Identification of a trypanosomatid *c*-type cytochrome maturation system and variant metabolism in other unicellular Eukaryotes

Asma Belbelazi

A thesis submitted to the University of Huddersfield in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Acknowledgements

I would like to express my sincere appreciation to my supervisor Prof. Michael Ginger for his endless support during my PhD study. His patience, inspiration, and great knowledge helped me throughout my study. I also would like thank my second supervisor Dr. Martin Carr for his help throughout my bioinformatics and computational analysis. Additionally, I would like to thank Dr. Jane Harmer for her insightful comments and continuous encouragement during my time in the lab.

Finally, I would like to thank my family for their unlimited support but most importantly, my sincere thanks goes to my great husband for his support and patience for the past years.

Abstract

Trypanosomatids are an unusual early-branching group of eukaryotic organisms in which many biological pathways differ extensively from the same pathways seen in other eukaryotes. Trypanosomatids include several medically relevant species such as *Trypanosoma brucei*, which is the causal agent of the tropical disease African sleeping sickness, and pathogenic *Leishmania* species causing different manifestations of leishmaniasis. New medicines are very much needed to treat these neglected tropical diseases. In eukaryotes, cytochromes *c* and c_1 are essential components of a typical mitochondrial electron transport chain, which is used for energy production. Generally, the activity of *c*-type cytochromes is determined by the covalent attachment of heme to two cysteine residues in a heme-binding motif. Mitochondrial cytochromes found in trypanosomatids are unique because they contain a single cysteine in the heme-binding motif. For the heme to become attached to cysteine(s) in a heme-binding motif, *c*-type cytochromes must undergo post-translation modification (or maturation). Previous genome analyses, in contrast to other eukaryotes, provided no evidence for the existence of the protein(s) responsible for catalysing this post-translational modification in trypanosomatids. A study that identified this maturation system would solve this puzzle and could provide an opportunity to identify a novel therapeutic target for a group of medically important parasitic protists.

In the mitochondrial intermembrane space of eukaryotes, other than trypanosomatids, either System I or III is used to attach heme to *c*-type cytochromes. One of my early analyses was to determine the distribution and key functional residues of both System I and III. The aim was to understand the types of motifs that might be essential for heme attachment in trypanosomatids' cytochrome *c* maturation system. The search for a candidate maturation protein was carried out using proteomic sift between trypanosomatids and *Phytomonas* proteomes. Candidate proteins conserved in trypanosomatids but not in *Phytomonas* were compiled and analysed to assess whether they contain conserved key motifs of either System I or III. In probing the evolutionary biochemistry of cytochrome *c* maturation, an unexpected variation in the distribution of Systems I versus III was noticed in different eukaryotic clades. Another significant outcome from the analysis was the identification of a candidate cytochrome *c* maturation in the trypanosomatids. The overall architecture of this candidate suggested a divergent System III. A combination of laboratory techniques was used to experimentally prove that this candidate protein functions as a catalyst for the maturation of the trypanosomatids *c*-type cytochrome. These techniques include protein expression and purification followed by heme staining and UV-Vis spectroscopy to detect heme attachment.

In the second part of my research, I focused on other aspects of the unusual metabolic biochemistry in other protists. In particular, I used bioinformatics and literature-led approaches to analyse mitochondrial metabolism in *Naegleria* species. *N. gruberi*, distantly related to the Euglenozoa is one of the few eukaryotes with extreme versatility to its energy metabolism and capable of apparent switching between oxidative phosphorylation, anaerobic respiration and fermentation linked to H_2 generation. The focus of the bioinformatics investigation was to examine the capability of *N. gruberi* to encode for two well-known anaerobic enzymes; acetyl-CoA synthetase and nitrite reductase. Findings have shown that *N. gruberi* candidate acetyl-CoA synthetase does not possess the functional key residues used to characterise the acetyl-CoA synthetase family. Thus, it is likely that *N. gruberi* candidate protein is involved in an alternative catalytic pathway. On the other hand, *N. gruberi* candidate nitrite reductase has all the essential residues and overall structure previously identified to be critical for nitrite reductase function. This indicates that it is likely to contribute to anaerobic respiration under appropriate environmental conditions. Another area of focus was to examine the cryptic peroxisomal targeting of the metabolic enzymes. The outcome of this analysis shows *N. gruberi* cryptic PTS1 motifs, in several metabolic enzymes including some of the main glycolytic enzymes, can confer peroxisomal targeting, which suggests that the function of peroxisomes can be more versatile than previously expected.

Contents

Abbreviations

ACD	Acetyl CoA synthetase (ADP-forming)
ACS	Acetyl CoA synthetase (AMP-forming)
AR	Aldehyde reductase
BLAST	Basic Local Alignment Search Tool
BLASTp	Protein-protein basic local alignment search tool
BLASTx	translate nucleotide- protein basic local alignment search tool
CCB	Cofactor assembly complex C
Ccm	Cytochrome *c* maturation
cDNA	Complementary deoxyribose nucleic acid
CS	Citrate synthase
DAPI	4',6-diamidino-2-phenylindole
DNA	Deoxyribonucleic acid
EGT	Endosymbiotic gene transfer
ETC	Electron transport chain
gDNA	Genomic deoxyribose nucleic acid
GFP	Green fluorescent protein
GPI	Glycosylphosphatidylinositol
HCCS	Holocytochrome *c* synthase
IDH2	Isocitrate dehydrogenase NADPH
IMS	Intermembrane space
IPTG	Isopropyl β-D-1- thiogalactopyranoside
kDNA	Kinetoplast deoxyribose nucleic acid
KO	Knockout
LECA	last eukaryotic common ancestor
LGT	Lateral Gene Transfer
MDH	Malate dehydrogenase
mRNA	Messenger ribonucleic acid
MROs	Mitochondrial-related organelles
NCBI	National Centre for Biotechnology Information

ORF	Open reading frame
PAM	Primary amoebic meningoencephalitis
PCR	Polymerase chain reaction
PGK	Phosphoglycerate kinase
RNA	Ribonucleic acid
ROS	Reactive oxygen species
RT-PCR	Reverse transcription polymerase chain reaction
SOD	Superoxide dismutase
*Tb*CYTC	T.brucei apocytochrome c
*Tb*pHCCS	T. brucei putative holocytochrome c synthase
TMBZ	3,3',5,5'-tetramethylbenzidine
TPI	Triosephosphate isomerase
UTR	Untranslated region
VSG	Variable surface glycoproteins
WHO	World Health Organization

Chapter 1

Introduction

In eukaryotes, cytochromes *c* and c_1 are essential components of typical mitochondrial electron transport chains (ETC) (Figure 1.1). Cytochrome *c* is a small soluble protein, which is required for transfer of electrons from complex III to complex IV. Cytochrome c_1 is a core subunit in complex III and it is involved in passing electrons from the Rieske subunit to cytochrome *c* (Rieske, 1976). The activity of a *c*-type cytochrome is determined by the covalent attachment of the prosthetic heme group (Fe-protoporphyrin IX) to the apocytochrome in a post-translation modification. The covalent bond is formed between the vinyl group of the porphyrin ring in the heme and the Cys(s) in the apocytochrome heme binding motif (Pettigrew and Moore, 2012, Sanders et al., 2010). In all eukaryotic lineages except one, the heme binding motif in the mitochondrial cytochromes contains two Cys residues (CXXCH motif where X is any amino acid except for Cys). The His in the CXXCH motif acts as an axial ligand to the iron atom in the heme (Verissimo and Daldal, 2014). The one eukaryotic group in which the mitochondrial *c*-type cytochromes have a single Cys in heme-binding motif is the Euglenozoa in which the parasitic trypanosomatids are found. Identifying how these atypical cytochromes are matured was a primary aim of my PhD research. Alongside this work I also looked at other aspects of organellar metabolism in protists including the composition and distribution of cytochrome *c* maturation systems in eukaryotes, a re-investigation of the apparent diversity of *Naegleria gruberi* metabolism and identification and characterisation of cryptic peroxisomal targeting determinant in various metabolic enzymes in *Naegleria* and other protists.

1.1 Function, structure, and biogenesis of *c*-type cytochromes

Cytochromes are a class of ubiquitous heme binding proteins, which facilitate electron transfer in several biologically important reactions, including aerobic and anaerobic respiration and photosynthesis. Their vital role inside the cell attracts the focus of many of multidisciplinary scientific groups. The crystal structure and characterisation of many cytochromes has already been studied (Riccio et al, 2002, Guo et al., 2004, Lange and Hunte, 2002). One approach to classify cytochromes is via the type of attached heme. Heme can be incorporated in a number of essential proteins including haemoglobin, myoglobin,

cytochromes, peroxidase and catalyses. The biosynthesis of the prosthetic heme is a highly conserved process, which contains 11 highly regulated enzymes found in all eukaryotes and many bacteria (Layer at al, 2010, Ponka et al., 1997). During heme biosynthesis the ferrous iron (Fe^{2+}) is inserted into protoporphyrin IX group via the activity of ferrochelatase. The pyrrole rings in the protoporphyrin supply the tetradentate ligands to the iron (Figure 1.3). In the heme macrocycle, two vinyl groups are found at positions 2nd and 4th and two propionate groups at positions 6th and 7th. Once the heme is synthesised, it is known as heme *b*. Heme *b* is the most common type of heme in cells and it is associated to proteins via a single coordinate bond (non-covalent bond), this bond is formed by an amino acid side chain in the protein (from a Cys, His or Met) and the heme iron (Bowman and Bren, 2008). Examples of proteins associated with heme *b* include haemoglobin, globins and peroxidase proteins. The attachment of heme *b* to proteins does not need maturation systems and it occurs naturally once the apoprotein is synthesised. Heme *b* can undergo covalent modification, which involves the attachment of a hydroxyethylfarnesyl group to the 2nd vinyl group in the heme. This attachment results in the formation of heme *a* and *o*. However, heme *a* is different from heme *o* due to the attachment of the formyl group at the 8th position of the heme. Like heme *b*, both hemes *a* and *o* are bound to proteins non covalently. Contrary to hemes *b*, *a* and *o*, the vinyl group of heme *c* is replaced by two thioether bonds (rarely one) covalently attached to the apoprotein (Figure 1.2). To form this covalent attachment, the cell requires an independent maturation system. Proteins bound to heme *c* are known as cytochromes *c*. Cytochromes *c* are an essential group of proteins which are often involved in energy production inside the cell. Biochemical and genetic studies have revealed the presence of four systems across prokaryotes and eukaryotes, and which catalyse the heme *c* attachment to cytochrome *c*. These are System I (cytochrome *c* maturation (Ccm) system), System II, System III (holocytochrome *c* synthase) and System IV. In eukaryotes, Systems I and III are known to catalyse heme attachment to CXXCH motif in mitochondrial cytochromes. The maturation of *c*-type cytochromes is thought to take place in the mitochondrial intermembrane space (IMS) (Giegé et al., 2008). The origin of System I can be traced back to either Bacteria or Archaea but System III is found only in eukaryotic lineages (Allen et al., 2008).

System II is found in many cyanobacteria and in the chloroplasts of phototrophic eukaryotes. System II is used to catalyse heme attachment to CXXCH motif in cytochrome c_6 and

cytochrome *f* (the sub-unit in the b_6f complex required for electron transport from photosystem II to photosystem I) (Kranz et al., 2009). At least one more system is known to catalyse a more distinctive attachment of heme to a single Cys within a single protein and that is System IV. This biogenesis system is conserved in oxygenic phototrophs (cyanobacteria, plants, green algae), and attaches heme by a single thioether linkage to cytochrome b_6 of the b_6f complex (Saint-Marcoux et al., 2009).

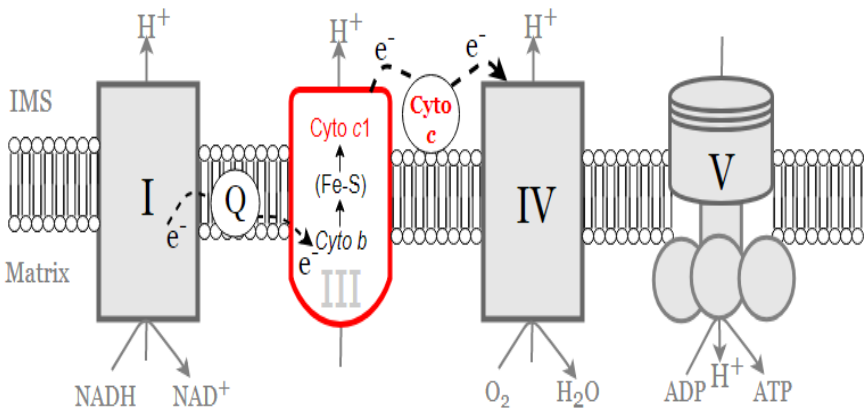


Figure 1.1 Mitochondrial electron transport chain

During glycolysis, the citric acid cycle, and other catabolic pathways, electron carriers (NAD^+, FAD) are continually reduced. In the inner mitochondrial membrane, complex I transfers electrons from NADH (and complex II transfers electrons from $FADH_2$, not shown) to the mobile electron carrier ubiquinone (Q). In turn, complex III transfers electrons from reduced ubiquinol to another mobile electron carrier, the small (~12 kDa) protein cytochrome *c*. Complex IV is the membrane-bound enzyme which, in turn, transfers electrons from reduced cytochrome *c* to the terminal electron acceptor O_2 forming H_2O. During the electron transfers catalysed by complexes I, III, and IV, the energy released is used to pump protons (H^+) into the mitochondrial IMS. This creates an ion electrochemical proton gradient which is exploited by complex V to generate energy in form of ATP. The dashed arrows indicate the transfer of electron by mobile electrons carriers.

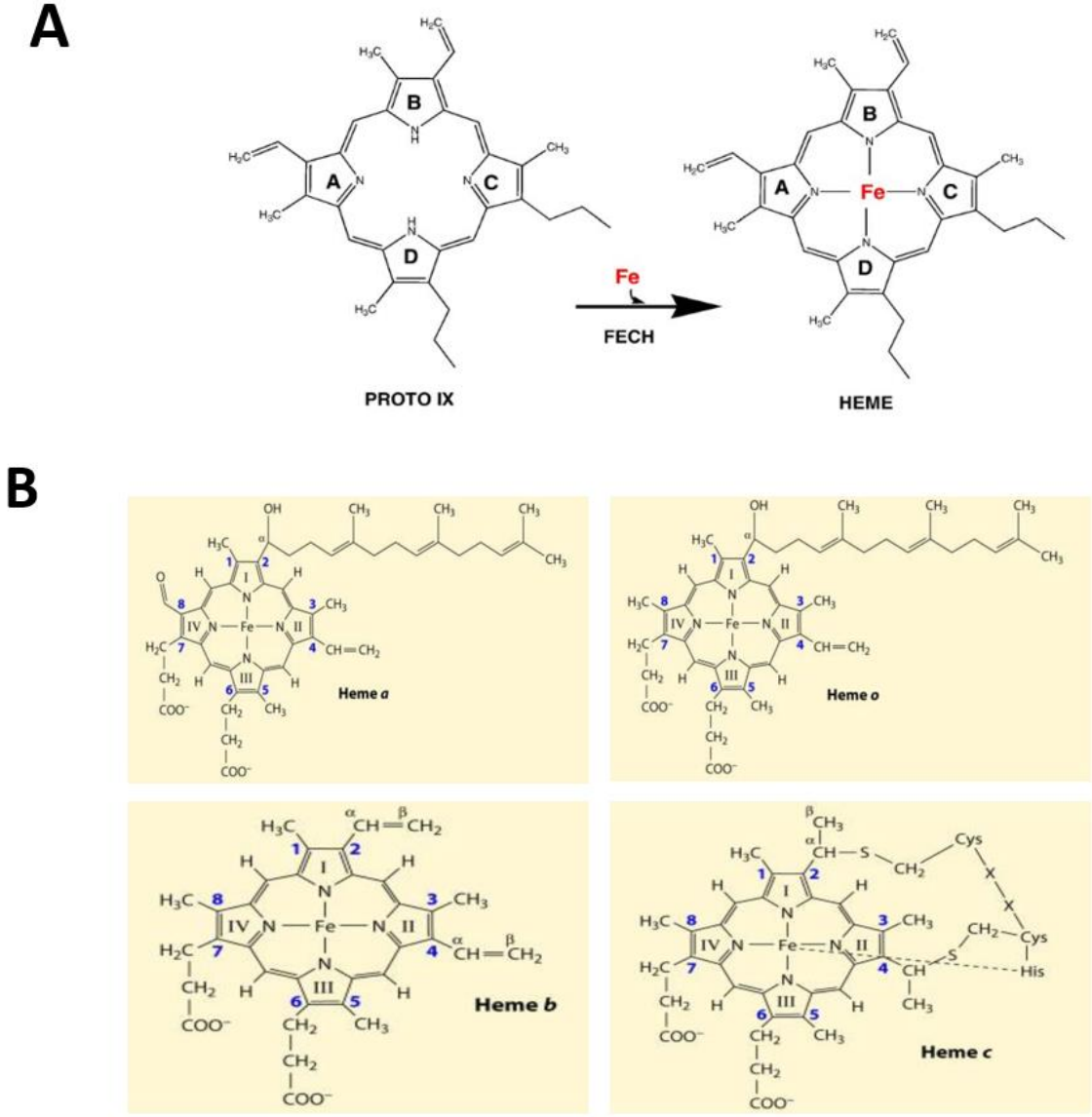


Figure 1.2. Biochemical structure of protoporphyrin IX and different heme classes

(A) The chemical structure of protoporphyrin IX (PROTO XI) and the formation of ferroprotoporphyrin IX (heme *b*). (B) Different types of heme. Heme *b* is the final product of heme biosynthesis. Heme *a, o* and *c* have a similar basic chemical structure of heme *b* but heme *a* and *o* have a covalent attachment of hydroxyethylfarnesyl group in the 2nd vinyl group. Heme *a* is different from heme *o* due to the attachment of the formyl group at the 8th position of the heme. Heme *c* forms two (and rarely a single) thioether bonds, which replace the two vinyl groups in the porphyrin ring of heme *b*. Image A was reproduced with permission from Ajioka et al. (2006) and Image B was reproduced with permission from Kranz et al (2009).

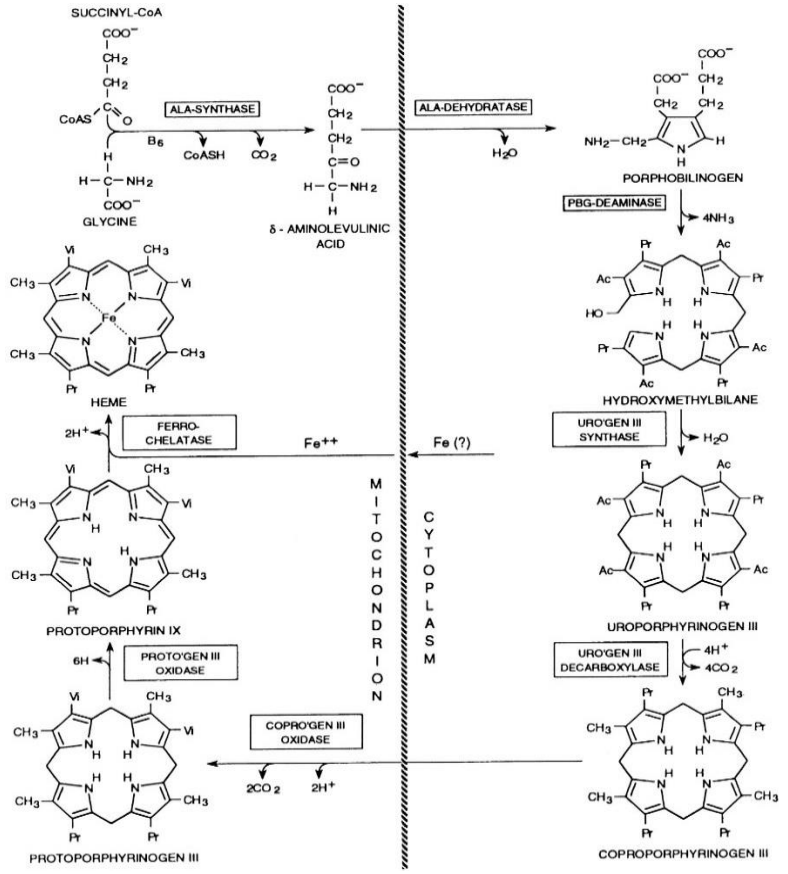


Figure 1.3. Biosynthetic pathway of heme *b* in eukaryotes
The first step of the heme biosynthesis is formation of 5-aminolevulinic acid (ALA), catalysed by ALA synthase. Next, two ALA molecules undergo serious of biochemical reactions which result in the formation of Uroporphyrinogen III and thence protoporphyrin IX. Heme formation is completed when Fe^{2+} is inserted into protoporphyrin IX by the activity of ferrochelatase enzyme. The diagram was reproduced with permission from Ponka et al (1997).

1.1.1 System I

System I is widely conserved and, to date, most comprehensively studied in bacteria. There have been comparatively very few studies on eukaryotic System I. Studies on the Systems I of *Rhodobacter capsulatus* have led to the discovery of nine main membrane and periplasmic proteins (CcmA to CcmI), plus several accessory proteins (*e.g.*DsbA, DsbD) (Thöny-Meyer et al., 1995, Sanders et al., 2010). CcmA and CcmB are ATP-dependent ABC transporter proteins which form a complex with CcmC and CcmD that is required for heme delivery to CcmE. The role of these proteins is still unknown, but studies suggest a role in releasing holoCcmE protein from CcmC using the energy released from ATP (Hamel et al., 2009, Stevens et al., 2011). CcmE is a soluble heme chaperone protein that binds covalently to heme vinyl groups by the

His in a conserved HXXXY motif. The attachment of CcmE protein to heme depends on the heme-interacting protein CcmC. The interaction between CcmC and CcmE is facilitated by the small integral membrane protein CcmD. CcmF is an integral membrane protein with at least 11 transmembrane helices; it is involved in relocalisation of the heme from CcmE to apocytochrome. For formation of the thioether linkage between the apocytochrome and heme, the Cys residues in CXXCH heme-binding motif must be reduced. The reduction of apocytochrome is facilitated by thioredoxin-like proteins CcmG and CcmH. CcmH recognises and reduces Cys sulphydryl groups in the CXXCH motif of the apocytochrome *c*. The RCXXC motif in CcmH is thought to be involved in the reduction of the Cys residues in the heme-binding motif (CXXCH) in the apocytochrome *c* (Verissimo and Daldal, 2014, San Francisco et al., 2014, Sanders et al., 2010, Stevens et al., 2011, Meyer et al., 2005) (Figure 1.4).

A common architecture to multi-component System I is found in many bacteria and this biogenesis system is used to attach heme to a wide variety of cytochromes. However, there are some components and potentially some catalytic differences seen between the Systems I of eukaryotes and bacteria. In eukaryotes, the CcmI, CcmG, CcmD and thioredoxin-like proteins DsdA and DsdD are apparently missing (Verissimo and Daldal, 2014). In eukaryotes with System I some proteins are integral within the inner mitochondrial membrane and in taxa reported thus far some System I components are encoded in the mitochondrial genome (Verissimo and Daldal, 2014).

1.1.2 System II

System II is found in many Gram-positive bacteria, some of proteobacteria, cyanobacteria and in chloroplasts of plants and algae (Kranz et al., 2009, Beckett et al., 2000). Many studies have focused on the structure of System II as it is found in *Chlamydomonas reinhardtii* (chlorophytes) and in some bacteria. Bacteria analysed in some depth include *Bacillus subtilis* and *Bordetella pertussis* (Xie et al., 1998, Hamel et al., 2003, Le Brun et al., 2000, Feissner et al., 2005). The structure of System II in some ways is similar to System I. According to studies on *B. subtilis*, System II is formed by four proteins; CcsX, CcsB, CcsA, and CcdA (Figure 1.4). The CcsB and CcsA (cytochrome *c* synthase) are involved in the attachment of heme. In some species, *e.g. Helicobacter hepaticus*, CcsB and CcsA are fused to be a single trans-membrane protein (also known as CcsBA). CcsA is analogous in function to CcmF and has a conserved highly rich tryptophan region (WWD domain). The WWD domain has two flanking conserved

His residues, for which the amino acid side-chains work as axial ligands to the heme and protects it from oxidation. Mutation of either of these His residues results in loss of function (Kranz et al, 2009). CcsX is thioredoxin-like protein, which is found in the outer side of the cytoplasmic membrane. In addition to CcsX, CcdA is used to specifically reduce the two-cysteinyl residues in the apocytochrome *c*. Similar to cyanobacteria, mutation studies on *C.reinhardtii* have shown that cytochromes in the plastid are matured using System II (Howe and Merchant, 1992; Gabilly and Hamel, 2017). The plastids in eukaryotes contain up to three (known) *c*-type cytochromes: cytochrome *f*, cytochrome c_6 and cytochrome c_{6A}. The membrane-bound cytochrome *f* is a subunit within the b_6f complex, while cytochrome c_6 is soluble protein, which transfers electrons from cytochrome *f* to photosystem I. The function of cytochrome c_{6A} is less clear (Reyes-Sosa et al., 2011). Both cytochrome *f* and c_6 have typical CXXCH heme-binding motifs; System II matures cytochromes *f* and c_6 on the positive side of the thylakoid membrane (the thylakoid lumen) (Gabilly and Hamel, 2017).

1.1.3 System III

System III comprises a single polypeptide also known as holocytochrome *c* synthase (HCCS). According to genome sequence analysis, HCCS is only found in eukaryotic lineages, including apicomplexans, stramenopiles, fungi, aerobic amoebozoans (*e.g. Dictyostelium discoideum*, *Acanthamoeba castellanii*), animals and numerous green algae. Fungi and some protists have two HCCS isoforms (HCCS and HCC1S), which attach heme to either cytochrome *c* (HCCS) or c_1 (HCC_1S), respectively. In contrast, animals and some protists have only one isoform, known as holo-cytochrome *c* synthase, which recognises both mitochondrial *c*-type cytochromes as a substrate (Allen, 2011). Biochemical and experimental studies of human HCCS have revealed it has four structural domains: I, II, III and IV. Domains I and II are required for heme binding and interaction: a domain II (His154Ala) mutant results in a steady decrease of the level of the complex formation between HCCS and the heme. Domain III contains the mitochondrial targeting signal and thus is involved in targeting HCCS to the mitochondrial IMS. Domain IV maintains HCCS integrity and stability (Babbitt et al., 2014).

In System III, as studied in the human HCCS, the maturation of *c*-type cytochromes via HCCS occurs via a four-step reaction. The first step is the attachment of heme to HCCS enzyme via His154 (based on numbering of human HCCS). The His154 works as the first axial ligand for the heme. The second step is the attachment of apocytochrome *c* to HCCS where the His19

in the CXXCH heme binding motif of human apocytochrome *c* works as a second axial ligand to the heme. The third step is the formation of two thioester-bonds between the heme and two Cys residues in CXXCH heme binding motif in apocytochrome *c*. This disturbs the HCCS-heme-apocytochrome *c* complex triggering the release of HCCS enzyme and the formation of holocytochrome *c* in step four (Figure 1.4) (Mendez et al., 2017, Babbitt et al., 2016).

1.1.4 System IV

Organisms that use oxygenic photosynthesis have two *c*-type cytochrome maturation systems in their chloroplast(s). Whilst System II is used to covalently attach heme to cytochrome *f* and soluble cytochrome c_6 via two thioester bonds, System IV is used solely to covalently attach heme to cytochrome b_6 (of the b_6f complex) via a single thioester-bond. The maturation of cytochrome *f* and cytochrome c_6 occur on the positive (the luminal or p-) side of the thylakoid membrane, whereas the maturation of cytochrome b_6 is located on the negative (stromal or n-) side of the membrane (Lezhneva et al., 2008). System IV, which is also known as CCB (cofactor assembly complex C), requires at least four nuclear-encoded proteins. These proteins, CCB1, CCB2, CCB3, and CCB4, are integral membrane proteins that are sparse in their annotated motif or domain architectures (Saint-Marcoux et al., 2009). Presently, a tentative model is used to explain how the CCB system matures cytochrome b_6 (Figure 1.4). First, cytochrome b_6 forms a transit non-covalent bond to CCB1. Cytochrome b_6 is then transferred to CCB3 to form CCB3/ b_6 complex. This enables the recruitment of CCB2 and CCB4, which then leads to the formation of a tetramer complex (CCB4/CCB2/CCB3/ b_6). This complex recruits heme to form heme ligation complex. The final step is the dissociation of the complex and the release of holocytochrome b_6 and CBB proteins (Saint-Marcoux et al., 2009; de Vitry, 2011).

To date, all eukaryotes (except those from one phylum) that use cytochrome *c* as part of their energy production have either System I or System III as a cytochrome *c* maturation system – one protist (*Ancoracysta twista)* may be in possession of both maturation systems (Janouškovec et al., 2017). The phylum Euglenozoa is the exceptional group, which uses mitochondrial *c*-type cytochromes to produce energy, but no maturation system is evident from genome analysis. Molecular and spectroscopy studies have shown that the mitochondrial *c*-type cytochromes in this phylum are unique. The heme attaches to the mitochondrial cytochromes *c* via a single Cys in a heme binding motif (an A(A/G)QCH in

cytochrome *c* and FAPCH in cytochrome c_1) (Fülöp et al., 2009, Allen et al., 2004, Pettigrew at al, 1975, Priest et al, 1992). Members of the phylum Euglenozoa include the parasitic trypanosomatids, free-living phagotrophic and phototrophic euglenoids and marine flagellated diplonemids. Notably some euglenozoans, *e.g. Euglena gracilis*, are mixotrophs and therefore they can also use photosynthesis to produce energy. The *c*-type cytochromes in the chloroplasts of these algae have typical CXXCH heme-binding motifs, which require System II for the cytochrome maturation (Ebenezer et al, 2019).

Trypanosomatids are a family of exclusively parasitic microorganisms which have significant socio-economic impact, hence they are the most studied group of protists. The next few sections summarise the life cycles, mitochondrial form, and the origin of the trypanosomatids relevant to the experimental work described in Chapters 3 and 4.

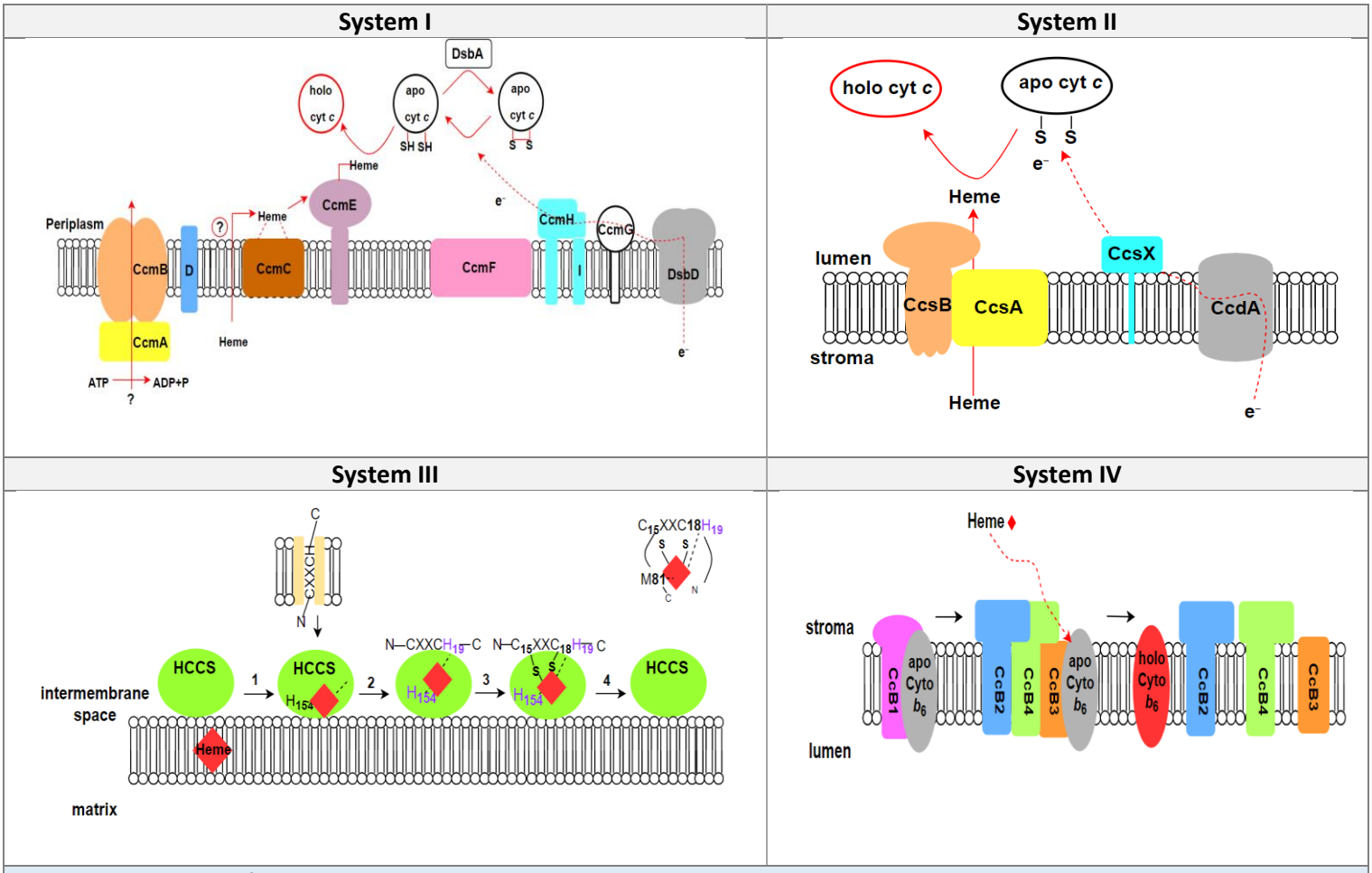


Figure 1.4. Overview of the cytochrome *c* maturation system pathways

The dashed arrows indicate the transfer of electrons or heme across the membrane. System I proteins are distributed across cytoplasmic membrane and periplasm in prokaryotes and mitochondrial inner membrane and IMS in eukaryotes. System II is found in cyanobacteria and viridiplantae and it is used to mature *f*-type and c_6-type cytochromes. System III operates in the IMS of mitochondria in most eukaryotes such as yeast, animals, invertebrates and some protists. System IV is found in cyanobacteria and viridiplantae and it is used to mature b_6 cytochromes.

1.2 Life cycles of parasitic Trypanosomatids

Trypanosoma brucei, *Leishmania* and *Phytomonas* species are examples of trypanosomatids that have dixenous life cycles.

1.2.1 Life cycle biology of *Trypanosoma brucei*

Trypanosoma brucei is the responsible agent for African trypanosomiasis, which is a fatal disease for both animal (nagana disease) and humans (sleeping sickness or human African trypanosomiasis (HAT)). Although medicines, surveillance, and diagnostics are able to control the periodic epidemics that arise, HAT is one of the most widely spread, potential fatal infectious diseases to be found in Africa; the parasite and its vector occur in 37 different sub-Saharan Africa countries (Baral, 2010). HAT can take two different forms depending on the causative agent. *Trypanosoma brucei gambiense* is responsible for a chronic infection with incidence spread across 24 countries in West and Central Africa. *T. b. gambiense* is responsible for the 97% of the reported cases of HAT (Cordon-Obras et al., 2015). Person can retain the parasites for several years before any symptoms are observed (Brun et al, 2010, Biteau et al., 2016). *Trypanosoma brucei rhodesiense* is responsible for an acute form of HAT where infection, to first presentation of symptoms, to death (in the absence of medical intervention) typically takes weeks to a few months. This form of the disease is detected in 13 countries across Eastern and Southern Africa. Sudden death can even occur after only a few days of the infection (Babokhov et al., 2013). Uganda is the only African country where cases of both human-infectious sub-species of *T. brucei* are recorded. During the 20th century, HAT was an epidemic disease where it was responsible for thousands of deaths each year (Goodhead et al., 2013). For these reasons, global efforts were intensified to decrease the number of HAT cases. According to the World Health Organization (WHO), the global target to eliminate HAT by 2020 is, in the view of some, within reach, as the number of cases reported dropped to just over 2000 cases in 2016. This vast drop in case numbers was attributed to the intensive surveillance and disease control activities (Franco et al., 2017, Hollingsworth, 2018).

The tsetse fly (*Glossina* genus) is the vector responsible for *T. brucei* transmission between mammals. Changes in the cell morphology, gene expression, position and the function of kinetoplast (mitochondria) can be clearly identified during different stages of the *T. brucei* life cycle. The parasite multiplies as procyclic trypomastigotes in the insect midgut which then migrate to its vector's salivary glands; this migration involves differentiation into

additional life cycle forms (Figure 1.5) and in the salivary glands epimastigotes ultimately differentiate into non-replicative infectious metacyclic trypomastigotes. The transmission occurs during the insect's bite which results in the transmission of metacyclic parasites into the dermis and bloodstream of the new mammalian host (Brun et al., 2010, Franco et al., 2017, Franco et al., 2014). Within the dermis and bloodstream, metacyclic parasites differentiate quickly into replicative long slender bloodstream trypomastigotes and remarkably evade the host's immune system. Central to this immune evasion, at low level antibody titres raised against the highly immunogenic variant surface glycoprotein (VSG) that completely enshrouds the parasite cell surface, long slender trypomastigotes internalise surface-bound antibodies at the flagellar pocket and degrade internalised antibodies within the lysosome. Directional movement of the antibodies across the parasite surface, induced by hydrodynamic flow, which in turn is dependent upon forward swimming motion, ensure swift removal of host antibodies. Antibody half-life on the cell surface is approximatelly 12.5 min (Engstler et al., 2007). This protective mechanism is not sufficient to protect parasites when the anti-VSG antibody titres increase beyond a threshold amount (Dean et al., 2007). However, by the time higher titres of antibodies are reached, some parasites are now expressing a different VSG, not previously seen by the host immune system (Pays and Steinert, 1988). The VSG is held in the outer leaflet of the parasite plasma membrane via a glycosylphosphatidylinositol (GPI) anchor. The switching of the VSG enables the parasite to escape recognition by antibodies (Wiederschain, 2009, Paulick and Bertozzi, 2008). There is an extensive repertoire of VSG genes found in the *T. brucei* genome; the repertoire is a combination of full-length, intact VSG genes and pseudogenes (VSG genes with one or more premature stop codons and/or other gene mutations or encoding proteins with atypical features). The variation of VSG can simply occur by switching to different full-length VSG genes or by recombination-driven mosaic VSG gene formation. This defined recombination-driven mosaic VSG formation means a near endless supply of new VSGs is theoretically possible (Matthews, 2005, McCulloch et al., 2017). Once the long-slender bloodstream trypanosome population reaches a critical density, parasites differentiate into non-dividing G1-arrested short stumpy form, which are pre-positioned for swift differentiation into procyclic trypomastigotes and thence infection in the tsetse midgut (Rico et al., 2013, Matthews et al., 2005). The procyclic trypomastigotes have a different surface protein to the VSG, still abundant in expressed copy number, it is one of four isoforms of procyclin (Roditi

and Liniger, 2002). The roles played by the different procyclin isoforms are not fully elucidated, but they play key roles in transmission competence (Vassella et al., 2009).

Bloodstream and procyclic *T. brucei* trypomastigotes have completely different metabolic profiles with regard to the ways in which each generate ATP. Thus, bloodstream *T. brucei* relies solely on glycolysis to produce ATP, but curiously the first seven glycolytic enzymes are targeted to peroxisomes (Figure 1.6). Approximately 90% of protein content within the matrix of peroxisomes in this life cycle stage is from glycolytic enzymes, giving rise to the re-naming of peroxisomes in trypanosomatids as glycosomes. In bloodstream *T. brucei* and presumably other trypanosomatids life cycle stages, placing glycolytic enzymes in a closed compartment behind the peroxisomal membrane is essential for growth and ATP production (Coley et al., 2011), albeit that in procyclic *T. brucei* and other trypanosomatids 3-phosphoglycerate kinase (PGK) tends to predominantly be a cytosolic, rather than glycosomal enzyme (Opperdoes and Borst, 1977; Michels et al., 2006). A summary of the glycolysis pathway as found in bloodstream *T. brucei* is summarised in Figure 1.6: the two ATPs consumed in the reactions catalysed by hexokinase and phosphofructokinase are balanced by the ATP generated from substrate level phosphorylation by PGK and net production of ATP occurs in the cytosol catalysed by pyruvate kinase.

To sustain NAD^+ concentrations in glycosomes and thence to sustain glycolytic flux, a glycerol-3-phosphate shunt operates between glycosomes and the mitochondrion. Oxidation of glycosomal NADH is achieved by the activity of NAD-dependent glycerol-3-phosphate dehydrogenase (GPDH) (enzyme 8 in Figure 1.6). Then, in the mitochondrial IMS, inner-mitochondrial membrane-bound FAD-dependent glycerol-3-phosphate dehydrogenase transfers electrons from G3P to ubiquinone generating DHAP, which diffuses or is translocated back in the glycosomal matrix. Electrons passed to ubiquinone are then used to directly reduce oxygen to water through the activity of an essential alternative oxidase Transiently, under hypoxic conditions, glycosomal NADH is re-oxidised via an alternative route which causes the production of glycerol (Pineda et al., 2018). The glycerol kinase is used to catalyse the conversion of G3P to glycerol, the glycerol is then excreted outside the cell. This reaction is highly controlled by the concentration of G3P and the ratio of ADP/ATP inside the glycosome. This hypoxic (or possibly anaerobic) oxidation of glycosomal NADH halves the potential ATP yield of glycolysis and bloodstream *T. brucei* can only survive for a few hours

before death (Hammond et al., 1985, Michels et al., 2006). Notably, all trypanosomatids have no known capacity for anaerobic growth (Van Weelden et al., 2003, Ginger, 2005).

In addition to several glycolytic enzymes and the glycerol kinase other enzymes of carbohydrate metabolism and enzymes from other metabolic processes – the pentose phosphate pathway; nucleotide salvage – are notably present in glycosomes of procyclic *T. brucei* and other trypanosomatids (Creek et al., 2015, Boitz et al., 2012, Michels et al., 2006). In contrast to bloodstream *T. brucei*, procyclic *T. brucei* utilises mitochondrial oxidative phosphorylation and carbon sources other than glucose to support ATP production (Bauer and Morris, 2017, Wargnies et al., 2018). This metabolic shift requires a marked change in mitochondrial activity and biogenesis. In contrast to the minimal (or repressed) mitochondrial profile of long slender bloodstream *T.* brucei, short stumpy bloodstream *T. brucei* cells exhibit some de-repression of mitochondrial activity, explaining why they are pre-adapted for cell differentiation and full mitochondrial activation in the low glucose environment of the tsetse fly midgut (Dewar et al., 2018). Once established in the tsetse mid-gut, procyclic *T. brucei* starts to rely on L-proline as a carbon source for ATP production (Lamour et al., 2005, Wargnies et al., 2018). Additional enzymes of carbohydrate metabolism present and prominent in the glycosomes of procyclic *T. brucei* include pyruvate phosphate dikinase (PPDK) and phosphoenolpyruvate carboxykinase (PEPCK). These enzymes are important for the production of glucose 6-phosphate (G6P) through gluconeogenesis (Wargnies et al., 2018).

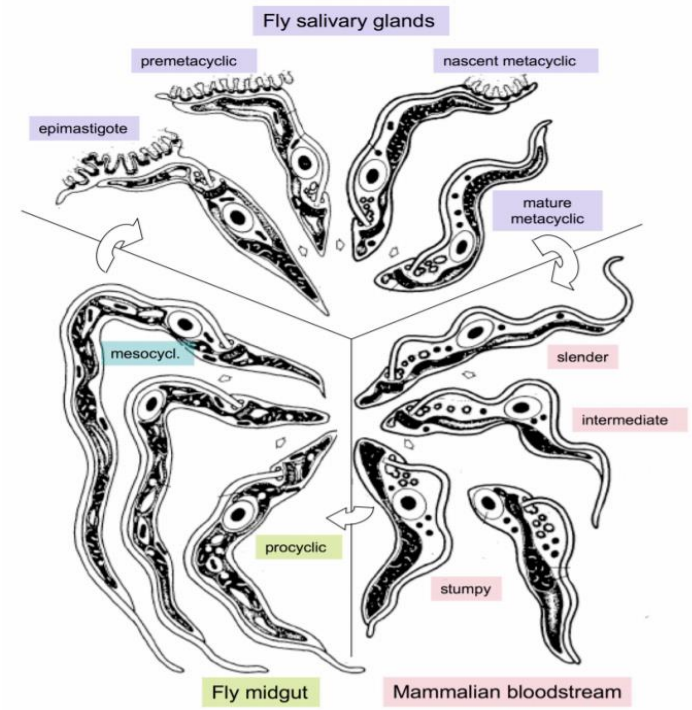


Figure 1.5 Different forms of *Trypanosome brucei* during its life cycle in tsetse fly and mammalian bloodstream

The diagram was reproduced with permission from Vickerman, K., (1985).

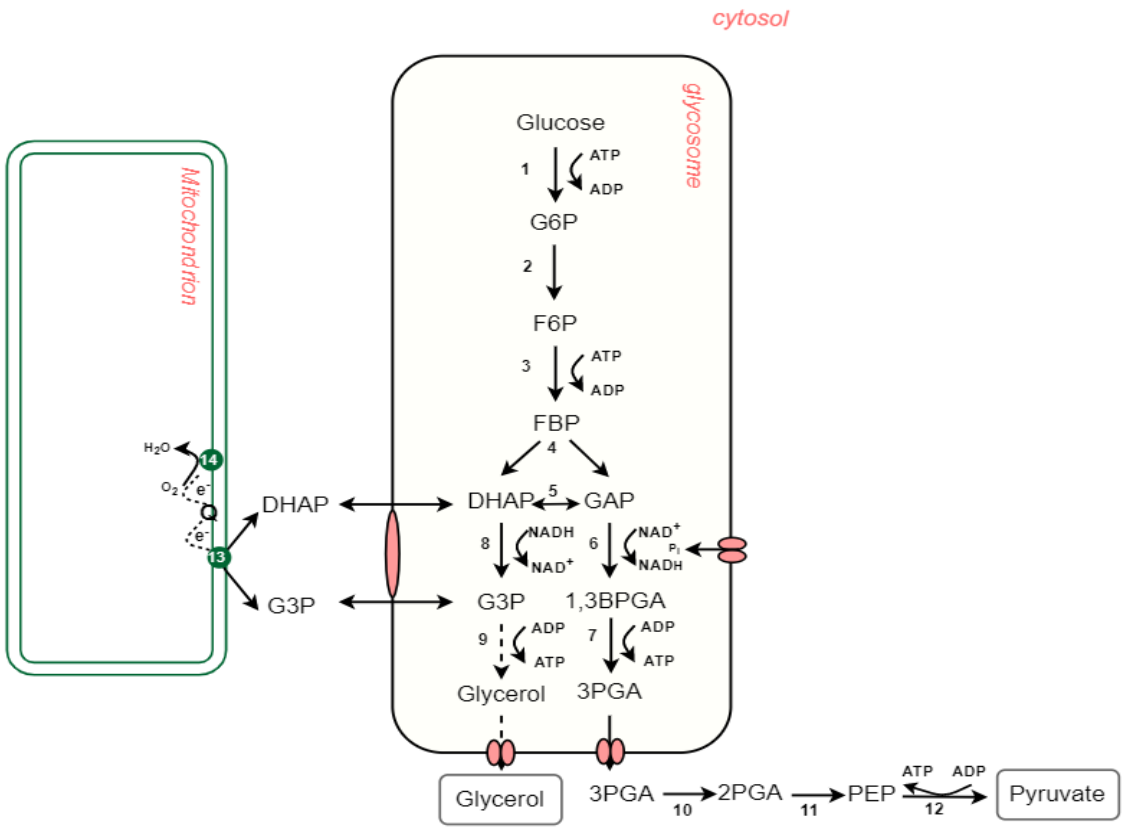


Figure 1.6. Glycolysis and ATP production in bloodstream form *T. brucei*

During aerobic conditions, the end product of glucose metabolism is pyruvate. The dashed arrow indicates that enzymatic activities are carried out under hypoxic or anaerobic conditions where the end product of glycolysis pathway is glycerol instead of pyruvate. The numbers represent different catalytic enzymes: 1, hexokinase; 2, glucose-6-phosphate isomerase; 3, phosphofructokinase; 4, aldolase; 5, triosephosphate isomerase; 6, glyceraldehyde-3-phosphate dehydrogenase; 7, phosphoglycerate kinase; 8, NAD-dependent glycerol-3-phosphate dehydrogenase; 9, glycerol kinase; 10, phosphoglycerate mutase; 11, phosphopyruvate hydratase; 12, pyruvate kinase; 13, FAD-dependent glycerol-3-phosphate dehydrogenase; 14, alternative oxidase.

Abbreviations: G-6-P, glucose 6-phosphate; F-6-P,fructose 6-phosphate; FBP, fructose 1,6-bisphosphate; DHAP, dihydroxyacetone phosphate; GAP, glyceraldehyde 3-phosphate; G-3-P, glycerol 3-phosphate; 1,3BPGA, 1,3-bisphosphoglycerate; 3-PGA, 3-phosphoglycerate; 2PGA, 2-phosphoglycerate; PEP, phosphoenolpyruvate; Pi, inorganic phosphate; Q, ubiquinone.

1.2.2 Life cycle biology of *Leishmania*

Human pathogenic *Leishmania* species cause a diverse spectrum of pathologies known collectively as leishmaniasis. Various forms of leishmaniasis are endemic through the majority of South and Central America, Africa, and Asia. There are an estimated 350 million people at risk of disease and records indicate that 70,000 deaths per year are caused by leishmaniasis disease (Reithinger et al., 2007). Leishmaniasis can be split broadly into three different forms

depending upon the identity of the infecting *Leishmania* species and on the immunity of the host: cutaneous leishmaniasis, mucocutaneous leishmaniasis, and visceral leishmaniasis (also known as kala-azar). Visceral leishmaniasis is considered the most severe form of the disease and is typically fatal if not treated. The disease is carried and transmitted via female sand flies (Phlebotominae subfamily) (Torres-Guerrero et al., 2017, Reithinger et al., 2007, Bravo and Sanchez, 2003). There are six genera and more than a thousand species which belong to Phlebotominae subfamily. Among them, female species of two genera, *Lutzomyia* (found in the New World) and *Phlebotomus* (found in the Old World), are known to be vectors for *Leishmania* species that are pathogenic for humans (Young and Lawyer, 1987, Berenger and Parola, 2017, Dostálová and Volf, 2012). *Leishmania* undergoes major morphogenetic changes between the mammalian host and the sand fly vector. In the mammalian host, an infected sand fly injects metacyclic promastigotes (infective forms) during feeding. The metacyclic promastigotes are promptly phagocytised by phagocytes where they differentiate from motile metacyclic promastigote with a long flagellum into immotile amastigotes with a short flagellum and more spherical cell body (Cuvillier et al., 2003). The shape of amastigotes might help them to minimise the surface area exposed to the hostile oxidative environment found in the parasitophorous vacuole (Mittra et al., 2017, Antoine et al., 1990, Sunter and Gull, 2017). Amastigotes are obligated to live inside phagocytes and replicate by simple division in the parasitophorous vacuole (Esch and Petersen, 2013). Ultimately phagocytes become overburdened with amastigotes which cause the cells to rupture and release amastigotes. The extracellular amastigotes re-infect further phagocytes causing subsequent cycles of infection (Walker et al., 2014). When the sand fly takes another blood meal, amastigotes differentiate and divide into procyclic promastigotes in the fly's midgut (Van Assche et al., 2011). The procyclic promastigotes undergo a series of differentiation stages before they eventually migrate to the salivary glands and transform into infectious metacyclic promastigotes where they become ready to infect another mammalian host (Sunter and Gull, 2017; Sacks and Noben-Trauth, 2002).

1.2.3 Life cycle biology of *Phytomonas*

Phytomonas was first isolated from *Euphorbia pilulifera* tissue back in 1909 and it was classified as a plant pathogen, which belongs to trypanosomatids (Donovan, 1909). One of the most distinctive features of *Phytomonas* sp. is the absence of some metabolic genes,

including cytochrome *c* and mitochondrial respiratory chain complexes III and IV, which are otherwise found in human pathogenic trypanosomatids. The loss of these genes is likely a consequence of the ecological niche occupied by these trypanosomatids. The phytophagous insects, which transmit *Phytomonas*, feed on rich carbohydrate plant sap. Accordingly, *Phytomonas* relies heavily on carbohydrate metabolism (Porcel et al., 2014, Sanchez-Moreno et al., 1992, Nawathean and Maslov, 2000). In fact, *Phytomonas* is one of the few aerobic eukaryotes known to be able to grow normally without heme: *b*-type and *c*-type cytochromes found in many eukaryotes are all absent from *Phytomonas* genomes (Kořený et al., 2012). With the lack of mitochondrial *c*- type cytochromes and with the genome sequence at hand, *Phytomonas* is an ideal model for the comparative genome sift versus *Trypanosome*/*Leishmania* to leave a pool of candidate genes from which a cryptic trypanosomatid cytochrome *c* maturation system might be sought.

Only two *Phytomonas* species are known to cause acute, deadly diseases in plants; *Phytomonas staheli*, which causes fatal wilt to coconut palm and slow wilt to oil palm, and *Phytomonas leptovasorum* which is known to cause phloem necrosis in Liberica and Arabica coffee plants (Parthasarathy et al., 1976, Di Lucca et al., 2013). Some countries (*e.g.* Brazil, Colombia and Suriname) rely heavily on exports of coffee beans and/or palm oil. Plant pathogenic *Phytomonas* species therefore pose significant economic risk to some countries (Jaskowska et al., 2015, Porcel et al., 2014). *Phytomonas* has been isolated from both extracellular and intracellular plant environments, including seeds, phloem and latex ducts. Typically, *Phytomonas* is isolated from hosts in a promastigote form, with an elongated cell body and single flagellum. The life cycle of *Phytomonas* is divided between plant and phytophagous insect, the exact infective stages of *Phytomonas* is poorly understood but there are some suggestions that *Phytomonas* is the only trypanosomatid with a digenetic life cycle that retains the promastigote form in both hosts (da Silva et al., 2013; Catarino et al., 2001, Vickerman and Preston, 1976, H Lopes et al., 2010, Jaskowska et al., 2015). After the insect vector feeds on infected plant fluid, the flagellated parasites increase in size in the insect digestive tract, but they do not begin division until they migrate to the salivary glands. In salivary glands, the flagellates return to its 'normal' size and then begin replication. The flagellates are transmitted to another plant during feeding of the infected insect (Dollet, 1984, Jaskowska et al., 2015).

1.3 Trypanosomatid evolution

In the 20th century, trypanosomatids were classified according to cell morphology, life cycle and host specificity. Nowadays, their classification is based on genome sequence, phylogenetic analysis and other biological characterisation (Hoare and Wallace, 1966, Kaufer et al., 2017). Three genes, 18S rRNA, 70 kDa heat-shock protein and glycosomal Glyceraldehyde 3-phosphate dehydrogenase (gGAPDH), are currently used for phylogenetic classification (Hutchinson and Stevens, 2018, Fraga et al., 2016, Qiu et al., 2019). The Trypanosomatidae family is currently known to have six subfamilies and 22 genera (Figure 1.7) (Maslov et al., 2018). Within the six subfamilies some have monoxenous life cycles and they only exist in insects (*e.g.* species from the genera *Borovskiya*, *Crithidia* and *Leptomonas*) and others (*e.g. Leishmania*, *Paraleishmania* and *Trypanosome*) have dixenous life cycles whereby the parasite undergoes patterns of complex life cycle biology in two hosts, typically an invertebrate 'vector' and either a vertebrate or plant host (Simpson et al., 2006). There is a long-standing debate that revolves around the evolution of monoxenous and dixenous life cycles. Today it is accepted that the dixenous life cycles were established and evolved independently from ancestrally monoxenous trypanosomatids. However, any dynamic interplay between these two life cycle modes or how the dixenous state arises remains poorly understood (Lukeš et al., 2014, Maslov et al., 2018).

More widely, trypanosomatids belong to the phylum Euglenozoa, which in turn are one of the small range of phyla that are collectively grouped into a eukaryotic supergroup known as Excavata. It is generally accepted that the last common ancestor of Excavata diverged at an early point in eukaryotic evolution. This could partially explain the wide differences in the cellular morphology of trypanosomatids and their often extreme biological characteristics relative to many other eukaryotes (Cavalier-Smith, 2009, Cavalier-Smith et al., 2014). Notable examples of extreme trypanosomatid biology are genome organisation, genome transcription, mRNA modification, Kinetoplast and RNA editing.

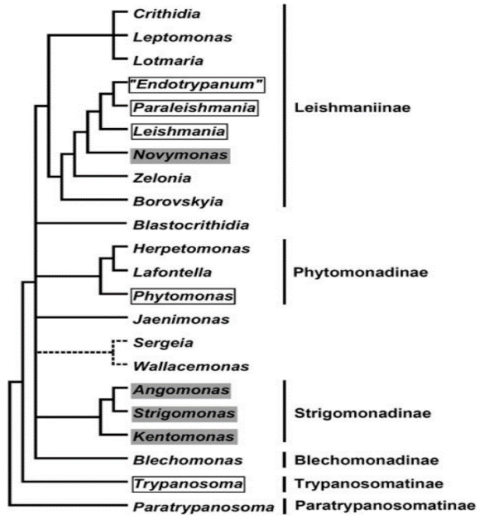


Figure 1.7. A phylogenetic tree showing the sub-families and the genera of Trypanosomatidae family
The dashed lines indicate that the phylogenetic position of the genera remains uncertain. The diagram was reproduced with permission from Maslov et al (2018).

1.4 Genome organisation and transcription

1.4.1 Kinetoplast and RNA editing

Trypanosomatids are part of the phylum Euglenozoa, but belong to the order Kinetoplastida. Kinetoplastids are defined by the presence of a kinetoplast, which is the name given to the highly unusual mitochondrial genome architecture found in these protists. Thus, the mitochondrial genome in trypanosomatids and their free-living kinetoplastid relatives is also known as kinetoplast DNA (or kDNA) (Parada, 2010). kDNA is composed of several thousands minicircle DNA molecules and 30-40 maxicircle DNA molecules. The size of maxicircle DNA varies depending on the taxon, which ranges from 20kb to 40kb. Maxicircles contain the genes encoding for a few well-known mitochondrial proteins such as cytochrome *c* oxidase (COXI, II, III subunits), ATP synthase, NADH dehydrogenase (ND1, 3, 5 and 7) and cytochrome *b* subunit of cytochrome bc_1 complex – in many other eukaryotes the genes encoding these proteins are also encoded in mitochondrial genomes, albeit mitochondrial genomes that are more conventionally-organised than kDNA. In addition to protein-coding genes of known function, kDNA also encodes a few open reading frames where the putative protein products are of unknown function. The mitochondrial ribosomal rRNA (12S and 9S) are also found in the genome sequence of maxicircles (De la Cruz et al., 1984, Gluenz et al., 2011). In contrast

to maxicircles, minicircles are smaller size, they normally vary in size from 1 to 2.5 kb. Minicircles encode for several hundred species of guide RNA (gRNA) which are used for RNA editing of mitochondrial genes. There are several forms of RNA editing that occur in a variety of eukaryotes; the unusual form of RNA editing that occurs in trypanosomatids results in insertion and/or deletion of uridine monophosphate into nascent mitochondrial gene transcripts. In this way nonsensical genes are decoded into translatable mRNA species. Trypanosomatid mitochondrial mRNA molecules are often edited by a number of gRNA molecules, with the editing process occurring 3' to 5' along the gene (Parada, 2010). In trypanosomatids, but not free-living kinetoplastids, minicircles and maxicircles form a catenated network of DNA molecules that forms a single tubular network disk-like morphology. *In situ*, the kinetoplast is 100 nm thick and about 650 nm in diameter and it is physically attached to an extra-mitochondrial cytoskeletal structure, the flagellar basal body, explaining the precise positioning of the kinetoplast in trypanosomatid cells (Ogbadoyi et al., 2003). Replication of kinetoplast occurs once per trypanosomatid cell cycle with the attachment of kinetoplast to the flagellar body ensuring that each daughter cell inherits one intact unit of mitochondrial genome post-cytokinesis (Gluenz et al., 2011, Li et al., 2007). The evolutionary origin of the Kinetoplast and the rationale behind its existence remains unclear (Lukeš et al., 2005; Lukeš et al., 2002).

1.5 Peroxisomes and targeting of peroxisomal proteins

Peroxisomes are single membrane-bound organelles found in many eukaryotes. Classically, peroxisomes are associated with metabolism of long-chain fatty acids and degradation of reactive oxygen species (*e.g.* metabolism of hydrogen peroxides by catalase, a classic peroxisome marker enzyme) (Poirier et al., 2006; Wanders et al., 2016). Peroxisomes in some eukaryotes carry out a wide variety of other metabolic pathways. In higher plants, peroxisomes can have various functions in different cell types. Hence, they can be subdivided into distinct classes: leaf peroxisomes, cotyledonary peroxisomes and glyoxysomes. For each class, a unique set of enzymes can be found with a particular role inside the cell. For example, glyoxysomes are found in the lipid-storing tissue of germinating seeds and they contain enzymes required for the glyoxylate cycle. The glyoxylate cycle is anaerobic pathway which centres around the conversion of lipids into carbohydrates in an environment where glucose is not available (Kanai et al., 2010, Gabaldón, 2010). In trypanosomatids, peroxisomes are

heavily repurposed towards supporting glycolytic flux and other aspects of carbohydrate and intermediary metabolism, hence they are known as glycosomes (Section 1.2 the life cycle biology of *T. brucei*).

Although the roles of peroxisomes inside cells can vary, there is conservation in the motifs that target proteins to peroxisomes. Generally, peroxisomal matrix proteins are synthesised in the cytoplasm by free ribosomes and are post-translationally targeted into peroxisomes. These proteins have a specific amino acid sequence (targeting signal) at their C-terminal (PTS1) or N-terminal (PTS2), which can be recognised by targeting receptors known as peroxins (Pex). The majority of peroxisomal matrix proteins have PTS1 signal and are recognised and targeted to peroxisomes via Pex5 receptor. Less peroxisomal matrix proteins have PTS2 signal and are recognised and targeted to the peroxisomes via Pex7 receptor (Kim and Hettema, 2015). Once peroxisomal matrix proteins are recognised by either Pex5 or Pex7 receptors, they are targeted to the peroxisomes where the Pex-loaded receptors dock onto the docking complex in the peroxisomal membrane. The docking complex in plants and mammals consists of Pex13 and Pex14, although in yeasts it is also composed of an additional protein, Pex17 (Figure 1.8) (Wang et al., 2019). The Pex receptors interact with Pex14 protein and the peroxisomal matrix proteins are released into the peroxisomal lumen (Schell-Steven et al., 2005). The receptors are recycled to the cytoplasm via receptor recycling complex which contains ubiquitin ligase complex (consists of Pex2, Pex10 and Pex12 proteins) and ubiquitin conjugation complex (consists of Pex4 and Pex22 proteins). In yeast, the receptor recycling complex is associated with docking complex via Pex8 protein. The receptor recycling complex monoubiquitinate the Pex5 receptor, which is then extracted from the peroxisomal membrane via the activity of the extraction complex (consists of Pex1, Pex6 and Pex15) (Farré et al., 2019). The exact mechanism of how the Pex7 receptor is recycled to the cytoplasm is less clear (Brown and Baker, 2008, Smith and Aitchison, 2009, Cross et al., 2016).

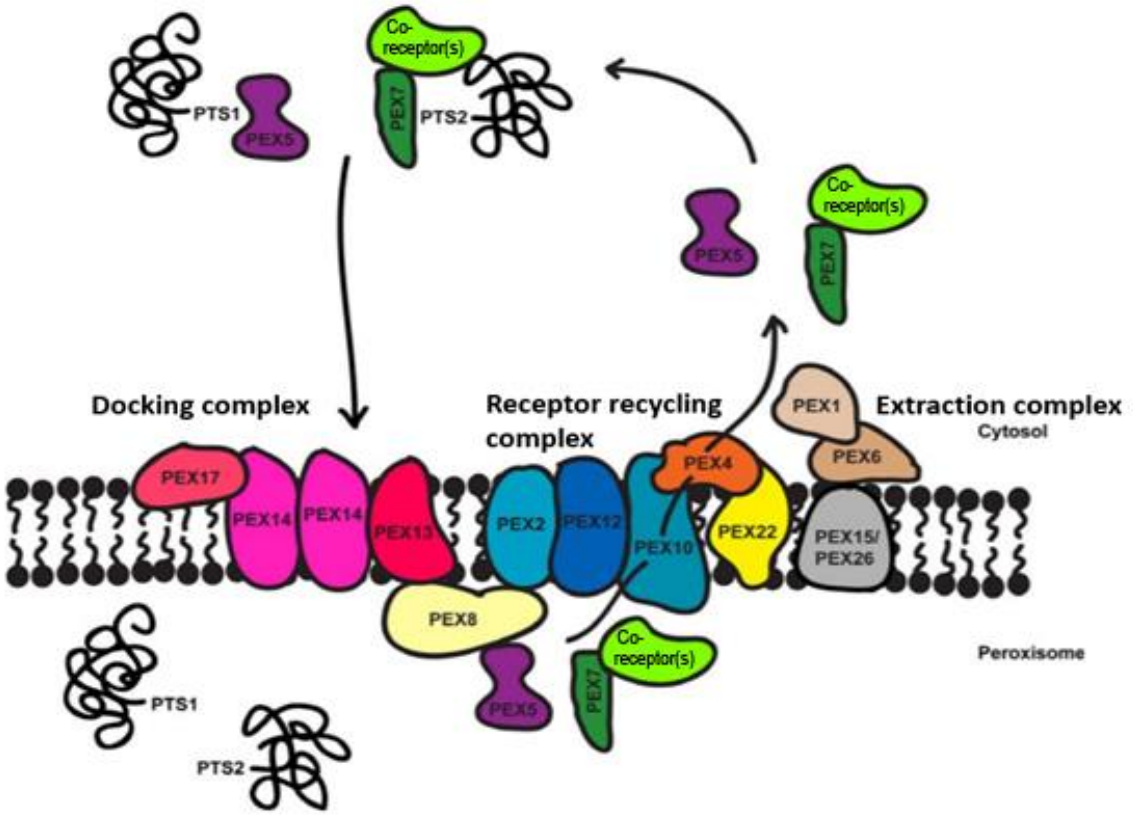


Figure 1.8. Outline of how peroxisomal matrix proteins import to the peroxisomal lumen
The diagram was reproduced with permission from Brown and Baker (2008).

1.6 Metabolic diversity among heterotrophic protists

In the presence of O_2, most eukaryotes use aerobic respiration to produce energy, and this involves four stages: glycolysis, pyruvate decarboxylation, Krebs cycle, and ETC (Lodish et al., 2000). However, a number of unicellular eukaryotes will have to perform an alternative route to produce energy under different environments. In a nutrient-rich environment, some eukaryotic cells rely only on the glycolysis pathway for energy production. These eukaryotic cells preform different metabolism reactions to oxidise NAD^+ molecule, which is needed to sustain the glycolysis pathway. One of these reactions is the fermentation of pyruvate to lactic acid, which regenerates NAD^+. This reaction can be observed in the protozoan malaria parasite during the merozoites stage (Elliott et al., 2001; Oyelade et al., 2016). Under low levels of oxygen (*e.g.* anaerobic and microaerophilic environments), some unicellular eukaryotes use various metabolic strategies to survive. For example, *Entamoeba* and *Giardia* use acetyl CoA synthetase [ADP forming] (ACD) to catalyse the conversion of acetyl CoA to acetate (Jones and Ingram-Smith, 2014). This reaction results in ATP production. *Naegleria*

genus is a rare example of a protist species suggested to have the ability of both aerobic and anaerobic respiration. This was revealed using bioinformatics analysis of *Naegleria gruberi* genomic DNA, which was first analysed in 2010 (Fritz-Laylin et al., 2010, Ginger et al., 2010, Opperdoes et al., 2011).

1.6.1 *Naegleria gruberi* metabolism

Species of the genus *Naegleria* are amoebo-flagellates and collectively these species are ubiquitous in freshwater and soil globally. There are 47 different species of *Naegleria*, each predating on bacteria (Majid et al., 2017, Schuster et al.,2004). Although some *Naegleria* species appear to have a capability for anaerobic metabolism, they are generally found growing in environments where oxygen is abundant and are considered to be obligate aerobes (Tsaousis et al., 2014, Fritz-Laylin et al., 2010). Recently, it was shown that in culture *N.* gruberi preferentially catabolises fatty acids for energy production, rather than glucose or other carbohydrates (Bexkens et al., 2018). Depending on the environmental conditions, *Naegleria* differentiates between three life cycle stages: replicative, predatory trophozoites; highly motile, non-replicative flagellates; and resting cysts. In nutrition-depleted conditions, trophozoites transform to flagellates, characterised by the presence of two flagella that together with nucleating basal bodies are assembled entirely *de novo* during the differentiation process. Flagellates swim around 100x faster than amoebae crawl; thus flagellate swimming allows cells relocate to more favourable environments whereon they transform back to the replicative trophozoite form (Heuser and Razavi, 1970, Johan, 2002). Despite the complexity of the transformation from amoebic trophozoites to swimming flagellates, *Naegleria* can complete this in a matter of 1-2 hours (De Jonckheere et al., 2001, Long et al., 2017) . In extreme environments where food scarcity is high or arid conditions, *N. gruberi* encyst (Fritz-Laylin et al., 2010).

The non-pathogenic *N. gruberi* is the most studied genus among *Naegleria* genera and it is recent used to give insight into the metabolic pathways that may be used by its most pathogenic relative *N. fowleri*. *N. fowleri* causes a fatal disease in human known as primary amoebic meningoencephalitis (PAM). Cases of this disease have been reported all over the world. Most commonly the pathogen enters the brain via the nasal cavity when swimming in warm freshwater (Hassan and Ali, 2018). There are very few experimental studies on the metabolic pathways of *Naegleria* genus despite the fact that many metabolic enzymes have

now been identified (Opperdoes et al., 2011, Ginger et al., 2010, Fritz-Laylin et al., 2010). However, it is well known that *N. gruberi* feeds on bacteria or yeast prey by phagocytosis. The presence of etherase and D-lactate dehydrogenase enzymes indicates that they are likely used in degrading the bacterial cell wall (Opperdoes et al., 2011). The etherase is used to hydrolyse the ether bond in the N-acetylmuramate 6-phosphate to produce N-acetylglucosamine 6-phosphate and D-lactate. The D-lactate is converted to pyruvate by the action of D-lactate dehydrogenase (Veiga-da-Cunha et al., 2009). *Naegleria* genome also has the sequence of the conventional metabolic enzymes, which include fructokinase/glucokinase, ribokinase and carbohydrate kinase. The presence of these enzyme sequences indicates that *Naegleria* can presumably catabolise a variety of monosaccharides as an energy source, despite recent reference to the importance of lipids as the main energy source for *N. gruberi*. Furthermore, the genome dataset indicates that a complete set of glycolytic enzymes is present in *Naegleria*. However, there are slight differences between glycolytic enzymes found in *Naeglaria* and other eukaryotes. One of these differences is the absence of hexokinase and the presence of a putative fructokinase/glucokinase (Milanes et al., 2019). The glucokinase sequence is similar to the one encoded by trypanosomatids and *Trichomonas vaginalis* all of which are closely related to bacterial glucokinase (Henze et al., 2001, Cáceres et al., 2007). Another variability is the replacement of ATP-dependent phosphofructokinase (PFK) by pyrophosphate-(PPi) dependent PFK. The latter is used to catalyse the second phosphorylation step in the glycolytic pathway (Mertens et al., 1993, Johan, 2002). *Naegleria* also encodes for two different enzymes to catalyse the last step in the glycolysis pathway. In addition to pyruvate kinase, which is used to catalyse the conversion of phosphoenolpyruvate to pyruvate, *Naegleria* encodes for pyruvate-phosphate dikinase (PPDK). One of the differences between pyruvate kinase and PPDK is the utilisation of AMP and PPi instead of ADP as a phosphoryl acceptor. PPDK can also catalyse the reverse reaction of phosphoenolpyruvate to pyruvate. Studies have indicated that the presence of PPDK and pyrophosphate- (PPi-)dependent PFK in *Naegleria* might be used for gluconeogenesis (Opperdoes et al., 2011, Fritz-Laylin et al., 2010).

Alongside the conventional aerobic metabolic pathways, *Naegleria* also has a number of anaerobic enzymes which provides strong evidence for anaerobic respiration abilities. Therefore, under particular anaerobic conditions *Naegleria* is still able to produce energy

(Wellings, 1979, Marciano-Cabral, 1988, Ginger et al., 2010). The metabolic flexibility of *Naegleria* species makes them a great candidate for studying the origin of metabolic versatility in early divergent eukaryotes (Fritz-Laylin et al., 2010).

1.7 Thesis overview

The primary objective in my research was to identify the mechanism by which trypanosomatids mature their mitochondrial cytochromes *c* and c_1. However, other aspects of metabolic variation in unicellular eukaryotes were also examined. Data collected during my research work are reported in four results chapters. In Chapter 3, I focus on the distribution and possible variation of known cytochrome *c* maturation systems in eukaryotes. Here, my objectives were to understand the types of motifs that might be essential for heme attachment in the cryptic trypanosomatid cytochrome *c* maturation system. In Chapter 4, I present the laboratory work which was used to test the candidature of a divergent-looking System III in trypanosomatids. In Chapters 5 and 6, the project was directed to study the adaptation of metabolic pathways in different eukaryotes. In Chapter 5, I outline a bioinformatics re-investigation of the existence of genes for some enzymes of anaerobic ATP production in *Naegleria gruberi*. Finally, in Chapter 6 I describe my study to map, and in select instances verify, the distribution of cryptic peroxisomal targeting determinants on some glycolytic and other metabolic enzymes in a variety of protists.

Chapter 2

Materials and Methods

2.1 Bioinformatics tools

2.1.1 Bioinformatics retrieval of protein sequences

To identify protein homologues in a variety of unicellular organisms, amino acid sequences of target proteins were used as queries for BLASTp analyses against NCBI non-redundant databases. The BLASTp analyses were run each time against specific organism database. To identify divergent-looking protein homologues, short conserved regions of target proteins were additionally used to refine the queries in BLASTp. Proteins identified in this way with expected values < e^{-20} were selected for further analysis.

2.1.2 Bioinformatics retrieval of Trypanosomatids data

The TriTryp database (http://tritrypdb.org/tritrypdb/) was used to access and interrogate different genes within Trypanosomatid genomes.

2.1.3 Local BLAST analyses

The genome sequences for some of the unicellular eukaryotes analysed during this work were not available from NCBI databases. Therefore, primary (or 'raw') DNA sequences (or 'reads') for these species were acquired from several publicly available web servers (Table 2.1). Then, local BLAST analyses were used to retrieve protein homology in those species. Local BLASTx query libraries were built manually and contained target protein sequences. The following command line was used to set up the queries: makeblastdb –in name_of_sequence_file.txt –parse_seqids –dbtype prot –out name_of_databasedb. Sequence data for several protists dispersed across eukaryotic phylogeny were used as an individual database for local BLASTx. The following command line was used to run the local BLASTx: Blastx -db name_of_databasedb –query nuc_sequence.txt -out whatever_db_name_DPout.txt -outfmt "6 qseqid sseqid evalue". Sequences with an expected values < e^{-20} were selected for further analyses.

Table 2.1. Different databases used in BLAST analysis to search for protein homologues.			
Organism	**Database**	**Website**	**Citation**
Fonticula alba	Ensembl protist	https://protists.ensembl.org/index.html	Kersey et al., 2018
Cyanophora paradoxa	Cyanophora paradoxa Genome Project	http://cyanophora.rutgers.edu/cyanophora/	Price et al., 2012
Fern/ Lycophyta/ Chlorophyta/ Streptophytina/ Chlorokybophyta/ Charophyta	1KP project	www.onekp.com	Matasci et al., 2014
Naegleria fowleri	AmoebaDB	https://amoebadb.org/amoeba/	Aurrecoechea et al.,2010
Toxoplasma gondii	ToxoDB	https://toxodb.org/toxo/	Gajria et al., 2007
Parvularia atlantis	MULTICELLGENOME LAB	http://multicellgenome.com/meet-our-organisms/parvularia-atlantis	López-Escardó et al., 2018
Sphaerothecum destruens	MULTICELLGENOME LAB	http://multicellgenome.com/meet-our-organisms/sphaerotecum-destruens	Torruella et al., 2015
Chromosphaera perkinsii	MULTICELLGENOME LAB	http://multicellgenome.com/meet-our-organisms/chromosphaera-perkinsii	Grau-Bové et al., 2017
Paramecium tetraurelia	ParameciumDB	http://paramecium-archive.i2bc.paris-saclay.fr/	Arnaiz and Sperling, 2011
All other species	NCBI	https://www.ncbi.nlm.nih.gov/pubmed/	Sayers et al., 2011

2.1.4 Sequence alignments

Multiple Sequence Alignment was used to align amino acid and nucleotide sequences. MAFFT was used to create Pearson/FASTA format alignments which were then used to create phylogenetic trees. The alignments were constructed with the default settings. To look for specific, conserved functional motifs within a protein sequence, ClustalW alignments were built using CLC sequence viewer 7.7.1 (CLC Bio, Aarhus, Denmark, http://www.clcbio.com/). Resultant ClustalW alignments were then analysed manually for motifs. Annotated protein sequences were entered into the programme using NCBI accession number whereas proteins acquired from local BLAST were entered manually.

2.1.5 ExPASy server

Various DNA and RNA sequences were translated to protein sequences using ExPASy server. Standard genetic code was used for all unicellular eukaryotic nucleotide sequences except for ciliates. In ciliates, the TAA and TAG stop codons can be translated to glutamine amino acid instead of inducing translation termination (Swart et al., 2016). Thus, when ciliate nucleotide sequences were translated, the genetic code was specified to 'Ciliate, Dasycladacean and Hexamita nuclear code'.

2.1.6 Phylogenetic tree constructions

Multiple sequence alignments in FASTA format were initially built using MAFFT software followed by manual editing (Table 2.2). To optimise alignments, redundant sequences and vaguely aligned positions were removed, leaving a final alignment which was then used to derive phylogenies. Maximum likelihood trees were generated using RAxML maximum likelihood analysis (Stamatakis, 2014). Initially, however, MrBayes was used to derive phylogenies (Ronquist et al., 2012). This was done to determine the best protein substitution matrix for RAxML analysis. The parameters of RAxML analysis were as follows: WAG as the best protein substitution matrix, CAT model and Rapid bootstrap values. Scale bars on phylogenetic trees represented the average number of amino acid substitution per site. RAxML analyses were run using an online Cipres web server (Stamatakis et al., 2008) (Table 2.2). Figtree v1.4.3 server was used to visualise trees (http://tree.bio.ed.ac.uk/software/figtree/).

2.1.7 Open Reading Frame Finder

To identify Open reading frames (ORF) in DNA sequences, ORFfinder server from NCBI resources was used (Table 2.2). The DNA sequence in FASTA format (size up to 50kb) was entered as a query. The settings were set up with a minimal ORF length of 600 nucleotides and start codon of only ATG. The genetic code was optimised according to species' identity.

2.1.8 Genome analysis to detect cryptic peroxisomal targeting signal

To identify cryptic peroxisomal targeting signal type I (cryptic PTS1) generated via ribosomal read-through of a stop codon or alternative splicing, first, the gene locus of the metabolic enzymes within each species' genome was located. The downstream sequence of each of these genes was screened for a potential PTS1 codon. For ribosomal read-through, putative cryptic PTS1 must be located in-frame downstream of the stop codon and be followed immediately by another stop codon. For alternative splicing, it was decided that the putative PTS1 motifs should be located within the 900 nucleotides downstream of the stop codon. Here, if a PTS1 motif was detected, the sequence (containing the putative PTS1) should translate from 5` to 3` and be preceded by a potential intron. Introns were identified by the presence of a sequence flanked by GT at the 5` end and AG at the 3`end (Piovesan et al., 2015). If in silico splicing of potential introns put the codons of the PTS1 in frame with the annotated open reading frame, then this was considered as a potential cryptic PTS1. The identity of PTS1 sequence was determined by existing literature (Reumann, 2014; Lingner et al., 2011; Emanuelsson et al., 2003; Ma and Reumann, 2008).

2.2 Proteomics Tools

Following the acquisition of protein sequences, various servers were used to analyse different aspects of the protein structure. Below is a description of each server; summary details can be found in Table 2.2.

2.2.1 Protein architectures

Three online applications were used to form a number of predictions on protein structure. These were TMHMM Server v. 2.0, Simple Modular Architecture Research Tool (SMART) and InterPro databases. The predictions cover protein secondary structure, functional domains, transmembrane helices, and signal peptides. Protein sequences in FASTA format were used as a query for each server. The default settings were used to run each of the programmes.

2.2.2 Protein disorder prediction

To predict the protein structure disorder, two servers were used; importantly each tool used a different algorithm to predict protein disorder. The first server was IUPredA which identifies protein disorder regions from the primary protein structure based on the measured pairwise energy content. The second server was MFDp2 which identifies protein disorder regions based on solvent accessibility, residue flexibility and B-factors. Protein sequences were submitted using FASTA format for both servers (Mizianty et al, 2013, Mészáros et al., 2018).

Table 2.2. Summary of web servers used to predict different aspects of protein and DNA sequences.			
Server name	Function	Web addresses	Reference
TMHMM	Predicts transmembrane domains	http://www.cbs.dtu.dk/services/TMHMM/	(Krogh et al., 2001)
SMART	Protein secondary structure and signal peptides	http://smart.embl-heidelberg.de/	(Letunic and Bork, 2017)
InterPro	Protein secondary structure and signal peptides	https://www.ebi.ac.uk/interpro/	(Mitchell et al., 2018)
ORFfinder	Identify Open reading frames	https://www.ncbi.nlm.nih.gov/orffinder/	(Sayers et al., 2011)
IUPred	Protein disorder prediction	https://iupred2a.elte.hu/	(Mészáros et al., 2018)
MFDp2	Protein disorder prediction	http://biomine.cs.vcu.edu/servers/MFDp2/	(Mizianty et al, 2013)
MAFFT	Sequence alignment	https://www.ebi.ac.uk/Tools/msa/mafft/	(Katoh et al., 2013)
Cipres	Phylogenetic tree construction	http://www.phylo.org/	(Stamatakis et al., 2008)

2.3 General laboratory solutions and buffers

- TAE Buffer (ThermoFisher): 40 mM Tris, 20 mM acetic acid and 1 mM EDTA.
- Elution Buffer (EB) (New England Biolabs): 10 mM Tris at pH 8-9.
- DNA gel loading dye (ThermoFisher): 10 mM Tris-HCl (pH 7.6), 0.03 % bromophenol blue, 0.03 % xylene cyanol FF, 60 % glycerol 60 mM EDTA.
- Lysogeny broth (Melford labs): per 1 L, 10 g tryptone, 5 g yeast extract, and 10 g NaCl.
- SDS loading buffer: 0.5 M of phosphate-buffered saline [PBS], 800 µL of 2% SDS and 0.1 M of DTT.
- 5×Running Buffer: per 1 L, 15.1 g Tris-base, 94 g glycine, 5 g SDS.
- Transfer Buffer: per 1 L, 5.8 g Tris-base, 29.0 g glycine, 0.37 g SDS, 200 ml absolute methanol.
- PBS- Tween (PBS-T): 1xPBS and 0.05% v/v Tween-20
- Blocking Buffer for Western blots: 5% (w/v) milk in PBS-T.
- Blocking buffer for microscopy slides: 1% bovine serum albumin (BSA) in PBS- Tween.
- Lysis buffer (under native conditions): 20 mM Tris-HCl pH 8.0, 500mM NaCl, 0.02% Triton-X, 20 mM imidazole and 10% glycerol.
- Wash buffer (under native conditions): 20 mM Tris-HCl pH8.0, 300mM NaCl, 0.02% triton-X, 20mM imidazole, 10% glycerol.
- Elution buffer (under native conditions): 20 mM Tris-HCl pH 8.0, 300 mM NaCl, 0.02% triton-X, 500 mM imidazole, 10% glycerol.
- Amoeba saline stock solution 1: 2 mM Na_2HPO_4, 2 mM KH_2PO_4.
- Amoeba saline stock solution 2: 32 µM $MgSO_4.7H_2O$, 54 µM $CaCl_2.2H_2O$, 4 mM NaCl.
- ZMG buffer: 8 mM Na_2HPO_4, 132 mM NaCl, 0.09 mM $Ca(CH_3COO)_2$, pH 7.0, 0.5 mM $Mg(CH_3COO)_2$, 1% (w/v) glucose.
- DNA extraction buffer: 50 mM Tris, 100mM NaCl, 1mM EDTA.
- 1.6x Cytomix: 200 mM Na_2HPO_4, 70 mM NaH_2PO_4, 15 mM KCl, 150 mM HEPES pH 7.3.
- Solution I in pyridine hemochrome assays: 0.2 M NaOH, 40% (v/v) pyridine, 500 µM potassium ferricyanide.
- Solution II in pyridine hemochrome assays: 0.5 M sodium dithionite in 0.5 M NaOH.
- Heme stain: 30 mL of 6.3mM TMBZ in absolute methanol, 70 mL of 0.25M sodium acetate (pH 5.0).

- ATCC medium: 1034 Modified PYNFH medium: per 1 L, 10 g Peptone, 10 g Yeast Extract, 1 g Yeast nucleic acid (Ribonucleic Acid, Type VI from Torula Yeast), 15 mg Folic acid, 0.4 mL of Hemin stock solution, 20 mL Buffer solution (18.1 g KH_2PO_4 , 25.0 g Na_2HPO_4).

2.3.1 Antibiotic stock solutions

Antibiotic	Stock solution	Final working solution
Ampicillin	100 mg/mL	100 µg/mL
Kanamycin	50 mg/mL	50 µg/mL
Streptomycin (sulfate form)	50 mg/ml	30 ug/ml

2.4 Molecular Biology Methods

2.4.1 Primers

All primers were ordered from Eurofins Genomics. Primers were re-suspended using EB to a stock concentration of 100 µM and stored at -20° C. Primer sequences generated during this study can be found in Appendices 2 and 3.

2.4.2 High fidelity PCR (HFPCR)

For HFPCR, proofreading Phusion high fidelity DNA polymerase (Thermo Scientific) was used. PCR mixes were set up in total volumes of either 50 µL or 20 µL and contained the following: 50 ng of DNA template, 200 nM of forward and reverse primers, 1X Phusion HF Buffer, 10 mM dNTPs and 0.02 U/µL Phusion DNA Polymerase. The PCR protocol and the thermocycling parameters were carried out as described in the manufacturer's protocol. Phusion polymerase yields blunt-ended PCR amplicons. Thus, for further TA cloning, an additional step of adding overhangs at 3` ends in the DNA double-strand was required. A-overhangs were added by mixing equal volumes of the PCR product (after PCR purification) to 2xPCR Master Mix (ThermoFisher, which contains Taq DNA polymerase) and incubation for 30 min at 72 °C.

2.4.3 Agarose gel electrophoresis

Agarose ME medium EEO (Melford labs) was used to make 0.8 - 1% (w/v) agarose gels to analyse PCR and restriction digest products. 3 µL of 10,000X SYBR DNA gel stain (Invitrogen) was added to gels for DNA visualisation under the UV radiation. Before loading DNA into

agarose gels, gel loading dye (6x) was added to the DNA samples. Gels were electrophoresed using 1x TAE buffer. New England Biolabs DNA ladder was used as a molecular weight marker to size DNA fragments. Gels were typically subjected to electrophoresis for 45 min at 100 V before visualisation using a UV transilluminator. When required, DNA bands were extracted using a Thermo Scientific™ GeneJET™ PCR Purification Kit as detailed in the manufacturer's manual.

2.4.4 Plasmid Constructions

For routine cloning of PCR amplicons, pGEM®-T Easy Vector was used as described in the Promega Corporation protocol. For protein expression, two types of vectors were used. The first type had one cloning site, which was used to express a single protein (pET28a, pET-15b and pNUS vectors). The second type had two multiple cloning sites which were used for co-expression of two different proteins (pCDFDuet vector). Restriction maps for plasmids used in this study can be found in Appendix 1. The ligation reactions using these plasmids were set up as follows: 1 µl of 10x rapid ligation buffer, 6 U/µL of T4 DNA ligase (Sigma-Aldrich). The ratio of plasmid DNA to DNA insert was calculated using online ligation calculator (http://www.insilico.uni-duesseldorf.de/Lig_Input.html). The final volume for ligation reactions was 10 µl. Ligation reactions were then left at room temperature for 12 hrs.

2.4.5 DNA Dephosphorylation

Generally, two different restriction enzyme sites were routinely used for the plasmid sub-cloning. Partially, this was done to allow for directional cloning of the DNA insert to the allocated plasmid. However, in some instances, only one restriction enzyme site was used for the plasmid sub-cloning of insert into a plasmid (vector) backbone. If this was the case, an additional step of DNA dephosphorylation was carried out to prevent re-ligation of the linearized plasmid. This was done using Shrimp Alkaline Phosphatase (rSAP) (New England Biolabs). The activity of rSAP catalyses nonspecifically the dephosphorylation of 5` ends of the DNA. The reaction had a total volume of 20 µL and contained the following: 1 to 2 µg of plasmid DNA, 5 units of rSAP and 1x CutSmart buffer. The reaction was incubated at 37°C for 30 min. The reaction was halted by an extra incubation step at 65°C for 5 min.

2.4.6 Annealing oligonucleotides

The following protocol was used to anneal two single-stranded DNA with complementary sequences. The oligonucleotides were dissolved in a sterilised EB to a concentration of 10 pmol/µl. Then 10 µl of each of equimolar oligonucleotides was added to a PCR tube. To anneal two complementary sequences, the mixed oligonucleotides mixture was heated slowly to 95°C and then allowed to cool slowly to 25°C. To allow for precision during this process, a thermocycler was used. The thermal profile was step up to heat the sample to 95°C (one degree per second) for 2 min and then to cool it down to 25°C over 45 min. Samples were then held at 4°C and at this stage, the double stranded DNA was ready to use for subsequent cloning. The annealed oligonucleotides were cloned to pNUS vector for the heterologous expression in *Crithidia fasciculata*. The list of single-stranded oligonucleotides that were annealed can be found in Appendix 3.

2.4.7 Restriction endonuclease digestions

There were two types of restriction digestion reactions. The first type was for plasmid cloning purposes. These digestion reactions had a total volume of 40 µL and contained the following: 5 units of restriction enzyme (NEB), 1xNEBuffer and 2 to 3 µg of plasmid DNA. The reactions were incubated overnight at 37°C. The second type of digested was for diagnostic purposes. Here, reactions were carried out in a total volume of 20 µL and contained: 1 unit of restriction enzyme, 1xNEBuffer and 1 µg of plasmid DNA. The reactions were incubated for 2 hrs at 37°C.

2.4.8 Culturing *Escherichia coli*

E. coli colonies were grown in solid media using LB media supplied with 1.5% agar (Melford labs). Liquid *E. coli* cultures, derived from single colonies, were typically grown overnight in LB at 37 ^{o}C with shaking at 180 rpm. To ensure aeration of these cultures, non-baffled conical flasks contained LB at no more than 10% of the total flask capacity.

2.4.9 *E. coli* transformation

XL-1 Blue Subcloning-grade competent *E. coli* was used for cloning purposes. The transformation reaction set up as described in Stratagene manual (https://www.chem-agilent.com/pdf/strata/200249.pdf). For gene expression, either BL21(DE3) or Rosetta™(DE3) competent *E. coli* cells were used. The BL21(DE3) and XL1 Blue were purchased from Agilent Technologies. The Rosetta™ cells were purchased from Novagen. For each transformation,

100 µl of competent cells were mixed on ice with 10 to 30 ng of ligated plasmid. The cells were incubated on ice for 30 min before the heat shock at 42°C for 1 min. If the transformed plasmid provided a resistance against ampicillin, the cells were incubated on ice for 2 min then were spread on an agar plate containing ampicillin. If the transformed plasmid provided a resistance to either kanamycin A or streptomycin, 1 mL of LB was added to the cells after the heat shock. The cells were incubated at 37°C and 180 rpm for 1 hr then were spread on an agar plate containing either kanamycin or streptomycin. The plates were incubated overnight at 37°C. Colonies were picked next morning and were used to inoculate 3 mL of LB media containing the antibiotic. Cells were then incubated overnight again at 37°C and 180 rpm.

2.4.10 Isolation of plasmid DNA

The plasmid DNA was isolated from overnight *E. coli* cultures using the Miniprep New England Biolabs kit and following the procedure as per the manufacturer's protocol. Isolated plasmids were stored at -20 °C in EB.

2.4.11 DNA Sequencing

Isolated DNA plasmids, diluted to 100 ng/µl, were sent for Sanger sequencing at Source BioScience. Oligonucleotide primers used for the sequencing were diluted to 3.2 pmol/µl and were sent together with the plasmids.

2.4.12 RNA isolation

Total RNA was isolated from *N. gruberi* cell pellets using TRI-zol reagent (Sigma-Aldrich). The culture (≈2x10^7 cells taken from a 20 mm x 90 mm petri-dish covered with active trophozoites) was centrifuged for 3 min at 2,000 x *g*. The supernatant was removed, and the pellet re-suspended into TRI-zol reagent (1 mL per 10^7 cells). The mixture was homogenised by repeated pipetting. To ensure total separation of nucleoprotein complexes, cell lysates were kept at room temperature for 5 min. The phase separation started when chloroform was added to the sample (0.2 mL per 1 mL of TRI-zol reagent). The sample was vortexed for 15 sec and it was allowed to stand at room temperature for 10 min. The sample was centrifuged at 12,000 x *g* for 15 min at 4°C. The centrifugation separated the sample into three phases: a lower red organic phase (cell proteins content), interphase (cellular DNA content) and a colourless, clear top phase (total RNA content). The uppermost phase was transferred to a

new tube and 2-isopropanol was added (0.5 mL per 1 mL of TRI-zol reagent). The sample was mixed and then placed at room temperature for 10 min. To precipitate RNA, samples were centrifuged at 12,000 x *g* for 10 min at 4°C. The supernatant was removed, and the RNA pellet re-suspended and washed using 75% ethanol (1 mL per 1 mL of TRI reagent). The samples were centrifuged at 7,500 x *g* for 5 min at 4°C; the ethanol was completely removed, and the RNA pellet was solubilised using 30 to 50 µl of RNase-free water. The RNA samples were then stored at -80°C.

2.4.13 Complementary DNA synthesis

Preparation of complementary DNA (cDNA) from RNA samples was carried out using SuperScript IV First-Strand Synthesis kit (ThermoFisher) as per the manufacturer's protocol. Each cDNA sample was prepared using 5 µg RNA. At the primer annealing step, 2 µM of gene-specific reverse primer was used. To remove any RNA contamination, 1 µl of RNase H was added to the cDNA mixture; the cDNA was incubated following this addition for 30 min at 37°C. At this stage, the cDNA was ready for use as a template for PCR analysis.

2.4.14 Isolation of genomic DNA

The genomic DNA (gDNA) was isolated from *N. gruberi*, *C. fasciculata* and *C, owczarzaki*. 10 mL of culture media were used for subsequent isolation, which generally contain 5-10×10^6 cells mL^{-1}. Cells were recovered by centrifugation at 6,000 x *g* for 5 min. The media was removed, and the cells were re-suspended using 500 µL of DNA extraction buffer, 25 µL of 20% SDS and 0.2 mg/mL of proteinase K. The re-suspended cells were either incubated at 37°C overnight or 56°C for 2 hrs. To increase the purity of the isolated gDNA, 500 µL of phenol: chloroform: isoamyl alcohol (25:24:1) was added. The sample was inverted a few times until a cloudy solution developed. After that, the sample was centrifuged at 12,000 *x g* for 2 min. The clear, colourless top layer was then transferred to a fresh Eppendorf tube and, to precipitate gDNA, 1 mL of absolute ethanol was added. The sample was placed on ice for 10 min or until all DNA had precipitated. The gDNA was then recovered using sterile plastic inoculation loops and washed six times by dipping in 70% ethanol. gDNA was eluted into EB containing 1 µL of RNase H (to degrade co-purifying rRNA); re-suspended gDNA was incubated for 1 hr at 37°C before long term storage at 4°C.

2.5 Recombinant protein-related methods

2.5.1 Induction of protein expression

For small scale protein induction, 100 µl from a transformed overnight bacterial culture was transferred into 5 mL fresh LB media and was allowed to grow aerobically and reach the logarithmic phase of growth (OD_{600} 0.4 – 0.6). Isopropyl β-D-1- thiogalactopyranoside (IPTG) was added to a final concentration of 1 mM to the bacterial suspension in order to induce protein expression. Following induction, bacterial pellets were collected by centrifugation at 8000 x *g* for 2 min at 0, 1, and 3 hrs post-induction. For induction of protein expression at larger scale, 2 L baffled conical flasks were filled with 500 mL of LB media along with 50 mL of transformed bacterial culture. These cultures were left to reach the logarithmic phase of growth before addition of IPTG to a final concentration of 1 mM (from a 1 M stock solution) and antibiotics to which resistance was conferred by the presence of the expression plasmid(s). Bacterial cultures were then incubated overnight at 18°C and 100 rpm. Following the overnight incubation, bacterial pellets were collected by centrifugation at 2500 x *g* and 4 °C for 25 min.

2.5.2 Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE)

The protocol to prepare SDS-PAGE gels was taken from *Molecular Cloning: a Laboratory Manual* (Sambrook et al., 1989). For SDS-PAGE, generally 12% polyacrylamide resolving gels were used. The loading buffer for protein samples was prepared and heated at 100°C for 5 min prior to use. Then, 100 µL of boiling loading buffer was used to re-suspend bacterial pellets from small-scale protein inductions (1 ml of bacterial culture was harvested every 0, 1, 3 hrs after the addition of IPTG). Protein within these samples was denatured by incubation at 100°C for 5 min; 10 µl aliquots of denatured samples were loaded onto SDS-PAGE stacking gels along with 6 µL of PageRuler™ Plus Prestained protein ladder (Bio-Rad). Gels were typically electrophoresed for 1 hr at 200 V and then stained with 10 ml of Instant Blue (Ambicon) for 15 min to visualize protein bands.

2.5.3 Electroblotting Transfer

In case the SDS-PAGE gel was intended to be used for Western blot, electrophoresed proteins were transferred onto a nitrocellulose membrane, instead of staining with Instant Blue. Prior to transfer, nitrocellulose membranes were hydrated by immersion in methanol for 10 sec

before incubation in transfer buffer for 10 min. Along with each nitrocellulose membrane, two filter papers and two fibre pads were also incubated in the transfer buffer for 10 min. Transfer cassettes were prepared by placing one fibre pad followed by one piece of filter paper on the cassette's black side. The resolving gel was next recovered from the SDS-PAGE apparatus and placed on top of the filter paper, followed by the pre-incubated nitrocellulose membrane. Finally, on top of the membrane, the second filter paper and fibre pad were placed. The cassette holder was tightly closed and placed in the transfer tank together with an ice pack. The tank was filled with the transfer buffer and electroblotting was allowed to occur at 100 V for 60 min. Following protein transfer, membranes were either placed directly into blocking solution (for probing with primary and secondary antibodies) or dried on filter paper for storage at 4°C. If the membranes were dried, then an extra rehydration step was required prior to antibody probing; This was done by immersing the membrane in 100% ethanol for 10 sec prior to incubation in 1 x PBS for 10 min.

2.5.4 Antibody Detection of Western blots

To prevent non-specific antibody binding, membranes were incubated in 10 mL blocking buffer for 1 hr on a rotating platform. After this, the blocking buffer was replaced with 5 mL fresh blocking buffer containing anti-Myc primary antibodies (Abcam) at a dilution of 1 in 100. Membranes were incubated with primary antibodies for 1 hr on the rotating platform. Next, the primary antibodies were removed, and the membrane washed three times in 10 mL PBS-T; each wash was for 10 min. Membranes were then incubated in 5 mL fresh blocking buffer containing anti-mouse IgG HRP-linked secondary antibodies (Cell signalling) diluted to 1 in 1000 for 1 hr on the rotating platform. Secondary antibodies were then removed, and the membrane washed three times in 10 mL PBS-T (10 min per wash). For detected of the HRP-conjugated secondary antibodies, membranes were incubated for 2 min with chemiluminescent HRP substrate (prepared according to the manufacturer's protocol (SuperSignal West Pico Chemiluminescent substrate Thermo-Scientific)). Excess substrate was then drained, and the membrane was wrapped with Saran-wrap, placed in a film cassette and exposed in the dark room to X-ray film for 10 sec. Remaining in the darkroom, the X-ray film was removed from the film cassette and placed into 'developer' solution for few sec. The film was then rinsed briefly in MilliQ water and placed in 'fixer' solution until bands appeared on the film.

2.5.5 Protein solubility

To test the solubility of recombinant proteins, BugBuster (Novagene) was initially used in accordance with the manufacturer's protocol. Subsequently, the use of Bugbuster was abandoned, and an alternative method using ultrasonication was adapted to assess protein solubility.

Thus, following protein induction, bacterial pellet was re-suspended with lysis buffer. The lysis buffer was added as 20 ml buffer per litre of bacterial culture grown. Then, 10 µl of 100 mM PMSF protease inhibitor was added per 1 ml of bacterial resuspension. The suspended bacterial pellet was then left on a shaker at room temperature for 30 min. Following this, the lysed bacterial suspension was disturbed by ultra-sonication by bursts of 5 sec on/, and 15 sec off for a duration of 10 min at an amplitude of 85%. The bacterial suspension was next centrifuged at 20,000 rpm for 30 min at 4°C. The supernatant was transferred to fresh tubes and both pellet and the supernatant were stored at -20 °C until analysis by SDS-PAGE.

2.5.6 Purification of His-tagged proteins under non-denaturing conditions

To prepare Ni^{+2}-NTA columns, a small amount of glass wool was placed into a 20 ml syringe barrel. The end of the syringe was capped using parafilm. Then, 1 ml of Amintra Ni-NTA Resin was set into the syringe. Next, 2.5 ml of lysis buffer was added and the syringe was inverted six times to equilibrate the resin. After this, the excess of lysis buffer was removed and the supernatant solution, which was separated from the bacterial pellet after the ultrasonication, added into the column. Columns were placed in the tube rotator and incubated overnight at 70 rpm and 4°C. The next day, the flow-through was collected from columns and kept as a flow-through fraction. The column was then washed twice with 4 ml of wash buffer. To elute His_6-tagged recombinant protein, 8 ml of elution buffer was added to the column; eluted protein solution was collected as 1 ml fractions. All fractions were kept at -20°C and protein purity was assessed by SDS-PAGE.

2.5.7 Protein precipitation

To increase protein concentration, two different methods were used. The first method used acetone to precipitate protein as detailed in the Thermo Fisher Scientific protocol (https://tools.thermofisher.com/content/sfs/brochures/TR0049-Acetone-precipitation.pdf).

The second method involved the usage of vivaspin 20 centrifugal concentrators (MWCO 3 kDa) as per the manufacturer's protocol (Sigma-Aldrich).

2.6 Heme detection related methods

2.6.1 Heme staining (TMBZ) of SDS-PAGE gels

Heme staining was performed using 3,3',5,5'-tetramethylbenzidine (TMBZ) based on the protocol by Thomas et al. 1976. Here, partially purified recombinant cytochrome *c* samples were re-suspended using SDS-PAGE loading buffer without any reducing agents (*e.g.* DTT). Re-suspended protein samples were loaded on 12% gels for SDS-PAGE along with a control sample. The control sample for heme staining was reduced horse heart cytochrome *c* (Sigma-Aldrich). The resolving gels were recovered from the SDS-PAGE apparatus and were rinsed with water for 5 min. In the dark room, gels were placed in 100 mL of heme stain and were incubated for 2 hrs. Next, 495 µL of 30% hydrogen peroxide was added to the gels and staining was completed within 15 min.

2.6.2 Pyridine Hemochromagen Assay of heme attachment to trypanosomatid cytochrome *c* determined by UV-Vis spectroscopy

The absorption spectrum of the partially purified recombinant cytochrome *c* was carried out using pyridine hemochromagen as described by Barr and Guo (Barr and Guo, 2015). To measure the oxidized heme spectra, the heme-containing sample was mixed 1:1 with solution I (containing pyridine, which serves as a ligand for the heme) and potassium ferricyanide (an oxidising agent). To measure the reduced heme spectra, 10 µl of solution II (containing sodium dithionite, a reducing agent) was added to the sample. Heme spectra were recorded every min for 5 min until the absorbance peak was stable. For UV-Vis spectroscopy, protein samples were placed in disposable UV cuvettes with a pathlength of 1 cm. The UV-Vis absorption spectra were recorded at room temperature using Cary 4000 UV/Vis spectrophotometer.

2.7 Protist culture

2.7.1 *Capsaspora owczarzaki*

Capsaspora owczarzaki (strain ATCC 30864) was cultured axenically at 25°C in ATCC medium: 1034 Modified PYNFH medium. The details of the components of this medium can be found

in section 2.3 (General laboratory solutions and buffers). The cells were passed weekly to a cell density of 5×10^5 cells mL^{-1}.

2.7.2 *Naegleria gruberi*

The phagotrophic NEG strain of *Naegleria gruberi* culture was grown xenically on amoeba saline plates with *Klebsiella pneumoniae* provided as the prey source. Amoeba saline plates were prepared by mixing 2.5 ml of amoeba saline stock solution 1 with 2.5 ml of amoeba saline stock solution 2 (components can be found in section 2.3). To solidify the culture media, 1 to 2% of agar No. 2 was added. The solution was then autoclaved and poured to petri dishes using aseptic techniques. The *K. pneumoniae* culture was grown on LB agar at 25°C. Once the culture was established, *K. pneumoniae* cells were transferred to amoeba saline plates. The suspended *N. gruberi* cysts were pipetted in the middle of the plate. To culture *N. gruberi* under microaerophilic conditions, plates containing cells were placed in 2.5 L Oxoid™ AnaeroJar™ containers with gas generation compact sachets (Oxoid™ CampyGen™ 2.5L Sachet). The plates were kept at 25°C. The *N. gruberi* cysts were excysted and they were transformed into trophozoites cells within 48 hrs.

2.7.3 *Crithidia fasciculata*

Crithidia fasciculata choanomastigotes were grown axenically at 26°C in a Brain-Heart Infusion (BHI) broth supplemented with 10% (v/v) heat-inactivated serum and 20 µg/ml haemin. The cells were passed once every 48 hrs to a cell density of 10^6 cells mL^{-1}.

2.8 *Crithidia fasciculata* electroporation

For transfection of *C. fasciculata*, $\approx2\times10^8$ cells were used at the late exponential-phase of their growth. Cells were recovered by centrifugation at 800 x *g* at 4°C for 10 min. The culture media was removed and the cell pellet was washed twice with ice-cold culture media. After the final wash, the cell pellet was re-suspended at 5×10^7 cells ml^{-1} using ice-cold ZMG buffer. Then, 0.5 ml of the re-suspended cells were transferred to a 0.4 cm electroporation cuvette along with 40 to 80 µg of plasmid DNA. The cuvette was then placed in Bio-Rad Genepulser and subjected to double pulses at 2500 V, 25 µF capacitance and 200 ohms resistance. After electroporation, transformed cells were placed in 5 ml fresh media and allowed to recover for 5 hrs. After recovery, 200 µg/ml of Hygromycin B was added to the culture. Stable transformants were obtained approximately 7 days' post-electroporation. Stable transformants were

cryopreserved by mixing logarithmic cultures with glycerol to a final concentration of 10% and stored in 0.5 ml aliquots at -80°C. For recovery of cryopreserved *Crithidia*, thawed frozen aliquots were added to 4.5 ml fresh culture medium and incubated at 26 °C.

2.8.1 *Crithidia fasciculata* protein sample preparation

To confirm heterologous gene expression in transfected *C. fasciculata* cell lines, Western blot analyses to detect expressed protein was vital. For this analysis, the protein content from 2x 10^7 cells was required. After cells were recovered by centrifugation, the resultant cell pellet was collected, washed twice with PBS, and re-suspended at 2×10^9 cells ml^{-1} using boiling SDS-PAGE loading buffer. The solution was then heated at 100°C for 5 min. The protein samples were stored at -20°C prior to SDS-PAGE and Western blotting.

2.9 Indirect immunofluorescence

Once stable *Crithidia* transformants reach a generation time of 4-6 hrs, microscope slides were prepared at an early point to look for the fluorescence signal from eGFP reporter expression. To prepare slides, ≈5x10^6 *C. fasciculata* cells were collected by centrifugation at 2,000 *x g* for 2 min. The supernatant was removed, and the cell pellet was washed twice with 1 mL PBS. The pellet was then re-suspended using 0.2-0.4 mL of PBS depending on the size of the cell pellet. The re-suspended cells were pipetted to a hydrophobic wall drawn on an uncoated slide. Cells were allowed to settle for 10-20 min before fixation with 4% paraformaldehyde (in PBS) for 10 min. Next, the slides were placed in absolute methanol for at least 10 min at -20°C – this began permeablisation of the fixed cells. Subsequently, slides were rehydrated by placement into PBS (in a coplin jar) for 5 min. In the case where only the detection of the GFP signal was required, one drop of Vectashield mounting medium containing 4',6-diamidino-2-phenylindole (DAPI) was added to the hydrated cells. A coverslip was then added, excess Vectashield removed, and the slide-coverslip interface sealed with nail varnish. If, however, detection of an indirect immunofluorescence signal was also required (to co-localise eGFP signal with peroxisomal marker proteins), then additional steps were carried out.

To detect the indirect immunofluorescent signal, a combination of primary and secondary antibodies was used. Here, after rehydration in PBS, slides were incubated in 0.2% Triton X-100 for 10 min, and to further increase cell permeability, the slides were then incubated in

0.5% Tween for 10 min. To limit non-specific binding of primary antibodies, slides were placed in a humid chamber and hydrophobic wells were covered with blocking buffer and left, in the humid chamber, for 1 hr at room temperature. Then, the blocking buffer was removed and replaced with 1 mL anti-GAPDH primary antibodies diluted 1 in 100; the primary anti-GAPDH antibodies were kindly supplied by Dr Frédéric Bringaud from University of Bordeaux. The slides were incubated in the humid chamber for 1 hr before washing three times with PBS-T in a coplin jar (10 min per wash). After the last wash, 1 mL of Goat Anti-Rabbit IgG (Rhodamine conjugate) secondary antibodies diluted 1 in 200 was added. Blocking buffer was used to dilute both the primary and secondary antibodies. The slides were incubated for a further 1 hr before three final 10 min washes with PBS-T. Finally, Vectashield mounting medium with DAPI was added to wells and the slides were mounted and sealed with a coverslip. Imaging was done using the AiryScan detector module of Carl Zeiss Ltd LSM880 confocal laser-scanning microscope at 60× magnification; images were processed using associated Zen blue software. For the filters used and their wavelengths see Table 2.3.

2.10 *Naegleria gruberi* cell live imaging

Amebae saline agar plate was prepared, then *K. pneumoniae* was spread across the entire plate. *Naegleria* cysts solution was then pipetted in the middle of the plate. The plate was incubated for 48 hr at 25°C. At this stage, the cysts had transformed into trophozoites. Once these amoebae had reached the edge of their plate and the plate was completely covered with active trophozoites, 4 mL of sterilised amoeba saline solution was then used to recover the amoebae from the plate. Next, 3 mL of re-suspended amoebae were transferred into a 20 ml hydrophobic surface tissue culture flask with a plugged cap (STARLAB) and 0.6 µL of deep red mitotracker (Thermo Fisher) was added to the flask. The cells were incubated at room temperature for 30 min before they were placed into a 24-well cell culture plastic bottom plate. Live-cell time-lapse images were recorded every second at 40× magnification using AiryScan detector module of the Zeiss confocal microscope.

Table 2.3. Filters used in the confocal microscopy and the operating wavelength for each filter.

Filter	Excitation (nm)	Emission (nm)
DAPI	360	455
YGFP	500	535
Rhodamine	570	640
MitoTracker	644	665

2.11 CRISPR-Cas9

This technique of genome manipulation is based on naturally occurring endonuclease enzyme known as Cas9. Cas9 cuts DNA at a specific location using a short-guided RNA sequence. By taking advantage of the cell DNA repair machinery, a foreign piece of DNA can replace an existing DNA sequence. This particular technique was used to knock out the candidate holocytochrome-*c* synthase gene from *Leishmania mexicana* and to replace it with a drug-resistant cassette which can be used for cell selection (Beneke et al., 2017, Ishemgulova et al., 2018). This aspect of the work was carried out in collaboration with Dr Rachel Neish in the group of Prof Jeremy Mottram at the University of York.

2.11.1 CRISPR/Cas9 constructs

To introduce the double strand breaks within a DNA sequence, two single guide RNAs (sgRNA) were utilised. These sgRNAs were used to cut the DNA sequence at 5` and 3` of the untranslated region (UTR) of the target gene's ORF. For effective gene knockout, two drug resistant cassettes were prepared. One cassette provided resistance to blasticidin S (BSD) and the other resistance to puromycin (PAC). The drug resistant cassettes were introduced at the cut sites to replace the two copies of the target gene via homologous recombination. The PCR based technique was used to generate both sgRNA and the drug resistant cassettes. The primers used in the PCR were designed using leishgedit tool (http://www.leishgedit.net/Home.html). The gene ID was required by the leishgedit tool to design primers that can be used to knockout a specific gene. The primers were based on pT and pPLOT plasmids. Different regions within each primer are shown in Figure 2.1. Tables 2.4 and 2.5 show the exact components used in the PCR. Tables 2.6 and 2.7 show the PCR thermocycling parameters. To amplify the BSD repair cassette, a pGL2662 template DNA vector was used in PCR. Also to amplify the PAC repair cassette, a pGL2667 template vector was used in PCR. Both pGL2662 and pGL2667 were kindly provided by Dr Rachel Neish. The standard Rev primer (OL6137) was used as DNA template in PCR to amplify both sgRNAs. In total, four PCR products were made, two for sgRNAs and two for the cassettes. The four PCR products were combined in one tube and purified using PCR purification kit. The purified PCR products were sent to the University of York to carry out the gene knockout experiment.

2.11.2 Cells culture

The gene knockout in *L. mexicana* were carried out at the University of York by Dr Rachel Neish. The promastigote form of *L. mexicana* cells were grown in M199 medium which was supplied with 0.005% haemin, 2.2 g/L $NaHCO_3$, 20% fetal calf serum (FCS) and 40mM HEPES Buffer at pH 7.4. *L. mexicana* JM6159 cell line was used for the DNA transfection.

2.11.3 Transfection and selection

For each transfection, *L. mexicana* culture containing $1x10^7$ cells were centrifuged at 1,800 x *g* for 10 min at 4°C. The media was removed, and the cell pellet was re-suspended with 126 µl of 1.6x Cytomix. The sterilised DNA constructs (sgRNA and drug resistance cassettes) were added to the cells. The cells were transferred to 2 mm cuvette and were electroporated by a single pulse using Nucleofector 2B (Program X-001). Then, the cells were transferred to 10 ml media and were allowed to recover for 24 hrs. After recovery, selective antibiotics were added to the media. A serial dilution was made from the population transfection to prepare a clonal transfection. The cells were placed in a 96-well plate (200 µl per well). The cells were allowed to grow for 10 to 20 days before the most diluted wells were transferred to 10 mL media with selected antibiotics. The genomic DNA was isolated once the cells reached a density of $7x10^6$ cell mL^{-1}.

Table 2.4. Solutions needed to generate sgRNA linear cassette. PCR had total volume of 20 µL.		
Solution	**Concentration stock**	**Final concentration**
OL6137	100 µM	2 µM
sgRNA primer (Fwd)	100 µM	2 µM
dNTP	10 mM	0.2 mM
Hifi polymerase (Q5)	3 Units/µl	1 Unit/µl
5x HiFi reaction buffer (with MgCl2)	5x	1x
ddH2O	-	-

Table 2.5. Solutions needed to generate the repair cassettes with either BSD or PAC resistances. PCR had total volume of 40 µL.		
Solution	**Concentration stock**	**Final concentration**
Template pGL2662/pGL2667	30 ng/ µL	30ng
dNTP	10mM	0.2mM
Hifi polymerase (Q5)	3Units/ µL	1Unit
5x HiFi reaction buffer (with MgCl2)	5x	1x
Forward Primer	100µM	2µM
Reverse Primer	100µM	2µM
ddH2O	-	-

Table 2.6. Phusion High-Fidelity PCR thermocycling steps used for sgRNA PCR amplification.			
Process	**Temperature**	**Time (Min)**	**Number of cycles**
Initial Denaturation	98	00:30	1
Denaturation	98	00:10	35
Annealing	60	00:30	
Extension	72	00:15	
Final Extension	72	5	1

Table 2.7. Phusion High-Fidelity PCR thermocycling steps used for repair cassette PCR amplification.			
Process	**Temperature**	**Time (Min)**	**Number of cycles**
Initial Denaturation	94	5	1
Denaturation	98	00:30	45
Annealing	65	00:30	
Extension	72	2:15	
Final Extension	72	7	1

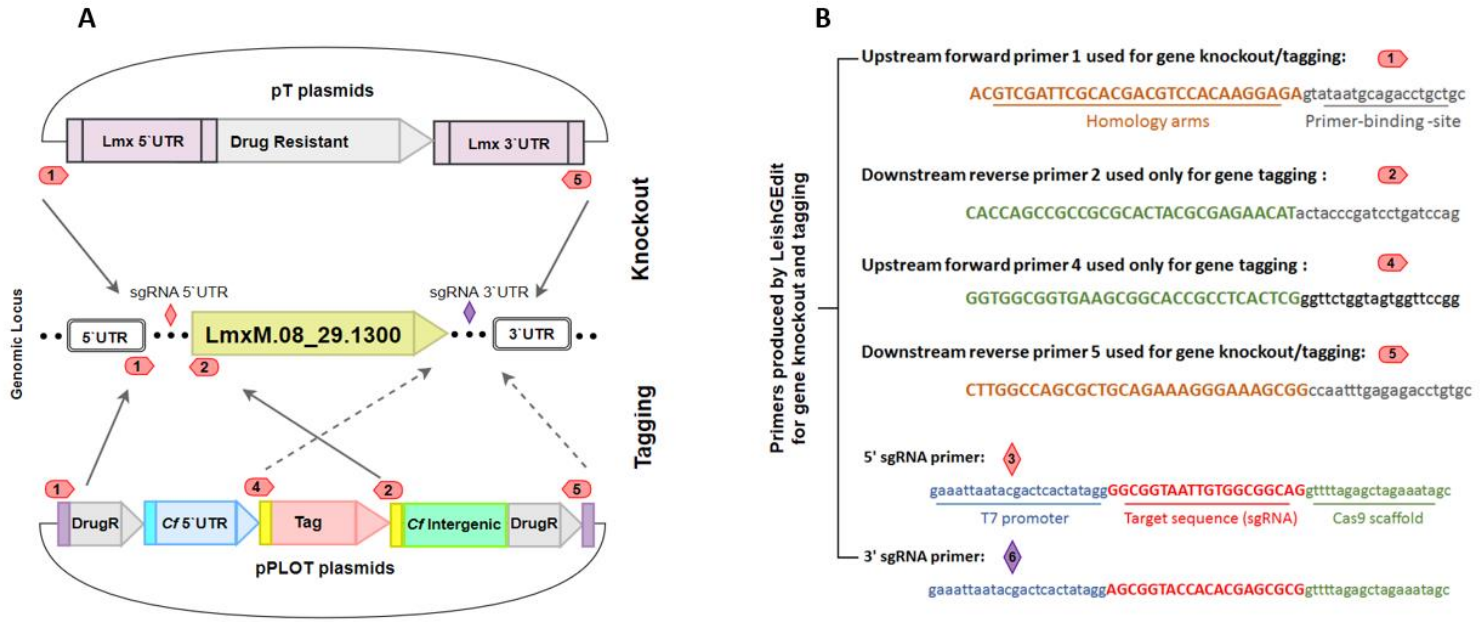


Figure 2.1. The technique of PCR-amplification to generate CRISPR/Cas9 DNA fragments

(A) A scheme diagram which shows how the DNA fragments were generated for gene modification. (B) The primer sequences used in the PCR. Primers 1 and 5 were used to generate gene knockout using pT plasmid as a DNA template. Primers 1 and 2 were used to tag the protein N-terminus (showed by grey arrows) and primers 4 and 5 were used to tag the protein C-terminus (showed by the grey dashed arrows). For protein tagging, the pPLOT plasmid was used as a DNA template. The drug resistant cassettes were used as a DNA donor to repair the double-strand breaks via homologues recombination. The sgRNAs were used to introduce double-strand breaks. For gene knockout, two sgRNA were used to specifically cut upstream (5`UTR) and downstream (3`UTR) off the target gene. For gene tagging, single sgRNA was used to either cut at 5`UTR or 3`UTR sites.

Chapter 3

Composition and distribution of cytochrome *c* maturation systems in eukaryotes

System I for *c*-type cytochrome maturation (or Ccm system) has most often been studied in gram-negative bacteria (Sanders et al., 2010, Stevens et al., 2011, Verissimo and Daldal, 2014, Shevket et al., 2018). In contrast, fewer studies have focused on the distribution and possible variations of the Ccm system in the mitochondria of eukaryotes that do not possess holo-cytochrome *c* synthetase (HCCS or System III). To provide some contaxt, variations in the composition of the Ccm system between prokaryotes and eukaryotes may occur. This is possible as (i) variations in the Ccm systems in Archaea and bacteria have been reported previously (Kletzin et al., 2015, Allen et al., 2006); (ii) the redox environments of the bacterial periplasm and mitochondrial intermembrane space – the sites of *c*-type cytochrome maturation in prokaryotes and eukaryotes respectively – are potentially distinct; (iii) in bacteria the Ccm system matures a wide variety of *c*-type cytochromes whereas mitochondrial cytochromes *c* and c_1, the sole *c*-type cytochromes found in non-photosynthetic eukaryotes, share a similar protein fold (Sharma et al., 2010, Bertini et al., 2006, Allen, 2011). As a prelude to looking bioinformatically for possible components of a trypanosomatid cytochrome *c* maturation system, one of my early analyses was to determine the distribution and possible variation of Ccm system in eukaryotes. The objective of this work was to potentially gain an understanding of the types of motifs that might be important for heme attachment in the cryptic trypanosomatid cytochrome *c* maturation system.

In bacteria, the Ccm system contains up to nine different proteins, CcmA to CcmI (as summarised in the introductory chapter), some of which are conserved in at least some land plant and red algal genomes (Ohta et al., 1998, Spielewoy et al., 2001). Using bioinformatics approaches, protein alignments, and by reference to published literature, I examined the occurrence and variations of System I and III for a broad range of unicellular eukaryotes, ferns, hornworts, and mosses. At the time that I carried out this analysis, many of the taxa for which genomes or transcriptomes became publicly available were previously (and are perhaps still) little known. Some of these taxa are evolved from ancestors believed to have diverged close

to some of the better known eukaryotic groups such as animals and fungi or that they were divergent at an early point during eukaryote evolution.

This chapter outlines the different stages of the analysis conducted. First, System I in protists and plants was studied extensively. Although previous studies have reported on the use of System I by some species, they generally did not explore the composition and the conserved functional motif of System I components, which is the main focus of this analysis. Next, the distribution of both Systems I and III across eukaryotic phylogeny was mapped. During this stage, unexpected and complex appearances of both systems across major eukaryotic groups were identified. Finally, this chapter concludes with the identification of a candidate protein used by trypanosomatids to mature their unusual *c*-type cytochrome.

3.1 Ccm system in eukaryotes - proteins required for the attachment of heme to CcmE

In the first steps of heme attachment to *c*-type cytochromes via System I, heme is covalently attached to the heme chaperone CcmE. In bacteria, this requires the activities of CcmA, B, C, D and E (Figure 3.1).

CcmA is an ATPase that belongs to the ABC transporter family. CcmA orthologues were previously identified in plants (*Arabidopsis thaliana*) and in the mitochondrial genomes of the red alga *Cyanidioschyzon merolae* and the jakobid *Reclinomonas americana* (Lang et al., 1997, Ohta et al., 1998, Rayapuram et al., 2007). To test the degree of conservation between CcmA orthologues from different taxa, a protein alignment was constructed using CLC sequence viewer. This alignment used CcmA protein sequences from *C. merolae*, *R. americana* and *A. thaliana* (Figure 3.2). From this alignment, it appeared that CcmA is not particularly well conserved between taxa except for a conserved C-terminal region (boxed in Figure 3.2). This conserved region was used as a query when mapping the distribution of System I versus System III components. When candidate CcmA sequences were identified in transcriptome or genome sequences, alignments were performed to confirm the presence of a second, less conserved region in the N-terminus of CcmA (red box in Figure 3.2). In this way, putative CcmA proteins from *Naegleria gruberi*, jakobids and ciliates were identified (accession numbers can be found in Appendix 4). In contrast to *C. merolae* and *R. americana*, CcmA in *N. gruberi* and ciliates was nuclear-encoded (Figure 3.5).

In plants and bacteria ,CcmA forms a complex structure with another ABC transporter called CcmB (Rayapuram et al., 2007, Dale et al., 2007). CcmB is widely found in bacteria. Candidate CcmB proteins are also annotated in the mitochondrial genomes of some protists (some red algae, jakobids) and land plants (Figure 3.5), but from failure to identify candidate orthologues by BLAST analyses CcmB is likely to be either absent from or at least highly divergent in the genomes of other eukaryotes in which System I is present (*e.g.* red alga *Galdieria sulphuraria*, *N. gruberi*, ciliates). In contrast to CcmA and B, the protein alignment of the mitochondrial encoded CcmC is more obviously conserved between *A. thaliana, C. merolae* and *R. americana* (Figure 3.2). A highly conserved region of CcmC (≈100 amino acids from the C-terminus) was used to refine the query for BLASTp (black box in Figure 3.2). As result from BLASTp, putative CcmC orthologues were identified in *N. gruberi* and *Malawimonas*. However, CcmC protein was not identified in ciliates (*e.g*. *Oxytricha trifallax, Paramecium tetraurelia, Tetrahymena thermophila* and *Stylonychia lemnae*). CcmC is used to deliver heme to CcmE which then forms a temporary complex (CcmC: heme: CcmE). The heme attachment to CcmC requires a conserved WWD motif and two His residues (Sanders et al., 2010). Additional conserved residues, Asp47, Gln50 and Arg55, were identified in *E. coli* CcmC and are involved in mediating the interaction between CcmC and CcmE (Verissimo and Daldal, 2014, Shevket et al., 2018). The residues needed for the heme attachment to CcmC are found to be widely conserved across eukaryotes, however, in *N. gruberi* and *N. fowleri,* the candidate CcmC was found to be lacking some of the key residues in WWD motif. Furthermore, the candidate CcmC in these species is found to lack all the residues required for CcmC-CcmE interaction (Figure 3.3).

The heme chaperone protein of the Ccm system is known as CcmE (Sanders et al., 2010). CcmE forms an intermediate complex with heme and CcmC. The conserved HXXXY motif in CcmE protein facilitates heme interaction. In eukaryotes examined to date, the HXXXY motif is found to be orientated in the inner membrane space of the mitochondria (IMS), where the His forms a temporary covalent-link to heme and Tyr works as an axial ligand to the heme (Sanders et al., 2010, Shevket et al., 2018). Aside from HXXXY motif, there are additional conserved residues in bacterial CcmE (Arg73, Asp101 and Glu105 in *E. coli* CcmE) which are known to interact with CcmC. In eukaryotes previously examined (plants and red algae), CcmE is a nuclear encoded gene and amino acid sequence of the gene product are well conserved

in these organisms. Full-length CcmE sequences from *A. thaliana* and *C. merolae* were used as queries in BLASTp. Candidate CcmE proteins were identified in all species where mitochondrial encoded Ccm components were found. Protein alignment was done using the identified candidate CcmE proteins sequences. From the protein alignment, the HXXXY motif in CcmE is found to be conserved in most eukaryotes, which use Ccm system (Figure 3.3). However, there are some exceptions, in *N. gruberi* and some ciliates (*e.g. Tetrahymena thermophila)*, only the His in the HXXXY motif was found to be conserved. In addition to this, the candidate CcmE in ciliates lacked most of the conserved residues known to interact with CcmC, which is expected because candidate CcmC proteins were not identified in ciliates (Figure 3.3).

CcmD is another protein that is required for the covalent heme attachment to CcmE. CcmD is a small protein of 69 amino acids in *E. coli* with no recognisable motifs and for which no candidate orthologues could be found in any eukaryotes.

3.2 Ccm system in eukaryotes - proteins required for heme transfer to apocytochrome *c*

The next steps in cytochrome *c* maturation by System I involve the heme delivery from CcmE to apocytochrome *c*. In bacteria, these steps require the activity of CcmF, G, H and I (Figure 3.1).

CcmF is mitochondrial encoded protein with at least 11 mitochondrial transmembrane helices. Studies have shown that CcmF has a conserved WWD motif along with three other conserved His residues (His179, His266 and His308 in *C. merolae* CcmF). CcmF is used to bind and reduce the heme from holoCcmE and transferring it to apocytochrome *c* (Rayapuram et al., 2008, Verissimo and Daldal, 2014). Candidate CcmF proteins were found to be annotated in the mitochondrial genomes of some protists (some red algae, *Malawimonas*, jakobids and *N. gruberi*) which use System I for cytochrome *c* maturation. Protein alignment was carried out using candidate CcmF protein sequences. From the alignment, the conserved WWD motif and His residues were identified in every ortholog (Figure 3.4). In previous literature, it was stated that CcmF gene in plants is split into two separate polypeptides; $CcmF_N$ and $CcmF_C$ where $CcmF_N$ encodes for the N-terminal region of CcmF and $CcmF_C$ encodes for the C-terminal region (Rayapuram et al., 2007, Giegé et al., 2004). However, the $CcmF_N$ in *A.*

thaliana was shown to split further to two separate polypeptides which means it has three sperate genes for CcmF in the mitochondrial genome ($CcmF_N1$, $CcmF_N2$ and $CcmF_C$) (Rayapuram et al., 2007). Despite that CcmF is commonly split into two separate genes in plants, it is encoded as a single polypeptide in the unicellular protists, which use System I for cytochrome *c* maturation.

The final bacterial Ccm protein that is conserved in eukaryotes is CcmH. CcmH is thiol-disulfide oxidoreductase enzyme and it is used to recognise and reduce the CXXCH motif in apocytochrome *c* in order for the heme to attach to apocytochrome. The conserved redox-active RCXXC motif in CcmH is responsible for the reduction of the disulfate in apocytochrome *c* (Di Matteo et al., 2007, Cramer and Toivo, 2016). To search for CcmH candidates in *N. gruberi*, algae and ciliates genomes, the protein sequence of previously identified *A. thaliana* CcmH was used as query in BLAST analyses (Meyer et al., 2005). CcmH orthologs were found in algae and ciliates but unexpectedly not in *N. gruberi* and *N. fowleri*. Amino acid sequence alignment between identified CcmH proteins was constructed. The conserved functional RCXXC motif was found to be conserved in most species except for some ciliates such as in *Paramecium tetraurelia, Stylonychia lemnae* and *Oxytricha trifallax* (Figure 3.4). In some bacteria, there is an additional Ccm protein known as CcmI which interacts with CcmF and H to form a multisubunit protein complex. CcmI is known to chaperone the CXXCH motif in apocytochrome *c* to the heme ligation site (Sanders et al., 2008, Verissimo et al., 2015). In eukaryotes, there is no trace of CcmI and studies had shown that CcmH interacts with CcmF to form a complex that is responsible for the heme transfer to the apocytochrome *c* (Rayapuram et al., 2008, Meyer et al., 2005).

CcmG protein is another important protein found to be conserved in bacteria and it is used to maintain the redox balance in periplasm space. No orthologues for the bacterial CcmG genes were found in eukaryotes.

Based on the above results, System I components are generally different between bacteria and eukaryotes. Notably, three main bacterial Ccm proteins, CcmD, CcmI and G were absent from eukaryotes or at least are too divergent to be identified using standard BLAST approaches. Even between eukaryotic lineages, Ccm components can be varied (Figure 3.5). However, CcmE, A and F were found to be present in all eukaryotic species that use System I for cytochrome *c* maturation. The conserved functional motifs in these protein sequences

were used in the attempt to identify possible components of a trypanosomatid cytochrome *c* maturation system (Section 3.5).

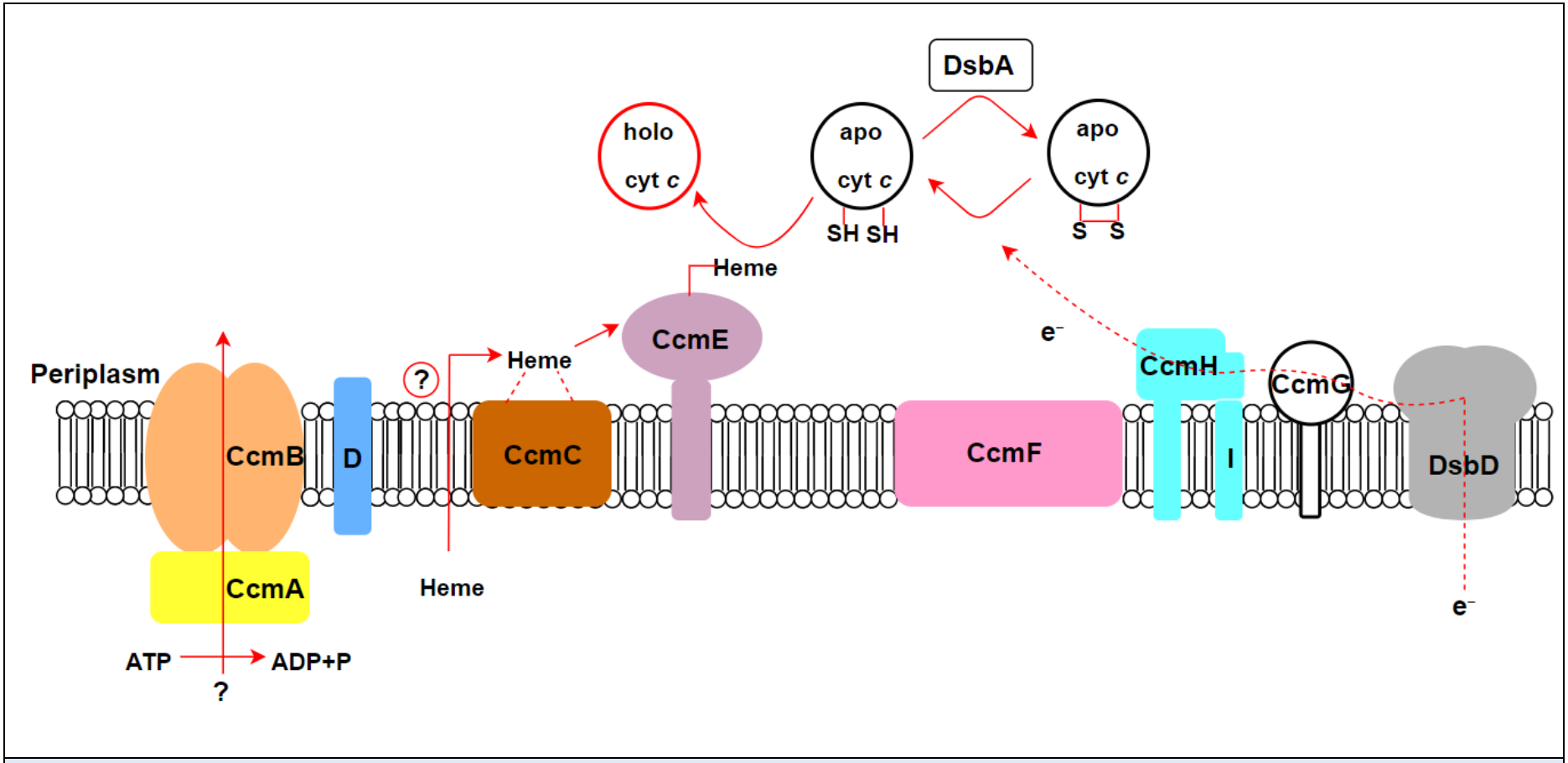


Figure 3.1. Overview of bacterial cytochrome *c* maturation system

In bacteria, CcmA, B, C, D and E are anticipated to be involved in heme delivery whereas CcmF, I, G and H are associated in the redox activity and heme ligation. The dashed arrows indicate the transfer of electrons or heme across the membrane. System I is found in the periplasmic space in most bacteria and the intermembrane space of mitochondria in plants and some red algae.

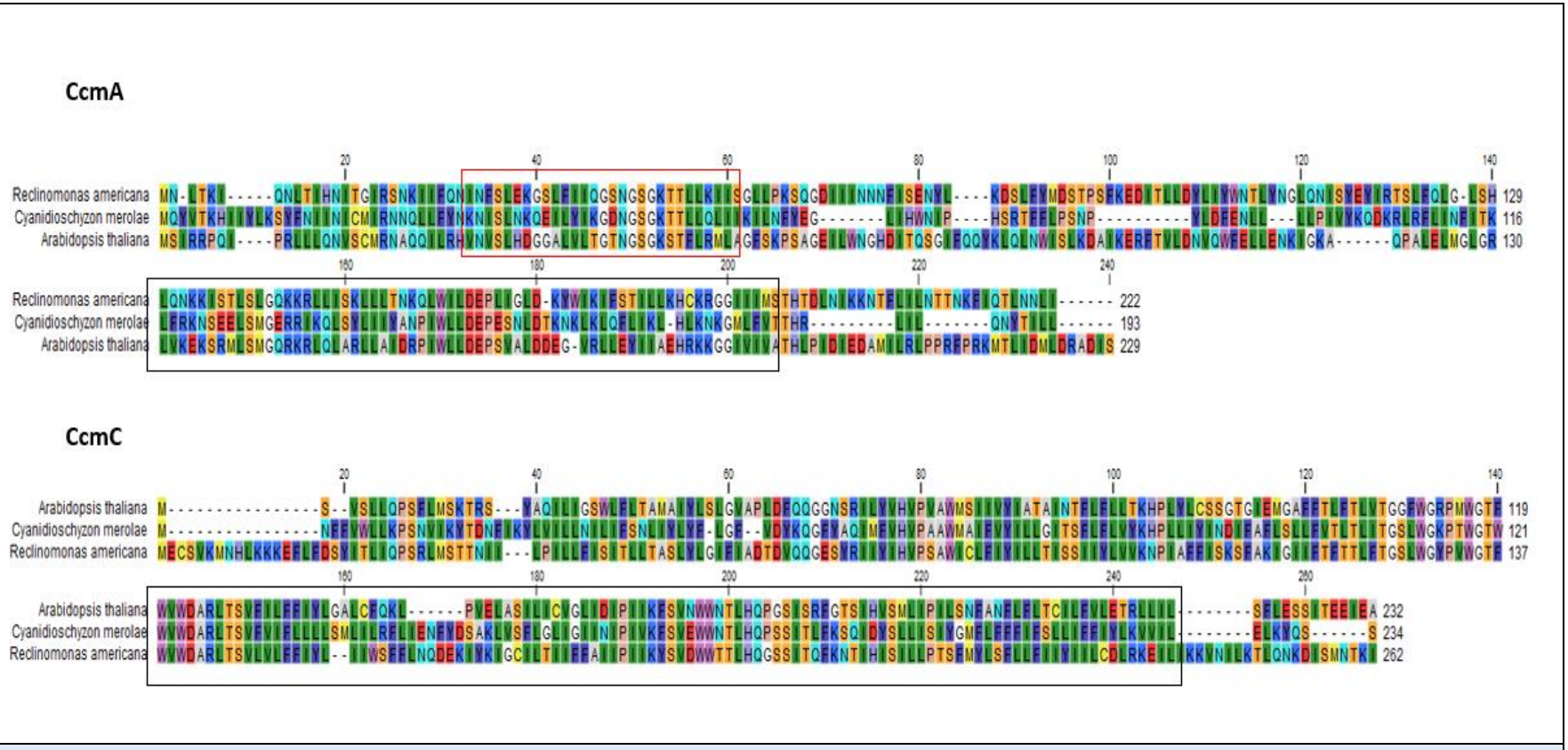


Figure 3.2. Protein alignment of full length protein sequences of CcmA and CcmC across red algae, jakobida and plants

The black boxes show the amino acids used to refine the BLAST query. The red box shows the amino acids used to validate the identity of the putative CcmA protein. The protein sequences were acquired from NCBI database and CLC sequence viewer was used to build the alignment. The amino acid colour scheme is according to RasMol 2.75 and it is based on traditional amino acid properties (Appendix 13). The number at the top of the alignment represents the amino acid position within the sequence including the inserted gaps.

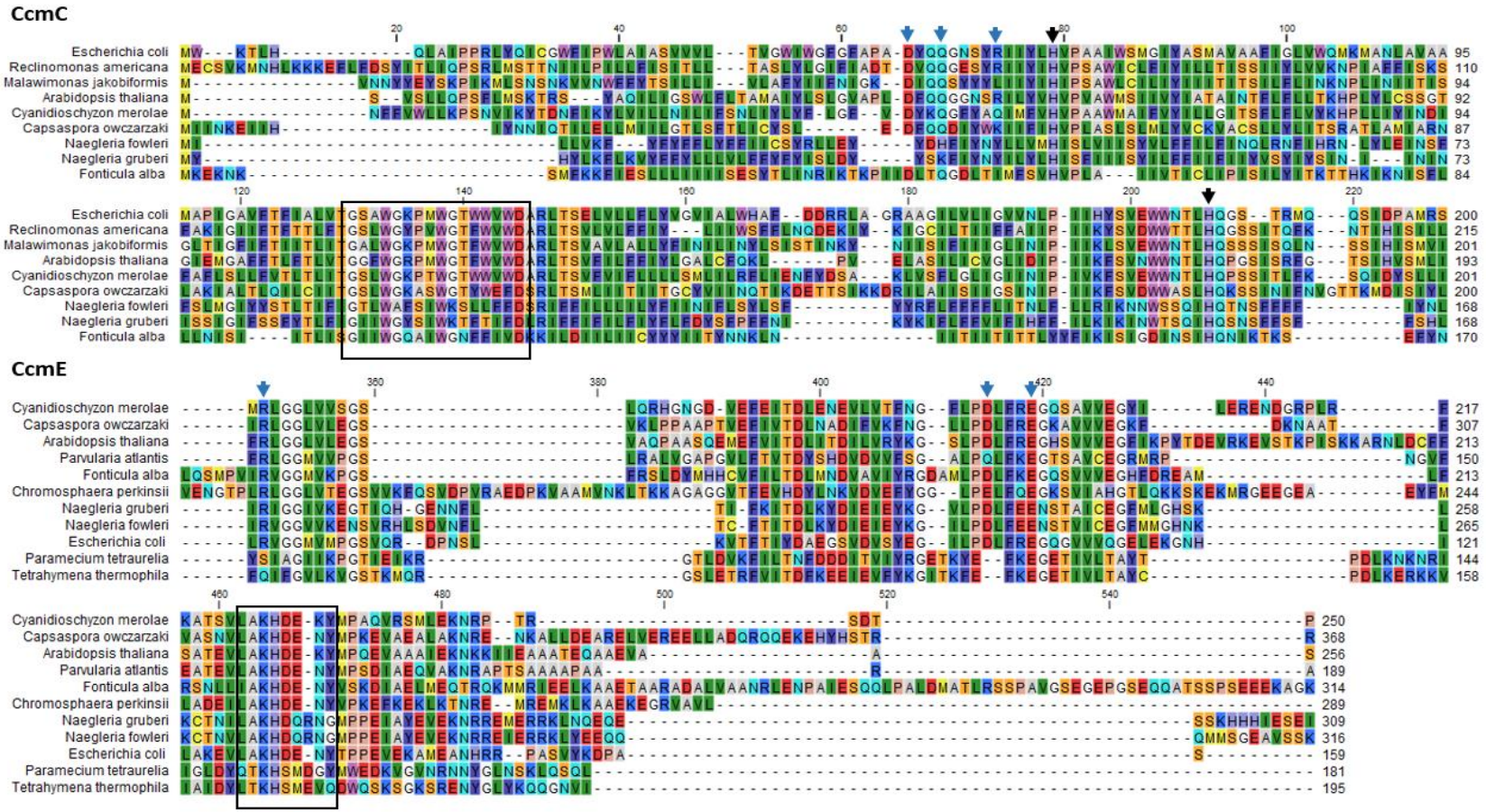


Figure 3.3. Protein alignment of conserved regions of CcmC and E from different eukaryotic lineages. The black arrows show the conserved functional residues of Ccm proteins and the blue arrows show the bacterial conserved residues used for the interaction between CcmE and C. The black boxes show the conserved functional motifs. The length and the accession number of Ccm proteins can be found in Appendix 4. The amino acid colour scheme is according to RasMol 2.75 and it is based on traditional amino acid properties (Appendix 13). The number at the top of the alignment represents the amino acid position within the sequence including the inserted gaps.

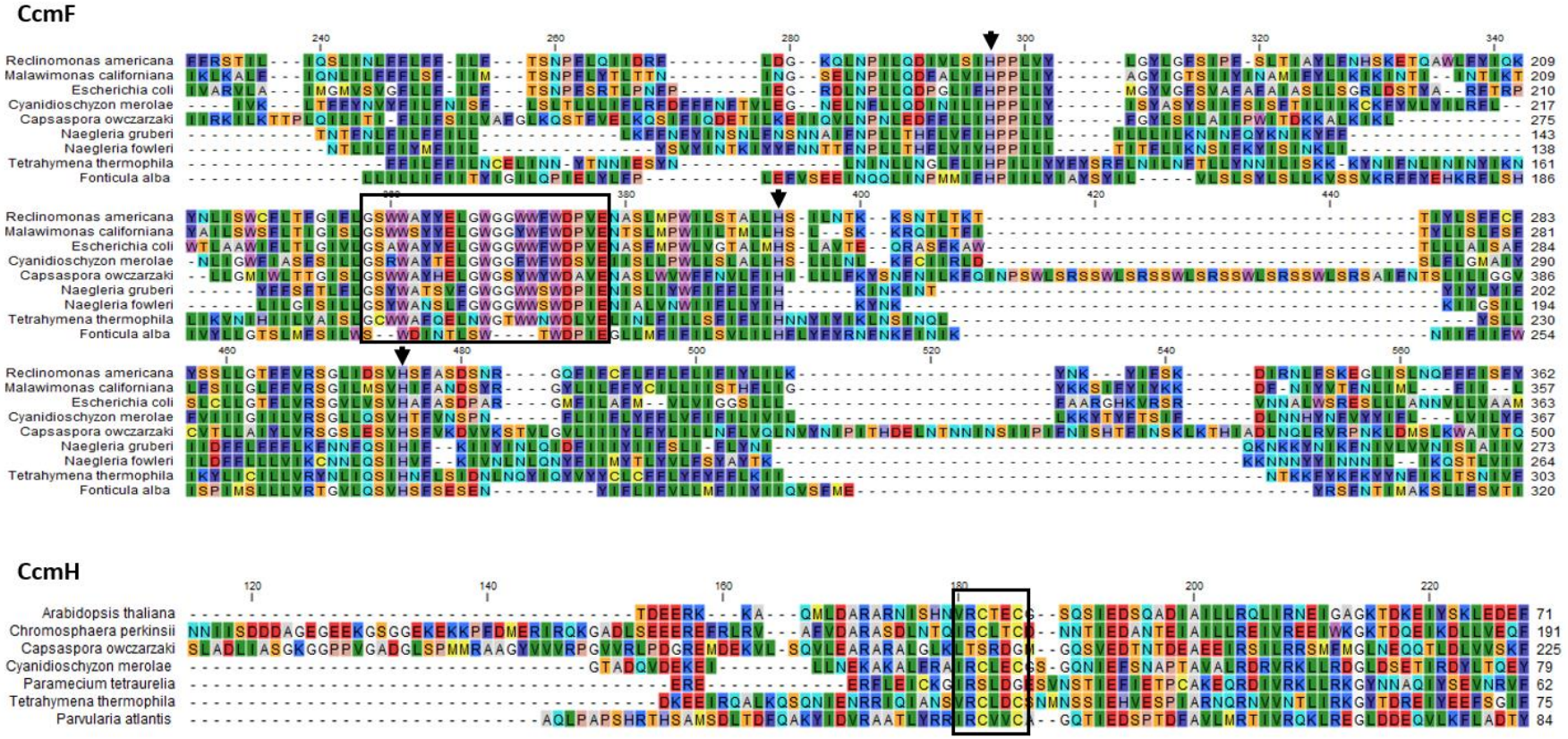


Figure 3.4. Protein alignment of conserved regions of CcmF and H from different eukaryotic lineages

The black arrows show the conserved functional residues of Ccm proteins and the black boxes show the conserved functional motifs. The length and the accession number of Ccm proteins can be found in Appendix 4. The amino acid colour scheme is according to RasMol 2.75 and it is based on traditional amino acid properties (Appendix 13). The number at the top of the alignment represents the amino acid position within the sequence including the inserted gaps.

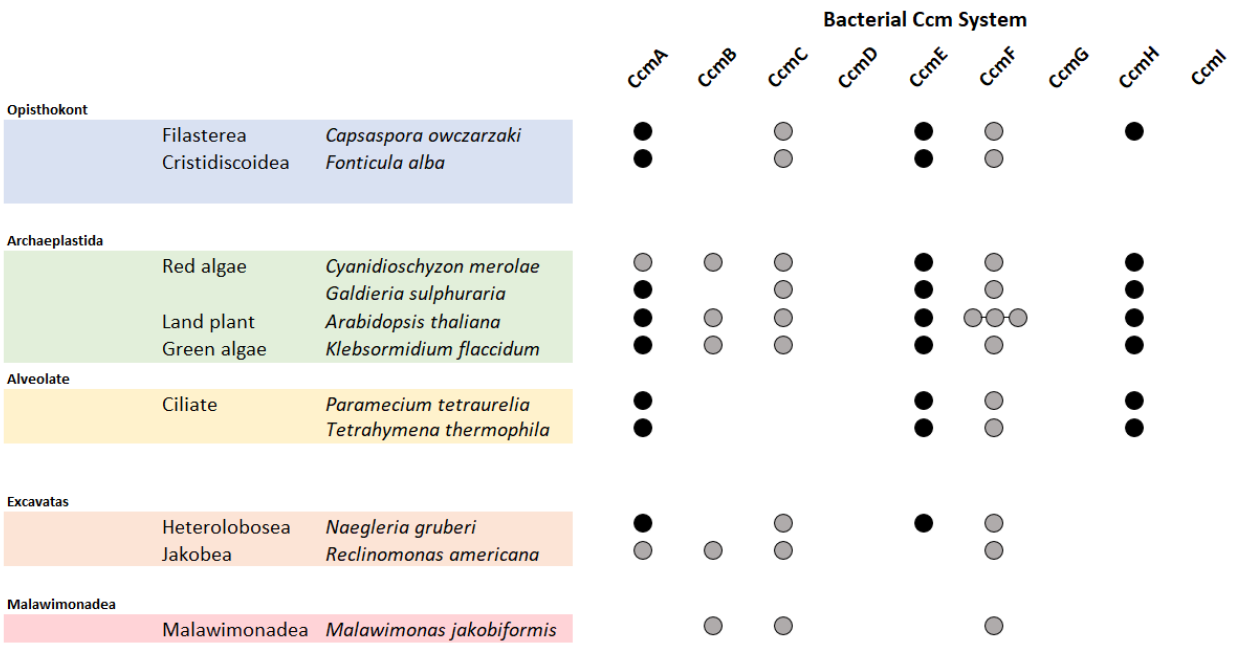


Figure 3.5. The variation of the Ccm components in different eukaryotic clades
In bacteria, there are nine Ccm proteins from CcmA to CcmI. Black circles represent the nuclear-encoded genes and the grey circles represent the mitochondrial encoded Ccm genes. For Jakobea and Malawimonadea only the mitochondria genome sequence is available. Therefore, the nuclear coded Ccm proteins were not been identified in these lineages.

3.3 Unexpected presence of Ccm system in protists related to animal and fungi

Opisthokonts (animals, fungi, and their respective unicellular protists) examined to date use holocytochrome *c* synthase (System III) to mature their *c*-type cytochromes (Figure 3.6) (Suga et al., 2013). When this analysis was started, there was no evidence that opisthokonts taxa use System I instead of System III for cytochrome *c* maturation. DNA data from a number of opisthokont taxa were used in a local BLAST to look for the presence of System I. Publicly available transcriptome or genome sequences of many opisthokonts from the Multicellgenome were downloaded and used as database for local BLASTx. The query library for the local BLASTx was designed manually: it was composed of full length Ccm proteins from *C. merolae* and *A. thaliana* and the two HCCS isoforms from *Saccharomyces cerevisiae*. A separate local BLAST was run for each unicellular opisthokonts where either the transcriptome or genome was publicly available. As predicted, the outcome of the local BLAST for most unicellular opisthokonts showed that they use System III (Figure 3.6). However, conserved System I components were unexpectedly identified in some opisthokont taxa

Fonticula alba, *Parvularia atlantis* (*Nuclearia sp*), *Capsaspora owczarzaki* and *Chromosphaera perkinsii* (protein sequences can be found in Appendix 5). In these species, typical *c*-type cytochromes were identified with CXXCH motif but orthologues of System III proteins were not found. For *C. owczarzaki*, both mitochondria and nuclear genomes were publicly available. Form the genome data, the mitochondrial encoded CcmC and CcmF were identified along with nuclear-encoded CcmA, E and H. Apart from CcmH, Ccm proteins in *C. owczarzaki* had all the conserved residues required for their function (Figures 3.3 and 3.4). For *C. perkinsii* and *P. atlantis*, only transcriptome data was available for download, therefore only the nuclear-encoded Ccm genes (CcmA, CcmE and CcmH) were detected. Despite the fact that both the mitochondria and nuclear genomes were available for download for *F. alba*, only the nuclear-encoded Ccm proteins were identified (CcmA and CcmE). At first glance, this suggested a possible minimal form of System I where only nuclear-encoded candidates CcmA and CcmE were detected. To test whether *F. alba* is truly missing mitochondrial encoded Ccm proteins, a full-length mitochondrial genome of *F. alba* was downloaded from the NCBI (accession number: NW_009243181). Open reading frames (ORF) were identified using ORFfinder server and were used as a database in another local BLAST. The local BLAST was performed using all known mitochondrial encoded Ccm proteins as queries. In this way, candidates CcmC and CcmF were found in *F. alba* (Figure 3.3 and 3.4).

To eliminate the possibility of bacterial contamination, nuclear-encoded Ccm proteins were mapped and upstream and downstream genes (5kb) were identified in the whole genome of *F. alba* and *C. owczarzaki* (Figure 3.7). BLASTx was carried out for the upstream and downstream ORFs to make sure that the BLAST outputs were belonging to either *F. alba* or *C. owczarzaki*. This step was not completed for *C. perkinsii and P. atlantis* because no annotated genome was available. To verify the presence of Ccm system in *C. owczarzaki*, additional laboratory work was carried out. This involve PCR mapping of Ccm genes from isolated genomic DNA. This was carried out to confirm that candidate Ccm genes belong to *C. owczarzaki* rather than bacterial prey or contaminants in the cultures from which the publicly available genome sequences were generated. Complete DNA was purified from axenic culture that is truly *C. owczarzaki* specific. This DNA was used as template for the PCR. Primers were designed to amplify each mitochondrial and nuclear coding sequence of Ccm genes. In addition to this, primers were also designed to amplify part of the genes that are flanking

each of the nuclear encoded Ccm genes. Figure 3.8 shows the amplified nuclear-encoded Ccm genes (CcmA, E, H) along with upstream and downstream flanking regions. The AT content of *C. owczarzaki* mitochondrial genome is high, which meant that designing primers for the mitochondrial System I genes was problematic (Suga et al., 2013b). Therefore, only part of the coding sequence of the mitochondrial encoded Ccm genes were amplified. The amplified genes were ligated to pGEM-T Easy plasmid and were sent for sequencing (pGEM-T Easy cloning and the generated Ccm sequences can be found in in Appendices 6, 7 and 8). The data from PCR mapping and gene sequencing confirmed that *C. owczarzaki* does indeed use System I instead of III to catalyse heme attachment to *c*-type cytochromes.

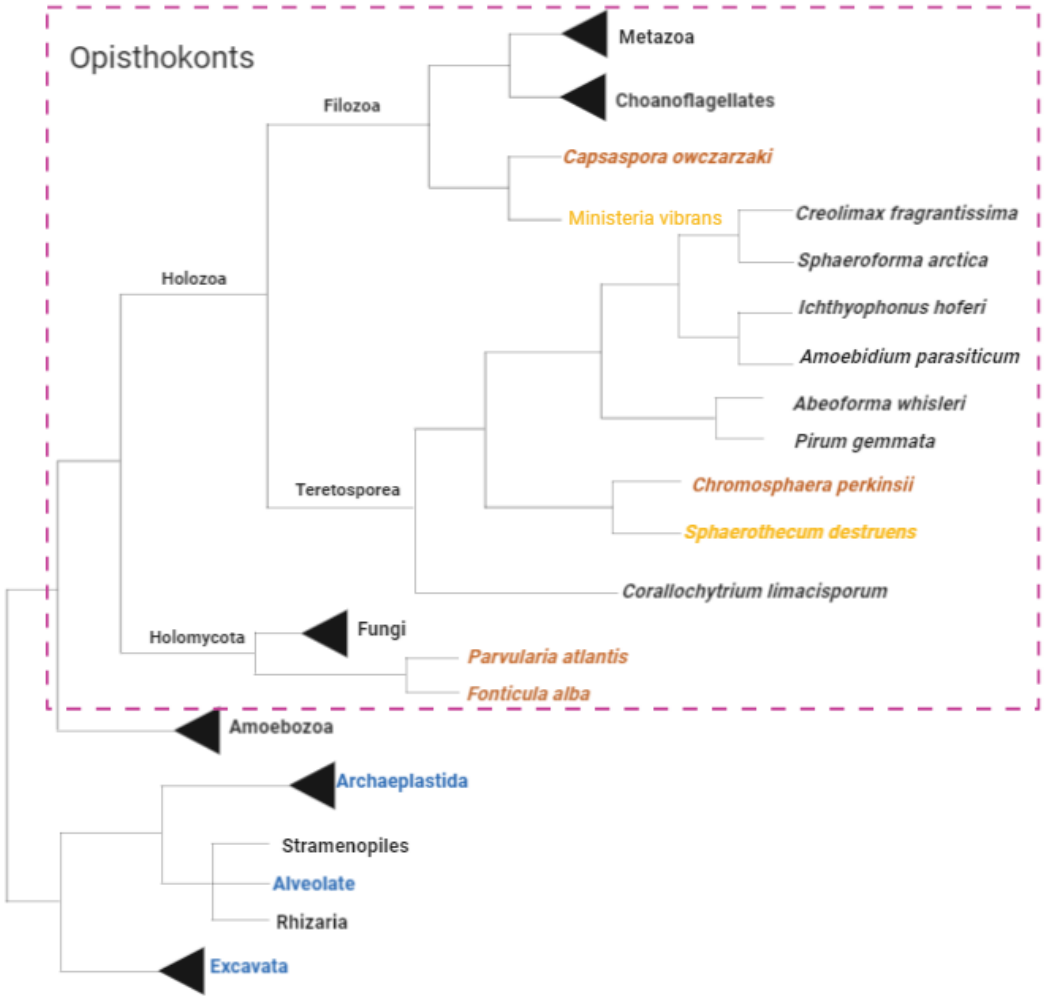


Figure 3.6. The phylogenetic distribution of Systems I and III in unicellular opisthokonts
The transcriptome and genome sequence of multiple unicellular opisthokonts were downloaded from Multicellgenome lab. All groups in black use System III. The brown coloured species use System I and the blue coloured groups shows the mixed usage of System I and III. Species in yellow had transcriptome sequences that were too short to indicate which system was used.

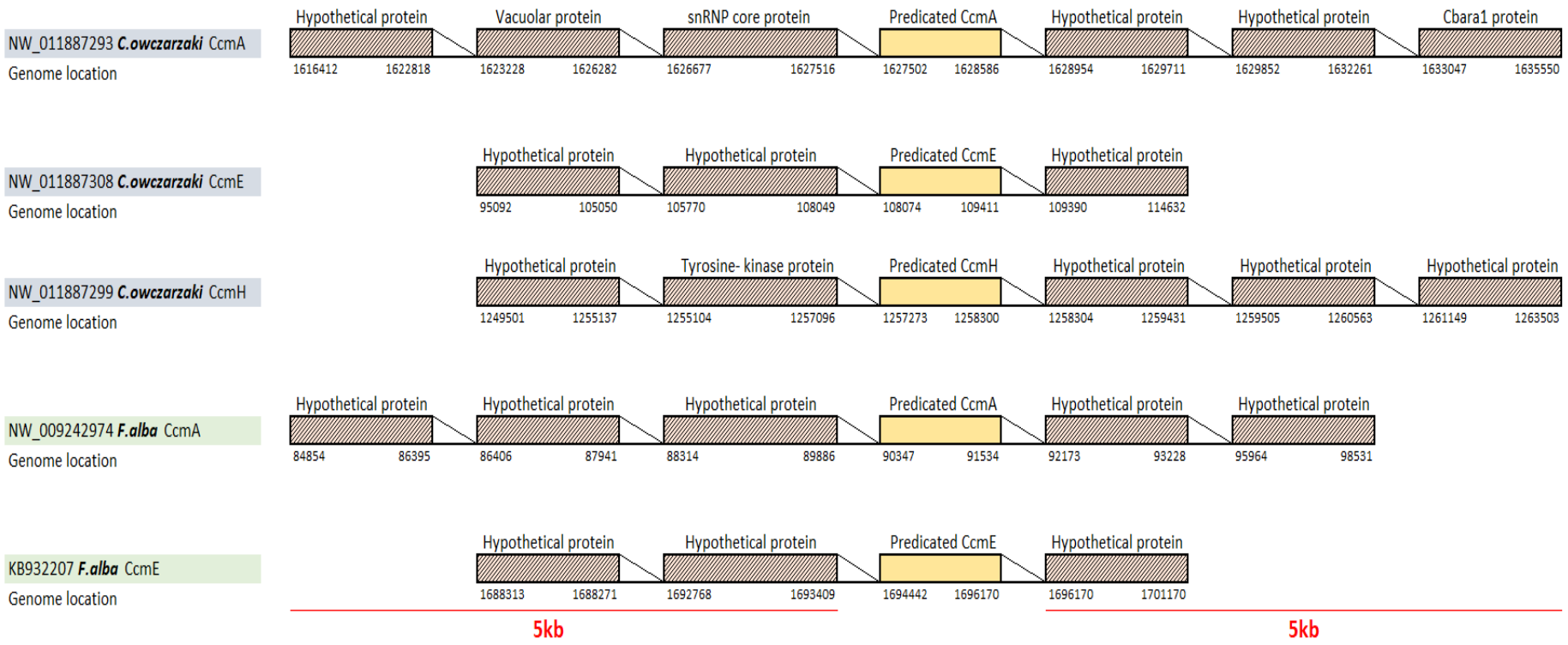


Figure 3.7. The genomic map of Ccm proteins from *Capsaspora owczarzaki* and *Fonticula alba*

The complete genome sequences were acquired from NCBI database. The nucleotides sequence of each of Ccm protein were obtained using tBLASTn. The location of each Ccm gene was mapped and the 5kb upstream and 5kb downstream were used as queries in BLASTx search. The genes upstream and downstream of the Ccm gene were identified and mapped in the complete genome.

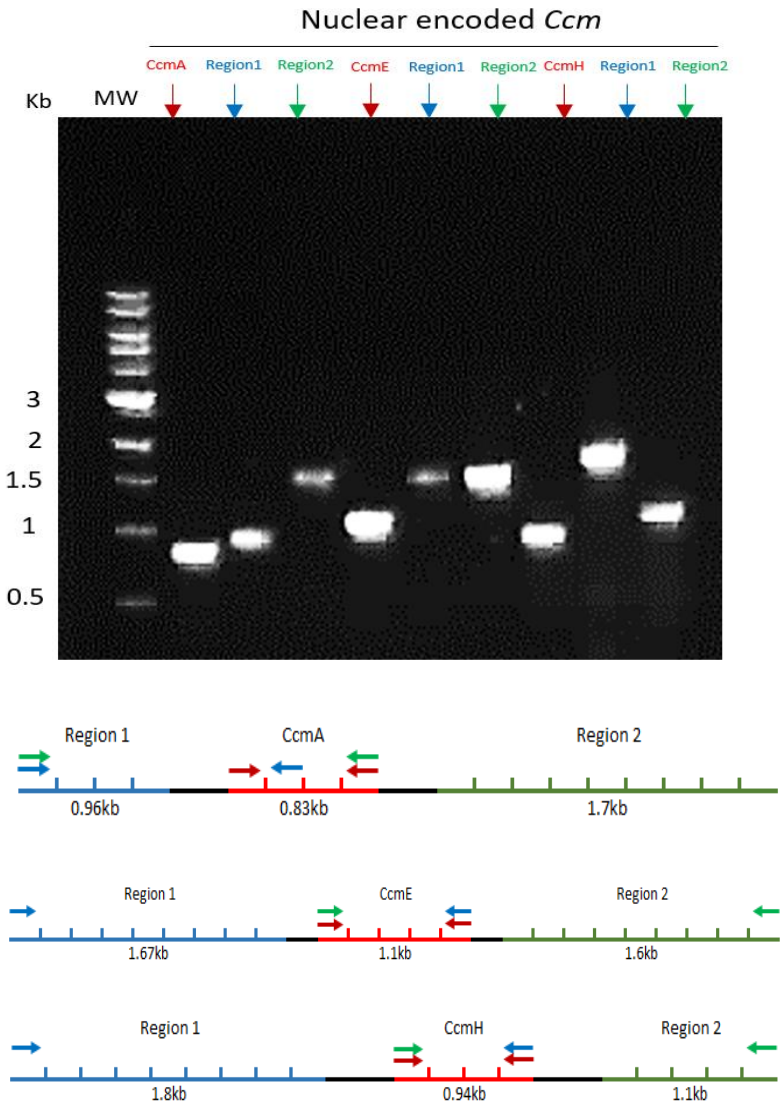


Figure 3.8. PCR mapping of *Capsaspora owczarzaki* nuclear System I genes

Complete gDNA was purified from axenic *C. owczarzaki* culture and was used in PCR. The PCR had a total volume of 25 µL: 50 ng of DNA template, 200 nM of forward and reverse primers, and 25 µL 2x PCR Taq mix. 5 µL of the PCR reactions were electrophoresed through a 0.7% agarose gel. MW: NEB 1 kb DNA Ladder.

3.4 Unexpected presence of System III in primitive plants, charophytes and malawimonads

Following the unexpected identification of System I in number of unicellular opisthokonts, further analysis was then undertaken to look for the distribution of System III in phyla where System I is predominantly found.

It is known that charophytes and land plants use System I instead of III for cytochrome *c* maturation. However, taking into account the recent substantial increase of Viridiplantae transcriptome data, bioinformatics search was undertaken to determine whether any charophytes or land plants do use System III. A tBLASTn search was run using HCCS protein sequences from *Chlamydomonas reinhardtii* as a query against Viridiplantae transcriptome data in the 1k Plants database. The outcome of the tBLASTn showed that System III can be found in a wide range of charophyte algae and `primitive` plants. Taking into account the possibility of fungi or insect contamination in this transcriptome data, the candidate System III proteins from charophytes and primitive plants were used as queries in a separate BLASTp against NCBI database. If the outcome results here showed that candidate System III completely matches the System III sequence from insect or fungus/yeast, then the candidate System III would be considered as probable contamination (Figure 3.10). However, if the BLAST outcome of the candidate System III showed a level of similarity to chlorophytes System III (chlorophytes are known to use System III for cytochrome *c* maturation) then the candidate System III would be regarded a genuine finding. From this analysis, System III orthologues were identified in some charophytes, ferns and club moss (lycophytes, liverworts and hornworts) (Figure 3.9). Aside from two club mosses (*Phylloglossum drummondii* and *Huperzia squarrosa*), none of the System I proteins were found in species where putative System III was present. *P. drummondii* has a partial sequence of the nuclear encoded Ccm genes (CcmH, CcmA and CcmE) and *H. squarrosa* has mitochondrial pseudogene CcmF.

To further increase confidence that the candidate System III do not belong to fungi or insects, phylogenetic analysis was performed to determine if candidate System III proteins from charophytes and primitive plants form a monophyletic group with other known chlorophytes System III. To do this, all the putative System III sequences from charophytes and primitive plants were aligned with System III from different eukaryotic lineages, which are known to use System III to mature their *c*-type cytochrome (accession numbers can be found in

Appendix 10). FASTA protein alignment was used to build phylogenetic tree using RAxML (Figure 3.11). RAxML is a programme that uses the Maximum Likelihood (ML) method to build phylogenetic trees. The ML method is regarded as one of the most accurate methods used to produce reliable results for phylogenetic analysis (Lees et al., 2018). RAxML was therefore selected for use in this study due to its accuracy and speed especially when working with large datasets (Zhou et al., 2018, Stamatakis, 2014). The resultant phylogenetic tree showed that all of the identified putative HCCS enzymes in charophyte and primitive plants form a monophyletic clade with HCCS enzymes from chlorophytes with supported bootstrap values ranging from 70% to 100% but with few exceptions that had weak bootstrap values ranging from 50% to 60%. Finally, by looking for fungi specific proteins in the transcriptome sequence of charophytes and primitive plants it was also possible to determine if there was contamination by fungi sequences. The protein sequence of chitin synthase 2 (chs2) and chitin synthase regulatory factor 4 (chr4) were used as queries in tBLASTn. The tBLASTn was specifically run against the genome or transcriptome database for each of the charophytes and primitive plants where System III was unexpectedly present (Appendix 9). Most of these taxa did not have a match for both chs2 and chr4, however, some had a match with a weak expected value (range from 0.003 to 0.0015).

In addition to charophytes and primitive plants, another divergent unicellular eukaryote belongs to malawimonads lineages, was found to encode for System III instead of System I. Mitochondrial genome data published prior to my study identified Malawimonads contain mitochondria encoded Ccm proteins (*e.g*. *Malawimonas jakobiformis* and *Malawimonas californiana*). The lack of publicly accessible complete nuclear sequence of these species makes it difficult to detect the presence of nuclear encoded Ccm proteins. However, a recent transcriptome sequence of one malawimonad taxon *Gefionella okellyi* was published (PRJNA400870) (Heiss et al., 2018). This was an opportunity to look for the nuclear-encoded Ccm proteins. The transcriptome data were downloaded from NCBI and was used as a database in local BLAST. The queries for the local BLASTx was designed manually and it was composed of *C. merolae* and *A. thaliana* Ccm proteins and HCCS isoforms from *S. cerevisiae*. According to the local BLAST, *G. okellyi* was unexpectedly found to encode for System III and there is no indication of the presence of System I genes in the *G. okellyi* transcriptome.

From the above analysis, the distribution of both Systems I and III is far more complicated than what was previously anticipated. Taxa from the same phylum can either use System I or III for cytochrome *c* maturation (Figure 3.12). This gives an indication that even though most species belonging to Excavata (*e.g.* Jakobea and Percolozoa) use System I for cytochrome *c* maturation, trypanosomatids could use a variant form of either System I or III to mature their unique looking mitochondrial *c*-type cytochromes. The next step of my analysis was to identify whether any of the known System I and III motifs could be identified within any trypanosomatids protein coding gene(s).

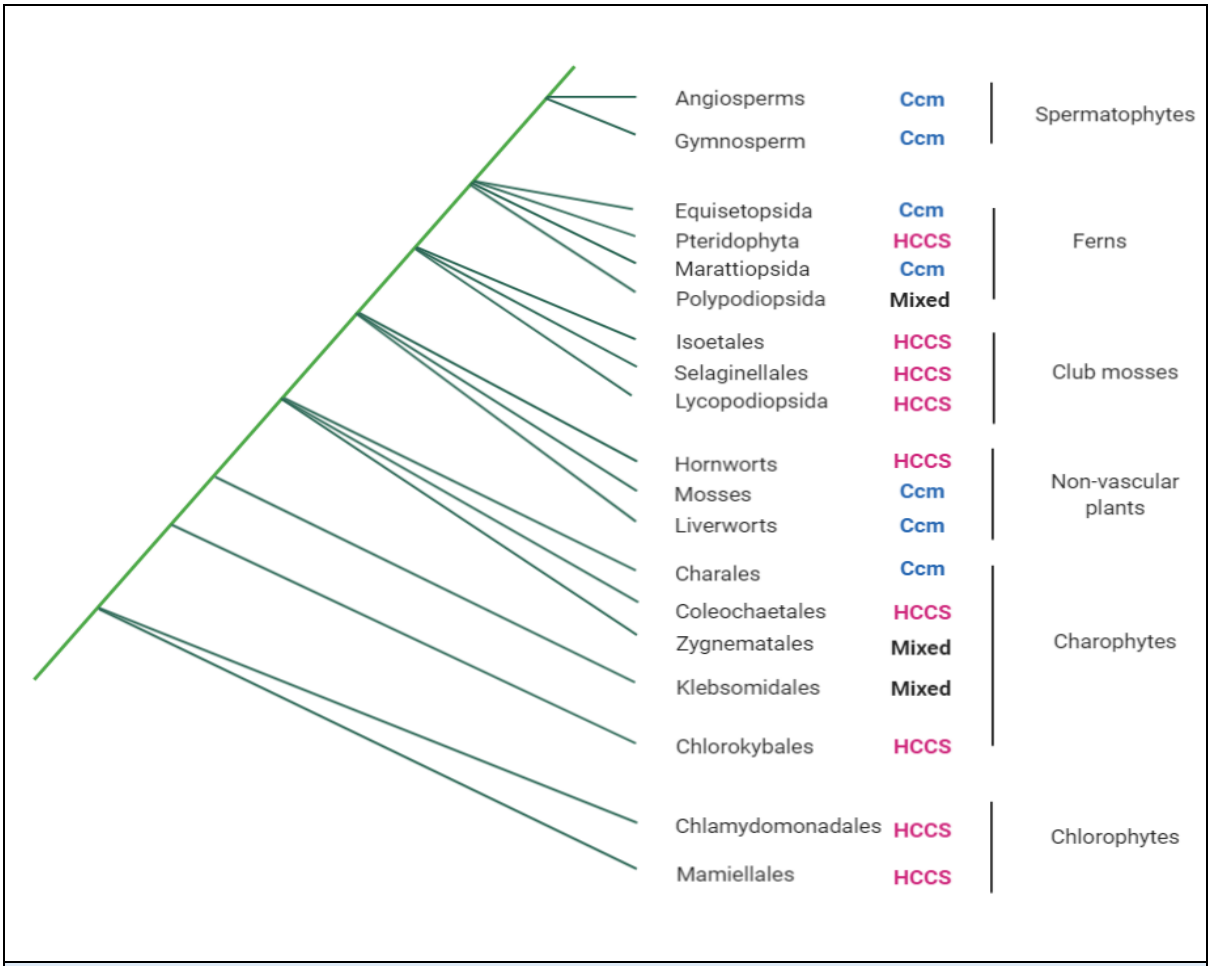


Figure 3.9. Mapping the distribution of cytochrome *c* maturation systems among Viridiplantae
'HCCS' indicates the presence of candidate System III in the transcriptome data obtained from 1k Plants database. 'Ccm' indicates the presence of System I in the genome obtained from the NCBI database. 'Mixed' indicates the presence of either candidate System III in the transcriptome data available in 1k Plants database or System I in the genome obtained from the NCBI database.

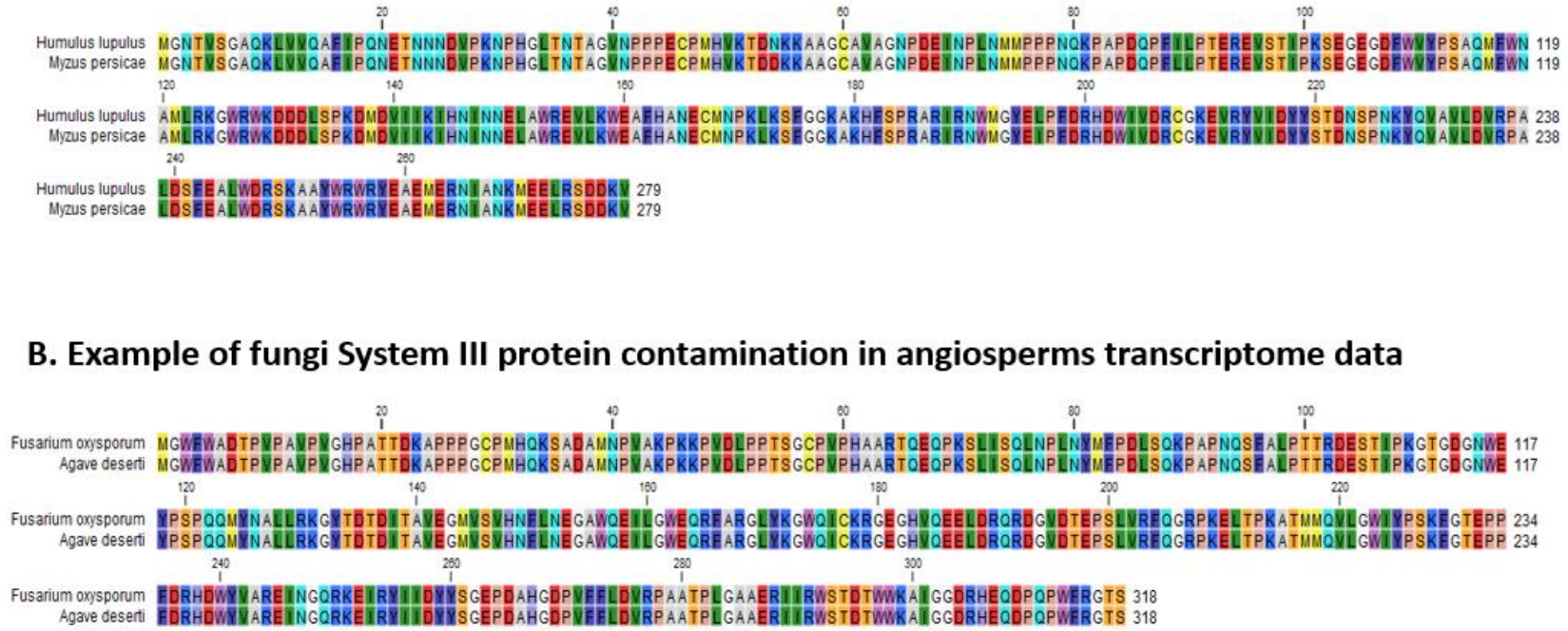


Figure 3.10. Examples of insect and fungi contamination in 1k Plants database

Agave deserti and *Humulus lupulus* belong to angiosperms. Transcriptomes of System III found in these species were obtained from 1k Plants database and were used as queries in BLASTx against NCBI database. The top hits retrieved by BLAST were *Myzus persicae* and *Fusarium oxysporum*, the former belongs to Arthropoda and the latter belongs to Fungi. The amino acid colour scheme is according to RasMol 2.75 and it is based on traditional amino acid properties (Appendix 13). The number at the top of the alignment represents the amino acid position within the sequence including the inserted gaps.

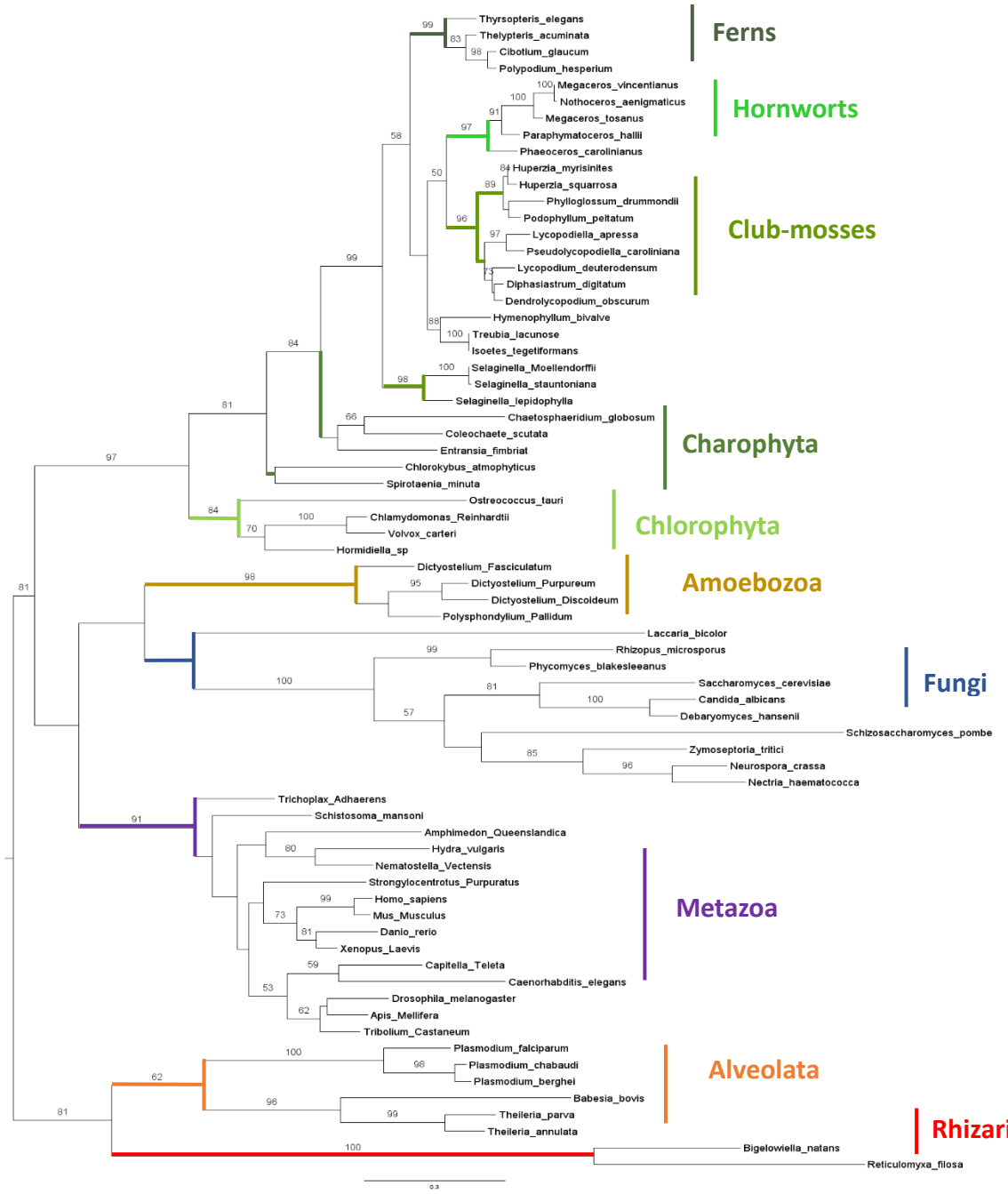


Figure 3.11. Phylogenetic tree of HCCS enzyme

Most of Fungi, Alveolata, Metazoa and Viridiplantae HCCS form a monophyletic clade with good bootstrap support values from 99 to 70% with exceptions in some lineages. This midpoint rooted tree is based on the RAxML maximum likelihood analysis. The parameters of RAxML analysis are as follows: WAG as the best protein substitution matrix, CAT model and Rapid bootstrap analysis (with 1000 standard bootstrap replicates). The scale bar of 0.3 represents the average number of amino acid substitution per site. The RAxML analysis was run on online Cipres web server (Stamatakis et al., 2008). Figtree v1.4.3 server was used to visualise the tree (http://tree.bio.ed.ac.uk/software/figtree/).

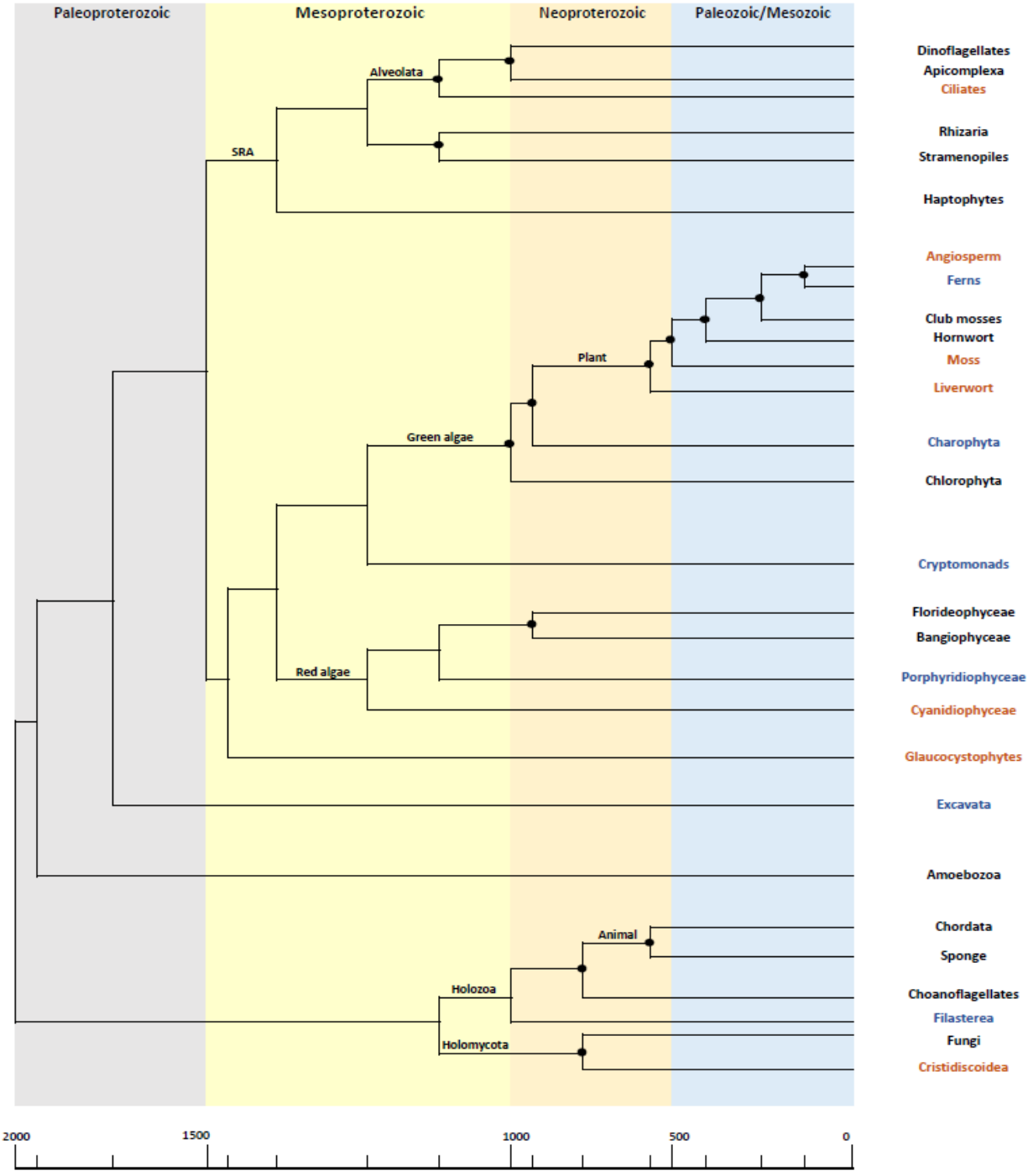


Figure 3.12. The phylogenetic distribution of System I and III and the estimate molecular clock for the major eukaryotic lineages

The time estimate for the divergence of major phyla was based on the following studies; Yang et al., 2016, Parfrey et al., 2011, Cunningham et al., 2017, Clarke et al., 2011, Berbee and Taylor, 2010 and Eme et al., 2014. In these studies, the time-calibrated phylogenetic analyses were calibrated using Phanerozoic (including: Paleozoic/Mesozoic eras) and Proterozoic fossils (including: Paleoproterozoic, Mesoproterozoic, Neoproterozoic eras) (black circles). The colour code represents which system the eukaryotic lineages use to mature their cytochrome *c*. The black code is for System III, the blue code is for mixed lineages where some use System I and some use System III and the brown code is for the System I. Appendix 12 shows what species within each group use System I or III.

3.5 Identification of candidate trypanosomatid cytochrome *c* maturation system

Trypanosomatid parasites of the phylum Euglenozoa are responsible for a spectrum of neglected tropical diseases (as discussed in the introductory chapter). Understanding their highly divergent biology will provide an opportunity to identify possible drug targets to develop new medicines. Uniquely, euglenozoan protists attach heme to mitochondrial *c*-type cytochromes by a single thioether linkage. From their genome analysis, there is no indication for the presence of HCCS or Ccm homologues. To identify candidate cytochrome *c* maturation system in trypanosomatids, bioinformatic proteomic sifting was carried out between trypanosomatids and *Phytomonas* proteomes. *Phytomonas* is plant-pathogenic trypanosomatids with a minimal respiratory chain that lacks ubiquinol:cytochrome *c* oxidoreductase, cytochrome *c* oxidase and cytochrome *c* (Čermáková et al., 2007). Indeed, *Phytomonas* is a rare, if not unique, example of an obligate aerobic eukaryote that grows in the absence of heme (Kořený et al., 2012). Thus, one would predict that the maturation system used for cytochrome *c* biogenesis in trypanosomatids possessing a typical cytochrome-dependent mitochondrial respiratory chain is absent from *Phytomonas* species. Based on this prediction I took two lists of 2194 genes in total encoding for candidate mitochondrial proteins expressed in *T. brucei* (Acestor et al., 2009, Niemann et al., 2013) and sifted for those genes which lacked an orthologue in *Phytomonas*. In this way, a list of 295 genes was produced. This list was sifted further to identify genes that lacked any functional annotation – *i.e*. encoded hypothetical proteins. This yielded a smaller list of 34 proteins (see Table 3.1) that were screened for the potential presence of at least one of the conserved motifs defining core components of the Ccm and HCCS systems (Table 3.2), reasoning that the mechanism of heme attachment to apo-cytochromes *c* and c_1 might utilise a conserved or similar motif. Figure 3.17 summrises the key steps taken to identify candidate maturation system in kinetoplastids. Two genes were identified as candidates: one (Tb927.11.16510) encoded for a protein with predicted limited identity to CcmE, but was very highly conserved amongst a core set of kinetoplastid species (Figure 3.13) and another gene encoding a protein with very limited identity to HCCS, but which was again well conserved amongst kinetoplastids (Figure 3.14 and 3.15). Neither protein in BLAST search was able to recover CcmE or HCCS. However, for Tb927.3.3890 the amino acid sequence identities that suggested the possibility of divergent HCCS domains I-IV occurred in the same relative order as HCCS

enzymes that add heme to a conventional CXXCH heme-binding motif. Moreover, when protein topology was further investigated it showed a similar tendency for a relative disorder prediction within the N-terminus of the protein versus the prediction for some order within the C-terminus (Figure 3.16). This tentative prediction of a candidate trypanosomatid cytochrome *c* maturation protein was supported by the presence of an orthologue in *Perkinsela* albeit that this orthologue was less well conserved when compared with other kinetoplastids (Figure 3.15). The significance observation about *Perkinsela* is that it is an early-branching kinetoplastid lineage that is symbiont of *Paramoeba pemaquidensis* and while it has undergone extensive genome reduction with respect to its cytoskeleton and metabolic capacities it retains a mitochondrion with a complete cytochrome *c*-dependent respiratory chain (Tanifuji et al., 2017). Finding an orthologue of this divergent HCCS in this species is an indication of wide conservation among kinetoplastids. The co-expression experiments of recombinant proteins in *E. coli* that were carried out to test the candidature of a divergent-looking HCCS from trypanosomatids are reported in the next chapter.

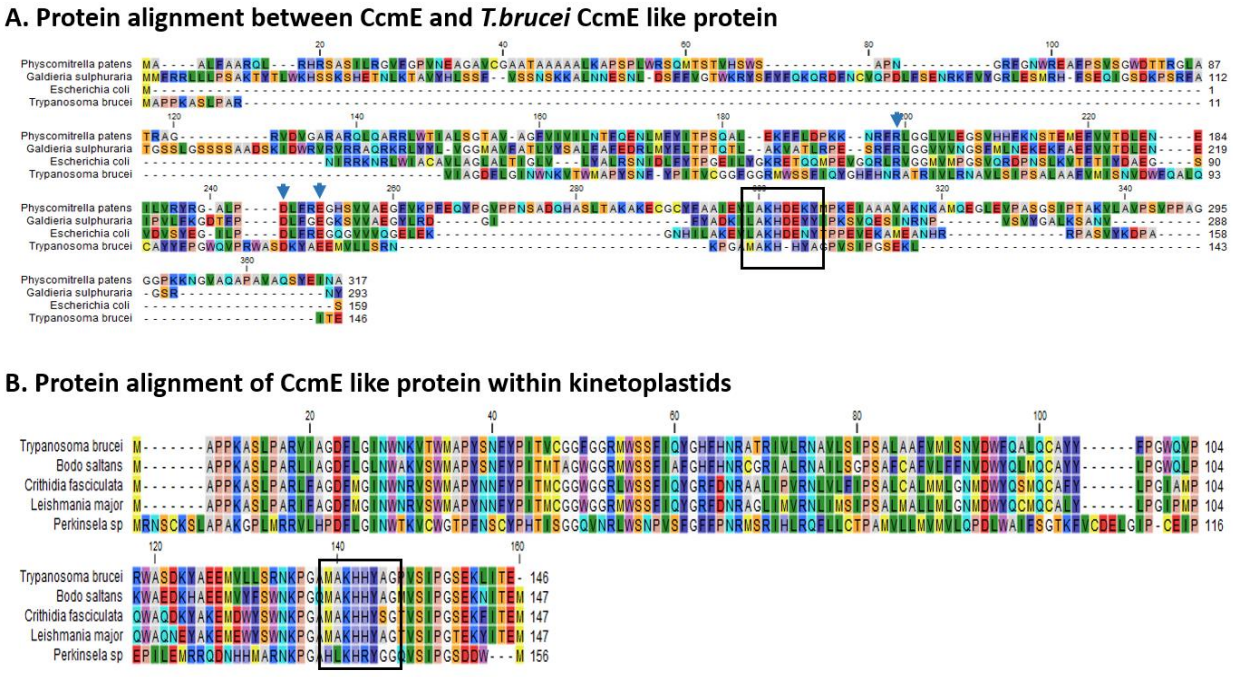


Figure 3.13. Protein alignment of CcmE like protein found in kinetoplastids
The black boxes show the conserved functional motifs of CcmE. The amino acid colour scheme is according to RasMol 2.75 and it is based on traditional amino acid properties (Appendix 13). The number at the top of the alignment represents the amino acid position within the sequence including the inserted gaps.

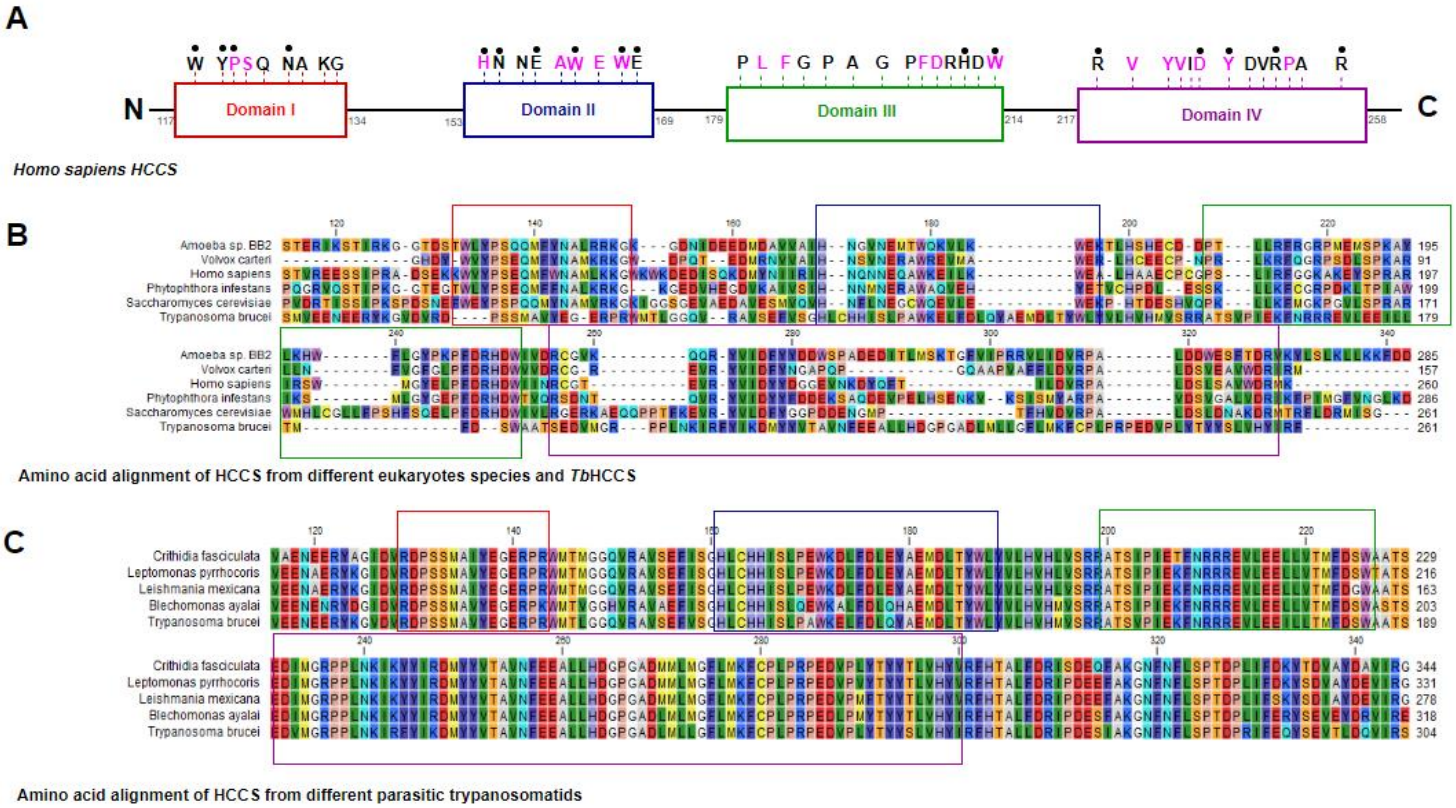


Figure 3.14. Structural and functional conservation of HCCS domains:

(A) Functional domains of the HCCS and their conserved residues. Black coloured amino acids indicate the wide conservation across different eukaryotic species. Pink coloured residues indicate the conservation in the *T. brucei* putative HCCS (*Tb*pHCCS). Black dots indicate residues that were substituted in previous studies. The conserved residues in domain I and II mediate the binding and the release of heme. The conserved residues in domain III are part of the mitochondrial targeting signal. The conserved residues in domain IV are involved in the folding and the stability of HCCS. The number at the top of the alignment represents the amino acid position within the sequence including the inserted gaps.

(B) Sequence alignment of the HCCS across different eukaryotic species and *Tb*pHCCS.

(C) Sequence alignment of the putative HCCS from different parasitic trypanosomatids.

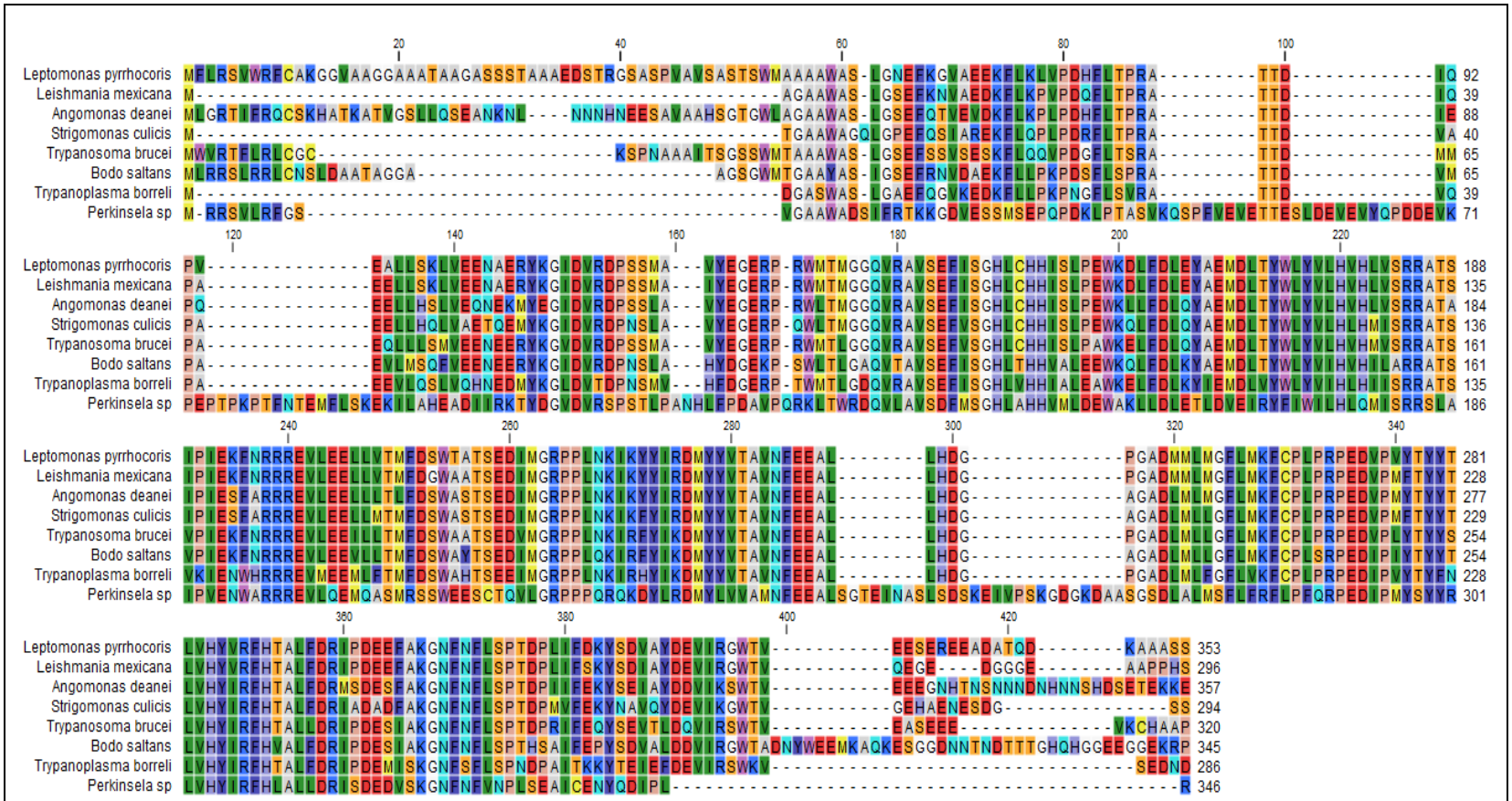


Figure 3.15. The CLC Sequence Viewer protein alignment of putative divergent holocytochrome *c* synthase protein from parasitic and free living Euglenozoa

The accession number for the putative holocytochrome *c* synthase proteins can be found in Appendix 11. The amino acid colour scheme is according to RasMol 2.75 and it is based on traditional amino acid properties (Appendix 13). The number at the top of the alignment represents the amino acid position within the sequence including the inserted gaps.

Figure 3.16. In silico prediction of protein structural disorder of HCCS from different eukaryotic lineages and the putative trypanosomatids HCCS protein

(A) Prediction of the proteins disorder using IUPredA server. IUPredA identifies protein disorder regions from the primary protein structure based on the measured pairwise energy content. (B) Predication of the protein disorder regions using MFDp2 web server. MFDp2 identify protein disorder regions based on solvent accessibility, residue flexibility and B-factors.

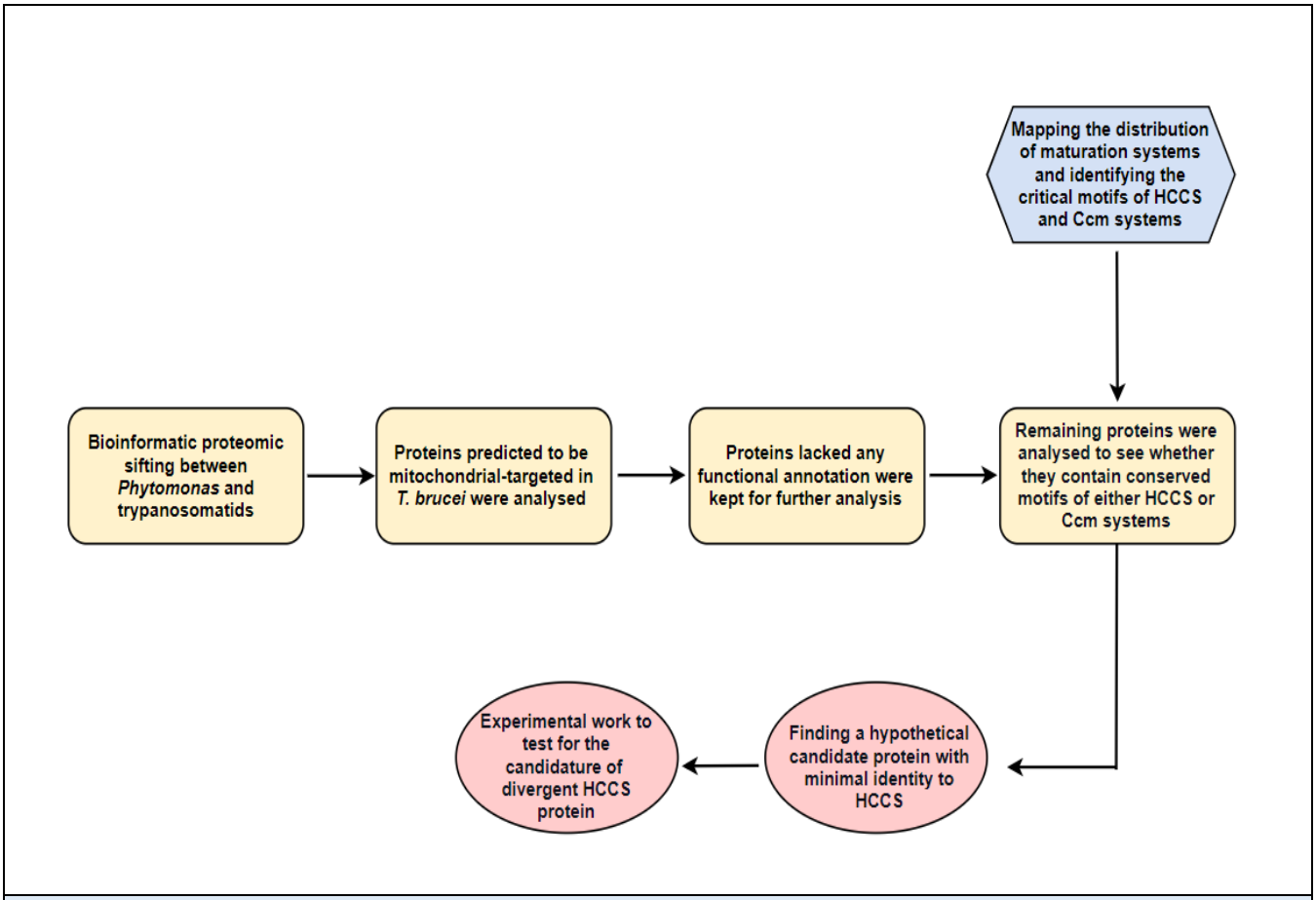


Figure 3.17. Flowchart summarising steps taken to identify putative divergent holocytochrome *c* synthase protein in kinetoplastids.

Table 3.1. Hypothetical candidate mitochondrial proteins of trypanosomatids absent from *Phytomonas*

IM indicates that the protein is likely to be located in the mitochondrial inner membrane. OM indicates that the protein is likely to be located in the mitochondrial outer membrane. Mixed indicate that the protein could be either located in the outer or inner mitochondrial membranes.

Candidate proteins	Length	No. of TMD	Location of TMD	Predicted location within the mitochondria	Domians/motifs	Location of the domain/motif
Tb927.10.5220	447	-	-	OM	Armadillo-like helical	168 - 301
Tb927.10.4240	120	1	78-100	Mixed	-	-
Tb10.v4.0039	990	-	-	OM	Signal peptide	1_15
Tb927.11.13270	270	-	-	OM	low complexity	-
Tb927.11.1290	398	-	-	OM	-	-
Tb927.3.2370	134	1	49-71	IM	-	-
Tb927.3.3570	162	1	111-133	Mixed	-	-
Tb927.3.3890	320	-	-	IM	low complexity	-
Tb927.4.860	300	-	-	Mixed	low complexity	-
Tb927.7.3140	186	-	-	OM	low complexity	-
Tb927.7.3200	354	-	-	OM	-	-
Tb927.7.4270	222	1	170-192	Mixed	Signal peptide	1_28
Tb927.7.6610	897	-	-	OM	Forkhead-associated	23 - 77
Tb927.8.2470	94	-	-	IM	-	-
Tb927.8.5150	95	1	5_22	Mixed	-	-
Tb927.8.8160	821	-	-	OM	low complexity	-
Tb927.3.3520	427	-	-	OM	low complexity	-
Tb927.11.15700	161	1	109-128	Mixed	-	-
Tb927.11.16710	118	1	38-60	Mixed	-	-
Tb927.11.2930	121	-	-	IM	-	-
Tb927.11.4850	443	1	324-343	Mixed	DnaJ domain	50-104
Tb927.11.7250	496	-	-	IM	low complexity	-
Tb927.10.13600	158	1	78-100	Mixed	Coiled coil	130-156
Tb927.10.14370	120	1	48-70	Mixed	-	-
Tb927.10.4240	120	1	78-100	Mixed	-	-
Tb927.10.5220	447	-	-	IM	Armadillo-like helical	168 - 301
Tb927.3.700	85	1	54-76	Mixed	-	-
Tb927.5.740	209	1	183-205	Mixed	low complexity	-
Tb927.11.16510	92	-	-	-	-	-
Tb927.10.5400	574	1	431-453	Mixed	Signal peptide	1-23
Tb927.2.2140	108	1	38-60	Mixed	low complexity	-
Tb927.6.2610	95	1	26-43	Mixed	-	-
Tb927.7.510	356	-	-	OM	low complexity	-

Table 3.2. List of the System I and III conserved functional motifs used to identify trypanosomatids cytochrome *c* maturation system					
Systems	Proteins	TMD	Functional motifs	Other critical residues	Accession number
System I	CcmC	6	W_{117}GX(F/W/Y)WXW(D/E)XR_{127}	H58/H182	NP_059382
	CcmF	11-13	W_{234}(S/A)YXXLGWGG(F/W/Y)WXWDXVEXXS(F/L)(L/M/I)WL_{260}	H179/ H266/H308	NP_059365
	CcmE	1	H_{225}XXXY_{229}	-	BAM80061
	CcmH	2	R_{32}CXXC_{36}	-	BAM80358
System III	HCCS	0	Domain I: W_{117}XXPS_{122} Domain II: $H_{153}N(X)_7W_{161}$ Domain III: P_{206}FDRHDW_{213} Domain IV: Y_{223}(V/I)D(F/Y)Y$X_{16}RP_{247}$	-	AAH01691

Chapter 4

Validation of a divergent HCCS in Trypanosomatids

4.1 Co-expression of *Tb*pHCCS and *Tb*^{His}CYTC substrates – strategy 1

Following bioinformatics-led identification of a candidate HCCS in trypanosomatids that was unrecognisable from comparative BLAST analyses, it was necessary to validate the candidature of this divergent-looking cytochrome *c* maturation system. Thus, to determine whether *Trypanosoma brucei* putative holocytochrome *c* synthase (*Tb*pHCCS) is responsible for heme attachment to *T. brucei* apocytochrome *c* substrate (*Tb*CYTC), a strategy of co-expression in *E. coli* of recombinant *Tb*pHCCS together with recombinant *Tb*CYTC substrates was adopted. The first stage was to amplify by PCR the coding sequences for *Tb*CYTC and *Tb*pHCCS for sub-cloning into *E. coli* expression plasmids pET-15b and pET-28a, respectively. In this way *Tb*CYTC would be expressed with a hexa-histidine (His_6) tag in the N-terminus (*Tb*^{His}CYTC) and *Tb*pHCCS expressed in an untagged, native form. Genomic DNA from *T. brucei* was used as template for PCR and a high-fidelity polymerase was used to maximise sequence fidelity for resultant PCR amplicons. Following PCR, amplicons corresponding to *Tb*^{His}CYTC and *Tb*pHCCS coding sequences were analysed by DNA agarose gel electrophoresis. Figure 4.1 shows two DNA bands sized around 0.9 kb and 0.4 kb which correspond to expected length of *Tb*pHCCS (963 bp) and *Tb*^{His}CYTC (355 bp) coding sequences respectively. The *Tb*pHCCS coding sequence was then digested with *Xho*I and *Nde*I for sub-cloning into *Xho*I and *Nde*I digested pET-28a vector; the *Tb*^{His}CYTC coding sequence was digested with *Bam*HI and *Nde*I for sub-cloning into *Bam*HI and *Nde*I digested pET-15b vector. (Confirmation of sub-cloning of the *Tb*^{His}CYTC into pET-15b and *Tb*pHCCS into pET-28 was carried out by diagnostic restriction digest using *Bam*HI/*Nde*I and *Xho*I/*Nde*I enzymes respectively. Figure 4.1 shows the pET vector backbone and the insert (*Tb*pHCCS and *Tb*^{His}CYTC). Plasmids were then sent for Sanger sequencing to confirm the correct integration of the insert. Following confirmation of the DNA sequences, pET28a and pET-15b plasmids containing their respective inserts were transformed into BL21 competent *E. coli*. The protein expression was initiated by the addition of IPTG (1 mM) to exponentially growing *E. coli*. To detect expression of *Tb*pHCCS and *Tb*^{His}CYTC recombinant proteins, bacterial pellets were collected at three different time points; prior to protein induction (t=0 hr) and then 1 hr and 3 hrs post induction. Following

the addition of SDS loading buffer, resuspended pellets were subjected to SDS-PAGE analysis. Figure 4.2 shows that *E. coli* readily expresses *Tb*pHCCS and *Tb*^{His}CYTC individually. The distinct protein bands at 37 kDa and 14 kDa, corresponding to *Tb*pHCCS (36.6 kDa) and *Tb*^{His}CYTC (13.5 kDa including the weight of His-tag) respectively, are only found within bacterial lysates where IPTG was added to the growing culture. After the trial inductions of *Tb*pHCCS and *Tb*^{His}CYTC separately, the next step was to co-express both proteins in the same expression strain of *E. coli*. The cloned pET-28a and pET-15b vectors were both transformed to the same aliquot of BL21 *E. coli*. Three colonies were picked to detect the protein expression. From the SDS-PAGE gel image in Figure 4.2, the *E. coli* did not express recombinant proteins when both plasmids were transformed into the same aliquot of competent *E. coli*. After going back to the literature, it likely was a consequence of a plasmid incompatibility, which occurs when two plasmids share a common origin of replication (Ori) and begin to compete for the DNA replication machinery of the host cell. Such incompatibility causes the inhibition of recombinant protein expression (Velappan et al., 2007). Both pET-15b and pET-28a share the same origin of replication (pBR322).

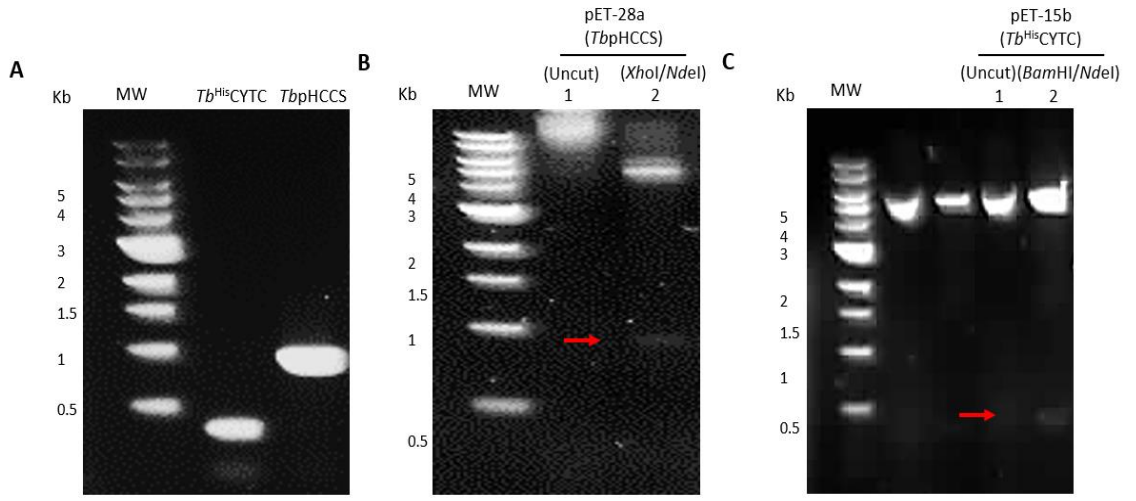


Figure 4.1. Cloning of *Tb*pHCCS and *Tb*^{His}CYTC coding sequences to pET-28a and pET-15b

(A) Agarose gel electrophoresis analysis of *Tb*^{His}CYTC and *Tb*pHCCS genes amplified by PCR. PCR yielded products of the predicted sizes. 5 µL of the PCR mixture was subjected to 0.7% agarose gel electrophoresis. (B-C) Agarose gel electrophoresis analysis of diagnostic restriction digest reaction of pET-28a and pET-15b vectors sub-cloned to *Tb*pHCCS and *Tb*^{His}CYTC inserts. (B) The diagnostic digestion of pET-28 cloned to *Tb*pHCCS. Lane 1 corresponds to undigested plasmid. Lane 2 corresponds to pET-28 digested using *Xho*I/*Nde*I. (C) The diagnostic digestion of pET-15 cloned to *Tb*^{His}CYTC. Lane 1 corresponds to undigested plasmid. Lane 2 corresponds to pET-15 digested using *Bam*HI/*Nde*I. The digestive solutions (20 µL) were subjected to electrophorese in a 0.7% agarose gel. MW: NEB 1 kb DNA Ladder.

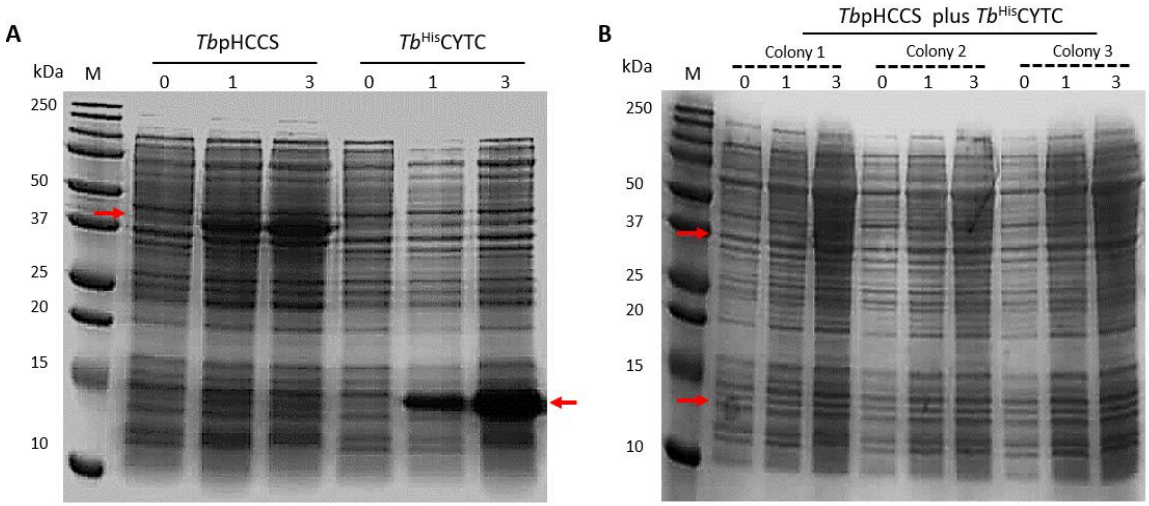


Figure 4.2. SDS-PAGE analysis of *Tb*pHCCS and *Tb*^{His}CYTC protein expression from the pET vectors

(A) SDS-PAGE analysis shows the expression of *Tb*^{His}CYTC and *Tb*pHCCS using separated aliquots of BL21 competent *E. coli*. (B) SDS-PAGE analysis shows the lack of co-expression of *Tb*^{His}CYTC and *Tb*pHCCS in the same aliquot of BL21 *E. coli*. Protein expression was induced by adding IPTG (1 mM). Bacterial pellets were collected 0, 1 and 3 hrs post induction and bacterial lysates were prepared by boiling SDS-PAGE loading buffer. Lysates were analysed on 10% polyacrylamide gels. Post SDS-PAGE, gels were stained with InstantBlue.

4.2 Co-expression of *Tb*pHCCS and *Tb*^{His}CYTC substrates – strategy 2

Following the realisation of the plasmid incompatibility issue, an alternative strategy to co-express *Tb*pHCCS and *Tb*^{His}CYTC was adapted. Here, I used pCDF-Duet vector available from Novagen. This plasmid has dual cloning sites which facilitate the expression of either one or simultaneously two distinct genes. The pCDF-Duet derived plasmids containing either *Tb*^{His}CYTC, *Tb*pHCCS, or both coding sequences were obtained via standard cloning procedures (Appendix 14). Cloned pCDF-Duet plasmids were transfected into Rosetta™(DE3) competent *E. coli* and protein expression was initiated by addition of IPTG (1 mM). Bacterial pellets were again collected at three different time points (0, 1, and 3 hrs) post induction and recombinant protein expression was assessed by SDS-PAGE. Figure 4.3 shows the accumulation of recombinant *Tb*^{His}CYTC and *Tb*pHCCS when each expressed individually (indicated by arrows). Three colonies were also picked, after the transformation of pCDF-Duet containing both *Tb*^{His}CYTC and *Tb*pHCCS coding sequences. Co-expression of both *Tb*^{His}CYTC and *Tb*pHCCS was successful, as shown in Figure 4.4 (indicated by arrows). However, here the accumulation of recombinant *Tb*^{His}CYTC appeared greater than *Tb*pHCCS.

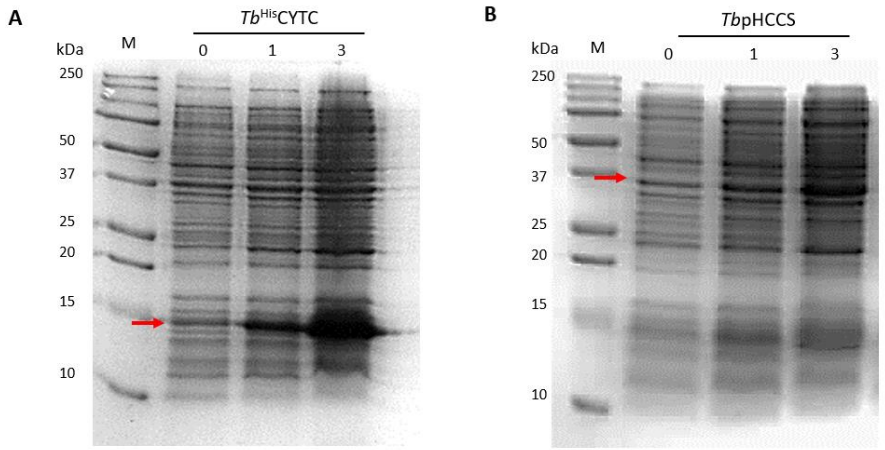


Figure 4.3. Independent expression of recombinant *Tb*^{His}CYTC and *Tb*pHCCS from pCDF-Duet vector

Protein expression was induced by adding IPTG (1 mM). Bacterial pellets were collected 0, 1 and 3 hrs post induction and bacterial lysates were prepared by boiling SDS-PAGE loading buffer. Lysates were analysed on 12% polyacrylamide gels. Post SDS-PAGE, gels were stained with InstantBlue. (A) SDS-PAGE analysis shows the expression of *Tb*^{His}CYTC. (B) SDS-PAGE analysis shows the expression of *Tb*pHCCS. The red arrows indicate the size of the induced proteins. M represents protein size marker in kDa.

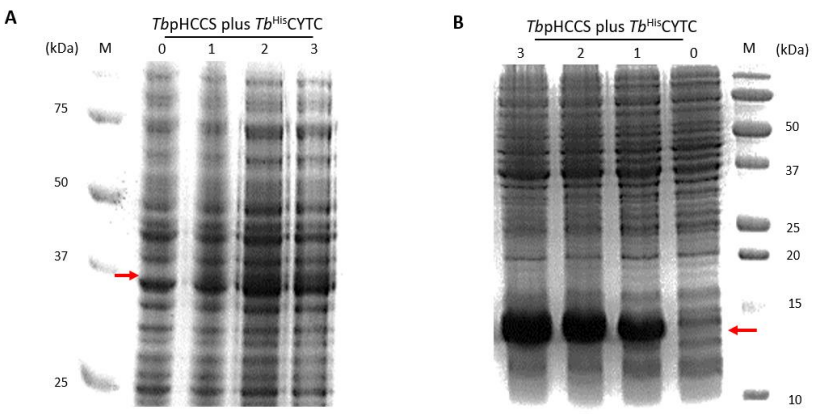


Figure 4.4. Co-expression of *Tb*^{His}CYTC and *Tb*pHCCS from the pCDF-Duet vector
The co-expression of pCDFDuet-1 sub-cloned to *Tb*^{His}CYTC and *Tb*pHCCS using Rosetta™(DE3) competent *E. coli*. Protein expression was induced by adding IPTG (1 mM). Bacterial pellets were collected 0, 1, 2 and 3 hrs post induction and bacterial lysates were prepared by boiling SDS-PAGE loading buffer. (A) The samples were subjected to 8% polyacrylamide gels to analyse the expression of recombinant *Tb*pHCCS protein. (B) The samples were subjected to 12% polyacrylamide gels to analyse the expression of recombinant *Tb*^{His}CYTC protein. InstantBlue used to stain the gels. M represents protein size marker in kDa. The red arrows indicate the size of the induced proteins.

4.2.1 Protein solubility

After confirming the recombinant co-expression of both *Tb*^{His}CYTC and *Tb*pHCCS, the next step was to test the solubility of *Tb*^{His}CYTC, in *E. coli* expressing both *Tb*^{His}CYTC and *Tb*pHCCS. Initially, Bugbuster (Novagene) was used to look for evidence of cytochrome *c* solubility (Method chapter section 2.5). As shown in Figure 4.5, the recombinant *Tb*^{His}CYTC appeared only as an insoluble protein. The apparent insolubility of *Tb*^{His}CYTC protein was then re-examined following ultra-sonication of harvested *E. coli* cultures (or pellets). As shown in Figure 4.6, *Tb*^{His}CYTC expressed partially in a soluble form when expressed along with *Tb*pHCCS. However, the level of solubility of *Tb*^{His}CYTC appear to differ when it was expressed on its own in *E. coli*. Expressed on its own, *Tb*^{His}CYTC was found to be both in soluble and insoluble forms in approximately equal amounts (Figure 4.6). The differences in solubility levels, might be due to the fact that some of *Tb*^{His}CYTC is in a complex with *Tb*pHCCS which causes some level of *Tb*^{His}CYTC insolubility.

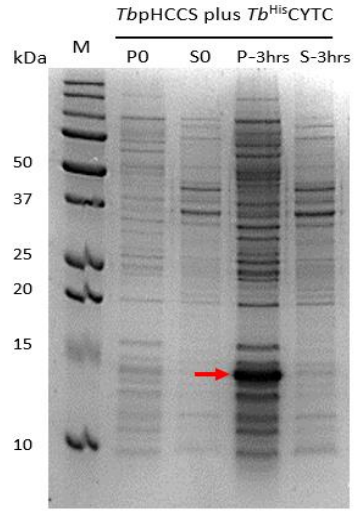


Figure 4.5. Protein solubility testing of *Tb*^{His}CYTC using Bigbuster

P0 and S0 correspond to pellet and supernatant respectively before protein induction. P-3hrs and S-3hrs correspond to bacterial pellet and supernatant (containing soluble proteins) respectively after 3 hrs post induction. The samples were analysed by SDS-PAGE using a 12% polyacrylamide gel. InstantBlue was used to stain the protein bands. M represents protein size marker in kDa.

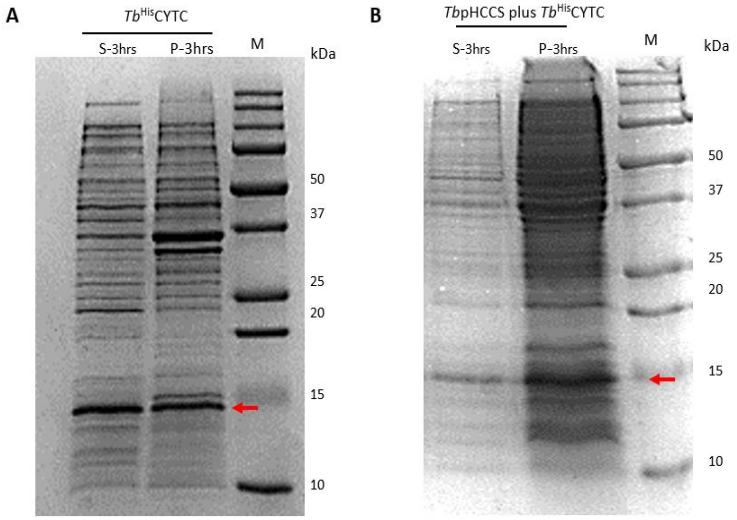


Figure 4.6. Protein solubility testing of *Tb*^{His}CYTC following ultra-sonication

(A) SDS-PAGE analysis shows *Tb*^{His}CYTC protein solubility test from a bacterial culture induced to expresses only *Tb*^{His}CYTC. (B) SDS-PAGE analysis shows *Tb*^{His}CYTC protein solubility test from a bacterial culture in which *Tb*^{His}CYTC and *Tb*pHCCS were co-expressed. P-3hrs and S-3hrs correspond to bacterial pellet and supernatant respectively after 3 hrs post induction. The samples were analysed by SDS-PAGE using a 12% polyacrylamide gel. InstantBlue was used to stain the protein bands. M represents protein size marker in kDa.

4.2.1 Protein purification

To determine whether heme was attached to *Tb*^{His}CYTC, this recombinant protein was purified using Ni^{+2}-NTA agarose affinity chromatography with purification suitable for native rather than denatured proteins (Method chapter section 2.5). *Tb*^{His}CYTC was purified from the large-scale growths of *E. coli* (1 L). Recombinant *Tb*^{His}CYTC was purified from a bacterial culture induced to expresses only *Tb*^{His}CYTC and from a bacterial culture in which *Tb*^{His}CYTC and *Tb*pHCCS were co-expressed. Following purification, the *Tb*^{His}CYTC proteins were subjected to acetone precipitation thereby readily increasing protein concentration for SDS-PAGE. The *Tb*^{His}CYTC protein pellets were resuspended using 100 µl of SDS-PAGE loading buffer, then 10 µl of the resuspended pellets were subjected to 12% polyacrylamide gels. As shown in Figure 4.7, the purified recombinant *Tb*^{His}CYTC proteins are only detectable after three sequential additions of 1 ml of elution buffer to the column (lanes 3 to 7 in Figure 4.7). The subsequent experiments were carried out to detect covalent heme attachment in the purified recombinant *Tb*^{His}CYTC.

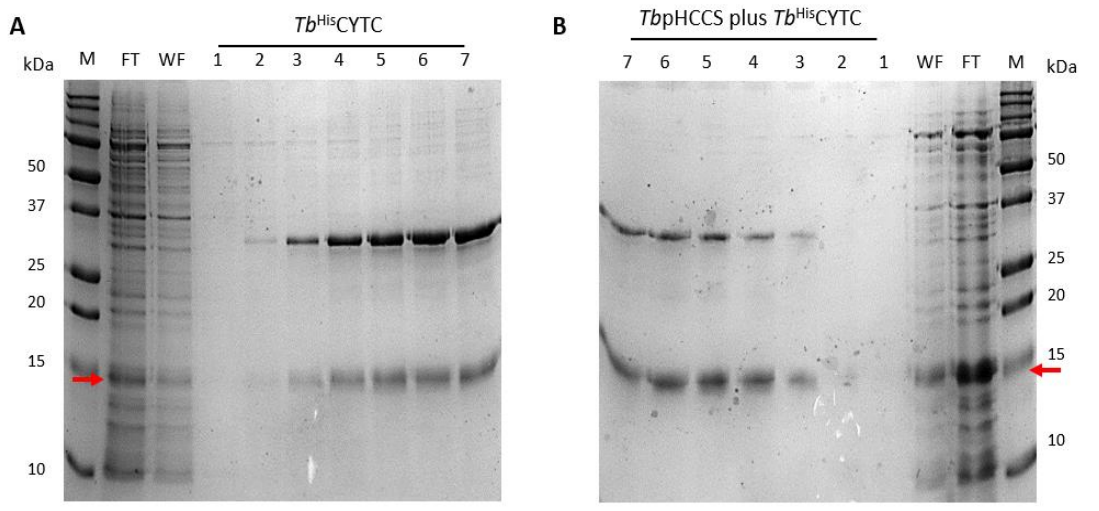


Figure 4.7. Purification of *Tb*^{His}CYTC using Ni^{+2}-NTA agarose affinity chromatography
Bacteria culture induced to express the recombinant *Tb*^{His}CYTC protein were lysed and bacterial supernatant was applied to Ni^{+2}-NTA column. The column was washed, and the recombinant protein was eluted using seven sequential additions of 1 ml of elution buffer (high imidazole concentration). (A) SDS-PAGE analysis shows *Tb*^{His}CYTC purification from bacterial culture that expresses pCDF-Duet containing *Tb*^{His}CYTC. (B) SDS-PAGE analysis shows *Tb*^{His}CYTC purification from the bacterial culture that expresses pCDF-Duet containing *Tb*^{His}CYTC and *Tb*pHCCS. FT corresponds to the column flow-through; WF corresponds to the samples collected after the addition of washing buffer; Lanes 1 to 7 correspond to samples after the addition of an elution buffer. Protein samples were then subjected to protein acetone precipitation. The protein pellets were resuspended using 100 µl of SDS-PAGE loading buffer. 10 µl of the samples were analysed by SDS-PAGE using a 12% polyacrylamide gel. M represents protein size marker in kDa.

4.3 Experimental methods to detect heme attachment to *Tb*HisCYTC

4.3.1 Heme staining of SDS-PAGE gel

The Heme 3,3',5,5'-Tetramethylbenzidine (TMBZ) staining technique was used to detect whether *Tb*pHCCS catalyses the heme attachment to *Tb*^{His}CYTC. Under mildly acidic conditions, holocytochrome *c* has peroxidase-like activity which can be detected using TMBZ staining technique. The peroxidase-like activity catalyses the reduction of hydrogen peroxidase (H_2O_2) to water by transferring $2H^+$ from TMBZ to H_2O_2. Oxidized TMBZ has a noticeable blue colour. This blue colour can be seen on SDS-PAGE gels (Miller and Nicholas, 1984).

Following the large-scale induction, the recombinant *Tb*^{His}CYTC was purified from a bacterial culture induced to expresses only *Tb*^{His}CYTC, and from a bacterial culture in which *Tb*^{His}CYTC and *Tb*pHCCS were co-expressed. The purified recombinant *Tb*^{His}CYTC proteins were then collected and concentrated by acetone precipitation. Precipitated purified *Tb*^{His}CYTC were resuspended using non-reducing SDS-PAGE loading buffer and were subjected to 12% non-reducing polyacrylamide gels. Reduced horse heart cytochrome *c* was used as control for heme staining. Figure 4.8 shows the SDS-PAGE gel images. Image A shows an InstantBlue stained SDS-PAGE gel to detect the complete bacterial protein content. Image B shows the gel stained using TMBZ stain; here only the presence of reduced holocytochrome *c* is detected. On the TMBZ stained gel image, bands were visible for horse heart cytochrome *c*, as well as the sample where *Tb*^{His}CYTC and *Tb*pHCCS were expressed simultaneously in *E. coli*. However, no bands were present on the sample where only *Tb*^{His}CYTC was expressed in *E. coli* even though image A clearly shows enough protein was present to be detected by heme stain if heme attachment had occurred. This provides further evidence that *Tb*pHCCS is used by trypanosomatids to mature their unusual-looking *c*-type cytochromes.

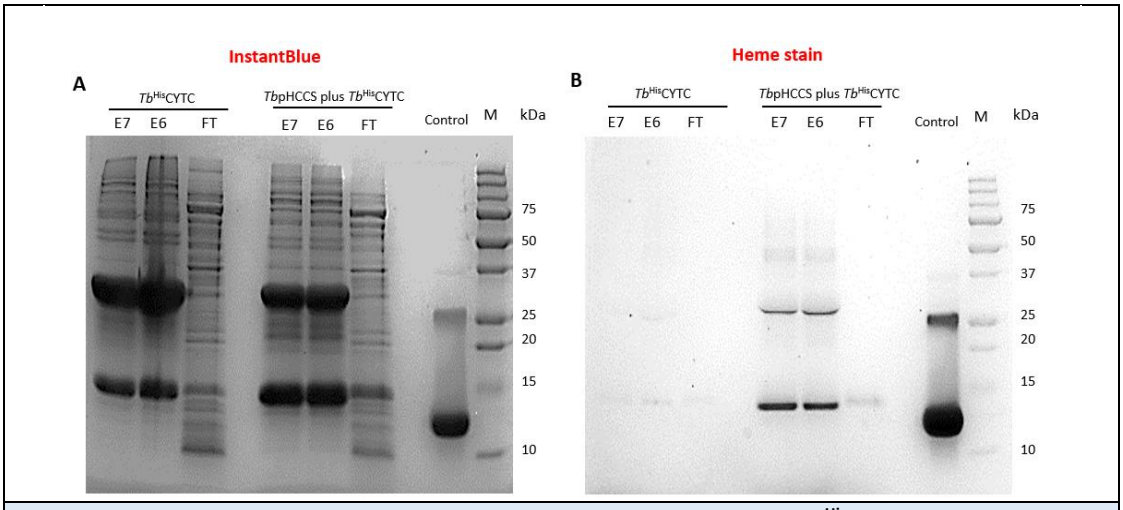


Figure 4.8. Detection of covalent heme attachment to purified soluble *Tb*^{His}CYTC

The control sample is horse heart cytochrome *c*. Soluble fractions of *Tb*^{His}CYTC were purified using Ni^{+2}-NTA agarose affinity chromatography. FT corresponds to the column flow-through; E6 and E7 correspond to eluted *Tb*^{His}CYTC fractions 6 and 7. Acetone precipitation was used to concentrate *Tb*^{His}CYTC, pellets were resuspended using 20 µl non-reducing SDS-PAGE loading buffer. Protein samples were subjected to non-reducing 12% polyacrylamide gel. (A) SDS-PAGE analysis shows the gel stained with InstantBlue and (B) shows the gel stained with TMBZ stain. *Tb*pHCCS plus *Tb*^{His}CYTC represent *Tb*^{His}CYTC isolated from bacteria culture expressing both *Tb*^{His}CYTC and *Tb*pHCCS genes. *Tb*^{His}CYTC represents *Tb*^{His}CYTC isolated from bacteria culture expressing *Tb*^{His}CYTC gene only. The protein bands around 25 kDa correlate to cytochrome *c* oligomers which are formed by successively swapping the C-terminal domains of cytochrome *c* to the equivalent domains of other cytochrome *c* molecules (Hirota et al., 2010). M represents the protein size marker in kDa.

4.3.2 UV-Vis spectroscopy to detect *Trypanosoma brucei* holocytochrome *c* formation

To characterise the covalent heme attachment to the recombinant *Tb*^{His}CYTC, pyridine hemochromagen assays were carried out as described by Berry and Trumpower (Berry and Trumpower, 1987). Initially, a control sample (horse heart cytochrome *c*) was used to test the assay sensitivity to holocytochrome *c*. A series of solutions with varying concentrations of the horse heart cytochrome *c* were prepared (5 µM, 15 µM, 20 µM and 40 µM). These solutions were used in the hemochromagen assay and analysed for their UV-Vis absorption spectra to detect the covalent heme attachment. Generally, the heme attaches to the *c*-type cytochromes via two thioether bonds in the heme binding motif. For heme in the control sample, the UV-Vis spectra show two absorption maxima, α-band maximum at 550 nm and β-band maximum at 520 nm. The diagnostic α and β absorption maxima of holocytochrome *c* were only detected when the concentration of horse heart cytochrome *c* was equal to or higher than 15 µM (Figure 4.9).

To compare relative sensitivity of TMBZ stain and UV-Vis spectroscopy to holocytochrome *c*, another TMBZ staining of SDS-PAGE gel was prepared and a series of different concentrations of horse heart cytochrome *c* (0.08 pM to 4 µM) were loaded into a 12% polyacrylamide gel. Figure 4.9 shows that TMBZ staining of SDS-PAGE gels can detect holocytochrome *c* of less than 0.1 µM. Thus, the TMBZ staining of SDS-PAGE gels appeared more sensitive to holocytochrome *c* than UV-Vis spectroscopy.

Consequently, a large amount of recombinant *Tb*^{His}CYTC was needed before carrying out pyridine hemochromagen assay. To produce enough *Tb*^{His}CYTC to be detected by UV-Vis spectroscopy, 8 L of bacterial culture induced to express both *Tb*^{His}CYTC and *Tb*pHCCS was prepared. *Tb*^{His}CYTC purification was completed as detailed previously (section 4.2). The purified *Tb*^{His}CYTC in a volume of 80 ml was concentrated to a volume of 0.5 ml in order to carry out the hemochromagen assay. Unlike TMBZ stain, acetone precipitation is not compatible with UV-Vis spectroscopy. Thus, another approach for protein concentration had to be adopted. Here eluted fractions of soluble *Tb*^{His}CYTC were concentrated using vivaspin 20 centrifugal concentrators (MWCO 3 kDa). Using a concentrator, 80 mL of purified *Tb*^{His}CYTC was concentrated to 0.5 ml. The concentrated *Tb*^{His}CYTC had a noticeable brown colour. An aliquot of the concentrated *Tb*^{His}CYTC was analysed by SDS-PAGE and TMBZ staining to confirm covalent heme attachment (Figure 4.10). The remainder of the *Tb*^{His}CYTC was then used for pyridine hemochromagen assay and UV-Vis spectroscopy. As shown in Figure 4.10, the horse heart cytochrome *c* exhibits the typical absorption maxima (α at 550 nm, β at 520 nm and soret at 414 nm) of heme with two covalent linkages in cytochrome *c* heme binding motif. In contrast, the absorption maxima of recombinant *Tb*^{His}CYTC was red shifted with α at 553 nm, β at 522 nm and soret at 415 nm. The red shifted α and β absorption maxima indicate that the *Tb*^{His}CYTC is attached to heme via a single thioether bond. These results confirm that trypanosomatids use a divergent HCCS to mature their mitochondrial cytochrome *c*.

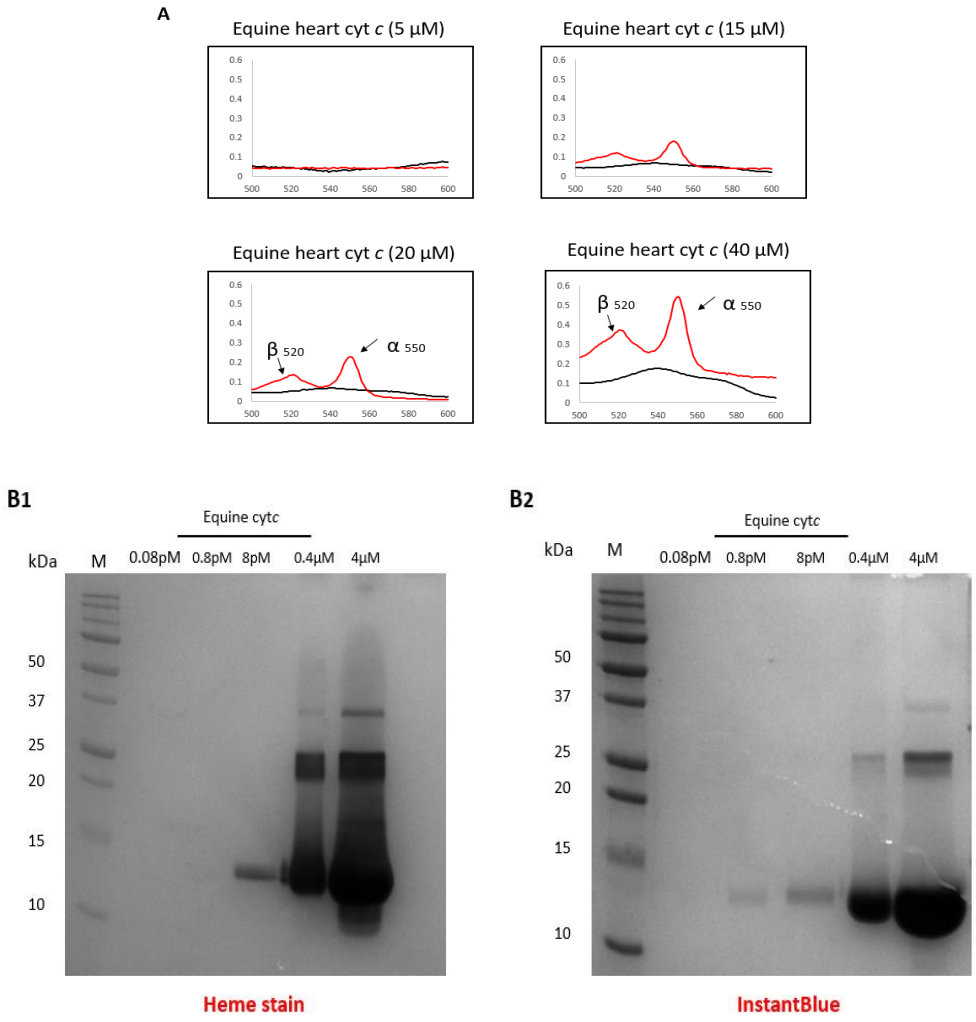


Figure 4.9. Comparing the sensitivity of TMBZ staining of SDS-PAGE gels and UV-Vis spectroscopy to holocytochrome *c*.

(A) UV-spectroscopy data show the absorption maxima of reduced and oxidized horse heart cytochrome *c* (equine cytc) using pyridine hemochromagen assay. Absorption spectra were collected using a Cary 4000 UV/Vis spectrophotometer. (B1) SDS-PAGE analysis shows the gel stained with heme (TMBZ) and (B2) SDS-PAGE analysis shows the gel stained with InstantBlue. Different concentrations of reduced horse heart cytochrome *c* were prepared using non-reducing SDS-PAGE loading buffer and were subjected to non-reducing 12% polyacrylamide gel. M represents protein size marker in kDa.

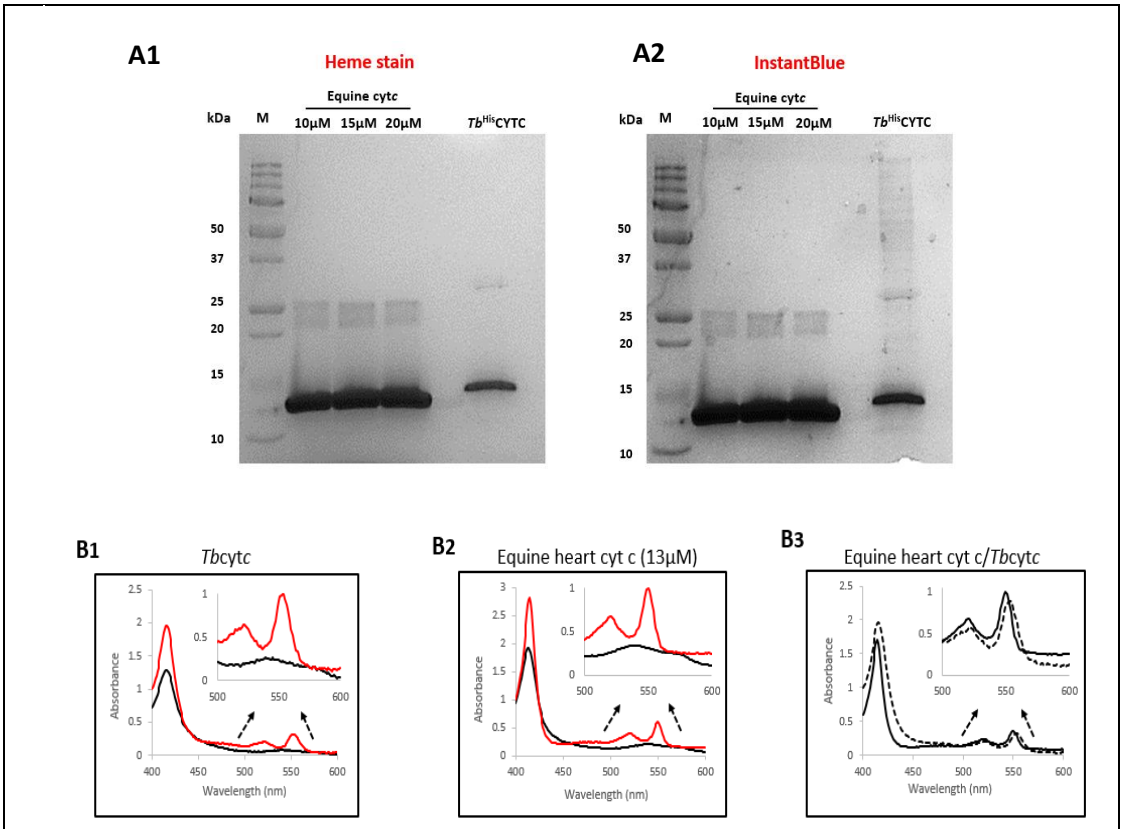


Figure 4.10. Detecting covalent heme attachment to the recombinant *Tb*^{His}CYTC

Large scale induction of *Tb*^{His}CYTC in bacterial culture expressing both *Tb*^{His}CYTC and *Tb*pHCCS was carried out. 80 ml of recombinant *Tb*^{His}CYTC was purified and concentrated to 0.5 ml. For *Tb*^{His}CYTC, 2 µl of concentrated *Tb*^{His}CYTC was mixed with 15 µl of non-reducing SDS-PAGE loading buffer. For the equine heart cytochrome *c*, different concentrations were prepared using non-reducing SDS-PAGE loading buffer. Samples were subjected to 12% polyacrylamide gels. (A1) SDS-PAGE analysis shows the gel stained with heme (TMBZ) and (A2) SDS-PAGE analysis shows the gel stained with InstantBlue. (B1) Absorption spectra of recombinant *Tb*^{His}CYTC, (B2) Absorption spectra of equine heart cytochrome *c*, and (B3) the spectra of both *Tb*^{His}CYTC and equine heart reduced cytochrome *c*. The inset shows the expanded regions of α and β absorption maxima. The red line shows the reduced spectra and the black line shows the oxidized spectra. The solid line shows the reduced spectra of equine heart cytochrome *c* and the dotted line shows the reduced spectra of *Tb*^{His}CYTC. Absorption spectra were collected using a Cary 4000 UV/Vis spectrophotometer.

4.4 Co-expression using pCDF-Duet vector without His-tag tail in recombinant *T. brucei* apocytochrome *c*

At this stage, the recombinant *T. brucei* apocytochrome *c* was expressed with N-terminal His_6-tag. The extra His residues in the *Tb*^{His}CYTC may affect the overall fold of the protein, which in turn could interfere with the interaction between *Tb*^{His}CYTC and *Tb*pHCCS. This may influence the activity level of *Tb*pHCCS and the formation of holocytochrome *c*. Additionally, His residues in His_6-tag can bind to the metal ion in the heme group, which may cause a false-positive result in TMBZ staining of SDS-PAGE gels. However, this possibility was eliminated by the fact that bands from TMBZ staining of SDS-PAGE gel were not observed when *Tb*^{His}CYTC was induced in bacteria culture expressing *Tb*^{His}CYTC only (Figures 4.8). Also, the absorption spectra result show that the heme is covalently attached via a single thioether bond in the heme-binding motif of *Tb*^{His}CYTC (Figure 4.10).

To try to improve the activity of *Tb*pHCCS and increase the rate of holocytochrome *c* formation, an attempt was made to express *Tb*CYTC without His_6-tag. The pCDF-Duet derived plasmids containing either *Tb*CYTC or both *Tb*CYTC and *Tb*pHCCS coding sequences were obtained via standard cloning procedures (Appendix 14). Cloned pCDF-Duet plasmids were transfected into Rosetta™(DE3) competent *E. coli* and protein expression was initiated by the addition of IPTG (1 mM). Bacterial pellets were collected at three different time points (0, 1, and 3 hrs) post-induction and recombinant protein expression was assessed by SDS-PAGE. Figure 4.11 shows a significant decrease in *Tb*CYTC expression relative to the level of *Tb*^{His}CYTC expression (Figure 4.4). The low level of *Tb*CYTC expression implies that measuring the activity rate of *Tb*pHCCS will be difficult. Thus, efforts were focused on cell culture methods to examine the phenotype(s) of pHCCS gene deletion on trypanosomatid parasites.

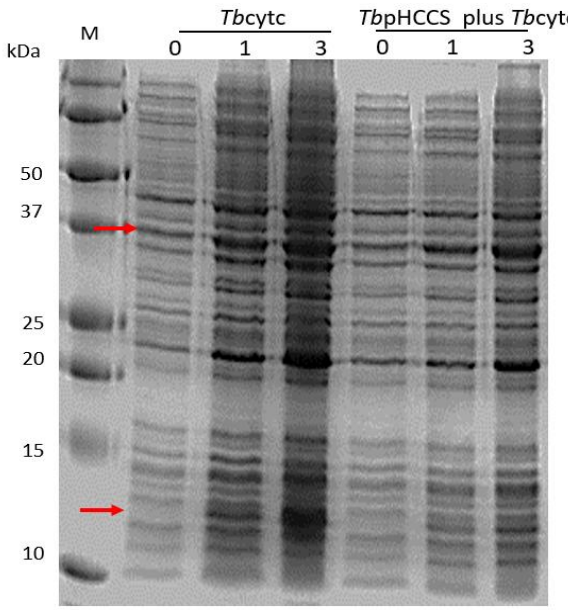


Figure 4.11. Co-expression of *Tb*CYTC (without His_6tag) and *Tb*pHCCS from the pCDF-Duet vector

The co-expression of pCDFDuet-1 subcloned to *Tb*CYTC and *Tb*pHCCS using Rosetta™(DE3) competent *E. coli*. Protein expression was induced by adding IPTG (1 mM). Bacterial pellets were collected 0, 1 and 3 hrs post induction and bacterial lysates were prepared by boiling SDS-PAGE loading buffer. InstantBlue was used to stain the protein content. M represents protein size marker in kDa. The red arrows indicate the position of the induced proteins.

4.5 Genome editing in *Leishmania mexicana* using CRISPR-Cas9

The deletion of the pHCCS gene in vitro will help determine whether this gene is vital for cell growth or whether it has a distinct phenotype that indicates its physiological role. If pHCCS gene is used to produce holocytochrome *c*, then the primary target of this gene is mitochondria and deleting it should be lethal for trypanosomatids. Using cell culture methods, an attempt to produce a null mutant of pHCCS gene in *Crithidia fasciculata* was made. To produce a null mutant of *pHCCS*, both copies of pHCCS gene need to be deleted and replaced by drug-resistant cassettes via homologous recombination. The drug-resistant cassettes are essential for cell selection during the transfection process. There were a few complications in the process of producing *pHCCS* null mutant in *C. fasciculata*. A key issue was that *C. fasciculata* is resistant to some routinely used selective drugs including puromycin, phleomycin and blasticidin. This was surprising given that there is literature report of using phleomycin to select for stable, genetically manipulated transformants (Tetaud et al, 2002). This made it difficult to find three selective drugs: two for use in producing the null double knockout, (*C. fasciculata* from genome analysis is diploid for single-copy pHCCS) and a third

drug-selectable marker for use to rescuing any null mutant phenotype or confirming essentiality of *Cf*HCCS following ectopic expression of a rescue copy. The second issue was to obtain the optimal parameters for electroporation. Different parameters were used in an attempt to produce the null mutant, however, *C. fasciculata* cells were either killed during the selection process or survived but kept both copies of pHCCS genes (data is not shown – and this issue was eventually resolved in obtaining transformants to look at the import capacity of cryptic 'PTS1' motifis as discussed in Chapter 6). Thus, an alternative strategy was commenced to test the essentiality (or otherwise) of pHCCS in a trypanosomatid parasite. This alternative strategy involved the application of the 'clustered regularly interspaced short palindromic repeats' (CRISPR) technique in *Leishmania mexicana*. A series of gene manipulation studies on *L. mexicana* using the CRISPR-Cas9 approach, were performed in the Centre of Infection and Immunity in the University of York by Dr Rachel Neish in the laboratory of Prof Jeremy Mottram. These experiments aim to GFP-tag and produce a null mutant of HCCS gene in *L. mexicana* (*Lm*pHCCS) using suitable selectable markers.

The following experiments were performed by Dr Rachel Neish using *L. mexicana* cells (promastigotes). To determine the location of *Lm*pHCCS in *L. mexicana*, *Lm*pHCCS was fused in-frame at its C-terminus to mNeon Green and transfected into *L. mexicana* cell line expressing CRISPR-Cas9 from an endogenous chromosomal locus. Once stable transformants were established, promastigotes were incubated with a Mitotracker Deep red^{633} and confocal fluorescence microscopy images were recorded. Figure 4.13 shows that the fluorescence signal for *Lm*pHCCS::Neon$_{Green}$ colocalised with Mitotracker Deep red^{633} . This mitochondrial localisation was consistent with previous reports of System III proteins being present in the mitochondrial intermembrane space of yeast where maturation of cytochrome *c* takes place (Diekert et al., 1999). The microscopy, thus, provided additional evidence that trypanosomatid pHCCS is used for cytochrome *c* maturation.

To examine the essentiality of pHCCS, a null mutant of pHCCS gene in *L. mexicana* was required. Two drug-resistant cassettes were used to replace the two copies of *Lm*pHCCS via homologous recombination, and two single guide RNAs (sgRNAs) were used to target Cas9 enzyme so it cuts at a specific site within the genome sequence. Although transfection of *Leishmania* parasites took place in York, I prepared the required DNA. The sgRNAs and the drug resistance cassettes were prepared by fusion PCR using primers generated using the

leishgedit tool (Method chapter section 2.10). The PCR products were combined to one tube and purified using PCR purification kit (Figure 4.12). Eluted DNA was then sent to the University of York for gene knockout experiments.

In York, a series of experiments are ongoing; the available data all point to the essentiality of pHCCS for promastigote *L. mexicana*. Outcomes from the initial experiments where I was involved in assessing or mapping the integration events are summarised below: (i) attempts to make a Lm*HCCS* null mutant using CRISPR-Cas9; (ii) attempts to make a chromosomal Lm*HCCS* null mutant with the phenotype of gene deletion rescued by expression of *Lm*HCCS::eGFP from an episome; and (iii) attempts to make a chromosomal Lm*HCCS* null mutant with the phenotype of gene deletion rescued by expression of trypanosome alternative oxidase (*Tb*AOX::eGFP) from an episome. This last experiment was an attempt to reengineer the *Leishmania* respiratory chain to function in the absence of *c*-type cytochromes but with an enzyme (alternative oxidase) capable of passing electrons to the terminal acceptor O_2 – *i.e.* production of a *Leishmania* mutant with a mitochondrial respiratory chain similar to that of *Phytomonas*.

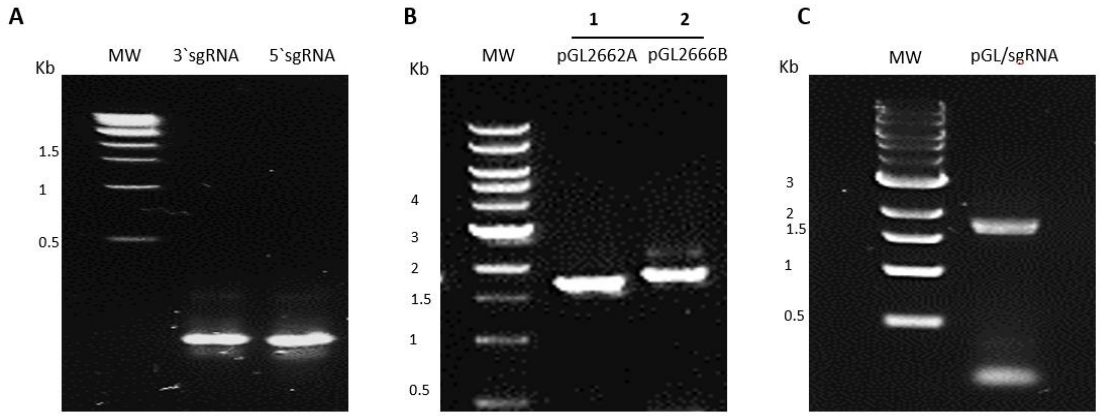


Figure 4.12. Agarose gel of CRISPR-Cas9 DNA constructs used to knockout pHCCS gene in *L. Mexicana*
(A) Agarose gel electrophoresis analysis shows the amplified 5` and 3` sgRNAs where 2 µl of the 20 µl PCR product were loaded on 2% agarose gel. (B) Agarose gel electrophoresis analysis shows the amplification of repair cassettes. 1 represents the blasticidin (BSD) repair cassette where pGL2662 plasmid was used as DNA template in the PCR. 2 represents the puromycin (BSD) repair cassette where pGL2666 plasmid was used as DNA template in the PCR. 2 µl from each of the 40 µl PCR product was loaded on 0.8% agarose gel. (C) Agarose gel electrophoresis analysis shows the presence of both sgRNAs and the repair cassette products (pGL) after the purification step using PCR purification kit.

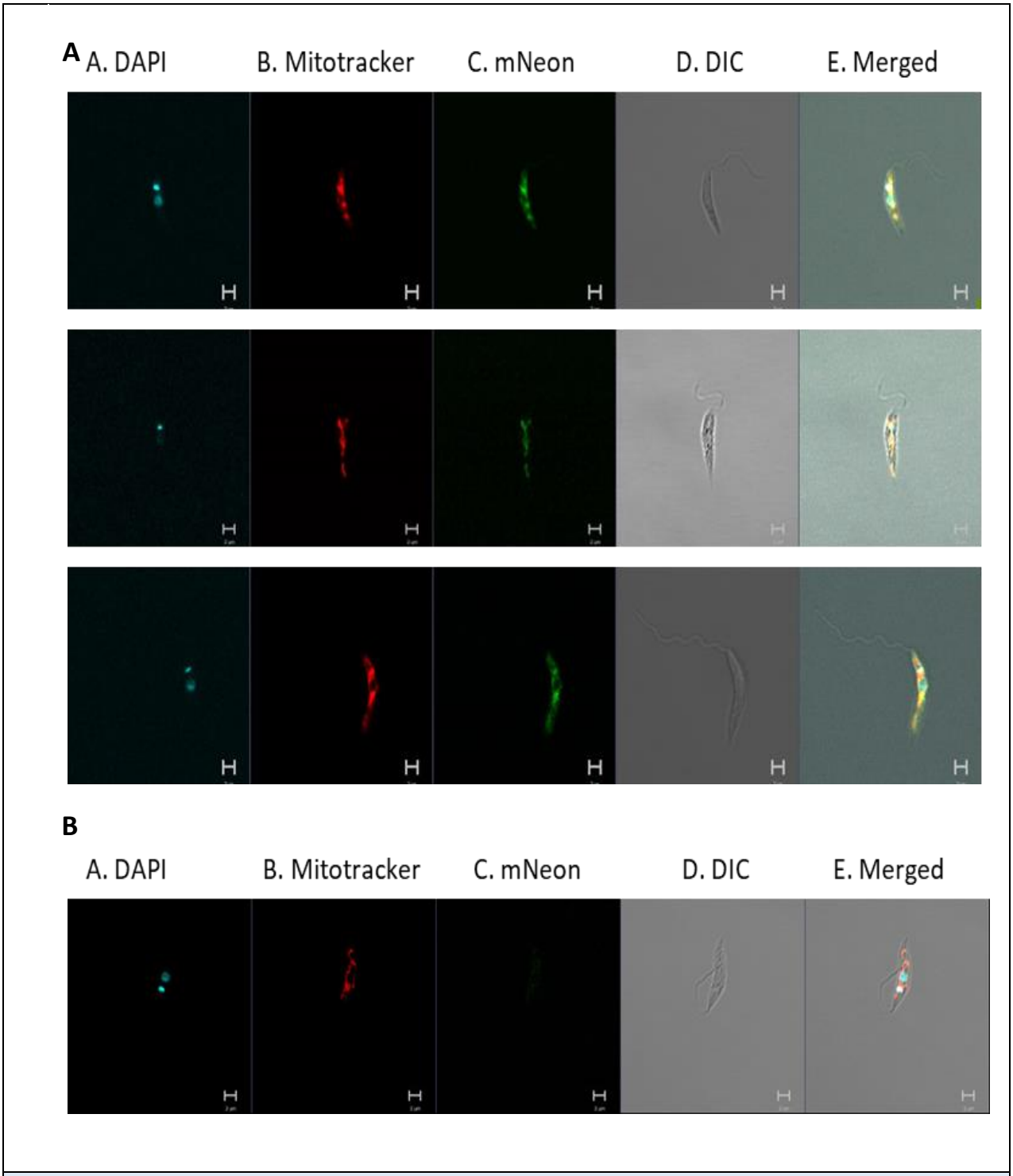


Figure 4.13. Targeting *Lm*pHCCS in *L. Mexicana*

Dr Rachel Neish generated the above images. (A) Mitochondrial localization of tagged *Lm*pHCCS. The *L. mexicana* cells were transfected with 3`sgRNA template and donor DNA (mNeonGreen+3Myc tag with BSD resistance – for C terminal tagging). (B) *L. mexicana* cell line expressing CRISPR-Cas9 stained with DAPI and mitochondria marker (control).

4.5.1 Diagnostic PCRs to check presence/absence of *Lm*pHCCS gene and the integration of drug resistant cassettes

Four independent transfections were prepared. Two weeks following transfection, cells started to steadily grow. All four independent transfections were growing at a normal rate. To inspect the presence and/or the absence of *Lm*pHCCS gene from the knockout cell lines, DNA templates were isolated from each transfection two weeks' post-electroporation and a series of PCR mappings were completed. The genomic DNA for wild type and two independent KO cell lines were received from Dr Rachel Neish. Primers were designed to exclusively map the replacement of *Lm*pHCCS gene by the drug resistance cassettes and they were used in four different PCRs (reactions A to D) (Figure 4.14). Primers within *Lm*pHCCS coding sequence were used in reactions A and B. If both copies of *Lm*pHCCS gene were deleted, PCR products for reaction A and B are expected only when the wild type genomic DNA is used and not when the KO genomic DNA is used. However, the gel image in Figure 4.14 shows that even when the KO genomic DNA was used, there were PCR products in reactions A and B, which indicate that the KO cell lines still obtain a copy of *Lm*pHCCS gene. The inability to produce null mutant in *L. mexicana* suggests that *Lm*pHCCS gene is essential. In reaction D, primers mapping the flanking region of *Lm*pHCCS gene were used. In theory, there should be a DNA product of reaction D when either the wild type or KO genomic DNA is used. However, there were PCR products in reaction D only when the genomic DNA of the KO cell lines were used (Figure 4.14). The PCR products from reaction D were sub-cloned to pGEM-T Easy and were sent for sequencing (Appendix 15). From the nucleotide sequence alignment in Appendix 15, the KO cell lines still obtain a copy of the *Lm*pHCCS gene. The presence of PCR products in reaction D when KO genomic DNA was used but not when the wild type genomic DNA was used, might indicate genome duplication in the KO cell lines. This behaviour of genome plasticity is very common and can be seen when pressure is applied to delete an essential gene in *L. mexicana* (Ishemgulova et al., 2018). To test this theory, flow cytometry was carried out by Dr Rachel Neish. The data shows genome duplication in the KOX and Y cell lines, but it was absent in the wild type (data not shown). This provides evidence that *Lm*pHCCS gene is critical for cell validity.

4.5.2 Attempt to facilitate the production of pHCCS null mutant in *L. mexicana* using episomal expressed pHCCS

An additional experiment was carried out to see whether the increase in *Lm*pHCCS gene expression using episomal vector would allow the production of null mutant. Episomal vector compatible with *L. mexicana* cells was supplied by Dr Rachel Neish (pGI2785 plasmid). High fidelity PCR was carried out to amplify the *Lm*pHCCS coding sequence from wild type *L. mexicana* genomic DNA. *Lm*pHCCS coding sequence was amplified, however, small fragments of DNA were observed in the agarose gel (Figure 4.15). To eliminate the possibility of ligating these small DNA fragments to pGI2785 plasmid instead of *Lm*pHCCS coding sequence, the PCR product was ligated into pGEM-T Easy first. Plasmids were isolated and a diagnostic digest using *Eco*RI was carried out to check for the release of insert DNA. The pGEM-T Easy plasmids containing *Lm*pHCCS coding sequence (0.89 Kb) were used for bulk digest using *Avr*II. Gel extraction was done to separate *Lm*pHCCS coding sequence from pGEM-T Easy vector. The *Lm*pHCCS coding sequence was then cloned into plasmid pGI2785 (Figure 4.15) and sent to Dr Rachel Neish to carry out the transfection experiments.

4.5.3 Attempt to facilitate the production of pHCCS null mutant in *L. mexicana* using episomal expressed *T. brucei* alternative oxidase

An additional experiment was carried out to test if the loss of *Lm*pHCCS function can be compensated by the expression of *T. brucei* alternative oxidase (*Tb*AOX). Due to the vital role of *Lm*pHCCS gene, the production of the double knockout was problematic. In order to produce double knockout, *Lm*pHCCS must become a nonessential gene to the cells. An attempt to express *T. brucei* alternative oxidase in *L. mexicana* prior to CRISPR-Cas9 knockout of *Lm*pHCCS was made. Alternative oxidase is an enzyme found as part of the mitochondrial electron transport chain in numerous protists including *T. brucei* bloodstream form. This enzyme provides an alternative way to reduce oxygen by making the cell reliant only on complex I for proton pumping to set mitochondrial inner membrane potential (Figure 4.16). Such ectopic expression of the alternative oxidase may allow *L. mexicana* to grow without the need for *c*-type cytochromes. If this was the case, then *Lm*pHCCS will become non-essential and can be deleted using CRISPR-Cas9. If the function of *Lm*pHCCS could be replaced by the activity of *Tb*AOX, this will provide evidence of the possibility of re-engineering the mitochondria electron transport chain in *Leishmania*. Episomal vector compatible with *L.*

*mexicana (*pGI2785 plasmid) cloned to *Tb*AOX coding sequences was obtained via standard cloning procedures (Figure 4.17). The cloned plasmid was sent to Dr Rachel Neish to test the effect of expressing *Tb*AOX on producing *Lm*pHCCS null mutant.

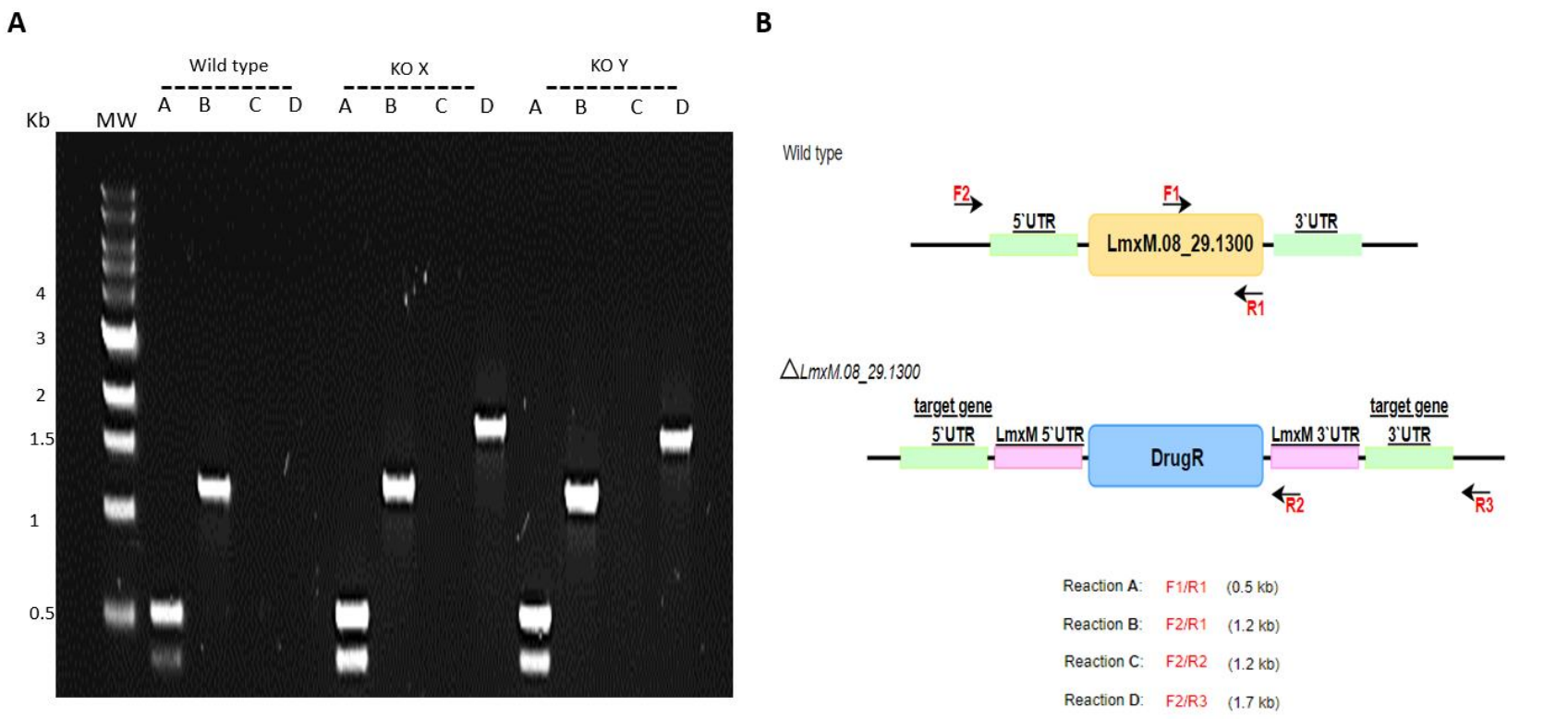


Figure 4.14. PCR Screening for the replacement of *Lm*pHCCS gene by drug resistance cassettes

(A) Agarose gel electrophoresis analysis shows DNA products for PCR mapping. Four PCRs (A, B, C and D) were carried out using DNA template of two independent knockout cell lines KOX and Y, and wild type genomic DNA. The 2 µl from each of the PCR products were loaded on 0.7% agarose gel. MW: NEB 1 kb DNA Ladder. (B) Schematic diagram shows the primers (arrows) used for PCR mapping and the size of the expected products. Reaction A and B contain reverse primer located within *Lm*pHCCS coding sequence (LmxM.08_29.130), reaction C contains reverse primer located within the resistance cassettes and reaction D contains reverse primer located outside the resistance cassettes. If the production of *Lm*pHCCS null mutant had been successful, PCR products would not be detected for reaction A and B using the gDNA from candidate null mutant KOX and Y.

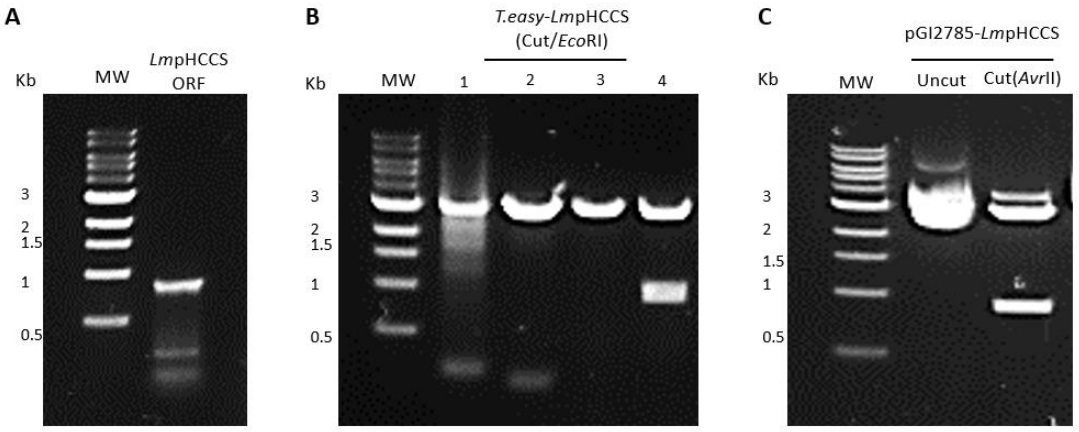


Figure 4.15. Cloning of *Lm*pHCCS coding sequence to the episomal pGI2785 plasmid

(A) PCR amplification of the *Lm*pHCCS coding sequence. High fidelity polymerase was used in PCR. To isolate *Lm*pHCCS coding sequence from non-specific DNA bands, the PCR product was cloned to pGEM-T Easy. (B) Diagnostic digest of cloned pGEM-T Easy vectors using *Eco*RI enzyme. Columns 1-4 represent different colonies used to isolate pGEM-T Easy plasmid. Plasmid isolated from colony 4 has an insert size of 0.89 kb which is the size of *Lm*pHCCS coding sequence. Bulk digest using *Avr*II enzyme was carried out and gel extraction was used to purify *Lm*pHCCS. The purified *Lm*pHCCS insert is then used for pGI2785 plasmid cloning. (C) Agarose gel electrophoresis shows a diagnostic digest using *Avr*II of the cloned pGI2785 plasmid. The cloned plasmid was sent for sequencing to confirm the sequence of *Lm*pHCCS coding sequence. The samples were subjected to electrophorese in a 0.8% agarose gel. MW: NEB 1 kb DNA Ladder.

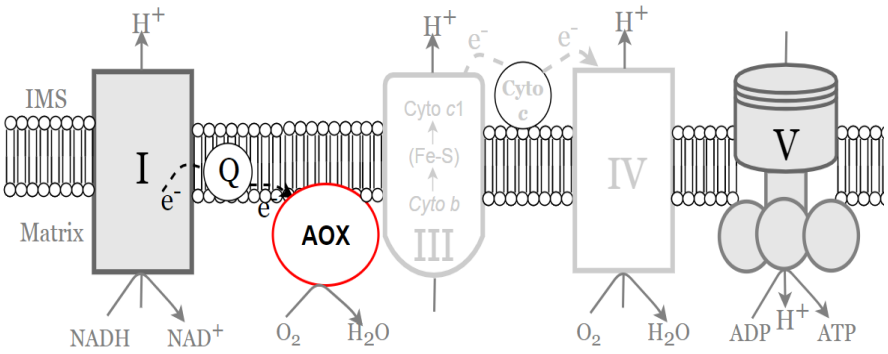


Figure 4.16. Modified mitochondrial electron transport chain showing the location of alternative oxidase (AOX)

The AOX will introduce an alternative pathway in which the reduced ubiquinone passes electrons to AOX instead of complex III. The modified electron transport chain will then be independent of complex III and IV.

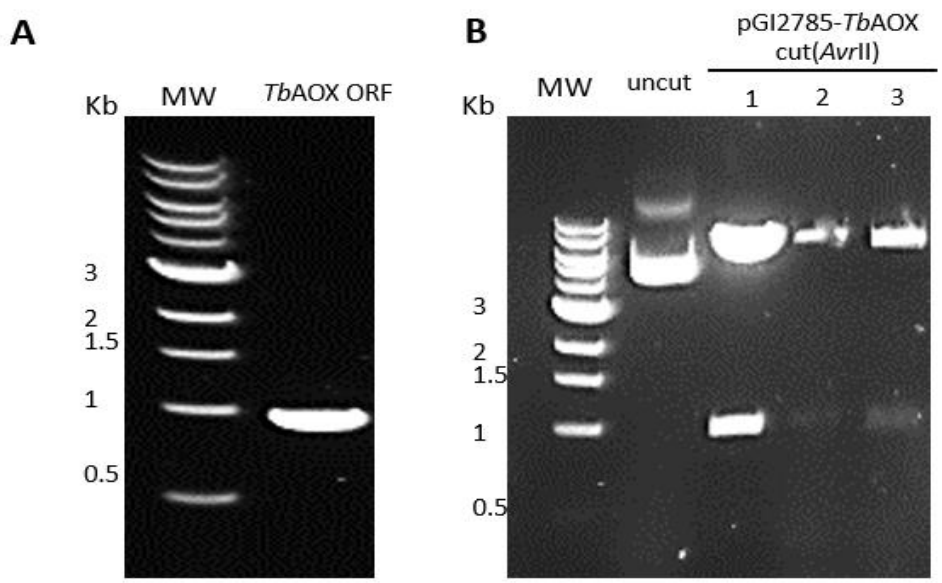


Figure 4.17. Cloning of *Tb*AOX coding sequence to the episomal pGI2785 plasmid

(A) The PCR amplification of the *Tb*AOX coding sequence. High fidelity polymerase was used in the PCR. (B) Agrose gel electrophoresis analysis of diagnostic digestion of cloned pGI2785 plasmid using *Avr*II. The cloned plasmid was sent for sequencing to confirm the sequence of *Tb*AOX. The samples were subjected to electrophorese in a 0.8% agarose gel. MW: NEB 1 kb DNA Ladder.

Chapter 5

Bioinformatics-led reassessment of atypical mitochondrial metabolism in *Naegleria gruberi*

The unexpected divergence of the trypanosomatid *c*-type cytochrome maturation system, coupled to the relative genetic intractability that I encountered for *Crithidia fasciculata* prompted me to investigate and screen for atypical metabolism in other eukaryotes. In the next two chapters, the investigation of unusual central energy metabolism in protists with particular focus on *Naegleria gruberi* is documented. *Naegleria* belongs to Heterolobosean amoebae which are evolutionarily divergent eukaryotes, distantly related to trypanosomatid parasites and other kinetoplastids (Yang et al. 2017).

One set of metabolic adaptations found recurrently in diverse microbial eukaryotes, secondarily adapted to grow under microaerophilic or anoxic conditions, is the reductive evolution and loss of the mitochondrial respiratory chain, the presence of pathways to oxidise NADH produced through catabolism with concomitant production of H_2, and the importance of acetate as end-products of energy metabolism (Müller et al., 2012, Stairs et al., 2018, Lewis et al., 2019). Thus far, H_2 generation in heterotrophic eukaryotes that are considered to grow aerobically appears extremely rare and is limited to *Naegleria* and *Acanthamoeba castellanii* (Leger et al., 2013, Tsaousis et al., 2014, Fritz-Laylin et al., 2010). Published *N. gruberi* genome data previously provided a unique and unanticipated insight into a heterotrophic metabolism that appeared capable of classic aerobic mitochondrial cytochrome *c*-dependent respiration, anaerobic NO_2^- metabolism, anaerobic FeFe-hydrogenase-dependent H_2 production, and production of acetate as a candidate end-product of energy metabolism, at least under some circumstances (Fritz-Laylin et al., 2010, Tsaousis et al., 2014). For other circumstances, CO_2 production as a consequence of an intact citric acid cycle was also apparent.

N. gruberi is a free-living predatory amoeba, common in moist soils and freshwater around the world. In response to different environmental stimuli, it can differentiate into a non-replicative motile flagellate form or encyst. This protist belongs to the phylum Heterolobosea and from an evolutionary sense, Heterolobosea is a member of the Discoba group (Burki et al., 2019). The results from genome-based phylogenetic analyses and cytoskeletal morphology, support an existing hypothesis that the phylogenetic position of Discoba lies

close to or within the eukaryotic last common ancestor (Yabuki et al., 2018, He et al., 2014, Cavalier-Smith, 2018). Thus, *N. gruberi* is considered as evolutionarily very divergent and is potentially descended from an early-branching eukaryotic lineage. A significant portion of the biology of *Naegleria* may therefore be reflective of the last common ancestor of all eukaryotes (Koonin, 2010). Thus, if the apparent metabolic flexibility of *N. gruberi* was backed by experimental data, *N. gruberi* could be an ideal candidate to study and understand the basic implications of the metabolic repertoire of the last common ancestor or other early-evolving eukaryotes.

Annotation of nuclear and mitochondrial genomes in *N. gruberi* suggested that under aerobic conditions, *N. gruberi* oxidises carbohydrates to CO_2 and H_2O via glycolysis and citric acid cycle functions (Fritz-Laylin et al., 2010). Amino acids and fatty acids provide alternative carbon sources for energy generation and for the provision of precursors for citric acid cycle activity. Yet, in conditions where oxygen is limited or absent, *N. gruberi* would partially rely on anaerobic glycolysis pathway to generate ATP, and in order to keep the glycolysis active, NADH needs to oxidise back to NAD^+ via a variety of possible pathways. For instance, cytosolically produced NAD^+ could likely be regenerated via phosphoenolpyruvate carboxykinase which converts phosphoenolpyruvate to oxaloacetate; oxaloacetate could then be reduced to malate by malate dehydrogenase; malate shuttled to mitochondria, and oxidised to acetyl-CoA with acetate production coupled to ATP production via a variety of candidate enzymes. For instance, Fritz-Laylin and co-workers have suggested that acetyl-CoA can be converted by *Naegleria* to acetate using an (ADP-forming) acetyl-CoA synthetase (ACD) (Figure 5.1) (Fritz-Laylin et al., 2010). For every one mole of acetate produced, one mole of ATP would also produced by substrate-level phosphorylation (Jones et al., 2017). In addition, [FeFe] hydrogenase activity could also contribute to redox (NAD:NADH) balance in cytosol and mitochondria under anaerobic conditions (Tsaousis et al., 2014; Fritz-Laylin et al., 2010; Opperdoes et al., 2011). *N. gruberi* nuclear genome annotation also revealed the presence of a gene coding for copper-containing nitrite reductase (CuNiR). CuNiRs reduces nitrite (NO_2^-) to nitric oxide (NO); in many bacteria this is a key enzyme in the denitrification process and anaerobic respiration. In some fungi, this enzyme can be used as a terminal electron acceptor in the ETC in environments where O_2 is limited (Kim et al., 2009). Outside of the fungi, reports

of CuNiR or other NO_2^- reductase in heterotrophic eukaryotes are very limited (Woehle et al., 2018).

In this chapter I revisit the predicted facultative anaerobic metabolism of *N. gruberi*. Utilising published literature and publicly available eukaryotic transcriptome and genome sequence data to first map the phylogenetic distribution of ACD and CuNirK. With genome and/or transcriptome data for a wide range of microbial eukaryotes now sampled, the aim was to revisit how common ACD and CuNirK might be among eukaryotes, particularly since ACD and CuNirK enzymes are widely distributed in bacteria and archaea. My analysis revealed ACD and CuNirK are sparsely distributed in microbial eukaryotes (Figure 5.2). I found both enzymes co-exist in some amoebozoans (*e.g. Entamoeba histolytica and Acanthamoeba castellanii*). However, only in *N. gruberi*, do both candidate enzymes appear in the same species. Moreover, in revisiting the distribution of ACD and looking closely at *Naegleria* orthologs I suggest the proposed activity, production of acetate from acetyl-CoA with the concomitant production of ATP, is not correct. I discuss an alternative view that this *Naegleria* protein acts as a protein acetyltransferase (or PAT).

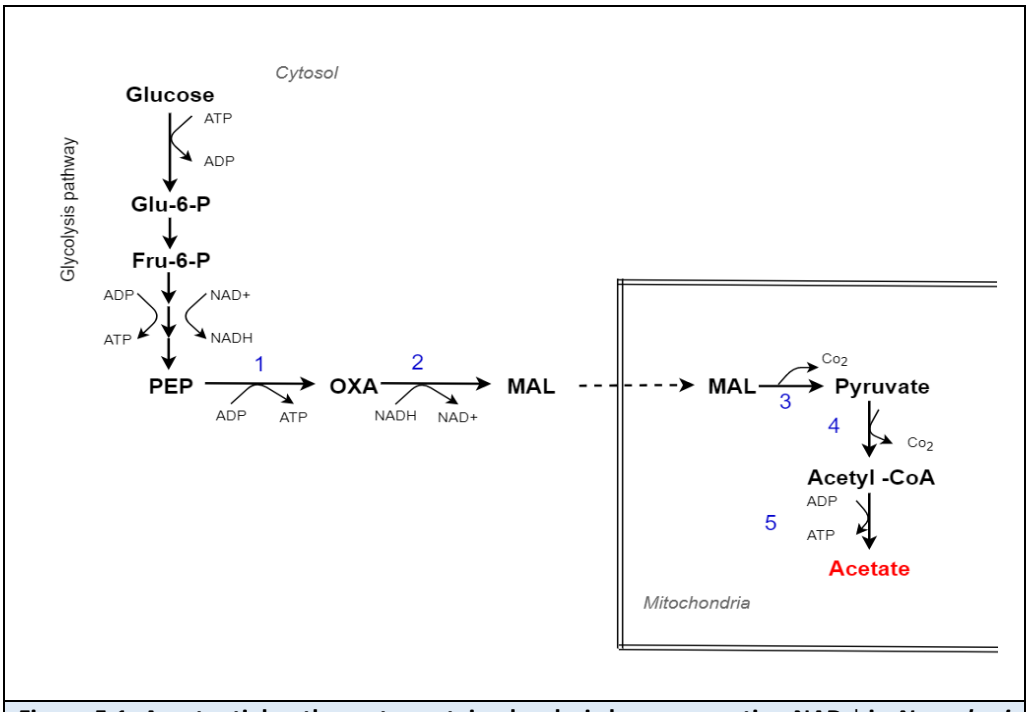


Figure 5.1. A potential pathway to sustain glycolysis by regenerating NAD $^+$ in *N. gruberi* during anaerobic conditions

Abbreviations: Glu-6-P, Glucose 6-phosphate; Fru-6-P, Fructose 6-phosphate; PEP, Phosphoenolpyruvate; OXA, Oxaloacetate; MAL, Malate; (1), Phosphoenolpyruvate carboxykinase; (2), Malate dehydrogenase; (3), Malic enzyme; (4), Pyruvate dehydrogenase complex; (5), Acetyl-CoA synthetase [ADP- forming].

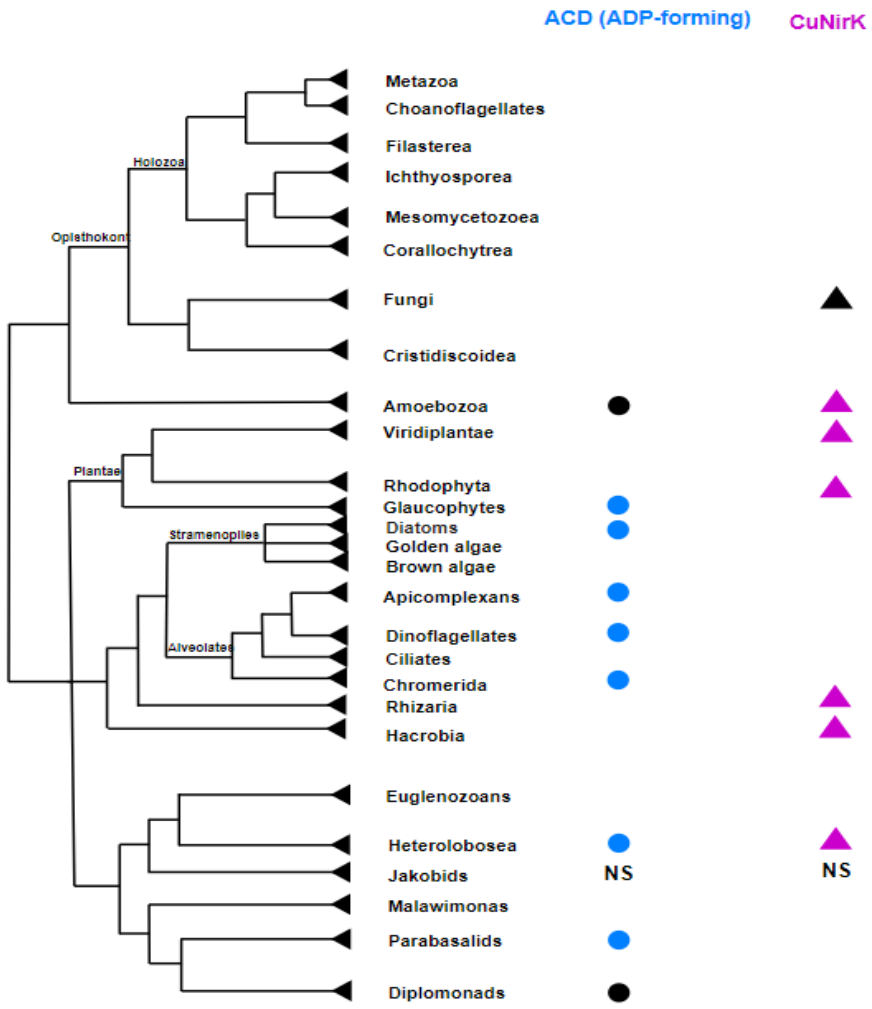


Figure 5.2. The phylogenetic occurrence of ACD and CuNirK in eukaryotes
The blue circles represent species where ACD was found to be present using genome sequences; black circles represent species where the presence of ACD was confirmed by experimental data. The purple triangles represent species where CuNirK was found to be present using genome sequence; black triangles represent species where the presence of CuNiK was confirmed by experimental data. (NS) No nuclear genome sequence is available.

5.1 Bioinformatics analysis of candidate *Naegleria* ACD function

NDP-forming acyl-CoA synthetases are a group of enzymes catalysing conversion of several CoA thioesters to their corresponding acids (Sánchez et al., 2000). This reaction is coupled

with the production of ATP or other NTPs via substrate-level phosphorylation. ACD belongs to this family of enzymes and specifically catalyses formation of acetate from acetyl-CoA and inorganic phosphate (P_i). Isoforms of ACD are widely conserved and have been biochemically characterised from archaea and some eukaryotes (*e.g. E. histolytica* and *Giardia lamblia*) (Jones and Ingram-Smith, 2014, Sanchez and Müller, 1996). The structure of ACDs has substantial heterogeneity amongst different species. For example, the ACDs from archaeal *Pyrococcus furiosus* and *Thermococcus kodakarensis* are heterotetramers that are composed of two alpha and two beta independent subunits. In *E. histolytica* and *G. lamblia* (similar to halophilic Archaea *Haloarcula marismortui*), the ACDs are homodimeric and are composed of fused alpha and beta domains linked by a hinge region (Bräsen et al., 2008, Sánchez et al., 2000, Jones et al., 2017). Previous studies on *P. furiosus* ACD had shown that the catalytic mechanism of ACD can be divided into four steps, and requires phosphorylation of two critical His residues found in the α, and β subunits (Weisse et al., 2016, Bräsen et al., 2008) (Figure 5.3). A recent study on *E. histolytica* ACD showed that the His residue (His252), which corresponds to the Hisα in *P. furiosus* ACD, is essential to the activity of ACD and mutation at this specific site will cause a complete loss of activity. However, the His residue in *E. histolytica* ACD (His533), which corresponds to the Hisβ in *P. furiosus* ACD, is important for the optimal activity of ACD but is not essential (Jones and Ingram-Smith, 2017). Two additional conserved residues in *E. histolytica* ACD, Glu213 and Asp674, were found to stabilise and interact with the catalytic Hisα during phosphorylation. Mutation at Glu213 site was found to stop the catalytic activity of ACD, whereas mutation at Asp674 caused a drastic decrease in the ACD activity (Jones and Ingram-Smith, 2017). A similar observation was detected in archaeal ACD (Weisse et al., 2016, Bräsen et al., 2008).

In order for *N. gruberi* candidate ACD to be functional with its postulated activity, one would anticipate that these conserved residues (two His, Glu and Asp residues) are present in the primary sequence. To assess this, protein alignments between candidate *N. gruberi*, and from recently released *N. fowleri* genome data, ACDs and *bona fide* ACD enzymes from different lineages were constructed. The candidate ACDs in *N. gruberi* and *N. fowleri* are predicted to be homodimers, encoded by a single gene. Thus, protein sequence alignments, were built using known homodimeric ACDs. In order to collect homodimeric ACD protein sequences from different taxa of bacteria, archaea and eukaryotes, the protein sequence of

the ACDs from *E. histolytica* and *G. lamblia* were used as queries in BLASTp. Amino acid sequence alignments were then built using CLC Sequence Viewer (protein accession numbers are provided in Appendices 16 and 17). Five separate alignments were constructed: (1) containing archaeal ACDs (Figure 5.4); (2) containing bacterial ACDs (Figure 5.5); (3) containing eukaryotic ACDs and candidate ACDs (Figure 5.6); (4) containing eukaryotic ACDs and candidate ACDs, including *N. gruberi* and *N. fowleri* sequences (Figure 5.7); (5) containing ACD sequences from taxa spread across the three life domains and candidate *N. gruberi* ACD (Figure 5.8). From the amino acid sequence alignments, it was apparent that the key His, Glu and Asp residues from the ACD found in *E. histolytica* were widely conserved across and between bacterial, archaeal and eukaryotic ACD candidates and known ACDs. Yet, neither conserved His residues were conserved in *N. gruberi* and *N. fowleri* candidate ACDs. This suggested either the ACDs might not be active as ACDs or that provided examples of divergent ACD enzymes.

Next, I analysed the domain architectures of atypical-looking *N. gruberi* and *N. fowleri* ACD candidates and compared their architectures to *bona fide* ACD enzymes. This was done using the SMART and InterPro servers whereby FASTA format of 'ACD' proteins were used as queries. ACD enzymes from archaea, bacteria and eukaryotes shared a common overall architecture with five conserved functional regions: a CoA-binding; two CoA-ligase; and two ATP-grasp domains (Appendix 18). The *N. gruberi* and *N. fowleri* candidate ACDs also exhibited conserved CoA_ligase and ATP-grasp domains. However, they lacked the CoA-binding domain and each contained a GNAT domain (Figure 5.9). Collectively, the absence of the functional conserved His residues and lack of the widely conserved CoA-binding domain suggest that classic ACD (ADP-forming) activity does not occur in *Naegleria* species or that the *N. gruberi* and *N. fowleri* proteins have a different role(s) within the cell. Although studies from a previous PhD student showed that the 60 amino acids in the N-terminus of *N. gurberi* candidate ACD are sufficient to direct GFP into the mitochondrion of procyclic *T. brucei* (Lynch, 2016).

According to a paper published in 2017, the conserved His252 residue in *E. histolytica* ACD (equivalent to Hisα in *P. furiosus* ACD) is critical for the function of ACD and is considered as a representative feature of acyl-CoA synthetase family (Jones et al., 2017). These authors also state that this conserved His can be used to distinguish between the ACD activity and the

protein acetyltransferase (PAT) activity (Jones et al., 2017). PAT is used to modulate target protein activity/activities by reversible addition of an acetyl group to specific lysine residues on target proteins (Wang et al., 2010). PAT shares a homology and same domains arrangement with ACD enzyme (Appendix 23). However, PAT lacks critical His residue and has a GNAT domain. The GNAT domain can be found in few ACDs and *Naegleria* putative 'ACDs' (Appendix 18). Nevertheless, only *N. gruberi* and *N. fowleri* 'ACDs' lack the characteristic His residue for ACD function. This suggests that *N. gruberi* and *N. fowleri* 'ACDs' are actually orthologous to PATs. To test this hypothesis, an alignment between *Naegleria* candidate proteins and a known bacterial PAT was carried out and the critical residues for PAT function were highlighted (Figure 5.10). Five residues in the N-terminal region of bacterial PAT are known to be important for catalytic function; these are (according to *Salmonella enterica* numbering) Asp165; Arg170; Ala220; Arg450 and Asp592. Mutations at these sites cause a decrease in acetyltransferase activity (Thao and Escalante-Semerena, 2012). In addition to these sites, PAT enzyme, which belongs to the acetyltransferase superfamily, has long conserved motifs in the C-terminal known as motif A. Motif A is considered as a hallmark for acetyltransferase superfamily. Within motif A, Glu residue (Glu809 in *S. enterica*) was identified to be critical for PAT structure. Mutation at this specific site caused a major structural changes of PAT (Thao and Escalante-Semerena, 2012). The *Naegleria* candidate proteins have some of the conserved N-terminal residues as well as most of the conserved residues in motif A, including the critical Glu residue (Figure 5.10). This indicates that the *N. gruberi* candidate protein might be used for lysine acetylation. In the future, a subsequent analysis could be to carry out in vitro to test PAT`s activity of either the *N. gruberi* or *N. fowleri* candidate proteins as described in the study of Thao and Escalante-Semerena (2012).

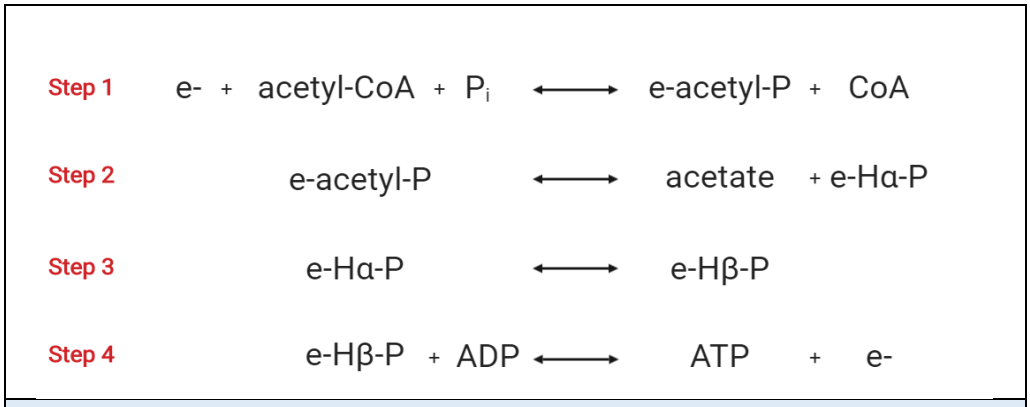


Figure 5.3. A proposed mechanism for how ACD converts acetyl CoA to acetate
The mechanism follows a four-step reaction. This reaction involves the phosphorylation and dephosphorylation of two highly conserved His residues (Hisα and Hisβ).

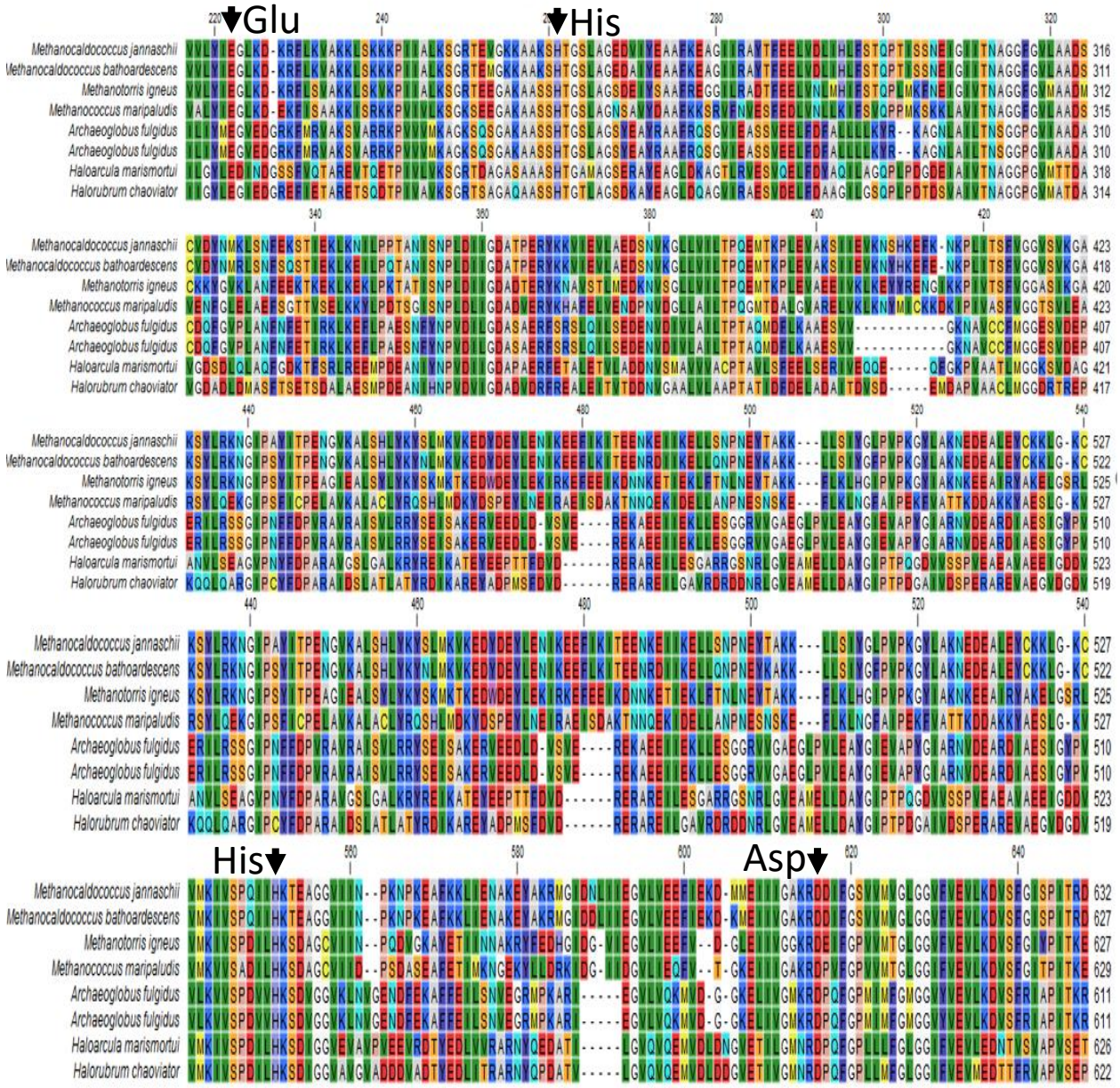


Figure 5.4. Protein alignment of conserved regions of acetyl-CoA synthetase (ADP-forming) between different archaeal species

CLC Sequence Viewer, using amino acid sequences, generated the alignment. The black arrows point to the functionally conserved residues (Hisα, Hisβ, Gluα and Aspβ in *P. furiosus* ACD). The amino acid colour scheme is according to RasMol 2.75 and it is based on traditional amino acid properties (Appendix 13). The number at the top of the alignment represents the amino acid position within the sequence including the inserted gaps.

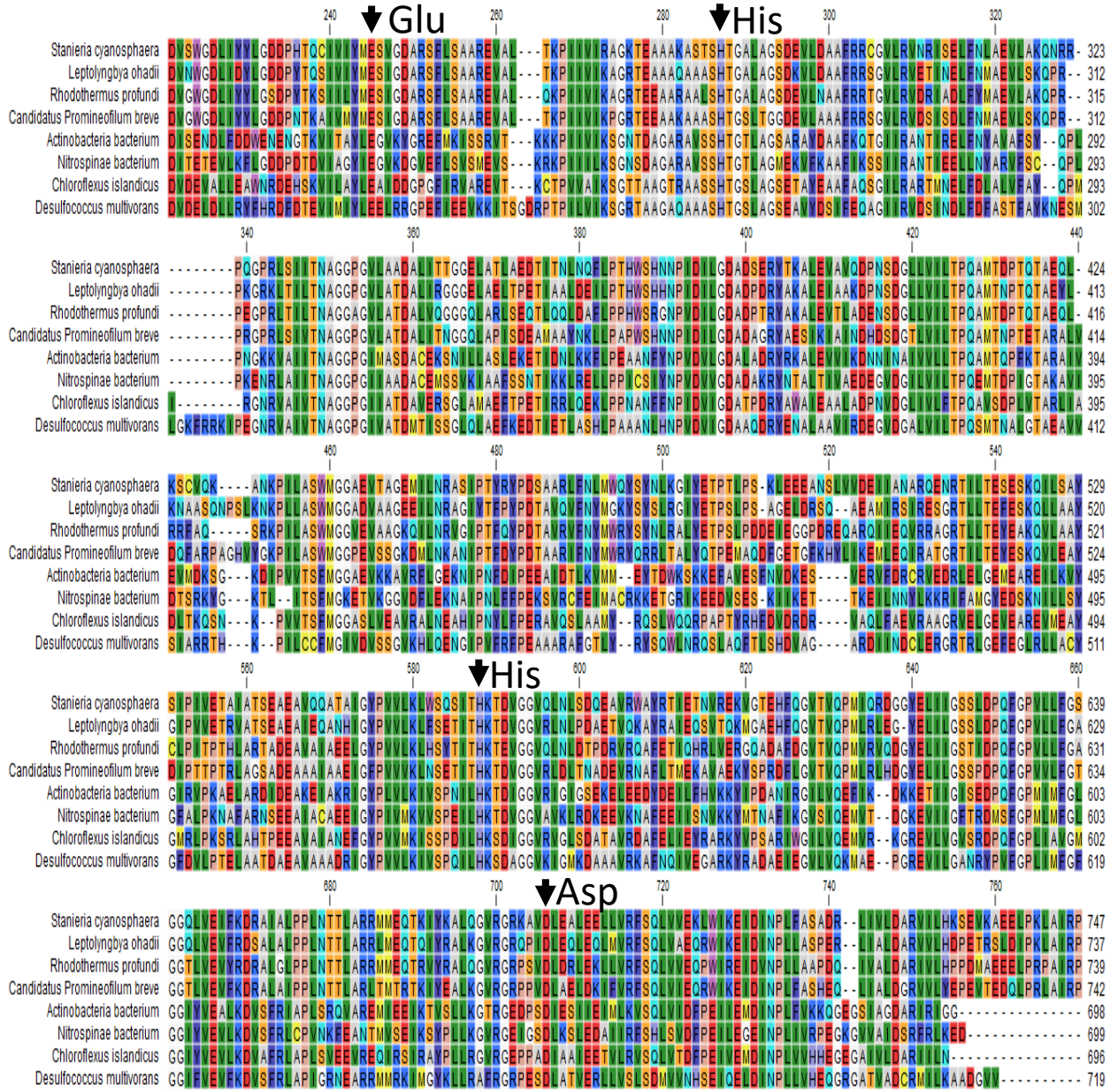


Figure 5.5. Protein alignment of conserved regions of acetyl-CoA synthetase (ADP-forming) between different bacterial species

CLC Sequence Viewer, using amino acid sequences, generated the alignment. The black arrows point to the functionally conserved residues (Hisα, Hisβ, Gluα and Aspβ in *P. furiosus* ACD). The amino acid colour scheme is according to RasMol 2.75 and it is based on traditional amino acid properties (Appendix 13). The number at the top of the alignment represents the amino acid position within the sequence including the inserted gaps.

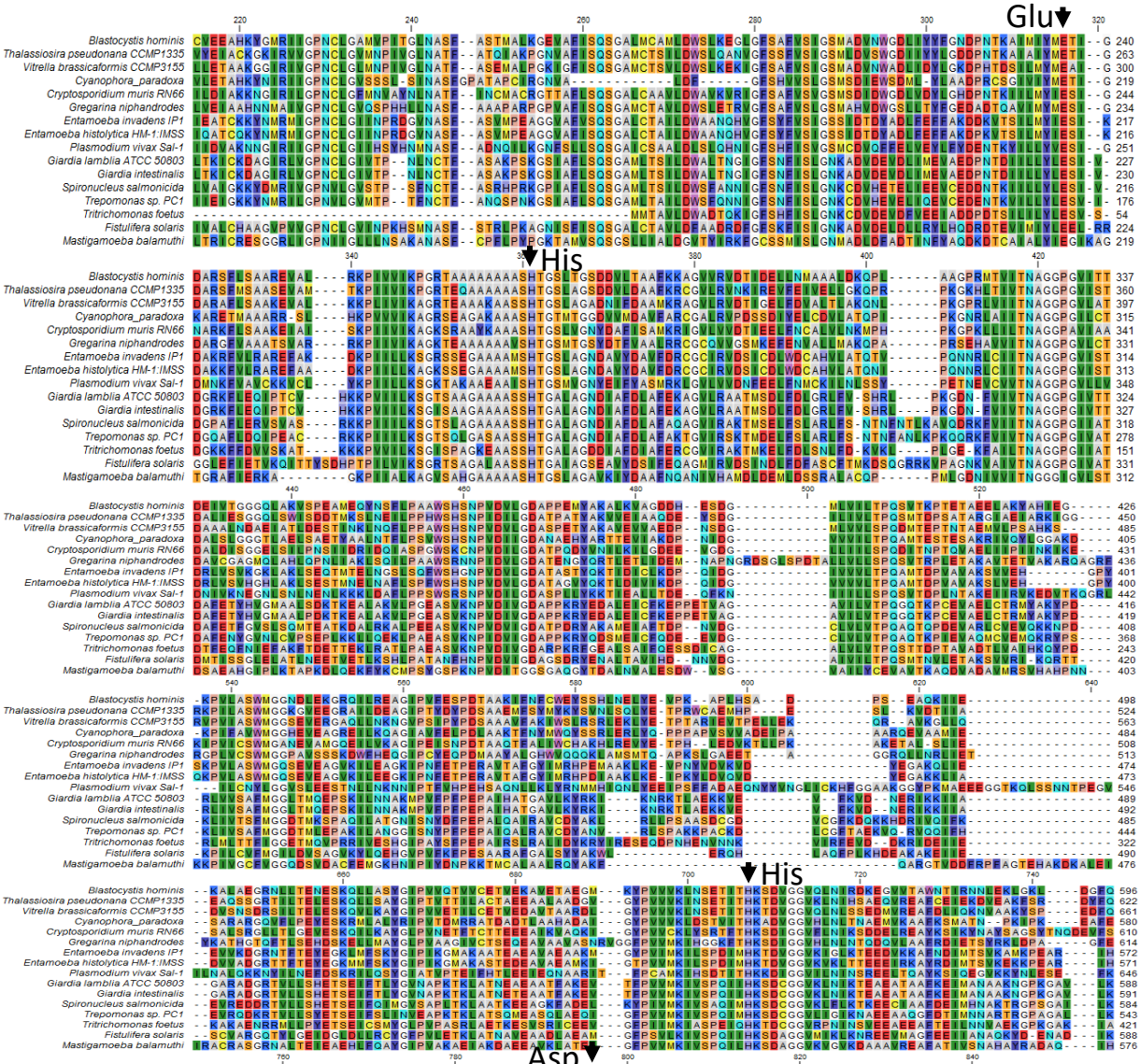


Figure 5.6. Protein alignment of conserved regions of acetyl-CoA synthetase (ADP-forming) between different eukaryote species without *Naegleria* putative ACDs

CLC Sequence Viewer, using amino acid sequences, generated the alignment. The black arrows point to the functionally conserved residues (Hisα, Hisβ, Gluα and Aspβ in *P. furiosus* ACD). The amino acid colour scheme is according to RasMol 2.75 and it is based on traditional amino acid properties (Appendix 13). The number at the top of the alignment represents the amino acid position within the sequence including the inserted gaps.

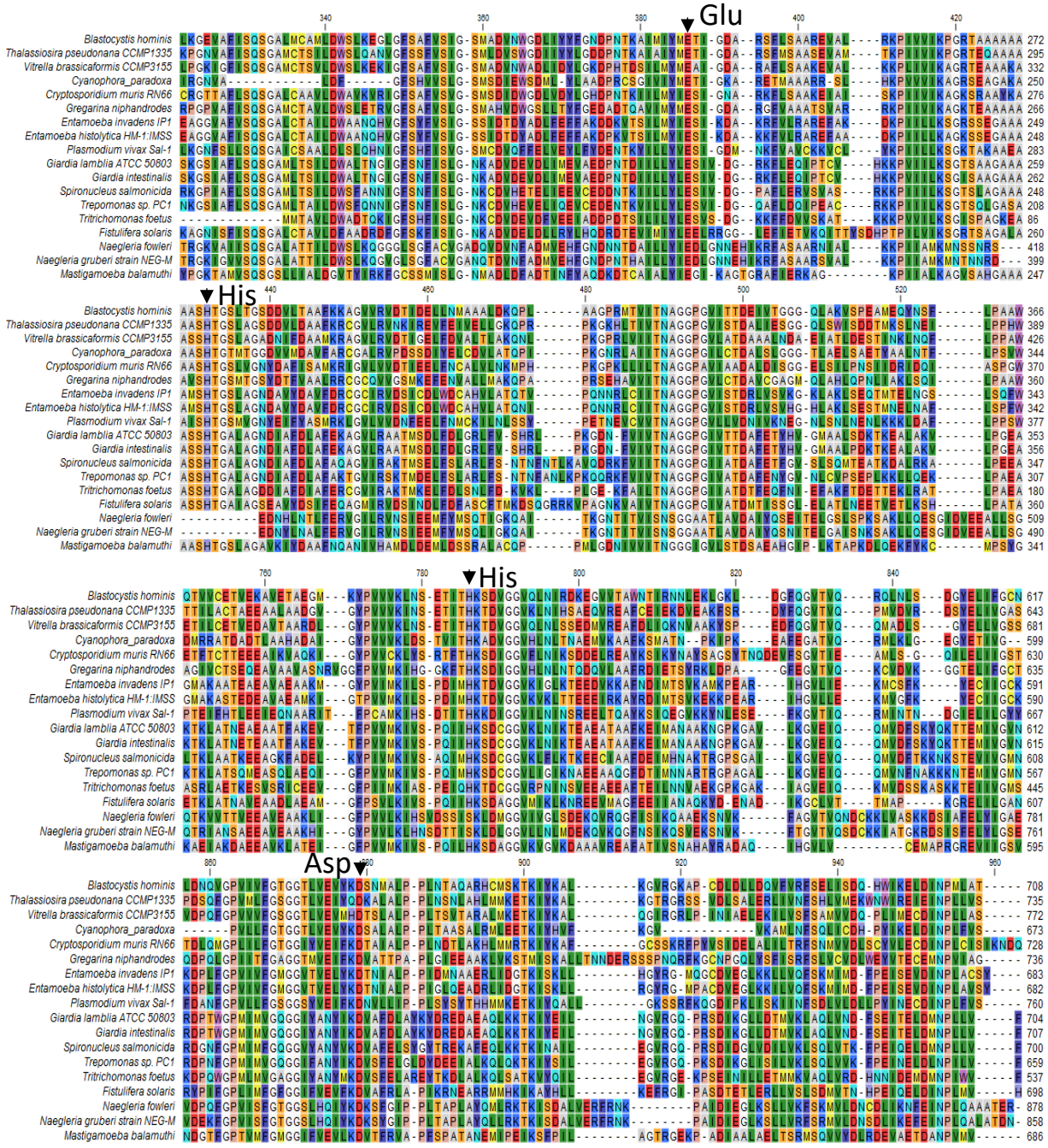


Figure 5.7. Protein alignment of conserved regions of acetyl-CoA synthetase (ADP-forming) between different eukaryote species and *Naegleria* putative ACDs

CLC Sequence Viewer, using amino acid sequences, generated the alignment. The black arrows point to the functionally conserved residues (Hisα, Hisβ, Gluα and Aspβ in *P. furiosus* ACD). The amino acid colour scheme is according to RasMol 2.75 and it is based on traditional amino acid properties (Appendix 13). The number at the top of the alignment represents the amino acid position within the sequence including the inserted gaps.

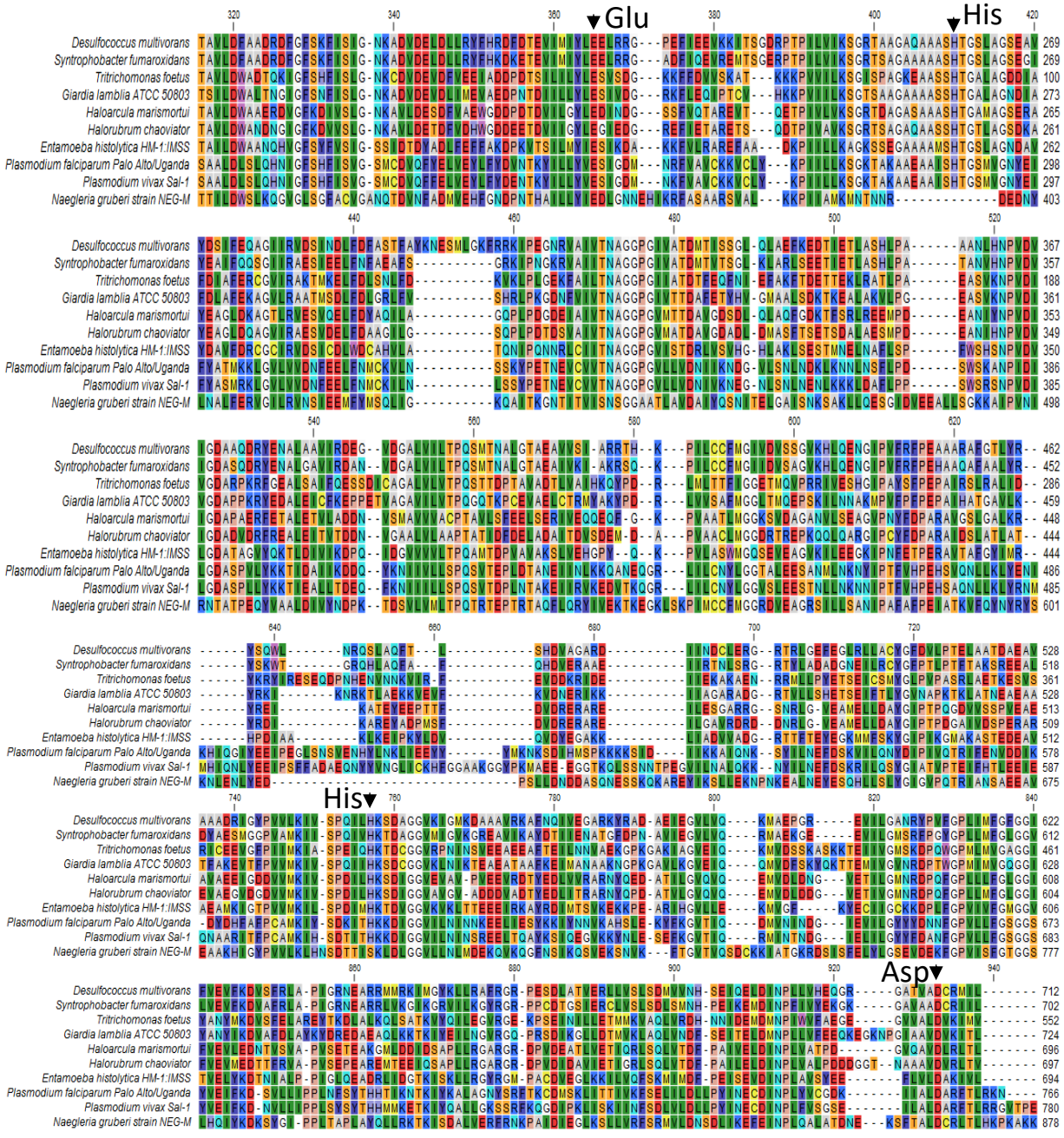


Figure 5.8. Protein alignment of conserved regions of acetyl-CoA synthetase (ADP-forming) between three life domains and *N. gruberi* candidate ACD

ACD protein sequences from archaea, bacteria and eukaryotes were collected and used to build an alignment along with *N. gruberi* candidate ACD. CLC Sequence Viewer, using amino acid sequences, generated the alignment. The black arrows point to the functionally conserved residues (Hisα, Hisβ, Gluα and Aspβ in *P. furiosus* ACD). The amino acid colour scheme is according to RasMol 2.75 and it is based on traditional amino acid properties (Appendix 13). The number at the top of the alignment represents the amino acid position within the sequence including the inserted gaps.

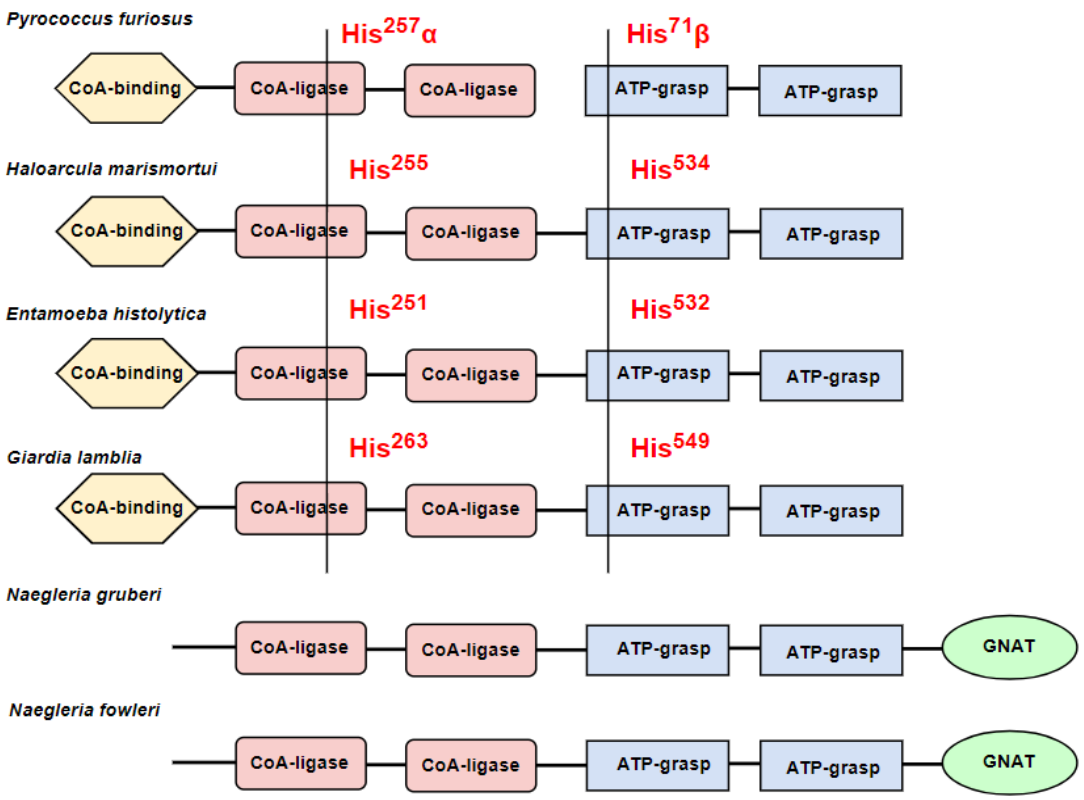


Figure 5.9. Overall structure of previously characterised ACD and *Naegleria* putative ACDs
ACD of *Pyrococcus furiosus* is a heterotetrameric complex that contains two alpha and two beta separate subunits. ACDs in *Haloarcula marismortui*, *Entamoeba histolytica* and *Giardia lamblia* are homodimer proteins that contain fused alpha and beta domains linked by a hinge region. The protein structure was analysed using InterPro and SMART servers.

Figure 5.10 Protein alignment of conserved regions of bacterial protein acetyltransferases (PAT) and *Naegleria* candidate 'ACDs'

The alignment was generated by CLC Sequence Viewer, using amino acid sequences. The black arrows point to the functionally conserved residues in PAT N-terminal region (according to *S. enterica* numbering: Asp165, Arg170, Ala220, Arg450 and Asp592). The black box highlights the region of motif A. Motif A is widely conserved among acetyltransferases family. The arrow in the black box points to the conserved functional residue within motif A (according to *S. enterica* numbering: Glu809). The number at the top of the alignment represents the amino acid position within the sequence including the inserted gaps.

5.2 Bioinformatics analysis of candidate *Naegleria* CuNiR function

The nuclear genome in *N. gruberi* contains a gene coding for a candidate nitrite reductase (NiR). NiR enzymes catalyse reduction of NO_2^- to NO. This reaction is considered a committed step in denitrification (leading to N_2 production), which occurs widely in facultatively anaerobic bacteria and archaea (Philippot and expression, 2002, Adhikari et al., 2017). In contrast, in eukaryotes NiR enzymes have been detected and characterised from a few fungi (*e.g. Fusarium oxysporum*), foraminifera and autotrophic green alga (*e.g. Chlamydomonas reinhardtii*) (Galván et al., 1991, Shoun and Tanimoto, 1991, Woehle et al., 2018). Two types of NiR enzymes are known: one group utilises a heme cofactor (cd_1NiRs); the second group utilises a copper cofactor (CuNiRs). Holo-CuNiR forms a homotrimer and typically, each subunit contains two Cu^+ cofactors, each occupying a different chemical environment within the folded protein (environments T1Cu and T2Cu). The two Cu^+ ions are linked by Cys-His bridge (Figure 5.11). T1Cu is coordinated by one Cys, one Met and two His residues. T2Cu is coordinated via one Asp and three His residues. CuNirK receives electrons from an electron donor (*e.g.* cytochrome *c*, azurin-like blue copper protein) at T1Cu and catalyses the reduction of NO_2^- to NO at T2Cu (Li et al., 2015, Boulanger et al., 2000, Kobayashi and Shoun, 1995). In a study published by Helen and co-workers in 2016, 267 full-length CuNiRs where analysed phylogenetically, dividing CuNiRs group into two main clades (Helen et al., 2016). Clade I contained only enzymes from α-, β-, and γ-proteobacteria; clade II enzymes originated from a greater range of taxonomic groups, including archaea, firmicute, actinobacteria and chloroflexi organisms.

The start point for my analysis was to analyse putative *Naegleria* CuNiRs to determine whether the predicted proteins contained the conserved, essential residues needed for catalysis. Accordingly, CuNiR sequences from bacteria, eukaryotes and archaea were collected using BLAST (for protein accession numbers see Appendices 19, 20 and 21). Using these CuNiR protein sequences, five different alignments were constructed: (1) an alignment containing only clade I CuNiRs (Figure 5.12); (2) an alignment containing only clade II CuNiRs (Figure 5.13); (3) an alignment of eukaryotic CuNiRs, including candidate *N. gruberi* and *N. fowleri* CuNiRs (Figure 5.14); (4) an alignment with eukaryotic CuNiRs, the *N. gruberi* candidate and some clade I CuNiRs (Figure 5.15); and (5) an alignment with eukaryotic CuNiRs, the *N. gruberi* candidate, and some clade II CuNiRs (Figure 5.16). Despite the observation that

the primary sequences of CuNiRs from different species are not particularly well conserved, residues needed for Cu^{+} coordination and catalysis are invariantly conserved between archaea, bacteria and eukaryotes. Most importantly, these conserved residues were identified in *N. gruberi* and *N. fowleri* candidate CuNiRs.

My next step was to examine the overall domain and motif architectures of CuNiRs from different clades and to compare them to *N. gruberi* and *N. fowleri* candidate CuNiRs. The domain architecture of CuNiRs was examined using SMART and InterPro servers. The outcome of this analysis showed that CuNiRs from clade I, clade II, a fungus, algae and *Naegleria* share similar architectures containing two conserved copper oxidase type I and II domains (Figure 5.17, Appendix 22). From this analysis it seems likely that the *Naegleria* CuNiRs are catalytically active and likely to contribute to anaerobic respiration under appropriate environmental conditions. As a final step in the analysis, I looked at the phylogenetic relationship between *Naegleria* CuNirKs and other eukaryotic CuNirKs. My final phylogenetic tree was inferred from a protein alignment constructed using full-length CuNiRK sequences. The phylogenetic tree was constructed as described in the Method chapter, section 2.1. Sequences for 71 CuNirK proteins from clades I, II and eukaryotes, including *Naegleria* putative CuNirKs were used. The low level of conservation of CuNirKs seen across different groups and the lack of CuNirk in most of the eukaryotic phyla rendered some of the bootstrap values in the resultant phylogenetic tree unreliable. However, from the phylogenetic tree, it appears that *Naegleria* CuNirKs grouped within a eukaryotic CuNirK which in turn cluster together with CuNirK clade II (Figure 5.18). The formation of a monophyletic group of eukaryotic CuNirK and CuNirK clade II was previously reported (Kim et al., 2009). Based on this and the identification of the highly conserved functional residues in *N. gruberi* and *N. fowleri* candidate CuNiRs and a common overall structure, I suggest *Naegleria* CuNiRs do have a physiological role to play during some conditions of hypoxia or anoxia.

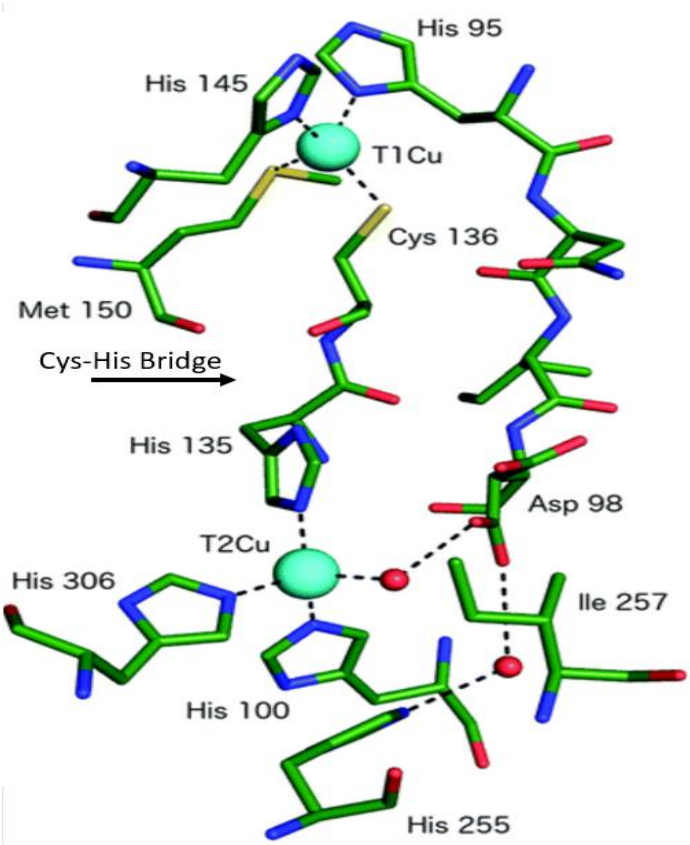


Figure 5.11. T1Cu and T2Cu centres of *Achromobacter cycloclastes* CuNirK
T1Cu centre is coordinated by one Cys, one Met and two His residues and is linked to T2Cu via Cys-His bridge. T2Cu centre is coordinated via one Asp and three His residues. The diagram was reproduced with permission from Horrell et al (2017).

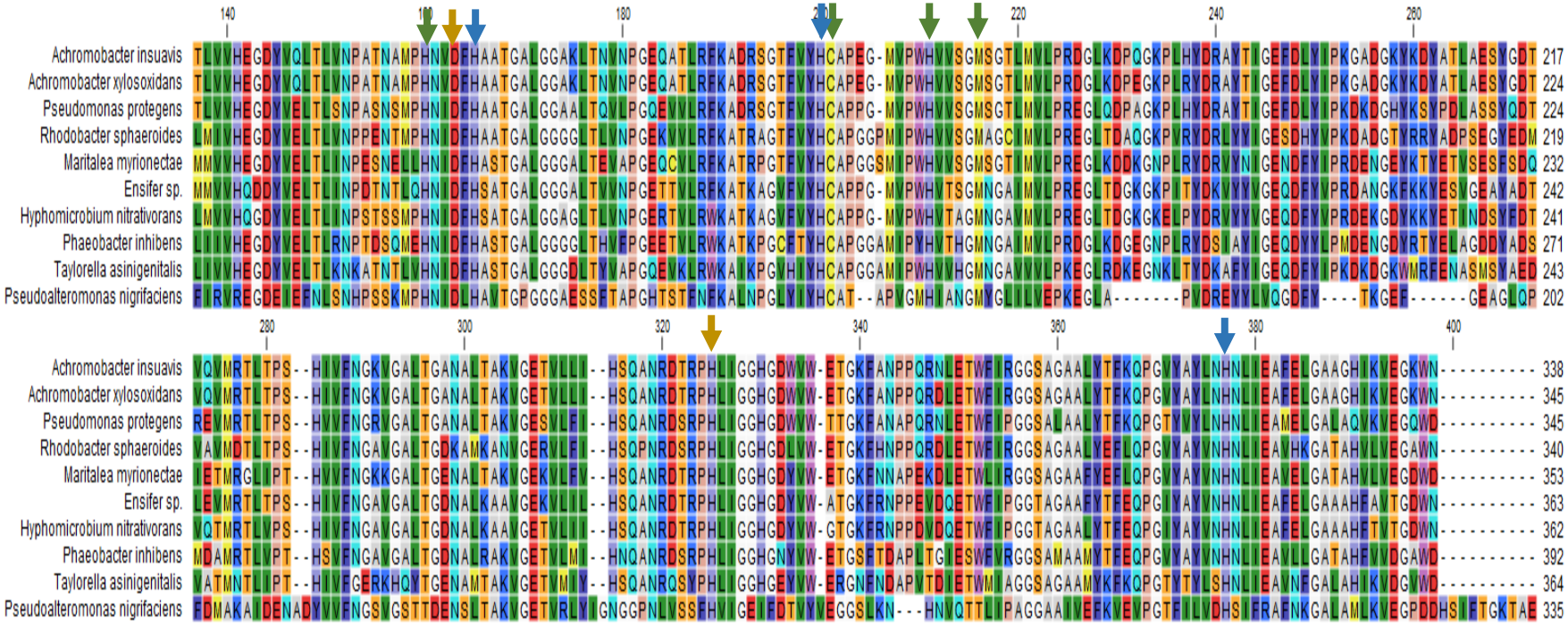


Figure 5.12. Protein alignment of conserved regions of CuNiR belonging to Clade I

The alignment was generated by CLC Sequence Viewer, using amino acid sequences. Green arrows label the conserved residues which bind to T1Cu (two His, Cys and Met residues), the blue arrows label the conserved residues which bind to T2Cu (three His residues) and the brown arrows label residues which found in the active site of CuNirK (Asp and His). The amino acid colour scheme is according to RasMol 2.75 and it is based on traditional amino acid properties (Appendix 13). The number at the top of the alignment represents the amino acid position within the sequence following the inserted gaps.

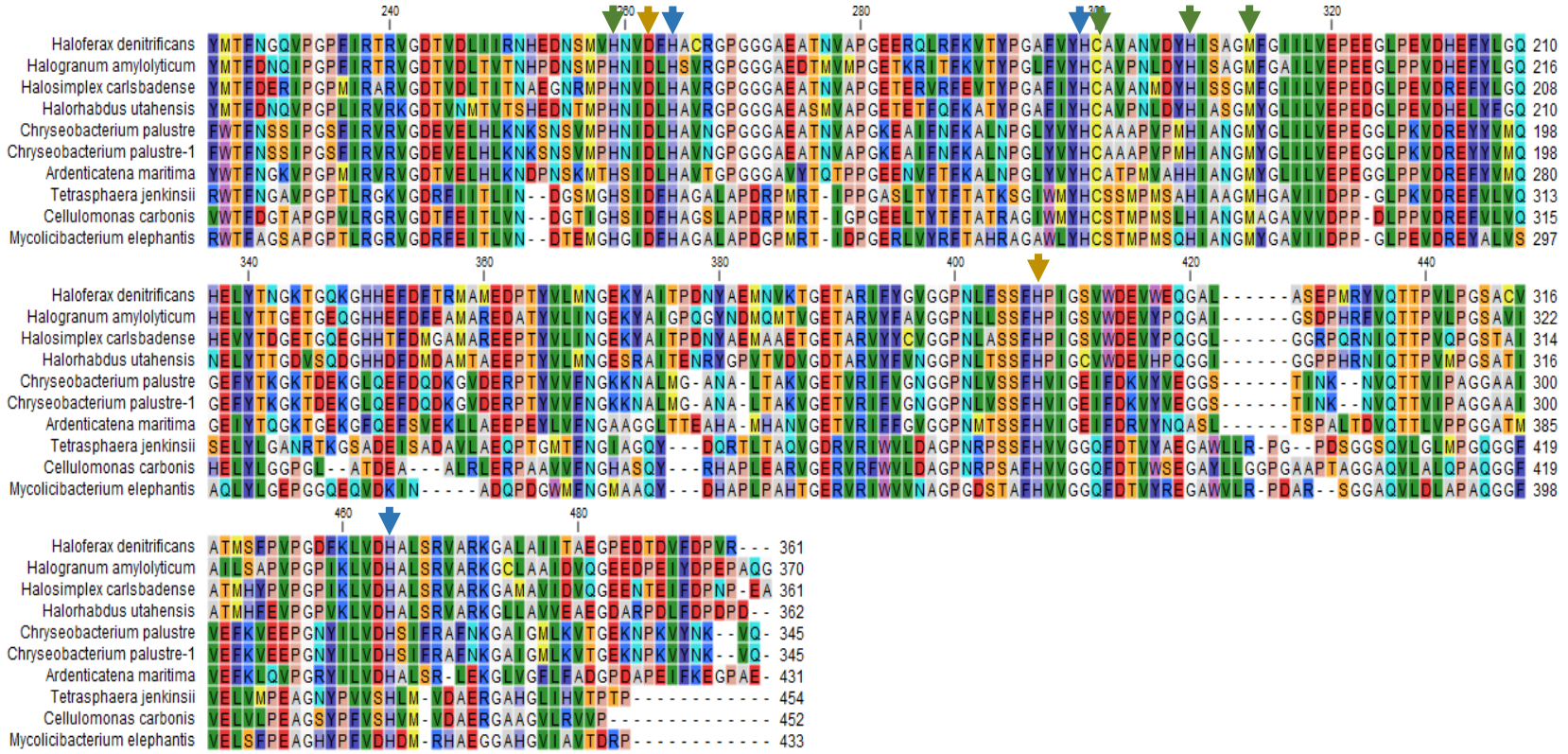


Figure 5.13. Protein alignment of conserved regions of CuNiR belonging to Clade II

The alignment was generated by CLC Sequence Viewer, using amino acid sequences. Green arrows label the conserved residues which bind to T1Cu (two His, Cys and Met residues), the blue arrows label the conserved residues which bind to T2Cu (three His residues) and the brown arrows label residues which found in the active site of NirK (Asp and His). The amino acid colour scheme is according to RasMol 2.75 and it is based on traditional amino acid properties (Appendix 13). The number at the top of the alignment represents the amino acid position within the sequence following the inserted gaps.

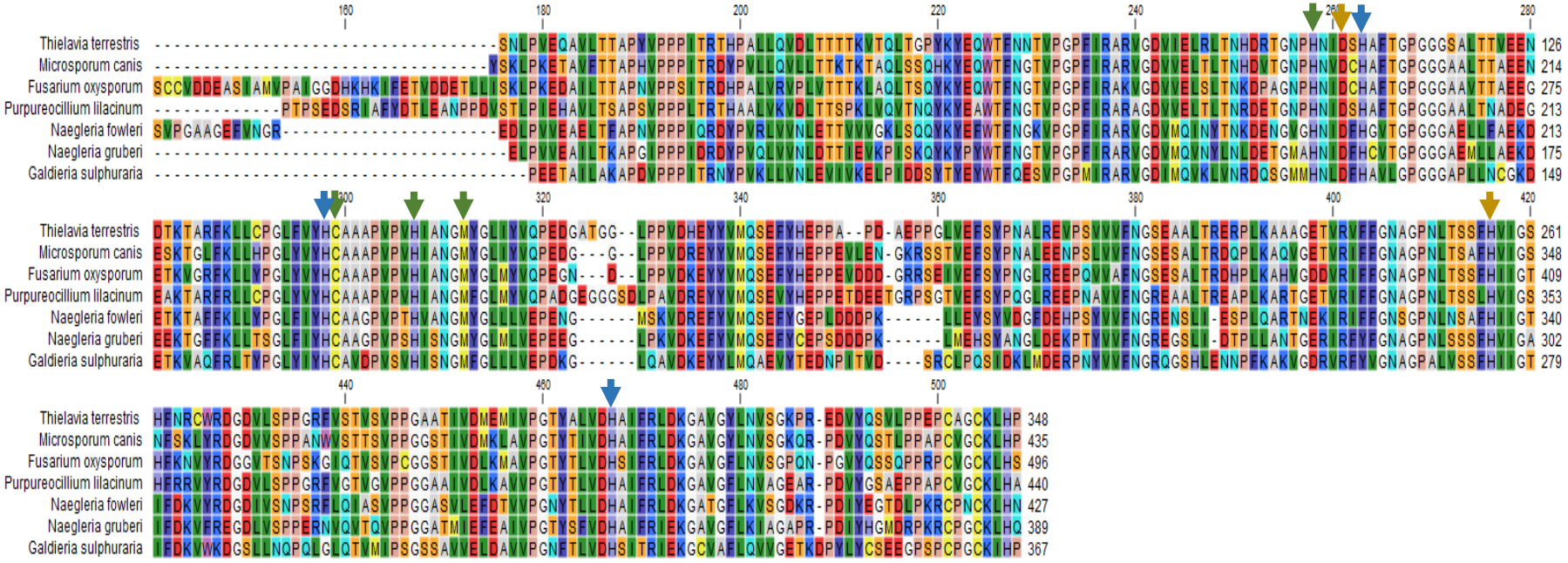


Figure 5.14. Protein alignment of conserved regions of CuNiR from eukaryotes including *Naegleria* putative CuNiRs

CLC Sequence Viewer generated the alignment of the amino acids sequence. Green arrows label the conserved residues which bind to T1Cu (two His, Cys and Met residues), the blue arrows label the conserved residues which bind to T2Cu (three His residues) and the brown arrows label residues which found in the active site of NirK (Asp and His). The amino acid colour scheme is according to RasMol 2.75 and it is based on traditional amino acid properties (Appendix 13). The number at the top of the alignment represents the amino acid position within the sequence following the inserted gaps.

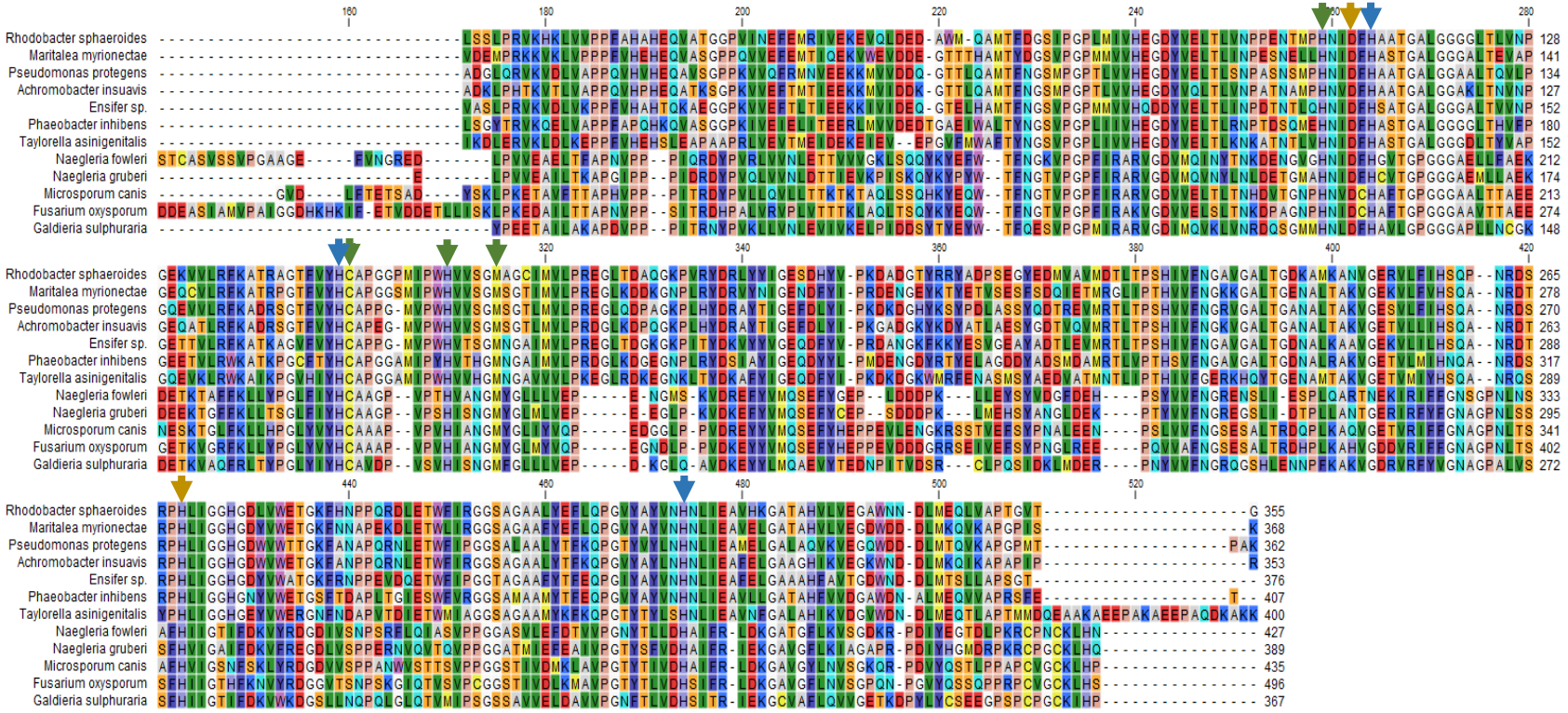


Figure 5.15. Protein alignment of conserved regions of CuNiR from eukaryotes including *Naegleria* putative CuNiRs and CuNiR from Clade I

The alignment was generated by CLC Sequence Viewer, using amino acid sequences. Green arrows label the conserved residues which bind to T1Cu (two His, Cys and Met residues), the blue arrows label the conserved residues which bind to T2Cu (three His residues) and the brown arrows label residues which found in the active site of NirK (Asp and His). The amino acid colour scheme is according to RasMol 2.75 and it is based on traditional amino acid properties (Appendix 13). The number at the top of the alignment represents the amino acid position within the sequence following the inserted gaps.

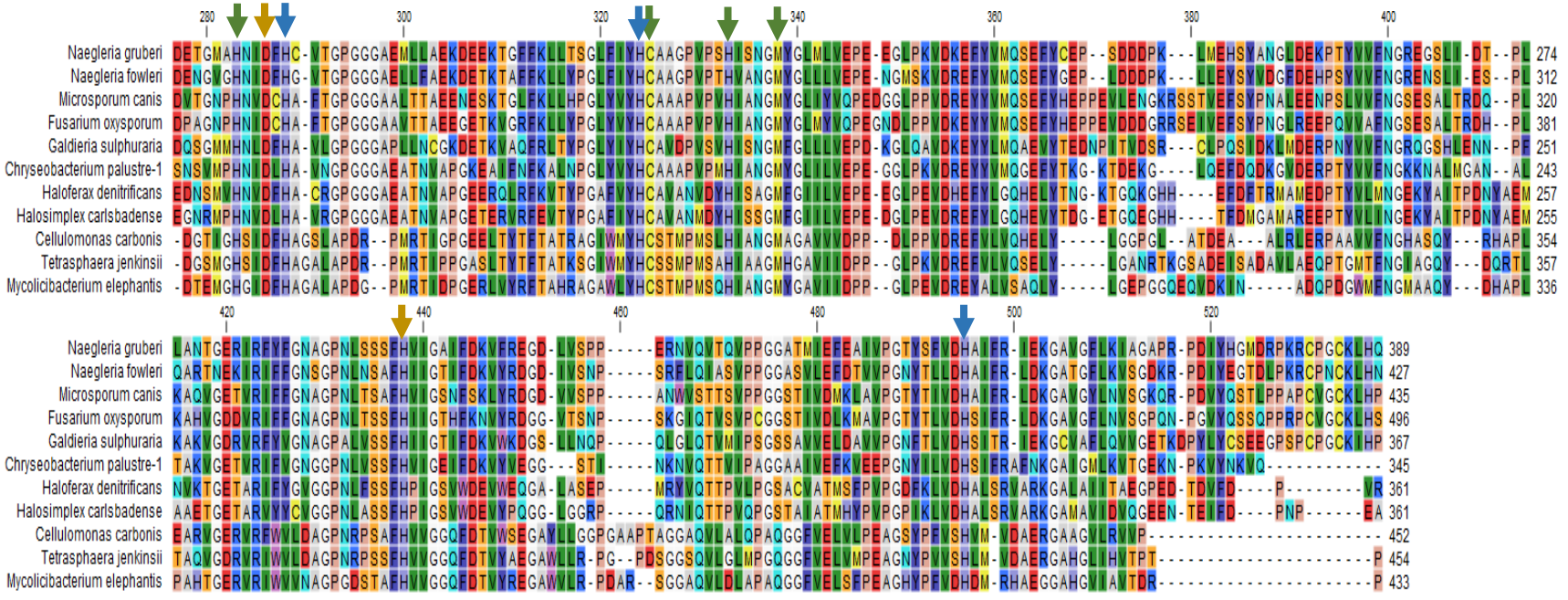


Figure 5.16. Protein alignment of conserved regions of CuNiR from eukaryotes including *Naegleria* putative CuNiRs and CuNiR from Clade II

The alignment was generated by CLC Sequence Viewer, using amino acid sequences. Green arrows label the conserved residues which bind to T1Cu (Two His, Cys and Met residues), the blue arrows label the conserved residues which bind to T2Cu (three His residues) and the brown arrows label residues which found in the active site of NirK (Asp and His). The amino acid colour scheme is according to RasMol 2.75 and it is based on traditional amino acid properties (Appendix 13). The number at the top of the alignment represents the amino acid position within the sequence following the inserted gaps.

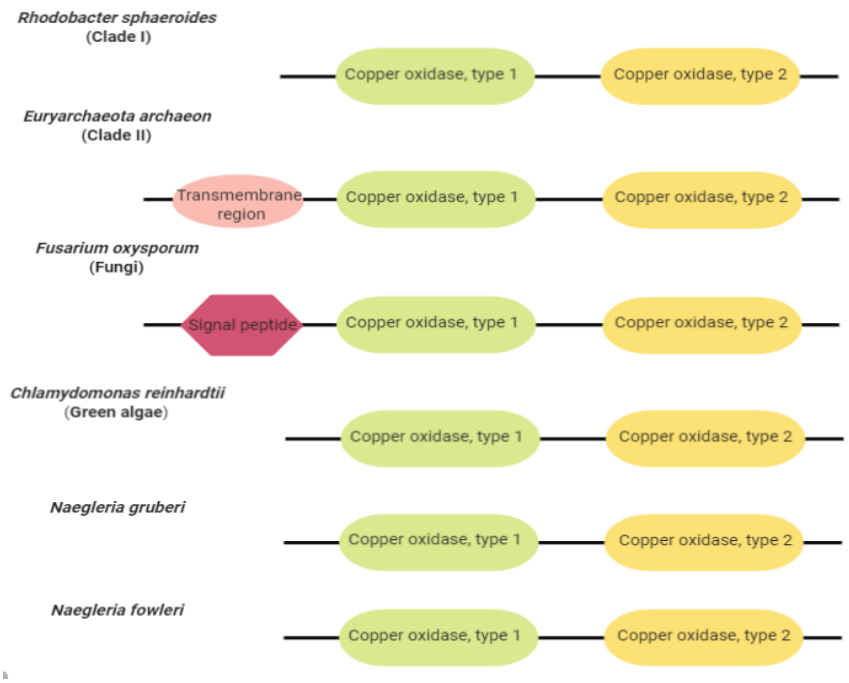


Fig 5.17. General structure of CuNirk in three life domains and putative *Naegleria* CuNirks

Rhodobacter sphaeroides represents the bacteria domain, *Euryarchaeota archaeon* represents the archaea domain and *Fusarium oxysporum* and *Chlamydomonas reinhardtii* represents the eukaryotes domain. Protein sequences in FASTA format were used as queries in the InterPro and SMART web servers to analyse protein structure.

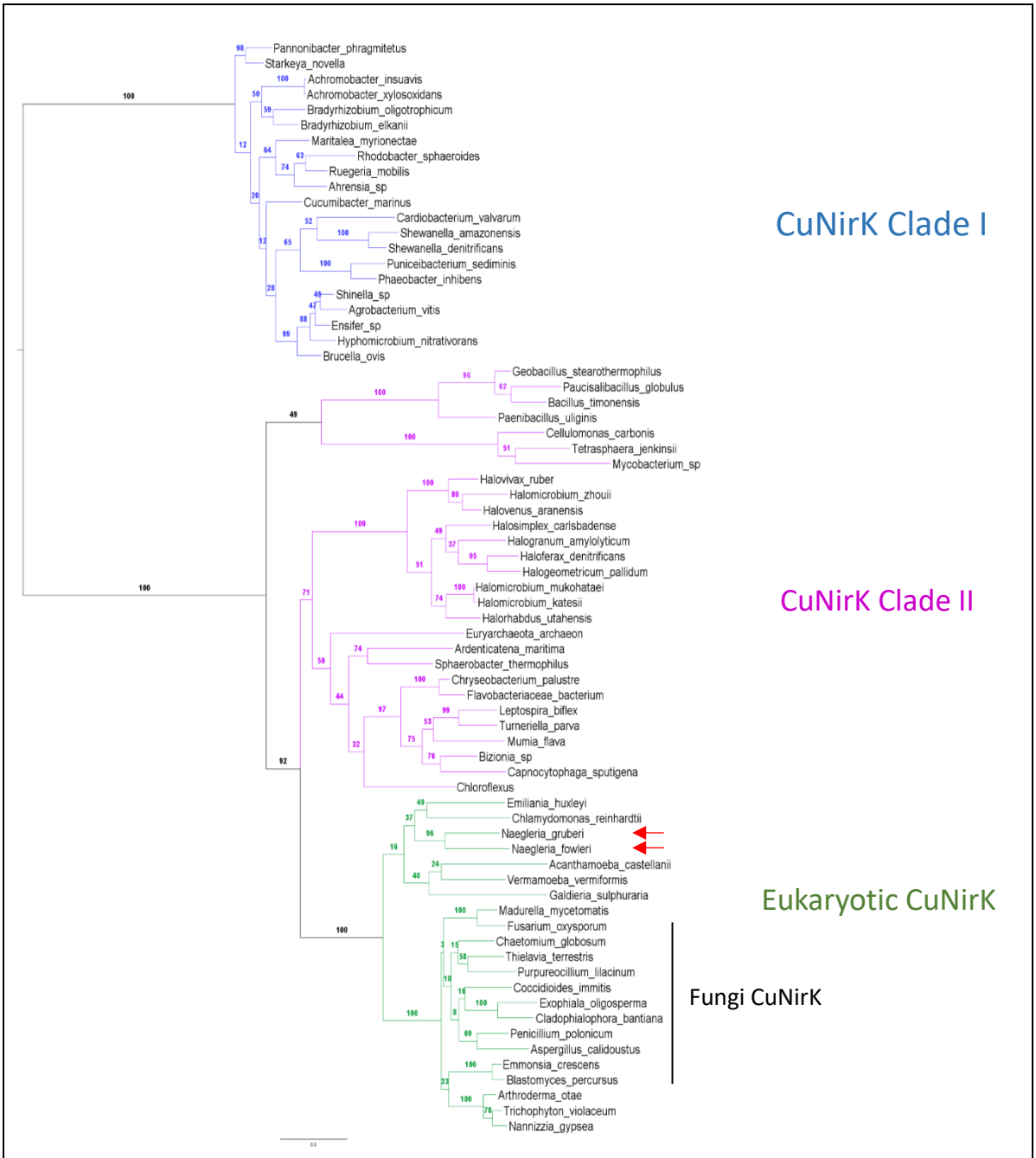


Figure 5.18. Phylogenetic tree of copper-containing nitrite reductase (CuNirK)
CuNirK clade I contain the protein sequences from α-, β-, and γ-proteobacteria, while CuNirK clade II contains protein sequences from archaea, firmicutes, actinobacteria and chloroflexi. Majority of eukaryotic CuNirK comprised of fungi, which grouped in the tree. The red arrows indicate the position of *Naegleria* putative CuNirKs. This midpoint-rooted tree is based on the RaxaMl maximum likelihood analysis. The parameters of RaxaMl analysis are as follows: WAG as the best protein substitution matrix, CAT model and Rapid bootstrap analysis (with 1000 standard bootstrap replicates). The scale bar of 0.5 represents the average number of amino acid substitution per site. The RaxaMl analysis was run on online Cipres webserver (Stamatakis et al., 2008). Figtree v1.4.3 server was used to visualise the tree (http://tree.bio.ed.ac.uk/software/figtree/).

5.3 Mitochondrial morphology in *N. gruberi*

With a unique metabolic profile, it was key to attempt to visualise and study *N. gruberi* mitochondria and its position within *N. gruberi* during the trophozoite stage by live-cell imaging. The trophozoite stage is the vegetative form of *N. gruberi*. By focusing on the movement of *N. gruberi* it was possible to obtain an insight into the number and organisation of mitochondria within the amoebae and how the organelles re-position during pseudopod-based movement. This allowed me to gain an understanding of the mitochondrial dynamics and mitochondria morphology. Thus, live active trophozoite cells were incubated with Deep Red MitoTracker (DRMT). The staining technique of DRMT is dependent on mitochondrial membrane potential changes. The cells were incubated with DRMT for 30 minutes prior to the live-cell imaging using the confocal Zeiss microscope. Figure 5.19 shows a high copy number of mitochondria, which look like small puncta that move through the *N. gruberi* cell body in coordination with pseudopod-based amoeboid movement. Notably, during *N. gruberi* movement, the leading edge of pseudopodia constantly lack mitochondria (pointed out by arrows in Figure 5.19).

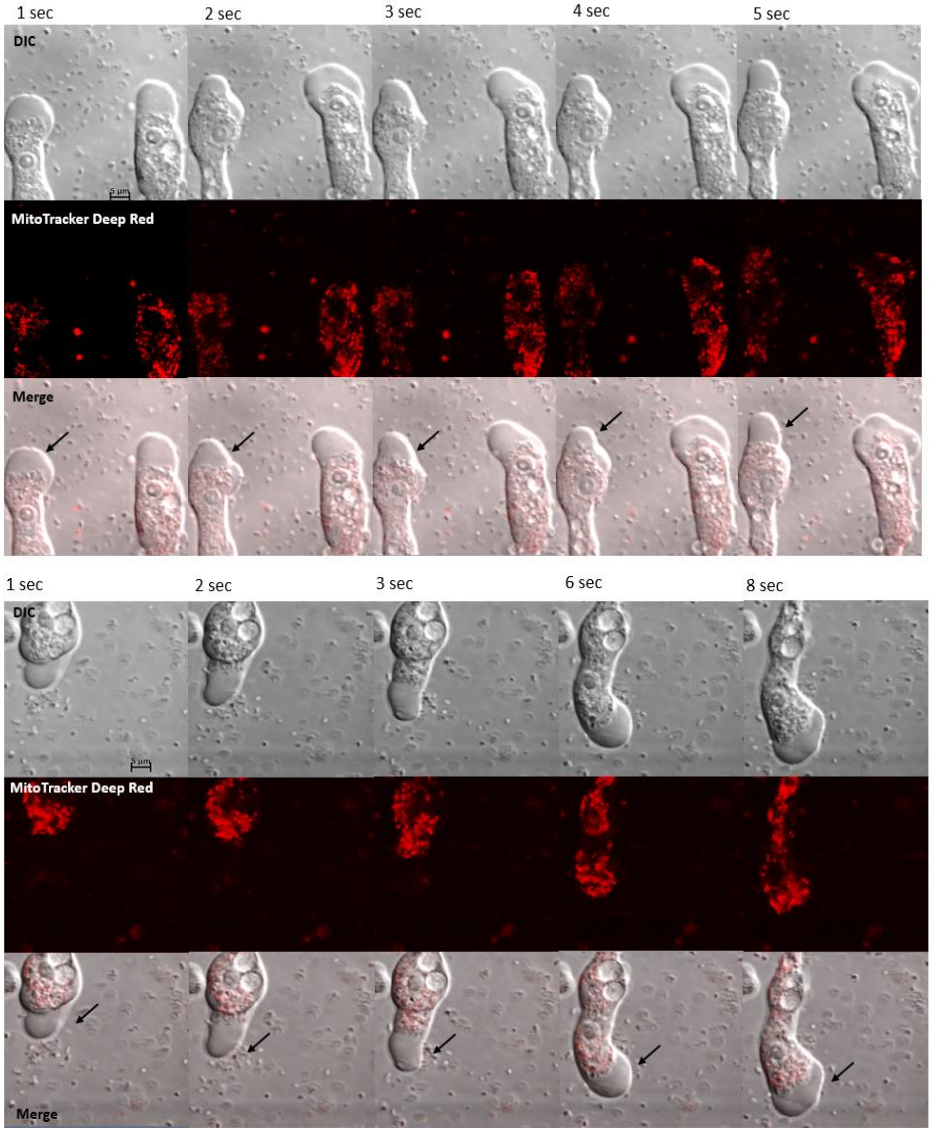


Figure 5.19. Mitochondrial morphology in *N. gruberi* during trophozoite stage

The deep red mitotraker (red colour) was used to stain mitochondria. Live-cell time-lapse images were recorded every second using Zesis confocal microscope. The black arrows point to the pseudopodia found in the direction of cytoplasm flow. The scale bar is 5 µm.

Chapter 6

Evidence of widespread cryptic peroxisomal targeting in protists

Peroxisomes are single membrane-bound organelles, which were first biochemically characterised from rat liver and *Tetrahymena* by Christian de Duve, Henri Beaufay and Pierre Baudhuin during the 1960s (Baudhuin et al., 1965, De Duve and Baudhuin, 1966, De Duve, 1965). Originally noted for their high catalase content, peroxisomes are now associated with a variety of biochemical reactions in many eukaryotes (as discussed in the introductory chapter) (De Duve and Baudhuin, 1966). The classic role(s) played by peroxisomes in many eukaryotes is the detoxification of hydrogen peroxide and β-oxidation of long chain fatty acids (Cooper et al., 2000, Gabaldón, 2010).

Peroxisomes lack DNA and a protein synthetic machinery; peroxisomal proteins are therefore synthesised in the cytoplasm and post-translationally directed into the organelle matrix (Gabaldón, 2010). Experimental studies with mammals, yeast, plants and trypanosomes, and bioinformatics studies of other eukaryotes, revealed that peroxisomes share evolutionarily conserved import mechanisms. These are used to target peroxisomal proteins into peroxisomal matrix and membrane. These conserved mechanisms utilise up to 34 highly specific proteins known as peroxins (or Pex proteins). Pex protein complexes work by recognising a specific motif present in most peroxisomal proteins which is known as a peroxisomal targeting signal (or PTS). There are two different types of PTS: the C-terminal PTS1 and the N-terminal PTS2. A PTS1 is the most common peroxisomal signal and it is formed by three amino acids with the consensus motif of S/A/C-K/R/H-L/M (Gould et al., 1989, Williams et al., 2012). Besides the conserved tripeptides, the 9 amino acids upstream of PTS1 have been found to modulate the targeting efficiency with which proteins are directed into peroxisomes (Brocard and Hartig, 2006). Proteins with PTS1 are recognised and targeted to the peroxisome via the activity of the Pex5 receptor. Fewer peroxisomal proteins have a PTS2 motifs, which are recognised by Pex7 receptor. The consensus sequence of the PTS2 motif is R/K-L/V/I-X5-H/Q-L/A (Lazarow, 2006).

Typically, the peroxisomal proteins involved in reactive oxygen species (ROS) metabolism and in the β-oxidation of long-chain fatty acids. The enzymes catalase, super

oxidise dismutase and peroxisomal acyl-CoA oxidase are often considered classic peroxisomal marker enzymes (Ast et al., 2013; Cross et al, 2016; Kim et al, 2015). Curiously, however, there is a study that shows in many fungi, there is dual targeting of some of the core glycolytic enzymes to peroxisomes and cytosol (Freitag et al., 2012). Additionally, other studies showed that there is cryptic peroxisomal targeting of other enzymes in animal cells (Stiebler et al., 2014, Fodor et al., 2012, Baron et al., 2016, Ast et al., 2013). Dual targeting is the mechanism by which protein isoforms, which are encoded by a single gene, can be located in different compartments within the cell (Kisslov et al., 2014). For example, isoforms of glyceraldehyde 3-phosphate dehydrogenase (GAPDH) and 3-phosphoglycerate kinase (PGK) can be found in both cytoplasm and peroxisome in the pathogenic fungus *Ustilago maydis* (Freitag et al., 2012). The GAPDH is used to catalyse the sixth step of glycolysis pathway whereas the PGK is used for the seventh step (Figure 6.1). The peroxisomal targeting of these isoforms is achieved by ribosomal read-through of the primary stop-codon for PGK and by alternative splicing in the example of GAPDH (Freitag et al., 2012). These enzymes catalyse reversible glycolytic reactions; therefore, they are involved in glycolysis and gluconeogenesis pathways. Since both enzymes utilise NAD^+/NADH and ATP/ADP, they might act in the intraperoxisomal homeostasis of these metabolites. Alternatively, targeting a fraction of these enzymes to the peroxisomes might help to limit futile cycles of ATP hydrolysis, which can take place if both glycolysis and gluconeogenesis run simultaneously (Freitag et al., 2012). The futile cycle occurs when two antagonistic reactions run concurrently with no obvious gain to the organism (Schwender et al., 2004).

The unexpected targeting of the core glycolytic enzymes to the peroxisome in many fungi, coupled to the dual cytosol-peroxisome targeting of several other metabolic enzymes in mammalian cells suggested a further screening for PTS motifs in metabolic enzymes within other eukaryotic lineages. Notably within lineages evolutionarily far from the opisthokonts that include animals and fungi. This chapter is focused on the search and validation of cryptic peroxisomal targeting signal type 1 (cryptic PTS1) motifs in a variety of glycolytic enzymes and other metabolic enzymes across diverse eukaryotic taxa.

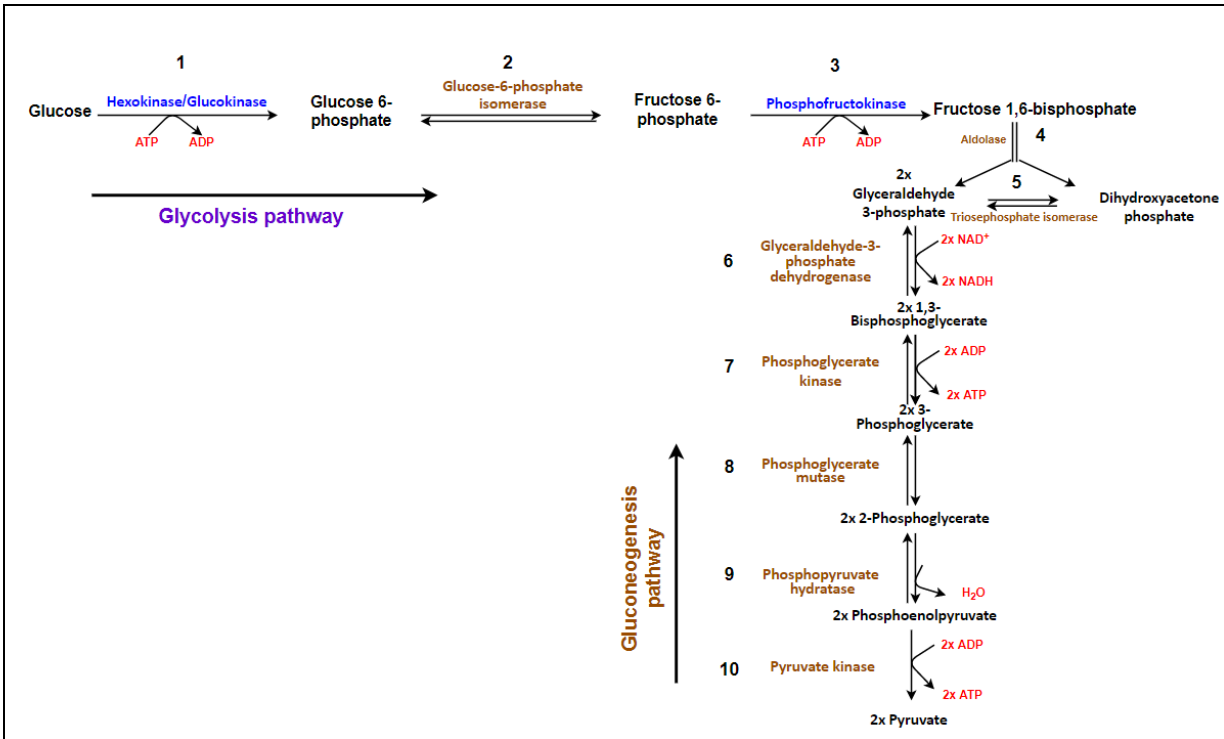


Figure 6.1. Schematic representation of the glycolysis pathway

The glycolytic pathway can be separated into three phases. The first phase is known as the `activation` phase where free glucose molecules are converted to fructose 1,6 bisphosphatase (steps 1 to 3). Two ATP molecules are invested during this phase where they are consumed during the irreversible first and third steps. The second phase is known as `splitting of the 6-carbon sugar` (steps 4 and 5). During this phase, the pentose chain in fructose 1,6 bisphosphatase is cleaved to two molecules of glyceraldehyde-3-phosphate. The final phase is known as the `extraction` phase where the glyceraldehyde-3-phosphate is converted to pyruvate (steps 6 to 10). During this phase, two molecules of ATP are formed via substrate-level phosphorylation. The extraction phase occurs twice per one molecule of glucose. Glycolysis and gluconeogenesis are coordinated inside the cell where one pathway is active and the other one remains inactive. Both pathways share several common enzymes (shown in brown), however, some enzymes catalyse physiologically non-reversible reactions and these are specific for one pathway only (shown in blue).

6.1 In silico identification of PTS1 tripeptides

A survey focused on looking for putative PTS1 in several metabolic enzymes across various unicellular eukaryotes was carried out. This survey focused on eukaryotic groups where putative PTS1 motifs in glycolytic enzymes had not been studied before (Figure 6.2). Although the primary focus was to look for cryptic PTS1 motifs in glycolytic enzymes, additional metabolic enzymes, which also exhibit unexpected PTS1-motifs in fungi and/or animal cells, were also included in this study (Ast et al., 2013). Three main strategies were considered for the identification of putative cryptic PTS1 motifs:

1) To check whether there were multiple genes present in each organism, each encoding different isoforms of the same enzyme and thus whether one or more of the predicted gene products have a putative PTS1 motif;
2) To determine if candidate PTS1 motifs could be detected via a potential for alternative splicing of transcripts encoding glycolytic or other enzymes;
3) To determine if candidate PTS1 motifs could occur as a consequence of ribosomal read-through of stop codons.

The analysis was carried out for twelve different protists from diverse lineages spanning the breadth of unicellular eukaryotic diversity: *Naegleria gruberi*, *N. fowleri*, *Capsaspora owczarzaki*, *Thecamonas trahens*, *Monosiga brevicollis*, *Dictyostelium discoideum*, *Toxoplasma gondii*, *Vitrella brassicaformis, Tetrahymena thermophila*, *Paramecium tetraurelia*, *Cyanoschizon merolae* and *Chlamydomonas reinhardtii* (Figure 6.2). The basis behind selecting these species was mainly because of their position within current views of eukaryotic phylogeny (Burki et al., 2019). For example, *Thecamonas trahens* belongs to Apusozoa phylum, which is considered as an early divergent unicellular sister group to opisthokonta (Cavalier-Smith and Chao, 2010). Moreover, *Dictyostelium discoideum* was selected because of its phylogenetic position, which is after the split between the last common ancestor of animals and plants but before the divergence of Fungi (Figure 6.2). In addition, its capacity to exchange between uni- and multicellularity, makes it an ideal model to study the genetic changes that took place during the evolution of multicellularity (Eichinger et al., 2005). *N. gruberi* was selected because it belongs to an early divergent group of protists, and it has unexpectedly flexible metabolic profile as revealed by analysis of annotated genome sequences (Fritz-Laylin et al., 2010, Opperdoes et al., 2011, Ginger at al., 2010). Understanding the molecular traits of this organism has potential to provide insights into the possible characteristics found in the eukaryotic last common ancestor. In addition to the evolutionary prospects, some species were selected because their peroxisomes have some unique features. For example, peroxisomes in *Toxoplasma gondii* were recently found to be active only at a specific time in the life cycle (Ludewig-Klingner et al., 2017). Moreover, peroxisomes in green algea *C. reinhardtii* are known to lack several peroxisomal enzymes, which are widely conserved in higher plants (*e.g.* peroxisomal enzymes associated with

photorespiration and lipid metabolism). Finally, red algea *C. merolae* is minimalist and one of the few species that contains a single peroxisome (Imoto et al., 2017, Hayashi et al., 2012).

For each taxon, the protein sequences corresponding to all glycolytic enzymes plus metabolic enzymes for which cryptic PTS1 motifs had recently been reported were collected (Ast et al., 2013). To identify in each species whether enzymes were present in multiple isoforms due to the presence of multiple discrete genes, BLASTp was employed against species-specific databases. If the BLAST outputs returned with one or more isoforms, the isoforms were then screened for a putative PTS1 motif at the C-terminus. If PTS1 motifs were not recognised in the primary sequence of the isoforms, then the potential for PTS1 addition by ribosomal read-through or alternative splicing was determined (as described in the method chapter section 2.1.8). If a potential PTS1 sequence was found, two PTS1 predictor programmes were then used to indicate the likelihood of peroxisomal targeting. These were Ppero and PTS1 predictor. PPero provided a numeric value between -3 and 3. Higher values indicated that the protein was more likely to be targeted to the peroxisome (Wang et al, 2017). The PTS1 predictor provided a score based on the query sequence. If the score was higher than or equal to 0, then the query sequence was considered as "targeted". If the score was between -10 and 0, then the query sequence was considered as “twilight zone”. A score lower than -10 means the query sequence is regarded as “not targeted” (Neuberger et al., 2003).

The outcome of my analysis revealed that the presence of cryptic putative PTS1 motifs in metabolic enzymes is notably common among unicellular eukaryotes. With the exception of three taxa, all other protists in my study had at least one glycolytic enzyme with a cryptic putative PTS1 motif. The three exceptions were *Tetrahymena thermophila*, *V. brassicaformis* and *C. reinhardtii* (Figure 6.2). In some glycolytic enzymes, the `cryptic` putative PTS1 motif was found in the protein primary sequence. However, most of the putative cryptic PTS1 motifs in the glycolytic enzymes were seen to be generated via ribosomal read-through or via alternative splicing (Appendices 24 to 35). The molecular technique used to generate the glycolytic isoforms with putative PTS1 motifs varied amongst the different eukaryotes analysed. For example, *N. gruberi*, *D. discoideum*, *M. brevicollis* and *Thecamonas trahens* encoded glycolytic enzymes where putative PTS1 motifs were revealed using ribosomal read-through of the stop codon. The efficiency of ribosomal read-through can be influenced by

many factors, one of which is the identity of the primary stop codon. In all of these species, the stop codon `TGA` was found to be often used in the glycolytic genes that have candidate PTS1 generated via ribosomal read-through. For other glycolytic genes the `TAA` was found to be the predominant stop codon (Appendices 24, 26, 27 and 28). For example, in *D. discoideum* both triosephosphate isomerase and phosphoglycerate kinase genes, which have cryptic putative PTS1 codons generated via ribosomal readthrough, have `TGA` stop codon whereas the rest of the glycolytic genes have `TAA` stop codon. Previous studies had shown that the stop codon `TGA` had the highest potential for ribosomal read through, whereas, `TAA` had the lowest (Dabrowski et al., 2015). The frequent usage of the `TGA` as a stop codon in some of the glycolytic genes with candidate PTS1 provide a level of confidence that these species might use ribosomal read through to generate peroxisomal isoforms for some of the glycolytic enzymes. Besides the identity of the stop codon, the nucleotides which are immediately downstream the stop codon could affect the rate of ribosomal read through. According to studies on mammalian cells and yeast, cytosine and thymine promote a higher rate of ribosomal read though compared to adenine and guanine. Furthermore, nucleotides at the 5` end of the stop codon can also determine the rate of ribosomal read through. The efficiency of ribosomal read through was increased when the position immediately upstream the stop codon was occupied by adenine or generally purine (Cridge et al., 2018, Loughran et al., 2014, Dabrowski et al., 2015). The presence of pyrimidines immediately downstream the stop codon and purine immediately upstream the stop codon could be detected in some of *N. gruberi* glycolytic genes that have putative PTS1 motif generated via ribosomal read through (Appendix 24).

In addition to the ribosomal read-through, alternative splicing was also found to be a potential mechanism used by some protists (*P. tetraurelia*, *Toxoplasma gondii*, *C. merolae* and *C. owczarzaki*) to generate glycolytic enzymes with candidate PTS1 motifs (Appendices 30, 33, 34 and 35). Within the glycolytic enzymes, the following were frequently found to have a potential PTS1: glucose-6-phosphate isomerase (GPI), triosephosphate isomerase (TPI), glyceraldehyde 3 phosphate dehydrogenase (GAPDH) and phosphoglycerate kinase (PGK) (Figure 6.3). TPI, PGK and GAPDH were previously shown to have PTS1 in several fungi. Both PGK and GAPDH are involved in glycolysis and gluconeogenesis pathways where they catalyse the reversible conversion of glyceraldehyde 3-phosphate to 3-phosphoglycerate (Figure 6.1).

These reactions require either the consumption or the production of ATP and NADH. There is a limited exchange between peroxisome and cytosol of NAD^+/NADH and ATP/ADP. Thus, targeting a fraction of GAPDH and PGK to peroxisomes have the potential to provide a mechanism to balance cofactors homeostasis (Freitag et al., 2012).

In addition to the glycolytic enzymes, isoforms of both catalase and citrate synthase were found to have putative PTS1 motifs in many species (Figure 6.4). This was expected, as both of these enzymes are often peroxisomal in eukaryotes analysed to date. In many eukaryotes, catalase is utilised as an antioxidant defense in the peroxisome; catalase breaks down hydrogen peroxide (a major contributor of intracellular ROS) to oxygen and water. In plants, the peroxisomal citrate synthase is involved in carbon transfer from peroxisome to the mitochondria during fatty acid respiration (Piovesan et al., 2004, Rodríguez-Serrano et al., 2007, Schrader and Fahimi, 2006, Tilbrook et al., 2014). Furthermore, superoxide dismutase, which is another enzyme provide ROS defenses, was also predicted to be targeted to peroxisomes in several taxa. This was not unexpected as isoforms of SOD enzymes are found in peroxisomes in several mammals, land plants and trypanosomatids (Kira et al., 2002, Schrader and Fahimi, 2006, Dufernez et al., 2006). The other two enzymes which were found to have putative PTS1 motifs in many species are malate dehydrogenase and isocitrate dehydrogenase. These enzymes are part of the shuttle systems for NADH and NADPH used to maintain interorganellar redox balance in some organelles including mitochondria and peroxisomes. Isoforms of these enzymes are known to be targeting peroxisome in fungi (Ast et al., 2013, Stiebler et al., 2014).

My findings indicate that cryptic targeting of a common suite of metabolic enzymes to peroxisomes is likely to be widespread across eukaryotic phylogeny. It also indicates that the metabolic profile of peroxisomes might be more versatile than often expected. The next stages of my analysis was to try to experimentally prove that some of the cryptic PTS1 motifs identified were indeed capable of conferring peroxisomal import to a protein.

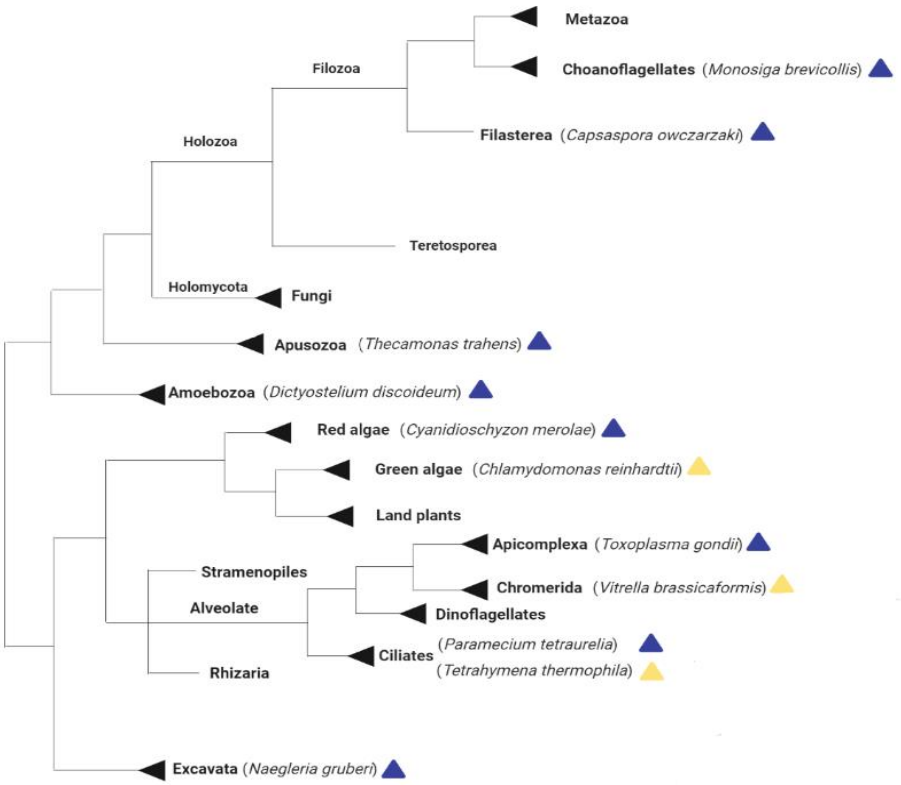


Figure 6.2. Distribution of putative cryptic PTS1 motifs in several metabolic enzymes in unicellular eukaryotes

Blue triangles represent species that were found to have putative cryptic PTS1 in the glycolytic enzymes. Yellow triangles represent species where none of the glycolytic enzymes were seen to have cryptic putative PTS1 motifs.

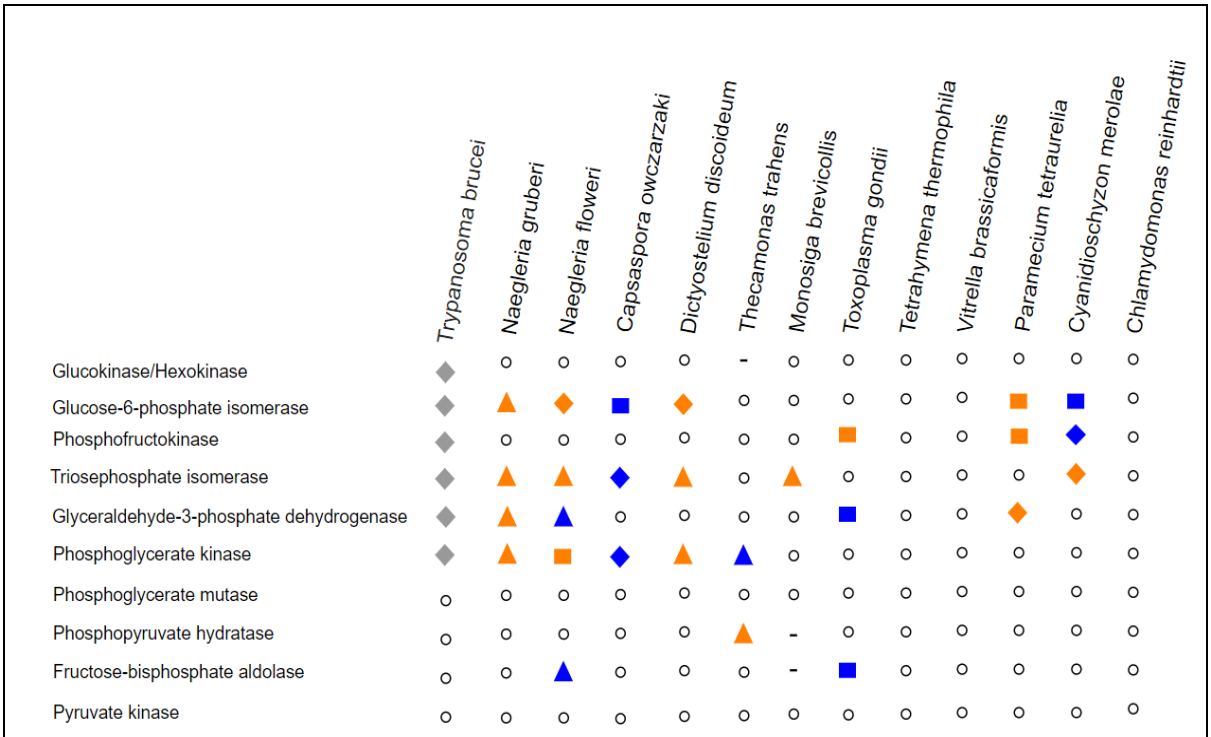


Figure 6.3. The distribution of glycolytic enzymes with cryptic putative PTS1 motifs across different protists Orange colour represents enzymes with a putative cryptic PTS1 as predicted by both PPero and PTS1 predictor web servers. Blue colour represents enzymes with putative PTS1 as predicated by either PPero or PTS1 predictor web servers. Different shapes denote different mechanisms used to generate the PTS1 motif: triangles, PTS1 motifs generated via ribosomal read-through of the stop codon; squares, PTS1 motifs generated via alternative splicing; diamonds, PTS1 motifs found in the protein primary sequence. Grey diamonds represent *T. brucei* glycolytic enzymes with PTS motifs found in the protein primary sequences which were experimentally shown to be targeted to the glycosome (Coley et al., 2011). White circles represent enzymes with no PTS1 motifs. Absence of enzyme from genomes are represented by dashes.

Figure 6.4. The distribution of various metabolic enzymes with cryptic PTS1 motifs across different protists
Orange colour represents enzymes with a putative cryptic PTS1 motifs as predicted by both PPero and PTS1 predictor web servers. Blue colour represents enzymes with putative PTS1 as predicated by either PPero or PTS1 predictor web servers. Different shapes denote different mechanisms used to generate the PTS1 motif: triangles, PTS1 motifs generated via ribosomal read-through of the stop codon; squares, PTS1 motifs generated via alternative splicing; diamonds, PTS1 motifs found in the protein primary sequence. White circles represent enzymes with no PTS1 motifs. Absence of enzyme from genomes are represented by dashes.

6.2 Looking for evidence of alternative splicing in *N. gruberi* using RT-PCR

One way to examine whether the cryptic peroxisomal targeting takes place is by analysing RNA transcripts, where cryptic putative PTS1 motifs are potentially added via alternative splicing. Even though the majority of mRNA will translate to yield the major cytosolic isoform, if alternative splicing takes place, then an alternative minor mRNA species will be generated and translated to yield the minor amount of the peroxisomal isoforms. Alternative splicing can be identified in target gene using RT-PCR assuming that alternative splicing takes place during the growth regimen employed prior the isolation of RNA. The RT-PCR is a laboratory technique which requires reverse transcription of RNA back to complementary DNA (cDNA). With cDNA as a template, PCR can be used to detect whether there are alternative transcripts of a specific gene.

Based on the bioinformatic analysis, several metabolic enzymes in *N. gruberi* were found to have candidate PTS1 motifs that would be revealed (in a translated protein) via alternative splicing of the transcript. These are malate dehydrogenase (MDH), $NADP^+$-dependent isocitrate dehydrogenase (IDH2), citrate synthase (CS), aldehyde reductase (AR), D-ribulose-5-phosphate 3-epimerase (PPE) and isocitrate lyase. Intriguingly, orthologous proteins MDH, IDH2 and AR were also predicted to have cryptic putative PTS1 motifs in a related pathogenic species *N. fowleri* (Appendices 24 and 25).

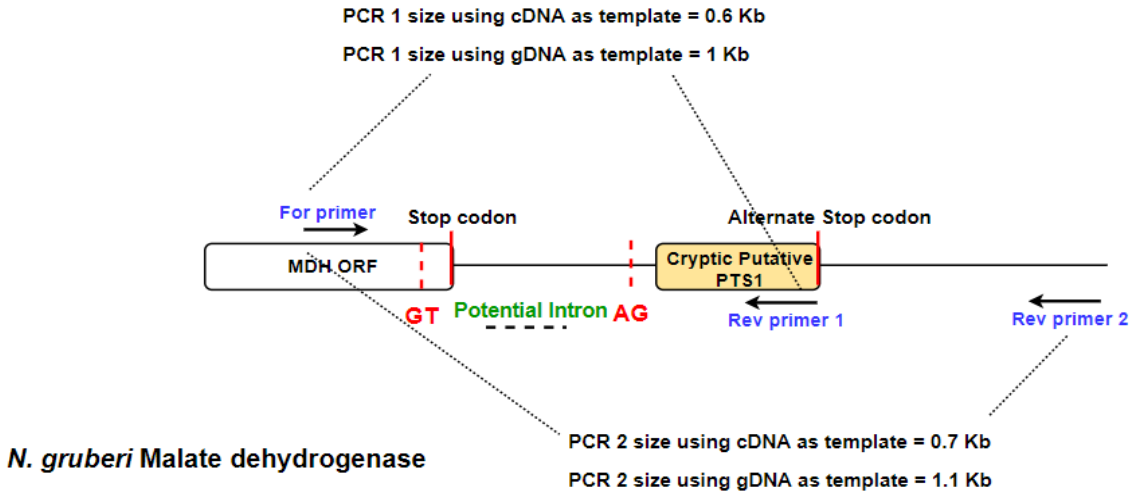
To test the hypothesis that transcripts in *N. gruberi* could be alternatively spliced, RT-PCR was carried out using specific primers that were designed to detect the presence of alternatively spliced MDH, IDH2, PPE and AR transcripts with candidate PTS1 codons. RNA was extracted from xenic *N. gruberi* trophozoites, which were grown in either aerobic or microaerophilic conditions, as described in Chapter 2. The two different growth conditions maximised the opportunity to detect alternatively spliced transcripts yielding a candidate cryptic PTS1 motifs. Using annotated genome sequence of *N. gruberi* (publicly available in NCBI database), one forward primer (For primer) and two reverse primers (Rev primer 1 and Rev primer 2) were designed for each of the MDH, IDH2, PPE and AR genes. The forward primer was located in the middle of the gene coding sequence. Rev primer 1 was located downstream a potential intron and the Rev primer 2 was located between 140 to 200 bp downstream Rev primer 1. The cDNA was synthesized from RNA using SuperScript IV Reverse Transcriptase. For reverse transcription of specific target RNA transcripts, a gene-specific

primer was used for the cDNA synthesis step (Rev primer 2 in Figures 6.5, 6.7, 6.9 and 6.11). Two PCRs were performed for each of MDH, IDH2, AR and PPE. PCR 1 involved the usage of the For primer and Rev primer 1 whereas PCR 2 involved the usage of the For primer and Rev primer 2 (Figures 6.5, 6.7, 6.9 and 6.11). Both PCRs were run using *N. gruberi* cDNA which was synthesized using RNA isolated from either aerobic or microaerophilic conditions. To ensure that primers are specific and mutually compatible, additional PCR was run for each of MDH, IDH2, PPE and AR using *N. gruberi* genomic DNA (gDNA).

For MDH, DNA bands (using cDNA as template) were detected in PCR 1 (microaerophilic conditions) and PCR 2 (in both aerobic and microaerophilic conditions) (Figure 6.5). To detect whether any of these DNA bands have MDH sequence with cryptic putative PTS1 codons, the DNA bands were cloned to pGEM-T Easy and they were sent for sequencing (Figure 6.6). Unfortunately, the outcome of the sequencing data indicates that most of the amplified DNA fragments either belonged to bacterial contamination or to a random region of *N. gruberi* genome (Appendix 40). Similar findings were observed for IDH2, AR and PPE (Figures 6.8, 6.10 and 6.12).

Assuming that the genes analysed are indeed alternatively spliced in *N. gruberi*, the unsuccessful attempt to detect transcripts with candidate PTS1 sequence by RT-PCR might be due to the low percentage of alternatively spliced mRNA relative to the total mRNA produced for the genes analysed. If this was the case, the primers will not be able to detect the peroxisomal transcripts from the crude mixture of cDNA. An alternative approach was used to test whether cryptic PTS1 motifs can confer peroxisomal targting. This approach involved the expression and the detection of peroxisomal localisation in *Crithidia fasciculata*.

A



B

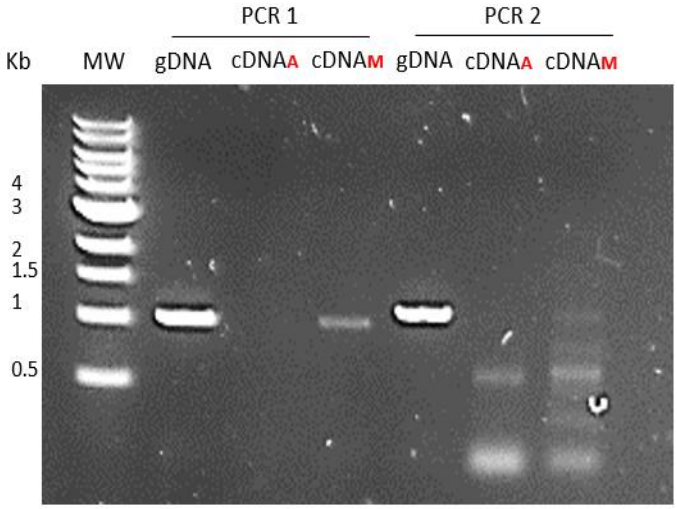


Figure 6.5. RT-PCR to detect potential alternative splicing in *N. gruberi* malate dehydrogenase

(A) Schematic diagram of malate dehydrogenase gene in *N. gruberi* showing primers location used for PCR. The red coloured nucleotide (GT/AG) represent the location of potential introns, which will result in an isoform with putative cryptic PTS1 motif if the predicted alternative splicing takes place. (B) Agarose gel electrophoresis analysis of PCR products of *N. gruberi* malate dehydrogenase. 'A' and 'M' denote cDNA synthesized using RNA isolated from cultures growing in aerobic and microaerophilic conditions respectively. The cDNA was used as a template to detect potential alternative splicing and gDNA was used as a control to show that the primer combinations will work in tandem. For/Rev1 primers were used in PCR 1 whereas For/Rev2 primers were used in PCR 2. 4 µL from each of the PCR products were loaded on 0.8% agarose gel. MW: NEB 1 kb DNA Ladder.

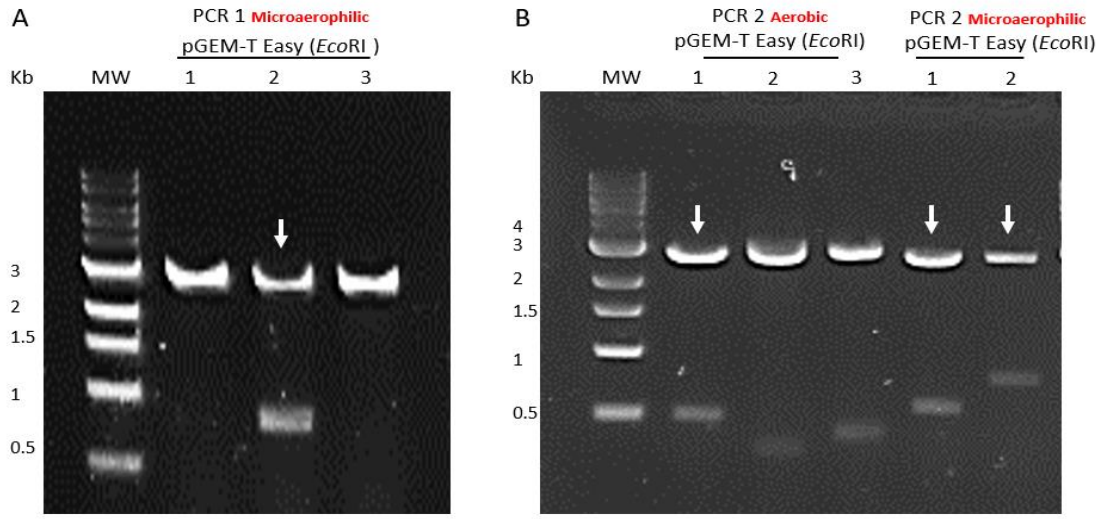
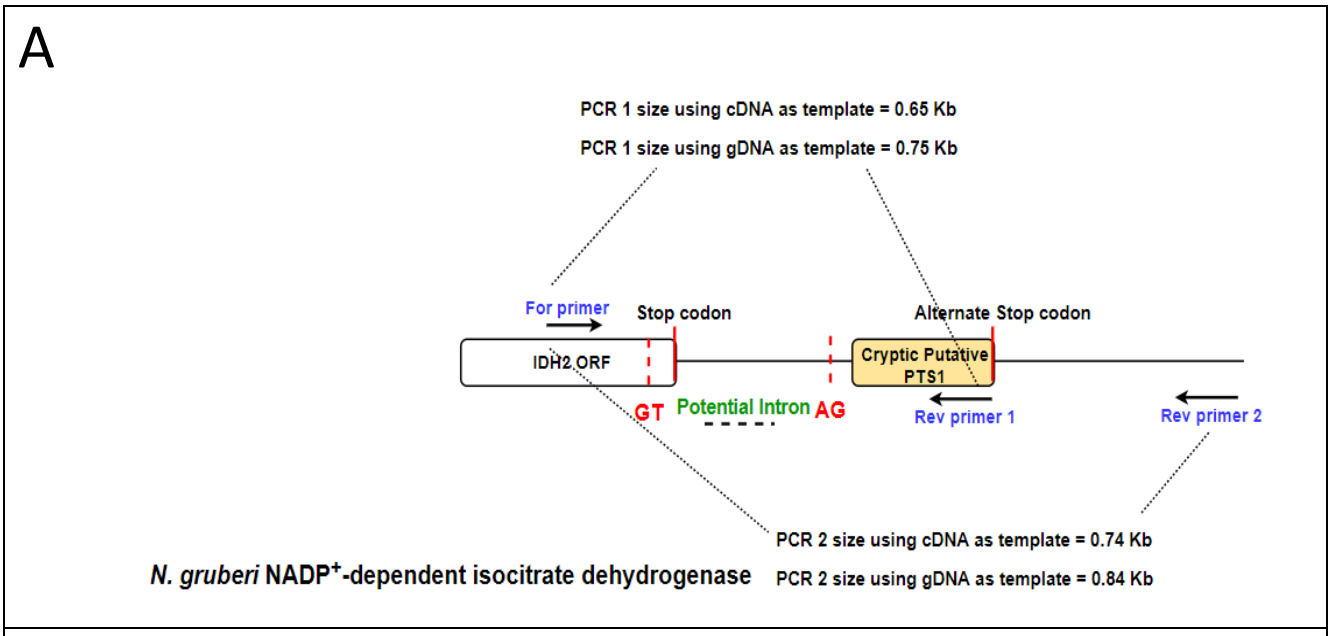


Figure 6.6. Cloning PCR products of *N. gruberi* malate dehydrogenase to pGEM-T Easy

The DNA bands from PCR 1 (microaerophilic) and PCR 2 (aerobic and microaerophilic) were cloned to pGEM-T Easy. Different column numbers represent the different colonies used to isolate pGEM-T Easy plasmid. *Eco*RI enzyme was used for the diagnostic digest of the isolated plasmids. (A) The diagnostic digestion of the pGEM-T Easy plasmids cloned to the DNA fragments from PCR 1. (B) The diagnostic digestion of the pGEM-T Easy plasmids cloned to the DNA fragments from PCR 2. The arrows point towards the plasmids, which were sent for sequencing. The digestive solutions (20 µL) were subjected to electrophorese in a 0.7% agarose gel. MW: NEB 1 kb DNA Ladder.



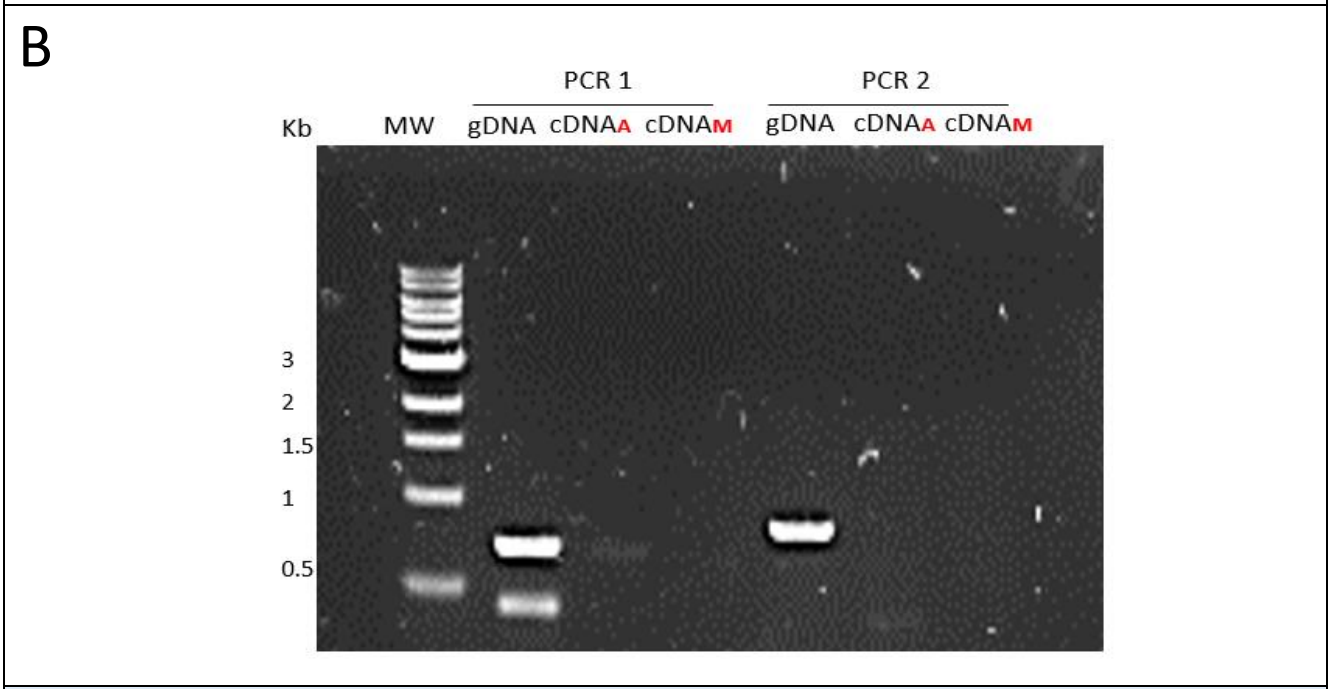


Figure 6.7. RT-PCR to detect potential alternative splicing in *N. gruberi* NADP$^+$-dependent isocitrate dehydrogenase
(A) Schematic diagram of NADP$^+$-dependent isocitrate dehydrogenasegene in *N. gruberi* showing primers location used for PCR. The red coloured nucleotide (GT/AG) represent the location of potential introns, which will result in an isoform with putative cryptic PTS1 motif if the predicted alternative splicing takes place. (B) Agarose gel electrophoresis analysis of PCR products of *N. gruberi* NADP$^+$-dependent isocitrate dehydrogenase. 'A' and 'M' denote cDNA synthesized using RNA isolated from cultures growing in aerobic and microaerophilic conditions respectively. The cDNA was used as a template to detect potential alternative splicing and gDNA was used as a control to show that the primer combinations will work in tandem. For/Rev1 primers were used in PCR 1 whereas For/Rev2 primers were used in PCR 2. 4 µL from each of the PCR products were loaded on 0.8% agarose gel. MW: NEB 1 kb DNA Ladder.

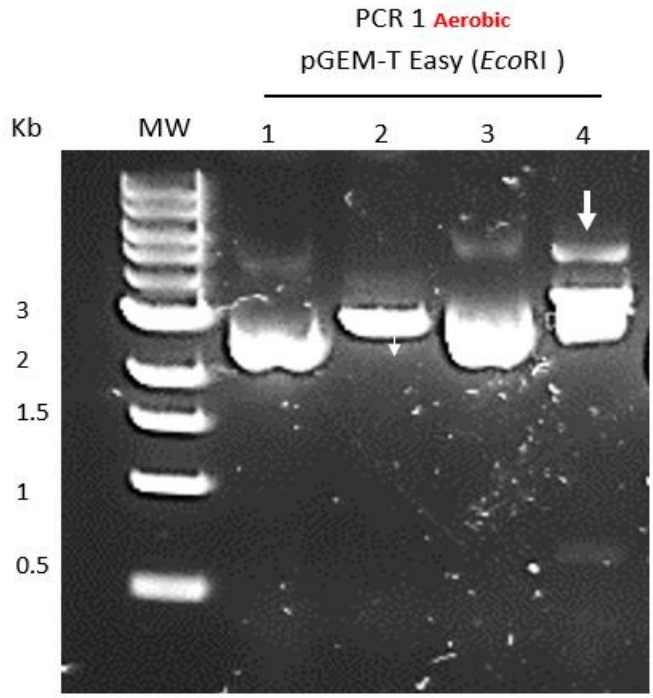
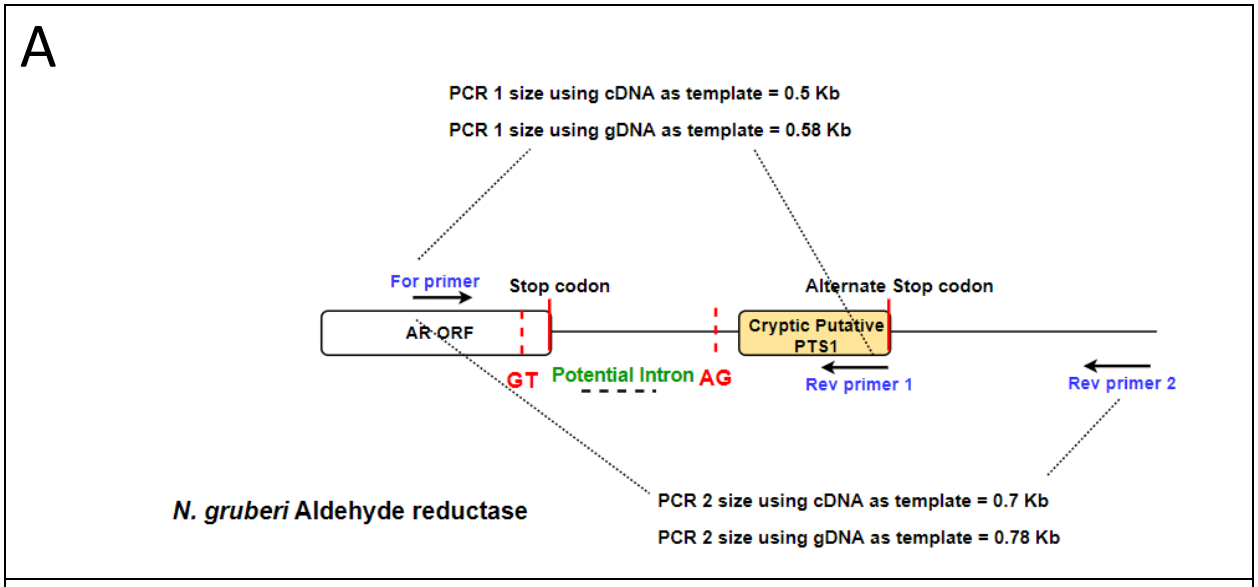


Figure 6.8. Cloning PCR products of *N. gruberi* isocitrate dehydrogenase to pGEM-T Easy

The DNA bands of PCR 1 (aerobic) was cloned to pGEM-T Easy. Different column numbers represent the different colonies used to isolate pGEM-T Easy plasmid. *Eco*RI enzyme was used for the diagnostic digest of the isolated plasmids. The arrow point towards the plasmid, which was sent for sequencing. The digestive solutions (20 µL) were subjected to electrophorese in a 0.7% agarose gel. MW: NEB 1 kb DNA Ladder.



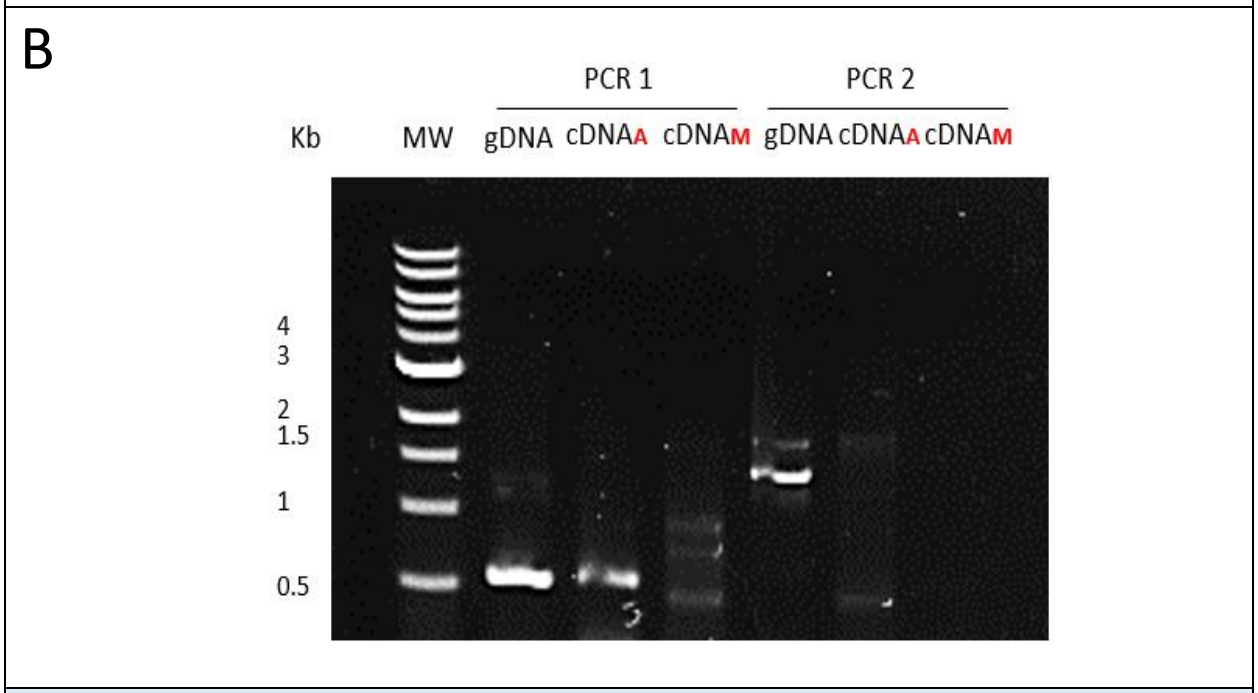


Figure 6.9. RT-PCR to detect potential alternative splicing in *N. gruberi* aldehyde reductase

(A) Schematic diagram of aldehyde reductase gene in *N. gruberi* showing primers location used for PCR. The red coloured nucleotide (GT/AG) represent the location of potential introns, which will result in an isoform with putative cryptic PTS1 motif if the predicted alternative splicing takes place. (B) Agarose gel electrophoresis analysis of PCR products of *N. gruberi* aldehyde reductase. 'A' and 'M' denote cDNA synthesized using RNA isolated from cultures growing in aerobic and microaerophilic conditions respectively. The cDNA was used as a template to detect potential alternative splicing and gDNA was used as a control to show that the primer combinations will work in tandem. For/Rev1 primers were used in PCR 1 whereas For/Rev2 primers were used in PCR 2. 4 µL from each of the PCR products were loaded on 0.8% agarose gel. MW: NEB 1 kb DNA Ladder.

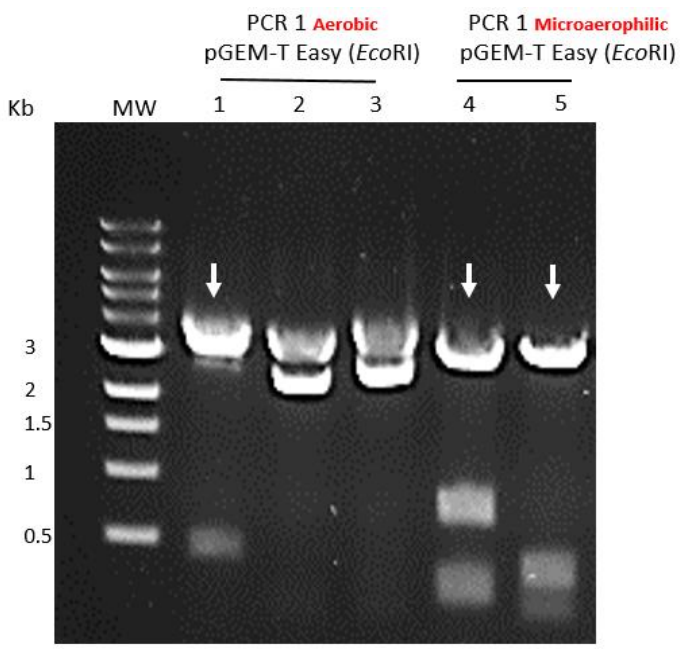
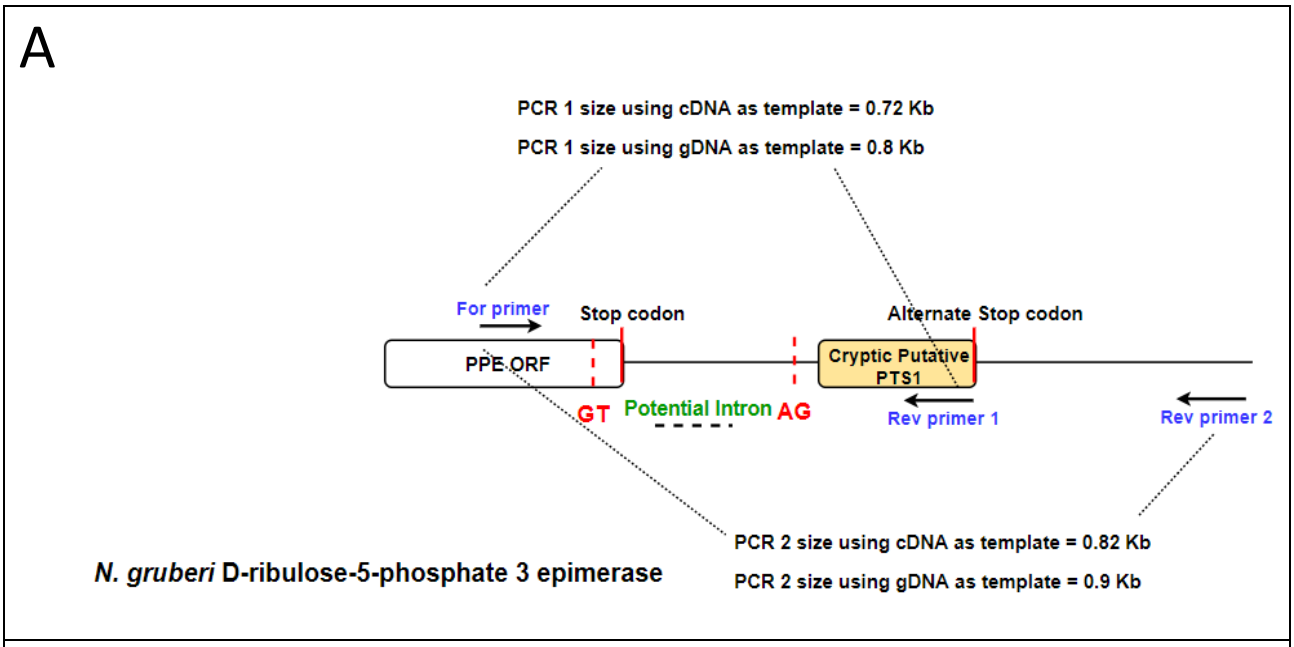


Figure 6.10. Cloning the PCR products of *N. gruberi* aldehyde reductase to pGEM-T Easy
The DNA bands from PCR 1 (aerobic and microaerophilic) were cloned to pGEM-T Easy. Different column numbers represent the different colonies used to isolate pGEM-T Easy plasmid. *Eco*RI enzyme was used for the diagnostic digest of the isolated plasmids. The arrows point towards the plasmids, which were sent for sequencing. The digestive solutions (20 µL) were subjected to electrophorese in a 0.7% agarose gel. MW: NEB 1 kb DNA Ladder.



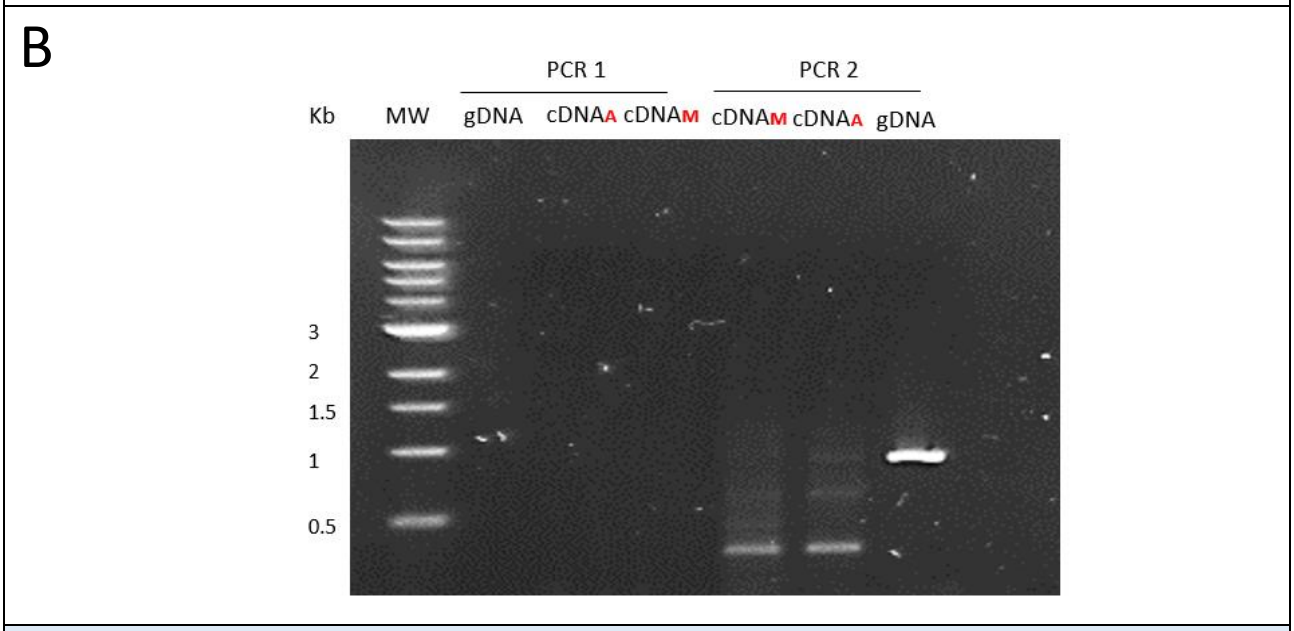


Figure 6.11. RT-PCR to detect potential alternative splicing in *N. gruberi* D-ribulose-5-phosphate 3 epimerase
(A) Schematic diagram of D-ribulose-5-phosphate 3 epimerase gene in *N. gruberi* showing primers location used for PCR. The red coloured nucleotide (GT/AG) represent the location of potential introns, which will result in an isoform with putative cryptic PTS1 motif if the predicted alternative splicing takes place. (B) Agarose gel electrophoresis analysis of PCR products of *N. gruberi* D-ribulose-5-phosphate 3 epimerase. 'A' and 'M' denote cDNA synthesized using RNA isolated from cultures growing in aerobic and microaerophilic conditions respectively. The cDNA was used as a template to detect potential alternative splicing and gDNA was used as a control to show that the primer combinations will work in tandem For/Rev1 primers were used in PCR 1 whereas For/Rev2 primers were used in PCR 2. 4 µL from each of the PCR products were loaded on 0.8% agarose gel. MW: NEB 1 kb DNA Ladder.

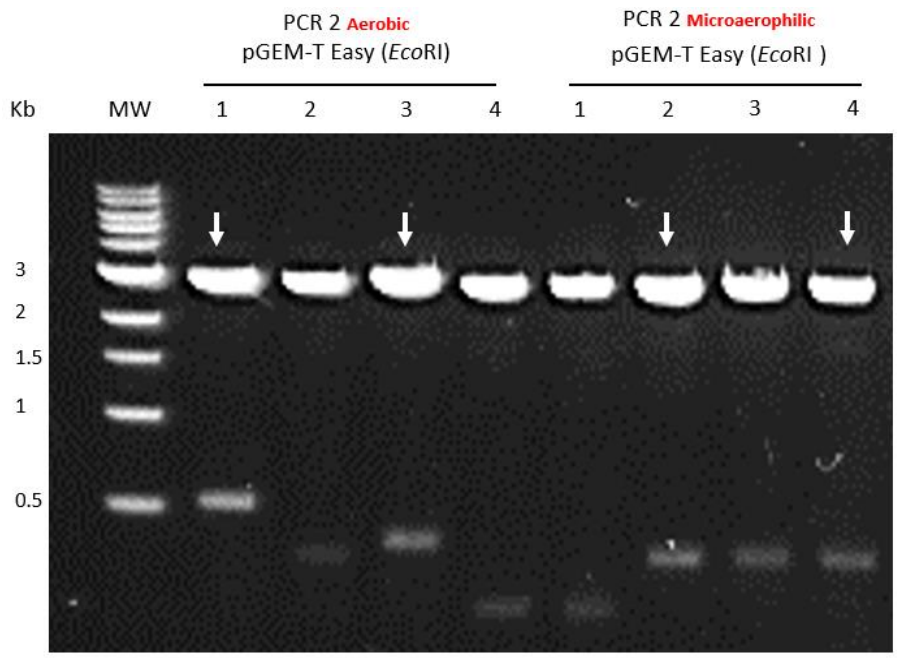


Figure 6.12. Cloning PCR products of *N. gruberi* D-ribulose-5-phosphate 3 epimerase to pGEM-T Easy
The DNA bands from PCR 2 (aerobic and microaerophilic) were cloned to pGEM-T Easy. Different column numbers represent the different colonies used to isolate pGEM-T Easy plasmid. *Eco*RI enzyme was used for the diagnostic digest of the isolated plasmids. The arrows point towards the plasmids, which were sent for sequencing. The digestive solutions (20 µL) were subjected to electrophorese in a 0.7% agarose gel. MW: NEB 1 kb DNA Ladder.

6.3 Peroxisomal targeting of GFP conferred by cryptic *N. gruberi* PTS1 motifs

Protein localisation via immunofluorescence was used to test whether *N. gruberi* cryptic PTS1 motifs in some of the metabolic enzymes can confer peroxisomal targeting. Four *N. gruberi* metabolic enzymes were selected, two of which are glycolytic enzymes (PGK and TPI); the remaining two were AR and SOD. These enzymes have putative PTS1 either by ribosomal read-through (for PGK, TPI and SOD) or via alternative splicing (for AR) (image A in Figures 6.18 to 6.21). All of these enzymes were identified to have candidate peroxisomal isoforms in *N. fowleri*, and were predicted by PPero programme to be targeted to the peroxisomes in *N. gruberi* (Appendix 41).

There is no reverse genetics for *N. gruberi* thus switching to heterologous expression in *Crithidia fasciculata* that is distantly related to *N. gruberi* was carried out assuming that the Pex5 in *C. fasciculata* will recognise the same PTS1 motifs as Pex5 in *N. gruberi*. The immunofluorescence technique was used to test whether the C-terminal extensions with cryptic putative PTS1 sequences from these enzymes, generated via ribosomal read-through or alternative splicing, could be used to target green fluorescent protein (GFP) into peroxisomes. For C-terminal extensions with candidate PTS1 generated via ribosomal read-through, the primary stop codons were replaced with *bona fide* codons as described by Dabrowski et al (Dabrowski et al, 2015). For PGK and TPI, the stop codon `TGA` was replaced by Arg codon and for SOD the stop codon `TAA` was replaced by Glu codon. The last 75 nucleotide sequences of each of the four enzymes' genes, which include the putative cryptic PTS1 codons, were fused in-frame at the N- terminal with Myc tag and GFP (Myc-GFP-$PTS1_{Extension}$). These gene constructs were cloned to pNUS vector and were transfected in *C. fasciculata*. First, the gene construct for PGK was sent to be synthetically synthesized (Myc-GFP-$PGK_{Extension}$). The synthesised gene was removed from the backbone vector via gel extraction and then it was cloned to pNUS vector (Figure 6.13). Following that, the cloned pNUS vector was used to prepare the gene constructs for SOD, TPI and AR. To achieve this, the C-terminal extension of PGK was removed from pNUS vector via gel extraction and it was replaced by a 75 bp of C-terminal extension of SOD, TPI or AR respectively. The last 75 nucleotide sequence of these genes were prepared by annealing two complementary oligonucleotides strands. In the process of designing these oligonucleotides, restriction

enzyme sites were inserted into each gene to allow directional cloning as well as restriction digest mapping. Since different GFP products would have the same size in Western blot, additional specific restriction sites were inserted to minimise the chances of mixing or exchanging plasmids. Figures 6.14 to 6.16 summarise the restriction digest mapping of the cloned vectors. All cloned vectors were sent for sequencing to confirm the correct integration (for DNA sequence alignments see Appendices 36-39).

The transfection method in *C. fasciculata* was not straightforward. To optimise the chances for successful transfection, different electroporation parameters and electroporation buffers (including cytomix, ZMG and warren culture medium) were used to find the settings that work the best with *C. fasciculata*. After finding the optimal conditions for *C. fasciculata* transfection (as detailed in the method section), the cloned pNUS vectors were transfected to wild type *C. fasciculata* cells. After transfection, the cells required seven days to reach their normal growth rate with generation time between 4 to 6 hours. Once the cells reached a healthy state, they were settled and fixed on microscope slides and prepared for indirect immunofluorescence using polyclonal antibodies detecting the peroxisomal enzyme GAPDH (peroxisomal marker). When viewed by Airyscan confocal microscopy a punctate GFP signal was observed for all transfected cells but not in the wild type cells (control sample) (Figures 6.18 - 6.23). As illustrated in Figures 6.18 to 6.22, there is co-localisation between the peroxisomal marker and the PTS1-bearing GFP in each transfected cell line. On a basis of these heterologous expression experiments, and coupled with the bioinformatics comparisons between *N. gruberi* and *N. fowleri* this suggested that isoforms of PGK, TPI, SOD and AR with putative PTS1 can be targeted to peroxisomes in *N. gruberi* following either stop codon read-through (for PGK, TPI and SOD) or alternative splicing of nascent RNA (for AR).

To confirm that proteins were translated and are stable with the expected molecular mass in *C. fasciculata*, Western blot analysis was essential. Using anti-Myc antibody (Abcam) and anti-mouse IgG HRP-linked as primary and secondary antibodies, a Western blot analysis was carried out to detect the protein expression of myc-GFP-PTS1$_{Extension}$. The lysates equivalent to 10^7 cells from the transformed populations or the wild type population (control) were analysed by SDS-PAGE. The first SDS-PAGE gel was stained using InstantBlue to show complete protein content. The second SDS-PAGE gel was used for immunoblotting; the resultant nitrocellulose membrane was probed with an anti-Myc monoclonal antibodies

followed by anti-IgG HRP-linked secondary antibodies. From image A in Figure 6.17, the SDS-PAGE gel stained with InstantBlue showed equal amounts of cell lysate were loaded between samples which can be observed in the wild type as well as in the transformed populations. However, when the protein bands were probed against Myc-tag, protein bands were only observed in the transformed populations but not in the wild type. The presence of proteins bands with relatively the same size as Myc-GFP-PTS1$_{Extension}$ (≈ 32 kDa) in the transformed populations confirms the protein expression (shown by the red arrow in Figure 6.17). Collectively, these results indicate that *N. gruberi* cryptic PTS1 motifs, in several metabolic enzymes including some of the main glycolytic enzymes, can confer peroxisomal targeting. Hence the functional properties of peroxisome organelle might be far more complicated than it was previously anticipated.

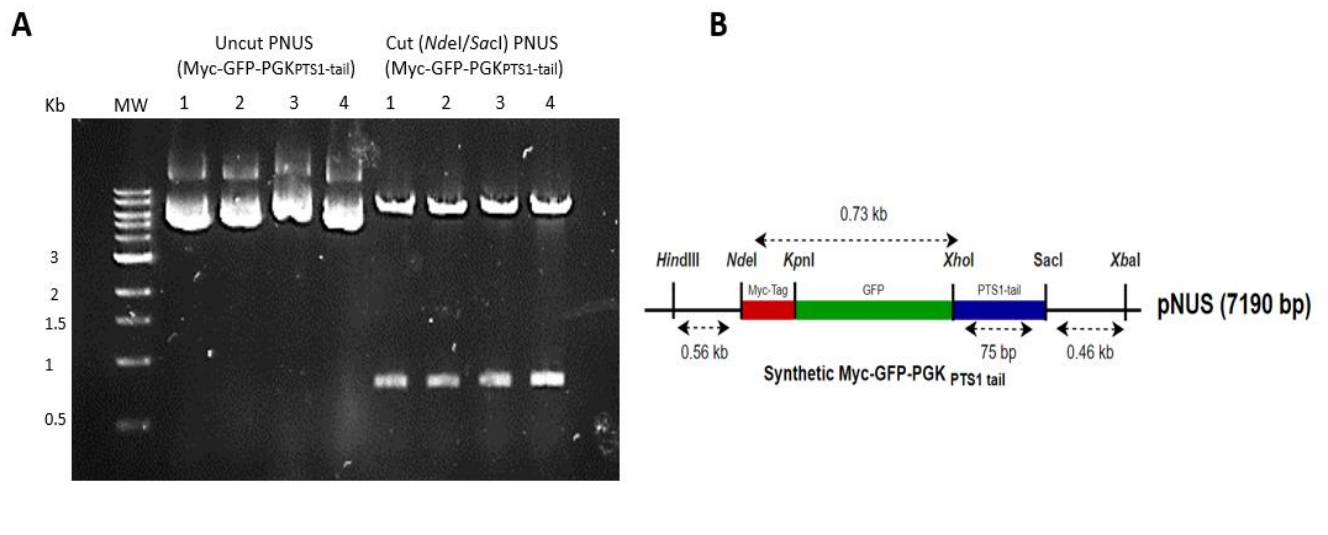


Figure 6.13. Cloning *N. gruberi* PGK synthetic gene construct into pNUS expression vector
The last 75 nucleotides of PGK were fused in-frame at the N-terminus with Myc tag and GFP sequence (Myc-GFP-PTS1$_{Extension}$). (A) Agarose gel electrophoresis analysis of diagnostic digest of the cloned pNUS vector. The diagnostic digest was carried out using *Nde*I and *Sac*I, which released the Myc-GFP-PTS1$_{Extension}$ gene (0.8 kb) from the vector backbone. The digestive solutions (20 µL) were subjected to electrophorese in a 0.7% agarose gel. MW: NEB 1 kb DNA Ladder. (B) A schematic diagram, which shows the location of different restriction enzyme sites within the gene construct. The plasmids were sent for DNA sequencing using a reverse primer to ensure correct integration of the DNA insert.

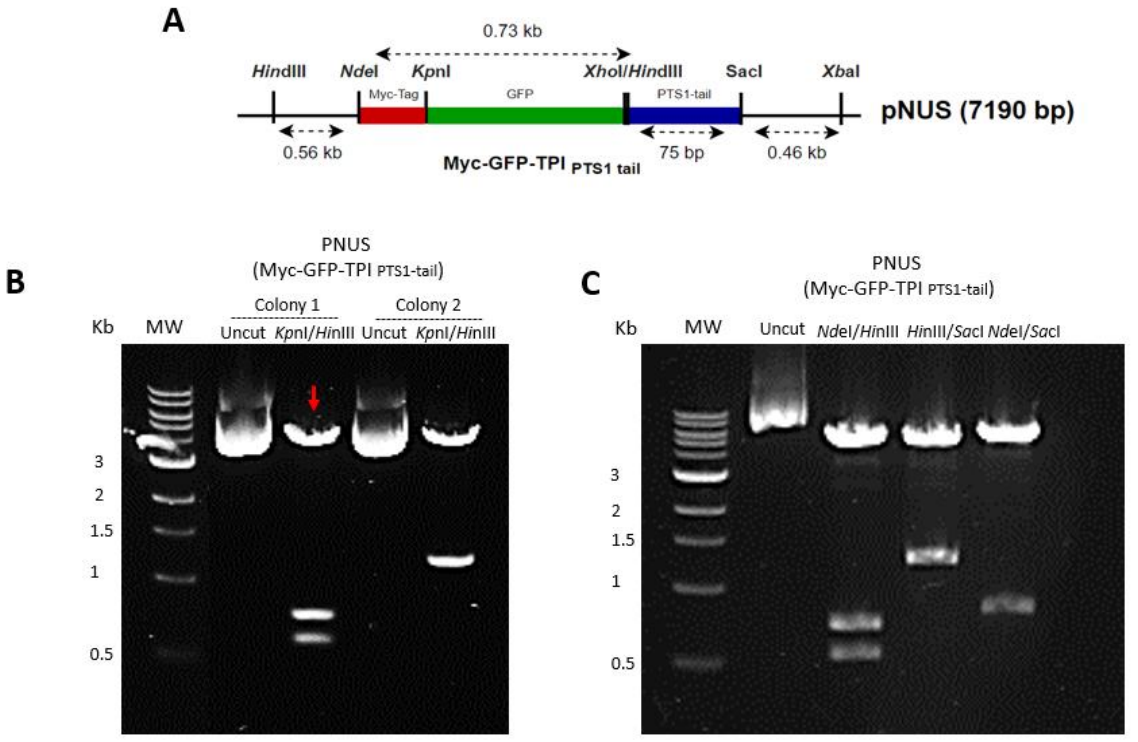


Figure 6.14. Cloning the annealed oligonucleotides of *N. gruberi* candidate PTS1 extension of TPI gene into pNUS vector

(A) Schematic diagram which shows the location of different restriction enzymes within the cloned plasmid. (B) Agarose gel electrophoresis of diagnostic digest of cloned pNUS vector. The diagnostic digest was carried out using *Kp*nI and *Hin*dIII, which released the GFP-PTS1$_{Extension}$ from the vector backbone. (C) Agarose gel electrophoresis analysis of further diagnostic digests of cloned pNUS vector colony 1 (red arrow). The diagnostic digest using *Nde*I/ *Hin*dIII results in two DNA bands (exclude the pNUS backbone) at 0.73 kb and 0.53 kb. The diagnostic digest using *Hin*dIII/*Sac*I results in two DNA bands at 1.3 kb and 75 bp. The diagnostic digest using *Nde*I/*Sac*I results in one band at 0.8 kb. The digestive solutions (20 µL) were subjected to electrophorese in a 0.7% agarose gel. MW: NEB 1 kb DNA Ladder. The plasmids were sent for DNA sequencing using a reverse primer to ensure correct integration of the DNA insert.

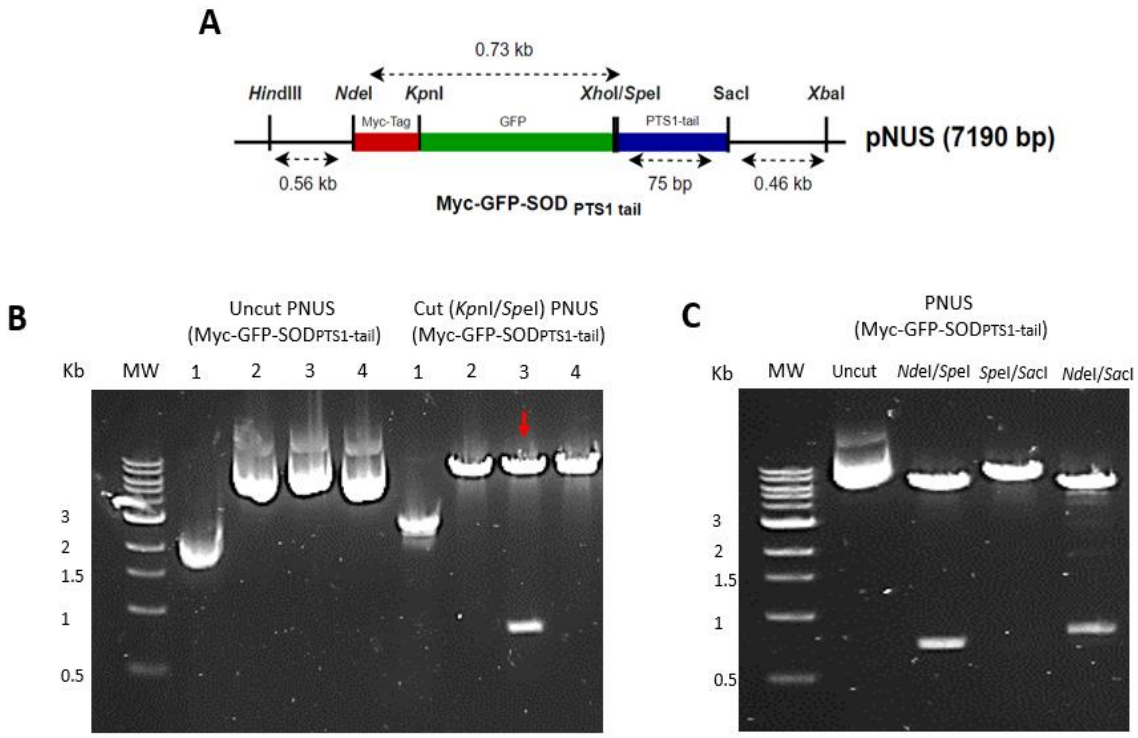


Figure 6.15. Cloning the annealed oligonucleotides of *N. gruberi* candidate PTS1 extension of SOD gene into pNUS vector

(A) Schematic diagram shows the location of different restriction enzymes within the cloned plasmid. (B) Agarose gel electrophoresis of diagnostic digest of cloned pNUS vector. The diagnostic digest was carried out using *Kpn*I and *Spe*I, which released the GFP-PTS1$_{Extension}$ gene from the vector backbone. (C) Agarose gel electrophoresis analysis of further diagnostic digests of cloned pNUS vector colony 3 (red arrow). The diagnostic digest using *Nde*I/*Spe*I results of two DNA bands (exclude the pNUS backbone) at 0.73kb and 0.53 kb. The diagnostic digest using *Spe*I/*Sac*I results of two bands at 1.3 kb and 75 bp. The diagnostic digest using *Nde*I/*Sac*I results of one band at 0.8 kb. The digestive solutions (20 µL) were subjected to electrophorese in a 0.7% agarose gel. MW: NEB 1 kb DNA Ladder. The plasmid was sent for DNA sequencing using a reverse primer to ensure the correct integration of the DNA insert.

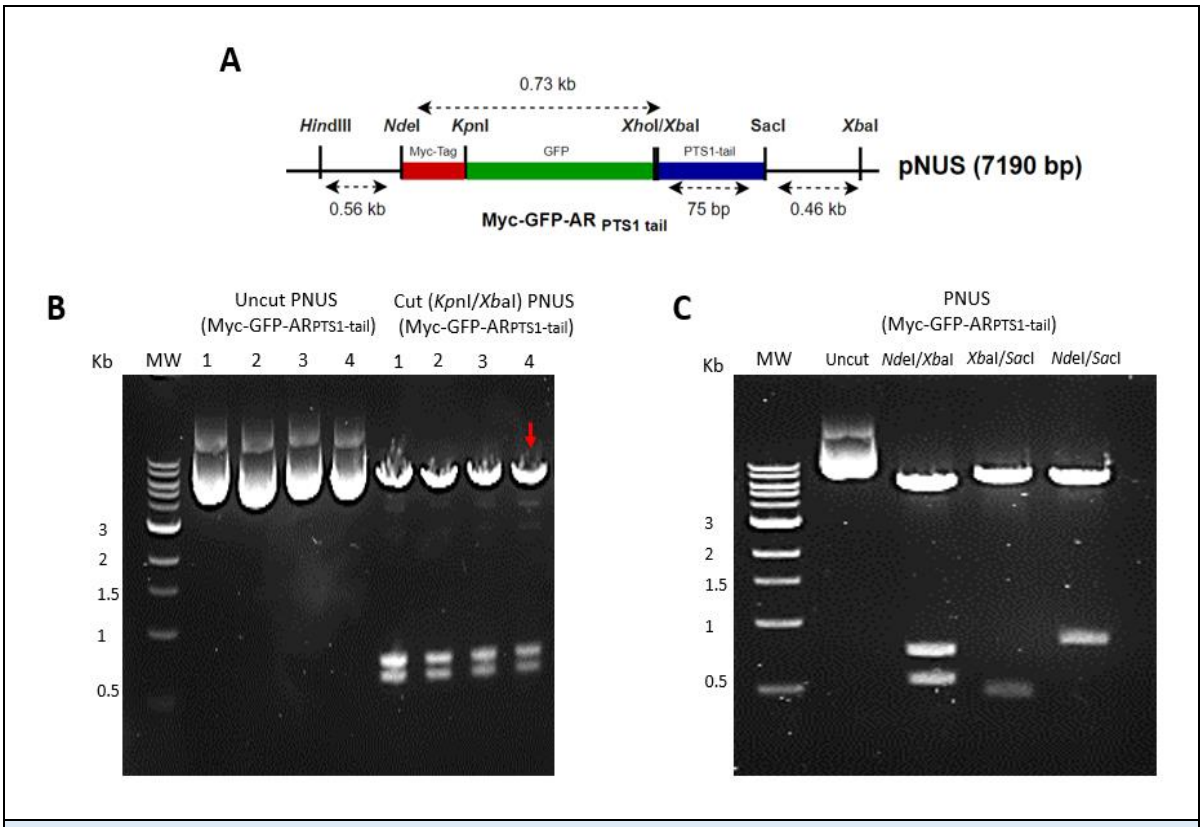


Figure 6.16. Cloning the annealed oligonucleotides of *N. gruberi* candidate PTS1 extension of AR gene into pNUS vector

(A) Schematic diagram shows the location of different restriction enzymes within the cloned plasmid. (B) Agarose gel electrophoresis analysis of diagnostic digest of cloned pNUS vector. The diagnostic digest was carried out using *Kp*nI and *Xb*aI which released the GFP-PTS1 Extension gene (0.77kb) from the vector backbone. (C) Agarose gel electrophoresis analysis of further diagnostic digests of cloned pNUS vector colony 4 (red arrow). The diagnostic digest using *Nd*eI/ *Xb*aI results of two DNA bands (exclude the pNUS backbone) at 0.73kb and 0.53kb. The diagnostic digest using *Xb*aI/*Sa*cI results of two bands at 1.3kb and 7bp. The diagnostic digest using *Nd*eI/*Sa*cI results of one band at 0.8 kb. The digestive solutions (20 µL) were subjected to electrophorese in a 0.7% agarose gel. MW: NEB 1 kb DNA Ladder. The plasmid was sent for DNA sequencing using a reverse primer to ensure correct integration of the DNA insert.

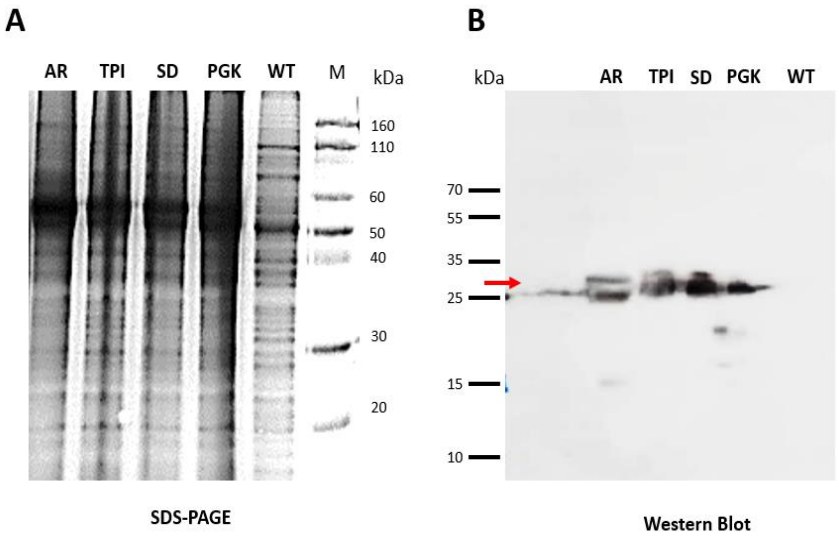


Figure 6.17. Western blot analysis of Myc-GFP-PTS1$_{Extension}$ expression in *Crithidia fasciculate*

C. fasciculata whole cell lysates (equivalent to 10^7) were analysed by SDS-PAGE using a 10% polyacrylamide gel. (A) The SDS-PAGE gel was stained with InstantBlue to show the complete protein content of the cell lysates. (B) Immunoblot detection of protein expression of the Myc-GFP-PTS1$_{Extension}$. The relative molecular mass of the protein construct is equivalent to ~32 kDa. PGK corresponds to phosphoglycerate kinase, TPI corresponds to triosephosphate isomerase, SOD corresponds to superoxide dismutase and AR corresponds to aldehyde Reductase. The wild type *C. fasciculata* (WT)was used as a control sample.

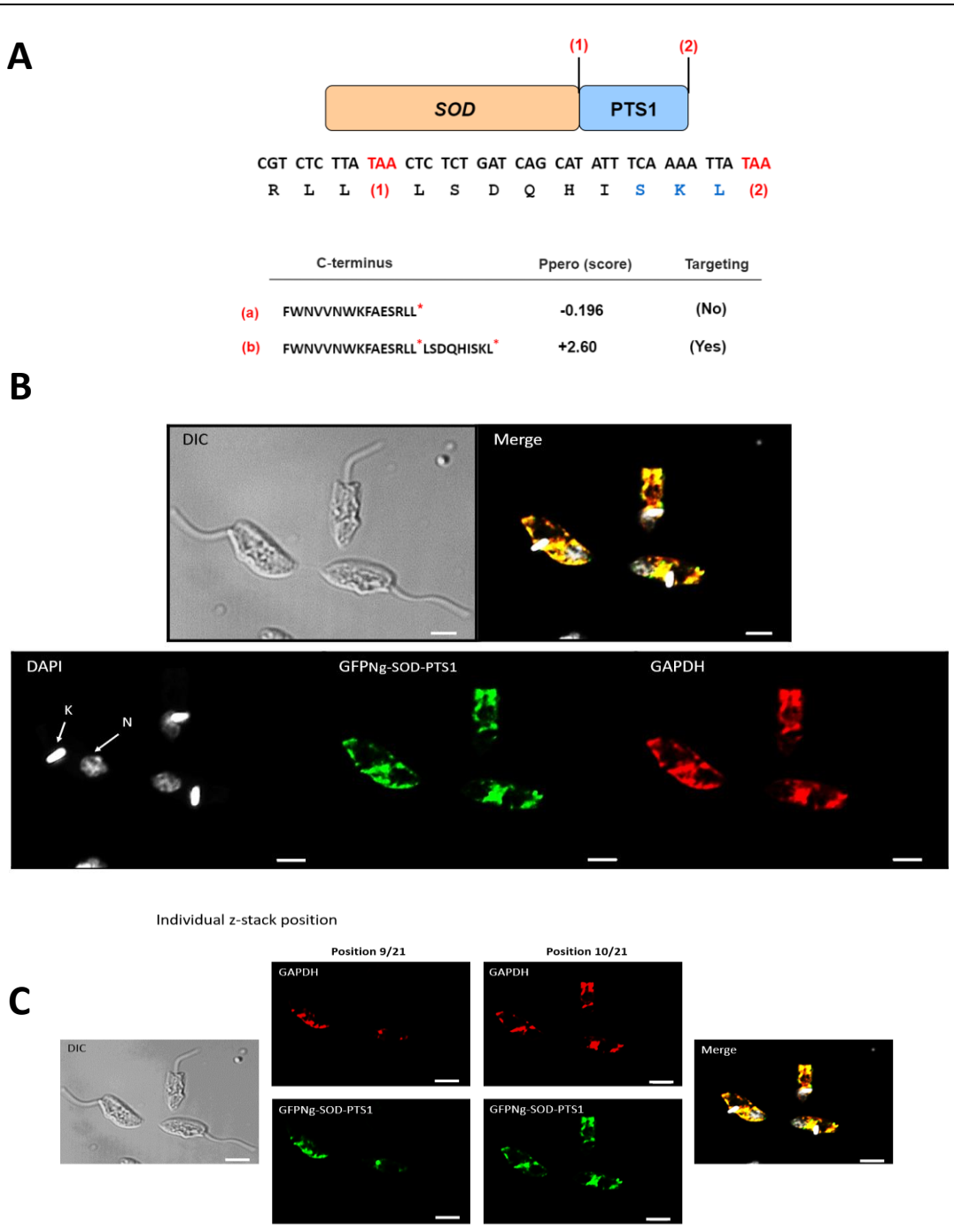


Figure 6.18. Gene expression in *C. fasciculata* of *N. gruberi* superoxide dismutase with a candidate cryptic PTS1 motif
(A) Cartoon schematic of *N. gruberi* superoxide dismutase gene showing the position of the stop codon (1) and (2) and a candidate cryptic PTS1 targeting motif available by ribosomal read-through. The DNA sequence corresponding to the C-terminal coding sequence through of stop codon (2) is also shown. The tripeptide predicted PTS1 is shown in blue. Prediction of peroxisomal targeting using stop codon (1) and (2) is provided. (B) Assessment of cryptic PTS1 candidature. The airyscan projection shown co-localisation of GFP fused to the C-terminal 25 amino acids of *N. gruberi* superoxide dismutase assuming translation read-through of the stop codon (1) (GFPNg-SOD-PTS1) with indirect immunofluorescence signal for polyclonal antisera detecting glycosomal GAPDH in paraformaldehyde-fixed *C. fasciculata*. Nuclear (N) and mitochondria (K) DNA were detected by staining with 4',6-diamidino-2-phenylindole (DAPI). (C) Individual z stacks from the projection shown in B also indirect co-localisation of GFPNg-SOD-PTS1 with glycosomal GAPDH. Scale bars, 2µm.

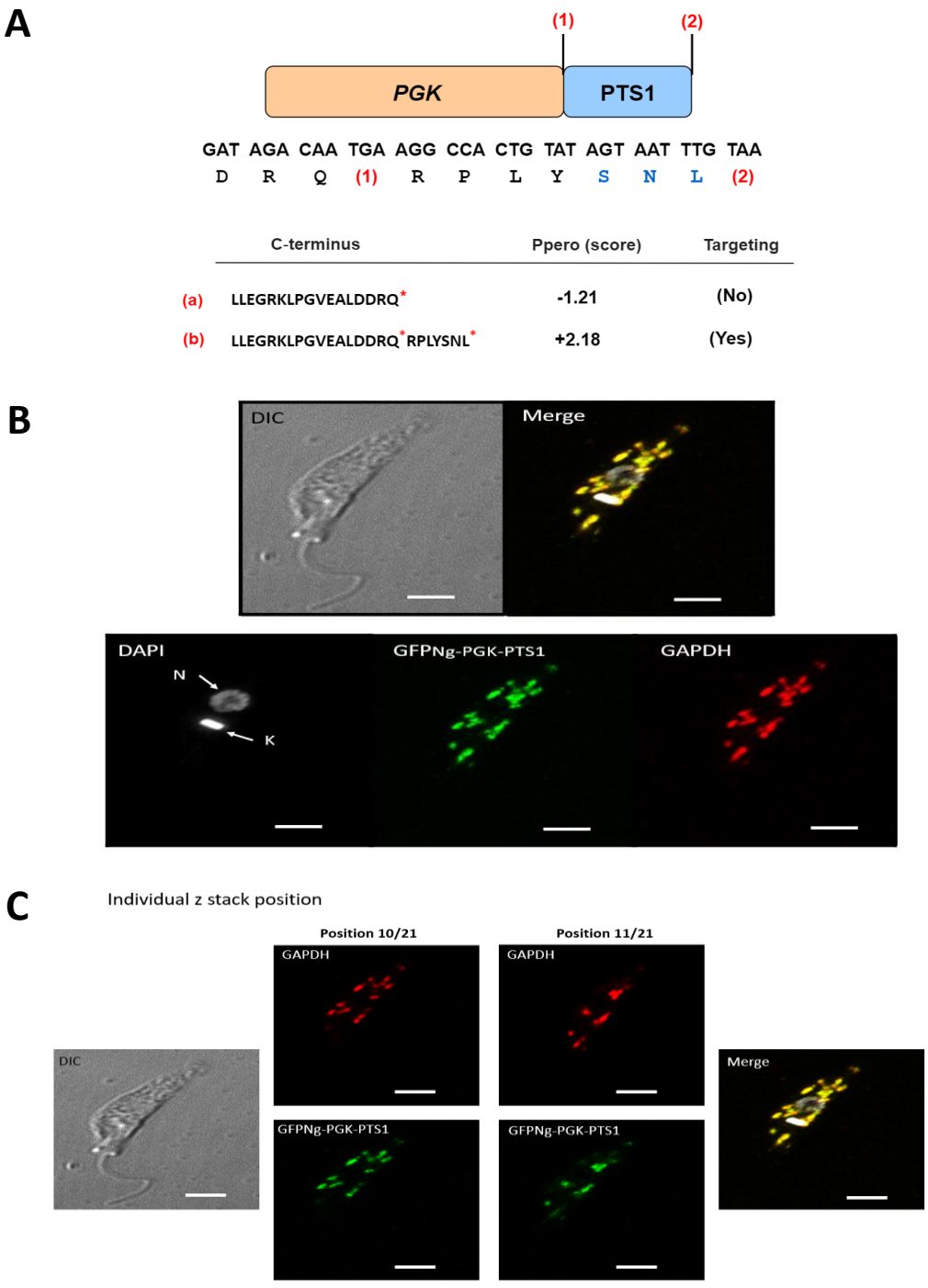


Figure 6.19. Gene expression in *C. fasciculata* of *N. gruberi* phosphoglycerate kinase with a candidate cryptic PTS1 motif. (A) Cartoon schematic of *N. gruberi* phosphoglycerate kinase gene showing the position of the stop codon (1) and (2) and a candidate cryptic PTS1 targeting motif available by ribosomal read-through. The DNA sequence corresponding to the C-terminal coding sequence through of stop codon (2) is also shown. The tripeptide predicted PTS1 is shown in blue. Prediction of peroxisomal targeting using stop codon (1) and (2) is provided. (B) Assessment of cryptic PTS1 candidature. The airyscan projection shown co-localisation of GFP fused to the C-terminal 25 amino acids of *N. gruberi* phosphoglycerate kinase assuming translation read-through of the stop codon (1) (GFPNg-PGK-PTS1) with indirect immunofluorescence signal for polyclonal antisera detecting glycosomal GAPDH in paraformaldehyde-fixed *C. fasciculata*. Nuclear (N) and mitochondria (K) DNA were detected by staining with DAPI. (C) Individual z stacks from the projection shown in B also indirect co-localisation of GFPNg-PGK-PTS1 with glycosomal GAPDH. Scale bars, 2um.

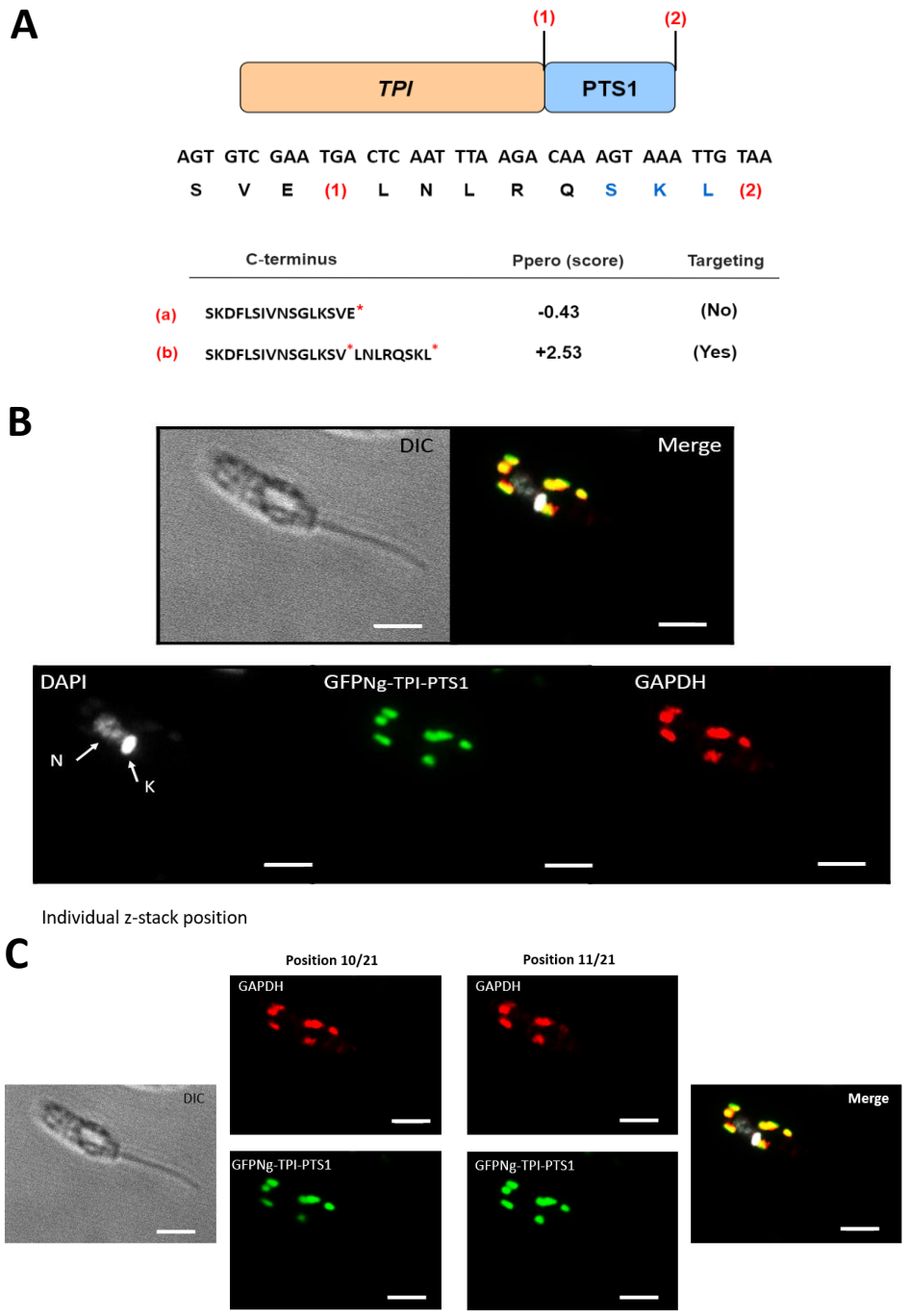


Figure 6.20. Gene expression in *C. fasciculata* of *N. gruberi* triosephosphate isomerase with a candidate cryptic PTS1 motif. (A) Cartoon schematic of *N. gruberi* triosephosphate isomerase gene showing the position of the stop codon (1) and (2) and a candidate cryptic PTS1 targeting motif available by ribosomal read-through. The DNA sequence corresponding to the C-terminal coding sequence through of stop codon (2) is also shown. The tripeptide predicted PTS1 is shown in blue. Prediction of peroxisomal targeting using stop codon (1) and (2) is provided. (B) Assessment of cryptic PTS1 candidature. The airyscan projection shown co-localisation of GFP fused to the C-terminal 25 amino acids of *N. gruberi* triosephosphate isomerase assuming translation read-through of the stop codon (1) (GFPNg-TPI-PTS1) with indirect immunofluorescence signal for polyclonal antisera detecting glycosomal GAPDH in paraformaldehyde-fixed *C. fasciculata*. Nuclear (N) and mitochondria (K) DNA were detected by staining with DAPI. (C) Individual z stacks from the projection shown in B also indirect co-localisation of GFPNg-TPI-PTS1 with glycosomal GAPDH. Scale bars, 2µm.

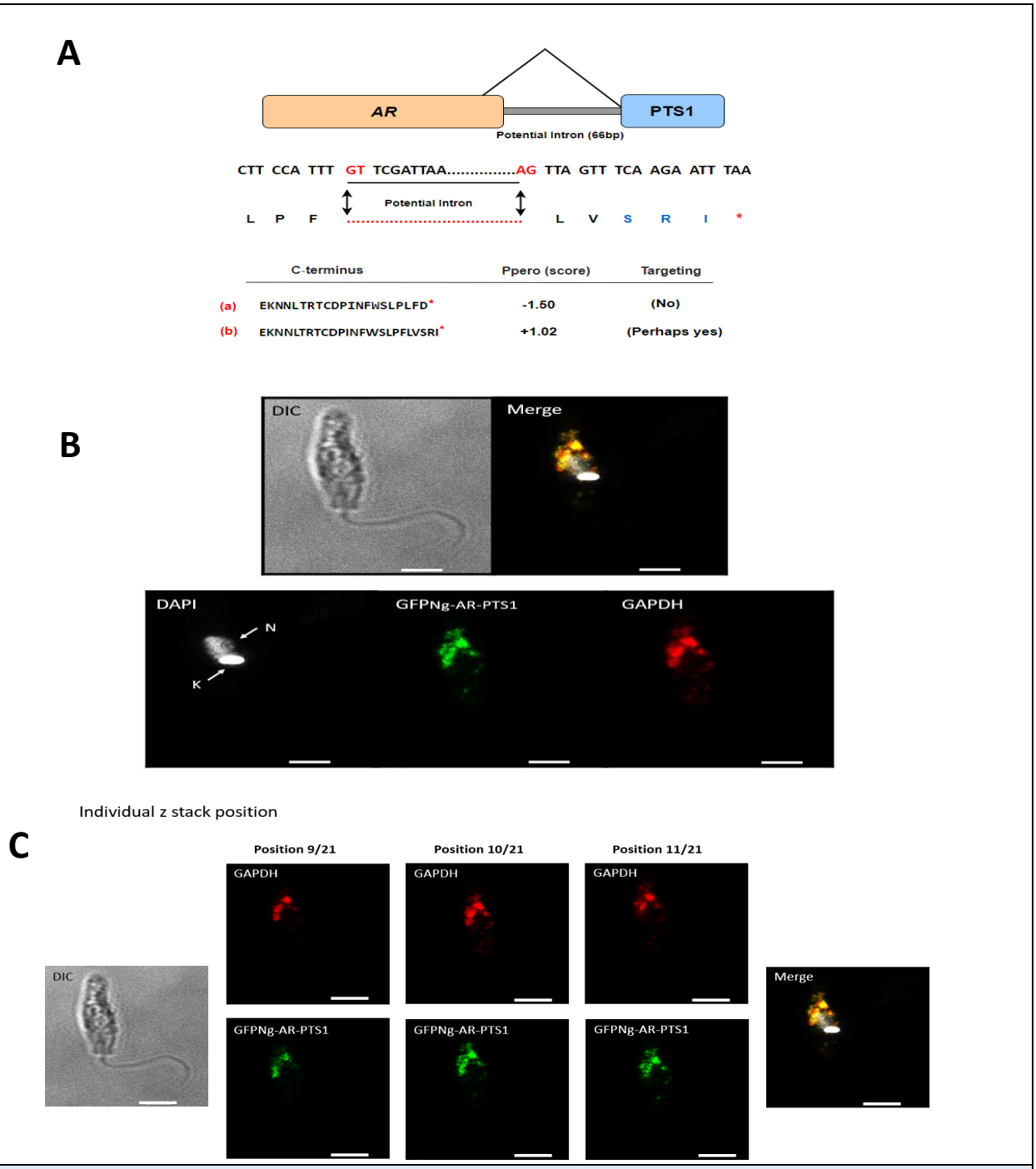


Figure 6.21. Gene expression in *C. fasciculata* of *N. gruberi* aldehyde reductase with a candidate cryptic PTS1 motif. (A) Cartoon schematic of *N. gruberi* aldehyde reductase gene showing a candidate cryptic PTS1 targeting motif available by alternative splicing. The protein sequence corresponding to the C-terminal coding sequence after alternative splicing is also shown. The tripeptide predicted PTS1 is shown in blue. Prediction of peroxisomal targeting before and after alternative splicing is provided. (B) Assessment of cryptic PTS1 candidature. The airyscan projection shown co-localisation of GFP fused to the C-terminal 25 amino acids of *N. gruberi* aldehyde reductase assuming alternative splicing of a potential intron (GFPNg-AR-PTS1) with indirect immunofluorescence signal for polyclonal antisera detecting glycosomal GAPDH in paraformaldehyde-fixed *C. fasciculata*. Nuclear (N) and mitochondria (K) DNA were detected by staining with DAPI. (C) Individual z stacks from the projection shown in B also indirect co-localisation of GFPNg-AR-PTS1 with glycosomal GAPDH. Scale bars, 2um.

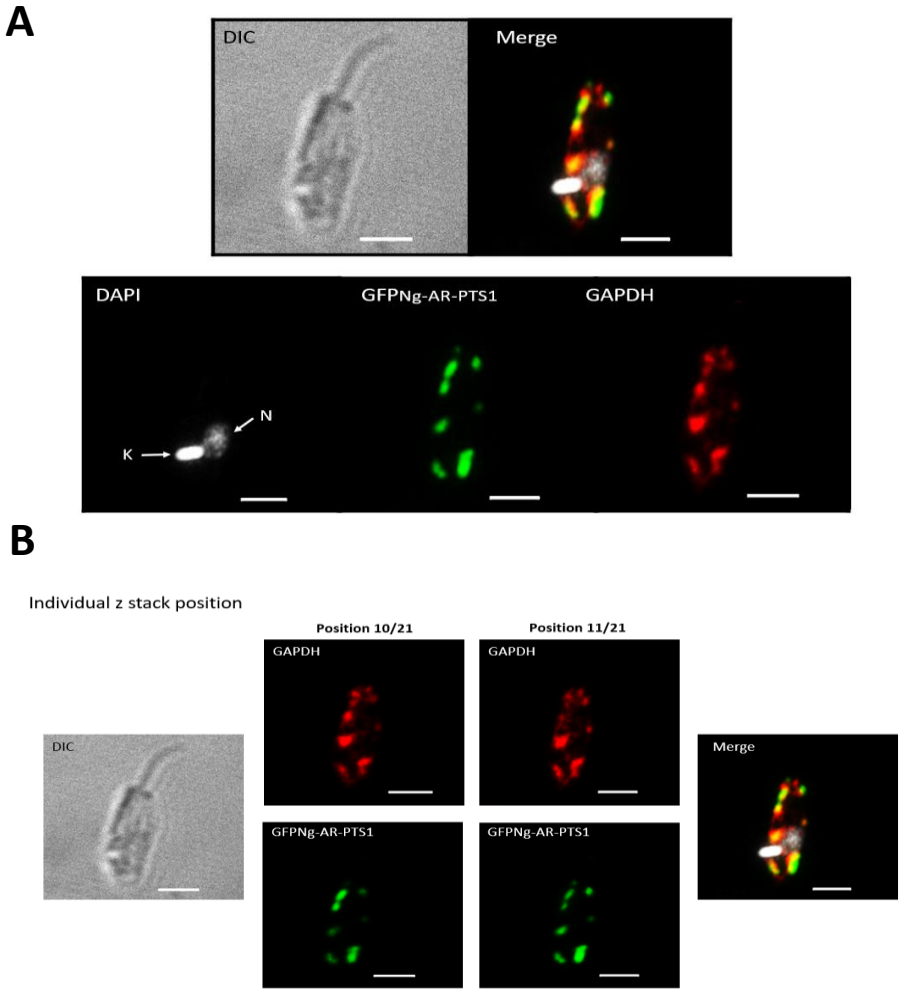


Figure 6.22. Another example of gene expression in *C. fasciculata* of *N. gruberi* aldehyde reductase with a candidate cryptic PTS1 motif.

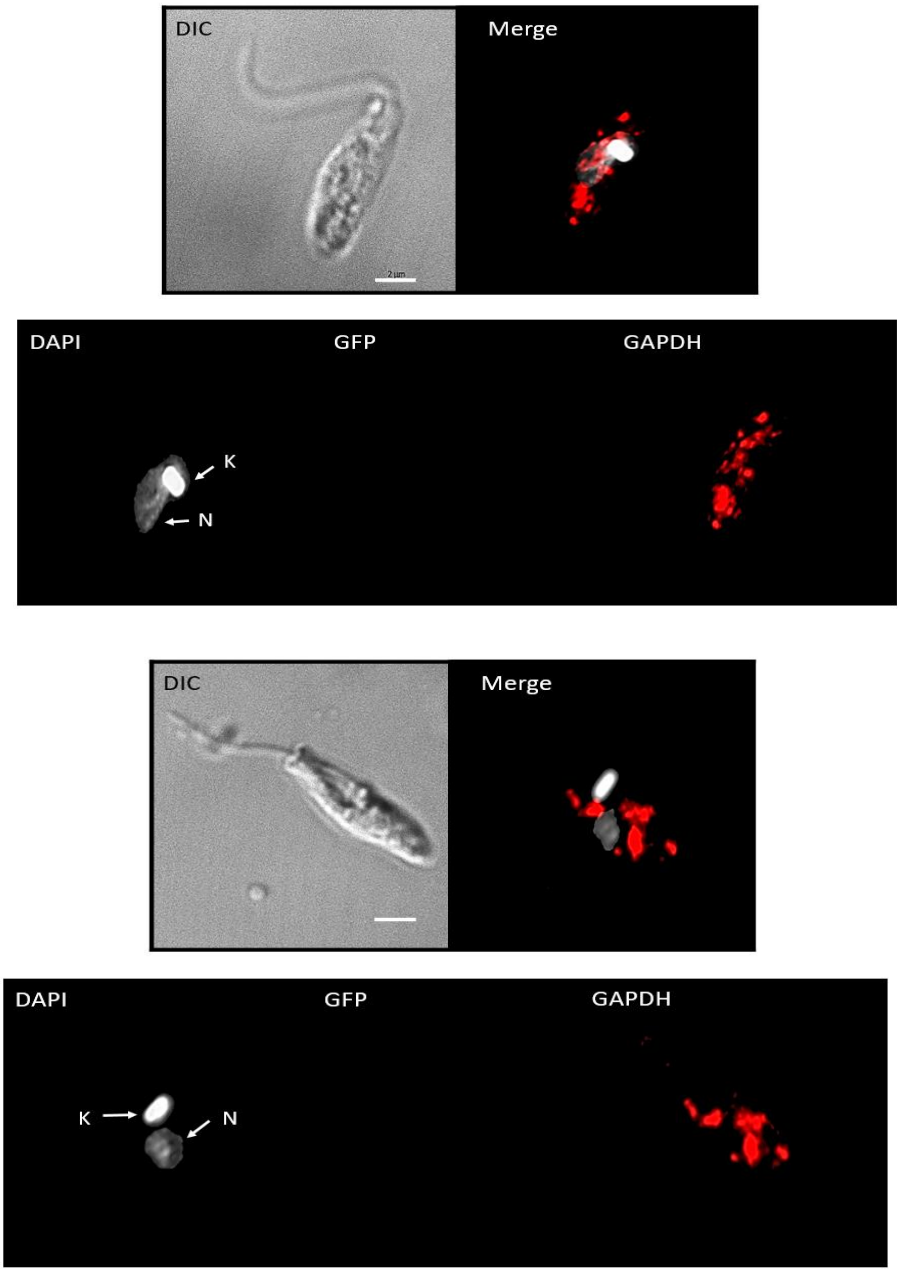


Figure 6.23. The lack of GFP signal in wild type *C. fasciculata* cells

Same parameters were used to detect GFP signal in both transfected cell lines and the wild type cells (control). The airyscan projection shown the indirect immunofluorescence signal for polyclonal antisera detecting glycosomal GAPDH and the absence of GFP signal in paraformaldehyde-fixed wild type *C. fasciculata*. Nuclear (N) and mitochondria (K) DNA were detected by staining with DAPI. Scale bars, 2µm.

Chapter 7

General Discussion and Conclusion

The metabolic capabilities of unicellular eukaryotes can vary in different habitats. For example, when glucose is abundant, some species solely rely on the glycolytic pathway for energy production (*e.g.* bloodstream form *T. brucei* and asexually replicating *P. falciparum* merozoites). Alternatively, some anaerobic, or microaerophilic protists rely on hydrogen production to produce energy (*e.g*. within the hydrogenosomes of *Trichomonas vaginalis*) (Oyelade et al., 2016, Schneider et al., 2011). Thus, metabolic pathways for ATP production deviate readily from, and can be far more elaborate than, standard pathways described in many standard textbooks (Ginger et al., 2010). In my PhD research, I combined bioinformatics analysis with experimental studies to examine several aspects of the metabolic versatility of different protist lineages. In particular, my studies led to (i) identification of a divergent System III used to mature mitochondrial cytochromes *c* in trypanosomatids and (ii) evidence that cryptic peroxisomal targeting of glycolytic and other metabolic enzymes is widespread in microbial eukaryotes: in particular, I characterised cryptic peroxisomal targeting in *Naegleria* amoebae.

7.1 Further consideration of the variation of System I in eukaryotes

Evidence from bioinformatics analysis has shown that System I (Ccm system) in eukaryotes is modified and seemingly simpler than in prokaryotes. Key differences between System I in bacteria and eukaryotes include the number of proteins required for the function of Ccm system and the conservation level of functional motifs of Ccm proteins. In bacteria, System I contains up to nine proteins (CcmA to CcmI) (Figure 7.1). However, I did not find orthologues of bacterial CcmD, G and I in any eukaryotic taxon, suggesting System I in eukaryotes contains a maximum of six Ccm proteins (CcmA, B, C, E, F and H). Moreover, a few eukaryotes do not appear to possess the full complement of six proteins. For example, CcmC was not found in ciliates, CcmH was not found in *N. gruberi*, and CcmB was either absent from or at least highly divergent in the genomes of several eukaryotic lineages. It is worth noting that CcmE is found in all eukaryotes which use System I. CcmE is a heme chaperone protein which forms an intermediate covalent link with heme. Biochemical evidence in *E. coli* has shown that CcmE can covalently attach to heme without the need for CcmA and B, however, CcmC is considered

as a pivotal component of System I and is essential for the heme transfer to CcmE (Shevket et al., 2018, Daltrop et al., 2002, Feissner et al., 2006). Hence, it is not clear how ciliates could compensate for the loss of CcmC, assuming that a divergent CcmC is not present. In species that use System I to mature *c*-type cytochromes, CcmC is encoded in the mitochondrial genome. There is a possibility that ciliates transferred CcmC gene from the mitochondrial genome to the nuclear genome and during this transfer, the sequence of CcmC could have been altered so that it was no longer possible to detect it using BLAST analysis.

A previous study on *Arabidopsis thaliana* had shown that heme transfer from holoCcmE to apocytochrome could take place only when both heme and the heme-binding motif in the apocytochrome were in the reduced state (Meyer et al., 2005). The CcmH is a thiol-disulfide oxidoreductase enzyme which is used to reduce the two Cys residues in the heme-binding motif in typical apocytochromes *c* (Figure 7.1). CcmH contains a conserved RCXXC motif where the Cys residues are known to be important for the reduction of apocytochromes (Meyer et al., 2005, Cramer and Toivo, 2016). A variant form of CcmH was identified in some ciliates species and in *C. owczarzaki* where the RCXXC motif was not conserved (sequence alignment in Chapter 3 Figure 3.4). Furthermore, no orthologues of CcmH was identified in the genome sequence of *N. gruberi*. A similar case was previously reported in some Archaea where CcmH was not detected. However, in these archaea a novel form of CcmE was identified. This variant form of CcmE does not have the highly conserved functional His residue in HXXXY motif and instead it is replaced by a Cys residue that occurs in a CXXXY motif (Allen et al., 2006). Studies had shown that this conserved His is critical for the intermediate covalent link between CcmE and heme (Mavridou et al., 2013, Enggist et al., 2003, Stevens et al., 2003). This variant form of CcmE can also be found in several proteobacteria species which also lack the conserved CcmH (Kletzin et al., 2015). Even though *N. gruberi* nuclear and mitochondrial genomes lack CcmH, its CcmE has the typically conserved His residue (sequence alignment in Chapter 3 Figure 3.3). The absence of CcmH in *N. gruberi* and the variant form of CcmH identified in some ciliates and *C. owczarzaki* could be explained by the presence of alternative redox-active protein in the mitochondrial IMS, which negates the requirement for CcmH. Nevertheless, there is still a possibility that one of the remaining conserved Ccm proteins in *N. gruberi* and the variant form of CcmH in ciliates

and *C. owczarzaki* both contain an atypical reductant-transfer motif which compensates for the reductive function of the CcmH or RCXXC motif.

The differences in Ccm proteins between bacteria and eukaryotes might be explained by the fact that System I in bacteria is used to attach heme to different types of *c*-type cytochromes, whereas in eukaryotes, System I is only used to mature mitochondrial cytochrome *c* and c_1, and these proteins share the same protein fold whereas bacterial *c*-type cytochromes are structurally very diverse (Allen et al., 2006, Kleingardner and Bren, 2011). Moreover, maturation of *c*-type cytochromes in bacteria takes place in the periplasm, while in eukaryotes, it takes place in the mitochondrial IMS. There are some environmental differences between bacterial periplasm and mitochondrial IMS (Manganas et al., 2017). Proteins imported into these compartments are in a reduced state, due to the presence of thioredoxin and glutaredoxin systems in the cytosol, which prevents the formation of disulfide bonds (Herrmann and Riemer, 2014). However, once in the bacterial periplasm, the presence of DsbA and DsbB causes the formation of a disulfide bond in the periplasm proteins (Kadokura and Beckwith, 2010). Similarly, the mitochondrial IMS has a mitochondrial import and assembly system (MIA) that is composed of two main proteins (Mia40 and Erv1). This system is used to import and introduce disulfide bonds to the mitochondrial IMS proteins (Backes and Herrmann, 2017, Banci et al., 2009). In some ways, at least, Mia40 and Erv1 are therefore equivalent to bacterial DsbA and DsbB. However, even though periplasm and mitochondrial IMS use similar systems to introduce the disulfide bonds, the reductive systems in these compartments are different. In the periplasm, there is a well-characterised reductive system, which is composed of two proteins (DsbD and DsbC). Along with CcmH and CcmG, this system is used to reduce the disulfide bonds in the heme and apocytochrome *c* (Verissimo et al., 2017). Compared to bacterial periplasm, there is no clear evidence for the presence of a reductive system in the mitochondrial IMS (Manganas et al., 2017). However, some studies on yeast proposed the presence of the cytosol reductive systems in the mitochondrial IMS (thioredoxin and glutaredoxin systems) (Kojer et al., 2015). Therefore, the differences in the reductive system indicate that the mechanism of reducing the disulfide bonds in the heme and apocytochrome *c* in these two compartments do vary. This might, in part, explain the variation of CcmH in some eukaryotes and the absence of CcmG in all eukaryotes that use System I for cytochrome *c* maturation.

There is also the possibility that eukaryotes do contain genes that code for missing Ccm proteins, but that these gene products are highly divergent relative to bacterial orthologues. For instance, divergent orthologous proteins might share a similar tertiary structure, but without the conservation of the primary structure. As a result, it will be difficult to identify them using bioinformatics analysis based on the sequence alignments.

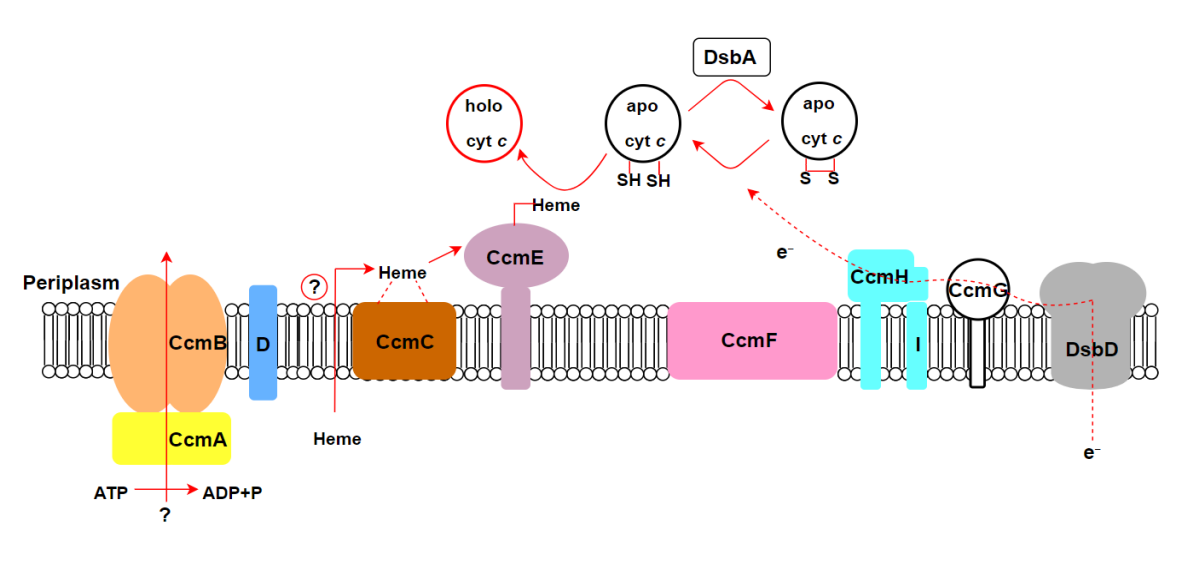


Figure 7.1. Overview of the cytochrome *c* maturation System I
In bacteria, CcmA, B, C, D and E are anticipated to be involved in heme delivery whereas CcmF, G and H are associated in the redox activity. The dashed arrows indicate the transfer of electrons or heme across the membrane. System I proteins are distributed across cytoplasmic membrane and periplasm in prokaryotes and mitochondrial inner membrane and IMS in eukaryotes

7.2 Divergent Trypanosomatid System III

Systems I and III are known to mature the mitochondrial *c*-type cytochromes with double Cys residues in the heme-binding motif. In euglenozoan protists, which include trypanosomatids, the heme-binding motif in the mitochondrial *c*-type cytochromes contains only a single Cys. How these unique mitochondrial cytochromes were matured was obscure (Allen et al., 2004, Ginger et al., 2012). In all trypanosomatids (except for dyskinetoplastic trypanosomes and bloodstream form of *T. brucei*), a mitochondrial ETC is used to produce ATP via oxidative phosphorylation. Therefore, it is essential to have a system to catalyse the heme attachment to *c*-type cytochromes. The notable exception to this rule are the plant pathogenic *Phytomonas*, which possess a truncated mitochondrial ETC missing complexes III and IV and *c*-type cytochromes. Here proton-pumping complex I establishes the ion electrochemical

gradient for ATP synthase activity; electrons transferred to ubiquinone by complex I are transferred to terminal electron acceptor O_2 by an alternative oxidase. The absence of *c*-type cytochromes, indeed the absence of all proteins known to associate with heme, suggested that the enigmatic cytochrome *c* machinery would also be absent from *Phytomonas*, thus prompting a proteomic sift between trypanosomatids and *Phytomonas* proteomes to identify proteins conserved in the former and absent from the latter. From this sift a candidate protein with limited motif similarity to and with overall structural topology similar to System III (holocytochrome *c* synthase) was identified. This hypothetic protein, termed 'putative holocytochrome *c* synthase' (or pHCCS), was well conserved across all kinetoplastids that are either free-living or parasitic, including in the minimalist intracellular parasite of amoebae *Perkinsela*, which despite striking reductive evolution of its metabolism nonetheless contains complexes III and IV and a cytochrome *c* (Tanifuji et al., 2017).

An earlier study had shown that *Saccharomyces cerevisiae* recombinant holocytochrome *c* synthase is able to mature *T. brucei* cytochrome *c* with very low efficiency (Fülöp et al., 2009). Furthermore, the crystal structure of *C. fasciculata* cytochrome *c* had near identical overall fold and heme attachment stereochemistry as other known *c*-type cytochromes, which have a typical heme-binding motif (Fülöp et al., 2009). These observations provided an additional level of assurance that trypanosomatids conceivably used a variant form of System III to mature their mitochondrial *c*-type cytochromes.

Following the bioinformatics screen, molecular laboratory work was completed to obtain evidence that this divergent System III is used to mature the *c*-type cytochromes in trypanosomatids. After an initial lack of success with a variety of alternative recombinant expression systems, *T. brucei* putative holocytochrome *c* synthase (*Tb*pHCCS) and *T. brucei* apocytochrome *c* substrate (*Tb*CYTC) ORFs were successfully cloned and expressed using pCDF-Duet plasmid and Rosetta competent *E. coli*. Soluble fractions of *Tb*CYTC, which expressed on its own or co-expressed with *Tb*pHCCS, were obtained. If *Tb*pHCCS is the system used by trypanosomatids to attach heme to the *c*-type cytochromes, then *Tb*CYTC fractions from the bacteria culture expressing both *Tb*CYTC and *Tb*pHCCS should be in the holo-cytochrome form. In contrast, the *Tb*CYTC fractions from the bacteria culture expressing only *Tb*CYTC gene should be in the apo-cytochrome form. After heme staining of SDS-PAGE gels, the heme, denoting covalent attachment of the heme to the apocytochrome, was detected

only in samples where both *Tb*CYTC and *Tb*pHCCS were expressed – *i.e.* not from *E. coli* were only *Tb*CYTC expressed. Additionally, using the UV spectroscopy, the heme absorption spectra clearly showed that the heme was attached to the *T. brucei* recombinant cytochrome *c* via a single Cys in the AXXCH heme-binding motif, whereas, the heme absorption spectra of the control sample (horse holocyotochorme *c*) showed that the heme was attached via two Cys residues in a CXXCH heme-binding motif. This work provided verification that trypanosomatids use a divergent System III to catalyse heme attachment to their mitochondrial cytochromes *c*.

There is no clear explanation to why one group of protists has evolved a variant form of mitochondrial *c*-type cytochromes matured by a divergent System III for the heme attachment. The identification of this system opens up many questions. For example, did trypanosomatids evolve by accident or selection the single Cys in the heme-binding motif of the mitochondrial *c*-type cytochromes first so that System III had to accommodate this variation during its evolution? If this is the case, then why did the *c*-type cytochromes maintain a single Cys in the heme-binding motif instead of two? Is it energetically favoured or is the biochemistry of the kinetoplast (mitochondria) is the driver behind the evolution of these variant *c*-type cytochromes? What if the origin of this variant form of System III can be traced back to the eukaryotic last common ancestor? and since trypanosomatids are an early divergent group of eukaryotes and they have a relatively different kinetoplast IMS compared to other unicellular protists, did they then retain divergent System III whereas other eukaryotes evolved the known form of System III? If the last hypothesis is true, then this divergent System III is expected to be found in other early divergent unicellular eukaryotes. Given that Amoeba 'BB2' and *Pharyngomonas kirbyi* encode for System III, however, this system was found to be relatively different from the divergent System III despite belonging to an early divergent heteroloboseans which is a sister group to the Euglenozoa (Chapter 3 Figure 3.14). Moreover, these species contain a typical mitochondrial cytochrome *c* with two Cys residues instead of one in the heme-binding motif. Even though at present, only species belonging to the Euglenozoa group are known to have these unique features, however, the continuous increase in the transcriptome data may reveal otherwise.

Another intriguing question is whether *Bacillus* also uses a divergent looking System III to mature their unusual cytochrome *b* (which is similar to b_6 but is part of a *bc* complex) which

has heme covalently attached to a single Cys (Yu and Le Brun, 1998, Allen et al., 2008). To date, there is no known mechanism used by *Bacillus* to mature this unique form of cytochrome *b*. From the initial bioinformatics analysis, homologues of trypanosomatids' divergent System III was not found in *Bacillus*. This raises the question whether *Bacillus* uses a distinct form of System III that is different from trypanosomatids' divergent System III; a variant form of System I; or a novel maturation system.

7.3 Possible origins of cytochrome *c* maturation systems

System I is used for cytochrome *c* maturation in archaea and bacteria (Allen et al., 2006). Variants of System I are found in eukaryotes. Coding of this system is partially mitochondrial in all protists where System I is present, indicating it was an ancestral feature of the *proto*-mitochondrion (Lee et al., 2007, Allen et al., 2008). System III (Holocytochrome *c* synthetase) is the alternative route for mitochondrial cytochrome maturation in eukaryotes and it has only been ever identified as a nuclear-encoded gene. The distribution of System III versus System I between different eukaryotic lineages is somewhat mosaic-like and remains poorly understood. Moreover, System III is one of the few eukaryotic proteins for which prokaryotic homologues are not (yet) evident (Allen, 2011). There are two hypotheses which explain the mosaic distribution of System III in eukaryotes:

1) System III was evolved from one ancestral eukaryotic lineage then spread into numerous other lineages by Lateral Gene Transfer (LGT). As System I is far more complex than System III in terms of the component proteins required, in species which acquired System III by LGT, mitochondrial and nuclear-encoded Ccm protein-coding genes would have been free to degenerate.
2) Both Systems I and III were present in the last eukaryotic common ancestor (LECA) and there was then differential gene loss of either System I or III (Cavalier-Smith, 2009, Allen, 2011).

The evolutionary origin of System III by a number of LGT events between distinct eukaryotic lineages is mentioned in some existing literature (Allen et al. 2008, Nishimura et al., 2016). LGT as an origin for System III is supported by the following observations: (i) System III composed of a single protein (HCCS), therefore, only one protein is needed to be transferred from one species to another for the system to be functional; (ii) HCCS only needs to be a

transcript at relatively small quantity to be functional, therefore, only a basic, inefficient promoter is required to drive gene expression of newly laterally acquired System III; (iii) there is a conserved mitochondrial import system across eukaryotes, therefore, there is likely less need for signal modification to target HCCS to the mitochondrial inter-membrane space. Thus, the presence of System III in early-divergent lineages such as Amoebozoa, many algae and other protists is readily explained by a secondary and tertiary endosymbiotic gene transfer (EGT) "you are what you eat" (Doolittle, 1998, Allen et al., 2008).

The increase in transcriptome data is useful in carrying out deeper bioinformatics analysis to better understand the origin of System III. The literature currently places almost all eukaryotic taxa into three main groups: (i) Amorphea, which include metazoans, some unicellular protists (*e.g.* choanoflagellates and filastereans such as *Capsaspora*) and the Amoebozoa; (ii) Diaphoretickes which include Viridiplantae and SAR protists (Stramenopiles, Alveolates, Rhizaria); (iii) Excavata which include euglenids, jakobids, heteroloboseids, metamonadsa. *Malawimonas* and ancyromonads, in contrast, are orphan taxa for which an evolutionary position in the tree of eukaryotes is not known (Adl et al., 2012, Cavalier-Smith, 2013). According to this taxonomy and from the bioinformatics data represented in this study, there is a mixed occurrence of System I and III in almost every eukaryotic group. For example, within Viridiplantae clade, the chlorophytes, for which there are transcriptomes or nuclear genomes sequenced, only System III was detected. On the other hand, flowering plants use System I for cytochrome *c* maturation. To date, there is no evidence for a flowering plant that uses System III instead of System I. In contrast, species belonging to charophytes and primitive plants were found to be either using System I or III to mature *c*-type cytochromes (Figure 7.2). A similar pattern of the mosaic distribution of System I and III were detected in red algae. The early-divergent red algae (Cyanidiales) uses System I, however, the Cyanidiophyceae shares a common ancestor with Porphyridiophyceae, Floridophyceae and Bangiophyceae which use System III (Figure 7.2).

This assorted occurrence of Systems I and III were also detected in excavate protists, which collectively were previously thought to possibly only encode for System I and, in the case of euglenozoans, a novel maturation system. The Excavata is an early divergent group of eukaryotes with an estimate divergent time of 1510–1699 Ma (Parfrey et al., 2011). A recent study had shown that System III is present in Amoeba 'BB2' and *P. kirbyi*, both of which belong

to Heterolobosea. *Naegleria* species, which belong to the same group, use System I for cytochrome *c* maturation. The presence of both systems across a wide variety of eukaryotic lineage and the identification of System III in several early divergent species suggest that for LGT hypothesis to be true it will require a large number of inter-lineage transfers. Therefore, the more favourable explanation of System III origin might be the LECA hypothesis and the independent loss of either System I or III. However, the LGT hypothesis for System III origin cannot be ruled out, if the LECA hypothesis is true, it does not easily explain the presence of complex System I instead of III in few unicellular species (*e.g*. *Capsaspora owczarzaki* and *Fonticula alba*) that in evolutionarily terms are relatively close to fungi and animals – *i.e.* System I has persisted for longer in the Amorphea than might have been anticipated until recently. However, if both systems can be traced back to the LECA, then the presence of species with both systems should perhaps be more common within the eukaryotic phylogeny. To date, only one eukaryotic species (*Ancoracysta twista*) was found to encode for both systems (Janouškovec et al., 2017). The findings from my analysis suggest that this complex distribution pattern of System III might have evolved via a combination of LGT and LECA events. If this was the case, then System III was likely to have evolved in ancient eukaryotic lineages. Then a combination of LGTs and differential gene loss would have taken place, which explains the mosaic distribution of System III in eukaryotes.

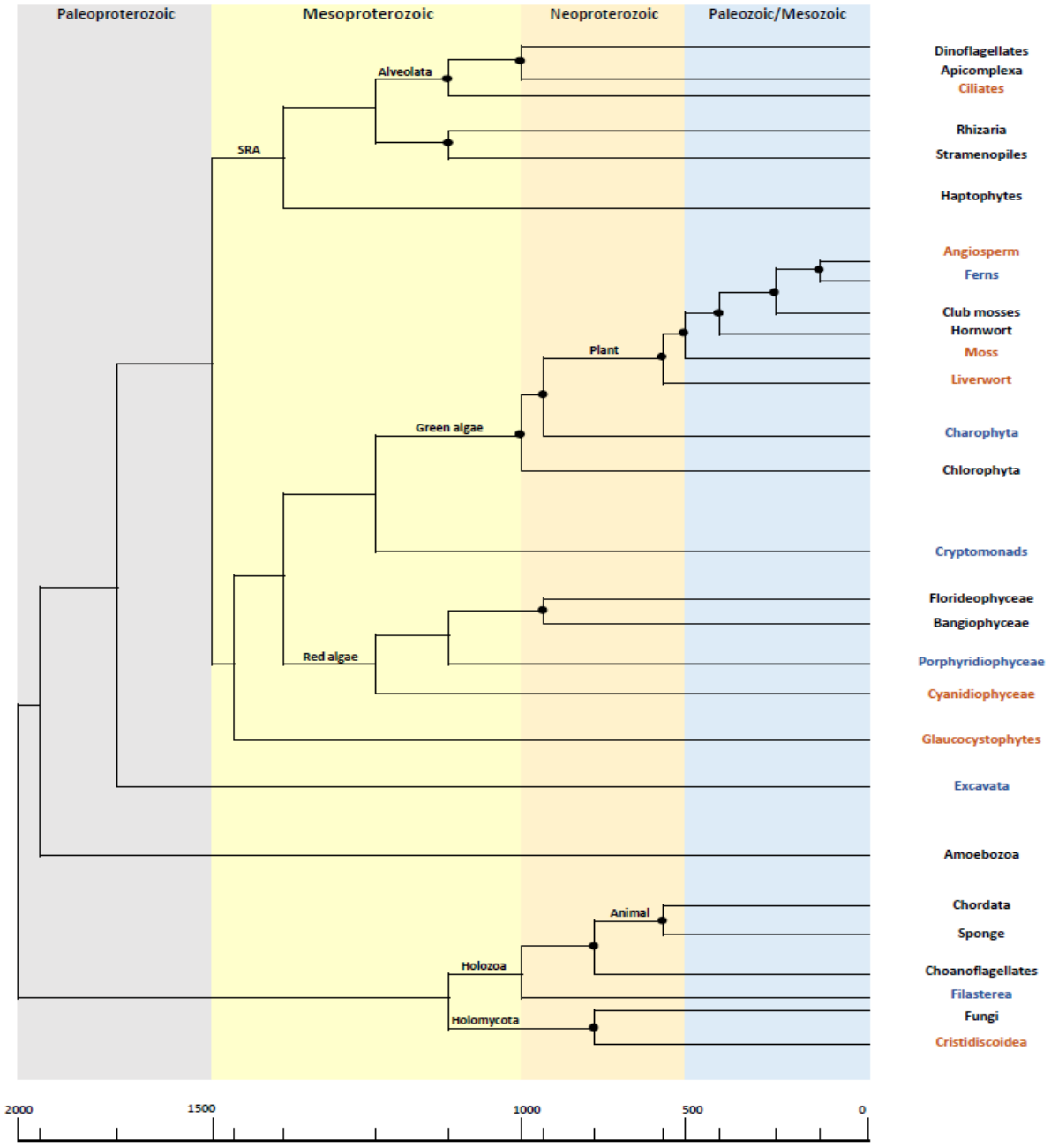


Figure 7.2. The phylogenetic distribution of System I and III and the estimate molecular clock for the major eukaryotic lineages

The time estimate for the divergence of major phyla was based on the following studies; Yang et al., 2016, Parfrey et al., 2011, Cunningham et al., 2017, Clarke et al., 2011, Berbee and Taylor, 2010 and Eme et al., 2014. In these studies, the time-calibrated phylogenetic analyses were calibrated using Phanerozoic (including: Paleozoic/Mesozoic eras) and Proterozoic fossils (including: Paleoproterozoic, Mesoproterozoic, Neoproterozoic eras) (black circles). The colour code represents which system the eukaryotic lineages use to mature their cytochrome *c*. The black code is for System III, the blue code is for mixed lineages where some use System I and some use System III and the brown code is for the System I. Appendix 12 shows what species within each group use System I or III.

7.4 Metabolic Versatility of *Naegleria gruberi*

Naegleria gruberi is a free-living amoeba that belongs to the phylum Heterolobosea. This phylum is considered as a sister to Euglenozoa within the supergroup Excavata. The published, annotated genome sequence of *N. gruberi* revealed the capability for aerobic and anaerobic metabolism with the capacity for hydrogen production (Fritz-Laylin et al., 2010). Species that can survive in anaerobic environments usually possess mitochondrial related organelles (MRO) such as mitosomes and hydrogenosomes (*e.g. Trichomonas* and *Giardia*) (Stairs et al., 2015). According to the genome sequence, *N. gruberi* is a rare example of a free-living, heterotrophic aerobic eukaryotes that presents classic-looking mitochondria, but shows an apparent ability to survive in anaerobic environment via energy metabolism more typically seen in obligate anaerobic (or microaerophilic) eukaryotes. Thus, the genome sequence of *N. gruberi* shows the presence of hydrogenosomes enzymes (*e.g.* [FeFe]-hydrogenase and hydrogenase maturases). These enzymes were experimentally verified to be used for hydrogen production as a mechanism for energy generation under anaerobic conditions (Tsaousis et al., 2014, Fritz-Laylin et al., 2010).

In addition to classic hydrogenosomal enzymes, the *N. gruberi* genome sequence also suggests that this organism encodes for two additional anaerobic enzymes (ACD and CuNirK). To date, there is no experimental data that shows *N. gruberi* utilises these enzymes to generate energy under a low level of oxygen. My study used bioinformatics analysis to examine whether these two enzymes have the characteristic features of ACD and CuNirK which include the overall structure of the protein and the presence of functional motifs. From this analysis, the *N. gruberi* putative ACD did have the majority of the conserved amino acid sequences found in other known ACD proteins. However, *N. gruberi* putative ACD lacked the conserved functional His residues and the acyl-CoA-binding domain. Many studies showed that mutation of the catalytic His residue causes complete loss of function. Hence, the His residue is considered as a hallmark for ACDs (Weiße et al., 2016, Jones and Ingram-Smith, 2014, Jones et al., 2017). Furthermore, it was previously suggested that conservation of this particular catalytic His residue can be used to distinguish between ACD and bacterial acetyltransferases (*i.e.* PAT proteins). PATs have a similar overall structure to ACDs but lack the catalytic His residue and have an GNAT domain (Jones et al., 2017, Dyda et al., 2000). These defining features of PAT were recognised in the putative *N. gruberi* and *N. fowleri* ACDs.

Amino acid sequence alignments between putative *Naegleria* ACDs and PATs revealed the *Naegleria* proteins possessed a majority of the functional residues which were identified as important for PAT activity (sequence alignment in Chapter 5 Figure 5.10). PAT can reversibly regulate the function of a catalytic enzyme by adding an acetyl group to specific lysine residues critical to enzyme function. This regulatory mechanism is known as protein acetylation (Starai and Escalante-Semerena, 2004, Cho et al., 2008).

There are two main types of protein acetylation. The first type involves the addition of the acetyl group at the α-amino group of the amino acid at the N-terminal tail of the target protein whereas the second type involves the addition of the acetyl group at an ε -amino group of internal lysine residues. PAT activities are associated with the second mode of reversible acetylation (Polevoda and Sherman, 2002, Soppa, 2010). Only recently, protein acetylation was identified in bacteria as a mechanism to control enzymatic activity. To date, only few enzymes were identified as targets for acetylation. In bacteria, PAT is known to modulate the activity of both chemotaxis protein and the acetyl-CoA synthetase (AMP-forming) (ACS) (Wang et al., 2010, Thao and Escalante-Semerena, 2012, Christensen et al., 2019). ACS catalyses the formation of acetyl-CoA from acetate through the consumption of ATP (Starai and Escalante-Semerena, 2004). In contrast to bacteria, protein acetylation is widely used in eukaryotic species to control the activity of various types of proteins. One of the major targets for protein acetylation in eukaryotes are histones. Histones are a group of alkaline proteins that bind to DNA and maintain it in a structure known as nucleosomes. Acetylation of these proteins can cause a decrease in the binding affinity to DNA and consequently change the rate of gene expression (Lee and Workman, 2007, Hildmann et al., 2007).

The N-terminal amino acids of *N. gruberi* 'ACD' are sufficient to target eGFP to trypanosome mitochondria (Lynch, 2016). Therefore, if this putative protein is used for protein acetylation in *N. gurberi*, it is likely to be used to regulate the activity of mitochondrial proteins. Based on the proteomic analysis, protein acetylation can be detected in a wide variety of mitochondrial proteins including some of the metabolic enzymes (Kim et al., 2006). Although acetylation appears common in mitochondrial proteins, the effect(s) of lysine acetylation on mitochondrial protein activities is(are) poorly understood. Furthermore, there is no known mechanism of how mitochondrial proteins are acetylated. There is some

evidence that acetylation of histones can occur in vitro without the need for enzymatic activity. Therefore, there is the possibility that acetylation of mitochondrial proteins may take place spontaneously. Nevertheless, there might be a class of mitochondrial acetyltransferases that is yet to be discovered (Scott et al., 2012). While the process for mitochondrial protein acetylation is not known, the deacetylation of mitochondrial proteins on the other hand is identified and well-studied. Mitochondrial protein deactylation is carried out by a group of proteins belonging to the sirtuins family (Nogueiras et al., 2012). For example, sirtuin 3 was identified to increase the activity of ACS via deacetylation (Schwer et al., 2006, Hallows et al., 2006).

The above discussion can be summarised in three key points. First, the enzymes responsible for the acetylation of mitochondrial proteins are unknown. Second, the activity of ACS can be increased in mammals by deacetylation (Schwer et al., 2006). Finally, the acetylation of the bacterial orthologs of ACS is carried out by PAT (Figure 7.3). All of these points support my hypothesis that *Naegleria* 'ACD's might be ortholgous to bacterial PAT, rather than being orthologous to the acetyl-CoA synethase (ADP-forming) that is a central feature of energy generation in a variety of anaerobic parasites (Jones and Ingram-Smith, 2017, Guo et al., 2015).

CuNirk is another anaerobic enzyme which was also identified by the nuclear genome annotation of *N. gruberi*. The gene sequence of CuNirk was previously identified in a wide range of fungi with denitrifying activity detected in fungus *Fusarium oxysporum* (Shoun et al., 2012). The CuNirk gene was also identified in few protists including green algae (*e.g. Chlamydomonas reinhardti*) and some species belonging to the phylum Amoebozoa (*e.g*. *Acanthamoeba castellani* and *Vermamoeba vermiformis*). The CuNirk gene in these protists forms a monophyletic group with fungal CuNirk. To date, there are no further studies on the structure and function of CuNirk gene in these protists (Kim et al., 2009). Although CuNirk gene was identified in few protists, the increase in transcriptome data of unicellular eukaryotes may show a wider distribution. For example, a recent study identified denitrification enzymes, including CuNirk, in two newly isolated *Globobulimina* species (unicellular protists belong to Rhizaria group) (Woehle et al., 2018). CuNirk is considered a key part of the global nitrogen cycle and it is used to reduce nitrite to nitric oxide. The denitrification process can be found in the mitochondria of some fungi where it is coupled to

the ETC to generate energy under anaerobic conditions (Kim et al., 2009) (Figure 7.4). The structure and the primary sequence of fungal CuNirk resembles its bacterial counterpart, which is grouped under the same branch in the phylogenetic tree (Kim et al., 2009). Bioinformatics analysis was conducted to detect the presence of functional motifs/domains in *N. gruberi* putative CuNirK, which are critical to its function. Moreover, a phylogenetic tree was constructed to examine the position of *N. gruberi* putative CuNirk within the CuNirK phylogeny. The outcome of this analysis shows that *N. gruberi* putative CuNirK does have all the necessary functional motifs and domains which are widely conserved in CuNirK found in bacteria, archaea and eukaryotes. Furthermore, this putative protein had formed a monophyletic group with other eukaryotic CuNirK proteins. This provides enough evidence to carry out experimental work to prove conclusively that *N. gruberi* putative CuNirK can generate nitric oxide under anaerobic conditions.

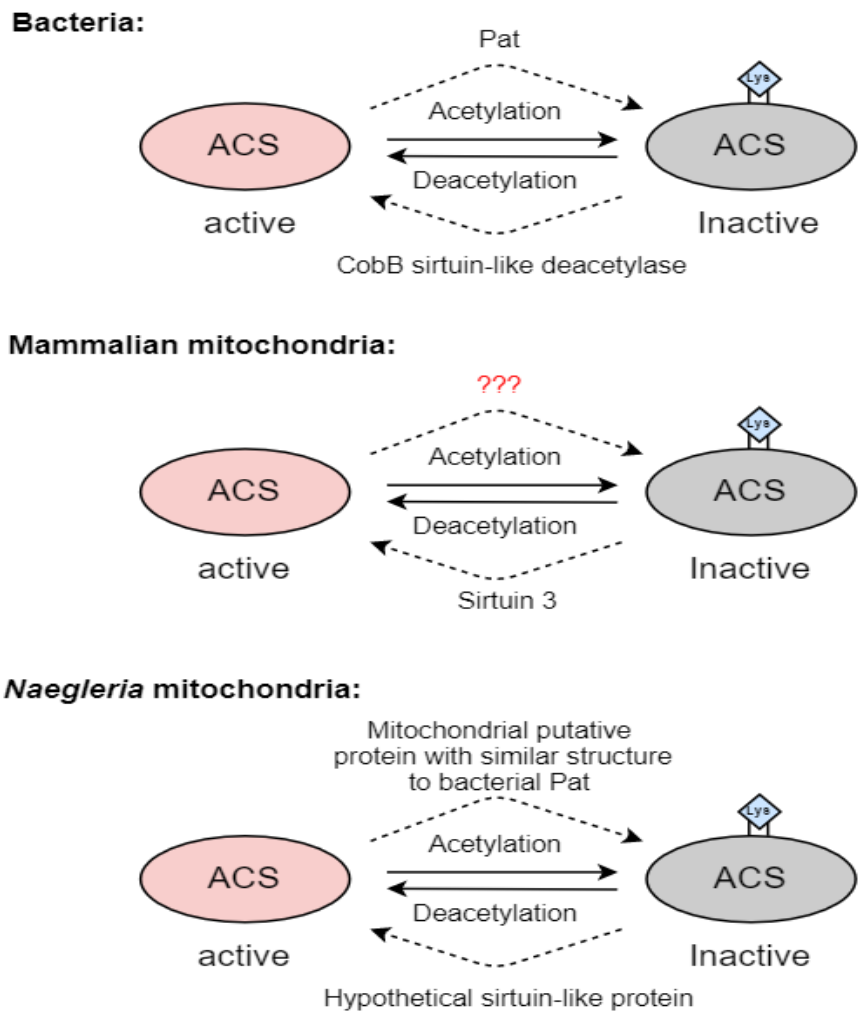


Figure 7.3. Proposed posttranslational modification of ACS via acetylation/deacetylation in bacteria, mammals and potentially in *Naegleria*.

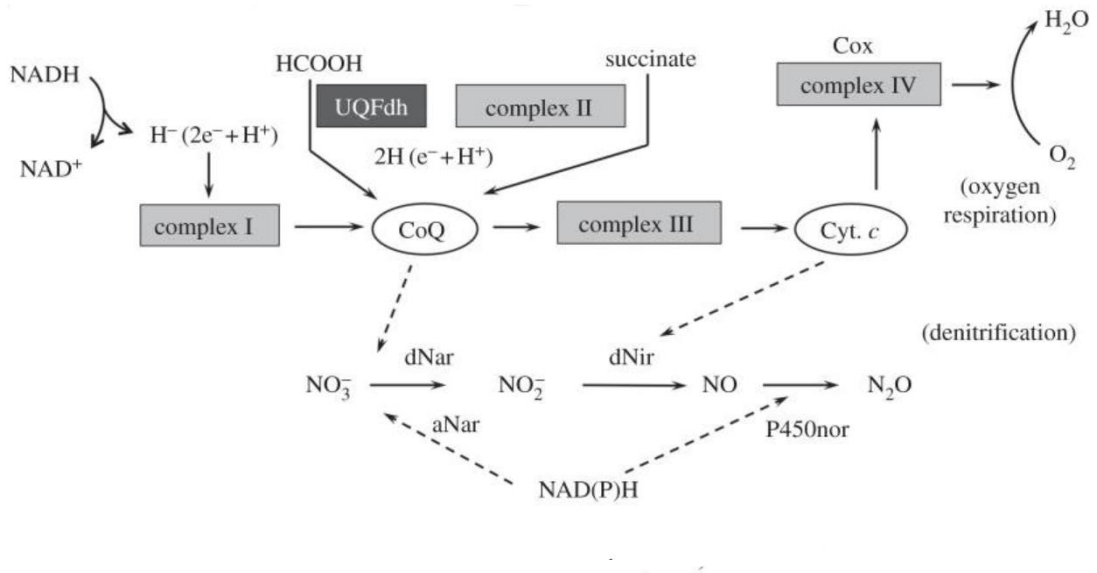


Figure 7.4. Denitrification process coupled with mitochondrial ETC in *Fusarium oxysporum* under anaerobic conditions

The final product of the denitrifying system in fungi is nitrous oxide (N_2O) instead of dinitrogen gas (N_2). This is due to the absence of fungal nitrous oxide reductase (Nos). Abbreviations: dNar, nitrate reductase; dNir, nitrite reductase; P450nor; nitric oxide reductase. The diagram was reproduced with permission from Shoun et al (2012).

7.5 Cryptic peroxisomal targeting of metabolic enzymes appears rather common among microbial eukaryotes

In eukaryotes, the presence of enclosed organelles aids the separation of different biochemical pathways in the context of a single cell. Several metabolic reactions are performed in a specific organelle, however, some reactions can be detected in more than one organelle (Hwang et al., 2004). Proteins are translated in the cytoplasm and then often directed to different organelles via specific signals found in the protein primary sequence. In case a specific protein is targeted to more than one organelle, then protein variants (isoforms) can be produced, each with a different signal (Gunning et al., 1998). These isoforms can be produced via various ways including gene duplication, ribosomal read-through of the stop codons and alternative splicing (Ast et al., 2014).

Studies had shown that metabolic enzymes, found in some fungi and animals and generally function in the cytoplasm and mitochondria, can be targeted to the peroxisomes. In some fungi, the glycolytic enzymes including glyceraldehyde-3-phosphate dehydrogenase

(GAPDH) and 3-phosphoglycerate kinase (PGK) can be detected in both the cytoplasm and the peroxisomes (Freitag et al., 2014). However, in animals, peroxisomal isoforms of NADP-dependent isocitrate dehydrogenase and pyrophosphatase were detected in *Drosophila melanogaster* and *Caenorhabditis elegans* respectively (Stiebler et al., 2014). Even though studies had shown that some metabolic enzymes can be targeted to the peroxisomes in fungi and animals, these studies only cover two major eukaryotic groups. To date, there is no indication of the distribution of this cryptic peroxisomal targeting of metabolic enzymes across unicellular protists. In this project, the usage of several bioinformatics techniques showed that some metabolic enzymes including core glycolytic enzymes might be targeted to the peroxisomes in several species across different eukaryotic lineages. In addition to this, the results from heterologous expression in *C. fasciculata* showed that the cryptic PTS1 motifs in four *N. gruberi* metabolic enzymes, which are generated either via ribosomal read through (for TPI, PGK,SOD) or alternative splicing (for AR), are capable to direct GFP to *C. fasciculata* glycosomes. The bioinformatics and cell biology analyses have shown that there is considerable scope for peroxisomal targeting of metabolic enzymes via cryptic PTS1 motifs in diverse range of eukaryotes, including some of the early divergent unicellular protists (*e.g*. species belonging to Amoebozoans and Heterolobosea). This distribution indicates that the function of peroxisome is likely more versatile than previously assumed.

Notably, not all species used in this study have the potential translation stop codon readthrough or alternative splicing to produce isoforms of the glycolytic enzymes with putative peroxisomal PTS1 import signals. Out of twelve taxa distributed widely across the eukaryotic phylogeny, only three were devoid of the presence of a PTS1 in any of the glycolytic enzymes. Two of these three species (*Vitrella brassicaformis* and *Tetrahymena thermophila*) belonged to the Alveolata. Within the 'superphylum' Alveolata are the Apicomplexa; some apicomplexans lack peroxisomes (*e.g. Plasmodium* and *Cryptosporidium* species); others have peroxisomal activities at certain stages of their life cycle (*e.g*. *Toxoplasma gondii*) (Gabaldón et al., 2016, Ludewig-Klingner et al., 2017).

One recent study suggested the process of ribosomal read-through of the stop codon and alternative splicing is generally non-native and it is completely due to a transcription error (Li and Zhang, 2019). If this was true, then an explanation is required for the various studies which showed that these molecular mechanisms are found inside the cells and are able to

produce a large proteomic diversity many of which are active and have a specific role within the cell (Namy et al., 2002, Freitag et al., 2012, Ast et al., 2013, Wang and Brendel, 2006, Blekhman et al., 2010, Chen et al., 2012). In this research, for example, several metabolic genes were identified to have a peroxisomal signal via alternative splicing and ribosomal read through. The latter was found to be frequently used in a several unicellular protists to generate the peroxisomal signal. It is therefore unlikely that the presence of peroxisomal signal sequence immediately after the primary stop codon followed by an additional stop codon in a specific gene is a case of random chance. There is the argument that eukaryotes tend to have additional stop codon after the primary stop codon as a mechanism of error proofreading strategy. This strategy provides another chance for translational termination in case the primary stop codon is bypassed during translation (Ho and Hurst, 2019). However, this does not explain easily the presence of putative peroxisomal signal sequence between the primary and the additional stop codons in multiple metabolic genes across different eukaryotic lineages. Therefore, a type of statistical analysis is needed to determine whether the presence of the peroxisomal signal generated by alternative splicing or ribosomal read through is not due to transcription error. Furthermore, most of the studies which concluded that ribosomal read-through and alternative splicing in the cell are not an adaptive processes, were primarily focused on multicellular species with one recent study focusing on yeast (Xu et al., 2019, Li and Zhang, 2019, Xu and Zhang, 2014). Thus, a similar analysis needs to be done on a wide variety of unicellular protists to see if a comparable statistically significant outcome can be obtained.

7.6 Conclusion

One of the primary aims of this project was to resolve the mystery of how trypanosomatids mature their *c* type cytochromes. Cytochrome *c* is one of the essential compounds of the ETC. The activity of mitochondrial *c*-type cytochromes is determined by the covalent attachment of heme at highly conserved heme-binding motif. This attachment is catalysed by the activity of either System I or III. The genomic data of several trypanosomatid taxa showed that any components of these systems were absent. An intensive study was carried out to investigate whether any of the critical motifs for both Systems I and III can be found in the divergent looking system in trypanosomatids. One of the major findings of this analysis was the identification of divergent System III which was well conserved across trypanosomatids.

Furthermore, during the same study of Systems I and III, unexpected findings were also observed. One of these findings was that System I components can differ between different eukaryotic groups and generally eukaryotic System I is simpler relative to System I in bacteria. The second finding was that the distribution of both Systems I and III in eukaryotic phylogeny is more complex than previously anticipated. These findings significantly impact the current understanding of the evolution of *c*-type cytochrome maturation systems.

The second part of the project focused on the different aspects of metabolic flexibility in different eukaryotic lineages. The ability to produce energy in anaerobic conditions is limited to few eukaryotic species, one of which is *N. gruberi*. The genome data of *N. gruberi* suggests that it is able to produce both ACD and CuNirK anaerobic enzymes. By conducting a detailed bioinformatics analysis focusing on the overall structure and the conservation levels of the functional motifs in these *N. gruberi* proteins, it was found that *N. gruberi* putative ACD did not have some of the essential features of the ACD structure. Additionally, this putative protein had a considerable level of conservation of functional motifs of another group of proteins belonging to the acetyltransferase family. Further bioinformatics analysis had shown that *N. gruberi* putative CuNirK did have the necessary functional motifs which are conserved in CuNirK across eukaryotes, bacteria and archaea. This indicates that this CuNirK could be used by *N. gruberi* for energy production in certain environments.

The final piece of this project was to survey the distribution of peroxisomal targeting of some of the metabolic enzymes across different eukaryotic groups. Peroxisomes are found in most eukaryotic organisms where they are generally used for fatty acids metabolism. However, recent studies had shown that some of the main glycolytic enzymes can be directed to the peroxisomes in some fungi and animals. In this study, the distribution of this phenomenon in other unicellular protozoa was carried out. Using different techniques of bioinformatics screening, the peroxisomal signal was detected in some of the glycolytic enzymes across various eukaryotic lineages. These results indicate that the function of peroxisomes can be more versatile than previously expected and that perhaps there is still much to learn about the partitioning compartmentalisation of central metabolism in eukaryotes.

References

Acestor, N., Panigrahi, A.K., Ogata, Y., Anupama, A. & Stuart, K.D., 2009. Protein composition of Trypanosoma brucei mitochondrial membranes. *Proteomics*, 9(24), pp.5497-5508.

Adhikari, U.K. & Rahman, M.M., 2017. Comparative analysis of amino acid composition in the active site of nirk gene encoding copper-containing nitrite reductase (CuNiR) in bacterial spp. *Computational biology and chemistry*, 67, pp.102-113.

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., 2012. The revised classification of eukaryotes. *Journal of Eukaryotic Microbiology*, 59(5), pp.429-514.

Ajioka, R.S., Phillips, J.D. and Kushner, J.P., 2006. Biosynthesis of heme in mammals. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1763(7), pp.723-736.

Allen, J. W. 2011. Cytochrome c biogenesis in mitochondria–Systems III and V. *FEBS Journal,* 278**,** 4198-4216.

Allen, J.W.A. & Ferguson, S.J., 2006. What is the substrate specificity of the System I cytochrome c biogenesis apparatus?. *Biochemical Society Transactions*, 34(1), pp.150-151.

Allen, J. W., Ginger, M. L. & Ferguson, S. J. 2004. Maturation of the unusual single-cysteine (XXXCH) mitochondrial c-type cytochromes found in trypanosomatids must occur through a novel biogenesis pathway. *Biochemical Journal,* 383**,** 537-542.

Allen, J.W., Harvat, E.M., Stevens, J.M. & Ferguson, S.J., 2006. A variant System I for cytochrome c biogenesis in archaea and some bacteria has a novel CcmE and no CcmH. *FEBS letters*, 580(20), pp.4827-4834.

Allen, J.W., Jackson, A.P., Rigden, D.J., Willis, A.C., Ferguson, S.J. & Ginger, M.L., 2008. Order within a mosaic distribution of mitochondrial c-type cytochrome biogenesis systems?. *The FEBS journal*, 275(10), pp.2385-2402.

Antoine, J.C., Prina, E., Jouanne, C. and Bongrand, P., 1990. Parasitophorous vacuoles of Leishmania amazonensis-infected macrophages maintain an acidic pH. *Infection and immunity*, 58(3), pp.779-787.

Arnaiz, O. & Sperling, L. 2011, "ParameciumDB in 2011: new tools and new data for functional and comparative genomics of the model ciliate Paramecium tetraurelia", *Nucleic acids research*, vol. 39, no. Database issue, pp. D632-D636.

Ast, J., Stiebler, A.C., Freitag, J. & Bölker, M., 2013. Dual targeting of peroxisomal proteins. *Frontiers in physiology*, 4, p.297.

Aurrecoechea, C., Barreto, A., Brestelli, J., Brunk, B.P., Caler, E.V., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G. & Harb, O.S., 2010. AmoebaDB and MicrosporidiaDB: functional genomic resources for Amoebozoa and Microsporidia species. *Nucleic acids research*, 39(suppl_1), pp.D612-D619.

Babbitt, S. E., Hsu, J. & Kranz, R. G. 2016. Molecular Basis Behind Inability of Mitochondrial Holocytochrome c Synthase to Mature Bacterial Cytochromes DEFINING A CRITICAL ROLE FOR CYTOCHROME c α HELIX-1. *Journal of Biological Chemistry,* 291**,** 17523-17534.

Babbitt, S. E., San Francisco, B., Bretsnyder, E. C. & Kranz, R. G. 2014. Conserved residues of the human mitochondrial holocytochrome c synthase mediate interactions with heme. *Biochemistry,* 53**,** 5261-5271.

Babokhov, P., Sanyaolu, A.O., Oyibo, W.A., Fagbenro-Beyioku, A.F. & Iriemenam, N.C., 2013. A current analysis of chemotherapy strategies for the treatment of human African trypanosomiasis. *Pathogens and global health*, 107(5), pp.242-252.

Backes, S. and Herrmann, J.M., 2017. Protein translocation into the intermembrane space and matrix of mitochondria: mechanisms and driving forces. *Frontiers in molecular biosciences*, 4, p.83.

Bakker, B.M., Mensonides, F.I., Teusink, B., van Hoek, P., Michels, P.A. and Westerhoff, H.V., 2000. Compartmentation protects trypanosomes from the dangerous design of glycolysis. *Proceedings of the National Academy of Sciences*, 97(5), pp.2087-2092.

Banci, L., Bertini, I., Cefaro, C., Ciofi-Baffoni, S., Gallo, A., Martinelli, M., Sideris, D.P., Katrakili, N. and Tokatlidis, K., 2009. MIA40 is an oxidoreductase that catalyzes oxidative protein folding in mitochondria. *Nature structural & molecular biology*, 16(2), p.198.

Baral, T.N., 2010. Immunobiology of African trypanosomes: need of alternative interventions. *BioMed Research International*, 2010.

Baron, M.N., Klinger, C.M., Rachubinski, R.A. and Simmonds, A.J., 2016. A Systematic Cell-Based Analysis of Localization of Predicted Drosophila Peroxisomal Proteins. *Traffic*, 17(5), pp.536-553.

Barr, I. & Guo, F., 2015. Pyridine hemochromagen assay for determining the concentration of heme in purified protein solutions. *Bio-protocol*, 5(18).

Bates, P.A., 2018. Revising Leishmania's life cycle. *Nature microbiology*, 3(5), p.529.

Baudhuin, P., Beaufay, H. and de Duve, C., 1965. Combined biochemical and morphological study of particulate fractions from rat liver: analysis of preparations enriched in lysosomes or in particles containing urate oxidase, D-amino acid oxidase, and catalase. *The Journal of cell biology*, 26(1), pp.219-243.

Bauer, S. & Morris, M.T., 2017. Glycosome biogenesis in trypanosomes and the de novo dilemma. *PLoS neglected tropical diseases*, 11(4), p.e0005333.

Beckett, C.S., Loughman, J.A., Karberg, K.A., Donato, G.M., Goldman, W.E. and Kranz, R.G., 2000. Four genes are required for the system II cytochrome c biogenesis pathway in Bordetella pertussis, a unique bacterial model. *Molecular microbiology*, 38(3), pp.465-481.

Beneke, T., Madden, R., Makin, L., Valli, J., Sunter, J. & Gluenz, E., 2017. A CRISPR Cas9 high-throughput genome editing toolkit for kinetoplastids. *Royal Society open science,* 4(5), p.170095.

Berbee, M.L. and Taylor, J.W., 2010. Dating the molecular clock in fungi–how close are we?. *Fungal Biology Reviews*, 24(1-2), pp.1-16.

Berenger, J.M. and Parola, P., 2017. Arthropod Vectors of Medical Importance. In Infectious Diseases (pp. 104-112). *Elsevier*.

Berry, E.A. & Trumpower, B.L., 1987. Simultaneous determination of hemes a, b, and c from pyridine hemochrome spectra. *Analytical biochemistry*, 161(1), pp.1-15.

Bertini, I., Cavallaro, G. & Rosato, A. 2006. Cytochrome c: occurrence and functions. *Chemical reviews*, 106, 90-115.

Bexkens, M.L., Zimorski, V., Sarink, M.J., Wienk, H., Brouwers, J.F., De Jonckheere, J.F., Martin, W.F., Opperdoes, F.R., Van Hellemond, J.J. & Tielens, A.G., 2018. Lipids are the preferred substrate of the protist Naegleria gruberi, relative of a human brain pathogen. *Cell reports*, 25(3), pp.537-543.

Birdsey, G.M., Lewin, J., Cunningham, A.A., Bruford, M.W. & Danpure, C.J., 2004. Differential enzyme targeting as an evolutionary adaptation to herbivory in carnivora. *Molecular biology and evolution*, 21(4), pp.632-646.

Biteau, N., Asencio, C., Izotte, J., Rousseau, B., Fevre, M., Pillay, D. and Baltz, T., 2016. Trypanosoma brucei gambiense infections in mice lead to tropism to the reproductive organs, and horizontal and vertical transmission. *PLoS neglected tropical diseases*, 10(1).

Blekhman, R., Marioni, J.C., Zumbo, P., Stephens, M. & Gilad, Y., 2010. Sex-specific and lineage-specific alternative splicing in primates. *Genome research*, 20(2), pp.180-189.

Boitz, J.M., Ullman, B., Jardim, A. and Carter, N.S., 2012. Purine salvage in Leishmania: complex or simple by design?. *Trends in parasitology*, 28(8), pp.345-352.

Boulanger, M.J., Kukimoto, M., Nishiyama, M., Horinouchi, S. & Murphy, M.E., 2000. Catalytic roles for two water bridged residues (Asp-98 and His-255) in the active site of copper-containing nitrite reductase. *Journal of Biological Chemistry*, 275(31), pp.23957-23964.

Bowman, S.E. and Bren, K.L., 2008. The chemistry and biochemistry of heme c: functional bases for covalent attachment. *Natural product reports*, 25(6), pp.1118-1130.

Bräsen, C., Schmidt, M., Grötzinger, J. and Schönheit, P., 2008. Reaction Mechanism and Structural Model of ADP-forming Acetyl-Coa Synthetase From The Hyperthermophilic Archaeon Pyrococcus Furiosus Evidence For A Second Active Site Histidine Residue. *Journal of Biological Chemistry*, 283(22), pp.15409-15418.

Bravo, F. and Sanchez, M.R., 2003. New and re-emerging cutaneous infectious diseases in Latin America and other geographic areas. *Dermatologic clinics*, 21(4), pp.655-68.

Brocard, C. & Hartig, A., 2006. Peroxisome targeting signal 1: is it really a simple tripeptide?. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1763(12), pp.1565-1573.

Brown, L.A. and Baker, A., 2008. Shuttles and cycles: transport of proteins into the peroxisome matrix. *Molecular membrane biology*, 25(5), pp.363-375.

Brun, R., Blum, J., Chappuis, F. & Burri, C., 2010. Human african trypanosomiasis. *The Lancet*, 375(9709), pp.148-159.

Burki, F., Roger, A.J., Brown, M.W. and Simpson, A.G., 2019. *The new tree of eukaryotes*. Trends in ecology & evolution.

Cáceres, A.J., Quiñones, W., Gualdrón, M., Cordeiro, A., Avilán, L., Michels, P.A. and Concepción, J.L., 2007. Molecular and biochemical characterization of novel glucokinases from Trypanosoma cruzi and Leishmania spp. *Molecular and biochemical parasitology*, 156(2), pp.235-245.

Cardona, T., 2018. Early Archean origin of heterodimeric Photosystem I. *Heliyon*, 4(3), p.e00548.

Catarino, L.M., Serrano, M.G., Cavazzana Jr, M., Almeida, M.L., Kaneshina, E.K., Campaner, M., Jankevicius, J.V., Teixeira, M.M. and Itow-Jankevicius, S., 2001. Classification of trypanosomatids from fruits and seeds using morphological, biochemical and molecular markers revealed several genera among fruit isolates. *FEMS Microbiology Letters*, 201(1), pp.65-72.

Cavalier-Smith, T. & Chao, E.E., 2010. Phylogeny and evolution of apusomonadida (protozoa: apusozoa): new genera and species. *Protist*, 161(4), pp.549-576.

Cavalier-Smith, T., 2009. Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree. *Biology letters*, 6(3), pp.342-345.

Cavalier-Smith, T., 2013. Early evolution of eukaryote feeding modes, cell structural diversity, and classification of the protozoan phyla Loukozoa, Sulcozoa, and Choanozoa. *European journal of protistology*, 49(2), pp.115-178.

Cavalier-Smith, T., 2018. Kingdom Chromista and its eight phyla: a new synthesis emphasising periplastid protein targeting, cytoskeletal and periplastid evolution, and ancient divergences. *Protoplasma*, 255(1), pp.297-357.

Cavalier-Smith, T., Chao, E.E., Snell, E.A., Berney, C., Fiore-Donno, A.M. and Lewis, R., 2014. Multigene eukaryote phylogeny reveals the likely protozoan ancestors of opisthokonts (animals, fungi, choanozoans) and Amoebozoa. *Molecular phylogenetics and evolution*, 81, pp.71-85.

ČErmáková, P., Verner, Z., Man, P., Lukeš, J. & Horváth, A. 2007. Characterization of the NADH: ubiquinone oxidoreductase (complex I) in the trypanosomatid Phytomonas serpens (Kinetoplastida). *FEBS Journal*, 274, 3150-3158.

Chaudhuri, M., Ott, R.D. & Hill, G.C., 2006. Trypanosome alternative oxidase: from molecule to function. *Trends in parasitology*, 22(10), pp.484-491.

Chen, L., Tovar-Corona, J.M. & Urrutia, A.O., 2012. Alternative splicing: a potential source of functional innovation in the eukaryotic genome. *International journal of evolutionary biology*, 2012.

Cho, C.C., Luo, C.W. & Hsu, C.H., 2008. Crystallization and preliminary X-ray diffraction analysis of PAT, an acetyltransferase from Sulfolobus solfataricus. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*, 64(11), pp.1049-1051.

Christensen, D.G., Xie, X., Basisty, N., Byrnes, J., McSweeney, S., Schilling, B. and Wolfe, A.J., 2019. Post-translational Protein Acetylation: An Elegant Mechanism for Bacteria to Dynamically Regulate Metabolic Functions. *Frontiers in microbiology*, 10, p.1604.

Clarke, J.T., Warnock, R.C. and Donoghue, P.C., 2011. Establishing a time-scale for plant evolution. *New Phytologist*, 192(1), pp.266-301.

Coley, A.F., Dodson, H.C., Morris, M.T. & Morris, J.C., 2011. Glycolysis in the african trypanosome: targeting enzymes and their subcellular compartments for therapeutic development. *Molecular biology international*, 2011.

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*, 2011.

Cordon-Obras, C., Rodriguez, Y.F., Fernandez-Martinez, A., Cano, J., Ndong-Mabale, N., Ncogo-Ada, P., Ndongo-Asumu, P., Aparicio, P., Navarro, M., Benito, A. & Bart, J.M., 2015. Molecular evidence of a Trypanosoma brucei gambiense sylvatic cycle in the human african trypanosomiasis foci of Equatorial Guinea. *Frontiers in microbiology*, 6, p.765.

Cramer, W.A. & Kallas, T. Eds., 2016. Cytochrome complexes: evolution, structures, energy transduction, and signaling (Vol. 41). *Springer*.

Creek, D.J., Mazet, M., Achcar, F., Anderson, J., Kim, D.H., Kamour, R., Morand, P., Millerioux, Y., Biran, M., Kerkhoven, E.J. & Chokkathukalam, A., 2015. Probing the metabolic network in bloodstream-form Trypanosoma brucei using untargeted metabolomics with stable isotope labelled glucose. *PLoS pathogens*, 11(3), p.e1004689.

Cridge, A.G., Crowe-McAuliffe, C., Mathew, S.F. and Tate, W.P., 2018. Eukaryotic translational termination efficiency is influenced by the 3′ nucleotides within the ribosomal mRNA channel. *Nucleic acids research*, 46(4), pp.1927-1944.

Cross, L.L., Ebeed, H.T. & Baker, A., 2016. Peroxisome biogenesis, protein targeting mechanisms and PEX gene functions in plants. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1863(5), pp.850-862.

Cunningham, J.A., Liu, A.G., Bengtson, S. and Donoghue, P.C., 2017. The origin of animals: can molecular clocks and the fossil record be reconciled?. *BioEssays*, 39(1), pp.1-12.

Cuvillier, A., Miranda, J.C., Ambit, A., Barral, A. and Merlin, G., 2003. Abortive infection of Lutzomyia longipalpis insect vectors by aflagellated LdARL-3A-Q70L overexpressing Leishmania amazonensis parasites. *Cellular microbiology*, 5(10), pp.717-728.

Da Silva, R.V., Malvezi, A.D., da Silva Augusto, L., Kian, D., Tatakihara, V.L.H., Yamauchi, L.M., Yamada-Ogatta, S.F., Rizzo, L.V., Schenkman, S. and Pinge-Filho, P., 2013. Oral exposure to Phytomonas serpens attenuates thrombocytopenia and leukopenia during acute infection with Trypanosoma cruzi. *PloS one*, 8(7), p.e68299.

Dabrowski, M., Bukowy-Bieryllo, Z. & Zietkiewicz, E., 2015. Translational readthrough potential of natural termination codons in eucaryotes–The impact of RNA sequence. *RNA biology*, 12(9), pp.950-958.

Dale, J.R., Wade, R. & Dichristina, T.J., 2007. A conserved histidine in cytochrome c maturation permease CcmB of Shewanella putrefaciens is required for anaerobic growth below a threshold standard redox potential. *Journal of bacteriology*, 189(3), pp.1036-1043.

Daltrop, O., Stevens, J.M., Higham, C.W. and Ferguson, S.J., 2002. The CcmE protein of the c-type cytochrome biogenesis system: unusual in vitro heme incorporation into apo-CcmE and transfer from holo-CcmE to apocytochrome. *Proceedings of the National Academy of Sciences*, 99(15), pp.9703-9708.

De Almeida, M.C., Vilhena, V., Barral, A. and Barral-Netto, M., 2003. Leishmanial infection: analysis of its first steps. A review. *Memorias do Instituto Oswaldo Cruz*, 98(7), pp.861-870.

De Duve, C., 1965. Functions of microbodies (peroxisomes). *J. Cell Biol*., 27, pp.25A-26A.

De Duve, C., 1969. Evolution of the peroxisome. *Annals of the New York Academy of Sciences*, 168(2), pp.369-381.

De Duve, C.A.B.P. & Baudhuin, P., 1966. Peroxisomes (microbodies and related particles). *Physiological reviews*, 46(2), pp.323-357.

De Jonckheere, J.F., Brown, S., Dobson, P.J., Robinson, B.S. & Pernin, P., 2001. The amoeba-to-flagellate transformation test is not reliable for the diagnosis of the genus Naegleria. Description of three new Naegleria spp. *Protist*, 152(2), pp.115-121.

De la Cruz, V.F., Neckelmann, N. and Simpson, L., 1984. Sequences of six genes and several open reading frames in the kinetoplast maxicircle DNA of Leishmania tarentolae. *Journal of Biological Chemistry*, 259(24), pp.15136-15147.

De Vitry, C., 2011. Cytochrome c maturation system on the negative side of bioenergetic membranes: CCB or System IV. *The FEBS journal*, 278(22), pp.4189-4197.

Dean, S.D. & Matthews, K.R., 2007. Restless gossamers: antibody clearance by hydrodynamic flow forces generated at the surface of motile trypanosome parasites. *Cell host & microbe*, 2(5), pp.279-281.

Dewar, C.E., MacGregor, P., Cooper, S., Gould, M.K., Matthews, K.R., Savill, N.J. & Schnaufer, A., 2018. Mitochondrial DNA is critical for longevity and metabolism of transmission stage Trypanosoma brucei. *PLoS pathogens*, 14(7), p.e1007195.

Di Lucca, A.G.T., Chipana, T., Fernando, E., Talledo Albújar, M.J., Dávila Peralta, W., Montoya Piedra, Y.C. and Arévalo Zelada, J.L., 2013. Slow wilt: another form of Marchitez in oil palm associated with trypanosomatids in Peru. *Tropical Plant Pathology*, 38(6), pp.522-533.

Di Matteo, A., Gianni, S., Schininà, M.E., Giorgi, A., Altieri, F., Calosci, N., Brunori, M. & Travaglini-AllocatelII, C., 2007. A Strategic Protein in Cytochrome c Maturation Three-Dimensional Structure Of CcmH And Binding To Apocytochrome c. *Journal of Biological Chemistry*, 282(37), pp.27012-27019.

Diekert, K., Kispal, G., Guiard, B. and Lill, R., 1999. An internal targeting signal directing proteins into the mitochondrial intermembrane space. *Proceedings of the National Academy of Sciences*, 96(21), pp.11752-11757.

Dollet, M., 1984. Plant diseases caused by flagellate protozoa (Phytomonas). *Annual review of phytopathology*, 22(1), pp.115-132.

Donovan, C., 1909. Kala-azar in Madras, especially with regard to its connexion with the dog and the bug (Conorrhinus). *The Lancet*, 174(4499), pp.1495-1496.

Doolittle, W. F. 1998. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends in Genetics*, 14, 307-311.

Dostálová, A. and Volf, P., 2012. Leishmania development in sand flies: parasite-vector interactions overview. *Parasites & vectors*, 5(1), p.276.

Dufernez, F., Yernaux, C., Gerbod, D., Noël, C., Chauvenet, M., Wintjens, R., Edgcomb, V.P., Capron, M., Opperdoes, F.R. and Viscogliosi, E., 2006. The presence of four iron-containing superoxide dismutase isozymes in trypanosomatidae: characterization, subcellular localization, and phylogenetic origin in Trypanosoma brucei. *Free Radical Biology and Medicine*, 40(2), pp.210-225.

Dyda, F., Klein, D.C. & Hickman, A.B., 2000. GCN5-related N-acetyltransferases: a structural overview. *Annual review of biophysics and biomolecular structure*, 29(1), pp.81-103.

Ebenezer, T.E., Zoltner, M., Burrell, A., Nenarokova, A., Vanclová, A.M.N., Prasad, B., Soukal, P., Santana-Molina, C., O'neill, E., Nankissoor, N.N. & Vadakedath, N., 2019. Transcriptome, proteome and draft genome of Euglena gracilis. *BMC biology*, 17(1), p.11.

Eichinger, L., Pachebat, J.A., Glöckner, G., Rajandream, M.A., Sucgang, R., Berriman, M., Song, J., Olsen, R., Szafranski, K., Xu, Q. & Tunggal, B., 2005. The genome of the social amoeba Dictyostelium discoideum. *Nature*, 435(7038), p.43.

Elliott, J.L., Saliba, K.J. and Kirk, K., 2001. Transport of lactate and pyruvate in the intraerythrocytic malaria parasite, Plasmodium falciparum. *Biochemical Journal*, 355(3), pp.733-739.

Emanuelsson, O., Elofsson, A., Von Heijne, G. & Cristobal, S., 2003. In silico prediction of the peroxisomal proteome in fungi, plants and animals. *Journal of molecular biology*, 330(2), pp.443-456.

Eme, L., Sharpe, S.C., Brown, M.W. and Roger, A.J., 2014. On the age of eukaryotes: evaluating evidence from fossils and molecular clocks. *Cold Spring Harbor Perspectives in Biology*, 6(8), p.a016139.

Enggist, E., Schneider, M.J., Schulz, H. & Thöny-Meyer, L., 2003. Biochemical and mutational characterization of the heme chaperone CcmE reveals a heme binding site. *Journal of bacteriology*, 185(1), pp.175-183.

Engstler, M., Pfohl, T., Herminghaus, S., Boshart, M., Wiegertjes, G., Heddergott, N. and Overath, P., 2007. Hydrodynamic flow-mediated protein sorting on the cell surface of trypanosomes. *Cell*, 131(3), pp.505-515.

Esch, K.J. and Petersen, C.A., 2013. Transmission and epidemiology of zoonotic protozoal diseases of companion animals. *Clinical microbiology reviews*, 26(1), pp.58-85.

Feissner, R.E., Beckett, C.S., Loughman, J.A. and Kranz, R.G., 2005. Mutations in cytochrome assembly and periplasmic redox pathways in Bordetella pertussis. *Journal of bacteriology,* 187(12), pp.3941-3949.

Feissner, R.E., Richard-Fogal, C.L., Frawley, E.R. and Kranz, R.G., 2006. ABC transporter-mediated release of a haem chaperone allows cytochrome c biogenesis. *Molecular microbiology*, 61(1), pp.219-231.

Fodor, K., Wolf, J., Erdmann, R., Schliebs, W. and Wilmanns, M., 2012. Molecular requirements for peroxisomal targeting of alanine-glyoxylate aminotransferase as an essential determinant in primary hyperoxaluria type 1. *PLoS biology*, 10(4).

Fothergill-Gilmore, L.A. & Michels, P.A., 1993. Evolution of glycolysis. *Progress in biophysics and molecular biology,* 59(2), pp.105-235.

Fraga, J., Fernández-Calienes, A., Montalvo, A.M., Maes, I., Deborggraeve, S., Büscher, P., Dujardin, J.C. and Van der Auwera, G., 2016. Phylogenetic analysis of the Trypanosoma genus based on the heat-shock protein 70 gene. Infection, *Genetics and Evolution*, 43, pp.165-172.

Franco, J. R., Cecchi, G., Priotto, G., Paone, M., Diarra, A., Grout, L., Mattioli, R. C. & Argaw, D. 2017. Monitoring the elimination of human African trypanosomiasis: update to 2014. *PLoS neglected tropical diseases,* 11**,** e0005585.

Franco, J.R., Simarro, P.P., Diarra, A. & Jannin, J.G., 2014. Epidemiology of human African trypanosomiasis. *Clinical epidemiology*, 6, p.257.

Freitag, J., Ast, J. & Bölker, M., 2012. Cryptic peroxisomal targeting via alternative splicing and stop codon read-through in fungi. *Nature*, 485(7399), p.522.

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., 2010. The genome of Naegleria gruberi illuminates early eukaryotic versatility. *Cell*, 140(5), pp.631-642.

Fülöp, V., Sam, K.A., Ferguson, S.J., Ginger, M.L. & Allen, J.W., 2009. Structure of a trypanosomatid mitochondrial cytochrome c with heme attached via only one thioether bond and implications for the substrate recognition requirements of heme lyase. *The FEBS journal*, 276(10), pp.2822-2832.

Gabaldón, T., 2010. Peroxisome diversity and evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1541), pp.765-773.

Gabaldón, T., 2010. Peroxisome diversity and evolution. Philosophical Transactions of the Royal Society B: *Biological Sciences*, 365(1541), pp.765-773.

Gabaldón, T., Ginger, M.L. & Michels, P.A., 2016. Peroxisomes in parasitic protists. *Molecular and biochemical parasitology*, 209(1-2), pp.35-45.

Gabilly, S.T. & Hamel, P.P., 2017. Maturation of plastid c-type cytochromes. *Frontiers in plant science*, 8, p.1313.

Gajria, B., Bahl, A., Brestelli, J., Dommer, J., Fischer, S., Gao, X., Heiges, M., Iodice, J., Kissinger, J.C., Mackey, A.J. and Pinney, D.F., 2007. ToxoDB: an integrated Toxoplasma gondii database resource. *Nucleic acids research*, 36(suppl_1), pp.D553-D556.

Galván, A., Córdoba, F., Cárdenas, J. & Fernández, E., 1991. Regulation of nitrite uptake and nitrite reductase expression in Chlamydomonas reinhardtii. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1074(1), pp.6-11.

Giegé, P., Grienenberger, J.M. and Bonnard, G., 2008. Cytochrome c biogenesis in mitochondria. *Mitochondrion*, 8(1), pp.61-73.

Ginger, M.L., 2005. Niche metabolism in parasitic protozoa. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1465), pp.101-118.

Ginger, M.L., Fritz-Laylin, L.K., Fulton, C., Cande, W.Z. & Dawson, S.C., 2010. Intermediary metabolism in protists: a sequence-based view of facultative anaerobic metabolism in evolutionarily diverse eukaryotes. *Protist*, 161(5), pp.642-671.

Ginger, M.L., Mcfadden, G.I. & Michels, P.A.M. 2010, "The evolution of organellar metabolism in unicellular eukaryotes", *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 365, no. 1541, pp. 693-698.

Ginger, M.L., Sam, K.A. & Allen, J.W., 2012. Probing why trypanosomes assemble atypical cytochrome c with an AxxCH haem-binding motif instead of CxxCH. *Biochemical Journal*, 448(2), pp.253-260.

Gluenz, E., Povelones, M.L., Englund, P.T. and Gull, K., 2011. The kinetoplast duplication cycle in Trypanosoma brucei is orchestrated by cytoskeleton-mediated cell morphogenesis. *Molecular and cellular biology*, 31(5), pp.1012-1021.

Goddard, A. D., Stevens, J. M., Rao, F., Mavridou, D. A., Chan, W., Richardson, D. J., Allen, J. W. & Ferguson, S. J. 2010. c-Type cytochrome biogenesis can occur via a natural Ccm system lacking CcmH, CcmG, and the heme-binding histidine of CcmE. *Journal of Biological Chemistry*, 285, 22882-22889.

Goodhead, I., Capewell, P., Bailey, J.W., Beament, T., Chance, M., Kay, S., Forrester, S., MacLeod, A., Taylor, M., Noyes, H. and Hall, N., 2013. Whole-genome sequencing of Trypanosoma brucei reveals introgression between subspecies that is associated with virulence. *MBio*, 4(4), pp.e00197-13.

Gould, S.J., Keller, G.A., Hosken, N., Wilkinson, J. & Subramani, S., 1989. A conserved tripeptide sorts proteins to peroxisomes. *The Journal of cell biology*, 108(5), pp.1657-1664.

Grau-Bove, X., Torruella, G., Donachie, S., Suga, H., Leonard, G., Richards, T.A. & Ruiz-Trillo, I., 2017. Dynamics of genomic innovation in the unicellular ancestry of animals. *Elife*, 6, p.e26036.

Guerra-Giraldez, C., Quijada, L. and Clayton, C.E., 2002. Compartmentation of enzymes in a microbody, the glycosome, is essential in Trypanosoma brucei. *Journal of Cell Science*, 115(13), pp.2651-2658.

Gunning, P., Weinberger, R., Jeffrey, P. & Hardeman, E., 1998. Isoform sorting and the creation of intracellular compartments. *Annual review of cell and developmental biology*, 14(1), pp.339-372.

Guo, F., Ortega-Pierres, G., Argüello-García, R., Zhang, H. and Zhu, G., 2015. Giardia fatty acyl-CoA synthetases as potential drug targets. *Frontiers in microbiology*, 6, p.753.

Guo, M., Bhaskar, B., Li, H., Barrows, T.P. & Poulos, T.L., 2004. Crystal structure and characterization of a cytochrome c peroxidase–cytochrome c site-specific cross-link. *Proceedings of the National Academy of Sciences*, 101(16), pp.5940-5945.

Hallows, W.C., Lee, S. & Denu, J.M., 2006. Sirtuins deacetylate and activate mammalian acetyl-CoA synthetases. *Proceedings of the National Academy of Sciences*, 103(27), pp.10230-10235.

Hamel, P., Corvest, V., Giegé, P. & Bonnard, G. 2009. Biochemical requirements for the maturation of mitochondrial c-type cytochromes. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research,* 1793**,** 125-138.

Hamel, P.P., Dreyfuss, B.W., Xie, Z., Gabilly, S.T. and Merchant, S., 2003. Essential histidine and tryptophan residues in CcsA, a system II polytopic cytochrome c biogenesis protein. *Journal of Biological Chemistry*, 278(4), pp.2593-2603.

Hammarton, T.C., 2007. Cell cycle regulation in Trypanosoma brucei. *Molecular and biochemical parasitology*, 153(1), pp.1-8.
Hammond, D.J., Aman, R.A. and Wang, C.C., 1985. The role of compartmentation and glycerol kinase in the synthesis of ATP within the glycosome of Trypanosoma brucei. *Journal of Biological Chemistry*, 260(29), pp.15646-15654.
Hannaert, V., Bringaud, F., Opperdoes, F.R. & Michels, P.A., 2003. Evolution of energy metabolism and its compartmentation in Kinetoplastida. *Kinetoplastid biology and disease*, 2(1), p.11.
Hassan, B.A. & Ali, S.A., 2018. Primary meningoencephalitis caused by Naegleria fowleri: a mini review. *Baqai Journal of Health Sciences*, 21(1), pp.42-48.
Hayashi, Y. & Shinozaki, A., 2012. Visualization of microbodies in Chlamydomonas reinhardtii. *Journal of plant research*, 125(4), pp.579-586.
He, D., Fiz-Palacios, O., Fu, C.J., Fehling, J., Tsai, C.C. and Baldauf, S.L., 2014. An alternative root for the eukaryote tree of life. *Current Biology*, 24(4), pp.465-470.
Hedges, S.B., Blair, J.E., Venturi, M.L. & Shoe, J.L., 2004. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC evolutionary biology*, 4(1), p.2.
Heiss, A.A., Kolisko, M., Ekelund, F., Brown, M.W., Roger, A.J. and Simpson, A.G., 2018. Combined morphological and phylogenomic re-examination of malawimonads, a critical taxon for inferring the evolutionary history of eukaryotes. *Royal Society open science*, 5(4), p.171707.
Helen, D., Kim, H., Tytgat, B. & Anne, W., 2016. Highly diverse nirK genes comprise two major clades that harbour ammonium-producing denitrifiers. *BMC genomics*, 17(1), p.155.
Henze, K., Horner, D.S., Suguri, S., Moore, D.V., Sánchez, L.B., Müller, M. and Embley, T.M., 2001. Unique phylogenetic relationships of glucokinase and glucosephosphate isomerase of the amitochondriate eukaryotes Giardia intestinalis, Spironucleus barkhanus and Trichomonas vaginalis. *Gene*, 281(1-2), pp.123-131.
Herrmann, J.M. and Riemer, J., 2014. Three approaches to one problem: protein folding in the periplasm, the endoplasmic reticulum, and the intermembrane space. *Antioxidants & redox signaling*, 21(3), pp.438-456.
Heuser, M. & Razavi, L. 1970. Amebo-flagellates as research partners: the laboratory biology of Naegleria and Tetramitus. *Methods in cell biology.* Elsevier.
Hildmann, C., Riester, D. & Schwienhorst, A., 2007. Histone deacetylases—an important class of cellular regulators with a variety of functions. *Applied microbiology and biotechnology*, 75(3), pp.487-497.
Hirota, S., Hattori, Y., Nagao, S., Taketa, M., Komori, H., Kamikubo, H., Wang, Z., Takahashi, I., Negi, S., Sugiura, Y. and Kataoka, M., 2010. Cytochrome c polymerization by successive domain swapping at the C-terminal helix. *Proceedings of the National Academy of Sciences*, 107(29), pp.12854-12859.
Ho, A.T. & Hurst, L.D., 2019. In eubacteria, unlike eukaryotes, there is no evidence for selection favouring fail-safe 3'additional stop codons. *PLoS genetics*, 15(9), p.e1008386.
Hoare, C.A. and Wallace, F.G., 1966. Developmental stages of trypanosomatid flagellates: a new terminology. *Nature*, 212(5068), pp.1385-1386.
Hollingsworth, T.D., 2018. Counting down the 2020 Goals for 9 Neglected Tropical Diseases: what have we learned from quantitative analysis and transmission modeling?. *Clinical Infectious Diseases*, 66(suppl_4), pp.S237-S244.
Horrell, S., Kekilli, D., Strange, R.W. and Hough, M.A., 2017. Recent structural insights into the function of copper nitrite reductases. *Metallomics*, 9(11), pp.1470-1482. Hutchinson, R. and Stevens, J.R., 2018. Barcoding in trypanosomes. *Parasitology*, 145(5), pp.563-573.
Hwang, Y.T., Pelitire, S.M., Henderson, M.P., Andrews, D.W., Dyer, J.M. & Mullen, R.T., 2004. Novel targeting signals mediate the sorting of different isoforms of the tail-anchored membrane protein cytochrome b5 to either endoplasmic reticulum or mitochondria. *The Plant Cell*, 16(11), pp.3002-3019.

Imoto, Y., Abe, Y., Okumoto, K., Honsho, M., Kuroiwa, H., Kuroiwa, T. & Fujiki, Y., 2017. Defining the dynamin-based ring organizing center on the peroxisome-dividing machinery isolated from Cyanidioschyzon merolae. *J Cell Sci*, 130(5), pp.853-867.

Ishemgulova, A., Hlaváčová, J., Majerova, K., Butenko, A., Lukeš, J., Votýpka, J., Volf, P. & Yurchenko, V., 2018. CRISPR/Cas9 in Leishmania mexicana: A case study of LmxBTN1. *PloS one*, 13(2), p.e0192723.

Ishida, T. & Kinoshita, K., 2007. PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic acids research*, 35(suppl_2), pp.W460-W464.

Janouškovec, J., Tikhonenkov, D.V., Burki, F., Howe, A.T., Rohwer, F.L., Mylnikov, A.P. & Keeling,P.J., 2017. A new lineage of eukaryotes illuminates early mitochondrial genome reduction. *Current Biology*, 27(23), pp.3717-3724.

Jaskowska, E., Butler, C., Preston, G. and Kelly, S., 2015. Phytomonas: trypanosomatids adapted to plant environments. *PLoS Pathogens*, 11(1), p.e1004484.

Jastroch, M., Divakaruni, A.S., Mookerjee, S., Treberg, J.R. & Brand, M.D., 2010. Mitochondrial proton and electron leaks. *Essays in biochemistry*, 47, pp.53-67.

Jerlström-Hultqvist, J., Einarsson, E., Xu, F., Hjort, K., Ek, B., Steinhauf, D., Hultenby, K., Bergquist, J., Andersson, J.O. & Svärd, S.G., 2013. Hydrogenosomes in the diplomonad Spironucleus salmonicida. *Nature communications*, 4, p.2493.

Johan, F., 2002. A century of research on the amoeboflagellate genus Naegleria. *Acta Protozool*, 41, pp.309-342.

Jones, C.P. & Ingram-Smith, C., 2014. Biochemical and kinetic characterization of the recombinant ADP-forming acetyl coenzyme A synthetase from the amitochondriate protozoan Entamoeba histolytica. *Eukaryotic cell*, 13(12), pp.1530-1537.

Jones, C.P., Khan, K. & Ingram-Smith, C., 2017. Investigating the mechanism of ADP-forming acetyl-CoA synthetase from the protozoan parasite Entamoeba histolytica. *FEBS letters*, 591(4), pp.603-612.

Kadokura, H. and Beckwith, J., 2010. Mechanisms of oxidative protein folding in the bacterial cell envelope. *Antioxidants & redox signaling*, 13(8), pp.1231-1246.

Kanai, M., Nishimura, M. & HayashI, M., 2010. A peroxisomal ABC transporter promotes seed germination by inducing pectin degradation under the control of ABI5. *The Plant Journal*, 62(6), pp.936-947.

Katoh, K. & Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4), pp.772-780.

Kaufer, A., Ellis, J., Stark, D. and Barratt, J., 2017. The evolution of trypanosomatid taxonomy. *Parasites & vectors*, 10(1), p.287.

Kent, W.S., 1883. *Handbook of the Marine and Freshwater Fishes of the British Islands:(including an Enumeration of Every Species).* W. Clowes and Sons.

Kersey, P.J., Allen, J.E., Allot, A., Barba, M., Boddu, S., Bolt, B.J., Carvalho-Silva, D., Christensen, M., Davis, P., Grabmueller, C., Kumar, N., Liu, Z., Maurel, T., Moore, B., Mcdowall, M.D., Maheswari, U., Naamati, G., Newman, V., Ong, C.K., Paulini, M., Pedro, H., Perry, E., Russell, M., Sparrow, H., Tapanari, E., Taylor, K., Vullo, A., Williams, G., Zadissia, A., Olson, A., Stein, J., Wei, S., Tello-Ruiz, M., Ware, D., Luciani, A., Potter, S., Finn, R.D., Urban, M., Hammond-Kosack, K.E., Bolser, D.M., De Silva, N., Howe, K.L., Langridge, N., Maslen, G., Staines, D.M. & Yates, A. 2018, "Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species", *Nucleic Acids Research*, vol. 46, no. D1, pp. D802-D808.

Kim, P.K. & Hettema, E.H., 2015. Multiple pathways for protein transport to peroxisomes. *Journal of molecular biology*, 427(6), pp.1176-1190.

Kim, P.K. and Hettema, E.H., 2015. Multiple pathways for protein transport to peroxisomes. *Journal of molecular biology*, 427(6), pp.1176-1190.

Kim, S.C., Sprung, R., Chen, Y., Xu, Y., Ball, H., Pei, J., Cheng, T., Kho, Y., Xiao, H., Xiao, L. & Grishin, N.V., 2006. Substrate and functional diversity of lysine acetylation revealed by a proteomics survey. *Molecular cell*, 23(4), pp.607-618.

Kim, S.W., Fushinobu, S., Zhou, S., Wakagi, T. & Shoun, H., 2009. Eukaryotic nirK genes encoding copper-containing nitrite reductase: originating from the protomitochondrion?. *Appl. Environ. Microbiol*., 75(9), pp.2652-2658.

Kira, Y., Sato, E.F. & Inoue, M., 2002. Association of Cu, Zn-type superoxide dismutase with mitochondria and peroxisomes. *Archives of biochemistry and biophysics*, 399(1), pp.96-102.

Kisslov, I., Naamati, A., Shakarchy, N. & Pines, O., 2014. Dual-targeted proteins tend to be more evolutionarily conserved. *Molecular biology and evolution*, 31(10), pp.2770-2779.

Kleingardner, J.G. & Bren, K.L., 2011. Comparing substrate specificity between cytochrome c maturation and cytochrome c heme lyase systems for cytochrome c biogenesis. *Metallomics*, 3(4), pp.396-403.

Kletzin, A., Heimerl, T., Flechsler, J., Van Niftrik, L., Rachel, R. & Klingl, A., 2015. Cytochromes c in Archaea: distribution, maturation, cell architecture, and the special case of Ignicoccus hospitalis. *Frontiers in microbiology*, 6, p.439.

Kletzin, A., Heimerl, T., Flechsler, J., Van Niftrik, L., Rachel, R. & Klingl, A., 2015. Cytochromes c in Archaea: distribution, maturation, cell architecture, and the special case of Ignicoccus hospitalis. *Frontiers in microbiology*, 6, p.439.

Kobayashi, M. and Shoun, H., 1995. The copper-containing dissimilatory nitrite reductase involved in the denitrifying system of the fungus Fusarium oxysporum. *Journal of Biological Chemistry*, 270(8), pp.4146-4151.

Kojer, K., Peleh, V., Calabrese, G., Herrmann, J.M. and Riemer, J., 2015. Kinetic control by limiting glutaredoxin amounts enables thiol oxidation in the reducing mitochondrial intermembrane space. *Molecular biology of the cell*, 26(2), pp.195-204.

Koonin, E.V., 2010. The incredible expanding ancestor of eukaryotes. *Cell*, 140(5), pp.606-608.

Kořený, L., Sobotka, R., Kovářová, J., Gnipová, A., Flegontov, P., Horváth, A., Oborník, M., Ayala, F. J. & Lukeš, J. 2012. Aerobic kinetoplastid flagellate Phytomonas does not require heme for viability. *Proceedings of the National Academy of Sciences*, 109, 3808-3813.

Kranz, R.G., Richard-Fogal, C., Taylor, J.S. and Frawley, E.R., 2009. Cytochrome c biogenesis: mechanisms for covalent modifications and trafficking of heme and for heme-iron redox control. *Microbiol*. Mol. Biol. Rev., 73(3), pp.510-528.

Krogh, A., Larsson, B., Von Heijne, G. & Sonnhammer, E. L. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology*, 305, 567-580.

Lamour, N., Rivière, L., Coustou, V., Coombs, G.H., Barrett, M.P. & Bringaud, F., 2005. Proline metabolism in procyclic Trypanosoma brucei is down-regulated in the presence of glucose. *Journal of Biological Chemistry*, 280(12), pp.11902-11910.

Lang, B. F., Burger, G., O'kelly, C. J. & Cedergren, R. 1997. An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature*, 387, 493.

Lange, C. & Hunte, C., 2002. Crystal structure of the yeast cytochrome bc1 complex with its bound substrate cytochrome c. *Proceedings of the National Academy of Sciences*, 99(5), pp.2800-2805.

Layer, G., Reichelt, J., Jahn, D. & Heinz, D.W., 2010. Structure and function of enzymes in heme biosynthesis. *Protein Science,* 19(6), pp.1137-1161.

Lazarow, P.B., 2006 .1. 7. The import receptor Pex7p and the PTS2 targeting sequence. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1763(12), pp.1599-1604.

Le Brun, N.E., Bengtsson, J. and Hederstedt, L., 2000. Genes required for cytochrome c synthesis inBacillus subtilis. *Molecular microbiology*, 36(3), pp.638-650.

Lee, J.H., Harvat, E.M., Stevens, J.M., Ferguson, S.J. & Saier JR, M.H., 2007. Evolutionary origins of members of a superfamily of integral membrane cytochrome c biogenesis proteins. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1768(9), pp.2164-2181.

Lee, K.K. & Workman, J.L., 2007. Histone acetyltransferase complexes: one size doesn't fit all. *Nature reviews Molecular cell biology*, 8(4), p.284.

Lees, J.A., Kendall, M., Parkhill, J., Colijn, C., Bentley, S.D. and Harris, S.R., 2018. Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study. *Wellcome open research*, 3.

Leger, M.M., Gawryluk, R.M., Gray, M.W. and Roger, A.J., 2013. Evidence for a hydrogenosomal-type anaerobic ATP generation pathway in Acanthamoeba castellanii. *PLoS One*, 8(9), p.e69532.

Letunic, I. & Bork, P., 2017. 20 years of the SMART protein domain annotation resource. *Nucleic acids research*, 46(D1), pp.D493-D496.

Lewin, A., Crow, A., Hodson, C.T., Hederstedt, L. and Le Brun, N.E., 2008. Effects of substitutions in the CXXC active-site motif of the extracytoplasmic thioredoxin ResA. *Biochemical Journal*, 414(1), pp.81-91.

Lewis, W.H., Lind, A.E., Sendra, K.M., Onsbring, H., Williams, T.A., Esteban, G., Hirt, R.P., Ettema, T.J. and Embley, T.M., 2019. Convergent evolution of hydrogenosomes from mitochondria by gene transfer and loss. *Molecular biology and evolution*.

Lezhneva, L., Kuras, R., Ephritikhine, G. and De Vitry, C., 2008. A novel pathway of cytochrome c biogenesis is involved in the assembly of the cytochrome b6f complex in Arabidopsis chloroplasts. *Journal of Biological Chemistry*, 283(36), pp.24608-24616.

Li, C. & Zhang, J., 2019. Stop-codon read-through arises largely from molecular errors and is generally nonadaptive. *PLoS genetics*, 15(5), p.e1008141.

Li, Y., Hodak, M. & Bernholc, J., 2015. Enzymatic mechanism of copper-containing nitrite reductase. *Biochemistry*, 54(5), pp.1233-1242.

Li, Y., Sun, Y., Hines, J.C. and Ray, D.S., 2007. Identification of new kinetoplast DNA replication proteins in trypanosomatids based on predicted S-phase expression and mitochondrial targeting. *Eukaryotic cell*, 6(12), pp.2303-2310.

Lingner, T., Kataya, A.R., Antonicelli, G.E., Benichou, A., Nilssen, K., Chen, X.Y., Siemsen, T., Morgenstern, B., Meinicke, P. & Reumann, S., 2011. Identification of novel plant peroxisomal targeting signals by a combination of machine learning methods and in vivo subcellular targeting analyses. *The Plant Cell*, 23(4), pp.1556-1572.

Lodish, H., Berk, A., Zipursky, S.L., Matsudaira, P., Baltimore, D. and Darnell, J., 2000. Molecular cell biology 4th edition. *National Center for Biotechnology Information, Bookshelf.*

Long, S. S., Prober, C. G. & Fischer, M. 2017. *Principles and practice of pediatric infectious diseases*, Elsevier Health Sciences.

López-Escardó, D., López-García, P., Moreira, D., Ruiz-Trillo, I. & Torruella, G. 2018, "Parvularia atlantis gen. et sp. nov., a Nucleariid Filose Amoeba (Holomycota, Opisthokonta)", *Journal of Eukaryotic Microbiology*, vol. 65, no. 2, pp. 170-179.

Loughran, G., Chou, M.Y., Ivanov, I.P., Jungreis, I., Kellis, M., Kiran, A.M., Baranov, P.V. and Atkins, J.F., 2014. Evidence of efficient stop codon readthrough in four mammalian genes. *Nucleic acids research*, 42(14), pp.8928-8938.

Lozupone, C.A., Knight, R.D. & Landweber, L.F. 2001, "The molecular basis of nuclear genetic code change in ciliates", *Current Biology*, vol. 11, no. 2, pp. 65-74.

Ludewig-Klingner, A.K., Michael, V., Jarek, M., Brinkmann, H. & Petersen, J., 2017. Distribution and evolution of peroxisomes in Alveolates (apicomplexa, dinoflagellates, ciliates). *Genome biology and evolution*, 10(1), pp.1-13.

Lukeš, J., Guilbride, D.L., Votýpka, J., Zíková, A., Benne, R. & Englund, P.T., 2002. Kinetoplast DNA network: evolution of an improbable structure. *Eukaryotic cell*, 1(4), pp.495-502.

Lukeš, J., Hashimi, H. and Zíková, A., 2005. Unexplained complexity of the mitochondrial genome and transcriptome in kinetoplastid flagellates. *Current genetics*, 48(5), pp.277-299.

Lukeš, J., Skalický, T., Týč, J., Votýpka, J. and Yurchenko, V., 2014. Evolution of parasitism in kinetoplastid flagellates. *Molecular and biochemical parasitology*, 195(2), pp.115-122.

Lynch, A.C., 2016. *Novel mitochondrial metabolism in excavate protists* (Doctoral dissertation, Lancaster University).

Ma, C. & Reumann, S., 2008. Improved prediction of peroxisomal PTS1 proteins from genome sequences based on experimental subcellular targeting analyses as exemplified for protein kinases from Arabidopsis. *Journal of experimental botany*, 59(13), pp.3767-3779.

Majid, M.A.A., Mahboob, T., Mong, B.G., Jaturas, N., Richard, R.L., Tian-Chye, T., Phimphila, A., Mahaphonh, P., Aye, K.N., Aung, W.L. & Chuah, J., 2017. Pathogenic waterborne free-living amoebae: an update from selected Southeast Asian countries. *PloS one*, 12(2), p.e0169448.

Manganas, P., Macpherson, L. & Tokatlidis, K., 2017. Oxidative protein biogenesis and redox regulation in the mitochondrial intermembrane space. *Cell and tissue research*, 367(1), pp.43-57.

Marciano-Cabral, F., 1988. Biology of Naegleria spp. *Microbiological Reviews*, 52(1), p.114.

Maslov, D. A., Opperdoes, F. R., Kostygov, A. Y., Hashimi, H., Lukeš, J. & Yurchenko, V. 2018. Recent advances in trypanosomatid research: genome organization, expression, metabolism, taxonomy and evolution. *Parasitology***,** 1-27, .

Matasci, N., Hung, L., Yan, Z., Carpenter, E.J., Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Ayyampalayam, S., Barker, M., Burleigh, J.G., Gitzendanner, M.A., Wafula, E., Der, J.P., Depamphilis, C.W., Roure, B., Philippe, H., Ruhfel, B.R., Miles, N.W., Graham, S.W., Mathews, S., Surek, B., Melkonian, M., Soltis, D.E., Soltis, P.S., Rothfels, C., Pokorny, L., Shaw, J.A., Degironimo, L., Stevenson, D.W., Villarreal, J.C., Chen, T., Kutchan, T.M., Rolf, M., Baucom, R.S., Deyholos, M.K., Samudrala, R., Tian, Z., Wu, X., Sun, X., Zhang, Y., Wang, J., Leebens-Mack, J. & Wong, G.K. 2014, "Data access for the 1,000 Plants (1KP) project", *GigaScience*, vol. 3, no. 1, pp. 17.

Matthews, K. R. 2005. The developmental cell biology of Trypanosoma brucei. *Journal of cell science,* 118**,** 283-290.

Mavridou, D.A., Clark, M.N., Choulat, C., Ferguson, S.J. & Stevens, J.M., 2013. Probing heme delivery processes in cytochrome c biogenesis system I. *Biochemistry*, 52(41), pp.7262-7270.

Mészáros, B., Erdős, G. & Dosztányi, Z., 2018. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic acids research*, 46(W1), pp.W329-W337.

Meyer, E. H., Giegé, P., Gelhaye, E., Rayapuram, N., Ahuja, U., Thöny-Meyer, L., Grienenberger, J.-M. & Bonnard, G. 2005. AtCCMH, an essential component of the c-type cytochrome maturation pathway in Arabidopsis mitochondria, interacts with apocytochrome c. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 16113-16118.

McCulloch, R., Cobbold, C.A., Figueiredo, L., Jackson, A., Morrison, L.J., Mugnier, M.R., Papavasiliou, N., Schnaufer, A. & Matthews, K., 2017. Emerging challenges in understanding trypanosome antigenic variation. *Emerging topics in life sciences*, 1(6), pp.585-592.

Mendez, D. L., Babbitt, S. E., King, J. D., D'alessandro, J., Watson, M. B., Blankenship, R. E., Mirica, L. M. & Kranz, R. G. 2017. Engineered holocytochrome c synthases that biosynthesize new cytochromes c. *Proceedings of the National Academy of Sciences,* 114**,** 2235-2240.

Mertens, E., De Jonckheere, J. & Van Schaftingen, E., 1993. Pyrophosphate-dependent phosphofructokinase from the amoeba Naegleria fowleri, an AMP-sensitive enzyme. *Biochemical journal*, 292(3), pp.797-803.

Michels, P.A., Bringaud, F., Herman, M. & Hannaert, V., 2006. Metabolic functions of glycosomes in trypanosomatids. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1763(12), pp.1463-1477.

Milanes, J.E., Suryadi, J., Abendroth, J., Van Voorhis, W.C., Barrett, K.F., Dranow, D.M., Phan, I.Q., Patrick, S.L., Rozema, S.D., Khalifa, M.M. & Golden, J.E., 2019. Enzymatic and structural

characterization of the Naegleria fowleri glucokinase. *Antimicrobial agents and chemotherapy*, 63(5), pp.e02410-18.
Miller, D. J. & Nicholas, D. 1984. 3, 3', 5, 5'-Tetramethylbenzidine/H2O2 staining is not specific for heme proteins separated by gel electrophoresis. *Analytical biochemistry*, 140, 577-580.
Miller, M.A., Pfeiffer, W. & Schwartz, T., 2010, November. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *In 2010 gateway computing environments workshop* (GCE) (pp. 1-8). Ieee.
Mitchell, A.L., Attwood, T.K., Babbitt, P.C., Blum, M., Bork, P., Bridge, A., Brown, S.D., Chang, H.Y., El-Gebali, S., Fraser, M.I. & Gough, J., 2018. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic acids research*, 47(D1), pp.D351-D360.
Mittra, B., Laranjeira-Silva, M. F., Miguel, D. C., De Menezes, J. P. B. & Andrews, N. W. 2017. The iron-dependent mitochondrial superoxide dismutase SODA promotes Leishmania virulence. Mizianty, M.J., Peng, Z. and Kurgan, L., 2013. MFDp2: Accurate predictor of disorder in proteins by fusion of disorder probabilities, content and profiles. Intrinsically Disordered Proteins, 1(1), p.e24428.*Journal of Biological Chemistry*, jbc. M116. 772624.
Müller, M., Mentel, M., van Hellemond, J.J., Henze, K., Woehle, C., Gould, S.B., Yu, R.Y., van der Giezen, M., Tielens, A.G. & Martin, W.F., 2012. Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. *Microbiol*. Mol. Biol. Rev., 76(2), pp.444-495.
Namy, O., Duchateau-Nguyen, G. & Rousset, J.P., 2002. Translational readthrough of the PDE2 stop codon modulates cAMP levels in Saccharomyces cerevisiae. *Molecular microbiology*, 43(3), pp.641-652.
Nawathean, P. and Maslov, D.A., 2000. The absence of genes for cytochrome c oxidase and reductase subunits in maxicircle kinetoplast DNA of the respiration-deficient plant trypanosomatid Phytomonas serpens. *Current genetics*, 38(2), pp.95-103.
Neuberger, G., Maurer-Stroh, S., Eisenhaber, B., Hartig, A. and Eisenhaber, F., 2003. Prediction of peroxisomal targeting signal 1 containing proteins from amino acid sequence. *Journal of molecular biology*, 328(3), pp.581-592.
Niemann, M., Wiese, S., Mani, J., Chanfon, A., Jackson, C., Meisinger, C., Warscheid, B. & Schneider, A., 2013. Mitochondrial outer membrane proteome of Trypanosoma brucei reveals novel factors required to maintain mitochondrial morphology. *Molecular & cellular proteomics*, 12(2), pp.515-528.
Nishimura, Y., Tanifuji, G., Kamikawa, R., Yabuki, A., Hashimoto, T. & Inagaki, Y., 2016. Mitochondrial genome of Palpitomonas bilix: derived genome structure and ancestral system for cytochrome c maturation. *Genome biology and evolution*, 8(10), pp.3090-3098.
Nogueiras, R., Habegger, K.M., Chaudhary, N., Finan, B., Banks, A.S., Dietrich, M.O., Horvath, T.L., Sinclair, D.A., Pfluger, P.T. & Tschöp, M.H., 2012. Sirtuin 1 and sirtuin 3: physiological modulators of metabolism. *Physiological reviews*, 92(3), pp.1479-1514.
Novikova, O.D. & Solovyeva, T.F., 2009. Nonspecific porins of the outer membrane of Gram-negative bacteria: Structure and functions. *Biochemistry (Moscow) Supplement Series A: Membrane and Cell Biology*, 3(1), pp.3-15.
Ogbadoyi, E.O., Robinson, D.R. and Gull, K., 2003. A high-order trans-membrane structural linkage is responsible for mitochondrial genome positioning and segregation by flagellar basal bodies in trypanosomes. *Molecular biology of the cell*, 14(5), pp.1769-1779.
Ohta, N., Sato, N. & Kuroiwa, T. 1998. Structure and organization of the mitochondrial genome of the unicellular red alga Cyanidioschyzon merolae deduced from the complete nucleotide sequence. *Nucleic acids research*, 26, 5190-5198.
Opperdoes, F.R. and Borst, P., 1977. Localization of nine glycolytic enzymes in a microbody-like organelle in Trypanosoma brucei: the glycosome. *FEBS letters*, 80(2), pp.360-364.
Opperdoes, F.R., De Jonckheere, J.F. & Tielens, A.G., 2011. Naegleria gruberi metabolism. *International journal for parasitology*, 41(9), pp.915-924.

Oyelade, J., Isewon, I., Rotimi, S. & Okunoren, I., 2016. Modeling of the glycolysis pathway in plasmodium falciparum using petri nets. *Bioinformatics and Biology insights*, 10, pp.BBI-S37296.
Parada, L., 2010. Kinetoplastids and their networks of interlocked DNA. *Nature Education*, 3(9), p.63.
Parfrey, L.W., Lahr, D.J., Knoll, A.H. & Katz, L.A., 2011. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proceedings of the National Academy of Sciences*, 108(33), pp.13624-13629.
Parsons, M., 2004. Glycosomes: parasites and the divergence of peroxisomal purpose. *Molecular microbiology*, 53(3), pp.717-724.
Parthasarathy, M.V., Van Slobbe, W.G. and Soudant, C., 1976. Trypanosomatid flagellate in the phloem of diseased coconut palms. *Science*, 192(4246), pp.1346-1348.
Paulick, M.G. & Bertozzi, C.R., 2008. The glycosylphosphatidylinositol anchor: a complex membrane-anchoring structure for proteins. *Biochemistry*, 47(27), pp.6991-7000.
Pays, E. and Steinert, M., 1988. Control of antigen gene expression in African trypanosomes. *Annual review of genetics*, 22(1), pp.107-126.
Pettigrew, G. W. & Moore, G. 2012. *Cytochromes c: biological aspects*, Springer Science & Business Media.
Pettigrew, G.W., Leaver, J.L., Meyer, T.E. & Ryle, A.P., 1975. Purification, properties and amino acid sequence of atypical cytochrome c from two protozoa, Euglena gracilis and Crithidia oncopelti. *Biochemical Journal*, 147(2), pp.291-302.
Philippot, L., 2002. Denitrifying genes in bacterial and archaeal genomes. *Biochimica et biophysica acta (BBA)-Gene structure and expression*, 1577(3), pp.355-376.
Pineda, E., Thonnus, M., Mazet, M., Mourier, A., Cahoreau, E., Kulyk, H., Dupuy, J.W., Biran, M., Masante, C., Allmann, S. & Rivière, L., 2018. Glycerol supports growth of the Trypanosoma brucei bloodstream forms in the absence of glucose: Analysis of metabolic adaptations on glycerol-rich conditions. *PLoS pathogens*, 14(11), p.e1007412.
Piovesan, A., Caracausi, M., Ricci, M., Strippoli, P., Vitale, L. & Pelleri, M.C., 2015. Identification of minimal eukaryotic introns through GeneBase, a user-friendly tool for parsing the NCBI Gene databank. *DNA Research*, 22(6), pp.495-503.
Poirier, Y., Antonenkov, V.D., Glumoff, T. and Hiltunen, J.K., 2006. Peroxisomal β-oxidation—a metabolic pathway with multiple functions. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1763(12), pp.1413-1426.
Polevoda, B. & Sherman, F., 2002. The diversity of acetylated proteins. *Genome biology*, 3(5), pp.reviews0006-1.
Ponka, P., 1997. Tissue-specific regulation of iron metabolism and heme synthesis: distinct control mechanisms in erythroid cells. *Blood*, 89(1), pp.1-25.
Porcel, B.M., Denoeud, F., Opperdoes, F., Noel, B., Madoui, M.A., Hammarton, T.C., Field, M.C., Da Silva, C., Couloux, A., Poulain, J. and Katinka, M., 2014. The streamlined genome of Phytomonas spp. relative to human pathogenic kinetoplastids reveals a parasite tailored for plants. *PLoS genetics*, 10(2), p.e1004007.
Price, D.C., Chan, C.X., Yoon, H.S., Yang, E.C., Qiu, H., Weber, A.P., Schwacke, R., Gross, J., Blouin, N.A., Lane, C. & Reyes-Prieto, A., 2012. Cyanophora paradoxa genome elucidates origin of photosynthesis in algae and plants. *Science*, 335(6070), pp.843-847.
Priest, J.W. & Hajduk, S.L., 1992. Cytochrome c reductase purified from Crithidia fasciculata contains an atypical cytochrome c1. *Journal of Biological Chemistry*, 267(28), pp.20188-20195.
Qiu, Y., Kajihara, M., Harima, H., Hang'ombe, B.M., Nakao, R., Hayashida, K., Mori-Kajihara, A., Changula, K., Eto, Y., Ndebe, J. and Yoshida, R., 2019. Molecular characterization and phylogenetic analysis of Trypanosoma spp. detected from striped leaf-nosed bats (Hipposideros vittatus) in Zambia. *International Journal for Parasitology: Parasites and Wildlife*, 9, pp.234-238.

Radwanska, M., Vereecke, N., Deleeuw, V., Pinto, J. & Magez, S., 2018. Salivarian Trypanosomosis: A Review of Parasites Involved, Their Global Distribution and Their Interaction With the Innate and Adaptive Mammalian Host Immune System. *Frontiers in immunology*, 9.

Rayapuram, N., Hagenmuller, J., Grienenberger, J.M., Bonnard, G. and Giegé, P., 2008. The three mitochondrial encoded CcmF proteins form a complex that interacts with CCMH and c-type apocytochromes in Arabidopsis. *Journal of Biological Chemistry*, 283(37), pp.25200-25208.

Rayapuram, N., Hagenmuller, J., Grienenberger, J.-M., Giegé, P. & Bonnard, G. 2007. AtCCMA interacts with AtCcmB to form a novel mitochondrial ABC transporter involved in cytochrome c maturation in Arabidopsis. *Journal of Biological Chemistry*, 282, 21015-21023.

Reithinger, R., Dujardin, J.C., Louzir, H., Pirmez, C., Alexander, B. and Brooker, S., 2007. Cutaneous leishmaniasis. *The Lancet infectious diseases*, 7(9), pp.581-596.

Reumann, S., 2004. Specification of the peroxisome targeting signals type 1 and type 2 of plant peroxisomes by bioinformatics analyses. *Plant physiology*, 135(2), pp.783-800.

Reyes-Sosa, F.M., Gil-Martínez, J. and Molina-Heredia, F.P., 2011. Cytochrome c 6-like protein as a putative donor of electrons to photosystem I in the cyanobacterium Nostoc sp. PCC 7119. *Photosynthesis research*, 110(1), p.61.

Riccio, A., Vitagliano, L., Di Prisco, G., Zagari, A. & Mazzarella, L., 2002. The crystal structure of a tetrameric hemoglobin in a partial hemichrome state. *Proceedings of the National Academy of Sciences*, 99(15), pp.9801-9806.

Rico, E., Rojas, F., Mony, B.M., Szoor, B., MacGregor, P. and Matthews, K.R., 2013. Bloodstream form pre-adaptation to the tsetse fly in Trypanosoma brucei. *Frontiers in cellular and infection microbiology*, 3, p.78.

Rieske, J. 1976. Composition, structure, and function of complex III of the respiratory chain. *Biochimica et Biophysica Acta (BBA)-Reviews on Bioenergetics,* 456**,** 195-247.

Roditi, I. and Liniger, M., 2002. Dressed for success: the surface coats of insect-borne protozoan parasites. *Trends in microbiology*, 10(3), pp.128-134.

Rodríguez-Serrano, M., Romero-Puertas, M.C., Pastori, G.M., Corpas, F.J., Sandalio, L.M., del Río, L.A. & Palma, J.M., 2007. Peroxisomal membrane manganese superoxide dismutase: characterization of the isozyme from watermelon (Citrullus lanatus Schrad.) cotyledons. *Journal of experimental botany*, 58(10), pp.2417-2427.

Ruberto, I. 2011, *Functional dissection of T. brucei Protein Tyrosine Phosphatase 1 and investigation of its development as a therapeutic target*, University of Edinburgh.

Sacks, D. & Noben-Trauth, N., 2002. The immunology of susceptibility and resistance to Leishmania major in mice. *Nature Reviews Immunology*, 2(11), p.845.

Saint-Marcoux, D., Wollman, F.A. and de Vitry, C., 2009. Biogenesis of cytochrome b6 in photosynthetic membranes. *The Journal of cell biology*, 185(7), pp.1195-1207.

Sambrook, J., Fritsch, E.F. & Maniatis, T., 1989. Molecular cloning: a laboratory manual (No. Ed. 2). Cold spring harbor laboratory press.

San Francisco, B., Sutherland, M. C. & Kranz, R. G. 2014. The CcmFH complex is the system I holocytochrome c synthetase: engineering cytochrome c maturation independent of CcmABCDE. *Molecular microbiology,* 91**,** 996-1008.

Sanchez, L.B. & Müller, M., 1996. Purification and characterization of the acetate forming enzyme, acetyl-CoA synthetase (ADP-forming) from the amitochondriate protist, Giardia lamblia. *FEBS letters*, 378(3), pp.240-244.

Sánchez, L.B., Galperin, M.Y. and Müller, M., 2000. Acetyl-CoA Synthetase from the Amitochondriate EukaryoteGiardia lamblia Belongs to the Newly Recognized Superfamily of Acyl-CoA Synthetases (Nucleoside Diphosphate-forming). *Journal of Biological Chemistry*, 275(8), pp.5794-5803.

Sanchez-Moreno, M., Lasztity, D., Coppens, I. and Opperdoes, F.R., 1992. Characterization of carbohydrate metabolism and demonstration of glycosomes in a Phytomonas sp. isolated from Euphorbia characias. *Molecular and biochemical parasitology*, 54(2), pp.185-199.

Sánchez-Pascuala, A., de Lorenzo, V. & Nikel, P.I., 2017. Refactoring the Embden–Meyerhof–Parnas pathway as a whole of portable GlucoBricks for implantation of glycolytic modules in gram-negative bacteria. *ACS synthetic biology*, 6(5), pp.793-805.

Sanders, C., Turkarslan, S., Lee, D.-W. & Daldal, F. 2010. Cytochrome c biogenesis: the Ccm system. *Trends in microbiology,* 18**,** 266-274.

Sanders, C., Turkarslan, S., Lee, D.W., Onder, O., Kranz, R.G. and Daldal, F., 2008. The cytochrome c maturation components CcmF, CcmH, and CcmI form a membrane-integral multisubunit heme ligation complex. *Journal of Biological Chemistry*, 283(44), pp.29715-29722

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M. & Federhen, S. 2011. Database resources of the national center for biotechnology information. *Nucleic acids research,* 39**,** D38-D51.

Schell-Steven, A., Stein, K., Amoros, M., Landgraf, C., Volkmer-Engert, R., Rottensteiner, H. and Erdmann, R., 2005. Identification of a novel, intraperoxisomal pex14-binding site in pex13: association of pex13 with the docking complex is essential for peroxisomal matrix protein import. *Molecular and cellular biology*, 25(8), pp.3007-3018.

Schneider, R.E., Brown, M.T., Shiflett, A.M., Dyall, S.D., Hayes, R.D., Xie, Y., Loo, J.A. and Johnson, P.J., 2011. The Trichomonas vaginalis hydrogenosome proteome is highly reduced relative to mitochondria, yet complex compared with mitosomes. *International journal for parasitology*, 41(13-14), pp.1421-1434.

Schrader, M. & Fahimi, H.D., 2006. Peroxisomes and oxidative stress. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1763(12), pp.1755-1766.

Schuster, F.L. & Visvesvara, G.S., 2004. Free-living amoebae as opportunistic and non-opportunistic pathogens of humans and animals. *International journal for parasitology*, 34(9), pp.1001-1027.

Schwender, J., Ohlrogge, J. & Shachar-Hill, Y., 2004. Understanding flux in plant metabolic networks. *Current opinion in plant biology*, 7(3), pp.309-317.

Schwer, B., Bunkenborg, J., Verdin, R.O., Andersen, J.S. & Verdin, E., 2006. Reversible lysine acetylation controls the activity of the mitochondrial enzyme acetyl-CoA synthetase 2. *Proceedings of the National Academy of Sciences*, 103(27), pp.10224-10229.

Scott, I., Anderson, K.A. & Hirschey, M.D., 2012. Mitochondrial protein acetylation regulates metabolism. *Essays in biochemistry*, 52, pp.23-35.

Soppa, J., 2010. Protein acetylation in archaea, bacteria, and eukaryotes. *Archaea*, 2010.

Sharma, S., Cavallaro, G. & Rosato, A., 2010. A systematic investigation of multiheme c-type cytochromes in prokaryotes. *JBIC Journal of Biological Inorganic Chemistry*, 15(4), pp.559-571.

Shevket, S.H., Gonzalez, D., Cartwright, J.L., Kleanthous, C., Ferguson, S.J., Redfield, C. & Mavridou, D.A., 2018. The CcmC–CcmE interaction during cytochrome c maturation by System I is driven by protein–protein and not protein–heme contacts. *Journal of Biological Chemistry*, 293(43), pp.16778-16790.

Shoun, H. & Tanimoto, T., 1991. Denitrification by the fungus Fusarium oxysporum and involvement of cytochrome P-450 in the respiratory nitrite reduction. *Journal of Biological Chemistry*, 266(17), pp.11078-11082.

Shoun, H., Fushinobu, S., Jiang, L., Kim, S.W. & Wakagi, T., 2012. Fungal denitrification and nitric oxide reductase cytochrome P450nor. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1593), pp.1186-1194.

Simpson, A.G., Stevens, J.R. and Lukeš, J., 2006. The evolution and diversity of kinetoplastid flagellates. *Trends in parasitology*, 22(4), pp.168-174.

Smith, J.J. and Aitchison, J.D., 2009. Regulation of peroxisome dynamics. *Current opinion in cell biology*, 21(1), pp.119-126.

Smith, T.K., Bringaud, F., Nolan, D.P. & Figueiredo, L.M., 2017. Metabolic reprogramming during the Trypanosoma brucei life cycle. *F1000Research*, 6.

Spielewoy, N., Schulz, H., Grienenberger, J. M., Thöny-Meyer, L. & Bonnard, G. 2001. CCME, a nuclear-encoded heme-binding protein involved in cytochrome c maturation in plant mitochondria. *Journal of Biological Chemistry*, 276, 5491-5497.

Stairs, C.W., Eme, L., Muñoz-Gómez, S.A., Cohen, A., Dellaire, G., Shepherd, J.N., Fawcett, J.P. and Roger, A.J., 2018. Microbial eukaryotes have adapted to hypoxia by horizontal acquisitions of a gene involved in rhodoquinone biosynthesis. *Elife*, 7, p.e34292.

Stairs, C.W., Leger, M.M. & Roger, A.J., 2015. Diversity and origins of anaerobic metabolism in mitochondria and related organelles. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678), p.20140326.

Stairs, C.W., Roger, A.J. & Hampl, V., 2011. Eukaryotic pyruvate formate lyase and its activating enzyme were acquired laterally from a firmicute. *Molecular biology and evolution*, 28(7), pp.2087-2099.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), pp.1312-1313.

Stamatakis, A., Hoover, P., Rougemont, J. & Renner, S. 2008. A rapid bootstrap algorithm for the RAxML web servers. *Systematic biology*, 57, 758-771.

Starai, V.J. & Escalante-Semerena, J.C., 2004. Identification of the protein acetyltransferase (Pat) enzyme that acetylates acetyl-CoA synthetase in Salmonella enterica. *Journal of molecular biology*, 340(5), pp.1005-1012.

Starai, V.J. and Escalante-Semerena, J.C., 2004. Acetyl-coenzyme A synthetase (AMP forming). *Cellular and Molecular Life Sciences CMLS*, 61(16), pp.2020-2030.

Stevens, J.M., Daltrop, O., Higham, C.W. & Ferguson, S.J., 2003. Interaction of heme with variants of the heme chaperone CcmE carrying active site mutations and a cleavable N-terminal His tag. *Journal of Biological Chemistry*, 278(23), pp.20500-20506.

Stevens, J.M., Mavridou, D.A., Hamer, R., Kritsiligkou, P., Goddard, A.D. & Ferguson, S.J., 2011. Cytochrome c biogenesis System I. *The FEBS journal*, 278(22), pp.4170-4178.

Stiebler, A.C., Freitag, J., Schink, K.O., Stehlik, T., Tillmann, B.A., Ast, J. & Boelker, M., 2014. Ribosomal readthrough at a short UGA stop codon context triggers dual localization of metabolic enzymes in fungi and animals. *PLoS genetics*, 10(10), p.e1004685.

Suga, H., Chen, Z., De Mendoza, A., Sebé-Pedrós, A., Brown, M.W., Kramer, E., Carr, M., Kerner, P., Vervoort, M., Sánchez-Pons, N. & Torruella, G., 2013. The Capsaspora genome reveals a complex unicellular prehistory of animals. *Nature communications*, 4, p.2325.

Sunter, J. & Gull, K., 2017. Shape, form, function and Leishmania pathogenicity: from textbook descriptions to biological understanding. *Open biology*, 7(9), p.170165.

Swenson, S.A., Moore, C.M., Marcero, J.R., Medlock, A.E., Reddi, A.R. and Khalimonchuk, O., 2020. From Synthesis to Utilization: The Ins and Outs of Mitochondrial Heme. *Cells*, 9(3), p.579.

Tanifuji, G., Cenci, U., Moog, D., Dean, S., Nakayama, T., David, V., Fiala, I., Curtis, B.A., Sibbald, S.J., Onodera, N.T. & Colp, M., 2017. Genome sequencing reveals metabolic and cellular interdependence in an amoeba-kinetoplastid symbiosis. *Scientific reports*, 7(1), p.11688.

Tetaud, E., Lecuix, I., Sheldrake, T., Baltz, T. and Fairlamb, A.H., 2002. A new expression vector for Crithidia fasciculata and Leishmania. *Molecular and biochemical parasitology*, 120(2), pp.195-204.

Teusink, B., Walsh, M.C., van Dam, K. & Westerhoff, H.V., 1998. The danger of metabolic pathways with turbo design. *Trends in biochemical sciences*, 23(5), pp.162-169.

Thao, S. & Escalante-Semerena, J.C., 2012. A positive selection approach identifies residues important for folding of Salmonella enterica Pat, an Nε-lysine acetyltransferase that regulates central metabolism enzymes. *Research in microbiology*, 163(6-7), pp.427-435.

Thomas, P.E., Ryan, D. & Levin, W., 1976. An improved staining procedure for the detection of the peroxidase activity of cytochrome P-450 on sodium dodecyl sulfate polyacrylamide gels. *Analytical biochemistry*, 75(1), pp.168-176.

Thöny-Meyer, L., 2000. Haem-polypeptide interactions during cytochrome c maturation. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1459(2-3), pp.316-324.

Thöny-Meyer, L., Fischer, F., Künzler, P., Ritz, D. & Hennecke, H. 1995. Escherichia coli genes required for cytochrome c maturation. *Journal of bacteriology,* 177**,** 4321-4326.

Tilbrook, K., Poirier, Y., Gebbie, L., Schenk, P.M., McQualter, R.B. & Brumbley, S.M., 2014. Reduced peroxisomal citrate synthase activity increases substrate availability for polyhydroxyalkanoate biosynthesis in plant peroxisomes. *Plant biotechnology journal*, 12(8), pp.1044-1052.

Torres-Guerrero, E., Quintanilla-Cedillo, M.R., Ruiz-Esmenjaud, J. and Arenas, R., 2017. Leishmaniasis: a review. *F1000Research*, 6.

Torruella, G., De Mendoza, A., Grau-Bové, X., Antó, M., Chaplin, M., Del Campo, J., Eme, L., Pérez-Cordón, G., Whipps, C., Nichols, K., Paley, R., Roger, A., Sitjà-Bobadilla, A., Donachie, S. & Ruiz-Trillo, I. 2015, "Phylogenomics Reveals Convergent Evolution of Lifestyles in Close Relatives of Animals and Fungi", *Current Biology*, vol. 25, no. 18, pp. 2404-2410.

Tsaousis, A.D., Nývltova, E., Šuták, R., Hrdý, I. & Tachezy, J., 2014. A nonmitochondrial hydrogen production in Naegleria gruberi. *Genome biology and evolution*, 6(4), pp.792-799.

Turmel, M., Otis, C. & Lemieux, C., 2003. The mitochondrial genome of Chara vulgaris: insights into the mitochondrial DNA architecture of the last common ancestor of green algae and land plants. *The Plant Cell*, 15(8), pp.1888-1903.

Van Assche, T., Deschacht, M., da Luz, R.A.I., Maes, L. & Cos, P., 2011. Leishmania–macrophage interactions: Insights into the redox biology. *Free Radical Biology and Medicine*, 51(2), pp.337-351.

Van der Giezen, M., 2011. Mitochondria and the rise of eukaryotes. *Bioscience*, 61(8), pp.594-601.

Van Weelden, S.W., Fast, B., Vogt, A., Van Der Meer, P., Saas, J., Van Hellemond, J.J., Tielens, A.G. & Boshart, M., 2003. Procyclic Trypanosoma brucei do not use Krebs cycle activity for energy generation. *Journal of Biological Chemistry*, 278(15), pp.12854-12863.

Vassella, E., Oberle, M., Urwyler, S., Renggli, C.K., Studer, E., Hemphill, A., Fragoso, C., Bütikofer, P., Brun, R. and Roditi, I., 2009. Major surface glycoproteins of insect forms of Trypanosoma brucei are not essential for cyclical transmission by tsetse. *PloS one*, 4(2).

Veiga-da-Cunha, M., Sokolova, T., Opperdoes, F. and Van Schaftingen, E., 2009. Evolution of vertebrate glucokinase regulatory protein from a bacterial N-acetylmuramate 6-phosphate etherase. *Biochemical Journal*, 423(3), pp.323-332.

Velappan, N., Sblattero, D., Chasteen, L., Pavlik, P. & Bradbury, A. R. 2007. Plasmid incompatibility: more compatible than previously thought? *Protein Engineering, Design and Selection*, 20, 309-313.

Verissimo, A. F. & Daldal, F. 2014. Cytochrome c biogenesis System I: An intricate process catalyzed by a maturase supercomplex? *Biochimica et Biophysica Acta (BBA)-Bioenergetics,* 1837**,** 989-998.

Verissimo, A.F., Khalfaoui-Hassani, B., Hwang, J., Steimle, S., Selamoglu, N., Sanders, C., Khatchikian, C.E. and Daldal, F., 2017. The thioreduction component CcmG confers efficiency and the heme ligation component CcmH ensures stereo-specificity during cytochrome c maturation. *Journal of Biological Chemistry*, 292(32), pp.13154-13167.

Verissimo, A.F., Shroff, N.P. and Daldal, F., 2015. During cytochrome c maturation CcmI chaperones the class I apocytochromes until the formation of their b-type cytochrome intermediates. *Journal of Biological Chemistry*, 290(27), pp.16989-17003

Vickerman, K. and Preston, T.M., 1976. Comparative cell biology of the kinetoplastid flagellates. *Biology of Kinetoplastida*, 1, pp.66-67.

Vickerman, K., 1985. Developmental cycles and biology of pathogenic trypanosomes. *British medical bulletin*, 41(2), pp.105-114.

Walker, D.M., Oghumu, S., Gupta, G., McGwire, B.S., Drew, M.E. and Satoskar, A.R., 2014. Mechanisms of cellular invasion by intracellular parasites. *Cellular and molecular life sciences*, 71(7), pp.1245-1263.

Wanders, R.J., Waterham, H.R. & Ferdinandusse, S., 2016. Metabolic interplay between peroxisomes and other subcellular organelles including mitochondria and the endoplasmic reticulum. *Frontiers in cell and developmental biology*, 3, p.83.

Wang, B.B. & Brendel, V., 2006. Genomewide comparative analysis of alternative splicing in plants. *Proceedings of the National Academy of Sciences*, 103(18), pp.7175-7180.

Wang, J., Wang, Y., Gao, C., Jiang, L. & Guo, D., 2017. PPero, a computational model for plant PTS1 type peroxisomal protein prediction. *PloS one*, 12(1), p.e0168912.

Wang, Q., Zhang, Y., Yang, C., Xiong, H., Lin, Y., Yao, J., Li, H., Xie, L., Zhao, W., Yao, Y. & Ning, Z.B., 2010. Acetylation of metabolic enzymes coordinates carbon source utilization and metabolic flux. *Science*, 327(5968), pp.1004-1007.

Wargnies, M., Bertiaux, E., Cahoreau, E., Ziebart, N., Crouzols, A., Morand, P., Biran, M., Allmann, S., Hubert, J., Villafraz, O. & Millerioux, Y., 2018. Gluconeogenesis is essential for trypanosome development in the tsetse fly vector. *PLoS pathogens*, 14(12), p.e1007502.

Weisse, R.H.J., Faust, A., Schmidt, M., Schönheit, P. & Scheidig, A.J., 2016. Structure of NDP-forming Acetyl-CoA synthetase ACD1 reveals a large rearrangement for phosphoryl transfer. *Proceedings of the National Academy of Sciences*, 113(5), pp.E519-E528.

Wellings, F.M., 1979. *Pathogenic Naegleria: distribution in nature* (Vol. 1). Environmental Protection Agency, Office of Research and Development,[Office of Health and Ecological Effects], Health Effects Research Laboratory.

Wiederschain, G.Y., 2009. Essentials of glycobiology. *Biochemistry (Moscow)*, 74(9), pp.1056-1056.

Wilkinson, S.R., Prathalingam, S.R., Taylor, M.C., Ahmed, A., Horn, D. & Kelly, J.M., 2006. Functional characterisation of the iron superoxide dismutase gene repertoire in Trypanosoma brucei. *Free Radical Biology and Medicine*, 40(2), pp.198-209.

Williams, C., Aksam, E.B., Gunkel, K., Veenhuis, M. & Van Der Klei, I.J., 2012. The relevance of the non-canonical PTS1 of peroxisomal catalase. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1823(7), pp.1133-1141.

Woehle, C., Roy, A.S., Glock, N., Wein, T., Weissenbach, J., Rosenstiel, P., Hiebenthal, C., Michels, J., Schönfeld, J. and Dagan, T., 2018. A novel eukaryotic denitrification pathway in foraminifera. *Current Biology*, 28(16), pp.2536-2543.

Xie, Z., Culler, D., Dreyfuss, B.W., Kuras, R., Wollman, F.A., Girard-Bascou, J. and Merchant, S., 1998. Genetic analysis of chloroplast c-type cytochrome assembly in Chlamydomonas reinhardtii: one chloroplast locus and at least four nuclear loci are required for heme attachment. *Genetics*, 148(2), pp.681-692.

Xu, C., Park, J.K. & Zhang, J., 2019. Evidence that alternative transcriptional initiation is largely nonadaptive. *PLoS biology*, 17(3), p.e3000197.

Xu, G. & Zhang, J., 2014. Human coding RNA editing is generally nonadaptive. Proceedings of the *National Academy of Sciences*, 111(10), pp.3769-3774.

Yang, E. C., Boo, S. M., Bhattacharya, D., Saunders, G. W., Knoll, A. H., Fredericq, S., Graf, L. & Yoon, H. S. 2016. Divergence time estimates and the evolution of major lineages in the florideophyte red algae. *Scientific reports*, 6.

Yabuki, A., Gyaltshen, Y., Heiss, A.A., Fujikura, K. and Kim, E., 2018. Ophirina amphinema n. gen., n. sp., a New Deeply Branching Discobid with Phylogenetic Affinity to Jakobids. *Scientific reports*, 8(1), p.16219.

Yang, J., Harding, T., Kamikawa, R., Simpson, A.G. & Roger, A.J., 2017. Mitochondrial genome evolution and a novel rna editing system in deep-branching heteroloboseids. *Genome biology and evolution*, 9(5), pp.1161-1174.

Young, D.G. and Lawyer, P.G., 1987. New World vectors of the leishmaniases. In Current topics in vector research (pp. 29-71). *Springer*, New York, NY.

Yu, J. and Le Brun, N.E., 1998. Studies of the Cytochrome Subunits of Menaquinone: Cytochromec Reductase (bc Complex) of Bacillus subtilis EVIDENCE FOR THE COVALENT ATTACHMENT OF HEME TO THE CYTOCHROMEb SUBUNIT. *Journal of Biological Chemistry*, 273(15), pp.8860-8866.

Zhao, Y., Wang, Z.B. & Xu, J.X., 2003. Effect of Cytochrome c on the Generation and Elimination of O and H2O2 in Mitochondria. *Journal of Biological Chemistry*, 278(4), pp.2356-2360.

Zhou, X., Shen, X.X., Hittinger, C.T. and Rokas, A., 2018. Evaluating fast maximum likelihood-based phylogenetic programs using empirical phylogenomic data sets. *Molecular biology and evolution*, 35(2), pp.486-503.

Appendices

Appendix 1

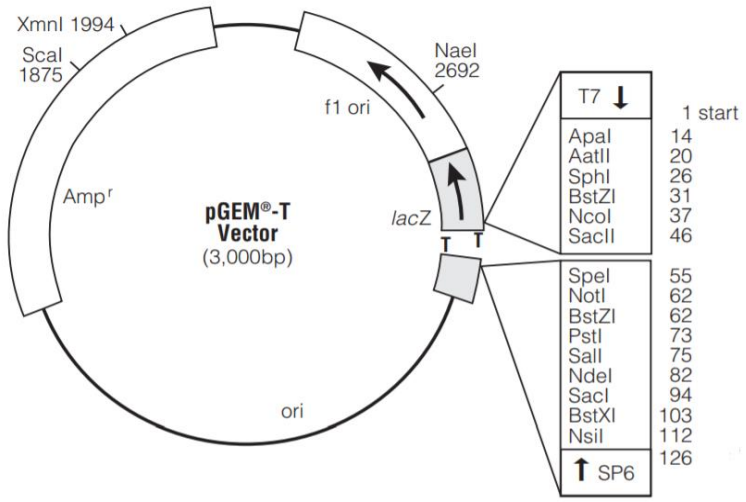


Figure 9.1. The pGEM®-T vector map and restriction enzyme reference points.

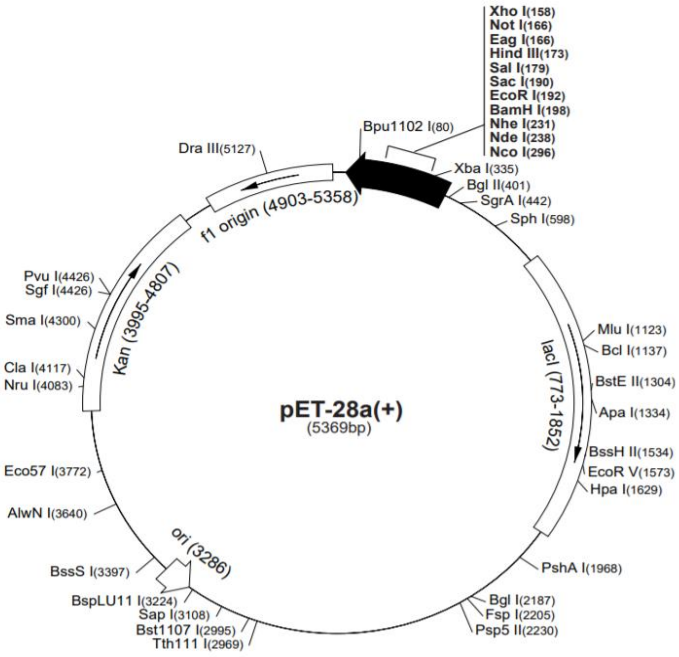


Figure 9.2. The pET-28a-c(+) vector map and restriction enzyme reference points.

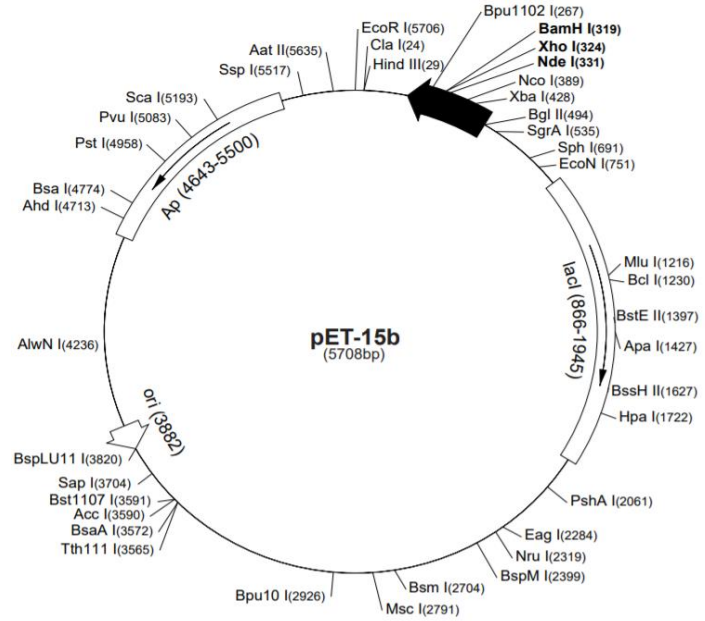


Figure 9.3. The pET-15b vector map and restriction enzyme reference points.

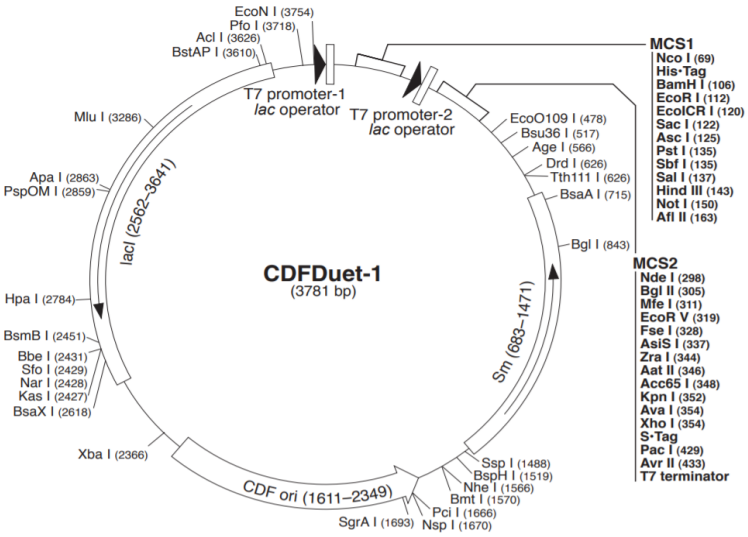


Figure 9.4. The pCDFDuet-1 vector map and restriction enzyme reference points.

Appendix 2

Table 9.1 List of the primers used in the PCR reaction to amplify Trypanosome cytochrome *c* and putative holocytochorme *c* synthetase.			
Primer ID	**Nucleotide sequence (5' -> 3')**	**Restriction site**	**Cloning Plasmid**
Cyto *c* Forward	TTGGATCC**GCATGCCACCAAAGGAGCGTGC**	*BamHI*	*pET-15b (His-tag)*
Cyto *c* Reverse	GCCATATG**TTAGTCCTTTAATGTCTCGAGG**	*NdeI*	
Cyto *c* Forward	TTGAATTC**GCATGCCACCAAAGGAGCGTGC**	*EcoRI*	*pCDFDuet-1(His-tag)*
Cyto *c* Reverse	GCAAGCTT**TTAGTCCTTTAATGTCTCGAGG**	*HindIII*	
Cyto *c* Forward	TTCCATGG**GCATGCCACCAAAGGAGCGTGC**	*NcoI*	*pCDFDuet-1(without His-tag)*
Cyto *c* Reverse	GCGGATCC**TTAGTCCTTTAATGTCTCGAGG**	*BamHI*	
*Tb*pHCCS Forward	CACCATATG**TGGGTGAGGACATTCCTGC**	*NdeI*	*pET-28a* and *pCDFDuet-1*
*Tb*pHCCS Reverse	AAGCTCGAG**TCACGGTGCCGCATGGCATTTTAC**	*XhoI*	

Table 9.2 List of the primers used in the PCR reaction to amplify system I proteins from *Capsaspora owczarzaki* genomic genomic DNA.	
Primer ID	**Nucleotide sequence (5' -> 3')**
CapCcmE Map1	AAAATGTCGATGCGGACTCTCC
CapCcmE Map2	GGTCGAGTGGTAATGCTCC
CapCcmE Map3	ATCTTGGGGTTGATGCTTACC
CapCcmE Map4	AATCACTCAGCCTTATCTGC
CapCcmH Map1	ATGCACTTTCTCAAGCACTTTGC
CapCcmH Map2	TCATTCACGAGCGAGAATCG
CapCcmH Map3	TGTGCTGCAGGAAAACAAAGAAGG
CapCcmH Map4	CGATTGTGCTTCATCAAGACG
CapCcmA Map1	ATGCGTAAGGACACGGCATTTTCG
CapCcmA Map2	AAACTCACTGCAATGGAGGAAGG
CapCcmA Map3	ATGGTAAAGTCTCTGATGTTTGTACC
CapCcmA Map4	ATACGCAATCCATTGCCTTGC
CapCcmC Map1	TTTATAATAATATACAGACC
CapCcmC Map2	CTAGGAAGAAGTAACACAGG
CapCcmF Map1	TGAGATGCAGTTGAAAATGC
CapCcmF Map2	CCTGGTATTATTTTAACTTTAACC

Table 9.3. Primers used for Cas9/Crispr to generate gene knockout of putative holocytochrome c synthase gene knockout in *Leishmania mexicana*.	
Primer ID	**Nucleotide sequence (5' -> 3')**
3' sgRNA primer	gaaattaatacgactcactataggAGCGGTACCACACGAGCGCGgttttagagctagaaatagc
5' sgRNA primer	gaaattaatacgactcactataggGGCGGTAATTGTGGCGGCAGgttttagagctagaaatagc
OL6137	aaaagcaccgactcggtgccacttttcaagttgataacggactagccttattttaacttgctatttctagctctaaaac
Forward primer for resistance cassettes	ACGTCGATTCGCACGACGTCCACAAGGAGAgtataatgcagacctgctgc
Reverse primer for resistance cassettes	CTTGGCCAGCGCTGCAGAAAGGGAAAGCGGccaatttgagagacctgtgc

Table 9.4 List of the primers used in the PCR reaction to T. brucei alternative oxidase (*Tb*AOX), *T. brucei* holocytochorme *c* synthetase (TbpHCCS) and *L. mexicana* holocytochorme *c* synthetase (LmpHCCS).			
Primer ID	**Nucleotide sequence (5' -> 3')**	**Restriction site**	**Cloning Plasmid**
*Tb*AOX Forward	AAGCCTAGG**ATGTTTCGTAACCACGCATCG**	*Avr*II	pGL2785
*Tb*AOX Reverse	TTCCCTAGG**TTACACGTGTTTGTTTACATTACTCG**	*Avr*II	
*Lm*pHCCS Forward	AAGCCTAGG**ATGGCGGGGGCGGCGTGGGCC**	*Avr*II	pGL2785
*Lm*pHCCS Reverse	TTCCCTAGG**CTACGAGTGAGGCGGTGCC**	*Avr*II	
*Tb*pHCCS Forward	AAGCCTAGG**ATGTGGGTGAGGACATTCC**	*Avr*II	pGL2785
*Tb*pHCCS Reverse	TTCCCTAGG**TCACGGTGCCGCATGGCATTTTACC**	*Avr*II	

Appendix 3

Below is a list of complementary oligonucleotides sequence of *Naegleria gruberi* metabolism enzymes with putative peroxisomal targeting:

Triosephosphate isomerase:

Forward primer TPI (XhoI/HindIII): TC GAG AAG CTT AGC AAG GAC TTC TTG TCA ATT GTT AAT TCT GGT TTG AAA AGT GTC GAA CGA CTC AAT TTA AGA CAA AGT AAA TTG TAA GAG CT

Reverse primer TPI (SacI): C TTA CAA TTT ACT TTG TCT TAA ATT GAG TCG TTC GAC ACT TTT CAA ACC AGA ATT AAC AAT TGA CAA GAA GTC CTT GCT AAG CTT C

Superoxide dismutase:

Forward primer SD (XhoI/SpeI): TC GAG ACT AGT TTC TGG AAT GTT GTC AAC TGG AAG TTT GCT GAA TCC CGT CTC TTA GAA CTC TCT GAT CAG CAT ATT TCA AAA TTA TAA GAG CT

Reverse primer SD (SacI): C TTA TAA TTT TGA AAT ATG CTG ATC AGA GAG TTC TAA GAG ACG GGA TTC AGC AAA CTT CCA GTT GAC AAC ATT CCA GAA ACT AGT C

Aldehyde reductase:

Forward primer AR (XhoI/XbaI): TC GAG TCT AGA GAA AAG AAC AAT CTC ACT AGA ACT TGT GAT CCA ATC AAT TTC TGG AGT CTT CCA TTT TTA GTT TCA AGA ATT TAA GAG CT

Reverse primer AR (SacI): C TTA AAT TCT TGA AAC TAA AAA TGG AAG ACT CCA GAA ATT GAT TGG ATC ACA AGT TCT AGT GAG ATT GTT CTT TTC TCT AGA C

Appendix 4

Table 9.5. The accession numbers of Ccm proteins from different lineages. Blue colour represents the nuclear encoded Ccm. Orange colour represents the mitochondrial encoded Ccm.

Organism	Product Annotation	GeneBank protein Accession No	Length
C. merolae	CcmA	NP_059366	197
C. merolae	CcmB	NP_059367	227
C. merolae	CcmC	NP_059382	234
C. merolae	CcmE	BAM80061	250
C. merolae	CcmF	NP_059365	620
C. merolae	CcmH	BAM80358	180
G. sulphuraria	CcmC	YP_009051209	185
G. sulphuraria	CcmE	XP_005706583	293
G. sulphuraria	CcmF	NP_059365	620
G. sulphuraria	CcmH	XP_005709016	190
R. americana	CcmA	NP_044792	222
R. americana	CcmB	NP_044791	224
R. americana	CcmC	NP_044790	262
R. americana	CcmF	NP_044789	637
N. gruberi	CcmA	XP_002683281	289
N. gruberi	CcmC	NP_066532	210
N. gruberi	CcmE	XP_002673590	309
N. gruberi	CcmF	NP_066539	474
M. californiana	CcmB	AAG13700	218
M. californiana	CcmC	AAG13701	234
M. californiana	CcmF	YP_009118114	628
T. thermophila	CcmA	XP_001462007	291
T. thermophila	CcmE	XP_001024126	195
T. thermophila	CcmF	NP_149387	518
T. thermophila	CcmH	XP_001014576	126
P .tetraurelia	CcmA	XP_001462007	242
P. tetraurelia	CcmE	XP_001437907	181
P. tetraurelia	CcmF	ACH80948	Partial
P. tetraurelia	CcmH	XP_001455766	104
A. thaliana	CcmA	NP_176516	229
A. thaliana	CcmB	AEK01274	206
A. thaliana	CcmC	CCMC_ARATH	232
A. thaliana	CcmE	NP_190747	256
A. thaliana	CcmFc CcmF1 CcmF2	CCMFC_ARATH/ CCMF1_ARATH/ CCMF2_ARATH	442 382 208
A. thaliana	CcmH	OAP13103	159
Capsaspora	CcmE	XP_011270861.1	368
Capsaspora	CcmF	AGE93635.1	848
Capsaspora	CcmH	XP_004347007.1	313
Capsaspora	CcmC	AGE93595.1	244
Capsaspora	CcmA	XP_004364551.1	195
F. alba	CcmE	XP_009496718.1	314
F. alba	CcmA	XP_009494751.1	238

Appendix 5

Below is the protein sequences of System I in *F.alba*, *P.atlantis* and *C.perkinsii* that were generate using local BLAST:

> *Fonticula alba* CcmC

MKEKNKSMFKKFIESLLLIIIISESYTLINRIKTKPIIDLTQGDLTIMFSVHVPLAIIVTICLIPISILYITKTTHKIKNISFLLLNIS
IITLISGIIWGQAIWGNFFIVDKKILDIILIICYYYIITYNNKLNIITIITITTLYYFIKISIGDINSIHQNIKTKSEFYNSIPSTYGD
MILNNITIATALGTLYLITT

> *Fonticula alba* CcmF

MTYSLQLLTLSIITLSPLFIKQNIKTLSSISILGLLINLIFYHDELTDLKKDNSIITFFNLIFNNFESIYLLYFIIIIISFIVRTHLIQE
LNDINSTNYKKELLILLIFIITYIGILQPIELYLFPLEFVSEEINQQLINPMMIFHPIILYIAYSYILVLSLSYLSLLKVSSVKRFF
YEHKRFLSHIVYLLGTSLMFSILWSWDINTLSWTWDPIEGLLMFIFILSVLILHFLYFYRNFNKFINIKNIIFIIFWISPIMS
LLLVRTGVLQSVHSFSESENYIFLIFVLLMFIIYIIQVSFMEYRSFNTIMAKSLLFSVTILLFLITLILFLVHALHPNMVGFKL
EPEDFESIVLLPLLIFTCFIMLIIPLLINSINNIKYISYIIILYISVYYTYNLSYIVYTSVSTIIISKLLKSKNFRIIHAMILGGLLYITI
ALCIKHEEEIAITVGQCLESEGYKLILYKLQTCKQEPYIYNIAQILYYNNKNTENIQLIIRNTENININSKSRIYNCILGHSDH
TIKLIINKNQQFLNIKQYYTPHIITIITIITLFKTIKKQHNNNININKEFIAQW

> *Parvularia atlantis* CcmE

APSALGIARPIAVRTAAGGAAVARVDRHAVHASADARRTRRFWTLGVSVVVAGAGVFGLLNAMGDTMMFYLTP
TQLLEKDPPMPSSKRFRLGGMVVPGSLRALVGAPGVLFTVTDYSHDVDVVFSGALPQLFKEGTSAVCEGRMRPNG
VFEATEVLAKHDENYMPSDIAEQVAKNRAPTSAAAAPAARA

> *Parvularia atlantis* CcmA
GMWLESICGVGLAFRRQHTLVLRDVHFAVRAGSALHVQGANGSGKTTLLKLIAGLIAPTNGGFVINGRAVPPGQQ
NEYMSGRTHYVAARTDGLDPHATVDFNLRFWNTLYRGTDEQLERAVERFDLRHLLSRPVSSLSTGQRRAISLARIS
AAYCQVWLLDEPVNALDRRMTAIVEDAIRAHRRAGG

> *Parvularia atlantis* CcmH
TDAQLPAPSHRTHSAMSDLTDFQAKYIDVRAATLYRRIRCVVCAGQTIEDSPTDFAVLMRTIVRQKLREGLDDEQV
LKFLADTYGPEVLYAGGLDTAVLVAAGVATAAAVGIG

> *Chromosphaera perkinsii* CcmE
MSGVFVTKRLLGGIQNFSRLSKSALSRSTAQITTKESLTIPSSSCSSTNQQQRSFSSSSRKLVMGNQEPVPYGREALT
QATMARRQKLQSARRLRKLGIMLGSLTGVGGVTYLALQYFEDSLQFYLTPTELLDNRKFKKFVENGTPLRLGGLVTE
GSVVKFQSVDPVRAEDPKVAAMVNKLTKKAGAGGVTFEVHDYLNKVDVEFYGGLPELFQEGKSVIAHGTLQKKSK
EKMRGEEGEAEYFMLADEILAKHDENYVPKEFKEKLKTNREMREMKLKAAEKEGRVAVL

> *Chromosphaera perkinsii* CcmH

MWRLLCSISSSKRIFTRCLHSTSSKSNRTLPIQASFGGGGGGGGGAGKRIFGGGAAVMAGGGVSGFVFDEKGVGK
GDSGNTNNNIISDDDAGEGEEKGSGGEKEKKPFDMERIRQKGADLSEEEREFRLRVAFVDARASDLNTQIRCLTCD
NNTIEDANTEIAILLREIVREEIWKGKTDQEIKDLLVEQFGEEVTFDAPESFRAIVYGVLPAVTGIAYFLIQTGIYKRSRP
HLYYAQLCRGFPSTHKELTAIKRLVALPPKNPQMYTAWMRFRETVLEKWLGLAPKGKLYPRWPSYGMRAPGGHR
AEKMVPYSIWEMPGPFSEKYIPPRTSSFRDVVR

> *Chromosphaera perkinsii* CcmA

MNRLCVRGGCKMGWEGLGRVLSEVSTGAAISGKKRHRRTGMGLGWVWSRGLCSEYKPENLTGLYARGRVSDSE
RLEMFDKANSINYLIRYDTPPWTPVLTARSIGYVVESDIIFENINFSLKPGGALLLQGPNGCGKSTLMNMILGQVKPT
VGAFTLNNIKIERPKRSPLRHTGFCQYLYAELLGMHPEFTVEENMGYYDAHMFASSDPSLALDALDVLHLRGRLFKF
CSMGQKKRCNLARMLHVAAPLWIMDEPTVGLDIPSIRILELMITHHRNRGGMVIVSSHQNLKLENATTLKFPGTFT
HKFIQ

Appendix 6

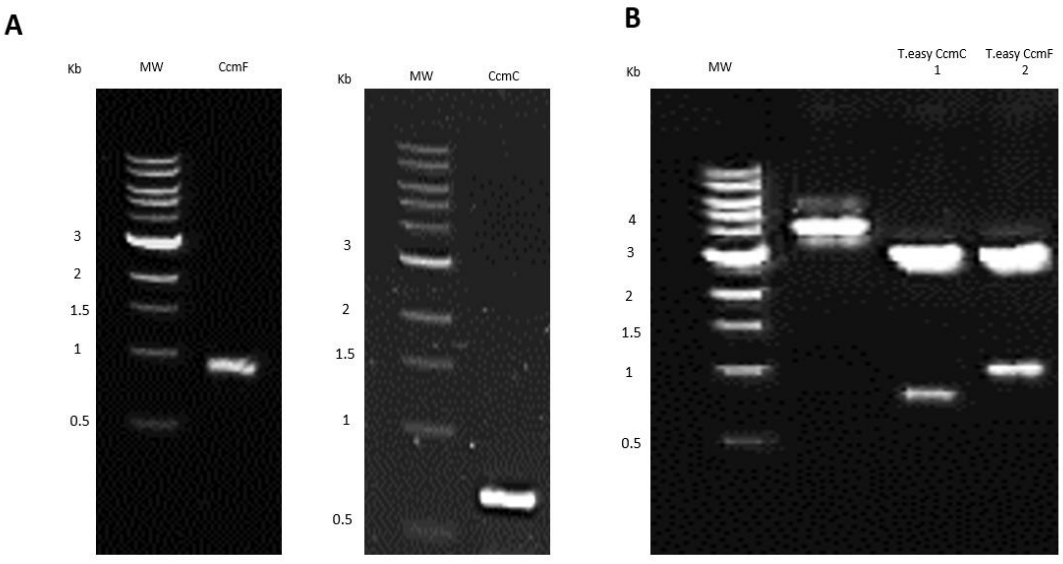


Figure 9.6. Cloning of *Capsaspora owczarzaki* mitochondrial Ccm genes into pGEM-T Easy. Image A reflects the PCR amplification of the Ccm genes. The size of CcmF gene is 0.977Kb and the size of CcmC is 0.706Kb. Image B shows a diagnostic restriction digest of ligated pGEM-T Easy. *Eco*RI restriction enzyme was used to separate Ccm genes from the pGEM-T Easy backbone. 1 shows the diagnostic digest of CcmC and 2 shows the diagnostic digest of CcmF.

Appendix 7

Sequence from SourceBioscience Sequencing Data of *Capsaspora owczarzaki* CcmC ligated to pGEM-T Easy:

NNNNNNNNNNNNNNNCTATAGATACTCAGCTATGCATCCAACGCGTTGGGAGCTCTCCCATATGGTCGACCTGCATGCGGCCGCGAATTCACTAGT
GATTTTTATAATAATATACAGACCATTTTAGAACTATTAATGATAATTTTAGGTACACTAAGTTTTACATTAATATGCTATAGTTTAGAAGATTTTCAACA
AGATATATATTGAAAAATAATTTTTATTCACGTTCCATTAGCCAGTTTATCATTAATGTTATATGTGTGTAAAGTCGCCTGTAGTTTATTGTACTTAATAA
CAAGTCGAGCTACTTTAGCTATGATAGCTAGAAATTTGGCTAAAATTGCATTAACATTACAAATTTTATGTATTATAACAGGGAGTTTATGAGGTAAGG
CATCATGAGGAACATACTGAGAATTTGATTCAAGATTGACATCAATGTTAATTATCACTATTATTACTGGATGCTATGTTATAATAAACCAAACTATAAA
AGATGAAACAACATCTATAAAAAAAGATAGAATATTAGCCATTATTTCAATTATAGGTAGTATTAATATACCAATAATTAAATTTAGTGTAGATTGATG
AGCAAGTTTACATCAAAAAAGTTCAATTAATATTTTTAATGTAGGTACAACTAAAAAGGATATTTCGATTTATTTGCCATTAATGATGGTTTTATTTGTAT
TTGTATTATGAATTATACACATGTTTAGCATATACATGATGATTGAACAATCGGAAAGAAAATTGCAAATTATACAAAAACAAATACCTGTGTTACCACT
TCTTCCTAGAATCGAATTCCCGCGGCCGCCATGGCGGCCGGGAGCATGCGACGTCGGGCCCAATTCGCCCTATAGTGAGTCGTATTACAATTCACTGG
CCNGTCGTTTTACAACGTCGTGACTGGGNAAAACCCTGGCAGTTACACAANTNAANCNGCCCTGGCANNANNTTCCCCGTTTNNNCAGCTG

Sequence b from SourceBioscience Sequencing Data *Capsaspora owczarzaki* CcmF ligated to pGEM-T Easy:

NNNNNNNNNNNNNNNCTATAGANACTCAGCTATGCATCCAACGCGTTGGGAGCTCTCCCATATGGTCGACCTGNNNGCGGCCGCGAATTCACTAG
TGATTATGTATTTTGTAATTATGAGTTTAATATCAATAATAACAATATGAAAAGTCATACATAAAAACATGGATTTAGATTTAGTAGAAAGATTATCGTC
TATTACACAGACAATTGAATATGCAATTAGTGGAATTTGAGGGAACCATGATGGGTCATTAATTTTGTTATTATTAATGTTAATTTTTTATCAAATCAAA
GGTAAAAATGAATCAAATATGCATGTTTTATCCACAACATTTTTATTATCTAATAATGAGCATTCTAATAATGAGCATTCTAATAATGAGCATTCTAATA
ATGAGCATTCTAATAATGAGCATTCTAATAATGAGCATTCTAATAATGAGCATTCTAATAATGAGCATTCTAATAATGAGCATTCTAATAATGAGCATTT
TAATAATAAGCATTCTAATAATAAGCATTCTAATAATAAGCATTCTAATAATAAGCATTCTAATAATGAGCATGTATCTAAAACATTAGTAACACCAAAA
TTTGTGATAAATGCGCCTAATAAAGCCAGAAAAGGCTTGAGAATAATAAGAAAGATTTTAAAAACAACCCCTTTACAAATATTAATAACAATATTTTTA
ATCTTTTCTATATTAGTAGCATTTGGTTTAAAACAAAGTACATTTGTAGAATTAAAACAATCAATTTTCATTCAAGATGAAACAATATTAAAAGAAATAA
TACAAGTTTTAAATCCTAATTTGGAAGATTTCTTCTTATTGATACACCCACCAATTTTATACTTCGGATATTTAAGCATATTAGCCATAATTCCATGGATT
ACAGATAAAAANGCATTAAAAATCAAATTATTGTTAGGTATGATATGATTAACTACTGGTATCAGTCTAGGTTCATGATGAGCCTACCATGAATTAGGA
TGAGGTAGTTATTGATATTGAGATGCAGTTGAAAATGCATCATTATGAGTTTGATTTTTTAATGTTTTATTTATACATATATTAATCGAATTCCCGCGGC
CGCCATGGCGGCCGGGAGCATGCGACGTCGGGCCCAATTCGCCCTATAGTGAGTCGTATTACAATTCACTGGCCGTCGTTTTACAACGTCNNGACTGG
GAAAACCCTGGCNTTACCCAACTTAATCGCCTTGCAGCNNATCCCCCTTTCGCCAGCTGGCGTAANANCNAAAANGGCCNNCACCGATNNCCCTTCCC
AACAGTTGCCNAGCCTGAATGGCAAATGGACCNNCCNTGTAACGGCCNATTA

Figure 9.7. DNA sequences of pGEM-T Easy ligated to *Capsaspora owczarzaki* Ccm genes from SourceBioscience Sequencing Data. Red colour indicates part of the open reading frame of Ccm genes, blue colour indicates the digest site of *Eco*RI enzyme and black is part of pGEM-T Easy backbone sequence.

Appendix 8

Figure 9.8. Pairwise nucleotide sequence alignment of the *C. owczarzaki* CcmC genes using MAFFT server. CcmC-T.easy is the sequence generated from the PCR and CcmC-Web-annota is the annotated sequence from NCBI.

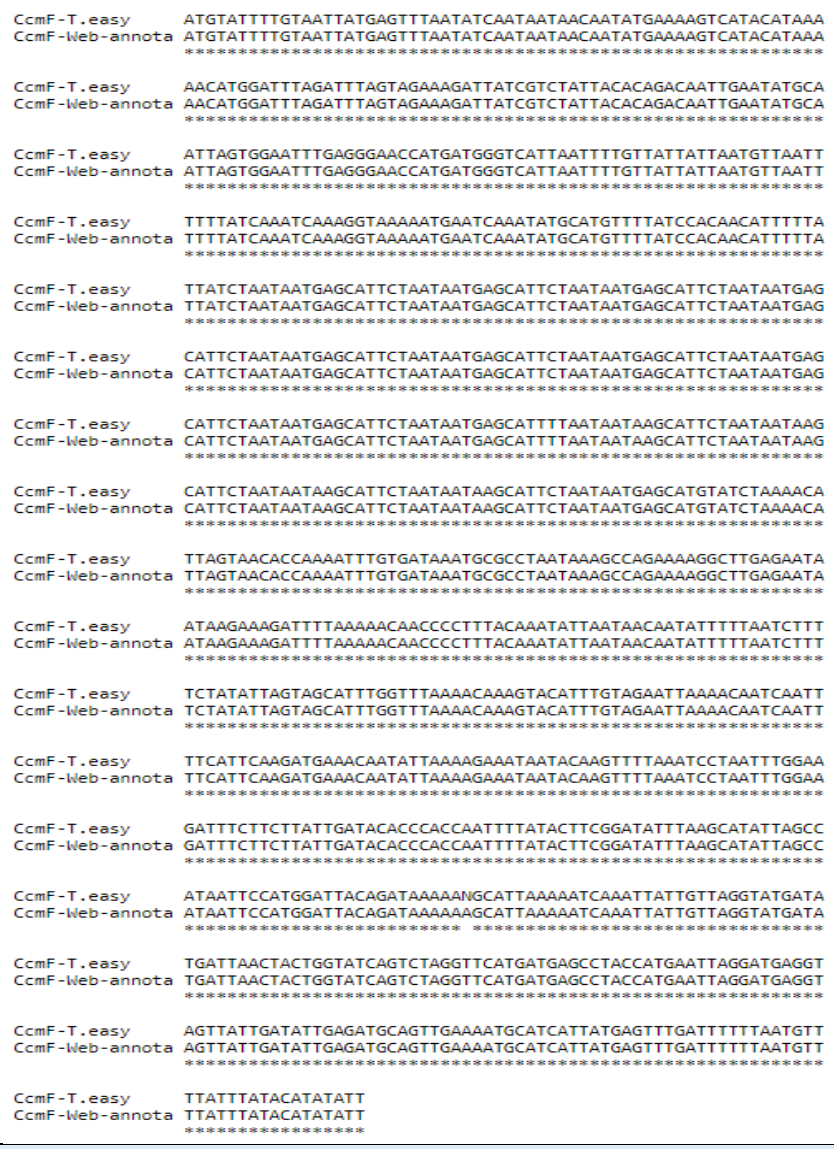


Figure 9.9. Pairwise nucleotide sequence alignment of the *C. owczarzaki* CcmF genes using MAFFT server. CcmF-T.easy is the sequence generated from the PCR and CcmF-Web-annota is the annotated sequence from NCBI.

Appendix 9

Table 9.6. Detecting the presence of fungi specific proteins in charophyte and primitive plants which found to use System III for cytochrome *c* maturation.

Species	Clade	Chitin synthase 2 CHS2	Chitin synthase regulatory factor 4
Cibotium_glaucum	Ferns	2e-03	2e-2
Thyrsopteris_elegans	Ferns	1e-03	-
Polypodium_hesperium	Ferns	-	-
Blechnum_spicant	Ferns	-	-
Thelypteris_acuminata	Ferns	-	-
Hymenophyllum_bivalve	Ferns	-	-
Ophioglossum_vulgatum	Ferns	9e-03	-
Megaceros_vincentianus	Hornwort	2e-03	2e-4
Nothoceros_aenigmaticus	Hornwort	1e-03	-
Megaceros_tosanus	Hornwort	-	-
Paraphymatoceros_hallii	Hornwort	-	-
Phaeoceros_carolinianus	Hornwort	-	-
Selaginella_stauntoniana	lycophyta	9e-03	-
Selaginella_stauntoniana	lycophyta	9e-03	-
Selaginella_lepidophylla	lycophyta	-	-
Isoetes_tegetiformans	lycophyta	6e-02	-
Lycopodium_deuterodensum Lycopodiella_apressa	lycophyta	-	-
Huperzia_squarrosa Huperzia_myrisinites	lycophyta	-	-
Diphasiastrum_digitatum	lycophyta	-	-
Pseudolycopodiella_caroliniana	lycophyta	-	-
Dendrolycopodium_obscurum	lycophyta	-	-
Phylloglossum_drummondii	lycophyta	-	-
Entransia_fimbriat	Charyophyta	2e-3	-
Coleochaete_scutata	Charyophyta	-	-
Spirotaenia_minuta	Charyophyta	3e-4	-
Coleochaete_irregularis	Charyophyta	-	-
Chaetosphaeridium_globosum	Charyophyta	-	4e-3
Chlorokybus_atmophyticus	Chlorokybaceae	7e-3	-
Hormidiella_sp	Klebsormidiales	-	-

Appendix 10

Table 9.7. List of holocytochrome *c* synthase and holocytochrome *c* $_1$ synthase proteins from different lineages. Species from the same eukaryotic group have the same background colour.

Species	Phylum	Protein accession number for system III	Length
Drosophila melanogaster	Arthropoda	NP_651003	281
Apis mellifera	Arthropoda	XP_006564658	276
Tribolium castaneum	Arthropoda	XP_966576	256
Homo sapiens	Chordata	NP_005324	268
Danio rerio	Chordata	NP_958859	269
Mus musculus	Chordata	AAB19008	272
Xenopus laevis	Chordata	NP_001080580	301
Caenorhabditis elegans	Nematoda	NP_496403	280
Schistosoma mansoni	Platyhelminthes	XP_018653883	249
Trichoplax adhaerens	Placozoa	XP_002114862	185
Amphimedon queenslandica	Porifera	XP_003385371	226
Monosiga brevicollis	Choanoflagellate (Class)	XP_001748514	188
Thecamonas trahens	Apusozoa	XP_013762314 XP_013757272	420 265
Dictyostelium discoideum	Amoebozoa	XP_643563	211
Dictyostelium purpureum	Amoebozoa	XP_003293567	196
Dictyostelium fasciculatum	Amoebozoa	XP_004366903	208
Acanthamoeba castellanii	Amoebozoa	XP_004334332	193
Polysphondylium pallidum	Amoebozoa	XP_020432267	223
Hydra vulgaris	Cnidaria	XP_004210211	256
Nematostella vectensis	Cnidaria	XP_001637179	241
Lottia gigantea	Mollusca	XP_009056954	237
Capitella teleta	Annelida	ELU04991	203
Strongylocentrotus purpuratus	Echinodermata	XP_011661593	443
Saccharomyces_cerevisiae	Fungi	AJO95760	269

		AJP39951	224
Phycomyces_blakesleeanus	Fungi	XP_018290546 XP_018286087	181 242
Schizosaccharomyces_pombe	Fungi	NP_594026 NP_596655	216 377
Nectria_haematococca	Fungi	XP_003052508 XP_003040024	322 262
Trichoderma_reesei	Fungi	XP_006965207 XP_006966990	330 290
Neurospora_crassa	Fungi	XP_960563 XP_959445	346 317
Zymoseptoria_tritici	Fungi	XP_003847843 XP_003854706	263 328
Candida_albicans	Fungi	KGQ91448 XP_722650	264 251
Debaryomyces_hansenii	Fungi	XP_459831 XP_462329	267 255
Kluyveromyces_lactis	Fungi	XP_452785 XP_456233	320 271
Eremothecium_gossypii	Fungi	NP_984478 NP_984251	281 252
Ustilago_maydis	Fungi	XP_011388050 XP_011391593	399 284
Coprinopsis_cinerea	Fungi	XP_001831313 XP_001833434	269 228
Rhizopus_microsporus	Fungi	ORE01114 ORE12443	269 241
Laccaria bicolor	Fungi	XP_001878634	212
Theileria_annulata	Apicomplexa	XP_951931 XP_953889	220 194
Babesia bovis	Apicomplexa	XP_001610811 XP_001612270	196 208
Plasmodium_chabaudi	Apicomplexa	XP_016654981 XP_745340	234 174
Plasmodium_falciparum	Apicomplexa	EWC87389 XP_001350445	246 184
Plasmodium_berghei	Apicomplexa	CDS52077 CDS45400	234 174
Theileria_ parva	Apicomplexa	XP_765351 XP_766407	207 194
Toxoplasma_gondii	Apicomplexa	CEL77642 KYK65410	431 215
Phytophthora_infestans	Heterokonta	XP_002909041 XP_002908738	410 228
Phytophthora_sojae	Heterokonta	XP_009529976 XP_009529594	229 413
Ectocarpus_siliculosus	Heterokonta	CBJ25846 CBN75788	510 455

Aureococcus anophagefferens	Heterokonta	XP_009039044	127
Phaeodactylum_tricornutum	Heterokonta	XP_002182758 XP_002184940	212 314
Guillardia theta	Cryptophyta	XP_005830278	232
Chrysochromulina_sp	Haptophyte	KOO26558 KOO29384	571 168
Volvox_carteri	Chlorophyta	XP_002952003 XP_002953420	170 293
Chlamydomonas reinhardtii	Chlorophyta	XP_001699246 XP_001696887	160 322
Ostreococcus_tauri	Chlorophyta	CEG00551 XP_003083917	244 292
Treubia_lacunose	Marchantiophyta	Transcriptome	237
Megaceros_vincentianus	Hornwort	Transcriptome	232
Nothoceros_aenigmaticus	Hornwort	Transcriptome	232
Selaginella moellendorffii	Lycophyte	XP_002982596	179
Isoetes_tegetiformans	Lycophyte	Transcriptome	237
Huperzia_squarrosa	Lycophyte	Transcriptome	338
Phylloglossum_drummondii	Lycophyte	Transcriptome	128
Hormidiella_sp	charophyte	Transcriptome	282
Chlorokybus_atmophyticus	charophyte	Transcriptome	350
Chondrus crispus	Red algae	XP_005713760	474
Plasmodiophora brassicae	Rhizaria	CEO98847	422
Reticulomyxa filosa	Rhizaria	ETO02116	290
Bigelowiella_natans	Rhizaria	Transcriptome	209 204
Amoeba sp. BB2	Heterolobosea	Transcriptome	144
Vitrella brassicaformis	Chromerida	CEL92815	252
Corallochytrium limacisporum	Holozoa	Transcriptome	231
Abeoforma whisleri's	Holozoa	Transcriptome	271 256
Pirum gemmata	Holozoa	Transcriptome	273
Sphaeroforma_arctica	Holozoa	XP_014157736 XP_014158860	299 269
Nutomonas longa	Apusozoa	Transcriptome	346

Appendix 11

Table 9.8. List of putative holocytochrome *c* synthase proteins from kinetoplastid species		
Organism	**TrypDB Protein Database Accession No**	**Length**
Crithidia fasciculata	CFAC1_200019900.1	369
Perkinsela sp	KNH04224.1	346
Leishmania major	LMJLV39_290019400.1	357
Trypanosoma brucei	Tb427.03.3890	320
Bodo saltans	BS59245 (Sanger database)	345

Appendix 12

Table 9.9. The phylogenetic distribution of System I and III within different eukaryotic lineages.

Species	Phylum	System I	System III	Accession number for system III	Length
Drosophila melanogaster	Arthropoda	No	Yes	NP_651003	281
Apis mellifera	Arthropoda	No	Yes	XP_006564658	276
Tribolium castaneum	Arthropoda	No	Yes	XP_966576	256
Homo sapiens	Chordata	No	Yes	NP_005324	268
Danio rerio	Chordata	No	Yes	NP_958859	269
Mus musculus	Chordata	No	Yes	AAB19008	272
Xenopus laevis	Chordata	No	Yes	NP_001080580	301
Caenorhabditis elegans	Nematoda	No	Yes	NP_496403	280
Schistosoma mansoni	Platyhelminthes	No	Yes	XP_018653883	249
Trichoplax adhaerens	Placozoa	No	Yes	XP_002114862	185
Amphimedon queenslandica	Porifera	No	Yes	XP_003385371	226
Monosiga brevicollis	Choanoflagellate (Class)	No	Yes	XP_001748514	188
Thecamonas trahens	Apusozoa	No	Yes	XP_013762314 XP_013757272	420 265
Dictyostelium discoideum	Amoebozoa	No	Yes	XP_643563	211
Dictyostelium purpureum	Amoebozoa	No	Yes	XP_003293567	196
Dictyostelium fasciculatum	Amoebozoa	No	Yes	XP_004366903	208
Acanthamoeba castellanii	Amoebozoa	No	Yes	XP_004334332	193

Polysphondylium pallidum	Amoebozoa	No	Yes	XP_020432267	223
Hydra vulgaris	Cnidaria	No	Yes	XP_004210211	256
Nematostella vectensis	Cnidaria	No	Yes	XP_001637179	241
Lottia gigantea	Mollusca	No	Yes	XP_009056954	237
Capitella teleta	Annelida	No	Yes	ELU04991	203
Strongylocentrotus purpuratus	Echinodermata	No	Yes	XP_011661593	443
Saccharomyces_cerevisiae	Fungi	No	Yes	AJO95760 AJP39951	269 224
Phycomyces_blakesleeanus	Fungi	No	Yes	XP_018290546 XP_018286087	181 242
Schizosaccharomyces_pombe	Fungi	No	Yes	NP_594026 NP_596655	216 377
Nectria_haematococca	Fungi	No	Yes	XP_003052508 XP_003040024	322 262
Trichoderma_reesei	Fungi	No	Yes	XP_006965207 XP_006966990	330 290
Neurospora_crassa	Fungi	No	Yes	XP_960563 XP_959445	346 317
Zymoseptoria_tritici	Fungi	No	Yes	XP_003847843 XP_003854706	263 328
Candida_albicans	Fungi	No	Yes	KGQ91448 XP_722650	264 251
Debaryomyces_hansenii	Fungi	No	Yes	XP_459831 XP_462329	267 255
Kluyveromyces_lactis	Fungi	No	Yes	XP_452785 XP_456233	320 271
Eremothecium_gossypii	Fungi	No	Yes	NP_984478 NP_984251	281 252
Ustilago_maydis	Fungi	No	Yes	XP_011388050 XP_011391593	399 284
Coprinopsis_cinerea	Fungi	No	Yes	XP_001831313 XP_001833434	269 228
Rhizopus_microsporus	Fungi	No	Yes	ORE01114 ORE12443	269 241
Laccaria bicolor	Fungi	No	Yes	XP_001878634	212
Theileria_annulata	Apicomplexa	No	Yes	XP_951931 XP_953889	220 194
Babesia bovis	Apicomplexa	No	Yes	XP_001610811 XP_001612270	196 208
Plasmodium_chabaudi	Apicomplexa	No	Yes	XP_016654981 XP_745340	234 174
Plasmodium_falciparum	Apicomplexa	No	Yes	EWC87389 XP_001350445	246 184
Plasmodium_berghei	Apicomplexa	No	Yes	CDS52077 CDS45400	234 174

Theileria_ parva	Apicomplexa	No	Yes	XP_765351 XP_766407	207 194
Toxoplasma_gondii	Apicomplexa	No	Yes	CEL77642 KYK65410	431 215
Phytophthora_infestans	Heterokonta	No	Yes	XP_002909041 XP_002908738	410 228
Phytophthora_sojae	Heterokonta	No	Yes	XP_009529976 XP_009529594	229 413
Ectocarpus_siliculosus	Heterokonta	No	Yes	CBJ25846 CBN75788	510 455
Aureococcus anophagefferens	Heterokonta	No	Yes	XP_009039044	127
Phaeodactylum_tricornutum	Heterokonta	No	Yes	XP_002182758 XP_002184940	212 314
Guillardia theta	Cryptophyta	No	Yes	XP_005830278	232
Chrysochromulina_sp	Haptophyte	No	Yes	KOO26558 KOO29384	571 168
Volvox_carteri	Chlorophyta	No	Yes	XP_002952003 XP_002953420	170 293
Chlamydomonas reinhardtii	Chlorophyta	No	Yes	XP_001699246 XP_001696887	160 322
Ostreococcus_tauri	Chlorophyta	No	Yes	CEG00551 XP_003083917	244 292
Treubia_lacunose	Marchantiophyta	No	Yes	Transcriptome	237
Megaceros_vincentianus	Hornwort	No	Yes	Transcriptome	232
Megaceros_tosanus	Hornwort	Yes	No	-	-
Nothoceros_aenigmaticus	Hornwort	No	Yes	Transcriptome	232
Selaginella moellendorffii	Lycophyte	No	Yes	XP_002982596	179
Isoetes_tegetiformans	Lycophyte	No	Yes	Transcriptome	237
Huperzia_squarrosa	Lycophyte	No	Yes	Transcriptome	338
Phylloglossum_drummondii	Lycophyte	Yes	Yes	Transcriptome	128
Hordeum vulgare	Angiosperms	Yes	No	-	-
Arachis ipaensis	Angiosperms	Yes	No	-	-
Arabidopsis thaliana	Angiosperms	Yes	No	-	-
Physcomitrella patens	Bryophyta	Yes	No	-	-
Hormidiella_sp	charophyte	No	Yes	Transcriptome	282
Chlorokybus_atmophyticus	charophyte	No	Yes	Transcriptome	350
Chondrus crispus	Red algae	No	Yes	XP_005713760	474

Galdieria sulphuraria	Red algae	Yes	No	-	-
Cyanidioschyzon merolae	Red algae	Yes	No	-	-
Plasmodiophora brassicae	Rhizaria	No	Yes	CEO98847	422
Reticulomyxa filosa	Rhizaria	No	Yes	ETO02116	290
Bigelowiella_natans	Rhizaria	No	Yes	Transcriptome	209 204
Naegleria gruberi	Heterolobosea	Yes	No	-	-
Naegleria fowleri	Heterolobosea	Yes	No	-	-
Amoeba sp. BB2	Heterolobosea	No	Yes	Transcriptome	144
Reclinomonas americana	Jakobida	Yes	No	-	-
Malawimonas californiana	Loukozoa	Yes	No	-	-
Palpitomonas bilix	Cryptists	Yes	No	-	-
Tetrahymena thermophila	Ciliophora	Yes	No	-	-
Paramecium tetraurelia	Ciliophora	Yes	No	-	-
Ichthyophthirius multifiliis	Ciliophora	Yes	No	-	-
Oxytricha trifallax	Ciliophora	Yes	No	-	-
Pseudocohnilembus persalinus	Ciliophora	Yes	No	-	-
Stylonychia lemnae	Ciliophora	Yes	No	-	-
Vitrella brassicaformis	Chromerida	No	Yes	CEL92815	252
Perkinsus marinus	Perkinsozoa	Yes	No	-	-
Corallochytrium limacisporum	Holozoa	No	Yes	Transcriptome	231
Parvularia atlantis	Holomyocota	Yes	No	-	-
Fonticula alba	Holomyocota	Yes	No	-	-
Chromosphaera perkinsii	Holozoa	Yes	No	-	-
Abeoforma whisleri's	Holozoa	No	Yes	Transcriptome	271 256
Pirum gemmata	Holozoa	No	Yes	Transcriptome	273
Capsaspora owczarzaki	Holozoa	Yes	No	-	-
Sphaeroforma_arctica	Holozoa	No	Yes	XP_014157736 XP_014158860	299 269
Nutomonas longa	Apusozoa	No	Yes	Transcriptome	346

Appendix 13

Table 9.10. Amino acid colour code scheme according to RasMol 2.75.	
Amino Acids	**Colour Name**
ASP, GLU	Bright Red
CYS, MET	Yellow
LYS, ARG	Blue
SER, THR	Orange
PHE, TYR	Mid Blue
ASN, GLN	Cyan
GLY	Light Grey
LEU, VAL, ILE	Green
ALA	Dark Grey
TRP	Purple
HIS	Pale Blue
PRO	Flesh

Appendix 14

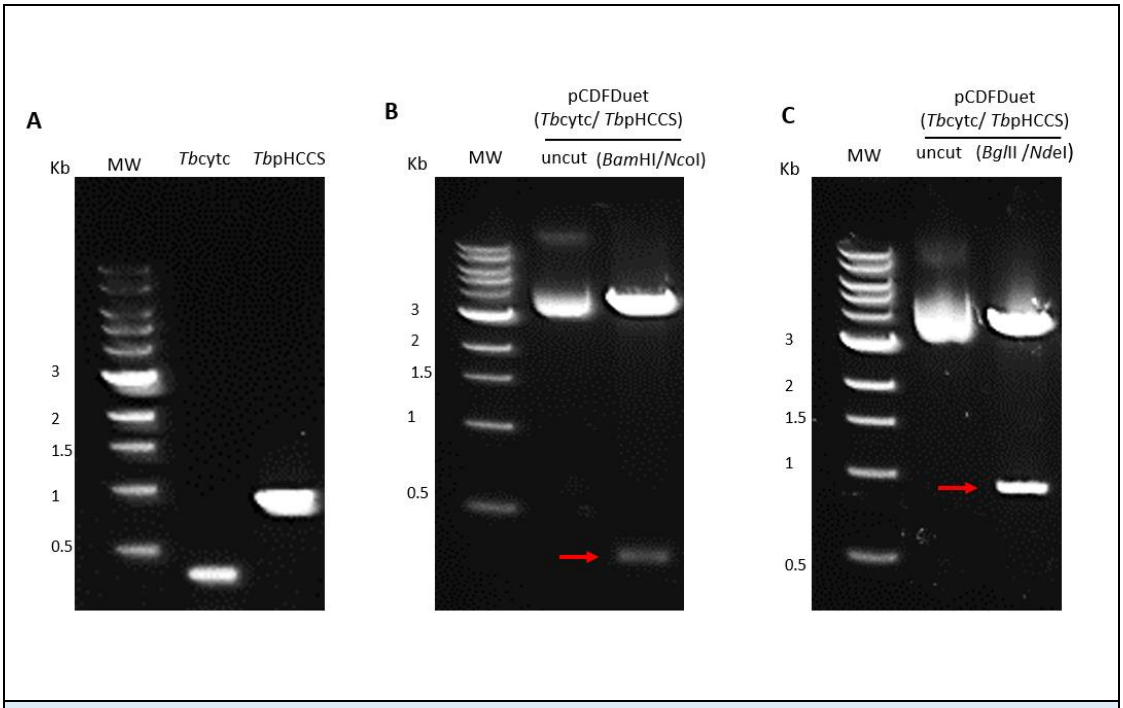


Figure 9.11. Cloning of *Tb*pHCCS and *Tb*CYTC (non-His tag) to pCDF-Duet plasmid. (A) Agarose gel electrophoresis analysis of *Tb*CYTC and *Tb*pHCCS coding sequence amplification. For *Tb*CYTC gene, the primers had restriction sites for *Bam*HI and *Nco*I. For *Tb*pHCCS gene, the primers had restriction sites of *Bgl*II and *Nde*I. (B) The diagnostic digestion of pCDF-Duet plasmid to detect the presence of *Tb*CYTC. Lane 1 corresponds to undigested plasmid. Lane 2 corresponds to pCDFDuet digested using *Bam*HI and *Nco*I. (C) The diagnostic digest of pCDF-Duet plasmid to detect the presence of *Tb*pHCCS. Lane 1 corresponds to undigested plasmid. Lane 2 corresponds to pCDFDuet digested using *Bgl*II and *Nde*I. The digestive solutions (20 µL) were subjected to electrophorese in a 0.7% agarose gel. MW: NEB 1 kb DNA Ladder. The red arrows point to the inserts.

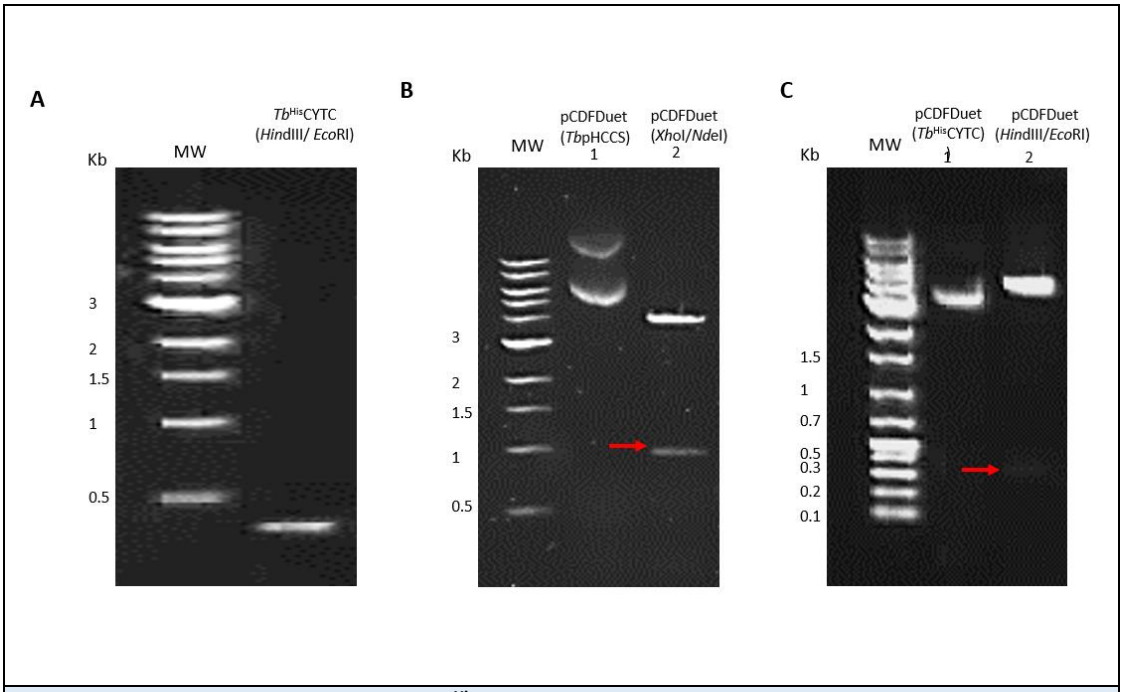


Figure 9.10. Cloning of *Tb*pHCCS and *Tb*^{His}CYTC (with His tag) to pCDF-Duet plasmid. (A) Agarose gel electrophoresis of *Tb*^{His}CYTC amplified by PCR using primers with restriction sites of *Hind*III and *EcoR*I. For *Tb*pHCCS, the same PCR product used in the cloning of *Tb*pHCCS into pET-28a, was used for pCDF-Duet cloning. (B) The diagnostic digestion of pCDF-Duet to detect the presence of *Tb*pHCCS. Lane 1 corresponds to the undigested plasmid. Lane 2 corresponds to pCDF-Duet digested using *Xho*I/*Nde*I. (C) The diagnostic digestion of pCDF-Duet to detect the presence of *Tb*^{His}CYTC. Lane 1 corresponds to the undigested plasmid. Lane 2 corresponds to pCDF-Duet digested using *Hind*III/*Eco*RI. The digestive solutions (20 µL) were subjected to electrophorese in a 0.7% agarose gel. MW: NEB 1 kb DNA Ladder.

Appendix 15

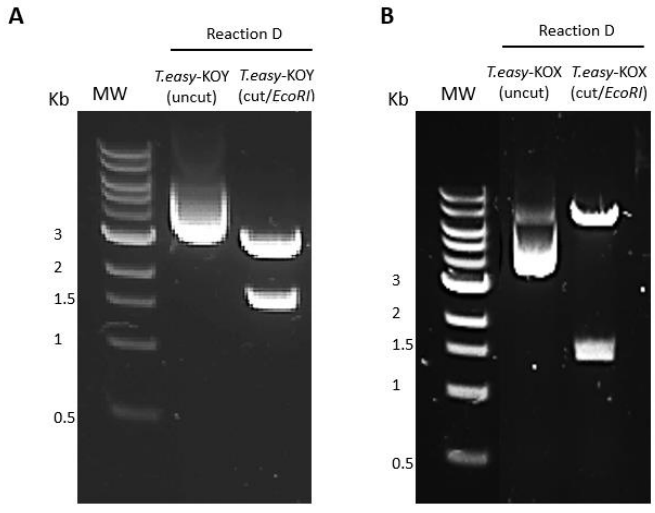


Figure 9.11. Diagnostic digest of pGEM-T Easy plasmid cloned to PCR products from reaction D. The PCR product from reaction D (Chapter 4 Figure 4.14) was cloned to pGEM-T Easy to obtain the sequence data. Images A and B show the diagnostic digest of pGEM-T Easy cloned to reaction D products from the KOY and KOX cell lines respectively. MW: NEB 1 kb DNA Ladder. The ligated vectors were sent to SourceBioscience for sequencing using M13R and M13F primers.

Figure 9.12. Pairwise nucleotide sequence alignment of the pGEM-T Easy cloned to reaction D product and web annotated wild type *Lm*pHCCS gene. Red indicates a part of *Lm*pHCCS coding sequence, green indicates the upstream and downstream sequence of *Lm*pHCCS gene, dark blue shows the sequence of 5` and 3` sgRNA and yellow highlight is the sequence of 5` and 3` UTR. The alignment was performed using MAFFT server.

Appendix 16

Table 9.11. The accession numbers of acetyl coenzyme A synthetase from bacteria and archaea.

Species	Phyla	Accession number
Haloarcula marismortui	Archaea	WP_011223298.1
Halorubrum chaoviator	Archaea	WP_089308072.1
Archaeoglobus fulgidus	Archaea	WP_048095590.1
Methanocaldococcus jannaschii	Archaea	WP_010870094.1
Methanocaldococcus bathoardescens	Archaea	WP_048201820.1
Methanotorris igneus	Archaea	WP_013799096.1
Methanococcus maripaludis	Archaea	WP_012067786.1
Actinobacteria bacterium	Actinobacteria	OFW49717.1
Nitrospinae bacterium	Nitrospinae	OGV97865.1
Chloroflexus islandicus	Chloroflexi	WP_066787342.1
Ardenticatena	Chloroflexi	CUS05644.1
Caldilinea aerophila	Chloroflexi	WP_044276212.1
Anaerolineae bacterium	Chloroflexi	KPK04638.1
Stanieria cyanosphaera	Cyanobacteria	WP_015193816.1
Calothrix sp	Cyanobacteria	BAZ39627.1
Anabaena sp	Cyanobacteria	WP_016949993.1
Cyanothece sp	Cyanobacteria	WP_013321343.1
Leptolyngbya ohadii	Cyanobacteria	WP_088894771.1
Desulfococcus multivorans	Proteobacteria	WP_020878498.1
Desulfobacteraceae bacterium	Proteobacteria	OQX25726.1
Desulfatirhabdium butyrativorans	Proteobacteria	WP_028324336.1
Desulfatitalea sp	Proteobacteria	KJS33360.1
Syntrophobacter fumaroxidans	Proteobacteria	WP_011698821.1
Rhodothermus profundi	Bacteroidete	WP_072714317.1
Rhodothermaceae bacterium	Bacteroidete	WP_068121898.1

Appendix 17

Table 9.12. The accession numbers of acetyl coenzyme A synthetase from eukaryotes.

Species	Rank	Accession number
Entamoeba invadens	Amoebozoa	XP_004259966.1
Entamoeba histolytica	Amoebozoa	XP_656290.1
Mastigamoeba balamuthi	Amoebozoa	AIW47214.1
Tritrichomonas foetus	Parabasalia	OHT12303.1
Naegleria fowleri	Heterolobosea	mRNA1_NF0124800
Naegleria gruberi	Heterolobosea	XP_002669321.1
Cyanidioschyzon merolae	Rhodophyta	XP_005537438.1
Cyanophora_paradoxa	Glaucophyta	ConsensusfromContig9202
Plasmodium falciparum	Alveolata	ETW54022.1
Plasmodium berghei	Alveolata	XP_675930.1
Plasmodium yoelii	Alveolata	ETB58353
Plasmodium vivax	Alveolata	XP_001616599.1
Cryptosporidium muris	Alveolata	XP_002140975.1
Gregarina niphandrodes	Alveolata	XP_011130974.1
Vitrella brassicaformis	Alveolata	CEL97119.1
Symbiodinium microadriaticum	Alveolata	OLQ13875.1
Giardia lamblia	Diplomonad	XP_001705744.1
Giardia intestinalis	Diplomonad	ESU37821.1
Spironucleus salmonicida	Diplomonad	AFV80087.1
Trepomonas sp	Diplomonad	AGV05441.1
Thalassiosira pseudonana	Stramenopiles	XP_002294673.1
Fistulifera solaris	Stramenopiles	GAX11266.1
Blastocystis hominis	Stramenopiles	XP_012899147

Appendix 18

Table 9.13: Domain architecture of the ACD across three life domains and *N. gruberi* putative ACD protein. The grey background indicates that species belong to archaea and the orange background indicates that species belong to bacteria and the green background indicates that species belong to eukaryotes.

Species	Type of domain	Domain(s) name	Sequence position	TMHMM	Sequence position
Haloarcula marismortui	Multiple domain	CoA_binding Succ_CoA_ligase ATP-grasp	7-103 154 – 291 477- 696	No	-
Halorubrum chaoviator	Multiple domain	CoA_binding Succ_CoA_ligase ATP-grasp	7 -99 150 - 287 471 - 697	No	-
Archaeoglobus fulgidus	Multiple domain	CoA_binding Succ_CoA_ligase ATP-grasp	4-97 148 – 285 464- 681	No	-
Caldilinea aerophila	Multiple domain	CoA_binding Succ_CoA_ligase ATP-grasp GNAT	22 - 117 168 - 305 497 – 718 751- 894	No	-
Calothrix sp	Multiple domain	CoA_binding Succ_CoA_ligase ATP-grasp GNAT	27- 122 173- 310 505- 726 759- 903	No	-
Desulfococcus multivorans	Multiple domain	CoA_binding Succ_CoA_ligase ATP-grasp	7 - 102 155 – 295 490- 712	No	-
Entamoeba histolytica	Multiple domain	CoA_binding Succ_CoA_ligase ATP-grasp	6 -100 151 -288 474- 694	No	-
Naegleria gruberi	Multiple domain	Succ_CoA_ligase ATP-grasp GNAT	301- 429 638- 871 945 -1,100	No	-
Cyanidioschyzon merolae	Multiple domain	CoA_binding Succ_CoA_ligase ATP-grasp GNAT	117- 211 265- 402 597- 820 856- 1006	No	-
Cyanophora_paradoxa	Multiple domain	CoA_binding Succ_CoA_ligase ATP-grasp	24 - 119 165 – 290 485- 686	No	-
Plasmodium falciparum	Multiple domain	CoA_binding Succ_CoA_ligase ATP-grasp	40 - 136 187 – 324 540- 763	No	-
Giardia lamblia	Multiple domain	CoA_binding Succ_CoA_ligase ATP-grasp	18 - 113 162 – 299 490- 724	No	-

Thalassiosira pseudonana	Multiple domain	CoA_binding Succ_CoA_ligase ATP-grasp	47- 147 198- 335 525- 748	No	-

Appendix 19

Table 9.14. The accession numbers of CuNirk from bacteria and archaea (cladeII)

Species	Phyla	Accession number
Euryarchaeota archaeon	Archaea	OGS48846.1
Haloferax denitrificans	Archaea	WP_049917686.1
Halosimplex carlsbadense	Archaea	WP_006882419.1
Halogranum amylolyticum	Archaea	SEP05421.1
Halomicrobium mukohataei	Archaea	WP_015761990.1
Halovivax ruber	Archaea	AGB17466.1
Halomicrobium katesii	Archaea	WP_026190161.1
Halomicrobium zhouii	Archaea	SFS07765.1
Halorhabdus utahensis	Archaea	WP_012795101.1
Halogeometricum pallidum	Archaea	ELZ32490.1
Halovenus aranensis	Archaea	SDJ68612.1
Chryseobacterium palustre	Bacteroidete	WP_027376684.1
Flavobacteriaceae bacterium	Bacteroidete	WP_088358735.1
Bizionia sp	Bacteroidete	WP_066255894.1
Capnocytophaga sputigen	Bacteroidete	WP_002678273.1
Turneriella parva	Spirochaetes	WP_014803251.1
Leptospira biflexa	Spirochaetes	WP_012476189.1
Ardenticatena maritima	Chloroflexi	GAP62154.1
Chloroflexus	Chloroflexi	WP_012257443.1
Sphaerobacter thermophilus	Chloroflexi	WP_012870640.1
Bacillus timonensis	Firmicutes	WP_010678487.1
Geobacillus stearothermophilus	Firmicutes	WP_033015213.1
Paenibacillus uliginis	Firmicutes	SMF90687.1
Paucisalibacillus globulus	Firmicutes	WP_026906831.1
Tetrasphaera jenkinsii	Actinobacteria	CCI53895.1
Mycobacterium sp	Actinobacteria	OBA69013.1
Cellulomonas carbonis	Actinobacteria	WP_081978963.1
Mumia flava	Actinobacteria	KHL07042.1

Appendix 20

Table 9.15. The accession numbers of CuNirk from proteobacteria (cladeI)		
Species	Phyla	Accession number
Rhodobacter sphaeroides	Alphaproteobacteria	ACM00517.1
Ensifer sp	Alphaproteobacteria	WP_029962470.1
Maritalea myrionectae	Alphaproteobacteria	WP_027835903.1
Hyphomicrobium nitrativorans	Alphaproteobacteria	WP_023787941
Ahrensia sp	Alphaproteobacteria	WP_026480238.1
Brucella ovis	Alphaproteobacteria	WP_006015294.1
Ruegeria mobilis	Alphaproteobacteria	WP_005620794.1
Pannonibacter phragmitetus	Alphaproteobacteria	WP_050473129.1
Cucumibacter marinus	Alphaproteobacteria	WP_051332723.1
Shinella sp	Alphaproteobacteria	EYR78640.1
Starkeya novella	Alphaproteobacteria	WP_013165971.1
Bradyrhizobium elkanii	Alphaproteobacteria	WP_028162236.1
Bradyrhizobium oligotrophicum	Alphaproteobacteria	WP_015664480.1
Agrobacterium_vitis	Alphaproteobacteria	WP_071201566.1
Phaeobacter inhibens	Alphaproteobacteria	WP_014889326.1
Achromobacter insuavis	Betaproteobacteria	EGP44463.1
Achromobacter xylosoxidans	Betaproteobacteria	WP_076414157.1
Castellaniella defragrans	Betaproteobacteria	WP_043686203.1
Taylorella asinigenitalis	Betaproteobacteria	CCG19001.1
Pseudomonas protegens	Gammaproteobacteria	AGL87193.1
Pseudoalteromonas nigrifaciens	Gammaproteobacteria	WP_089368261.1
Shewanella denitrificans	Gammaproteobacteria	WP_011497898.1
Shewanella amazonensis	Gammaproteobacteria	WP_011760789.1
Cardiobacterium valvarum	Gammaproteobacteria	WP_006986126.1

Appendix 21

Table 9.16. The accession numbers of CuNirk from eukaryotes.		
Species	**Rank**	**Accession number**
Acanthamoeba castellanii	Amoebozoa	ABD46578.1
Physarum polycephalum	Amoebozoa	ABD46592.1
Vermamoeba vermiformis	Amoebozoa	ABD46561.1
Fusarium oxysporum	Fungi	ABU88100.1
Arthroderma otae	Fungi	XP_002846079.1
Trichophyton violaceum	Fungi	OAL71010.1
Penicillium polonicum	Fungi	OQD65046.1
Uncinocarpus reesii	Fungi	XP_002542581.1
Nannizzia gypsea	Fungi	XP_003171471.1
Cladophialophora bantiana	Fungi	XP_016615461.1
Chaetomium globosum	Fungi	XP_001227922.1
Thielavia terrestris	Fungi	XP_003650933.1
Purpureocillium lilacinum	Fungi	XP_018178962.1
Aspergillus calidoustus	Fungi	CEL11877.1
Madurella mycetomatis	Fungi	KXX72853.1
Emmonsia crescens	Fungi	KKZ67494.1
Blastomyces percursus	Fungi	OJD23300.1
Exophiala oligosperma	Fungi	XP_016269148.1
Coccidioides immitis	Fungi	XP_001240576.2
Naegleria fowleri	Heterolobosea	mRNA1_NF0120230
Naegleria gruberi	Heterolobosea	XP_002674759.1
Emiliania huxleyi	Isochrysidales	XP_005783766.1
Galdieria sulphuraria	Rhodophyta	XP_005707361.1
Chlamydomonas reinhardtii	Chlorophyta	PNW79625.1

Appendix 22

Table 9.17: Domain architecture of the CuNirK across three life domains and *N. gruberi* putative CuNirK protein. The yellow background indicates that species belong to bacteria and the blue background indicates that species belong to archaea and the green background indicates that species belong to eukaryotes. Star symbol represents partial CuNirK.

Species	Type of domain	Domain(s) name	Sequence position	TMHMM	Sequence position
Rhodobacter sphaeroides	Multiple domain	Copper oxidase, type 1 - like domain Copper oxidase, type 2-like domain	57-174 187-339	No	-
Shewanella denitrificans	Multiple domain	Transmembrane region Copper oxidase, type 1 - like domain Copper oxidase, type 2-like domain	20-42 93-210 223-375	Yes	20-42
Achromobacter insuavis	Multiple domain	Copper oxidase, type 1 - like domain Copper oxidase, type 2-like domain	41-172 185-337	No	-
Euryarchaeota archaeon	Multiple domain	Transmembrane region Copper oxidase, type 1 - like domain Copper oxidase, type 2-like domain	27-49 220-331 337-482	Yes	27-49
Bacillus timonensis	Multiple domain	Copper oxidase, type 1 - like domain Copper oxidase, type 2-like domain	71-188 209-343	No	-
Mycobacterium sp	Multiple domain	Copper oxidase, type 1 - like domain Copper oxidase, type 2-like domain	236-285 290-430	No	-
Vermamoeba vermiformis	Multiple domain	Copper oxidase, type1 - like domain Copper oxidase, type 2-like domain	33-145 150-265	No	-
Fusarium oxysporum	Multiple domain	Copper oxidase, type 1 - like domain Copper oxidase, type 2-like domain	203- 318 324-479	No	-
Naegleria gruberi	Multiple domain	Copper oxidase, type 1 - like domain Copper oxidase, type 2-like domain	104-218 223-372	No	-
Emiliania huxleyi	Multiple domain	Copper oxidase, type 1 - like domain Copper oxidase, type 2-like domain	24- 127 157-239	No	-
Galdieria sulphuraria	Multiple domain	Copper oxidase, type 1 - like domain Copper oxidase, type 2-like domain	78- 192 197-347	No	-
*Chlamydomonas reinhardtii**	Multiple domain	Copper oxidase, type 1 - like domain Copper oxidase, type 2-like domain	24- 136 141-271	No	-

Appendix 23

Table 9.18: Secondary sequence of Pat, an N(ε)-lysine acetyltransferase protein predicted by Smart.

Species	Type of domain	Domain(s) name	Sequence position	TMHMM	Sequence position
Escherichia coli	Multiple domain	CoA_binding Succ_CoA_ligase ATP-grasp GNAT	9- 104 147- 280 472- 694 725- 864	No	-
Salmonella enterica	Multiple domain	CoA_binding Succ_CoA_ligase ATP-grasp GNAT	9- 104 147- 280 472- 694 725- 864	No	-
Pantoea sp	Multiple domain	CoA_binding Succ_CoA_ligase ATP-grasp GNAT	9- 104 147- 280 472- 694 725- 864	No	-
Vibrio cholerae	Multiple domain	CoA_binding Succ_CoA_ligase ATP-grasp GNAT	7- 102 153- 290 483- 706 716- 876	No	-
Shigella flexneri	Multiple domain	CoA_binding Succ_CoA_ligase ATP-grasp GNAT	9- 104 147- 280 472- 694 725- 864	No	-
Dickeya zeae	Multiple domain	CoA_binding Succ_CoA_ligase ATP-grasp GNAT	9- 104 147- 281 473- 695 727- 866	No	-
Naegleria gruberi	Multiple domain	Succ_CoA_ligase ATP-grasp GNAT	301- 429 638- 871 949- 1090	No	-
Naegleria fowleri	Multiple domain	Succ_CoA_ligase ATP-grasp GNAT	320- 448 658- 891 968- 1100	No	-

Appendix 24

Table 9.19. *Naegleria gruberi* glycolysis enzyme. Red coloured proteins indicate the possibility of PTS1 in the protein C-terminus.

Glycolysis enzyme	Accession number in *N. gruberi*	Location in the genome	PTS1 in ORF	Mechanism for dual localization	Stop codon
Glucokinase	XP_002672587.1	108429- 109864 GB:NW_003163275.1	No	No	**G**-TAA-**A**
Glucose-6-phosphate isomerase	XP_002680064.1	189908- 191455 GB:NW_003163316.1	No	Yes, Ribosomal read-through	**A**-TAA-**G**
Phosphofructokinase	XP_002680445.1	291695- 293176 GB:NW_003163318.1	No	No	**A**-TAA-**T**
Triosephosphate isomerase	XP_002683097.1	439537- 440295 GB:NW_003163326.1	No	Yes, Ribosomal read-through	**A**-TGA-**C**
Glyceraldehyde-3-phosphate dehydrogenase	XP_002669989.1	214017- 215018 GB:NW_003163251.1	No	Yes, Ribosomal read-through	**A**-TAA-**C**
Phosphoglycerate kinase	XP_002672486.1	259781- 261033 GB:NW_003163274.1	No	Yes, Ribosomal read-through	**A**-TGA-**A**
Phosphoglycerate mutase	XP_002682122.1	428704- 429496 GB:NW_003163323.1	No	No	**A**-TAA-**A**
Phosphopyruvate hydratase	XP_002680773.1	737020- 738555 GB:NW_003163319.1	No	No	**A**-TAA-**T**
Fructose-bisphosphate aldolase	XP_002670041.1	212044- 213114 GB:NW_003163251.1	No	No	**T**-TAA-**A**
Pyruvate kinase	XP_002681063.1	308231- 310234 GB:NW_003163320.1	No	No	**T**-TAA-**T**
	XP_002677247.1	332986-334527 GB:NW_003163303.1	No	No	**T**-TAA- **A**

Table 9.20. *Naegleria gruberi* proteins homologues to the proteins mentioned in Stiebler et al, 2014 and Ast et al,2013. Red coloured proteins indicate the possibility of PTS1 in the protein C-terminus. Dash symbol represents the failure to recover protein homologues using a BLAST search.

Enzyme (ORF)	Accession number in *Naegleria gruberi*	Location in the genome	~PTS1 motif in ORF	Mechanism for dual localization
Malate dehydrogenase	XP_002678973.1	210379- 211394 GB:NW_003163312.1	No	No
	XP_002680884.1	214958- 216643 GB:NW_003163319.1	No	Yes, Alternative splicing frame 3 after 465 bp of the ORF stop codon
	XP_002670559.1	29617-30603 GB:NW_003163257.1	No	No
Isocitrate dehydrogenase NADPH	XP_002673738.1	189664- 190952 GB:NW_003163280.1	No	Yes, Alternative splicing frame 3 after 150bp of the ORF stop codon
	XP_002673333.1	189664- 190952 GB:NW_003163280.1	No	No
	XP_002674910.1	354155-355592 GB:NW_003163290.1	No	No
Citrate synthase	XP_002678141.1	452601- 454002 GB:NW_003163308.1	No	No
	XP_002669424.1	73950_ 75461 GB:NW_003163244.1	No	Yes, Alternative splicing frame 3 after 85bp of the ORF stop codon
	XP_002668999.1	17977_19738 GB:NW_003163237.1	No	Yes, Alternative splicing frame 2 after 200bp of the ORF stop codon
6-phosphogluconate dehydrogenase	XP_002681464.1	218485_ 220105 GB:NW_003163321.1	No	No
Hydroxypyruvate reductase	XP_002681144.1	675115- 675951 GB:NW_003163320.1	No	No
Superoxide dismutase	XP_002679302.1	19966- 20578 GB:NW_003163313.1	No	Yes, Ribosomal read-through
	XP_002669988.1	207758 - 208465 GB:NW_003163251.1	No	No
	XP_002674960.1	186880 - 187704 GB:NW_003163290.1	No	No
Hydroxymethylglutaryl-CoA lyase	XP_002673715.1	243055- 245261 GB:NW_003163282.1	No	No

Catalase	XP_002679157.1	465062 - 467114 GB:NW_003163312.1	No	No
	XP_002671483.1	75027 - 76702 GB:NW_003163266.1	No	No
Epoxide hydrolase	EFC41014.1	182780 - 184376 GB:NW_003163283.1	No	No
Alanine-glyoxylate aminotransferase	XP_002682173.1	633568- 635308 GB:NW_003163323.1	No	No
glycerol-3-phosphate dehydrogenase (NAD+)	XP_002675914.1	24251- 25473 GB:NW_003163296.1	No	No
	EFC49996.1	863005- 864792 GB:NW_003163325.1	No	No
glucose-6-phosphate dehydrogenase	XP_002681321.1	220664 - 222959 GB:NW_003163321.1	No	No
NAD-dependent L-lactate dehydrogenase A	-	-	-	-
Aldehyde reductase	XP_002668932.1	45147 - 46258 GB:NW_003163235	No	Yes, Alternative splicing frame 3 after 66bp of the ORF stop codon
	XP_002677708.1	232903 - 234046 GB:NW_003163306.1	No	No
D-ribulose-5-phosphate 3-epimerase	XP_002681532.1	544840- 545547 GB:NW_003163321.1	No	Yes, Alternative splicing frame 2 after 85bp of the ORF stop codon
Isocitrate lyase	XP_002675800.1	236251- 237326 GB:NW_003163295.1	No	Yes, Alternative splicing frame 1 after 376bp of the ORF stop codon
Malate synthase	-	-	-	-

Appendix 25

Table 9.21. *Naegleria fowleri* glycolysis enzyme. Red coloured proteins indicate the possibility of PTS1 in the protein C-terminus.

Glycolysis enzyme	Accession number in *Naegleria fowleri*	PTS1 in ORF	Mechanism for dual localization	Stop codon
Glucokinase	NF0035880	No	No	**G**-TAA-**C**
Glucose-6-phosphate isomerase	NF0070790	Yes	No	**G**-TAA-**A**
Phosphofructokinase	NF0074270	No	No	**A**-TGA- **A**
Triosephosphate isomerase	NF0071610	No	Yes, Ribosomal read-through	**A**-TGA- **A**
Glyceraldehyde-3-phosphate dehydrogenase	NF0055660	No	Yes, Ribosomal read-through	**A**-TAA- **C**
Phosphoglycerate kinase	NF0025440	No	Yes, Alternative splicing frame 1 after 197 bp of the ORF stop codon	**G**-TAA-**A**
Phosphoglycerate mutase	NF0048460	No	No	**G**-TAA- **A**
Phosphopyruvate hydratase	NF0118810	No	No	**A**-TAA- **A**
Fructose-bisphosphate aldolase	NF0055670	No	Yes, Ribosomal read-through	**C**-TAA- **A**
Pyruvate kinase	NF0099830	No	No	**T**-TGA- **G**
	NF0102490	No	No	**A**-TAA- **A**

Table 9.22. *Naegleria fowleri* proteins homologues to the proteins mentioned in Stiebler et al, 2014 and Ast et al,2013. Red coloured proteins indicate the possibility of PTS1 in the protein C-terminus. Dash symbol represents the failure to recover protein homologues using a BLAST search.

Enzyme (ORF)	Accession number in *Naegleria fowleri*	~PTS1 motif in ORF	Mechanism for dual localization
Malate dehydrogenase	NF0114780	No	No
	NF0101270	No	No
	NF0021050	No	Yes, Alternative splicing frame 3 after 257 bp of the ORF stop codon
Isocitrate dehydrogenase NADPH	NF0066410	Yes	No
	NF0063250	No	No
	NF0016700	No	No
Citrate synthase	NF0117150	No	No
	NF0062500	No	No
	NF0123360	No	No
6-phosphogluconate dehydrogenase	NF0099360	No	No
Hydroxypyruvate reductase	NF0111240	No	No
Hydroxymethylglutaryl-CoA lyase	NF0057980	No	No
Superoxide dismutase	NF0020800	No	Yes, Ribosomal read-through
Catalase	NF0104910	Yes	No
	NF0121200	No	No
Epoxide hydrolase	NF0062220	No	No

Alanine-glyoxylate aminotransferase	NF0049330	No	No
glycerol-3-phosphate dehydrogenase (NAD+)	NF0014410	No	No
glucose-6-phosphate dehydrogenase	NF0099370	No	No
NAD-dependent L-lactate dehydrogenase A	-	-	-
Aldehyde reductase	NF0019910	No	Yes, Alternative splicing frame 2 after 363bp of the ORF stop codon
	NF0118010	No	No
D-ribulose-5-phosphate 3-epimerase	NF0059920	No	No
Isocitrate lyase	-	-	-
Malate synthase	-	-	-

Appendix 26

Table 9.23. *Dictyostelium discoideum* glycolysis enzyme. Red coloured proteins indicate the possibility of PTS1 in the protein C-terminus**.**

Glycolysis enzyme	Accession number in *Dictyostelium*	Location in the genome	PTS1 in ORF	Mechanism for dual localization	Stop codon
Glucokinase	XP_640672	4486206-4487897 GB:NC_007089.4	No	No	**C**-TAA-**A**
Glucose-6-phosphate isomerase	XP_638957	954966- 956651 GB: NC_007090.3	Yes	No	**T**-TAA-**A**
Phosphofructokinase	XP_003287952	26866-29280 GB:NW_003519454.1	No	No	**T**-TAA-**A**
	XP_644162	4510635-4513139 GB:NC_007088.5	No	No	**A**-TAA-**A**
Triosephosphate isomerase	XP_644150	4481626-4482490 GB:NC_007088.5	No	Yes, Ribosomal read-through	**T**-TGA-**A**
	XP_003283377	39879-40816 GB:NW_003519344.1	No	No	**C**-TGA-**A**
Glyceraldehyde-3-phosphate dehydrogenase	XP_643857	5324765- 5325948 GB: NC_007088.5	No	No	**T**-TAA-**A**
	XP_003285647	44991- 46293 GB: NW_003519387.1	No	No	**C**-TAA-A
Phosphoglycerate kinase	XP_637130	494968-496392 GB:NC_007091.3	No	Yes, Ribosomal read-through	**A**-TGA-**A**
Phosphoglycerate mutase	XP_641003	1432985- 1434625 GB: XP_641003.1	No	No	**A**-TAA-**T**
Phosphopyruvate hydratase	XP_639231	225309- 226956 GB: NC_007090.3	No	No	**T**-TAA-**T**
	XP_647650	1697493- 1698824 GB: NC_007087.3	No	No	**T**-TAA-**T**
	XP_003294597	2984- 3620 GB: NW_003519923.1	No	No	**T**-TAA-**A**
Fructose-bisphosphate aldolase	XP_643874	3777877- 3779287 GB: NC_007088.5	No	No	**T**-TAA-**A**
Pyruvate kinase	XP_639190	344914- 346711 GB: NC_007090.3	No	No	**A**-TAA-**A**

Table 9.24. *Dictyostelium discoideum* proteins homologues to the proteins mentioned in Stiebler et al, 2014 and Ast et al,2013. Red coloured proteins indicate the possibility of PTS1 in the protein C-terminus. Dash symbol represents the failure to recover protein homologues using a BLAST search.

Enzyme (ORF)	Accession number in *Dictyostelium*	Location in the genome	~PTS1 motif in ORF	Mechanism for dual localization
Malate dehydrogenase	XP_641333	3171359-3172577 GB: XP_641333	No	No
	XP_635832.1	3795307-3796614 GB: NC_007091.3	No	No
	XP_629516.1	1826095-1827141 GB: NC_007092.3	No	No
Isocitrate dehydrogenase NADPH	XP_645283.1	1670009-1671569 GB:NC_007088.5	No	Alternative transcription frame 3 after 186 bp of the ORF stop codon
	XP_645284.1	1672253 - 1673699 GB: NC_007088.5	No	No
Citrate synthase	XP_643860.1	81916- 85716 GB:NC_007088.5	No	No
	XP_642824.1	7216750_7218586 GB:NC_007088.5	No	No
	XP_647596.1	1569279_1571100 GB:NC_007087.3	No	No
6-phosphogluconate dehydrogenase	XP_642122.1	273541_ 275022 GB: NC_007089.4	No	No
Hydroxypyruvate reductase	XP_629831.1	1190069- 1191163 GB: NC_007092.3	No	No
Superoxide dismutase	XP_645815.1	35093-36256 GB: NC_007088.5	No	No
	XP_642754.1	7803232-7804074 GB: NC_007088.5	No	No
	XP_647129.1	338107-338878 GB: NC_007087.3	No	No
	XP_635813.1	3992463-3992921 GB: NC_007091.3	Yes	No
	XP_640693.1	4537876-4538331 GB: NC_007089.4	Yes	No
	XP_640716.1	4612258-4612782 GB: NC_007089.4	Yes	No
Hydroxymethylglutaryl-CoA lyase	XP_638884.1	1145359- 1146736 GB: NC_007090.3	No	No
Catalase	XP_643894.1	3823148-3824719 GB: NC_007088.5	Yes	No
	XP_646153.1	3244376- 3246648 GB: NC_007087.3	No	No
Epoxide hydrolase	XP_637802.2	4234344- 4235648 GB: NC_007090.3	No	No
Alanine-glyoxylate aminotransferase	XP_001134511.1	4489219- 4490816 GB: NC_007091.3	No	No
glycerol-3-phosphate dehydrogenase (NAD+)	-	-	-	-
glucose-6-phosphate dehydrogenase	XP_644436.1	3207904 - 3209791 GB: NC_007088.5	No	No

NAD-dependent L-lactate dehydrogenase A	-	-	-	-
Aldehyde reductase	XP_628918.1	3404974 - 3406128 GB: NC_007092.3	No	No
D-ribulose-5-phosphate 3-epimerase	XP_642255.1	603034- 603814 GB: NC_007089.4	No	No
Isocitrate lyase	XP_644989.1	2236569- 2237975 GB: NC_007088.5	No	No
Malate synthase	XP_643552.2	5903117 - 5904832 GB: NC_007088.5	Yes	No
	XP_639287.1	72206 - 73834 GB: NC_007090.3	Yes	No

Appendix 27

Table 9.25. *Monosiga brevicollis* glycolysis enzyme. Red coloured proteins indicate the possibility of PTS1 in the protein C-terminus. Dash symbol represents the failure to recover protein homologues using a BLAST search.

Glycolysis enzyme	Accession number in *Monosiga brevicollis*	Location in the genome	PTS1 in ORF	Mechanism for dual localization	Stop codon
Glucokinase	XP_001750338.1	298402 - 300899 GB:NW_001865076.1	No	No	**T**-TAA-**A**
Glucose-6-phosphate isomerase	XP_001749895.1	351756- 356219 GB:NW_001865071.1	No	No	**G**-TAA-**A**
Phosphofructokinase	XP_001742461.1	3369718- 3374398 GB:NW_001865039.1	No	No	**C**-TGA-**A**
Triosephosphate isomerase	XP_001744473.1	375957- 377327 GB:NW_001865043.1	No	Yes, Ribosomal read-through	**C**-TAG-**G**
Glyceraldehyde-3-phosphate dehydrogenase	XP_001747972.1	593440- 595204 GB:NW_001865056.1	No	No	**G**-TAA-**A**
Phosphoglycerate kinase	XP_001745299.1	27276- 30105 GB:NW_001865045.1	No	No	**A**-TAA-**G**
Phosphoglycerate mutase	XP_001747083.1	700107 - 702156 GB:NW_001865052.1	No	No	**G**-TAG-**A**
Phosphopyruvate hydratase	-	-	-	-	-
Fructose-bisphosphate aldolase	-	-	-	-	-
Pyruvate kinase	XP_001743459.1	1751193 - 1753809 GB:NW_001865040.1	No	No	**C**-TAA-**G**

Table 9.26. *Monosiga brevicollis* proteins homologues to the proteins mentioned in Stiebler et al, 2014 and Ast et al,2013. Red coloured proteins indicate the possibility of PTS1 in the protein C-terminus. Dash symbol represents the failure to recover protein homologues using a BLAST search.

Enzyme (ORF)	Accession number in *Monosiga brevicollis*	Location in the genome	~PTS1 motif in ORF	Mechanism for dual localization
Malate dehydrogenase	XP_001744311.1	787052- 788650 GB:NW_001865042.1	No	Yes, Ribosomal read-through
	XP_001750413.1	291340 - 293037 GB:NW_001865077.1	No	No
Isocitrate dehydrogenase NADPH	XP_001743390.1	1210723 - 1214627 GB:NW_001865040.1	No	No
Citrate synthase	XP_001743409.1	1332321 - 1338224 GB:NW_001865040.1	No	No
6-phosphogluconate dehydrogenase	XP_001743751.1	1537280_ 1540041 GB:NW_001865041.1	No	No
Hydroxypyruvate reductase	XP_001742459.1	3355588- 3357421 GB:NW_001865039.1	No	No
Superoxide dismutase	XP_001750075.1	154907- 156591 GB:NW_001865073.1	Yes	No
	EDQ87759.1	275644- 280995 GB: CH991557.1	No	No
Hydroxymethylglutaryl-CoA lyase	XP_001748375.1	154129- 156916 GB:NW_001865059.1	No	No
Catalase	XP_001744568.1	1209815 - 1211815 GB:NW_001865043.1	No	No
Epoxide hydrolase	XP_001749708.1	197495- 198619 GB:NW_001865069.1	Yes	No
Alanine-glyoxylate aminotransferase	XP_001749244.1	391801- 393703 GB:NW_001865065.1	No	No
	XP_001750972.1	134046- 135432 GB:NW_001865089.1	No	No
	XP_001746664.1	149213- 152083 GB:NW_001865050.1	No	No
glycerol-3-phosphate dehydrogenase (NAD+)	XP_001746022.1	75965- 78238 GB:NW_001865048.1	No	No
glucose-6-phosphate dehydrogenase	XP_001745653.1	418191 - 420866 GB:NW_001865046.1	No	No
NAD-dependent L-lactate dehydrogenase A	XP_001745899.1	166028 - 168098 GB:NW_001865047.1	No	Yes, Alternative splicing frame 3 after 531 bp of the ORF stop codon
Aldehyde reductase	XP_001745978.1	898619- 901044 GB:NW_001865047.1	No	No
	XP_001750041.1	898619- 901044 GB:NW_001865072.1	No	No
	XP_001744280.1	426067- 427612 GB:NW_001865042.1	No	No
D-ribulose-5-phosphate 3-epimerase	-	-	-	-
Isocitrate lyase	XP_001745522.1	508847- 512163 GB:NW_001865046.1	No	No

Malate synthase	EDQ90026.1	1108719 - 1110029 GB: CH991549.1	No	No
	EDQ85515.1	174333- 178150 GB: CH991573.1	No	No

Appendix 28

Table 9.27. *Thecamonas trahens* glycolysis enzyme. Red coloured proteins indicate the possibility of PTS1 in the protein C-terminus. Dash symbol represents the failure to recover protein homologues using a BLAST search.

Glycolysis enzyme	Accession number in *T. trahens*	Location in the genome	PTS1 in ORF	Mechanism for dual localization	Stop codon
Hexokinase	-	-	-	-	-
Glucokinase	-	-	-	-	-
Glucose-6-phosphate isomerase	XP_013762808	505554- 507353 GB: NW_013657678.1	No	No	**C**-TAA-**A**
Phosphofructokinase	XP_013761043.1	98645 - 101024 GB: XP_013761043.1	No	No	**T**-TGA-**T**
Triosephosphate isomerase	XP_013760960	498068- 498914 GB: NW_013657671.1	No	No	**G**-TAG-**C**
Glyceraldehyde-3-phosphate dehydrogenase	XP_013760051.1	79178- 82921 GB: NW_013657668.1	No	No	**G**-TAG-**C**
	XP_013756240.1	238059- 239526 GB: NW_013657646.1	No	No	**G**-TAG-**C**
	XP_013760315.1	122269- 123888 GB: NW_013657669.1	No	No	**G**-TAA-**G**
Phosphoglycerate kinase	XP_013759818.1	105762- 107686 GB: XP_013759818.1	No	Yes, Ribosomal read-through	**G**-TGA-**G**
	XP_013752926.1	74531- 75897 GB: NW_013657610.1	No	No	**G**-TAG-**G**
Phosphoglycerate mutase	XP_013759515.1	451639 - 452604 GB: NW_013657665.1	No	No	**C**-TAG-**C**
	XP_013760621.1	286923 - 288058 GB: NW_013657670.1	No	No	**G**-TGA-**G**
Phosphopyruvate hydratase	XP_013752878.1	106834- 108328 GB: NW_013657608.1	No	No	**G**-TAA-**C**
	XP_013755533.1	110716- 112066 GB: NW_013657641.1	No	Yes, Ribosomal read-through	**C**-TAA-**G**
Fructose-bisphosphate aldolase	XP_013758981.1	3777877- 3779287 GB: NW_013657663.1	No	No	**C**-TGA-**G**
Pyruvate kinase	XP_013753091.1	20581 - 22860 GB: NW_013657614.1	No	No	**G**-TAG-**T**
	XP_013756197.1	115062-116670 GB: NW_013657646.1	No	No	**G**-TAG-**C**

Table 9.28. *Thecamonas trahens* proteins homologues to the proteins mentioned in Stiebler et al, 2014 and Ast et al,2013. Red coloured proteins indicate the possibility of PTS1 in the protein C-terminus. Dash symbol represents the failure to recover protein homologues using a BLAST search.

Enzyme (ORF)	Accession number in *T. trahens*	Location in the genome	~PTS1 motif in ORF	Mechanism for dual localization
Malate dehydrogenase	XP_013754559	193224 - 194249 GB:NW_013657632.1	Yes	No
	XP_013759779.1	3616 - 4739 GB:NW_013657667.1	No	No
Isocitrate dehydrogenase NADPH	XP_013755294.1	135615- 137330 GB:NW_013657639.1	No	No
Citrate synthase	XP_013756098.1	178427- 180458 GB:NW_013657645.1	No	No
	XP_013756103.1	7216750_7218586 GB:NW_013657645.1	Yes	No
6-phosphogluconate dehydrogenase	XP_013754421.1	117322_ 119095 GB:NW_013657631.1	No	No
Hydroxypyruvate reductase	XP_013756837.1	5548- 8640 GB:NW_013657651.1	No	No
Superoxide dismutase	XP_013761803.1	89221- 90224 GB:NW_013657675.1	No	NA
	XP_013753261.1	15448- 16041 GB:NW_013657617.1	No	No
	XP_013759336.1	465332- 466252 GB:NW_013657664.1	No	Yes, by Ribosomal read-through
	XP_013759345.1	482022-483288 GB:NW_013657664.1	No	No
	XP_013754534.1	117546-119592 GB:NW_013657632.1	No	No
	XP_013762572.1	641703-644344 GB:NW_013657677.1	No	No
Hydroxymethylglutaryl-CoA lyase	XP_013754841.1	87641- 89169 GB:NW_013657635.1	No	No
Catalase	XP_013754323.1	97425- 100078 GB:NW_013657630.1	Yes	No
	XP_013754281.1	51990- 54257 GB:NW_013657629.1	No	No
Epoxide hydrolase	XP_013760667.1	374740 - 376021 GB:NW_013657670.1	Yes	No
Alanine-glyoxylate aminotransferase	XP_013753060.1	97523- 99124 GB:NW_013657613.1	No	No
	XP_013758727.1	97523- 99124 GB:NW_013657662.1	No	No
	XP_013758726.1	139743- 142517 GB:NW_013657662.1	No	No
glycerol-3-phosphate dehydrogenase (NAD+)	XP_013761830.1	374740 - 376021 GB:NW_013657675.1	Yes	-
glucose-6-phosphate dehydrogenase	XP_013754640.1	104115 - 105749 GB:NW_013657633.1	Yes	-
NAD-dependent L-lactate dehydrogenase A	-	-	-	-

Aldehyde reductase	XP_013754132.1	58048- 58968 GB:NW_013657628.1	No	No
	XP_013757368.1	186321- 187382 GB:NW_013657654.1	No	No
	XP_013756146.1	330885- 331846 GB:NW_013657645.1	No	No
D-ribulose-5-phosphate 3-epimerase	XP_013761450.1	552652- 553686 GB:NW_013657673.1	No	Yes, Alternative splicing frame 2 after 108 bp of the ORF stop codon
Isocitrate lyase	-	-	-	-
Malate synthase	XP_013752844.1	62398- 68889 GB:NW_013657607.1	No	No

Appendix 29

Table 9.29. *Vitrella brassicaformis* glycolysis enzyme. Red coloured proteins indicate the possibility of PTS1 in the protein C-terminus.

Glycolysis enzyme	Accession number in *V. brassicaformis*	Location in the genome	PTS1 in ORF	Mechanism for dual localization
Hexokinase	CEL99948.1	89924- 94238 GB: CDMY01000292	No	No
Glucose-6-phosphate isomerase	CEL97776.1	608789- 612033 GB: CDMY01000255	No	No
Phosphofructokinase	CEM01769.1	113765- 119822 GB: CDMY01000309	No	No
	CEL94756.1	99489- 106161 GB: CDMY01000226	No	No
Triosephosphate isomerase	CEM22592.1	297801- 299440 GB:CDMY01000553	No	No
	CEM14871.1	9705- 11063 GB:CDMY01000460	No	No
Glyceraldehyde-3-phosphate dehydrogenase	CEM36876.1	181725- 184268 GB: CDMY01000908	No	No
	CEM11865.1	25463- 27816 GB: CDMY01000427	No	-
Phosphoglycerate kinase	CEM12713.1	303671 - 306368 GB:CDMY01000436	No	No
	CEM26305.1	135345 - 137393 GB:CDMY01000613	No	No
Phosphoglycerate mutase	CEM32665.1	110056- 111663 GB: CDMY01000759	No	No
	CEM05605.1	35367- 38566 GB: CDMY01000361	No	No
Phosphopyruvate hydratase	CEL93148.1	5500 - 7273 GB: CDMY01000144	No	No
	CEL96994.1	95779 - 97918 GB: CDMY01000251	No	No
Fructose-bisphosphate aldolase	CEM05410.1	64014- 66015 GB: CDMY01000357	No	No
	CEM21222.1	19055- 20983 GB: CDMY01000538	No	No
	CEM04838.1	30529- 32293 GB: CDMY01000356	No	No
	CEM17675.1	63525- 65131 GB: CDMY01000493	No	No
Pyruvate kinase	CEL92797.1	22553- 24649 GB: CDMY01000113	No	No
	CEL94407.1	126230- 129092 GB: CDMY01000218	No	No

Table 9.30. *Vitrella brassicaformis* proteins homologues to the proteins mentioned in Stiebler et al, 2014 and Ast et al,2013. Red coloured proteins indicate the possibility of PTS1 in the protein C-terminus. Dash symbol represents the failure to recover protein homologues using a BLAST search.

Enzyme (ORF)	Accession number in *V. brassicaformis*	Location in the genome	~PTS1 motif in ORF	Mechanism for dual localization
Malate dehydrogenase	CEL99279.1	9120- 11864 GB: CDMY01000277	No	No
	CEM30509.1	82998- 85019 GB: CDMY01000698	Yes	No
	CEL96690.1	579440- 580883 GB: CDMY01000243	No	No
Isocitrate dehydrogenase NADPH	CEL96761.1	148709- 150391 GB: CDMY01000244	No	No
Citrate synthase	CEM24004.1	37120- 39320 GB: CDMY01000575	No	No
	CEL98102.1	134286- 137765 GB: CDMY01000262	No	No
	CEM02703.1	83227- 86219 GB: CDMY01000333	No	No
6-phosphogluconate dehydrogenase	CEL92579.1	39684- 41294 GB: CDMY01000091	No	No
	CEM25244.1	18373- 19972 GB: CDMY01000593	No	No
	CEL95590.1	113670- 115550 GB: CDMY01000231	No	No
Hydroxypyruvate reductase	CEL95534.1	111975- 114784 GB: CDMY01000228	No	No
	CEM11643.1	46678- 49764 GB: CDMY01000421	No	No
	CEL95648.1	138266- 140140 GB: CDMY01000234	No	No
	CEL95650.1	143319- 145163 GB: CDMY01000234	No	No
Hydroxymethylglutaryl-CoA lyase	AND95672.1	184- 1218 GB: KR704762	No	No
Superoxide dismutase	CEM02146.1	47910- 48923 GB: CDMY01000319	No	No
	CEM24553.1	29891- 32755 GB: CDMY01000582	No	No
	CEM23297.1	12186- 13088 GB: CDMY01000563	No	No
	CEM28879.1	25043- 26143 GB: CDMY01000663	No	No
	CEM36361.1	96917- 98957 GB: CDMY01000887	No	No
	CEM37809.1	148920- 149947 GB: CDMY01000954	No	No
Catalase	CEM23795.1	193703 - 5898 GB: CDMY01000571	No	No
Epoxide hydrolase	-	-	-	-

Alanine-glyoxylate aminotransferase	CEL94492.1	57928- 60360 GB: CDMY01000220	No	No
	CEM36795.1	17293- 19003 GB: CDMY01000908	No	No
	CEM28848.1	13128- 16100 GB: CDMY01000662	No	No
glycerol-3-phosphate dehydrogenase (NAD+)	CEM04071.1	170561- 172295 GB: CDMY01000347	No	No
	CEM03367.1	38139- 39982 GB:CDMY01000339	No	No
	CEM31911.1	194002- 195944 GB:CDMY01000738	No	No
glucose-6-phosphate dehydrogenase	CEM02434.1	68804- 71981 GB:CDMY01000326	No	No
	CEM39713.1	206479- 211445 GB:CDMY01001045	No	No
NAD-dependent L-lactate dehydrogenase A	-	-	-	-
Aldehyde reductase	CEL92451.1	138287- 139978 GB:CDMY01000077	No	No
	CEL97053.1	115156- 116990 GB:CDMY01000252	No	No
	CEL97018.1	23466- 24848 GB:CDMY01000252	No	No
	CEM29212.1	269770- 271024 GB:CDMY01000666	No	No
	CEM18750.1	20556- 22063 GB:CDMY01000508	No	No
D-ribulose-5-phosphate 3-epimerase	CEL97044.1	90659- 91612 GB: CDMY01000252	No	No
	CEL94013.1	96067- 97117 GB: CDMY01000201	No	No
	CEL93920.1	169303- 170346 GB: CDMY01000198	No	No
Isocitrate lyase	CEM20334.1	20791- 23921 GB:CDMY01000525	No	No
Malate synthase	CEM21864.1	1720- 4773 GB: CDMY01000544	Yes	No
	CEM35133.1	12880- 23921 GB:CDMY01000842	No	No

Appendix 30

Table 9.31. *Paramecium tetraurelia* glycolysis enzyme. Red coloured proteins indicate the possibility of PTS1 in the protein C-terminus.

Glycolysis enzyme	Accession number in *P. tetraurelia*	Location in the genome	PTS1 in ORF	Mechanism for dual localization
Glucokinase	XP_001456377.1	268318- 269558 GB: NW_001799622	No	No
Glucose-6-phosphate isomerase	XP_001443059.1	422458- 424308 GB: NW_001799178	No	Yes, Alternative splicing frame 3 after 433 bp of the ORF stop codon
Phosphofructokinase	XP_001444382.1	392280- 394046 GB: NW_001799222	No	No
	XP_001439166.1	9347- 11114 GB: NW_001799076	No	No
	XP_001457102.1	80297- 82011 GB: NW_001799626	No	Yes, Alternative splicing frame 3 after 110 bp of the ORF stop codon
	XP_001454725.1	43898- 45612 GB: NW_001799588	No	No
Triosephosphate isomerase	XP_001434096.1	52095- 52999 GB: NW_001799017	No	No
	XP_001425288.1	234030- 234934 GB: NW_001798964	No	No
	XP_001430966.1	52661- 53456 GB: NW_001798997	No	No
Glyceraldehyde-3-phosphate dehydrogenase	XP_001449912.1	294007- 295249 GB:NW_001799400	Yes	No
	XP_001445566.1	330452- 331731 GB:NW_001799266	Yes	No
Phosphoglycerate kinase	XP_001427410.1	89985 - 91340 GB: NW_001798976	No	No
	XP_001457525.1	64647 - 66008 GB: NW_001799628	No	No
	XP_001461438.1	219628 - 221138 GB: NW_001799645	No	No
Phosphoglycerate mutase	XP_001437741.1	10315- 10990 GB: NW_001799049	No	No
	XP_001445811.1	185584- 186313 GB: NW_001799277	No	No
Phosphopyruvate hydratase	XP_001452406.1	342060 - 343540 GB: NW_001799499	No	No
	XP_001423791.1	444582 - 446103 GB: NW_001798956	No	No
Fructose-bisphosphate aldolase	XP_001458636.1	88744- 89961 GB:NW_001799632.1	No	No
	XP_001460782.1	43253- 44530 GB:NW_001799642	No	No
	XP_001457442.1	296034- 297364 GB:NW_001799627	No	No
Pyruvate kinase	XP_001444997.1	397336- 398961 GB: NW_001799244	No	No

	XP_001439246.1	118778- 120435 GB: NW_001799078	No	No
	XP_001425796.1	231167- 232863 GB: NW_001798967	No	No

Table 9.32. *Paramecium tetraurelia* proteins homologues to the proteins mentioned in Stiebler et al, 2014 and Ast et al,2013. Red coloured proteins indicate the possibility of PTS1 in the protein C-terminus. Dash symbol represents the failure to recover protein homologues using a BLAST search.

Enzyme (ORF)	Accession number in *P. tetraurelia*	Location in the genome	~PTS1 motif in ORF	Mechanism for dual localization
Malate dehydrogenase	XP_001431831.1	109885- 110990 GB: NW_001799001	No	No
	XP_001446847.1	245170- 246313 GB: NW_001799300	No	No
	XP_001442138.1	288917- 289982 GB: NW_001799166	No	No
	XP_001434590.1	100718- 101857 GB:NW_001799021	No	No
	XP_001456694.1	65006- 66179 GB:NW_001799624	No	No
	XP_001424905.1	172730- 173916 GB:NW_001798962	No	No
	XP_001425443.1	200489- 201701 GB:NW_001798965	No	No
	XP_001451278.1	363582- 364799 GB:NW_001799455	No	No
	XP_001451278.1	207266- 207924 GB:NW_001798988	No	No
Isocitrate dehydrogenase NADPH	XP_001460949.1	315903-317026 GB:NW_001799642	No	No
	XP_001457274.1	8515-9679 GB:NW_001799627	No	No
	XP_001444582.1	219563-220790 GB:NW_001799233	No	No
	XP_001445240.1	276578-277773 GB:NW_001799255	No	No
	XP_001450353.1	162942-164158 GB:NW_001799422	No	No
Citrate synthase	XP_001426183.1	207705- 209222 GB: NW_001798968	No	No
	XP_001458389.1	40727- 42292 GB: NW_001799631	No	No
	XP_001447471.1	262245- 263601 GB: NW_001799322	Yes	No
	XP_001458120.1	317012- 318486 GB: NW_001799630	No	No
	XP_001453830.1	225617-227125 GB:NW_001799544	No	No
6-phosphogluconate dehydrogenase	XP_001428786.1	177583- 178461 GB:NW_001798983	No	No
	XM_001460070.1	233648- 234553 GB:NW_001799639	No	No
Hydroxypyruvate reductase	XP_001439222.1	68728- 69899 GB:NW_001799078	No	No
Hydroxymethylglutaryl-CoA lyase	XP_001426433.1	83721- 84832 GB: NW_001798970	No	No
Superoxide dismutase	XP_001425110.1	203607- 204364 GB: NW_001798963	No	No

	XP_001461921.1	81451- 82107 GB: NW_001799648	No	No
	XP_001433955.1	176296- 177001 GB: NW_001799015	No	No
Catalase	XP_001453137.1	381481 - 383025 GB: NW_001799511	Yes	No
Epoxide hydrolase	-	-	-	-
Alanine-glyoxylate aminotransferase	XP_001454869.1	282783- 284211 GB: NW_001799588	No	No
	XP_001457245.1	311807- 313203 GB: NW_001799626	No	No
glycerol-3-phosphate dehydrogenase (NAD+)	XP_001448978.1	363382- 364490 GB: NW_001799377	No	No
glucose-6-phosphate dehydrogenase	-	-	-	-
NAD-dependent L-lactate dehydrogenase	-	-	-	-
Aldehyde reductase	XP_001347021.1	429088- 429992 GB:NC_006058	No	No
D-ribulose-5-phosphate 3-epimerase	XP_001459009.1	19965- 20612 GB: NW_001799634	No	No
Isocitrate lyase	XP_001453133.1	374698- 376406 GB: NW_001799511	No	No
Malate synthase	-	-	-	-

Appendix 31

Table 9.33. *Chlamydomonas reinhardtii* glycolysis enzyme. Red coloured proteins indicate the possibility of PTS1 in the protein C-terminus**.**

Glycolysis enzyme	Accession number in *C. reinhardtii*	Location in the genome	PTS1 in ORF	Mechanism for dual localization
Glucokinase	XP_001689605.1	2823654- 2827142 GB: NW_001843471	No	No
Glucose-6-phosphate isomerase	XP_001703279.1	635107 - 641699 GB: NW_001843987	No	No
Phosphofructokinase	XP_001696306.1	1422492 - 1427381 GB: NW_001843733	No	No
	XP_001694148	1569856 - 1576122 GB: NW_001843643	No	No
	XP_001696305	1422492 - 1427381 GB: NW_001843733	No	No
	XP_001701413.1	179253 - 187961 GB: NW_001843880	No	No
Triosephosphate isomerase	XP_001690035	2419745 - 2421552 GB: NW_001843471	No	No
Glyceraldehyde-3-phosphate dehydrogenase	XP_001702068.1	113583 - 117293 GB: NW_001843892	No	No
	XP_001689871.1	99942- 102567 GB: NW_001843471	No	No
	XP_001703200.1	214981- 218199 GB: NW_001843980	No	No
Phosphoglycerate kinase	XP_001699523.1	209885 - 213534 GB: NW_001843825	No	No
	XP_001702049.1	117430- 121899 GB: NW_001843892	No	No
Phosphoglycerate mutase	XP_001702950.1	163281- 164764 GB:NW_001843951	No	No
Phosphopyruvate hydratase	XP_001702971.1	17749 - 21610 GB: NW_001843955	No	No
Fructose-bisphosphate aldolase	XP_001700659.1	489705- 491316 GB: NW_001843863	No	No
	XP_001700318.1	518552- 524317 GB: NW_001843843	No	No
	XP_001701797.1	2606845- 2610546 GB: NW_001843888	No	No
	XP_001699879.1	2606845- 2610546 GB: NW_001843834	No	No
Pyruvate kinase	XP_001693008.1	846204- 852147 GB:NW_001843629	No	No
	XP_001700637.1	71711- 84137 GB:NW_001843863	No	No
	XP_001695738.1	1147789- 1162124 GB:NW_001843709	No	No
	XP_001694900.1	1007238- 1013529 GB:NW_001843688	No	No
	XP_001695439.1	331202- 337261 GB:NW_001843705	No	No

Table 9.34. *Chlamydomonas reinhardtii* proteins homologues to the proteins mentioned in Stiebler et al, 2014 and Ast et al,2013. Red coloured proteins indicate the possibility of PTS1 in the protein C-terminus. Dash symbol represents the failure to recover protein homologues using a BLAST search.

Enzyme (ORF)	Accession number in *C. reinhardtii*	Location in the genome	~PTS1 motif in ORF	Mechanism for dual localization
Malate dehydrogenase	XP_001694886.1	761660- 764513 GB:NW_001843688.1	No	No
	XP_001696786.1	375660- 381169 GB: NW_001843734	No	No
Isocitrate dehydrogenase NADPH	XP_001694857.1	339920- 344294 GB: NW_001843688	No	No
	XP_001697281.1	1112368- 1117334 GB: NW_001843747	No	Yes, ribosomal read through
	XP_001701074.1	2911607- 2917269 GB: NW_001843867	No	No
Citrate synthase	XP_001702983.1	213754- 219633 GB: NW_001843955	No	No
	XP_001695571.1	383281- 388201 GB: NW_001843709	No	No
6-phosphogluconate dehydrogenase	XP_001698018.1	725919- 728871 GB: NW_001843783	No	No
	XP_001691921.1	1274264- 1276107 GB: NW_001843572	No	No
Hydroxypyruvate reductase	XP_001691480.1	2196553- 2201874 GB: NW_001843537	No	No
Hydroxymethylglutaryl-CoA lyase	XP_001703205.1	277166- 280578 GB: NW_001843980	No	No
Superoxide dismutase	XP_001699077.1	424939- 427558 GB: NW_001843809	No	No
	XP_001700058.1	2771090- 2773307 GB: NW_001843834	No	Yes, ribosomal read through
	XP_001690591.1	1209718- 1211033 GB: NW_001843472	No	No
	XP_001690936.1	2013106- 2015867 GB: NW_001843510	No	No
	XP_001695947.1	901389- 904798 GB: NW_001843718	No	No
Catalase	XP_001696763.1	1312932 - 1317626 GB: NW_001843734	No	No
Epoxide hydrolase	-	-	-	-
Alanine-glyoxylate aminotransferase	XP_001698245.1	631741 - 636836 GB: NW_001843787	No	Yes, Alternative splicing frame 3 after 424bp of the ORF stop codon
Glycerol-3-phosphate dehydrogenase (NAD+)	XP_001690876.1	1272836 - 1276118 GB: NW_001843510	No	No
	XP_001689796.1	5448829 - 5460120 GB: NW_001843471	No	No

glucose-6-phosphate dehydrogenase	PNW80105.1	3891413 - 3898200 GB: CM008969	No	No
NAD-dependent L-lactate dehydrogenase A	-	-	-	-
Aldehyde reductase	XP_001694768.1	531642 - 535754 GB: NW_001843677	No	No
	PNW77353.1	2040956 - 2044928 GB: CM008971.1	No	No
	XP_001699353.1	286218- 289494 GB: NW_001843817	No	No
	XP_001696169.1	24338- 26609 GB: NW_001843733	No	No
D-ribulose-5-phosphate 3-epimerase	PNW87282.1	1336314 - 1339173 GB: CM008963.1	No	No
	XP_001691071.1	6497140 - 6499069 GB: NW_001843510	No	No
Isocitrate lyase	XP_001695331.1	521128 - 525138 GB: NW_001843705	No	No
Malate synthase	XP_001695632.1	1164993 - 1170435 GB: NW_001843709	Yes	No

Appendix 32

Table 9.35. *Tetrahymena thermophila* glycolysis enzyme. Red coloured proteins indicate the possibility of PTS1 in the protein C-terminus**.**

Glycolysis enzyme	Accession number in *T. thermophila*	Location in the genome	PTS1 in ORF	Mechanism for dual localization
Glucokinase	XP_001019344.2	279859- 281736 GB:NW_002476358	No	No
Glucose-6-phosphate isomerase	XP_001030231.2	81528- 84150 GB:NW_002476179	No	No
	XP_001030227.2	69993- 70689 GB:NW_002476179	No	No
	XP_001030229.2	75790- 77195 GB:NW_002476179	No	No
Phosphofructokinase	XP_001008692.1	222597- 225496 GB:NW_002476554	No	No
	XP_001017609.2	397653- 400156 GB:NW_002476380	No	No
	XP_001017610.3	400678- 403430 GB:NW_002476380	No	No
Triosephosphate isomerase	XP_001008794.2	626978- 628222 GB:NW_002476554	No	No
Glyceraldehyde-3-phosphate dehydrogenase	XP_001009209.2	208118- 209903 GB:NW_002476542	No	No
	XP_001009209.2	255402- 256663 GB:NW_002476542	No	-
Phosphoglycerate kinase	XP_001025708.1	161326 - 163320 GB:NW_002476082	No	No
Phosphoglycerate mutase	XP_001021628.2	573855- 574970 GB:NW_002476317	No	No
	XP_001021629.1	575640- 577033 GB:NW_002476317	No	No
Phosphopyruvate hydratase	XP_001014643.1	323515 - 325912 GB:NW_002476426	No	No
Fructose-bisphosphate aldolase	XP_001032921.1	489705- 491316 GB:NW_002476301	No	No
Pyruvate kinase	XP_001010760.3	1023523- 1026100 GB:NW_002476512	No	No
	XP_001011238.2	867824- 870918 GB:NW_002476507	No	No
	XP_001029957.2	23858- 26519 GB:NW_002476197	No	No

Table 9.36. *Tetrahymena thermophila* proteins homologues to the proteins mentioned in Stiebler et al, 2014 and Ast et al,2013. Red coloured proteins indicate the possibility of PTS1 in the protein C-terminus. Dash symbol represents the failure to recover protein homologues using a BLAST search.

Enzyme (ORF)	Accession number in *T. thermophila*	Location in the genome	~PTS1 motif in ORF	Mechanism for dual localization
Malate dehydrogenase	XP_001014922	1370305- 1371608 GB:NW_002476426	No	No
	XP_001024780	72900- 73964 GB:NW_002476157	No	No
	XP_001015159	86916- 89864 GB:NW_002476424	No	No
	XP_976916.1	702946- 704700 GB:NW_002476434	No	No
Isocitrate dehydrogenase NADPH	XP_001031817	160564-162953 GB:NW_002476146	No	No
	XP_001018448	284020-286760 GB:NW_002476368	No	No
Citrate synthase	XP_001032690	109883- 112143 GB:NW_002476236	No	No
	XP_001021439	599605- 601435 GB:NW_002476319	No	No
	XP_001032196	130501-132632 GB:NW_002476216	No	No
	XP_001014802	918510-920989 GB:NW_002476426	No	No
	XP_001023516	332771-334294 GB:NW_002476209	No	Yes, Alternative splicing frame 2 after 352 bp of the ORF stop codon
6-phosphogluconate dehydrogenase	XP_001032191	119961- 121029 GB:NW_002476216	No	No
Hydroxypyruvate reductase	-	-	-	-
Hydroxymethylglutaryl-CoA lyase	XP_001025212	289494- 291161 GB:NW_002476149	No	No
Superoxide dismutase	XP_001010506	1804498- 1805371 GB:NW_002476463	No	No
	XP_001033543	651263- 651817 GB:NW_002476212	No	No
	XP_001032187	110723- 111551 GB:NW_002476216	No	No
	XP_001007667	807053- 808243 GB:NW_002476567	No	No
Catalase	XP_001026590	3087 - 5898 GB:NW_002476012	Yes	No
Epoxide hydrolase	-	-	-	-
Alanine-glyoxylate aminotransferase	XP_001025527	9429- 11219 GB:NW_002476089	No	No
	XP_001022163	175626- 177122 GB:NW_002476266	No	No
glycerol-3-phosphate dehydrogenase (NAD+)	XP_012655009	253380- 254833 GB:NW_002476236	No	No

	XP_001032725	250123- 252535 GB:NW_002476236	No	Yes, Alternative splicing frame 2 after 52bp of the ORF stop codon
glucose-6-phosphate dehydrogenase	-	-	-	-
NAD-dependent L-lactate dehydrogenase	-	-	-	-
Aldehyde reductase	More than 20 different homologous genes	-	-	-
D-ribulose-5-phosphate 3-epimerase	XP_001023165.2	161833- 162825 GB:NW_002476226	No	No
Isocitrate lyase	XP_001030129.2	314936- 316722 GB:NW_002476399	No	No
Malate synthase	XP_001016067.2	77659- 80055 GB:NW_002476416	No	No

Appendix 33

Table 9.37. *Toxoplasma gondii* glycolysis enzyme. Red coloured proteins indicate the possibility of PTS1 in the protein C-terminus.

Glycolysis enzyme	Accession number in Toxoplasma gondii	PTS1 in ORF	Mechanism for dual localization
Hexokinase	TGVEG_265450	No	No
Glucose-6-phosphate isomerase	TGVEG_283780	No	No
Phosphofructokinase	TGVEG_226960	No	Yes, Alternative splicing frame 2 after 586 bp of the ORF stop codon
	TGVEG_240890	No	No
	TGVEG_281400A	No	No
Triosephosphate isomerase	TGVEG_225930	No	No
	TGVEG_233500	No	No
Glyceraldehyde-3-phosphate dehydrogenase	TGVEG_289690	No	No
	TGVEG_269190	Yes	Yes, Alternative splicing frame 2 after 327 bp of the ORF stop codon
Phosphoglycerate kinase	TGVEG_318230	No	No
	TGVEG_222020	No	No
Phosphoglycerate mutase	TGVEG_273030	No	No
	TGVEG_222910	No	No
Phosphopyruvate hydratase	TGVEG_268850	No	No
	TGVEG_268860	No	No
Fructose-bisphosphate aldolase	TGVEG_236050	No	Yes, Alternative splicing frame 1 after 282 bp of the ORF stop codon
	TGVEG_236040	No	No
Pyruvate kinase	TGVEG_256760	No	No
	TGVEG_299070	No	No

Table 9.38. *Toxoplasma gondii* proteins homologues to the proteins mentioned in Stiebler et al, 2014 and Ast et al,2013. Red coloured proteins indicate the possibility of PTS1 in the protein C-terminus. Dash symbol represents the failure to recover protein homologues using a BLAST search.

Enzyme (ORF)	Accession number in *Naegleria fowleri*	~PTS1 motif in ORF	Mechanism for dual localization
Malate dehydrogenase	TGVEG_318430	No	Yes, Alternative splicing frame 3 after 485 bp of the ORF stop codon
Isocitrate dehydrogenase NADPH	TGVEG_266760	No	No
	TGVEG_313140	No	No
Citrate synthase	TGVEG_268890	No	No
	TGVEG_203110	No	No
	TGVEG_263130	Yes	No
6-phosphogluconate dehydrogenase	TGVEG_307850	No	Yes, Alternative splicing frame 3 after 545 bp of the ORF stop codon
	TGVEG_242600	No	No
Hydroxypyruvate reductase	TGVEG_239820	No	No
Hydroxymethylglutaryl-CoA lyase	TGVEG_204460	No	No
Superoxide dismutase	TGVEG_316330	No	No
	TGVEG_316190	No	Yes, ribosomal read through
	TGVEG_316310	No	No
Catalase	TGVEG_232250	No	No
Epoxide hydrolase	TGVEG_261940	No	No
Alanine-glyoxylate aminotransferase	TGVEG_269110	No	No
glycerol-3-phosphate dehydrogenase (NAD+)	TGVEG_210260	No	No
	TGVEG_307575	No	Yes, Alternative splicing frame 3 after 773 bp of the ORF stop codon
glucose-6-phosphate dehydrogenase	TGVEG_278830	No	No
	TGVEG_294200	No	Yes, Alternative splicing frame 3 after 803bp of the ORF stop codon
NAD-dependent L-lactate dehydrogenase A	TGVEG_291040	No	No
	TGVEG_232350	No	No
Aldehyde reductase	TGVEG_273920	No	Yes, Alternative splicing frame 2 after 363bp of the ORF stop codon
D-ribulose-5-phosphate 3-epimerase	TGVEG_247670	No	No
Isocitrate lyase	TGVEG_293810	No	No
Malate synthase	-	-	-

Appendix 34

Table 9.39. *Cyanidioschyzon merolae* glycolysis enzyme. Red coloured proteins indicate the possibility of PTS1 in the protein C-terminus.

Glycolysis enzyme	Accession number in *C.owczarzaki*	Location in the genome	PTS1 in ORF	Mechanism for dual localization
Hexokinase	XP_004346803.2	619057-621083 GB:NW_011887299	No	No
Glucose-6-phosphate isomerase	XP_004343157.1	345661-348280 GB:KE346374.1	No	Yes, , Alternative splicing frame 2 after 122 bp of the ORF stop codon
Phosphofructokinase	KJE93532.1 One protein	1043897-1046976 GB:KE346365.1	No	No
Triosephosphate isomerase	XP_004345684.1	935354-937349 GB:NW_011887301.1	Yes	No
Glyceraldehyde-3-phosphate dehydrogenase	KJE89507.1	3091534-3094233 GB:KE346360.1	No	No
Phosphoglycerate kinase	XP_004364164.1	3091534-3094233 GB:KE346360.1	Yes	No
Phosphoglycerate mutase	KJE93296.1	418948- 420735 GB:KE346365.1	No	No
Phosphopyruvate hydratase	XP_004346165.1	318668- 321013 GB:KE346368.1	No	No
Fructose-bisphosphate aldolase	XP_011270262.1	643438-645465 GB:KE346363.1	No	No
Pyruvate kinase	XP_004345801.1 XP_004349468.1	2275100-2277729 GB: KE346362.1	No	No

Table 9.40. *Cyanidioschyzon merolae* proteins homologues to the proteins mentioned in Stiebler et al, 2014 and Ast et al,2013. Red coloured proteins indicate the possibility of PTS1 in the protein C-terminus. Dash symbol represents the failure to recover protein homologues using a BLAST search.

Enzyme (ORF)	Accession number in *C. merolae*	Location in the genome	~PTS1 motif in ORF	Mechanism for dual localization
Malate dehydrogenase	XP_005539548	1552156-1553199 GB:NC_010146.1	No	No
	XP_005537858	488848-490107 GB:NC_010142.1	No	No
Isocitrate dehydrogenase NADPH	XM_005539184	548334-549812 GB:NC_010146.1	No	No
	XP_005538902	683783-684949 GB:NC_010145.1	Yes	No
Citrate synthase	XP_005538125	469819-471267 GB:NC_010143.1	No	No
	XP_005535043	115333-116844 GB:NC_010127.1	Yes	No
	XP_005536987	144658-146655 GB:NC_010139.1	No	No
6-phosphogluconate dehydrogenase	XP_005538849	517364-518860 GB:NC_010145.1	No	No
	XP_005536718	127508-129433 GB:NC_010138.1	No	No
	XP_005537109	571763-573688 GB:NC_010139.1	No	No
	XP_005536703	81691-82878 GB:NC_010138.1	No	No
	XP_005536704	82571-83398 GB:NC_010138.1	No	No
Hydroxypyruvate reductase	XP_005538203	746208-747227 GB:NC_010143.1	No	No
	XP_005539020	1059938-1061107 GB:NC_010145.1	Yes	No
Hydroxymethylglutaryl-CoA lyase	XP_005535064	213411-214472 GB:NC_010127.1	No	No
Superoxide dismutase	XP_005538444	392571-393380 GB:NC_010144.1	No	No
	XP_005537203	51145-51819 GB:NC_010140.1	No	No
	XP_005539094	70562-71509 GB:NC_010146.1	No	Yes, Alternative splicing frame 2 after 340 bp of the ORF stop codon
Catalase	XP_005536240	147075-148565 GB:NC_010135.1	No	No
Epoxide hydrolase	-	-	-	-
Alanine-glyoxylate aminotransferase	XP_005538023	141166-142638 GB:NC_010143.1	No	No
glycerol-3-phosphate dehydrogenase (NAD+)	XP_005535472	284340-285413 GB:NC_010130.1	No	No
	XP_005536841	524890-526959 GB:NC_010138.1	No	No

glucose-6-phosphate dehydrogenase	XP_005538340	37543-39369 GB:NC_010144.1	No	No
	XP_005536380	605422-607164 GB:NC_010135.1	No	No
NAD-dependent L-lactate dehydrogenase A	XP_005536444.	14006-15070 GB:NC_010137.1	No	No
	XP_005534764	4509-5339 GB:NC_010136.1	No	No
Aldehyde reductase	XP_005537159	748367-749362 GB:NC_010139.1	No	No
	XP_005539308	769487-770809 GB:NC_010146.1	No	No
D-ribulose-5-phosphate 3-epimerase	XP_005537614	597500-598234 GB: NC_010141.1	No	Yes, Alternative splicing frame 2 after 82bp of the ORF stop codon
	XP_005539563	1607440-1608279 GB:NC_010146.1	No	No
	XP_005536266	242614-243240 GB:NC_010135.1	No	No
Isocitrate lyase	-	-	-	-
Malate synthase	-	-	-	-

Appendix 35

Table 9.41. *Capsaspora owczarzaki* glycolysis enzyme. Red coloured proteins indicate the possibility of PTS1 in the protein C-terminus.

Glycolysis enzyme	Accession number in *C.owczarzaki*	Location in the genome	PTS1 in ORF	Putative PTS1 via Alternative splicing	Putative PTS1 via Ribosomal read-through
Hexokinase	XP_004346803.2	619057-621083 GB:NW_011887299	No	No	No
Glucose-6-phosphate isomerase	XP_004343157.1	345661-348280 GB:KE346374.1	No	Yes frame 2 after 122 bp of the ORF stop codon	No
Phosphofructokinase	KJE93532.1 One protein	1043897-1046976 GB:KE346365.1	No	No	No
Triosephosphate isomerase	XP_004345684.1	935354-937349 GB:NW_011887301.1	Yes	No	No
Glyceraldehyde-3-phosphate dehydrogenase	KJE89507.1	3091534-3094233 GB:KE346360.1	No	No	No
Phosphoglycerate kinase	XP_004364164.1	3091534-3094233 GB:KE346360.1	Yes	No	No
Phosphoglycerate mutase	KJE93296.1	418948- 420735 GB:KE346365.1	No	No	No
Phosphopyruvate hydratase	XP_004346165.1	318668- 321013 GB:KE346368.1	No	No	No
Fructose-bisphosphate aldolase	XP_011270262.1	643438-645465 GB:KE346363.1	No	No	No
Pyruvate kinase	XP_004345801.1 XP_004349468.1	2275100-2277729 GB: KE346362.1	No	No	No

Table 9.42. *Capsaspora owczarzaki* proteins homologues to the proteins mentioned in Stiebler et al, 2014 and Ast et al,2013. Red coloured proteins indicate the possibility of PTS1 in the protein C-terminus.

Enzyme (ORF)	Accession number in *C.owczarzaki*	~PTS1 motif	Mechanism for dual localization	Location in the genome
Malate dehydrogenase	XP_004343487.1	No	No	489559-491509 GB:KE346373
Isocitrate dehydrogenase NADPH	KJE94760.1	Yes	Gene duplication	1343776-1348135 GB:KE346367
Citrate synthase	XP_004348777.1	Yes	Alternative splicing frame 3 after 154 bp of the ORF stop codon	81916- 85716 GB:KE346362
6-phosphogluconate dehydrogenase	XP_004343090.1	No	No	97536_99885 GB:KE346374
Hydroxypyruvate reductase	XP_004349870.1	No	No	553044-554517 GB: KE346361.1
	XP_004347486.2	No	No	788798-790163 GB: KE346366.1
Glyceraldehyde-3-phosphate dehydrogenase	XP_004365835.1	No	No	3091534-3094233 GB:KE346360
Superoxide dismutase	XP_004365015.1			507102-508478 GB: KE346360.1
Hydroxymethylglutaryl-CoA lyase	XP_004364803	NA	NA	2422416- 2423503 GB: KE346361.1
Catalase	XP_004349387	Yes	Alternative splicing frame 3 after 282 bp of the ORF stop codon	2018139- 2020772 GB: KE346362.1
Epoxide hydrolase	XP_004363725.2	No	No	264525-267198 GB: KE346364.1
Alanine-glyoxylate aminotransferase	XP_004363417	No	No	806337-809032 GB: KE346364.1
	XP_004346261	No	No	648554-650647 GB: KE346368.1
glycerol-3-phosphate dehydrogenase (NAD+)	XP_004364069.2	No	No	1420384-1423597 GB: KE346363.1
	XP_004349769.2	No	No	219880-221489 GB: KE346361.1
glucose-6-phosphate dehydrogenase	XP_004364168.2	No	No	1728303-1731563 GB: KE346363.1
NAD-dependent L-lactate dehydrogenase A	XP_004344973.1	No	No	40811-43490 GB: KE346370.1
NADH-dependent aldehyde reductase	XP_004344325.2	No	No	353133-355497 GB: KE346371.1
D-ribulose-5-phosphate 3-epimerase	XP_004346411.1	No	No	1151252-1152712 GB: KE346368.1
Isocitrate lyase	XP_004343864.1	No	No	685712-688381 GB: NW_011887304.1
Malate synthase	XP_004343865.1	Yes	Gene duplication	689250-692016 GB: NW_011887304.1

Appendix 36

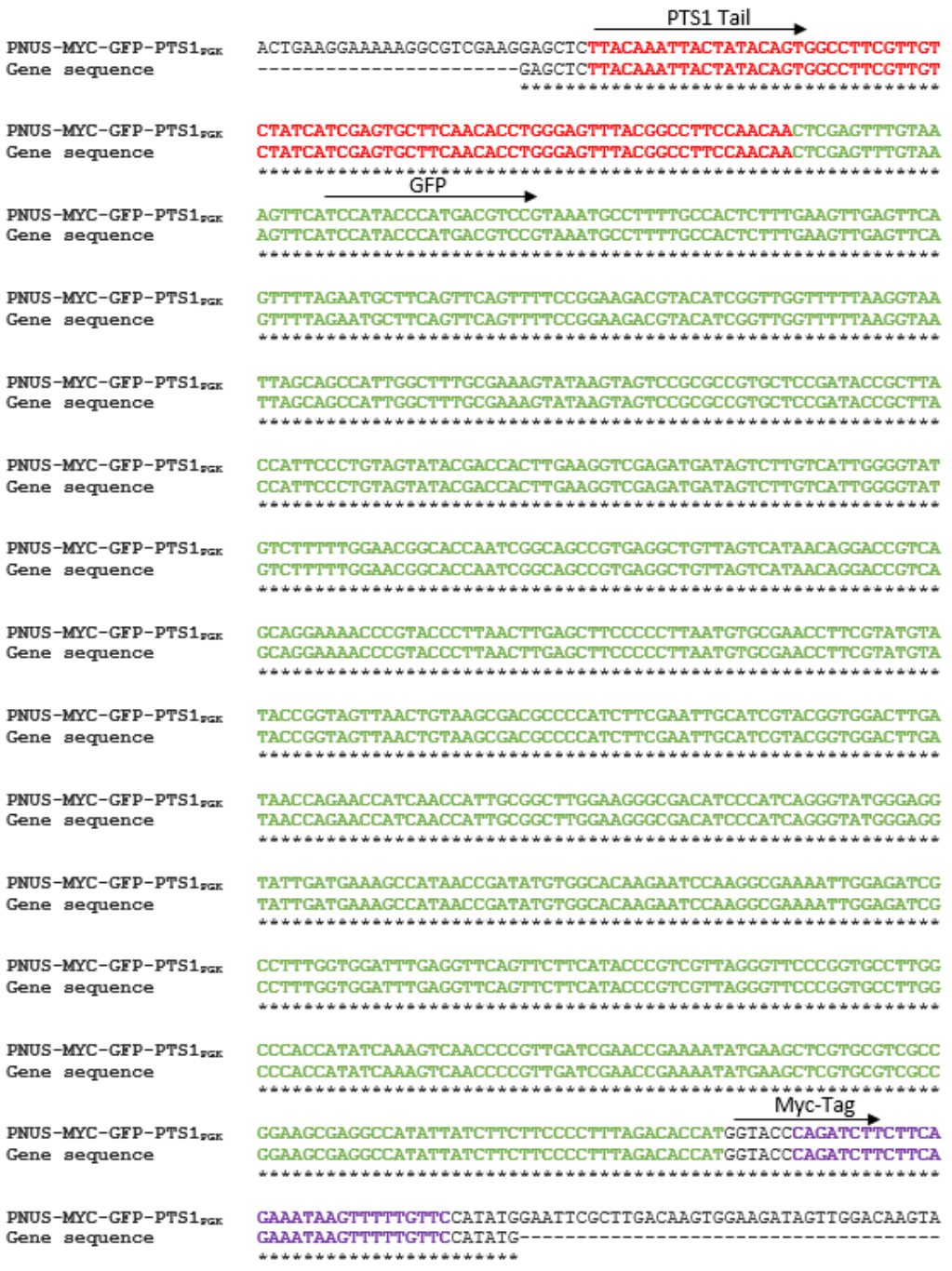


Figure 9.12. DNA alignment between pNUS vector ligated to Myc-GFP-PTS1 PGK extension and the synthetic gene construct sequence. The cloned pNUS vector was send to SourceBioscience for the Sequencing Data using reverse primer. Red colour shows the sequence of the last 75 nucleotide sequence of PGK, green colour shows the nucleotide sequence of GFP and the purple colour shows the nucleotide sequence of Myc epitope.

Appendix 37

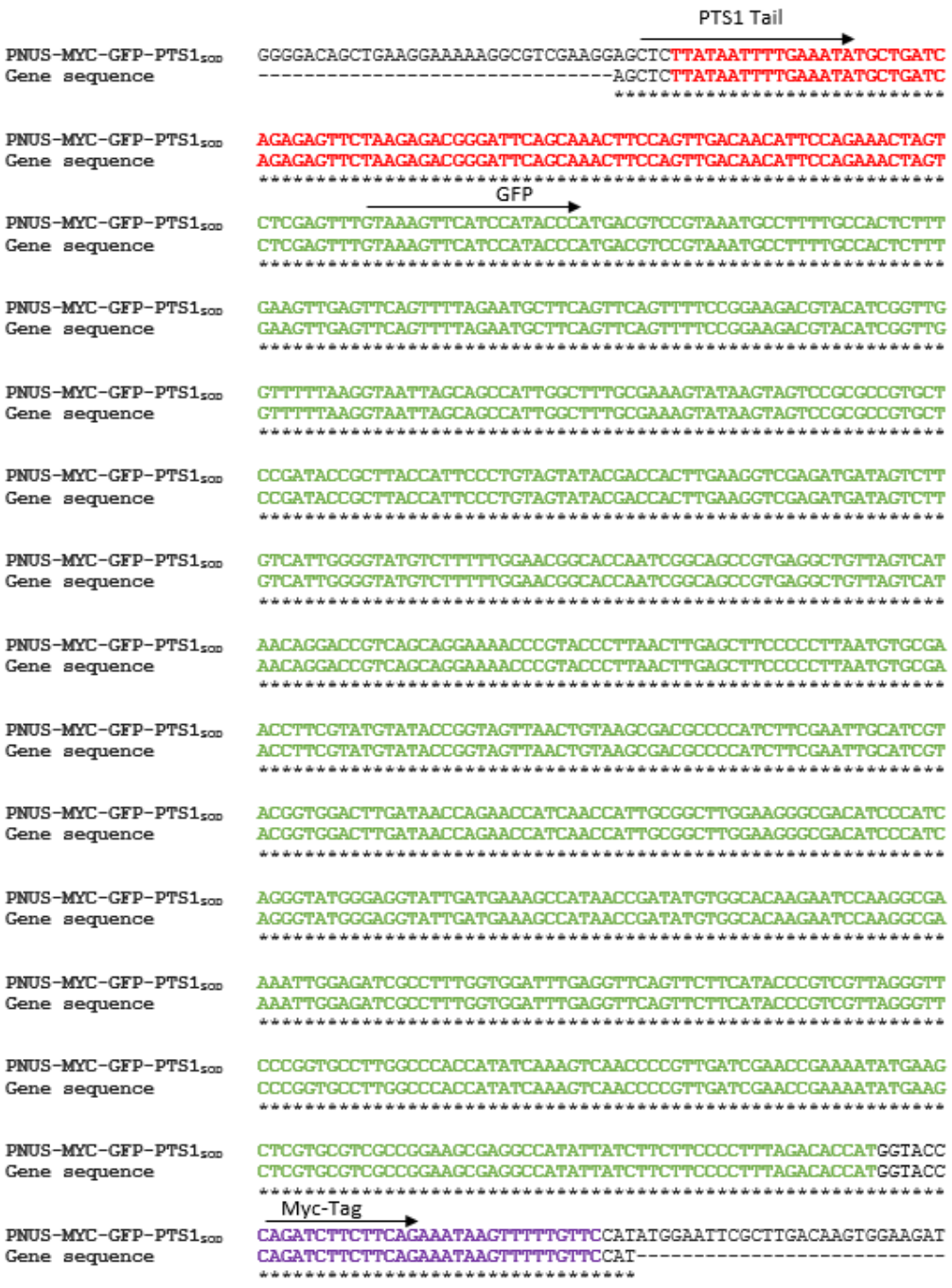


Figure 9.13. DNA alignment between pNUS vector ligated to Myc-GFP-PTS1 $_{SOD\ extension}$ and the gene construct sequence. The cloned pNUS vector was send to SourceBioscience for the Sequencing Data using reverse primer. Red colour shows the sequence of the last 75 nucleotide sequence of SOD, green colour shows the nucleotide sequence of GFP and the purple colour shows the nucleotide sequence of Myc epitope.

Appendix 38

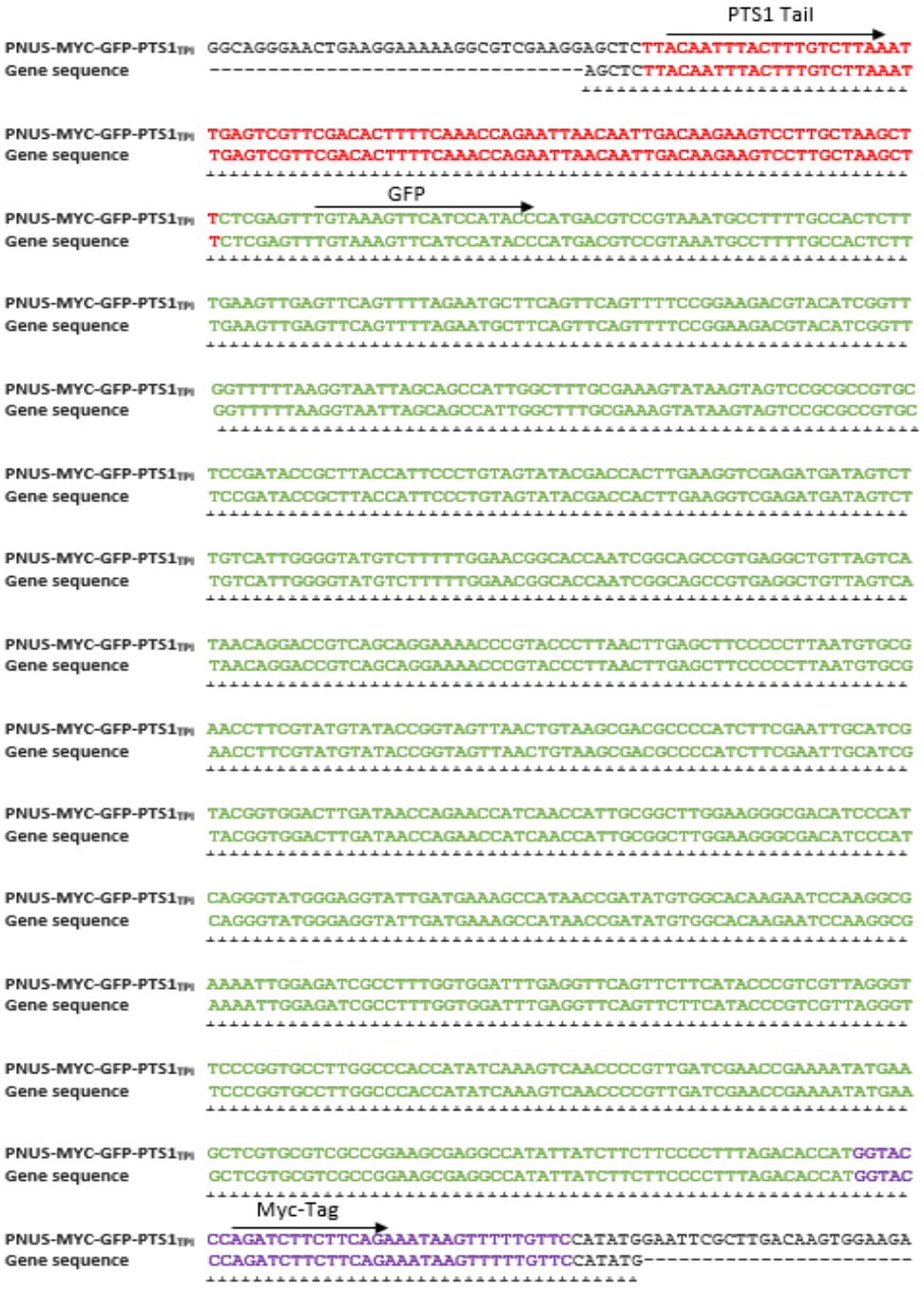


Figure 9.14. DNA alignment between pNUS vector ligated Myc-GFP-PTS1 $_{TPI\ extension}$ and the synthetic gene construct sequence. The cloned pNUS vector was send to SourceBioscience for the Sequencing Data using reverse primer. Red colour shows the sequence of the last 75 nucleotide sequence of TPI, green colour shows the nucleotide sequence of GFP and the purple colour shows the nucleotide sequence of Myc epitope.

Appendix 39

Figure 9.15. DNA alignment between pNUS vector ligated Myc-GFP-PTS1 $_{AR\ extension}$ and the synthetic gene construct sequence. The cloned pNUS vector was send to SourceBioscience for the Sequencing Data using reverse primer. Red colour shows the sequence of the last 75 nucleotide sequence of AR, green colour shows the nucleotide sequence of GFP and the purple colour shows the nucleotide sequence of Myc epitope.

Appendix 40

A Example of cDNA sequence conserved with a random region within *N .gruberi* genome

```
N.gruberi-hypot TTGGTTCCTCCAAATCCATGGACAAGTTCTACCGCTGGTGGTTACATCGCTCTTAAAGTT
cDNA-MLD-sequen TTGGTTCCTCCAAATCCATGGACAAGTTCTACCGCTGGTGGTTACATCGCTCTTAAAGTT
                ************************************************************

N.gruberi-hypot TTAATAATGAGAACAAGAGGTTCTCACTTACAGGCCTATACACTAAAGTCTGCCGAATTG
cDNA-MLD-sequen TTAATAATGAGAACAAGAGGTTCTCACTTACAGGCCTATACACTAAAGTCTGCCGAATTG
                ************************************************************

N.gruberi-hypot AACGATCTTTTATCTGGTTTAAATACACTCAATGAAACTAGATGGAGAATCAATCCAACA
cDNA-MLD-sequen AACGATCTTTTATCTGGTTTAAATACACTCAATGAAACTAGATGGAGAATCAATCCAACA
                ************************************************************

N.gruberi-hypot GTATATGAAACAGTTCTTAAAGTTTGGAATTCAGGAGGTGGATTGGGAGAATTGCCATCA
cDNA-MLD-sequen GTATATGAAACAGTTCTTAAAGTTTGGAATTCAGGAGGTGGATTGGGAGAATTGCCATCA
                ************************************************************

N.gruberi-hypot AGATTTGACGATACTCTTCCAGAAGAACCAGAAACTGATGATCCAGAAATTAAGAAAAAA
cDNA-MLD-sequen AGATTTGACGATACTCTTCCAGAAGAACCAGAAACTGATGATCCAGAAATTAAGAAAAAA
                ************************************************************

N.gruberi-hypot TGGAAACGTACAGTAGAAAAGATTAAGACTCACAATAGAAACTTACACTCTCTCAGATGT
cDNA-MLD-sequen TGGAAACGTACAGTAGAAAAGATTAAGACTCACAATAGAAACTTACACTCTCTCAGATGT
                ************************************************************

N.gruberi-hypot GACATGCTTCTCAAACTTAAAGTTGCTGAAGATTTCAAGGATAACACTTTCTATTTCCCA
cDNA-MLD-sequen GACATGCTTCTCAAACTTAAAGTTGCTGAAGATTTCAAGGATAACACTTTCTATTTCCCA
                ************************************************************

N.gruberi-hypot CACAACATCGATTTTAGAGGAAGAGCTTATCCAATCCCTCCTCATTTGAACCATCTTGGT
cDNA-MLD-sequen CACAACATCGATTTTAGAGGAAGAGCTTATCCAATCCCTCCTCATTTGAACCATCTTGGT
                ************************************************************

N.gruberi-hypot GCTGATATCAACAGAGGTTTACTCACTTTTGAAAAGGGAATGCCAATTGGTAAGAGGGGT
cDNA-MLD-sequen GCTGATATCAACAGAGGTTTACTCACTTTTGAAAAGGGAATGCCAATTGGTAAGAGGGGT
                ************************************************************
```

B Example of cDNA sequence conserved with a random region within bacterial (*Bacteroides fragilis*) genome

```
B.fragilis-hypo ----------------TTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTC
cDNA-MLD-sequen GAGCTAACTCACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTC
                                ****************************************

B.fragilis-hypo GTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTTCGGATTGGGCG
cDNA-MLD-sequen GTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCG
                ********************************************** ** ********

B.fragilis-hypo CTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGGCGTTCGGCTGCGGCGAGCGGT
cDNA-MLD-sequen CTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGT
                ************************************ ********************

B.fragilis-hypo ATCAGCTCACTCAAAGGCGGGAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAA
cDNA-MLD-sequen ATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAA
                ******************* ****************************************

B.fragilis-hypo GAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGC
cDNA-MLD-sequen GAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGC
                ************************************************************

B.fragilis-hypo GTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAG
cDNA-MLD-sequen GTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCANAG
                ************************************************************ **

B.fragilis-hypo GTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGT
cDNA-MLD-sequen GTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGT
                ************************************************************

B.fragilis-hypo GCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGG
cDNA-MLD-sequen GCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGG
                ************************************************************

B.fragilis-hypo AAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCG
cDNA-MLD-sequen AAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCG
                ************************************************************
```

Figure 9.16. Examples of bacterial contamination and unspecific *N. gruberi* genome amplification during RT-PCR. (A) DNA alignment between the amplified *N. gruberi* cDNA (cDNA-MLD-sequen) and random hypothetical *N. gruberi* gene (N. gruberi-hypot). (B) DNA alignment between the amplified *N. gruberi* cDNA (cDNA-MLD-sequen) and random hypothetical *Bacteroides fragilis* (B.fragilis-hypo) gene.

Appendix 41

Table 9.43. Metabolic enzymes with candidate peroxisomal isoforms in both *N. gruberi* and *N. fowleri*. The red colour represents the putative PTS1. Brown colour represents amino acids involved in pex5 interaction.

	C-terminal dodecapeptide		PTS1 predicted using Ppero		Total positive	
Enzyme	*N.gruberi*	*N.fowleri*	*N.gruberi*	*N.fowleri*	*N.gruberi*	*N.fowleri*
Glucose-6-phosphate isomerase	QSQFVYQFVSKL	RTEFVYQFISKL	Yes (2.48)	Yes (2.51)	8/12	9/12
Triosephosphate isomerase	KSVELNLRQSKL	QIAKSKSKISKL	Yes (2.53)	Yes (2.54)	12/12	10/12
Glyceraldehyde-3-phosphate dehydrogenase	LSLATNKFYSDL	QLSNNHDFYSNM	Yes (2.06)	No (0.698)	10/12	9/12
Phosphoglycerate kinase	LDDRQRPLYSNL	TKRIFRNVKSNL	Yes (2.18)	Yes (2.156)	10/12	8/12
Malate dehydrogenase	VIISSLMGISRV	GSFHHHSFNSKI	No (0.728)	Perhaps yes (1.158)	8/12	11/12
Isocitrate dehydrogenase NADPH	LLNLPKQRASKL	LLKPNSHRVSKL	Yes (2.573)	Yes (2.552)	9/12	8/12
Citrate synthase	FISPMSKQRSKI	No	Yes (1.222)	-	9/12	-
Superoxide dismutase	RLLLSDQHISKL	LLLVKSDRNSKL	Yes (2.607)	Yes (2.593)	9/12	9/12
Aldehyde reductase	NFWSLPFLVSRI	SDGCKLCRLANL	Perhaps yes (1.02)	No (0.461)	9/12	9/12
D-ribulose-5-phosphate 3-epimerase	DSVNFFQLGSNL	No	Yes (2.131)	-	8/12	-
Isocitrate lyase	EYKNNWPTRHKL	-	No (0.604)	-	8/12	-

Copyright Statement