5 | EG_transcript_579 | 9.3333 | 20.6667 | 11.3333 | 36.6667 | 0.8235 | 0.5636 | nan | 1.0443 | nan | 1.4611 |
| 0 | 42 | EG_transcript_5 | EG_transcript_583 | 7.1667 | 27.8333 | 12.8333 | 35.1667 | 0.5584 | 0.7915 | 1.0237 | nan | nan | 0.7056 |
| 0 | 43 | EG_transcript_5 | EG_transcript_664 | 8.5 | 29.5 | 11 | 37 | 0.7727 | 0.7973 | nan | nan | nan | 0.9692 |
| 0 | 44 | EG_transcript_5 | EG_transcript_680 | 8.3333 | 20.6667 | 13.1667 | 34.8333 | 0.6329 | 0.5933 | 1.3929 | 1.1743 | 1.1861 | 1.0668 |
| 0 | 45 | EG_transcript_5 | EG_transcript_729 | 7.3333 | 22.6667 | 11 | 37 | 0.6667 | 0.6126 | 1.6479 | 1.273 | 1.2946 | 1.0882 |
| 0 | 46 | EG_transcript_5 | EG_transcript_744 | 9.1667 | 25.8333 | 12.6667 | 35.3333 | 0.7237 | 0.7311 | 2.5124 | 2.762 | 0.9097 | 0.9898 |
| 0 | 47 | EG_transcript_5 | EG_transcript_790 | 9.6667 | 23.3333 | 11.5 | 36.5 | 0.8406 | 0.6393 | nan | 1.4347 | nan | 1.3149 |
| 0 | 48 | EG_transcript_5 | EG_transcript_791 | 8.8333 | 27.1667 | 11.3333 | 36.6667 | 0.7794 | 0.7409 | nan | 3.3096 | nan | 1.052 |
| 0 | 49 | EG_transcript_5 | EG_transcript_795 | 8.8333 | 27.1667 | 10.8333 | 37.1667 | 0.8154 | 0.7309 | nan | 2.7544 | nan | 1.1155 |
| 0 | 50 | EG_transcript_5 | EG_transcript_854 | 7.6667 | 26.3333 | 10.6667 | 37.3333 | 0.7188 | 0.7054 | 2.3835 | 2.116 | 1.1264 | 1.019 |
| 0 | 51 | EG_transcript_5 | EG_transcript_862 | 10 | 29 | 10.8333 | 37.1667 | 0.9231 | 0.7803 | nan | nan | nan | 1.183 |
| 0 | 52 | EG_transcript_5 | EG_transcript_901 | 10.167 | 27.8333 | 11 | 37 | 0.9242 | 0.7523 | nan | nan | nan | 1.2286 |
| 0 | 53 | EG_transcript_5 | EG_transcript_910 | 11 | 26 | 12.1667 | 35.8333 | 0.9041 | 0.7256 | nan | 2.5685 | nan | 1.246 |
| 0 | 54 | EG_transcript_5 | EG_transcript_929 | 8.5 | 23.5 | 11 | 37 | 0.7727 | 0.6351 | nan | 1.4072 | nan | 1.2166 |
| 0 | 55 | EG_transcript_5 | EG_transcript_944 | 7.3333 | 24.6667 | 11.1667 | 36.8333 | 0.6567 | 0.6697 | 1.5633 | 1.6756 | 0.933 | 0.9806 |
| 0 | 56 | EG_transcript_5 | EG_transcript_950 | 6.1667 | 31.8333 | 10.3333 | 37.6667 | 0.5968 | 0.8451 | 1.1911 | nan | nan | 0.7061 |
| 0 | 57 | EG_transcript_5 | EG_transcript_1016 | 9.8333 | 27.1667 | 11.3333 | 36.6667 | 0.8676 | 0.7409 | nan | 3.3096 | nan | 1.1711 |
| 0 | 58 | EG_transcript_5 | EG_transcript_1056 | 7.1667 | 29.8333 | 11 | 37 | 0.6515 | 0.8063 | 1.5226 | nan | nan | 0.808 |
| 0 | 59 | EG_transcript_5 | EG_transcript_1073 | 6 | 25 | 11.5 | 36.5 | 0.5217 | 0.6849 | 0.8922 | 1.8335 | 0.4866 | 0.7617 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 60 | EG_transcript_5 | EG_transcript_1079 | 8.1667 | 24.8333 | 11.6667 | 36.3333 | 0.7 | 0.6835 | 2.031 | 1.817 | 1.1178 | 1.0242 |
| 0 | 61 | EG_transcript_5 | EG_transcript_1100 | 9 | 22 | 11.5 | 36.5 | 0.7826 | 0.6027 | nan | 1.2209 | nan | 1.2984 |
| 0 | 62 | EG_transcript_5 | EG_transcript_1102 | 5.5 | 27.5 | 10.5 | 37.5 | 0.5238 | 0.7333 | 0.899 | 2.855 | 0.3149 | 0.7143 |
| 0 | 63 | EG_transcript_5 | EG_transcript_1121 | 6.5 | 30.5 | 9.6667 | 38.3333 | 0.6724 | 0.7957 | 1.7015 | nan | nan | 0.8451 |
| 0 | 64 | EG_transcript_5 | EG_transcript_1120 | 4.1667 | 25.8333 | 12.3333 | 35.6667 | 0.3378 | 0.7243 | 0.449 | 2.5302 | 0.1775 | 0.4664 |
| 0 | 65 | EG_transcript_5 | EG_transcript_1126 | 8.5 | 26.5 | 10.8333 | 37.1667 | 0.7846 | 0.713 | nan | 2.257 | nan | 1.1004 |
| 0 | 66 | EG_transcript_5 | EG_transcript_1139 | 6.8333 | 26.1667 | 10.1667 | 37.8333 | 0.6721 | 0.6916 | 1.6988 | 1.915 | 0.8871 | 0.9718 |
| 0 | 67 | EG_transcript_5 | EG_transcript_1144 | 7.3333 | 25.6667 | 11.3333 | 36.6667 | 0.6471 | 0.7 | 1.4894 | 2.031 | 0.7333 | 0.9244 |
| 0 | 68 | EG_transcript_5 | EG_transcript_1145 | 10 | 25 | 12 | 36 | 0.8333 | 0.6944 | nan | 1.952 | nan | 1.2 |
| 0 | 69 | EG_transcript_5 | EG_transcript_1151 | 9.5 | 26.5 | 12.1667 | 35.8333 | 0.7808 | 0.7395 | nan | 3.204 | nan | 1.0558 |
| 0 | 70 | EG_transcript_5 | EG_transcript_1177 | 9.6667 | 26.3333 | 12.1667 | 35.8333 | 0.7945 | 0.7349 | nan | 2.9282 | nan | 1.0812 |
| 0 | 71 | EG_transcript_5 | EG_transcript_1208 | 9.1667 | 21.8333 | 10.8333 | 37.1667 | 0.8462 | 0.5874 | nan | 1.1468 | nan | 1.4404 |
| 0 | 72 | EG_transcript_5 | EG_transcript_1221 | 9.8333 | 27.1667 | 11.6667 | 36.3333 | 0.8429 | 0.7477 | nan | 4.3425 | nan | 1.1273 |
| 0 | 73 | EG_transcript_5 | EG_transcript_1236 | 7 | 21 | 11.1667 | 36.8333 | 0.6269 | 0.5701 | 1.3551 | 1.0709 | 1.2654 | 1.0995 |
| 0 | 74 | EG_transcript_5 | EG_transcript_1241 | 5 | 26 | 10.1667 | 37.8333 | 0.4918 | 0.6872 | 0.7998 | 1.8604 | 0.4299 | 0.7156 |
| 0 | 75 | EG_transcript_5 | EG_transcript_1245 | 4.6667 | 28.3333 | 10.1667 | 37.8333 | 0.459 | 0.7489 | 0.7101 | 4.8927 | 0.1451 | 0.6129 |
| 0 | 76 | EG_transcript_5 | EG_transcript_1330 | 8.5 | 27.5 | 11.1667 | 36.8333 | 0.7612 | 0.7466 | nan | 4.0486 | nan | 1.0195 |
| 0 | 77 | EG_transcript_5 | EG_transcript_1350 | 6.1667 | 26.8333 | 11.8333 | 36.1667 | 0.5211 | 0.7419 | 0.8902 | 3.3994 | 0.2619 | 0.7024 |
| 0 | 78 | EG_transcript_5 | EG_transcript_1352 | 10.833 | 21.1667 | 13.1667 | 34.8333 | 0.8228 | 0.6077 | nan | 1.2464 | nan | 1.354 |
| 0 | 79 | EG_transcript_5 | EG_transcript_1356 | 8.1667 | 27.8333 | 11.3333 | 36.6667 | 0.7206 | 0.7591 | 2.429 | nan | nan | 0.9493 |
| 0 | 80 | EG_transcript_5 | EG_transcript_1372 | 10.333 | 24.6667 | 13.1667 | 34.8333 | 0.7848 | 0.7081 | nan | 2.1642 | nan | 1.1083 |
| 0 | 81 | EG_transcript_5 | EG_transcript_1453 | 8 | 18 | 12 | 36 | 0.6667 | 0.5 | 1.6479 | 0.824 | 2 | 1.3333 |
| 0 | 82 | EG_transcript_5 | EG_transcript_1470 | 8.3333 | 24.6667 | 11.5 | 36.5 | 0.7246 | 0.6758 | 2.5401 | 1.735 | 1.4641 | 1.0723 |
| 0 | 83 | EG_transcript_5 | EG_transcript_1490 | 7.1667 | 23.8333 | 12.1667 | 35.8333 | 0.589 | 0.6651 | 1.1542 | 1.6341 | 0.7063 | 0.8856 |
| 0 | 84 | EG_transcript_5 | EG_transcript_1495 | 10 | 29 | 10.5 | 37.5 | 0.9524 | 0.7733 | nan | nan | nan | 1.2315 |
| 0 | 85 | EG_transcript_5 | EG_transcript_1523 | 7.6667 | 22.3333 | 11.8333 | 36.1667 | 0.6479 | 0.6175 | 1.4955 | 1.3002 | 1.1502 | 1.0492 |
| 0 | 86 | EG_transcript_5 | EG_transcript_1535 | 8.3333 | 25.6667 | 11.5 | 36.5 | 0.7246 | 0.7032 | 2.5401 | 2.0806 | 1.2209 | 1.0305 |
| 0 | 87 | EG_transcript_5 | EG_transcript_1536 | 6.5 | 23.5 | 10.8333 | 37.1667 | 0.6 | 0.6323 | 1.2071 | 1.3889 | 0.8691 | 0.9489 |
| 0 | 88 | EG_transcript_5 | EG_transcript_1548 | 11.5 | 23.5 | 13.1667 | 34.8333 | 0.8734 | 0.6746 | nan | 1.7234 | nan | 1.2946 |
| 0 | 89 | EG_transcript_5 | EG_transcript_1590 | 6.8333 | 30.1667 | 11 | 37 | 0.6212 | 0.8153 | 1.3214 | nan | nan | 0.7619 |
| 0 | 90 | EG_transcript_5 | EG_transcript_1597 | 7.6667 | 26.3333 | 11.6667 | 36.3333 | 0.6571 | 0.7248 | 1.5668 | 2.544 | 0.6159 | 0.9067 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 91 | EG_transcript_5 | EG_transcript_1613 | 8.6667 | 20.3333 | 11.1667 | 36.8333 | 0.7761 | 0.552 | nan | 0.999 | nan | 1.4059 |
| 0 | 92 | EG_transcript_5 | EG_transcript_1615 | 9.6667 | 25.3333 | 13 | 35 | 0.7436 | 0.7238 | 3.5716 | 2.516 | 1.4196 | 1.0273 |
| 0 | 93 | EG_transcript_5 | EG_transcript_1653 | 9.3333 | 24.6667 | 12 | 36 | 0.7778 | 0.6852 | nan | 1.8364 | nan | 1.1351 |
| 0 | 94 | EG_transcript_5 | EG_transcript_1656 | 9.8333 | 25.1667 | 12.5 | 35.5 | 0.7867 | 0.7089 | nan | 2.1784 | nan | 1.1097 |
| 0 | 95 | EG_transcript_5 | EG_transcript_1663 | 9 | 26 | 11.8333 | 36.1667 | 0.7606 | 0.7189 | nan | 2.387 | nan | 1.058 |
| 0 | 96 | EG_transcript_5 | EG_transcript_1784 | 7.5 | 26.5 | 11.3333 | 36.6667 | 0.6618 | 0.7227 | 1.605 | 2.4856 | 0.6457 | 0.9156 |
| 0 | 97 | EG_transcript_5 | EG_transcript_1831 | 7.8333 | 26.1667 | 11.8333 | 36.1667 | 0.662 | 0.7235 | 1.6068 | 2.5073 | 0.6409 | 0.915 |
| 0 | 98 | EG_transcript_5 | EG_transcript_1851 | 7.6667 | 20.3333 | 11.3333 | 36.6667 | 0.6765 | 0.5545 | 1.7418 | 1.0086 | 1.727 | 1.2199 |
| 0 | 99 | EG_transcript_5 | EG_transcript_1877 | 10.5 | 20.5 | 11.5 | 36.5 | 0.913 | 0.5616 | nan | 1.0363 | nan | 1.6257 |
| 0 | 100 | EG_transcript_5 | EG_transcript_1888 | 9.8333 | 23.1667 | 10.6667 | 37.3333 | 0.9219 | 0.6205 | nan | 1.3175 | nan | 1.4856 |

**Note:** The table shows the top 100 comparisons from the 1962511 cumulative comparisons. For additional comparions, see Appendix III, Table 2. For full description of Sd, Sn, S, N, ps, pn, ds, dn, ds/dn, and ps/pn, see Korber, 2000.

## 6.3 COMPARISON OF TRANSCRIPTOMIC AND PROTEOMIC DATA

Comparison of light and dark transcriptomic and proteomic experiments demonstrates a re-arrangement in the structural and functional organization of the photosynthetic and heterotrophic apparatus in *E. gracilis* (Figure 6.6, 6.7, 6.9). Comparing the 41045 transcript hits with the 4681 protein groups' quantified shows an overlap of 4287 (Figure 6.6, 6.9). Correlation analysis between the transcript and protein abundance shows an extremely poor correlation (Figure 6.9). The majority of the 380 infinite protein changes are not present in the transcritptome light and dark analysis. BLAST interrogation against the *E. gracilis* organellar genome (Hallick, *et al.,* 2013, Dobakova, *et al.,* 2015) showed that a considerable number of the hits with differential abundance and correlation are in the light regime and encoded in the chloroplast genome. These include photosystem I proteins such as P700 chlorophyll apoprotein A1, and the translation elongation facto EF-Tu which are significant. The reason why many of these correlated transcripts/protein groups are restricted to the light regime may be understandable. This is because during photosynthesis (light regime), a considerable amount of the photosynthetic apparatus is switched on. This is in contrast to the heterotrophic apparatus (such as mitochondria) that is switched on during heterotrophism (dark regime). It has been reported that the nuclear translation elongation factors are not influenced by switching cell growth conditions from dark to light (Montandon, *et al.,* 1989), which is consistent with our data that shows no differential expression of the nuclear EF-1α while both the chloroplastic EF-Tu protein and corresponding transcript [EG_transcript_1495] are highly upregulated in the light.

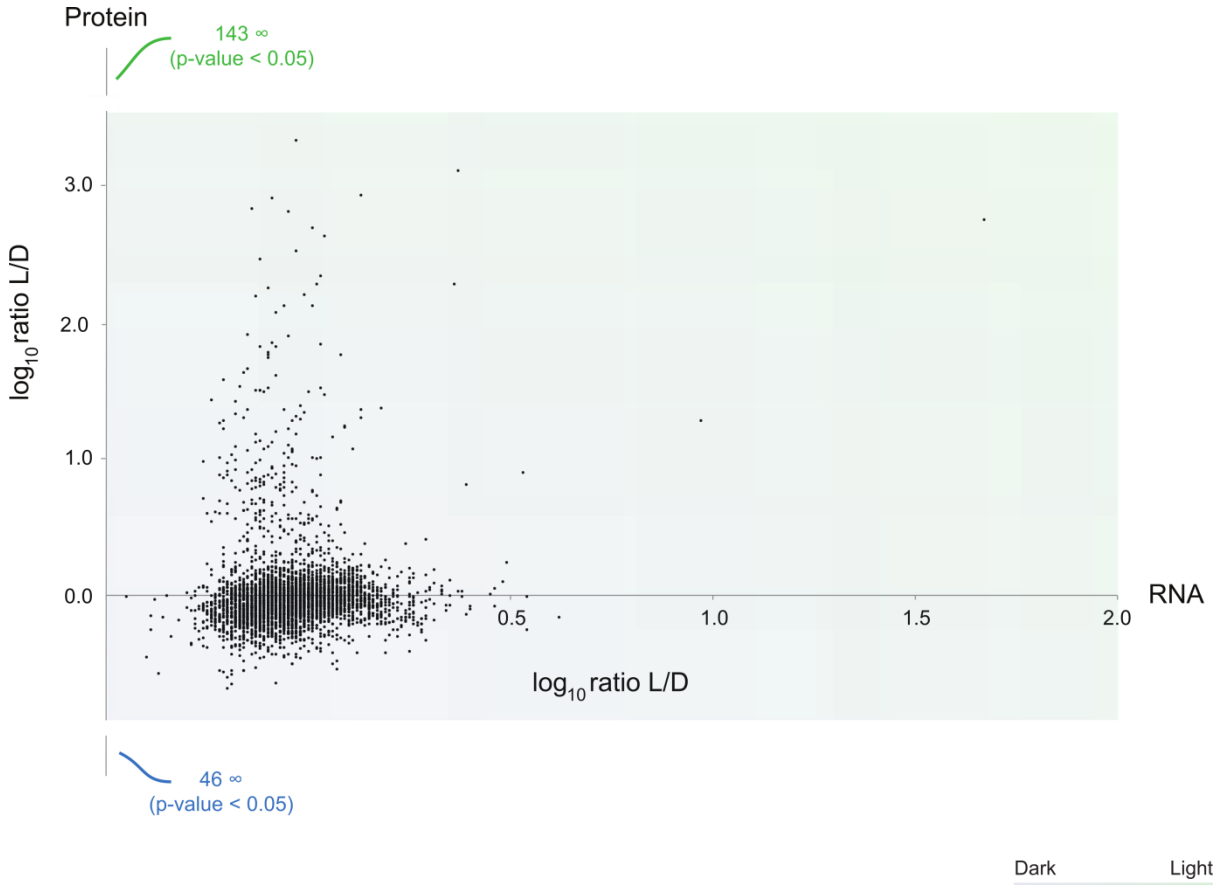



**Figure 6.9: Altered transcriptomic and proteomic analysis are not correlated.** The volcano plot shows a comparative analysis of transcriptome and proteome changes. Comparative volcano plot proteome and transcript distribution of light/dark adapted cells. Y-axis = log10 ration light/dark (protein) while X-axis = log10 ratio light/dark (RNA). The plot does not show correlation of the transcriptomic and proteomic experiment except for 4 transcripts/proteins. P-value is less than 0.05 for both protein and transcriptome with 143 ∞ and 46 ∞ values respectively. Thanks to Martin Zoltner for his kind help in generating this figure using Microsoft Excel and for Mark Field's help in reviewing it.

*E. gracilis* requires a light/dark regime for a productive photosynthesis. Light regime is needed for the photochemical production of adenosine triphosphate (ATP) which is an energy carrier within the cell and nicotinamide adenine dinucleotide phosphate (NADPH) which is a membrane-bound enzyme complex, while the dark regime is required for biochemical phase synthesis of essential molecules for growth (Al-Qasmi, *et al.,* 2012). It has been shown that increased levels of light intensity and light duration was associated with increased saturated fatty acids (SFA), decreased monounsaturated fatty acids (MUFA) and polyunsaturated fatty acids (PUFA) which are of utility in metabolic engineering (biotechnology). For instance, the efficiency of biomass production in *E. gracilis* can be improved by genetically modulating photosynthetic capacity (under photoautotrophic cultivation), resulting in the enhanced production of wax esters (Ogawa, *et al.,* 2015).

One unique finding in this experiment is that cell growth is not completely dependent on plastid metabolism. While there were marked cell growth differences between light and dark culture cells, growths were not completely reduced or inhibited in the dark regime. However, in some members of the euglenids, plastid metabolism is crucial for cell growth. For instance, in a study to evaluate the role of bleaching in *E. longa* evolution, it was demonstrated that bleaching *E. gracilis* using streptomycin and ofloxacin produces plastid gene deletions without affecting cell growth, while Streptomycin and ofloxacin inhibited *E. longa* growth indicating that it requires plastid genes to survive (Hadariová, *et al.,* 2016). This suggested that evolutionary divergence of *E. longa* from *E. gracilis* was triggered by the loss of a cytoplasmic metabolic activity also occurring in the plastid (Hadariová, *et al.,* 2016).

## 6.4 ELECTRON MICROSCOPY OF *E. gracilis* CELLS UNDER LIGHT AND DARK CONDITIONS

The result presented here shows the electron microscopy fine structure of *E. gracilis* samples adapted to light and dark environmental conditions - thanks to Dr. Sue Vaughan and Anna Burrel at the Oxford Brooks University, UK, for the Transmission Electron Microscopy (TEM) analysis. There were significant physiological and morphological ultrastructural differences between the light and dark adapted cells. Cells grown in the dark showed an increased paramylon (Pa) content in terms of numbers and size than those cultivated in the light – with cells in both conditions containing mitochondria (Mi) having narrow cristae (Figure 6.10). There were several amounts of variably shaped chloroplasts present in the light adapted samples with the elongated lamellar evident, while the presence of proplastids and the absence of thylakoids could be observed in the dark adapted samples. Overall, the chloroplast in the light adapted samples were larger than the proplastids in the dark adapted samples, and in most cases the later were invisible (data not shown). There were not significant differences in the nuclei, flagella, and Golgi apparatus of both dark and light adapted samples respectively. There is the presence of nucleoli surrounded by the nucleoplasm as well as the presence of paraflaegllar body and numerous narrow cristae in the flagellar and Golgi apparatus respectively (Figure 6.10).

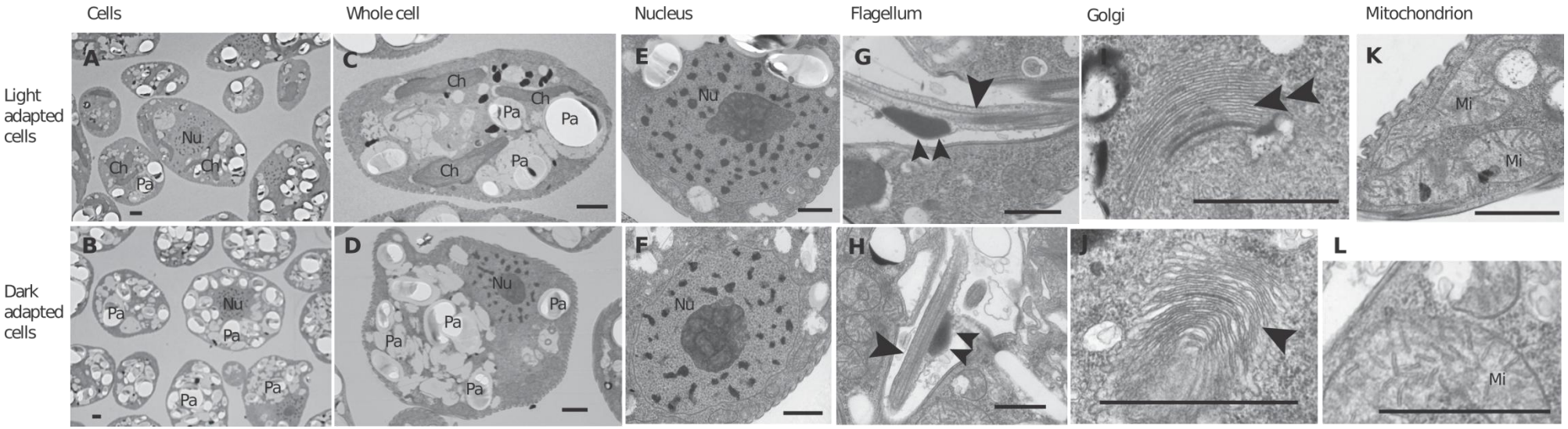



**Figure 6.10: Chloroplast and mitochondria are morphologically influenced by light or dark.** The diagram depicts the ultrastructure of dark and light adapted *E. gracilis.* Transmission electron micrographs comparing the ultrastructure of cells taken from dark adapted *E. gracilis* culture with that of cells taken from light adapted *E. gracilis* culture. Both light and dark adapted cells contained numerous paramylon granules (Pa) (A,B) but these were most extensive in the dark-adapted cells (B). Light adapted cells contained multiple mature chloroplasts (Ch) (C) which were absent from the dark-adapted cells (D). Both light and dark adapted cells contained nuclei with prominent nucleoli and many smaller electron dense foci in the surrounding nucleoplasm (Nu) (E,F). Both light and dark adapted cells contained flagella (arrow) with associated paraflagellar body (double arrow) (G,H). Both light and dark adapted cells contained Golgi apparatus with numerous narrow cisterna (arrow) (I,J). Both light and dark adapted cells contained mitochondria with narrow cristae (Mi) (K,L). Scale bars in all panels ~1µm

## 6.5 DISCUSSION AND CONCLUSION

The *E. gracilis* cells showed pronounced growth rate in light adapted cultures than in dark adapted cultures (Figure 6.1). Microscopic, spectrophotometric, and electrophoresis evidence confirmed the transition states from light to dark conditions and vice versa (Figure 6.2). This is consistent with published literatures where *E. gracilis* tend to loose its chloroplast and chlorophyll under dark conditions (Gibbs, 1960; Wolken, 1956; Seigesmund, 1962; Pellegrini, 1980; Rocchetta, *et al.,* 2007; Kivic and Vesk, 1972; Ben-Shaul, *et. al.,* 1963; Lyman, *et. al.,* 1961; Mollenhauer, *et. al.,* 1967). A complete recovery of chloroplast structure occurred with the reassociated thylakoids and clear partitions (data not shown). The different growth conditions (light and dark) did not, in anyway, affect the inherent sequencing statistics or characteristics of the *E. gracilis* samples (Table 6.2). For instance, the number of reads, total clean bases, Q20 %, and GC % were not markedly different in the light and dark replicate samples respectively. It may well be possible that the GC content of *E. gracilis* may be gene expression independent, or perhaps, that *E. gracilis* maintains a GC % between 50 % and 60 % irrespective of environmental influence. It is quite unclear if this plays any physiological or functional role. Quality analysis of the correlation of the light and dark replicates showed consistency within the replicates of the light and dark samples with a minimum and maximum correlation coefficient of 0.93 and 0.94 respectively (Figure 6.2).

There were significant changes in the protein and RNA levels with respect to light or dark conditions. Proteins and RNA levels of light adapted samples were much higher and significantly expressed when compared to dark adapted cultures (Figure 6.3, 6.4). This is presumable due to the increased utility of the photosynthetic apparatus in the light adapted cultures. In a related study, similar results, though less pronounced, were found in the DNA analyses - total cell DNA, estimated by two different techniques, was about 3.5 pg per cell when Euglena was grown in the dark; but 4.4 pg per cell (range 4.0-4.8) when grown in the light (Cook, 1972).

The proteomics analysis showed wider expression of peptides (Fugure 6.4) in the light adapted samples when compared with the transcriptomics, with even transcript expression in the light and dark samples (Figure 6.3), suggesting evidence of posttranscriptional events (Vesteg, *et al.,* 2009). This posttranscriptional event is

further supported by the presence of expressed peptides in the proteomic analysis which is absent in the transcriptomic analysis (Vesteg, *et al.,* 2009). The infinite changes (232 light vs 152 dark) in the proteomic analysis (Figure 6.5) illustrate the exclusive expression of these peptides in their respective states or conditions, corresponding to peptides which are exclusive to chloroplast and mitochondria machineries (Urbaniak, *et al.,* 2012). Analysis of the dN/dS ratios of transcripts suggest a synonymous substitution for protein coding genes in *E. gracilis* with a purifying selection. The dN/dS values (0.87 and 0.99 respectively) (Figure 6.8) recorded implies that this degree of selections are close to neutrality or no selection occured (dN/dS = 1) suggesting that the coding sequences may have evolved naturally (Kryazhimskiy and Plotkin, 2008; Jeffares, *et al.,* 2015). Furthermore, comparison of the proteomic and transcriptomic analysis (Figure 6.5) showed poor correlation with only about 4 peptides/transcripts showing strong correlation, corresponding to transcripts associated with the photosynthetic machinery. It has also been proven that dark adapted cells have limited, or entirely lacking, ribosomes (Russell and Draffan, 1978). This is consistent with our data where there are limited chloroplast associated ribosomes (e.g. 70S ribosome; Appendix III) in the dark samples which are essential for the manufacture of proteins within the plastid.

In most model eukaryotes including land plants, many signaling pathways are ultimately directed toward modulating gene expression by influencing transcription factor recruitment to promoter elements (Vesteg, *et al.,* 2009). Most genes and even non-coding sequences are constitutively transcribed in trypanosomatids and their expression is regulated post-transcriptionally by modulating mRNA stabilization and translation (Furger, *et al.,* 1997; Hotz, *et al.,* 1997; Teixeira, 1998). In trypanosomatids, the predominantly polycistronic transcription with subsequent resolution of nascent RNAs through trans-splicing and polyadenylation (Benz, *et al.,* 2005; Campbell, *et al.,* 2003; Clayton, 2002) reduces the influence of promoters as the main players in controlling gene expression, with major potential impact on signaling systems (Field, 2005). In *E. gracilis*, evidences suggest that regulation of gene expression occurs at the post-transcriptional rather than at transcriptional level (Hoffmeister, *et al.,* 2004; Keller, *et al.,* 1991; Kishore and Schwartzbach, 1992; Madhusudhan, *et al.,* 2003; Levasseur, *et al.,* 1994; Saint-Guily, *et al.,* 1994; Vacula, *et al.,* 2001; Dos Santos, *et al.,* 2007). Therefore, this analysis further supports this

evidence, as well as suggests that light or dark environmental conditions have minimal influence on the transcript or peptide expression of nuclear genes as previously report by Vesteg, *et al.,* 2009. This also suggest that *E. gracilis* is more similar to the trypanosomatids in terms of gene expression mechanism. Also, the addition of SL-RNA leaders to cytosolic mRNAs via trans-splicing is one of the features common to Euglenozoa (Frantz, *et al.,* 2000), which suggests that all members of the Euglenozoa, including euglenids and kinetoplastids, would likely share similar mechanism of gene expression.

It is evident from the electron microscopy that significant differences were observed in physiological processes and morphological structures associated in the transition from phototrophic to heterotrophic lifestyle and vice versa. This includes the photosynthetic and heterotrophic apparatus such as the chloroplast and paramylon respectively. This is also consistent with observations in published literatures (Gibbs, 1960; Wolken, 1956; Seigesmund, 1962; Pellegrini, 1980; Rocchetta, *et. al,* 2007; Kivic and Vesk, 1972; Ben-Shaul, *et. al.,* 1963; Lyman, *et. al.,* 1961; Mollenhauer, *et. al.,* 1967). Other organelles such as the flagella and Golgi apparatus remain morphological unchanged in the light and dark samples.

Strong reliance on post-transcriptional processes in the regulation of gene expression has been long recognized as a feature of *Euglena* (Saint-Guily, *et al.,* 1994), but there also is mounting evidence that post-transcriptional, translational and degradation regulation are crucial determinants of cellular protein abundances (Vogel and Marcotte, 2012) in general. Recent studies comparing the mRNA levels with the levels of protein in yeast (Foss, *et al.,* 2007) and *Arabidopsis* (Fu, *et al.,* 2009) suggest modest concordance. The partial correlation also suggests that cytosolic and organellar gene expressions are differently organized.

### *6.5.1 Significance of findings – adaptations, functions and applications*

*E. gracilis* gene expression studies was produced under light and dark environmental conditions. The results in this chapter presents the biological findings associated with the transition from light to dark and vice versa. These include an increased growth rate in light versus dark regime, presence of chloroplast/chlorophyll in light versus presence of proplastids in the dark regime, greater number of transcripts than

protein-coding genes, uniform transcript differential expression between light and dark regime, widespread of peptides/proteins in light regime when compared to dark regime, unique protein changes or presence of unique peptides between light and dark regime respectively, presence of peptides which present in proteomics analysis and absent in transcriptome analysis, non correlation of the proteomics and transcriptomic experiments, and cellular morphological changes (e.g. organelles and surface). In this section, the significance of these findings will be discussed with reference to adaptions, functions (biology), and applications.

*Light is a catalyst for growth, chloroplast to proplastid transitions are oxygen dependent:* Light is essential for photosynthetic growth in *E. gracilis*, and the mechanism for this is linked to the rate of DNA synthesis. It has been proven that in either light or dark conditions**,** the initiation of DNA synthesis in *E. gracilis* for cell division cycle (mitosis) is light dependent, and once cells commence it in the light, they complete the cycle in the dark (Yee, and Bartholomew, 1988). Hence, an increased growth rate is correlated with increased DNA and RNA synthesis and expression (Figure 6.1, 6.2, 6.4, 6.5). *E. gracilis* is also a good model for studying the biochemical and molecular mechanisms in the transition from chloroplast to proplastids (which is oxygen dependent) and vice versa (Sumida, *et al.,* 2007).

*Transcriptomics and proteomics studies are not mutually exclusive.* In Next Generation Sequencing (NGS) studies, it is relevant to carry out proteomics analysis and transcriptomics analysis simultenously. As evident from the analysis (Figure 6.4, 6.5, 6.9), there is a high transcript to protein-coding genes ratio (Figure 6.4, 6.5, and Appendix II & III), uniform expression in transcript distribution (Figure 6.4), wider distribution of peptide expressions (Figure 6.5), evidence of untranslated transcripts arising from some transcripts not having a corresponding proteomics candidate peptide, presence of some peptides in the proteomics experiment which are absent in the transcriptomics experiment (Figure 6.5). These suggests that not all transcripts were possess an ORF, and indicates, that protein manufacture in *E. gracilis* are posttranscriptionally directed.

*Energy conversion in E. gracilis is a function of two evolutionarily derived organelles:* In eukaryotes, a set of reactions occur in the cytosol, energy derived from the partial oxidation of energy-rich carbohydrate molecules is used to form ATP, the chemical

energy currency of cells (Alberts, *et al.,* 2002; Igamberdiev and Kleczkowski, 2015). Evolutionarily, an efficient method of energy generation involved membrane-bound organelles such as chloroplast and mitochondrial. The former is central to the conversion of light energy into chemical bond energy in photosynthesis, while the later is central to the aerobic respiration that requires the exchange of oxygen and carbon (IV) oxide (Alberts, *et al.,* 2002). In this studies, in the light regime, light is the primary source of energy, while in the dark regime glucose is the primary source of energy. Correlation were observed in 4 – 5 proteins corresponding to PSI, PSII, and LHC. This analysis further supports the evidences that in *E. gracilis*, as in other photosynthetic and heterotrophic systems, the chloroplast drives phototropism while the mitochondria drives heterotropism. This is also evident following the observation of unique protein changes or presence of unique peptides in light or dark regime (Figure 6.9) – these are peptides that are found exclusively either in light or dark regime.

*Cellular morphological changes are strongly influenced by light and dark conditions:* In the present studies, the major significant changes (increase in sizes and numbers) within the organelles are the paramylon and the chloroplasts which correspond to adaptations to dark and light conditions respectively. For instance, in the absence of light, and the presence of glucose as the carbon source, *E. gracilis* fixes $CO_2$ to produce the polysaccharide paramylon, a β-1,3-glucan (Calvayract, *et al.,* 1981), with yet an unknown mechanism of formation (Tanaka, *et al.,* 2017). Paramylon stored under aerobic conditions is degraded during anaerobic cultivation to produce a wax ester consisting of saturated fatty acids and alcohol chains (Inui, *et al.,* 1984). The results presented in this section further supports that paramylon synthesis is influenced by dark conditions as well as advance this understanding (by the utility of NGS approaches) and their applications in biotechnology (biofuels). It also suggests that in looking for clues on the influence of light and dark conditions in *E. gracilis*, other organelles (aside paramylon and chloroplast) will not provide suitable and sufficient information.

## 7.0 CONCLUSION

In this study, the specific aspects of the biology of *E. gracilis* have been investigated in the *E. gracilis* genome and transcriptome, with significant resolution into (amongst others): Protein trafficking, surfaceome, nuclear cohorts, and transporters (see Chapter 3 – 6 for full list). Their distribution was studied using computational as well as manual curation, which not only offers the catalogue of these specific biological aspects in the draft transcriptome, but also provide interesting insights into the evolution and functions of their protein families in eukaryotes (O'Neill, *et al.,* 2015). In some instances where phylogenetic trees are utilized, and in most instances, the protein phylogenetic analyses show moderate statistical support, suggesting that additional data may be required to address these gaps, or perhaps, that these proteins have diverged in *E. gracilis*. The presence of the core aspects of the biology of living organisms show that the transcriptome is a good dataset for further interrogation of protein families and genes in eukaryotes while the genome does not support extensive annotation analysis. The presence of novel splicing mechanisms, unusual surface and complex trafficking system, hybrid genome, and the possession of conventional and kinetoplastid chromosome segregation machineries, and posttranscriptional events, characterizes the complex biology of this unique organism (Figure 7.1). Still yet, there are some aspects of the *E. gracilis* biology that still remain unanswered until a complete genome sequence is available, and these include: the full gene compliment, the evolutionary origins of these genes as well as the complete splicing mechanism.

### 7.1 CONTRIBUTION OF FINDINGS TO SCIENTIFIC KNOWLEDGE

The findings in this work has provided specific contributions to scientific knowledge. While the findings (e.g. information processing, plastid, mitochondria) from majority of the biological aspects investigated are consistent or conventional with other eukaryotic genomes, there is however specific findings with significance to scientific knowledge as described below.

*Large genome size:* Most prokaryotes and eukaryotes possess genome sizes that are smaller than the human genome. However, there are a handful of eukaryotic genomes that are larger than the human genome (3 Gbp), and include amongst

others, *Picea abies* (Norway spruce; 19.6 x $10^9$ Gbp), *Psilotum nudum* (Whisk fern; 2.5 x $10^{11}$ Gbp), *Paris japonica* (canopy plant; 150 Gbp), and *Amoeba dubia* (670 Gbp). The evidences in this work has potentially captured *E. gracilis* as one of the few organisms with genome size either close to, or larger than, the human genome (Figure 7.1, and Ebenezer, *et al.,* 2017). These evidences are also consistent with the scientific knowledge that genome size is not a function of biological complexity, and that genome sizes more, or less, correlate with number of genes.

*Shared genes and hybrid genome:* Several research papers have predicted the hybrid nature of *E. gracilis*. For instance, Ahmadinejad, *et al.,* 2007, predicted and concluded that the genome of *E. gracilis* is a hybrid of photosynthetic and heterotrophic genomes using information from the nearest neighbours of its genes. The gene catalogues in *E. gracilis*, they posited, will consist of four main gene classes: 1) *E. gracilis* specific genes, 2) Kinetoplastid-specific genes, 3) eukaryotic genes that are spread in other eukaryotes (eukaryotic conserved genes), and 4) genes acquired during the secondary endosymbiotic relationship. Since this publication (Ahmadinejad, *et al,* 2007) over a decade ago, there has not been any molecular data (NGS) to support these predictions. Using NGS approaches, the data in this work further supports this predictions, and beyond. For instance, in this work, are Euglena genes that are conserved across eukaryotes, Euglena-specific genes, Euglena genes shared with Kinetoplastids, and Euglena genes shared with green algae, taking into account the results from the orthologous clustering analysis. Beyond the predictions of Ahmadinejad, *et al.,* 2007, are genes that are also shared specifically with red algae, land plants, free living Bodonids, and other individual selected eukaryotes. This analysis further extends the endosymbiotic status and nature of *E. gracilis* beyond that previously known and described in Ahmadinejad, *et al.,* 2007. However, it's not yet known if all of the hybrid biological nature observed in this work were acquired during the endosymbiotic events that occurred with the cyanobacteria and proteobacteria, or if these were as a result of gene loss and gains across selected eukaryotes. I hope that a Lateral Gene Transfer (LGT) analysis will provide answers to this question.

*Complex endomembrane and heterochromatin organization:* In several eukaryotes, including kinetoplastids, the endomembrane system and heterochromatin

organization is simple, with one or two copies of trafficking genes (i.e having simple trafficking routes) and the presence of peripheral heterochromatin around the nuclear envelope respectively. In *E. gracilis*, the situation seems to be different, there are multiple routes of trafficking as well as the absence of peripheral heterochromatin around the nuclear envelope (Figure 7.1), demonstrating its complexity, and appears to be one of the few eukaryotic organism to demonstrate complexity in these biology. This suggests active protein trafficking activities and might explain the absence of a nuclear lamin in *E. gracilis* respectively. Similarly, several literatures have reported a condensed heterechromatin in *E. gracilis* (see Chapter 1), however, this data does not support this observation as the *E. gracilis* heterchromatin appears to be less condensed (see Chapter 4).

*Conventional and non conventional splicing and kinetochores:* In most eukaryotes, the splicing mechanism and kinetochores are conventional except in kinetoplastids where the kinetochores are non conventional (Akiyoshi and Gull, 2013, 2016; D'Archivio and Wickstead, 2016). In *E. gracilis*, the splicing mechanism and kinetochores are both conventional and non conventional, suggesting an unusual feature for most eukaryotes (Figure 7.1). This further suggests that *E. gracilis* will serve as a suitable model for studies involving splicing and chromosome segregation.

*Presence of DBP5 and absence of nuclear lamin:* The analysis in this work suggests that the nuclear system in *E. gracilis* is conserved, consistent and conventional with those found in other eukaryotes such as *T. brucei, A. thaliana, S. cerevisiae,* and *H. sapiens* (Figure 7.1). However, a major strinking difference is the presence of DBP5 and the absence of a nuclear lamin in *E. gracilis* (see Chapter 5). While DBP5 is absent in *T. brucei* and present in other eukaryotes, nuclear lamins are present in most eukaryotes (including the reference four above). The absence of DBP5 in *T. brucei* (which is present in many eukaryotes) and it's presence in *E. gracilis* suggests that this protein extends to the base of the Euglenozoa and may have been lost in *T. brucei* – this raises the question: What substitutes for the role of mRNA export mechanism in *T. brucei*? Similarly, the absence of the nuclear lamins in *E. gracilis* may be unexpected and raises the question about the origin of nuclear lamins and what performs its functions in *E. gracilis*. This also correlates with the results of

published work on the relationship between the heterochromatin and the nuclear lamin, which are both connected by the protein HP1 and PRR14, and ensures that the heterochromatins in eukaryote are peripheral around the nuclear envelope (see Chapter 5). Since the heterochromatin organization in *E. gracilis* is non peripheral around the nuclear envelope, the absence of the nuclear lamins may explain this situation.

*Unusual surface:* When the family of a closely related glycoprotein, Amastin, was discovered in *T. cruzi* (Teixeira, *et al.,* 1994), and subsequently in *L. major* (Jackson, 2010), it was observed that amastin repertoire is much larger in *Leishmania* relative to *Trypanosoma* (Jackson, 2010). The study further highlighted four distinct amastin subfamilies within the trypanosomatids with diversification which occurred after the origin of Leishmania, suggesting that some amastin genes evolved novel functions crucial to cell function in leishmanial parasites after the acquisition of a vertebrate host. In the present work, the amastin gene family and repertoire are quite different from those of the kinetoplastids, suggesting a cellular function of *E. gracilis* to adaptation to environmental factors. The presence and uniqueness of the amastin gene family in *E. gracilis* (in this work) further extends it's origin to the base of the Euglenozoa. This may mean that the amastin gene family may have initially diversified from the base of the Euglenozoa rather than from the base of the trypanosomatids (*Leishmania spp*).

*Post-transcriptional regulation:* The data in this work supports the observations that posttranscriptional regulation (Figure 7.1) of gene expression is an ancient activity of earliest eukaryotes, considering the position of *E. gracilis*. The evidence for post-transcriptional regulation in *E. gracilis* are seen in the light and dark experiments (see Chapter 6), as well as the presence of the regulatory elements and activities such as RNA binding protein (RBP), alternative splicing, nuclear degradation (exosomes), editosomes, RNA recognition motif (RRM), and nuclear export (see Chapter 5 for details). The data in this work also correspond with gene expression studies in trypanosomes, where virtually all control of gene expression is posttranscriptional (Droll, *et al.,* 2013). Posttranscriptional regulation may also play adaptive roles to external environmental factors, such as heat shock (HSP), or to host-parasite interactions as seen in trypanosomes (Droll, *et al.,* 2013).

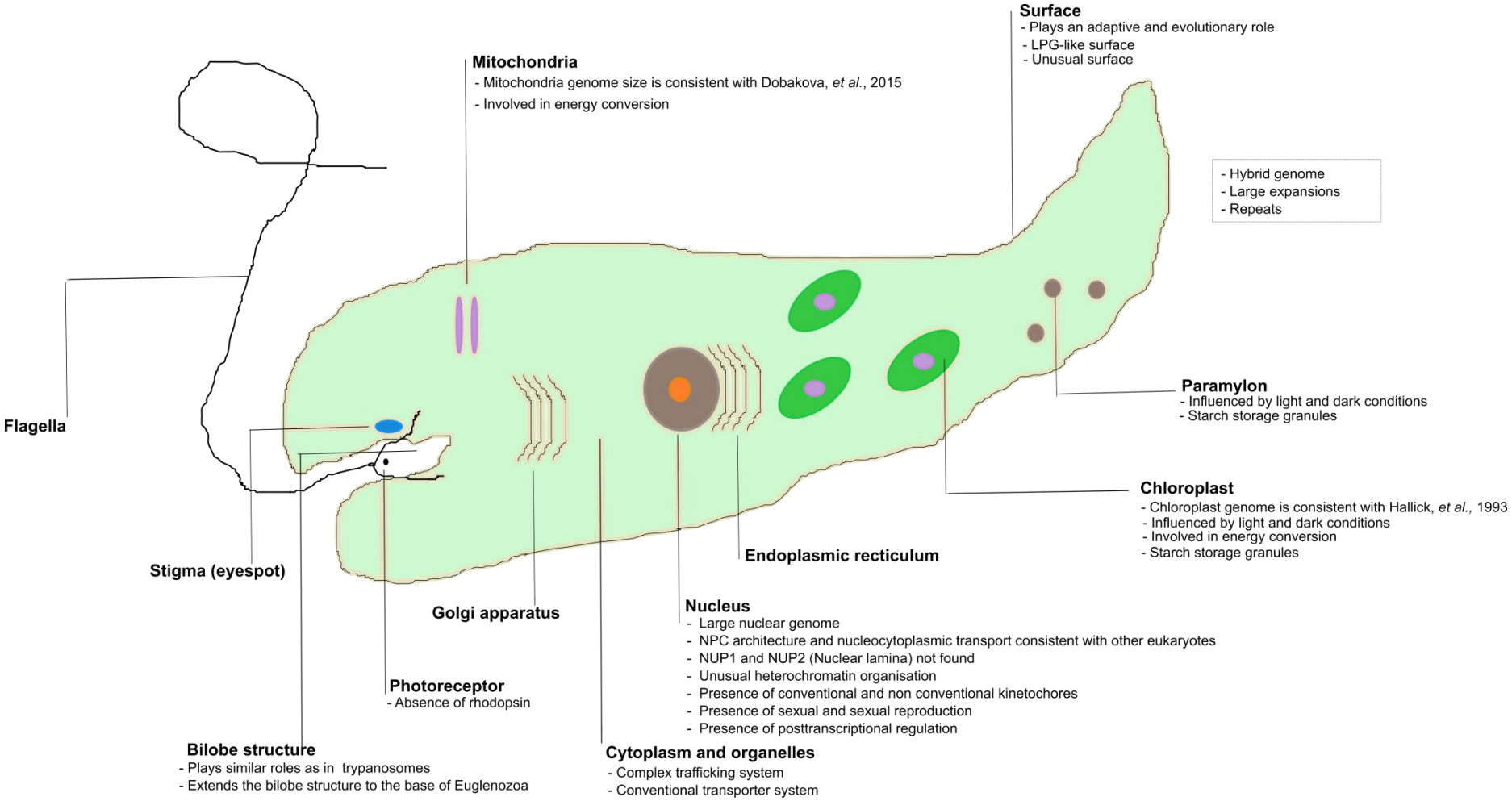

Surface
- Plays an adaptive and evolutionary role
- LPG-like surface
- Unusual surface
Mitochondria
- Mitochondria genome size is consistent with Dobakova, *et al.*, 2015
- Involved in energy conversion
- Hybrid genome
- Large expansions
- Repeats
Flagella
Paramylon
- Influenced by light and dark conditions
- Starch storage granules
Chloroplast
- Chloroplast genome is consistent with Hallick, *et al.*, 1993
- Influenced by light and dark conditions
- Involved in energy conversion
- Starch storage granules
Stigma (eyespot)
Endoplasmic recticulum
Golgi apparatus
Nucleus
- Large nuclear genome
- NPC architecture and nucleocytoplasmic transport consistent with other eukaryotes
- NUP1 and NUP2 (Nuclear lamina) not found
- Unusual heterochromatin organisation
- Presence of conventional and non conventional kinetochores
- Presence of sexual and sexual reproduction
- Presence of posttranscriptional regulation
Photoreceptor
- Absence of rhodopsin
Bilobe structure
- Plays similar roles as in trypanosomes
- Extends the bilobe structure to the base of Euglenozoa
Cytoplasm and organelles
- Complex trafficking system
- Conventional transporter system

**Figure 7.1. *E. gracilis* possesses a complex biology.** The diagram shows the schematic summarized representation of findings and conclusion. It shows the *E. gracilis* cell with associated organelle and cellular component (lines with title highlights). Annotated organelles and cellular components delineates the major associated findings. Text in box highlight findings that is general to the biology of *E. gracilis.*

## 7.2 RECOMMENDATIONS AND FUTURE WORKS

The analysis in this work has provided further platform and framework for additional scientific investigations into the biology of *E. gracilis*. This genome project has been propelled by gene discovery, functional and comparative genomics, with futuristic significance to the wider society in drug discovery, biotechnology, and environmental studies, amongst others. While significant amount of efforts and progress has been made in deciphering the genome and transcriptome sequence of *E. gracilis*, the production of a complete genome for *E. gracilis* has been quite challenging, particularly for technical reasons: Short read lengths and multiple repeat elements. The analysis in this work still leave several questions unanswered as well as provide a platform for further interrogation of specific areas of interests as will be discussed below.

*Recommendations:* The analysis of the genome of *E. gracilis* raises the question of whether there are expected to be potential breakthroughs with current sequencing technologies and bioinformatics algorithms? While it's been previously thought about the potential negative impact of Base J (see introduction) on *E. gracilis* DNA sequencing, it is not expected that this will be the case considering that the proportion of Base J in *E. gracilis* is the same in *T. brucei* – and the current nucleic acid extraction technique for *E. gracilis* have proven to produce high quality DNA (see Chapter 2). The two strong candidate recommendations for a complete *E. gracilis* genome will involve sequencing large volumes of long reads of > 100 kbp in length using PacBio/Nanopore technologies and the development of bioinformatics algorithm that can deal with high level of polyploidy in eukaryotic genome assembly. Considering the error rate for PacBio, the strongest candidate for this would be Nanopore – but this raises the question of cost since a huge amount of data is required to be generated to attain the level of genome coverage for a complete genome sequencing. It is also not yet known how long it will take for a complete *E. gracilis* genome sequence considering that some potential challenges may be unforeseen. To overcome these challenges will require strong bioinformatics collaborations with existing genome centers and model organism database communities (e.g. EuPath, EBI-EMBL, Joint Genome Institute) in what I will otherwise call the Euglena Genome Consortium (EGC). This is required to provide

both technical and administrative support, for instance, in the development of assembly algorithms and pipelines adapted to *E. gracilis*. The choice of this consortium is due to the relevance of *E. gracilis* genome to the biomedical, environmental, biotechnology, and energy community.

*Future directions:* The studies in this work provides a platform for further interrogation of specific area of the biology of *E. gracilis* as well as it's utility and contribution to the wider society through its technological applications. There are several applications and biological aspects (wet lab) of *E. gracilis* which has been stalled due to the absence of molecular data (genome and transcriptome). The production of a high quality transcriptome means that researches in *E. gracilis* using proteomics methods is now possible. Furthermore, gene expression studies is now possible with the availability of a reference transcriptome and proteome, particularly looking into environmental (response to stress), population diversity, and signal transduction (circadian rythms). While several of the analysis of these biological aspects of *E. gracilis* are possible with a complete transcriptome and proteome, some specific biological aspects such as genetics and there applications in biotechnology and biofuel may still be challenging until a complete genome sequence is available.

So, what will be the sequence of events? Additional funding will be sourced for the *E. gracilis* genome project and the production of a complete genome will progress in parallel with other studies such as population diversity, gene expression under differing environmental conditions, and signal transduction. Some specific questions are required to be answered when the complete genome of *E. gracilis* becomes available. This include: The full gene compliments and the sources of these genes in evolutionary timelines, splicing mechanisms and comprehensive genomic architecture, the evolution of introns and twintrons and their roles. This will also reveal the actual size for the genome, and provide further evidence for such a large genome size when compared to kinetoplastids which are their sister relatives.

## 7.3 CONCLUDING REMARKS

The genome and transcriptome of *E. gracilis* has now been sequenced with the assembly of the data into working and final drafts respectively. While the transcriptome can support extensive annotation analysis, the genome cannot.

Evolutionary, functional and comparative analysis of the transcriptome suggests a complex biology for this unique organism. While some specific biological questions (such as trafficking, nuclear system, gene expression) and their potential applicability were able to be answered, which could be inferred from the predicted proteome of the transcriptome, some specific biological questions such as full gene compliment, splicing mechanism, and actual genome size may still remain elusive until a complete genome sequence becomes available.

## 8.0 APPENDICES

**APPENDIX I:** COMMUNITY-BASED ANNOTATION

**Table 1:** Accessions of genes associated with specific cellular functions, metabolic pathways and organelles.

**A:** https://www.dropbox.com/s/ls5pn3be4o1fz6t/Appendix%20I-a.xlsx?dl=0

**B:** https://www.dropbox.com/s/67jhfo6ijbfqipa/Appendix%20I-b.xlsx?dl=0

**APPENDIX II:** DIFFERENTIAL EXPRESSION ANALYSIS FILES

https://www.dropbox.com/s/aybw5wno79e697t/Appendix%20II.xlsx?dl=0

**APPENDIX III:** LIGHT AND DARK TRANSCRIPTOMIC AND PROTEOMIC ANALYSIS

**Table 1: Raw data for proteomics and transcriptomics of *E. gracilis* under adaptive conditions.**

A: https://www.dropbox.com/s/lupj2j7vzffc78y/Appendix%20III-A.xlsx?dl=0

B: https://www.dropbox.com/s/e9xgnz8v0cpga04/Appendix%20III-B.xlsx?dl=0

C: https://www.dropbox.com/s/v32kuns5or3t695/Appendix%20III-C.xlsx?dl=0

Cells were grown under dark or light conditions and described in methods and subjected to protein or RNA extraction and analysed by mass spectrometry or RNAseq. Each condition was analysed in triplicate (n = 3) and data for individual samples together with the merged data are provided as Appendix IV-A and B, together with BLAST annotation of altered transcripts (Appendix IV-C). Data are presented graphically in Chapter 6.

**Table 2: Summary of the synonymous and nonsynonymous information comparisons across 196251 comparisons**

D: https://www.dropbox.com/s/4pxiksqjfyf5pmz/Appendix%20III-D.txt?dl=0

## 8.0 REFERENCES


Abbasi, K., DuBois, K. N., Dacks, J. B., Field, M. C. (2011). A Novel Rho-Like Protein TbRHP Is Involved in Spindle Formation and Mitosis in Trypanosomes. *PLoS ONE,* **6(11):** e26890.

Acestor, N., Zikova, A., Dalley, R. A., Anupama, A., Panigrahi, A. K., Stuart, K. D., (2011) Trypanosoma brucei mitochondrial respiratome: composition and organization in procyclic form. *Mol. Cell. Proteomics,* **10 (M110):**006908.

Adams, M. D. Dubnick, M., Kerlavage, A. R., Moreno, R., Kelley, J. M., Utterback, T. R., Nagle, J. W., Fields, C. and Venter, J. C. (1992). Sequence identification of 2,375 human brain genes. *Nature,* **355:**632–634.

Adl, S. M., Simpson, A. G. B., Farmer, M. A., Andersen, R. A., Anderson, O. R., Barta, J. R., Bowser, S. S., Brugerolle, G., Fensome, R. A., Fredericq, S., James, T. Y., Karpov, S., Kugrens, P., Krug, J., Lane, C. E., Lewis, L. A., Lodge, J., Lynn, D. H., Mann, D. G., Mccourt, R. M., Mendoza, L., Moestrup, O., Mozley- Standridge, S. E., Nerad, T. A., Shearer, C. A., Smirnov, A.V., Spiegel, F. W. & Taylor, M. F. J. R. (2005). The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J. Eukaryot. Microbiol.,***52:** 399 – 451.

Adl, S. M., Simpson, A. G., Lane, C. E., Lukeš, J., Bass, D., Bowser, S. S., Brown, M. W., Burki, F., Dunthorn, M., Hampl, V., Heiss, A., Hoppenrath, M., Lara, E., Le Gall, L., Lynn, D. H., McManus, H., Mitchell, E. A., Mozley-Stanridge, S. E., Parfrey, L. W., Pawlowski, J., Rueckert, S., Shadwick, R. S., Schoch, C. L., Smirnov, A., Spiegel, F. W. (2012). The revised classification of eukaryotes. *J Eukaryot Microbiol.* **59(5):**429-93. doi: 10.1111/j.1550-7408.2012.00644.x.

Adung'a, V. O., Gadelha, C. and Field, M. C. (2013). Proteomic Analysis of Clathrin Interactions in Trypanosomes Reveals Dynamic Evolution of Endocytosis. *Traffic,* 1**4:** 440–457

Aebi, M. (2013). N-linked protein glycosylation in the ER. *Biochimica et Biophysica Acta*, **1833:**2430–2437.

Afek, A. and Lukatsky, D. B. (2013). Genome-Wide Organization of Eukaryotic Preinitiation Complex IsInfluenced by Nonconsensus Protein-DNA Binding.Biophysical Journal, **104:**1107–1115.

Agrawal, G. K., Jwa, N. S., Lebrun, M. H., Job, D., and Rakwal, R (2010). Plant secretome: unlocking secrets of the secreted proteins. *Proteomics,* **10** (4): 799–827. *doi*:*10.1002/pmic.200900514*.

Ahmadinejad N., Dagan, T., Martin, W. (2007). Genome history in the symbiotic hybrid *Euglena gracilis. Gene,* **402:** 35–39.

Akiyoshi, B. and Gull, K. (2013). Evolutionary cell biology of chromosome segregation: insights from trypanosomes. *Open Biol.,* **3:** 130023. http://dx.doi.org/10.1098/rsob.130023.

Akiyoshi, B. and Gull, K. (2014).Discovery of Unconventional Kinetochores in Kinetoplastids. *Cell,* **156:**1247–1258.

Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B. T., Rout, M. P. and Sali, A.(2007). Determining the architectures of macromolecular assemblies. *Nature,* **450(7170):**683–94. Epub 2007/11/30. [pii] doi: 10.1038/nature06404 PMID: 18046405.

Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B. T., Sali, A. and Rout, M. P. (2007). The molecular architecture of the nuclear pore complex. *Nature,* **450(7170):**695–701.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). Molecular Biology of the Cell. 4th edition. New York: Garland Science. *The Genetic Systems of Mitochondria and Plastids.* Available from: http://www.ncbi.nlm.nih.gov/books/NBK26924/

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). Molecular Biology of the Cell. 4th edition. New York: Garland Science;

Chapter 14, Energy Conversion: Mitochondria and Chloroplasts. Available from: https://www.ncbi.nlm.nih.gov/books/NBK21063/

Alegre, C. and Subirana, J. A. (1989). The diameter of chromatin fibres depends on linker length. *Chromosoma*, **98:** 77-80

Allen, C. A., Jackson, A. P., Rigden, D. J., Willis, A. C., Ferguson, S. J. and Ginger, M. L. (2008). Order within a mosaic distribution of mitochondrial c-type cytochrome biogenesis systems? *FEBS J.,* **275:**2385–2402.

Allen, J. F., de Paula, W. B. M., Puthiyaveetil, S. and Nield, J. (2011). A structural phylogenetic map for chloroplast photosynthesis. *Trends in Plant Science,* **16 (12):** 645 – 655. doi:10.1016/j.tplants.2011.10.004.

Al-Qasmi, M., Raut, N., Talebi, S., Al-Rajhi, S., Al-Barwani, T. (2012). A Review of Effect of Light on Microalgae Growth. Proceedings of the World Congress on Engineering 2012 Vol I WCE 2012, July 4 - 6, 2012, London, U.K.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J.,Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.*Nucleic Acids Research,* **25(17):** 3389–3402.

Ambrósio, D. L., Badjatia, N. and Günzl, A. (2015). The spliceosomal PRP19 complex of trypanosomes. *Molecular Microbiology*, **95(5):** 885–901. doi:10.1111/mmi.12910.

Amos, W. B. and Grimstone, A. V. (1968). Intercisternal material in the Golgi body of *Trichomonas. J. Cell Biol.,* **38:**466-471.

Ana Maria Rojas, Gloria Fuentes, Antonio Rausell, and Alfonso Valencia (2012). The Ras protein superfamily: Evolutionary tree and role of conserved amino acids. JCB: Review, Evolution. *J. Cell Biol.,* **196(2):** 189–201.

Analia Tolivia, Visitacion Conforti, Osvaldo Cordoba, Lujan Flores (2013). Chemical constituents and biological activity of *Euglena gracilis* extracts. *Journal of pharmacy research,* **7:** 209 – 214.

Angeler, D. G., Mullner, A. N., Schagerl, M. (1999). Comparative ultrastructure of the cytoskeleton and nucleus of Distigma (Euglenozoa). *Europ. J. Protistol.,* **35:**309 – 318.

Annie Regnault, Francoise Piton and Regis Calvayrac (1990). Growth, Proteins and Chlorophyll in *Euglena* Adapted to Various C/N Balances. *Phytochemistry,* **29(12):** 3711-3715

Anthony M. Bolger, Marc Lohse, and Bjoern Usade (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, **pp** 1–7. doi:10.1093/bioinformatics/btu170 http://hannonlab.cshl.edu/fastx_toolkit/

AOKI, K. F. & KANEHISA, M. (2005). Using the KEGG database resource, chapter 1.12 in Current Protocols in Bioinformatics, Current Protocols (Miranker, L., series ed.) (John Wiley & Sons, August, 2005).

Aphasizhev, R., and Aphasizheva, I. (2011). Uridine insertion/deletion editing in trypanosomes: a playground for RNA-guided information transfer. *Wiley Interdiscip. Rev. RNA*, **2:**669–685.

Aphasizhev, R., Aphasizheva, I., and Simpson, L. (2003c). A tale of two TUTases. *Proc. Natl. Acad. Sci. USA,* **100:**10617–10622.

*Arabidopsis* Genomics Initiative (AGI) (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature,* **408:**796 – 815.

Archer, S.K., Luu, V.D., de Queiroz, R.A., Brems, S., and Clayton, C. (2009). Trypanosoma brucei PUF9 regulates mRNAs for proteins involved in replicative processes over the cell cycle. *PLoS Pathog.,* **5:** e1000565.

Archibald, J. M. (2005). Jumping genes and shrinking genomes—probing the evolution of eukaryotic photosynthesis with genomics. IUBMB Life, **57:**539–547.

Archibald, J. M., Rogers, M. B., Toop, M., Ishida, K., Keeling, P. J. (2003). Lateral gene transfer and the evolution of plastid-targeted proteins in the

secondary plastid-containing alga *Bigelowiella natans*. *Proc. Natl. Acad. Sci. U. S. A.* **100:**7678–7683.

Armbrust, E. V., Berges, J. A., Bowler, C., Green, B. R., Martinez, D., Putnam, N. H., Zhou, S., Allen, A. E., Apt, K. E., Bechner, M., Brzezinski, M. A., Chaal, B. K., Chiovitti, A., Davis, A. K., Demarest, M. S., Detter, J. C., Glavina, T., Goodstein, D., Hadi, M. Z., Hellsten, U., Hildebrand, M., Jenkins, B. D., Jurka, J., Kapitonov, V. V., Kroger, N., Lau, W. W., Lane, T. W., Larimer, F. W., Lippmeier, J. C., Lucas, S., Medina, M., Montsant, A., Obornik, M., Parker, M. S., Palenik, B., Pazour, G. J., Richardson, P. M., Rynearson, T. A., Saito, M. A., Schwartz, D. C., Thamatrakoln, K., Valentin, K., Vardi, A., Wilkerson, F. P. & Rokhsar, D. S. (2004). The genome of the diatom *Thalassiosira pseudonana*: *ecology, evolution, and metabolism. Science,* **306:**79 – 86.

Armstrong, J. R., Glasscock, J., Schillebeeckx, I. and Dadighat, N. (2012). Strategies for *De Novo* Assembly of Genomes and Transcriptomes Using Combined Illumina and Roche 454 Sequencing Data. Cofactor Genomics, St. Louis, MO. http://cofactorgenomics.com/strategies-de-novo-assembly-genomes-and-transcriptomes-using-combined-illumina-and-roche-4/

Arnez, J. G. and Steitz, T. A. (1994). Crystal structure of unmodified tRNAGln complexed with glutaminyltRNA synthetase and ATP suggests a possible role for pseudo-uridines in stabilization of RNA structure. Biochemistry, **33:**7560–7567.

Arnon, D. I. and Crane, F. L. (1965). In Biochemistry of Quinones, p. 433. Ed. By Morton, R. A. New York and London: Academic Press Inc.

Arnott, H. J. and Waine, P. L. (1967). Observations on the fine structure of the pellicle pores of *Euglena granulata. Protoplasma,* **64:** 330-344.

Asaul, Z. I. (1975). Viznachnik evlenovikh vodorostey Ukrainskoy R. S. R. [Survey of the euglenophytes of the Ukrainian SSR]. *Naukova Dumka, Kiev*. 407 **pp.** [in Ukrainian].

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics,* **25:**25 – 29. doi:10.1038/75556

Aslett, M., Aurrecoechea, C., Berriman, M., Brestelli, J., Brunk, B. P., Carrington, M., Depledge, D. P., Fischer, S., Gajria, B., Gao, X., Gardner, M. J., Gingle, A., Grant, G., Harb, O. S., Heiges, M., Hertz-Fowler, C., Houston, R., Innamorato, F., Iodice, J., Kissinger, J. C., Kraemer, E., Li, W., Logan, F. J., Miller, J. A., Mitra, S., Myler, P. J., Nayak, V., Pennington, C., Phan, I., Pinney, D. F., Ramasamy, G., Rogers, M. B., Roos, D. S., Ross, C., Sivam, D., Smith, D. F., Srinivasamoorthy, G., Stoeckert, C. J. Jr, Subramanian, S., Thibodeau, R., Tivey, A., Treatman, C., Velarde, G. and Wang, H. (2010). TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Research,* **38(D457 – D462):** doi:10.1093/nar/gkp851.

Atema, J. (1973). Microtubule theory of sensory transduction. *J. theor. Biol.,* **38:** 181–190.

Atkinson, G. C., Baldauf, S. L., and Hauryliuk, V. (2008). Evolution of nonstop, no-go and nonsense-mediated mRNA decay and their termination factor-derived components. *BMC Evol. Biol.,* **8:**290-308.

Attardi, G. and Schatz, G. (1988). Biogenesis ofmitochondria. *Annu Rev Cell Biol.,***4:** 289 – 333.

Auffinger, P. and Westhof, E. (1998). Effects of pseudouridylation on tRNA hydration and dynamics: a theoretical approach. In: *Modification and Editing of RNA* (Grosjean, H. & Benne, R., eds), **pp.** 103 – 112, ASM Press, Washington, DC.

Auld, D. S., Kawaguchi, H., Livingston, D. and Vallee, B. L. (1974a). Reverse transcriptase from avian myeloblastosis virus: a zinc metalloenzyme. *Biocliem. biophys. Res. Commun.* **57:** 967-972.

Auld, D. S., Kawaguchi, H., Livingston, D. and Vallee, B. L. (1974). RNA dependent DNA polymerase (reverse transcriptase) from avian myeloblastosis virus: a zinc metalloenzyme. *Proc. natn. Acad. Sci., U.S.A.* **71:** 2091-2095.

Avadhani, N. G. and Buetow, D. E. (1972). Isolation of active polyribosomes from the cytoplasm, mitochondria and chloroplasts of *Euglena gracilis*. *Biochem. J.,* **128:**353–365.

Avadhani, N. G. and Buetow, D. E. (I972b). Isolation of active polyribosomes from the cytoplasm, mitochondria and chloroplasts of *Euglena gracilis*. *Biochemical Journal.,* **128:**353 - 365.

Aykut, A. O., Atilgan, A. R., Atilgan, C. (2013). Designing Molecular Dynamics Simulations to Shift Populations of the Conformational States of Calmodulin. *PLoSComput Biol,***9(12):** e1003366. doi:10.1371/journal.pcbi.1003366

Ayub, M. J., Lapadula, W. J., Hoebeke, J. and Smulski, C. R. (2012). Ribosomes from Trypanosomatids: Unique Structural and Functional Properties. *"Cell-Free Protein Synthesis",* book edited by Manish Biyani, ISBN 978-953-51-0803-0 DOI: 10.5772/48336.

Bachellerie, J. P., Cavaille, J. and Huttenhofer, A. (2002). The expanding snoRNA world. *Biochimie*, **84(8):**775–790.

Bachvaroff, T. R., Sanchez Puerta, M. V. and Delwiche, C. F. (2005). Chlorophyll c-containing plastid relationships based on analyses of a multigene data set with all four chromalveolate lineages. *Mol. Biol. Evol.,* **22:** 1772–1782.

Bailey, T. L., Bodén, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., Noble, W. S. (2009). "MEME SUITE: tools for motif discovery and searching", *Nucleic Acids Research*, **37:**W202-W208

Bak, A. L., Zeuthen, J. and Crick, F. H. C. (1977). Higher-order structure of human mitotic chromosomes. *Proc. Natl. Acad. Sci. USA,* **74:**1595-1599.

Baker, C. L., (1933). Studies on the cytoplasmic components of Euglena gracilis Klebs. *Arch. Protistenk.,* **80:**434-468

Bakker, B. M., Michels, P. A. M., Opperdoes, F. R., and Westerhoff, H. V. (1997). Glycolysis in bloodstream form *Trypanosoma brucei* can be understood in terms of the kinetics of the glycolytic enzymes. *J. Biol. Chem.,* **272**:3207–3215.

Banchetti, R., Rosati, G. and Verni, F. (1980). Cytochemical analysis of the photoreceptor in *Euglena gracilis* Klebs (Flagellata Euglenoidina). *Monitore Zool. Ital.,* **14:** 165–171

Barbosa, R. L., Legrand, P., Wien, F., Pineau, B., Thompson, A., Guimaraes, B. G. (2014). RRP6 from Trypanosoma brucei: Crystal Structure of the Catalytic Domain, Association with EAP3 and Activity towards Structured and Non-Structured RNA Substrates. *PLoS ONE,* **9(2):** e89138. doi:10.1371/journal.pone.0089138.

Bark, I.C., Hahn, K.M., Ryabinin, A.E., Wilson, M.C. (1995). Differential Expression of SNAP-25 Protein Isoforms During Divergent Vesicle Fusion Events of Neural Development. *Proc. Natl. Acad. Sci. USA.,* **92(5):** 1510-1514. DOI: 10.1073/pnas.92.5.1510.

Barkan, A. (2011) Expression of plastid genes: organelle-specific elaborations on a prokaryotic scaffold. *Plant Physiol.,* **155:**1520–1532

Barkan, A. and Stern, D. B. (1998). Chloroplast mRNA processing: Intron splicing and 3'-end metabolism, in: Bailey-Serres J., Gallie D. R. (Eds.), A Look Beyond Transcription: Mechanisms Determining mRNA Stability and Translation in Plants, American Society of Plant Physiologists, Rockville, MD, **pp.** 162–173.

Barneche, F., Gaspin, C., Guyot, R. and Echeverría, M. (2001). Identification of 66 box C/D snoRNAs in Arabidopsis thaliana: extensive gene duplications

generated multiple isoforms predicting new ribosomal RNA 2(-O-methylation sites). *J. Mol. Biol.* **311:**57–73.

Barnes, R. L., Shi, H., Kolev, N. G., Tschudi, C., and Ullu, E. (2012). Comparative Genomics Reveals Two Novel RNAi Factors in *Trypanosoma brucei* and Provides Insight into the Core Machinery. *PLoS Pathog,* **8(5):** e1002678. doi:10.1371/journal.ppat.1002678.

Baroin, A., Perasso, R., Qu, L.H., Brugerolle, G., Bachellerie, J. P. and Adoutte, A. (1988). Partial phylogeny of the unicellular eukariotes based on rapid sequencing of a portion of 28S ribosomal RNA. *Proc. natn. Acad. Sci. U.S.A.,* **85:** 3474–3478.

Barras, D. R., and Stone, B. A. (1968). Carbohydrate composition and metabolism in Euglena. In: Edited by Buetow DE. *The Biology of Euglena* vol. 2. Academic Press, New York **pp**149–191.

Barsanti, L., Bastianini, A., Passarelli, V., Tredici, M. R., Gualtieri, P. (2000). Fatty acid content in wild type and WZSL mutant of *Euglena gracilis*. *J Appl Phycol.,* **12:**515 - 520.

Barsanti, L., Coltelli, P., Evangelista, V., Passarelli, V., Frassanito, A. M., Vesentini, N., Gualtieri, P. (2008). Low-resolution characterization of the 3D structure of the *Euglena gracilis* photoreceptor. *BBRC,* **375:** 471–476.

Barsanti, L., Passarelli, V., Walne, P. L. and Gualtieri, P. (1997). In vivo photocycle of the *Euglena gracilis* photoreceptor. *Biophys. J.,* **72:**545–553.

Barsanti, L., Passarelli, V., Walne, P. L. and Gualtieri, P. (2000). The photoreceptor protein of *Euglena gracilis*. *FEBS Letters,* **482:** 247 – 251.

Barsanti, L., Vismara, R., Passarelli, V., Gualtieri, P. (2001). Paramylon (b-1,3-glucan) content in wild type and WZSL mutant of *Euglena gracilis*. Effects of growth condition. *J Appl Phycol.,* **13:**59-65.

Bartlett, G. J., Walne, P. L., Schwartz, J. A. and Brown, D. H. (1972). Large scale isolation and purification of eyespot granules from *Euglena gracilis* var. *bacillaris. Pl. Physiol.,* **49:** 881–885.

Batista, T. M. and Marques, J. T. (2011). RNAi pathways in parasitic protists and worms. *Journal of Proteomics*, **74:**1504 – 1514.

Batista, T. M., and Marques J. T. (2011). RNAi pathways in parasitic protists and worms. *Journal of Proteomics,* **74:**1504 – 1514.

Batra, P. P. and Tollin, G. (1964). Phototaxis in *Euglena.* I Isolation of the eyespot granules and identification of the eyespot pigments. *Biochem. biophys. Acta,* **79:** 371–378.

Baum, D. A. and Baum, B. (2014). An inside-out origin for the eukaryotic cell. *BMC Biol.,* **12:**76.

Becker, T., Franckenberg, S., Wickles, S., Shoemaker, C. J., Anger, A. M., Armache, J. P., Sieber, H., Ungewickell, C., Berninghausen, O., Daberkow, I., Karcher, A., Thomm, M., Hopfner, K. P., Green, R., and Beckmann, R. (2012). Structural basis of highly conserved ribosome recycling in eukaryotes and archaea. *Nature,* **482:**501-506.

Beech, P. L., Nheu,T., Schultz,T., Herbert,S., Lithgow,T., Gilson,P. R., McFadden,G. I. (2000). Mitochondrial FtsZ in a chromophyte alga. *Science*, **287(5456):** 1276–1279.

Behnke, O. and Forer, A. (1967). Evidence for four classes of microtubules in individual cells. *J. Cell Sci.,* **2:**169-192.

Belasco, J. G. (1993). mRNA degradation in prokaryotic cells: An overview, in: Belasco J. G., Brawerman G. (Eds.), Control of messenger RNA stability, Academic Press, Inc., New York, NY, **pp.** 3–12.

Belhadri, A. (1995). Presence of centrin in the human parasite *Giardia*: a further indication of its ubiquity in eukaryotes. *Biochem. Biophys. Res. Commun.,* **214:**597– 601.

Belhadri, A. and Brugerolle, G. (1992). Morphogenesis of the feeding apparatus of Entosiphon sulcatum: an immunofluorescence and ultrastructural study. *Protoplasma* 168:125–135.

Belhadri, A., Bayle, D. and Brugerolle, G. (1992). Biochemical and immunological characterization of intermicrotubular cement in the feeding apparatus of phagotrophic euglenoids: Entosiphon, Peranema, and Ploeotia. *Protoplasma*, **168:**113 -124.

Ben Shaul, Y., Schiff, J. A. and Epstein, H. T. (1964). Studies of chloroplast development in *Euglena.* VII. Fine structure of the developing plastid. *Plant Physioh,* **39:** 231- 240.

Benedetti , P. A. and Lenci, F. (1977). *In vivo* microspectrofluorometry of photoreceptor pigments in *Euglena gracilis. Photochem. Photobiol*., **26:**315 – 318

Benedetti, P.A., Bianchini, G., Checcuci , A., Ferrara, R., Grassi, S., Percival, D. (1976). Spectroscopic properties and related functions of the stigma measured in living cells of *Euglena gracilis. Arch. Microbiol.,* **111:**73–76

Bennett, M. S., Wiegert, K. E., Triemer, R. E. (2012). Comparative chloroplast genomics between Euglena viridis and Euglena gracilis (Euglenophyta). *Phycologia*, **51:**711–718.

Benning, C. and Ohta, H. (2005). Three enzyme systems for galactoglycerolipid biosynthesis are coordinately regulated in plants. *J Biol Chem.,* **280(4):**2397 – 2400.

Ben-Shaul, Y., Epstein, H. T., Schiff, J. A. (1965) Studies of chloroplast development in Euglena. The return of the chloroplast to the proplastid condition during dark adaptation. *Canadian Journal of Botany*, **43:** 129–136.

Bentolila, S., Heller, W.P., Sun, T., Babina, A.M., Friso, G., van Wijk, K.J., and Hanson, M.R. (2012). RIP1, a member of an*Arabidopsis* protein family, interacts with the protein RARE1and broadly affects RNA editing. *Proc. Natl Acad. Sci. U S A.,***109:**E1453–E1461.

Berget, S. M., Moore, C., and Sharp, P. A. (1977). Spliced segments at the 5´ terminus of adenovirus 2 late mRNA. *Proc. Natl Acad. Sci. USA,* **74:**3171–3175.

Berridge, M. J. (2014). Cell Signalling Biology; doi:10.1042/csb0001001

Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D. C., Lennard, N. J., Caler, E., Hamlin, N. E., Haas, B., Böhme, U., Hannick, L., Aslett, M. A., Shallom, J., Marcello, L., Hou, L., Wickstead, B., Alsmark, U. C., Arrowsmith, C., Atkin, R. J., Barron, A. J., Bringaud, F., Brooks, K., Carrington, M., Cherevach, I., Chillingworth, T. J., Churcher, C., Clark, L. N., Corton, C. H., Cronin, A., Davies, R. M., Doggett, J., Djikeng, A., Feldblyum, T., Field, M. C., Fraser, A., Goodhead, I., Hance, Z., Harper, D., Harris, B. R., Hauser, H., Hostetler, J., Ivens, A., Jagels, K., Johnson, D., Johnson, J., Jones, K., Kerhornou, A. X., Koo, H., Larke, N., Landfear, S., Larkin, C., Leech, V., Line, A., Lord, A., Macleod, A., Mooney, P. J., Moule, S., Martin, D. M., Morgan, G. W., Mungall, K., Norbertczak, H., Ormond, D., Pai, G., Peacock, C. S., Peterson, J., Quail, M. A., Rabbinowitsch, E., Rajandream, M. A., Reitter, C., Salzberg, S. L., Sanders. M., Schobel, S., Sharp, S., Simmonds. M., Simpson, A. J., Tallon, L., Turner, C. M., Tait, A., Tivey, A. R., Van Aken, S., Walker, D., Wanless, D., Wang, S., White, B., White, O., Whitehead, S., Woodward, J., Wortman, J., Adams, M. D., Embley, T. M., Gull, K., Ullu, E., Barry, J. D., Fairlamb, A. H., Opperdoes, F., Barrell, B. G., Donelson, J. E., Hall, N., Fraser, C. M., Melville, S. E., El-Sayed, N. M. (2005). The genome of the African trypanosome *Trypanosoma brucei*. *Science,* **309 (5733):**416 - 22.

Bertaux, O., Valencia, R. (1975). Aspects structuraux et macromoleculaires de l'avitaminose B12 ches Euglena en culture synchrone. *Colloq Int Cent Natl Rech Sci.,* **240:** 331-343

Bhasin, M., Reinherz, E. L. and Reche, P. A. (2006). "Recognition and classification of histones using support vector machine". *Journal of Computational Biology*, **13(1):** 102–12. *doi*:*10.1089/cmb.2006.13.102*. *PMID* *16472024*.

Biggins, S. (2013). The composition, functions, and regulation of the budding yeast kinetochore. *Genetics,* **194:**817–846.

Biitschli, O. (1884). The nervous Rndorgane at the probes and the Chilognathen their relationships to those of certain Insects. *Zrlbl. Biol.* **4:** 113-116

Biitschli, O. (1884). Thoughts on the morphological significance of the so-called polar bodies. *Ztrbl. Biol.,* **4:** 6-12.

Bikle, D., Tilney, L. G. and Porter, K. R. (1966). Microtubules and pigment migration in the melanophores of *Fundulus heteroclitus* L. *Protoplasma,* **61:**322-345.

Blackwell, T. K. and Walker, A. K. (2006). Transcription mechanisms. *WormBook*, ed. *The C. elegans Research Community, WormBook,* doi/10.1895/wormbook.1.121.1, http://www.wormbook.org

Blinks, L. R., in "Autotrophic Micro-organisms" (Fry, B. A. and Peel, J. L., eds.), **pp.** 224. Cambridge Univ. Press, 1954.

Blobel, G. and Dobberstein, B. (1975). Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma*. J Cell Bio.,* **67(3):** 835–51. *doi*:*10.1083/jcb.67.3.835*.

Blum, B., Bakalara, N., and Simpson, L. (1990). A model for RNA editing in kinetoplastid mitochondria: ''guide'' RNA molecules transcribed from maxicircle DNA provide the edited information. *Cell,* **60:**189–198.

Bodył, A., Mackiewicz, P. and Milanowski, R. (2010). Did trypanosomatid parasites contain an eukaryotic alga-derived plastid in their evolutionary past? *J. Parasitol.,* **96(2):**465–75.

Boehm, M. and Bonifacino, J. S. (2001). Adaptins: the final recount. *Molecular Biology of the Cell*, **12:**2907–2920.

Boesger, J., Wagner, V., Weisheit, W., Mittag, M. (2009). Analysis of flagellar phosphoproteins from *Chlamydomonas reinhardtii*. *Eukaryot Cell,* **8:**922–932.

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics; Genome Analysis,* **27(4):**578–579. *doi:10.1093/bioinformatics/btq683*.

Bokov, K. and Steinberg, S. V. (2009). A hierarchical model for evolution of 23S ribosomal RNA. Nature, **457:**977–980.

Bolger, A. M., Lohse, M., and Usade, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, **pp** 1–7. doi:10.1093/bioinformatics/btu170 http://hannonlab.cshl.edu/fastx_toolkit/

Bolige, A., Hagiwara, S. Y., Zhang, Y., Goto, K., (2005a). Circadian G2 arrest as related to circadian gating of cell population growth in Euglena. *Plant Cell Physiol*., **46:**931–936.

Bolige, A., Kiyota, M., Goto, K. (2005b). Circadian rhythms of resistance to UV-C and UV-B radiation in Euglena as related to ‘‘escape from light’’ and ‘‘resistance to light’’. *J. Photochem. Photobiol. B.,* **81:**43 – 54.

Bolte, K., Bullmann, L., Hempel, F., Bozarth, A., Zauner, S. and Maier, U. (2009). Protein Targeting Into Secondary Plastids. *J. Eukaryot. Microbiol.,* **56(1):** 9–15.

Bonen, L. (1993). Trans-splicing of pre-mRNA in plants, animals, and protists. *FASEB J.,* **7:**40–46.

Bonham-Smith, P. C. and Bourque, D. P. (1989). Translation of chloroplast-encoded mRNA: potential initiation and termination signals. *Nucleic Acids Res.,* **17:**2057 – 2080.

Bonifacino, J.S. and Lippincott-Schwartz, J. (2003). Coat proteins: shaping membrane transport. *Nature Reviews Molecular Cell Biology,* **4:**409–414.

Bonora, A., Pancaldi, S., Gualandri, R., Fasulo, M. P. (2000). Carotenoid and ultrastructure variations in plastids of Arum italicum Miller fruit during maturation and ripening. *Journal of Experimental Botany*, **51:** 873 – 884.

Borst, P. and Van Leeuwen, F. (1997). beta-D-glucosyl-hydroxymethyluracil, a novel base in African trypanosomes and other Kinetoplastida. *Mol. Biochem. Parasitol.,* **90:**1–8.

Boto, L. (2014). Horizontal gene transfer in the acquisition of novel traits by metazoans. *Proc. Biol. Sci.,* **281:**20132450.

Bouck, G. B., Rogalski, A. (1978). Valaitis Surface organization and composition of *Euglena*. II Flagellar mastigonemes. *Cell Biol.,* **77:** 805–826

Bouck, G. B., Rosiere T.K., Levasseur, P. J. (1990). *Euglena gracilis*: A model for flagellar surface assembly, with reference to other cells that bear flagellar mastigonemes scales. R. A. Bloodgood (Ed.), *Ciliary* and *Flagellar Membrane*, Plenum Press, New York, **pp.** 65 – 90.

Bouck, G. B. and Ngo, H. (1996). Cortical structure and function in euglenoids with reference to trypanosomes, ciliates, and dinoflagellates. *Int. Rev. Cytol.,* **169**:267–318.

Bouck, G. B. and Ngô, H. (1996). Cortical structure and function in euglenoids with reference to trypanosomes, ciliates, and dinoflagellates. *Int. Rev. Cytol.* **169:**267-318.

Bouck, G. B., Green, P. M. (1976). Paracrystals and mastigonemes are directly attached to the *Euglena* axonemal microtubules. *J. Cell. Biol.,* **70:** 156a

Bouck, G. B., Rogalski, A., Valaitis, A. (1978). Surface organization and composition of Euglena. II. Flagellar mastigonemes. *J Cell Biol.,* **77(3):**805-26.

Bourelly, P. (1985). *Les algues d'eaux douce. Tom III: Les algues bleues et rouges. Les Eugleniens, Peridiniens et Cryptomonadines*. *N.Boube e et Cie, Paris* (In French).

Bovarnick, J. G., Chang, S. W., Schiff, J. A. and Schwartzbach, S. D. (1974). Events surrounding the early development of Euglena chloroplasts: experiments with streptomycin in non-dividing cells. *Journal of General Microbiology,* **83:**5I - 62.

Bovarnick, J. G., Schiff, J . A., Freedman, Z. and Jun, J. M. E. (1974). Events Surrounding the Early Development of Euglena Chloroplasts: Cellular

Origins of Chloroplast Enzymes in Euglena. *Journal of General Microbiology,* **83: 63-71.**

Bovee, E.C., Jahn, T. L. (1972). A theory of piezoelectric activity and ion-movements in the relation of flagellar structures and their movements to the phototaxis of *Euglena. J. theor. Biol.,* **35:**259–273

Bowman, L. H., Goldman, W. E., Goldberg, G. I., Herbert, M. B., Schlessinger, D. (1983). Location of the initial cleavage sites in mouse pre-rRNA. *Mol. Cell. Biol.* 3(8):1501-1510.

Bowman, L. H., Rabin, B., Schlessinger, D. (1981). Multiple ribosomal RNA cleavage pathways in mammalian Cells. *Nucleic Acids Res.,* **9:**4951-4966.

Bradbeecr and Woodrowm, L. (1976). Transport of vitamin $B_{12}$ i n *Escherichia coli:* energy dependence. *Journal of Bacteriology,* **128(1):**99-104.

Bradbeer, C. (1971). Transport of vitamin $B_{12}$ in *Ochromunas malhamensis. Archives of Biochemistry and Biophysics,* **144:** 184-197.

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J. A., Chapuis, G., Chikhi, R., Chitsaz, H., Chou, WC, Corbeil, J., Del Fabbro, C., Docking, T. R., Durbin, R., Earl, D., Emrich, S., Fedotov, P., Fonseca, N. A., Ganapathy, G., Gibbs, R. A., Gnerre, S., Godzaridis, É., Goldstein, S., Haimel, M., Hall, G., Haussler, D., Hiatt, J. B., Ho, I. Y., Howard, J., Hunt, M., Jackman, S. D., Jaffe, D. B., Jarvis, E. D., Jiang, H., Kazakov, S., Kersey, P. J., Kitzman, J. O., Knight, J. R., Koren, S., Lam, TW, Lavenier, D., Laviolette, F., Li, Y., Li, Z., Liu, B., Liu, Y., Luo, R., MacCallum, I., MacManes, M. D., Maillet, N., Melnikov, S., Naquin, D., Ning, Z., Otto, T. D., Paten, B., Paulo, O. S, Phillippy, A. M., Pina-Martins, F., Place, M., Przybylski, D., Qin, X., Qu, C., Ribeiro, F. J., Richards, S., Rokhsar, D. S., Ruby, J. G., Scalabrin, S., Schatz, M. C., Schwartz, D. C., Sergushichev, A., Sharpe, T., Shaw, T. I., Shendure, J., Shi, Y., Simpson, J. T., Song, H., Tsarev, F., Vezzi, F., Vicedomini, R., Vieira, B. M., Wang, J., Worley, K. C., Yin, S., Yiu, SM, Yuan, J., Zhang, G.,

Zhang, H., Zhou, S. and Korf, I. F. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience,* **2:**10 doi:10.1186/2047-217X-2-10

Bradnam, K. R., Fass, J. N., Korf, I. F. (2013). CEGMA gene predictions for Assemblathon 2 entries. GigaScience Database: http://dx.doi.org/10.5524/100061.

Braun, L., Cannella, D., Ortet, P., Barakat, M., Sautel, C. F., Kieffer, S., Garin, J., Bastien, O., Voinnet, O., Hakimi, M. (2010). A complex small RNA repertoire is generated by a plant/fungal-like machinery and effected by a metazoan-like Argonaute in the single-cell human parasite Toxoplasma gondii. *PLoS Pathog.,* **6:**e1000920.

Bray, D. (1979). Cytochalasin action. *Nature*, **282:** 671.

Bre, M. H., El Ferjani, E. and Lefort-Tran, M., (1981). Sequential protein-dependent steps in the cell cycle. Initiation and completion of division in vitamin $B_{12}$ replenished *Euglena gracilis. Protoplasma*, **108:** 301-318.

Breckenridge, D. G., Watanabe, Y-I., Greenwood, S. J., Gray, M. W., and Schnare, M. N. (1999). U1 Small Nuclear Rna And Spliceosomal Introns in *Euglena gracilis. Proc. Natl. Acad. Sci. USA,* **96:**852–856.

Breglia, S. A., Slamovits, C. H., Leander, B. S. (2007). Phylogeny of phagotrophic euglenids (Euglenozoa) as inferred from hsp90 gene sequences. *J Eukaryot Microbiol*., **54:**86–92.

Bricheux, G. and Brugerolle, G. (1986). The membrane cytoskeleton complex of euglenoids. I. Biochemical and immunological characterization of the epiplasmic proteins of *Euglena acus*. *Europ. J. Cell Biol.,* **40**:150–159.

Brighouse A., Dacks, J. B. and Field, M. C. (2010). Rab protein evolution and thehistory of the eukaryotic endomembrane system. *Cell Mol Life Sci.,* **67(20):**3449 – 3465.

Brighouse, A., Dacks, J. B., and Field, M. C. (2010). Rab protein evolution and the history of the eukaryotic endomembrane system. *Cell Mol. Life Sci.,* **67(20):** 3449–3465.

Brighouse, A., Dacks, J. B., Field, M. C. (2010). Rab protein evolution and the history of the eukaryotic endomembrane system. Cell Mol Life Sci., **67:**3449 – 3465.

Bright, L. J., Kambesis, N., Nelson, S. B., Jeong, B. and Turkewitz, A. P. (2010). Comprehensive analysis reveals dynamic and evolutionary plasticity of Rab GTPases and membrane traffic in Tetrahymena thermophila. *PLoS Genet.,* **6:** e1001155.

Brocker, C., Engelbrecht-Vandre, S., Ungermann, C. (2010). Multisubunit tethering complexes and their role in membrane fusion. *Current Biology,* **20(21):** 943-952. Doi: 10.1016/j.cub.2010.09.015.

Brodhun, B. and Hader, D. P. (1990). Photoreceptor protein and pigments in the paraflagellar body of the flagellate *Euglena gracilis. Photochem. Photobiol.,* **52:** 865–871

Brodie, A. F. (1965). In: Biochemistry of Quinones, p.356. Ed. By Morton, R. A. New York and London: Academic Press Inc.

Bronstrup, U., Hachtel, W. (1989). Cytochrome c oxidase of *Euglena gracilis*: purification, characterization, and identification of mitochondrially synthesized subunits. *J. Bioenerg. Biomembr*., **21:**359–373.

Brown, J. W. S., Clark, G. P., Leader, D. J., Simpson, C. G. and Lowe, T. (2001). Multiple snoRNA gene clusters from Arabidopsis. *RNA,* **7:**1817–1832.

Brown, R. D. and Haselkorn, R. (1971a). Chloroplast RNA Populations in Dark-Grown, Light-Grown, and Greening *Euglena gracilis*. *Proc. Natn. Acad. Sci. USA,* **68:**2536 -9.

Brown, R. D. and Haselkorn, R. (1971b). Synthesis and maturation of cytoplasmic trna in *Euglena gracilis*. *J. Mol. Biol.,* **59:**491 – 503.

Brown, R. W. B., Collingridge, P. W., Gull, K., Rigden, D. J., and Ginger, M. L. (2014). Evidence for Loss of a Partial Flagellar Glycolytic Pathway during Trypanosomatid Evolution. *PLoS ONE*, **9(7):** e103026. doi:10.1371/journal.pone.0103026.

Brown, S. S. and Spudich, J. A. (1979). Cytochalasin inhibits the rate of elongation of actin filament fragments. *J. Cell Biol.,* **83:** 657 – 662.

Brugerolle, G. (1992). Flagellar apparatus duplication and partition, flagellar transformation during division in Entosiphon sulcatum. *BioSystems*, **28:**203–209.

Brugerolle, G. (1993). Evolution and diversity of amitochondrial zooflagellates. *J. Eukaryot. Microbiol.,* **40**:616 – 618.

Brugerolle, G., and Mignot, J. P. (2003). The rhizoplast of chrysomonads, a basal body-nucleus connector that polarises the dividing spindle. *Protoplasma,* **222**:13–21.

Brugerolle, G., Bricheux, G. and Coffe, G. (2000). Centrin protein and genes in *Trichomonas vaginalis* and close relatives. *J. Eukaryot. Microbiol.,* **47:**129–138.

Brunold, C. and Schiff, J. A. (1976). Studies of sulfate utilization of algae: 15. Enzymes of assimilatory sulfate reduction in Euglena and their cellular localization. *Plant Physiol.,* **57:** 430-436.

Buell, G., Schulz, M.F., Arkinstall, S.J., Maury, K., Missotten, M., Adami, N., Talabot, F., Kawashima, E. (1992). Molecular characterisation, expression and localisation of human neurokinin-3 receptor. *FEBS Lett.,* **299 (1):**90–95.

Buell, G., Schulz, M.F., Arkinstall, S.J., Maury, K., Missotten, M., Adami, N., Talabot, F., and Kawashima, E. (1992). Molecular characterisation, expression and localisation of human neurokinin-3 receptor. *FEBS Lett.,* **299(1):** 90–95.

Buetow, D. E. (1968b). The Biology of *Euglena*, Vol. II Academic Press, New York

Buetow, D. E. (1968). Morphology and ultrastructure *of Euglena.* In: Buetow D-E (ed) The Biology of *Euglena. Vol. I, General Biology and ultrastructure. Academic Press, New York, London,* **pp.** 109-184

Buetow, D. E. (1968a). The Biology of *Euglena*, Vol. I Academic Press, New York

Buetow, D. E. (1982). The Biology of *Euglena*, Vol. III Academic Press, New York.

Buetow, D. E. (1982). The Biology of *Euglena*, Vol. III Academic Press, New York

Buetow, D. E. (1989). The Biology of *Euglena*, Vol. IV Academic Press, New York (1989)

Buetow, D. E. (1989). The mitochondrion, p 247–314. *In* Buetow DE (ed), *The biology of Euglena, vol 4. Academic Press,* San Diego, CA.

Buetow, D. E. (1989). The Mitochondrion. In Buetow DE (ed) The Biology of Euglena, Vol. IV. Academic Press, San Diego, **pp** 247 – 314.

Buetow, D. E. (1989). The mitochondrion. In: Buetow DE (ed) The biology of *Euglena* vol IV. *Academic Press Inc, San Diego,* **pp**. 247–314.

Buetow, D. E. (2011). Euglena. In: eLS. *John Wiley & Sons Ltd*, Chichester.

Buetow, D. E., ed. Bouck, G. B., Kempner, E. S., Bovee, E. C., Leedale, G. F., Colombetti, G., Lefort-Tran, M., Diehn, B., Lenci, F., Dubertret, G., Schiff, J. A., *et al.* (1982). The Biology of *Euglena*. Academic Press, New York.

Bugreev, D. V., Pezza, R. J., Mazina, O. M., Voloshin, O. N., Camerini-Otero, R. D., Mazin, A. V. (2011). The resistance of DMC1 D-loops to dissociation may account for the DMC1 requirement in meiosis. *Nat Struct MolBiol.,* **18:**56 – 61.

Buhrman, H., van der Guliky P. T. S., Severiniz, S. and Speijer, D. (2013). A mathematical model of kinetoplastid mitochondrial gene scrambling advantage. .arXiv:1307.1163v1 [q-bio.PE].

Bumbulis, M., J. and Balog, B M. (2013). UV-C Exposure Induces an Apoptosis-Like Process in *Euglena gracilis. ISRN Cell Biology* , Article ID **869216,** 6 pages.

Bunning, E. (1960). Circadian rhythms and the time measurement in photoperiodism. *Cold Spring Harb. Symp. Quant. Biol.,* **25:** 249 – 256.

Burger, G., Lang, B. F., Reith, M., and Gray, M. W. (1996). Genes encoding the same three subunits of respiratory complex II are present in the mitochondrial DNA of two phylogenetically distant eukaryotes. *Proc. Natl. Acad. Sci. U S A.***93:**2328–2332.

Burki, F. and Keeling, P. J. (2014). Rhizaria. *Current Biology,* **24(3):**R103-R107.

Burri, L. and Keeling, P. J. (2007). Protein targeting in parasites with cryptic mitochondria. *International Journal for Parasitology*, **37:** 265–272.

Burri, L., Keeling P. J. (2007). Protein targeting in parasites with cryptic mitochondria. *International Journal for Parasitology,* **37:**265–272. doi:10.1016/j.ijpara.2006.12.004

Burton, P. R. and Fernandez, H. L. (1973). Delineation by lanthanum staining of filamentous elements associated with the surfaces of axonal microtubules. *J. Cell Sci.,* **12:**567-583.

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.

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

Busse, I. and Preisfeld, A. (2002b). Unusually expanded SSU ribosomal DNA of primary osmotrophic euglenids: molecular evolution and phylogenetic inference. *J. Mol. Evol.,* **55,** 757–767.

Busse, I., Patterson, D. J., Preisfeld, A. (2003). Phylogeny of phagotrophic euglenids (‘‘Euglenozoa’’): a molecular approach based on culture material and environmental samples. *J. Phycol.,* **39:** 828 – 836.

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.,* **25:**1915–1927.

Cachon, J., Cachon, M., Cosson, M. P., and Cosson, J. (1988). The Paraflagellar Rod: a structure in search of a function. *Biol. Cell,* **63:** 169–181

Cai, H., Brown, J. D., Deng, C. and Oakes, M. A. (2007). Self-esteem and culture: Differences in cognitive self-evaluations or affective self-regard? *Asian Journal of Social Psychology,* **10:162**–170.

Cai, H., Reinisch, K., Ferro-Novick, S. (2007). Coats, tethers, Rabs, and SNAREs work together to mediate the intracellular destination of a transport vesicle. Dev. Cell, **12:**671–682.

Calvayrac, R. and Lefort-Tran, M. (1976). Organisation spatiale dea chloroplastes chez 1' *Euglena* a l'aide de coupes strides semi-fines. *Protoplasma,* **89:** 353-358.

Calvayrac, R., Bomsel, J. L., Laval-Martin, D. (1979) Analysis and characterization of DCMU-resistant *Euglena.* I - Growth, metabolic and ultrastructural modifications during adaptation to different doses of DCMU. *Plant Physiol.,* **63:**857-865.

Calvayrac, R., Laval-Martin, D., Briand, j., and Farineau, J. (1981). Paramylon synthesis by *Euglena gracilis* photoheterotrophically grown under low $O_2$ pressure: Description of a mitochloroplast complex. *Planta,* **153:** 6 – 13.

Calvayrac, R., Laval-Martin, D., Briand, J., and Farineau, J. (1981). Paramylon synthesis by *Euglena gracilis* photoheterotrophically grown under low O2 pressure: description of a mitochloroplast complex. *Planta,* **153:**6–13. doi: 10.1007/BF00385311

Canaday, J., Tessier, L. H., Imbault, P., and Paulus, F. (2001). Analysis of Euglena gracilis alpha-, beta- and gamma-tubulin genes: introns and premRNA maturation. *Mol. Genet. Genomics,* **265:**153 – 160.

Cann, J. P., Penning, N. C. (1986). Observations on *Petalomonas cantuscyfni*, n.sp., a new halo-tolerant strain*. Arch. Protistenkd.,* **132:** 63–71

Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., Henrissat, B. (2009). *Nucleic Acids Res.,* **37:**D233–D238.

Capella-Gutierrez, S., Silla-Martinez, J. M., Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25:** 1972-1973.

Caragea, C., Sinapov, J., Silvescu, A., Dobbs, D. and Honavar, V. (2007). Glycosylation site prediction using ensembles of Support Vector Machine classifiers. *BMC Bioinformatics,* **8:**438 doi:10.1186/1471-2105-8-438.

Carell, E. F (1969). Studies on chloroplast development and replication in Euglena. *J Cell Biol.,* **41:** 431-440.

Carell, E. F (1975). Modification by vitamin B12 of the cell cycle in Euglena: studies on the induction and recovery from B12 deficiency. *Colloq. Int. Cent. Natl. Rech. Sci.,* **240:** 321-329.

Carell, E. F. and Seeger, J. W., Jr. (1980). Ribonucleotide Reductase Activity In Vitamin B12-Deficient *Euglena gracilis*. *Biochem. J.,* **188:**573-576.

Carnes, J., Trotter, J.R., Ernst, N.L., Steinberg, A., Stuart, K. (2005). An essential RNase III insertion editing endonuclease in *Trypanosoma brucei*. *Proc. Natl. Acad. Sci. USA,* **102:**16614–16619.

Carninci, P., Waki, K., Shiraki, T., Konno, H., Shibata, K., Itoh, M., Aizawa, K., Arakawa, T., Ishii, Y., Sasaki, D., Bono, H., Kondo, S., Sugahara, Y., Saito, R., Osato, N., Fukuda, S., Sato, K., Watahiki, A., Hirozane-Kishikawa, T., Nakamura, M., Shibata, Y., Yasunishi, A., Kikuchi, N., Yoshiki, A., Kusakabe, M., Gustincich, S., Beisel, K., Pavan, W., Aidinis, V.,

Nakagawara, A., Held, W. A., Iwata, H., Kono, T., Nakauchi, H., Lyons, P., Wells, C., Hume, D. A., Fagiolini, M., Hensch, T. K., Brinkmeier, M., Camper, S., Hirota, J., Mombaerts, P., Muramatsu, M., Okazaki, Y., Kawai, J., Hayashizaki, Y. (2003). Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. *Genome Res.,* **13:**1273–1289.

Carpenter, B. G., Baldwin, J. R., Bradbury, M., Ibel, K. (1976). Organization of subunits in chromatin. *Nucleic Acids Res,* **3:**1739-1746

Carter, M.S., and Krause, J. E. (1990). Structure, expression, and some regulatory mechanisms of the rat preprotachykinin gene encoding substance P, neurokinin A, neuropeptide K, and neuropeptide gamma. *J. Neurosci.*, **10(7):** 2203–2214.

Carter, S. B. (1967). Effects of cytochalasin on mammalian cells. *Nature,* **213:** 26l-264.

Casella, J. F., Flanagan, M. D., Lin, S. (1981). Cytochalasin D inhibits actin polymerization and induces depolymerization of actin filaments formed during plateled shape change. *Nature,* **293:** 302- 304.

Castro-Guerrero, N. A., Jasso-Chavez, R., Moreno-Sanchez, R. (2005). Physiological role of rhodoquinone in *Euglena gracilis* mitochondria. *Biochim. Biophys. Acta,* **1710**:113–121.

Castro-Guerrero, N. A., Krab, K., Moreno-Sanchez, R. (2004). The alternative respiratory pathway of *Euglena* mitochondria. *J. Bioenerg. Biomembr.,* **36**:459–469.

Cavalier-Smith, T. (1981). "Eukaryote kingdoms: Seven or nine?" *Bio Systems,* **14 (3–4):** 461–81.

Cavalier-Smith, T. (1987a). The origin of cells: a symbiosis between genes, catalysts, and membranes. *Cold Spring Harb Symp Quant Biol,* **52**: 805 - 824.

Cavalier-Smith, T. (1987b). Bacterial DNA segregation: its motors and positional control. *J Theor Biol,* **127**: 361 - 372.

Cavalier-Smith, T. (1987c). The origin of eukaryote and archaebacterial cells. *Ann N Y Acad Sci,* **503**: 17 - 54.

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

Cavalier-Smith, T. (1987e). The simultaneous symbiotic origin of mitochondria, chloroplasts, and microbodies. *Ann NY Acad Sci,* **503**: 55 - 71.

Cavalier-Smith, T. (1987f). Glaucophyceae and the origin of plants. *Evol. Trends Plants,* **2**: 75 - 78.

Cavalier-Smith, T. (1987g). The origin of Fungi and pseudofungi. In *Evolutionary Biology of the Fungi* (Symposium of the British Mycological Society no. 13), *pp.* 339 - 353. Edited by A. D. M. Rayner, C. M. Brasier & D. Moore. Cambridge: Cambridge University Press.

Cavalier-Smith, T. (1993). "Kingdom protozoa and its 18 phyla". *Microbiological reviews* **57 (4):** 953 – 94.

Cavalier-Smith, T. (1993). "The Protozoan Phylum Opalozoa". *The Journal of Eukaryotic Microbiology,* **40 (5):** 609–15.

Cavalier-Smith, T. (1995). Membrane heredity, symbiogenesis, and the multiple origins of algae, in: Arai R., Kato M., Doi Y. (Eds.), *Biodiversity and Evolution*, The National Science Museum Foundation, Tokyo **pp.** 69 - 107.

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

Cavalier-Smith, T. (1999). Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. *J Eukaryot Microbiol.* **46:**347–366.

Cavalier-Smith, T. (2002). The phagotrophic origin of eukaryotes and phylogenetic classification of protozoa. *Int. J. Syst. Evol. Biol.,* **52:**297–354.

Cavalier-Smith, T. (2003). "Protist phylogeny and the high-level classification of Protozoa". *European Journal of Protistology*, **39 (4):** 338. doi:10.1078/0932-4739-00002

Cavalier-Smith, T. (2010). .Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree. *Biol. Lett.* **6:**342–345.

Cech, T. R. (1990). Self-splicing of group I introns. *Annual Review of Biochemistry*, **59:** 543–568.

Cerutti, H., and Casas-Mollano, J. A. (2006). On the origin and functions of RNA mediated silencing: from protists to man. *Curr. Genet.,* **50:** 81–99.

Cervantes, C., Campos-Garcia, J., Devars, S., Gutiérrez-Corona, F., Loza-Tavera, H., Torres-Guzman, J. C., Moreno-Sánchez, R. (2001). Interactions of chromium with microorganisms and plants, *FEMS Microbiol. Rev.,* **25:**335–347.

Cetkovic, H., Mikoc, A., Muller, W. E. G., Gamulin, V. (2007). Ras-like Small GTPases Form a Large Family of Proteins in the Marine Sponge Suberites domuncula. *J. Mol. Evol.,* **64:**332 – 341.

Cetkovic, H., Mikoc, A., Müller, W. E. G. and Gamulin, V. (2007). Ras-like Small GTPases Form a Large Family of Proteins in the Marine Sponge *Suberites domuncula. Journal of Molecular Evolution*, **64(3):** 332 - 341.

Chaal, B. K. and Green, B. R. (2005). Protein import pathways in ‘complex’ chloroplasts derived from secondary endosymbiosis involving a red algal ancestor. *Plant Mol. Biol.,* **57:**333–342.

Chadefaud, M. (1940). Phénomènes sexuels chez les Euglènes. *Rev. Sci.,* **78:** 179-180.

Chan, S. H., Heckers, L. I., Brum, C. K., Schnabel, J.J., Heckmant, J.E., Silberklangt, M., Raj, U. L. Bhandary, T. and Barnett, W. E. (1981). The nucleotide sequence of Euglena cytoplasnic phenylabnine transfer RNA. Evidence for possible classification of Euglena among the animal rather than the plant kingdom. *Nucleic Acids Research*, **9(13):** 1981

Charette, J. M., Gray, M. W. (2009). U3 snoRNA genes are multi-copy and frequently linked to U5 snRNA genes in *Euglena gracilis*. *BMC Genomics*, **10:**528.

Charette, M. and Gray, M. W. (2000). Pseudouridine in RNA: what, where, how, and why. *IUBMB Life,* **49:**341–351.

Chasey, D. (1974). The three-dimensional arrangement of radial spokes in the flagella of *Chlamydomonas Reinhardii*. *Exp. Cell Res.,* **84:** 374–380

Chaudhuri, M., Ott, R. D., Hill, G. C. (2006). Trypanosome alternative oxidase: from molecule to function. *Trends Parasitol.,* **22:**484 – 491.

Chauhan, J. S., Rao, A. and Raghava, G. P. S. (2013). In silico Platform for Prediction of N-, O- and C-Glycosites in Eukaryotic Protein Sequences. *PLoS ONE,* **8(6):** e67008. doi:10.1371/journal.pone.0067008

Chauhan, J. S., Rao, A., and Raghava, G. P. S. (2013). In silico Platform for Prediction of N-, O- and C-Glycosites in Eukaryotic Protein Sequences. *PLoS ONE,* **8(6):**e67008. doi:10.1371/journal.pone.0067008.

Chavez, L. A., W. Balamuth, and T. Gong. (1986). A light and electron microscopical study of a new, polymorphic free-living amoeba, *Phreatamoeba balamuthi* n.g., n.*sp. J. Protozool.,* **33**:397–404.

Cheeseman, I. M. and Desai, A. (2008). Molecular architecture of the kinetochore-microtubule interface. *Nat. Rev.Mol. Cell Biol.,* **9:**33–46.

Chelm, B. K., Hoben, P. J. and Hallick, R. B. (1977). Cellular content of chloroplast DNA and chloroplast ribosomal RNA genes in *Euglena gracilis* during chloroplast development. *Biochemistry*, **16:**782–786.

Chen, C. L., Liang, D., Zhou, H., Zhuo, M., Chen, Y. Q. and Qu, L. H. (2003). The high diversity of snoRNAs in plants: identification and comparative study of 120 snoRNA genes from Oryza sativa. *Nucleic Acids Res.,* **31:**2601–2613.

Chen, C.-L., Chen, C.-J., Vallon, O., Huang, Z.-P. (2008). Genomewide analysis of box C/D and box H/ACA snoRNAs in Chlamydomonas reinhardtii reveals

an extensive organization into intronic gene clusters. *Genetics,* **179(1):**21-30. doi: 10.1534/genetics.107.086025.

Chen, Y. K., Leng, C. H., Olivares, H, Lee, M. H., Chang, Y. C., Kung, W. M., Ti, S. C., Lo, Y. H., Wang, A. H., Chang, C. S., Bisho, D. K., Hseuh, Y. P. and Wang, T. F. (2004). Heterodimeric complexes of Hop2 and Mnd1 function with Dmc1 to promote meiotic homolog juxtaposition and strand assimilation. *Proc Natl Acad Sci U S A.*, **101:** 10572–10577.

Cheung A.Y., Mcnellis T., Piekos B. (1993). Maintenance Of Chloroplast Components During Chromoplast Differentiation In Tomato Mutant GREEN FLESH. *Plant Physiology*, **101:** 1223–1229.

Chi, J., Mahe, F., Loidl, J., Logsdon, J., and Dunthorn, M. (2013). Meiosis Gene Inventory of Four Ciliates Reveals the Prevalence of a Synaptonemal Complex-Independent Crossover Pathway. *Mol. Biol. Evo.,* **31(3):**660–672 doi:10.1093/molbev/mst258.

Cholnoky, L., Gyorgyfy, C., Nagy, E., and Pancziel, M., (1956). Function of carotenoids in chlorophyll containing organs. *Nature,* **178:** 410-11.

Chow, L. T., Gelinas, R. E., Broker, T. R., and Roberts, R. J. (1977). An amazing sequence arrangement at the 5´ ends of adenovirus 2 messenger RNA. *Cell,* **12:**1–8.

Christian Frantz, Chantal Ebel, Francois Paulus, Patrice Imbault (2000). Characterization of trans-splicing in Euglenoids. *Curr Genet,* **37:**349 – 355.

Ciska, M., Masuda, K. and de la Espina, M. D. S. (2013). Lamin-like analogues in plants: the characterization of NMCP1 in *Allium cepa. J Exp Bot.,* **64:**1553-1564.

Clarkson, B. K., Gilbert, W. V. and Doudna, J. A. (2010). Functional overlap between eIF4G isoforms in *Saccharomyces cerevisiae*. *PLoS One,* **5:**e9114. http://dx.doi.org/10.1371/journal.pone.0009114.

Clayton Christine, Usler Thomas Ha, and Blattner Judith (1995). Protein Trafficking in Kinetoplastid Protozoa. *Microbiological Reviews*, **p.** 325–344

Clayton, C. E., and P. Michels (1996). Metabolic compartmentation in African trypanosomes. *Parasitol. Today,* **12**:465–471.

Clayton, C., and Estevez, A. (2010). The exosomes of trypanosomes and other protists. In: RNA Exosome. Jensen, T. (ed.). Austin, TX: Landes Bioscience, **pp.** 39–49.

Cloonan, N., Forrest, A. R. R., Kolle, G., Gardiner, B. B. A., Faulkner, G. J., Brown, M. K., Taylor, D. F., Steptoe, A. L., Wani, S., Bethel, G., Robertson, A. J., Perkins, A. C., Bruce, S. J., Lee, C. C., Ranade, S. S., Peckham, H. E., Manning, J. M., McKernan, K. J., & Grimmond, S. M. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods,* **5:**613 – 619.

Coakley, G., Maizels, R. M., and Buck, A. H. (2015). Exosomes and Other Extracellular Vesicles: The New Communicators in Parasite Infections. Trends in Parasitology, **31(10):** 477 – 480.

Cocquyt, E., Verbruggen, H., Leliaert, F., Zechman, F. W., Sabbe, K., and De Clerck, O.(2009). Gain and loss of elongation factor genes in green algae. *BMC Evol. Biol.,* **9:**39.

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry, C. E., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M. A., Rajandream, M. A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J. E., Taylor, K., Whitehead, S., Barrell, B. G. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature,* **393:**537– 544.

Coley, A. F., Dodson, H. C., Morris, M. T., and Morris, J. C. (2011). Glycolysis in the African Trypanosome: Targeting Enzymes and Their Subcellular Compartments for Therapeutic Development. *Molecular Biology International*, **(ID 123702):** 10 pages doi:10.4061/2011/123702.

Collins, J. E., White, S., Searle, S. M. and Stemple, D. L. (2012). Incorporating RNA-seq data into the zebrafish Ensembl genebuild. *Genome Res.,* **22:**2067–2078.

Colombetti , G., Petracchi, D. (1989). Photoresponse mechanisms in flagellated algae. *CRC Crit. Rev. Plant Sci.,* **8:** 309–355

Comings, D. E., (1967). Histones of genetically active and inactive chromatin. *J cell boil.,* **35(3):**699–708

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., and Robles, M (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics,* **21(18):**3674–3676.

Conforti, V. and Tell, G. (1989). Ultrastructure of the pellicle and the envelope of some euglenoid flagellates from Argentina by means of S.E.M. Nova Hedwigia, **48:** 187-206.

Contreras, V., Richardson, M. A., Hao, E. and Keiper, B. D. (2008). Depletion of the cap-associated isoform of translation factor eIF4G induces germline apoptosis in *C. elegans*. *Cell Death Differ.,* **15:**1232–1242. http://dx.doi.org/10.1038/cdd.2008.46.

Cook, J. R. and Roxby, R. (1985). Physical properties of a plasmid-like DNA from *Euglena gracilis. Biochemica et Biophvsica Acta,* **824:** 80 83.

Cook, J. R. (1972). Ultraviolet inactivation of *Euglena*chloroplasts. I. Effect of light intensity of culture. *Biophys. J . ,* **12:** 1467-1473.

Cook, J. R. (1973). Unbalanced growth and replication of chloroplast populations in *Euglena gracilis. J. gen. Microbiol.,* **75:** 51 – 60

COOK, J. R. and Li, T. C. C. (1973). Influence of culture pH on chloroplast structure in *Euglena gracilis. J. Protozool.,* **20:** 652-653.

Cook, J. R., Haggard, S. S., and Harris, P. (1976). Cyclic changes in chloroplast structure in synchronized *Euglena gracilis. J. Protozool.* **23:** 368-373

Cooper Geoffrey, M. (2000). The Cell: A Molecular Approach. 2nd edition. Sunderland (MA): Sinauer Associates; ISBN-10: 0-87893-106-6.

Copeland, M. (1974). The cellular responses to cytochalasin B. A critical review. *Cytologia*. **39:** 709- 727.

Copertino, D. W., Hallick, R. B. (1993). Group II and group III introns of twintrons: potential relationships with nuclear pre-mRNA introns. *Trends in Biochemical Sciences,* **18:** 467–471.

Corliss, J. O. (1984). The kingdom Protista and its 45 phyla. *BioSystems,* **17:**87-126.

Couvillion, M. T., Lee, S. R., Hogstad, B., Malone, C. D., Tonkin, L. A., Sachidanandam, R., Hannon, G. J., and Collins K. (2009). Sequence, biogenesis, and function of diverse small RNA classes bound to the Piwi family proteins of Tetrahymena thermophila. *Genes Dev.,* **23:**2016–32.

Covello, P. S., Gray, M. W. (1993). On the evolution of RNA editing. *Trends Genet.* **9:**265 – 268 .

Cox, M., Nelson, D. R. and Lehninger, A. L.(2005).Lehninger Principles of Biochemistry. *San Francisco: W.H. Freeman. ISBN 0-7167-4339-6*.

Creutz, C., Diehn, B. (1976). Motor response to polarized light and gravity sensing in *Euglena gracilis. J. Protozool.,* **23:** 552 – 560.

Croan, D. G., Morrison, D. A., Ellis, J. T. (1997). Evolution of the genus Leishmania revealedby comparison of DNA and RNA polymerase gene sequences. *Mol BiochemParasitol.,* **89:**149–59.

Cronshaw, J. M., Krutchinsky, A. N., Zhang, W., Chait, B. T. and Matunis, M. J. (2002). Proteomic analysis of the mammaliannuclear pore complex. *J Cell Biol.,* **158(5):**915–27.

Crooks, G., Hon, G. (2004). WebLogo: a sequence logo generator. *Genome Res.* **14:**1188–1190.

Crouse, E. J., Vandrey, J. P., Stutz, E. (1974). Hybridization studies with RNA and DNA isolated from Euglena gracilis chloroplasts and mitochondria. *FEBS Lett.,* **42:**262–266.

Cull,B., Godinho,J. L. P., Rodrigues,J. C. F., Frank,B., Schurigt,U., Williams,R. A. M., Coombs,G. H., and Mottram,J. C. (2014). Glycosome turnover in Leishmania major is mediated by autophagy. *Autophagy,* **10(12):**2143 - 2157.

Cunningham, F. X. and Schiff, J. A. (1986). Chlorophyll-protein complexes from *Euglena gracilis* and mutants deficient in chlorophyll b. *Plant Physiol.,* **80:**231-238.

Cunningham, W. P., Morre, D. J. and Mollenhauer, H. H. (1966). Structure of isolated plant Golgi apparatus revealed by negative staining. *J. Cell Biol.,* **28:**169-179.

Curtis, B. A., Tanifuji, G., Burki, F., Gruber, A., Irimia, M., Maruyama, S., Arias, M. C., Ball, S. G., Gile, G. H., Hirakawa, Y., Hopkins, J. F., Kuo, A., Rensing, S. A., Schmutz, J., Symeonidi, A., Elias, M., Eveleigh, R. J. M., Herman, E. K., Klute, M. J., Nakayama, T., Obornık, M., Reyes-Prieto, A., Armbrust, E. V., Aves, S. J., Beiko, R. G., Coutinho, P., Dacks, J. B., Durnford, D. G., Fast, N. M., Green, B. R., Grisdale, C. J., Hempel, F., Henrissat, B., Hoppner, M. P., Ishida, K., Kim, E., Koreny, L., Kroth, P. G., Liu, Y., Malik, S., Maier, U. G., McRose, D., Mock, T., Neilson, J. A. D., Onodera, N. T., Poole, A. M., Pritham, E. J., Richards, T. A., Rocap, G., Roy, S. W., Sarai, C., Schaack, S., Shirato, S., Slamovits, C. H., Spencer, D. F., Suzuki, S., Worden, A. Z., Zauner, S., Barry, K., Bell, C., Bharti, A. K., Crow, J. A., Grimwood, J., Kramer, R., Lindquist, E., Lucas, S., Salamov, A., McFadden, G. I., Lane, C. E., Keeling, P. J., Gray, M. W., Grigoriev, I. V. , and Archibald, J. M. (2012). Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature*, **492(7427):**59-65. Doi: 10.1038/nature11681.

Cushman, J. C. and Price, C. A. (1986). Synthesis and Turnover of Proteins in Proplastids and Chloroplasts of Euglena gracilis. *Plant Physiol.,* **82:** 972-977.

D'Onofrio, D. J. and An, G. (2010). A comparative approach for the investigation of biological information processing: An examination of the structure and function of computer hard drives and DNA.*Theoretical Biology and Medical Modelling,* **7:**3. http://www.tbiomed.com/content/7/1/3.

Dacks Joel B., Poon Pak P., and Field Mark C. (2007). Phylogeny of endocytic components yields insight into the process of non endosymbiotic organelle evolution. *PNAS,* **105(2):** 588 – 593.

Dacks, J. and Field, M. (2007). Evolution of the eukaryotic membrane-trafficking system: origin, tempo and mode. *Journal of cell science,* **120 (Pt 17):** 2977-85. Doi: 10.1242/jcs.013250.

Dacks, J. and Roger, A. J. (1999). The first sexual lineage and the relevance of facultative sex. *J Mol Evol.,* **48:**779–783.

Dacks, J. B., Poon, P. P., and Field, M. C. (2007). Phylogeny of endocytic components yields insight into the process of non endosymbiotic organelle evolution. *PNAS,* **105(2):** 588 – 593.doi: 10.1073/pnas.0707318105.

Dagan, T., Martin, W. (2006). The tree of one percent. *Genome. Biol.,* **7:**118.

Daiker Viktor, Häder Donat-P., Richter Peter R., Lebert Michael (2011). The involvement of a protein kinase in phototaxis and gravitaxis of *Euglena gracilis. Planta,* **233:**1055–1062.

Daiker, V., Häder, D-P., Richter, P. R. and Lebert, M. (2011). The involvement of a protein kinase in phototaxis and gravitaxis of *Euglena gracilis. Planta,* **233:**1055–1062 DOI 10.1007/s00425-011-1364-5.

Daiker, V., Lebert, M., Richter, P., Hader, DP. (2010). Molecular characterization of a calmodulin involved in the signal transduction chain of gravitaxis in *Euglena gracilis*. *Planta*, **231:**1229–1236. DOI 10.1007/s00425-010-1126-9.

Dang, Y., Yang, Q., Xue, Z., and Liu, Y. (2011). RNA Interference in Fungi: Pathways, Functions, and Applications. *Eukaryotic Cell,* 1148–1155.

Daniel R. Zerbino and Ewan Birney (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.,* **18:** 821-829.

DaRocha, W. D., Otsu, K., Teixeira, S. M., Donelson, J. E. (2004). Tests of cytoplasmic RNA interference (RNAi) and construction of a tetracycline-inducible T7 promoter system in *Trypanosoma cruzi*. *Mol Biochem Parasitol,***133:**175–86.

Davies, W. H., Mercer, E. I., and Goodwin, T. W. (1965). The occurrence and intracellular distribution of the plant sulpholipid in maize, runner beans, plant tissue cultures and *Euglena gracilis*. *Phytochemistry,* **4**: 741–749.

Davis, D. R. (1995). Stabilization of RNA stacking by pseudouridine. *Nucleic Acids Res.* **23:**5020–5026.

Davis, D. R. (1998). Biophysical and conformational properties of modified nucleosides in RNA (nuclear magnetic resonance studies). In: *Modification and Editing of RNA* (Grosjean, H. & Benne, R., eds), **pp.** 85–102, ASM Press, Washington, DC.

Davis, D. R., Veltri, C. A. and Nielsen, L. (1998). An RNA model system for investigation of pseudouridine stabilization of the codon–anticodon interaction in tRNALys, tRNAHis and tRNATyr. *J. Biomol. Struct. Dyn.,* **15:**1121–1132.

Davis, E. A. and Epstein, H.T. (1971). Some factors controlling step-wise variation of organelle number in *Euglena gracilis. Exp. CellRes.,***65:** 273-280.

De Gaudenzi, J. G., Noe, G., Campo, V. A., Frasch, A. C., and Cassola, A. (2011). Gene expression regulation in trypanosomatids. *Essays Biochemistry,* **51:**31-46. doi: 10.1042/bse0510031

de Graffenried, C. L., Ho, H. H. and Warren, G. (2008). Polo-like kinase is required for Golgi and bilobe biogenesis in *Trypanosoma brucei*. *J Cell Biol.,* **181:** 431–438.

de Koning, H. P. (2001). Transporters in African trypanosomes: role in drug action and resistance. International Journal for Parasitology, **31:**512 –522.

de Lederkremer, R. M. and Colli, W. (1995). Galactofuranose-containing glycoconjugates in trypanosomatids. *Glycobiology,* **5:**547–552.

de Lederkremer, R. M., Casal, O. L., Alves, M. J., Colli, W. (1980). Evidence for the presence of D-galactofuranose in the lipopeptidophosphoglycan from Trypanosome cruzi: Modification and tritium labeling. *FEBS Lett.,* **116:**25–29.

de Lederkremer, R. M., Lima, C. E., Ramirez, M. I., Goncalvez, M. F., Colli, W. (1993). Hexadecylpalmitoylglycerol or ceramide is linked to similar glycophosphoinositol anchor-like structures in *Trypanosoma cruzi*. *Eur J Biochem.,* **218:**929–936.

de Lederkremer, R. M., Lima, C., Ramirez, M. I., Ferguson, M. A., Homans, S. W., Thomas-Oates, J. (1991). Complete structure of the glycan of lipopeptidophosphoglycan from *Trypanosoma cruzi* epimastigotes. *J Biol Chem.,* **266:**23670–23675.

De Luca, F., Reyes, A., Veronico, P., Di Vito, M., Lamberti, F. and De Giorgi, C. (2002). Characterization of the (GAAA) microsatellite region in the plant parasitic nematode Meloidogyne artiellia. *Gene,* **293:**191–198.

de Paiva, R. M. C., Grazielle-Silva, V., Cardoso, M. S., Nakagaki, B. N., Mendonça-Neto, R. P., Canavaci, A. M. C., Melo, N. S., Martinelli, P. M., Fernandes, A. P., daRocha, W. D., Teixeira, S. M. R. (2015). Amastin Knockdown in *Leishmania braziliensis* Affects Parasite-Macrophage Interaction and Results in Impaired Viability of Intracellular Amastigotes. *PLOS Pathogens,* **11(12):** e1005296. https://doi.org/10.1371/journal.ppat.1005296.

de Smit, M. H., and van Duin, J. (1990). Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. *Proc. Natl. Acad. Sci. U. S. A.,* **87(19):**7668 – 7672.

Deane, M. P., Jansen, A. M. (1988). From a mono to a digenetic life-cycle: how was thejump for flagellates of the family Trypanosomatidae? *Mem Inst Oswaldo Cruz*, **83:**273–5.

Debrabant, A., Lee, N., Bertholet, S., Duncan, R., and Nakhasi, H. L. (2003). "Programmed cell death in trypanosomatids and other unicellular organisms," *International Journal for Parasitology*, **33(3):** 257–267.

Decatur, W. A. and Fournier, M. J. (2002). rRNA modifications and ribosome function. *Trends Biochem. Sci.,* **27:**344–351.

Decatur, W. A., Fournier, M. J. (2003). RNA-guided nucleotide modification of ribosomal and other RNAs. *J. Biol. Chem.,* **278(2):**695–698.

Dechat, T., Pfleghaar, K., Sengupta, K., Shimi, T., Shumaker, D. K., Solimando, L. and Goldman, R. D. (2008). Nuclear lamins: major factors in the structural organization and function of the nucleus and chromatin. *Genes Dev*., **22:**832-853.

Dehesh K., Apel, K. (1983). The function of proteases during the light-dependent transformation of etioplasts to chloroplasts in barley (Hordeum vulgare L.). *Planta*, **157:** 381-383.

Delmas, P. (2004). Polycystins: from mechanosensation to gene regulation. *Cell,***118:**145-148.

Delpech, S., Bre, M. H., Mazen, A., de Murcia, G., Champagne, M., Lefort-Tran, M. (1982). "Electron microscopic visualization of nucleosomal organization in B12 starved and control *Euglena* chromatin," *Cell Biology International Reports*, **6(2):** 197–203.

Depledge, D. P., MacLean, L. M., Hodgkinson, M. R., Smith, B. A., Jackson, A. P., Ma, S., Uliana, S. R. B., and Smith, D. F. (2010). Leishmania-Specific

Surface Antigens Show Sub-GenusSequence Variation and Immune Recognition. *PLoS Negl Trop Dis.,* **4(9):** e829. doi:10.1371/journal.pntd.0000829.

Deponte, M. (2008). “Programmed cell death in protists,” *Biochimicaet Biophysica Acta*, **1783(7):** 1396–1405.

DeRobertis, D., Schmidt, O. and Soll, D. (1980). Two control regions for eukaryotic tRNA gene transcription. *Proc. Natl. Acad. Sci. USA,* **77:** 3365 – 3368.

Deschamps, P., Lara, E., Marande, W., Lopez-Garcıa, P., Ekelund, F., and Moreira, D. (2011). Phylogenomic Analysis of Kinetoplastids Supports That Trypanosomatids Arose from within Bodonids. *Mol. Biol. Evol.,* **28(:1)** 53 - 58. doi:10.1093/molbev/msq28.

Desmond, E., Brochier-Armanet, C., Forterre, P., Gribaldo, S. (2011). On the last common ancestor and early evolution of eukaryotes: Reconstructing the history of mitochondrial ribosomes. *Res. Microbiol.,* **162:**53–70.

Devic, M., Schantz, R. (1983). Light-induced biosynthesis of chlorophyll-binding proteins in Euglena: Correlation between the level of mRNA and the accumulation of proteins in the thylakoids. Abstracts, Sixth International Congress of Photosynthesis, Vol 1, Brussels ABS 103-14.

Devos, D. and Valencia, A. (2001). Intrinsic errors in genome annotation. *Trends in Genetics,* **(17)8:**429 – 4431.

Devos, D., Dokudovskaya, S., Alber, F., Williams, R., Chait, B. T., Sali, A., and Rout, M. P. (2004). Components of coated vesicles and nuclear pore complexes share a common molecular architecture. *PLoS Biol.,* **2:**e380.

Devos, D., Dokudovskaya, S., Williams, R., Alber, F., Eswar, N., Chait, B. T., Rout, M. P. and Sali, A. (2006). Simple fold composition and modular architecture of the nuclear pore complex. *Proc Natl Acad Sci U S A*., **103(7):**2172–7.

Dhir, V., Goulding, D. and Field, M. C.(2004). TbRAB1 and TbRAB2 mediatetrafficking through the early secretory pathway of *Trypanosoma brucei*. Mol. Biochem.Parasitol., **137:**253 - 265.

Dhir, V., Goulding, D., Field, M. C. (2004). TbRAB1 and TbRAB2 mediate trafficking through the early secretory pathway of *Trypanosoma brucei*. Mol Biochem Parasitol., **137:**253–265.

Di Giulio, M. (2007) The universal ancestor and the ancestors of Archaea and Bacteria were anaerobes whereas the ancestor of the Eukarya domain was an aerobe. *J Evol Biol.,* **20:**543–548.

Dieci, G., Preti, M., Montanini, B. (2009). Eukaryotic snoRNAs: a paradigm for gene expression flexibility. *Genomics*, **94(2):**83–88.

Diekmann, Y., Seixas, E., Gouw, M., Tavares-Cadete, F., Seabra, M.C. and Pereira-Leal, J. B. (2011). Thousands of Rab GTPases for the Cell Biologist. *PLoS Comput. Biol.,* **7:**e1002217. doi:10.1371/journal.pcbi.1002217

Dien, B. (1979). Photic responses and sensory transduction in motile protists. H. Autrum (Ed.), Handbook of Sensory Physiology, VII/6aSpringer Verlag, Berlin, **pp.** 24–55

*Differentiation II, Origin and Continuity of Cell Organelles* (ed. J. Reinert & H. Ursprung), **pp.** 222-260. Berlin: Springer-Verlag.

Digirolampo., M. and Bradbeecr (1971). Transport of vitamin B12 in *Escherichia coli. Journal of Bacteriology,* **106:** 745-750.

Dobakova, E., Flegontov, P, Skalicky, T, and Lukes, J. (2015). Unexpectedly Streamlined Mitochondrial Genome of the Euglenozoan *Euglena gracilis*. Genome Biol. Evol. 7(12):3358–3367. doi:10.1093/gbe/evv229

Dobell, C. (1932). Antony van Leeuwenhoek and his ‘little animals’: being some account of the father of protozoology and bacteriology and his multifarious discoveries in these disciplines. *Constable, London, UK. Reprinted 1958. Russell and Russell, New York, New York, USA.*

Dobrowolski, M. P., Tommerup, I. C., Blakeman, H. D. and O'Brien, P. A. (2002). Non-Mendelian inheritance revealed in a genetic analysis of sexual progeny of Phytophthora cinnamomi with microsatellite markers. *Fungal Genet. Biol.,* **35:**197–212.

Dodge, J. D. (1973). The Fine Structure of Algal Cells. *Academic Press, New York*

Doege, M., Ohman, E., Tschiersch, H. (2000) Chlorophyll fluorescence in the alga *Euglena gracilis*. *Photosynthesis Research*, **63:** 159–170.

Doetsch, N. A., Thompson, M. D., Favreau, M. R., Hallick, R. B. (2001). Comparison of psbK operon organization and group III intron content in chloroplast genomes of 12 Euglenoid species. *Molecular and General Genetics,* **264:**682–690.

Doetsch, N. A., Thompson, M. D., Hallick, R. B. (1998). A maturase-encoding group III twintron is conserved in deeply rooted euglenoid species: are group III introns the chicken or the egg? *Molecular Biology and Evolution,* **15:** 76 – 86.

Doetsch, N. A., Thompson, M. D., Hallick, R. B. (1998). A maturase-encoding 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. DOI 10.1007/s00438-010-0585-9.

doi/10.1895/wormbook.1.77.1, http://www.wormbook.org.

Dolezel, J., Bartos, J., Voglmayr, H., Greilhuber, J. (2003). Nuclear DNA and genome size of trout and human, Cytometry, **51:**127–128.

Dolphin, W. D. (1970). Photoinduced Carotenogenesis in Chlorotic *Euglena gracilis*. *Plant Physiol.,* **46:**685-691.

Donaldson, J. G. and Jackson, C. L. (2011). ARF family G proteins and theirregulators: roles in membrane transport, development and disease. *Nat. Rev. Mol. Cell Biol.,* **12:**362–375.

Dooijes Dennis, Chaves Inês, Kieft Rudo, Dirks-Mulder Anita, Martin William, and Borst Piet (2000). Base J originally found in Kinetoplastida is also a minor constituent of nuclear DNA of *Euglena gracilis*. *Nucleic Acids Res.,* **28(16):** 3017 – 3021.

Dooijes, D., Chaves, I., Kieft, R., Dirks-Mulder, A., Martin, W. and Borst, P. (2000). Base J originally found in Kinetoplastida is also a minor constituent of nuclear DNA of *Euglena gracilis. Nucl. Acids Res.,* **28 (16):** 3017-3021. doi: 10.1093/nar/28.16.3017

Dorough, G. D., and Calvin, M., J. (1951). The path of oxygen in photosynthesis. *J. Am. Chem. Sot.,* **73:** 2362-2365.

dos Santos Ferreira, V., Rocchetta, I., Conforti, V., Bench, S., Feldman, R., and Levin, M. J. (2007). Gene expression patterns in *Euglena gracilis*: insights into the cellular response to environmental stress. *Gene,* **389:**136 - 145.

Doughty, M. J., Grieser, R., Diehn, B. (1980). Photosensory transduction in the flagellated alga, *Euglena gracilis*. 2 Evidence that blue light effects alteration in Na/K permeability of photoreceptor membrane, *Biochem. biophys. Acta,* **588:**148–168

Douglas, S., Zauner, S., Fraunholz, M., Beaton, M., Penny, S., Deng, L. T., Wu, X., Reith, M., Cavalier-Smith, T. & Maier, U. G. (2001). The highly reduced genome of an enslaved algal nucleus. *Nature,* **410:** 1091–1096.

Droll, D., Archer, S., Fenn, K., Delhi, P., Matthews, K., and Clayton, C. (2010). The trypanosome Pumilio-domain protein PUF7 associates with a nuclear cyclophilin and is involved in ribosomal RNA maturation. *FEBS Lett.,* **584:** 1156–1162.

Droll, D., Minia, I., Fadda, A., Singh, A., Stewart M, Queiroz, R. (2013) Post-Transcriptional Regulation of the Trypanosome Heat Shock Response by a Zinc Finger Protein. *PLoS Pathog.,* **9(4):** e1003286. doi:10.1371/journal.ppat.1003286.

Droll, D., Minia, I., Fadda, A., Singh, A., Stewart, M., Queiroz, R., and Clayton, C. (2013). Post-transcriptional regulation of the trypanosome heat shock response by a zinc finger protein. *PLoS Pathog.,* **9:** e1003286.

Dubertret, G., Lefort-Tran, M. (1982). Chloroplast molecular structure with particular reference to thylakoids and envelopes. In DE Buetow, ed, The Biology of Euglena, Vol 3. Academic Press, New York, **pp** 253-312.

DuBois, K. N., Alsford, S., Holden, J. M., Buisson, J., Swiderski, M., Bart, J., Ratushny, A. V., Wan, Y., Bastin, P., Barry, J. D,Navarro, M., Horn, D., Aitchison, J. D., Rout, M. P. and Field, M. C. (2012). NUP-1 Is a Large Coiled-Coil Nucleoskeletal Protein in Trypanosomes with Lamin-Like Functions. *PLoS Biol.,* **10(3):** e1001287. doi:10.1371/journal.pbio.1001287.

Dubreuil, R. R. and Bouck, G. B. (1985). The membrane skeleton of a unicellular organism consists of bridged articulating stripes. *J. Cell Biol.,* **101**:1884–1896.

Dubreuil, R. R., Marrs, J. A., Bouck, G. B. (1992). The cytoskeleton of euglenoids; cell form, surface motility, and cell replication are based on a membrane skeleton of repeating strips. In: Menzel D editor, *The cytoskeleton of the algae (Vol. 4) CRC Press, Inc.* **pp.** 59 – 78.

Dudov, K. P., Dabeva, M. D., Hadjiolov, A. A., Todorov, B. N. (1978). Processing and migration of ribosomal rna in the nucleolus and nucleoplasm of rat liver nuclei. *Biochem. J.* 171:375-384.

Dujardin, F. (1841). Histoire Naturelle des Zoophytes. Infusoires, comprenant la Physiologie et la Classification de ces Animaux, et la Manière de les Étudier a l'aide du Microscope. Paris. **pp.** 358.

Dulubova, .I, Yamaguchi, T., Arac, D., Li, H., Huryeva, I., Min, S.W., Rizo, J., Sudhof, T.C. (2002). Convergence and divergence in the mechanism of SNARE binding by Sec1/Munc18-like proteins. *Proc Natl Acad Sci U S A,* **100(1):**32-7.

Dunbar, D. A., Wormsley, S., Lowe, T. M., Baserga, S. J. (2000). Fibrillarinassociated box C/D small nucleolar RNAs in Trypanosoma brucei. Sequence conservation and implications for 20-O-ribose methylation of rRNA. *J. Biol. Chem.,* **275(19):**14767–14776.

Dunlap, J. R., Walne, P. L. and Bentley, J. (1983). Microarchitecture and elemental spatial segretation of envelopes of *Trachelomonas lefevrei* (Euglenophyceae). *Protoplasma,* **117:** 97-106.

Dunoyer, P., Brosnan,C. A, Schott, G., Wang, Y., Jay, F., Alioua, A., Himber, C. and Voinnet, O. (2010). An endogenous, systemic RNAi pathway in plants.*The EMBO Journal,* **29:**1699–1712.

Durand-Dubief, M., and Bastin, P. (2003). TbAGO1, an argonaute protein required for RNA interference, is involved in mitosis and chromosome segregation in *Trypanosoma brucei*. *BMC Biol.,* **1:** 2.

Durant, P. C. and Davis, D. R. (1999). Stabilization of the anticodon stem–loop of tRNALys,3 by an A+-C basepair and by pseudouridine. *J. Mol. Biol.,* **285:**115–131.

Durnford, D. G. and Gray, M. W. (2006). Analysis of *Euglena gracilis* Plastid-Targeted Proteins Reveals Different Classes of Transit Sequences. *Eukaryotic Cell,* 2079–2091.

Dworkin, M., (1959). Function of Carotenoids in Photosynthetic Bacteria. *Nature,* 184:1891 - 1892.

Dynesius, R. A., Walne, P. L. (1975). Ultrastructure of the reservoir and the flagella in *Phacus Pleuronectes* (Euglenophyceae). *J. Phycol.,* **11:**125–130

Dziembowski, A., Lorentzen, E., Conti, E., and Seraphin, B. (2007). A single subunit, Dis3, is essentially responsible for yeast exosome core activity. *Nat Struct Mol Biol.,* **14:**15–22.

Eakin, R. M. (1968). Evolution of Photoreceptors. *Evol. Biol.,* **2:** 194 – 240

Earl, D., Bradnam, K., St. John, J., Darling, A., Lin, D., Fass, J., Yu, H. O. K., Buffalo, V., Zerbino, D. R., Diekhans, M., Nguyen, N., Ariyaratne, P. N., Sung, W., Ning, Z., Haimel, M., Simpson, J. T., Fonseca, N. A., Birol, I., Docking, T. R., Ho, I. Y., Rokhsar, D. S., Chikhi, R., Lavenier, D., Chapuis, G., Naquin, D., Maillet, N., Schatz, M. C., Kelley, D. R., Phillippy, A. M., Koren, S., Yang, S. P., Wu, W., Chou, W. C., Srivastava, A., Shaw, T. I., J. Ruby, G., Skewes-Cox, P., Betegon, M., Dimon, M. T., Solovyev, V., Seledtsov, I., Kosarev, P., Vorobyev, D., Ramirez-Gonzalez, R., Leggett, R., MacLean, D., Xia, F., Luo, R., Li, Z., Xie, Y., Liu, B., Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F. J., Yin, S., Sharpe, T., Hall, G., Kersey, P. J., Durbin, R., Jackman, S. D., Chapman, J. A., Huang, X., DeRisi, J. L., Caccamo, M., Li, Y., Jaffe, D. B., Green, R. E., Haussler, D., Korf, I., and Paten, B. (2011). Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome research,* **21:** 2224–2241. doi: 10.1101/gr.126599.111

Ebel, C, Frantz, C., Paulus, F., and Imbault, P. (1999). Trans-splicing and cis-splicing in the colourless Euglenoid, Entosiphon sulcatum. *Curr. Genet.,* **35:**542–550.

Ebenezer, T. E., Zoltner, M., Burrel, A., Nenarokova, A., Novák Vanclová, A. M. G., Prasad, B., Soukal, P., Santana-Molina, C., O'Neill, E., Nankissoor, N. N., Vadakedath, N., Daiker, V., Obado, S., Jackson, A. P., Devos D., Lukeš, J., Lebert, M., Vaughan, S., Hampl, V., Carrington, M., Ginger, M. L., Dacks, J. B., Kelly, S., and Field, M. C., (2018) 'Unlocking the biological potential of *Euglena gracilis*: evolution, cell biology, and significance to parasitism.' bioRxiv. doi: https://doi.org/10.1101/228015

Ebersberger, I., Strauss, S., von Haeseler, A. (2009). HaMStR: profile hidden Markov model based search for orthologs in ESTs*. BMC Evol. Biol.* **9:**157.

Ebersberger, I., Strauss, S., von Haeseler, A. (2009). HaMStR: profile hidden Markov model based search for orthologs in ESTs. *BMC Evol. Biol.,* **9:**157.

Eddy, S. R. (2004). What is a hidden Markov model? *Nature Biotechnology,* **22:**1315 – 1316 doi:10.1038/nbt1004-1315.

Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Inform.,* **23(1):**205-11.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Comp. Biol.*, 7:e1002195

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res, **32(5):**1792–1797.

Edmunds Jr., L. N. (1988). Cellular andMolecular Bases of Biological Clocks. *Springer- Verlag New York, New York,* **pp.** 497.

Egel-Mitani M, Olson LW, Egel R. 1982. Meiosis in Aspergillus nidulans: another example for lacking synaptonemal complexes in the absence of crossover interference. *Hereditas,* **97:**179–187.

Eggimann, G., A., Sweeney, K., Bolt, H. L., Rozatian, N., Cobb, S. L., and Paul W. Denny (2015). The Role of Phosphoglycans in the Susceptibility of *Leishmania mexicana* to the Temporin Family of Anti-Microbial Peptides. *Molecules,* **20:** 2775-2785. doi:10.3390/molecules20022775.

Ehara, T., Shishira-Ishikawa, I., Osafune, T., Hase, E., Ohkuro, I. (1975) Some ultrastructural characteristics of chloroplast degeneration in cells of *Euglena gracilis* Z during their heterotrophic growth in darkness. *Journal of Electron Microscopy,* **24:** 253 – 261.

Ehrenberg, C. (1830). Organisation, Systematik und geographisches Verhältnifs der Infusionsthierchen. *Berlin,* 2:58-9

Ehrenberg, C. G. (1830a). Neue Beobachtungen uber blutartige Erscheinungen in Agypten, Arabienund Sibirien, nebst einer ¨ Ubersicht und Kritikder fruher bekannten. *Pogg Annalender Physik Chemie*, **18:**477–514 .

Ehrenberg, C. G. (1830b). Beitrage zur Kenntnisder Organisation der Infusorien und ihrer geographischen Verbreitung, besonders in Sibirien. *AbhK onigl Akad Wissensch Berlin,* **1830:**1–88.

Ehrenberg, M., Hauryliuk, V., Crist, C. G., and Nakamura, Y. (2007). Translation termination, the prion [PSI+], and ribosomal recycling, in *Translational control in biologyand medicine* (Mathews, M. B., Sonenberg, N., and Hershey, J. W. B., Eds.), **pp** 173-196, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Eichler, D. C., Craig, N. (1994). Processing of eukaryotic ribosomal RNA. *Prog. Nucleic Acid Res. Mol. Biol.,* **49:**197-239.

Ekanayake, D. and Sabatini, R. (2011). Epigenetic Regulation of Polymerase II Transcription Initiation in *Trypanosoma cruzi*: Modulation of Nucleosome Abundance, Histone Modification, and Polymerase Occupancy by O-Linked Thymine DNA Glucosylation. *Eukaryotic Cell,* **pp.** 1465–1472. doi:10.1128/EC.05185-11.

Eklund, H., Uhlin, U., Fa¨rnegårdh, M., Logan, D. T., and Nordlund, P. (2001). Structure and function of the radical enzyme ribonucleotide reductase. *Prog. Biophys. Mol. Biol.,* **77:**177–268.

Elbourne, L. D. H., Tetu, S., Hassan, K. and Paulsen I. T. (2017). TransportDB 2.0: a database for exploring membrane transporters in sequenced genomes from all domains of life. *Nucleic Acid Research,* **45(D1):** D320 - D324. doi: 10.1093/nar/gkw1068.

Elford, H. L., Bonner, E. L., Kerr, B. H., Hanna, S. D. and Smulson, M. (1977). Effect of methotrexate and 5-fluorodeoxyuridine on ribonucleotide reductase activity in mammalian cells. *Cancer Res.,* **37:**4389-4394.

Elias, M., Brighouse, A., Gabernet-Castello, C., Field, M. C., and Dacks, J. B. (2012). Sculpting the endomembrane system in deep time: high resolution phylogenetics of Rab GTPases. *J Cell Sci.,* **125(10):** 2500 – 2508.

Eliaz, D., Kannan, S., Shaked, H., Arvatz, G., Tkacz, I. D., Binder, L., Ben-Asher, W. H., Okalang, U., Chikne, V., Cohen-Chalamish, S., Michaeli, S. (2017). Exosome secretion affects social motility in *Trypanosoma brucei*. *PLoS Pathog.,* **13(3):** e1006245. doi:10.1371/journal. ppat.1006245.

Ellis, R. W., DeFeo, D., Shih, T. Y., Gonda, M. A., Young, H. A., Tsuchida, N., Lowy, D. R., Scolnick, E. M. (1981). The p21srcgenes of Harvey and Kirsten sarcoma viruses originate from divergentmembers of a family of normal vertebrate genes. *Nature,* **292:**506 – 511.

El-Metwally, S., Hamza, T., Zakaria, M., Helmy, M. (2013). Next-Generation Sequence Assembly: Four Stages of Data Processing and Computational Challenges. *PLoS Comput Biol.,* **9(12):**e1003345.

El-Sayed, N. M., Myler, P. J., Bartholomeu, D.C., Nilsson, D., Aggarwa, G., Tran, A., Ghedin, E., Worthey, E. A., Delcher, A. L., Blandin, G., Westenberger, S. J., Caler, E., Cerqueira, G. C., Branche, C., Haas, B., Anupama, A., Arner, E., Åslund, L., Attipoe, P., Bontempi, E., Bringaud, F., Burton, P., Cadag, E., Campbell, D. A., Carrington, M., Crabtree, J., Darban, H., da Silveira, J. F., de Jong, P., Edwards, K., Englund, P. T., Fazelina, G., Feldblyum, T., Ferella, M., Frasch, A. C., Gull, K., Horn, D., Hou, L., Huang, Y., Kindlund, E., Klingbeil, M., Kluge, S., Koo, H., Lacerda, D., Levin, M. J., Lorenzi, H., Louie, T., Machado, C. R., McCulloch, R., McKenna, A.,Mizuno, Y., Mottram, J. C., Nelson, S., Ochaya, S., Osoegawa, K., Pai, G.,Parsons, M., Pentony, M., Pettersson, U., Pop, M., Ramirez, J. L., Rinta, J., Robertson, L., Salzberg, S. L., Sanchez, D. O., Seyler, A., Sharma, R., Shetty, J., Simpson, A. J., Sisk, E., Tammi, M. T., Tarleton, R., Teixeira, S., Aken, S. V., Vogt, C., Ward, P. N., Wickstead, B., Wortman, J., White, O., Fraser, C. M., Stuart, K. D., Andersson, B.(2005). The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science,* **309:**409–415.

*Emanuelsson, O, Nielsen, H., Brunak, S. and von Heijne, G.* (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence.*J. Mol. Biol.,* **300:** 1005-1016.

Emanuelsson, O., Nielsen, H. and von Heijne, G. (1999). ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci.,* **8(5):**978-84.

Embly, T. M. and Martin, W. (2006). Eukaryotic evolution changes and challenges. *Nature*, **440:**623–630.

Emms, D. M. and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons. dramatically improves orthogroup inference accuracy. *Genome Biology,* **16:**157. DOI 10.1186/s13059-015-0721-2.

Epstein, H. T. and Allaway, E. (1967). Properties of selectively starved *Euglena. Biochem. Biophys. Acta.,* **142:** 195 - 207.

Erben, E., Chakraborty, C., and Clayton, C. (2014). The CAF1-NOT complex of trypanosomes. *Front Genet.,* **4:** 299.

Ernst, N.L., Panicucci, B., Carnes, J., and Stuart, K. (2009). Differential functions of two editosome exoUases in *Trypanosoma brucei*. *RNA,* **15:**947–957.

Ernst, N.L., Panicucci, B., Igo Jr., R.P., Panigrahi, A.K., Salavati, R., and Stuart, K., (2003). TbMP57 is a 30 terminal uridylyl transferase (TUTase) of the Trypanosoma brucei editosome. *Mol. Cell,* **11:**1525–1536.

Esson, H. J., Morriswood, B., Yavuz, S., Vidilaseris, K., Dong, G., and Warren, G. (2012). Morphology of the Trypanosome Bilobe, a Novel CytoskeletalStructure. *Eukaryotic Cell,* **11:**761–772. doi:10.1128/EC.05287-11.

Estevez, A. M., Kempf, T., and Clayton, C. (2001). The exosome of Trypanosoma brucei. *EMBO J.,* **20:**3831–3839.

Estevez, A. M., Lehner, B., Sanderson, C. M., Ruppert, T., and Clayton, C. (2003). The roles of intersubunit interactions in exosome stability. *J Biol. Chem.,* **37:** 34943–34951.

Eyre-Walker, A. and Hurst, L. D. (2001). OPINION: The evolution of isochores. *Nat. Rev. Genet.,* **2:**549–555 doi:10.1038/35080577.

Fadda, A., Ryten, M., Droll, D., Rojas, F., Färber, V., Haanstra, J. R., Merce, C., Bakker, B. M., Matthews, K. and Clayton, C. (2014). Transcriptome-wide analysis of trypanosome mRNA decay reveals complex degradation kinetics and suggests a role for co-transcriptional degradation in determining mRNA levels. *Molecular Microbiology,* **94(2):**307–326 doi:10.1111/mmi.12764.

Fadda, A., Färber, V., Droll, D., and Clayton, C. (2013). The roles of 3′-exoribonucleases and the exosome in trypanosome mRNA degradation. *RNA,* **19:** 937–947.

Faktorová, D., Dobáková, E., Peña-Diaz, P. and Lukeš, J. (2016). From simple to supercomplex: mitochondrial genomes ofeuglenozoan protists [version 1; referees: 2 approved] *F1000. Research,* **5(F1000 Faculty Rev):392** (doi:10.12688/f1000research.8040.1).

Faktorová, D., Dobáková, E., Peña-Diaz, P. and Lukeš, J. (2016). From simple to supercomplex: mitochondrial genomes of euglenozoan protists [version 1; referees: 2 approved]. *F1000Research,* **5(F1000 Faculty Rev):**392 (doi:10.12688/f1000research.8040.1).

Falchuk, K. H., Fawcett, D. W. and Vallee, B. L. (1975). Role of Zinc in Cell Division of *Euglena gracilis. J. Cell Sci.,* **17:** 57-78

Falchuk, K. H., Gordon, P. R., Stankiewicz, A., Hilt, K. L., and Vallee, B**.** L. (1986). *Euglena gracilis* Chromatin: Comparison of Effects of Zinc, Iron, Magnesium, or Manganese Deficiency and Cold Shock. *Biochemistry*, **25:**5388-5391.

Falchuk, K. H., Mazus, B., Ulpino, L., Vallee, B. L. (1976). *Euglena gracilis* DNA Dependent RNA Polymerase II: A Zinc Metalloenzyme. *Biochemistry,* **15(20):** 4468 – 4475.

Farmer, M. A. and Triemer, R. E. (1988). Flagellar systems in the euglenoid flagellates. *BioSystems,* **21:**283–291.

Farmer, M. A., Triemer, R. E. (1988). Flagellar systems in the euglenoid flagellates. *BioSistems,* **21:** 283–291.

Fatica, A., and Tollervey, D. (2002). Making ribosomes. *Curr. Opin. CellBiol.,* **14:**313–318. http://dx.doi.org/10.1016/S0955-0674(02)00336-8.

Feagin, J. E. (2000). Mitochondrial genome diversity in parasites. *Int. J. Parasitol.,* **30:**371–390.

Feinleib, M. E. (1985). Behavioral studies of free-swimming photoresponsive organisms. G. Colombetti, F. Lenci, S. Song (Eds.), Sensory Perception and Transduction in Aneural Organisms, Plenum Press, New York, **pp.** 119–146.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution,* **39(4):**783–791.

Fermino, B. R., Viola, L. B., Paiva, F., Garcia, H. A., de Paula, C. D., Botero-Arias, R., Takata, C. S. A., Campaner, M., Hamilton, P. B., Camargo, E. P. and Teixeira, M. M. G. (2013). The phylogeography of trypanosomes from South American alligatorids and African crocodilids is consistent with the geological history of South Ameri-can river basins and the transoceanic dispersal of Crocodylus at the Miocene. *Parasit Vectors,* **6:**313.

Fernandes, A. P., Nelson, K., Beverley, S. M. (1993). Evolution of nuclear ribosomal RNAsin kinetoplastid protozoa: perspectives on the age and origins of parasitism. *Proc. Natl. Acad. Sci. USA*, **90:**11608–12.

Ferrara, R. and Banchetti, R. (1976). Effect of streptomycin on the structure and function of the photoreceptor apparatus of *Euglena gracilis. J. exp. Zool.,* **198:**393 – 402

Ferroni, L., Baldisserotto, C., Pantaleoni, L., Fasulo, M. P., Fagioli, P. and Pancaldi, S. (2009). Degreening of The Unicellular Alga *Euglena gracilis*: Thylakoid Composition, Room Temperature Fluorescence Spectra and Chloroplast Morphology. *Plant Biology,* **11:** 631–641

Fey, E. G., Krochmalnic, G., Penman S. (1986). The nonchromatin substructures of the nucleus: the ribonucleoprotein (RNP)-containing and RNPdepleted matrices analyzed by sequential fractionation and resinless section electron microscopy. *J. Cell Biol.,* **102:**1654-1665.

Fiebig Michael, Gluenz Eva, Carrington Mark, Kelly Steven (2014). SLaP mapper: A webserver for identifying and quantifying spliced-leader addition and polyadenylation site usage in kinetoplastid genomes. Molecular & Biochemical Parasitology, **196:** 71–74

Field, H., Farjah, M., Pal, A., Gull, K., and Field, M. C. (1998). Complexity of trypanosomatid endocytosis pathways revealed by Rab4 and Rab5 isoforms in *Trypanosoma brucei*. *J Biol Chem.,* **273:**32102 – 32110.

Field, M. C., Sali, A., Rout, M. P. (2011). Evolution: on a bender—BARs, ESCRTs, COPs, and finally getting your coat. J Cell Biol. 193:963–972.

Figueiredo, L.M., Cross, G. A. M., and Janzen, C. J. (2009). Epigenetic regulation in African trypanosomes: a new kid on the block.*Nat. Rev. Microbiol.,* **7:**504–513, http://dx.doi.org/10.1038/nrmicro2149.

Finch, J. T., Klug, A. (1976). Solenoid model for superstructure in chromatin. *Proc. Natl. Acad. Sci. USA,* **73:**1897-1901

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm,L., Mistry, J., Sonnhammer, E. L. L., Tate, J., and Punta, M. (2014). Pfam: The protein families database. *Nucleic Acids Res.,* **42:**D222–D230.

Fish L. E., Deshaies, R., Jagendorf, A. T. (1983). A $Mg^{2+}$ requirement for rapid ATP-driven protein synthesis by intact pea chloroplasts. *Plant Sci Lett.,* **31:** 139 - 1

Fisher, K. A. and Stoeckenius, W. (1977). Freezing fractured purple membrane particles: protein content Science, 197, **pp.** 72–74

Flegontov Pavel, Gray Michael W., Burger Gertraud, Lukes Julius (2011). Gene fragmentation: a key to mitochondrial genome evolution in Euglenozoa? *Curr Genet.,* **57:**225 – 232 DOI 10.1007/s00294-011-0340-8.

Flegontov, P. and Lukeš, J. (2012). Mitochondrial Genomes of Photosynthetic Euglenids and Alveolates. *Advances in Botanical Research*, **63:** 127 – 153 http://dx.doi.org/10.1016/B978-0-12-394279-1.00006-5.

Flegontov, P., Gray, M. W., Burger, G., and Lukes, J. (2011). Gene fragmentation: a key to mitochondrial genome evolution in Euglenozoa? *Curr Genet.,* **57:**225 – 232 DOI 10.1007/s00294-011-0340-8.

Flegontov, P., Votypka, J., Skalicky, T., Logacheva, M. D., Penin, A. A., Tanifuji, G., Onodera, N. T., Kondrashov, A. S., Volf. P., Archibald, J. M., Lukeš, J. (2013). Paratrypanosoma is a novel early-branching trypanosomatid. *Curr Biol.,* **23:**1787–93.

Fleischmann, R. D. *et al. (1995).* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science,* **269:**496–512.

Follmann, H. (1974). Ribonucleotide Reductases: Radical Chemistry and Inhibition at the Active Site. *Angew. Chem. Int. Ed. Engl.,* **13:**569-579.

Fontana, D. R. and Devreotes, P. N. (1984). cAMP-stimulated adenylate cyclase activation in Dictyostelium discoideum is inhibited by agents acting at the cell surface. *Develop. Biol.,* **106:**76–82.

Foster, K. W., Smyth, R. D. (1980). Light antennas in phototactic algae. *Microbiol. Rev.,* **44:**572–630

Frantz, C., Ebel, C., Paulus, F. and Imbault, P. (2000). Characterization of transsplicing in Euglenoids. *Curr. Genet.,* **37:**349–355.

Frantz, C., Ebel, C., Paulus, F., Imbault, P. (2000). Characterization of trans-splicing in Euglenoids. *Curr Genet.,* **37:** 349 –355.

Fraser, C. *et al. (1995).* The minimal gene complement of *Mycoplasma genitalium*. *Science,* **270:**397– 403.

Freire, E. R., Dhalia, R., Moura, D. M., da Costa Lima, T. D., Lima, R. P., Reis, C. R., Hughes, K., Figueiredo, R. C., Standart, N., Carrington, M., de Melo Neto, O. P. (2011). The four trypanosomatid eIF4E homologues fall into two separategroups, with distinct features in primary sequence and biologicalproperties. *Mol. Biochem. Parasitol.*, **176:**25–36. http://dx.doi.org/10.1016/j.molbiopara.2010.11.011.

Frenster, J. H. (1969). Handbook of molecular cytology (ed A Lima-de-Faria) v. 251. North-Holland. Amsterdam.

Fritz-Laylin, L. K., Prochnik, S. E., Ginger, M. L., Dacks, J. B., Carpenter, M. L., Field, M. C., Kuo, A., Paredez, A., Chapman, J., Pham, J. Shu S, Neupane R, Cipriano M, Mancuso J, Tu H, Salamov A, Lindquist E, Shapiro H, Lucas S, Grigoriev I. V., Cande W. Z., Fulton, C., Rokhsar, D. S., Dawson, S. C. (2010). The genome of Naegleria gruberi illuminates early eukaryotic versatility. *Cell,* **140:**631- 642.

Frommolt, R., Werner, S., Paulsen, H., Goss, R., Wilhelm, C., Zauner, S., Maier, U. G., Grossman, A. R., Bhattacharya, D., Lohr, M. (2008). Ancient recruitment by chromists of green algal genes encoding enzymes for carotenoid biosynthesis. *Mol Biol Evol,* **25(12):**2653-2667.

Fukui, Y., Katsumaru, H., (1980). Dynamics of nuclear actin bundle induction by dimethyl sulfoxide and factors affecting its development. *J. Cell Biol.,* **84:**131 - 140.

Furuhashi, T., Ogawa, T., Nakai, R., Nakazawa, M., Okazawa, A, Padermschoke, A., Nishio, K., Hirai, M. Y., Arita, M., Ohta, D. (2015). Wax ester and lipophilic compound profiling of *Euglena gracilis* by gas chromatography-mass spectrometry: toward understanding of wax ester fermentation under hypoxia. *Metabolomics,* 11: 175. doi:10.1007/s11306-014-0687-1.

Gabaldón, T., Ginger, M. L. and Michels, P. A. M. (2016). Peroxisomes in parasitic protists. *Mol Biochem Parasitol,* **209(1-2):**35-45 http://dx.doi.org/10.1016/j.molbiopara.2016.02.005

Gabernet-Castello, C., O'Reilly, A. J., Dacks, J. B. and Field, M. C. (2013). Evolution of Tre-2/Bub2/Cdc16 (TBC) Rab GTPase-activating proteins. *Mol. Biol. Cell,* **24(10):**1574 - 83. doi: 10.1091/mbc.E12-07-0557.

Galland, P., Keiner, P., Dornemann, D., Senger, H., Brodhuhn, B., Hader, D.P. (1990). Pterin- and flavin-like fluorescence associated with isolated flagella of *Euglena gracilis. Photochem. Photobiol.,* 51, **pp.** 675–680

Gallie, D. R. and Browning, K. S. (2001). eIF4G functionally differs from eIFiso4G in promoting internal initiation, cap-independent translation, and translation of structured mRNAs. *J. Biol. Chem.,* **276:**36951–36960. http://dx.doi.org/10.1074/jbc.M103869200.

Gallo, J. M. and Schrevel, J. (1982). Euglenoid movement in Distigma proteus I. Cortical rotational movement. *Biol. Cell,* **44:**139–148.

Galtier, N., Piganeau, G., Mouchiroud, D. and Duret, L. (2001). GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. *Genetics,* **159:**907–911.

Garber, R. L. and Gage, L. P. (1979). Transcription of a cloned Bombyx mori $tRNA^{Ala}$ gene nucleotide sequence of the tRNA precursor and its processing in vitro. *Cell,* **17:** 389 – 397.

Garcia Silva, M. R., Tosar, J. P., Frugier, M., Pantano, S., Bonilla, B., Esteban, L., Serra, E., Rovira, C., Robello, C. and Cayota, A. (2010). Cloning, characterization and subcellular localization of a *Trypanosoma cruzi*

argonaute protein defining a new subfamily distinctive of trypanosomatids. *Gene,* **466:**26–35.

Gees, M., Colsoul, B., and Nilius, B. (2010). The Role of Transient Receptor Potential Cation Channels in Ca2+ Signaling. *Cold Spring Harbor Perspectives in Biology,* **2(10):**a003962.

Gehring, W. J. (2005). New perspectives on eye development and the evolution of eyes and photoreceptors. *J. Hered.,* **96:**171–184.

Geimer, S., Belicová, A., Legen, J., Sláviková, S., Herrmann, R. G., Krajbovib, J. (2009). Transcriptome analysis of the *Euglena gracilis* plastid chromosome. *Curr Genet,* **55:**425–438. DOI 10.1007/s00294-009-0256-8.

Gerard, N. P., Garraway, L. A., Eddy Jr., R. L., Shows, T. B., Iijima, H., Paquet, J. L., and Gerard, C. (1991). Human substance P receptor (NK-1): organization of the gene, chromosome localization, and functional expression of cDNA clones. *Biochemistry,* **30 (44):** 10640–10646.

Gerard, N.P., Eddy Jr., R.L., Shows, T. B., and Gerard, C.(1990). The human neurokinin A (substance K) receptor. Molecular cloning of the gene, chromosome localization, and isolation of cDNA from tracheal and gastric tissues. *J. Biol. Chem.,* **265(33):** 20455–20462.

Gerbi, S. A., Savino, R., Stebbins-Boaz, B., Jeppsen, C., Rivera-Leon, R. (1990). In: W. E. Hill, A. Dahlberg, A.R. Garrett, P.B. Moore, D. Schlessinger, J.R. Warner (Eds.). (1990). The Ribosome. Structure, Function, and Evolution, American Society for Microbiology, Washington, DC, **pp.** 452 - 469.

Gheiratmand, L., Brasseur, A., Zhou, Q. and He, C. Y. (2013). Biochemical Characterization of the Bi-lobe Reveals a Continuous Structural Network Linking the Bi-lobe to Other Single-copied Organelles in *Trypanosoma brucei. The Journal of Biological Chemistry,* **288(5):**3489–3499.

Ghetti, F., Colombetti, G., Lenci, F., Campani, E., Polacco, E., Quaglia, M. (1985). Fluorescence of *Euglena gracilis* photoreceptor pigment: an *in vivo* microspectrofluorimetric study. *Photochem. Photobiol.,* **42:**29–33

Ghildiyal, M., and Zamore, P. D. (2009). Small silencing RNAs: an expanding universe. *Nat Rev Genet,* **10:**94–108.

Gibbs, S. P. (1960). The fine structure of *Euglena gracilis* with special reference to the Chloroplasts and pyrenoids. *J. Ultrastruct. Res.,* **4:** 127-148.

Gibbs, S. P. (1978). The chloroplast of Euglena may have evolved from symbiotic green algae. *Can. J. Bot.,* **56:**2883–2889.

Gibbs, S. P. (1978). The chloroplasts of Euglena may have evolved from symbiotic green algae. *Can J Bot*., **56:** 2883–2889.

Gile, G. H., Faktorová, D., Castlejohn, C. A., Burger, G., Lang, B. F., Farmer, M. A., Lukes, J., and Keeling, P. J. (2009). Distribution and phylogeny of EFL and EF1alpha in Euglenozoa suggest ancestral co-occurrence followed by differential loss. *PLoS ONE,* **4:**e5162.

Gile, G. H., Novis, P. M., Cragg, D. S., Zuccarello, G. C., and Keeling, P. J. (2009) Thedistribution of Elongation Factor-1 Alpha (EF1alpha), Elongation Factor-Like (EFL), anda non-canonical genetic code in the ulvophyceae: discrete genetic characters support aconsistent phylogenetic framework. *J. Eukar. Microbiol., 56:* 367-372.

Gilson, P. R., Su, V., Slamovits, C. H., Reith, M. E., Keeling, P. J., McFadden, G. I. (2006). Complete nucleotide sequence of the chlorarachniophyte nucleomorph: Nature's smallest nucleus. *Proc. Nat. Acad. Sci. USA*, **103:**9566 – 9571.

Ginger, M. L., Fritz-Laylin, L. K., Fulton, C., Cande, W. Z., and Dawson, S. C. (2010). Intermediary Metabolism in Protists: a Sequence-based View of Facultative Anaerobic Metabolism in Evolutionarily Diverse Eukaryotes. *Protist,* **161(5):** 642–671. doi:10.1016/j.protis.2010.09.001.

Gitelman, S. E., Witman, G. B. (1980) Purification of calmodulin from Chlamydomonas: calmodulin occurs in cell bodies and flagella. *J Cell Biol.,* **98:**764–770.

Gleason, F. K. and Hogenkamp, H. P. (1970). Ribonucleotide reductase from *Euglena gracilis*, a deoxyadenosylcobalamin-dependent enzyme. *J. Biol. Chem.,* **245:**4894 – 4899.

Gnerre, S., MacCallum. I., Przybylski, D., Ribeiro, F., Burton, J., Walker, B., Sharpe, T., Hall, G., Shea, T., Sykes, S., Berlin, A., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E. S., and Jaffe, D. B. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data *Proceedings of the National Academy of Sciences USA,* **108(4):**1513-1518).

Gnerre, S., MacCallum. I., Przybylski, D., Ribeiro, F., Burton, J., Walker, B., Sharpe, T., Hall, G., Shea, T., Sykes, S., Berlin, A., Aird, D., Costello, M., Daza, R., Williams, L., Nicol, R., Gnirke, A., Nusbaum, C., Lander, E. S., and Jaffe, D. B. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences USA,* **108(4):**1513-1518.

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

Godjics, M. (1953). The genus *Euglena*. *Madison University,Wisconsin Press*

Godlerer, G., Loidol, P. and Grbbner, P. (1991). Cell cycledependent ADP-ribosylation of the nuclear matrix. *Eur. J. Cell Biol.,* **55:**183-185.

Goedl, C., Sawangwan, T., Mueller, M., Schwarz, A., Nidetzky, B. (2008). A high-yielding biocatalytic process for the production of 2-O-(alpha-D-glucopyranosyl)-sn-glycerol, a natural osmolyte and useful moisturizing ingredient. *Angew. Chem., Int. Ed.*47:10086–10089.

Goffeau, A. *et al.* (1996). Life with 6000 genes. The description of how the first eukaryotic genome was sequenced and annotated. *Science,* **274:**546.

Gojdics, M. (1953). *The Genus Euglena.* Madison, Wisconsin: Univ. Wisconsin Press.

Gold, J. C. and Spremulli, L. L. (1985). *Euglena gracilis* chloroplate initiation factor 2: Identification and initial characterization. *Journal of Biological Chemistry*, **260(28):**14897-14900.

Gomez-Silva, B., Timko, M. P. and Schiff, J. A. (1985). Chlorophyll biosynthesis from glutamate or 5-aminolevulinate in intact Euglena chloroplasts. *Planta,* **165(1):**12-22.

Gommers-Ampt, J. H., Van Leeuwen, F., De Beer, A. L. J., Vliegenthart, F. G., Dizdaroglu, M., Kowalak, J. A., Crain, P. F. and Borst, P. (1993). *Cell,* **75:**1129–1136.

Gordeeva, A. V., Labas, Y. A., and Zvyagilskaya, R. A. (2004). "Apoptosis in unicellular organisms: mechanisms and evolution," *Biochemistry*, **69(10):** 1055–1066.

Goringer, H. U. (2012). 'Gestalt,' Composition and Function of the *Trypanosoma brucei* Editosome. *Annu. Rev. Microbiol.,* **66:**65–82. 10.1146/annurev-micro-092611-150150.

Goto, K. and Beneragama, C. K. (2010). Circadian clocks and antiaging: Do non-aging microalgae like Euglena reveal anything? *Ageing Research Reviews,* **9:** 91–100.

Goto, K., Laval-Martin, D. M. and Edmunds Jr., L. N. (1985). Biochemical modeling of an autonomously oscillatory circadian clock in Euglena. *Science,* **228:**1284–1288.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F, Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.,* **29(7):**644-52. doi: 10.1038/nbt.1883. PubMed PMID: 21572440.

Gradstein, F. M., Ogg, J. G., Smith, A. G. (2004). A geologic time scale. Cambridge, UK: Cambridge University Press; **pp.** 589

Grain, J. (1986). The cytoskeleton in protists: nature, structure and functions. *Int. Rev. Cytol.,* **104**:153–249.

Grain, J., Mignot, J. P., de Puytorac, P. (1988). Ultrastructures and evolutionary modalities of flagellar and ciliary systems in protists. *Biol. Cell,* **63:**219–237

Grant, B. D. and Sato, M. (2006). Intracellular trafficking. WormBook, ed. *The C. elegans Research Community*, WormBook, doi/10.1895/wormbook.1.77.1, http://www.wormbook.org.

Grant, B. D. and Sato, M. (2006). *Intracellular trafficking. WormBook*, ed. The *C. elegans* Research Community, WormBook,

Grassé, P. P. (1952): Ordre des Trichomonadines, Ordre des Joeniides, Ordre des Lophomonadines, Ordre des Tricho-nymphines, Ordre des Spirotrichonymphines. *In: Grassé P. P. (éd.): Traité de Zoologie, Flagellés.,* **1:**704–962, Masson et Cie, Paris.

Gray, M. W. (1992). The endosymbiont hypothesis revisited. *Int. Rev. Cytol.,* **141:**233–357.

Gray, M. W. and Doolittle, W. F. (1982). Has the endosymbiont hypothesis been proven? *Microbiol Rev.***46:**1–42.

Gray, M. W., Lang, B. F. and Burger, G. (2004). Mitochondria of protists. *Annu. Rev. Genet.,* **38:**477 – 524.

Gray, M. W., Lukes J., Archibald, J. M., Keeling, P. J., Doolittle, W. F. (2010). Irremediable complexity? *Science,* **330:**920–921.

Green, D. E. and Brierley, G. P. (1965). In Biochemistry of Quinone, p. 405. Ed. By Morton, R. A. New York and London: Academic Press Inc.

Green, L. S., Li, Y., Emerich, D. W., Bergersen, F. J., Day, D. A. (2000). Catabolism of _-ketoglutarate by a sucA mutant of *Bradyrhizobium japonicum*:

evidence for an alternative tricarboxylic acid cycle. *J. Bacteriol*., **182**:2838–2844.

Greenberg, B. M., Narita, J. O., DeLuca-Flaherty, C., Gruissem, W., Rushlow, K. A. and Hallick, R. B. (1984). Evidence for two RNA polymerase activities in *Euglena gracilis* chloroplasts. *J. Biol. Chem.*, **259:**14880–14887.

Greenwood Spencer J., Schnare Murray N., Cook James R. and Gray Michael W. (2001). Analysis of intergenic spacer transcripts suggests 'read-around' transcription of the extrachromosomal circular rDNA in *Euglena gracilis*. *Nucl. Acids Res.,* **29 (10):** 2191-2198. doi: 10.1093/nar/29.10.2191

Greenwood, S. J., Schnare, M. N., Cook, J. R. and Gray, M. W. (2001). Analysis of intergenic spacer transcripts suggests'read-around'transcription of the extrachromosomal circular rDNA in Euglena gracilis. *Nucleic Acids Res.,* **29:**2191–2198.

Greer, E. L. and Brunet, A. (2008). Signaling networks in aging. *J. Cell Sci.,* **121:**407–412.

Gregory, T. R. (2005). Synergy between sequence and size in large genome-scale genomics. *NATURE REVIEWS* | *GENETICS,* **6:699**-708.

Gremme, G., Steinbiss, S. and Kurtz, S. (2013). *GenomeTools*: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **10(3):**645–656.

Grimma, P., Rissea, J. M., Cholewaa, D., Müllera, J. M., Beshayb, U., Friehsa, K., Flaschel, E. (2015). Applicability of *Euglena gracilis* for biorefineries demonstrated by theproduction of α-tocopherol and paramylon followed by anaerobic digestion. *Journal of Biotechnology*, **215:**72–79. http://dx.doi.org/10.1016/j.jbiotec.2015.04.004.

Grobner, P. and Loidol, P. (1985). ADP-ribosyltransferase in isolated nuclei during the cell cycle of *Physarum polycepharum. Biochem. J.,* **232:**214.

Grosjean, H., Breton, M., Sirand-Pugnet, P., Tardy, F., Thiaucourt, F.,Citti, C., Barre, A., Yoshizawa, S., Fourmy, D., de Crecy-Lagard, V., Blanchard, A.(2014). Predicting the Minimal Translation Apparatus: Lessons from the Reductive Evolution of Mollicutes. *PLoS Genet.,* **10(5):** e1004363. doi:10.1371/journal.pgen.1004363.

Gruissem, W. and Prescott, D. M. (1982). Transcription of *E. coli* and *Euglena* Chloroplast tRNA Gene Clusters and Processing of Polycistronic Transcripts in Hela Cell-Free System. *Cell,* **30:** 81 – 92.

Gruning, N. M., Rinnerthaler, M., Bluemlein, K., Mulleder, M., Wamelink, M. M., Lehrach, H., Jakobs, C., Breitenbach, M., Ralser, M. (2011). Pyruvate kinase triggers a metabolic feedback loop that controls redox metabolism in respiring cells. *Cell Metab.,* **14:**415–427. (doi:10.1016/j.cmet.2011.06.017).

Gualdron-Lopez, M., Brennand, A., Hannaert, V., Quinones, W., Caceres, A. J., Bringaud, F., Concepcion, J. L., Michels, P. A. (2012). When, how and why glycolysis became compartmentalised in the Kinetoplastea: a new look at an ancient organelle. *Int. J. Parasitol.,* **42:**1–20.

Gualtieri, P. (1991). Microspectroscopy of photoreceptor pigments in flagellated algae. *CRC Crit. Rev. Plant Sci.,* **9(6):** 475-495 DOI:10.1080/07352689109382302

Gualtieri, P., Barasanti, L., Passarelli, V. (1989). Absorption spectrum of a single isolated paraflagellar swelling of *Euglena gracilis. Biochim. biophys. Acta,* **993:**293 – 296

Gualtieri, P., Barsanti, L. G. (1986). Rosati Isolation of the photoreceptor (paraflagellar body) of the phototactic flagellate *Euglena gracilis. Arch. Microbiol.,* **145:**303–305

Gualtieri, P., Barsanti, L., Passarelli, V., Verni, F. (1988). Morphological investigation of the *Euglena gracilis* isolated paraflagellar body. *Micron Microsc. Acta,* **19:**241–246

Gualtieri, P., Barsanti, L., Passarelli, V., Verni, F., Rosati, G. (1990). A look into the reservoir of *Euglena gracilis*: SEM investigations of the flagellar apparatus. *Micron Microsc. Acta*, **21(3):**131–138

Gualtieri, P., Passarelli, V., Barsanti, L. (1989). A simple instrument to perform *in vivo* absorption spectra of pigmented cellular organelles. *Micron Microsc. Acta,* **20:** 107–110

Guindon Staphane and Gascuel Olivier (2003). "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood." *Systematic Biology,* **52(5):**696-704.

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59:**307–321.

Gull, K. (2001). Protist tubulins: new arrivals, evolutionary relationships and insights to cytoskeletal function. *Curr Opin Microbiol*, **4:** 427 - 432.

Gunderson, J. H., Elwood, H., Ingold, A., Kindle, K., Sogin, M. L. (1987). Phylogenetic relations between chlorophytes, chrysophytes and oomicetes. *Proc. natn. Acad. Sci. U.S.A.,* **84:**5823–5827

Guo, M., Yang, X. L., and Schimmel, P. (2010). New functions of aminoacyl-tRNAsynthetases beyond translation. *Nature Rev. Mol. Cell Biol.,* **11:**668-674.

Gupta, S. K., Hury, A., Ziporen, Y., Shi, H., Ullu, E., Michaeli, S. (2010). Small nucleolar RNA interference in *Trypanosoma brucei*: mechanism and utilization for elucidating the function of snoRNAs. *Nucleic Acids Res.,* **28(15):**2855–2861.

Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, **29(8):** 1072-1075. doi: 10.1093/bioinformatics/btt086

Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies, *Bioinformatics*, **29 (8):** 1072-1075. doi: 10.1093/bioinformatics/btt086

Gurevitz, M., Kratz, H., Ohad, I. (1977). Polypeptides of chloroplastic and cytoplastic origin required for development of photosystem II activity and chlorophyll-protein complexes in *Euglena gracilis* Z. chloroplast membranes. *Biochim. Biophys. Acta,* **461:** 475 – 488.

Gurney, T. (1985). Characterization of mouse 45*S* ribosomal subspecies suggests that the first processing cleavage occurs 600 +/- nucleotides from the 5' end and the second 500 +/- nucleotides from the 3'd end of a 13.9 kb precursor. *Nucleic Acids Res.,* **13:**4905-4919.

Guttman, H. N. (1971). Internal Cellular Details of Euglena gracilis Visualized by Scanning Electron Microscopy. *Science, New Series,* **171(3968):** 290-292.

Guttman, H. N., Ziegler, H. (1974). Clarification of structures related to function in *Euglena gracilis. Cytobiologie*, **9:**10–22.

Haapal. A. K. and Soyer, M. O. (1975). Organization of chromosome fibrils in *Euglena gracilis. Hereditas,* **80:**185-194.

Hackett, J. D., Yoon, H. S., Li, S., Reyes-Prieto, A., Rümmele, S. E. and Bhattacharya D. (2007). Phylogenomic Analysis Supports the Monophyly of Cryptophytes and Haptophytes and the Association of Rhizaria with Chromalveolate. *Mol. Biol. Evol.,* **24(8):**1702–1713. doi:10.1093/molbev/msm08

Hadariová, L., Vesteg, M., Bircák, E., Schwartzbach, S. D., Krajcovic, J. (2016). An intact plastid genome is essential for the survival of colorless *Euglena longa* but not *Euglena gracilis. Curr. Genet.,* **63(2):**331-341. DOI 10.1007/s00294-016-0641-z

Hader, D. P., Lebert, M., Di Lena, M. R. (1987). New evidence for the mechanism of phototactic orientation of *Euglena gracilis. Curr. Microbiol.,* **14:**157–163.

Häder, D. P., Richter, P., Schuster, M., Daiker, V., Lebert, M. (2009). Molecular analysis of the graviperception signal transduction in the flagellate *Euglena gracilis*: involvement of a transient receptor potential-like channel and a calmodulin. *Advances in Space Research,* **43:** 1179–1184.

Hader, D-P, Hemmersbach, R., Lebert, M. (2005a). Gravity and the behavior of unicellular organisms. Cambridge University Press, Cambridge.

Hagiwara, S. Y., Bolige, A., Zhang, Y., Takahashi, M., Yamagishi, A., Goto, K. (2002). Circadian gating of photoinduction of commitment to cell-cycle transitions in relation to photoperiodic control of cell reproduction in Euglena. *Photochem. Photobiol*., **76:** 105–115.

Hagiwara, S., Takahashi, M., Yamagishi, A., Zhang, Y., Goto, K. (2001). Novel findings regarding photoinduced commitments of G1-, S- and G2-phase cells to cellcycle transitions in darkness and dark-induced G1-, S- and G2-phase arrests in Euglena. *Photochem. Photobiol.,* **74:** 726–733.

Haile, S. and Papadopoulou, B. (2007). Developmental regulation of gene expression in trypanosomatid parasitic protozoa. *Curr. Opin. Microbiol.,*

Haile, S., Cristodero, M., Clayton, C., Estevez, A. M. (2007). The subcellular localization of trypanosome RRP6 and its association with the exosome. *Mol Biochem Parasitol,* **151:**52–58.

Haimo, L.T. and Rosenbaum,J. L. (1981). Cilia, flagella, and microtubules. *J. Cell Biol.,* **91:**125s–130s.

Hajduk, S. L., Harris, M. E., Pollard, V. W. (1993). RNA editing in kinetoplastid mitochondria. *FASEB J.,* **7:**54–63

Halldal, P., (1958). Action Spectra of Phototaxis and Related Problems in Volvocales, Ulva-Gametes and Dinophyceae. *Physiol. Plantarum,* **11:** 118-153.

Hallick, R. B., Hong, L., Drager, R. G., Favreau, M. R., Monfort, A., Orsat, B., Spielmann, A. and Stutz, E. (1993). Complete sequence of *Euglena gracilis* chloroplast DNA. *Nucleic Acids Res.,* **21:**3537–3544.

Hameroff, S. R. (1987). Ultimate Computing. North-Holland, Amsterdam

Hamilton, F. D. (1974). Ribonucleotide reductase from *Euglena gracilis*. *J. Biol. Chem.,* **249:**4428–4434.

Hammond, S. M., Bernstein, E., Beach, D., Hannon, G. J. (2000). An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. *Nature*, **404:** 293–296.

Hampl, V., Hug, L., Leigh, J. W., Dacks, J. B., Lang, B. F., Simpson, A. G. & Roger, A. J. (2009). Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic"supergroups". *Proc. Natl. Acad. Sci.* USA, **106:**3859–3864.

Hannaert, V., and Michels P. (1994). Structure, function and biogenesis of glycosomes in Kinetoplastida. *J. Bioenerg. Biomembr.,* **20**:205–212.

Hannaert, V., Brinkmann, H., Nowitzki, U., Lee, J. A., Albert, M., Sensen, C. W., Gaasterland, T., Muller, M., Michels, P., and Martin, W. (2000). Enolase from Trypanosoma brucei, from the amitochondriate protest Mastigamoeba balamuthi, and from the chloroplast and cytosol of Euglena gracilis: pieces in the evolutionary puzzle of the eukaryotic glycolytic pathway. *Mol. Biol. Evol.,* **17:**989–1000. doi:10.1093/oxfordjournals.molbev.a026395.

Hannaert, V., Saavedra, E., Duffieux, F., Szikora, J. P., Rigden, D. J., Michels, P. A., Opperdoes, F. R. (2003). Plantlike traits associated with metabolism of Trypanosoma parasites. *Proc. Natl. Acad. Sci. USA,* **100:**1067–1071.

Hansen, J. L., Schmeing, T. M., Moore, P. B., and Steitz, T. A. (2002). Structural insights into peptide bond formation. *Proc. Natl. Acad. Sci. U. S. A.,* **99:** 11670–11675. http://dx.doi.org/10.1073/pnas.172404099.

Hanson, S. J., Schurko, A. M., Hecox-Lea, B., Welch, D. B. M., Stelzer,C., and Logsdon, J. M. Jr. (2013). Inventory and Phylogenetic Analysis of Meiotic Genes in Monogonont Rotifers. *Journal of Heredity,* **1 – 14.** doi:10.1093/jhered/est011

Hanson, S. J., Stelzer, C., Welch, D. B. M. and Logsdon, J. M. Jr (2013). Comparative transcriptome analysis of obligately asexual and cyclically sexual rotifers reveals genes with putative functions in sexual reproduction, dormancy, and asexual egg production. Hanson et al. *BMC Genomics,* **14:**412. DOI: 10.1186/1471-2164-14-412

Haribabu, B., Hook, S. S., Selbert, M. A., Goldstein, E. G., Tomhave, E. D., Edelman, A. M., Snyderman, R., Means, A. R. (1995). "Human calcium-calmodulin dependent protein kinase I: cDNA cloning, domain structure and activation by phosphorylation at threonine-177 by calcium-calmodulin dependent protein kinase I kinase". *EMBO J.,* **14(15):** 3679-3686.

Harri Lempiainen and David Shore (2009). Growth control and ribosome biogenesis. *Current Opinion in Cell Biology,* **21:**855 – 863.

Hartl, D. L., Freifelder, D. and Snyder, L. A. (1988). Basic Genetics. *Boston: Jones and Bartlett Publishers.*ISBN 0-86720-090-1*.*

Hartshorne, T., Agabian, N. (1993). RNA B is the major nucleolar trimethylguanosine-capped small nuclear RNA associated with fibrillarin and pre-rRNAs in *Trypanosoma brucei*. *Mol. Cell. Biol.,* 13:144 -154.

Hasegawa, K., Kikuchi, H., Ishizaki, S., Tamura, A., Tsukahara, Y., Nakaoka, Y., Iwai, E., Sato, T., 1999. Simple fluctuation of Ca2+ elicits the complex circadian dynamics of cyclic AMP and cyclic GMP in Paramecium. *J. Cell Sci.,* **112:**201–207.

Havilio, M., Levanon, E. Y., Lerman, G., Kupiec, M. and Eisenberg, E. (2005). Evidence for abundant transcription of non-coding regions in the *Saccharomyces cerevisiae* genome. *BMC Genomics*, **6:**93 doi:10.1186/1471-2164-6-93.

Hayashi, H., Narumi, I., Wada, S., Kikuchi, M., Furuta, M., Uehara, K., and Watanabe, H. (2004). Light dependency of resistance to ionizing radiation in *Euglena gracilis*. *Journal of Plant Physiology*, **161:**1101-1106.

Hayashi-Isimaru, Y., Ueda, K., Nonaka, M. (1993). Detection of DNA in the nucleoids of chloroplasts and mitochondria in *Euglena graciIis* by immunoelectron microscopy. *J. Cell Sci.,* **105:**1159 – 1164.

Hazkani-Covo, E., Zeller, R. M. and Martin, W. (2010) Molecular poltergeist: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet.,* **6:**e1000834.

He, C. Y., Pypaert, M., Warren, G. (2005). Golgi duplication in *Trypanosoma brucei* requires Centrin2. *Science,* **310**:1196 –1198.

Heijde, M. and Ulm, R. (2012). UV-B photoreceptor-mediated signaling in plants. *Trends Plant Sci.,* **17:** 230–237.

Heim, K. E., Tagliaferro, A. R., Bobilya, D. J. (2002). Flavonoid antioxidants: chemistry, metabolism and structure-activity relationships. *J. Nutr. Biochem.,* **13:**572 - 584.

Heitman, J. (2006). Sexual reproduction and the evolution of microbial pathogens. Curr. Biol.,**16:**R711–R725.

Heitman, J. (2010). Evolution of eukaryotic microbial pathogens via covert sexual reproduction. *Cell Host Microbe,* **8:**86–99.

Hellis, D. V., Hellis, P. F., Kernick, W. A., Phillips, G. O. (1980). The stigma of *Euglena gracilis* strain Z: an investigation into the possible occurrence of carotenoproteins and nucleid acids. *Cytobios.,* **29:** 135–140.

Hellis, D. V., Kernick, W., Phillips, G.O., Davies, K. (1979). Separation and identification of the carotenoid pigments of stigmata isolated from light-grow cells of *Euglena gracilis* strain. Z. *Arch. Microbiol.,* **121:**207 – 215.

Helmy, M., Sugiyama, N., Tomita, M., Ishihama, Y. (2012) Mass spectrum sequential subtraction speeds up searching large peptide MS/MS spectra datasets against large nucleotide databases for proteogenomics. *Genes Cells*, **17:** 633–644.

Helmy, M., Tomita, M., Ishihama, Y. (2011). Peptide identification by searching large-scale tandem mass spectra against large databases: bioinformatics methods in proteogenomics. *Genes, Genomes and Genomics,* **6:** 76–85.

Henderson, R. (1977). The purple membrane from *Halobacterium halobium. A. Rev. Bioeng.,* **6:** 87–109

Henze, K., A. Badr, M. Wettern, R. Cerff, and W. Martin (1995). A nuclear gene of eubacterial origin in *Euglena gracilis* reflects cryptic endosymbioses during protist evolution. Proc. Natl. Acad. Sci. USA, **92**:9122–9126.

Henze, K., Badr, A.,Wettern, M., Cerff, R., Martin,W. (1995). A nuclear gene of eubacterial origin in Euglena gracilis reflects cryptic endosymbioses during protist evolution. *Proc. Natl. Acad. Sci. U. S. A.,* **92:**9122–9126.

Hepler, P. K. and Fosket, D. E. (1971). The role of microtubules in vessel member differentiation in *Coleus. Protoplasma* **72:**213-236.

Herlan, G., Ekert, W. A., Kaffenberger, W., Wunderlich, F. (1979). Isolation and characterization of an RNA-containing nuclear matrix from *Tetrahymena* macronuclei, *Biochemistry,* **18:**1782-1778.

Hernández, G. (2012). On the Emergence and Evolution of the Eukaryotic Translation Apparatus, Cell-Free Protein Synthesis, Prof. Manish Biyani (Ed.), InTech, DOI: 10.5772/48568. Available from: http://www.intechopen.com/books/cell-free-protein-synthesis/on-the-emergence-and-evolution-of-the-eukaryotic-translation-apparatus

Hernández, G. (2012). On the Emergence and Evolution of the Eukaryotic Translation Apparatus. *Cell-Free Protein Synthesis*, 31 – 50. http://dx.doi.org/10.5772/48568.

Hernández, G., Altmann, M., and Lasko, P. (2010) Origins and evolution of the mechanisms regulating translation initiation in eukaryotes. *Trends Biochem. Sci.* **35:**63-73.

Hernández, G., Proud, C. G., Preiss, T., and Parsyan, A. (2012). On the Diversification of the Translation Apparatus across Eukaryotes. *Comparative and Functional Genomics,* **2012(256848):** 14 pages. doi:10.1155/2012/256848

Hershey, J. W. B., and Merrick, W. C. (2000). Pathway and mechanism of initiation of protein synthesis, in *Translational control of gene expression* (Sonenberg, N., Hershey, J. W. B., and Mathews, M. B., Eds.), **pp** 33-88, Cold Spring Harbor Laboratory press, Cold Spring Harbor, New York.

Hess, W. R. and Borner, T. (1999). Organellar RNA polymerases of higher plants. *Int. Rev. Cytol.*, **190**, 1–59.

Hilenski, L. L. and Walne, P. L. (1985). Ultrastructure of the flagella of the colorless phagotroph Peranema trichophorum (Euglenophyceae. II. Flagellar roots). *J Phycol*., **21:**125–134.

Hilenski, L. L., Walne, P. L. (1985). Ultrastructure of the flagella of the colorless phatotroph *Peranema trichophorum* (Euglenophyceae). I: Flagellar mastigonemes, *J. Phycol.,* **21:**114–125

Hileski, L. L., Walne, P. L., Synder, F. (1976). Aliphatic chains of esterifield lipids in isolated eyespot of *Euglena gracilis* var. *bacillaris. Pl. Physiol.,* **57:**645–646.

Hill, H. Z., Epstein, H. T.and Schiff, J.A. (1966). Studies of chloroplast development in *Euglena* XIV. Sequential interactions of ultraviolet light and photoreactivating light in green colony formation. *Biophys. J .,***6:** 135 - 144.

Hiller, R. G., Wrench, P. M., Sharples, F. P. (1995). The light-harvesting chlorophyll a-c-binding protein of dinoflagellates: a putative polyprotein. *FEBS Lett.*, **363:**175–178.

Hinnebusch, A. G. (2011) Molecular mechanism of scanning and start codon selection in eukaryotes. *Microbiol. Mol. Biol. Rev.,**75*:434-467.

Hirst, J., Barlow, L. D., Francisco, G. C., Sahlender, D. A., Seaman, M. N. J., Dacks, J. B., Robinson, M. S. (2011). The Fifth Adaptor Protein Complex. *PLoS Biology,* **9(10):** e1001170

Hirst, J., Barlow, L. D., Francisco, G.C., Sahlender, D. A., Seaman, M.N.J., Dacks, J. B., Robinson, M. S. (2011). The Fifth Adaptor Protein Complex. *PLOS Biology.* 10(3):10. doi:10.1371/journal.pbio.1001170.

Hirst, J., Barlow, L.D., Francisco, G.C., Sahlender, D.A., Seaman, M.N.J., Dacks, J.B., Robinson, M. S. (2011). The fifth adaptor protein complex. *PLoS Biol,* **9:** e1001170.

Hirst, J., Schlacht, A., Norcott, J., Traynor, D., Bloomfield, G., Antrobus, R., Kay, R. R., Dacks, J. B, Robinson, M. S. (2014). Characterization of TSET, an ancient and widespread membrane trafficking complex. *eLife,* **3:**1-18. doi: 10.7554/eLife.02866.

Hirt, R. P., Alsmark, C. and Embley, T. M. (2015). Lateral gene transfers and the origins of the eukaryote proteome: a view from microbial parasites. *Curr. Opin. Microbiol.,* **23:**155–162.

Hodge, C. A., Tran, E. J., Noble, K. N., Alcazar-Roman, A. R., Ben-Yishay, R., Scarcelli, J. J., Folkmann, A. W., Shav-Tal, Y., Wente, S. R., and Cole, C. N. (2011). The Dbp5 cycle at the nuclear pore complex during mRNA export I: dbp5 mutants with defects in RNA binding and ATP hydrolysis define key steps for Nup159 and Gle1. Genes & Development, **25:**1052–1064.

Hoffmeister, M., van der Klei, A., Rotte, C., van Grinsven, K. W. A., van Hellemond, J. J., Henze, K., Tielens, A. G. M. and Martin, W. (2004). *Euglena gracilis* rhodoquinone: ubiquinone ratio and mitochondrial proteome differ under aerobic and anaerobic conditions. *J. Biol. Chem.,* **279:**22422 – 22429.

Hoffmeister, M., van der Klei, A., Rotte, C., van Grinsven, K. W., van Hellemond, J. J., Henze, K., Tielens, A. G., Martin, W. (2004). *Euglena gracilis*

rhodoquinone:ubiquinone ratio and mitochondrial proteome differ under aerobic and anaerobic conditions. *J. Biol. Chem.,* **279**:22422–22429.

Hofmann, C., Bouck, B. (1976). Immunological and structural evidence for patterned intussusceptive surface growth in a unicellular organism*. J. Cell Biol.,* **69:** 693- 715.

Holden, J. M., Koreny, L., Obado, S., Ratushny, A. V., Chen, W., Chiang, J., Kelly, S., Chait, B. T., Aitchison, J. D., Rout, M. P., and Field, M. C. (2014). *Mol. Biol. Cell,* **25(9):**1421-36. doi: 10.1091/mbc.E13-12-0750.

Hollande, A. (1942). *Arch. zool. exp. et gen.* **83:** 1

Holt, S. C., and Stern, A. I. (1970). The effect of 3-(3,4-dichlorophenyl)-l,l-dimethylurea on chloroplast development and maintenance in *Euglena gracilis. Plant Physiol.,* **45:** 475 - 483.

Hong, Y., Maeda, Y., Watanabe, R., Ohishi, K., Mishkind, M., Riezman, H. and Kinoshita, T. (1999). Pig-n, a Mammalian Homologue of Yeast Mcd4p, Is Involved in Transferring Phosphoethanolamine to the First Mannose of the Glycosylphosphatidylinositol. *The Journal of Biological Chemistry,* **274 (49): (Issue of December 3):** 35099–35106.

Honigberg B. M. (1963). Evolutionary and systematic relationships in the flagellate Order Trichomonadida Kirby. *J. Protozool.,* **10:** 20–63.

Honts, J. E. and Williams, N. E. (1990). Tetrins: polypeptides that form bundled filaments in *Tetrahymena*. *J. Cell Sci.,* **96**:293–302.

Hori, T., and Fukagawa, T. (2012). Establishment of the vertebrate kinetochores. *Chromosome Res.,* **20:**547–561.

Horner, D. S., Hirt, R. P., Embley, T. M. (1999). A single eubacterial origin of eukaryotic pyruvate:ferredoxin oxidoreductase genes: implications for the evolution of anaerobic eukaryotes. *Mol. Biol. Evol.,* **16**:1280 –1291.

Hornett, E. A. and Wheat, C. W. (2012). Quantitative RNA-Seq analysis in non-model species: assessing transcriptome assemblies as a scaffold and the

utility of evolutionary divergent genomic reference species. *BMC Genomics,* **13:361** doi: 10.1186/1471-2164-13-361

Hou, Y. and Lin, S (2009). Distinct Gene Number-Genome Size Relationships for Eukaryotes and Non-Eukaryotes: Gene Content Estimation for Dinoflagellate Genomes. *PLoS One,* **4(9):** e6978.

Houlné, G. and Schantz*, R.* (1993). Expression of Polyproteins in Euglena. *Critical Reviews in Plant Sciences,* **12(1/2):** 1 – 17.

Houlne, G., Schantz, R. (1987). Molecular analysis of the transcripts encoding the light-harvesting chlorophyll a/b protein in *Euglena gracilis*: unusual size of the mRNA. *Curr Genet.,* **12:** 611–616.

Howard, A. and Pelc, S. R. (1953). *Symposium on Chromosome Breakage* (supplement to *Heredity,* vol. 6). Springfield, Illinois: Thomas

Howard-Till, R. A., Lukaszewicz, A., Novatchkova, M., Loidl, J. (2013). A single cohesin complex performs mitotic and meiotic functions in the protist Tetrahymena. *PLoS Genet.,* **9:**e1003418.

Hrda, S., Fousek, J., Szabova, J., Hampl, V., Vlcek, C. (2012). The plastid genome of Eutreptiella provides a window into the process of secondary endosymbiosis of plastid in euglenids. *PloS ONE*, **7(3):** e33746. [doi: 10.1002/9780470015902.a0001964.pub3

http://web.bio.utk.edu/bcmb513-1/bcmb513-1.htm

Huang, H. J., Fu, S. F., Tai, Y. H., Chou, W. C., and Huang, D. D. (2002). Expression of *Oryza sativa* MAP kinase gene is developmentally regulated and stress-responsive. Physiologia Plantarum, **114:**572-580.

Huang, J. (2013). Horizontal gene transfer in eukaryotes: the weak-link model. *Bioessays,* **35:**868–875.

Huang, K., Diener, D. R., Mitchell, A., Pazour, G. J., Witman, G. B., and Rosenbaum, J. L. (2007). Function and dynamics of PKD2 in *Chlamydomonas reinhardtii* flagella. The Journal of Cell Biology, **179(3):**501-514.

Huang, Z. P., Zhou, H., Liang, D. and Qu, L. H. (2004). Different expression strategy: multiple intronic gene clusters of box H/ACA snoRNA in *Drosophila melanogaster*. *J. Mol. Biol.,* 3**41:**669–683.

Huelsenbeck, J. P., Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics, **17(8):**754-5.

Hug, L. A., Stechmann, A., Roger, A. J. (2010). Phylogenetic distributions and histories of proteins involved in anaerobic pyruvate metabolism in eukaryotes. *Mol. Biol. Evol.,* **27**:311–324.

Hulanicka, D., Erwin, J., and Bloch, K. (1964). *Lipid Metabolism of Euglena gracilis. J. Biol. Chem.*, **239:** 2778 – 2787.

Hulbert, A. J. (2005). On the importance of fatty acid composition of membranes for aging. *J. Theor. Biol.,* **234:**277–288.

Hulbert, A. J., Pamplona, R., Buffenstein, R., Buttemer, W. A. (2007). Life and death: metabolic rate, membrane composition, and life span of animals. Physiol. Rev., **87:**1175 –1213.

Humbeck, K. and Krupinska, K. (2003). The abundance of minor chlorophyll a／b-binding proteins CP29 and LHCI of barley (Hordeum vulgare L.) during leaf senescence is controlled by light. *Journal of Experimental Botany,* **54:** 375 – 383.

Humbeck, K., Quast, S., Krupinska, K. (1996) Functional and molecular changes in the photosynthetic apparatus during senescence of flag leaves from field-grown barley plants. *Plant, Cell and Environment*, **19:** 337–344.

Hunter, S. H. and Provasoli, L. (1964). *Nutrition of Algae. Annu. Rev. Plant Physiol.* **15:**37–56.

Hury, J., Nagaswamy, U., Larios-Sanz, M. and Fox, G. E. (2006). Ribosome origins: the relative age of 23S rRNA domains. *Orig. Life Evol. Biosphere,* **36:**421–429.

Hutagalung, A.H. and Novick, P.J. (2011). Role of Rab GTPases in Membrane Traffic and Cell Physiology. *Physiol Rev.,* **91:**119–149. doi:10.1152/physrev.00059.2009.

Hutner, S. H., Zahalsky, A. C., Aaronson S.,Baker, H., and Frank, O. (1966). *In Methods in Cell Physiology.* D. M. Prescott, editor. Academic PressInc., New York and London, **2:**217.

Huttenlauch, I. and Peck, R. K. (1991). The membrane skeleton of *Pseudomicrothorax*. II. Biochemical and immunological characterization of the glycosylated cytoskeletal proteins. *J. Cell Sci.,* **100**:707–715.

Huttenlauch, I., Peck, R. K. and Stick, R. (1998). Articulins and epiplasmins: two distinct classes of cytoskeletal proteins of the membrane skeleton in protists. *Journal of Cell Science,* **111:**3367-3378.

Hyams, J. S. (1982). The *Euglena* paraflagellar rod: structure, relationship to the other flagellar components and preliminary biochemical characterization. *J. Cell Sci.,* **55:**199–210

Ibarra, A. and Hetzer, M. W. (2015). Nuclear pore proteins and the control of genome functions. *Genes & Development,* **29:**337–349.

Igamberdiev, A. U. and Kleczkowski, L. A. (2015). Optimization of ATP synthase function in mitochondria and chloroplasts via the adenylate kinase equilibrium. *Frontiers in Plant Science,* **6(10):**1-8.

Illumina coverage: http://support.illumina.com/downloads/sequencing_coverage_calculator.html

Inagaki, J., Fujita, Y., Hase, T. and Yamamoto, Y. (2000). Protein translocation within chloroplast is similar in Euglena and higher plants. *Biochem. Biophys. Res. Commun.,* **277:**436–442.

Inagaki, Y., and Doolittle, W. F. (2001). Class I release factors in ciliates with variant genetic codes. *Nucleic Acids Res.,* **29:**921-927.

International Human Genome Sequencing Consortium (IHGSC) (2001). Initial sequencing and analysis of the human genome. A landmark paper describing the human 'rough draft' (public version) and its annotation. *Nature,* **409:**860 – 921.

International Transporter Consortium, Giacomini, K. M., Huang, S. M., Tweedie, D. J, Benet, L. Z., Brouwer, K. L., Chu, X., Dahlin, A., Evers, R., Fischer, V., Hillgren, K. M., Hoffmaster, K. A., Ishikawa, T., Keppler, D., Kim, R. B., Lee, C. A., Niemi, M., Polli, J. W., Sugiyama, Y., Swaan, P. W., Ware, J. A., Wright, S. H., Yee, S. W., Zamek-Gliszczynski, M. J. and Zhang, L. (2010). Membrane transporters in drug development. *Nat. Rev. Drug Discov.,* **9(3):**215-36. doi: 10.1038/nrd3028.

Inui, H., Miyatake, K., Nakano, Y., and Kitaoka, S. (1984). Fatty acid synthesis in mitochondria of *Euglena gracilis*. *Eur. J. Biochem.,* **142:**121–126. doi: 10.1111/j.1432-1033.1984.tb08258.x

Inui, H., Miyatake, K., Nakano, Y., Kitaoka S. (1984). Fatty acid synthesis in mitochondria of *Euglena gracilis*. *Eur. J. Biochem.,* **142**:121–126.

Inui, H., Miyatake, K., Nakano, Y., Kitaoka, S. (1982). Wax ester fermentation in *Euglena gracilis*. *FEBS Lett.,* **150**:89 –93.

Inui, H., Miyatake, K., Nakano, Y., Kitaoka, S. (1983). Production and composition of wax esters by fermentation of *Euglena gracilis*. *Agric. Biol. Chem.,* **47**:2669 – 2671.

Inui, H., Miyatake, K., Nakano, Y., Kitaoka, S. (1985). The physiological role of oxygen-sensitive pyruvate dehydrogenase in mitochondrial fatty acid synthesis in *Euglena gracilis*. *Arch. Biochem. Biophys.,* **237**:423– 429.

Inui, H., Ono, K., Miyatake, K., Nakano, Y., Kitaoka, S. (1987). Purification and characterization of pyruvate:NADP oxidoreductase in *Euglena gracilis*. *J. Biol. Chem.,* **262**:9130 –9135.

Inui, H., Yamaji, R., Saidoh, H., Miyatake, K., Nakano, Y., Kitaoka, S. (1991). Pyruvate:NADP oxidoreductase from *Euglena gracilis*: limited proteolysis of the enzyme with trypsin. *Arch. Biochem. Biophys.,* **286**:270 –276.

Irm Huttenlauch and Reimer Stick (2003). Occurrence of Articulins and Epiplasmins in Protists. *J. Eukaryot. Microbiol.,* **50(1):**15–18.

Isegawa, U., Nakano, Y., and Kitaoka, S. (1984). Conversion and Distribution of Cobalamin in *Euglena gracilis Z*, with Special Reference to Its Location and Probable Function within Chloroplasts. *Plant Physiol.,***76:** 814-818.

Iseki, M., Matsunaga, S., Murakami, A., Ohno, K., Shiga, K., Yoshida, K., Sugai, M., Takahashi, T., Hori, T. and Watanabe, M. (2002) A blue-light-activated adenylyl cyclase mediates photoavoidance in Euglena gracilis. *Nature*, **415:** 1047–1051.

Iseki, M., Matsunaga, S., Murakami, A., Ohno, K., Shiga, K., Yoshida, C., Sugai, M., Takahashi, T., Hori, T., Watanabe, M. (2002). A blue-lightactivated adenylyl cyclase mediates photoavoidance in *Euglena gracilis*. *Nature,* **415:**1047–1051.

Itzen, A. and Goody, R. S. (2011). GTPases involved in vesicular trafficking: Structures and mechanisms. *Seminars in Cell & Developmental Biology,* **22(1):** 48–56.

Ivens, A. C., Peacock, C. S., Worthey, E. A., Murphy, L., Aggarwal, G., Berriman, M., Sisk, E., Rajandream, M. A., Adlem, E., Aert, R., Anupama, A., Apostolou, Z., Attipoe, P., Bason, N., Bauser, C., Beck, A., Beverley, S. M., Bianchettin, G., Borzym, K., Bothe, G., Bruschi, C. V., Collins, M., Cadag, E., Ciarloni, L., Clayton, C., Coulson, R. M., Cronin, A., Cruz, A. K., Davies, R. M., De Gaudenzi, J., Dobson, D. E., Duesterhoeft, A., Fazelina, G., Fosker, N., Frasch, A. C., Fraser, A., Fuchs, M., Gabel, C., Goble, A., Goffeau, A., Harris, D., Hertz-Fowler, C., Hilbert, H., Horn, D., Huang, Y., Klages, S., Knights, A., Kube, M., Larke, N., Litvin, L., Lord, A., Louie, T., Marra, M., Masuy, D., Matthews, K., Michaeli, S., Mottram, J. C., Müller-Auer, S., Munden, H., Nelson, S., Norbertczak, H., Oliver, K., O'neil, S.,

Pentony, M., Pohl, T. M., Price, C., Purnelle, B., Quail, M. A., Rabbinowitsch, E., Reinhardt, R., Rieger, M., Rinta, J., Robben, J., Robertson, L., Ruiz, J. C., Rutter, S., Saunders, D., Schäfer, M., Schein, J., Schwartz, D. C., Seeger, K., Seyler, A., Sharp, S., Shin, H., Sivam, D., Squares, R., Squares, S., Tosato, V., Vogt, C., Volckaert, G., Wambutt, R., Warren, T., Wedler, H., Woodward, J., Zhou, S., Zimmermann, W., Smith, D. F., Blackwell, J. M., Stuart, K. D., Barrell, B. and Myler, P. J. (2005). *The genome of the kinetoplastid parasite, Leishmania major*. Science (New York, N.Y.), **309 (5733):**436 – 42.

Iyer, L. M., Anantharaman, V., Wolf, M. Y., Aravind, L. (2008). Comparative genomics of transcription factors and chromatin proteins in parasitic protists and other eukaryotes. *International Journal for Parasitology*, **38:** 1–31.

Jack, A. P. M., and Hake, S. B. (2014). Getting down to the core of histone modifications. *Chromosoma,* **123(4):**355-71 DOI 10.1007/s00412-014-0465-x

Jackson, A. P., Otto T. D., Aslett, M., Armstrong, S. D., Bringaud, F., Schlacht, A., Hartley, C., Sanders, M., Wastling, J. M., Dacks, J. B., Acosta-Serrano, A., Field, M. C., Ginger, M. L., and Berriman, M. (2016). Kinetoplastid Phylogenomics Reveals the Evolutionary Innovations Associated with the Origins of Parasitism. *Current Biology,***26:** 161–172.

Jackson, A. P. (2010). The Evolution of Amastin Surface Glycoproteins in Trypanosomatid Parasites. *Mol. Biol. Evol.,* **27(1):**33–45.

Jackson, A. P. (2016). Gene family phylogeny and the evolution of parasite cell surfaces. *Molecular & Biochemical Parasitology,* **(1-2):**64-75. doi: 10.1016/j.molbiopara.2016.03.007.

Jackson, A. P., Allison, H. C., Barry, J. D., Field, M. C., Hertz-Fowler, C., Berriman, M. (2013). A Cell-surface Phylome for African Trypanosomes*. PLoS Negl. Trop. Dis.,* **7(3):** e2121. doi:10.1371/journal.pntd.0002121.

Jackson, A. P., Berry, A., Aslett, M., Allison, H. C., Burton, P., Vavrova-Anderson, J., Brown, R., Browne, H., Corton, N., Hauser, H., Gamble, J., Gilderthorp, R., Marcello, L., McQuillan, J., Otto, T. D., Quail, M. A., Sanders, M. J., van Tonder, A., Ginger, M. L., Field, M. C., Barry, J. D., Hertz-Fowler, C., and Berriman, M. (2012). Antigenic diversity is generated by distinct evolutionary mechanisms in African trypanosome species. *PNAS*, **109(9):** 3416–3421.

Jackson, A. P., Otto, T. D., Aslett, M., Armstrong, S. D., Bringaud, F., Schlacht, A., Hartley, C., Sanders, M., Wastling, J. M., Dacks, J. B., Acosta-Serrano, A., Field, M. C., Ginger, M. L., and Berriman, M. (2016). Kinetoplastid Phylogenomics Reveals the Evolutionary Innovations Associated with the Origins of Parasitism. Current Biology, 26, 161–172.

Jackson, A. P., Otto, T. D., Darby, A., Ramaprasad, A., Xia, D., Echaide, I. E., Farber, M., Gahlot, S., Gamble, J., Gupta, D., Gupta, Y., Jackson, L., Malandrin, L., Malas, T. B., Moussa, E., Nair, M., Reid, A. J., Sanders, M., Sharma, J., Tracey, A., Quail, M. A., Weir, W., Wastling, J. M., Hall, N., Willadsen, P., Lingelbach, K., Shiels, B., Tait, A., Berriman, M., Allred, D. R. and Pain, A. (2015). The evolutionary dynamics of variant antigen genes in Babesia reveal a history of genomic innovation underlying host–parasite interaction. *Nucleic Acids Research,* **42(11):**7113–7131.

Jackson, A. P., Quail, M. A., Berriman, M. (2008). Insights into the genome sequence of a free-living Kinetoplastid: *Bodo saltans* (Kinetoplastida: Euglenozoa), **9:**594.

Jackson, A. P., Sanders, M., Berry, A., McQuillan, J., Aslett, M. A., Quail, M. A., Chukualim, B., Capewell, P., MacLeod, A., Melville, S. E., Gibson, W., Barry, J. D., Berriman, M. and Hertz-Fowler, C. (2010). The genome sequence of Trypanosoma brucei gambiense, causative agent of chronic human african trypanosomiasis. *PLoS neglected tropical diseases,***4(4):**e658

Jackson, A. P., Vaughan, S. and Gull, K. (2006). Evolution of Tubulin Gene Arrays in Trypanosomatid parasites: genomic restructuring in *Leishmania*.*BMC Genomics,* 7:261.

Jackson, R. J., Hellen, C. U. T., and Pestova, T. V. (2012). Termination and postterminationevents in eukaryotic translation. *Advances Prot. Chem. Struct. Biol.,* **86:**45-93.

Jackson, R. J., Hellen, C. U., and Pestova, T. V. (2010). The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat. Rev. Mol. Cell Biol.,* **11:**113-127.

Jacob, S. W., Rosenbaum, E. E., Wood, D. C. (1971). Dimethyl sulfoxide. *New York: Marcel Dekker Inc.*

Jaé, N., Preußer, C., Krüger, T., Tkacz, I. D., Engstler, M., Michaeli, S., and Bindereif,A. (2011). snRNA-specific role of SMN in trypanosome snRNP biogenesis in vivo. *RNA Biology,* **8(1):** 90-100, DOI: 10.4161/rna.8.1.13985.

Jahn, T. (1946). The Euglenoid Flagellates. *The Quarterly Review of Biology, 21*(3), 246-274. Retrieved from http://www.jstor.org/stable/2812264

James, R. Y. Rawson, Virginia K. Eckenrode, Cindy L. Boerma, and Stephanie Curtis (1979). DNA Sequence Organization in the Alga *Euglena gracilis. Biochimica et Biophysica Acta,* **563:** 1 – 16.

Jardine, N. J. and Leaver, J. L. (1978). “The fractionation of histones isolated from *Euglena gracilis*,” *Biochemical Journal*, **169(1):** 103–111.

Jeffares D. C., Tomiczek B., Sojo V., dos Reis M. (2015). A Beginners Guide to Estimating the Non-synonymous to Synonymous Rate Ratio of all Protein-Coding Genes in a Genome. In: Peacock C. (eds) Parasite Genomics Protocols. *Methods in Molecular Biology,* vol 1201. Humana Press, New York, NY. https://doi.org/10.1007/978-1-4939-1438-8_4

Jennings, H. S. (1906). Behaviour of the Lower Organisms. Columbia University Press, New York

Jensen, M. K., Vogt, J. K., Bressendorff, S., Seguin-Orlando, A., Petersen, M., Sicheritz-Pontén, T., Mundy, J. (2015) Transcriptome and Genome Size Analysis of the Venus Flytrap. *PLoS ONE*, **10(4):** e0123887. doi:10.1371/journal.pone.0123887\

Jensen, R. E. and Englund, P. T. (2012). Network news: the replication of kinetoplast DNA. *Annu Rev Microbiol,* **66:**473–91.

Ji, P., Liu, G., Xu J, Wang X, Li J, Zhao, Z., Zhang, X., Zhang, Y., Xu, P., Sun, X. (2012). Characterization of Common Carp Transcriptome: Sequencing, De Novo Assembly, Annotation and Comparative Genomics. *PLoS ONE,* **7(4):** e35152.

Jianfan Wen (2000). The nuclear matrix of *Euglena gracilis* (Euglenophyta): A stage of nuclear matrix evolution? *Biology of the Cell,* **92:**125-131

Jiang, H. and Wong, W.H. (2009). Statistical inferences for isoform expression in RNA-seq.
*Bioinformatics,* **25:**1026 –1032.

Jo, B. and Choi, S. S. (2015). Introns: The Functional Benefits of Introns in Genomes. *Genomics Inform.,* **13(4):**112-118.

Johmann, C, Eckhardt, R. A. and Gorovsky, M. A., (1973). The histones associated with condensed and extended chromatin of mouse liver. *J cell boil.,* 58:119-125.

John, D. M., Whitton, B. A. and Brook, A. J. (eds) (2002). *The Freshwater Algae of the British Isles An Identification Guide to Freshwater and Terrestrial Algae.* Cambridge University Press, Cambridge. 702 **pp.**

Johnson, L. S., Eddy, S. R., Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM searches procedure. *BMC Bioinformatics*. **11:**431

Jones, P., Binns, D., Chang, HY., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, SY, Rodrigo, L., and Hunter, S.

(2014**).** InterProScan 5: genome-scale protein function classification. *Bioinformatics,* **30(9):**1236-40. *doi:10.1093/bioinformatics/btu031.*

Jordan, A. and Reichard, P. (1998). Ribonucleotide reductases. *Annu. Rev. Biochem.* 67:71–98.

Joshua-Tor, L. and Hannon, G. J. (2011). Ancestral roles of small RNAs: an ago-centric perspective. *Cold Spring Harb Perspect Biol* 2010 published online September 1st 2011.

Junier, T., Zdobnov, E. M. (2010). The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics,* **26(13):**1669-70. doi: 10.1093/bioinformatics/btq243.

Jurica, M. S. and Moore, M. J. (2003). Pre-mRNA splicing: awash in a sea of proteins. *Mol. Cell,* **12:**5–14.

Kabachinski, G. and Schwartz, T. U. (2015). The nuclear pore complex—structure and function at a glance. *J. Cell Sci.,* **128(3):**423–9.

Kamikawa, R., Inagaki, Y., and Sako, Y. (2008). Direct phylogenetic evidence for lateraltransfer of elongation factor-like gene. *Proc. Natl. Acad. Sci. U. S. A.,* **105:** 6965-6969.

Kamikawa, R., Sakaguchi, M., Matsumoto, T., Hashimoto, T., and Inagaki, Y. (2010) Rooting for the root of elongation factor-like protein phylogeny. *Mol. Phylogenet. Evol.,* **56:**1082-1088.

Kamikawa, R., Yabuki, A., Nakayama, T., Ishida, K., Hashimoto, T., and Inagaki, Y.(2011). Cercozoa comprises both EF1α-containing and EFL-containing members. *Eur. J.Protistol.,* 47:24-28.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.,* **38:** D355–D360. doi:10.1093/nar/gkp896.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.,* **34:**D354–D357.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res.,* **32:** D277–D280. doi:10.1093/nar/gkh063.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res.,* **32:** D277–D280. doi:10.1093/nar/gkh063.

Kang, X., Rogers, K., Gao, G., Falick, A.M., Zhou, S., Simpson, L. (2005). Reconstitution of uridine-deletion precleaved RNA editing with two recombinant enzymes. *Proc. Natl. Acad. Sci. USA,* **102:**1017–1022.

Kapp, L. D., and Lorsch, J. R. (2004). The molecular mechanics of eukaryotic translation. *Annu. Rev. Biochem., **73:***657-704.

Karpov, S. A. (1990) *System of Protista. Biol.* Nauch. Issled. Inst., Omsk, Russia (In Russian with English summary)

Karrer, P., Jucker, E., Rutschmann, J., and Steinlin, K., Helv. Chim. (1945). α-Carotin-mono-epoxyd und Flavochrom. *Acta,* 26: 1146.

Kartenbeck, J. and Franke, W. W. (1971). Dense cytoplasmic aggregates associated with Golgi apparatus cisternae of rat hepatocytes. *Protoplasma,* **72:**49 - 53.

Katherine, P., Lemon and Alan D., Grossman (2001). The extrusion-capture model for chromosome partitioning in bacteria. *Genes and Development,* **15:**2031–2041

Kathrin Bolte, Lars Bullmann, Franziska Hempel, Andrew Bozarth, Stefan Zauner and Uwe-G. Maier (2009). Protein Targeting Into Secondary Plastids. *J. Eukaryot. Microbiol.,* **56(1):** 9–15

Katoh, K. and Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.,* **30(4):**772–780 doi:10.1093/molbev/mst010

Kawai, G., Yamamoto, Y., Kamimura, T., Masegi, T., Sekine, M., Hata, T., Iimori, T., Watanabe, T., Miyazawa, T., and Yokoyama, S. (1992). Conformational rigidity of specific pyrimidine residues in tRNA arises from posttranscriptional modifications that enhance steric interaction between the base and the 2′-hydroxyl group. *Biochemistry,* **31:**1040–1046.

Kawasaki, H. and Kretsinger, R. H. (1995). Calcium-binding proteins 1:EF-hands. *Protein profile,* **2(4):**297-490.

Keeling, P. J. (2004). Diversity and evolutionary history of plastids and their hosts. *Am. J. Bot.,* **91(10):**1481–1493.

Keeling, P. J., and Inagaki, Y. (2004). A class of eukaryotic GTPase with a punctuate distribution suggesting multiple functional replacements of translation elongation factor 1alpha. *Proc. Natl. Acad. Sci. U. S. A.,* **101:**15380-15385.

Keeney, S., Giroux, C. N. and Kleckner, N. (1997) Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. Cell, **88:**375–384.

Keibler, E. and Brent, M. R. (2003). Eval: A software package for analysis of genome annotations**.** *BMC Bioinformatics*, **4:50** doi:10.1186/1471-2105-4-50

Keiper, B. D., Lamphear, B. J., Deshpande, A. M., Jankowska-Anyszka, M.,Aamodt, E. J., Blumenthal, T. and Rhoads, R. E. (2000). Functional characterizationof five eIF4E isoforms in *Caenorhabditis elegans*. *J. Biol. Chem.*, **275:**10590–10596. http://dx.doi.org/10.1074/jbc.275.14.10590.

Keith Connolly, Jason P. Rife, and Gloria Culver (2008). Mechanistic insight into the ribosome biogenesis functions of the ancient protein KsgA. *Mol. Microbiol.* **70(5):** 1062–1075. doi:10.1111/j.1365-2958.2008.06485.x.

Keith, G., Ebel, J. P. and Dirheimer, G. (1974). The primary structure of tRNA trp from brewer's yeast. I. Complete digestion with pancreatic ribonuclease and T 1 ribonuclease. *FEBS Letts.,* **48:**50-52.

Keith, G., Picaud, F., Weissenbach, J., Ebel, J. P., Petrissant, G. and Dirheimer, G. (1973). The primary structure of rabbit liver tRNA Phe and its comparison with known tRNA Phe sequences. *FEBS Letts.,* **31:**345-347.

Keller, M., Chan, R. L., Tessier, L. H., Weil, J. H. and Imbault, P. (1991). Post-transcriptional regulation by light of the biosynthesis of Euglena ribulose-1,5-bisphosphate carboxylase/oxygenase small subunit. *Plant Mol. Biol.,* **17:**73–82.

Keller, O., Kollmar, M., Stanke, M., Waack, S. (2011). A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics*, **27(6):**757-63. doi: 10.1093/bioinformatics/btr010.

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols,* **10 (6):**845 – 858.

Kelly, S., Wickstead, B. and Gull, K. (2005). An in silico analysis of trypanosomatid RNA polymerases: insights into their unusual transcription. *Biochem. Soc. Trans.,* **33:**1435–1437. http://dx.doi.org/10.1042/BST20051435.

Kent, W. S. (1880-1882). A manual of the infusoria. D. Bogue, London.

Kerr, S. F. (2006). Molecular trees of trypanosomes incongruent with fossil records of hosts. *Mem Inst Oswaldo Cruz,* **101:**25–30.

Keskitalo, J., Bergquist, G., Gardestrom, P., Jansson, S. (2005). A cellular timetable of autumn senescence. *Plant Physiology,* **139:** 1635 – 1648.

Khurana, T., Brzostowski, J. A., Kimmel, A. R. (2005). A Rab21/LIM-only/CH-LIM complex regulates phagocytosis via both activating and inhibitory mechanisms. EMBO J., **24:**2254–2264.

Kiefel, B. R., Gilson, P. R. and Beech, P. L. (2004). Diverse eukaryotes have retained mitochondrial homologues of the bacterial division protein FtsZ. *Protist,* **155(1):** 105-115.

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

Kim, J. T., Boo S. M., Zakrys (1998). Floristic and Taxonomic Accounts of the Genus Euglena(Euglenophyceae) from Korean Fresh Waters. *Algae,* **13(2):**173-197

Kim, K., Lee, Y. S., and Carthew, R. W. (2007). Conversion of pre-RISC to holo-RISC by Ago2 during assembly of RNAi complexes. *RNA*, **13:** 22–29.

Kim, O. T., Yura, K., Go, N., and Harumoto, T. (2005). Newly sequenced eRF1s from ciliates: the diversity of stop codon usage and the molecular surfaces that are important for stop codon interactions. *Gene,* **346:**277-286.

Kindman, L. A., Cohen, C. E., Zeldin, M. H., Ben-Shaul, Y. and Schiff, J. A. (1978). Events Surrounding The Early Development of *Euglena* Chloroplasts. 12. Spectroscopic Examination Of The Protochlorophyll(Ide) Phototransformation in Intact Cells. *Photochem. Photobiol.,* **27:**787-794.

Kirkwood, T. B. L. (2005). Understanding the odd science of aging. *Cell*, **120:**437–447.

Kishore, R. and Schwartzbach, S. D. (1992). Translational control of the synthesis of the Euglena light harvesting chlorophyll a/b binding protein of photosystem II. *Plant Sci.,* **85:**79–89.

Kiss, T. (2001). Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. *EMBO J.*, **20(14):**3617–3622.

Kiss, T. (2002). Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. *Cell,* **109(2):**145–148.

Kister, K. P., Muller, B., Eckert, W. A. (1983). Complex endonudeolytic cleavage pattern during early events in the processing of pre-rRNA in the lower eukaryote, *Tetrahymena thermophila. Nucleic Acids Res.,* **11:**3487 - 3502.

Kitaoka, S., Y. Nakano, K. Miyatake, and Yokota, A. (1989). Enzymes and their functional location. **pp.** 2 – 135 *in* D. E. Buetow, ed. The biology of *Euglena,* Vol. 6. Subcellular biochemistry and molecular biology. Academic Press, San Diego.

Kivic, P. A. and P. L. Walne (1984). An evaluation of a possible phylogenetic relationship between the Euglenophyta and Kinetoplastida. *Origins Life,* **13:** 269 - 288.

Kivic, P. A. and Walne, P. L. (1983). Algal photosensory apparatus probably represent multiple parallel evolution. *Biosystems,* **16:** 31–38

Kivic, P. A. and Walne, P. L. (1984). An evaluation of a possible phylogenetic relationship between the Euglenophyta and Kinetoplastida. *Origins Life,* **13:**269-288.

Kivic, P. A., Vesk, M. (1974). The structure of the eyespot apparatus in bleached strains of *Euglena gracilis Cytobiology,* **10:**88–101

Kivic, P. A., Vesk, M. (1972). Structure and function in the euglenoid eyespot apparatus: the fine structure, and response to environmental changes. *Planta,* 105, **pp.** 1–14

Kivic, P. A., Walne, P. L. (1984). An evaluation of a possible phylogenetic relationship between the euglenophyta and kinoplastida. *Origins of life*, **13:**269–288.

Klebs, G. (1883). Über die Organization einiger Flagellatengruppenund ihre Beziehungen zu Algen and Infusorien.*Untersuchungen Bot Inst Tübingen*, 1:233–362 (in German)

Klein, D. J., Moore, P. B., and Steitz, T. A. (2004). The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit. *J. Mol. Biol.,* **340:**141–177. http://dx.doi.org/10.1016/j.jmb.2004.03.076.

Klein, S., Schiff, J. A, and Holowinsky, A. W. (1971) Events surrounding the early development of *Euglena* chloroplasts. 2. Normal development of fine structure and the consequences of pre-illumination. *Plant Physiol*., **45:** 339 - 347.

Klein, S., Schifv, J. A., and Holowinsiiy, A. W. (1971). Events surrounding the early development of *Euglena* chloroplasts. 2. Normal development of fine structure and the consequences of preillumination. *Plant Physiol.,* **45:** 339 - 347.

Klinger, C., Klute, M., and Dacks, J. (2013). Comparative Genomic Analysis of Multi-Subunit Tethering Complexes Demonstrates an Ancient Pan-Eukaryotic Complement and Sculpting in Apicomplexa. *PLoS ONE*. **8(9):**1-15. Doi: 10.1371/journal.pone.0076278.

Knowles, A. and Dartnall, H. J. A. (1977). In: H. Davson (Ed.), The Eye, 2B Academic Press, New York, **pp**. 74–100

Koch, M. H. J., Sayers, Z., Michon, A. M., Marquet, R., Houssier, C., Willfuhr, J. (1988). The superstructure of chromatin and its condensation mechanism. V. Effect of linker length, condensation by multivalent cations, solubility and electric dichroism properties. *Eur Biophys, J.,* **16:**177-185.

Kohli, J., and Bahler, J. (1994). Homologous recombination in fission yeast-absence of crossover interference and synaptonemal complex. *Experientia,* **50:**295–306.

Kolev, N. G., Ramey-Butler, K., Cross, G. A., Ullu, E., and Tschudi, C. (2012). Developmental progression to infectivity in *Trypanosoma brucei* triggered by an RNA-binding protein. *Science,* **338:** 1352–1353.

Kolev, N. G., Ullu,E., and Tschudi, C.(2014). The emerging role of RNA-binding proteins in the life cycle of *Trypanosoma brucei*. *Cellular Microbiology,* **16(4):**482–489.

Komachi, Y., Hatakeyama, S., Motomatsu, H., Futagami, T., Kizjakina, K., Sobrado, P., Ekino, K., Takegawa, K., Goto, M., Nomura, Y., and Oka, T. (2013). *gfsA* encodes a novel galactofuranosyltransferase involved in biosynthesis of galactofuranose antigen of *O*-glycan in *Aspergillus nidulans* and *A. fumigates. Mol Microbiol*., **90(5):** 1054–1073. doi:10.1111/mmi.12416.

Koo, J. S., and Spremulli, L. L. (1994). Effect of the secondary structure in the *Euglena gracilis* chloroplast ribulose-bisphosphate carboxylase/oxygenase messenger RNA on translational initiation. *J. Bid. Chem.,* **269:**7501-7508.

Koonin, E. V. (2009). Darwinian evolution in the light of genomics. *Nucl Acids Res.,* **37:**1011–1034.

Koonin, E. V. (2010). Preview. The incredible expanding ancestor of eukaryotes. *Cell,***140:** 606-608.

Korber B. (2000). HIV Signature and Sequence Variation Analysis. Computational Analysis of HIV Molecular Sequences, Chapter 4, pages 55-72. Allen G. Rodrigo and Gerald H. Learn, eds. Dordrecht, Netherlands: Kluwer Academic Publishers.

Koreny, L and Field, M. C. (2016). Ancient eukaryotic origin and evolutionary plasticity of nuclear lamina. *Genome Biol Evol.,* 8 (9): 2663-2671.

Koreny, L. and Field, M. (2017). Evidence for an early eukaryotic origin for nuclearlamins. *PNAS.*

Koreny, L. and Obornik, M. (2011). Sequence Evidence for the Presence of Two Tetrapyrrole Pathways in *Euglena gracilis*. *Genome Biol. Evol.,* **3:**359–364.

Kotani, H., Hoshimaru, M., Nawa, H., and Nakanishi, S. (1986). Structure and gene organization of bovine neuromedin K precursor. *Proc. Natl. Acad. Sci.USA*, **83(18):** 7074–7078.

Koumandou, V., Dacks, J., Coulson, R. and Field, M. (2007). Control systems for membrane fusion in the ancestral eukaryote; evolution of tethering complexes and SM proteins. *BMC evolutionary biology,* **7:**29. doi: 10.1186/1471-2148-7-29.

Kova, E. D., Flegontov, P., Tomas, S., and Lukes, J. (2015). Unexpectedly Streamlined Mitochondrial Genome of the Euglenozoan *Euglena gracilis. Genome Biol. Evol.,* **7(12):**3358–3367. doi:10.1093/gbe/evv229.

Kozak, M. (1989). The scanning model for translation: an update. *J. Cell Biol.,* ***108*:**22S 241.

Koziol, A. G. and Durnford, D. G. (2007). Euglena Light-Harvesting Complexes Are Encoded by Multifarious Polyprotein mRNAs that Evolve in Concert. *Mol. Biol. Evol.* **25(1):**92–100. 2008. doi:10.1093/molbev/msm232.

Koziol, A. G., Borza, T., Ishida, K. I., Keeling, P., Lee, R. W., Durnford, D. G. (2007). Tracing the evolution of the lightharvesting antennae in chlorophyll *a*/*b*-containing organisms. *Plant Physiology*, **143:** 1802–1816.

Krajcovic, J., Ebringer, L., Schwartzbach, S. D. (2002). Reversion of endosymbiosis? In Seckbach J, ed. Symbiosis: Mechanisms and Models. Cellular Origin in Extreme Habitats, Vol. 4. Dordrecht: Kluwer Academic Publisher. **pp** 185–206.

Krajcovic, J., Vesteg, M., and Schwartzbach, S. D. (2015). Euglenoid flagellates: a multifaceted biotechnology platform. *J Biotechnol.,* **202:**135–145.

Kramer, S., Piper, S., Estevez, A., and Carrington, M. (2016). Polycistronic trypanosome mRNAs are a target for the exosome. *Molecular & Biochemical Parasitology,* **205:**1–5. http://dx.doi.org/10.1016/j.molbiopara.2016.02.009

Kraus, B. L. and Spremulli, L. L. (1986). Chloroplast initiation factor 3 from *Euglena gracilis*. Identification and initial characterization. *J. Bid. Chem.,* **261:**4781 – 4784.

Krauspe, R., Lerbs, S., Parthier, B. and Wollgiehn, R. (1987). Light induction of translatable messenger rnas for the chloroplastic leucyl and valyl transfer rna synthetases of *Euglena gracilis*. *J. Plant Physiol.,* **130:**327-342.

Krawczyk, E., Rojek, J., Kowalkowska, A. K., Kapusta, M., Znaniecka, J., and Minasiewicz, J. (2016). Evidence for mixed sexual and asexual reproduction in the rare European mycoheterotrophic orchid *Epipogium aphyllum*, Orchidaceae (ghost orchid). *Ann Bot.,* **118(1):**159–172. 10.1093/aob/mcw084.

Krawiec, S., Eisenstadt, J. M. (1970b). Ribonucleic acids from the mitochondria of bleached Euglena gracilis. II. Characterization of highly polymeric ribonucleic acids. *Biochim Biophys Acta,* **217:**132–141.

Kreimeyer, A., Wielckens, K., Adamietz, P. and Hiltz, H. (1984). DNA repair-associated ADP-ribosylation *in vivo:* modification of histone HI differs from that of the principle acceptor protein. *J. biol. Chem.,* **259:** 890 - 896.

Krister Wennerberg, Kent L. Rossman, and Channing J. Der **(2005).** The Ras superfamily at a glance. *Journal of Cell Science,* **118:** 843 – 846.

Krnacova, K., Rýdlová, I., Vinarčíková, M., Krajčovič, J., Vesteg, M., Horváth, A. (2015). Characterization of oxidative phosphorylation enzymes in Euglena gracilis and its white mutant strain W(gm)ZOflL*. FEBS Lett.,* **589:**687–694.

Krnacova, K., Vesteg, M., Hampl, V., Vlcek, C., Horvath, A. (2012). *Euglena gracilis* and Trypanosomatids Possess Common Patterns in Predicted Mitochondrial Targeting Presequences. *J. Mol. Evol.,* **75 (3-4):**119-129 DOI 10.1007/s00239-012-9523-2.

Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol.,* **305(3):**567-80.

Kryazhimskiy, S., Plotkin, J. B. (2008). The Population Genetics of dN/dS. *PLOS Genetics*, **4(12):** e1000304. https://doi.org/10.1371/journal.pgen.1000304

Kuhaudomlarp, S. S., O'Neill, E. C. and Field, R. (Abstract). Identification of carbohydrate-active enzymes from *Euglena gracilis* for beta-1,3-glucan production.https://www.biochemistry.org/Portals/0/Conferences/Abstracts/SA169/SA169P008.pdf

Kunda, P. and Baum, B. (2009). The actin cytoskeleton in spindle assembly and positioning. *Trends in Cell Biology,* **19(4):** 174 – 179. DOI: http://dx.doi.org/10.1016/j.tcb.2009.01.006

Kuo, R. C., Zhang, H., Zhuang, Y., Hannick, L., Lin, S. (2013). Transcriptomic Study Reveals Widespread Spliced Leader Trans-Splicing, Short 59-UTRs and Potential Complex Carbon Fixation Mechanisms in the Euglenoid Alga Eutreptiella sp. *PLoS ONE,* **8(4):** e60826. doi:10.1371/journal.pone.0060826.

Kura-Hotta M., Satoh K., Katoh S. (1987). Relationship between photosynthesis and chlorophyll content during leaf senescence of rice seedlings. *Plant and Cell Physiology*, **28:** 1321–1329.

Kuroiwa, T., Nishida, K., Yoshida, Y., Fujiwara, T., Mori, T., Kuroiwa, H. and Misumi O. (2006). Structure, function and evolution of the mitochondrial division apparatus. *Biochim Biophys Acta,* **1763(5-6):**510–521.

Kurtz, M.M., Wang, R., Clements, M., Cascieri, M., Austin, C., Cunningham, B., Chicchi, G., Liu, Q. (2002). Identification, localization and receptor characterization of novel mammalian substance P-like peptides. *Gene,* **296(1–2):** 205–212.

Kuznicki, L., Mikolajczyk, E., Walne, P. L. (1990). Photobehavior of euglenoid flagellates: theorethical and evolutionary perspective. *Crit. Rev. Plant Sci.,* **9:**343–369

Lafontaine, D. L. and Tollervey, D. (2001). The function and synthesis of ribosomes. *Natl Rev Mol. Cell Biol.,* **2(7):**514–520.

Lagesen, K., Hallin, P. F., Rødland, E., Stærfeldt, H. H., Rognes, T. Ussery D. W. **(2007).** RNammer: consistent annotation of rRNA genes in genomic sequences. *Nucl. Acids Res.,* **35(9):** *3100-3108. doi: 10.1093/nar/gkm160*

Lai, Y. and Sun, F. (2004). Sampling distribution for microsatellites amplified by PCR: mean field approximation and its applications to genotyping. *J. Theor. Biol.,* **228:**185 – 194.

Lambowitz, A. M., Zimmerly, S. (2004). Mobile group II introns. *Annual Review of Genetics,* **38:** 1–35.

Lambowitz, A. M., Zimmerly, S. (2011). Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harbor Perspectives in Biology*, **3:** a003616.

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., Karthikeyan, A. S., Lee, C. H., Nelson, W. D., Ploetz, L., Singh, S., Wensel, A. and Huala, E. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools., *Nucleic acids research*, **40:**D1202-10

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., *et al* (2001). Initial sequencing and analysis of the human genome. Nature **409:** 860–921 doi:10.1038/35057062.

Lander, E. S., Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, **2(3):** 231-239.

Lane, N. J. and Treherne, J. E. (1970). Lanthanum staining of neurotubules in axons from cockroach ganglia. *J. Cell Sci.,* **7:**217-231.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, **23:**2947-2948.

Lassalle, F., Périan, S., Bataillon, T., Nesme, X., Duret, L., Daubin, V. (2015). GC-Content Evolution in Bacterial Genomes: The Biased Gene Conversion Hypothesis Expands. *PLoS Genet.,* **11(2):** e1004941. https://doi.org/10.1371/journal.pgen.1004941.

Latorre, A. and Silva, F. J. (2013). What is behind the genome? New applications for genomics. *Metode* 2013 – 32.

Lauberth, S. M., Nakayama,T., Wu, X., Ferris, A. L., Tang, Z., Hughes, S. H. and Roeder, R. G. (2013). H3K4me3 Interactions with TAF3Regulate Preinitiation ComplexAssembly and Selective Gene Activation. *Cell,* **152:**1021–1036.

Laun, P., Bruschi, C. V., Dickinson, J. R., Rinnerthalerm, M., Heeren, G., Schwimbersky, R., Rid, R., Breitenbach, M. (2007). Yeast mother cell-specific ageing, genetic (in)stability, and the somatic mutation theory of ageing. *Nucleic Acids Res.***, 35:** 7514 – 7526.

Laval-Martin, D., Dubertret, G., Calvayrac, R. (1977). Photosynthetic properties of a DCMU-resistant strain of *Euglena gracilis* Z. *Plant Sci. Lett.,* **10:**185-195.

Lawen, A. (2003). “Apoptosis - an introduction,” *BioEssays*, **25(9):** 888 - 896.

Lawrence, C. C., and Stubbe, J. (1998). The function of adenosylcobalamin in the mechanism of ribonucleoside triphosphate reductase from *Lactobacillus leichmannii*. *Curr. Opin. Chem. Biol.,* **2:**650–655.

Le, S.Q. and Gascuel, O. (2008). An improved general amino acid replacement matrix. *Mol. Biol. Evol*., **25:**1307–1320.

Leander, B. S. (2004). Did trypanosomatid parasites have photosynthetic ancestors? Trends Microbiol. **12(6):**251–258.

Leander, B. S. and Farmer, M. A. (2001). Comparative morphology of the euglenid pellicle. II. Diversity of strip substructure. *J Eukaryot Microbiol*., **48:**202–217.

Leander, B. S., Esson, H. J., and Breglia, S. A. (2007). Macroevolution of complex cytoskeletal systems in euglenids. *BioEssays,* **29:**987–1000.

Leander, B. S., Esson, H. J., and Breglia, S. A. (2007). Macroevolution of complex cytoskeletal systems in euglenids. *Bioessays,* **29(10):**987-1000.

Leander, B. S., Farmer, M. A. (2000). Comparative morphology of the euglenid pellicle. I. Patterns of strips and pores. *J Eukaryot Microbiol.,* **47:**469–479.

Leander, B. S., Triemer, R. E., Farmer, M. A. (2001) Character evolution in heterotrophic euglenids. *Eur J Protistol*., **37:** 337–356.

Leander, B. S., Triemer, R. E., Farmer, M. A. (2001). Character evolution in heterotrophic euglenoids. *Eur. J. Protistol.,* **37:**337–356.

Ledbetter, M. C. and Porter, K. R. (1963). A 'microtubule' in plant cell fine structure. *J. Cell Biol.,* **19:**239-250.

Lee, G.J., Sohn, E.J., Lee, M.H. and Hwang, I. (2004). The Arabidopsis Rab5 homologs Rha1 and Ara7 localize to the prevacuolarcompartment. *Plant Cell Physiol.,* **45:**1211–1220.

Lee, J. M. and Sonnhammer, E. L. (2003). Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res.,* **13:**875–82.

Lee, T. Y., Chang, W. C., Hsu, J. B. K., Chang, T. H., Shien, D. M. (2012). GPMiner: an integrated system for mining combinatorial cis-regulatory elements in mammalian gene group. BMC Genomics, **13(Suppl 1):S3** http://www.biomedcentral.com/1471-2164/13/S1/S3

Leedale G. F. (1985). Order 3. Euglenoidina Butschii. J.J. Lee, S.H. Hutner, E.C. Bovee (Eds.), Illustrated Guide to the Protozoa, **pp.** 41–65

Leedale, G. F. (1968). The biology of Euglena (ed D E Buetow) vol. 1, p. 185. Academic Press, New York.

Leedale, G. F. (1958). Mitosis and chromosome numbers in the Euglenineae (Flagellata). *Nature (London)*, **181:**502.

Leedale, G. F. (1958). Nuclear structure and mitosis in the Euglenineae. *Arch Mikrobiol*, **32:**32 - 64.

Leedale, G. F. (1967). The Euglenoid Flagellates. N. J. Prentice Hall, Englewood Cliffs

Leedale, G. F. (1968). The genus in Euglena - *In: Buetow, D.E. (Ed.): The biology of Euglena.***I:** I85-f4f., Academic Press, New York, London.

Leedale, G. F. (1968). The nucleus in *Euglena.* In: Buetow DE (ed) *The biology of Euglena*, vol 1. Academic Press, New York, **pp.** 185-243.

Leedale, G. F. (1974). Preliminary observations on nuclear cytology and ultrastructure in carbon-starved streptomycin-bleached *Euglenagracilis. Colloq. Int. C. N. R. S . Cycles Celluluires et Leur Blockage,* Gif-sur-Yvette, France, **pp.** 285-290.

Leedale, G. F. (1975). Envelope formation and structure in the euglenoid genus *Trachelomonas*. – *Br. Phycol. J.,* **10:** 279–294.

Leedale, G. F. 1967. Euglenoid Flagellates. Prentice Hall, Englewood Cliffs, NJ. 242 pp

Leedale, G. F., Leadbeater, B. S. C. and Massalski, A. (1970). The intracellular origin of flagellar hairs in the *Chrysophyceae* and *Xanthophyceae. J. Cell Sci.* **6:**701-719.

Leedale, G. F., Meeuse, B. J. D., E. G. (1965). Pringsheim Structure and physiology of *Euglena spirogyra*, I and II: *Arch. Fur Microbiologie*, **50:**68–102

Leedale, G.F. (1962). The Evidence of a Meiotic Process in the Euglenineae. *Archiv. Mikrobiol.,* **42:** 237-245. https://doi.org/10.1007/BF00422042

Lefort-Tran, M. (1974). Mitochondries et chloroplastes chez *Euglena* en culture synchrone. In *Les Cycles cellulaires et lew Blocage chez plusieurs Protistes. Colloques Internationaux du C.N.R.S.,* **240:** 297 - 308, Gif-sur-Yvette, 28-30 aout 1974.

Lefort-Tran, M., Bre, M. H., Pouphile, M., and Manigault, P. (1987**).** DNA Flow Cytometry of Control *Euglena* and Cell Cycle Blockade of Vitamin Bl2-Starved Cells. *Cytometry,* **8:** 46 – 54.

Lefort-Tran, M., Bre, M. H., Ranck, J. L., Pouphile, M., (1980). *Euglena* plasma membrane during normal and vitamin B12 starvation growth. *J. Cell Sci.* **41:** 245-262.

Lehner, C. F., Stick, R. (1987). Eppenberger H.M., Nigg E.A., Differential expression of nuclear lamin proteins during chicken development. *J. Cell Biol.,* **105:**577 - 587.

Leipe, D. D., Wolf, Y. I., Koonin, E. V. and Aravind, L. (2002). Classification and evolution of P-loop GTPases and related ATPases. *J Mol Biol*., **317:** 41-72. doi: 10.1006/jmbi.2001.5378.

Leipe, D. D., Wolf, Y. I., Koonin, E. V., and Aravind, L. (2002). Classification and evolution of P-loop GTPases and related ATPases. *J Mol Biol.,* **317:**41–72.

Leitsch, D., Kolarich, D., Duchêne, M. (2010). The flavin inhibitor diphenyleneiodonium renders *Trichomonas vaginalis* resistant to metronidazole, inhibits thioredoxin reductase and flavin reductase, and shuts off hydrogenosomal enzymatic pathways. *Mol. Biochem. Parasitol.,* **171**:17–24.

Lenci, F. and Ghetti, F. (1989). Photoreceptor pigments for photomovements of microorganisms. Some spectroscopic and related studies. *J. Photochem. Photobiol.,* **3:** 1–16

Lepage, T., Bryant, D., Philippe, H., Lartillot, N. (2007). A general comparison of relaxedmolecular clock models. *Mol Biol Evol,* **24:**2669–80

Lerbs-Mache, S. (2011) Function of plastid sigma factors in higher plants: regulation of gene expression or just preservation of constitutive transcription? *Plant Mol. Biol.,* **76:**235–249

Lerner, M. R., Boyle, J. A., Mount, S. M., Wolin, S. L., and Steitz, J. A. (1980). Are snRNPs involved in splicing? *Nature,* **283:** 220–224.

Letunic, I. and Bork, P. (2006). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation*. Bioinformatics,* **23(1)**:127-8

Letunic, I. and Bork, P. (2011). Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy*. Nucleic Acids Res.,* **39(Web Server issue):**W475-8 doi: 10.1093/nar/gkr201.

Levasseur, P. J., Meng, Q. and Bouck, G. B. (1994). The Tubulin Genes in the Algal Protist *Euglena gracilis*. *J. Euk. Microbiol.,* **41(5):** 468 – 477

Levy, Y. Y., Lai, E. Y., Remillard, S. P., Heintzelman, M. B. and Fulton, C. (1996). Centrin is a conserved protein that forms diverse associations with centrioles and MTOCs in *Naegleria* and other organisms. *Cell Motil. Cytoskelet.,* **33**:298–323.

Lezzi, M, (1969). *Physiol them physics.* **1:** 447.

Li, B., Ruotti, V., Stewart, R. M., Thomson, J.A. and Dewey, C. N. (2010). RNA-seq gene expression estimation with read mapping uncertainty. *Bioinformatics,* **26:**493–500.

Li, C.-H., Irmer, H., Gudjonsdottir-Planck, D., Freese, S., Salm, H., Haile, S., Estévez, A. M., Clayton, C. (2006). Roles of a *Trypanosomabrucei* 5′->3′ exoribonuclease homologue in mRNA degradation. *RNA,* **12:** 2171–2186.

Li, F., Ge, P., Hui, W.H., Atanasov, I., Rogers, K., Guo, Q., Osato, D., Falick, A.M., Zhou, Z. H., Simpson, L. (2009). Structure of the core editing complex (L-complex) involved in uridine insertion/deletion RNA editing in trypanosomatid mitochondria. *Proc. Natl. Acad. Sci. USA,* **106:**12306–12310.

Li, F., Li, C., Wang, M., Webb, G. I., Zhang, Y., Whisstock, J. C., and Song, J.(2015). GlycoMine: a machine learning-based approach for predicting N-, C- and

O-linked glycosylation in the human proteome. *Bioinformatics,* **31(9):** 1411–1419 doi: 10.1093/bioinformatics/btu852.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Li, S., Yang, H., Wang, J., and Wang, J. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.,* **20:**265-272.

Liang, W. Q., Clark, J. A., Fournier, M. J. (1997). The rRNA-processing function of the yeast U14 small nucleolar RNA can be rescued by a conserved RNA helicase-like protein. *Mol. Cell Biol.,* **17(7):** 4124 –4132.

Liang, Y., Li, S. and Chen, L. (2015). The physiological role of drug transporters. *Protein Cell,* **6(5):**334–350.

Lin, S., Zhang, H. (2010). Dinoflagellate meta-transcriptomics enabled by spliced leader. In: Ho CK, editor. 13th International Conference on Harmful Algae. Hong Kong. 166–170.

Lin, S., Zhang, H., Zhuang, Y., Tran, B., Gill, J. (2010). Spliced leader–based metatranscriptomic analyses lead to recognition of hidden genomic features in dinoflagellates. *PNAS*, **107:** 20033–20038.

Lincoln Stein (2001). Genome annotation: from sequence to biology. *Nature Reviews Genetics,* **2:**493 - 503 | doi:10.1038/35080529

Link, G. (2003). Redox regulation of chloroplast transcription. Antioxid. *Redox Signal.,* **5:**79–87

Linton, E. W., Karnkowska-Ishikawa, A., Kim, J. I., Shin, W., Bennett, M. S., Kwiatowski, J., Zakrys, B. and 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. doi: 10.1016/j.protis.2010.02.002

Linton, E. W., Nudelman, M. A., Conforti, V., Triemer, R. E. (2000). A molecular analysis of the euglenophytes using SSUr DNA. *J Phycol.,* **36:**740–746.

Linton, E., Hittner, D., Lewandowski, C. L., Auld, T., Triemer, R. E. (1999). A molecular study of euglenoid phylogeny using small subunit rDNA. *J. Eukaryot Microbiol*., **46:**217 – 223

Lipscomb, D. (1991). Broad classification: the kingdoms and the protozoa, **pp.** 81-136. In J. P. Kreier and J. R. Baker (ed.), *Parasitic protozoa,* 2nd ed., vol. 1. Academic Press, Inc., San Diego, Calif.

Lira-Silva, E., Ramírez-Lima, I. S., Olín-Sandoval, V., García-García, J. D., García-Contreras, R., Moreno-Sánchez, R., Jasso-Chávez, R. (2011). Removal, accumulation and resistance to chromium in heterotrophic Euglena gracilis. *Journal of Hazardous Materials,* **193:** 216– 224

Liu, B., Molina, H., Kalume, D., Pandey, A., Griffith, J. D., Englund, P. T. (2006). Role of p38 in replication of *Trypanosoma brucei* kinetoplast DNA. *Mol Cell Biol.,* **26(53):**82–93.

Liu, Q, and Paroo, Z. (2010). Biochemical principles of small RNA pathways. *Annu Rev Biochem,* **79:**295–319.

Liu, Q., Greimann, J. C., and Lima, C. D. (2006). Reconstitution, activities, and structure of eukaryotic RNA exosome. *Cell,***127:**1223–1237.

Liu, S., Ji, D., Cliffe, L., Sabatini, R., Wang, Y. (2014). Quantitative Mass Spectrometry-Based Analysis of β-D-Glucosyl-5-Hydroxymethyluracil in Genomic DNA of *Trypanosoma brucei. J. Am. Soc. Mass Spectrom*., **25(10):**1763–1770. doi:10.1007/s13361-014-0960-6.

Liu, X. and Han, B. (2009). Evolutionary conservation of neighbouring gene pairs in plants. *Gene,* **437:**71–9.

Liu, X-Q. and Jagendorf, A. T. (1984). ATP-dependent proteolysis in pea chloroplasts. *FEBS Lett.,* **166:** 248 – 252.

Lombard, V., Ramulu, H. G., Drula, E.,Coutinho, P. M. and Henrissat, B.(2013). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Research,* **42:**D490–D495. doi:10.1093/nar/gkt1178.

Lonergan, T. A. (1984). Regulation of cell shape in Euglena gracilis: II. The effect of altered extra- and intracellular Ca2? Concentrations and the effect of calmodulin antagonists. *J Cell Sci.* **71:**37–50.

Lonergan, T. A. (1985). Regulation of cell shape in *Euglena gracilis*. IV. Localization of actin, myosin and calmodulin. *J Cell Sci.,***77:**197-208.

Lonergan, T. A. (1990). Role of actin, myosin, microtubules, and calmodulin in regulating the cellular shape of Euglena gracilis. In: O'Day DH (ed) Calcium as an intracellular messenger in eukaryotic microbes. *American Society for Microbiology, Washington, DC,* **pp** 258–277.

Long, E. O. and Dawid, I. B. (1980). Repeated genes in eukaryotes. *Annu. Rev. Biochem.* 49:727 - 764.

Lopez-Bellido, R., Barreto-Valer, K. and Rodrıguez, R. E. (2013). Expression of tachykinin receptors (tacr1a and tacr1b) in zebrafish: influence of cocaine and opioid receptors.*Journal of Molecular Endocrinology,* **50:**115–129

Losh, J. S., and van Hoof, A. (2015). Gateway Arch to the RNA Exosome. *Cell,* **162:** 941.

Lowdon, M. and Vitols, E. (1973). Ribonucleotide reductase activity during the cell cycle of Saccharomyces cerevisiae. *Arch. Biochem. Biophys.*, **158:**177-184.

Lowe, T. M. and Eddy, S. R. (1997). tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucl. Acids Res.,* **25(5):** 0955-964. *doi:10.1093/nar/25.5.0955*

Lueong, S., Merce, C.,Fischer, B., Hoheisel, J. D. and Erben, E. D. (2016). Gene expression regulatory networks in *Trypanosoma brucei*: insights into the role of the mRNA-binding proteome. *Molecular Microbiology,* **100(3):** 457–471 doi:10.1111/mmi.13328.

Lukes J., Hashimi, H., Zıkova, A. (2005). Unexplained complexity of the mitochondrial genome and transcriptome in kinetoplastid flagellates. *Curr Genet.,* **48:** 277 – 299.

Lukes, J., Archibald, J. M., Keeling, P. J., Doolittle, W. F., Gray, M. W. (2011). How a neutral evolutionary ratchet can build cellular complexity. *IUBMB Life*, **63(7):**528-37 doi: 10.1002/iub.489.

Lukes, J., Guilbride, D. L., Votypka, J., Zıkova, A., Benne, R., Englund, P. T. (2002). Kinetoplast DNA network: evolution of an improbable structure. *Eukaryot. Cell,* **1:** 495—502.

Lukes, J., Leander, B. S., and Keeling, P. J. (2009). Cascades of convergent evolution: The corresponding evolutionary histories of euglenozoans and dinoflagellates. *PNAS,* **106(1):**9963–9970.

Lukes, J., Leander, B. S., Keeling, P. J. (2009). Cascades of convergent evolution: the cor-responding evolutionary histories of euglenozoans and dinoflagellates. *Proc. Natl. Acad. Sci. USA,* **106(Suppl. 1):**9963–70.

Lukes, J., Skalicky, T., Tyc, J., Votypka, J., and Yurchenko, V. (2014). Evolution of parasitism in kinetoplastid flagellates. *Mol Biochem Parasitol*., **195(2):** 115 - 122 http://dx.doi.org/10.1016/j.molbiopara.2014.05.007.

Lukes, J., Skalicky, T., Tyc, J., Votypka, J., Yurchenko, V. (2014). Evolution of parasitism in kinetoplastid flagellates. *Mol Biochem Parasitol*., **195(2):** 115 – 122.

Lukeš,J.(2014). Mitochondrial Genomes of Excavata. In: *Molecular Life Sciences*, **pages 1 – 7.** DOI 10.1007/978-1-4614-6436-5_118-2.

Luo, Y., Vassilev, P. M., Li, X., Kawanabe, Y. and Zhou, J. (2003). Native polycystin 2 functions as a plasma membrane $Ca^{2+}$-permeable cation channel in renal epithelia. *Mol. Cell. Biol.,* **23:**2600-2607.

Lwoff, A. (1932). Recherches biochemiques sur la nutrition des Protozoaires. Le pouvoir synthese. Monographies de l'Institute Pasteur (Paris, Masson et Cie), 158 **pp**.

Lye, L. F., Owens, K., Shi, H., Murta, S. M., Vieira, A. C., Turco, S. J., Tschudi, C., Ullu, E. and Beverley, S. M. (2010). Retention and loss of RNA interference pathways in trypanosomatid protozoans. *PLoS Pathog,* 6:e1001161.

Lynch, M. J., Leake, R. E., O'connell, K. M. and Buetow, D. E. (1974). Isolation, Fractionation and Template Activity of the Continuously-Condensed Chromatin of *Euglena gracilis. Experimental Cell Research*, **91:** 349 – 351

Lynch, M. (2007). The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc Natl Acad Sci USA,* **104:**8597 – 8604.

Lynch, M., Conery, J. S. (2003). The origins of genome complexity. *Science*, **302:**1401–1404.

Ma, L., and Spremulli, L. L. (1990). Identification and characterization of large, complex forms of chloroplast translational initiation factor 2 from *Euglena gracilis*. *J. Bid. Chem.,* **268:**13560 – 13565.

Ma, L., and Spremulli, L. L. (1992). Immunological characterization of the complex forms of chloroplast translational initiation factor 2 from *Euglena gracilis*. *J. Biol. Chem.,* **267:**18356 – 18360.

MacLeod, A., Tweedie, A., McLellan, S., Taylor, S., Cooper, A., Sweeney, L., Turner, M. R. C. and Tait, A. (2005). Allelic segregation and independent assortment in T. brucei crosses: Proof that the genetic system is Mendelian and involves meiosis. *Mol Biochem. Parasitol.,* **143:**12–19.

Maddison, W. P. and D.R. Maddison. 2015. Mesquite: a modular system for evolutionary analysis. Version 3.03 http://mesquiteproject.org

Maddox, P. S., Corbett, K. D., and Desai, A. (2012). Structure, assembly and reading of centromeric chromatin. *Curr. Opin. Genet. Dev.,* **22:**139–147.

Maden, B. E. (1990). The numerous modified nucleotides in eukaryotic ribosomal RNA. *Prog Nucleic Acid Res Mol Biol.,* **39:**241–303.

Maden, B. E. H. (1990). The numerous modified nucleotides in eukaryotic ribosomal RNA. *Prog. Nucleic Acid Res. Mol. Biol.,* **39:**241–303.

Madeo, F., Herker, E., Wissing, S., Jungwirth, H., Eisenberg, T., and Fröhlich, K. U. (2004). “Apoptosis in yeast,” *Current Opinion in Microbiology*, **7:**655 - 660.

Madhusudhan, R., Ishikawa, T., Sawa, Y., Shigeoka, S. & Shibata, H. (2003). Post-transcriptional regulation of ascorbate peroxidase during light adaptation of *Euglena gracilis*. *Plant Sci.,* **165:**233–238.

Mahapatra, D. M., Chanakya, H. N. & Ramachandra, T. V. (2013). *Euglena* sp. as a suitable source of lipids for potential use as biofuel and sustainable wastewater treatment. *J Appl Phycol.,* **25:** 855. doi:10.1007/s10811-013-9979-5.

Maier, R. M., Zeltz, P., Kossel, H., Bonnard, G., Gualberto, J. M. (1996). Grienenberger J.M., RNA editing in plant mitochondria and chloroplasts, *Plant Mol. Biol.,* **32:**343–365.

Maier, U.G., Bozarth, A. ,Funk, H. T., Zauner, S., Rensing, S. A., Schmitz-Linneweber, C., Börner, T. and Tillich, M. (2008) Complex chloroplast RNA metabolism: just debugging the genetic programme? *BMC Biol.,* **6:**9.

Mair, G., Shi, H., Li, H., Djikeng, A., Aviles, H. O., Bishop, J. R., Falcone, F. H., Gavrilescu, C., Montgomery, J. L., Santori, M. I., Stern, L. S., Wang, Z., Ullu, E., Tschudi, C. (2000). A new twist in trypanosome RNA metabolism: cis-splicing of pre-mRNA. *RNA*, **6:**163–169.

Maiti, M., Lee, H. C., and Liu, Y. (2007). QIP, a putative exonuclease, interacts with the Neurospora Argonaute protein and facilitates conversion of duplex siRNA into single strands. *Genes Dev*., **21:**590–600.

Malek, L., Bogorad, L., Ayers, A. L., Goldberg, A. L. (1984). Newly synthesized proteins are degraded by an ATP-stimulated proteolytic process in isolated pea chloroplasts. *FEBS Lett.,* **166:** 253-257.

Manful, T., Fadda, A., and Clayton, C. (2011) The role of the 5′-3′ exoribonuclease XRNA in transcriptome-wide mRNA degradation. *RNA,* **17:** 2039–2047.

Mani, J., Guttinger, A., Schimanski, B., Heller, M., Acosta-Serrano, A., Pescher, P., Späth, G., Roditi, I. (2011). Alba-domain proteins of *Trypanosoma brucei* are cytoplasmic RNA-binding proteins that interact with the translation machinery. *PLoS One,* **6:** e22463.

Manna, P. T., Kelly, S., and Field, M. C. (2013). Adaptin evolution in kinetoplastids and emergence of the variant surface glycoprotein coat in African trypanosomatids. *Mol. Phylogenet Evol.,* **67(1):** 123–128. doi: 10.1016/j.ympev.2013.01.002.

Manning, G, Whyte, D. B., Martinez, R., Hunter, T., Sudarsanam, S. (2002b).The protein kinase complement of the human genome. *Science (New York, N.Y.),* **298(5600):**1912-1934

Manning, G., Plowman, G. D., Hunter, T., and Sudarsanam, S. (2002a). Evolution of protein kinase signaling from yeast to man. *Trends in Biochemical Sciences,* **27(10):**514-520

Manning, J. E., Wolstenholme, D. R., Ryan, R. S., Hunter, J. A., Richards, O. C. (1971). Circular chloroplast DNA from *Euglena gracilis*. *Proc. Natl. Acad. Sci. USA,* **68:**1169–1173.

Marande, W., Lukes, J. and Burger, G. (2005).Unique mitochondrial genome structure in diplonemids, the sister group of kinetoplastids. *Eukaryot Cell.,* **4(6):**1137–46.

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., Geer, R. C., He, J., Gwadz, M., Hurwitz, D. I., Lanczycki, C. J., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Wang, Z., Yamashita, R. A., Zhang, D., Zheng, C., Bryant, S. H. (2015). CDD: NCBI's conserved domain database. *Nucleic Acids Res.,* **43(Database issue):**D222-6. doi: 10.1093/nar/gku1221.

Maree, J. P. and Patterton, H. (2014). The epigenome of *Trypanosoma brucei*: A regulatory interface to an unconventional transcriptional machine. *Biochimica et Biophysica Acta,* **1839:** 743–750

Marguerat, S. and Bahler, J. (2012). Coordinating genome expression with cell size. *Trends in Genetics,* **28(11):**560-565.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, YJ, Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, JB, Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F. and Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature,* 437(7057):376-80. doi:10.1038/nature03959

Marin, B. (2004). Origin and fate of chloroplasts in the Euglenoida. *Protist,* 155:13–14.

Marin, B. (2004). Origin and fate of chloroplasts in the euglenoida. *Protist,* **155:**13 – 14.

Marin, B., Palm, A., Klingberg, M., 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:**99–145.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.,* **18:**1509–1517.

Mark C. Field (2005). Signalling the genome: the Ras-like small GTPase family of trypanosomatids. *TRENDS in Parasitology,* **21(10):** 447 – 450.

Marks, B., Stowell, M. H., Vallis, Y., Mills, I. G., Gibson, A., Hopkins, C. R., and McMahon, H. T. (2001). GTPase activity of dynamin and resulting conformation change are essential for endocytosis. Nature, **410(6825):** 231-235.

Marsden, M. P. F., Laemmli, U. K. (1979). Metaphase chromosome structure: Evidence for a radial loop model. *Cell*, 17: 849-858

Martin, J. A. and Wang, Z. (2011). Next-generation transcriptome assembly. *Nature Reviews Genetics,* **12:**671-682. doi:10.1038/nrg3068.

Martin, J. A. and Wang, Z. (2011). Next-generation transcriptome assembly. *Nature Reviews* | *Genetics,* **12:** 671 – 682.

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

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

Martin, W. F., Garg, S., and Zimorski, V. (2015). Endosymbiotic theories for eukaryote origin. *Phil. Trans. R. Soc. B*., **370:** 20140330. http://dx.doi.org/10.1098/rstb.2014.0330.

Martin, W., Brinkmann, H., Savonna, C., Cerff, R. (1993). Evidence for a chimeric nature of nuclear genomes — eubacterial origin of eukaryotic

glycerinaldehyde-3-phosphate dehydrogenase genes. *Proc. Natl. Acad. Sci. U. S. A.,* **90:**8692–8696.

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

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

Martínez-Duncker, I., Díaz-Jímenez, D. F., and Mora-Montes, H. M. (2014). Comparative Analysis of Protein Glycosylation Pathways in Humans and the Fungal Pathogen *Candida albicans. International Journal of Microbiology*, **267497:** 16 pages http://dx.doi.org/10.1155/2014/267497.

Maruyama, S., Suzaki, T., Weber, A P. M., Archibald, J. M. and Nozaki, H. (2011). Eukaryote-to-eukaryote gene transfer gives rise to genome mosaicism in euglenids. *BMC Evolutionary Biology,* **11:**105

Maslov, D. A., Podlipaev, S. A., Lukes, J. (2001). Phylogeny of the kinetoplastida: taxonomicproblems and insights into the evolution of parasitism. *Mem Inst Oswaldo Cruz*, **96:**397–402.

Mast, F., Barlow, L., Rachubinski, R., Dacks, J. (2014). Evolutionary mechanisms for establishing eukaryotic cellular complexity. *Trends in Cell Biology,* **24(7):**435-442. Doi: 10.1016/j.tcb.2014.02.003.

Mast, S. O. (1911). Light and the Behaviour of Lower Organisms. John Wiley and Sons, New York

Masuda, S. (2013). Light Detection and Signal Transduction in the BLUF Photoreceptors.*Plant Cell Physiol.,* **54(2):** 171–179 doi:10.1093/pcp/pcs173.

Matera, G. and Wang, Z. (2014). A day in the life of the spliceosome. *Nature Reviews |Molecular Cell Biology,* **15:**108 – 121.

Mathews, M. B., Sonenberg, N., and Hershey, J. W. B. (2000) Origins and principles of translational control, in *Translational control of gene expression* (Sonenberg, N., Hershey, J. W. B., and Mathews, M. B., Eds.), **pp 1-31,** Cold Spring Harbor Laboratory press, Cold Spring Harbor, New York.

Mathews, M. B., Sonenberg, N., and Hershey, J. W. B. (2007) Origins and principles of translational control, in *Translational control in biology and medicine* (Mathews, M. B., Sonenberg, N., and Hershey, J. W. B., Eds.), **pp 1-40,** Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Mathews, M. B., Sonenberg, N., and Hershey, J. W. B., (Eds.) (2007). *Translational control in biology and medicine*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NewYork.

Matsuda, F., Hayashi, M., Kondo, A. (2011). Comparative profiling analysis of central metabolites in *Euglena gracilis* under various cultivation conditions. *Biosc Biotechnol and Biochem*., **75(11):**2253 - 2256.

Matsunaga, E., Higuchi, Y., Mori, K., Yairo, N., Oka, T., Shinozuka, S., Tashiro, K., Izumi, M., Kuhara, S., and Takegawa, K. (2015). Identification and Characterization of a Novel Galactofuranose-Specific β-D-Galactofuranosidase from Streptomyces Species. *PLoS ONE,* **10(9):** e0137230. doi:10.1371/journal.pone.0137230

Matsuura, H., Sokabe, T., Kohno, K., Tominaga, M., Kadowaki, T. (2009). Evolutionary conservation and changes in insect TRP channels. *BMC Evol Biol.,***9:**228.

Matsuyama, S. and Tsuyama, S. (1991). Mono-ADP-ribosylation in brain: purification and characterization of ADP-ribosyltransferases affecting actin from rat brain. *J. Neurochem.,* **57:** 1380 - 1387.

Matthews, M. M., and Sistrom, W. R., (1959). Function of carotenoid pigments in non-photosynthetic bacteria. *Nature,* **184:** 1892-1893.

Matvienko, O. M. (1938). Materyiali do vivcheniya vodorostej URSR. I. Vodorosteji Klyukvennogo bolota. Uchen. Zap. *Kharkyiv. derzh. Unyiv. Kharkyiv,* **14:**29-78. (in Ukrainian).

May, P., Christian J-O, Kempa S, Walther D. ChlamyCyc: an integrative systems biology database and web-portal for Chlamydomonas reinhardtii. *BMC Genomics,* **10:**209.

Mazus, B., Falchuk, K. H., and Vallee, B. L. (1984). Histone Formation, Gene Expression, and Zinc Deficiency in *Euglena gracilis*. *Biochemistry,* **23:** 42 - 47.

Mcalister, L. and M. J. Holland (1982). Targeted deletion of a yeast enolase structural gene. *J. Biol. Chem.,* **257**:7181–7188.

McAllaster, M. R., Ikeda, K. N., Lozano-Núñez, A., Anrather, D., Unterwurzacher, V., Gossenreiter, T., Perry, J. A., Crickley, R., Mercadante, C. J., Vaughan, S., and de Graffenried, C. L. (2015). Proteomic identification of novel cytoskeletal proteins associated with TbPLK, an essential regulator of cell morphogenesis in *T. brucei. Mol. Biol. Cell,* **26(17):**3013 - 3029. doi: 10.1091/mbc.E15-04-0219.

Mccarthy, S. A. and Schwartzbach, S. D. (1984). Absence of photoregulation of abundant mRNA levels in Euglera. *Plant Sci. Lett.,* **35:** 61-66.

McCormack, E. and Braam, J. (2003). Calmodulins and related potential calcium sensors of Arabidopsis. *New Phytologist,***159**: 585–598.

McDowall, J. (2016). "Calmodulin"*.InterPro Protein Archive.* Retrieved 19 May, 2016*.*

McFadden, G. I., (2001). Primary and secondary endosymbiosis and the origin of plastids. *J. Phycol*. **37:**951–959.

McIntosh, J.R., Volkov,V., Ataullakhanov,F.I. and Grishchuk,E. L. (2010). Tubulin depolymerization may be an ancient biological motor. *J. CellSci.,* **123:**3425–3434.

McKean, P. G., Vaughan, S. and Gull, K. (2001). The extended tubulin superfamily. *J Cell Sci*, **114:** 2723 - 2733.

McMahon, H. T. and Boucrot, E. *(2011).* Molecular mechanism and physiological functions of clathrin-mediated endocytosis. *Nature Reviews* | *Molecular Cell Biology*, **12:**517 –533.

McManus, M.T., Shimamura, M., Grams, J., Hajduk, S. L. (2001). Identification of candidate mitochondrial RNA editing ligases from *Trypanosoma brucei*. *RNA,* **7:**167–175.

Medawar, P. B. (1952). An Unsolved Problem of Biology. Lewis, London.

Mehlert, A., Zitzmann, N., Richardson, J. M., Treumann, A., Ferguson, M. A. (1998). The glycosylation of the variant surface glycoproteins and procyclic acidic repetitive proteins of *Trypanosoma brucei*. *Mol. Biochem. Parasitol.,* **91(1):**145-52.

Melkonian, M., Robenek, H., Rassat, J. (1982). Flagellar membrane specialization and their relationship to mastigonemes and microtubules in *Euglena gracilis. J. Cell Sci.,* **55:**115–135.

Meraldi, P., McAinsh, A.D., Rheinbay, E., and Sorger, P.K. (2006). Phylogenetic and structural analysis of centromeric DNA and kinetochore proteins. *Genome Biol.* **7:**R23.

Mercatelli, R., Quercioli, F., Barsanti, L., Evangelista, V., Coltelli, P., Passarelli, V., Frassanito, A. M., Gualtieri, P. (2009). Intramolecular photo-switching and intermolecular energy transfer as primary photoevents in photoreceptive processes: The case of *Euglena gracilis*. *Biochemical and Biophysical Research Communications,* **385:**176–180.

Mermelstein, C. S., Rodrigues, A. P. M., Einicker-Lamas, M., Navarrete de , B. R. E., Farina, M., Costa, M. L. (1998). Distribution of F-actin, α- actinin, tropomyosin, tubulin and organelles in Euglena gracilis by immunofluorescence microscopy. *Tissue & Cell,* **30 (5):**545 - 553

Mertens, E. (1993). ATP versus pyrophosphate: glycolysis revisited in parasitic protists. *Parasitol. Today,* **9**:122–126.

Mertens, E., E. Van Schaftingen, And M. Muller. (1992). Pyruvate kinase from *Trichomonas vaginalis,* an allosteric enzyme stimulated by ribose 5-phosphate and glycerate 3-phosphate. *Mol. Biochem. Parasitol.,* **54**:13–20.

Miazek, K., Iwanek, W., Remacle, C., Richel, A., and Goffin, D. (2015). Effect of Metals, Metalloids and Metallic Nanoparticles on Microalgae Growth and Industrial Product Biosynthesis: A Review. *Int J Mol Sci.,* **16(10):** 23929–23969.

Michael J. Bumbulis and Brian M. Balog (2013). UV-C Exposure Induces an Apoptosis-Like Process in *Euglena gracilis. ISRN Cell Biology*, **2013(869216):** 6 pages

Michel, H. (1990) in: Crystallization of Membrane Proteins (Michel, H., Ed.), **pp.** 73 - 89, CRC Press, Boca Raton, FL.

Michels, P. A. M., and V. Hannaert. (1994). The evolution of kinetoplastid glycosomes. *J. Bioenerg. Biomembr.* **26**:213–219.

Michels, P. A., Bringaud, F., Herman, M., and Hannaert, V. (2006). Metabolic functions of glycosomes in trypanosomatids. *Biochim. Biophys. Acta.,* **1763:**1463−1477.

Miersch, I., Heise, J., Zelmer, I., Humbeck, K. (2000) Differential degradation of the photosynthetic apparatus during leaf senescence in barley (Hordeum vulgare L.). *Plant Biology,* **2:** 618–623.

Mignot, J. P. (1966). Structure et ultrastructure de quelques Euglónomonadines. *Prostologica,* **2:**51–117

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:**721–730.

Milanowski Rafal, Karnkowska Anna, Ishikawa Takao, and Zakrys Bozena (2013). Distribution of Conventional and Nonconventional Introns in Tubulin (α and β) Genes of Euglenids. *Mol. Biol. Evol.* **31(3):**584–593 doi:10.1093/molbev/mst227.

Milanowski, R., Gumińska, N., Karnkowska, A., Ishikawa, T. and Zakryś, B. (2016). Intermediate introns in nuclear genes of euglenids – are they a distinct type? *BMC Evolutionary Biology*, **16:** 49 DOI 10.1186/s12862-016-0620-5.

Milanowski, R., Karnkowska, A., Ishikawa, T., and Zakryś, B. (2014). Distribution of Conventional and Nonconventional Introns in Tubulin (α and β) Genes of Euglenids. *Mol Biol Evol.,* **31(3):** 584–593.

Milanowski, R., Kosmala, S., Zakrys, B., Kwiatowski, J. (2006). Phylogeny of photosynthetic euglenophytes based on combined chloroplast and cytoplasmic SSUr DNA sequence analysis. *J. Phycol.* 42:721–730.

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

Miller, K. G., and B. Soliner-Webb. (1981). Transcription of mouse rRNA gene by RNA polymerase I: in vitro and in vivo initiation and processing sites. *Cell,* **27:**165-174.

Miller, K. R. and Miller, G. J. (1978). Organization of the cell membrane in *Euglena. Protoplasma,* **95:** 11 - 24.

Miller, K. R. and Staehelin, L. A. (1973). Fine Structure of the Chloroplast Membranes of *Euglena gracilis* as Revealed by Freeze-Cleaving and Deep-Etching Techniques. *Protoplasma*, **77:** 55 – 78

Miller, K. R., Miller, G. J. (1978). Organization of the cell membrane in *Euglena. Protoplasma,* **95:** 11–24.

Miller, M. E. and Price, C. A. (1982). Protein synthesis by developing plastids isolated from *Euglena gracilis*. *FEBS Lett.,* **147:** 156 – 160.

Miller, M. E., Jurgenson, J. E., Reardon, E. M., Price, C. A. (1983). Plastid translation in organello and in vitro during light-induced development in Euglena. *J Biol Chem*, **258:** 14478 – 14484.

Milner, J. J., Hershberger, C. L. and Buetow, D. E. (1979). *Euglena gracilis* chloroplast DNA codes for polyadenylated RNA. *Plant Physiol.,* **64:**818–821.

Minchin, E. A. (1908) Investigations on the development of trypanosomes in tsetseflies and other Diptera. *Q J Microsc Sci*., **52:**159–260

Minguez, A., Franca, S., de la Espina, S. M. D. (1994). Dinoflagellates have a eukaryotic nuclear matrix with lamina-lake proteins and topoisomerase II. *J. Cell Sci.,* **107:**2862-2837.

Mirskv. A. E., Burdick. C. J., Davidson. E. H. and Littau, V. C. (1968). The Role Of Lysine-Rich Histone In The Maintenance Of Chromatin Structure In Metaphase Chromosomes. *Proc. natl. acad. sci. USA*., **61:** 592.

Mishra, N. S., Tuteja, R., Tuteja, N. (2006). Signaling through MAP kinase networks in plants. *Archives of Biochemistry and Biophysics,* **452:**55-68.

Missel, A., Souza, A. E., Norskau, G., Goringer, H. U. (1997). Disruption of a gene encoding a novel mitochondrial DEAD-box protein in *Trypanosoma brucei* affects edited mRNAs. *Mol. Cell. Biol.,* **17:**4895–4903.

Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., Punta, M. (2013). Challenges in Homology Search: HMMER3 and Convergent Evolution of Coiled-Coil Regions. *Nucleic Acids Research*, **41:**e121

Mitchelson, K. R., Bekers, A. G. M., Wanka, F. (1979). Isolation of a residual protein structure from nuclei of the myxomycete *Physarum polycephalum, J. Cell Sci.,* **39:**247 - 256.

Miwa, I., Izumo, T., Sonoda, T. (1996). Cytoplasm rescues an arrhythmic mutant on the circadian rhythm of mating reactivity in *Paramecium bursaria*. *J. Eukaryot. Microbiol.,* **43:**231–236.

Miwa, I., Yajima, H. (1995). Correlation of the period length of circadian rhythms with the length of immaturity in Paramecium bursaria. *Zool. Sci.,* **12:** 53–59.

Miyagishima, S. Y. (2011). Mechanism of plastid division: From a bacterium to an organelle. *Plant Physiol.,* **155(4):**1533–1544.

Mizoguchi, T., Gotoh, Y., Nishida, E., Yamaguchi-Shinozaki, K., Hayashida, N., Iwasaki, T., Kamada, H., Shinozaki, K. (1994). Characterization of two cDNAs that encode MAP kinase homologues in *Arabidopsis thaliana* and analysis of the possible role of auxin in activating such kinase activities in cultured cells. *The Plant Journal*, **5:**111-122.

Mochizuki, K., Fine, N. A., Fujisawa, T., and Gorovsky, M. A. (2002). Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in tetrahymena. *Cell,* **110:**689–99.

Moestrup, O. (1982). Flagellar structure in algae: a review, with new observations particularly on the Crysophyceae, Phaeophyceae (Fucophyceae), Euglenophyceae and Reckertia. *Phycologia,* **21:** 427–528

Moestrup, O. (1978). On the phylogenetic validity of the flagellar apparatus in green algae and other chlorophyll a and b containing plants. *BioSystems,* **10:**117–144

Mollenhauer, H. H. (1965). An Intercisternal Structure in the Golgi Apparatus. *J. Cell Biol.,* **24:**504-511.

Mollenhauer, H. H. (1974). Distribution of Microtubules in the Golgi Apparatus of *Euglena gracilis. J. Cell Sci.,* **15:**89 – 97.

Mollenhauer, H. H. and Morre, D. J. (1966). Golgi apparatus and plant secretion. *A. Rev. Plant Physiol.,* **17:**27-46.

Mollenhauer, H. H. and Morre, D. J. (1972). Intercisternal substances of the Golgi apparatus, zones of exclusion, and other 'invisible' structures that contribute to subcellular compartmentalization. *What's New in Plant Physiol.,* **4:** 1-4.

Mollenhauer, H. H. and Morre, D. J. (1973). Intercisternal substances of plant dictyosomes. Unstacking of plant dictyosomes using chaotropic agents. *Protoplasma,* **78:**443-461

Mollenhauer, H. H., Totten, C. and Acuff, K. (1971). Separation of Golgi apparatus cisternae by solubilization of intercisternal bonding substances. *Abstr. nth A. Meeting Am. Soc. Cell Biol.,* **pp.** 345.

Mollin, D. L., Anderson, B. B., and Burman, J. F. (1976). The serum vitamin B12 level: its assay and significance. *Clin. Haematol.,* **5:**521–546.

Monde, R. A., Gadi Schuster, David B. Stern (2000). Processing and degradation of chloroplast mRNA. *Biochimie,* **82:** 573 – 582.

Monroy, A. F. and Schwartzbach, S. D. (1984). Catabolite repression of chloroplast development in Euglena. *Proc. Natl. Acad. Sci. USA,* **81:**2786-2790.

Monroy, A. F. and Schwartzbach, S. D. (1984). Influence of photosynthesis and chlorophyll synthesis on polypeptide accumulation in greening euglena. *Plant Physiol.,* **77:**811-816.

Monroy, A. F., Gomez-Silva, B., Schwartzbach, S. D. and Schiff, J. A. (1986). Photocontrol of chloroplast and mitochondrial polypeptide levels in Euglena. *Plant Physiol.,* **80:**618-622.

Monroy, A. F., McCarthy, S. A. and Schwartzbach, S. D. (1987). Evidence for translational regulation of chloroplast and mitochondrial biogenesis in *Euglena. Plant Sci.,* **51:** 61-76.

Montandon, P. E. and Stutz, E. (1990). Structure and expression of the *Euglena gracilis* nuclear gene coding for the translation elongation factor EF-1 alpha. *Nucleic Acids Res.,* **18(1):** 75–82.

Montandon, P. E. and Stutz, E. (1990). Structure and expression of the *Euglena gracilis* nuclear gene coding for the translation elongation factor EF-1 alpha. *Nucleic Acids Res.,* **18(1):** 75–82.

Moore, A. N., Russell, A. G. (2012). Clustered organization, polycistronic transcription, and evolution of modification-guide snoRNA genes in Euglena gracilis. *Mol. Genet. Genomics,* **287:**55–66.

Moore, M. S. and Blobel, G. (1993). The GTP-binding protein Ran/TC4 is required for protein import into the nucleus. *Nature,* **365:**661–663.

Moore-Ede, M. C., Sulzman, F. M., Fuller, C. A. (1982). The Clocks That Time Us. *Harvard University Press, Cambridge, MA*, **pp.** 448.

Morales, J., Hashimoto, M., Williams, T. A., Hirawake-Mogi, H., Makiuchi, T., Tsubouchi, A., Kaga, N., Taka, H., Fujimura, T., Koike, M., Mita, T., Bringaud, F., Concepcion, J. L., Hashimoto, T., Embley, T. M. and Nara, T. (2016). Differential remodelling of peroxisome function underpins the environmental and metabolic adaptability of diplonemids and kinetoplastids. *Proc. R. Soc. B*., **283:** 20160520. http://dx.doi.org/10.1098/rspb.2016.0520

Morales, J., Mogi, T., Mineki, S., Takashima, E., Mineki, R., Hirawake, H., Sakamoto, K., Omura, S., Kita, K. (2009). Novel mitochondrial complex II isolated from Trypanosoma cruzi is composed of 12 peptides including a heterodimeric Ip subunit. *J. Biol. Chem.,* **284:**7255 –7263.

Morales-Calderón, L. S., Armenta-Ortiz, N., Méndez-Trujillo, V., Ruiz-Sanchez, E., González-Mendoza, D., Grimaldo-Juarez, O., Cervantes-Diaz, and Aviles-Marin, M. (2012). Copper induced biofilm formation and changes on photosynthetic pigment in *Euglena gracilis. African Journal of Microbiology Research ,* **6(8):**1833-1836.

Moreira D., Lopez-Garcıa P., Vickerman K. (2004). An updated view of kinetoplastid phylogeny using environmental sequences and a closer outgroup: proposal for a new classification of the class Kinetoplastea. *Int J Syst Evol Microbiol.,* **54:**1861–1875.

Moreira, D., Kervestin, S., Jean-Jean, O., and Philippe, H. (2002). Evolution of eukaryotictranslation elongation and termination factors: variations of evolutionary rate andgenetic code deviations. *Mol. Biol. Evol.,* **19:**189-200.

Moremen,K. W., Tiemeyer,M., and Nairn,A. V. (2012). Vertebrate protein glycosylation: diversity, synthesis and function.*Nat. Rev. Mol. Cell Biol*. **13(7):** 448–462. doi:10.1038/nrm3383.

Moreno-Sánchez, R., Covian, R., Jasso-Chavez, R., Rodriguez-Enriquez, S., Pacheco-Moises, F., Torres-Marquez, M. E. (2000). Oxidative phosphorylation supported by an alternative respiratory pathway in mitochondria from Euglena. *Biochim. Biophys. Acta,* **1457:**200–210.

Moreno-Sanchez, R., Raya, J. C. (1987). Preparation of Coupled Mitochondria from Euglena by Sonication. Plant Science, **48:** 151 – 157.

Moreno-Sánchez, R., Rodríguez-Enríquez, S., Jasso-Chávez, R., Saavedra, E., García-García, J. D. (2017). "Biochemistry and Physiology of Heavy Metal Resistance and Accumulation in *Euglena"*. In: S. Schwartzbach, S. Shigeoka (eds.), *Euglena: Biochemistry, Cell and Molecular Biology*. Advances in Experimental Medicine and Biology, **979:**91-121. Springer International Publishing AG. DOI: 10.1007/978-3-319-54910-1_7.

Moriya, Y. K. (2007). In automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.,* **35:**W182–W185.

Morre, D. J., Mollenhauer, H. H. and Bracker, C. E. (1971). Origin and continuity of Golgi apparatus. In *Results and Problems in Cell Differentiation II, Origin and Continuity of Cell Organelles* (ed. J. Reinert & H. Ursprung), **pp.** 82-126. Berlin: Springer-Verlag.

Morriswood, B., Havlicek, K., Demmel, L., Yavuz, S., Sealey-Cardona, M., Vidilaseris, K., Anrather, D., Kostan, J., Djinovic´-Carugo, K., Roux, K. J., Warren, G.(2013). Novel Bilobe Components in *Trypanosoma brucei* Identified Using Proximity-Dependent Biotinylation. *Eukaryotic Cell,* **12(2):**356–367.

Morriswood, B., He, C. Y., Sealey-Cardona, M., Yelinek, J., Pypaert, M., Warren, G. (2009). The bilobe structure of *Trypanosoma brucei* contains a MORN-repeat protein. *Mol. Biochem. Parasitol.,* **167:**95–103.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods,* **5:**621–628.

Mortimer, S. A. and Weeks, K. M. (2007). A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J. Am. Chem. Soc.,* **129:**4144–4145.

Morton, B. R. (1998). Selection on the codon bias of chloroplast and cyanelle genes in different plant and algal lineages. *J. Mol. Evol.,* **46:** 449 – 459.

Morton, B. R. (1999). Strand asymmetry and codon usage bias in the chloroplast genome of *Euglena gracilis. Proc. Natl. Acad. Sci. USA,* **96(9):** 5123 – 5128.

Moyne, G., Bertaux, O., Puvion, E. (1975). The nucleus of *Euglena I.* An ultracytochemical study of the nucleic acids and nucleoproteins of synchronized *Euglena gracilis* Z. *J. Ultrastruct. Res.* **52:**362 – 376.

Muchal, U. S. and S. D. Schwartzbach (1992). Characterization of a *Euglena* gene encoding a polyprotein precursor to the light-harvesting chlorophyll a/b-binding protein of photosystem II. *Plant Mol. Biol.* **18:**287-299.

Mukherjee, S. B., Das, M., Sudhandiran, G., and Shaha, C. (2002). “Increase in cytosolic $Ca^{2+}$ levels through the activation of nonselective cation channels induced by oxidative stress causes mitochondrial depolarization leading to apoptosis-like death in *Leishmania donovani* promastigotes,” *The Journal of Biological Chemistry*, **277 (27):** 24717 - 24727.

Müller Miklós, Mentel Marek, van Hellemond Jaap J., Henze Katrin, Woehle Christian, Gould Sven B., Yu Re-Young, van der Giezen Mark, Tielens Aloysius G. M., and Martin William F. (2012). Biochemistry and Evolution of Anaerobic Energy Metabolism in Eukaryotes. Microbiology and Molecular Biology Reviews, **76(2):**444 – 495.

Mullner, A. N., Angler, D. G., Samuel, R., Linton, E. W., Triemer, R. E. (2001). Phylogenetic analysis of phagotrophic, phototrophic, and osmotrophic

euglenoids by using the nuclear 18S rDNA sequence. Int. J. *Syst. Evol. Microbiol*., **51:**783–791.

Munday, J. C., Settimo, L. and de Koning, H. P. (2015).Transport proteins determine drug sensitivity and resistance in a protozoan parasite, *Trypanosoma brucei*. *Frontiers in Pharmacology* | *Drug Metabolism and Transport*, **6(32):** 1 – 10.

Murakami, K., Elmlund, H., Kalisman, N., Bushnell, D. A., Adams, C. M., Azubel, M., Elmlund, D., Levi-Kalisman, Y., Liu, X., Levitt, M., and Kornberg, R. D. (2013). Architecture of an RNA polymerase II transcription pre-initiation complex. *Science*. **342(6159):** 1238724. doi:10.1126/science.1238724.

Murphree, S., Stubblefield, E. and Moore, E. C. (1969). Synchronized mammalian cell cultures. 3. Variation of ribonucleotide reductase activity during the replication cycle of Chinese hamster fibroblasts. *Exp. Cell Res*., **58:**118-124.

Murray, J. M. (1981). Control of cell shape by calcium in the *Euglenophyceae. J. Cell Sci.,* **49:** 99 - 118.

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M, Fasulo, D. P., Flannigan, M. J., Kravitz, S. A., Mobarry, C. M, Reinert, K. H., Remington, K. A., Anson, E. L., Bolanos, R. A, Chou, H. H., Jordan, C. M., Halpern, A. L., Lonardi, S., Beasley, E. M., Brandon, R. C., Chen, L., Dunn, P. J., Lai, Z., Lian, Y., Nusskern, D. R., Zhan, M., Zhang, Q., Zheng, X., Rubin, G. M., Adams, M. D., and Venter, J. C. (2000). A Whole-Genome Assembly of Drosophila. *Science,* **287:**2196-2204.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., Snyder, M. *(2008).* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science,* **320:**1344–1349.

Nakano, Y., Urade, Y., Urade, R., Kitaoka, S. (1987). Isolation, purification, and characterization of the pellicle of *Euglena gracilis Z. J. Biochem.,* **102:**1053–1063

Nakazawa, M., Inui, H., Yamaji, R., Yamamoto, T., Takenaka, S., Ueda, M., Nakano, Y. and Miyatake, K. (2000). The origin of pyruvate: NADP1 oxidoreductase in mitochondria of *Euglena gracilis*. *FEBS Lett.,***479:**155–156.

Nakazawa, M., Inui, H., Yamaji, R., Yamamoto, T., Takenaka, S., Ueda, M., Nakano, Y. and Miyatake, K. (2000). The origin of pyruvate: NADP1 oxidoreductase in mitochondria of *Euglena gracilis*. *FEBS Lett.,* **479:**155–156.

Nakazawa, M., Nishimura, M., Inoue, K., Ueda, M., Inui, H., Nakano, Y. and Miyatake, K. (2011). Characterization of a Bifunctional Glyoxylate Cycle Enzyme, Malate Synthase/Isocitrate Lyase, of *Euglena gracilis*. *J. Eukaryot. Microbiol.,* **58(2):**128–133.

Nakazawa, M., Takenaka, S., Ueda, M., Inui, H., Nakano, Y., Miyatake, K. (2003). Pyruvate:NADP_oxidoreductase is stabilized by its cofactor, thiamin pyrophosphate, in mitochondria of *Euglena gracilis*. *Arch. Biochem. Biophys.,* **411**:183–188.

Nass, M. M. K., Schori, L., Ben-Shaul, Y. and Edelman, M. (1974). Size and Configuration of Mitochondrial DNA in *Euglena gracilis. Biochimica Et Biophysica Acta,* **374:** 283-291.

Nass, M. M., Schori, L., Ben-Shaul, Y., Edelman, M. (1974). Size and configuration of mitochondrial DNA in Euglena gracilis. *Biochim Biophys Acta,* **374:** 283 – 291.

Nassoury, N., Cappadocia, M. and Morse, D. (2003). Plastid ultrastructure defines the protein import pathway in dinoflagellates. *J. Cell Sci.,* **116:**2867–2874.

Nawa, H., Kotani, H., Nakanishi, S. (1984). Tissue-specific generation of two preprotachykinin mRNAs from one gene by alternative RNA splicing. *Nature,* **312 (5996):** 729–734.

Neiman, M., Hehman, G., Miller, J. T., Logsdon, J. M., and Taylor, D. R.. (2009). Accelerated Mutation Accumulation in Asexual Lineages of a Freshwater Snail**.** *Molecular Biology and Evolution*, **27(4):**954-63 DOI: 10.1093/molbev/msp300.

Neumann, N., Lundin, D., Poole, A. M. (2010). Comparative Genomic Evidence for a Complete Nuclear Pore Complex in the Last Eukaryotic Common Ancestor. *PLoS ONE,* **5(10):** e13241. doi:10.1371/journal.pone.0013241.

Neumann, N., Lundin, D., Poole, A. M. (2010). Comparative genomic evidence for a complete nuclear pore complex in the last eukaryotic common ancestor. *PLoS One,* **5:**e13241.

Newcomb, E. H. (1969). *Plant microtubules. A. Rev. Plant Physiol.* **20:**253-288.

Newton, A. C. (2001). Protein kinase C: structural and spatial regulation by phosphorylation, cofactors, and macromolecular interactions. *Chem. Rev.,***101:**2353–2364

Newton, A. C. (2003). Regulation of the ABC kinases by phosphorylation: protein kinase C as a paradigm. *Biochem. J.,* **370:**361–371.

Newton, A. C. (2003). Regulation of the ABC kinases by phosphorylation: protein kinase C as a paradigm. *Biochem. J.,* **370:**361–371.

Nickel, W. and Seedorf, M. (2008). Unconventional mechanisms of protein transport to the cell surface of eukaryotic cells. *Ann. Rev. of Cell and Dev. Biol.,* **24**: 287–308. *doi*:*10.1146/annurev.cellbio.24.110707.175320*.

Niedringhaus, T. P., Milanova, D., Kerby, M. B., Snyder, M. P., Barron, A. E. (2011). Landscape of next-generation sequencing technologies. *Anal. Chem.,* **83:** 4327–4341.

Nigg, E. (1992). Assembly-disassembly of the nuclear envelope. *Curr. Opin. Cell Biol.,* **4:**105-109.

Nikolenko, S. I., Korobeynikov, A. I. and Alekseyev, M. A. (2013). BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics,***14(1):**S7

Nishant, K. T., Chen, C., Shinohara, M., Shinohara, A., Alani, E. (2010). Genetic analysis of baker's yeast Msh4-Msh5 reveals a threshold crossover level for meiotic viability. *PLoS Genet.,* **6(8):**e1001083.

Nishimoto, M., and Kitaoka, M. (2007). Identification of N-acetylhexosamine 1-kinase in the complete lacto-N-biose I/galacto-N-biose metabolic pathway in Bifidobacterium longum. *Biosci., Biotechnol., Biochem.*, **71:**2101–2104.

Nissen, P., Hansen, J., Ban, N., Moore, P. B., and Steitz, T. A. (2000). The structural basis of ribosome activity in peptide bond synthesis. *Science,* **289:**920–930. http://dx.doi.org/10.1126/science.289.5481.920.

Nivison, H. T. and Jagendorf, A. T. (1984). Factors permitting prolonged translation by isolated pea chloroplasts. *Plant Physiol,* **75:** 1001-1008.

Noble, G. P., Rogers, M. B., and Keeling, P. J. (2007) Complex distribution of EFL andEF1alpha proteins in the green algal lineage.*BMC Evolutionary Biology,* ***7*:**82.

Nooden, L. D., Guiamet, J. J., John, I. (1997). Senescence mechanisms. *Physiologia Plantarum*, **101:** 746–753.

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

Nott, A., Jung, H.-S., Koussevitzky, S. and Chory, J. (2006). Plastid-tonucleus retrograde signaling. *Annu. Rev. Plant Biol.,* 57:739–759.

Noyes, H. (1998). Implications of a Neotropical origin of the genus Leishmania. *MemInst Oswaldo Cruz*, **93:**657–61.

Noyes, H. A., Arana, B. A., Chance, M. L., Maingon, R. (1997). The *Leishmania hertigi* (Kine-toplastida; Trypanosomatidae) complex and the lizard Leishmania: theirclassification and evidence for a neotropical origin of the Leishmania-Endotrypanum clade. *J Eukaryot Microbiol.,* **44:**511–7.

Ntefidou, M. and Hader, DP. (2005). Photoactivated adenylyl cyclase (PAC) genes in the flagellate *Euglena gracilis* mutant strains. *Photochem. Photobiol. Sci.*, **4:**7 3 2 – 7 3 9

NteWdou, M., Häder, D-P. (2005). Photoactivated adenylyl cyclase (PAC) genes in the Xagellate *Euglena gracilis* mutant strains. *Photochem. Photobiol. Sci.,* **4:**732–739.

NteWdou, M., Iseki, M., Watanabe, M., Lebert, M., Häder, D-P. (2003). Photoactivated adenylyl cyclase controls phototaxis in the Xagellate *Euglena gracilis*. *Plant Physiol.,* **133:**1517–1521.

Nunez, P. A., Romero, H., Farber, M. D., and Rocha, E. P. C. (2012). Natural Selection for Operons Depends on Genome Size. *Genome Biol. Evol.,* **5(11):**2242–2254. doi:10.1093/gbe/evt174.

O'Neil, S. T. and Emrich, S. J. (2013). Assessing *De Novo* transcriptome assembly metrics for consistency and utility. *BMC Genomics*, **14**:465.

O'Neill, E. C. and Field, R. A. (2015). Enzymatic synthesis using glycoside phosphorylases. *Carbohydrate Research,* **403:**23–37.

O'Neill, E. C., Trick, M., Henrissatd, B. and Field, R. A. (2015). *Euglena* in time: Evolution, control of central metabolic processes andmulti-domain proteins in carbohydrateand natural product biochemistry. *Perspectives in Science*, **6:**84 - 93. http://dx.doi.org/10.1016/j.pisc.2015.07.002.

O'Neill, E. C., Trick, M., Hill, L., Rejzek, M., Dusi, R. G., Hamilton, C. J., Zimba, P. V., Henrissat, B., and Field, R. A. (2015). The transcriptome of *Euglena gracilis* reveals unexpected metabolic capabilities for carbohydrate and natural product biochemistry. *Mol. BioSyst.,***11:** 2808.

O'Sborn, M., Weber, K. (1980). Dimethylsulfoxide and the ionophore A 23187 affect the arrangement of actin and induce nuclear actin paracrystals in PtK2 cells. *Exp. Cell Res.,* **129:** 103-114.

Obado, S. O., Brillantes, M., Uryu, K., Zhang, W., Ketaren, N. E., Chait, B. T., Field, M. C. and Rout, M. P. (2016). Interactome Mapping Reveals the Evolutionary History of the Nuclear Pore Complex. *PLoS Biol.,* **14(2):** e1002365. doi:10.1371/journal.pbio.1002365.

O'Brien, T. P. (1967). Cytoplasmic microtubules in the leaf glands of *Phaseolus vulgaris. J. Cell Sci.,* **2:**557-562.

O'Donnell, E. H. J. (1965). Nucleolus and Chromosomes in *Euglena gracilis*. *Cytologia*, **30(2):**118-154.

of eukaryotic pre-initiation complexes. *NATURE,* **483:** 295 – 301.

Ofengand, J. (2002). Ribosomal RNA pseudouridines and pseudouridine synthases. *FEBS Lett.,* 514:17–25.

Ogawa, T., Tamoi, M., Kimura, A., Mine, A., Sakuyama, H., Yoshida, E., Maruta, T., Suzuki, K., Ishikawa, T. and Shigeoka, S. (2015). Enhancement of photosynthetic capacity in *Euglena gracilis* by expression of cyanobacterial fructose-1,6-/sedoheptulose-1,7-bisphosphatase leads to increases in biomass and wax ester production. *Biotechnology for Biofuels*, **8:**80.

Okada, T. A. and Comings, D. E. (1979). Higher order structure of chromosomes. *Chromosoma,* **72:**1-14.

Okamura, K., Ishizuka, A., Siomi, H., and Siomi, M. C. (2004). Distinct roles for Argonaute proteins in small RNA-directed RNA cleavage pathways. *Genes Dev.,* **18(14):**1655-66.

O'Kelly, C. J. and Nerad, T. A. (1999). *Malawimonas jakobiformis* n. g., n. sp. (Malawimonadidae n. fam.), a Jakoba-like heterotrophic nanoflagellate with discoidal mitochondrial cristae. *The Journal of Eukaryotic Microbiology*, **46**: 522-531.

Omer, A. D., Ziesche, S., Decatur, W. A., Fournier, M. J., Dennis, P. P. (2003). RNA-modifying machines in archaea. *Mol. Microbiol*., **48(3):** 617–629.

Ometto, L., Stephan, W., and De Lorenzo, D. (2005). Insertion/Deletion and Nucleotide Polymorphism Data Reveal Constraints in *Drosophila melanogaster* Introns and Intergenic Regions. *Genetics,* **169(3):** 1521–1527. doi: 10.1534/genetics.104.037689

Omodeo, P. (1980). The photoreceptive apparatus of flagellated algal cells: comparative morphology and some hypotheses of functioning. F. Lenci, G. Colombetti (Eds.), Photoreception and Sensory Transduction in Aneural Organisms, Plenum Press, New York, **pp.** 127–153

Ophir, I., Talmon, A., Polak-Charcon, S., Ben-Shaul, Y. (1975). Aspects of structure and photosynthetic competence of Euglena plastid under condition of greening and degreening. *Protoplasma,* **84:** 283–295.

Opperdoes, F. R. (1984). Localization of the initial steps in alkoxyphospholipid biosynthesis in glycosomes (microbodies) of *Trypanosoma brucei*. *FEBS Lett.,* **169:** 35−39.

Opperdoes, F. R., and Borst, P. (1977). Localization of nine glycolytic enzymes in a microbody-like organelle in *Trypanosoma brucei*: the glycosome. *FEBS Lett.,* **80:**360–364. (doi:10.1016/0014-5793(77)80476-6).

Opperdoes, F. R., and Szikora, J. P. (2006). In silico prediction of the glycosomal enzymes of Leishmania major and trypanosomes. *Mol. Biochem. Parasitol*., **147:**193−206.

Orcival-Lafont, AM. and Calvayrac, R. (1974). Le plastidome chez *Euglena gracilis Z. J. Phycol.,* **10:** 300 - 307.

Orlean, P. and Menon, A. K. (2007). GPI anchoring of protein in yeast and mammalian cells, or: How we learned to stop worrying and love glycophospholipids. *Journal of Lipid Research*, **48(5):** 993-1011. DOI: 10.1194/jlr.R700002-JLR200

Osafune, T., Schiff, J. A., Hase, E. (1990). Immunogold localization of Lhcp-II apoprotein in the golgi of euglena. *Cell Struct Funct.,* **15:**99–105.

Osafune, T., Sumida, S., Schiff, J. A., and Hase, E. (1991). Immunolocalization of LHCPII apoprotein in the Golgi during light-induced chloroplast development in non-dividing *Euglena* cells. *J. Electron Microsc.,* **40:**41–47.

Osato, D., Rogers, K., Guo, Q., Li, F., Richmond, G., Klug, F., and Simpson, L. (2009). Uridine insertion/deletion RNA editing in trypanosomatid mitochondria: in search of the editosome. RNA, **15:**1338–1344.

Osato, D., Rogers, K., Guo, Q., Li, F., Richmond, G., Klug, F., Simpson, L. (2009). Uridine insertion/deletion RNA editing in trypanosomatid mitochondria: in search of the editosome. *RNA,* **15:**1338–1344.

Oshlack, A. and Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct,* **4:**14.

Osteryoung, K. W. and Nunnari, J. (2003). The division of endosymbiotic organelles. *Science,* **302(5651):**1698–1704.

Ostlund, G., T. Schmitt, K. Forslund, T. Köstler, D. N. Messina, S. Roopra, O. Frings, and E. L. Sonnhammer **(2010)**. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* **38 (Database issue)**:D196–D203. http://dx.doi.org/10.1093/nar/gkp931

Owens, K. J., Farmer, M. A., Triemer, R. E. (1988). The flagellar apparatus and reservoir/canal cytoskeleton of *Cryptoglena pigra* (Euglenophyceae). *J. Phycol.,* **24:** 520–528

Owens, L. and Malham, S. (2015). Review of the RNA Interference Pathway in Molluscs Including Some Possibilities for Use in Bivalves in Aquaculture. *J. Mar. Sci. Eng.,* **3:**87 – 99.

Page, F. C. and Blanton, R. L. (1985). The Heterolobosea (Sarcodina: Rhizopoda), a new class uniting the Schizopyrenida and the Acrasidae (Acrasida). *Protistologica,* **21**: 121-132.

Page, N. M. (2005). New challenges in the study of the mammalian tachykinins. *Peptides,* **26(8):**1356–1368.

Page, N. M., Woods, R.J., Gardiner, S.M., Lomthaisong, K., Gladwell, R.T., Butlin, D.J., Manyonda, I.T., Lowry, P. J. (2000). Excessive placental secretion of neurokinin B during the third trimester causes pre-eclampsia. *Nature,* **405(6788):** 797–800.

Page, N.M., Bell, N.J., Gardiner, S.M., Manyonda, I.T., Brayley, K.J., Strange, P.G., Lowry, P. J. (2003). Characterization of the endokinins: human tachykinins with cardiovascular activity. *Proc. Natl. Acad. Sci. USA,* **100(10):** 6245–6250.

Pagni, P. G., Walne, P., Wehry, E. L. (1976). Fluorimetric evidence for flavin in isolated eyespots of *Euglena gracilis* var Bacillaris. *Photochem. Photobiol.,* **24:**373–375

Palenchar, J. B. and Bellofatto, V. (2006). Gene transcription in trypanosomes. *Mol. Biochem. Parasitol.,* **146:**135–141.

Palmer, J. D., Delwiche, C. F. (1996). Second-hand chloroplasts and the case of the disappearing nucleus. *Proc. Natl Acad. Sci. USA,* **93:** 7432–7435.

Pancaldi, S., Baldisserotto C., Ferroni L., Bonora A., Fasulo M. P. (2002). Room temperature microspectrofluorimetry as a useful tool for studying the assembly of the PSII chlorophyll-protein complexes in single living cells of etiolated *Euglena gracilis* Klebs during the greening process. *Journal of Experimental Botany,* **53:** 1753–1763.

Panigrahi, A. K., Zikova, A., Dalley, R. A., Acestor, N., Ogata, Y., Anupama, A., Myler, P. J., Stuart, K.D. (2008). Mitochondrial complexes in Trypanosoma brucei: a novel complex and a unique oxidoreductase complex. *Mol. Cell. Proteomics,* **7:**534–545.

Paolo Gualtierl, Laura Barsantl, Vincenzo Passarelll, Franco Verne And Giovanna Rosate (1990). A Look Into The Reservoir Of *Euglena gracilis:* Sem Investigations Of The Flagellar Apparatus. *Micron and Microscopica .4cta,* **21(3):** 131 – 138

Park, J. H., Schofield, P. J., and Edwards, M. R. (1997). Pyruvate kinase is present in *Giardia intestinalis. Exp. Parasitol.,* **87**:153–156.

Parle-McDermott, A., McManus, E. J., Mills, J. L., O'Leary, V. B., Pangilinan, F., Cox, C., Weiler, A., Molloy, A. M., Conley, M., Watson, D., Scott, J. M., Brody, L. C., and Kirke, P. N. (2003). *Mol. Genet. Metab.,* **80:**463–468.

Parodi, A. J. (1993). N-glycosylation in trypanosomatid protozoa. *Glycobiology,* **3:**193–199.

Parra, G., Bradnam, K. and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics,***23:**1061–1067.

Parsons, M. (2004). Glycosomes: parasites and the divergence of peroxisomal purpose. *Mol. Microbiol.,* **53:**717−724.

Parthier, B. (1982). In: *Biochemistry and Physiology of Protozoa*, vol. 4, 2nd edn (Levandowsky, M. and Hutner, S.H. eds) **pp.** 261-300, Academic Press, New York.

Parthier, B. (1982). The cooperation of nuclear and plastid genomes in plastid biogenesis and differentiation. *Biochem Physiol Pflanz,* **177:** 283-317.

Pasquale, S. M., Goodenough, U. W. (1988). Calmodulin sensitivity of the flagellar membrane adenylate cyclase and signaling of motile responses by cAMP in gametes of Chlamydomonas reinhardtii. *Bot Acta*, **101:**118–122.

Passos-Silva, D. G., Rajao, M. A., de Aguiar, P. H. N., Vieira-da-Rocha, J. P., Machado,C. R., and Furtado,C. (2010). Overview of DNA Repair in *Trypanosoma cruzi*, *Trypanosoma brucei* and *Leishmania major*. *Journal of Nucleic Acids,* **840768, 14 pages** doi:10.4061/2010/840768

Patel, R. K. and Jain, M. (2012). NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. *PLoS ONE,* **7(2):** e30619. doi:10.1371/journal.pone.0030619

Patrick, K. L., Shi, H., Kolev, N. G., Ersfeld, K., Tschudi, C., and Ullu, E. (2009) Distinct and overlapping roles for two Dicer-like proteins in the RNA interference pathways of the ancient eukaryote Trypanosoma brucei. *Proc. Natl. Acad. Sci. USA*, **106:**17933–17938.

Patron, N. J., Waller, R. F., Archibald, J. M. and Keeling, P. J. (2005). Complex protein targeting to dinoflagellate plastids. *J. Mol. Biol.,* **348:**1015 – 1024.

Patterson, D. J. (1988). The evolution of Protozoa. Mem. Inst. Oswaldo Cruz, **83(Suppl. 1):**580-600.

Peacock, L., Ferris, V., Sharma, R., Sunter, J., Bailey, M., Carrington, M., and Gibson, W. (2011). Identification of the meiotic life cycle stage of *Trypanosoma brucei* in the tsetse fly. *PNAS,* **108(9):** 3671–3676.

Peck, R. K. (1977). The cortical ultrastructure of the somatic cortex of *Pseudomicrothorax dubius*: structure and function of the epiplasm in ciliated protozoa. *J. Cell Sci.,* **25**:367–385.

Peck, R. K. (1986). The trace element selenium induces trichocyst formation in the ciliated protozoan *Pseudomicrothorax dubius*. *Eur. J. Cell Biol.,* **41**:174–181.

Peck, R. K., Duborgel, F., Huttenlauch, I. and de Haller, G. (1991). The membrane skeleton of *Pseudomicrothorax*, I. Isolation, structure, and composition. *J. Cell Sci.,* **100**:693–706.

Peculis, B. A. and Steitz, J. A. (1993). Conserved boxes C and D are essential nucleolar localization elements of U14 and U8 snoRNAs. *Cell,* **73:**1233 - 1245.

Pedersen, S. F., Owsianik, G. and Nilius, B. (2005). TRP channels: an overview. *Cell calcium,* **38:**233-252.

Pellegrini, M. (1980). Three-Dimensional Reconstruction of Organelles in *Euglena Gracilis* Z. I. Qualitative and Quantitative Changes of Chloroplasts and Mitochondrial Reticulum in Synchronous Photoautotrophic Culture. *J. Cell Sci.,* **43:** 137 – 166

Pellegrini, M. (1980b). Three-dimensional reconstruction of organelles in *Euglena gracilis* Z. II. Qualitative and quantitative changes of chloroplasts and mitochondrial reticulum in synchronous cultures during bleaching. *J. Cell Sci.,* **46:**313—340.

Penkina, M. V., Karpova, O. I. and Bogdanov, Y. F. (2002). Synaptonemal complex proteins: Specific proteins of meiotic chromosomes. *Mol. Biol.,* **36:**304–313.

Pennefather, J.N., Lecci, A., Candenas, M.L., Patakd, E., Pintoc, F.M., Maggi, C. A.(2004). Tachykinins and tachykinin receptors: a growing family. *Life Sci.,* **74(12):** 1445–1463.

Peregrín-Alvarez, J. M., Sanford, C., and Parkinson, J. (2009). The conservation and evolutionary modularity of metabolism. *Genome Biology,* **10:**R63 (doi:10.1186/gb-2009-10-6-r63).

Pereira-Leal, J. B. (2008). The Ypt/Rab family and the evolution of trafficking in fungi. *Traffic,* **9:**27-38.

Pereira-Leal, J. B. and Seabra, M. C. (2001). Evolution of the Rab family of small GTP-binding proteins. *J. Mol. Biol.,* **313:**889-901.

Perez, E., Lapaille, M., Degand, H., Cilibrasi, L., Villavicencio-Queijeiro, A., Morsomme, P., González-Halphen, D., Field, M. C., Remacle, C., Baurain, D., Cardol, P. (2014). The mitochondrial respiratory chain of the secondary green alga *Euglena gracilis* shares many additional subunits with parasitic Trypanosomatidae. *Mitochondrion* xxx (2014) xxx–xxx. http://dx.doi.org/10.1016/j.mito.2014.02.001.

Perez, E., Lapaille, M., Degand, H., Cilibrasi, L., Villavicencio-Queijeiro, A., Morsomme, P., González-Halphen, D., Field, M. C., Remacle, C., Baurain, D., Cardol, P. (2014). The mitochondrial respiratory chain of the secondary green alga *Euglena gracilis* shares many additional subunits with parasitic Trypanosomatidae. *Mitochondrion,* **19 Pt B:**338-49. doi: 10.1016/j.mito.2014.02.001.

Perry, R. P. (1976). Processing of RNA. *Annu. Rev. Biochem*. **45:**605 - 629.

Petersen, J., Teich, R., Brinkmann, H., Cerff, R. (2006) A “green” phosphoribulokinase in complex algae with red plastids: evidence for a single secondary endosymbiosis leading to haptophytes, cryptophytes, heterokonts, and dinoflagellates. *J Mol Evol,* **62(2):**143-157.

Petersen, T. N., Brunak, S., von Heijne, G., Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods,* **8:**785–786 doi:10.1038/nmeth.1701.

Petersen, T. N., Brunak, S., von Heijne, G.andNielsen, H.(2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, **8**:785-786.

Petukhova, G. V., Pezza, R. J., Vanevski, F., Ploquin, M., Masson, J. Y., Camerini-Otero, R. D. (2005). The Hop2 and Mnd1 proteins act in concert with Rad51 and Dmc1 in meiotic recombination. *Nat. Struct. Mol. Biol.,* **12:**449–453.

Peyretaillade, E., Parisot,N., Polonais,V., Terrat,S., Denonfoux,J., Dugat-Bony,E., Wawrzyniak, I., Biderre-Petit, C., Mahul, A., Rimour, S., Gonçalves, O., Bornes S., Delbac, F., Chebance, B., Duprat, S., Samson, G., Katinka, M., Weissenbach, J., Wincker, P. and Peyret, P.(2012). Annotation of microsporidian genomes using transcriptional signals. *Nature Communications,* **3:**1137

Piccinni, E. and Omodeo, P. (1975). Photoreceptors and phototactic programs in Protista. *Boll. Zool.,* **42:** 57–79

Piccinni, E., Albergoni , V. (1973). ATPase activity in isolated flagella from *Euglena gracilis. J. Protozool.,* **20:**456–458

Piccinni, E., Albergoni, V., Coppellotti, O. (1975). ATPase activity in flagella from *Euglena gracilis.* Localization of the enzyme and effects of detergents. *J. Protozool.,* **22:**331– 335

Piccinni, E., Mammi, M. (1978). Motor apparatus of *Euglena gracilis*: ultrastructure of the basal portion of the flagellum and the para-flagellar body. *Boll. Zool.,* **45:**405–414.

Pienta, K. J., Coffey, D. S. (1984). A structural analysis of the role of the nuclear matrix and DNA loops in the organization of the nucleus and chromosome. *J Cell Sci Suppl.,* **1:**123-35.

Pingault, L., Choulet, F., Alberti, A., Glover, N., Wincker, P., Feuillet, C. and Paux, E. (2015). Deep transcriptome sequencing provides new insights into the structural and functional organization of the wheat genome. *Genome Biology,* **6**:29

Pinheiro, H., Samalova, M., Geldner, N., Chory, J., Martinez, A., and Moore, I. (2009). Genetic evidence that the higher plant Rab-D1 and Rab-D2 GTPases exhibit distinct but overlapping interactions in the early secretory pathway. *J. Cell Sci.,* **122:**3749–3758.

Plaumann, M., Pelzer-Reith, B., Martin, W., and Schnarrenberger, C. (1997). Cloning of fructose-1,6-bisphosphate aldolases from *Euglena gracilis*: multiple recruitment of class I aldolase to chloroplasts and eubacterial origin of eukaryotic class II aldolase genes. *Curr. Genet.,* **31**:430–438.

Plaxton, W. C. (1996). The organization and regulation of plant glycolysis. *Annu. Rev. Plant Physiol. Plant Mol. Biol.,* **47:**185–214. (doi:10.1146/annurev.arplant.47.1.185).

Pohl, P. (1973). Light-induced Changes of Radioactivities in the 14C-labeled Lipids and Fatty Acids of Dark Grown *Euglena gracilis. Z. Naturforsch.,* **28c:**264-269.

Pohl, P. (1973). Some Evidence for Light-induced Transfers of Fatty Acids in *Euglena gracilis. Z. Naturforschung,* **28c:**270-284.

Poinar, Jr. G. (2007). Early Cretaceous trypanosomatids associated with fossil sand flylarvae in Burmese amber. *Mem Inst Oswaldo Cruz,* **102:**635–7.

Poinar, Jr. G., Poinar, R. (2004). Evidence of vector-borne disease of Early Cretaceousreptiles. *Vector Borne Zoon Dis,* **4:**281-4.

Poleshko, A. and Katz, R. A. (2014). Specifying peripheral heterochromatin during nuclear lamina reassembly. *Nucleus,* **5(1):**32–39.

Polonyi, J., Ebringer, L., Dobias, J. and Krajcovic, J. (1998). Giant mitochondria in chloroplast-deprived Euglena gracilis late after N-succinimidylofloxacin treatment. *Folia Microbiol.,* **43:**661–666.

Pombert, J-F., James, E. R., Janous kovec, J., Keeling, P. J. (2012) Evidence for Transitional Stages in the Evolution of Euglenid Group II Introns and Twintrons in the Monomorphina aenigmatica Plastid Genome. *PLoS ONE,* **7(12):** e53433. doi:10.1371/journal.pone.0053433.

Porter, K. R. (1966). Cytoplasmic microtubules and their function. In Ciba Fdn Symp. *Principles of Biomolecular Organization* (ed. G. E. W. Wolstenholme & M. O'Connor), **pp.** 308-345. London: Churchill.

Porterfield, D. M. (1997). Orientation of motile unicellular algae to oxygen: oxytaxis in *Euglena*. *Biol. Bull.,***193:**229–230.

Poss, Z. C., Ebmeier, C. C. and Taatjes, D. J. (2013). The Mediator complex and transcription regulation.*Critical Reviews in Biochemistry and Molecular Biology,* **48:**6, 575-608.

Potashkin, J. A., Zeigel, R. F. Huberman J.A. (1984). Isolation and initial characterization of residual nuclear structure from yeast, Exp. Cell Res. **153:**374 - 388.

Potten, C. and Wilson, J. (2004). *Apoptosis: The Life and Death of Cells*, Cambridge University Press, Cambridge, UK. **pp.** 127–153.

Poxleitner, M. K., Carpenter, M. L., Mancuso, J. J., Wang, C. J., Dawson, S. C., Cande, W. Z. (2008). Evidence for karyogamy and exchange of genetic material in the binucleate intestinal parasite Giardia intestinalis. *Science*, **319:**1530–1533.

Praefcke, G. J., and McMahon, H. T. (2004). The dynamin superfamily: universal membrane tubulation and fission molecules?*Nature reviews Molecular cell biology,* **5(2):**133-147.

Prask, J. A. and Plocke, D. J. (1971). A role for zinc in the structural integrity of the cytoplasmic ribosomes of *E. gracilis. Plant Physiol., Lancaster* **48:** 150-155

Pray, L. (2008). Eukaryotic genome complexity. *Nature Education,* **1(1):**96.

Preisfeld, A., Berger, S., Busse, I., Liller, S., Ruppel, H. G. (2000). Phylogenetic analyses of various euglenoid taxa (Euglenozoa) based on18S rDNA sequence data. *J Phycol.,* **36:**220 – 226.

Preußer, C., Jae, N., Gunzl, A. and Bindereif, A. (ed.) (2012). RNA Metabolism in Trypanosomes. *Nucleic Acids and Molecular Biology*, **28:**49–77 DOI 10.1007/978-3-642-28687-2_3.

Preußer,C., Jae, N., Günzl, A., and Bindereif, A. (2012). Pre-mRNA Splicing in *Trypanosoma brucei*: Factors, Mechanisms, and Regulation. A. Bindereif (ed.), RNA Metabolism in Trypanosomes, *Nucleic Acids and Molecular Biology* 28, DOI 10.1007/978-3-642-28687-2_3.

Previato, J. O., Gorin, P. A., Mazurek, M., Xavier, M. T., Fournet, B., Wieruszesk, J. M., and Mendonca-Previato, L. (1990). Primary structure of the oligosaccharide chain of lipopeptidophosphoglycan of epimastigote forms of Trypanosoma cruzi. *J Biol. Chem.*, **265:**2518–2526.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 -- Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE,* **5(3):**e9490. doi:10.1371/journal.pone.0009490.

Pringsheim E. G, (1953). Salzwasser-Eugleninen. *Arch. Mikrobiol.* **18:** 149 – 164

Pringsheim, E. G., and Hovasse, R. (1950). 'Les relations de parente entre Euglenacees et Astasiacees.' *Arch. Zool. exp. gen.,* **86:**499 – 549.

Pringsheim, E. G. (1962). *Chlamydomonas pallens*, a new organism proposed for assays of vitamin B12. *Nature,* **195:**604.

Prole, D. L., and Taylor, C. W. (2013). Identification and Analysis of Putative Homologues of Mechanosensitive Channels in Pathogenic Protozoa. *Langsley G, ed. PLoS ONE*, **8(6):** e66068.

Provasoli, L., Hutner, S. H. and Schatz, A. (1948). Streptomycin-induced chlorophyll-less races of Euglena. *Proceedings of the Society for Experimental Biology and Medicine,* **69:** 279 – 282.

Prucca, C. G., Slavin, I., Quiroga R, Elias EV, Rivero FD, Saura A, Carranza,P. G., andLuján,H. D. (2008). Antigenic variation in Giardia lamblia is regulated by RNA interference. *Nature,* **456:**750–4.

Ptushkina, M., Malys, N. and McCarthy, J. E. (2004). eIF4E isoform 2 in *Schizosaccharomyces pombe* is a novel stress-response factor*. EMBO Rep.,* **5:**311–http://dx.doi.org/10.1038/sj.embor.7400088.

Purkanti, R. and Thattai, M. (2015). Ancient dynamin segments capture early stages of host–mitochondrial integration. *Proceedings of the National Academy of Sciences,* **112(9):** 2800-2805.

Puthiyaveetil, S., Kavanagh, T. A., Cain, P., Sullivan, J. A., Newell, C. A., Gray, J. C., Robinson, C., van der Giezen, M., Rogers, M. and Allen, J. F. (2008). The ancestral symbiont redox sensor kinase CSK couples photosynthesis to gene transcription in chloroplasts. *Proc. Natl. Acad. Sci. USA,* **105:**10061–10066.

Rajan Sankaranarayanan and Dino Moras (2001). The fidelity of the transla tion of the genetic code. *Acta Biochimica Polinica,* 48**(2):** 323–335.

Ram Kishore and Steven D. Schwartzbach (1992). Translational Control of the Synthesis of the *Euglena* Light Harvesting Chlorophyll A/B Binding Protein of Photosystem II. *Plant Science,* **85:**79 – 89.

Ramadas, R. and Thattai, M. (2013). New organelles by gene duplication in a biophysical model of eukaryote endomembrane evolution. *Biophys. J.,* **104(11):**2553–2563.

Ramesh, M. A., Malik, S. B. and Logsdon, J. M., Jr. (2005) A phylogenomic inventory of meiotic genes: Evidence for sex in Giardia and an early eukaryotic origin of meiosis. *Curr. Biol.,* **15:**185–191.

Rapoport, T. (2007). "Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes". *Nature,* **450** (7170): 663–9. *doi*:*10.1038/nature06384*.

Ratcliffe, H. (1927). Mitosis and Cell Division in *Euglena spirogyra Ehrenberg*. *Biological Bulletin, 53*(2), 109-122. doi:10.2307/1537032

Ravel-Chapuis, P. (1988). Nuclear rDNA in *Euglena gracilis*: paucity of chromosomal units and replication of extrachromosomal units. *Nucleic Acids Res.* **16:**4801 - 4810.

Ravel-Chapuis, P., Flamant, F., Nicolas, P., Heizmann, P. and 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(2):**1039-48.

Ravel-Chapuis, P., Nicolas, P., Nigon, V., Neyret, O., Freyssinet, G. (1985). Extrachromosomal circular nuclear rDNA in *Euglena gracilis*. *Nucleic Acids Res.,* **13:**7529 - 7537.

Rawson, J. R. and Boerma, C. L. (1976) Influence of growth condition upon the number of chloroplast DNA molecules in *Euglena gracilis*. *Proc. Natl. Acad. Sci. USA,* **73:**2401–2404.

Rawson, J. R. Y. (1975). The characterization of *Euglena gracilis* DNA by its reassociation kinetics. *Biochim. Biophys. Acta*, **402:**171-178.

RayChaudhuri, A., Hait, N. C., Dasgupta, S., Bhaduri, T. J., Deb, R., and Majumder, A. L. (1997). L-myo-lnositol 1-Phosphate Synthase from Plant Sources (Characteristics of the Chloroplastic and Cytosolic Enzymes). *Plant Physiol.,* 115: 727-736.

Reardon, E. M., Price, C. A. (1983). Cytoplasmic regulation of chloroplast translation in Euglena gracilis. *Arch Biochem Biophys,* **226:** 433 – 440.

Reed, J. L., Famili, I., Thiele, I., Palsson, B. O. (2006). Towards multidimensional genome annotation. *Nat. Rev. Genet.,* **7(2):**130-41.

Reeder, R. H. (1990). rRNA synthesis in the nucleolus. *Trends Genet.,* **6:**390-395.

Reichard, P. (1993). From RNA to DNA, why so many ribonucleotide reductases? *Science,* **260:**1773 – 1777.

Reinbothe, S., Reinbothe, C., Krauspe, R., Parthier, B. (1991). Changing gene expression during dark-induced chloroplast dedifferentiation in *Euglena gracilis*. *Plant Physiology and Biochemistry*, **29:** 309–318.

Ren, Q., Chen, K. and Paulsen, I. T. (2007). TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic Acids Research,* **35:**D274–D279

Ren, X. Y. and Fiers, M. W., Stiekema, W. J., Nap, J. P. (2005). Local coexpression domains of two to four genes in the genome of Arabidopsis. *Plant Physiol.,* **138:**923–34.

Ren, X. Y., Stiekema, W. J., Nap, J. P. (2007). Local coexpression domains in the genome of rice show no microsynteny with Arabidopsis domains. *Plant Mol Biol.,* **65:**205–17.

Reyes-Prieto, A., Hackett, J. D., Soares, M. B., Bonaldo, M. F., Bhattacharya, D. (2006). Cyanobacterial contribution to algal nuclear genomes is primarily limited to plastid functions. *Curr. Biol.,* **16:**2320–2325.

Rhee, H. S. and Pugh, B. F. (2012). Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature,* **483:** 295 – 301.

Richards, O. C. (1967). Hybridization of *Euglena gracilis* chloroplast and nuclear DNA. *Biochemistry,* **57:** 156 – 163.

Richter, J. D. and Sonenberg, N. (2005). Regulation of cap-dependent translation by eIF4E inhibitory proteins. *Nature,* **433:**477–480. http://dx.doi.org/10.1038/nature03205.

Richter, P. R., Liu, Y., An, Y., Li, X., Nasir, A., Strauch, S. M., Becker, I., Krüger, J., Schuster, M., Ntefidou, M., Daiker, V., Haag, F. W. M., Aiach, A., Lebert, M. (2015). Amino acids as possible alternative nitrogen source for growth of *Euglena gracilis* Z in life support systems. *Life Sciences in Space Research,* **4:** 1–5.

Richter, P. R., Schuster, M., Lebert, M., Streb, C. and Häder, D-P. (2007). Gravitaxis of *Euglena gracilis* depends only partially on passive buoyancy. *Adv Space Res.,* **39:**1218–1224.

Richter, P. R., Streb, C., NteWdou, M., Lebert, M., Häder, D-P. (2003). High light-induced sign change of gravitaxis in the Xagellate *Euglena gracilis* is mediated by reactive oxygen species. *Acta Protozool.,* **42:**197–204.

Richter, P., Lebert, M., Korn, R. and Häder, D-P. (2001a). Possible involvement of the membrane potential in the gravitactic orientation of *Euglenagracilis*. *J Plant Physiol.,* **158:**35–39.

Richter, P., Lebert, M., Tahedl, H. and Häder D-P (2001b). Calcium is involved in the gravitactic orientation in colorless fagellates. *J. Plant Physiol.,* **158:**689–697.

Ridgley, E. L., Xiong, Z. H., and Ruben, L. (1999). “Reactive oxygen species activate a $Ca^{2+}$-dependent cell death pathway in the unicellular organism *Trypanosoma brucei brucei*”. *Biochemical Journal*, **340(1):** 33 - 40.

Ringertz, N. R. (1969). Handbook of molecular cytology (ed A Lima-de-Faria) **pp.** 656. North-Holland, Amsterdam.

Ris, H. (1966). Fine structure of chromosomes. *Proc R Soc Lond [Biol],* **164:**246-257

Rizzo, P. J. (1981). Comparative aspects of basic chromatin proteins in dinoflagellates. *BioSystems*, **14:**433-443.

Robenek, H., Melkonian, M. (1983). Structural specialization of the paraflagellar body membrane of *Euglena.* Protoplasma, **117:**154–157

Roberts, K. R., and Roberts, J. E. (1991). The flagellar apparatus and cytoskeleton of the dinoflagellates: a comparative overview. *Protoplasma,* **164**:105–122.

Robinson, M. S. (2004). Adaptable adaptors for coated vesicles. *Trends in Cell Biology,* **14(4):**167-174. Doi: 10.1016/j.tcb.2004.02.002.

Rochaix J. D. (1996). Post-transcriptional regulation of chloroplast gene expression in *Chlamydomonas*. *Plant Mol. Biol.,* **32:**327 – 341.

Rodríguez-Zavala, J. S., García-García, J. D., Ortiz-Cruz, M. A., Moreno-Sánchez, R. (2007). Molecular mechanisms of resistance to heavy metals in the protist Euglena gracilis. *J Environ Sci Health A Tox Hazard Subst Environ Eng.,* **42(10):**1365-78.

Rodríguez-Zavala, J. S., García-García, J. D., Ortiz-Cruz, M. A., Moreno-Sánchez, R. (2000). Molecular mechanisms of resistance to heavy metals in the protist *Euglena gracilis*. *J. Environ. Sci. Health part A. Toxicol. Hazard. Subst. Environ. Eng.,* **42:**1365–1378.

Rodriguez-Zavala, J. S., Ortiz-Cruz, M. A., Mendoza-Hernandez, G., Moreno-Sanchez, R. (2010). Increased synthesis of alpha-tocopherol, paramylon and tyrosine by Euglena gracilis under conditions of high biomass production. *J Appl Microbiol.,* **109(6):** 2160 - 2172.

Roe, B. A., Anandaraj, M. P. J. S., Chia, L. S. Y., Ran5trath, X., Gupta, R. C. and Randerath, K. (1975). *Biochem. Biophys. Res. Comm.,* **66:** 1096-1105.

Rogalski, A. A., Bouck, G. B. (1980). Characterization and localization of a flagellar-specific membrane glycoprotein in *Euglena. J. Cell. Biol.,* **86:**424–435

Rogalski, A. A., Bouck, G. B. (1982). Flagellar surface antigenes in *Euglena*: Immunological evidence for an external glycoprotein pool and its transfer to the regenerating flagellum. *J. Cell Biol.,* **93:**758–766

Roger, A. J. and Simpson, A. G. (2009). Evolution: revisiting the root of the eukaryote tree. *Curr. Biol.,* **19:**R165-R167.

Rogers, K., Gao, G., Simpson, L.(2007). Uridylate-specific 30–50-exoribonucleases involved in uridylate-deletion RNA editing in trypanosomatid mitochondria. *J. Biol. Chem.,* **282:**29073–29080.

Rogers, M. B., Gilson, P. R., Su, V., McFadden, G. I. and Keeling, P. J. (2007). The complete chloroplast genome of the chlorarachniophyte *Bigelowiella natans*: evidence for independent origins of chlorarachniophyte and euglenid secondary endosymbionts. *Mol. Biol. Evol.,* **24:**54–62.

*Rogers, M. E.* (2012). The role of *Leishmania* proteophosphoglycans in sand fly transmission and infection of the mammalian host. *Frontiers in Microbiology | Microbial Immunology,* **3(223):** 1 – 13.

Rojas, A. M., Fuentes, G., Rausell, A., and Valencia, A. (2012). The Ras protein superfamily: Evolutionary tree and role of conserved amino acids. *J. Cell Biol.,* **196(2):**189–201

Rojas, A. M., Fuentes, G., Rausell, A., and Valencia, A.(2012). The Ras protein superfamily: Evolutionary tree and role of conserved amino acids. *J. Cell Biol.,* **196(2):**189–201 doi/10.1083/jcb.201103008.

Ron Caspi, Tomer Altman, Joseph M. Dale, Kate Dreher, Carol A. Fulcher, Fred Gilham, Pallavi Kaipa, Athikkattuvalasu S. Karthikeyan, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Lukas A. Mueller, Suzanne Paley, Liviu Popescu, Anuradha Pujar, Alexander G. Shearer, Peifen Zhang and Peter D. Karp (2010). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research,* **38(Database Issue)**:D473–D479.

Rosati, G., Verni, F., Barsanti, L., Passarelli, V. and Gualtieri, P. (1991). Ultrastructure of the Apical Zone of *Euglena gracilis:* Photoreceptors And Motor Apparatus. *Electron Microsc. Rev.,* **4:** 319 – 342.

Rosati, G., Verni, F., Barsanti, L., Passarelli, V. and Gualtieri, P. (1991). Ultrastructure of the Apical Zone of *Euglena gracilis:* Photoreceptors And Motor Apparatus. *Electron Microsc. Rev.,* **4:** 319 – 342

Rosati, G., Verni, F., Barsanti, L., Passarelli, V., Gualtieri, P. (1991). Ultrastructure of the apical zone of *Euglena gracilis*: photoreceptors and motor apparatus. *Electron Microsc. Rev*., **4:** 319 –342.

Rosembaum, J. L., Child, F. M. (1967). Flagellar regeneration in protozoan flagellates. *J. Cell Biol.,* **34:**345–364

Rosenberg, A. (1967). Galactosyl Diglycerides: Their Possible Function in Euglena Chloroplasts. *Science,* **157 (3793):** 1191-1196. DOI: 10.1126/science.157.3793.1191.

Ross, G. I. M. (1952). Vitamin BI2 assay in body fluids using *Euglena gracilis. Journal of Clinical Pathology,* **5:** 250-256.

Rostislav Vacula, Silvia Sláviková, and Steven D. Schwartzbach (2007). Protein Trafficking to the Complex Chloroplasts of *Euglena. Methods in Molecular Biology, Vol. 390: Protein Targeting Protocols: Second Edition* Edited by: M. van der Giezen © Humana Press Inc., Totowa, NJ.

Roth, J. R., Lawrence, J. G., and Bobik, T. A. (1996). Cobalamin (coenzyme B12): synthesis and biological significance. *Annu. Rev. Microbiol.,* **50:**137–181.

Rotte, C., Stejskal, F., Zhu, G., Keithly, J. S., Martin, W. (2001). Pyruvate: NADP_oxidoreductase from the mitochondrion of *Euglena gracilis* and from the apicomplexan *Cryptosporidium parvum*: a biochemical relic linking pyruvate metabolism in mitochondriate and amitochondriate protists. *Mol. Biol. Evol.,* **18**:710 –720.

Rout, M. P., Aitchison, J. D., Suprapto, A., Hjertaas, K., Zhao, Y. and Chait, B. T. (2000). The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J Cell Biol.,* **148(4):**635–51.

Roy, J., Faktorovab, D., Lukes, J., and Burger, G. (2007). Unusual Mitochondrial Genome Structures throughout the Euglenozoa. *Protist,* **158:**385 – 396.

Roy, J., Faktorovab, D., Lukes, J., and Burger, G. (2007). Unusual Mitochondrial Genome Structures throughout the Euglenozoa. *Protist,* **158:**385 - 396

Ruan, J. P., Arhin, G. K., Ullu, E. and Tschudi, C. (2004) Functional characterization of a *Trypanosoma brucei* TATA-binding protein-related factor points to a universal regulator of transcription in trypanosomes. *Mol. Cell. Biol.,* **24:** 9610–9618.

Ruhela, D., Banerjee, P., and Vishwakarma, R. A. (2012). Chemistry and biology of glycosylphosphatidylinositol molecules. *Current Science (00113891),* **102 (2):** 194 -211.

Rusche, L.N., Huang, C.E., Piller, K.J., Hemann, M., Wirtz, E., Sollner-Webb, B., (2001). The two RNA ligases of the *Trypanosoma brucei* RNA editing complex: cloning the essential band IV gene and identifying the band V gene. *Mol. Cell. Biol.,* **21:**979–989.

Russel, A., Watanabe, Y., Charette, J. and Gray, M. (2005). Unusual features of fibrillarin cDNA and gene structure in Euglena gracilis: evolutionary conservation of core protein sand structural predictions formethylationguidebox C/D snoRNPs throughout the domain Eucarya. *Nucleic Acids Res.,* **33:**2781–2791.

Russell, A. G., Schnare, M. N., Gray, M. W. (2006). A large collection of compact box C/D snoRNAs and their isoforms in Euglena gracilis: Structural functional and evolutionary insights. *J. Mol. Biol.,* **357(5):**1548–1565.

Russell, A. G.,Watanabe, Y., Charette, J. M., Gray, M. W. (2005). Unusual features of fibrillarin cDNA and gene structure in *Euglena gracilis*: evolutionary conservation of core proteins and structural predictions for methylation-guide box C/D snoRNPs throughout the domain. *Eucarya. Nucleic Acids Res.,* **33:**2781–2791.

Rust Alistair G., Mongin Emmanuel and Birney Ewan (2012). Genome annotation techniques: new approaches and challenges. Reviews: Information biotechnology supplement, **7(11):** S70 – S76.

Sadgopal, A and Bonner, J. (1970). *Biochim biophys acta,* **207:**106.

Saidha, T. and Schiff, J. A. (1986). *Plant Physiol.* 80S: 84

Saidha, T., Stern, A. I., Lee, D. H. and Schiff, J. A. (1985). Localization of a sulphate-activating system within Euglena Mitochondria. *Biochem. J.* **232:** 357-365.

Saier, M. H. Jr. (2000) A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiol. Mol. Biol. Rev.,* **64:**354–411.

Saint-Guily, A., Schantz, M. L. and Schantz, R. (1994). Structure and expression of a cDNA encoding a histone H2A from Euglena gracilis. *Plant Mol. Biol.,* **24:**941–948.

Sakaguchi, M., Takishita, K., Matsumoto, T., Hashimoto, T., and Inagaki, Y. (2009). Tracing back EFL gene evolution in the cryptomonads-haptophytes assemblage: separate origins of EFL genes in haptophytes, photosynthetic cryptomonads, and goniomonads. *Gene,* 441:126-131.

Sakato, M., Sakakibara, H., King, S. M. (2007). *Chlamydomonas* outer arm dynein alters conformation in response to $Ca^{2+}$. *Mol Biol Cell,* **18:**3620–3634.

Sakyiama,J., Zimmer,S. L., Ciganda,M., Williams,N., and Read,L. K. (2013). Ribosome biogenesis requires a highly diverged XRNfamily 5′→3′ exoribonuclease for rRNA processingin *Trypanosoma brucei*.*RNA,* **19:**1–13.

Salavati, R., Moshiri, H., Kala, S. and Najafabadi, H. S. (2012). Inhibitors of RNA editing as potential chemotherapeutics against trypanosomatid pathogens. International Journal for Parasitology: *Drugs and Drug Resistance*, **2:**36–46.

Salisbury, J. L., Floyd, G. L. (1978). Calcium-induced contraction of the rhizoplast of a quadriflagellate green alga Science, **202:** 975–977

Samorodnitsky, E. and Pugh, B. F. (2010). Genome-Wide Modeling of Transcription Preinitiation Complex Disassembly Mechanisms using ChIP-chip Data. *PLoS Comput. Biol*., **6(4):** e1000733. doi:10.1371/journal.pcbi.1000733.

Sanchez-Perez, G. F., Hampl, V., Simpson, A. G. B., Roger, A. J. (2008). A new divergent type of eukaryotic methionine adenosyltransferase is present in multiple distantly related secondary algal lineages. *J Eukaryot Microbiol.,* **55(5):**374-381.

Sanger, J. W., Ganger, J. M., Kreis, T. E., Jockusch, B. M. (1980). Reversible translocation of cytoplasmic actin into the nucleus caused by dimethylsulfoxide. *Proc. nat. Acad. Sci. U.S.A.,* **77:** 5268 - 5272.

Santaguida, S., and Musacchio, A. (2009). The life and miracles of kinetochores. *EMBO J.,* **28:**2511–2531.

Santis-MacIossek, G., Kofer, W., Bock, A., Schoch, S., Maier, R. M., Wanner, G., Rudiger, W., Koop, H. U. and Herrmann, R. G. (1999). Targeted disruption of the plastid RNA polymerase genes rpoA, B and C1: molecular biology, biochemistry and ultrastructure. *Plant J.*, **18:**477–489.

Satheesh, K. S., Gokulasuriyan, R. K. and Monidipa, G. (2014). Comparative in-silico genome analysis of Leishmania (Leishmania) donovani: A step towards its species specificity. *Meta Gene,* **2:** 782–798.

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., and Tabata, S. (1999). Complete Structure Of The Chloroplast Genome of *Arabidopsis Thaliana. DNA RESEARCH,* **6:**283-290.

Savino, R., Gerbi, S. A. (1990). *EMBO J.* **9:**2299 - 2308.

Scaife, R. M., Wilson, L., and Purich, D. L. (1992). Microtubule protein ADP-ribosylation *in vivo* leads to assembly inhibition and rapid depolymerization. *Biochemistry,* **31:**310 - 316.

Schantz, M. and Schantz, R. (1989). Sequence of a cDNA clone encoding β-tubnlin from *Euglena gracilis. Nucleic Acids Research,***17(16):** 6727.

Scheer, A., Parthier, B. (1982). Dark-induced chloroplast dedifferentiation in *Euglena gracilis*. *Planta,* **156:**274–281.

Scheuerlein, R., Treml, S., Thar, B., Tirlapur, U. K., and Hader, D. P. (1995). "Evidence for UV-B-induced DNA degradation in *Euglena gracilis* mediated by activation of metal-dependent nucleases,"*Journal of Photochemistry and Photobiology B*, **31(3):** 113 - 123.

Schiff, A. J. and Epstein, H. T. (1968). In The ***Biology of Euglena*** Vol. II (Buetow, D. E., ed.), **pp.** 286. Academic Press, London.

Schiff, J. A. (1971). The informational and nutritional requirement of cellular organelles. *Stadler Symp.,* **3:** 89 -114.

Schiff, J. A. (1973). The development, inheritance and origin of the plastid in *Euglena.* In *Advances in Morphogenesis,* vol. 10 (ed. M. Abercrombie, J. Brachet & T. J. King), **pp.** 265 - 312. New York: Academic Press.

Schiff, J. A., Epstein, H. T. (1968). The continuity of the chloroplast in Euglena. In: Buetow D.E. (Ed.), The Biology of *Euglena*. *Biochemistry, Academic Press, London*: 286–333.

Schiff, J. A., Lyman, H. and Epstein, H. T. (1961). Studies of chloroplast development in *Euglena.* II. Photoreversal of the U.V. inhibition of green colony formation. *Biochim. biophys. Acta,* **50:** 310-318.

Schiff, J. A., Lyman, H. and Russel, G. A. (1971). Isolation of mutants from *Euglena gracilis*. *Meth. Enzymol.,* **23A:**143–162.

Schiff, J. A., Zeldin, M. H. and Rubman, J. (1967). Chlorophyll formation and photosynthetic competence in Euglena during light-induced chloroplast development in the presence of 3 (3,4-dichlorophenyl) r,i-dimethyl urea (DCMU). *Plant Physiology*, **42:**1716 - 1725.

Schlacht, A. and Dacks, J. B. (2015). Unexpected Ancient Paralogs and an Evolutionary Model for the COPII Coat Complex.*Genome Biol. Evol.,* **7(4):**1098–1109. doi:10.1093/gbe/evv045

Schlacht, A., Herman E. K., Klute, M. J., Field, M. C. and Dacks, J. B. (2014). Missing Pieces of an Ancient Puzzle: Evolution of the Eukaryotic Membrane-Trafficking System. *Cold Spring Harb Perspect Biol.* **6:**a016048. doi: 10.1101/cshperspect.a016048.

Schmidt, M., Geßner, G., Lu, V. M., Heiland, I., Wagner, V., Kaminski, M., Geimer, S., Eitzinger, N., Reißenweber, T., Voytsekh, O., Fiedler, M., Mittag, M., Kreimer, G. (2006). Proteomic analysis of the eyespot of *Chlamydomonas reinhardtii* provides novel insights into its components and tactic movements. *Plant Cell*, 18:1908–1930.

Schnare, M. N. and Gray, M. W. (1990). Sixteen discrete RNA components in the cytoplasmic ribosome of *Euglena gracilis*. *J. Mol. Biol.,* **215:**73–83.

Schnare, M. N. and Gray, M. W. (2011). Complete Modification Maps for the Cytosolic Small and Large Subunit rRNAs of *Euglena gracilis*: Functional and Evolutionary Implications of Contrasting Patterns between the two rRNA Components. *J. Mol. Biol.,* **413:**66–83. doi:10.1016/j.jmb.2011.08.037.

Schnare, M. N., Cook, J. R. and Gray, M. W. (1990). Fourteen internal transcribed spacers in the circular ribosomal DNA of *Euglena gracilis*. *J. Mol. Biol.,* **215:**85–91.

Schnare, M. N., Gray, M. W. (1990). Sixteen discrete RNA components in the cytoplasmic ribosome of Euglena gracilis. *J Mol Biol.* 215:73–83.

Schnare, M. N., Gray, M. W. (2011). Complete modification maps for the cytosolic small and large subunit rRNAs of *Euglena gracilis*: functional and evolutionary implications of contrasting patterns between the two rRNA components. *J Mol Biol.,* **413(1):**66-83. doi: 10.1016/j.jmb.2011.08.037.

Schnarrenberger, C., Jacobshagen, S., Muller, B., and Kruger, I. (1990). Evolution of isozymes of sugar metabolism in green algae. **pp.** 743 –746 *in* S. Ogita and J. S. Scandalios, eds. Isozymes: structure, function, and use in biology and medicine. Wiley-Liss, New York.

Schnaufer, A., Panigrahi, A.K., Panicucci, B., Igo Jr., R.P., Wirtz, E., Salavati, R., Stuart, K. (2001). An RNA ligase essential for RNA editing and survival of the bloodstream form of Trypanosoma brucei. *Science,* **291:**2159–2162.

Schneider, A., Bursa, D., Lithgow, T. (2008). The direct route: a simplified pathway for protein import into the mitochondrion of trypanosomes. *Trends Cell Biol.,* **18:**12–18.

Schneider, A., Martin, J., Agabian, N. (1994). A nuclear encoded tRNA of Trypanosoma brucei is imported into mitochondria. *Mol Cell Biol.,* **14:** 2317–2322.

Schneider, T., Betz, A. (1985). Wax ester fermentation in *Euglena gracilis* T. Factors favouring the synthesis of odd-numbered fatty acids and alcohols. *Planta,* **166:**67–73.

Schnepf, E. and Deichgraber, G. (1972). Tubular inclusions in the endoplasmic reticulum of the gland hairs of *Ononis repens* L. (Fabaceae). *J. Microscopie,* **14:**361-374.

Schulz, M. H., Zerbino, D. R., Vingron, M. and Birney, E. (2012). *Oases:* robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics; Genome Analysis,* **28(8):**1086–1092 doi:10.1093/bioinformatics/bts094

Schurko, A. M. and Logsdon, J. M. (2008). Using a meiosis detection toolkit to investigate ancient asexual “scandals” and the evolution of sex. *Bioessays,* **30:**579–589.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. and Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature,* **473:** 337–342.

Schwartzbach, S. D. and Schiff, J. A. (1983). Control of plastogenesis in Euglena. In: Shropshire W. Jr, Mohor H. (Eds), *Encyclopaedia of Plant Physiology,* New Series. Springer Verlag, Berlin **(16A):** 312–335.

Schwartzbach, S. D., Freyssinet, G. and Schiff, J. A. (1973). Binding of dihydrostreptomycin to Euglena chloroplast ribosomes prepared by an improved procedure. *Plant Physiology,* **51:** S27.

Scott, E. C., Rodney, L. H. and Ira, F. G. (2005). Microsatellite variation and evolution in the Peromyscus maniculatus species group. *Mol. Phylogen. Evol.,* **34:**408 – 415.

Scott, J. (1995). A place in the world for RNA editing. *Cell,* **81:**833 – 836.

Scott, M. (2008). Drugs and Transporters in Kinetoplastid Protozoa. In: Majumder H.K. (eds) Drug Targets in Kinetoplastid Parasites. Advances In Experimental Medicine And Biology, vol 625. Springer, New York, NY.

Scrutton, M. C, Wu, C. W. and Goldthwait, D. A. (1971). The presence and possible role of zinc in RNA polymerase obtained from *E. coli. Proc. natn. Acad. Sci. U.S.A.,* **68:** 2497-2501

Secundino, N., Kimblin, N., Peters, N. C., Lawyer, P., Capul, A. A., Beverley, S. M., Turco, S. J., and Sacks, D. (2010). Proteophosphoglycan confers resistance of *Leishmania major* to midgut digestive enzymes induced by blood feeding in vector sand flies.*Cell Microbiol*., **12(7):** 906–918. doi:10.1111/j.1462-5822.2010.01439.x

Secundino, N., Nicola, K., Peters, N. C., Lawyer, P., Capul, A. A., Beverley, S. M., Turco, S. J., and Sacks, D.(2010). Proteophosphoglycan confers resistance of *Leishmania major* tomidgut digestive enzymes induced by blood feeding in vector sand flies.*Cell Microbiol*., **12(7):** 906–918. doi:10.1111/j.1462-5822.2010.01439.x.

Segev, N. (2011). Coordination of intracellular transport steps by GTPases. *Seminars in Cell & Developmental Biology,* **22:**33–38.

Segev, N. (2011). GTPases in intracellular trafficking: An overview. *Seminars in Cell & Developmental Biology,* **22:**1–2.

Segura-Valdez, M. L., Agredano-Moreno, L. T.,Nepomuceno-Mejía, T., Fragoso-Soriano, R., Álvarez-Fernández, G., Zamora-Cura, A., Lara-Martínez, R. andJiménez-García, L. F. (2013). Cell Nanobiology. In: *Oxidative Stress and Chronic Degenerative Diseases - A Role for Antioxidants,* Dr. Jose Antonio Morales-Gonzalez (Ed.), *InTech*, DOI: 10.5772/52003. Available from: http://www.intechopen.com/books/oxidative-stress-and-chronic-degenerative-diseases-a-role-for-antioxidants/cell-nanobiology.

Seiwert, S.D., Heidmann, S., and Stuart, K., (1996). Direct visualization of uridylate deletion in vitro suggests a mechanism for kinetoplastid RNA editing. *Cell,* **84:**831–841.

Selvapandiyan, A., Kumar, P.,Morris, J. C.,Salisbury, J. L.,Wang, C. C.,and Nakhasi, H. L. (2007). Centrin1 is required for organelle segregation and cytokinesis in *Trypanosoma brucei*. *Mol. Biol.Cell,* **18**:3290 –3301.

Shanley, D. P., Kirkwood, T. B. L. (2000). Calorie restriction and aging: a life history analysis. *Evolution*, **54:** 740 – 750.

Shao, P., Yang, J. H., Zhou, H., Guan, D. G. and Qu, L. H. (2009). Genome-wide analysis of chicken snoRNAs provides unique implications for the evolution of vertebrate snoRNAs. *BMC Genomics,* **10:**86. **DOI:** 10.1186/1471-2164-10-86

Sharma, M. R., Wilson, D. N., Datta, P. P., Barat, C., Schluenzen, F., Fucini, P. and Agrawal, R. K. (2007). Cryo-EM study of the spinach chloroplast ribosome reveals the structural and functional roles of plastid-specific ribosomal proteins. *Proc. Natl. Acad. Sci. U.S.A.,* **104:**19315–19320.

Sharpless, T. K., Butow, R. A. (1970a). An inducible alternate terminal oxidase in *Euglena gracilis* mitochondria. *J. Biol. Chem*., **245:**58 –70.

Shawn T O'Neil and Scott J Emrich (2013). Assessing *De Novo* transcriptome assembly metrics for consistency and utility. *BMC Genomics*, **14**:465

Shay, J. W. (1972). Electron microscope studies of spermatozoa of *Rhynchosciara* sp. *J. Cell Biol.,* **54:**598-608.

Shehata, T. E., Kempner, E. S. (1978). Sequential changes in cell volume distribution during vitamin B12 starvation of Euglena gracilis. *J Bacteriol.,* **133:** 396-398.

Sheveleva, E. V., Giordani, N. V. and Hallick, R. B. (2002). Identification and comparative analysis of the chloroplast α-subunit gene of DNA-dependent RNA polymerase from seven Euglena species. *Nucleic Acids Research,* **30(5):** 1247–1254.

Shi, H., Djikeng, A., Mark, T., Wirtz, E., Tschudi, C., Ullu, E. (2000). Genetic interference in *Trypanosoma brucei* by heritable and inducible double-stranded RNA. *RNA,* **6:**1069–76.

Shi, H., Djikeng, A., Tschudi, C., Ullu, E. (2004). Argonaute protein in the early divergent eukaryote Trypanosoma brucei: control of small interfering RNA accumulation and retroposon transcript abundance. *Mol Cell Biol.,* **24:** 420–427.

Shi, H., Tschudi, C., and Ullu, E. (2006) Functional replacement of *Trypanosoma brucei* Argonaute by the human slicer Argonaute2. *RNA,* **12:** 943–947.

Shi, H., Tschudi, C., and Ullu, E. (2006). An unusual Dicer-like1 protein fuels the RNA interference pathway in *Trypanosoma brucei*. *RNA*, **12:** 2063–2072.

Shi, J., Franklin, J. B., Yelinek, J. T.,Ebersberger, I., Warren, G., He, C. Y. (2008). Centrin4 coordinates cell and nuclear division in *T. brucei*. *J. Cell Sci.,* **121**:3062–3070.

Shigeo Takenaka, Junko Inagaki, Shingo Tsuyama, Kazutaka Miyatake and Yoshihisa Nakano (1995). Cell cycle dependent ADP-ribosylation of a unicellular organism, *Euglena gracilis Z. Comp. Biochem. Physiol.,* **II1B(2):**277-282.

Shimojima, M. (2011). Biosynthesis and functions of the plant sulfolipid. *Prog Lipid Res.,* **50(3):**234-9. doi: 10.1016/j.plipres.2011.02.003.

Shin, W. G., Brosnan, S., Triemer, R. E. (2002). Are cytoplasmic pockets (MTR/ pocket) present in all photosynthetic euglenoid genera? *J Phycol.,* **38:**790–799.

Shin, W., Boo, S. M., and Triemer, R. E. (2001). Ultrastructure of the basal body complex and putative vestigial feeding apparatus in Phacus pleuronectes (Euglenophyceae). *J. Phycol.,* **37:**913 - 921.

Shin, Y. A. and Eichhorn, G. L. (1968a). Interaction of metal ions with polynucleotides and related compounds. *Biochemistry, N. Y.* **7:** 3003-3006.

Shin, Y. A. and Eichhorn, G. L. (1986). Reversible unwinding and rewinding of deoxyribonucleic acid by zinc. II. Ions through temperature manipulation. *Biochemistry, N.Y.* **7:** 1026-1032.

Shine, J., and Dalgarno, L. (1974). The 3′-Terminal Sequence of *Escherichia coli* 16S Ribosomal RNA: Complementarity to Nonsense Triplets and Ribosome Binding Sites. *Pm.* Natl. *Acad.* Sei. *U. S.* A. **71:**1342-1346.

Sieber, K. B., Bromleya, R. E., Hotopp, J. C. D. (2017). Lateral gene transfer between prokaryotes and eukaryotes. *Experimental Cell Research*, pii: S0014-4827(17)30054-X. http://dx.doi.org/10.1016/j.yexcr.2017.02.009.

Siegel, T. N., Hekstra, D. R., Wang, X., Dewell, S. and Cross, G. A. (2010). Genome-wide analysis of mRNA abundance in two life-cycle stages of *Trypanosoma brucei* and identification of splicing and polyadenylation sites. Nucleic Acids Res. **38**, 4946–4957.

Simpson, A. G. B. (1997). The identity and composition of the Euglenozoa. *Arch. Protistenk.,* **148:** 318-328..

Simpson, A. G. B. (2003). Cytoskeletal organization, phylogenetic affinities and systematics in the contentious taxon Excavata (Eukaryota). *Int J Syst Evol Microbiol.,* **53:**1759–1777. DOI: 10.1099/ijs0.02578-0.

Simpson, A. G. B., Gill, E. E., Callahan, H. A., Litaker, R. W., and Roger, A. J. (2004). “Early evolution within kinetoplastids (Euglenozoa), and the late emergence of trypanosomatids,” *Protist*, **155(4):** 407–422.

Simpson, A. G. B., Inagaki, Y. and Roger, A. J. (2006). Comprehensive multigene phylogenies of excavate protists reveal the evolutionary positions of “primitive” eukaryotes. *Mol. Biol. Evol.,* **23:**615–625.

Simpson, A. G. B., Stevens, J. R. and Lukes, J. (2006). The evolution and diversity of kinetoplastid flagellates. *Trends Parasitol*., **22:**168–174.

Simpson, J. T. and Durbin, R. (2011). Efficient de novo assembly of large genomes using compressed data structures. *Genome Research,* **22:**549–556.

Simpson, L., Sbicego, S., and Aphasizhev, R. (2003). Uridine insertion/deletion RNA editing in trypanosome mitochondria: a complex business. *RNA,* **9:**265–276.

Simpson, L., Thiemann, O. H. (1995). Sense from nonsense: RNA editing in mitochondria of kinetoplastid protozoa and slime molds. *Cell,* **81:**837–840.

Simpson, L., Thiemann, O. H., Savill, N. J., Alfonzo, J. D., Maslov, D. A. (2000). Evolution of RNA editing in trypanosome mitochondria. *Proc Natl Acad Sci USA,* **97:**6986–6993

Sims, D., Sudbery, I., Ilott, N. E., Heger, A. and Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics,* **15**:121–132. doi:10.1038/nrg3642

Sirlin, J. L. (1972). Biology of RNA, Academic Press, New York, **pp.** 158.

Sistrom, W. R., Griffiths, M., and Stanier, R. Y. .T. (1956). *The biology* of photosynthetic bacterium which lacks colored carotenoids. *Cellular Comp. Physiol.* 46: 459.

Sjoberg, B. M. (1997). Ribonucleotide reductases — a group of enzymes with different metallosites and a similar reaction mechanism. *Structure Bonding,* **88:**139 – 173.

Sjoberg, B. M., Reichard, P., Graslund, A. and Ehrenberg, A. (1978). Mechanism of Ribonucleoside Diphosphate Reductase from *Escherichia coli. J. Biol. Chem.,* **253:**6863–6865.

Sjostrand, F. S. and Hanzon, V. (1954). Ultrastructure of Golgi apparatus of exocrine cells of mouse pancreas. *Expl Cell Res.,* **7:**415-429.

Skippington, E., Barkman, T. J., Rice, D. W., Palmer, J. D. (2015). Miniaturized mitogenome of the parasitic plant Viscum scurruloideum is extremely divergent and dynamic and has lost all nad genes. *Proc Natl Acad Sci U S A.,* **112:**E3515–E3524.

Slater, J. P., Mildvan, A. S. and Loeb, L. A. (1971). Zinc in DNA polymerases. *Biochem. biophys. Res. Commun.,* **44:** 37-43

Slavikova, S., Vacula, R., Fang, Z., Ehara, T., Osafune, T. and Schwartzbach, S. D. (2005). Homologous and heterologous reconstitution of Golgi to chloroplast transport and protein import into the complex chloroplasts of Euglena. *J. Cell Sci.,* **118:**1651 – 1661.

Smallman, D. S., Schnare, M. N., Gray, M. W. (1996). RNA:RNA interactions in the large subunit ribosomal RNA of *Euglena gracilis. Biochimica et Biophysica Acta,* **1305:** 1-6.

Smarda, P., Bures, P., Horová, L., Leitch, I. J., Mucina, L., Pacini, E., Tichý, L., Grulich, V., and Rotreklová, O. (2014). Ecological and evolutionary significance of genomic GC content diversity in monocots. PNAS, E4096–E4102.

Smillie, R. M., Graham, D., Dwyer, M. R., Grieve, A. and Tobin, N. F. (1967). Evidence for the synthesis in vivo of proteins of the Calvin cycle and of the photosynthetic electron-transfer pathway on chloroplast ribosomes. *Biochemical and Biophysical Research Communications*, **28:** 604-610.

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

Smith, W., Galland, P., Senger, H., Furuya, M. (1990). Microspectrophotometry of *Euglena gracilis. Planta,* **192:** 375 – 381.

Smith-Sonneborn, J. (1971). Age correlated sensitivity to ultraviolet radiation in Paramecium. *Rad. Res.,* **46:** 64–69.

Smith-Sonneborn, J. (1979). DNA repair and longevity assurance in *Paramecium tetraurelia*. *Science,* **203:** 1115–1117.

Smith-Sonneborn, J. (2005). The Myth and reality of reversal of aging by hormesis. *Ann. N. Y. Acad. Sci.,* **1057:** 165–176.

Smith-Unna, R. D., Boursnell, C., Patro, R., Hibberd, J. M., and Kelly, S. (2014). Transrate v1.0.0.beta2 (2014). http://hibberdlab.com/transrate. DOI: 10.5281/zenodo14897.

Smith-Unna, R. D., Boursnell, C., Patro, R., Hibberd, J. M., and Kelly, S. (2016). TransRate: reference free quality assessment of *de-novo* transcriptome assemblies. *Genome Res.,* **26(8):**1134-44. doi: 10.1101/gr.196469.115.

Smyth, R. D., Saranak, J., Foster, K. W. (1988). Algal visual systems and their photoreceptor pigments. W. Round, I. Chapman (Eds.), *Progress in Psychological Research, Biopress Ltd, New York,* **pp.** 255–286

Snowden, T., Acharya, S., Butz, C., Berardini, M., Fishel, R. (2004). hMSH4-hMSH5 recognizes Holliday junctions and forms a meiosis-specific sliding clamp that embraces homologous chromosomes. *Mol Cell.,***15:** 437–451.

Sogin, M. L., Elwood, H. J., Gunderson, J. H. (1986). Evolutionary diversity of eukaryotic small-subunit rRNA genes. *Proc Natl Acad Sci USA,* **83:**1383–1387.

Sogin, M. L., Gunderson, J. H., Elwood, H. J., Alonson, R. A., Peattie, D. A. (1989). Phylogenetic meaning of the kingdom concept: an unusual ribosomal RNA from *Giardia lamblia*. *Science*, **243:**75–77.

Sollner-Webb, B., Mougey, E. B. (1991). News from the nucleolus: rRNA gene expression. *Trends Biochem. Sci.,* **16:**58-62.

Sollner-Webb, B., Tycowski, T., Steitz, J. A (1995). In: R.A. Zimmermann, A.E. Dahlberg (Eds.), Ribosomal RNA: Structure, Evolution, Processing, and Function in Protein Biosynthesis, CRC Press, Boca Raton, FL, **pp.** 469-490.

Solomon, J. A. (1986). Euglenoid flagellar root systems: ultrastructure and phylogenetic significance. Ph. D. Thesis University of Tennessee, Knoxville

Solomon, J. A., Walne, P. L., Kivic, P. A. (1987). Entosiphon sulcatum (Euglenophyceae): flagellar roots of the basal body complex and reservoir regions). *J Phycol.,* **23:**85–98.

Sommer, D. D., Delcher, A. L., Salzberg, S. L., Pop, M. (2007). Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics,* **26(8):**64.

Sonenberg, N., and Hinnebusch, A. (2009). Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell, **136*:**731-745.

Sonenberg, N., Hershey, J. W. B., and Mathews, M. B., (Eds.) (2000). *Translational control of gene expression*, Cold Spring Harbor Laboratory press, Cold Spring Harbor, NewYork.

Song, P. S., Walker, E. B. (1981). M. Levandonsky, S.N. Hutner (Eds.), Molecular aspects of photoreceptors in protozoa and other microorganisms, *Biochemistry and Physiology of Protozoa, 4, Academic Press, New York,* **pp.** 191–220

Sonneborn, T. M. (1954). The relation of autogamy to senescence and rejuvenescence in Paramecium aurelia*. J. Protozool.* **1:** 38–53.

Soucy, S. M., Huang, J. and Gogarten, J. P. (2015).Horizontal gene transfer: building the web of life. *NATURE REVIEWS* **|** *GENETICS,* **16:** 472 – 482.

Souza, R. T., Lima, F. M., Barros, R. M., Cortez, D. R., Santos, M. F., Cordero, E. M., Ruiz, J. C., Goldenberg, S., Teixeira, M. M. G., da Silveira, J. F. (2011). Genome Size, Karyotype Polymorphism and Chromosomal Evolution in

*Trypanosoma cruzi*. *PLoS ONE***6(8):** e23042. doi:10.1371/journal.pone.0023042

Spãth, G.F., Epstein, L., Leader, B., Singer, S.M., Avila, H.A., Turco, S.J., and Beverley, S. M. (2000). Lipophosphoglycan is a virulence factor distinct from related glycoconjugates in the protozoan parasite Leishmania major. *Proc. Natl. Acad. Sci. USA,* **97:**9258–9263.

Spencer J. Greenwood, Michael W. Gray (1998). Processing of precursor rRNA in *Euglena gracilis*: identification of intermediates in the pathway to a highly fragmented large subunit rRNA. Biochimica et Biophysica Acta, **1443:** 128 – 138.

Spencer, D. F. and Gray, M. W. (2011). Ribosomal RNA genes in Euglena gracilis mitochondrial DNA: fragmented genes in a seemingly fragmented genome. *Mol. Genet. Genomics.,* **285:**19–31.

Spencer, D. F., Gray, M. W. (2012). Ribosomal RNA genes in Euglena gracilis mitochondrial DNA: fragmented genes in a seemingly fragmented genome. *Mol Genet Genomics,* **285:**19–31.

Spencer, D. F., Gray, M. W. (2012). Ribosomal RNA genes in Euglena gracilis mitochondrial DNA: fragmented genes in a seemingly fragmented genome. *Mol Genet Genomics,* **285:**19–31.

Spitzer, M., Lorkowski, S., Cullen, P., Sczyrba, A., and Fuellen, G. (2006). IsoSVM-distinguishing isoforms and paralogs on the protein level. *BMC Bioinformatics,* **7:**110.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics,* **30 (9):** 1312-1313*. doi:10.1093/bioinformatics/btu033*.

Stanke Mario and Waack Stephan (2003). Gene Prediction with a Hidden-Markov Model and a new Intron Submodel. *Bioinformatics*, **19(2):** ii215 - ii225.

Stanke, M., Diekhans, M., Baertsch, R., Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, **24(5):**637-44. doi: 10.1093/bioinformatics/btn013.

Stankiewicz, A. J., Falchuk, K. H., and Vallee, B. L. (1983). Composition and Structure of Zinc-Deficient *Euglena gracilis* Chromatin. *Biochemistry, **22:***5150 – 5156.

Steffen Reinbothe and Benno Parthier (1990). Translational regulation of plastid gene expression in *Euglena gracilis.* FEBS 08468, 265(1,2, 7-11).

Stein, L. D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J. E., Harris, T. W., Arva, A., and Lewis, S. (2002). The Generic Genome Browser: A Building Block for a Model. *Genome Res.,* **12:**1599 - 1610 doi:10.1101/gr.403602.

Steinway, S. N., Dannenfelser, R., Laucius, C. D., Hayes, J. E. and Nayak, S. (2010). JCoDA: a tool for detecting evolutionary selection. *BMC Bioinformatics,* **11:**284 https://doi.org/10.1186/1471-2105-11-284

Steitz, T. A. and Steitz, J. A. (1993) A general two-metal-ion mechanism for catalytic RNA. *Proc. Natl. Acad. Sci*. USA, **90:**6498–6502.

Stenmark, H. (2009). Rab GTPases as coordinators of vesicle traffic. *Nature Reviews. Molecular Cell Biology,* **10(8):**513-525. Doi: 10.1038/nrm2728.

Stevens, J. R. (2014). Free-living bodonids and derived parasitic trypanosomatids: butwhat lies in between? *Trends Parasitol*., **30:**113–4.

Steward, C. and Burke, B. (1987). Teratocarcinoma stem cells and early mouse embryos contain only a single major lamin polypeptide closely resembling lamin B. *Cell*, **51:**383-392.

Stewart, E. J., Madden, R., Paul, G., Taddei, F. (2005). Aging and death in an organism that reproduces by morphologically symmetric division. *PLoS Biol.,* 3: e45.

Stick, R. (1985). Hausen P., Changes in the nuclear lamina composition during early development of *Xenopus laevis. Cell*, **41:**191 - 200.

Stick, R. (1992). The gene structure of *Xenopus* nuclear lamin A: A model for the evolution of A-type from B-type lamins by exon shuffling. *Chromasoma,* **101:**566 - 574.

Stoco, P. H., Aresi, C., Lückemeyer, D. D., Sperandio, M. M., Sincero, T. C. M., Steindel, M., Miletti, L. C., Grisard, E. C. (2012). *Trypanosoma rangeli* expresses a b-galactofuranosyl transferase. *Experimental Parasitology,* **130:**246–252.

Stoco, P. H., Aresi, C., Lückemeyer, D. D., Sperandio, M. M., Sincero, T. C. M., Steindel, M., Miletti, L. C., and Grisard, E. C. (2012). *Trypanosoma rangeli* expresses a b-galactofuranosyl transferase. *Experimental Parasitology,* **130:**246–252.

Stoebe, B., Maier, U. G. (2002). One, two, three: nature's toolbox for building plastids. *Protoplasma*, **219:** 123–130.

Stoltzfus, A. (1999). On the possibility of constructive neutral evolution. *J. Mol. Evol.,* **49:**169–181.

Stöver, B. C. and Müller, K. F.(2010). TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinformatics,* **11:**7. http://www.biomedcentral.com/1471-2105/11/7.

Streb, C., Richter, P., Ntefidou, M., Lebert, M., Hader, D-P. (2002). Sensory transduction of gravitaxis in *Euglena gracilis*. *J. Plant Physiol.***159:**855–862.

Streb, C., Richter, P., NteWdou, M., Lebert, M., Häder, D-P. (2002). Sensory transduction of gravitaxis in *Euglena gracilis*. *J. Plant Physiol.,* **159:**855–862.

Strother, G. K., Wolken, J. J. (1961). *In vivo* absorption spectra of *Euglena*: chloroplast and eyespot. *J. Protozool.,* **8:** 261–265

Stuart, K., Panigrahi, A. K. (2002). RNA editing: complexity and complications. *Mol Microbiol*., **45:**591–596.

Stubbe, J. (2003). Di-iron-tyrosyl radical ribonucleotide reductases. *Curr. Opin. Chem. Biol.,* **7:**183 – 188.

Sturm, N.R. and Simpson, L. (1990). Kinetoplast DNA minicircles encode guide RNAs forediting of cytochrome oxidase subunit III mRNA. *Cell,* **61:**879–884.

Sturm, N.R., and Simpson, L. (1990). Kinetoplast DNA minicircles encode guide RNAs for editing of cytochrome oxidase subunit III mRNA. *Cell,* **61:**879–884.

Subota, I., Rotureau, B., Blisnick, T., Ngwabyt, S., Durand-Dubief, M., Engstler, M., and Bastin, P. (2011). ALBA proteins are stage regulated during trypanosome development in the tsetse fly and participate in differentiation. *Mol. Biol. Cell,* **22:** 4205–4219.

Sukumaran, J. and Holder, M. T. (2015). The DendroPy Phylogenetic Computing Library Documentation. Retrieved June 04, 2015, from http://dendropy.org/.

Sukumaran, J. and Holder, M. T.. (2010). DendroPy: A Python library for phylogenetic computing. *Bioinformatics* 26: 1569-1571.

Sul, SJ and Williams, T. L. (2009). "An Experimental Analysis of Consensus Tree Algorithms for Large-Scale Tree Collections," 5th Intl. Symposium on Bioinformatics Research and Applications (ISBRA'09), **pp.** 100-111.

Sulli, C., Fang, Z., Muchhal, U. and Schwartzbach, S. D. (1999). Topology of Euglena chloroplast protein precursors within endoplasmic reticulum to Golgi to chloroplast transport vesicles*. J. Biol. Chem.,* **274:**457–463.

Sumida, S., Lyman, H., Kiyohara, N. and Osafune, T. (2007). Mechanism of Conversion from Heterotrophy to Autotrophy in *Euglena gracilis*. *Cytologia,* **72(4):** 447–457.

Surek, B., Melkonian, M. (1986). A cryptic cytostome is present in *Euglena. Protoplasma*, **133:** 39–59

Surek, B., Melkonian, M. (1986). A cryptic cytostome is present in Euglena. *Protoplasma,* **133:**39 - 49.

Suzaki, T. and Williamson, R. E. (1986). Cell surface displacement during eugienoid movement and its computer simulation. *Cell Motil. Cytoskeleton,* **6:**186-192.

Suzaki, T., Williamson, R. E. (1985). Euglenoid movement in *Euglena fusca*: evidence for sliding between pellicular strips. *Protoplasma,* **124:**137–146.

Suzaki, T., Williamson, R. E. (1986). Pellicular ultrastructure and euglenoid movements in Euglena ehrenbergii Klebs and Euglena oxyuris Schmarda. *J. Protozool.,* **33:**165 – 171.

Suzaki, T., Williamson, R. E. (1986). Reaction of Euglenoid movements and flagellar beating in detergent extracted cells of *Astasia longa*: Different mechanisms of force generation are involved. *J. Cell Biol.,* **80:** 75–89

Suzaki, T., Williamson, R. E. (1986). Ultrastructure and sliding of pellicular structures during euglenoid movement in Astasia longa Pringsheim (Sarcomastigophora, Euglenida). *J. Protozool.,* **33:**179–184.

Suzuki, H., Ito, Y., Yamazaki, Y., Katsuhiko, M, Masami Uji, Kazuhiro Abe, Kazutoshi Tani, Yoshinori Fujiyoshi and Sachiko Tsukita (2013). The four-transmembrane protein IP39 of Euglena forms strands by a trimeric unit repeat. *Nature Communications*, **4:**1766 DOI: 10.1038/ncomms2731 | www.nature.com/naturecommunications.

Swithers, K. S., Soucy, S. M. and Gogarten, J. P. (2012). The role of reticulate evolution in creating innovation and complexity. *Int. J. Evol. Biol.,* **2012:**418964.

Szathmary, E. and Smith, J. M. (1995). The major evolutionary transitions. *Nature*, **374(6519):**227–32. doi: 10. 1038/374227a0 PMID: 7885442.

Szmant, H. H. (1975). Physical properties of dimethyl sulfoxide and its function in biological systems. *Ann. N.Y. Acad. Sci.,* **243:** 20 - 23.

Szymański, M., Deniziak, M., and Barciszewski, J. (2000). The new aspects of aminoacyltRNAsynthetases. *Acta Biochim. Polonica,* **47:**821-834.

T. Osafune, J.A. Schiff (1980). Stigma and flagellar swelling in relation to light and carotenoids in *Euglena gracilis* var. *bacillaris. J. Ultrastruct. Res.,* **73:**336–349

Tahedl, H., Richter, P., Lebert, M., Häder DP. (1998). cAMP is involved in gravitaxis signal transduction of *Euglena gracilis*. *Microgravity Sci. Technol.,* **11:**173 – 178.

Taillon, B. E., Adler, S. A., Suhan, J. P., Jarvik, J. W. (1992). Mutational analysis of centrin: an EF-hand protein associated with three distinct contractile fibers in the basal body apparatus of *Chlamydomonas*. *J. Cell Biol.,* **119**: 1613–1624.

Takahashi, K., Tanaka, A., Hara, M., and Nakanishi, S.(1992). The primary structure and gene organization of human substance P and neuromedin K receptors. *Eur. J. Biochem.,* **204(3):**1025–1033.

Takeda, T., Nakano, Y., Takahashi, M., Konno, N., Sakamoto, Y.,Arashida, R., Marukawa, Y., Yoshida, E., Ishikawa, T., Suzuki,K. (2015). Identification and enzymatic characterization of anendo-1,3-b-glucanase from *Euglena gracilis*. *Phytochemistry,* **116:**21–27.

Takeda, Y., Chou, K.B., Takeda, J., Sachais, B.S., and Krause, J. E.(1991). Molecular cloning, structural organization and functional expression of the human substance P receptor. *Biochem. Biophys. Res. Commun.,* **179(3):**1232–1240.

Takenaka, M. (2014). How Complex Are the Editosomes in Plant Organelles? *Molecular Plant,* 7(4): 582–585.

Takenaka, M., Zehrmann, A., Härtel, B., Kugelmann, M., Verbitskiy, D., and Brennicke, A. (2012). MORF family proteins are requiredfor RNA editing in mitochondria and plastids of plants. *Proc.Natl Acad. Sci. U S A.,* **109:**5104–5109.

Talbert, P. B. and Henikoff, S. (2013). Phylogeny as the basis for naming histones. *Trends in Genetics,* **29(9):**499 – 500.

Talbert, P. B., Ahmad, K., Almouzni, G., Ausió, J., Berger, F., Bhalla, P. L., Bonner, W. M, Cande, W. Z., Chadwick, B. P., Chan, S. W. L., Cross, G. A. M., Cui, L., Dimitrov, S. I., Doenecke, D., Eirin-López, J. M., Gorovsky, M. A., Hake, S. B., Hamkalo, B. A., Holec, S., Jacobsen, S. E., Kamieniarz, K., Khochbin, S., Ladurner, A. G., Landsman, D., Latham, J. A., Loppin, B., Malik, H. S., Marzluff, W. F., Pehrson, J. R., Jan, P., Schneider, R., Singh, M. B., Smith, M. M., Thompson, E., Torres-Padilla, M., Tremethick, D. J., Turner, B. M., Waterborg, J. H., Wollmann, H., Yelagandula, R., Zhu,B. and Henikoff, S.(2012). A unified phylogeny-based nomenclature for histone variants. *Epigenetics & Chromatin,* **5(7):**1-19.

Talen, J. L., Sanders, J. P. M., Flavell, R. A. (1974). Genetic complexity of mitochondrial DNA from Euglena gracilis. *Biochim Biophys Acta,* **374:**129–135.

Talen, J. L., Sanders, J. P., Flavell, R. A. (1974). Genetic complexity of mitochondrial DNA from *Euglena gracilis*. *Biochim Biophys Acta*, **374:**129—135.

Talke, S., Preisfeld, A. (2002). Molecular evolution of euglenozoan paraxonemal rod genes par1 and par2 coincides with phylogenetic reconstruction based on small subunit rDNA data. *J. Phycol*., **38:**995–1003.

Tamura, K. and Hara-Nishimura, I. (2013). The molecular architecture of the plant nuclear pore complex. *Journal of Experimental Botany,* **64(4):**823–32.

Tamura, K., Fukao, Y., Iwamoto, M., Haraguchi, T., Hara-Nishimura, I. (2010). Identification and characterization of nuclear pore complex components in Arabidopsis thaliana. *The Plant cell.,* **22(12):**4084–97.

Tamura, K.,Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Mol Biol Evol.,* **30(12):** 2725–2729.

Tanaka, Y., Ogawa, T., Maruta, T., Yoshida, Y., Arakawa, K., Ishikawa, T. (2017). Glucan synthase-like 2 is indispensable for paramylon synthesis in *Euglena gracilis*. *FEBS Lett.,* **591(10):**1360-1370. doi: 10.1002/1873-3468.12659.

Tang, S., Lomsadze, A. and Borodovsky, M. (2015). Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res.,* **248:**1–18.

Tang, S., Lomsadze, A., Borodovsky, M. (2015). Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res*., **43(12):**e78. doi: 10.1093/nar/gkv227

Tang, Y., Wen, X., Lu, C. (2005). Differential changes in degradation of chlorophyll-protein complexes of photosystem I and photosystem II during flag leaf senescence of rice. *Plant Physiology and Biochemistry,* **43:** 193–201.

Tannreuther, G. W. (1923). Nutrition and Reproduction in Euglena. *Archiv für Entwicklungsmechanik der Organismen,* **52 (3-4):** 367 - 383. https://doi.org/10.1007/BF02093633

Tanuma S. and Kanai Y. (1982). Poly (ADP-ribos)ylation of chromosomal proteins in the HeLa S3 cell cycle. *J. biol. Chem.,* **257:** 6565 – 6570.

Tanuma, S. (1989). Evidence for a novel metabolic pathway of (ADP-ribose): pyrophosphorolysis of ADP-ribose in HeLa S3 cell nuclei. *Biochem. biophys. Res. Commun.,* **163:**1047 - 1055.

Tanuma, S. and Johnson, G. S. (1983). ADP-ribosylation of nonhistone high mobility group proteins in intact cells. *J. biol. Chem.,* **258:**4067 - 4070.

Teerawanichpan, P., Qiu, X. (2010). Fatty acyl-CoA reductase and wax synthase from *Euglena gracilis* in the biosynthesis of medium-chain wax esters. *Lipids,* **45**:263–273.

Tefsen, B., Ram, A. F. J., van Die, I., and Routier, F. H. (2012). Galactofuranose in eukaryotes: aspects of biosynthesis and functional impact.*Glycobiology,* **22(4):**456–469 doi:10.1093/glycob/cwr144.

Tefsen, B., Ram, A. F. J., van Die, I., and Routier, F. H.(2012). Galactofuranose in eukaryotes: aspects of biosynthesis and functional impact. *Glycobiology,* **22(4):** 456–469. doi:10.1093/glycob/cwr144

Teich, R., Zauner, S., Baurain, D., Brinkmann, H., Petersen, J. (2007). Origin and distribution of Calvin cycle fructose and sedoheptulose bisphosphatases in plantae and complex algae: a single secondary origin of complex red plastids and subsequent propagation via tertiary endosymbioses. *Protist,* **158(3):**263-276.

Teixeira, S. M. R., Russell, D. G., Kirchhoff, L. V., and Donelson, J. E. (1994). A Differentially Expressed Gene Family Encoding "Amastin", a Surface Protein of *Dypanosoma cruxi* Amastigotes. The Journal of Biological Chemistry, **269(32):**20509-20516.

Tekchand SAIDHA, Song-Qing NA, Jiayang LI and Jerome A. Schiff (1988). A sulphate metabolizing centre in *Euglena* mitochondria. *Biochem. J.* **253:** 533 – 539

Telewski, F. W. (2006). A unified hypothesis of mechanoperception in plants. American *Journal of Botany,* **93:**1466-1476.

Tessier, L. H., van der Speck H., Gualberto. J. M., Grienenberger, J. M. (1997). The cox1 gene from Euglena gracilis: a protist mitochondrial gene without introns and genetic code modifications. *Curr. Genet.,* **31:** 208–213.

Thatcher, T. H.and Gorovsky, M. A. (1994). Phylogenetic analysis of the core histones H2A, H2B, H3, and H4. *Nucleic Acids Research*, **22(2):**2174-179.

The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana. Nature,* **408:**796-815doi:10.1038/35048692.

The *C. elegans* Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. The sequencing and annotation of the *Caenorhabditis elegans* genome. *Science,* **282:**2012–2018.

The Gene Ontology Consortium, Ashburner,M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nat Genet.,* **25(1):** 25–29. doi:10.1038/75556.

Thomas, H., Ougham, H. J., Wagstaff, C., Stead, A. D. (2003). Defining senescence and death. *Journal of Experimental Botany*, **54:** 1127 – 1132.

Thompson, M. D., Copertino, D. W., Thompson, E., Favreau, M. R. and Hallick, R. B. (1995). Evidence for the late origin of introns in chloroplast genes from an evolutionary analysis of the genus *Euglena. Nucleic Acids Research,* **23(23):**4745 – 4752.

Threlfall, D. R. and Goodwin, T. W. (1967). Nature, Intracellular Distribution and Formation Of Terpenoid Quinones in *Euglena Gracilis*. *Biochem. J.* **103:** 573.

Tiller, N. and Bock, R. (2014). The Translational Apparatus of Plastids and Its Role in Plant Development. *Molecular Plant,* **7(7):**1105–1120.

Tilney, L. G. (1971). Origin and continuity of microtubules. In *Results and Problems in Cell*

Timmis, J. N., Ayliffe, M. A., Huang, C. Y., Martin, W. (2004). Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat. Rev., Genet.,* **5:**123–135.

Toda, H., Yazawa, M., Yagi, K. (1992) Amino acid sequence of calmodulin from *Euglena gracilis*. *Eur, J. Biochem.,* **205:**653–660.

Tomoko Ehara, Ikuko Shihira-Ishikawa, Tetsuaki Osafune, Eiji Hase And Isamu Ohkuro (1975) . Some Structural Characteristics of Chloroplast Degeneration in Cells of *Euglena gracilis* Z during Their Heterotrophic Growth in Darkness. *Journal of Electron Microscopy,* **24(4):**253 – 261.

Tong, I., Edmunds, L. N. Jr. (1993). Role of cyclic GMP in the mediation of circadian rhythmicity of the adenylate cyclase-cyclic AMP phosphodiesterase system in Euglena. *Biochem Pharmacol.,* **45:**2087–2091

Torrents, E., Aloy, P., Gibert, I., and Rodriguez-Trelles, F. (2002). Ribonucleotide reductases: divergent evolution of an ancient enzyme. *J. Mol. Evol.,* **55:**138–152.

Torrents, E., Eliasson, R., Wolpher, H., Graslund, A. and Reichard, P. (2001). *J. Biol. Chem.,* **276:**33488–33494.

Torrents, Eduard, Trevisiol Chiara, Rotte Carmen, Hellman Ulf, Martin William and Reichard Peter (2005). DNA: Replication, Repair, And Recombination: *Euglena gracilis* Ribonucleotide Reductase: The Eukaryote Class II Enzyme and the Possible Antiquity of Eukaryote B12 Dependence. *J. Biol. Chem.,* **281:**5604-5611.

Torruella, G., et al. (2012). Phylogenetic relationships within the Opisthokonta based on phylogenomic analyses of conserved single-copy protein domains. *Mol. Biol. Evol.,* **29(2):** 531–544.

Toyota, M., and Gilroy, S. (2013). Gravitropism and mechanical signaling in plants. *American Journal of Botany,* **100:**111-125.

Trapnell, C. *et al* (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols,* **7:**562–578.

Trapnell, C. Williams, B. A, Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. and Pachter, L. (2010). Transcript assembly

and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.,* **28:**511 – 515.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R, Pimentel, H., Salzberg, S. L., Rinn, J. L. and Pachter,. L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols,* **7(3):** 562 – 578.

Treangen, T. J. and Rocha, E. P. C. (2011). Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet.,* **7:**e1001284.

Triemer, R. E. and Farmer, M. A. (1991). An ultrastructural comparison of the mitotic apparatus, feeding apparatus, flagellar apparatus and cytoskeleton in euglenoids and kinetoplastids. *Protoplasma,* **164:**91–104.

Triemer, R. E., and Ott D. W. (1990). Ultrastructure of Diplonema ambulator Larsen & Patterson (Euglenozoa) and its relationship to Isonema. *Eur. J. Protistol.* **25:**316-320.

Trotter, J.R., Ernst, N.L., Carnes, J., Panicucci, B., Stuart, K.(2005). A deletion site editing endonuclease in *Trypanosoma brucei*. *Mol. Cell,* **20:**403–412.

*Trypanosoma brucei*. *J. Am. Soc. Mass Spectrom,* **25:**1763-1770 DOI: 10.1007/s13361-014-0960-6.

Tsaousis, A. D., Kunji, E. R., Goldberg, A. V., Lucocq, J. M., Hirt, R. P., Embley, T. M. (2008). A novel route for ATP acquisition by the remnant mitochondria of *Encephalitozoon cuniculi*. *Nature,* **453:**553 – 556.

Tsiokas, L., Kim, S. and Ong, E. C. (2007). Cell biology of polycystin-2. *Cell Signal.,* **19:**444-453.

Tucci, S., Vacula, R., Krajcovic, J., Proksch, P., Martin, W. (2010). Variability of wax ester fermentation in natural and bleached *Euglena gracilis* strains in response to oxygen and the elongase inhibitor flufenacet. *J. Eukaryot. Microbiol.,* **57**:63 – 69.

Tucker, J. B. (1972). Microtubule-arms and propulsion of food particles inside a large feeding organelle in the ciliate *Phascolodon vorticella. J. Cell Set.,* **10:**883-903.

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. *Mol Biol Evol.,* **26:** 631–48.

Turnock, D. C., Izquierdo, L., Ferguson, M. A. (2007). The de novo synthesis of GDP-fucose is essential for flagellar adhesion and cell growth in *Trypanosoma brucei. J. Biol. Chem.,* **282:**28853−28863.

Tzagoloff, A. and Myers, A. M. (1986). Genetics of mitochondrial biogenesis. *Annu. Rev. Biochem.,***55:**249–285.

Ueda, K. and Hayaishi O. (1985). ADP-ribosylation. *A. Rev. Biochem.* **54:**73 100.

Ueda, K., (1958). Structure of plant cells with special reference to lower plants. III. A cytological study of *Euglena gracilis*. *Cytologia (Tokyo)* **23:** 56-67

Ueda, K., Hayashi-Ishimaru, Y. (1996). Localization of DNA in the condensed interphase chromosomes of *Euglena. Chromosoma.,* **104:**380 – 385.

Uliel, S., Liang, X. H., Unger, R., Michaeli, S. (2004). Small nucleolar RNAs that guide modification in trypanosomatids: repertoire, targets, genome organisation, and unique functions. *Int. J. Parasitol.,* **34(4):**445–454.

Ulvskov, P., Paiva, D. S., Domozych, D., Harholt, J. (2013). Classification, Naming and Evolutionary History of Glycosyltransferases from Sequenced Green and Red Algal Genomes. *PLoS ONE,* **8(10):** e76511. doi:10.1371/journal.pone.0076511.

Umaer, K., and Williams, N. (2015). Kinetoplastid Specific RNA-Protein Interactions in *Trypanosoma cruzi* Ribosome Biogenesis.

Umaer, K., Ciganda, M. and Williams, N. (2014). Ribosome Biogenesis in African Trypanosomes Requires Conserved and Trypanosome-Specific Factors. *Eukaryotic Cell,* **13(6):** 727–737.

Urbaniak, M. D., Guther, M. L. S., and Ferguson, M. A. J. (2012). Comparative SILAC Proteomic Analysis of *Trypanosoma brucei* Bloodstream and Procyclic Lifecycle Stages. *PLoS ONE,* **7(5):** e36619. doi:10.1371/journal.pone.0036619.

Vacula, R., Steiner, J. M., Krajcovic, J., Ebringer, L. and Loffelhardt, W. (2001). Plastid state- and light-dependent regulation of the expression of nucleus-encoded genes for chloroplast proteins in the flagellate *Euglena gracilis*. *Folia Microbiol.,* **46:**433–441.

Vallee, B. L., and Falchuk, K. H. (1981). Zinc and gene expression. *Philosophical Transactions of the Royal Society, London B.* **1(294):**185–196.

van Besouw, A. and Wintermans, J. F. (1978). Galactolipid formation in chloroplast envelopes. I. Evidence for two mechanisms in galactosylation. *Biochim Biophys Acta.,* **529(1):**44-53.

van Dam, T. J. P., Townsend, M. J., Turk, M., Schlessinger, A., Sali, A., Field, M. C., and Huynen, M. A. (2013). Evolution of modular intraflagellar transport from a coatomer-like progenitor. *PNAS,* **110 (17):** 6943–6948

van der Horst, M.A. and Hellingwerf, K. (2004) Photoreceptor proteins, 'star actors of modern times': a review of the functional dynamics in the structure of representative members of six different photoreceptor families. *Acc. Chem. Res.*, **37:** 13–20.

van Dooren, G. G., Schwartzbach, S. D., Osafune, T., and McFadden, G. I. (2001). Translocation of proteins across the multiple membranes of complex plastids. *Biochim. Biophys. Acta* **1541**, 34–53.

Van Haastert, P. J. M. (1985). cAMP activates adenylate and guanylate cyclase of Dictyostelium discoideum cells by binding to different classes of cell-

surface receptors. A study with extracellular Ca2?. *Biochim Biophys Acta,* **846:**324–333

van Leeuwen, F., Taylor, M. C., Mondragon, A., Moreau, H., Gibson, W., Kieft, R. and Borst, P. (1998). *b*-d-glucosylhydroxymethyluracil is a conserved DNA modi®cation in kinetoplastid protozoans and is abundant in their telomeres. *Proc. Natl Acad. Sci. USA,* **95:**2366–2371.

Van Leeuwen, F., Wijsman, E. R., Kieft, R., van der Marel, G. A., Van Boom, J. H. and Borst, P. (1997). Localization of the modified base. J in telomeric VSG gene expression sites of Trypanosoma brucei. *Genes Dev.,* **11:**3232–3241.

Van Leeuwen, F., Wijsman, E. R., Kuyl-Yeheskiely, E., van der Marel, G. A., Van Boom, J. H. and Borst, P. (1996). The telomeric GGGTTA repeats of Trypanosoma brucei contain the hypermodified base J in both strands. *Nucleic Acids Res.,* **24:**2476–2482.

Van Loy, T., Vandersmissen, H.P., Poels, J., Van Hiel, M.B., Verlinden, H., Vanden Broeck, J. (2010). Tachykinin-related peptides and their receptors in invertebrates: a current view. *Peptides,* **31(3):**520–524.

Vannini, G. L. and Poli, F. (1983). Binucleation and Abnormal Chromosome Distribution in *Euglena gracilis* Cells Treated with Dimethyl Sulfoxide. *Protoplasma,* **114:**62 – 66.

Varmat, . N. S., Abrahama., and Hanseni,. A. (1961). Accumulation of 58C0 vitamin B12 by *Euglena gracilis. Journal of Protozoology,* **8:** 212 - 216.

Vasconcelos, A., Pollack, M., Mendiola, L. R., Hoffmann, H. P., Brown, D. H., Price, C. A. (1971). Isolation of Intact Chloroplasts from *Euglena gracilis* by Zonal Centrifugation. *Plant Physiol.,* **47:** 217-221

Vasquez, J. J., Hon, C. C., Vanselow, J. T., Schlosser, A., Siegel, T. N. (2014).Comparative ribosome profiling reveals extensive translational complexityin different *Trypanosoma brucei* life cycle stages. *Nucleic Acids Res.,* **42:**3623–3637. http://dx.doi.org/10.1093/nar/gkt1386.

Vazquez-Acevedo, M., Cardol, P., Cano-Estrada, A., Lapaille, M., Remacle, C., González-Halphen, D. (2006). The mitochondrial ATP synthase of chlorophycean algae contains eight subunits of unknown origin involved in the formation of an atypical stator-stalk and in the dimerization of the complex. *J Bioenerg Biomembr.*, **38:**271–282.

Veldman, G. M., Klootwijk, J., van Heerikhuizen, H. (1981). The nucleotide sequence of the intergenic region between the 5.8S and 26S rRNA genes of the yeast ribosomal RNA operon. Possible implications for the interaction between 5.8S and 26S rRNA and the processing of the primary transcript. *Planta, Nucleic Acids Res.* **9:**4847 - 4862.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru

S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. (2001). The Seqquence of the Human Genome. *Science.* **291(5507):**1304-51. Erratum in: *Science,* **292(5523):**1838.

Vesteg, M., Vacula, R., Burey, S., Ffelhardt, W. L., Drahovska, H., Martin, W. and Ovic, J. K. (2009). Expression of Nucleus-Encoded Genes for Chloroplast Proteins in the Flagellate *Euglena gracilis*. *J. Eukaryot. Microbiol.,* **56(2):**159–166.

Vesteg, M., Vacula, R., Steiner, J. M., Mateasikova, B., Loffelhardt, W., Brejova, B., Krajcovic, J. (2010). A possible role for short introns in the acquisition of stroma-targeting peptides in the flagellate *Euglena gracilis*. *DNA Res.,* **17:**223–231.

Vigues, B., Metenier, G. and Senaud, J. (1984). The sub-surface cytoskeleton of the ciliate *Polyplastron multivesiculatum*: isolation and major protein components. *Europ. J. Cell Biol.,* **35**:336–342.

Vismara, R., Barsanti, L., Lupetti, P., Passarelli, V., Mercati, D., Dallai, R., Gualtieri, P. (2000). Ultrastructure of the pellicle of *Euglena gracilis. Tissue & Cell.,* **32 (6):** 451-456.

Voelkerding, K. V., Dames, S. A., Durtschi, J. D. (2009). Next-generation sequencing from basic research to diagnostics. *Clin Chem*, **55:** 641–658.

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

Votypka, J., Klepetková, H., Jirku, M., Kment, P., Lukes, J. (2012). Phylogenetic relation-ships of trypanosomatids parasitising true bugs (Insecta: Heteroptera) insub-Saharan Africa. *Int J Parasitol*, **42:**489–500.

Wacker, W. E. C, Kornicker, W. and Pothier, L. (1965). Peptide-RNA complexes in *E. gracilis. Proc. Am. chem. Soc. 150th Meeting,* 88c.

Wacker, W. E. C. & Vallee, B. L. (1959). Nucleic acids and metals. I. Chromium, manganese, nickel, iron and other metals in ribonucleic acid from diverse biological sources. *J. biol. Chem.* **234:** 3257-3262.

Wacker, W. E. C. (1962). Nucleic acids and metal. III. Changes in nucleic acid, protein and metal content as a consequence of zinc deficiency in *Euglena gracilis. Biochemistry, N.Y.* **1:** 859-865.

Wagner, V., Boesger, J., Mittag, M. (2009). Sub-proteome analysis in the green Xagellate alga *Chlamydomonas reinhardtii*. *J Basic Microbiol.,* **49:**32–41.

Wakabayashi, KI., Ide, T., Kamiya, R. (2009). Calcium-dependent Xagellar motility activation in *Chlamydomonas reinhardtii* in response to mechanical agitation. *Cell Motil Cytoskel.,* **66:**736–742.

Wald, G., (1945 - 46). *Harvey Lectures, Ser.* **41:** 117

Walder, J., Eder, P., Engman, D., Brentano, S., Walder, R., *et al* (1986). The 35-nucleotide spliced leader sequence is common to all trypanosome messenger RNA's. *Science*, **233:** 569–571.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., Earl, A. M. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS ONE*, **9(11):** e112963. doi:10.1371/journal.pone.0112963

Waller, R. F., Jackson, C. J. (2009). Dinoflagellate mitochondrial genomes: stretching therules of molecular biology. *Bioessays,* **31:** 237–45.

Walne, P. L. (1980). Euglenoid Flagellates. E.R. Cox (Ed.), Phytoflagellates, Elsevier/North Holland, New York, **pp.** 165–205

Walne, P. L. and Kivic, P. A. (1989). *Phylum Euglenida. InHandbook of Protoctista,* Vol. 1, (Margulis, L., Corliss, J. O., Melkonian, M. and Chapman, D. J. (eds) **pp.** 270–287. Jones and Bartlett, Boston, MA.

Walne, P. L., H. J. Arnott (1967). The comparative ultrastructure and possible function of eyespots: *Euglena granulata* and *Chlamydomonas eugametos. Planta,* 77, pp. 325–353

Walne, P. L., Moestrup, O., Norris, R. E., Ettl, H. (1986). Light and electron microscopical studies of *Eutreptiella eupharyngea* sp. nov. (Euglenophyceae) from Danish and American waters. *Phycologia,* **25:**109–126

Walne, P.L., Passarelli, V., Barsanti, L. and Gualtieri, P. (1998). *Crit. Rev. Plant Sci.* **17:**559 – 574.

Wang, C.C., Roney, W. B., Alston, R. L., and Spremulli, L. L. (1989). Initiation complex formation on Euglena chloroplast 30S subunits in the presence of natural mRNAs. *Nucleic Acids Res.,* **17:**9735 – 9747.

Wang, M., Gheiratmand, L. and He, C. Y. (2012). An interplay between Centrin2 and Centrin4 on the bi-lobed structure in Trypanosoma brucei. *Molecular Microbiology,* **83(6):** 1153 – 1161.

Wang, Y., Wang, M. and Field, M. C. (2010). *Trypanosoma brucei*: Trypanosome-specific endoplasmic reticulum proteins involved in variant surface glycoprotein expression. *Experimental Parasitology,* **125:**208–221.

Warner, J. R. (1999). The economics of ribosome biosynthesis in yeast. *Trends Biochem Sci*., **24:**437 - 440.

Watanabe, F., Abe, K., Tamura, Y., and Nakano, Y. (1996). Adenosylcobalamin-dependent methylmalonyl-COA mutase isoenzymes in the photosynthetic protozoa *Euglena gracilis*. *Microbiology,* **142:** 2631–2634.

Watanabe, F., Ito, T., Tabuchi, T., Nakano, Y., and Kitaoka, S. (1988a). Isolation of pellicular cobalamin-binding proteins of the cobalamin uptake system of *Euglena gracilis. Journal of General Microbiology,* **134:** 67-74.

Watanabe, F., Nakano, Y., Stupperich, E. (1992). Different Corrinoid Specificities for Cell Growth and the Cobalamin Uptake System in *Euglena gracilis* Z. *Journal of General Microbiology*, **138:** 1807-1 813.

Weale, R. A. (1976). Ocular optics and evolution. *J. Opt. Soc. Am.,* **66:** 1053–1054

Weatherbee, J. A. (1981). Membranes and cell movement: interactions of membranes with proteins of the cytoskeleton. *Int. Rev, Cytol.* **12:** 113-176.

Webb, M. S. and Green, B. R. (1991). Biochemical and biophysical properties of thylakoid acyl lipids. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, **1060 (2):** 133 -158.

Weber, M. J. (2006). Mammalian small nucleolar RNAs are mobile genetic elements. *PLoS Genet.,* **2:e205:** *Erratum in PLoS Genet. (2007), 3, e36.*

Weerakoon, N. D., Harper,J. D. I., Simpson, A. G. B., Patterson, D. J. (1999). Centrin in the groove: immunolocalisation of centrin and microtubules in the

putatively primitive protist *Chilomastixcuspidata* (Retortamonadida). *Protoplasma,* **210:**75– 84.

Wehland, J., Weber, K., Osborn, M., (1980). Translocation of actin from the cytoplasm into the nucleus in mammalian cells exposed to dimethylsulfoxide. *Biol. Cell.,* **39:** 109 -111.

Weinrich, D. H. (1924). Studies on *Euglenamorpha hegeneri* n.g. n. sp., a Euglenoid flagellates found in tadpoles. *Biol. Bull.,* **47:**149–174

Weissenbach, J. (1993). A second generation linkage map of the human genome based on highly informative microsatellite loci. *Gene,* **135:**275 – 278.

Weizhong, Li and Adam Godzik (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics,* **22:**1658-9.

Wellauer, P. K., Dawid, I. B., Kelley, D. E., Perry, R.P. J. (1974). Secondary structure maps of ribosomal RNA. II. Processing of mouse L-cell ribosomal RNA and variations in the processing pathway. *Mol. Biol.,* **89:**397-407.

Wen, J. (2000). The nuclear matrix of *Euglena gracilis* (Euglenophyta): A stage of nuclear matrix evolution? *Biology of the Cell,* **92:**125 -131.

Wen, J. F., Li, J. Y. (1992). The nucleoskeleton of special dinoflagellate, *Oxyrrhis marina. Zool. Res.,* **13:**89 - 94.

Wen, J. F., Li, J. Y. (1998). Nuclear matrix of the most primitive eukaryote archezoa. *Sci. China,* **41:**479-487.

Westermann, S., and Schleiffer, A. (2013). Family matters: structural and functional conservation of centromere-associated proteins from yeast to humans. *Trends Cell Biol.*, **23:**260–269.

Westhorpe, F. G., and Straight, A. F. (2013). Functions of the centromere and kinetochore in chromosome segregation. *Curr. Opin. Cell Biol.,* **25:**334–340.

Wettstein, D., Rasmussen, S. W., Holm, P. B. (1984). The synaptonemal complex in genetic segregation. *Annu Rev Genet.,* **18:**331–413.

Wheeler, G. L., and Brownlee, C. (2008). $Ca^{2+}$ signaling in plants and green algae – changing channels. *Trends in Plant Science,***13(9):** 506-514.

Wheeler, T. J., and Eddy S. R. (2013). nhmmer: DNA Homology Search With Profile HMMs. *Bioinformatics*, **29:**2487-2489.

Wickstead, B. and Gull, K. (2011). The evolution of the cytoskeleton. *J. Cell Biol.,* **194(4):**513–525.

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). *J. Eukaryot Microbiol.,* **60:**214–221.

Will, C. L., and Luhrmann, R. (2011). Spliceosome structure and function*. Cold Spring Harb. Perspect. Biol.,* **3:**a003707.

Wille Jr., J. J., Ehret, C. F. (1968). Light synchronization of an endogenous circadian rhythm of cell division in Tetrahymena. *J. Protozool.,* **15:** 785–789.

Willey, R. L., Wibel, R. G. (1985). A cytostome/cytopharynx in green euglenoid flagellates (Euglenales) and its phylogenetic implications. *BioSystems*, **18:**369–376.

Willey, R. L., Wibel, R. G. (1987). Flagellar roots and the reservoir cytoskeleton of *Colacium libellae* (Eglenophyceae). *J. Phycol.,* 23:283–288

Williams, E. J. and Bowles, D. J. (2004). Coexpression of neighboring genes in the genome of Arabidopsis thaliana. *Genome Res.,* **14:**1060–7.

Williams, T. A., Foster, P. G., Cox, C. J and Embley, T. M. (2013). An archaeal origin of eukaryotes supports only two primary domains of life. *Nature,* **504(7479):**231–236.

Wilson, D. N., and Doudna Cate, J. H. (2012). The structure and function of the eukaryotic ribosome. *Cold Spring Harb. Perspect.Biol.,* **4:**1–17. http://dx.doi.org/10.1101/cshperspect.a011536.

Wilson, K. L. and Dawson, S. C. (2011). Functional evolution of nuclear structure. *J. Cell Biol.,***195(2):** 171–181. www.jcb.org/cgi/doi/10.1083/jcb.201103171.

Winicov, I. (1976). Alternate temporal order in ribosomal RNA maturation. *J. Mol. Biol.* **100:**141-155.

Witkus, E. R., Grillo, R. S. and Smith, W. J. (1969). Microtubule bundles in the hindgut epithelium of the woodlouse *Oniscus ascellus. J. Ultrastruct. Res.,* **29:**182-190.

Wittinghofer, A. and Vetter, I. R. (2011). Structure-Function relationships of the G domain, a canonical switch motif. *Ann. Review of Biochemistry,* **80:** 943-971.

Wolken, J. J. (1961). An Experimental Organism for Biochemical and Biophysical Studies. *Quinn and Boden Company, Rahway (NJ).*

Wolken, J. J. (1967). Euglena. Appleton-Century-Crofts, New York

Wolken, J. J. (1975). Photoprocess, Photoreceptors and Evolution. Academic Press, New York

Wolken, J. J. (1977). *Euglena*: the photoreceptor system for phototaxis*. J. Protozool.,* 24, **pp.** 518–522.

Wolowski, K. (1998a). Taxonomic and environmental studies on euglenophytes of the Kraków-Czestochowa Upland (Southern Poland). *Fragmenta Floristica et Geobotanica Suppl.,* **6:**1 – 192.

Wolowski, K. (1998b). Euglenophyta. In: Whitton B. A., John D. M., Johnson L. R., Boulton P. N. G., Kelly M. G, Haworth E. Y (eds) A Coded List of Freshwater Algae of the British Isles. NERC, LOIS Publication Number 222. *Institute of Hydrology, Wallingford:* 27 – 32.

Wu, H., Zhang, Z., Hu, S. and Yu, J. (2012). On the molecular mechanism of GC content variation among eubacterial genomes. *Biology Direct,* **7:**2 **DOI:** 10.1186/1745-6150-7-2.

Wuichet K. and Søgaard-Andersen, L. (2014). Evolution and Diversity of the Ras Superfamily of Small GTPases in Prokaryotes. *Genome Biol. Evol.,*

Wurst, M., Seliger, B., Jha, B. A., Klein, C., Queiroz, R., and. Clayton, C. (2012). Expression of the RNA recognition motif protein RBP10 promotes a bloodstream-form transcript pattern in *Trypanosoma brucei*. *Mol. Microbiol.,* **83:**1048–1063.

Wuthisuthimethavee, S., Lumubol, P., Vanavichit, A. and Tragoonrung, S. (2003). Development of microsatellite markers in black tiger shrimp (Penaeus monodon Fabricius). *Aquaculture,* **224:**39–50. doi/10.1083/jcb.201103008.

Wynn-Williams, D. D., Edwards, H. G. M., Newton, E. M., Holder, J. M. (2002). Pigmentation as a survival strategy for ancient and modern photosynthetic microbes under high ultraviolet stress on planetary surfaces. *Int. J. Astrobiol*., **1:**39–49.

Xu, Y., Liu, L., Lopez-Estrano, C., Michaeli, S. (2001) Expression studies on clustered trypanosomatid box C/D small nucleolar RNAs. *J. Biol. Chem.,* **276(17):**14289–14298.

Xu, Z., Kohel, R. J., Song, G., Cho, J., Alabady, M., Yu J, Koo, P., Chu, J., Yu, S., Wilkins, T. A., Zhu, Y., Yu, J. Z. (2008). Gene-rich islands for fiber development in the cotton genome. *Genomics,* **92:**173–83.

Yabuki, A., Nakayama, T., Yubuki, N., Hashimoto, T., Ishida, K. and Inagaki, Y. (2011). Tsukubamonas globosa n. g., n. sp.; a novel excavate flagellate

possibly holding a key for the early evolution in“Discoba”. *J. Eukaryot. Microbiol.,* **58:**319 – 331.

Yabuta, Y., Takamatsu, R., Kasagaki, S. and Watanabe, F. (2013). Isolation and Expression of a cDNA Encoding Methylmalonic Aciduria Type A Protein from *Euglena gracilis* Z. *Metabolites*, ***3*:**144-154. doi:10.3390/metabo3010144.

Yamada, K. M., Spooner, B. S. and Wessells, N. K. (1971). Ultrastructure and function of growth cones and axons of cultured nerve cells. *J. Cell Biol.,* **49:**614-635.

Yamada, K., Suzuki, H., Takeuchi, T., Kazama, Y., Mitra, S., Abe, T., Goda, K., Suzuki, K. and Iwata, O. (2016). Efficient selective breeding of live oil-rich *Euglena gracilis* with fluorescence-activated cell sorting. *Scientific Reports*, 6:26327). doi:10.1038/srep26327.

Yamaguchi, S., Hong, K., Liu, R., Shen, L., Inoue, A., Diep, D., Zhang, K., and Zhang, Y. (2012). Tet1 controls meiosis by regulating meiotic gene expression. *Nature,* **492:** 443 – 447.

Yandell, M. and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics,* **13:**329 - 342 doi:10.1038/nrg3174

Yang, L., Guan, T., Gerace, L. (1997). Lamin-bind fragment of LAP2 inhibits increase in nuclear volume during the cell cycle and progression into S phase. *J. Cell Biol.* **139:**1077-1087.

Yang, X., Feng, Y., He, Z., Stoffella, P. J. (2005). Molecular mechanisms of heavy metal hyperaccumulation and phytoremediation. *J. Trace Elem. Med. Biol.,* **18:**339–353.

Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.,* **10:**1396–1401.

Yang, Z., Rannala, B. (2006). Bayesian estimation of species divergence times under amolecular clock using multiple fossil calibrations with soft bounds. *Mol Biol Evol.,* **23:**212–26.

Yarian, C. S., Basti, M. M., Cain, R. J., Ansari, G., Guenther, R. H., Sochacka, E. *et al.* (1999). Structural and functional roles of the N1- and N3-protons of Ψ at tRNA's position 39. *Nucleic Acids Res.,* **27:**3543–3549.

Yasuhira, S., Simpson, L. (1997). Phylogenetic affinity of mitochondria of *Euglena gracilis* and kinetoplastids using cytochrome oxidase I and hsp60. *J Mol Evol.,* **44:** 341—347.

Yasuhira, S., Simpson, L. (1997). Phylogenetic affinity of mitochondria of *Euglena gracilis* and kinetoplastids using cytochrome oxidase I and hsp60. *J Mol Evol.,* **44:**341–347

Yee, M-C and Bartholomew, J. C. (1988). Light Regulation of the Cell Cycle in Euglena gracilis bacillaris. *Cytometry,* **9:**387-393.

Yin, J. and Wang, G.(2014). The Mediator complex: a master coordinator of transcription and cell lineage development. Development, **141:** 977-987 doi:10.1242/dev.098392.

Yoffe, Y., Léger, M., Zinoviev, A., Zuberek, J., Darzynkiewicz, E., Wagner, G.,Shapira, M. (2009). Evolutionary changes in the *Leishmania* eIF4F complexinvolve variations in the eIF4E-eIF4G interactions. *Nucleic Acids Res.,* **10:**3243–3253. http://dx.doi.org/10.1093/nar/gkp190.

Yoffe, Y., Zuberek, J., Lerer, A., Lewdorowicz, M., Stepinski, J., Altmann, M., Darzynkiewicz, E., Shapira, M. (2006). Binding specificities and potentialroles of isoforms of eukaryotic initiation factor eIF4E in *Leishmania*. *Eukaryot. Cell,* **5:**1969–1979. http://dx.doi.org/10.1128/EC.00230-06.

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

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

Yorimitsu, T., Sato, K. and Takeuchi, M. (2014). Molecular mechanisms of Sar/Arf GTPases in vesicular trafficking in yeast and plants. *Frontiers in plant science,* **5:**1-12. Doi:10.3389/fpls.2014.00411.

Yoshida Y, Tomiyama T, Maruta T, Tomita M, Ishikawa T, Arakawa K (2016). De novo assembly and comparative transcriptome analysis of Euglena gracilis in response to anaerobic conditions. *BMC Genomics.,* **17:**182. doi: 10.1186/s12864-016-2540-6.

Yoshida,Y., Tomiyama,T., Maruta,T., Tomita,M., Ishikawa,T. and Arakawa,K.(2016). De novo assembly and comparative transcriptome analysis of Euglena gracilis in response to anaerobic conditions. BMC Genomics, **17:**182DOI 10.1186/s12864-016-2540-6.

Yoshizawa, A.C., Kawashima, S., Okuda, S., Fujita, M., Itoh, M., Moriya, Y., Hattori, M. and Kanehisa, M. (2006). Extracting sequencemotifs and the phylogenetic features of SNARE-dependent membrane traffic. *Traffic*, **7(8):**1104–1118. doi: 10.1111/j.1600-0854.2006.00451.x.

Yu, B., Fey, P., Kestin-Pilcher, K. E., Fedorov, A., Prakash, A., Chisholm, R. L.,Wu, J. Y. (2011). Spliceosomal genes in the *D. discoideum* genome: a comparison with those in *H. sapiens*, *D. melanogaster*, *A. thaliana* and *S. cerevisiae*. *Protein Cell,* **2(5):** 395–409. DOI 10.1007/s13238-011-1052-z.

Yubuki, N., Edgcomb, V. P., Bernhard, J. M. and Leander, B. S. (2009). Ultrastructure and molecular phylogeny of Calkinsia aureus: cellularidentity of a novel clade of deep-sea euglenozoans with epibiotic bacteria. *BMC Microbiol.,* **9:**16.

Záhonová, K., Hadariová, L., Vacula R., Yurchenko, V., Eliáš, M., Krajčovič, J., Vesteg, M. (2014). A small portion of plastid transcripts is polyadenylated in the flagellate *Euglena gracilis*. *FEBS Letters,* **588:**783–788.

Zakrys, B. (1986). Contribution to the Monograph of Polish Members of the Genus Euglena Ehrenberg 1830. *Nova Hedw.,* **42:**491-540.

Zakrys, B. (1988). The nuclear DNA level as a potential taxonomic character in Euglena EHR. (Euglenophyceae). *Algological Studies,* **49:**483 - 504.

Zakrys, B. and Walne, P. L. (1994). Floristic, taxonomic and phytogeographic studies of green Euglenophyta from the Southeastern United States, with emphasis on new and rare species. *Algol. Stud.* **72:**71-114.

Zakrys, B. and Walne, P. L. (1994). Floristic, taxonomic and phytogeographic studies of green Euglenophyta from the Southeastern United States, with emphasis on new and rare species. *Algol. Stud.,* **72:**71-114.

Zakrys, B. and Walne, P.L. (1998). Ultrastructure of mucocysts in *Euglena oxyuris* Schmarda (Euglenophyceae). *Algolog. St.,* **88:** 125-133.

Zemann, A., op de Bekke, A., Kiefmann, M., Brosius, J. and Schmitz, J. (2006). Evolution of small nucleolar RNAs in nematodes. *Nucleic Acids Res.,* **34:**2676–2685.

Zhan, S., Horrocks, J., Lukens, L. N. (2006). Islands of co-expressed neighbouring genes in *Arabidopsis thaliana* suggest higher-order chromosome domains. *Plant J.,* **45:**347–57.

Zhang, H., Campbell, D. A., Sturm, N. R., Lin, S. (2009). Dinoflagellate spliced leader RNA genes display a variety of sequences and genomic arrangements. *Mol Biol Evol.,* **26:** 1757–1771.

Zhang, H., Ehrenkaufer, G. M., Pompey, J. M., Hackney, J. A., Singh, U. (2008). Small RNAs with 5′-polyphosphate termini associate with a Piwi-related protein and regulate gene expression in the single-celled eukaryote Entamoeba histolytica. *PLoS Pathog.,* **4:**e1000219.

Zhang, K., Barron, T., Turco, S. J., and Beverley, S. M.(2004). The LPG1 gene family of Leishmania major. *Mol. Biochem. Parasitol.,***136:**11–23.

Zhang, T., Liu, Y., Yang, T., Zhang, L., Xu, S., Xue, L., An, L. (2006). Diverse signals converge at MAPK cascades in plant. Plant Physiology and Biochemistry, **44:**274-283.

Zhang, T., Zhang, X., Hu, S., and Yu, J. (2011). An efficient procedure for plant organellar genome assembly, based on whole genome data from the 454 GS FLX sequencing platform. *Plant Methods,* **7:** 38.

Zhang, WJ., Yang, J., Yu, YH and Shen, YF (2005). Identification, Cloning, and Characterization of Microsatellite DNA in *Euglena gracilis*. *J. Eukaryot. Microbiol.,* **52(4):**356–359.

Zhang, X., Cui, J., Nilsson, D., Gunasekera, K., Chanfon, A., Song, X., Wang, H., Xu, Y., and Ochsenreiter, T.(2010). The *Trypanosoma brucei* MitoCarta and its regulation and splicing pattern during development. *Nucleic Acids Research,* **38(21):**7378-7387. doi: 10.1093/nar/gkq618.

Zhang, Y., Lu, L., Furlonger, C., Wu, G. E., Paige, C. J. (2000). Hemokinin is a hematopoietic-specific tachykinin that regulates B lymphopoiesis. *Nat. Immunol.,* 1: 392–397.

Zhou X, Ren L, Meng Q, Li Y, Yu Y, Yu, J. (2010). The next-generation sequencing technology and application. *Protein Cell,* **1:**520–536.

Zhou, H. & Qu, L.-H. (2008). Genomewide analysis of box C/D and box H/ACA snoRNAs in Chlamydomonas reinhardtii reveals an extensive organization into intronic gene clusters. *Genetics,* **179:** 21–30.

Zhou, Q., Gheiratmand, L., Chen, Y., Lim, T. K., Zhang, J., Li, S., Xia, N., Liu, B., Lin, Q., He, C. Y. (2010). A comparative proteomic analysis reveals a new bi-lobe protein required for bi-lobe duplication and cell division in *Trypanosoma brucei*. *PLoS One,* **5(3):**e9660. doi:10.1371/journal.pone. 0009660.

Zhou, W., Li, S., Liu, Y., Qi, X., Chen, H., Cheng, C. H. K., Liu, X., Zhang, Y., Lin, H.(2012). The evolution of tachykinin/tachykinin receptor (TAC/TACR) in vertebrates and molecular identification of the TAC3/TACR3 system in

zebrafish (*Danio rerio*). *Molecular and Cellular Endocrinology,* **361:**202–212.

Zíková, A., Hampl, V., Paris, Z., Týč, J., Lukeš, J. (2016). Aerobic mitochondria of parasitic protists: Diverse genomes and complex functions. *Mol Biochem Parasitol.,* **209(1-2):**46-57. doi: 10.1016/j.molbiopara.2016.02.007.

Zikova, A., Panigrahi, A. K., Uboldi, A. D., Dalley, R. A., Handman, E., Stuart, K. (2008). Structural and functional association of Trypanosoma brucei MIX protein with cytochrome c oxidase complex. Eukaryot. *Cell,* **7:**1994–2003.

Zikova, A., Schnaufer, A., Dalley, R. A., Panigrahi, A. K., Stuart, K. D. (2009). The F(0)F(1)-ATP synthase complex contains novel subunits and is essential for procyclic *Trypanosoma brucei*. *PLoS Pathog.* **5:**e1000436.

Zimba, P. V., Moeller, P. D., Beauchesne, K., Lane, H. E., Triemer, R. E. (2010). Identification of euglenophycin – A toxin found in certain euglenoids. *Toxicon,* **55:** 100 – 104.

Zinoviev, A. and Shapira, M. (2012). Evolutionary conservation and diversificationof the translation initiation apparatus in trypanosomatids. *Comp.Funct. Genomics,* **2012:**813718. http://dx.doi.org/10.1155/2012/813718.

Zomer, A. W., Opperdoes, F. R., and van den Bosch, H. (1995). Alkyl dihydroxyacetone phosphate synthase in glycosomes of *Trypanosoma brucei*. *Biochim. Biophys. Acta,* **1257:**167−173.

Zuo, Y., and Deutscher, M. P. (2001). Exoribonuclease superfamilies: structural analysis and plylogenetic distribution. *Nucleic Acids Res*, **29:**1017–1026.