THE GENOME OF *BLASTOCYSTIS* SP. ISOLATED FROM THE ORIENTAL COCKROACH

by

Sarah Shah

Submitted in partial fulfilment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
August 2018

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Blastocystis spp. are unicellular anaerobic stramenopiles that inhabit the colons of a wide range of animals. Previous genome studies of *Blastocystis* were restricted to mammalian and avian-colonizing subtypes. To investigate a deeper-branching subtype, I used next generation sequencing technologies to characterize the genome of a *Blastocystis* species from the Oriental cockroach. The ≈40kb mitochondrion-related organelle (MRO) genome was larger than other *Blastocystis* subtypes but had mostly conserved gene content and order. The nuclear genome was 17.1 Mbp in length, 19.9% GC and differed from other *Blastocystis* subtypes in gene content by 15-27%. Amongst the encoded proteins are key enzymes of anaerobic ATP generation. Laterally-acquired genes previously described in *Blastocysti*s sp. ST1 were also identified. Unexpectedly, 40% of the nuclear genes possessed homopolymer trinucleotide insertions encoding polyasaparagines. If these mutations are slightly deleterious, they were possibly fixed in the population by genetic drift due to a small effective population size.

LIST OF ABBREVIATIONS USED

[FeFe]	Iron-iron
ADP	Adenosine diphosphate
ASCT1C	Acetate: succinyl-CoA transferase subtype 1C
AT	Adenine-thymine
ATP	Adenosine triphosphate
BLAST	Basic Local Alignment Search Tool
bp	Base pair
cDNA	Complementary DNA
CIA	Cytosolic iron-sulfur assembly
CoA	Coenzyme A
CTAB	Cetrimonium bromide
EDTA	Ethylenediaminetetraacetic acid
ETC	Electron transport chain
evalue	Expect value
$FADH_2$	Flavin adenine dinucleotide (hydroquinone form)
Gb	Gigabase
GC	Guanine-cytosine
gDNA	Genomic DNA
HMM	Hidden Markov Model
HSP	High-scoring pair
Hsp	Heat shock protein

IBS	Irritable Bowel Syndrome
IGR	Intergenic region
ISC	Iron-sulfur cluster
kb	Kilobase
LGT	Lateral gene transfer
LINE	Long Interspersed Nuclear Element
LTR	Long Terminal Repeat
Mb	Megabase
MRO	Mitochondria Related Organelle
mtDNA	Mitochondrial DNA
N-mer	Polymer (made of N number of monomers)
NADH	Nicotinamide adenince dinucleotide (reduced)
NUMT	Nuclear mitochondrial DNA
ORF	Open reading frame
PCR	Polymerase Chain Reaction
PFO	Pyruvate:ferredoxin oxidoreductase
rpl	Large subunit ribosomal protein
rps	Small subunit ribosomal protein
rRNA	Ribosomal RNA
RT-PCR	Reverse Transcription Polymerase Chain Reaction
SAR	Stramenopile-Alveolata-Rhizaria
SDS	Sodium Dodecyl Sulfate
SINE	Short Interspersed Nuclear Element

SNP	Single Nucleotide Polymorphism
SSU rRNA	Small-subunit ribosomal RNA
ST	Subtype
SUF	Sulfur mobilization
TCA	Tricarboxylic acid
TE	Transposable element
tRNA	Transfer RNA
TXNDC12	Thioredoxin-domain-containing protein 12
VatC	V-type proton ATPase subunit C

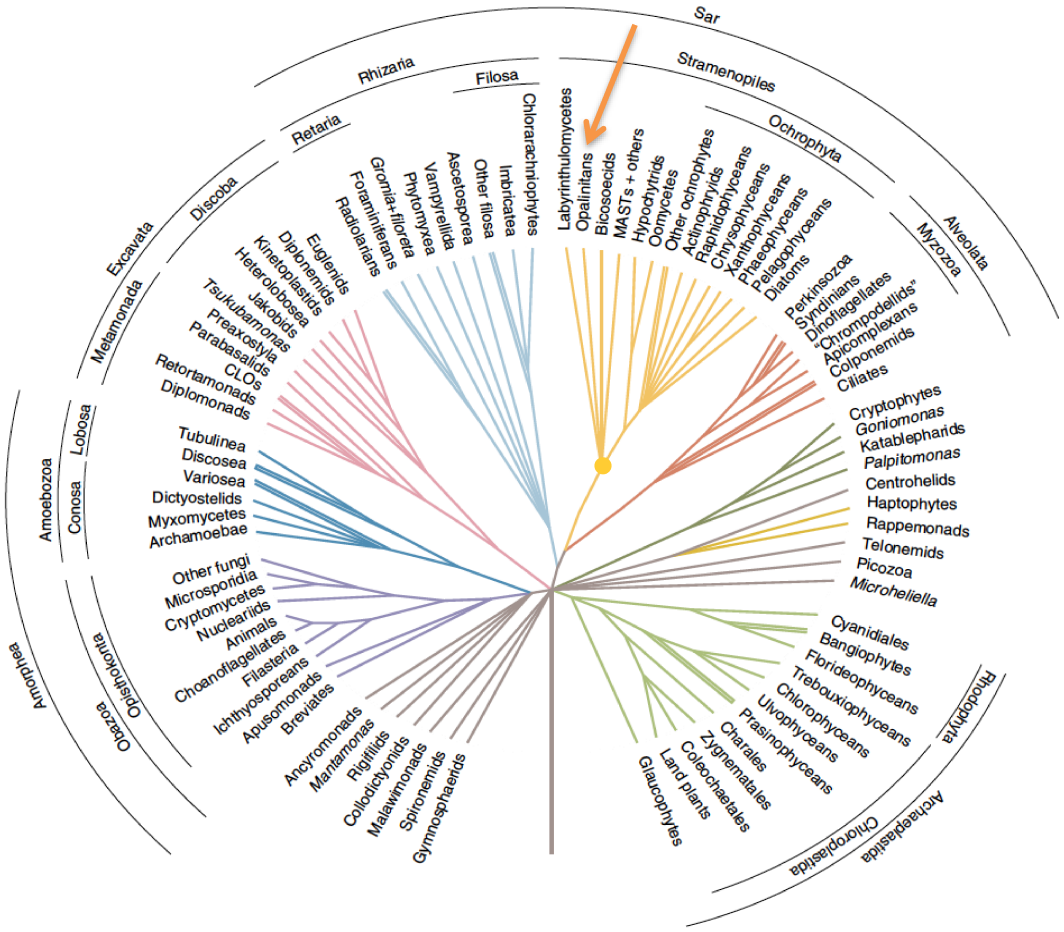
ACKNOWLEDGEMENTS

I would first like to thank my supervisor Andrew J. Roger for taking me in and convincing me that protist genomics are much more interesting than human genomics. He encouraged me every step through this project – when I started, I didn't even know that eukaryotic life existed outside of plants and animals. I would like to extend my gratitude to our collaborator C. Graham Clark of London School of Hygiene and Tropical Medicine for sending me cultures of *Blastocystis* despite the many customs issues and letting me base my project on them. To my co-supervisors Dayana Salas-Leiva and Jon Jerlström-Hultqvist, thank you for mentoring me these two years and always being there, during holidays and past working-hours, to answer my questions. Special thanks to Bruce Curtis who provided many scripts and shared his expertise on dealing with sequencing data. I am eternally grateful to former Roger Lab members Laura Eme and Michelle M. Leger for introducing me to the world of computer clusters and scripting. Many thanks to those who helped me with wet-lab work: Roger Lab technician Marlena Dlutek and former Honours students Kate Glennon and Shelby Williams, I apologize for exposing you to the stench of my *Blastocystis* cultures. I am also thankful to André M. Comeau of CGEB-IMR for sequencing my difficult samples. Finally, I would like to thank my supervisory committee members John M. Archibald and Claudio Slamovits for their helpful input on my project, and everyone else in CGEB and the Faculty of Graduate Studies for their advice on academic matters to Canadian winters, you will be missed!

CHAPTER 1 INTRODUCTION

Blastocystis is a genus of unicellular eukaryotes that colonize the lower intestines of diverse animals, including humans. They are usually observed to be spherical, ranging from ≈10-40 μm in diameter, contain one or more nuclei, a large central vacuole, and lack flagella (Tan, 2008). *Blastocystis* has been controversial in medical and taxonomic fields ever since its discovery. Alexieff (1911) first described it as a yeast under the name *Blastocystis enterocola*, found in the intestines of rats, guinea pigs, chickens, reptiles, and leeches. Brumpt (1912) renamed it as *Blastocystis hominis* and Zierdt et al. (1967) reclassified it under the parasitic protistan group known as the Apicomplexa that includes organisms such as the malaria parasite *Plasmodium* and the agent of toxoplasmosis, *Toxoplasma gondii*. Apicomplexans belong to the Alveolata group within the stramenopiles-alveolate-Rhizaria (SAR) supergroup (Adl et al., 2012). In the mid-1990s, the first molecular data was obtained, and based on the phylogeny of small-subunit ribosomal RNA (SSU rRNA) genes, *Blastocystis* was shown instead to belong to the stramenopiles (Silberman et al., 1996), a group containing diverse photosynthetic, heterotrophic, and parasitic eukaryotes (Figure 1.1).

Figure 1.1: Tree of Life of eukaryotes reflecting the diversity of protists, based on molecular and phylogenomic analyses, reproduced from Simpson & Eglit (2016) permission from Elsevier. *Blastocystis* is a member of the 'Opalinitans' (yellow arrow) within stramenopiles, a major eukaryotic clade that includes kelps, diatoms, and the oomycete that causes potato blight. The phylogenetic position of their most recent common ancestor is represented by the yellow dot.



The broad genetic diversity discovered amongst the SSU rRNA sequences of various strains (Clark, 1997) and their wide host ranges (Abe, 2004) led researchers to devise a new classification scheme for *Blastocystis* 'types'. Groups of *Blastocystis* whose SSU rRNA sequences differed from each other by ≥ 4% identity were renamed as '*Blastocystis* sp.' followed by an assigned subtype (ST) as suggested by Stensvold et al. (2007). Within each *Blastocystis* subtype clade in the phylogenetic tree, SSU rRNA genes

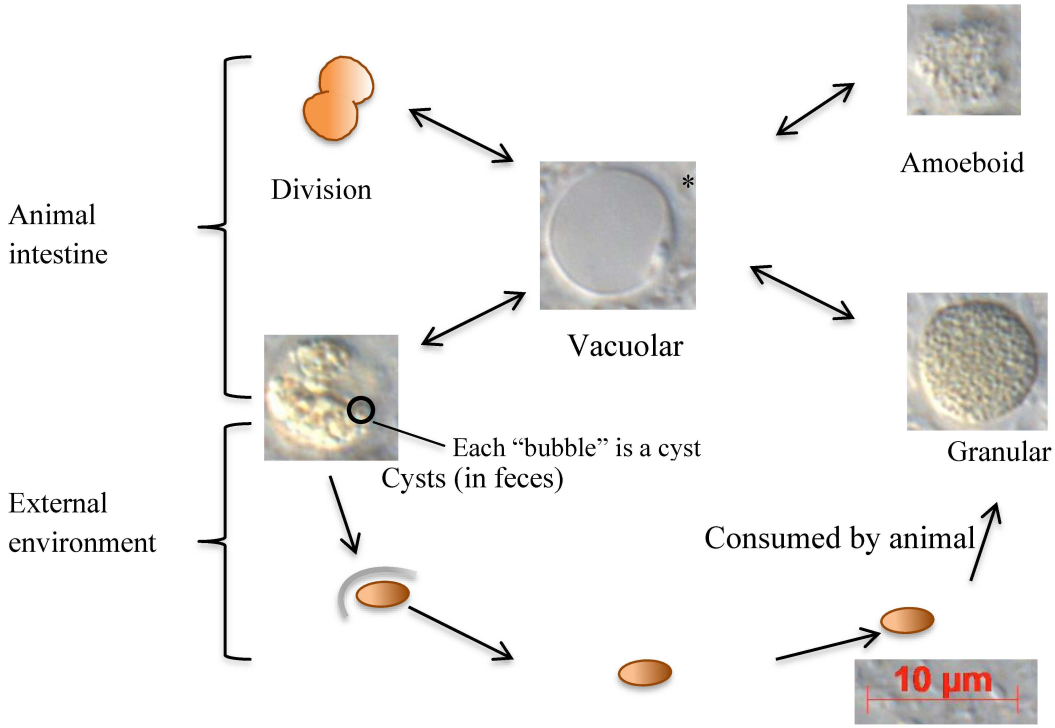
usually vary by 1-2% (Stensvold et al., 2007). This provisional taxonomy helped eradicate misleading binominal names (e.g., the now obsolete name '*Blastocystis hominis*') that falsely suggested host specificity.

There are at least 17 different subtypes described in the literature (ST1 to ST17) that can be broadly described as isolates from mammalian or avian hosts (Clark et al., 2013). The deeper branching amphibian/reptile/insect-infecting *Blastocystis* strains are not yet classfied into subtypes (Betts et al., 2017). Humans commonly host ST1-4, livestock host ST5, and birds usually have ST6-7, although there are exceptions to these general patterns (e.g., toad sequence grouping within ST5 (Yoshikawa et al., 2004)). The closest relatives of the *Blastocystis* clade are *Proteromonas lacertae*, a lizard-hosted gut stramenopile, and its sister taxon *Opalina* sp., found in frog guts (Betts et al., 2017). These 'Opalinitans' (see Figure 1.1) are heterotrophic, anaerobic, and flagellated (*Opalina* spp. have multiple flagella). Indeed, all other observed stramenopiles (i.e., excluding organisms that have been classified into this group by sequence alone, such as the marine stramenopiles (MAST) clades from metagenomic data) have a flagellum in at least one stage of their life cycle (Fu et al., 2014), which makes *Blastocystis* an oddity, having lost this feature secondarily (Yubuki et al., 2016).

The life cycle of *Blastocystis* is poorly understood. In its smallest form, usually in the external environment, a cell ('cyst') reaches 1-5 μm and it is presumably infectious. If ingested, once it arrives in the host gut it can become vacuolar, granular, or amoeboid (Tan, 2008) increasing its diameter from 3 to 120 μm (Lee & Stenzel, 1999) (see Figure 1.2). There is still some debate as to whether the granula or amoeboid stages occur *in vivo*, or are, instead, a consequence of oxygen exposure during microscopy (Stenzel &

Boreham, 1991; Vdovenko, 2000; Clark et al., 2013). The vacuolar form has been observed to predominantly divide by binary fission in humans, although recently, it was proposed that it is capable of reproducing sexually, as homologs of genes encoding proteins involved in meiosis were found in ST1 (Gentekaki et al., 2017). All known *Blastocystis* subtypes are strict anaerobes and a pure, i.e., axenic culture must be maintained in the absence of oxygen (Zierdt & Williams, 1974). *Blastocystis* subtypes are osmotrophic heterotrophs during most of their life cycle and are thought to metabolize fucose and sialic acid using enzymes encoded by genes acquired from bacteria through lateral gene transfer (LGT) (Eme et al., 2017).

Figure 1.2: Depiction of the putative life cycle of *Blastocystis* spp. Cysts are ingested into the animal intestine where they transform into the vacuolar form and divide. They may switch back and forth between vacuolar, granular, and amoeboid forms. The new cells eventually turn into cysts, and in humans a thick fibrillar layer has been observed around them, which is shed during passage into the external environment. Micrographs are of cultured *Blastocystis* sp. isolated from the cockroach *Blatta orientalis*. Scale bar applies to all micrographs. *The vacuolar form in this isolate are ≈5-20 µm.



Blastocystis cells have evolved serveral adaptations to the anerobic environment, including two key enzymes involved in anaerobic ATP production, pyruvate:ferredoxin oxidoreductase (PFO) and iron-only [FeFe] hydrogenase that were observed to localize to *Blastocystis* mitochondria. This localization suggests that these organelles function like hydrogenosomes, modified mitochondria that generate ATP via substrate-level phosphorylation, releasing hydrogen gas in the process (Stechmann et al., 2008; Denoeud et al., 2011). However, direct evidence of hydrogen gas production has not been observed for *Blastocystis*. Gentekaki et al. (2017) proposed an alternative pathway in which molecules other than hydrogen ions act as electron acceptors. Hence *Blastocystis* mitochondria are called mitochondrion-like or mitochondrion-related organelles (MLOs or MROs) (Stairs et al., 2014), a term used to describe any organelle that is mitochondrion-like but has different features/functions from the canonical aerobic mitochondria. Most hydrogenosomes of trichomonads and fungi lack a genome (Embley, 2006), but *Blastocystis* spp. do possess MRO genomes; the latter will be the focus of Chapter 2.

Since its discovery in the early 20th century *Blastocystis* was considered an intestinal parasite that caused diarrhea (Fantham, 1916; Lynch, 1917; Stabler, 1941). It is extremely prevalent in humans; up to 1 billion people worldwide (Stensvold, 2012) are estimated to harbour a subtype of *Blastocystis*. Colonization with this parasite has been termed "Blastocystosis" and it is sometimes associated with irritable bowel syndrome (IBS) symptoms, a general diagnosis that involves a group of gastrointestinal symptoms such as recurrent abdominal pain, diarrhea, and constipation (Grundmann & Yoon, 2010), which reportedly affects around 11% of the global population (Canavan et al.,

2014). Rostami et al. (2017) found a positive association of *Blastocystis* sp. with IBS, and Ajjampur & Tan (2016) compiled evidence that suggests the parasite may have the ability to damage intestinal epithelial cells and modulate host immunity depending on the subtype. However, Scanlan et al. (2014) reported that symptoms of blastocystosis can be attributed to other microorganisms. Beghini et al. (2017) also suggested that *Blastocystis* only becomes pathogenic through interaction with certain microorganisms or environmental factors, depending on host immunity status. Stensvold & Clark (2016) have suggested that studies linking gut diseases to *Blastocystis* frequently lack appropriate control groups and fail to consider other possibilities of the cause of the disease.

Several lines of research are currently investigating the phenotypic and genotypic diversity of *Blastocystis*. Approaches such as traditional PCR-based methods (Forsell et al., 2016), metagenomic analysis of fecal microbiota (Forsell et al., 2017; Siegwald et al., 2017)), protein activity studies (Armengaud et al., 2017), or genomic studies (Gentekaki et al.; Deneoud et al., 2011) can be used to investigate the suspected pathogenicity of the various *Blastocystis* subtypes. Such studies have uncovered interesting features such as the lack of canonical stop codons for at least 15% of nuclear protein-coding genes in ST1 and ST7 (Klimes et al., 2014). These genes are transcribed and complete ORFs are generated through creation of stop codons by the polyadenylation of transcripts (Klimes et al., 2014). That is, the first one or two nucleotides of the poly(A) tail added to the transcript end up forming the final or last two nucleotides of the stop codons UAA or UGA. While this phenomenon is known in the mitochondria of mammals (Anderson et al., 1981; Chang & Tong, 2012), dinoflagellates, apicomplexa (Nash et al., 2008) and

euglenids (Kiethega et al., 2013), it has never before been found in a eukaryotic nuclear genome. Other intriguing features of *Blastocystis* genomes include: large genetic variation among subtypes in terms of number of introns and gene content (9-20% of the gene set of each subtype is unique), extreme divergence in amino acid sequence among orthologous proteins (39-41% dissimilarity on average), and mitochondrial genes that have lost a start codon (e.g., in rps4) or possess in-frame stop codons (e.g., in "orf160") (Jacob et al., 2016).

The gene content differences between subtypes are substantial. For example, ST1 was found to have an expansion of STE20/7 kinase gene family, while calcium-calmodulin-dependent-like kinases were mostly exclusive to ST4 and ST7 (Gentekaki et al., 2017). ST4 completely lacked the M23 family of metallopeptidases found in ST1 and ST7. These enzymes are involved in lysing peptidoglycans in bacterial cell walls (Genetakaki et al., 2017).

While the debate continues on whether *Blastocystis* subtypes are harmful, beneficial, or have no effect on animal health, another, more fundamental, question arises: how did they become such ubiquitous gut symbionts in so many metazoans? The most recent common ancestor of stramenopiles is thought to have been free-living, biflagellated, and bacterivorous (Shiratori et al., 2015; Yubuki et al., 2016). Some claim that *Blastocystis* is bacterivorous (Stenzel & Boreham, 1996; Tan & Suresh, 2006), but it is neither free-living nor flagellated. Interestingly, *Blastocystis* subtypes have lost all parts of basal bodies and lack a microtubular cytoskeleton in their cytoplasms (Yubuki et al., 2016). This stands in contrast to some eukaryotic gut microbial parasites that have developed cytoskeletal adaptations (i.e., microtubule aggregations) to attach to the

intestinal wall of their hosts, such as the 'ventral disc' or the 'rostellum' that have been described in *Giardia* (Brugerolle & König, 1997) and in oxymonads (Holberton, 1973), respectively. Previous *in-vitro* and *in-vivo* studies on mammalian and avian isolates of *Blastocystis* spp. have shown that secreted cysteine proteases, legumain and cathepsin B, help them invade the gut epithelium (Ajjampur & Tan, 2016), much like the dysentry-causing amoeba *Entamoeba histolytica*, which produces lectin that binds to colonic mucin and epithelial cells (Singh et al., 2016). In-depth genome studies of more *Blastocystis* isolates might reveal new insights that may have aided their transition from free-living lifestyle to becoming dependent on the animal gut.

Blastocystis spp. are prevalent not only in birds and mammals, but in reptiles, amphibians, and insects, based on SSU rRNA detection (Cian et al., 2017). However, previous genomic studies only focused on mammalian and avian isolates. Among the non-mammals, the insects are the most undersampled: to date, the only known insect-hosted *Blastocystis* sequences are SSU rRNAs from the American cockroach, *Periplaneta americana* (Yoshikawa et al., 2007). Blattidae, the cockroach family which includes the common house pests *P. americana* and *Blatta orientalis*, is an older clade than primates (i.e., the former emerged ≈192 million years ago (mya) (Legendre et al., 2015) whereas the latter originated ≈65mya (Williams et al., 2010)). Cockroaches have different thermoregulation; they are poikilothermic (i.e., internal temperature varies) as opposed to being homeothermic (i.e., constant internal temperature) as in birds and mammals. Their gut microbiome has a different microbial community profile; for example, the proportion of Firmicutes: Bacteroidetes: Proteobacteria in a cockroach colon is ~5:4:1 vs. ~20:3:1 in healthy human colons (Schauer et al., 2012; Chen et al.,

2012; Sender et al., 2016). The ability of *Blastocystis* spp. to colonize broadly diverse hosts with distinctive metabolisms and gut microbiota make early divergent representatives of the clade good candidates to identify gene sets and features of genome structure that allowed *Blastocystis* spp. to become successful parasites. For these reasons, I chose to characterize the genome of a *Blastocystis* isolate from the Oriental cockroach species *Blatta orientalis,* a cosmopolitan pest that prefers the warm indoors (Arnett, 2000). If this isolate is found to possess the adaptations found in *Blastocystis* spp. from mammalian isolates, it will push back the timing of their acquisition. If absent, then it is possible these features evolved later in the *Blastocystis* lineage in adaptation to particular features of the vertebrate hosts they infect. It is also of interest to discover what unique features this isolate may have evolved in adaptation to the cockroach gut.

In this thesis, I present the findings of my research divided into four chapters. After this introductory chapter, in Chapter 2 I present the MRO genome of the cockroach *Blastocystis* isolate (hereby abbreviated to "BBO") and compare it to MRO genomes of mammalian isolates. In Chapter 3, I present an analysis of the nuclear genome of BBO and its predicted proteome including an investigation of the polyadenylation mechanism and LGTs previously found in mammalian isolates. Chapter 4 summarizes the findings of this project and traces the evolutionary changes that occurred to the BBO lineage, and proposes some ideas for future studies.

CHAPTER 2 THE MRO GENOME OF *BLASTOCYSTIS* SP.

2.1 INTRODUCTION TO MROS

Mitochondria are organelles best known for their role of providing energy, in the form of ATP, to the eukaryotic cell. These organelles first originated by the endosymbiosis of a bacterium within an archaeon-related host lineage. The endosymbiont is usually though to have evolved from within the Alphaproteobacteria (Yang et al., 1985) but recent analyses suggest it may have arisen from an immediate sister group of the Alphaproteobacteria (Martijn et al., 2018). The typical aerobic mitochondrion of model system eukaryotes is bound by two membranes, the inner one folded into cristae, in which various embedded structures work together to generate ATP. Pyruvate is actively transported from the cytosol into mitochondria and is oxidatively decarboxylated by the pyruvate dehydrogenase complex into acetyl-CoA, which enters the TCA cycle. The reducing equivalents generated by this cycle (NADH and $FADH_2$) are oxidized by the electron transport chain (ETC) complexes I and II. The shuttling of electrons through the ETC leads to the translocation of protons to the intermembrane space creating a proton gradient that drives the final complex V (ATP synthase) to phosphorylate ADP to ATP in the mitochondrial matrix. The electrons are ultimately passed to complex IV that reduces oxygen and protons to water.

The mitochondrion carries out a number of other essential cellular functions in addition to ATP synthesis. For example, mitochondria also house the highly-conserved iron–sulfur cluster assembly (ISC) machinery, which produces coordinated iron-sulfur clusters that are essential cofactors in a number of mitochondrial proteins (including

enzymes in the TCA cycle, complexes I-III, lipoate and heme synthesis pathways). The ISC system also produces a reduced sulfur-containing compound that is used by the cytosolic iron-sulfur assembly (CIA) system which in turn builds iron-sulfur clusters for proteins involved in cellular translation, nuclear DNA replication and repair, among others (Lill et al., 2015). Other functions of canonical aerobic mitochondria include storage of calcium ions, which regulates calcium levels across the mitochondrial membrane potential and controls the release of neurotransmitters or hormones (Miller, 1998; Santulli et al., 2015), and intracellular apoptosis (programmed cell death) in which stress induces apoptotic proteins to be released from the intermembrane space of the mitochondrion, changing its membrane potential, leading to its shutdown followed by cell death (Wang & Youle, 2009). However, mitochondrial functions vary markedly across eukaryotic diversity, especially in those organisms adapted to low oxygen conditions.

Organelles that are missing canonical features of model system aerobic mitochondria are collectively referred to as mitochondrion-related organelles (MRO) (Stairs et al., 2014). Mitochondrial functions have been reduced or replaced in many eukaryotes that have adapted to low-oxygen environments (e.g., animal guts or in anaerobic sediments). Reductive evolution has resulted in loss of parts of the electron transport chain (e.g., *Nyctotherus ovalis* (de Graaf et al., 2011)), the use of different molecules as terminal electron acceptors in the ETC (e.g., fumarate in *Euglena gracilis* or H^+ in the case of hydrogenosomes (Müller et al. 2012)), the loss of cristae (e.g., Loricifera (Danovaro et al., 2010)), loss of the mitochondrial genome (e.g., in some hydrogenosomes and mitosomes (see Müller et al. 2012, Stairs et al. 2015, Leger et al.,

2017, Roger et al., 2017)), or even the loss of the entire organelle as in *Monocercomonoides* sp. (Karnkowska et al., 2016). Protists can also acquire genes by lateral gene transfers (LGTs) whose products are targeted to MROs. For example, in the MRO of the breviate amoeboid flagellate *Pygsuia biforma*, the ISC system has been replaced by archaeal sulfur mobilization (SUF) system (Stairs et al., 2014).

In the case of *Blastocystis*, despite living in a low-oxygen environment, the mitochondria of ST1 and ST7 resemble canonical aerobic mitochondria in possessing a mitochondrial (MRO) genome, some enzymes of the TCA cycle, amino acid metabolism, the ISC system (Tsaousis et al., 2012), and mitochondrial protein import (Tsaousis et al., 2011; Gentekaki et al., 2017). However, these *Blastocystis* subtypes have completely lost complexes III to V of the ETC and appear to possesses a complete 'hydrogenosomal' pyruvate metabolism pathway expressing genes encoding the anaerobic enzymes pyruvate:ferredoxin oxidoreductase (PFO) and iron-only [FeFe] hydrogenase that are targeted to the MROs. In these respects the *Blastocystis* organelles appear to resemble the 'hydrogen-producing mitochondria' of the rumen ciliate *Nyctotherus ovalis* (de Graaf et al., 2011; Stechmann et al., 2008).

Mitochondrial genomes exhibit wide ranges of architecture, size, and nucleotide composition. They can be linear, circular, have multiple chromosomes (e.g., the green alga *Polytomella piriformis* has two linear chromosomes, parasitic protist trypanasomes have intertwining networks of thousands of circular chromosomes), range from 6 kb in some apicomplexans to 11 Mb in the angiosperm *Silene conica*, with GC content ranging from 12.6% in the yeast *Candida castelli* to 68% in the lycophyte *Selaginella moellendorffii* (see Smith & Keeling (2015) for review). Previous studies of *Blastocystis*

MRO genomes focused on mammalian/avian isolates which displayed synteny in their MRO chromosomes, had similar genome size and structure (27.7-29.3kb, circular), similar coding density (77-78%), and GC content (18.8-22.7%). All possessed the same set of genes: 27 coding for proteins, 2 for rRNAs and 16 for tRNAs (Jacob et al., 2016). Some of the MRO genomes are missing one or more of these genes: *Blastocystis* sp. ST1 strain NandII (GenBank: EF494740.3) and all strains of *Blastocystis* sp. ST3 (GenBank: HQ909887.2, HQ909886.2, HQ909888.2) do not possess the protein-coding gene rps7, and *Blastocystis* sp. ST9 (GenBank: KU900239.1) is missing one of the three tRNA-Met genes. In contrast, their closest stramenopile relative, *Proteromonas lacertae*, an anaerobic and heterotrophic flagellate isolated from lizard intestines, has a linear mitochondrial genome of 48.7 kb. This genome consists of two identical large inverted repeats flanking a central unique region, with a total of 27 protein-encoding genes, 2 rRNAs, and 23 tRNAs genes (Pérez-Brocal et al., 2010). The arrangement of these genes shows no similarity with the *Blastocystis* MROs genomes (i.e., not syntenic). It is evident that there are no major differences within mammalian/avian isolates of *Blastocystis* and thus an earlier-diverging lineage would help better characterize the ancestral MRO genome of the *Blastocystis* clade. Here, I characterized the MRO genome of a *Blastocystis* sp. isolated from the Oriental cockroach (abbreviated to "BBO") to assess gene losses, gains, or gene transfers to the nucleus, and compare it to the genomes of other MROs in *Blastocystis* subtypes and to the mitochondrial genome of *Proteromonas lacertae*.

2.2 METHODS

2.2.1 Culturing

A non-axenic culture of *Blastocystis* sp. (BBO) isolated from the gut of the Oriental cockroach, *Blatta orientalis*, was established using cells obtained from the lab of Graham Clark at London School of Hygiene and Tropical Medicine. These were sub-cultured every two weeks in LYSGYM medium which was prepared as follows: 1.4 g dibasic potassium phosphate, 0.2 g monophasic potassium phosphate, 3.75 g sodium chloride, 0.25 g neutralized liver digest, 1.25 g yeast extract, and 0.05 g porcine gastric mucin dissolved in 475 ml of water, with 25 ml of heat-inactivated adult bovine serum added after autoclaving (Clark & Stensvold, 2016). See Supplementary Figures S1 and S2 for photos of cells observed under light microscopy.

2.2.2 Nucleic acid extraction and sequencing

Cells were collected by centrifugation at 275×g for 5 minutes. The cell pellet was resuspended in 1× PBS and then layered on top of Histopaque (Sigma, Cat. No. 10771) in a 1:9 volume ratio (cell:Histopaque). This layered mixture was subjected to centrifugation at 2,000×g for 20 minutes to separate the bacteria and BBO via density gradient centrifugation. Genomic DNA (gDNA) was extracted from the upper BBO-enriched layer using a phenol/chloroform-based method (described in detail by Clark, 1992). Cells were resuspended in a lysis buffer (0.1M EDTA, 0.25% SDS) and incubated at 55ºC for ≈2 hours. NaCl solution was added to a final concentration of 0.7M and cetrimonium bromide (CTAB) was added to 1% of total volume to complex DNA. This was followed by 1:1 (v/v) DNA mixture to phenol/chloroform extractions (centrifugation

at 14,000×g for 10 minutes with subsequent rounds of extraction of aqueous layers). DNA was precipitated by adding isopropanol up to 70% of the total volume and the DNA pellet was washed with 70% ethanol. The resulting DNA was cleaned by gravity and anion-exchange column chromatography according to the manufacturer's instructions (Genomic-tip 20/G, QIAGEN, Cat. No. 10223). RNA was extracted from the Histopaque-separated BBO cells using Trizol (ThermoFisher).

Two different sequencing methods were used: short read paired-end and long read. For short paired-end sequencing, 315 ng of DNA was sent to Genome Quebec Innovation Centre (gqinnovationcenter.com). The library was prepared using Nextera XT Library Prep Kit (obtained library size: 473 bp); the library was run on the Illumina HiSeq X Ten with 150 bp paired-end reads. RNA (3.4 μg) was sent to Genome Quebec Innovation Centre, where the NEBNext® Ultra™ Directional Library Prep Kit was used for strand-specific library preparation with polyA selection (obtained library size: 345 bp), and then sequenced using the Illumina HiSeq4000 with 100 bp paired-end reads.

In the case of long-reads, the whole procedure was carried out in-house by preparing a library with the 1D-ligation Kit (kit type: SQK-LSK108, Cat. No. EXP-NBD103), which was then sequenced on a single MinION flow cell (model R9.4, Oxford Nanopore Technologies) using 900 ng of DNA. DNA used for MinION sequencing was from a different set of extractions than those used for Illumina sequencing having been extracted from BBO cultures six months (≈12 generations or "passaging") after the Illumina sequencing. Raw data obtained from the three types of sequencing are as follows: Illumina DNA: 86 million reads representing 26 gigabases, Illumina RNA: 40

million reads representing 8 gigabases, Nanopore DNA: 526,000 reads representing 3 gigabases.

After assembling the high throughput data (see section 2.2.3), some assembled mitochondrial segments exhibited duplicated fragments that needed to be re-assessed to rule out assembly artefacts resulting from software miscalling bases from homopolymer or single nucleotide polymorphism (SNP)-rich areas. For this, several oligonucleotide primer pairs were designed based on the draft assembly (see Table 2.1 for primer sequences that successfully produced PCR products – other primers have failed possibly due to difficulties in amplifying extremely AT-rich regions, and are not reported here) and were amplified with LongAmp® *Taq* PCR Kit (New England Biolabs®) and Phusion Hot Start II High-Fidelity PCR Kit (Thermo Scientific™). The default reaction mixtures and thermal profiles were used as described by the manufacturers' protocols, using 5-10 ng of DNA per reaction, with the following exceptions: for the LongAmp® pair, the annealing temperature was set at 48°C, the extension time at 5 min., the amplification cycle at ×30, and for the Phusion Hot Start pair, the annealing temperature was set at 50°C, the extension time at 10 min., the amplification cycle at ×35. The PCR products were sequenced with Illumina MiSeq with 2 × 300 bp paired-end reads following an amplicon protocol (cgeb-imr.ca/protocols.html) conducted by the Centre for Comparative Genomics and Evolutionary Bioinformatics Integrated Microbiome Resource (CGEB-IMR; cgeb-imr.ca).

Table 2.1. Primer sequences used to amplify the duplication regions.

Name	Sequence (5' to 3')	Kit
t10_44kc_f	CTAAACAAGTAGAAACATTATATATGTC	LongAmp®
t10_50kc_r	GTTGTTATACCATAATGGAAAC	
t10_start	ACAGTATATATCATAAAGGAGTGGTAG	Phusion Hot Start
t10_rev	ACAAATTTTGGTGTTTCT	

2.2.3 Assembly and annotation of the mitochondrial genomes

The data generated by the MinION sequencer was base-called by Albacore (Vera, 2017), and adaptor sequences were trimmed by Porechop (Wick, 2017a) using default parameter settings for both programs. Read statistics generated by NanoPlot (de Coster, 2017) reported a mean read length of 6,201 bp and N50 of 9,174 bp (N50: shortest read length shared by 50% of all reads). Canu (Kolen et al., 2017) was used to assemble the reads with a metagenomics setting and AT-rich bias (corMaxEvidenceErate of 0.15, corOutCoverage of 999, high corMhapSensitivity, estimated genomeSize of 20 Mb, and setting to 10 of the variables corMaxEvidenceCoverageLocal,corMinCoverage and corMaxEvidenceCoverageGlobal). The assembly was polished by Nanopolish (Simpson, 2017) and further corrected by short reads using Unicycler (Wick, 2017b) using default parameters. Canu labels contigs as linear or circular; the circular contig with the highest coverage was regarded as the mitochondrial genome. For Illumina-generated reads trimming was done with Trimmomatic v0.36 (Bolger et al., 2014) (quality trimming cutoff value of 25 and reads longer than 40) and these reads were used for correction as mentioned above.

The mitochondrial genome was first annotated using MFannot (Beck & Lang, 2010). RNA and DNA short reads were mapped onto it using HISAT2 (Kim et al., 2015) (strand-specific mapping, intron length of 20-100) and Bowtie2 (Langmead & Salzberg, 2012), respectively. Regions that could not be resolved due to ambiguous open reading frames (ORFs) (e.g., genes whose start codon was predicted at a different location than that found by MFannot's built-in HMM search) were manually corrected – bases were inserted or deleted based on which nucleotide was represented by the majority of RNA and DNA Illumina reads at a location (see Supplementary Figure S3). To study the duplicated region of the MRO, the Illumina-sequenced amplicons were trimmed using Trimmomatic (Bolger et al., 2014) and assembled by PEAR (Zhang et al., 2014) (default settings with minimum assembly length of 100). The assembled amplicons were aligned to the genome using Sequencher® v5.4.6 (Gene Corps Corporation, 2017) employing its built-in assembly algorithm. In the case of unannotated ORFs, its translated nucelotide sequences were used as queries in BLASTp (Gish & States, 1993) searches against the nr database or further narrowed down to a database of *Blastocystis* MRO genomes to infer their identity. Secondary structures were predicted using the online service PSIPRED (bioinf.cs.ucl.ac.uk/psipred/) for all ORFs that were not identifiable using homology-based approaches. These were compared against the secondary structure predictions of mitochondrial rps7 and rps11 from *Blastocystis* ST6, ST8, and *Phytophthora sojae*. Additionally, HMM profiles of mitochondrial rps7 and rps11 were downloaded from Pfam (pfam.xfam.org) and the ORFs were inspected with the program hmmscan from HMMER 3.1b2 (hmmer.org). Illumina reads coverage statistics of gDNA and RNA were generated using samtools (samtools.sourceforge.net/). Cloverleaf structures of tRNAs

were predicted by Predict a Secondary Structure Web Server on rna.urmc.rochester.edu/RNAstructureWeb.

2.2.4 MRO phylogeny

Complex I genes, or *nad* genes, were found to be highly conserved in previously sequenced *Blastocystis* MROs (Jacob et al., 2016). To explore the phylogenetic placement of BBO, a tree was reconstructed using a concatenated matrix containing the predicted mitochondrial proteins nad1, nad2, nad3, nad4, nad4L, nad5, nad6, nad7 and nad9 from all *Blastocystis* MROs sequenced to date, several other eukaryotes (mostly stramenopiles) and bacteria. Amino acid sequences were aligned with MAFFT-linsi (Katoh et al., 2017) complemented by manual trimming of ambiguously aligned regions using AliView (ormbunkar.se/aliview) yielding a concatenated matrix of 2,873 amino acid sites. The software IQ-TREE (Nguyen et al., 2015) was used for maximum-likelihood phylogenetic estimation using the substitution model LG+C20+F+Γ (Yang, 1994; Le & Gascuel, 2008; Le et al., 2008) with statistical support for branches assessed by 1000 replicates of the ultrafast bootstrap approximation method (Minh et al., 2013).

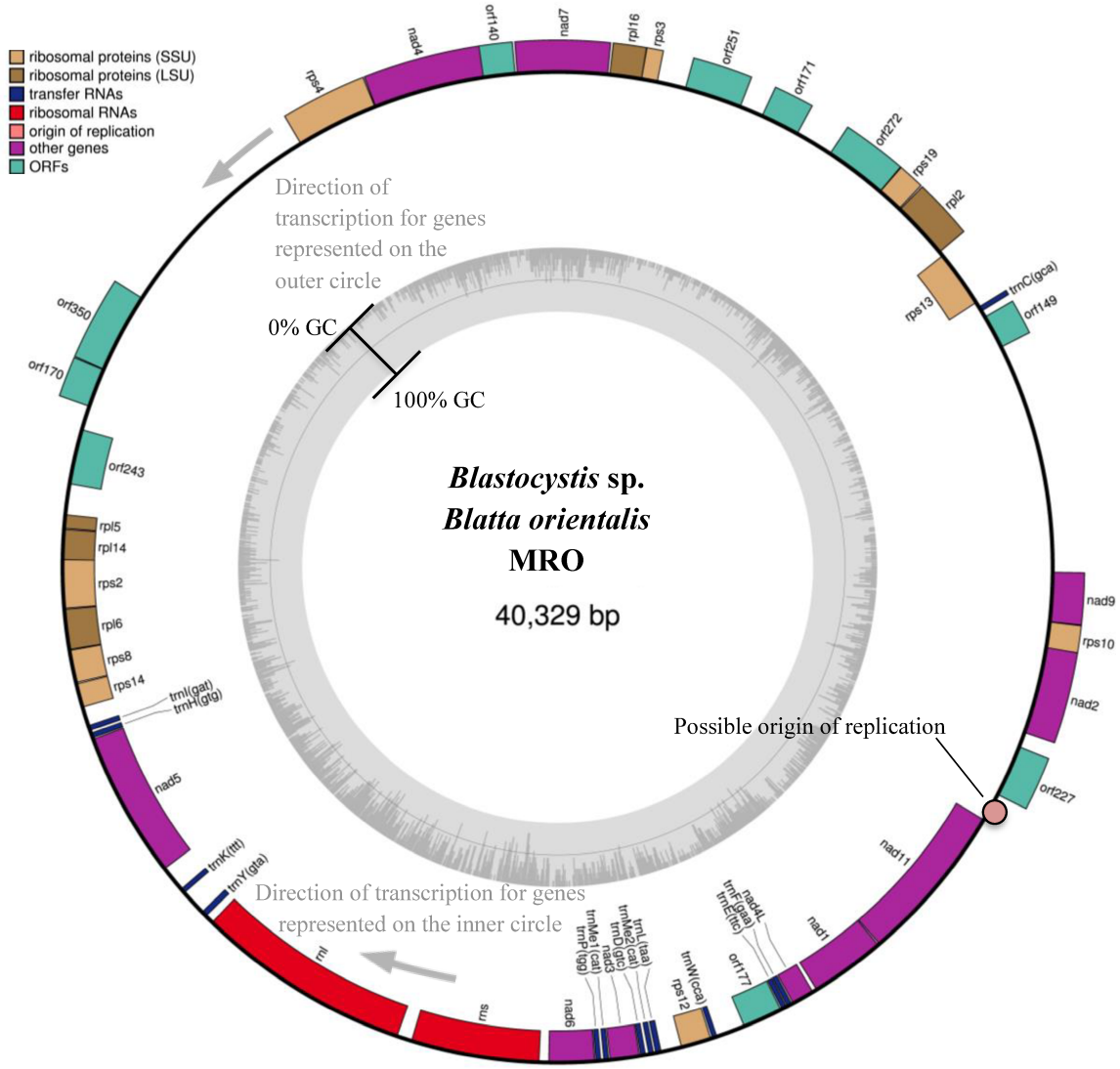
2.3 RESULTS

2.3.1 The MRO genome assembly of BBO

The first draft assembly of the BBO MRO genome showed a duplicated region (see Supplementary Figure S4) at each end of a single contig which elevated the size of the putative MRO chromosome to 50 kb long. These repeats were not inverted and appeared as though each was “copied-and-pasted” at each end of the contig. Since this

pattern has not been observed in any other MRO genomes of other *Blastocystis* subtypes or *Proteromonas lacertae*, validation of this region was needed. After designing a pair of oligonucleotide PCR primers (Table 2.1) targeting the ends of the chromosome and sequencing the PCR products, the region flanked by primers t10_44kc_f and t10_50kc_r was confirmed to be the genuine sequence (see Supplementary Figure S4). As the software Canu does not circularize contigs and it may assemble reads into different locations when they are not similar enough (this is presumably the result of an attempt to minimize SNPs or errors in a region as discussed in canu.readthedocs.io/en/latest/faq.html), I reassessed the duplicated areas comparatively by analyzing if this phenomenon of duplication also appears with other mitochondrial long read data assembled with Canu. For doing this, the published MRO genome of *Blastocystis* sp. ST6 strain SSI-754 (accession number KU900237.1) and a Canu long read assembly for the same strain produced in-house were compared. Interestingly, a falsely duplicated region of about 8kb in the long-read assembly was also observed only in the long-read assembly (i.e., the extra region was a non-inverted duplication of the sequence spanning from rps11 to tRNA-Phe). This duplication pattern looked similar to the BBO MRO genome draft sequence, hence the latter was subjected to additional inspection. After finding that 1) the amplified and re-sequenced fragment corresponded to only one of the duplicated regionS, and 2) both DNA and RNAseq Illumina paired-end reads aligned more evenly to this area than they did to the region that was not amplified, the duplicated regions were ruled as a bioinformatics artifact. Using all the gathered evidence, the regions were manually overlapped and re-aligned producing the mitochondrial genome of 40,329 bases schematically depicted in Figure 2.1.

Figure 2.1: A schematic of the MRO genome of *Blastocystis* sp. isolated from *Blatta orientalis*. OGDraw v1.2 (Lohse et al., 2013) was used to draw the annotated MRO genome. Inner grey circular graph shows GC content with 0% on the outside and 100% on the inside. The centre line represents 50% GC.



The MRO genome was recovered in one single contig. The breadth of coverage for gDNA and RNA (by at least 5× depth) was 70.3% and 40.5% of the genome respectively, with the average gDNA coverage depth of 184×. The BBO MRO genome is characterized by an extremely low GC content (13.6%) with a coding density of 59%. The GC content within coding regions is also low (see Supplementary Table S1), and while the intergenic regions (IGR) regions are large, there is evidence of overlapping

genes (Table 2.2). The genome encodes 13 tRNAs, 2 rRNAs, and, if the standard 'universal' genetic code is employed, 34 protein-coding genes are predicted; 10 of these were unidentified ORFs. Out of the total 49 genes, 35 had 1× or higher RNA coverage per base, indicating that they may be actively transcribed under the conditions in which the cultures were maintained. In an effort to discern if a different genetic code was being used, the BBO MRO genome was re-annotated using alternative genetic codes (e.g., yeast mitochondrial code, denoted as "translation table 3" by Elzanowski & Ostell (2016)). These alternative annotations resulted in fewer genes or truncated genes (i.e., these proteins did not align to their *Blastocystis* homologs at their N- or C- termini).

Two genes in other characterized *Blastocystis* subtype MRO genomes have non-canonical codon usage: orf160 has in-frame stop codons and rps4 uses alternative start codons (i.e., not ATG) (Jacob et al., 2016). The BBO orf140 may be homologous to orf160 in other *Blastocystis* MROs because, like orf160, it is located between the nad7 and nad4 genes. However, the putative amino acid sequences of orf140 and *Blastocystis* sp. ST1 NandII orf160 share 80% identity over only 20 residues (covering only about 35% of the length of orf140), which is not statistically significantly similar (cutoff is usually set at 30% identity over 100 residues (Pearson, 2013)). A protein vs. protein shuffle (PRSS) analysis (fasta.bioch.virginia.edu) between orf140 and orf160 also yielded no significant similarity. PRSS estimates the statistical significance by shuffling the second sequence up to 1000 times and calculating the alignment score from each shuffle. The BBO orf140 does not contain any in-frame stop codons. Similarly, rps4 in the BBO MRO genome has a standard start codon, in contrast to most other rps4 homologs in the *Blastocystis* subtypes which are missing their start codons.

Table 2.2. Genome statistics of *Blastocystis* spp. MRO genomes. All unassigned ORFs were treated as protein-coding genes. Analyses for all subtypes except BBO were done by Jacob et al. (2016).

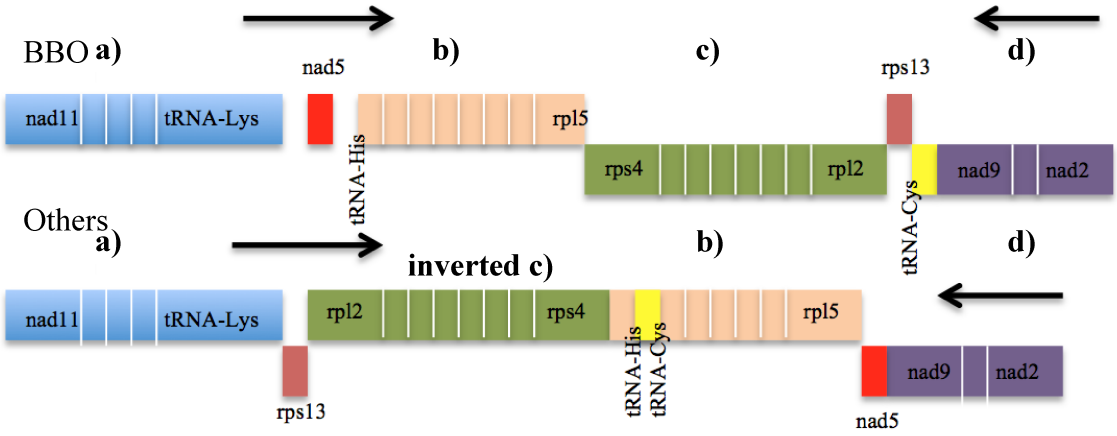
Genome Characteristics	GenBank Accession Number	Genome size (bp)	Coding density (%)	Intergenic region (IGR) content (%)	Sum of IGR sizes (bp)	Overlapped genes	Total length of overlap (bp)	Protein coding genes	tRNAs	GC content (%)
ST1 Nand II	EF494740	28,385	77.5	4.1	1,165	6	115	27	16	19.9
ST2 Flemming	KU900235	28,305	78.0	3.7	1,044	8	163	26	16	19.7
ST3 DMP/08-326	HQ909886	28,243	77.5	3.8	1,081	7	113	27	16	21.6
ST4 DMP/02-328	EF494739	27,718	77.1	3.5	964	8	126	27	16	21.9
ST6 SSI:754	KU900237	28,806	77.0	4.1	1,167	11	176	26	16	18.9
ST7 B	CU914152	29,270	77.1	4.8	1,399	7	193	26	16	20.1
ST8 DMP/08-128	KU900238	27,958	77.0	3.7	1,043	9	237	27	16	22.7
ST9 F5323	KU900239	28,788	77.3	4.1	1,179	11	204	26	15	18.8
BBO	To be determined	40,329	59.0	20.4	8,214	9	43	34	13	13.6

Eight out of the 13 tRNAs showed the typical cloverleaf structure: tRNA-Pro, tRNA-Trp, tRNA-Glu, tRNA-Phe, tRNA-His, tRNA-Lys, tRNA-Leu, and tRNA-Tyr, the latter six of which showed one or two non-canonical (i.e., non-Watson-Crick) G:U base pairing in stem regions. The large and small subunit rRNA genes were in a syntenic block in the same strand as other MROs, and were separated by a 183 bp intergenic region; other *Blastocystis* MROs (except for ST9) have a third tRNA-Met located in between these two genes (see GenBank accession IDs: EF494740, KU900235-KU900238, HQ909886-HQ909888, NC_027962, KU90023).

2.3.2 Synteny comparison of the MRO genome

Except for nad5 and tRNA-Cys, the blocks containing a) nad11 to tRNA-Lys, b) rpl5 to tRNA-His, c) rps4 to rps13, and d) nad9 to nad2 display synteny to other subtypes (Figure 2.2). Remarkably, the genes rps7, rps11, tRNA-Ala, tRNA-Asn, and a third tRNA-Met that are found usually in other *Blastocystis* MRO genomes are absent from the BBO MRO genome. Other MRO genomes possess three tRNA-Met, two of which are elongators: first between tRNA-Leu and tRNA-Asp, the second between nad3 and tRNA-Pro, and the third is an initiator tRNA-Met between rns and rnl. The unassigned ORF, orf177 (see Figure 2.1) located within what appears to be a syntenic block was suspected to be rps7. Its protein sequence was compared to the HMM profile of rps7 from other *Blastocystis* spp. and stramenopiles, but no evidence for homology was obtained. Its secondary structures also showed no similarity to rps7 from closely-related stramenopiles. A similar approach was used to see if the remaining unassigned ORFs (except orf140) were rps7 or rps11 homologs, but again no positive matches were obtained (Supplementary Table S2 and Supplementary Figure S5).

Figure 2.2: Synteny comparison of the BBO MRO genome (top) and a representative *Blastocystis* sp. MRO genome (bottom), linearized, not to scale. Blocks of conserved synteny are labelled alphabetically. The arrows show transcriptional direction, with blocks above the centre line transcribed towards the right and vice versa. Only genes on the borders of each block are shown. Unassigned ORFs, intergenic regions, and missing genes are not shown. The blocks a), b), and d) are conserved in synteny and transcriptional direction. The BBO block c) is inverted, shifted, and on the opposite strand. tRNA-Cys appeared to have moved out of b) block and moved to the start of d) block in BBO. The nad5 gene is located upstream of b) block in BBO rather than at the start of d) block in other subtypes.



2.3.3 Identifying the unassigned ORFs

I compared the RNA read coverage of the unidentified ORFs with the coverage of annotated genes (Table 2.3). An extremely low coverage suggests that orf170, orf171, orf243, and orf251 are poorly expressed. It is possible these are not real genes, although other known annotated genes had even lower coverage (e.g., nad6). Since the RNA used to generate the RNA-seq data was poly(A)-selected, MRO genes are not expected to be well-represented and should have markedly lower coverage than nuclear genes. Therefore, the lack of RNA coverage alone is not proof that these are pseudogenes. Homologs of tRNA-Asn and tRNA-Ala were found in the nuclear genome (by using TRNAscan-SE (Lowe & Eddy, 1997) and InterProScan (Jones et al., 2014)). A BLASTn (Altschul et al., 1990) search against the nt database revealed that the nuclear-encoded

tRNA-Asn and tRNA-Ala were cytosolic and not of the mitochondrial type (best hits were to homologs in nuclear genomes of other eukaryotes). Rps7 and rps11 homologs were found in the BBO transcriptome, but a BLASTp (Gish & States, 1993) search against the nr database showed that they were most similar to cytosolic versions. To assess if these homologs carried mitochondrial targeting peptides, the online service TargetP (http://www.cbs.dtu.dk/services/TargetP/) was used. The probabilities of these rps7 and rps11 homologs carrying any type of targeting peptides were below 0.5, indicating that they are unlikely to be secreted or targeted to any organelles. This supports the BLASTp result that they are likely cytosolic.

Table 2.3. RNA-seq coverage data over selected MRO genes. Genes are arranged in order of descending read coverage.

Gene	Gene start	Gene End	Direction	Reads/bp
rpl2	4,425	5,183	+	484.1
rnl	25,284	28,133	-	434.3
nad7	9,429	10,610	+	344.2
rns	28,318	30,011	-	39.08
trnF(gaa)	33,340	33,412	-	14.83
orf272	5,528	6,346	+	13.57
rpl14	19,667	20,071	-	13.52
$trnM_{e2}$(cat)	31,432	31,504	-	10.53
orf177	32,720	33,253	-	8.70
nad11	34,781	36,904	-	7.66
orf149	2,954	3,403	+	6.91
nad9	39,618	40,268	+	6.56
orf140	10,630	11,052	+	6.08

Table 2.3. (Cont.)

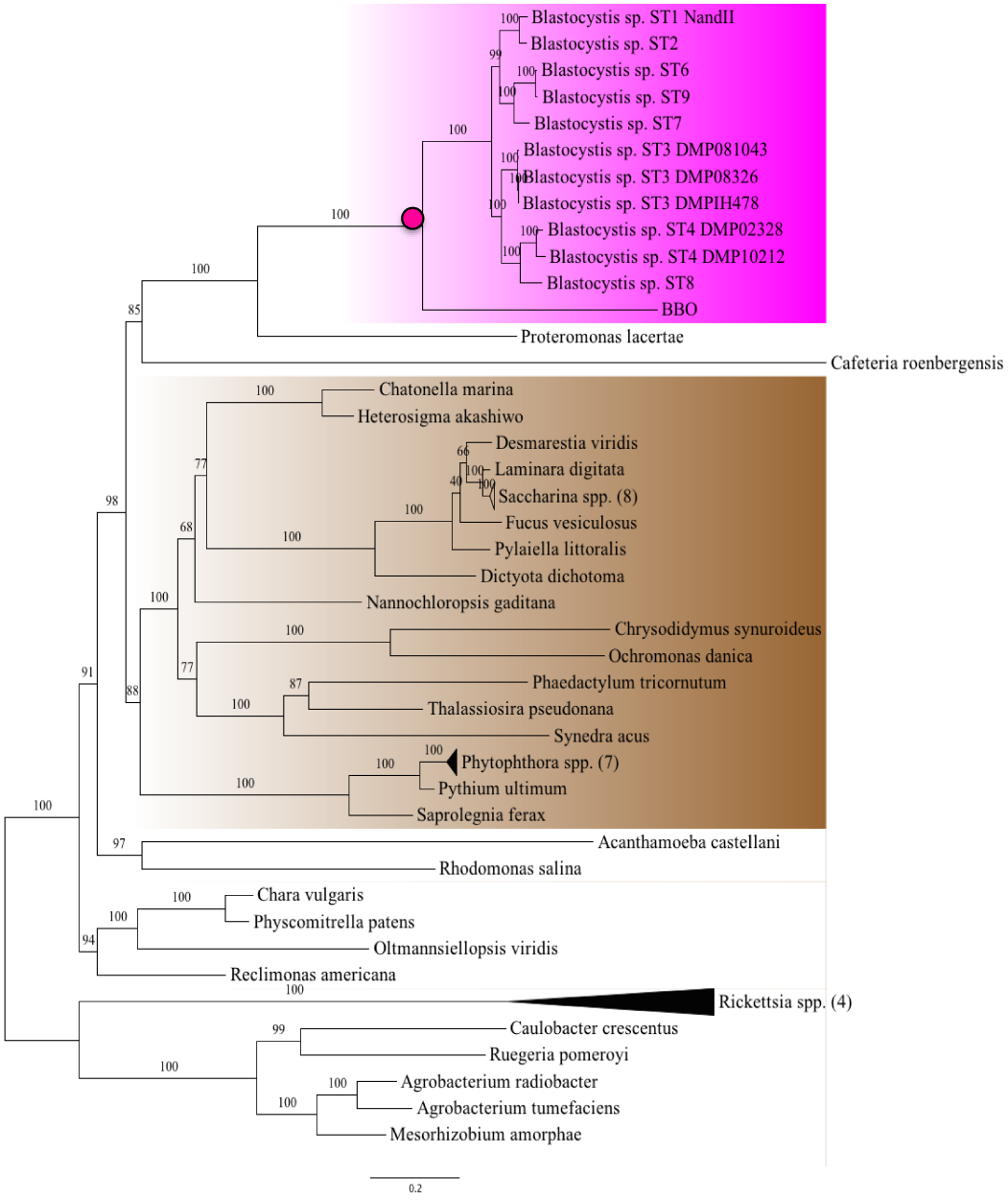
Gene	Gene start	Gene End	Direction	Reads/bp
rps10	39,277	39,621	+	5.71
orf227	37,269	37,952	+	3.97
rps4	12,505	13,584	+	2.23
orf350	16,476	17,528	+	2.18
nad5	22,434	24,368	-	2.17
nad3	30,924	31,319	-	0.86
orf171	6,838	7,353	+	0.61
orf251	7,656	8,411	+	0.61
orf170	17,531	18,043	+	0.20
orf243	18,352	19,083	-	0.10
trnC(gca)	3,485	3,554	+	0.03
trnK(ttt)	24,697	24,769	-	0.00
nad6	30,127	30,732	-	0.00
$trnM_{e1}$(cat)	30,843	30,914	-	0.00

The nuclear genome was searched using BLASTn (Altschul et al., 1990) with the MRO genome as the query to check if any of its fragments were present in the nuclear genome as a nuclear mitochondrial DNA (NUMT), but results were negative. NUMTs are mitochondrial DNA (mtDNA) that have been transferred to the nuclear genome via nonhomologous recombination of nuclear DNA with leaked mtDNA from damaged mitochondria (Henze & Martin, 2001; Mourier et al., 2001; Woischnick & Moraes, 2002). NUMTs can be whole or highly fragmented and rearranged, exist in one or more copies, remain unexpressed or function as they did in the mitochondrial genome, or even create functional novel genes (Richly & Leister, 2004; Noutsos et al., 2007).

2.3.4 The phylogenetic placement of the BBO MRO

A comparison of the MRO nad genes of BBO with all known *Blastocystis* MRO genomes supports the prediction that it is deeper-branching than the mammalian/avian subtypes (Figure 2.3). As shown by Jacob et al. (2016) the mammalian *Blastocystis* spp. diverge into four main groups: 1) ST1 and ST2, 2) ST4 and ST8, 3) ST6, ST7, and ST8, and 4) all strains of ST3. BBO diverges from a deeper node before the branch leading to their most recent common ancestor. *P. lacertae* branches outside of the whole *Blastocystis* clade (the branching order is supported by bootstrap values (BVs) of 99-100%). The deeper branching positions of the outgroup taxa are well-supported as well (BVs=85-100%), but the inner branches within the distantly-related stramenopiles are less resolved (BVs=66-100%). The branch length of BBO to the common ancestor of the *Blastocystis* spp., which represents the number of substitutions per amino acid site that have accrued along this lineage, is nearly twice as long as the other subtypes, implying that BBO has the fastest rate of evolution among them.

Figure 2.3: Maximum-likelihood phylogenetic tree of concatenated *nad* genes from mitochondrial genomes of stramenopiles, and their homologs from several other eukaryotes and prokaryotes. The substitution model LG+C20+F+Γ was used for a concatenated matrix of 2873 amino acid sites. Bootstrap values are shown above each branch. The *Blastocystis* clade is highlighed in magenta and distantly related stramenopiles in brown. The common ancestor of all *Blastocystis* spp. is represented by the magenta circle.



2.4 DISCUSSION

2.4.1 Characteristics unique to the BBO MRO genome

The MRO genome of BBO was at least 10 kb bigger than other *Blastocystis* MRO genomes due to bigger intergenic regions (IGR) (20.4%, see Table 2.2). This case is similar to nonbilaterian animal mitochondrial (mt) DNA, in which most of the size variation comes from the number of noncoding nucleotides (Lavrov & Pett, 2016). The bigger IGR could be the result of an expansion unique to BBO, or that BBO has a more conserved form of a large ancestral MRO genome, whereas the other mammalian isolates shrank as they lost most of the IGRs. The deeper-branching *Proteromonas lacertae* mitochondrial genome is 48.7 kb, of which 13.2% is IGR (Pérez-Brocal et al., 2010), lending support to the latter hypothesis.

The GC content of mitochondrial genomes is generally low with an average of 35% GC (Smith, 2012); the lowest GC content is 12.6% from the yeast *Candida castelli* (Bouchier et al., 2009). The low GC content of mitochondrial genomes is hypothesised to be the result of increased chances of replication errors due to mtDNA undergoing multiple rounds of replication per cell division (Birky, 2001), spontaneous deamination of cytosine to uracil (Kennedy et al., 2013), and the high concentrations of reactive oxygen species within the mitochondrion which promote the oxidative conversion of guanine to 8-oxo-guanine (excision and repair of the latter results in a guanine→thymine change) (Shokolenko et al., 2009). The BBO MRO genome has the lowest GC content among *Blastocystis* spp. (13.6% vs. 18.8-22.7%). This may be the consequence of a faster rate of evolution, having accumulated the highest amount of mutations among all known *Blastocystis* MRO genomes (see branch length of BBO in Figure 2.3).

The reason why BBO might have accumulated higher AT content in its MRO genome than other species could be related to its population genetic history. GC to AT nucleotide replacements can be deleterious if they are non-synonymous substitutions that switch out important amino acid residues in the cores or catalytic domains of proteins, with residues of different physicochemical properties that render the proteins non-functional, or produce truncated proteins by introducing in-frame stop codons. Many non-synonymous mutations in proteins will not completely destroy protein function, but will slightly lower their stability or catalytic efficiency and are therefore only slightly deleterious. Residue changes to surface loops not involved in binding or folding may not affect the protein much at all (Goo et al., 2004). The fates of these kinds of mutations in populations depend on how much they lower the fitness of the carrier and the effective population size of the organism. Slightly deleterious mutations can become fixed (i.e., eventually rise to 100% representation) in populations with a small effective population size (i.e., a low number of individuals producing offspring or a population that goes through a temporary bottleneck). This depends on the absolute value of the selection coefficient *s* for the mutant allele in the Wright-Fisher model (Fisher, 1930; Wright, 1931). If *s* is smaller than the inverse of the effective population size N_e (i.e., $|s|<\frac{1}{N_e}$), then the mutation is said to be neutral and, other things being equal, its dynamics in the population are mostly determined by random genetic drift (Kimura, 1968; King & Jukes, 1969). Thus, the smaller the population, the higher the probability a neutral or slightly deleterious mutation will get fixed (see Popadin et al. (2007) for examples of this phenomenon in mammalian populations). We suggest that the extremely high AT content and high evolutionary rate of the MRO genes in the BBO lineage may be the result of

either a population bottleneck or a generally smaller population of the parasite in the Oriental cockroach relative to mammalian/avian *Blastocystis* subtypes. The number of bacterial cells in the human colon is estimated at 4×10^{13} (Sender et al., 2016) versus 4×10^{8} (Schauer et al., 2012) in the colon of the Turkestan cockroach, *Shelfordella lateralis* (member of the family Blattidae, which includes the Oriental cockroach). If the 10^{5} fold decrease in bacterial community size is also found in BBO population sizes, this is consistent with a smaller effective population size if the cockroach population itself is not 10^{5} fold larger than humans. However, this remains highly speculative and must be further explored. I discuss potential future studies in chapter 3.4.3.

Some synteny was observed between the mtDNA of BBO and those of other *Blastocystis* subtypes. The exception is an rps4 to rpl2 "block" of genes that appears to have moved and inverted (Figure 2.2). Gene rearrangement is common in mitochondrial genomes across animals, plants, and protists (Dowton et al., 2009; Liu & Cui, 2010; Sloan, 2013; Gray et al., 2004). Short lengths (2-4 genes in a row) of protein-coding and ribosomal subunit genes of the BBO MRO genome show synteny to the *P. lacertae* mitochondrial genome, with the largest conserved block being the 7 kb stretch from the large subunit rRNA to rps12 (which partially corresponds to the repeated region in *P. lacertae*), but the tRNAs are completely rearranged. The MRO genomes of other subtypes share no more synteny with the MRO genome of *P. lactertae* than that of BBO. More sampling of earlier-branching *Blastocystis* subtypes would be required to reliably infer the ancestral *Blastocystis* MRO genome structure.

The predicted secondary structures of BBO tRNAs were similar to the typical cloverleaf structures of tRNAs, though some of their stem regions showed a non-

canonical nucleotide pairing of G:U. These non-Watson-Crick pairings arise when the different edges of RNA bases interact in *cis* or *trans* orientations (Leontis et al., 2002). This pairing has also been detected in other *Blastocystis* MRO genomes (Jacob et al., 2016), as well as in tRNAs of *B. taurus* and *T. thermophilus* (Chawla et al., 2015). No BBO-specific mitochondrial tRNAs nor specific idiosyncratic tRNA structural features were found.

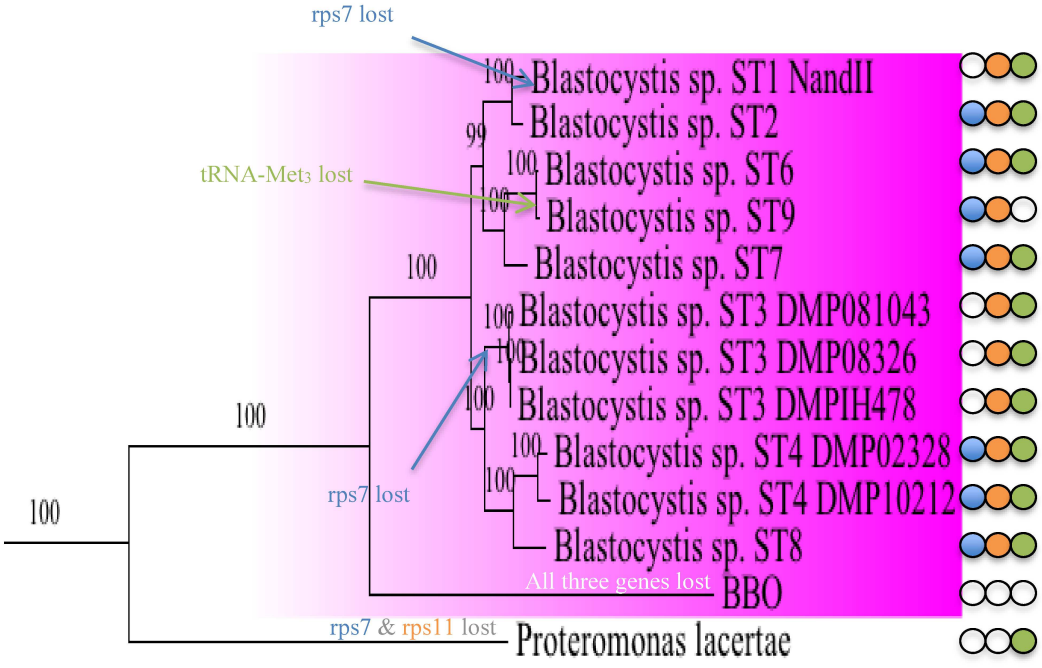
2.4.2 Gene losses

Genomes can lose genes through several mechanisms. Physical "deletions" can happen by unequal crossing over during meiosis. Transposable elements (see Chapter 3.1) can result in genes being cut out or interrupted. Another mechanism is through adaptive selection or neutral drift. Accumulation of nonsense mutations can produce truncated proteins and insertions/deletions (indels) that cause framshifts (see Albalat & Cañestro, 2016). As discussed in the previous section, the gene losses in the BBO genome likely happened through non-physical mechanisms. i.e., by selection or neutral drift.

The putative absence of mitochondrial rps7, rps11, tRNA-Asn, tRNA-Ala, and the third tRNA-Met is not specific to the BBO MRO genome. Homologs of rps7, tRNA-Asn, and tRNA-Ala were found in the BBO nuclear genome, but their protein products were predicted to be cytosolic and not MRO-targeted. In angiosperms, the loss of mitochondrial rps11 is believed to be an ancient event that occurred after its transfer to the nuclear genome, while the loss of mitochondrial rps7 has been observed in several lineages without evidence of transfer to the nuclear genome (Adams et al., 2002). Rps7 is missing from *Blastocystis* ST1 NandII MRO (GenBank: EF494740.3) and all strains of

Blastocystis sp. ST3 MRO (GenBank: HQ909887.2, HQ909886.2, HQ909888.2), and the third tRNA-Met is missing from *Blastocystis* sp. ST9 MRO (GenBank: KU900239.1). Both rps7 and rps11 are absent from the *P. lacertae* (GenBank: GU563431.1) MRO genome. This suggests that many eukaryotic lineages have independently lost mitochondrial rps7, rps11, and tRNA-Met$_3$, and therefore BBO is not out of the ordinary in this respect (gene losses are illustrated in Figure 2.4). In summary, assuming that mitochondrial rps7, rps11, and tRNA-Met$_3$ were present in the common ancestor of the opalinitans, independent losses of one or more of these genes occurred in some lineages as they diverged.

Figure 2.4: Gene losses of rps7, rps11, and the third tRNA-Met among all *Blastocystis* spp. MRO genomes and in the deeper-branching reptile gut parasite *Proteromonas lacertae*. Tree branch lengths not to scale. Circles in blue: rps7, orange: rps11, green: tRNA-Met$_3$, white: absence of gene. *Blastocystis* spp. highlighted in magenta.



Rps7 initiates the assembly of the SSU rRNA along with rps4 (Maguire & Zimmermann, 2001); in organisms which completely lost rps7, I speculate that its function has been taken up by another gene, or solely by rps4. Rps11 plays a role in mitochondrial translation (UniProtKB: P82912), but its function may have been replaced by one of the many other ribosomal proteins involved in this process as well. As for the missing mitochondrial tRNA-Asn and tRNA-Ala, these could also be independent losses – several protist lineages have lost all mitochondrial genome-encoded tRNAs, such as apicomplexans (Feagin, 2000) and trypanosomes (Simpson et al., 1989). In trypanosomes, nuclear-encoded cytosolic tRNAs are hypothesized to be imported into the mitochondria as "naked molecules" (i.e., no carrier molecules are bound to them) through the same translocase of the outer membrane (TOM) complex that other proteins use (Niemann et al., 2017). Studies of the tRNA import system in *Blastocystis* MROs could give insights into mitochondrial genome reduction in parasitic protists.

Comparisons of HMMs and secondary structure predictions of the nine unassigned ORFs did not reveal any putative homologs with known functions. These ORFs may be pseudogenes, novel genes, or simply too divergent to be recognized as homologs of existing genes. It is not uncommon for protist mitochondrial genomes to have unidentified ORFs (Gray et al., 2004). Sampling more deeply-branching *Blastocystis* MROs could help determine if they are evolutionarily conserved amongst *Blastocystis* spp. or reveal homologs in other taxa. Organelle purification followed by mass spectrometry proteomic analyses have been successful in identifying novel mitochondrial proteins from the protist *Tetrahymena thermophila* (Smith et al., 2007).

This approach may prove useful in determining if there are proteins expressed from these unassigned ORFs in the various *Blastocystis* mitochondrial genomes.

CHAPTER 3 THE NUCLEAR GENOME OF *BLASTOCYSTIS* SP.

3.1 INTRODUCTION TO NUCLEAR GENOMES OF *BLASTOCYSTIS*

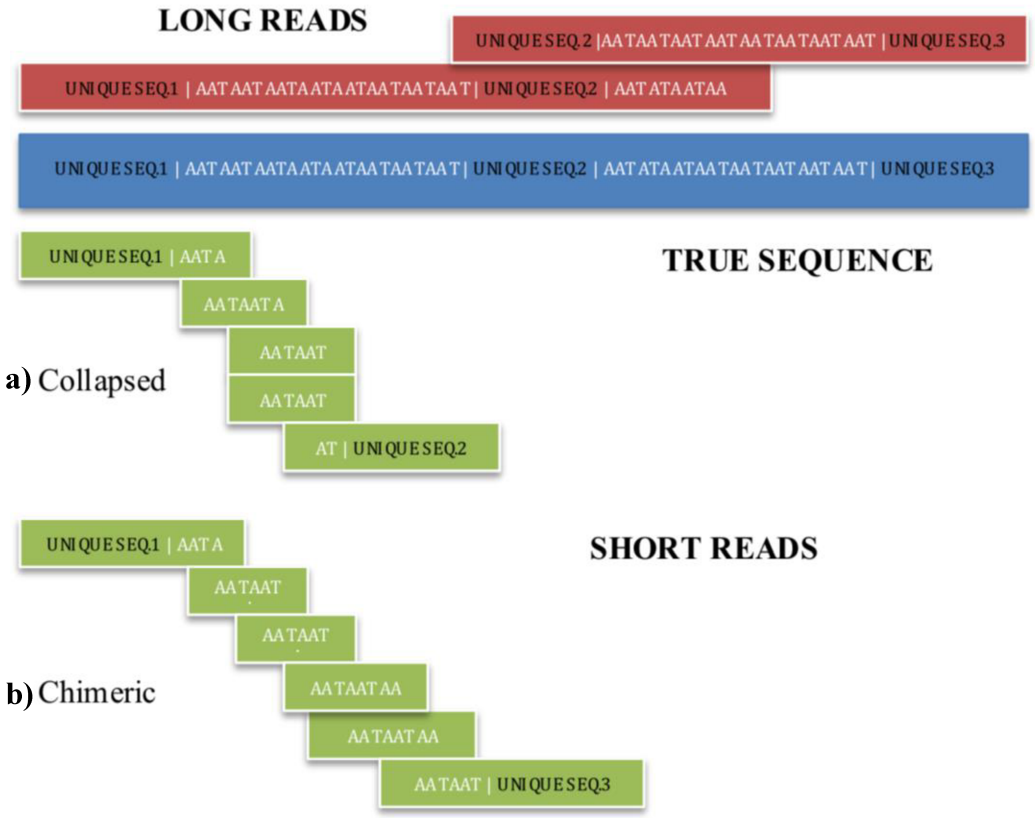
Nuclear genomes of eukaryotes are rife with repeated sequences. Repeats can be categorized into three major groups: transposable elements (TE) which are interspersed throughout the genome, short or long tandem repeats which sit next to each other either directly or inverted (e.g., microsatellites), and structural repeats with known functions such as telomeric and centromeric repeats (Biscotti et al., 2015). TEs invade genomes and replicate in various ways: a retrotransposon is transcribed into RNA, reverse transcribed into DNA, which is then inserted into a different location in the host genome. A DNA transposon is cut out of the genome and is ligated back at a different site. A helitron, also a type of DNA transposon, can replicate via a "rolling circle" mechanism, where the donor sequence rolls into a circle then is copied into a target on the genome (Kapitonov & Jurka, 2007). Each type can be autonomous, i.e., code for their own reverse transcriptase or transposase, or be dependent on other TEs if they lack these enzymes themselves. Over two-thirds of the human nuclear genome is comprised of TEs, mostly from retrotranscribed repeats such as LINEs, SINEs (**l**ong/**s**hort **i**nterspersed **n**uclear **e**lements), and LTRs (**l**ong **t**erminal **r**epeats, thought to have originated from exogenous retroviruses that integrated into the genome of germ cells) (de Koning et al., 2011).

Another mechanism by which repeated regions arise in eukaryotic genomes is by whole or partial genome duplication events. This has been extensively studied in the case of the yeast *Saccharomyces cerevisiae* (Wolfe, 2015) and in animals, plants, and other microbial eukaryotes (Dehal & Boore, 2005; Jiao et al., 2011; Aury et al., 2006). This is

particularly relevant to the genomes of *Blastocystis* spp., since Deneoud et al. (2011) found evidence of whole genome duplication in *Blastocystis* sp. ST7; 39% of the genome was composed of two or more copies of blocks of paralogous genes.

Repeated sequences in genomes can be hard to analyze using short read sequencing data. Although Illumina sequencing generates short and accurate reads of up to 300 bp, assembling them properly into a full genome sequence can be problematic if the genome has an extremely high or low GC content (Chen et al., 2013) or if there are many repetitive regions longer than the fragments that are produced by Illumina library preparation. This problem is illustrated in Figure 3.1 that depicts how repeat regions can be "collapsed" or form chimeric scaffolds (Treangen & Salzberg, 2011). Long-read sequencing data, by contrast, can sometimes span repeat regions to anchor them in unique sequence on either side avoiding the repeat "collapsing" and chimera formation problems. The downsides are that current long-read sequencing technology requires higher quality and larger amounts of DNA as input, has a high intrinsic error rate (up to 20% (Weirather, 2017)) and has difficulty in determining homopolymer regions (i.e., repeats of the same nucleotide) during base calling. Basecalling using Oxford Nanopore technology is the process of assigning nucleobases identitites to electrical signals that are emitted from pores on the flow cell as single strands of DNA pass through the pores, disrupting the ion gradient across the membrane. Homopolymers induce many repeats of the same electrical signals, making it difficult for basecalling software tools to precisely determine their length.

Figure 3.1: Two examples of problems encountered with short read sequencing data. The reads can collapse into a false short sequence if **a)** overlaps are not found, and/or form **b)** chimeric sequences by joining wrong repeat units. Long reads can resolve these problems by spanning the repeat region and its adjacent unique regions. Red: long reads, blue: true sequence, green: short reads.



To improve the error-prone long-read data, further polishing and correcting steps are taken. Polishing improves the assembly by evaluating the probability of observing a sequence of electric signals for a given nucleotide sequence and modifies the assembly iteratively (Loman et al., 2015). Correcting the assembly involves aligning accurate short reads to the assembly and fixing indels, substitutions, or gaps. Thus short reads and long reads can together produce a high-quality genome assembly. Previous in-depth genomic studies of *Blastocystis* subtypes have involved assembling genomes from either short reads or Sanger sequencing: *Blastocystis* sp. ST7 was sequenced solely from Sanger sequencing (Denoeud et al., 2011), and Illumina HiSeq was used for *Blastocystis* sp. ST1

NandII (Gentekaki et al., 2017) and *Blastocystis* sp. ST4-WR1 (Wawrzyniak et al., 2015). Here I aimed to use a combination of long reads and short reads to produce a high-quality assembly of the BBO nuclear genome.

In this chapter I present the results from my preliminary characterization of the BBO nuclear genome. After presenting some basic nuclear genome statistics of BBO, I compare its gene set and degree of amino acid divergence in orthologous proteins with *Blastocystis* ST1, ST4, and ST7. Gentekaki et al. (2017) found that the amino acid identities between orthologs from subtype pairs ranged from 59%-61% and that 6-20% of their gene sets were unique in pairwise comparisons (i.e., not found in the other subtype). BBO, being an earlier branching subtype, can be expected to have a larger set of unique genes and a higher amino acid sequence divergence between orthologs. I have also conducted a concatenated phylogenomic analysis with nuclear-encoded protein-coding genes to definitively place BBO in particular, and the opalinitans in general, in the stramenopile tree.

Four genes of evolutionary significance were searched for in the predicted gene set of BBO. Two of them, pyruvate:ferredoxin oxidoreductase (PFO) and iron-only [FeFe] hydrogenase, are enzymes involved in anaerobic ATP production. They have been found in *Blastocystis* sp. ST1, ST4, and ST7. The other two are cyanobacteria-derived genes found in ST1 and ST7 (Denoeud et al., 2011): phosphoglycerate kinase and 6-phosphogluconate dehydrogenase. Denoeud et al. (2011) implied that their presence in the genomes of *Blastocystis* may have been a result of secondary loss of plastid in the *Blastocystis* lineage. This may indicate that the common ancestor of all stramenopiles possessed a plastid derived from secondary endosymbiosis.

I also analyse several cases of lateral gene transfer (LGT) in *Blastocystis* BBO. Lateral gene transfer (also known as horizontal gene transfer) is "*the movement of genetic information across normal mating barriers, between more or less distantly related organisms, and thus stands in distinction to the standard vertical transmission of genes from parent to offspring*" (Keeling & Palmer, 2008, p.605). It has been well-characterized among prokaryotes, a famous example being bacterial drug resistance transfer (Gyles & Boerlin, 2013). However, the mechanism and occurrence of LGT among eukaryotes is poorly understood (Husnik & McCutcheon 2018; Leger et al., 2018). Proposed mechanisms include viral transduction and plasmid transfer via conjugation or transformation (Husnik & McCutcheon 2018).

A previous study has shown that 74 genes from a total of 6,544 genes in *Blastocystis* sp. ST1 were laterally acquired from prokaryote and eukaryote donors (Eme et al., 2017). I searched for homologs of a selection of these previously-reported LGTs in BBO and its opalinitan relatives: *Blastocystis* subtypes KOZE1a (isolated from the Solomon Islands skink, *Corucia zebrata*) and TEHE3a (isolated from Hermann's tortoise, *Testudo hermannii*), *Opalina* sp., a gut parasite of frogs, and in *Proteromonas* sp., a gut parasite of lizards. I also searched in lineages closely related to the opalinitans. These lineages are free-living marine isolates, grouped into facultative aerobes and obligate anaerobes. The former group includes: *Cafeteria* sp. (Caron, 2000), *Incisomonas marina* (Cavalier-Smith & Scoble, 2013), and the halophilic *Halocafeteria seosinensis* (Park et al., 2006). *Rictus lutensis* (Yubuki et al., 2010) and *Cantina marsupialis* (Noguchi et al., 2015) are included in the latter group. The goal is to determine the relative timings of acquisition of these LGTs along the opalinitan stramenopile lineage, and to determine if

any of them were inherited from a common ancestor of all known *Blastocystis* lineages before they diversified.

3.2 METHODS

See Chapter 2.2.1-2.2.2 for culturing and nucleic acid extractions of BBO. For phylogenomic and LGT analyses (sections 3.2.5 and 3.2.10) *Blastocystis* sp. KOZE1a and TEHE3a were added. They were grown axenically in Ivan Čepička's lab at Charles University, Prague. RNA was extracted and sequenced employing the same methods as described for BBO in section 2.2.2. The transcripts were assembled using Trinity (Grabherr et al., 2011) with default parameter settings.

3.2.1 Illumina-only assembly

As Illumina short-read data was the first dataset available for this project (obtained 6 months before Nanopore long-read data), it was assembled for preliminary analysis. Adapter sequences were trimmed away from the raw reads using Trimmomatic (Bolger et al., 2014). The trimmed reads were then assembled using SPAdes (Bankevich et al., 2012) with default parameter settings.

An SSU rRNA phylogenetic tree was estimated for the parabasalid contaminant (see preliminary results in section 3.3.1). First, the SSU rRNA sequence was found in the parabasalid contigs using RNAMMER (Lagesen et al., 2007). It was aligned to parabasalid SSU rRNA sequences, including unpublished sequences obtained from Dr. Gillian Gile's laboratory (gilliangile.com) at Arizona State University, using MAFFT (Katoh et al., 2017). The alignments were then trimmed using trimAl (Capella-Gutierrez et al., 2009) with default settings. Then a maximum-likelihood tree was estimated using

IQ-TREE with an automatic substitution model search setting (the GTR+R7 model (Tavaré, 1986; Yang, 1995; Soubrier et al., 2012) was selected as optimal), with a 1000-replicate ultrafast bootstrap approximation for branch support (Minh et al., 2013).

3.2.2 Decontamination of the genome

Data from MinION sequencing were assembled, polished, and corrected as described in Chapter 2.2.3 with different Canu settings: corOutCoverage was set to "all", corMinCoverage at 0, genomeSize at 5 Mb, and correctedErrorRate at 0.105. This was used to ensure the maximum amount of sequencing data was used and that no organism, including any bacterium, was missed. The BBO genome was separated from prokaryotic sequences using the following approach:

1) The sequences of contigs from the polished and corrected assembly were used as queries in BLASTn (Altschul et al., 1990) searches against the comprehensive "nt" database (National Center for Biotechnology Information, 1988). Contigs that hit bacteria and were around the average size of a bacterial genome (5Mb (Land et al., 2015)) were discarded. All other contigs were kept. For this sequencing run, the bacterial contigs constituted 78% of the total assembly size.
2) The contigs retained from step 1 were visualized using Anvi'o (Eren et al., 2015). Anvi'o splits each contig into several pieces and then executes DIAMOND BLASTx (Buchfink et al., 2015) searches against the non-redundant (nr) database. Anvi'o then clusters these contig pieces into "splits" according to criteria set by the user, such as GC content and coverage information. Anvi'o also integrates Barrnap (Seemann, 2013) to identify rRNA genes using HMM models from bacteria. Any contig that contained bacterial rRNA was flagged. Contigs that

contained splits that had hits to organisms other than *Blastocystis* were also flagged.

3) RNAMMER (Lagesen et al., 2007) was used to predict eukaryotic rRNA genes in the eukaryotic contigs. The sequences identified here were searched via BLASTn (Altschul et al., 1990) against the nt database, and if they were not best matches to *Blastocystis* rRNA seqences in the database, they were flagged.
4) Read coverage information was obtained by mapping reads to the contigs flagged from step 2 and step 3 using Bowtie2 (Langmead & Salzberg, 2012) for short reads, and minimap2 (Li, 2017) for long reads.
5) The flagged contigs were kept if the majority of the splits hit *Blastocystis*. If the contaminating split was surrounded by *Blastocystis* splits, and had average or above-average short-read and long-read coverage, the contig was also kept.

The quality of the final genome assembly was assessed using QUAST (Gurevich et al., 2013). Depth and breadth of coverage of sequence reads over the genome were calculated using samtools (samtools.sourceforge.net/).

3.2.3 Gene prediction

Bruce Curtis and Dayana Salas-Leiva, postdoctoral researchers at Roger Lab in Dalhousie University, contributed to some of the scripts and workflows used in this section.

RNA reads were mapped onto the assembled genome by HISAT2 (Langmead & Salzberg, 2015) with standard splice junction settings (i.e., GT-AG sequences for the exon-intron junction) and the resulting .bam file was checked by eye using IGV to ensure that the reads were mapped correctly. MapSplice (Wang et al., 2010) was used to

corroborate the results from HISAT2. The depth of coverage of RNA-seq reads mapping onto the genome was generated by samtools (samtools.sourceforge.net/). Then an in-house script developed by Bruce Curtis was used to assess reliability of the predicted introns (e.g., if the first few bases into a putative intron showed a coverage of 1 or 2, but the flanking exons were covered by more than 10, said bases were considered an intron, i.e., the ratio of coverage between flanking exons and the intron must be over 10). The script then created a "hints" file containing intron information. Next, a three-step gene prediction workflow developed by Dayana Salas-Leiva was performed:

1) The hints file was fed into GeneMark-ET (Lomsadze et al., 2014), a program that uses unsupervised training, a form of machine learning, to recognize gene features such as exons, introns, and ORFs on the genome assembly. Here, canonical start and stop codons were used.
2) The predicted genes from the previous step were provided to AUGUSTUS (Keller et al., 2011), where they were sub-sampled into "buckets" to iteratively train the program to detect more genes. Predicted genes with introns >2kb in length were manually inspected to check for potential mis-predicted "chimeric" genes; the latter were subsequently split into separate predicted genes using an in-house script.
3) Assembled transcripts (assembled using Trinity (Grabherr et al., 2011) with the strand-specific setting: --SS_lib_type RF) were mapped onto the predicted genes from the previous step using PASA (Haas et al., 2003), and genes were validated if they coincided with transcripts.

These triple-validated genes were used for subsequent analyses. Statistics of repeated amino acids (homopolymers) were generated for the predicted genes of BBO, *Blastocystis* sp. ST1, ST4-WR1, and ST7 using an in-house script developed by Bruce Curtis.

3.2.4 Searching for genes of evolutionary significance

Four genes of particular interest, PFO, [FeFe] hydrogenase, phosphoglycerate kinase, and 6-phosphogluconate dehydrogenase, were retrieved from the BBO predicted gene set using BLASTp (Gish & States, 1993) searches with ST1 orthologs as queries. The retrieved BBO genes were then used as queries in BLASTp (Gish & States, 1993) searches against the nr database to confirm that their best matches were indeed to these protein families. Single gene phylogenetic trees were constructed from the latter two proteins to determine their origin. They were each aligned to their first 1000 best-matching sequences in the nr database based on BLASTp (Gish & States, 1993) searches using MAFFT (Katoh et al., 2017). The alignments were then trimmed using trimAl (Capella-Gutierrez et al., 2009) with default settings. Single gene trees were estimated by maximum-likelihood using IQ-TREE with the substitution model LG4X (Le et al., 2012) with a 1000-replicate ultrafast bootstrap approximation for branch support (Minh et al., 2013). To assess if these homologs carried targeting peptides, the online service TargetP (http://www.cbs.dtu.dk/services/TargetP/) was used. If TargetP showed inconclusive results, MitoProt (https://ihg.gsf.de/ihg/mitoprot.html) was then used to predict targeting peptides.

3.2.5 Phylogenomic analysis

A dataset of 351 *Arabidopsis thaliana* proteins (Brown et al., 2018; Kang et al., 2017), which are well-conserved among eukaryotes, was used as queries to retrieve homologs from the predicted gene set of BBO. An in-house pipeline was used to construct a multigene phylogenomic tree. In brief, the "pipeline" performed the following series of steps: BLASTp (Gish & States, 1993) searches were used to retrieve homologs of the 351-gene dataset. The retrieved sequences were then aligned to a database of homologs (collected in-house and curated by Brown et al., (2018)) from various fungi, plants, animals, and protists using MAFFT-einsi (Katoh et al., 2017), and the alignment was trimmed using BMGE (Criscuolo & Gribaldo, 2010) using default settings. Single-protein trees were then made for each of the proteins using IQ-TREE (Nguyen et al., 2015) using a 1000-replicate ultrafast bootstrap approximation (Minh et al., 2013) under the substitution model LG4X (Le et al., 2012). The trees were analyzed by eye and unreliable sequences – those corresponding to extremely long branches on the tree and those occupying obviously wrong positions based on well-accepted knowledge of the eukaryote tree (i.e., contaminants or paralogs) were removed. Finally, the sequences were concatenated into a supermatrix of 81 taxa with 62,839 aligned amino acid sites, and a phylogenomic tree was estimated by maximum likelihood using IQ-TREE with the LG+C20+F+Γ (Yang, 1994; Le & Gascuel, 2008; Le et al., 2008) substitution model, and branch support evaluated with a 1000-replicate ultrafast bootstrap approximation (Minh et al., 2013).

3.2.6 Comparative gene set analysis

The shared orthologous gene set between the the total predicted gene sets of BBO and *Blastocystis* sp. ST1 were determined using reciprocal BLASTp (Gish & States, 1993) with an evalue cutoff of $1x10^{-5}$ and query coverage >30%. This was repeated between pairs of BBO and the other subtypes (ST4 and ST7). Low stringency settings were selected here because BBO is very divergent from the other subtypes, and loose thresholds will enable more distantly-related homologs to be identified. To retrieve more potential homologs, the BBO genome was searched using tBLASTn (i.e., a six-frame translation of the BBO genome was searched), using the predicted protein sequences of other subtypes as queries. If matches were found that were absent from the BLASTp step, they were added to the shared gene sets between each pair of subtypes. A plot comparing the orthologous gene set shared by each pair of subtypes was generated by Circos (Krzywinski et al., 2009).

3.2.7 Amino acid sequence divergence

Orthologous genes (genes of common ancestral origin due to speciation) between pairs of *Blastocystis* spp. were selected using orthoparahomlist.pl (Stanke, 2011) with default parameters. The script uses reciprocal BLASTp (evalue cutoff: $1x10^{-10}$) (Gish & States, 1993) to determine the best hit and outputs the percent identity of the first high scoring pair (HSP) for each ortholog. The median sequence identity was calculated for each pair of subtypes.

3.2.8 tRNA search and codon usage

The programs ARAGORN (Laslett & Canback, 2004) and tRNAscan-SE (Lowe & Eddy, 1997) were used to predict all tRNAs from the nuclear genome. To corroborate the anticodons of the retrieved tRNAs with actual codons in the predicted proteins, codon usage for BBO, *Blastocystis* ST1, ST4-WR1, and ST7 was calculated by EMBOSS: cusp (Rice et al., 2000).

3.2.9 Polyasparagine validation

The following procedure was performed by the Roger Lab technician Marlena Dlutek.

To test that polyasparagine tracts in predicted proteins were present in mature transcripts, I sequenced a part of a highly conserved gene, VatC, that contained a 35 consecutive asparagine codons in its sixth exon. The BBO RNA was reverse transcribed into cDNA using the Omniscript RT Kit (QIAGEN, Cat no.: 205111) following the manufacturer's instructions. A negative control was set up using the same method but excluding the reverse transcriptase enzyme. PCR was performed on the cDNA with two sets of VatC primers (forward pair 1: 5'-GTTCAGATCCGAATGTTAAT-3', forward pair 2: 5'-CAAGAACAAACTTTCTTACAAAC-3'; reverse for both pairs: 5'-GGTAATACAGGTGGACAATAGTTATT-3'; pair 1 product size: 142 bp, pair 2: 179 bp) under default thermocycler conditions as recommended by Phusion Hot Start II DNA Polymerase kit (Thermo Scientific™), with the annealing temperature set at 59.9°C. The PCR products and the negative control from the reverse transcription step were visualized on a 5% agarose gel. The PCR products were then sequenced using the Sanger method, and the fragments were assembled in Sequencher® v5.4.6 (Gene Corps Corporation, 2017).

3.2.10 LGT analysis

Homologs of eight selected LGTs out of the 74 reported in *Blastocystis* sp. ST1 NandII by Eme et al. (2017) were retrieved from the predicted protein set of BBO by BLASTp (Gish & States, 1993) searches with default parameter settings, using NandII homologs as the query sequences. The LGT candidate genes in BBO were examined for the presence of introns and eukaryotic flanking genes of eukaryotic origin to confirm they were not contaminants.

For *Blastocystis* KOZE1a, *Blastocystis* TEHE3a, *Opalina* sp., *Proteromonas* sp., *Rictus lutensis*, *Cafeteria* sp., *Cantina marsupialis*, *Halocafeteria seosinensis*, only transcriptomes were available (see Supplementary Table S3 for sources of transcriptome data). The *Blastocystis* spp. and *Rictus lutensis* samples were multiplexed in their respective sequencing runs (see Supplementary Table S3). Thus their transcriptomes were cleaned using WinstonCleaner (Nenarokov, 2018) from potential cross-contamination. Next, the transcriptomes were decontaminated from bacterial sequences using Anvi'o (Eren et al., 2015). Then genes were predicted from 6-frame translations of their transcriptomes using TransDecoder v5.0.2 (Haas, 2017). For *Incisomonas marina*, protein sequences predicted by Derelle et al. (2016) were used. For the ASCT1C gene phylogeny, the sequence for *Cantina marsupialis* was taken from Noguchi et al. (2015) as theirs was better curated (i.e., already confirmed via phylogeny).

Homologs of the selected *Blastocystis* sp. ST1 NandII LGTs were then identified in the predicted protein sets using BLASTp (Gish & States, 1993). Each LGT gene set was aligned using MAFFT (Katoh et al., 2017) and then trimmed by trimAl (Capella-Gutierrez et al., 2009) (both using default settings). Single-gene phylogenetic trees were

estimated for each LGT based on maximum likelihood using IQ-TREE with the substitution model LG4X (Le et al., 2012). This was to confirm that LGT candidates clustered within stramenopiles and not with unrelated bacterial or archaeal sequences. If the latter occurred the candidates were either likely to be paralogs (genes related due to duplication within a genome) or bacterial/archaeal contamination, and were removed from the alignment. For *Blastocystis* sp. ST2, ST3 strain ZGR, ST6, ST8, and ST9, homologs were identifed via reciprocal BLASTx (Gish & States, 1993) from their draft genome assemblies available on NCBI (NCBI genome ID: 13540) as their predicted proteins are not yet available. Because the *Blastocystis* sp. ST2 genome was missing the ASCT1C gene, another form, ASCT1B, was searched. Using the ASCT1B protein sequence from *Stygiella incarcerata* as the query, a tBLASTn search of the *Blastocystis* sp. ST2 draft genome revealed a good match of 56% sequence identity covering 94% of the query length (evalue: 9×10^{51}).

3.3 RESULTS

3.3.1 Contamination of sequencing data with prokaryotic DNA

A large proportion of the original sequencing data did not correspond to eukaryotic genomic data. There was heavy contamination of bacterial and archaeal DNA in nucleic acid preparations even after careful separation: 82.5% of Oxford Nanopore reads (mapped using minimap2 (Li, 2017)), 59.0% of Illumina RNA reads, and 88.0% of Illumina DNA reads (the latter two sets of reads were aligned to the assembly using PLAST (Nguyen & Lavenier, 2009) with an evalue cutoff of 10). The prokaryotic sequencing data was carefully removed prior to further analysis.

3.3.2 Genomic data revealed multiple eukaryotes in the BBO culture

The first sequencing dataset was obtained from short-read Illumina sequencing technology. This data was assembled and was visualized using Anvi'o (Eren et al., 2015) (see Supplementary Figure S7). After contaminating bacterial contigs were removed, a total of 12,138 eukaryotic contigs remained. The latter grouped into three different categories or "bins": a *Blastocystis* bin with low GC content (21% GC, total size of 13.9 Mb), a *Blastocystis* bin with high GC content (59% GC, total size of 21.5 Mb), and unexpectedly, a bin with sequences that were best matches for parabasalid homologs in the nr database. The parabasalid bin was likely a consequence of an unknown parabasalid growing in the cultures. An SSU rRNA phylogenetic tree was constructed using a homolog extracted from the parabasalid bin. This phylogeny showed that the parabasalid most closely related to an unidentified American cockroach gut symbiont with a highly-supported bootstrap value of 100% (Supplementary Figure S6). The closest deeper-branching clade to these two were a clade of termite gut symbionts with bootstrap support of ≥98% (e.g., *Cthylla microfasciculumque* and *Cthulu macrofasciculumque*). Given that the BBO culture was isolated from feces out the gut of the Oriental cockroach, it is reasonable to assume that this cockroach hosted a parabasalid alongside *Blastocystis*.

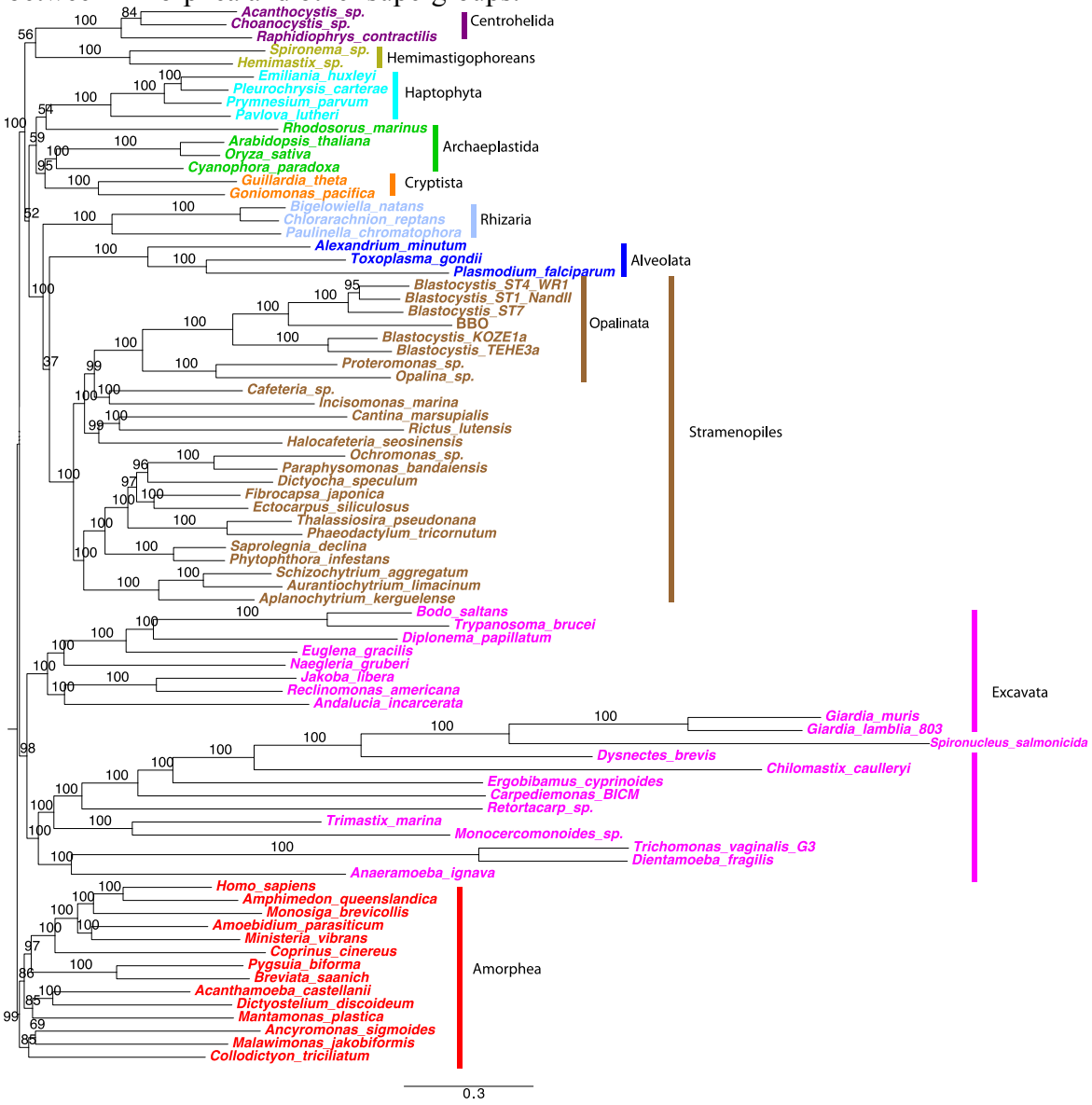
Curiously, inspection of the Nanopore reads did not reveal the presence of any parabasalid-like sequences (Supplementary Figure S7). It seems that the unidentified parabasalid was eliminated by subculturing during the six months between the time of Illumina sequencing and Nanopore sequencing. This time period corresponds to ≈12 generations of BBO growth assuming a doubling time of ≈2 weeks.

However, the Nanopore-based long-read assembly of the genome data still contained another type of contamination. Visualized using Anvi'o (Eren et al., 2015), 11 contigs with a total length of 233 kb were found to have a much higher GC content (60% GC vs. 19.9% GC) and had deeper Illumina DNA coverage (263 vs. 172 reads/base) than the rest of the genome. The deeper coverage of the higher GC content may be the result of mapping programs having difficulties in aligning reads to low-complexity regions. It is stated on the Minimap2 website (github.com/lh3/minimap2) that *"Minimap2 may produce suboptimal alignments through long low-complexity regions where seed positions may be suboptimal"*. The more extreme the GC/AT-bias, the lower the complexity becomes (i.e., sequences become less unique). These anomalous contigs were compared against the Illumina-only assembly (described in the previous section) using BLASTn (Altschul et al., 1990), and found to be identical to the "High GC Blastocystis" bin. A multi-gene phylogenomic analysis of the predicted genes corresponding to this high GC bin placed them at a distinct position as a sister group to the low GC *Blastocystis* bin with a bootstrap support of 100% (Supplementary Figure S8). Therefore I determined that the long-read sequencing had partially covered a different strain of *Blastocystis*. This also suggests that two different strains of *Blastocystis* spp. co-existed in the cell culture; one with a genome of high GC content (60% GC) and the other with a low GC content (19.9% GC). Because this high GC bin was much smaller (233 kb) than the expected genome size of any *Blastocystis* subtype (12-20 Mb on average), I concluded that the high GC-content *Blastocystis* strain was not completely covered. It was subsequently removed and the following analyses focus exclusively on the more complete, low GC-content genome.

3.3.3 Phylogenomic analysis confirms BBO's deeper-branching position

As expected, based on the mitochondrial gene phylogenetic analysis in Chapter 1, phylogenomic analysis of nuclear genes shows that the BBO lineage splits prior to the clade of *Blastocystis* sp. ST1, ST4, and ST7 isolates (Figure 3.2). The *Blastocystis* KOZE1a and TEHE3a subtypes are even more deeper-branching than BBO. *Proteromonas* sp. and *Opalina* sp. form a group that is sister to the *Blastocystis* spp. clade; collectively all of these lineages make up the Opalinata clade. The closest free-living relatives of this clade are two marine bacterivores *Cafeteria* sp., isolated from pelagic water column (Caron, 2000), and *Incisomonas marina*, isolated from sandy littoral (Cavalier-Smith & Scoble, 2013). All of these branches are well-supported by bootstrap values of 99-100%. The internal branches of the other labelled groups are also well-supported (bootstrap values ≥95%). However, the divisions between centrohelids, hemimastigophoreans, archaeplastida, and cryptista do not receive strong support in agreement with other phylogenomic analyses (e.g., Brown et al., 2018).

Figure 3.2: Maximum-likelihood phylogenomic tree of major protist groups. The substitution model LG+C20+F+Γ was used from a concatenated matrix of 62,839 amino acid sites. Bootstrap values are shown above each branch. The tree is arbitrarily rooted between Amorphea and other supergroups.



3.3.4 General statistics of the BBO nuclear genome and its predicted genes

BBO displays similar genomic features to other subtypes such as in total genome size, the number of predicted protein-coding genes, gene size, and number of introns (Table 3.1). The most striking differences are that BBO exhibits a very low GC content,

larger intron sizes, and higher intron size variation (the mode length, 49 bp, is found in only 11% of all intron sizes, vs. the mode length representing 21-54% of introns in other subtypes). The average depth of long-read coverage was 26 reads per base, with the breadth of coverage being 96.4% of the genome with above 10× coverage.

Table 3.1. Genome statistics of BBO, *Blastocystis* sp. ST1 NandII, ST4-WR1, and ST7. Data for the latter three subtypes obtained from Gentekaki et al. (2017).

Genomic features	**BBO**	**ST1**	**ST4**	**ST7**
Genome assembly size (Mb)	17.1	16.5	12.9	18.8
Scaffolds	166	580	1,301	54
GC content (%)	19.9	54.6	39.6	45.2
Number of protein-coding genes	6,732	6,544	5,713	6,020
Average gene size (bp)	1,915	1,760	1,386	1,296
Genes/kb	0.39	0.39	0.44	0.32
Genes with introns (%)	87.9	94.6	92.7	84.6
Average exon per gene	5.75	6.45	5.06	4.58
Mean length of introns (bp)	109*	50	33	50
Mode length of introns (bp)	49 (11%)**	30 (54%)	30 (36%)	30 (21%)
Number of introns	30,786	35,412	24,093	18,200
Average length of proteins	435	499	416	359
Number of forward genes	3,394	3,261	2,815	3,005
Number of reverse genes	3,338	3,283	2,898	3,015

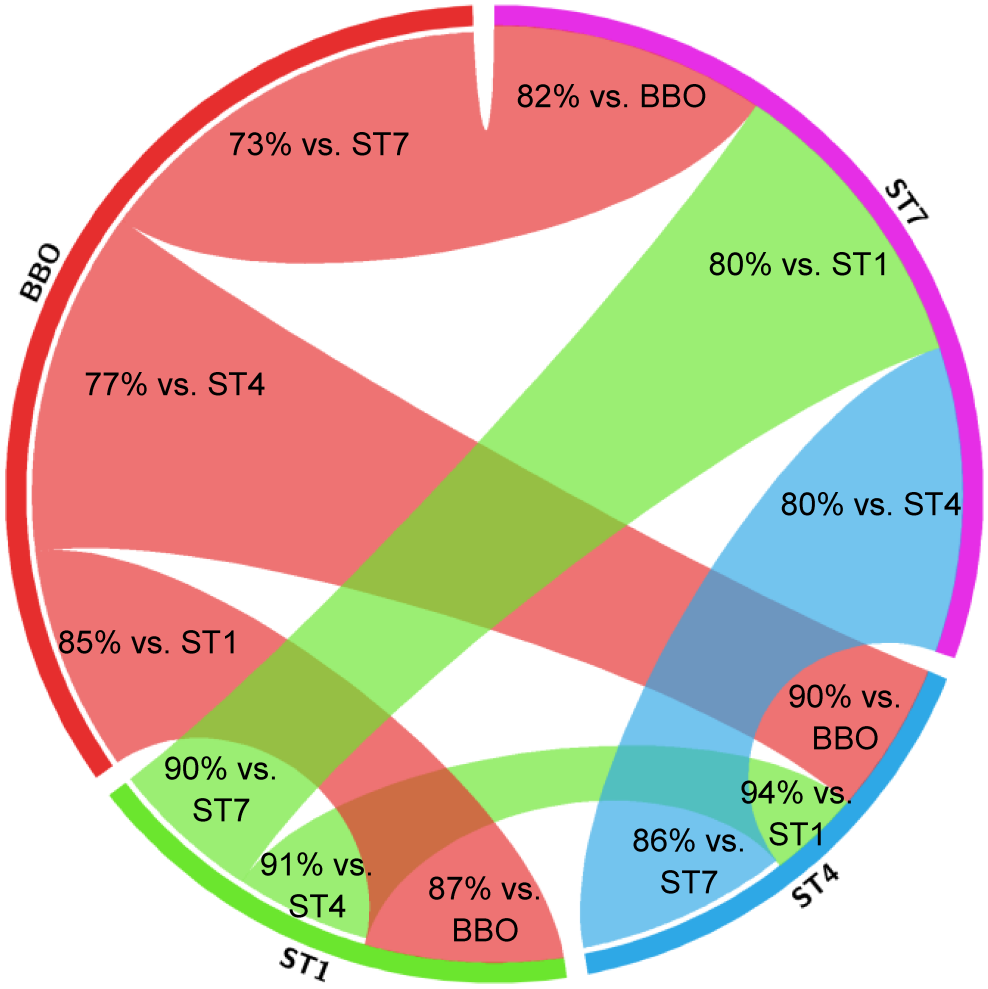
*See Supplementary Figure S9 for the intron length distribution of all intron-containing genes in the BBO genome.
**The percent of all introns that are of this length is shown in parentheses.

The amino acid sequence divergence between orthologous gene sets of *Blastocystis* sp. ST1 NandII, ST4-WR1, ST7, and BBO are presented in Table 3.2. The median sequence identities between BBO and other subtypes ranges from 46-48% (vs. 59-61% among ST1, ST4, and ST7). The lower identities were expected; the mammalian subtypes are more closely related to each other than to BBO. Between BBO and the three other subtypes, the proportion of genes unique to BBO was 15-27%, generally higher than among the other three subtypes comparisons (9-20%) (Figure 3.3).

Table 3.2. Median sequence identity of orthologous proteins from pairs of *Blastocystis* spp.

Pair	Median sequence identity (%)
BBO-ST1	46
BBO-ST4	47
BBO-ST7	48

Figure 3.3: Gene set comparisons between BBO, *Blastocystis* sp. ST1, ST4-WR1, and ST7 (obtained using BLASTp with evalue cutoff: $1x10^{-5}$, query coverage >30%). Percentages represent shared genes out of the whole set of predicted genes in each subtype. For example, 73% of BBO genes are shared with the gene set of ST7 (i.e., 27% of BBO genes are unique compared to ST7). Note that the analysis between the latter three subtypes were done using more stringent BLAST settings (evalue cutoff:$1x10^{-30}$, query cover >50% (Gentekaki et al., 2017)).



3.3.5 Polyadenylation

No instances of the polyadenylation-mediated stop codon generation phenomenon that was described in ST1 and ST7 (Klimes et al., 2014; Gentekaki et al., 2017) were found in the BBO genome. If this phenomenon were occurring, ORF predictions based on canonical genome-encoded codons would not encounter the appropriate stop codons, resulting in aberrantly long, chimeric, and/or overlapping genes. The latter cases are

usually flagged during gene prediction for genomes but in the case of BBO none were observed. I searched the BBO predicted gene set for a subset of homologs of genes whose stop codons were generated via polyadenylation in *Blastocystis* ST1 NandII. Using IGV (Thorvaldsdóttir, et al., 2013), I visualized the genes with the RNA reads mapped onto them. The reads aligned at normal stop codon positions and did not have the telltale stretch of adenines at the 3' end of the ORF. I also searched for the highly-conserved motif (TGTTTGTT) that, in the genome sequences of ST1 and ST7 (Gentekaki et al., 2017), is usually found 5 bases downstream of where the poly-A tail is added to transcripts (see Supplementary Figure S10). The motif itself was found individually in ≈2000 locations on the BBO genome, but when a subset of them were manually inspected, no poly-A-tailed transcripts aligned near them.

3.3.6 Enzymes of evolutionary significance

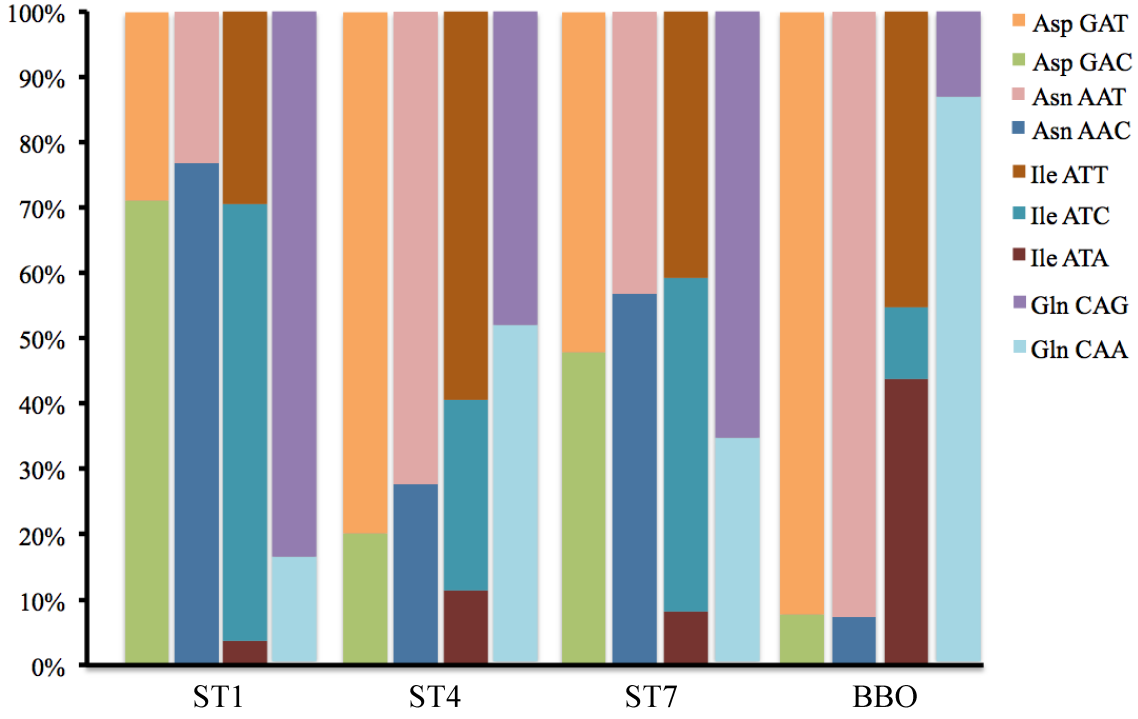
Genes encoding the two key enzymes involved in anaerobic ATP production, PFO and [FeFe] hydrogenase, were present among the predicted genes of BBO. The two genes of possible cyanobacterial origin described by Denoeud et al. (2011), phosphoglycerate kinase and 6-phosphogluconate dehydrogenase, were also present in the BBO predicted gene set, and their single gene trees confirmed that the BBO homologs branched within the stramenopile homologs (Supplementary Figure S11). PFO, [FeFe] hydrogenase, and phosphoglycerate kinase were predicted to possess mitochondrial-targeting peptides (TargetP prediction probability of 0.878, 0.899, and 0.726 respectively), suggesting these enzymes likely function within the MROs of BBO.

3.3.7 tRNAs and codon usage

tRNAs corresponding to most of the amino acids, including selenocysteine, were retrieved from the BBO genome. Only tRNA-Cys could not be found, but it is likely that its sequence is too divergent to be detected by the tools used here, or that its sequence was not assembled into one contig and is split over an end of a contig and the start of another. This latter problem can apply to any type of gene; better scaffolding would solve this problem, either by the use of better assembly algorithms or by obtaining deeper coverage of the genome by re-sequencing.

The genetic code of the BBO genome is canonical (see Supplementary Table S4 for all codon usage of the four *Blastocystis* spp.) with the codon usage being biased towards AT-rich codons. For example, alanine codons can be GCA, GCT, GCC, or GCG. Yet 57% and 35% of all alanine codons in the BBO genome are encoded by GCA and GCT respectively, with less than 5% composed of the GCC or GCG codons. The other *Blastocystis* subtypes have a more even distribution of alanine codon usage. The codon usage for asparagine is particularly interesting (Figure 3.4). Although both asparagine codons, AAT and AAC, are represented in the genomes, all tRNA-Asn from the studied *Blastocystis* subtypes present the anticodon GTT, which canonically binds to AAC.

Figure 3.4: Codon usage of asparagine (Asn), glutamine (Gln), isoleucine (Ile), and aspartic acid (Asp) in *Blastocystis* sp. ST1, ST4-WR1, ST7, and BBO.



Several other interesting codon usage patterns – namely those encoding glutamine, isoleucine and aspartic acid – are shown in Figure 3.4. Along with asparagine, these are involved in unusual homopolymer tracts found in BBO genes. Homopolymer tracts are long stretches of the same monomer repeated ≥6 times in a row (see the next section for a detailed description of amino acid homopolymers in the BBO predicted gene set). The highly AT-biased codon usage for these amino acids (AAT for asparagine, CAA for glutamine, and GAT for aspartic acid) demonstrates the peculiarity and divergent nature of BBO compared to other subtypes.

3.3.8 Amino acid homopolymers

A total of 2,640 of the predicted genes in BBO had long stretches (over 6 in a row) of trinucleotide repeats of the codon AAT (coding for asparagine) inserted in

various positions in the corresponding predicted protein. Figure 3.5 shows the number of genes containing polyasparagine tracts distributed over various lengths. In addition, there were stretches of homopolymers of isoleucine codons (repeats of ATA), and shorter stretches of glutamine (CAA) and aspartic acid (GAT) codons. Some of these homopolymer tracts occur in the same genes (Table 3.3). The codons making up these repeats (except ATA encoding isoleucine) are also over-represented in the BBO genome (Figure 3.4). Isoleucine codons are biased in favour of both ATA and ATT to the exclusion of ATC, but curiously, there are many fewer ATT homopolyer tracts detected (10 out of a total of 185 polyisoleucine tracts).

Figure 3.5: Polyasparagine length distribution in BBO predicted genes. Each length category represents a length variation of 18-mers.

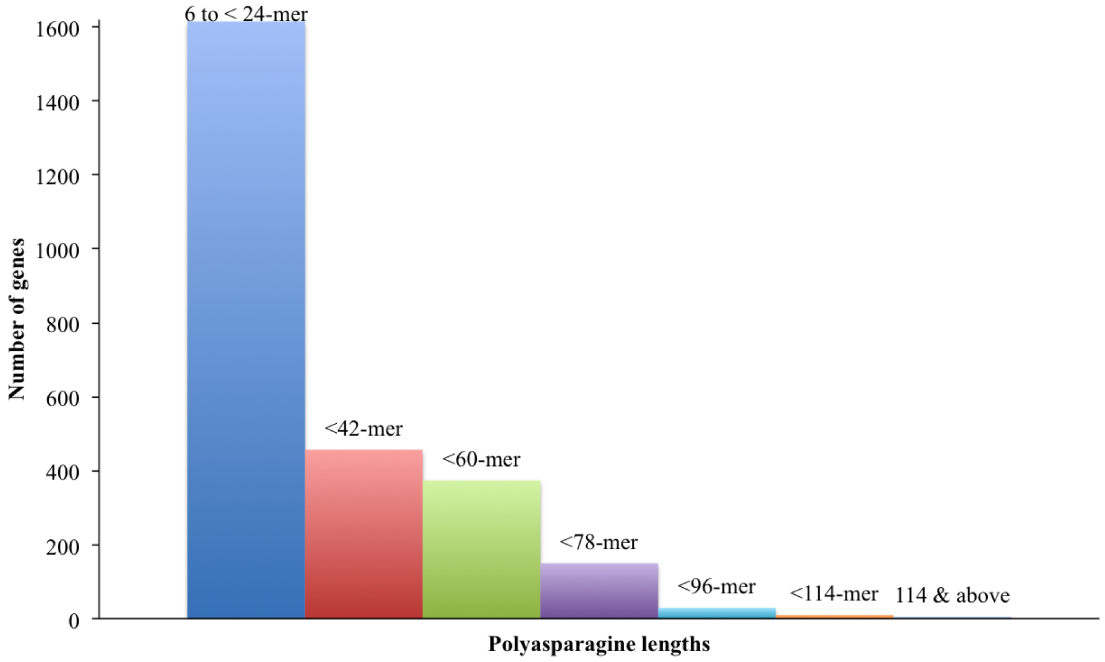


Table 3.3. Occurrences of homopolymer amino acids (≥ 6-mer) in BBO. The fourth column shows the number of genes containing a combination of a pair of homopolymers, e.g., 3 genes have polyaspartic acid and polyglutamine.

Amino Acid	Maximum length of homopolymer	Number of genes	Number of genes that contain two kinds of homopolymers			
			D	Q	I	N
Aspartic acid (D)	17	177	-	3	5	103
Glutamine (Q)	30	40	-	-	0	20
Isoleucine (I)	60	185	-	-	-	169
Asparagine (N)	129	2,640	-	-	-	-

By manual inspection of mapped RNA-seq data, I could not identify clear splice junctions near the homopolymeric regions suggesting that they are unlikely to be introns that were mispredicted as exons. I further confirmed that they were present in mature mRNAs by performing PCR experiments using cDNA as a template. I targeted a fragment that encoded a continuous stretch of 35 asparagine residues in the well-conserved V-type proton ATPase subunit C (VatC) gene. The resulting consensus sequence of the PCR product was 95.3% identical to the VatC gene over 128 bp, encompassing the polyasparagine tract (Supplementary Figure S12). Cloning the PCR product using plasmid vectors may have produced the ideal 100% identity. Nevertheless, the PCR products had unique flanking sequences that aligned to the regions surrounding the polyasparagine tract (albeit being of low-quality; they were trimmed after alignment). Therefore I concluded that this result was sufficient to validate the existence of the polyasparagine tract in VatC mature transcripts of BBO.

Using SWISS-MODEL (swissmodel.expasy.org) (Waterhouse et al., 2018) I predicted a 3D model of the BBO VatC protein. In this model, the polyasparagine stretch tract appeared to form a surface loop compared to the structurally aligned homolog from the yeast *Saccharomyces cerevisiae* S288C (Supplementary Figure S13). In other

eukaryotes, such as in *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, humans, and *Blastocystis* ST1, the polyasparagine tract is absent (UniProt IDs: Q9SDS7, P31412, P21283, and A0A196S6Q1 respectively; see Supplementary Figure S14 for the protein alignment of these homologs).

I also investigated another example of a predicted BBO protein with a polyasparagine tract. This protein, an ortholog of arginine transporter Can1, has a 10-mer asparagine tract near the C-terminus of the predicted protein. It was evenly covered by RNA-seq reads and the only part of its amino acid sequence that did not align to homologs in a BLASTp (Gish & States, 1993) search of the nr database was the polyasparagine tract (Supplementary Figure S15).

All genes containing 6-mer or longer polyasparagine coding regions were automatically annotated using InterProScan (Jones et al., 2014). The annotations suggest the gene functions range from repetitive structural motifs such as WD40 repeats and zinc fingers to highly conserved eukaryotic proteins including heatshock proteins and ATPases (Supplementary Table S5).

Amino acid homopolymers were also found in proteins of the other subtypes (ST1, ST4-WR1, ST7), but the number of genes containing them were low (<400 per subtype) and the homopolymers were shorter than 20-mers. Examples of such repeats are: 1) a 15-mer of glutamine near the C-terminus of a nuclear mitotic apparatus protein from ST1 (GenBank ID: OAO12395.1), in which the homopolymer region itself is neither inside any domains nor aligns to other homologs in the nr database, and 2) a 15-mer of glutamic acid inside the C-terminal domain of RPA1 subunit of eukaryotic RNA polymerase I (RNAP I) from ST7 (NCBI reference sequence: XP_012895430.1), in

which only the homopolymer region remained unaligned to any homologs in the nr database.

3.3.9 The presence of *Blastocystis* sp. ST1 LGTs in BBO and other related stramenopiles

Eight out of the 74 LGTs previously reported in *Blastocystis* sp. ST1 NandII by Eme et al. (2017) were analysed in this chapter. They were selected because of their key roles in anaerobic ATP production (RquA and ASCT1C), immune evasion (TXNDC12, glycosyltransferase GTd1, and sialidase), carbohydrate scavenging (fucose dehyrogenase), iron-sulfur cluster production (SufCB), and in eliciting host immune response (O-methyltransferase). Table 3.4 describes their specific functions and origins.

Table 3.4. Selected LGTs, their associated functions and ancestral origin.

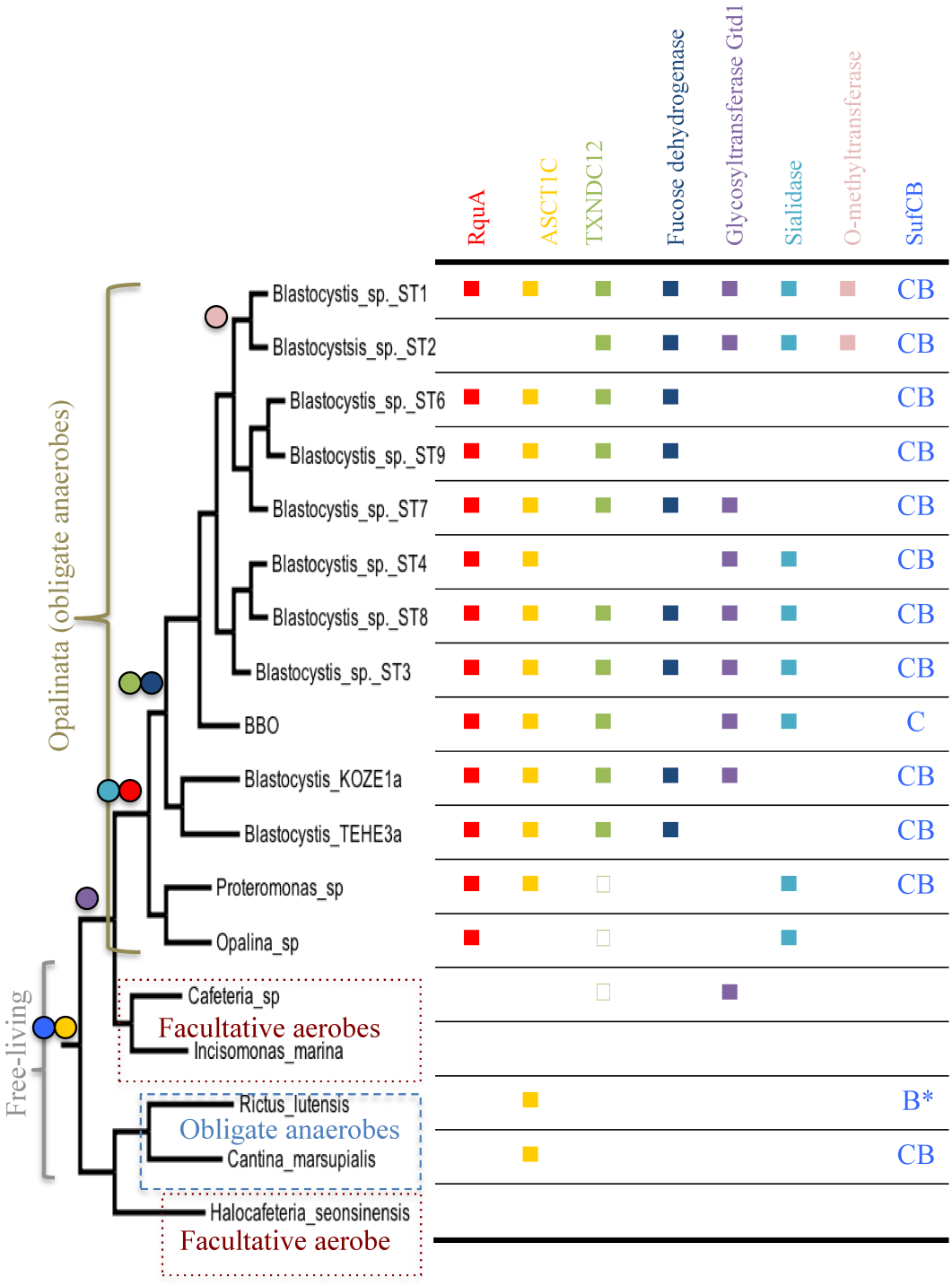
Gene name	Function	Origin
RquA	Biosynthesis of rhodoquinone (Stairs et al., 2014), a compound that passes electrons from Complex I to Complex II (in *Blastocystis*, this is fumarate dehydrogenase that runs in reverse to the typical succinate dehydrogenase of aerobic eukaryotes) (Stechmann et al., 2008). Involved in anaerobic ATP production.	Bacteria or eukaryote
Acetate: succinyl-CoA transferase subtype 1C (ASCT1C)	Converts Acetyl-CoA into succinyl-CoA, which then participates in substrate-level phosphorylation of ADP to ATP (Gentekaki et al., 2017). Involved in anaerobic ATP production.	Bacteria or eukaryote
Thioredoxin-domain-containing protein 12 (TXNDC12)	Redox regulation or defense against oxidative stress (Matsuo et al, 2001), such as during host immune response to *Blastocystis* infection (Chandramathi et al., 2009).	Metazoa
Fucose dehydrogenase	Releases fucose monomers from mucin glycans produced by the intestinal epithelium for energy (Hooper & Gordon, 2001).	Bacteria

Gene name	Function	Origin
Glycosyltransferase Gtd1	Scavenges carbohydrates to express them on cell surface to mimic antigens for immune evasion as done by *Helicobacter pylori*, a bacterial gut pathogen (Moran, 2008; Eme et al., 2017).	Bacteria
Sialidase	Scavenges for sialic acid. The latter is then expressed on the *Blastocystis* cell surface (Lanuza et al., 1996), possibly for immune evasion (Eme et al., 2017).	Bacteria
O-methyltransferase	Biosynthesis of polyketides, which induces intestinal inflammation. Also present in the *Trichomonas vaginalis* genome (Denoeud et al., 2011; Poirier et al., 2012; Eme et al., 2017). It may also be involved in producing an antibiotic, as a closely related *Streptomyces* homolog tcmP participates in the biosynthesis of the antibiotic tetracenomycin C (Decker et al., 1993; Eme et al., 2017).	Bacteria
SufCB	A sulfur mobilization system (SUF) in the cytosol that synthesizes iron-sulfur (Fe-S) clusters, replacing (or in *Blastocystis* spp., co-existing) with the iron-sulfur-cluster assembly machinery located in the mitochondria in other typical aerobic eukaryotes (Tsaousis et al., 2012). Also found in *Pygsuia biforma*, a deep-branching anaerobic protist (Stairs et al., 2014), and in *Stygiella incarcerata*, a microaerophilic Jakobid (Leger et al., 2016). This version is a fusion of SufC, a member of the ATP-binding cassette transporter family, though likely no longer part of a transporter (TIGR01978), and SufB, a sulfur-acceptor (Hirabayashi et al., 2015).	Archaea

The presence/absence of LGTs of the eight selected genes was mapped on the phylogeny of opalinitans and closely related stramenopiles (Figure 3.5). RquA was not found in the transcriptome of *Proteromonas* sp., but was identified by Stairs et al. (2018) in the genome (the full genome is not yet published but is available in NCBI). Note that for *Blastocystis* sp. KOZE1a, *Blastocystis* sp. TEHE3a, *Opalina* sp., *Proteromonas* sp.,

Cafeteria sp., *Halocafeteria seosinensis*, *Cantina marsupialis*, and *Rictus lutensis*, the 'absence' of some genes may be because they were not highly expressed and therefore may not have been covered by RNA-seq. When their genomes become available, gene prediction may reveal the presence of some of these 'absent' genes. In *lieu* of that data, a tentative recontruction of the origins of the genes was attempted (Figure 3.6). An in-depth discussion of this pattern as it relates to each gene is presented in section 3.4.5. Phylogenetic trees for each LGT (except for TXNDC12; see Figure 3.8) are represented in Supplementary Figure S16.

Figure 3.6: Presence of LGTs in *Blastocystis* spp. and in closely related species. The phylogenetic tree on the left is based on Figure 3.2. Branch lengths not to scale. Genes detected in *Blastocystis* ST1 but missing from other subtypes/species are represented. Squares indicate presence of gene. Stars indicate presence of gene from a different donor. CB: SufC and B found fused, B: only SufB found, C: only SufC found. Data for *Blastocystis* ST1-9 for all genes except SufCB obtained from Eme et al. (2017). Coloured dots (corresponding to the font colours of the genes names) on the tree indicate the first point where the LGT may have been gained. Incomplete sequences are indicated by asterisks.


RquA
ASCT1C
TXNDC12
Fucose dehydrogenase
Glycosyltransferase Gtd1
Sialidase
O-methyltransferase
SufCB
Opalinata (obligate anaerobes)
Free-living
Blastocystis_sp._ST1
Blastocystsis_sp._ST2
Blastocystis_sp._ST6
Blastocystis_sp._ST9
Blastocystis_sp._ST7
Blastocystis_sp._ST4
Blastocystis_sp._ST8
Blastocystis_sp._ST3
BBO
Blastocystis_KOZE1a
Blastocystis_TEHE3a
Proteromonas_sp
Opalina_sp
Cafeteria_sp
Facultative aerobes
Incisomonas_marina
Rictus_lutensis
Obligate anaerobes
Cantina_marsupialis
Halocafeteria_seonsinensis
Facultative aerobe
CB
CB
CB
CB
CB
CB
CB
C
CB
CB
CB
B*
CB

3.4 DISCUSSION

3.4.1 Limitations of long-read sequencing and bioinformatics tools

Ideally, the combination of short and long-read sequencing datasets should produce large chromosome-level contigs. However, the number of scaffolds in BBO was at least 13 times the number of chromosomes found in human *Blastocystis* isolates (Carbajal et al., 1997). To obtain longer reads that can assemble into bigger contigs, BBO DNA could have been fragmented to a larger size, or this fragmentation step could have been skipped entirely during library preparation for Nanopore sequencing. In this project, because DNA extractions had a very low yield (<100 ng for every 40 ml of culture), the input DNA had to be sheared to a rather small fragment length (≈8 kb on average) to allow the flow cell to work efficiently; the activity of the pores in the flow cell is proportional to the number of DNA fragments ligated with adapter protein (Judd, 2017). Hence the lower the amount of starting DNA, the more it needs to be fragmented. Because BBO was cultivated non-axenically, an extra amount of starting DNA had to be collected (at least 5μg) to compensate for the high proportion of bacterial DNA. Long-read sequencing has a long way to go in terms of minimizing input requirements compared to Illumina sequencing, which can be as low as 50 ng input DNA per library (Illumina Inc., 2018).

Another concern was the errors in long read data that might lead to protein misprediction. As most errors are insertions or deletions (indels), accurate gene prediction can be seriously compromised if these errors are not corrected. Watson (2018) found that 33% of protein-coding genes from Oxford Nanopore data in a human chromosome were mispredicted even after signal-level polishing and short read

correction. Therefore thorough manual inspection and correction using short read data mapped onto the draft genome is needed to produce an accurate genome. For the BBO genome, out of the 166 contigs, only those containing genes for this project's analyses were manually inspected (e.g., those containing genes with homopolymer stretches, LGT candidates, and the genes used for phylogenomic analysis). Although a few indels were observed (< 5) they were located in noncoding regions where corrections would not result in frameshifts. For more fine-scale analyses in the future I suggest that every gene/region of interest be scrutinized by eye.

During the third step of gene prediction (section 3.2.3), assembled transcripts were found to have highly uneven coverage over genes with long amino acid homopolymers. For example, the transcripts that mapped to the VatC gene were split into two parts: those mapping upstream versus those mapping downstream of the polyasparagine tract. No transcripts spanned across the repeated sequence, making validation of the gene prediction difficult. Closer inspection revealed that the transcriptome assembler (Trinity (Grabherr et al., 2011)) had discarded RNA reads that contained repeats and failed to assemble them, a problem that was previously observed by Lima et al. (2017). Lima et al. (2017) developed a method to flag transcripts with repeated sequences, which can then be collected and reassembled. Therefore care must be taken when using bioinformatics tools to analyse repetitive sequences.

3.4.2 Four key enzymes conserved in BBO

The presence of PFO and [FeFe] hydrogenase in the predicted gene set of BBO, both found with mitochondrial-targeting peptides, implies that the anaerobic ATP production pathway in *Blastocystis* sp. ST7 described by Stechmann et al. (2008) and

Denoeud et al. (2011) is conserved in BBO and likely functions in its MROs. This pathway, in brief, converts pyruvate into acetyl-coenzyme A (acetyl-CoA) through the activity of PFO, reducing ferredoxin in the process. [FeFe] hydrogenase reduces H+ ions to hydrogen gas while oxidizing ferredoxin. Acetyl-CoA is converted into acetate by acetyl:succinate CoA tranferase (ASCT). One of the forms of ASCT is present in BBO and is likely targeted to the MRO (TargetP probability: 0.874; this gene was included in the LGT analysis. See Figure 3.6 and discussion in section 3.4.5). ASCT catalyzes the transfer of the CoA moeity to succinate, generating succinyl-CoA. This succinyl-CoA is converted back to succinate by succinyl-CoA synthetase, phosphorylating ADP into ATP in the process (see Supplementary Figure S17). Succinyl-CoA synthetase is made up of an alpha and a beta subunit. Both homologs were found in the BBO predicted gene set. The alpha subunit had a mitochondrial-targeting peptide (MitoProt probability of 0.939, TargetP probability of 0.490) but the predictions of a targeting peptide on the beta subunit were inconclusive (MitoProt probability of 0.222, TargetP probability of 0.374).

The presence of two genes of cyanobacterial origin, phosphoglycerate kinase and 6-phosphogluconate dehydrogenase, tentatively supports the hypothesis suggested by Denoeud et al. (2011) that *Blastocystis* spp. may have lost a secondary plastid, and that some of the plastid genes have moved into the nuclear genome. The phosphoglycerate kinase protein sequence was found to have a mitochondrial-targeting peptide, suggesting that this enzyme may have been co-opted to fulfill its role in glycolysis in an ancestral stramenopile after the plastid was lost. Indeed, Nakayama et al. (2012) found evidence that suggested that in stramenopiles, including *Blastocystis* sp. ST7, part of the glycolytic pathway occurs in mitochondria. An alternative interpretation for the evolutionary origins

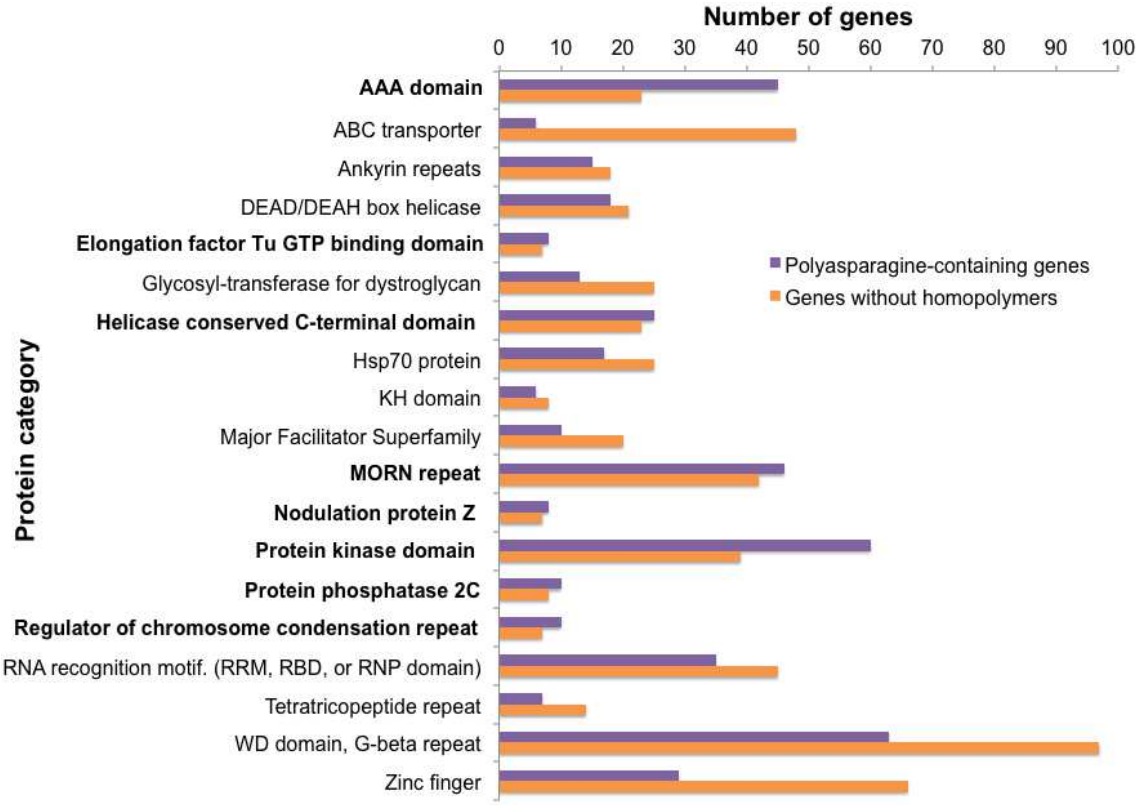
of these two genes is that they are LGTs in eukaryotes from cyanobacteria independent of plastid origins. Phylogenetic analyses by Andersson & Roger (2002) could not resolve a single point of acquisition among all the eukaryotes that possess these genes. One approach to confirm the origin of these genes in *Blastocystis* subtypes is to conduct a focused search for other potentially plastid-derived genes and analyse their phylogeny to see if multiple independent genes in these organisms have a similar cyanobacterial/plastid signal. The latter case would support the origin of these genes from a plastid-containing ancestor (either by primary or secondary endosymbiosis).

3.4.3 Amino acid homopolymers and codon usage

About 40% of the predicted genes contained polyasparagine runs (6-mers or more) coded by the trinucleotide repeat AAT. A similar phenomenon was reported in the genome of the slime mold amoeba, *Dictyostelium discoideum*. It has a low GC content (GC 22.4%) and 34% of its predicted proteins contain polyasparagine (encoded by AAT) or polyglutamine (encoded by CAA) tracts of ≥20-mers (Eichinger et al., 2005; Scala et al., 2012). Similarly, in the genome of the malaria parasite, *Plasmodium falciparum* (GC 19.4%), 35% of the encoded proteins have polyasparagine tracts (6-mers or more, coded by AAT) (Singh et al., 2004). In *P. falciparum*, the tracts are observed in proteins of all functional classes except for molecular chaperones and ribosomal proteins (Singh et al., 2004). In BBO, the polyasparagine-containing genes include chaperones and ribosomal proteins, in addition to highly-conserved genes such as those encoding MutS (proteins that repair mismatches during DNA replication) and ATPases (Supplementary Table S5). A comparison of related genes with polyasparagine tracts and ones without homopolymers showed that homopolymer-containing genes are overrepresented in the

following proteins/structural motifs: AAA domain, elongation factor Tu GTP binding domain, helicase conserved C-terminal domain, MORN repeat, nodulation protein Z, protein kinase domain, protein phosphatase 2C, and regulator of chromosome condensation repeat (Figure 3.7). These eight protein categories have diverse functions and it appears that there is no common feature among them (i.e., no identity in sequence, secondary or tertiary structure).

Figure 3.7: InterProScan annotation (no manual curation) of related protein families with polyasparagine tracts and ones without any homopolymers in BBO. Only groups of proteins and structural motifs with frequency > 5 are shown. Proteins in bold are overrepresented in the polyasparagine-containing genes.



The overall low GC content of the BBO genome may have restricted codon variation. This may have led to a higher chance of consecutive codons being identical, thus producing trinucleotide repeats. Another phenomenon surrounding codon usage is the isoleucine involved in polyisoleucine tracts. Most of these tracts were encoded by

ATA, which makes up about 45% of all isoleucine codons overall in the BBO genome (Figure 3.4). However, another 45% of the overall isoleucine codons in the BBO genome is composed of the ATT codon. This raises the question of why polyisoleucine tracts favour the ATA codon over the ATT codon. I observed that many of the polyisoleucine tracts occur at the ends of longer polyasparagine tracts. This may be a consequence of uncorrected Nanopore sequences causing a frameshift in the protein translation. For example, a deletion of adenine is indicated by the vertical line in this sequence: AATAATAATAAT|ATAATA. This sequence would be translated to NNNNII, but the corrected sequence would translate to NNNNNN. To ensure that these polyisoleucine tracts are genuine, deeper-coverage Nanopore sequencing data must be obtained from BBO.

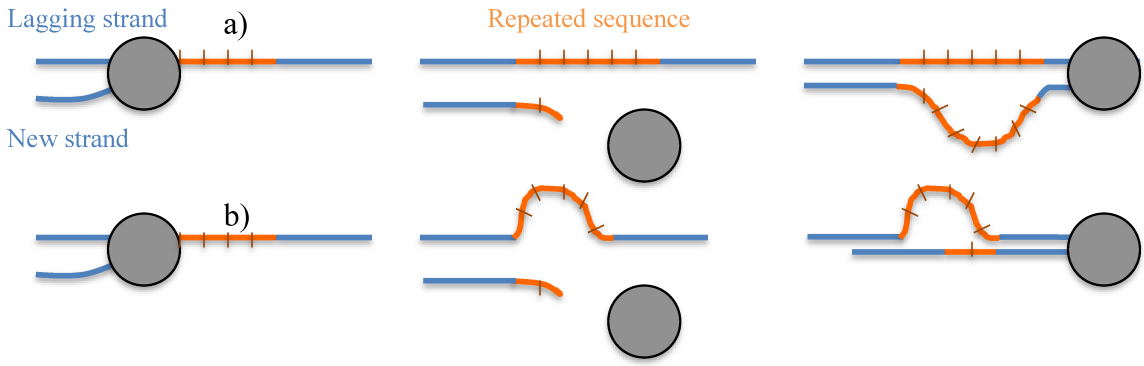
Proteins with asparagine-rich repeats tend to form amyloids and prions (Halfmann et al., 2011) which have been associated with neurodegenerative diseases (Chiti & Dobson, 2006). However, protein aggregation is not always deleterious; in yeast, it mediates inheritance of several phenotypes, and in mammals, it maintains synaptic facilitation and activates antiviral immunity (Patino, 1996; Si et al., 2010; Hou et al., 2011). In *D. discoideum*, under normal conditions, most of its homopolymer-containing proteins do not aggregate. The few that do misfold are compartmentalized to the nucleus and degraded by a ubiquitin-proteasome system (Malinovska et al., 2015). Under heat stress, more proteins aggregate, but Hsp100 disaggregase reverses aggregation (Malinovska et al., 2015). As such, I recommend future studies of molecular chaperones and homopolymer protein localization in BBO cells.

In the BBO VatC gene example discussed earlier in this chapter, its 35-mer polyasparagine tract appears to form a surface loop, which likely explains why the protein can continue to fold and function. A study of a deubiquitinating enzyme of *P. falciparum* also revealed that a polyasparagine tract (26-mers) in the middle of the protein sequence did not interfere with its catalytic activity (Artavanis-Tsakonas et al., 2006). Its predicted 3D model showed that the polyasparagine tract was located away from the active site (Artavanis-Tsakonas et al., 2006).

I propose two possible alternative mechanisms of how these trinucleotide homopolymers proliferated in the BBO genome in the first place: 1) replication slippage or 2) transposable-element-like movement. During DNA replication if the DNA polymerase pauses at a repeated sequence on the lagging strand it dissociates after it partially synthesises the repeated sequence (Viguera et al., 2001). Because the lagging strand is synthesized in short, separated fragments, i.e., Okazaki fragments, the chance of pausing increases if the repeated sequence forms a secondary structure such as a hairpin. The DNA polymerase complex then attempts to reassemble at the end of the repeated sequence, but in doing so, backtracks and inserts the previously added nucleotides. The newly synthesized repeated sequences from before and after the pause become ligated during the nucleotide excision repair process. This results in an expansion of the repeated sequence in the daughter strand (Viguera et al., 2001). This mechanism is illustrated in Figure 3.8a. Dinucleotide and trinucleotide microsatellites are thought to expand in this manner (Hartl & Ruvolo, 2012). Replication slippage has been implicated in genetic diseases such as the expansion of CAG/CTG trinucleotide repeats that causes Huntington's disease (Petruska et al., 1998). Note that replication slippage can also

shrink a repeated sequence (e.g., microsatellite removals in *Drosophila* spp. (Harr et al., 2000)) if the Okazaki fragment that contains the partial repeated sequence moves forward and aligns to the end of the repeated sequence (illustrated in Figure 3.8b). In the case of BBO, *P. falciparum*, and *D. discoideum* genomes, their low GC content likely led to low variation in codon usage. This increased the chance of consecutive codons being identical, which may have led to replication slippage occurring more frequently.

Figure 3.8: Replication slippage process. Top strand represents DNA template of the lagging strand, bottom strand represents the newly synthesised DNA strand. a) DNA polymerase inserting the same repeated nucleotides after re-binding to the template causes an expansion of the repeated sequence. b) Repeated sequence in the Okazaki fragment erroneously aligns to the end of the template's repeated sequence and causes a reduction in the repeated sequence in the daughter strand. Normal DNA sequence in blue, repeated sequence in orange. Grey circle represents DNA polymerase. Ticks on repeated sequence represent one repeat unit.



A second possible explanation for the prevalence of repeated sequences in the BBO genome involves transposons. Transposable element insertions usually occur in noncoding regions of the genome and can affect gene expression (e.g., leading to kernel colour changes in maize (McClintock, 1950)). If they occur inside coding regions, they can either cause diseases, e.g., LINE1 insertion into an exon in the factor VIII gene resulting in hemophilia (Kazazian et al., 1988), or they can drive novel gene formation/function such as in retrotransposon-derived placental genes (Rawn & Cross,

2008). However, trinucleotide repeats are not a characteristic of transposable elements. Indeed, a search for 20-mer repeats of AAT in Dfam (Hubley et al., 2016), a database of repetitive DNA elements, yielded no candidates. Thus the replication slippage explanation is the more plausible mechanism for the trinucleotide repeats observed in the BBO genome.

Whatever the mechanism for their origin, the long amino acid homopolymers in BBO proteins have not been removed by purifying selection. If they are deleterious, they have been fixed in the population as a result of a small effective population size or population bottleneck in the BBO lineage (see section 2.4 for discussion on how deleterious mutations become fixed in a population). On the other hand, possible functions for homopolymer insertions include transcriptional activation or intermolecular binding (Jordan & Kajava, 2010). An *in silico* approach to test if these homopolymer tracts serve any functions is to predict their interactions with other molecules in their respective biochemical pathways. Currently, molecular dynamic simulations are computationally heavy (e.g., it takes ≈60 hours to predict allosteric sites on a T4-lysozyme using 16 cores of 2.3 GHz CPUs (Greener et al., 2017)). Therefore it would be impractical to assess all of the 2,640 genes containing polyasparagine. If this method is pursued, it should focus on a subset of polyasparagine-containing genes whose homologs are highly-conserved and well-studied among eukaryotes.

The energy consumption of a cell is proportional to the total mass of proteins it synthesizes (Li et al., 2014). Keeping all other variables of energy consumption constant (expression level of genes, total number of proteins synthesized, etc.) a BBO cell would produce a larger mass of proteins than a *Blastocystis* sp. ST1 cell due to the

homopolymer insertions. If these insertions were each slightly deleterious to protein function, it would be especially difficult to rationalize why more cellular resources were devoted to their synthesis, as this creates an additional burden and disadvantage relative to non-insertion containing cells. This again suggests an explanation for these insertions that involves an important role for random genetic drift as a result of a small effective population size. To test this hypothesis, more in-depth investigations into the population genetic history of BBO are needed. One of the methods used to assess effective population size of a species is to look for signatures of translational selection. In a large effective population, such as in *E. coli*, the highly-expressed genes have a greater codon usage bias compared to genes with low expression (Bennetzen & Hall, 1982; Ikemura, 1985). The codon usage bias is matched to the more abundant tRNA which provides a slight advantage to translational efficiency and accuracy (Bennetzen & Hall, 1982; Ikemura, 1985). In organisms which do not undergo competitive exponential growth (and therefore have a lower effective population size) such as in *Helicobacter pylori*, this signature is absent from their genomes because selection is too weak to fix substitutions that improve translational efficiency (Lafay et al., 2000). Therefore we can predict that if BBO has experienced a lowered effective population size relative to other *Blastocystis* subtypes, then signatures of translational selection will be much more pronounced in the latter relative to BBO. A comparison of translational selection in the genomes of BBO and human-infecting *Blastocystis* sp. subtypes should be conducted to address this question.

It is interesting that BBO, *P. falciparum*, and *D. discoideum* all have extremely AT-rich genomes and possess trinucleotide-encoded amino acid homopolymers in over

30% of their predicted proteins. Their strong bias for AT-rich codons may have increased the frequency of replication slippage; A:T base pairs are held together by only two hydrogen bonds (versus three in G:C pairs) and thus DNA strands enriched in them are more easily dissociated.

3.4.4 The tRNA wobble effect in amino acid homopolymer tracts

Although the polyasparagines were coded by trinucleotide repeats of AAT, the only type of tRNA-Asn sequences retrieved from the genome possessed the anticodon GTT. This codon canonically binds to another asparagine codon, AAC. Only a few asparagine residues, those not involved in homopolymer tracts, were coded by AAC (only 7.4% of all asparagine codons). This suggests that one or more tRNA-Asn in BBO binds non-canonically to the codon AAT. This may be attibuted to the wobble effect: the third codon position is degenerate and has the freedom to form non-Watson-Crick pairs such I:U, I:A (I: inosine, deaminated adenine), or G:U (this is likely the case with the one tRNA-Asn in BBO) (Crick, 1966). The tRNA wobble effect is important in many cellular functions. In the yeast *S. cerevisiae*, tRNA wobble modification (such as the wobble U converted to 5-methyl-2-thio-U) was demonstrated to affect not only codon reading, but also ribosomal translation rate and the coordination between cell homeostasis and stress adaptation (Ranjan & Rodnina, 2016). Other *Blastocystis* subtypes, ST1, ST4-WR1, and ST7, also only have tRNA-Asn-GTT while still using the codon AAT (albeit in smaller proportions, see Figure 3.4). The presence of the same phenomenon in BBO suggests that the wobble effect in this tRNA has been acquired early on in *Blastocystis* lineage. The G:U wobble pairing of tRNA-Asn in *Blastocystis* spp. may be catalyzed by tRNA-guanine transglycosylase (tgt). Tgt converts guanine to queuosine in the first anticodon

position of tRNA-Asn, tRNA-His, tRNA-Asp, and tRNA-Tyr. Unlike guanine, which canonically binds to cytosine, queuosine has a slightly higher affinity to bind to uracil (Harada & Nishimura, 1972). Tgt homologs were found in the predicted gene sets of *Blastocystis* sp. ST1 NandII, ST4-WR1, ST7, and in BBO. If this phenomenon is also detected in the genomes of closely-related stramenopiles, I suggest a further investigation into tgt expression levels in relation to codon usage bias. This may reveal an evolutionary impact of tRNA-Asn modification to ribosomal translation rates as seen in *Drosophila* (Chiari et al., 2010).

3.4.5 LGT acquisition in the opalinitans and close relatives

The presence and absence of the eight selected LGTs revealed an interesting history of the Opalinata lineage and their free-living relative *Rictus lutensis*. Starting with the most ubiquitous LGT, the ASCT1C gene was found in all of the obligate anaerobes except in *Blastocystis* sp. ST2 and in *Opalina* sp.. This makes it the "earliest" LGT acquisition among all the species studied here. ASCT1C is a key gene involved in substrate-level phosphorylation during anaerobic ATP production, thus its absence in these two anaerobic organisms is curious. ASCT1C has another form: ASCT1B, found in the genome of the microaerophilic jakobid *Stygiella incarcerata* and in some of the other *Blastocystis* spp., *Rictus lutensis*, *Cantina marsupialis*, and in *Halocafeteria seosinensis* (see Supplementary Figure S16h for phylogenetic tree). The presence of ASCT1B in the latter is intriguing, as it is a facultative aerobe. *Blastocystis* sp. ST2 also had the ASCT1B form. As suggested by Leger et al. (2016) the presence of the different forms of ASCT in divergent protists indicates convergent adaptation or differential loss from a common ancestor possessing both subtypes 1B and 1C. The latter is a more parsimonious

explanation for the absence of ASCT1C in *Blastocystis* sp. ST2; the 1C form was independently lost while the 1B remained. When the complete genome of *Blastocystis* sp. ST2 becomes available, a phylogenetic analysis of the ASCT1B gene is needed to test this hypothesis. With the available data, it appears that the common ancestor of all the species studied here acquired both ASCT1B and ASCT1C. Upon diversification, some of the lineages kept both or either one. In the case of *Opalina* sp., not all genes may have been covered by the transcriptomic data used in this analysis. If this absence represents a genuine independent loss in *Opalina* sp. however, a close look at its anaerobic ATP production pathways may reveal novel enzymes and/or mechanisms.

Another gene acquired relatively around the same time as the ASCTs is the SufCB. The fused SufCB is an archaeal-derived gene involved in iron-sulfur cluster synthesis. Only the SufC unit is present in BBO; its sequence was very divergent from the other homologs and thus had to be removed from the phylogenetic tree shown in Supplementary Figure S16g. The BBO SufB unit may have accumulated so many mutations that it was unrecognizable by BLAST or HMMscan. In contrast, *Rictus lutensis* only has the SufB unit. Its sequence branched between *Protermonas* sp. and the *Blastocystis* spp. clade (Supplementary Figure S16g). This tree topology is different from the one obtained in the phylogenomic analysis, in which *Rictus lutensis* is deeper-branching than the opalinitans (Figure 3.2). This branching pattern could mean that the SufB unit of *Rictus lutensis* was a recent independent gain or it could reflect some kind of phylogenetic artefact. The latter hypothesis is supported by the fact that the *R. lutensis* SufB transcript used for this analysis was incomplete (i.e., no start codon) and so the C domain may actually be present in the encoded gene. Further evidence supporting this

hypothesis is that the fused SufCB gene was found in *Cantina marsupialis*, an obligate anaerobe that branches together with *Rictus lutensis*. *Opalina* sp. appears to have lost SufCB independently, although again, this absence may be an artefact of incomplete transcriptome data.

The second earliest (in relative timing) LGT in the opalinitans and their close relatives is the gene encoding glycosyltransferase Gtd1. Glycosyltransferases are ubiquitous among both eukaryotes and prokaryotes (Breton et al., 2012), but the glycosyltransferase Gtd1 acquired by LGT is believed to play a role in host antigen mimicry (Moran, 2008; Eme et al., 2017). This gene may have contributed to the success of *Blastocystis* spp. in colonizing animal guts. Surprisingly, *Cafeteria* sp. retained this gene, even though it is free-living. Its closest relative, *Incisomonas marina*, which is also free-living, does not possess this gene, indicating that it was lost independently. This gene from *Cafeteria* sp. possessed a secretory pathway signal peptide (TargetP probability of 0.909), suggesting that the protein product is secreted out of the cell. I speculate that this gene was retained in the *Cafeteria* sp. lineage because it helps them evade predation by other protists/animals (e.g., dinoflagellates) by mimicking the latter's antigens.

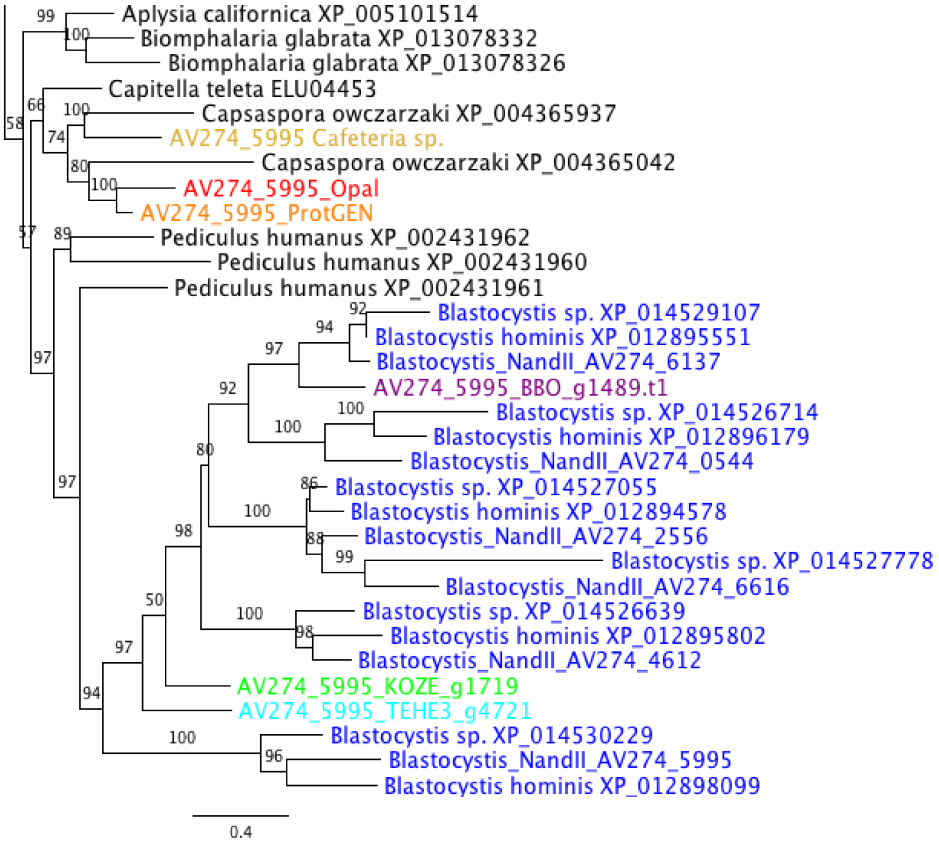
The two genes RquA and sialidase were acquired by the common ancestor of all opalinitans. RquA is present in all opalinitans except in *Blastocystis* sp. ST2. Again, its absence in *Blastocystis* sp. ST2 may be a consequence of working with an incomplete draft genome. RquA is a gene involved in the biosynthesis of rhodoquinone. Rhodoquinone has a lower redox potential than ubiquinone, which is necessary for the MRO complex II to run "backwards" for anaerobic ATP production (Stechmann et al.,

2008). Its presence in all the opalinitans implies that they share the same mechanism for anaerobic ATP production and it was an ancestral adaptation to anaerobiosis.

Sialidase was found in most of the opalinitans. *Blastocystis* sp. ST6, ST7, ST8, KOZE1a, and TEHE3a, appear to have lost it independently. Again, its absence from the latter two subtypes cannot be confirmed because they were based on transcriptome data. Sialidases are involved in immune evasion. In *Streptococcus* bacteria (GBS COH1) they cleave sialic acids present in the human respiratory mucosal lining and co-opt them to mimic neutrophil receptors. This dampens the host innate immune response (Carlin et al., 2009). *Clostridium perfringens* type D, which are strains of bacteria that cause enteritis in livestock by producing toxins, express a sialidase that modifies the host cell surface and enhances its binding to toxins, thereby prolonging toxin delivery (Li et al., 2011). Along with glycosyltransferase Gtd1, this gene may have contributed to the success of opalinitans in colonizing animals.

Two LGTs unique to the *Blastocystis* spp. clade are the genes encoding fucose dehydrogenase and the metazoan-derived TXNDC12. The latter participates in redox regulation or defense against oxidative stress (Matsuo et al, 2001). Interestingly, *Proteromonas* sp., *Opalina* sp., and *Cafeteria* sp. appear to have gained TXNDC12 from a filasterian (Figure 1.1, Chapter 1; purple line, deeper branching than animals and choanoflagellates) donor rather than from a metazoan donor (Figure 3.9). However, this branching order is not highly supported (minimum BV=57%). Also, filasterians and metazoans are closely related opisthokonts. To confirm the differences in origins of this gene a more accurate phylogeny must be estimated (e.g., by using predicted genes from genomes instead of transcriptomes).

Figure 3.9: Close-up of the single gene tree of TXNDC12. Red: *Opalina* sp., orange: *Proteromonas* sp., green: *Blastocystis* sp. KOZE1a, cyan: *Blastocystis* sp. TEHE3a, purple: BBO, blue: other *Blastocystis* spp. Maximum-likelihood tree by IQ-TREE using substitution model LG4X (Le et al., 2012) from 93 informative sites. The closest relative to *Proteromonas* sp., *Opalina* sp. and *Cafeteria* sp. is *Capsaspora owczarzaki*, a single-celled eukaryote isolated from the haemolymph of the fresh-water snail *Biomphalaria glabrata* (Stibbs et al., 1979).



The absence of fucose dehydrogenase in the BBO predicted gene set seems to be an independent loss, as *Blastocystis* sp. ST3 has also lost it. Though insects do not have mucus lining in the midgut, their microvilli are covered by a viscous glycoprotein-glycolipid matrix known as glycocalyx (Bignell, 1982; Nation, 2015, p. 41). Insect gut epithelial cells are also covered in glycan epitopes (Walski et al., 2017). Not much is known about the fucosylation levels of glycocalyx or glycan epitopes in insects. It can be

hypothesized that the absence of the gene encoding fucose dehydrogenase in the BBO genome is related to the loss of a fucose source in the cockroach gut.

As expected, all the species analyzed here are missing O-methyltransferase, a gene involved in polyketide biosynthesis; Eme et al. (2017) inferred that this was a recent acquisition unique to *Blastocystis* sp. ST1 and ST2. Eme et al. (2017) suggested that this gene is involved in the biosynthesis of tetracenomycin C, an antibiotic also produced by the bacterium *Streptomyces glaucescens*. This compound has a broad cytotoxic activity against the actinomycete bacteria (Hutchinson, 1997). Thus only these two lineages of *Blastocystis* subtypes may have acquired it because they encountered actinomycetes, and/or lived in close proximity to *Streptomyces glaucescens*.

In summary, ASCT1C and SufCB appear to have been gained by the common ancestor of all the species in this analysis. They were conserved in obligate anaerobic lineages, but were lost in the facultative aerobes. Next, the common ancestor of the opalinitans and their closest facultative aerobe lineages acquired glycosyltransferase Gtd1 (i.e., *Cafeteria* sp. and *Incisomona marina*, although the latter lost it independently). Then, the common ancestor of the opalinitans acquired RquA and sialidase. This was followed by the common ancestor of all *Blastocystis* spp. acquiring TXNDC12 and fucose dehydrogenase. *Cafeteria* sp., *Opalina* sp., and *Proteromonas* sp. also independently acquired TXNDC12 from a different donor.

I would like to caution against extrapolating these results to all heterotrophic stramenopiles. There are uncultured marine stramenopiles (MASTs) that branch between the species studied here, whose oxygen tolerance have not yet been identified. The clades MAST-12 and MAST-3 in particular were found to be the closest to the opalinitans in an

SSU rRNA phylogenetic analysis (Yubuki et al., 2010). Obtaining high-quality genomes from all known heterotrophic stramenopiles and MASTs will help resolve the relative timings of the acquisition of these genes.

CHAPTER 4: CONCLUSION

The MRO genome and the nuclear genome were obtained from a *Blastocystis* sp. isolated from the Oriental cockroach. Phylogenomic analysis showed that this species is a new subtype of *Blastocystis* (tentatively named "BBO") and it is deeper-branching than the previously studied mammalian/avian subtypes. This study demonstrated that a combination of long-read and short-read sequencing can isolate a near-complete eukaryotic genome from a heavily-contaminated non-axenic culture. Parts of this workflow can be applied to metagenomic studies, though it does not take into account abundance of organisms because it is based on a culture and not an environmental sample.

The MRO genome of BBO is mostly conserved in gene content and synteny relative to mammalian and avian *Blastocystis* subtypes, although it has nine potentially novel genes of unknown functions. The nuclear genome of BBO on the other hand is highly divergent: orthologs have diverged in amino acid sequence by 52-54% in comparison to other subtypes (vs. 39-41% among *Blastocystis* sp. ST1, ST4-WR1, and ST7) and 15-27% of genes predicted from BBO are unique among *Blastocystis* spp. (vs. 9-20% among the other subtypes).

The most unexpected feature of the BBO nuclear genome is the presence of amino acid homopolymers in 40% of its genes from a variety of functional categories (e.g., ATPases, ribosomal and heatshock proteins). These homopolymers do not seem to destroy the functions of the proteins they 'inhabit'. Extrapolating based on the one case studied here (e.g., VatC), it is possible that these homopolymer regions frequently form surface loops. Further studies, similar to the protein aggregation studies in *D. discoideum*,

need to be pursued to determine what effect these homopolymer tracts have on BBO cells under different conditions.

The BBO nuclear genome does not share the polyadenylation-mediated stop codon generation mechanism found in other subtypes. Whether this is an independent loss or a more recent feature that evolved in the common ancestor of mammalian/avian subtypes will become clearer when genomes of earlier-branching *Blastocystis* spp., such as KOZE1a (isolated from a skink) and TEHE3a (isolated from a tortoise) become available.

Another striking characteristic of both the mitochondrial and nuclear genomes of BBO is that they have extremely low GC content (13.6% and 19.9% respectively). I suggested this might be the result of the accumulation of slightly deleterious or neutral spontaneous deamination and replication errors by DNA polymerase. Our hypothesis is that such mutations have been fixed by genetic drift in BBO. This was possible, we suggest, because of a decrease in the effective population size in ancestors of the BBO lineage that lowered the power of purifying selection and allowed the fixation of slightly deleterious mutations. Whether the BBO genome was actually affected by adaptive selection or neutral drift needs further investigation.

Among the eight LGTs analysed in this project, the two genes RquA and sialidase appear to have been acquired by the common ancestor of the Opalinata lineage. These genes are involved in anaerobic ATP production and immune evasion respectively. The combined presence of these genes in the Opalinata clade indicate their importance in adaptation to the animal gut. To better understand the adaptations of opalinitans to the animal gut, I suggest sequencing the genomes of their closest free-living sister taxa:

MAST-12 and MAST-3. More general patterns with respect to adaptation to the vertebrate gut may be revealed through comparisons of new opalinitan genomes and predicted proteomes to those of gut parasites from other eukaryote groups such as *Giardia* (a metamonad) and *Entamoeba histolytica* (an amoebozoan).

ASCT1C and SufCB were acquired by the common ancestor of all the species studied here (i.e., heterotrophic stramenopiles). They were conserved in both the "parasitic" and free-living obligate anaerobic lineages. This suggests that these two genes were crucial for anaerobic adaptation. A comparative study of the anaerobic ATP production pathway and iron-sulfur cluster synthesis among these species may reveal if they conserve the same pathways.

Finally, I would like to emphasize the importance of basic research – most transformative medicines emerged out of research addressing fundamental questions (Spector et al., 2018). I believe the further investigation of the *Blastocystis* BBO genome and proteome will contribute to our understanding of the role of *Blastocystis* in animal gut health and disease. BBO could even provide insight into prion formation through further research of its polyasparagine-rich proteins.

APPENDIX A – SUPPLEMENTARY TABLES

Table S1. Comparison of GC content (%) of genes across selected *Blastocystis* subtypes. Genes missing from the sp. *Blatta orientalis* are omitted.

Genes	ST4 DMP/02-328	ST1 NandII	ST7-B	BBO
nad1	27.4	27.3	26.0	25.9
nad11	22.0	20.3	21.8	13.0
nad2	17.7	16.7	18.0	15.5
nad3	25.0	25.6	25.0	18.5
nad4	21.0	21.1	22.5	17.8
nad4L	16.2	17.8	16.7	12.1
nad5	25.1	25.1	24.4	18.2
nad6	20.3	17.4	17.7	11.6
nad7	27.8	26.1	27.1	25.0
nad9	22.8	20.3	20.6	13.5
orf160	17.0	11.4	12.1	orf140=7.0
orf149	NA	NA	NA	7.6
orf170	NA	NA	NA	5.9
orf171	NA	NA	NA	7.8
orf177	NA	NA	NA	6.4
orf227	NA	NA	NA	12.0
orf243	NA	NA	NA	4.9
orf251	NA	NA	NA	4.5
orf272	NA	NA	NA	8.4
orf350	NA	NA	NA	6.1
rnl	27.6	28.7	28.8	25.0

Genes	ST4 DMP/02-328	ST1 NandII	ST7-B	BBO
rns	27.7	29.3	26.4	24.3
rpl14	20.7	17.2	17.9	17.1
rpl16	25.7	20.7	21.2	17.8
rpl2	27.4	23.2	23.8	21.8
rpl6	14.5	10.9	12.8	7.9
rps10	19.8	11.1	13.3	9.9
rps12	28.8	25.4	26.3	22.5
rps13	16.8	14.9	14.7	9.9
rps14	18.8	14.2	15.2	12.5
rps19	19.8	18.4	18.9	15.4
rps3	17.1	10.9	12.8	11.5
rps4	15.0	11.0	11.6	10.3
rps8	21.1	16.4	15.4	11.7
trnC(gca)	27.4	27.4	27.4	24.6
trnD(guc)	29.7	36.5	39.2	26.4
trnE(uuc)	38.9	40.3	40.3	34.3
trnF(gaa)	37.0	37.0	37.0	34.7
trnH(gug)	28.4	31.1	28.4	28.4
trnI(gau)	33.8	36.5	33.8	32.9
trnK(uuu)	36.5	43.2	37.8	31.9
trnL(uaa)	28.6	31.3	32.1	33.3
trnM(cau)e1	33.8	33.8	33.8	21.1
trnM(cau)e2	28.8	27.8	29.2	33.3
trnP(ugg)	39.7	39.7	34.2	31.5

Genes	ST4 DMP/02-328	ST1 NandII	ST7-B	BBO
trnW(cca)	28.6	29.6	33.8	25.7
trnY(gua)	39.1	39.5	39.1	42.2

Table S2. HMMscan results. Unassigned ORFs were scanned for signals of HMM profiles of missing genes.

HMM Profile Identifier	Database	Gene	Signal detected in any of the unassigned ORFs?
PF00177.20	Pfam	rps7	No
PF00411.18	Pfam	rps11	No
PTHR11759.orig.30.pir	PANTHER	rps11	No

Table S3. Sources of transcriptome data used for LGT analysis

Transcriptome	Source
Proteromonas sp.	In-house
Opalina sp.	In-house
Blastocystis sp. TEHE3a	In-house, multiplexed with other organisms (not shown here)
Blastocystis sp. KOZE1a	
Rictus lutensis	In-house, multiplexed with other organisms (not shown here)
Halocafeteria seosinensis	In-house
Cafeteria sp.	MMETSP1104 (Caron, 2000)
Cantina marsupialis	NCBI SRA: DRX027417 (Noguchi et al., 2015)

Table S4. Codon usage (%) of *Blastocystis* ST1 NandII, ST4-WR1, ST7, and BBO.

Amino acid	Codon	ST1	ST4	ST7	BBO
Ala	GCA	8.5	29.7	13.6	57.0
	GCC	30.5	14.3	23.1	4.5
	GCG	48.6	7.8	38.8	3.6
	GCT	12.4	48.1	24.5	34.8
Cys	TGC	77.5	20.7	61.6	8.6
	TGT	22.5	79.3	38.4	91.4
Asp	GAC	71.1	20.0	47.8	7.6
	GAT	28.9	80.0	52.2	92.4
Glu	GAA	19.1	55.7	42.3	88.4
	GAG	80.9	44.3	57.7	11.6
Phe	TTC	81.6	50.0	64.0	23.3
	TTT	18.4	50.0	36.0	76.7
Gly	GGA	20.1	47.3	39.2	40.3
	GGC	51.5	8.4	29.2	4.8
	GGG	13.3	7.7	12.9	1.9
	GGT	15.0	36.7	18.6	53.0
His	CAC	78.2	24.7	56.0	12.6
	CAT	21.8	75.3	44.0	87.4
Ile	ATA	3.6	11.3	8.1	43.8
	ATC	66.9	29.3	51.1	10.9
	ATT	29.5	59.5	40.8	45.3
Lys	AAA	10.8	38.5	33.6	76.5
	AAG	89.2	61.5	66.4	23.5
Leu	CTA	2.4	9.0	4.5	9.9
	CTC	22.1	12.9	21.8	1.8
	CTG	55.0	15.6	34.1	3.5
	CTT	7.7	22.4	13.3	7.8
	TTA	1.3	14.5	7.0	60.2
	TTG	11.6	25.6	19.3	16.9
Met	ATG	100.0	100.0	100.0	100.0
Asn	AAC	76.7	27.7	56.8	7.4
	AAT	23.3	72.3	43.2	92.6
Pro	CCA	7.1	29.7	13.6	68.4

Amino acid	Codon	ST1	ST4	ST7	BBO
Pro (cont.)	CCC	41.0	11.3	29.3	2.0
	CCG	36.5	7.7	29.8	2.2
	CCT	15.4	51.3	27.2	27.4
Gln	CAA	16.1	51.7	34.4	86.9
	CAG	83.9	48.3	65.6	13.1
Arg	AGA	10.3	15.8	12.9	60.3
	AGG	6.6	5.4	5.4	5.4
	CGA	9.0	28.2	20.3	16.2
	CGC	45.7	7.2	30.9	1.9
	CGG	15.5	7.7	12.2	0.5
	CGT	13.0	35.8	18.2	15.6
Ser	AGC	24.8	5.7	18.7	4.6
	AGT	9.0	22.9	11.6	35.5
	TCA	4.1	21.7	7.9	35.1
	TCC	27.5	11.1	19.1	3.5
	TCG	24.6	11.3	25.8	5.1
	TCT	10.0	27.4	16.8	16.3
Thr	ACA	9.5	37.0	15.9	62.4
	ACC	29.4	13.5	25.3	5.9
	ACG	51.3	13.1	38.8	5.2
	ACT	9.8	36.5	20.0	26.5
Val	GTA	3.5	21.0	9.5	33.9
	GTC	17.5	12.9	19.1	5.0
	GTG	69.2	33.3	51.1	14.3
	GTT	9.7	32.7	20.3	46.7
Trp	TGG	100.0	100.0	100.0	100.0
Tyr	TAC	78.9	27.3	53.7	8.2
	TAT	21.1	72.7	46.3	91.8
STOP	TAA	41.4	41.5	39.0	65.2
	TAG	32.3	26.4	26.8	18.0
	TGA	26.3	32.1	34.2	16.8

Table S5. InterProScan annotation of polyasparagine-containing genes, grouped into general domains/subunits/functions. Note that these are not curated.

Pfam/TIGFRAM annotation	Frequency
Zinc finger	65
WD repeat	63
Protein kinase domain	60
MORN repeat	46
RNA recognition motif	38
Helicase domain	32
AAA domain	27
Elongation factor	23
ATPase family	22
DEAD/DEAH box helicase	18
Ankyrin repeat	17
Hsp70 protein	17
MutS domain	17
Myb-like DNA-binding	15
tRNA synthetase	15
Glycosyltransferase	14
Kinesin motor domain	13
Protein phosphatase	12
Ribosomal protein	12
SNF family	12
Vacuolar sorting protein	12
Ubiquitin hydrolase	11
ABC transporter	10
Chromosome condensation repeat	10
Major Facilitator Superfamily	10
PPR repeat	10
Glycosyl hydrolase	9
Carbamoyl-phosphate synthase subunit	8
GTP-binding protein domain	8
Nodulation protein Z	8
RNA polymerase domain	8
DnaJ domain	7
PH domain	7
PHD-finger	7
Ring finger domain	7
Tetratricopeptide repeat	7
DNA polymerase	6
KH domain	6

Pfam/TIGFRAM annotation	Frequency
Leucine Rich repeat	6
MULE transposase domain	6
Poly(A) polymerase domain	6
Ribonuclease domain	6
AMP-binding enzyme	5
Beta-ketoacyl synthase domain	5
C2 domain	5
Coiled-coil domain	5
Cyclin domain	5
DEAD_2	5
Eukaryotic translation initiation factor	5
Mob1/phocein family	5
MyTH4 domain	5
Nucleotide-diphospho-sugar transferase	5
Peptidase family	5
Ras family	5
RhoGAP domain	5
RNA helicase	5
Tesmin/TSO1-like CXC cysteine-rich domain	5
Acyltransferase	4
AIR synthase protein	4
Chromo domain	4
DHHC palmitoyltransferase	4
Dual specificity phosphatase domain	4
EF-hand domain pair	4
Fibronectin type III domain	4
GRIP domain	4
HMG box	4
IKI3 family	4
LNR domain	4
MIF4G domain	4
RecF/RecN/SMC N terminal domain	4
Replication factor domain	4
SH3 domain	4
tRNA methyltransferase	4
tRNA selenium transferase	4
Tub family	4
Adaptin N terminal region	3
Adenosine deaminase	3
AFG1-like ATPase	3

Pfam/TIGFRAM annotation	Frequency
Alpha-kinase family	3
Aminotransferase	3
Anaphase-promoting complex	3
Anticodon-binding domain	3
ARID/BRIGHT DNA binding domain	3
AT hook motif	3
BRCT domain, a BRCA1 C-terminus domain	3
Chorein	3
Coatomer WD associated region	3
Collagen triple helix repeat	3
CPSF A subunit region	3
DNA repair helicase	3
Dynamin domain	3
Eukaryotic glutathione synthase	3
Exonuclease	3
Glucosamine-6-phosphate deaminase	3
Histidine kinase	3
Histone-binding protein RBBP4 or subunit C of CAF1 complex	3
HSF-type DNA-binding	3
Hydroxymethylglutaryl-coenzyme A reductase	3
Metallo-beta-lactamase superfamily	3
NLI interacting factor-like phosphatase	3
Oligonucleotide/oligosaccharide-binding fold	3
Phosphatidylinositol-4-phosphate 5-Kinase	3
Phospholipid-translocating P-type ATPase C-terminal	3
Phosphotyrosyl phosphate activator protein	3
Protein tyrosine kinase	3
Protein-only RNase P	3
Pyruvate:ferredoxin flavodoxin oxidoreductase	3
Rab-GTPase-TBC domain	3
RNA pol I transcription initiation factor	3
Sec7 domain	3
Serine dehydratase	3
Sir2 family	3
Subtilase family	3
Translation initiation factor	3
Ubiquitin family	3
XRN 5'-3' exonuclease N-terminus	3
Acetyl-CoA carboxylase	2
Adenosylmethionine decarboxylase	2

Pfam/TIGFRAM annotation	Frequency
Alanine dehydrogenase/PNT domain	2
ALG6, ALG8 glycosyltransferase family	2
Alpha galactosidase A	2
Arginine-tRNA ligase	2
ATP-dependent DNA helicase, RecQ family	2
ATP-grasp domain	2
Beta-Casp domain	2
Biotin carboxylase domain	2
BRCA1 C Terminus domain	2
Bromodomain	2
BTB/POZ domain	2
Calcineurin-like phosphoesterase	2
Calponin homology domain	2
Carbohydrate phosphorylase	2
CCR4-Not complex component	2
Cell division protein	2
Chaperone protein DnaK	2
CHORD	2
COG4 transport protein	2
Conserved oligomeric complex COG6	2
Cyclin-dependent kinase regulatory subunit	2
Cysteine-tRNA ligase	2
Diacylglycerol kinase catalytic domain	2
Diphthamide synthase	2
DNA gyrase	2
DNA topoisomerase	2
Dullard-like phosphatase domain	2
E1-E2 ATPase	2
E2F/DP family winged-helix DNA-binding domain	2
EGF-like domain	2
Endonuclease	2
Endonuclease/Exonuclease/phosphatase family	2
Est1 DNA/RNA binding domain	2
FANCI solenoid	2
FATC domain	2
Fe-S dicluster domain	2
FHA domain	2
Formin Homology 2 Domain	2
FtsH family	2
FtsJ-like methyltransferase	2

Pfam/TIGFRAM annotation	Frequency
Fumble	2
G-patch domain	2
Galactosyltransferase	2
GDP dissociation inhibitor	2
GHMP kinases N terminal domain	2
Glucosamine-6-phosphate isomerases	2
Glutamine amidotransferase	2
Glycogen debranching enzyme	2
Glycogen debranching enzyme	2
GNT-I family	2
GWT1	2
GYF domain	2
HECT-domain	2
Histidine phosphatase superfamily	2
Inositol hexakisphosphate	2
Insulinase	2
JAB1/Mov34/MPN/PAD-1 ubiquitin protease	2
LIM domain	2
Lipase	2
Lysine methyltransferase	2
Methyltransferase TYW3	2
Mitochondrial carrier protein	2
Mono-functional DNA-alkylating methyl methanesulfonate	2
MYND finger	2
NAD dependent epimerase/dehydratase family	2
NADH-quinone oxidoreductase subunit	2
NADP transhydrogenase subunit	2
NOA36 protein	2
NUC153 domain	2
Nucleotide-sugar transporter	2
Oxidoreductase NAD-binding domain	2
PAP2 superfamily	2
pfkB family carbohydrate kinase	2
PHD-like zinc-binding domain	2
Phosphatidylinositol 3- and 4-kinase	2
Phosphoinositide 3-kinase family	2
Phospholipase/Carboxylesterase	2
Phospholipid-translocating ATPase N-terminal	2
PIF1-like helicase	2
PPPDE putative peptidase domain	2

Pfam/TIGFRAM annotation	Frequency
Pre-mRNA processing factor	2
PUB domain	2
Pumilio-family RNA binding repeat	2
Pyridine nucleotide-disulphide oxidoreductase	2
Pyruvate phosphate dikinase	2
Rad51	2
Radical SAM superfamily	2
Rapamycin-insensitive companion of mTOR, domain 5	2
RAVE protein 1 C terminal	2
RecQ mediated genome instability protein	2
RecQ zinc-binding	2
RhoGEF domain	2
Ribosomal RNA large subunit methyltransferase E	2
RIO1 family	2
RNA cap guanine-N2 methyltransferase	2
RNB domain	2
Rtf2 RING-finger	2
Scd6-like Sm domain	2
Sec1 family	2
Serine aminopeptidase, S33	2
SET domain	2
Signal peptide peptidase	2
Sin3 binding region of histone deacetylase	2
SPFH domain / Band 7 family	2
SpoU rRNA Methylase family	2
SPRY domain	2
Stealth protein	2
TCP-1/cpn60 chaperonin family	2
ThiF family	2
Thioredoxin	2
TMEM154 protein family	2
Toprim domain	2
Transcription factor TFIIB repeat	2
Trehalose-phosphatase	2
Twin BRCT domain	2
Type III restriction enzyme	2
Tyrosyl-DNA phosphodiesterase	2
U-box domain	2
Ubiquitin-conjugating enzyme	2
Utp14 protein	2

Pfam/TIGFRAM annotation	Frequency
Variant SH3 domain	2
XPG N-terminal domain	2
Xylanase inhibitor	2
Zinc-ribbon	2
ZIP Zinc transporter	2
Abhydrolase domain	1
ADP-ribosylglycohydrolase	1
Alcohol dehydrogenase domain	1
Alpha/beta hydrolase family	1
Amidohydrolase family	1
Amidophosphoribosyltransferase	1
Amino acid permease	1
Amylo-alpha-1,6-glucosidase	1
Apurinic endonuclease	1
Apyrase	1
Armadillo/beta-catenin-like repeat	1
Asparagine synthase	1
ATP-dependent zinc metalloprotease FtsH	1
ATP-NAD kinase	1
Bacterial Ig-like domain	1
BAG domain	1
Beige/BEACH domain	1
Beta-acetyl hexosaminidase like	1
Beta2-adaptin appendage	1
Biotin ligase	1
Biotin-requiring enzyme	1
Biotin/lipoate A/B protein ligase family	1
Brf1-like TBP-binding domain	1
Bromodomain transcription regulation	1
bZIP transcription factor	1
C-5 cytosine-specific DNA methylase	1
C-terminal associated domain of TOPRIM	1
CAP-Gly domain	1
Carboxyl transferase domain	1
Casein kinase II regulatory subunit	1
Cation transport ATPase	1
CBF/Mak21 family	1
CDC45-like protein	1
Chitin synthase	1
CLASP N terminal	1

Pfam/TIGFRAM annotation	Frequency
Class II Aldolase and Adducin N-terminal domain	1
Clathrin adaptor complex small chain	1
Cleft lip and palate transmembrane protein 1	1
CoA binding domain	1
CobB/CobQ-like glutamine amidotransferase domain	1
Coiled coil protein 84	1
COMM domain	1
Condensin complex subunit 2	1
Conserved hypothetical ATP binding protein	1
Copine	1
CRAL/TRIO domain	1
Ctf8	1
CUE domain	1
Cyclophilin type peptidyl-prolyl cis-trans isomerase	1
Cytidine and deoxycytidylate deaminase zinc-binding region	1
Cytosolic iron-sulfur protein assembly protein	1
DALR anticodon binding domain	1
DDE superfamily endonuclease	1
dDENN domain	1
DEK C terminal domain	1
DENN domain	1
Diaphanous FH3 Domain	1
Dihydrouridine synthase	1
Dimerisation and cyclophilin-binding domain of Mon2	1
Dinuclear metal center protein	1
DIRP	1
Divergent CRAL/TRIO domain	1
DNA mismatch repair protein	1
DNA repair protein Ercc1	1
Dyggve-Melchior-Clausen syndrome protein	1
EF hand	1
Electron transfer flavoprotein FAD-binding domain	1
Electron transfer flavoprotein-ubiquinone oxidoreductase	1
ELM2 domain	1
EMG1/NEP1 methyltransferase	1
Endopeptidase La	1
ER-Golgi trafficking complex subunit	1
ERCC3/RAD25/XPB C-terminal helicase	1
Eukaryotic initiation factor 4E	1
Eukaryotic membrane protein family	1

Pfam/TIGFRAM annotation	Frequency
Eukaryotic porin	1
Exportin 1-like protein	1
F/Y rich C-terminus	1
F/Y-rich N-terminus	1
FAD binding domain	1
FAD dependent oxidoreductase	1
FERM central domain	1
FGAM synthetase	1
Flavodoxin	1
Formyl transferase	1
FtsX-like permease family	1
Galactose mutarotase-like	1
Glutamate-cysteine ligase	1
Glutathione S-transferase domain	1
Glycogen/starch/alpha-glucan phosphorylases	1
GNAT acetyltransferase 2	1
GNS1/SUR4 family	1
GTP-binding protein LepA C-terminus	1
Guanine nucleotide exchange factor in Golgi transport	1
GUCT domain	1
HAD-hyrolase-like	1
Haloacid dehalogenase-like hydrolase	1
HEAT repeat	1
Helix-hairpin-helix motif	1
Histone RNA hairpin-binding protein RNA-binding domain	1
Histone-like transcription factor	1
Homeobox domain	1
Homeodomain-like domain	1
Hormone receptor domain	1
HRDC domain	1
Hsp20/alpha crystallin family	1
Hyaluronan / mRNA binding family	1
Importin-beta N-terminal domain	1
Inorganic H+ pyrophosphatase	1
Inositol polyphosphate kinase	1
Inositol-pentakisphosphate 2-kinase	1
IPP transferase	1
Isoleucine-tRNA ligase	1
Ketoacyl-synthetase C-terminal extension	1
KOW motif	1

Pfam/TIGFRAM annotation	Frequency
KR domain	1
L-fucokinase	1
L28: ribosomal protein bL28	1
La domain	1
Las1-like	1
Legume-like lectin family	1
Leo1-like protein	1
Lipin, N-terminal conserved region	1
LNS2 (Lipin/Ned1/Smp2)	1
Lon protease (S16) C-terminal proteolytic domain	1
Lysine-tRNA ligase	1
MA3 domain	1
Mago binding	1
Mago nashi protein	1
Maintenance of mitochondrial morphology protein 1	1
Maintenance of mitochondrial structure and function	1
Man1-Src1p-C-terminal domain	1
MCM2/3/5 family	1
MED6 mediator sub complex component	1
Mediator complex subunit 27	1
Memo-like protein	1
Methylmalonic aciduria and homocystinuria type C family	1
Mevalonate kinase	1
MGS-like domain	1
Misato Segment II tubulin-like domain	1
Mitochondrial dehydrogenase kinase	1
Mitochondrial pyridine nucleotide transhydrogenase	1
MJ0570_dom: MJ0570-related uncharacterized domain	1
MT-A70	1
MuDR family transposase	1
N-acetylglucosamine-6-phosphate deacetylase	1
N-acetyltransferase B complex subunit	1
N2,N2-dimethylguanosine tRNA methyltransferase	1
N2227-like protein	1
Na+/H+ antiporter family	1
NAD kinase	1
NEMP family	1
Neurochondrin	1
NGG1p interacting factor 3	1
Nin one binding Zn-ribbon like	1

Pfam/TIGFRAM annotation	Frequency
NMD3 family	1
Non-SMC mitotic condensation complex subunit 1	1
NRDE-2 for RNA interference	1
NUC173 domain	1
Nuclear fragile X mental retardation-interacting protein	1
Nucleolar complex-associated protein	1
Nucleopolyhedrovirus protein	1
Nucleoporin autopeptidase	1
Nucleotidyltransferase domain	1
NUDIX domain	1
OB-fold nucleic acid binding domain	1
Origin recognition complex subunit	1
Paf1	1
Partial alpha/beta-hydrolase lipase region	1
Patatin-like phospholipase	1
PD-(D/E)XK nuclease superfamily	1
Pentacotripeptide-repeat region of PROPR	1
PEP-utilising enzyme, mobile domain	1
Pep3/Vps18/deep orange family	1
Peptidase M16 domain	1
Peptidyl-tRNA hydrolase PTH2	1
Pescadillo homolog	1
Pescadillo N-terminus	1
Phorbol esters/diacylglycerol binding domain	1
Phosphate transporter family	1
Phosphatidylinositol N-acetylglucosaminyltransferase	1
Phosphatidylserine decarboxylase	1
Phosphopantetheine attachment site	1
Phosphopantetheinyl transferase	1
Phosphoribosylformylglycinamidine synthase	1
Phosphorylated CTD interacting factor 1 WW domain	1
PIN domain of ribonuclease	1
Plasma-membrane choline transporter	1
POLO box duplicated region	1
Polyketide synthase dehydratase	1
Polyprenyl synthetase	1
PP-loop family	1
Pre-rRNA-processing protein TSR2	1
Prolyl oligopeptidase family	1
Protein prenyltransferase alpha subunit repeat	1

Pfam/TIGFRAM annotation	Frequency
PS_decarb: phosphatidylserine decarboxylase	1
PSP1 C-terminal conserved region	1
PUL domain	1
PX domain	1
Pyridoxal-dependent decarboxylase, C-terminal sheet domain	1
Pyridoxal-dependent decarboxylase, pyridoxal binding domain	1
Pyruvate:ferredoxin oxidoreductase	1
Queuine tRNA-ribosyltransferase	1
Rab3 GTPase-activating protein catalytic subunit	1
Rad52/22 family double-strand break repair protein	1
Radical SAM methylthiotransferase	1
Rapamycin-insensitive companion of mTOR, N-term	1
Raptor N-terminal CASPase like domain	1
Ras-induced vulval development antagonist	1
Region in Clathrin and VPS	1
Regulated-SNARE-like domain	1
Repeating coiled region of VPS13	1
Response regulator receiver domain	1
Retinoblastoma-associated protein B domain	1
Retinoic acid induced 16-like protein	1
Rft protein	1
RING/Ubox like zinc-binding domain	1
RNA pol II transcription elongation factor	1
RNA pseudouridylate synthase	1
RNA-Pol-II transcription regulator	1
RQC domain	1
Rrp44-like cold shock domain	1
RWD domain	1
S1 domain	1
SAC3/GANP family	1
SacI homology domain	1
Sad1 / UNC-like C-terminal	1
Scavenger mRNA decapping enzyme	1
Sec23-binding domain of Sec16	1
Serine incorporator	1
SGF29 tudor-like domain	1
SGT1 protein	1
Short chain dehydrogenase	1
SHQ1 protein	1
Sin-like protein	1

Pfam/TIGFRAM annotation	Frequency
SIS domain	1
Small nuclear ribonucleoprotein	1
SMC proteins Flexible Hinge Domain	1
Snare region anchored in the vesicle membrane C-terminus	1
snRNA-activating protein of 50kDa MW C terminal	1
Sodium/hydrogen exchanger family	1
Solute carrier family 35	1
Spc97 / Spc98 family	1
SprT-like family	1
Ssl1-like	1
STAG domain	1
Succinyl-CoA ligase like flavodoxin domain	1
Surfeit locus protein 2	1
Surp module	1
SWIB/MDM2 domain	1
TAF6 C-terminal HEAT repeat domain	1
TATA box binding protein associated factor	1
TatD related DNase	1
Telomerase activating protein Est1	1
Telomere stability and silencing	1
Telomeric single stranded DNA binding	1
TENA/THI-4/PQQC family	1
Tetraspanin family	1
TFIIB zinc-binding	1
TFIIH C1-like domain	1
Thg1 C terminal domain	1
Thiamine pyrophosphate enzyme	1
Thioredoxin-like domain	1
TLD	1
Transcription factor ssl1	1
Transcription initiation factor	1
Transcriptional repressor TCF25	1
Translationally controlled tumour protein	1
Transmembrane Fragile-X-F protein	1
Transmembrane protein 43	1
Transmembrane receptor	1
Transposase	1
Triose-phosphate Transporter family	1
tRNA guanylyltransferase	1
tRNA intron endonuclease	1

Pfam/TIGFRAM annotation	Frequency
tRNA-guanine family transglycosylase	1
U3 small nucleolar RNA-associated protein 6	1
Ubiquitin binding	1
Ubiquitin elongating factor core	1
Ubiquitin fusion degradation protein	1
uDENN domain	1
UDP-galactose transporter	1
Ulp1 protease family, C-terminal catalytic domain	1
Utp21 specific WD40 associated putative domain	1
V-ATPase subunit C	1
V-type H translocating pyrophosphatase	1
Vacuolar protein C-terminal binding	1
VHS domain	1
Villin headpiece domain	1
VRR-NUC domain	1
WGR domain	1
WW domain	1
Wyosine base formation	1
XPG I-region	1
Xylose isomerase-like TIM barrel	1
Yippee zinc-binding/DNA-binding centromere assembly	1
Zinc carboxypeptidase	1
Zn-dependent metallo-hydrolase RNA domain	1

APPENDIX B – SUPPLEMENTARY FIGURES

Figure S1: A multinucleated *Blastocystis* sp. cell of about 20μm isolated from *Blatta orientalis*, observed under a ZEISS inverted light microscope. The smaller rod-shaped and oval-shaped cells in the background are bacteria. Blue arrows = *Blastocystis* nucleus, red arrows = bacteria

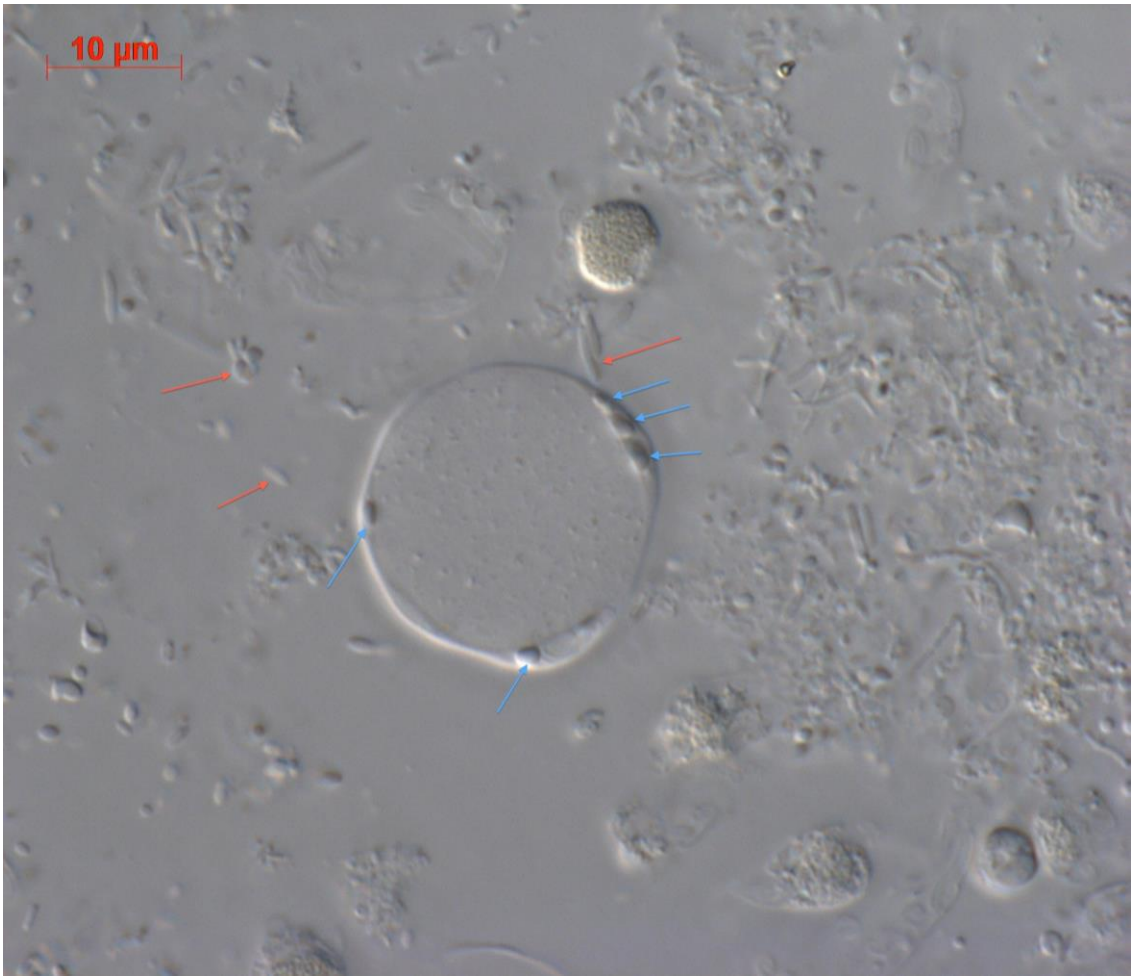


Figure S2: Encysted *Blastocystis* cells. Each large cell contains many smaller cysts. One such cell can be seen bursting open (circled in red), cysts ready to exit the gut and go into the external environment were it not in a culture tube. Two normal, vacuolar *Blastocystis* cells are also present (indicated by red arrows), their vacuole filling up most of their internal space as is typical of them.

Figure S3: RNA reads (top panel) and short DNA Illumina reads (centre panel) mapped onto the *Blastocystis* sp. MRO genome (bottom sequence), viewed with IGV (Thorvaldsdóttir, et al., 2013). Grey blocks represent read coverage; the taller the block, the higher the number of reads. Note the purple line indicating an insertion of a base (moving the mouse cursor over it in IGV will reveal the inserted base. In this case, it was adenine).

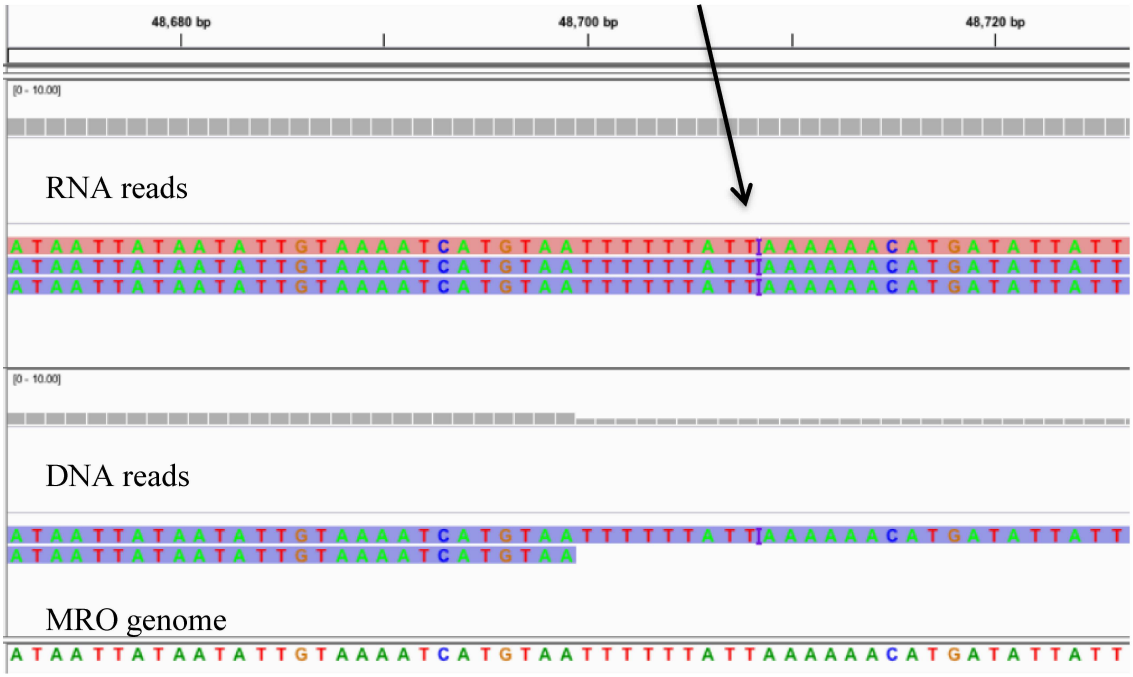


Figure S4: The first draft sequence of the BBO MRO genome. The dotted arcs indicate the duplicated regions. Some genes are missing/unannotated here as this draft has not been fully manually corrected. Oligonucleotide PCR primer pairs are represented in red, with arrows showing direction of amplifications. Outer and inner grey arrows show directions of transcription for the genes represented on the outer and inner rings respectively.

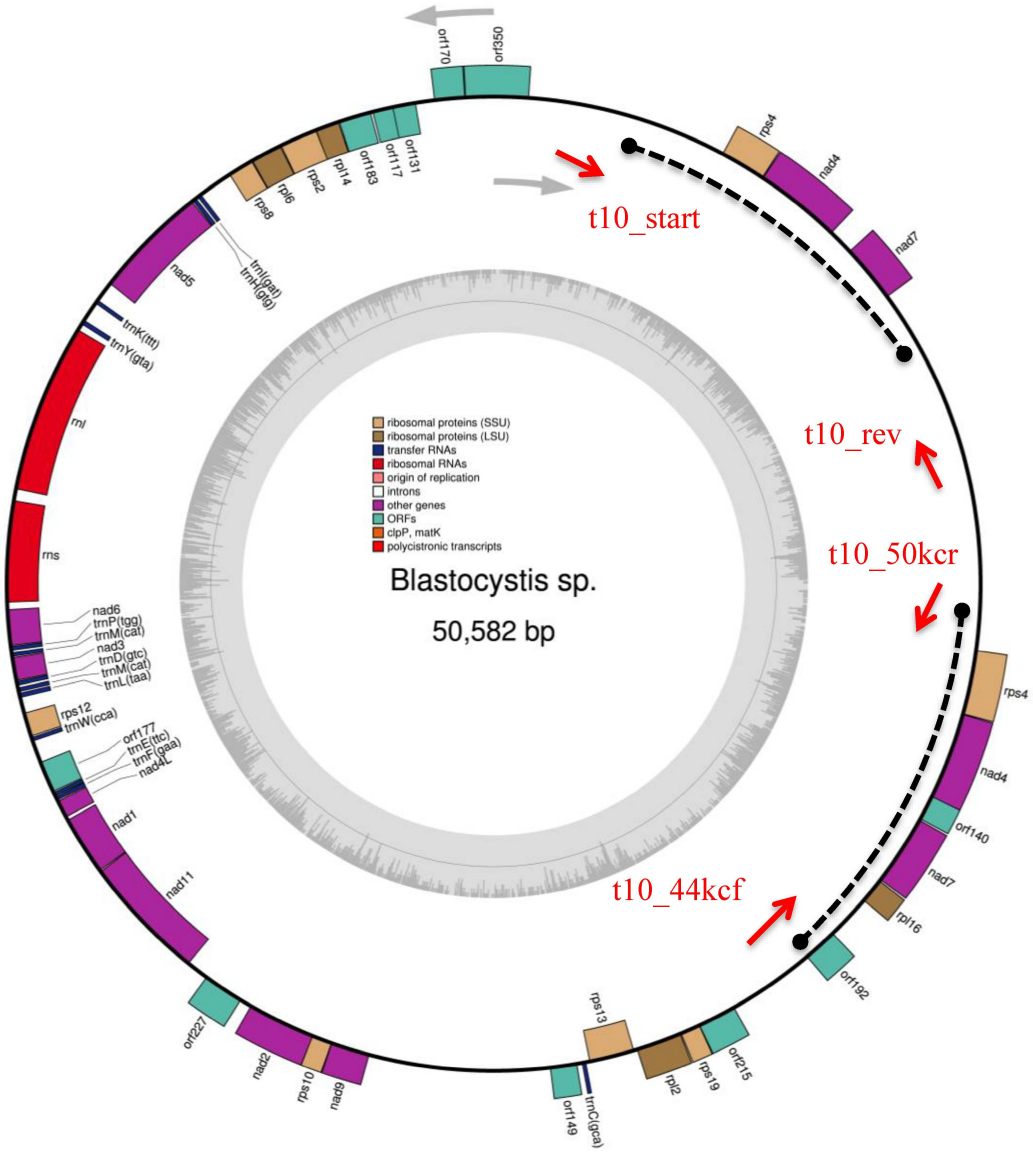


Figure S5: Secondary structure predictions of rps7 (left) from the ST8 MRO as a representative, and orf177 (right) from the *Blastocystis* sp. *Blatta orientalis* MRO, predicted using PSIPRED (Buchan et al., 2013).

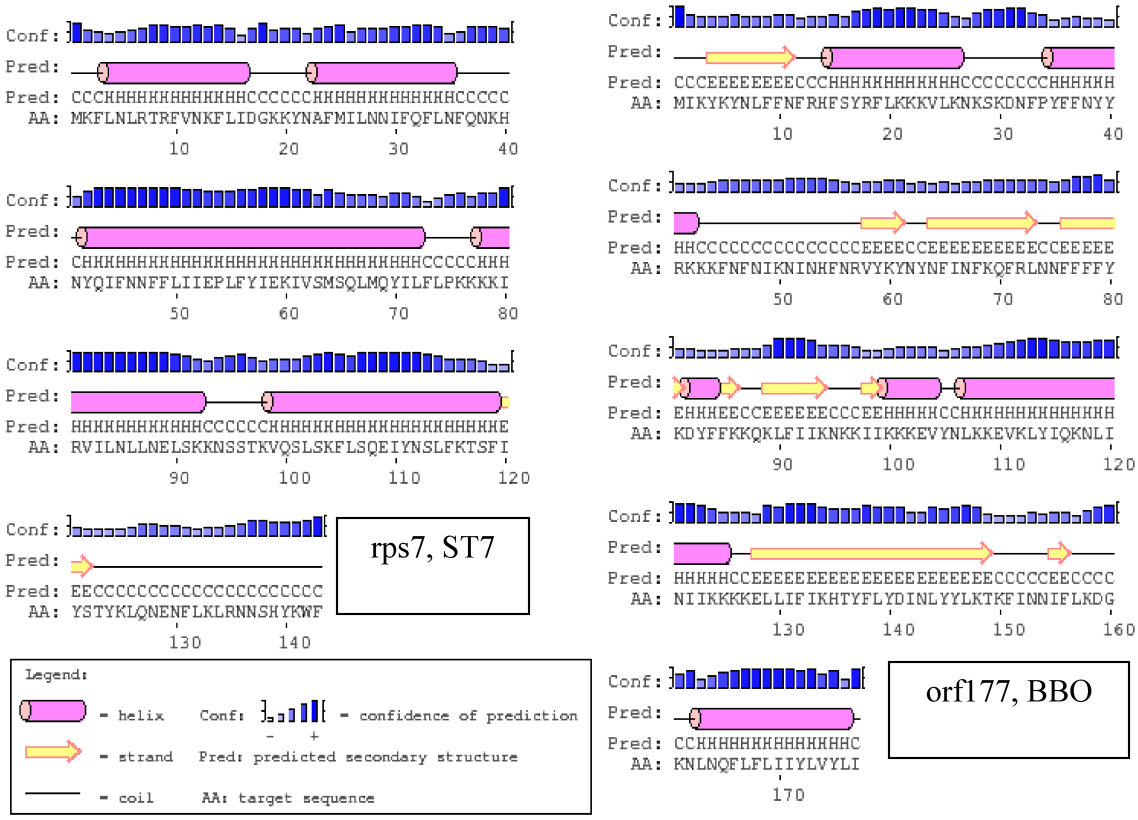


Figure S6: A close-up of the SSU rRNA phylogenetic tree of parabasalids. The contaminant parabasalid isolated from the BBO culture is in magenta. Maximum-likelihood tree estimated using IQ-TREE with GTR+R7 model (Tavaré, 1986; Yang, 1995; Soubrier et al., 2012). 982 nucleotide sites used, with a 1000-replicate ultrafast bootstrap approximation for branch support (Minh et al., 2013). *The branch length of the *Trichonympha* clade was halved to fit into this diagram.

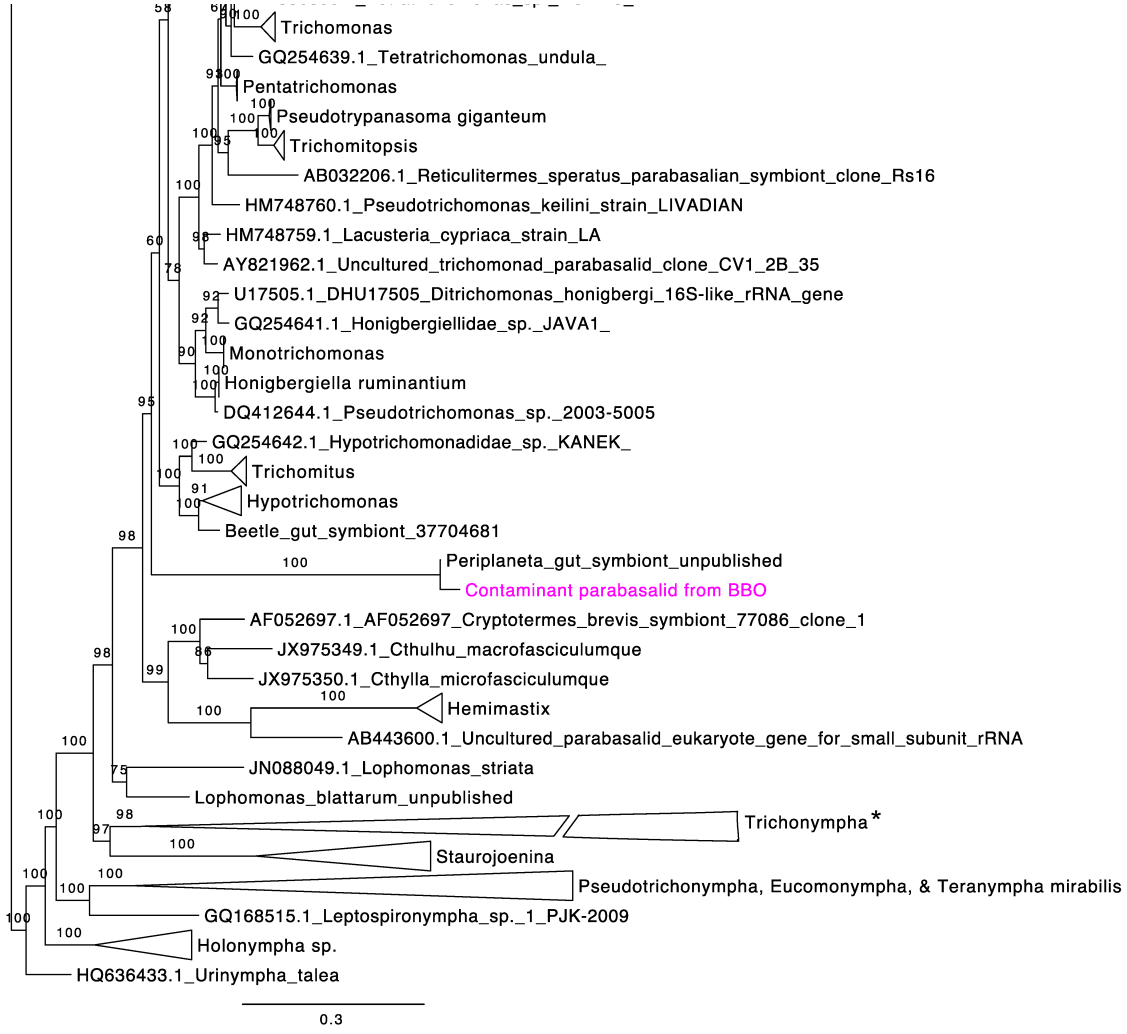
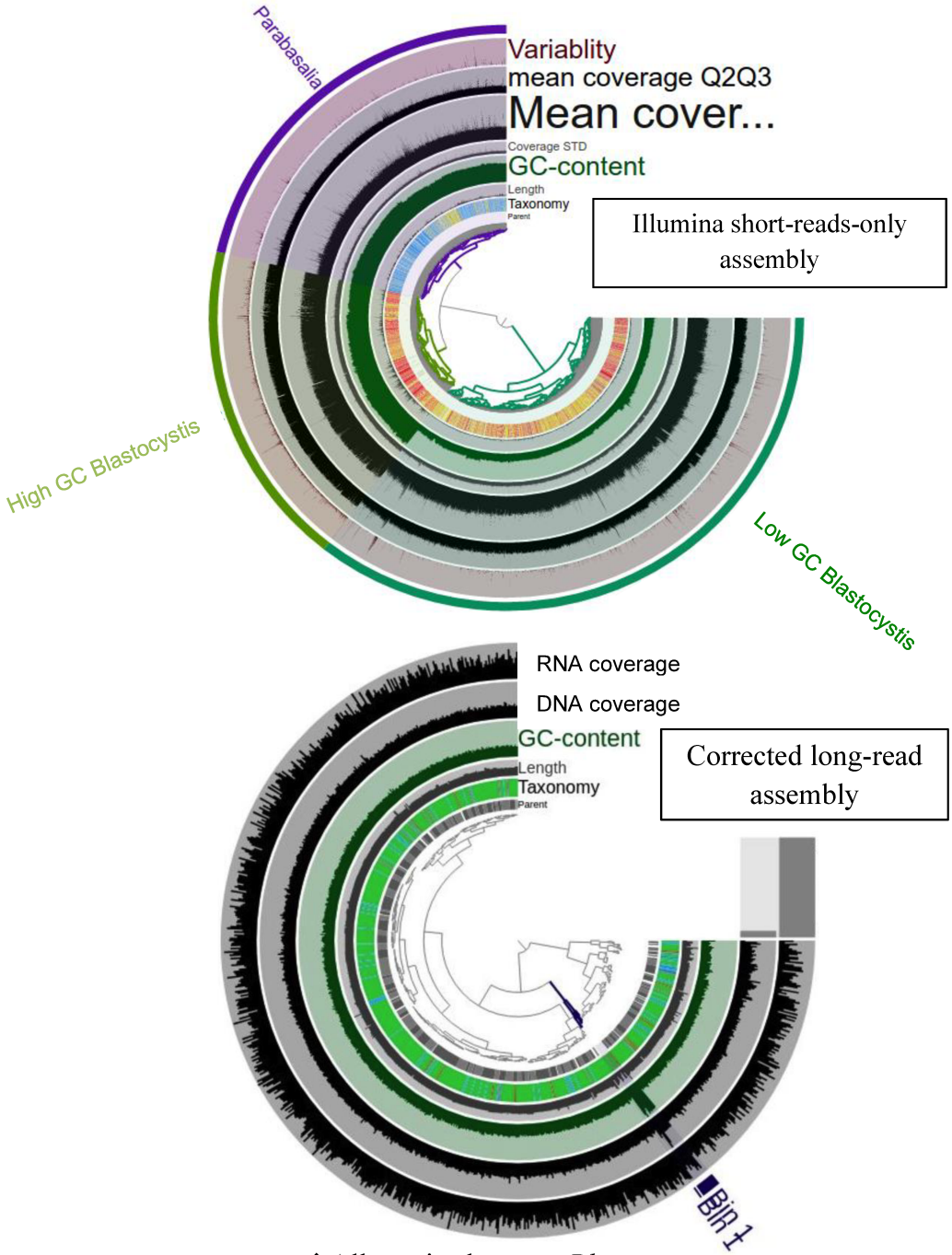


Figure S7: Anvi'o (Eren et al., 2015) display of the polished and corrected long-read BBO genome assembly (bottom) and the Illumina-only assembly (top). The latter divided into three distinct categories or "bins": Parabasalia (possibly a contamination in cell culture), High GC Blastocystis, and Low GC Blastocystis. See table for details on this assembly. For the long-read assembly, the outermost layer displays RNA-seq coverage, followed by Illumina DNA coverage. Note the spike in GC content (third layer) and DNA coverage of Bin 1.

Illumina-only assembly	Number of contigs	GC content (%)	Size (Mb)
High GC Blastocystis	2905	59	21.5
Low GC Blastocystis	5765	21	13.9
Parabasalia	3468	52	13.1



↑All contigs here are *Blastocytis* sequences.

Figure S8: Close-up of the maximum-likelihood tree of the "High GC Blastocystis" bin (in orange) among stramenopiles. BBO (the lower GC-content genome) in magenta. Made using IQ-TREE, substitution model LG4X, 13 genes, 3373 informative sites.

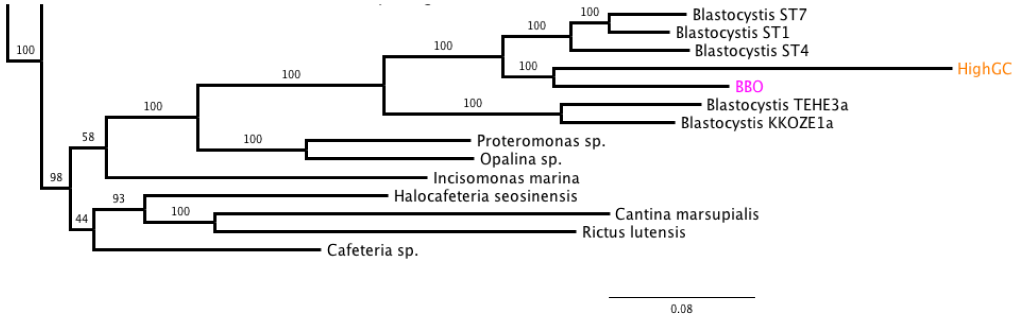


Figure S9: Graph of intron length distribution of all genes with at least one intron. 300 bases was chosen as the cutoff for analysis because the frequency tapers off at this point, and longer introns were not backed up by RNA-seq data.

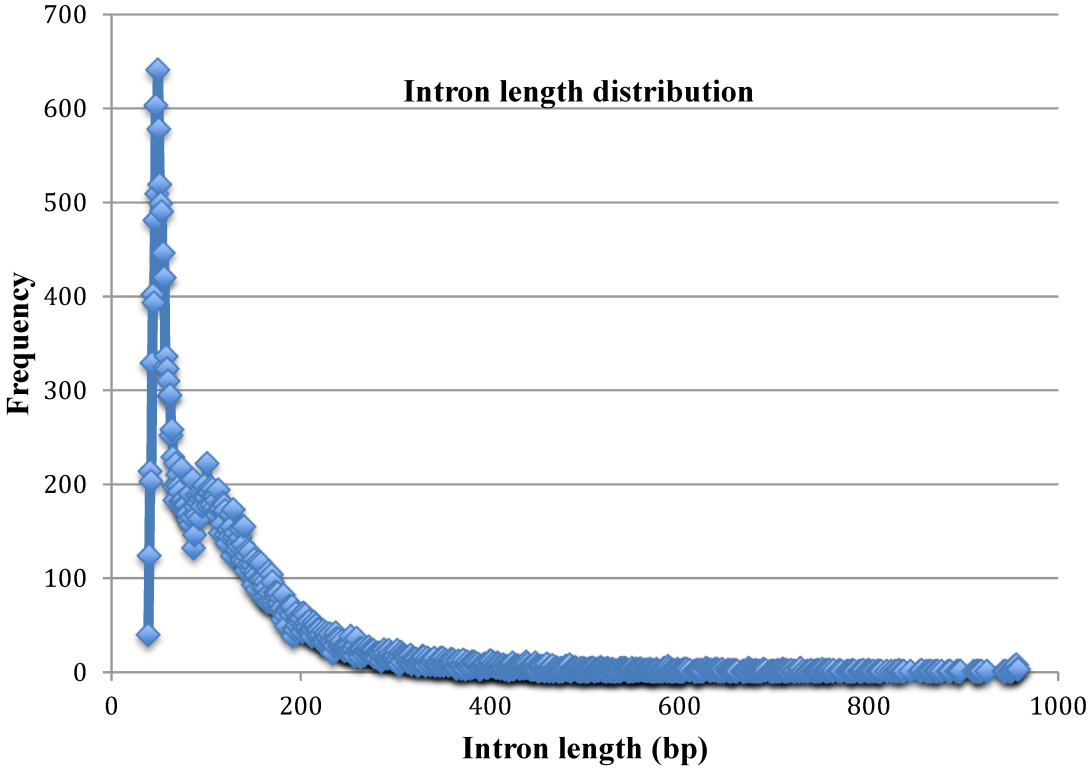


Figure S10: An example of a BBO homolog of a gene from *Blastocystis* ST1 NandII in which its stop codon is generated by polyadenylation, displaying a standard stop codon (black vertical line indicates end of stop codon) and normal RNA read coverage, showing that it does not have this mechanism. In addition, there was no polyadenylation-stop motif (TGTTTGTT) downstream of the stop codon, which is highly conserved in *Blastocystis* ST1, ST4, and ST7 (Gentekaki et al., 2017). Reads past the stop codon make up the 3' untranslated region. The gene shown here is a calcium-binding mitochondrial carrier protein Aralar 2 (according to best BLASTp hit against nr; accession number OAO15013.1).

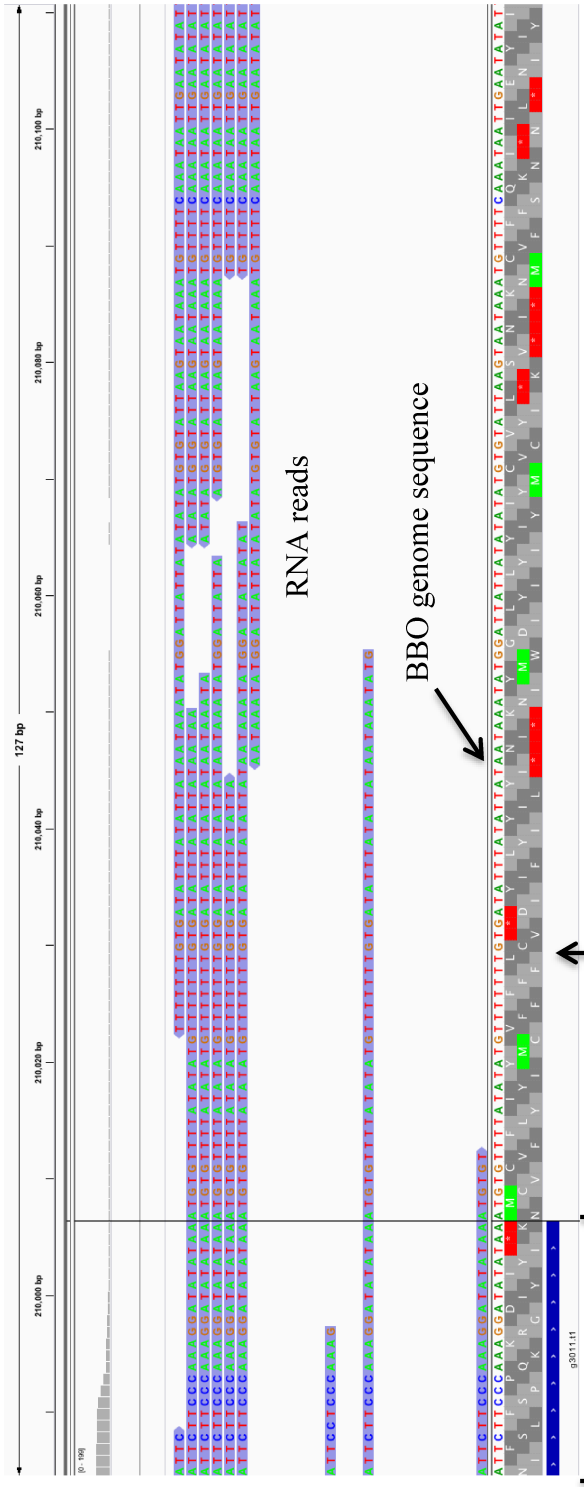
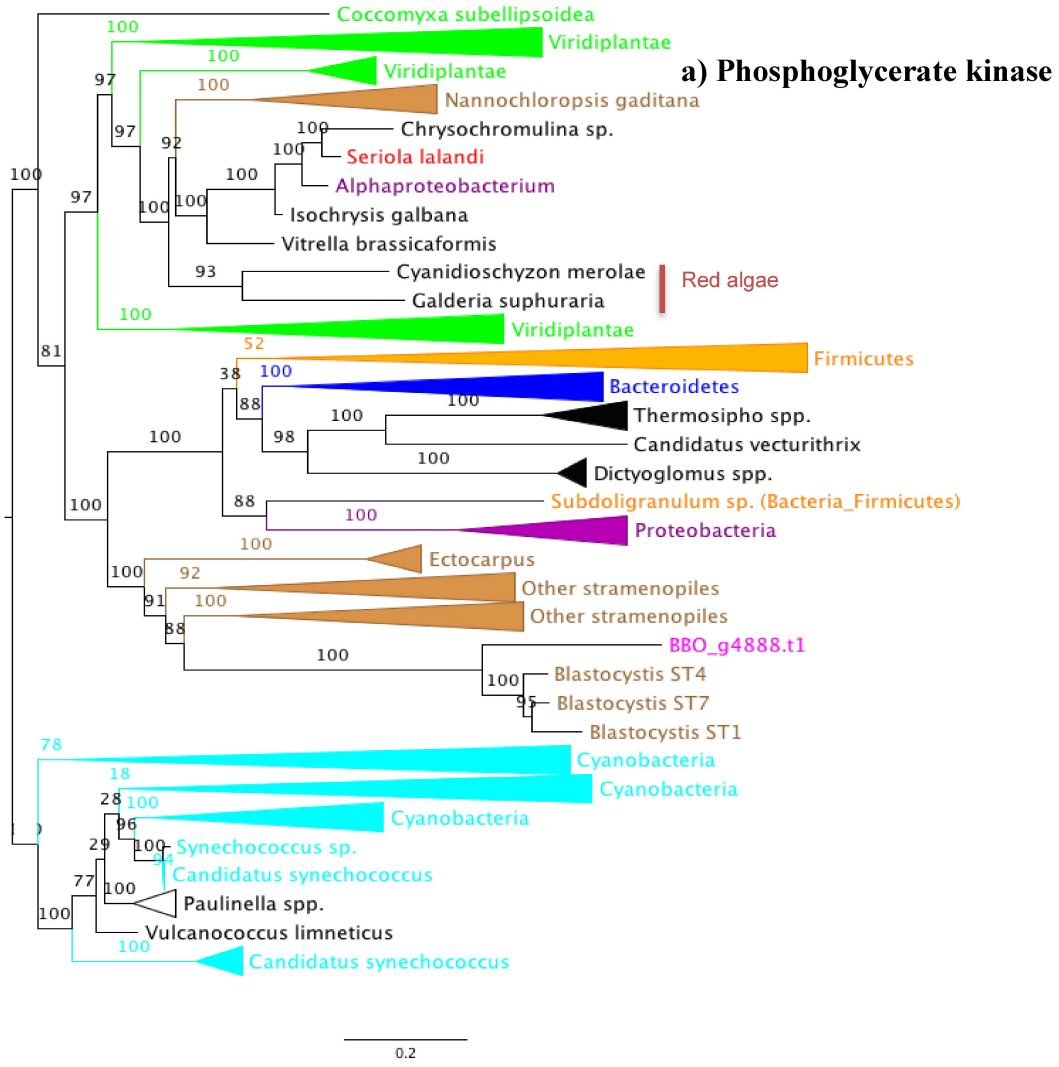
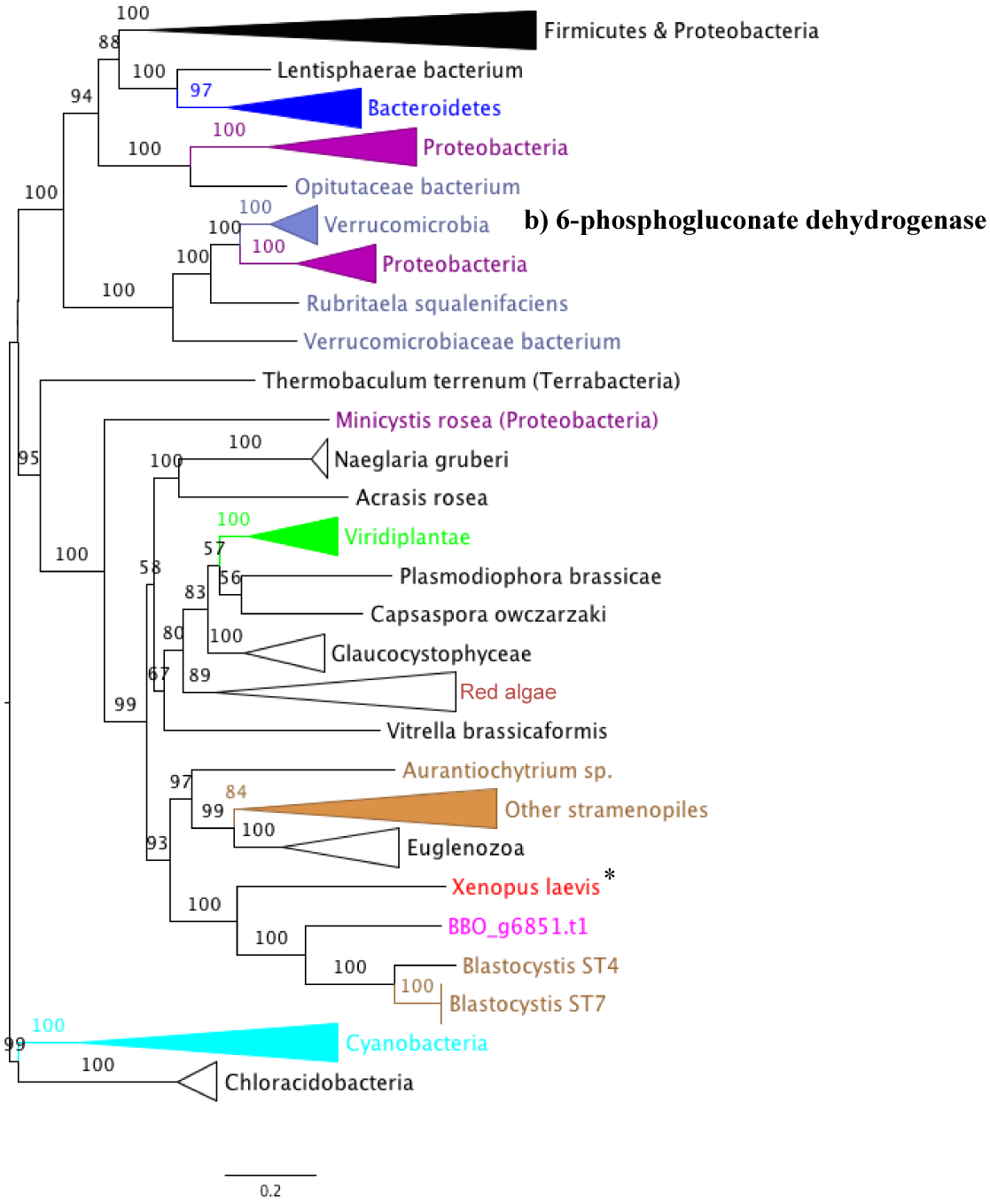

RNA reads
BBO genome sequence
End of BBO gene
(see 1st row translation)
3-frame forward translation

Figure S11: Single gene trees of a) phosphoglycerate kinase and b) 6-phosphogluconate dehydrogenase. Maximum-likelihood tree by IQ-TREE, substitution model LG4X (Le et al., 2012), 1000-replicate ultrafast bootstrap approximation. Cyan: cyanobacteria, purple: Proteobacteria, orange: Firmicutes, blue: Bacteroidetes, dark blue: Verrucomicrobia, red: metazoa, green: viridiplantae, brown: stramenopiles, magenta: BBO, black: other taxa. Although these trees are not rooted, previous studies have established that plastids in Viridiplantae, also known as Chloroplastida, originated from primary endosymbiosis of cyanobacteria (Kim & Archibald, 2009), and in turn, plastids in photosynthetic stramenopiles originated from secondary (or higher) endosymbiosis of red algae (Dorell & Bowler, 2017), and thus the trees were rooted between the cyanobacterial clade and the rest of taxa.





b) 6-phosphogluconate dehydrogenase

*This African clawed frog sequence may be a contaminant as no other metazoan sequence are present in this tree. It is likely a sequence from an opalinatan that lived in this frog's gut.

Figure S12: **a)** Sequence alignment of VatC cDNA RT-PCR product from Sanger sequencing against its RNA-seq counterpart. Screenshot from a Sequencher® v5.4.6 (Gene Corps Corporation, 2017) project. The top underlined sequence: reference gene, three sequences in centre: fragments of the PCR product (low quality bases trimmed), bottom nucleotide sequence: consensus sequence with its amino acid translation shown below (in the same frame as the predicted VatC gene). Lighter shades of blue indicate better base call quality. Primer directions shown by orange arrows. Blue dots at the end of the consensus sequence represent introns. **b)** Gel image of PCR products shown on the right. +λ: Lambda DNA for PCR positive control (1.1kb), -λ: negative control for PCR (water instead of DNA), -RNA: RT-PCR product without reverse transcriptase, to ensure there was no DNA carried over from before reverse transcription. U.L: Ultra Low Range DNA Ladder (ThermoFisher cat.: 10597012), 1kb: GeneRuler 1kb DNA Ladder (ThermoFisher cat.: SM1331), Pair1: PCR product from first pair of VatC primers (142bp), Pair2: PCR product from second pair of VatC primers (179bp). U.L. ladder sizes on left, 1kb ladder sizes on right.

a)
PCR sequences
Forward1 primer
Reverse primer
Intron
b)
U.L.
Pair1
+λ
-λ
-RNA
U.L.
1kb
Pair2
20kb
1.5kb
300bp
100bp
10bp

Figure S13: 3D model prediction of VatC using SWISS-MODEL (swissmodel.expasy.org). The amino acid sequence alignment on the right shows the BBO sequence ("Model_01") aligned to its homolog from the the yeast *Saccharomyces cerevisiae* S288C ("5voy.1.O"; PDBe database ID: 5voy). Blue: hydrophilic, red: hydrophobic. The degree of mismatch is represented by faded colours. The yellow square highlights the surface loop containing the polyasparagine tract.

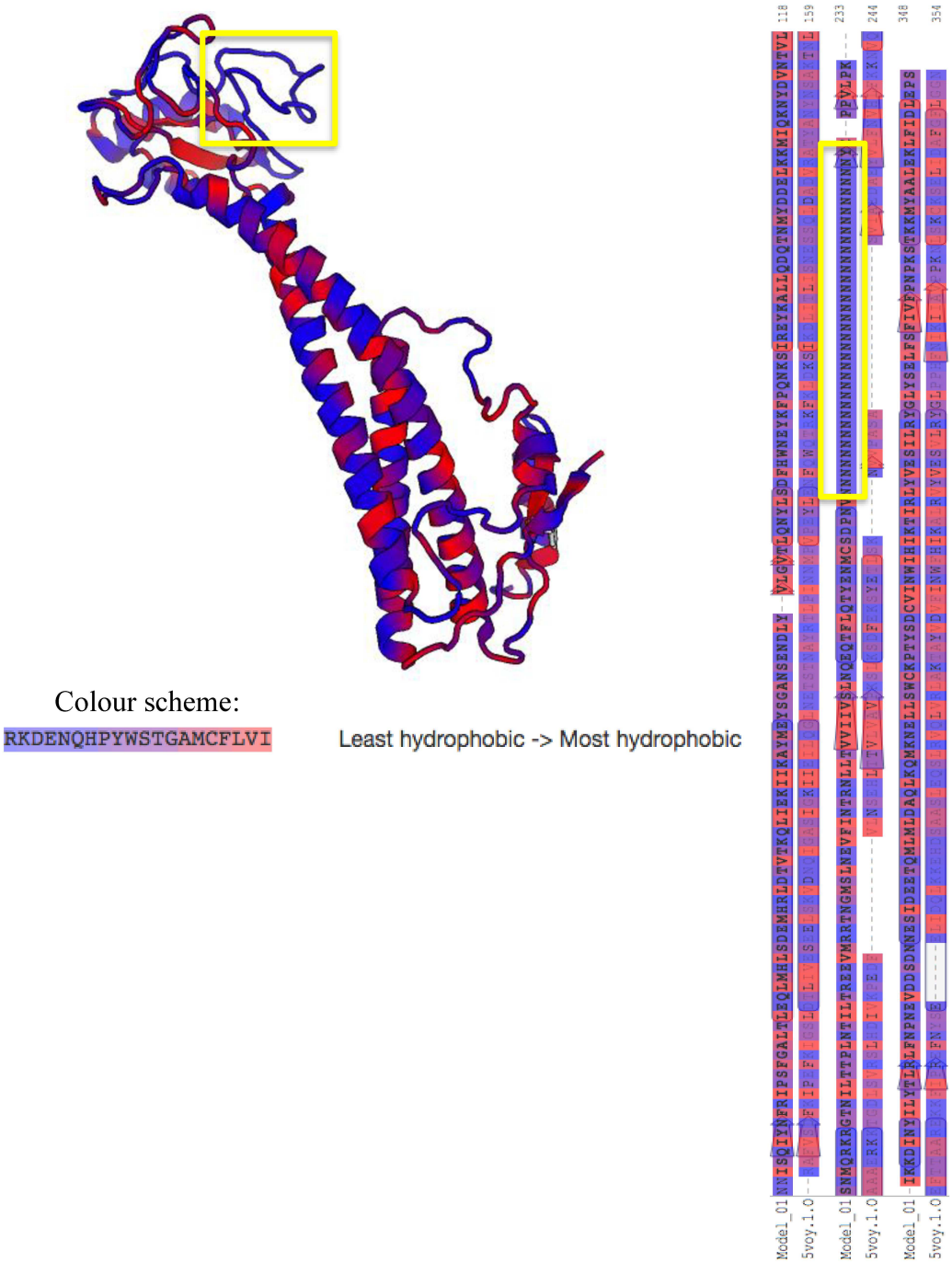


Figure S14: Part of the protein alignment of homologs of the VatC gene using MAFFT-einsi (Katoh et al., 2017). Only the BBO sequence contains the polyasparagine tract.

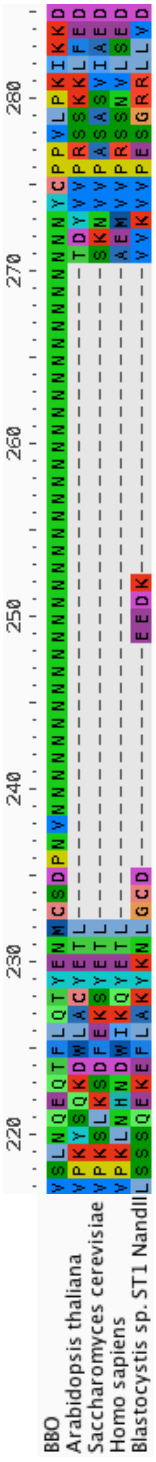
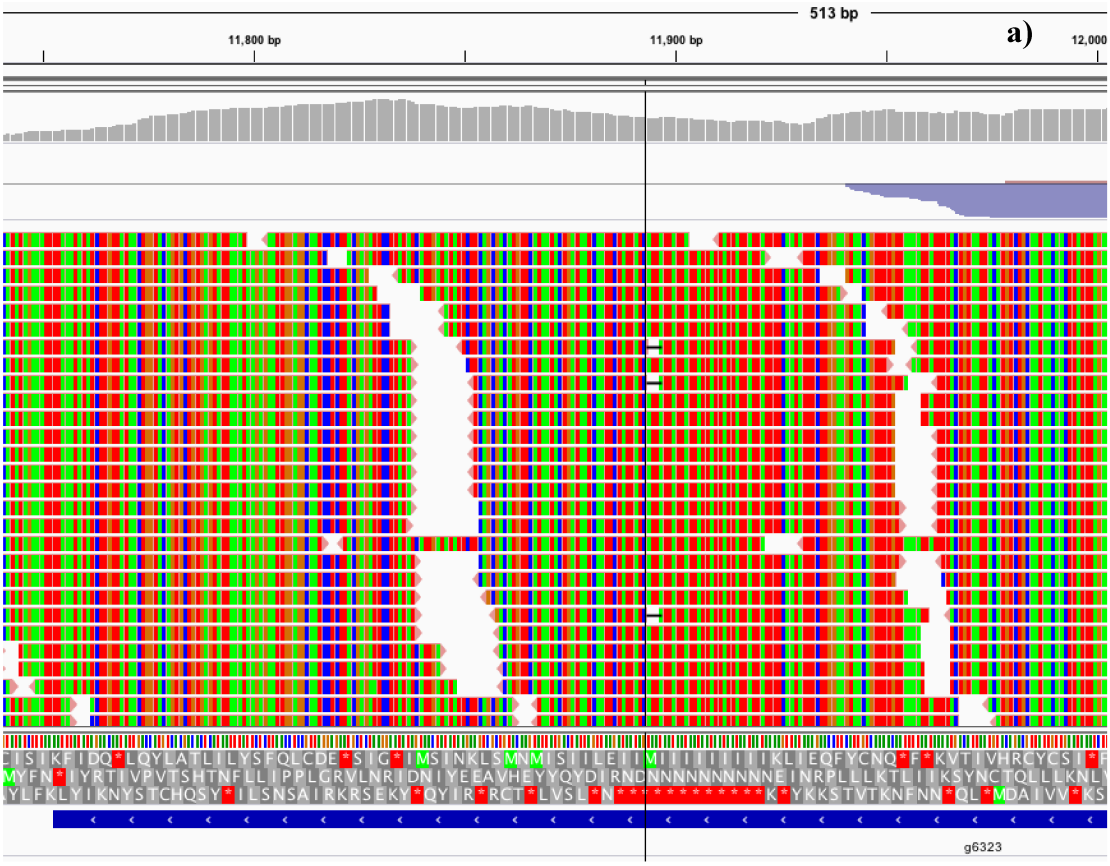


Figure S15: **a)** RNA coverage and **b)** the best BLASTp result for arginine transporter Can1 gene. The vertical black line marks the end of the 10-mer asparagine stretch (in the second frame translation, second line from the bottom).



arginine transporter Can1 [Blastocystis sp. ATCC 50177/Nand II]

b)

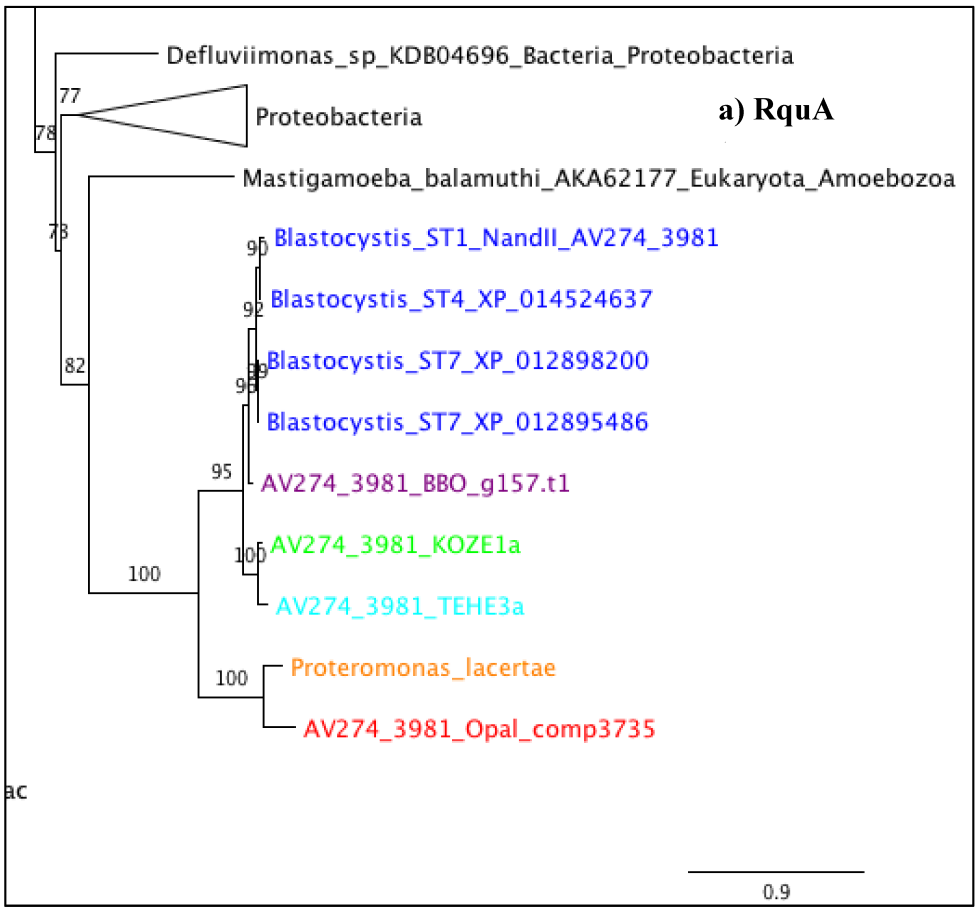
Sequence ID: OAO18008.1 **Length:** 830 **Number of Matches:** 1

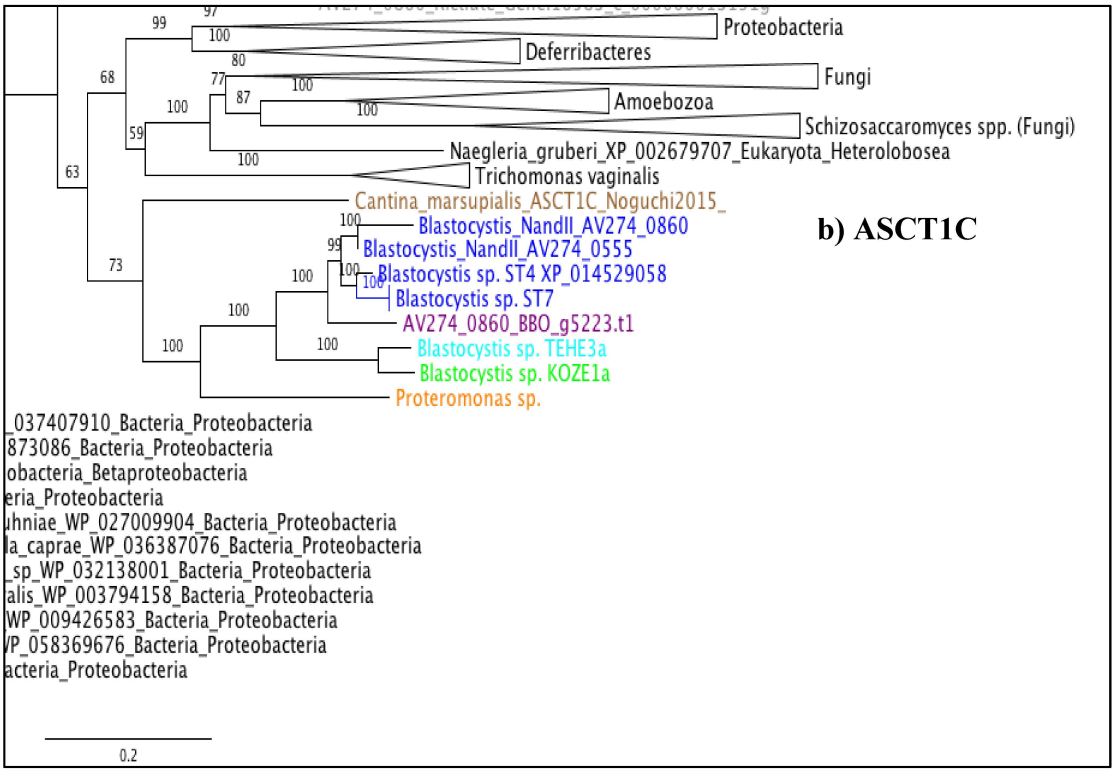
Range 1: 553 to 830 GenPept Graphics ▼ Next Match ▲ Previous Match

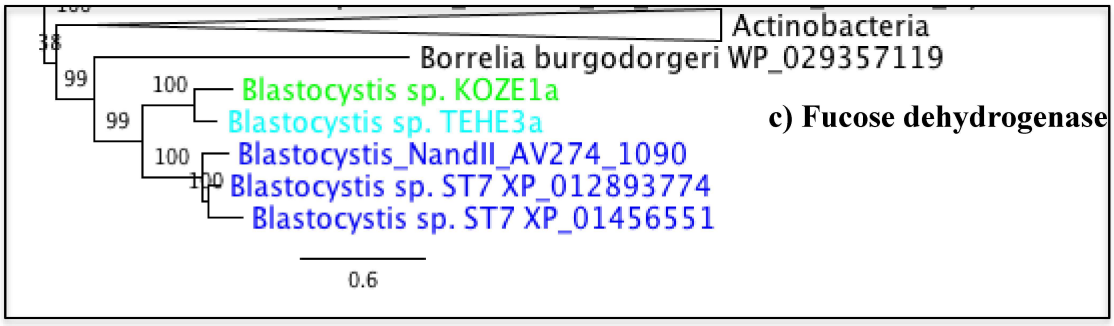
Score	Expect	Method	Identities	Positives	Gaps
188 bits(478)	2e-50	Compositional matrix adjust.	114/304(38%)	179/304(58%)	27/304(8%)

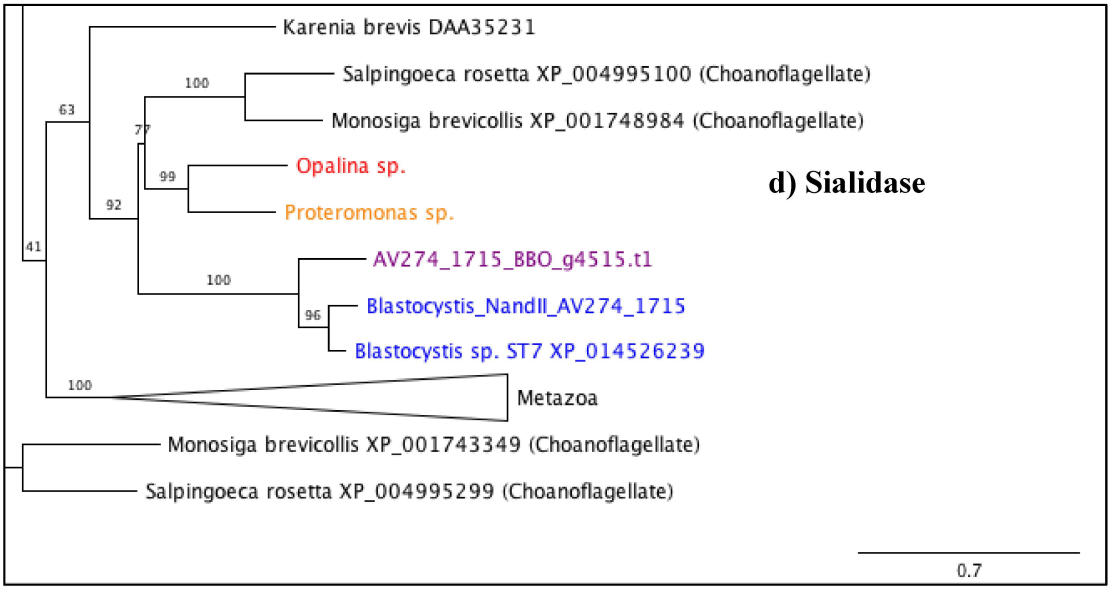
```
Query  2    GLQNLITNVGIGALKPNTVVIPLEDEQDCDCDHRSVTSMELLGSIKGVTLDDAANISMPF  61
            GL +++   G+GAL+PNT+V+PLE  +  +       +S+E+L  ++G+   +   +S+P
Sbjct  553  GLHDMVATCGLGALRPNTIVLPLERARGREVS----SSVEILHRLRGMAEAELPLVSVPL  608

Query  62   PCKLTGHDYLHIIYDSLKLKRNVAVAANYEKISDRTTKLVWEGVKENKKALEGEFIDMVI  121
            P  L+  ++  ++ D L L++N+ VAANYEKI    T+L W  + E +   E   IDMVI
Sbjct  609  PSMLSDEEFCRMVNDVLVLQKNIVVAANYEKIQPHGTELTWAAMGEVRAHPELHTIDMVI  668

Query  122  VGDWTWDIEFKNYLSVILQYGYIINRNKVWNKYPLRIIQVVYDKPDNYDTKEEEQRLDDL  181
            VG W W++EF NYLSVILQYGYII+RNK W  + LR++Q++ D+ +  + +EE+++L++L
Sbjct  669  VGHWDWNVEFSNYLSVILQYGYIIHRNKNWQGFTLRLLQILPDREEGENLEEEKKKLEEL  728

Query  182  LEGVRINAKKLVIQLPA-ISIKFTVHKETSFEELSLNEKCLVLNKLLLQTCNYSKIILTK  240
            LE VR+ A  +V  +P  +   +  ++ S   L  +    V+N LL + C  S++IL K
Sbjct  729  LEEVRVEATTVVRGVPQHVFDPYREREDFSVHVLPFDVLAAVVNPLLKELCPASEMILMK  788

Query  241  LLLPRNIENNNNNNNNNNDNRIDYQYYEHVAEEYINDIRNLVRGLPPILLFNTHSTVPVI  300
            L  P                        E  A EY+  +R LV+ LPP+LL N+ + +PVI
Sbjct  789  LPSPE-----------------------EGKAGEYMKALRALVKDLPPMLLMNSSTHLPVI  826

Query  301  TRYI  304
            TRYI
Sbjct  827  TRYI  830
```

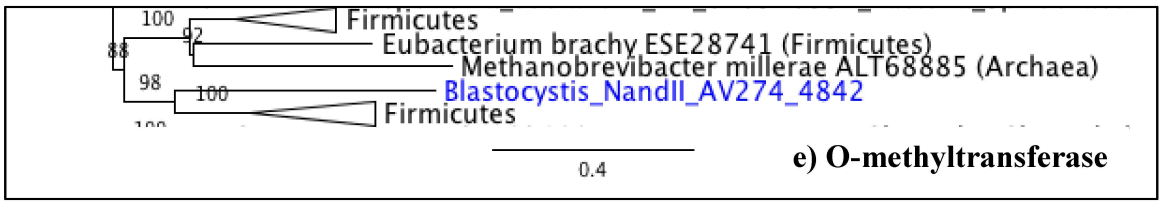
Figure S16: Close-up views of the phylogentic trees of each gene in the LGT analysis. Constructed by maximum-likelihood estimation using IQ-TREE with the substitution model LG4X (Le et al., 2012) with a 1000-replicate ultrafast bootstrap approximation for branch support (Minh et al., 2013). a) RquA b) ASCT1C c) Fucose dehydrogenase d) Sialidase e) O-methyltransferase f) Glycosyltransferase Gtd1 g) SufCB. Taxa colours: Red: *Opalina* sp., orange: *Proteromonas* sp., green: *Blastocystis* sp. KOZE1a, cyan: *Blastocystis* sp. TEHE3a, purple: BBO, blue: other *Blastocystis* spp., yellow: *Cafeteria* sp., grey: *Rictus lutensis,* brown: *Cantina marsupialis*, pink: *Halocafeteria seosinensis*.



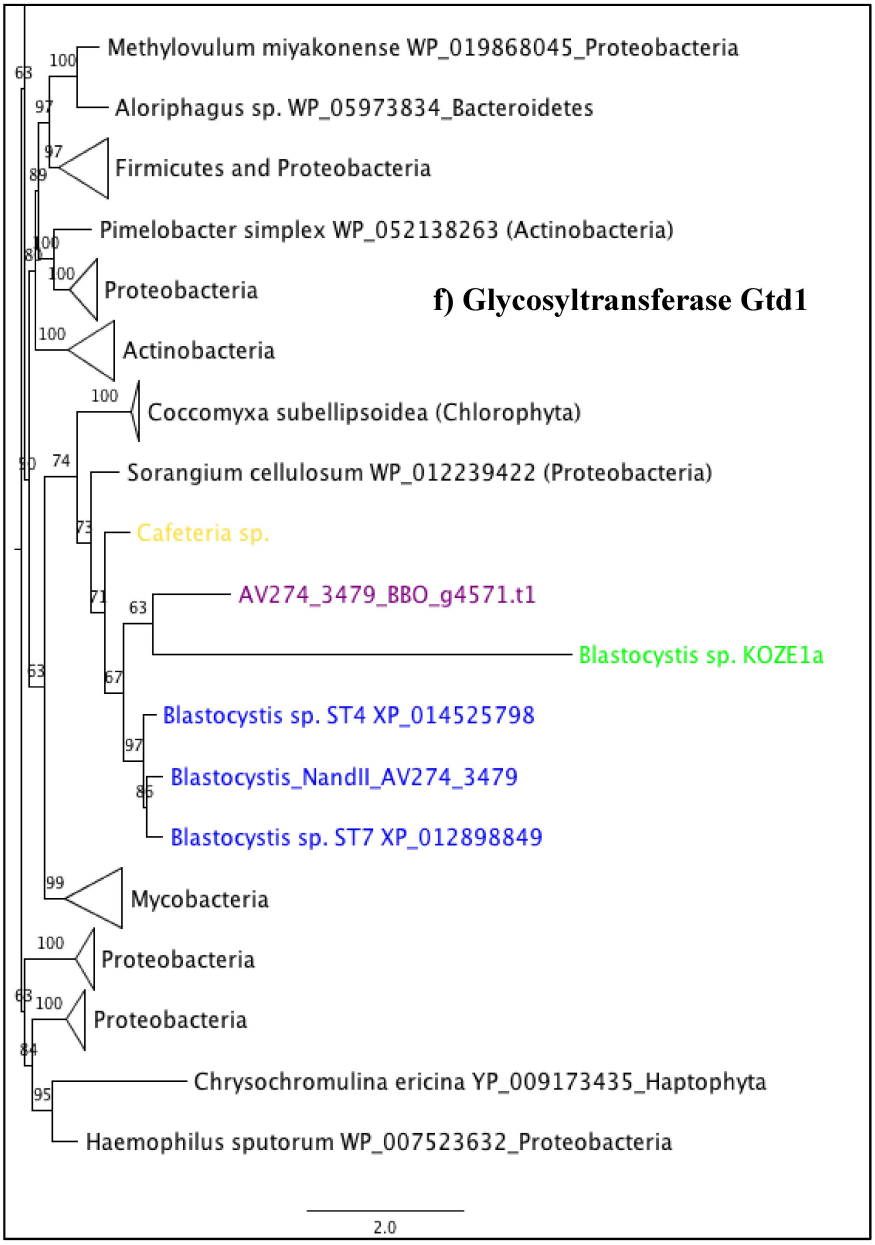

Proteobacteria
Deferribacteres
Fungi
Amoebozoa
Schizosaccaromyces spp. (Fungi)
Naegleria_gruberi_XP_002679707_Eukaryota_Heterolobosea
Trichomonas vaginalis
Cantina_marsupialis_ASCT1C_Noguchi2015_
Blastocystis_NandII_AV274_0860
Blastocystis_NandII_AV274_0555
Blastocystis sp. ST4 XP_014529058
Blastocystis sp. ST7
AV274_0860_BBO_g5223.t1
Blastocystis sp. TEHE3a
Blastocystis sp. KOZE1a
Proteromonas sp.
b) ASCT1C
_037407910_Bacteria_Proteobacteria
873086_Bacteria_Proteobacteria
obacteria_Betaproteobacteria
eria_Proteobacteria
uhniae_WP_027009904_Bacteria_Proteobacteria
la_caprae_WP_036387076_Bacteria_Proteobacteria
_sp_WP_032138001_Bacteria_Proteobacteria
alis_WP_003794158_Bacteria_Proteobacteria
WP_009426583_Bacteria_Proteobacteria
/P_058369676_Bacteria_Proteobacteria
acteria_Proteobacteria
0.2


Actinobacteria
Borrelia burgodorgeri WP_029357119
Blastocystis sp. KOZE1a
Blastocystis sp. TEHE3a
Blastocystis_NandII_AV274_1090
Blastocystis sp. ST7 XP_012893774
Blastocystis sp. ST7 XP_01456551
c) Fucose dehydrogenase
0.6

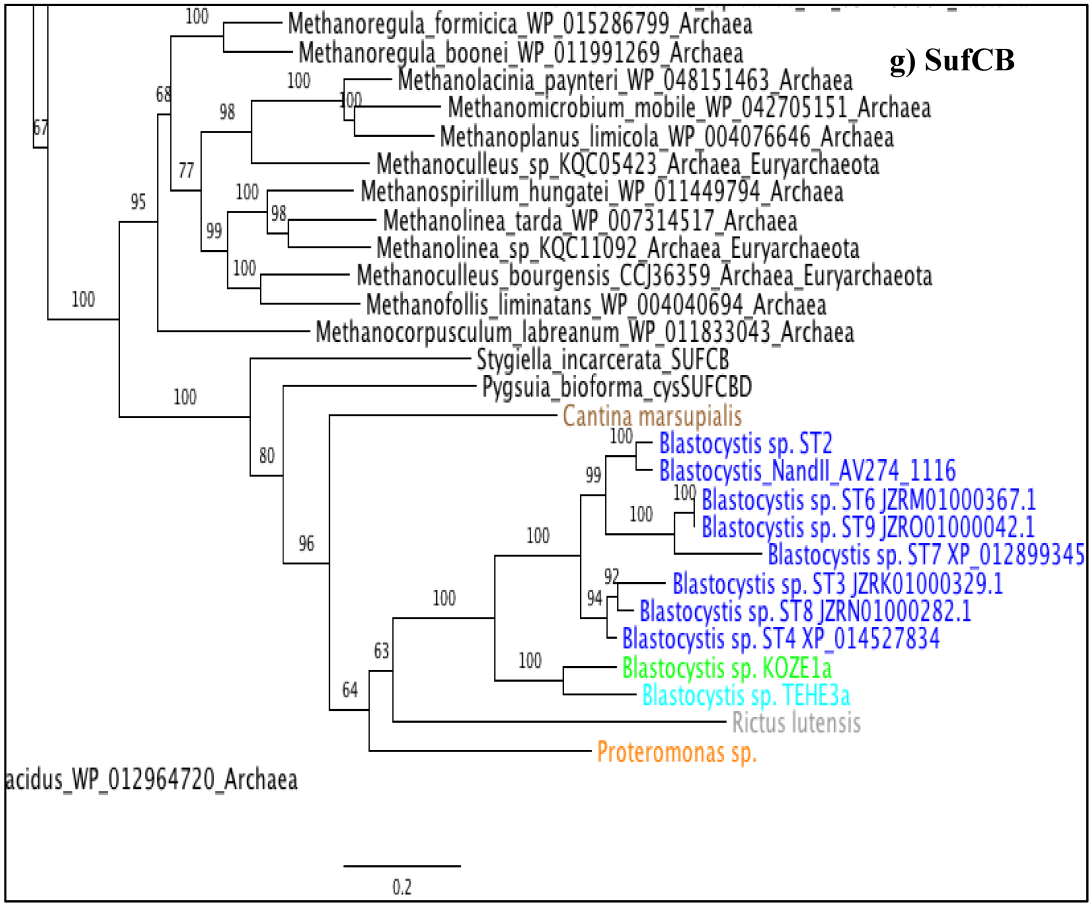

Karenia brevis DAA35231
Salpingoeca rosetta XP_004995100 (Choanoflagellate)
Monosiga brevicollis XP_001748984 (Choanoflagellate)
Opalina sp.
Proteromonas sp.
AV274_1715_BBO_g4515.t1
Blastocystis_NandII_AV274_1715
Blastocystis sp. ST7 XP_014526239
Metazoa
Monosiga brevicollis XP_001743349 (Choanoflagellate)
Salpingoeca rosetta XP_004995299 (Choanoflagellate)
d) Sialidase
0.7



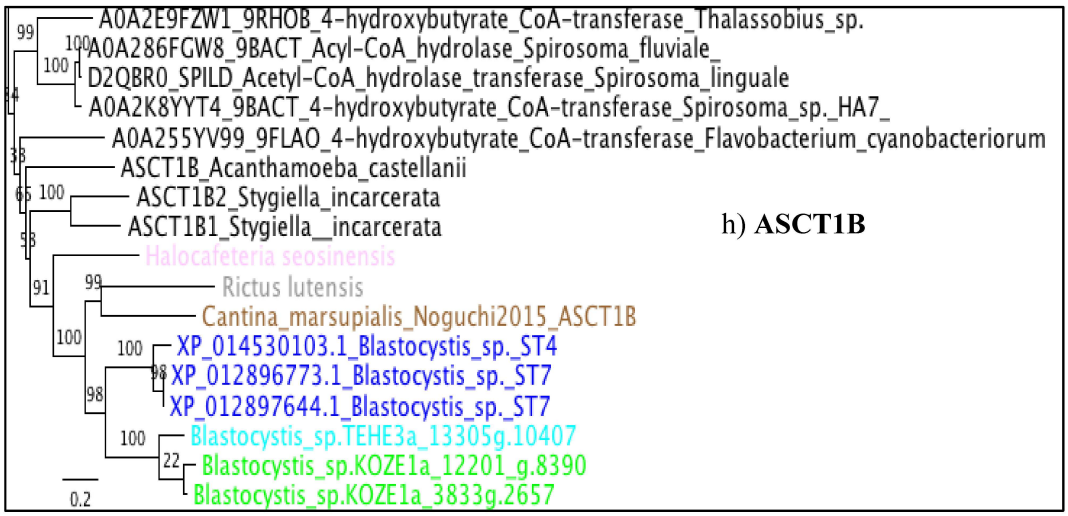
e) O-methyltransferase



f) Glycosyltransferase Gtd1

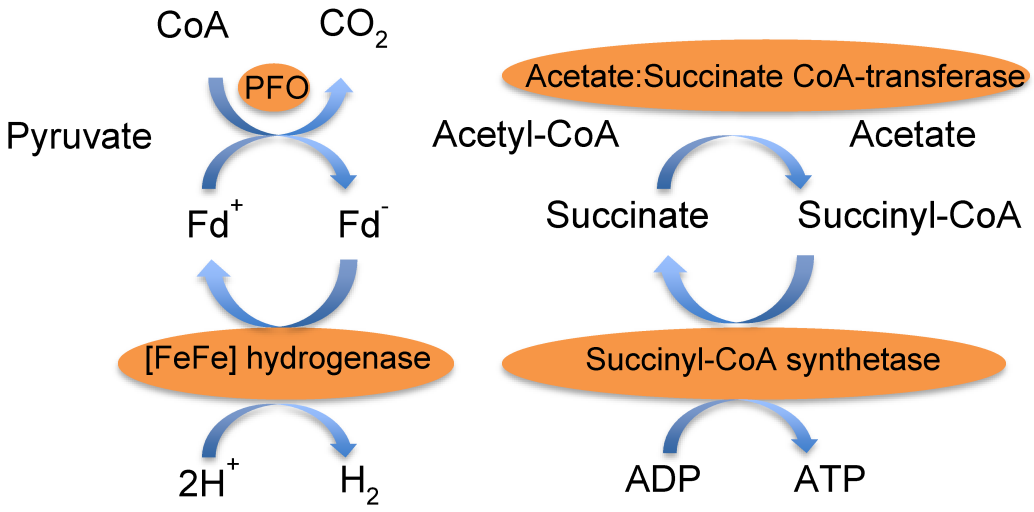


*The *Rictus lutensis* sequence used for this analysis was a partial gene, containing only the SufB unit.


A0A2E9FZW1_9RHOB_4-hydroxybutyrate_CoA-transferase_Thalassobius_sp.
A0A286FGW8_9BACT_Acyl-CoA_hydrolase_Spirosoma_fluviale_
D2QBR0_SPILD_Acetyl-CoA_hydrolase_transferase_Spirosoma_linguale
A0A2K8YYT4_9BACT_4-hydroxybutyrate_CoA-transferase_Spirosoma_sp._HA7_
A0A255YV99_9FLAO_4-hydroxybutyrate_CoA-transferase_Flavobacterium_cyanobacteriorum
ASCT1B_Acanthamoeba_castellanii
ASCT1B2_Stygiella_incarcerata
ASCT1B1_Stygiella_incarcerata
Halocafeteria seosinensis
Rictus lutensis
Cantina_marsupialis_Noguchi2015_ASCT1B
XP_014530103.1_Blastocystis_sp._ST4
XP_012896773.1_Blastocystis_sp._ST7
XP_012897644.1_Blastocystis_sp._ST7
Blastocystis_sp.TEHE3a_13305g.10407
Blastocystis_sp.KOZE1a_12201_g.8390
Blastocystis_sp.KOZE1a_3833g.2657
0.2

h) **ASCT1B**

Figure S17: A schematic representation of the anaerobic ATP production pathway in *Blastocystis* spp. Enzymes are represented in orange. All four homologs were found in the BBO nuclear genome and predicted to possess mitochondrial-targeting peptides.



APPENDIX C – COPYRIGHT PERMISSIONS

ELSEVIER LICENSE
TERMS AND CONDITIONS

Apr 26, 2018

This Agreement between Ms. Sarah Shah ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	4335980731337
License date	Apr 25, 2018
Licensed Content Publisher	Elsevier
Licensed Content Publication	Elsevier Books
Licensed Content Title	Encyclopedia of Evolutionary Biology
Licensed Content Author	A.G.B. Simpson,Y. Eglit
Licensed Content Date	Jan 1, 2016
Licensed Content Pages	17
Start Page	344
End Page	360
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables /illustrations	1
Format	electronic
Are you the author of this Elsevier chapter?	No
Will you be translating?	No
Original figure numbers	Figure 3
Title of your thesis/dissertation	Genome of Blastocystis sp. isolated from the cockroach Blatta orientalis
Expected completion date	Oct 2018

full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.

9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.

10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.

12. No Amendment Except in Writing: This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).

13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

14. Revocation: Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.

LIMITED LICENSE

The following terms and conditions apply only to specific license types:

15. **Translation**: This permission is granted for non-exclusive world **English** rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article.

16. **Posting licensed content on any Website**: The following terms and conditions apply as follows: Licensing material from an Elsevier journal: All content posted to the web site must maintain the copyright information line on the bottom of each image; A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx or the Elsevier homepage for books at

http://www.elsevier.com; Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.
Licensing material from an Elsevier book: A hyper-text link must be included to the Elsevier homepage at http://www.elsevier.com . All content posted to the web site must maintain the copyright information line on the bottom of each image.

Posting licensed content on Electronic reserve: In addition to the above the following clauses are applicable: The web site must be password-protected and made available only to bona fide students registered on a relevant course. This permission is granted for 1 year only. You may obtain a new license for future website posting.
17. **For journal authors:** the following clauses are applicable in addition to the above:
Preprints:
A preprint is an author's own write-up of research results and analysis, it has not been peer-reviewed, nor has it had any other value added to it by a publisher (such as formatting, copyright, technical enhancement etc.).
Authors can share their preprints anywhere at any time. Preprints should not be added to or enhanced in any way in order to appear more like, or to substitute for, the final versions of articles however authors can update their preprints on arXiv or RePEc with their Accepted Author Manuscript (see below).
If accepted for publication, we encourage authors to link from the preprint to their formal publication via its DOI. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help users to find, access, cite and use the best available version. Please note that Cell Press, The Lancet and some society-owned have different preprint policies. Information on these policies is available on the journal homepage.
Accepted Author Manuscripts: An accepted author manuscript is the manuscript of an article that has been accepted for publication and which typically includes author-incorporated changes suggested during submission, peer review and editor-author communications.
Authors can share their accepted author manuscript:

- immediately
  - via their non-commercial person homepage or blog
  - by updating a preprint in arXiv or RePEc with the accepted manuscript
  - via their research institute or institutional repository for internal institutional uses or as part of an invitation-only research collaboration work-group
  - directly by providing copies to their students or to research collaborators for their personal use
  - for private scholarly sharing as part of an invitation-only work group on commercial sites with which Elsevier has an agreement
- After the embargo period
  - via non-commercial hosting platforms such as their institutional repository
  - via commercial sites with which Elsevier has an agreement

In all cases accepted manuscripts should:

- link to the formal publication via its DOI
- bear a CC-BY-NC-ND license - this is easy to do

- if aggregated with other manuscripts, for example in a repository or other site, be shared in alignment with our hosting policy not be added to or enhanced in any way to appear more like, or to substitute for, the published journal article.

Published journal article (JPA): A published journal article (PJA) is the definitive final record of published research that appears or will appear in the journal and embodies all value-adding publishing activities including peer review co-ordination, copy-editing, formatting, (if relevant) pagination and online enrichment.
Policies for sharing publishing journal articles differ for subscription and gold open access articles:
Subscription Articles: If you are an author, please share a link to your article rather than the full-text. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help your users to find, access, cite, and use the best available version.
Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.
If you are affiliated with a library that subscribes to ScienceDirect you have additional private sharing rights for others' research accessed under that agreement. This includes use for classroom teaching and internal training at the institution (including use in course packs and courseware programs), and inclusion of the article for grant funding purposes.
Gold Open Access Articles: May be shared according to the author-selected end-user license and should contain a CrossMark logo, the end user license, and a DOI link to the formal publication on ScienceDirect.
Please refer to Elsevier's posting policy for further information.
18. **For book authors** the following clauses are applicable in addition to the above:
Authors are permitted to place a brief summary of their work online only. You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version. **Posting to a repository:** Authors are permitted to post a summary of their chapter only in their institution's repository.
19. **Thesis/Dissertation**: If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for Proquest/UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission. Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

Elsevier Open Access Terms and Conditions
You can publish open access with Elsevier in hundreds of open access journals or in nearly 2000 established subscription journals that support open access publishing. Permitted third party re-use of these open access articles is defined by the author's choice of Creative Commons user license. See our open access license policy for more information.
Terms & Conditions applicable to all Open Access articles published with Elsevier:
Any reuse of the article must not represent the author as endorsing the adaptation of the article nor should the article be modified in such a way as to damage the author's honour or reputation. If any changes have been made, such changes must be clearly indicated.

REFERENCES

Abe, N. (2004). Molecular and phylogenetic analysis of *Blastocystis* isolates from various hosts. *Vet. parasitol., 120*, 235-242. doi:10.1016/j.vetpar.2004.01.003

Adams, K., Qiu, Y., Stoutemyer, M., & Palmer, J. (2002). Punctuated Evolution of Mitochondrial Gene Content: High and Variable Rates of Mitochondrial Gene Loss and Transfer to the Nucleus during Angiosperm Evolution. *PNAS, 99*(15), 9905-9912. doi:10.1073/pnas.042694899

Adl, S. M., Simpson, A. G. B., Lane, C. E., Lukeš, J., Bass, D., Bowser, S. S., . . . Spiegel, F. W. (2012). The Revised Classification of Eukaryotes. *Journal of Eukaryotic Microbiology, 59*(5), 429-514.

Ajjampur, S. S. R., & Tan, S. W. T. (2016). Pathogenic mechanisms in *Blastocystis* spp. - Interpreting results from in vitro and in vivo studies. *Parasitology International, 65*(6), 772-779. doi:10.1016/j.parint.2016.05.007

Albalat, R. & Cañestro, C. (2016). Evolution by gene loss. *Nat. Rev. Genet. 17*, 379-391. doi:10.1038/nrg.2016.39

Alexieff A (1911). "Sur la nature des formations dites "kystes de Trichomonas intestinalis"". *CR Soc Biol. 71*, 296–298.

Alfellani, M., Taner-Mulla, D., Jacob, A., Imeede, C., Yoshikawa, H., Stensvold, C., & Clark, C. (2013). Genetic Diversity of *Blastocystis* in Livestock and Zoo Animals. *Protist*, *140*(8), 497-509. doi:10.1016/j.protis.2013.05.003

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*(3), 430-410. doi:10.1016/S0022-2836(05)80360-2

Anderson, S., Bankier, A. T., Barrell, B. G., De Bruijn, M. H. L., Coulson, A. R., Drouin, J., . . . Young, I. G. (1981). Sequence and organization of the human mitochondrial genome. *Nature, 290*(5806), 457-65.

Armengaud, J., Pible, O., Gaillard, J., Cian, A., Gantois, N., Tan, K., . . . Viscogliosi, E. (2017). Proteogenomic Insights into the Intestinal Parasite *Blastocystis* sp. Subtype 4 Isolate WR1. *PROTEOMICS, 17*(21). doi:10.1002/pmic.201700211

Arnett, R. H., Jr. (2000). Dictyoptera. In *American Insects: A Handbook of the Insects of America North of Mexico* (2nd ed., p. 195). Boca Raton, Florida: CRC Press.

Artavanis-Tsakonas, K., Misaghi, S., Comeaux, C. A., Catic, A., Spooner, E., Duraisingh, M. T., & Ploegh, H. L. (2006). Identification by functional proteomics of a deubiquitinating/deNeddylating enzyme in *Plasmodium falciparum. Molecular Microbiology, 61*(5), 1187-1195. doi:10.1111/j.1365-2958.2006.05307.x

Aury, J.-C., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B. M., . . . Wincker, P. (2006). Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature, 444*(7116), 171-178. doi:10.1038/nature05230

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A., Dvorkin, M., Kulikov, A. S., . . . Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology, 19*(5), 455-477. doi:10.1089/cmb.2012.0021

Beghini, F., Pasolli, E., Truong, T. D., Putignani, L., Cacciò, S. M., & Segata N. (2017). Large-scale comparative metagenomics of *Blastocystis*, a common member of the human gut microbiome. *The ISME Journal, 11*(12), 2848-2863. doi:10.1038/ismej.2017.139

Beck, N. & Lang, B. (2010). MFannot, organelle genome annotation webserver. http://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl.

Bennetzen, J. L. & Hall, B. D. (1982). Codon selection in yeast. *J. Biol. Chem., 257*, 295-307.

Betts, E. L., Gentekaki, E., Thomasz, A., Breakell, V., Carpenter, A. I., Tsaousis, A. D. (2017). Genetic diversity of *Blastocystis* in non-primate animals. Parasitology. doi:10.1017/S0031182017002347

Biscotti, M., Olmo, A., & Heslop-Harrison, E. (2015). Repetitive DNA in eukaryotic genomes. *Chromosome Research, 23*(3), 415-420. doi:10.1007/s10577-015-9499-z

Bignell, D. E. (1982). Nutrition and digestion. In W. D. Bell & K. G. Adiyodi (Eds.), *The American Cockroach* (pp. 57-86). New York, NY: Chapma and Hall.

Birky, C. W. (2001). The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. *Annu. Rev. Genet*., *35*, 125–148. doi:10.1146/annurev.genet.35.102401.090231

Bolger, A., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120. doi:10.1093/bioinformatics/btu170

Bouchier, C., Ma, L., Créno, S., Dujon, B., & Fairhead, C. (2009). Complete mitochondrial genome sequences of three *Nakaseomyces* species reveal invasion by palindromic GC clusters and considerable size expansion. *FEMS Yeast Research*, *9*(8), 1283-1292. doi:10.1111/j.1567-1364.2009.00551.x

Breton, C., Fournel-Gigleux, S., & Palcic, M. M. (2012). Recent structures, evolution and mechanisms of glycosyltransferases. *Curr. Opin. Struct. Biol., 22*(5), 540-549. doi:10.1016/j.sbi.2012.06.007

Brown, M., Heiss, A., Kamikawa, R., Inagaki, Y., Yabuki, A., Tice, A., . . . Roger, A. (2018). Phylogenomics Places Orphan Protistan Lineages in a Novel Eukaryotic Super-Group. *Genome Biology and Evolution, 10*(2), 427-433. doi:10.1093/gbe/evy014

Brugerolle, G., & König, H. (1997). Ultrastructure and organization of the cytoskeleton in Oxymonas, an intestinal flagellate of termites. *J. Eukaryot. Microbiol. 44*(11), 305–313. doi:10.1111/j.1550-7408.1997.tb05671.x

Brumpt, E. (1912). "*Blastocystis hominis* n. sp. et formes voisines". *Bulletin of the Exotic Pathology Society, 5*, 725–30.

Buchan, D. W. A., Minneci, F., Nugent, T. C. O., Bryson, K., & Jones, D. T. (2013). Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Research, 41* (W1), W340-W348.

Buchfink, B., Xie, C., & Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nature Methods, 12*(1), 59-60. doi:10.1038/nmeth.3176

Canavan C, West J, & Card T. (2014). The epidemiology of irritable bowel syndrome. *Clinical Epidemiology, 2014*, 71-80. https://doaj.org/article/16584cc4bca34f8c940ca59add994b2c

Capella-Gutiérrez, S., Silla-Martínez, J., & Gabaldón, T. (2009). TrimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics, 25*(15), 1972-1973. doi:10.1093/bioinformatics/btp348

Carbajal, J. A., Castillo, L. D., Lanuza, M. D., Villar, J., & Borrás, R. (1997). Karyotypic diversity among *Blastocystis hominis* isolates. *International Journal for Parasitology, 27*(8), 941-945. doi:10.1016/S0020-7519(97)00042-8

Carlin, A., Uchiyama, S., Chang, Y., Lewis, A., Nizet, V., & Varki, A. (2009). Molecular mimicry of host sialylated glycans allows a bacterial pathogen to engage neutrophil Siglec-9 and dampen the innate immune response. *Blood, 113*(14), 3333-6. doi:10.1182/blood-2008-11-187302

Caron, D. A. (2000). Symbiosis and mixotrophy among pelagic microorganisms. In D.L. Kirchman (Ed.), *Microbial ecology of the oceans* (pp. 495-523). New York: John Wiley & Sons, Inc.

Cavalier-Smith, T., & Scoble, J. M. (2013). Phylogeny of Heterokonta: Incisomonas marina, a uniciliate gliding opalozoan related to Solenicola (Nanomonadea), and evidence that Actinophryida evolved from raphidophytes. *European Journal of Protistology, 49*(3), 328-353. doi:10.1016/j.ejop.2012.09.002

Chandramathi, S., Suresh, K., Shuba, S., Mahmood, A., & Kuppusamy, U. R. (2010). High levels of oxidative stress in rats infected with *Blastocystis hominis*. *Parasitology, 137*(4), 605-611. doi:10.1017/S0031182009991351

Chawla, M., Oliva, R., Bujnicki, J., & Cavallo, L. (2015). An atlas of RNA base pairs involving modified nucleobases with optimal geometries and accurate energies. *Nucleic Acids Research, 43*(19), 9573. doi: 10.1093/nar/gkv606

Chen, Y.-C., Liu, T., Yu, C.-H., Chiang, T.-Y., & hwang, C.-C. (2013). Effects of GC Bias in Next-Generation-Sequencing Data on *De Novo* Genome Assembly. *PLoS ONE, 8*(4), E62856. doi:10.1371/journal.pone.0062856

Chang, J. H., & Tong, L. (2012). Mitochondrial poly(A) polymerase and polyadenylation. *Biochim Biophys Acta., 1819*, 992–997. doi:10.1016/j.bbagrm.2011.10.012

Chen, W., Liu, F., Ling, Z., Tong, X., Xiang, C., & Moschetta, A. (2012). Human Intestinal Lumen and Mucosa-Associated Microbiota in Patients with Colorectal Cancer (Colorectal Cancer Microbiota). *PLoS ONE, 7*(6), E39743. doi:10.1371/journal.pone.0039743

Chiari, Y., Dion, K., Colborn, J., Parmakelis, A., & Powell, J. R. (2010). On the possible role of tRNA base modification in the evolution of codon usage: queosine and Drosophila. J. Mol. Evol., 70(4), 339-345. doi:10.1007/s00239-010-9329-z

Chiti, F., & Dobson, C. (2006). Protein Misfolding, Functional Amyloid, and Human Disease. *Annual Review of Biochemistry, 75*, 333-366. doi:10.1146/annurev.biochem.75.101304.123901

Cian, A., El Safadi, D., Osman, M., Moriniere, R., Gantois, N., Benamrouz-Vanneste, S., . . . Viscogliosi, E. (2017). Molecular Epidemiology of *Blastocystis* sp. in Various Animal Groups from Two French Zoos and Evaluation of Potential Zoonotic Risk. *PLoS ONE, 12*(1), E0169659. doi:10.1371/journal.pone.0169659

Clark, C. G. (1992). DNA purification from polysaccharide-rich cells. In J. J. Lee & A. T. Soldo (Eds.), *PROTOCOLS IN PROTOZOOLOGY* (pp. D-3.1-D-3.2). Lawrence, Kansas: Allen Press.

Clark, C. G. (1997). Extensive genetic diversity in *Blastocystis hominis*. *Mol. Biochem. Parasitol.*, *87,* 79-83. doi:10.1016/S0166-6851(97)00046-7

Clark, C. G., van der Giezen, M., Alfellani, M. A., Stensvold, C. R. (2013). Recent developments in *Blastocystis* research. Advances in Parasitology, 82. doi:10.1016/B978-0-12-407706-5.00001-0

Clark, C. G., & Stensvold, C. R. (2016) *Blastocystis*: Isolation, xenic cultivation, and cryopreservation. *Curr. Protoc. Microbiol., 43*, 20A.1.1-20A.1.8. doi:10.1002/cpmc.18

Crick, F. H. C. (1966). Codon-anticodon pairing. The wobble hypothesis. *Journal of Molecular Biology, 19* (2), 548-555. doi:10.1016/S0022-2836(66)80022-0

Criscuolo, A., & Gribaldo, S. (2010). BMGE (Block Mapping and Gathering with Entropy): A new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evolutionary Biology, 10*, 210. doi:10.1186/1471-2148-10-210

Danovaro, R., Dell'Anno, A., Pusceddu, A., Gambi, C., Heiner, I., & Kristensen, R., M. (2010). The first metazoa living in permanently anoxic conditions. *BMC Biology, 8*, 30. doi:10.1186/1741-7007-8-30

de Coster, W. (2017). NanoPlot: plotting scripts for long read sequencing data. Retrieved from https://github.com/wdecoster/NanoPlot

de Graaf, R., Ricard, G., Van Alen, T., Duarte, I., Dutilh, B., Burgtorf, C., . . . Hackstein, J. (2011). The Organellar Genome and Metabolic Potential of the Hydrogen-Producing Mitochondrion of Nyctotherus ovalis. *Mol. Biol. Evol. 28,* 2379–2391. doi:10.1093/molbev/msr059

Dehal, P. & Boore, J. L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLos Biol., 3*(10), e314. doi:10.1371/journal.pbio.0030314

De Koning, A., Gu, W., Castoe, T., Batzer, M., Pollock, D., & Copenhaver, G. (2011). Repetitive Elements May Comprise Over Two-Thirds of the Human Genome (Repeat-Derived Dark Matter of the Human Genome). *PLoS Genetics, 7*(12), E1002384. doi:10.1371/journal.pgen.1002384

Decker, H., Motamedi, H., & Hutchinson, C. R. (1993). Nucleotide sequencs and heterologous expression of tcmG and tcmP, biosynthetic genes for tetracenomycin C in *Streptomyces glaucescens*. *J. Bacteriol., 175*, 3876-3886.

Denoeud, F., Roussel, M., Noel, B., Wawrzyniak, I., Da Silva, C., Diogon, M., . . . El Alaoui, H. (2011). Genome sequence of the stramenopile *Blastocystis*, a human anaerobic parasite. *Genome Biology, 12*(3), R29. doi:10.1186/gb-2011-12-3-r29

Derelle, R., López-García, P., Timpano, H., Moreira, D. (2016). A phylogenomic framework to study the diversity and evolution of stramenopiles (=Heterokonts). *Mol. Biol. Evol., 33*(11), 2890-2898. doi:10.1093/molbev/msw168

Dorrell, R. G. & Bowler, C. (2017). Secondary Plastids of Stramenopiles. Advances in Botanical Research, 84, 57-103. doi:10.1016/bs.abr.2017.06.003

Dowton, M., Cameron, S., Dowavic, J., Austin, A., & Whiting, M. (2009). Characterization of 67 Mitochondrial tRNA Gene Rearrangements in the Hymenoptera Suggests That Mitochondrial tRNA Gene Position Is Selectively Neutral. *Molecular Biology and Evolution, 26*(7), 1607-1617. doi:10.1093/molbev/msp072

Eichinger, L., Pachebat, J. A., Glöckner, G., Rajandream, M.-A., Sucgang, R., Berriman, M., … Kuspa, A. (2005). The genome of the social amoeba *Dictyostelium discoideum. Nature, 435*(7038), 43-57. doi:10.1038/nature03481

Elzanowski, A. & Ostell, J. (2016). The Genetic Codes. Retrieved from https://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=cgencodes

Embley, M. T. (2006). Multiple secondary origins of the anaerobic lifestyle in eukaryotes. *Philos Trans R Soc Lond B Biol Sci., 361*(1470), 1055-1067. doi:10.1098/rstb.2006.1844

Eme, L., Gentekaki, E., Curtis, B., Archibald, J. M., & Roger, A. J. (2017). Lateral Gene Transfer in the Adaptation of the Anaerobic Parasite *Blastocystis* to the Gut. *Current Biology, 27*(6), 807-820. 10.1016/j.cub.2017.02.003

Eren, A., Esen, &., Quince, C., Vineis, J., Morrison, H., Sogin, M., & Delmont, T. (2015). Anvi'o: An advanced analysis and visualization platform for 'omics data. *PeerJ, 3*, E1319. doi:10.7717/peerj.1319

Esseiva, A. C., Schneider, A., Naguleswaran, A., & Hemphill, A. (2004). Mitochondrial tRNA import in *Toxoplasma gondii*. *Journal of Biological Chemistry, 279*(41), 42363-42368. doi:10.1074/jbc.M404519200

Fantham, H. B. (1916). The nature and distribution of the parasites observed in the stools of 1305 dysenteric patients. *Lancet i(June 10)*, 1165-1166.

Feagin, J. E. (2000). Mitochondrial genome diversity in parasites. *International Journal for Parasitology, 30*(4), 371-390. doi:10.1016/S0020-7519(99)00190-3

Fisher, R.A. (1930). The genetical theory of natural selection. Oxford: Clarendon.

Forsell, J., Bengtsson-Palme, J., Angelin, M., Johansson, A., Evengård, B., & Granlund, M. (2017). The relation between *Blastocystis* and the intestinal microbiota in Swedish travellers. *Bmc Microbiology, 17*, BMC Microbiology, 2017, Vol.17. doi:10.1186/s12866-017-1139-7

Forsell, J., Granlund, M., Samuelsson, L., Koskiniemi, S., Edebro, H., & Evengård, B. (2016). High occurrence of *Blastocystis* sp subtypes 1-3 and Giardia intestinalis assemblage B among patients in Zanzibar, Tanzania. *Parasites &Amp; Vectors, 9*, Parasites & Vectors, 2016, Vol.9. doi:10.1186/s13071-016-1637-8

Fu, G., Nagasato, C., Oka, S., Cock, J. M., & Motomura, T. (2014). Proteomics analysis of heterogenous flagella in brown algae. *Protist, 165*(5), 662-675. doi:10.1016/j.protis.2014.07.007

Gene Corps Corporations (2017). Sequencher® version 5.4.6 DNA sequence analysis software. Retrieved from http://www.genecodes.com

Gentekaki, E., Curtis, B. A., Stairs, C. W., Klimes, V., Elias, M., Salas-Leiva, D. E., . . . Roger, A., J. (2017). Extreme genome diversity in the hyper-prevalent parasitic eukaryote *Blastocystis*. *Plos Biology, 15*(9). doi:10.1371/journal.pbio.2003769

Gish, W., & States, D. J. (1993). Identification of protein coding regions by database similarity search. *Nature Genetics, 3*(3), 266-72. doi:10.1038/ng0393-266

Grabherr, M. G, Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D.A., Amit, I., … Regev, A. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology, 29*(7), 644-52. doi:10.1038/nbt.1883

Gray, M. W., Lang, B. F., & Burger, G. (2004). Mitochondria of protists. *Annual Review of Genetics, 38*, 477-524. doi:10.1146/annurev.genet.37.110801.142526

Greener, J. G., Filippis, I., & Sternberg, M. J. E. (2017). Predicting protein dynamics and allostery using multi-protein atomic distance constraints. *Structure, 25*(3), 546-558. doi:10.1016/j.str.2017.01.008

Grundmann, O., & Yoon, S. (2010). Irritable bowel syndrome: Epidemiology, diagnosis and treatment: An update for health care practitioners. *Journal of Gastroenterology and Hepatology, 25*(4), 691-699. doi:10.1111/j.1440-1746.2009.06120.x

Guo, H. H., Choe, J., & Loeb, L. A. (2004). Protein tolerance to random amino acid change. *PNAS, 101*(25), 9205-9210. doi:10.1073/pnas.0403255101

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics, 29*(8), 1072-1075. doi:10.1093/bioinformatics/btt086

Gyles, C., & Boerlin, P. (2014). Horizontally Transferred Genetic Elements and Their Role in Pathogenesis of Bacterial Disease. *Veterinary Pathology, 51*(2), 328-340. doi:10.1177/0300985813511131

Haas, B, J. (2017). TransDecoder. Retrieved from https://github.com/TransDecoder/TransDecoder

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K., Hannick, L. I., . . . White, O. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research, 31*(19), 5654-66. doi:10.1093/nar/gkg770

Halfmann, R., Alberti, S., Krishnan, R., Lyle, N., O'Donnell, C. W., . . . Lindquist, S. (2011). Opposing Effects of Glutamine and Asparagine Govern Prion Formation by Intrinsically Disordered Proteins. *Molecular Cell, 43*(1), 72-84. doi:10.1016/j.molcel.2011.05.013

Harada, E. & Nishimura, S. (1972). Possible anticodon sequences of $tRNA^{His}$, $tRNA^{Asn}$, and $tRNA^{Asp}$ from *Escherichia coli B*. Universial presence of nulceoside Q in the first position of the anticodon of three transfer ribonulceic acids. *Biochemistry, 11*, 301–308.

Harr, B., Zangerl, B., & Schlötterer, C. (2000). Removal of Microsatellite Interruptions by DNA Replication Slippage: Phylogenetic Evidence from *Drosophila*. *Molecular Biology and Evolution*, *17*(7), 1001-1009. doi:10.1093/oxfordjournals.molbev.a026381

Hartl, D., & Ruvolo, M. (2012). *Genetics : Analysis of genes and genomes* (8th ed.). Burlington, MA: Jones & Bartlett Learning.

Henze, K., & Martin, W. (2001). How do mitochondrial genes get into the nucleus? *Trends in Genetics, 17*(7), 383-387. doi:10.1016/S0168-9525(01)02312-5

Hirabayashi, K., Yuda, E., Tanaka, N., Katayama, S., Iwasaki, K., Matsumoto, T., . . . Wada,K. (2015). Functional Dynamics Revealed by the Structure of the SufBCD Complex, a Novel ATP-binding Cassette (ABC) Protein That Serves as a Scaffold for Iron-Sulfur Cluster Biogenesis. *The Journal of Biological Chemistry, 290*(50), 29717-31. doi:10.1074/jbc.M115.680934

Holberton, D. V. (1973). Fine structure of the ventral disk apparatus and the mechanism of attachment in the flagellate giardia muris. *Journal of Cell Science, 13*(1), 11-41. http://jcs.biologists.org/content/13/1/11

Hooper, L. V. & Gordon, J. I. (2001). Glycans as legislators of host-microbial interactions: Spanning the spectrum from symbiosis to pathogenicity. *Glycobiology, 11*(2), 1R-10R. doi:10.1093/glycob/11.2.1R

Hou, F., Sun, L., Zheng, H., Skaug, B., Jiang, Q., & Chen, Z. J. (2011). MAVS Forms Functional Prion-like Aggregates to Activate and Propagate Antiviral Innate Immune Response. *Cell, 146*(5), 841. doi:10.1016/j.cell.2011.08.013

Hubley, R., Finn, R. D., Clements, J., Eddy, S. R., Jones, T. A., Bao, W., … Wheeler, T. J. (2016). The Dfam database of repetitive DNA families. Nucleic Acids Research, 44(Database issue), D81-9. doi:10.1093/nar/gkv1272

Husnik, F. & McCutcheon, J. P. (2018). Functional horizontal gene transfer from bacteria to eukaryotes. *Nat. Rev. Microbiol., 16*, 67-79. doi:10.1038/nrmicro.2017.137

Hutchinson, C. R. (1997). Biosynthetic Studies of Daunorubicin and Tetracenomycin C. *Chem. Rev., 97*(7), 2525-2536. doi:10.1021/cr960022x

Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Bio. Evol. 2*, 389-409. doi:10.1093/oxfordjournals.molbev.a040335

Illumina Inc. (2018). Input requirements. Retrieved from https://support.illumina.com/sequencing/sequencing_kits/nextera_dna_kit/input_req.html

Jackson, C., Norman, J., Schnare, M., Gray, M., Keeling, P., & Waller, R. (2007). Broad genomic and transcriptional analysis reveals a highly derived genome in dinoflagellate mitochondria. *BMC Biology, 5*, 41. doi:10.1186/1741-7007-5-41

Jacob, A., Andersen, L., Pavinski Bitar, P., Richards, V., Shah, S., Stanhope, M., . . . Clark, C. (2016). *Blastocystis* mitochondrial genomes appear to show multiple independent gains and losses of start and stop codons. *Genome Biology and Evolution, 8*(11), 3340-3350. doi:10.1093/gbe/evw255

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., . . . Depamphilis, C. W. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature, 473*(7345), 97-100. doi:10.1038/nature09916

Jones, P., Binns, D., Chang, H., Fraser, M., Li, W., McAnulla, C., . . . Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics, 30*(9), 1236-1240. doi:10.1093/bioinformatics/btu031

Jordan, J. & Kajava, A. V. (2010). Protein homorepeats: sequences, structures, evolution, and functions. *Adv Protein Chem Struct Biol., 79*, 59-88. doi:10.1016/S1876-1623(10)79002-7

Judd, L. (2017, August 15). Whole genome viral DNA – to fragment…or not to fragment [Blog post]. Retrieved from https://community.nanoporetech.com/posts/whole-genome-viral-dna-t

Kang, S., Tice, A., Spiegel, F., Silberman, J., Pánek, T., Čepička, I., . . . Brown, M. (2017). Between a Pod and a Hard Test: The Deep Evolution of Amoebae. *Molecular Biology and Evolution, 34*(9), 2258-2270. doi:10.1093/molbev/msx162

Kapitonov, V. V. & Kurka. J. (2007). Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends in Genetics, 23*(10), 521-529. doi:10.1016/j.tig.2007.08.004

Karnkowska, A., Vacek, V., Zubáčová, Z., Treitli, S.C., Petrželková, R., Eme, L., … Hampl, V. (2016). A Eukaryote without a Mitochondrial Organelle. *Current Biology, 26*(10), 1274-1284. doi:10.1016/j.cub.2016.03.053

Katoh, K., Rozewicki, J., & Yamada, K. (2017). MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics,* 1-7. doi:10.1093/bib/bbx108

Kazazian, H. H., Wong, C., Youssoufian, H., Scott, A. F., Phillips, D. G., & Antonarakis, S. E. (1988). Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature, 332*(6160), 164-6.

Keeling, P. J. & Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics, 9*(8), 605-18. doi:10.1038/nrg2386

Keller, O., Kollmar, M., Stanke, M., & Waack, S. (2011). A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics, 27*(6), 757-763. doi:10.1093/bioinformatics/btr010

Kennedy, S., Salk, J., Schmitt, M., & Loeb, L. (2013). Ultra-Sensitive Sequencing Reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage: E1003794. *PLoS Genetics, 9*(9). doi:10.1371/journal.pgen.1003794

Kiethega, G. N., Yan, Y., Turcotte, M., & Burger, G. (2013). RNA-level unscrambling of fragmented genes in Diplonema mitochondria. *RNA Biol., 10*, 301–313. doi:10.4161/rna.23340

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature Methods, 12*(4), 357-360. doi:10.1038/nmeth.3317

Kim. E. & Archiblad, J. M. (2009). Diversity and Evolution of Plastids and Their Genomes. In A. S. Sandelius & H. Aronsson, *The Chloroplast* (pp. 1-39). Berlin, Heidelberg: Springer.

Kimura, M. (1968). Evolutionary Rate at the Molecular Level. *Nature, 217*, 624-626.

King, J. L. & Jukes, T. H. (1969). Non-Darwinian evolution. *Science, 164*(3881), 788-798.

Klimes, V., Gentekaki, E., Roger, A., & Eliás, M. (2014). A Large Number of Nuclear Genes in the Human Parasite *Blastocystis* Require mRNA Polyadenylation to Create Functional Termination Codons. *Genome Biology and Evolution, 6*(8), 1956-1961. doi:10.1093/gbe/evu146

Koren, S., Walenz, B., Berlin, K., Miller, J., Bergman, N., & Phillippy, A. (2017). Canu: Scalable and accurate long-read assembly via adaptive -mer weighting and repeat separation. *Genome Research, 27*(5), 722-736. doi:10.1101/gr.215087.116

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., . . . Marra, M. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research, 19*(9), 1639-45. doi:10.1101/gr.092759.109

Lafay, B., Atherton, J., Sharp, P. (2000). Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*. *Microbiology*, 146, 851-860. doi:10.1099/00221287-146-4-851

Lagesen, K., Hallin, P., Rødland, E., Staerfeldt, H., Rognes, T., & Ussery, D. (2007). RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research, 35*(9), 3100-8. doi:10.1093/nar/gkm160

Land, M., Hauser, L., Jun, S., Nookaew, I., Leuze, M., Ahn, T., . . . Ussery, D. (2015). Insights from 20 years of bacterial genome sequencing. *Functional & Integrative Genomics, 15*(2), 141-161. doi:10.1007/s10142-015-0433-4

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods, 9*(4), 357-9. doi:10.1038/nmeth.1923

Lanuza, M. D., Carbajal, J. A., & Borrás, R. (1996). Identification of surface coat carbohydrates in *Blastocystis hominis* by lectin probes. *International Journal for Parasitology, 26*(5), 527-532. doi:10.1016/0020-7519(96)00010-0

Laslett, D, & Canbäck, Björn. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research, 32*(1), 11-16. doi:10.1093/nar/gkh152

Lavrov, D. V. & Pett, W. (2016). Animal mitochondrial DNA as We Do Not Know It: mt-Genome Organization and Evolution in Nonbilaterian Lineages. *Genome Biology and Evolution, 8*(9), 2896-2913. doi:10.1093/gbe/evw195

Le, S. Q., & Gascuel, O. (2008). An Improved General Amino Acid Replacement Matrix. *Molecular Biology and Evolution, 25*(7), 1307-1320. doi:10.1093/molbev/msn067

Le, S., Q., Dang, C., & Gascuel, O. (2012). Modeling Protein Evolution with Several Amino Acid Replacement Matrices Depending on Site Rates. *Molecular Biology and Evolution, 29*(10), 2921-2936. doi:10.1093/molbev/mss112

Le, S. Q., Gascuel, O., & Lartillot, N. (2008). Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics, 24*(20), 2317-2323. doi:10.1093/bioinformatics/btn445

Lee, M. G. & Stenzel, D. J. (1999). A survey of *Blastocystis* in domestic chickens. *Parasitol. Res., 85*, 109-117. doi:10.1007/s004360050518

Leger, M. M, Eme, L., Hug, L., & Roger, A. J. (2016). Novel Hydrogenosomes in the Microaerophilic Jakobid *Stygiella incarcerata*. *Molecular Biology and Evolution, 33*(9), 2318-2336. doi:10.1093/molbev/msw103

Leger, M. M., Kolisko, M., Kamikawa, R., Stairs, C. W., Kume, K., Čepička, I., … Roger, A. J. (2017). Organelles that illuminate the origins of Trichomonas hydrogenosomes and Giardia mitosomes. Nature Ecology & Evolution, 1. doi: 10.1038/s41559-017-0092

Leger, M. M., Eme, L., Stairs, C. W., & Roger, A. J. (2018). Demystifying eukaryote lateral gene transfer (Response to Martin 2017 DOI:10.1002/bies.201700115). *Bioessays, 40*(5). doi:10.1002/bies.201700242

Legendre, F., Nel, A., Svenson, G., Robillard, T., Pellens, R., & Grandcolas, P. (2015). Phylogeny of Dictyoptera: Dating the Origin of Cockroaches, Praying Mantises and Termites with Molecular Data and Controlled Fossil Evidence: E0130127. *PLoS ONE, 10*(7). doi: 10.1371/journal.pone.0130127

Leontis, N., Stombaugh, J., & Westhof, E. (2002). The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Research, 30*(16), 3497-531.

Li, H. (2017). Minimap2: Versatile pairwise alignment for nucleotide sequences. Retrieved from arXiv:1708.01492

Li, G-W., Burkhardt, D., Gross, C., Weissman, J. S. (2014). Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell, 157*(3), 624-635. doi:10.1016/j.cell.2014.02.033

Li, J., Sayeed, S., Robertson, S., Chen, J., McClane, B., & Koehler, T. (2011). Sialidases Affect the Host Cell Adherence and Epsilon Toxin-Induced Cytotoxicity of *Clostridium perfringens* Type D Strain CN3718. *PLoS Pathogens, 7*(12), E1002429. doi:10.1371/journal.ppat.1002429

Lill, R., Dutkiewicz, R., Freibert, S. A., Heidenreich, T., Mascarenhas, J., Netz, D. J., . . . Mühlenhoff, U. (2015). The role of mitochondria and the CIA machinery in the maturation of cytosolic and nuclear iron–sulfur proteins. *European Journal of Cell Biology, 94*(7-9), 280-291. doi:10.1016/j.ejcb.2015.05.002

Lima, L., Sinaimeri, B., Sacomoto, G., Lopez-Maestre, H., Marchet, C., Miele, V., . . . Lacroix, V. (2017). Playing hide and seek with repeats in local and global de novo transcriptome assembly of short RNA-seq reads. *Algorithms for Molecular Biology : AMB, 12*(1), 2. doi:10.1186/s13015-017-0091-2

Liu, Y., & Cui, Z. (2010). Complete mitochondrial genome of the Asian paddle crab *Charybdis japonica* (Crustacea: Decapoda: Portunidae): Gene rearrangement of the marine brachyurans and phylogenetic considerations of the decapods. *Molecular Biology Reports, 37*(5), 2559-2569. doi:10.1007/s11033-009-9773-2

Lohse, M., Drechsel, O., Kahlau, S., & Bock, R. (2013). OrganellarGenomeDRAW--a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Research, 41*(Web Server issue), W575-81. doi:10.1093/nar/gkt289

Loman, N. J., Quick, J., & Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods, 12*(8), 733-735. doi:10.1038/nmeth.3444

Lomsadze A., Burns P.D. & Borodovsky M. (2014). Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research, 42* (15), e119. doi:10.1093/nar/gku557

Lowe, T. M., & Eddy, S. R. (1997). TRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research, 25*(5), 955-64.

Lynch, K. M. (1917). *Blastocystis hominis*: its characteristics and its prevalence in intestinal contents and feces in South Carolina. *J. Bacteriol., 2*, 369-377. http://jb.asm.org/content/2/4/369

Maguire, B. A. & Zimmermann, R. A. (2001). The ribosome in focus. Cell, 104(6), 813-816. doi:10.1016/S0092-8674(01)00278-1

Malinovska, L., Palm, S., Gibson, K., Verbavatz, J., & Alberti, S. (2015). *Dictyostelium discoideum* has a highly Q/N-rich proteome and shows an unusual resilience to protein aggregation. *PNAS, 112*(20), E2620-9. doi:10.1073/pnas.1504459112

Martijn, J., Vosseberg, J., Guy, L., Offre, P., & Ettema, T. J. G. (2018). Deep mitochondrial origin outside the sampled alphaproteobacteria. Nature. doi: 10.1038/s41586-018-0059-5

Matsuo, Y., Akiyama, N., Nakamura, H., Yodoi, J., Noda, M., & Kizaka-Kondoh, S. (2001). Identification of a novel thioredoxin-related transmembrane protein. *The Journal of Biological Chemistry, 276*(13), 10032-8. doi:10.1074/jbc.M011037200

McClintock, B. The origin and behavior of mutable loci in maize. *PNAS, 36*(6), 344-355. doi:10.1073/pnas.36.6.344

Miller, R. J. (1998). Mitochondria – the kraken awakens! *Trends in Neurosciences, 21*(3), 95-97. doi:10.1016/S0166-2236(97)01206-X

Minh, B., Nguyen, M., & Von Haeseler, A. (2013). Ultrafast Approximation for Phylogenetic Bootstrap. *Molecular Biology and Evolution, 30*(5), 1188-1195. doi:10.1093/molbev/mst024

Moran, A. P. (2008). Relevance of fucosylation and Lewis antigen expression in the bacterial gastroduodenal pathogen Helicobacter pylori. *Carbohydrate Research, 343*(12), 1952-1965. doi:10.1016/j.carres.2007.12.012

Mourier, T. J., Hansen, A. J., Willerslev, E., & Arctander, P. (2001). The human genome project reveals a continuous transfer of large mitochondrial fragments to the nucleus [3]. *Molecular Biology and Evolution, 18*(9), 1833-1837. doi:10.1093/oxfordjournals.molbev.a003971

Müller, M., Mentel, M., van Hellemond, J.J., Henze, K., Woehle, C., Gould, S.B., … Martin, W.F. (2012). The organellar genome and metabolic potential of the hydrogen-producing mitochondrion of *Nyctotherus ovalis*. *Mol. Biol. Evol.*, *28,* 2379–2391. doi:10.1128/MMBR.05024-11

Nakayama, T., Ishida, K., & Archibald, J. M. (2012). Broad Distribution of TPI-GAPDH Fusion Proteins among Eukaryotes: Evidence for Glycolytic Reactions in the Mitochondrion? *PLoS ONE, 7*(12), E52340. doi:10.1371/journal.pone.0052340

Nash, E. A., Nisbet, R. E., Barbrook, A. C., & Howe, C. J. (2008). Dinoflagellates: a mitochondrial genome all at sea. *Trends Genet. 24*, 328-335. doi:10.1016/j.tig.2008.04.001

Nation, J. L., Sr. (2015). *Insect Physiology and Biochemistry (3rd ed.)*. Gainesville, Florida: CRC Press.

National Center for Biotechnology Information (1988). Retrieved from https://www.ncbi.nlm.nih.gov/

Nenarokov, S. (2018). WinstonCleaner. Retrieved from https://github.com/kolecko007/WinstonCleaner

Nguyen, V. H. & Lavenier, D. (2009). PLAST: parallel local alignment search tool for database comparison. *BMC Bioinformatics*, *10* (329). doi:10.1186/1471-2105-10-329

Nguyen, L., Schmidt, H., Von Haeseler, A., & Minh, B. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution, 32*(1), 268-74. doi:10.1093/molbev/msu300

Niemann, M., Harsman, A., Mani, J., Peikert, C., Oeljeklaus, S., Warscheid, B., . . . Schneider, A. (2017). TRNAs and proteins use the same import channel for translocation across the mitochondrial outer membrane of trypanosomes. *PNAS, 114*(37), E7679-E7687. doi:10.1073/pnas.1711430114

Noguchi, F., Simamura, S., Nakayama, T., Yazaki, E., Yabuki, A., Hashimoto, T., … Takishita, K. (2015). Metabolic capacity of mitochondrion-related organelles in the free-living anaerobic stramenopile *Cantina marsupialis. Protist, 166*(5), 534-550. doi:10.1016/j.protis.2015.08.002

Noutsos, C., Kleine, T., Armbruster, U., Dalcorso, G., & Leister, D. (2007). Nuclear insertions of organellar DNA can create novel patches of functional exon sequences. *Trends in Genetics, 23*(12), 597-601. doi:10.1016/j.tig.2007.08.016

Park, J. S., Cho, B. C., & Simpson, A. G. (2006). *Halocafeteria seosinensis* gen. et sp. nov. (Bicosoecida), a halophilic bacterivorous nanoflagellate isolated from a solar saltern. *Extremophiles, 6*, 493-504. doi:10.1007/s00792-006-0001-x

Patino, M., Liu, J., Glover, J., & Lindquist, S. (1996). Support for the prion hypothesis for inheritance of a phenotypic trait in yeast. *Science (New York, N.Y.), 273*(5275), 622-6. doi:10.1126/science.273.5275.622

Pearson, W. R. (2013). An introduction to sequence similarity ("Homolog") searching. *Curr Protoc Bioinformatics*. doi:10.1002/0471250953.bi0301s42

Pérez-Brocal, V., Shahar-Golan, R., & Clark, C. G. (2010). A Linear Molecule with Two Large Inverted Repeats: The Mitochondrial Genome of the Stramenopile *Proteromonas lacertae*. *Genome Biol Evol., 2*(1), 257-66. doi:10.1093/gbe/evq015

Petruska, J. J., Hartenstine, M. F., & Goodman, M. (1998). Analysis of strand slippage in DNA polymerase expansions of CAG/CTG triplet repeats associated with neurodegenerative disease. *Journal of Biological Chemistry, 273*(9), 5204-5210. doi:10.1074/jbc.273.9.5204

Poirier, P., Wawrzyniak, I., Vivarès, C., Delbac, F., El Alaoui, H., & Knoll, L. (2012). New Insights into *Blastocystis* spp.: A Potential Link with Irritable Bowel Syndrome. *PLoS Pathogens, 8*(3), E1002545. doi:10.1371/journal.ppat.1002545

Popadin, K., Polishchuk, L. V., Mamirova, L., Knorre, D., & Gunbin, K. (2007). Accumulation of slightly deleterious mutations in mitochondrial protein-coding genes of large versus small mammals. *PNAS, 104*(33), 13390-13395. doi:10.1073/pnas.0701256104

Ranjan, N., & Rodnina, M. (2016). TRNA wobble modifications and protein homeostasis. *Translation, 4*(1). doi:10.1080/21690731.2016.1143076

Rawn, S., & Cross, J. (2008). The Evolution, Regulation, and Function of Placenta-Specific Genes. *Annual Review of Cell and Developmental Biology, 24*(1), 159-181. doi:10.1146/annurev.cellbio.24.110707.175418

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics, 16*(6), 276-277. doi:10.1016/S0168-9525(00)02024-2

Richly, E., & Leister, D. (2004). NUMTs in Sequenced Eukaryotic Genomes. *Molecular Biology and Evolution, 21*(6), 1081-1084. doi:10.1093/molbev/msh110

Roger, A. J., Muñoz-Gómez, S. A., & Kamikawa, R. (2017). The Origin and Diversification of Mitochondria. *Current Biology, 27*(21), R1177-R1192. doi:10.1016/j.cub.2017.09.015

Rostami, A., Riahi, S., Haghighi, M., Saber, V., Armon, B., & Seyyedtabaei, S. (2017). The role of *Blastocystis* sp . and *Dientamoeba fragilis* in irritable bowel syndrome: A systematic review and meta-analysis. *Parasitology Research, 116*(9), 2361-2371. doi:10.1007/s00436-017-5535-6

Santulli, G., Pagano, G., Sardu, C., Xie, W., Reiken, S., D'Ascia, S., . . . Marks, A. (2015). Calcium release channel RyR2 regulates insulin release and glucose homeostasis. *The Journal of Clinical Investigation, 125*(5), 1968-78. doi:10.1172/JCI79273

Scala, C., Tian, X., Mehdiabadi, N. J., Smith, M. H., Saxer, G., Stephens, K., … Queller, D. C. (2012). Amino acid repeats cause extraordinary coding sequence variation in the social amoeba *Dictyostelium discoideum. PLos One, 7*(9), e46150. doi:10.1371/journal.pone.0046150necess

Scanlan, P., Stensvold, C. R., Rajilić- Stojanović, M., Heilig, H., De Vos, W., O'Toole, P., & Cotter, P. (2014). The microbial eukaryote *Blastocystis* is a prevalent and diverse member of the healthy human gut microbiota. *FEMS Microbiology Ecology, 90*(1), 326-330. doi:10.1111/1574-6941.12396

Schauer, C., Thompson, C. L., & Brune, A. (2012). The Bacterial Community in the Gut of the Cockroach *Shelfordella lateralis* Reflects the Close Evolutionary Relatedness of Cockroaches and Termites. *Applied and Environmental Microbiology, 78*(8), 2758-2767. doi:10.1128/aem.07788-11

Sender, R., Fuchs, S., & Milo, R. (2016). Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLOS Biology, 14*(8). doi:10.1371/journal.pbio.1002533

Seeman, T. (2013). Barrnap - Bacterial ribosomal rRNA predictor. Retrieved from https://github.com/tseemann/barrnap

Shiratori, T., Nakayama, T., & Ishida, K.-I. (2015). A New Deep-branching Stramenopile, *Platysulcus tardus* gen. nov., sp. nov. *Protist, 166*(3), 337-348. doi:10.1016/j.protis.2015.05.001

Shokolenko, I., Venediktova, N., Bochkareva, A., Wilson, G. L., & Alexeyev, M. F. (2009). Oxidative stress induces degradation of mitochondrial DNA. *Nucleic Acids Res.*, *37*, 2539–2548. doi:10.1093/nar/gkp100

Si, K., Choi, Y., White-Grindley, E., Majumdar, A., & Kandel, E. R. (2010). Aplysia CPEB Can Form Prion-like Multimers in Sensory Neurons that Contribute to Long-Term Facilitation. *Cell, 140*(3), 421-435. doi:10.1016/j.cell.2010.01.008

Siegwald, L., Audebert, C., Even, G., Caboche, S., Viscogliosi, E., & Chabé, M. (2017). Targeted metagenomic sequencing data of human gut microbiota associated with *Blastocystis* colonization. *Scientific Data, 4*. doi:10.1038/sdata.2017.81

Simpson, A. G. B., & Eglit, Y. (2016). Protist Diversification. In R. M. Kliman, *The Encyclopedia of Evolutionary Biology* (pp. 344-369). Watham, MA: Academic Press. doi:10.1016/B978-0-12-800049-6.00247-X

Simpson, A. M., Suyama, Y., Dewcs, H., Campbell, D. A., & Simpson, L. (1989). Kinetoplastid mitochondria contain functional tRNAs which are encoded in nuclear DNA and also contain small mlnltircJe and maxicircle transcripts of unknown function. *Nucleic Acids Research, 17*(14), 5427–5446. doi:10.1093/nar/17.14.5427

Simpson, J. (2017). Nanopolish. Rerieved from https://github.com/jts/nanopolish

Silberman, J. D., Sogin, M. L., Leipe, D. D., & Clark, C. G. (1996). Human parasite finds taxonomic home. *Nature, 380*(6573), 398. doi:10.1038/380398a0

Singh, G. P., Chandra, B. R., Bhattacharya, A., Akhouri, R. R., Singh, S. K., & Sharma, A. (2004). Hyper-expansion of asparagines correlates with an abundance of proteins with prion-like domains in *Plasmodium falciparum*. *Molecular & Biochemical Parasitology, 137*(2), 307-319. doi:10.1016/j.molbiopara.2004.05.016

Singh, R. S., Walia, A. K., & Kanwar, J. R. (2016). Protozoa lectins and their role in host–pathogen interactions. *Biotechnology Advances, 34*(5), 1018-1029. doi: 10.1016/j.biotechadv.2016.06.002

Sloan, D. (2013). One ring to rule them all? Genome sequencing provides new insights into the 'master circle' model of plant mitochondrial DNA structure. *New Phytologist, 200*(4), 978-985. doi:10.1111/nph.12395

Smith, D. R. (2012). Updating our view of organelle genome nucleotide landscape. *Frontiers in Genetics.* doi:10.3389/fgene.2012.00175

Smith, D. G. S., Gawryluk, R. M. R., Spencer, D. F., Pearlman, R. E., Siu, K. W. M., & Gray, M. W. (2007). Exploring the Mitochondrial Proteome of the Ciliate Protozoon *Tetrahymena thermophila*: Direct Analysis by Tandem Mass Spectrometry. *Journal of Molecular Biology, 374*(3), 837-863. doi:10.1016/j.jmb.2007.09.051

Smith, D. R., & Keeling, P. J. (2015). Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. *PNAS, 112*(33), 10177-84. doi:10.1073/pnas.1422049112

Soubrier, J., Steel, M., Lee, M. S. Y., Sarkissian, C. D., Guindon, S., … Copper, A. (2012). The influence of rate hererogeneity among sites on the time dependence of molecular rates. *Mol. Biol. Evol., 11*(1), 3345-3358. doi:10.1093/molbev/mss140

Spector, J., Harrison, R., & Fishman, M. (2018). Fundamental science behind today's important medicines. *Science Translational Medicine, 10*(438). doi:10.1126/scitranslmed.aaq1787

Stabler, R. M. (1941). Intestinal protozoa in 106 parasitology students. *J. Parasitol., 27*, 90.

Stanke, M. (2011). Orthoparahomlist.pl. Retrieved from https://github.com/goshng/RNASeq-Analysis/blob/master/pl/orthoparahomlist.pl

Stairs, C., Eme, L., Muñoz-Gómez, S., Cohen, A., Shepherd, J., Fawcett, J., & Roger, A. J. (2018). Microbial eukaryotes have adapted to hypoxia by horizontal acquisitions of a gene involved in rhodoquinone biosynthesis. *ELife, 7*. doi:10.7554/eLife.34292

Stairs, C. W., Eme, L., Brown, M. W., Mutsaers, C., Susko, E., Dellaire, G., . . . Roger, A. J. (2014). A SUF Fe-S Cluster Biogenesis System in the Mitochondrion-Related Organelles of the Anaerobic Protist *Pygsuia*. *Current Biology, 24*(11), 1176-1186. doi:10.1016/j.cub.2014.04.033

Stechmann, A., Hamblin, K., Pérez-Brocal, V., Gaston, D., Richmond, G.S., van der Giezen, M., Clark, C.G., & Roger, A.J. (2008). Organelles in *Blastocystis* that Blur the Distinction between Mitochondria and Hydrogenosomes. *Curr. Biol., 18*, 580-585. doi:10.1016/j.cub.2008.03.037

Stensvold, C. R. (2012). Thinking *Blastocystis* outside the box. *Trends in Parasitoloty, 28*(8), 305. doi:10.1016/j.pt.2012.05.004

Stensvold, C. R., & Clark, C. G. (2016). Current status of *Blastocystis*: A personal view. *Parasitology International, 65*(6), 763-771. doi:10.1016/j.parint.2016.05.015

Stensvold, C. R., Suresh, G. K., Tan, K. S. W., Thompson, R. C. A., Traub, R. J., Viscogliosi, E., . . . Clark, C. G. (2007). Terminology for Blastocystis subtypes – a consensus. *Trends in Parasitology, 23*(3), 93-96. doi:10.1016/j.pt.2007.01.004

Stenzel, D.J. & Boreham, P.F.L. (1991). A cyst-like stage of *Blastocystis hominis*. *Int. J. Parasitol.*, *21*, 613–615.

Stenzel, D. J., & Boreham, P. F. L. (1996). *Blastocystis hominis* revisited. *Clinical Microbiology Reviews, 9*(4), 563-84. http://cmr.asm.org/content/9/4/563

Stibbs, H. H., Owczarzak, A., Bayne, C. J., & DeWan, P. (1979). Schistosome sporocyst-killing Amoebae isolated from *Biomphalaria glabrata*. *Journal of Invertebrate Pathology, 33* (2), 159–170. doi:10.1016/0022-2011(79)90149-6

Swart, E. C., Serra, V., Petroni, G., & Nowacki, M. (2016). Genetic Codes with No Dedicated Stop Codon: Context-Dependent Translation Termination. *Cell, 166*(3), 691-702. doi:10.1016/j.cell.2016.06.020

Tan, K. S. W. (2008). New Insights on Classification, Identification, and Clinical Relevance of *Blastocystis* spp. *Clinical Microbiology Reviews, 21*(4), 639-65. doi:10.1128/CMR.00022-08

Tan, T., & Suresh, K. (2006). Amoeboid form of *Blastocystis hominis* - a detailed ultrastructural insight. *Parasitology Research, 99*(6), 737-42. doi:10.1007/s00436-006-0214-z

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *American Mathematical Society, 17*, 57-86.

Thorvaldsdóttir, H., Robinson, J., & Mesirov, J. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics, 14*(2), 178-192. doi:10.1093/bib/bbs017

Treangen, T. D. & Salzberg, S. L. (2011). Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nature Reviews Genetics, 13*(1), 36-46. doi:10.1038/nrg3117

Tsaousis, A. D., Gaston, D., Stechmann, A., Walker, P. B., Lithgow, T., & Roger, A. J. (2011). A Functional Tom70 in the Human Parasite *Blastocystis* sp.: Implications for the Evolution of the Mitochondrial Import Apparatus. *Molecular Biology and Evolution, 28*(1), 781-791. doi:10.1093/molbev/msq252

Tsaousis, A. D., Ollagnier de Choudens, S., Gentekaki, E., Long, S., Gaston, D., Stechmann, A., . . . Roger, A. J. (2012). Evolution of Fe/S cluster biogenesis in the anaerobic parasite *Blastocystis*. *PNAS, 109*(26), 10426-31. doi:10.1073/pnas.1116067109

Vdovenko, A.A. (2000). *Blastocystis hominis*: origin and significance of vacuolar and granular forms. *Parasitol. Res.*, *86* (1), 8–10.

Vera, D. (2017). Albacore. Retrieved from https://github.com/dvera/albacore

Viguera, E., Canceill, D., & Ehrlich, S. D. (2001). Replication slippage involves DNA polymerase pausing and dissociation. *The EMBO Journal, 20*, 2587-2595.

Walski, T., De Schutter, K., Cappelle, K., Van Damme, E., & Smagghe, G. (2017). Distribution of glycan motifs at the surface of midgut cells in the cotton leafworm (*Spodoptera littoralis*) demonstrated by lectin binding. FRONTIERS IN PHYSIOLOGY, 8, 1020. doi:10.3389/fphys.2017.01020

Wang, C. & Youle, R. J. (2009). The role of mitochondria in apoptosis. *Anu Rev Genet.*, *43*, 95-118. doi:10.1146/annurev-genet-102108-134850

Wang, K., Singh, D., Zeng, Z., Coleman, S., Huang, Y., Savich, G., . . . Liu, J. (2010). MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research, 38*(18), E178. doi:10.1093/nar/gkq622

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., . . . Schwede, T. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Research,* gky427. doi:10.1093/nar/gky427

Watson, M. (2018). Mind the gaps - ignoring errors in long read assemblies critically affects protein prediction. bioRxiv 10.1101/285049

Wawrzyniak, I., Courtine, D., Osman, M., Hubans-Pierlot, C., Cian, A., Nourrisson, C., … Delbac, F. (2015). Draft genome sequence of the intestinal parasite *Blastocystis* subtype 4-isolate WR1. *Genomics Data, 4*, 22-23. doi:10.1016/j.gdata.2015.01.009

Weirather, J. L., De Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X., . . . Au, K. F. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research, 6*, 100. doi:10.12688/f1000research.10571.2

Wick, R. (2017a). Porechop. Retrieved from https://github.com/rrwick/Porechop

Wick, R. (2017b). Unicycler. Retrieved from https://github.com/rrwick/Unicycler

Williams, B., Kay, R., & Kirk, E. (2010). New perspectives on anthropoid origins. *PNAS, 107*(11), 4797-804. doi:10.1073/pnas.0908320107

Woischnik, M., & Moraes, C. T. (2002). Pattern of organization of human mitochondrial pseudogenes in the nuclear genome. *Genome Research, 12*(6), 885-93. doi:10.1101/gr.227202

Wolfe, K. (2015). Origin of the yeast whole-genome duplication. *PLos Biol., 13*(8), e1002221. doi:10.1371/journal.pbio.1002221

Wright, S. (1931). Evolution in Mendelian populations. *Genetics, 16*, 97-159.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution, 39*(3), 306-314. doi:10.1007/BF00160154

Yang, Z. (1995). A space-time process model for the evolution of DNA sequences. *Genetics, 139*(2), 99-1005. doi:

Yang, D., Oyaizu, Y., Oyaizu, H., Olsen, G. J., & Woese, C. R. (1985). Mitochondrial origins. *PNAS, 82*(13), 4443-4447. doi:10.1073/pnas.82.13.4443

Yoshikawa, H., Morimoto, K., Wu, Z., Singh, M., & Hashimoto, T (2004). Problems in speciation in the genus *Blastocystis*. *Trends Parasitol. 20*, 251–255. doi:10.1016/j.pt.2004.03.010

Yoshikawa, H., Wu, Z., Howe, J., Hashimoto, T., Geok- Choo, N., & Tan, K. (2007). Ultrastructural and Phylogenetic Studies on *Blastocystis* Isolates from Cockroaches. *Journal of Eukaryotic Microbiology, 54*(1), 33-37. doi:10.1111/j.1550-7408.2006.00141.x

Yubuki, N., Leander, B. S., & Silberman, J. D. (2010). Ultrastructure and Molecular Phylogenetic Position of a Novel Phagotrophic Stramenopile from Low Oxygen Environments: *Rictus lutensis* gen. et sp. nov. (Bicosoecida, incertae sedis). *Protist, 161*(2), 264-278. doi:10.1016/j.protis.2009.10.004

Yubuki, N., Čepička, I., & Leander, B. S. (2016). Evolution of the microtubular cytoskeleton (flagellar apparatus) in parasitic protists. *Molecular & Biochemical Parasitology, 209*(1-2), 26-34. doi:10.1016/j.molbiopara.2016.02.002

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics, 30*(5), 614-620. doi:10.1093/bioinformatics/btt593

Zierdt, C.H., & Williams, R. L. (1974). *Blastocystis hominis*: axenic cultivation. *Exp Parasitol, 36*(2), 233-243.

Zierdt, C. H., Rude, W. S., & Bull, B. S. (1967). Protozoan characteristics of *Blastocystis hominis*. *Am J Clin Pathol, 48*(5), 495-501.