FACULTY OF SCIENCE AND TECHNOLOGY

MASTER'S THESIS

Study program/specialization: Biological Chemistry	The autumn semester, 2023 Open
Author: Ine Holand	I. Holand (author signature)
Supervisor at UiS: Mark van der Giezen	
Thesis title: Exploring Pathogenicity in *Blastocystis* ST7 and ST4: A comparative analysis of virulence factors from known intestinal parasites for pathogenicity assessment in *Blastocystis* spp.	
Credits (ECTS): 60stp.	
Keywords: *Blastocystis* Intestinal parasites Virulence factors Pathogenicity Bioinformatics	Pages: 78 + appendix: 12 Stavanger, October 2023

Acknowledgment

What a year. Thank you to my friends and family for being there for me during this unexpected and intense year. Thank you to my classmates for always supporting me through the tough periods and for all the coffee breaks and motivation you have given me.

I would like to express my gratitude to Mark for his supervision throughout this period. Additionally, I would like to thank Kari for her support, assistance, and the time she dedicated to helping me during the last period.

Au revoir!

Abstract

Blastocystis, a unicellular parasite found in the human gastrointestinal tract, has been a topic of study due to its potential pathogenicity. This study identified potential virulence factors within *Blastocystis* subtypes ST7 and ST4 from the established virulence factors of *Giardia intestinalis*, *Entamoeba histolytica*, and *Cryptosporidium pavrum*. Bioinformatics tools have been applied to analyze structural and characteristic differences. Certain motifs were found in *Blastocystis* which could impact its pathogenicity, as well as many conserved regions like in the known virulence factors. These include the motifs QxVxG, CxxC, RGD, ERFNIN, and GNFD, as well as a possible occluding loop. In terms of the active sites found, *Blastocystis* ST7 has all three active sites forming a catalytic dyad, while *Blastocystis* ST4 only had two active sites. This could indicate that *Blastocystis* ST7 has a higher enzymatic activity. The findings could have an impact on the regulation of the protease activity, modulating protein function, regulation of biological processes, and stability and folding of proteins.

The project also includes a molecular approach, where *Blastocystis* ST7 was meant to be cloned and determine the expression of genes 60SRPL32 and PC1A. Despite the challenges faced during the laboratory work of the project, not all experiments were completed. Therefore, a methodology of how it can be performed is explained. By completing this research, it could lead us to an improved understanding of *Blastocystis* and its behavior. The findings suggest similarities in the mechanisms and functions of these virulence factors, indicating the potential pathogenic nature of *Blastocystis*.

This study contributes to the growing field of knowledge in parasitology by giving important insights into the complicated world of parasitic infections. As scientists continue to examine the complexities of *Blastocystis*, this study provides the framework for eventually improving global health outcomes in the challenge of parasitic illnesses.

Abbreviations

A	Adenine
Asn	Asparagine
BLAST	Basic Local Alignment Search Tool
BP	Base Pair
C	Cytosine
CatB	Cathepsin B
CP	Cysteine Protease
cPCR	colony Polymerase Chain Reaction
CXCL	C-X-C Motif Chemokine Ligand
Cys	Cysteine
Cystatin	Cysteine protease inhibitor
EhCP	*Entamoeba histolytica* cysteine proteinase
FP	Forward Primer
G	Guanine
His	Histidine
IgA	Immunoglobulin A
IgG	Immunoglobulin G
IL	Interleukin
LB	Lysogeny Broth
NFκB	Nuclear factor kappa B
PC1A	Peptidase C1A
PCR	Polymerase Chain Reaction
PLP	Papain-like Proteases
PMN	Polymorphonuclear neutrophils
R	Arginine
RP	Reverse Primer
ST4	Subtype 4
ST7	Subtype 7
T	Thymine
T_m	Melting temperature
TNF-α	Tumor Necrosis Factor-alpha
60SRPL32	60S Ribosomal Protein L32

Table of Contents

1 Introduction

Parasitic infections remain a significant global health concern, affecting millions of individuals annually (Cummings & van Die, 2015). Among the diverse array of parasitic pathogens, *Giardia intestinalis*, *Entamoeba histolytica*, and *Cryptosporidium parvum* have gathered considerable attention due to their virulence factors and the associated diseases they cause (Argüello-García et al., 2023). Behind these considerably studied parasites lies *Blastocystis*, a protist that is not fully understood by researchers, in terms of genomic exploration and virulence factor characterization (Melo et al., 2021). *Blastocystis*, comprising multiple subtypes, presents a challenging yet intriguing area of research, with subtype 7 (ST7) and subtype 4 (ST4) emerging as subjects of particular interest in this project.

By comparing known virulence factors in other parasites to *Blastocystis* ST7 and ST4, we can potentially gain insights into its interactions with the host, as well as its involvement in different biological functions and pathways.

This thesis aims to study genes of interest of *Blastocystis* ST7 and ST4, comparing them with selected virulence factors of *Giardia intestinalis*, *Entamoeba histolytica*, and *Cryptosporidium parvum*. These organisms were chosen due to their similarities to *Blastocystis* like symptoms and environments. Although there are significant functional differences, studying these organisms can help improve our understanding and identification of potential drug targets. By focusing on the genomic complexities of *Blastocystis*, this study aims to contribute to the broader understanding of parasitic pathogenesis and insights into the less-explored aspects of parasitology.

1.1 *Blastocystis* – a pathogen?

Blastocystis is a microscopic parasite, strictly anaerobic, and can be found in the gastrointestinal tract of insects, birds, and mammals (Parija & Jeremiah, 2013). The parasite has been found in the stools of some individuals who experience symptoms like diarrhea, stomach pain, or other gastrointestinal issues, but researchers are unsure whether *Blastocystis* contributes to disease development at all (Andersen & Stensvold, 2016).

1.1.1 History of *Blastocystis* spp.

Due to not having enough documentation for studies of *Blastocystis* in the 1800s, the exact date of the initial discovery of the parasite is uncertain (Zierdt, 1991). It was initially discovered in diarrheal patients' feces by the Russian phycologist and protozoologist A. G. Alexeieff in 1911 (Alexeieff, 1911). He designated the organism as yeast and gave it the name *Blastocystis enterocola*. He presented an illustration of *Blastocystis* through its life cycle, shown in Figure 1, and it is the very first documentation of the parasite. A year later, Brumpt discovered it after looking at a human fecal sample and named the microbe *Blastocystis hominis* (Stenzel & Boreham, 1996). It was not considered a pathogenic until a study by Zierdt in the 1970s where the characterization of the organism at the microscopic level was described, providing valuable awareness of its morphology and cellular composition (Sienzel et al., 1991).

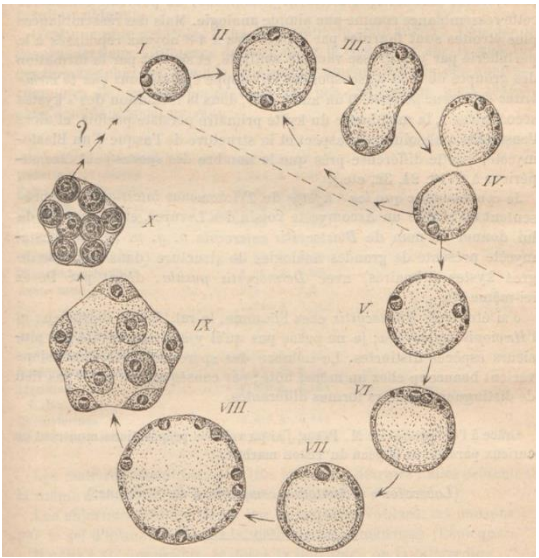


Figure 1 The first documented illustration of the life cycle of *Blastocystis* created by Alexeieff (Alexeieff, 1911).

1.1.2 Characteristics and classification

The different characteristic forms of *Blastocystis* are vacuolar, granular, ameboid, and its vegetative forms are avacuolar and multi-vacuolar forms (Parija & Jeremiah, 2013). These different forms are illustrated in Figure 2.

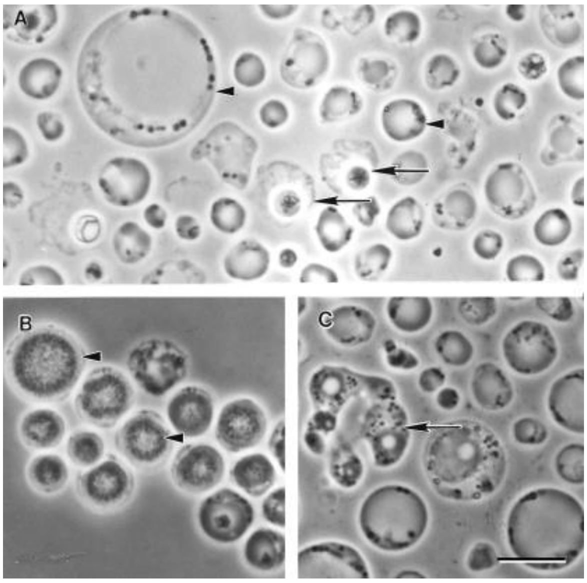


Figure 2 Photo illustrating forms of *Blastocystis* (Bar, 10 µm) produced by Kevin S. W Tan. A) Vacuolar form. It shows a big, central vacuole with a wide range of size variations. B) Granular form. Within the central vacuole, the granular form has discrete granular inclusions. C) Ameboid form. This can occasionally be observed in culture, where the cytoplasmic extensions resemble pseudopods (Tan, 2008).

Researchers have discovered several variants of the parasite, including various strains or separate species. The classification and understanding of *Blastocystis* has undergone several revisions over the years (Noël et al., 2005; Parija & Jeremiah, 2013). The term "blastocystosis" refers to a *Blastocystis* infection, and the current scientific designation is *Blastocystis* spp., which stands for "many species" (Boorom et al., 2008). Recent research has led to a new classification system based on genetic analysis, which divides *Blastocystis* into distinct genetic clusters known as "ST" (subtype types) based on molecular analysis of the small subunit ribosomal RNA gene (SSU rRNA) (Ajjampur & Tan, 2016). Currently, 40 STs have been described, and at least 28 subtypes (STs, ST1-ST17, ST21, ST23–32) have been identified in humans, other mammals, and birds (Martín-Escolano et al., 2023; Stensvold et al., 2023). These STs show distinct geographical and host-specific distribution patterns and may have different clinical and pathological implications (Jiménez et al., 2022). However, the relationship between *Blastocystis* subtypes and pathogenicity is not yet fully understood and needs more research.

1.1.3 Transmission and pathogenesis

For many years, *Blastocystis* was considered a harmless commensal organism, meaning it simply coexisted with its host without causing significant harm (Lepczyńska et al., 2017). However, in the 1990s, studies began to suggest that *Blastocystis* might be involved in causing gastrointestinal symptoms in some individuals, particularly those with chronic diarrhea, abdominal pain, and other digestive issues. *Blastocystis* can spread by food, drink, contact with human or animal waste, and other contaminated sources (Wawrzyniak et al., 2013). People who work with animals, live or travel in developing countries tend to be more susceptible to *Blastocystis* infection (Rajah Salim et al., 1999). Numerous animal species have the microbial parasite *Blastocystis* invade their large intestines, and mounting evidence connects *Blastocystis* infection to enteric illnesses, which can manifest as stomach pain, constipation, diarrhea, nausea, vomiting, and flatulence (Yoshikawa et al., 2004). Additionally, ST7 is now known to play a significant role in the host's intestinal microbiota (Deng et al., 2022). Although significant progress in our understanding of *Blastocystis* cell biology and host-parasite interactions, a new tool for genetic alteration has been introduced in the past year (Li et al., 2019). Currently, *Blastocystis* is recognized as one of the most common parasites found in human stools worldwide, with a prevalence of up to 60% in developing countries (Soghra et al., 2022). Clinical manifestations linked to *Blastocystis* infection exhibit a broad spectrum, encompassing asymptomatic carriage, acute or chronic gastrointestinal symptoms like diarrhea, abdominal pain, bloating, and nausea, as well as extraintestinal effects such as urticaria, irritable bowel syndrome (IBS) and chronic fatigue syndrome (Tan et al., 2010). However, the pathogenic mechanisms underlying *Blastocystis*-associated diseases remain unclear (Beyhan et al., 2015). The diagnosis of *Blastocystis* infection relies on the detection of the parasite in fecal samples by microscopy or molecular methods, although the sensitivity and specificity of these tests vary widely. The treatment of *Blastocystis* infection is also debatable, as some studies suggest that the parasite may have intrinsic resistance to some antiparasitic drugs (Mirza et al., 2011).

1.2 *Giardia intestinalis*

Giardia intestinalis is a pathogenic protozoan, also known as *Giardia duodenalis* and *Giardia lamblia*. The parasite is characterized by unicellular flagella, illustrated in a 3D model in Figure 3. The disease caused by *G. intestinalis* often starts as an acute condition but can also develop into a chronic condition (Rumsey & Waseem, 2023). It is typically contracted through contact with polluted water and is transferred fecal-orally (Dixon, 2021). Although some infected people may not exhibit any symptoms, the most typical signs and symptoms include aqueous diarrhea, oily stools, nausea, abdominal

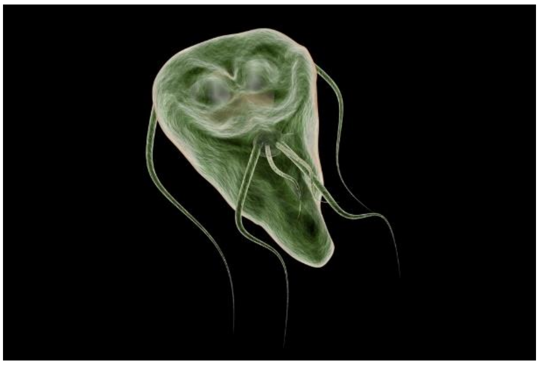


Figure 3 Illustration of *Giardia intestinalis* in a 3D version. Copyright: fotovapl / Shutterstock (Robertson, 2019)

pain, vomiting, and weight loss (Vivancos et al., 2018). Giardiasis is also linked to IBS (Løge, 2012). Metronidazole is the first-line treatment; however, there are other possibilities, and the disease is typically self-limiting (Petri, 2005; Rumsey & Waseem, 2023).

Different assemblages of *Giardia* are categorized as types A through H, with types A and B appearing in humans and animals and types C to H occurring only in animals (Zajaczkowski et al., 2021). It displayed proteolytic activity in *G. intestinalis*, where cysteine proteases (CPs) have an important role in the parasite's virulence (Liu, 2019). The breakdown of the intestinal epithelial junctional complex, intestinal epithelial cell death, and degradation of host immunological components like chemokines and immunoglobulins are all caused directly by *Giardia* CPs (Allain et al., 2019).

Virulence factors of *G. intestinalis* include energy metabolism enzymes, proteinases, high-cysteine membrane proteins (HCMPs), and variant surface proteins (VSPs) (Liu, 2019). In this study, the focus is mainly on cysteine proteinases that impact the interaction with *Giardia* on host cells. (Argüello-García & Ortega-Pierres, 2021; Peirasmaki et al., 2020)

1.3 *Entamoeba histolytica*

The protozoan *Entamoeba histolytica* causes intestinal amebiasis and is commonly found in countries with inadequate socioeconomic conditions and reduced public health (Chou & Austin, 2023). The parasite is a global health issue, and the transmission is usually through contaminated food or water sources, where the consumption of amebic cysts through fecal-oral contact (Kantor et al., 2018). *E. histolytica* can exist in two forms: the active and invasive trophozoite stage and the cyst form, which can live in the environment for a long time. Trophozoites, which can invade and penetrate the intestinal mucosa and damage epithelial cells and inflammatory cells, may occur after the ingestion of the cyst form (Chou & Austin, 2023). This is illustrated in Figure 4.

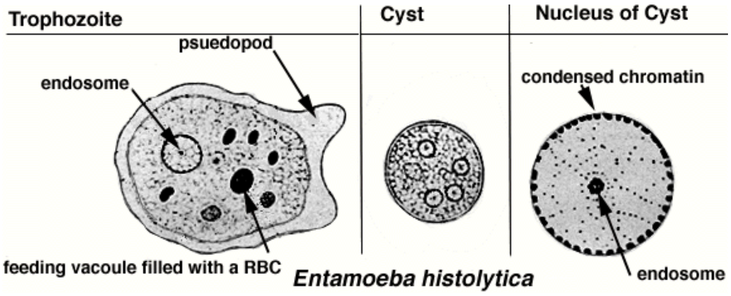


Figure 4 Illustration of the two forms of *E. histolytica*; trophozoite and cyst. Illustration by G. Karki (Karki, 2017).

Three main virulence factors, Gal/GalNAc lectin, amebapore, and proteases, are used by pathogenic amoebas to lyse, phagocytose, kill, and destroy a range of cells and tissues in the host (Padilla-Vaca & Anaya-Velazquez, 2010).

1.4 *Cryptosporidium pavrum*

Cryptosporidiosis is a global infection caused by *Cryptosporidium pavrum*, a type of protozoan parasite that can infect various vertebrate species, including humans (Current & Garcia, 1991). This infection leads to symptoms such as acute gastroenteritis, abdominal pain, and diarrhea (Gerace et al., 2019). The primary mode of transmission for cryptosporidiosis is through the fecal-oral route, which means that it is mainly contracted by ingesting viable oocysts (infectious forms) from contaminated food or water of animal or human origin (Fayer et al., 2000). Whilst waterborne transmission of this pathogen is well-documented, the natural reservoir and the exact route of infection for *Cryptosporidium* are not yet fully understood (Khalil et al., 2018).

The life cycle of *C. pavrum* begins with the ingestion of its hardy oocysts, illustrated in Figure 5. The oocysts release infective sporozoites that invade the host's small intestine cells (Smith et al., 2005). Within the cells, trophozoites multiply asexually, producing daughter cells called merozoites (Leitch & He, 2012). These merozoites cause damage as they invade new cells. Some merozoites differentiate into sexual forms, undergo fertilization, and form oocysts. The oocysts are excreted in feces and can contaminate water or food, transmitting the parasite to new hosts. The life cycle takes about 7-14 days (Lendner & Daugschies, 2014).

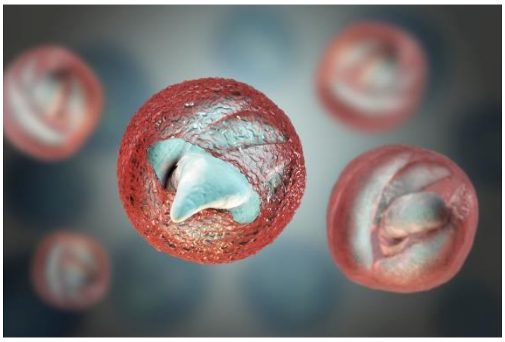


Figure 5 Illustration of *Cryptosporidium pavrum* oocysts. Copyright © 2018 Kateryna Kon/Shutterstock

However, humans are only infected by a limited number of *Cryptosporidium* species, with *C. parvum* and *Cryptosporidium hominis* being the most found (Bouzid, 2014; Bouzid et al., 2013). These species are known for their remarkable resistance, primarily attributed to their environmentally durable oocysts, which can withstand most water disinfection procedures and endure in aquatic environments for several months (Adeyemo et al., 2019; Venczel et al., 1997). The oocysts possess a spherical shape with a diameter of 4 to 6 µm and exhibit a sturdy wall composition rich in complex polysaccharides (Jenkins et al., 2010; Rossle & Latif, 2013). Even a minimal ingestion of 30 *Cryptosporidium* oocysts can lead to the onset of profuse watery diarrhea, and the infection tends to be far more severe in individuals with compromised immune systems (Guerrant, 1997).

1.5 Protein structure

Proteins are complex molecules composed of amino acids that play important roles in cell structure, enzymatic reactions, transportation, immune defense, and numerous other biological processes (Sanvictores & Farci, 2023). By knowing the function of one protein in a specific organism, one can get an insight into possible similar functions of proteins in other organisms based on the identification of their consensus sequences and comparing their positions of motifs and domains (Schaeffer et al., 2016; Xiong, 2006).

1.5.1 Domains

A structural domain refers to a distinct component within a protein's overall structure, characterized by its stability and ability to fold independently from the rest of the protein chain (Wang et al., 2021). While many domains are not exclusive to the products of a single gene, they can be found in various proteins. Proteins that share multiple common domains are typically encoded by genes belonging to evolutionarily related families, known as gene families (Bergtrom, 2022).

The nomenclature of domains often derives from their significant biological functions within the proteins to which they belong. For example, a domain may be named after its prominent role, such as the cathepsin propeptide inhibitor domain (I29), or it may be named after its discoverer. Domains can include for example transmembrane domains and ligand-binding domains. The process of domain swapping is a natural genetic phenomenon that gives rise to the formation of gene families and superfamilies (Libretexts, 2021).

1.5.2 Motifs

Protein motifs are conserved regions within the three-dimensional structure or amino acid sequence of proteins that are shared among different proteins, but they are shorter than domains. They represent identifiable patterns within protein structures, which may or may not be determined by a distinctive chemical or biological function (Xiong, 2006). Examples of motifs are ERFNIN and GNDF which exhibit conservation within the cathepsin L sub-family of proteases belonging to the papain family. These motifs serve as crucial mediators of prodomain inhibitory activity (Pandey et al., 2009).

1.5.3 Virulence factors

The ability of an organism to infect the host and spread disease is described as virulence (Leitão, 2020). The chemicals known as virulence factors help the pathogen colonize the host at a cellular level (Brock et al., 2003). These microbial components can have a secretory, membrane-related, or cytosolic character (Sharma et al., 2017). Certain enzymes that are considered virulence factors actively target and harm the host's components, leading to tissue damage and an environment contributing to microbial infections (Nash et al., 2015). Enzymes with virulence factors, such as proteases, neuraminidases, and phospholipases, cause cellular damage and break down substances into smaller components that microbes can use as nutrients (Zachary, 2017). Additionally, these enzymes modify the host's cellular receptors, disrupting the binding of their usual ligands like complement. This alteration affects microbial behavior, promoting invasiveness, serum resistance, and evasion of the host's immune system (Casadevall & Pirofski, 2009).

1.5.4 Proteases

Proteolytic enzymes, also known as proteases, play a vital role in degrading proteins and can be found in various organisms, including viruses, humans, animals, and plants (Razzaq et al., 2019). These enzymes have crucial functions in protein synthesis and turnover, allowing proteins to regulate biological processes (Mótyán et al., 2013). In the case of pathogens, proteases are essential for their biological processes and life cycles as they participate in the conversion of newly formed molecules into active forms and the inhibition of protein activity (Figaj et al., 2019). As a result, proteases hold significant relevance in the realm of medical and pharmacological research and development (Fairlie et al., 2000; Wlodawer, 2002). Similar types of proteases are used by infectious organisms like viruses, bacteria, and other parasites, and establish around 1% of their genomes. Once they are present in infected mammalian hosts, the proteases compete for host resources and work with cellular machinery to prolong infectivity (Tyndall et al., 2005). One strategy for battling infectious disease is to selectively suppress foreign proteases within host cells, which will slow the rate of reproduction of infectious organisms and help the body's immune system fight them off (Barrett, 2000). To avoid inhibiting highly similar host proteases necessary for normal host physiology, the goal of protease inhibitor design is to construct powerful inhibitors of the harming foreign or mammalian protease (Ranasinghe & McManus, 2017; Tyndall et al., 2005). This highlights their prevalence and importance in the context of infectious diseases. Furthermore, proteases find extensive use as enzymes in various fields of biotechnology and industry, and their function is required in numerous research applications. The versatility and widespread presence of proteases make them valuable tools for scientific investigations and practical applications alike (Mótyán et al., 2013). Proteases are categorized as exoproteases or endoproteases based on their cleavage location in proteins. Exoproteases, located at the N- and C-terminal ends, remove one amino acid at a time, stopping protein activity, while endoproteases cut internally near specific amino acids (van der Velden & Hulsmann, 1999).

Additionally, proteases help defensive mechanisms because they stimulate the immune system when they recognize peptide fragments of foreign proteins (López-Otín & Bond, 2008).

Cysteine proteases exhibit cytopathic effects on host cells and are regarded as virulence factors in various protozoan parasites (Yang et al., 2023). They play a significant role in differentiation, development, and pathogenicity (Puthia et al., 2005). Cysteine proteases are mostly found in the lysosomes and participate in phagocytosis, whose function is the removal and digesting of extracellular material in cells (Britannica, 2022). For the family C1 peptidases, the lysosomal system of eukaryotic cells and the digestive vacuoles of protozoa both benefit from proteolytic action (Rawlings, 2018).

1.5.5 Cysteine protease in *Blastocystis* spp.

Even though *Blastocystis* spp., was discovered a very long time ago and has been studied for many years, there is limited data on the pathogenic effects on the host cells. Several studies in the field of parasitology have given information on the signaling mechanisms triggered by cysteine proteases produced by *Blastocystis* (Ajjampur & Tan, 2016; Tan, 2008). The studies have explored the effect of cysteine proteases on mRNA expression of certain inflammatory cytokines, such as interleukin (IL)-1β and Tumor Necrosis Factor-alpha (TNF-α), and their relationship with mitogen-activated protein kinases (Lim et al., 2014). Pathogenicity of different subtypes varies, and several studies have revealed that ST7 exhibits significantly higher cysteine protease activity when compared to ST4 (Mirza & Tan, 2009; Wu et al., 2014). Furthermore, another recent study investigated the molecular mechanisms by which *Blastocystis* activates IL-8 gene expression in human colonic epithelial cells. In the context of the mentioned study, cysteine proteases derived from a zoonotic isolate *Blastocystis ratti* WR1 (ST4) were shown to activate IL-8 gene expression where it was also identified the involvement of NFκB activation in the production of IL-8 (Puthia et al., 2008). *Blastocystis* ST7 has also shown greater resistance to both anti-parasitic drugs and the host's innate immune response compared to ST4 (Mirza et al., 2011; Yason et al., 2016; Yason et al., 2018)

1.5.6 Cysteine protease in *Giardia intestinalis*

Studies conducted on *Giardia intestinalis* have revealed that cathepsins possess the ability to dampen a specific aspect of the proinflammatory response elicited by the host, which is initiated by an independent proinflammatory stimulus (Cotton et al., 2014). Cathepsin B (catB) belongs to the papain family of lysosomal cysteine proteases (Cavallo-Medved et al., 2011). Its role includes intracellular protein degradation. However, under specific circumstances, it might participate in various physiological processes, including antigen processing in immune responses, hormone activation, and regulation of bone turnover (Mort & Buttle, 1997). Research has shown results that suggest initiating apoptosis, CPs in *G. intestinalis* also affect the permeability of the intestinal epithelial barrier of the host, leading to its harm (Cotton et al., 2014; Liu, 2019). Additionally, it was established that these modifications may be mediated by caspase-3 (A. C. Chin et al., 2002). Research has also indicated that CPs in *Giardia*, through rearranged junctional proteins and chemokine degradation, may disrupt intestinal epithelial barriers and modulate the immune response. (Liu, 2018). Cysteine proteases exhibit the presence of a catalytic dyad composed of active-site cysteine (Cys) and histidine (His) residues, but also as a catalytic triad including the active site of an asparagine (Asn) (Drag, 2013; Kermasha & Eskin, 2021). CatB proteases can also have an additional segment called the occluding loop, which consists of 20 amino acids and makes it unique compared to other cysteine cathepsins (Illy et al., 1997; Renko et al., 2010).

1.5.7 Cysteine protease in *Entamoeba histolytica*

Cysteine proteinases serve as an important component in the virulence of *E. histolytica* (referred to as EhCPs), contributing to several roles in infection and invasion (Betanzos et al., 2019). *E. histolytica's* genome has 80 genes that code for proteases, including 50 CPs from the papain superfamily (He, 2010). Among these genes, research has observed that the main cysteine proteases in *E. histolytica*, EhCP1, EhCP2, EhCP5, and EhCP7 are strongly expressed in *E. histolytica* exhibit the highest level of up-regulation and play a critical role as virulence factors (Irmer et al., 2009). Through their capacity to break down extracellular matrix proteins as well as mucin 2, the main component of colon mucus, CPs have a direct role in tissue invasion (Que & Reed, 2000; Thibeaux et al., 2014). By weakening host antibodies and complement, they are crucial for immunological evasion (Faust & Guillen, 2012). The most likely candidate among the *E. histolytica* CPs implicated in the pathogenic process is EhCP5, given that it is specific to the parasite, localizes at the amoebic surface, and participates in human colon invasion (Ankri et al., 1999). The propeptide region of EhCP5 has an RGD integrin-binding motif (Arg-Gly-Asp), which has also been discovered in the proregion of cathepsin X from higher eukaryotes (Lechner et al., 2006). RGD motifs act as ligand recognition sites for cell surface receptors like the integrins in cell adhesion proteins like fibronectin (Marquay Markiewicz et al., 2011). EhCP2 is a membrane-associated cysteine protease and has the structure of cathepsin L (Que et al., 2002). Phagocytosis results in the passive release of EhCP1, EhCP2, and EhCP5. Given that the release of all these three proteinases was increased when EhRab11B was overexpressed, this release is most likely a component of the recently identified EhRab11B-associated secretory pathway (Meléndez-López et al., 2007).

The various and important roles of cysteine proteases during infections include helping the attachment of the pathogen, degradation of the tissue structure, breaking down host proteins to evade the immune system, activating proteolytic processes in host cells, and assisting the spread of infection to produce new infection sites (Cuellar et al., 2017; He, 2010; Que & Reed, 2000). The cysteine proteases evade the host immune response by cleaving secretory immunoglobulin A (sIgA), and immunoglobulin G (IgG) , and activating complement mechanisms (Que & Reed, 2000).

1.5.8 Cysteine protease in *Cryptosporidium pavrum*

There are about 20 genes that encode clan CA CPs (papain-like cysteine proteases) in *C. pavrum* Iowa strain, where 3 of them are cathepsin L-like and 2 are cathepsin B-like members from the C1 family, collectively referred to as "Cryptopains" (Abrahamsen et al., 2004). The cysteine protease functions in *C. pavrum* have been detected in various developmental phases, potentially playing a role in the parasite's excystation and invasion of host cells. Cryptopain-1 shares structural characteristics with enzymes from the papain family, specifically cathepsin L-like enzymes, and has been shown to have virulent functions (Na et al., 2009). This area of virulence factors in *C. pavrum* has not received extensive research attention and requires further investigation to gain a deeper understanding.

1.6 Virulence factors

The genes studied in this project all fall under the category of papain-like proteases (PLP), displaying both structural and enzymatic similarities with papain. Papain-like proteases possess a shared catalytic dyad active site, characterized by a cysteine amino acid residue that functions as a nucleophile (Novinec & Lenarčič, 2013). The prodomain of eukaryotic cathepsins possesses the following two distinct and well-defined functions. Firstly, to preserve the enzyme in an inactive form (zymogen) until it reaches the appropriate site for protease activity, and secondly to serve as a structural template that ensures proper folding during translation (Coulombe et al., 1996).

1.6.1 Cathepsin B in *Giardia intestinalis*

Exposure of trophozoites to host-derived soluble factors (HSF) has been shown to up-regulate the expression of the cysteine protease like CatB (GL50803_16779) as studied as a virulence factor in this project (Argüello-García & Ortega-Pierres, 2021; Emery et al., 2016). Research has revealed that catB degrades the proinflammatory secretion of CXCL8 (C-X-C Motif Chemokine Ligand 8) from intestinal epithelial cells when exposed to proinflammatory stimuli derived from either the host or pathogens. Additionally, the degradation of CXCL8 leads to a reduction in CXCL8-induced chemotaxis of polymorphonuclear neutrophils (PMNs) (Cotton et al., 2014).

The expression of catB, along with cystatin, an inhibitor of cathepsin proteases, is increased in the presence of released host factors (Emery et al., 2016). Cystatins have been observed to regulate the host immune response in parasitic nematodes, suggesting their potential involvement in controlling host defenses (Ochieng & Chaudhuri, 2010). Figure 6 shows the structure of a cystatin and the binding to a cysteine protease. Furthermore, cystatins may also act as internal regulators of parasite cathepsins (Khatri et al., 2020). Remarkably, the *Giardia* cystatin exhibits a phylogenetically simple nature compared to eukaryotic cystatins and shows a closer resemblance to bacterial counterparts (Emery et al., 2016).

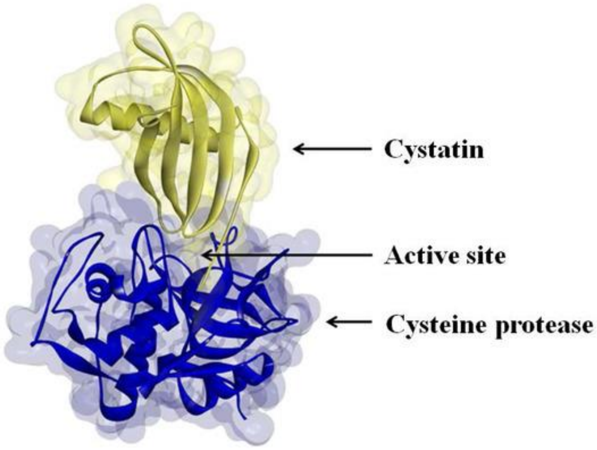


Figure 6 Structure of cysteine protease and its inhibitor cystatin binding to its active site (Vorster et al., 2013).

Cathepsin B has been identified as a virulence factor in *Blastocystis* ST7. Studies have demonstrated an improved paracellular permeability of intestinal Caco-2 cell monolayers in response to this discovery (Nourrisson et al., 2016). No research has been documented for CatB in *Blastocystis* ST4, making it interesting to compare it with Cathepsin B from *G. intestinalis* and *Blastocystis* ST7.

1.6.2 Cysteine proteinase 2 and 5 in *Entamoeba histolytica*

The virulence factors studied in this project from *Entamoeba histolytica* are EhCP2 and EhCP5. Both have been demonstrated through research to promote invasion and virulence, initiating a pro-inflammatory response in macrophages, breaking down tight junctions as well as degrading extracellular matrix components like fibronectin, collagen, and laminin (Espinosa-Cantellano & Martínez-Palomo, 2000; Siqueira-Neto et al., 2018). From a study of

E. histolytica, researchers discovered that EhCP2 is expressed at high levels and is actively incorporated into phagocytic vesicles (Bruchhaus et al., 2003; Que et al., 2003).

EhCP2 has also been shown to be accountable for cleaving CXCL8 into a more potent isoform that intensifies PMN chemotaxis which can damage host tissues (Junger, 2008; Pertuz Belloso et al., 2004). However, the overexpression of EhCP2 led to a significant increase in the destruction of *in vitro* monolayers (Hellberg et al., 2001). These findings support the important role of these molecules in initiating cell damage caused by *E. histolytica*.

EhCP5 includes roles such as actively breaking down the colon's protective mucin barrier, facilitating tissue invasion through the degradation of extracellular matrix proteins, and serving as a pivotal player in immune evasion by breaking down host antibodies and complement factors (Hou et al., 2010). From a previous study on *E. histolytica* in mice, it was observed that EhCP5 has an impact on proinflammatory responses and the modification of tight junction permeability (Kissoon-Singh et al., 2013). They also found that the data further indicates that changes in intestinal permeability in the absence of direct contact are mediated primarily by the virulence factor EhCP5, which is secreted by live parasites.

Research done on a cysteine proteinase-deficient amoeba showed unsuccessful induction of intestinal epithelial cell production of the inflammatory cytokines IL-1β and IL-8 and also caused significantly less gut inflammation and damage to the intestinal permeability barrier (Zhang et al., 2000). These studies suggest that the acute host response and amebic invasion result from a complex interaction of parasite virulence factors and host defenses.

1.6.3 Cryptopain-1 in *Cryptosporidium*

From two studies (Na et al., 2009; Siqueira-Neto et al., 2018) cryptopain-1 was assumed to be a virulence factor with a suggested factor in host cell invasion and is used in this study to compare with *Blastocystis* ST7 and ST4. In comparison to cathepsin L-like cysteine proteases, cryptopain-1 possesses distinctive structural and biochemical characteristics. Notably, it does not require a pro-domain for proper folding, setting it apart from related enzymes (Na et al., 2009). The selective breakdown of collagen and fibronectin suggests that cryptopain-1 likely has a biological function in facilitating the invasion and release of the parasite from host cells (Argüello-García et al., 2023).

Cryptopain-1 displays unique characteristics that distinguish it from typical papain-family enzymes. These include an extensive pro-domain, a transmembrane domain near the beginning of the protein, and unique insertions at the start of the mature domain sequence

(Argüello-García et al., 2023; Na et al., 2009). In contrast, conventional papain-family cysteine proteases share similar structural properties, consisting of a signal peptide, a pro-peptide, and a catalytic domain that represents an active form of the enzyme (Faheem et al., 2016). Figure 7 shows a model of the protein taken from the mentioned study by Argüello-García et al. (2023). Although the enzymes' pro-domain sequences have greater variability than the mature domain, specific regions within the pro-domain (ERFNIN and GNFD motifs) show relatively high conservation. These motifs play a significant role in these enzymes' processing and folding mechanisms (Na et al., 2009).

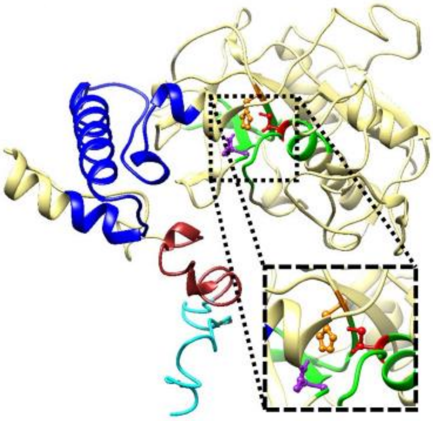


Figure 7 Protein model of Cryptopain-1. The catalytic triad Cys-His-Asn is displayed in ball-and-stick conformation and is magnified within dotted squares. From analyses, Cryptopain-1 are predicted to be membrane anchored (Argüello-García et al., 2023).

1.7 Bioinformatics

1.7.1 NCBI

The National Center for Biotechnology Information (NCBI) is an integral component of the United States National Library of Medicine (NLM) and plays a pivotal role in the realm of molecular biology. As a significant resource for researchers and scientists, the NCBI offers an extensive collection of databases, tools, and resources that are indispensable for investigating, accessing, and analyzing biological information and are frequently used in this study.

1.7.2 Databases

The GenBank database, which acts as a huge compilation of publicly accessible DNA sequences collected from various organisms, is one of the NCBI's main services. This comprehensive collection enables researchers to study genetic information, trace evolutionary relationships, and gain insights into various biological processes. Additionally, the NCBI provides access to databases like PubMed, which contains a wealth of biomedical literature, and the Protein Database (PDB), which houses three-dimensional structures of proteins. The NCBI's suite of tools includes the widely used BLAST (Basic Local Alignment Search Tool, https://blast.ncbi.nlm.nih.gov/Blast.cgi), which enables researchers to perform sequence similarity searches, helping in the identification of genes, proteins, and other biological molecules (McGinnis & Madden, 2004).

The online database MEROPS (https://www.ebi.ac.uk/merops/index.shtml) can be used to find information about proteases (Rawlings, 2018). SMART (Simple Modular Architecture Research Tool) is a database (http://smart.embl-heidelberg.de/) used to identify and annotate genetic mobile protein domains and analyze protein domain architectures (Letunic I & Bork.). MEGA (Molecular Evolutionary Genetics Analysis) is a software offering many tools, including sequence alignment and the ability to create phylogenetic trees (Tamura K, 2021).

2 Methods

2.1 Bioinformatic part

2.1.1 Collecting relevant genes

Virulence factors from specific organisms were found in research and articles from PMC (PubMed Central) and other online websites. NCBI is used to access information about genes and sequences. GenBank was used to identify both protein and genomic sequences of interest as well as to retrieve the sequences in a FASTA format (NCBI, n.d.). These sequences were chosen from *Giardia intestinalis*, *Entamoeba histolytica,* and *Cryptosporidium pavrum* for additional examination and comparison with the organisms of interest; *Blastocystis* ST7 and ST4.

2.1.2 Search for proteins in other organisms

The protein sequence of interest was put in a BLAST search (Altschul et al., 1990), to determine if present in the *Blastocystis* genome. To accomplish this, the "tblasn" tool, which performs a search by translating the protein sequence into nucleotides, was utilized. The search is specifically directed towards the organism *Blastocystis* (taxid: 12967) to exclusively identify proteins within that particular organism. Upon obtaining potential matches, the subsequent phase entails analyzing the results using the expected value (e-value), also referred to as the false-positive rate, as well as the query score. A lower e-value signifies a greater degree of similarity between the proteins. The threshold for the e-value is established at 10. The query cover measures the proportion of the query sequence that overlaps the reference sequence. For the initial triage, a high query cover value is between 70% and 80%. Should a match meet the specified requirements, the corresponding encoded gene in *Blastocystis* is confirmed as the sought-after protein by cross-referencing the protein code and excluding *Blastocystis* from the search. This confirmation process ensures the accuracy of the identified protein of interest. The graphic summary of the search outcomes organizes the aligned sequences into distinct color-coded segments based on their alignment scores. Each bar corresponds to a segment of another sequence that exhibits similarity to the query sequence. A red bar signifies the highest degree of similarity (≥200), followed by pink indicating moderately good matches (80-200), green for less impressive matches (50-80), blue for the lowest scores (40-50), and finally, a black bar for poor hits (<40).

The proteins of interest from the specific organisms were searched for in the other organism's genome as well as *Blastocystis* for comparison.

2.1.3 Identification of domains and signal peptides

To find domains, the SMART database was used (http://smart.embl-heidelberg.de/). The protein sequence was put in search and with PFAM domains setting.

The online database SignalP 5.0 was used to determine signal peptides within the protein sequences. These short sequences at the N-terminal of proteins carry information for protein secretion and protein target location (Owji et al., 2018).

2.1.4 Identification of motifs and active sites

MEME database (https://meme-suite.org/meme/tools/meme) with a motif width of 8 and 15 residues respectively, and a maximum number of 15 motifs, keeping the rest of the parameters at default. To determine active sites, present in the protein structure, the protein database ScanProsite was used (https://prosite.expasy.org/). The sequence of interest was put in the search and active sites as well as other motifs were found.

2.1.5 Sequence alignment and phylogenetic tree

All protein accession codes for each protein search were put together in MEGA to look for conserved regions and to perform sequence alignment with MUSCLE. Then, the program BioEdit (Hall, 1999) was used to edit the sequences.

To determine the evolutionary relationships within the proteins of interest, molecular sequences were analyzed using an online platform phylogeny.fr (Dereeper et al., 2008). Firstly, various organisms were selected from the outcomes of the BLAST searches conducted for each protein from *G. intestinalis*, *E. histolytica*, and *C. pavrum*. Targeted searches were performed within specific kingdoms to discover their relationships, enabling a comparison of the organisms identified in specific and nonspecific hits. The chosen proteins were exported in FASTA formats, and a phylogenetic tree was constructed in phylogeny.fr with the mode "One click". Furthermore, the trees were edited using the online tool iTOL (Letunic & Bork, 2021) and colors were modified in the program Adobe Acrobat.

2.1.6 Vector Construction

The vectors utilized in the molecular part of this study were built and optimized by PhD candidate Mitchellrey Toleco. Three vectors were constructed: pMRTω, pMRTτ, and pMRTφ. The idea was to first clone each promoter into vector pMRTω, and the terminator into vector pMRTτ. If both are successful, then the final cloning into pMRTφ with both promoter and corresponding terminator.

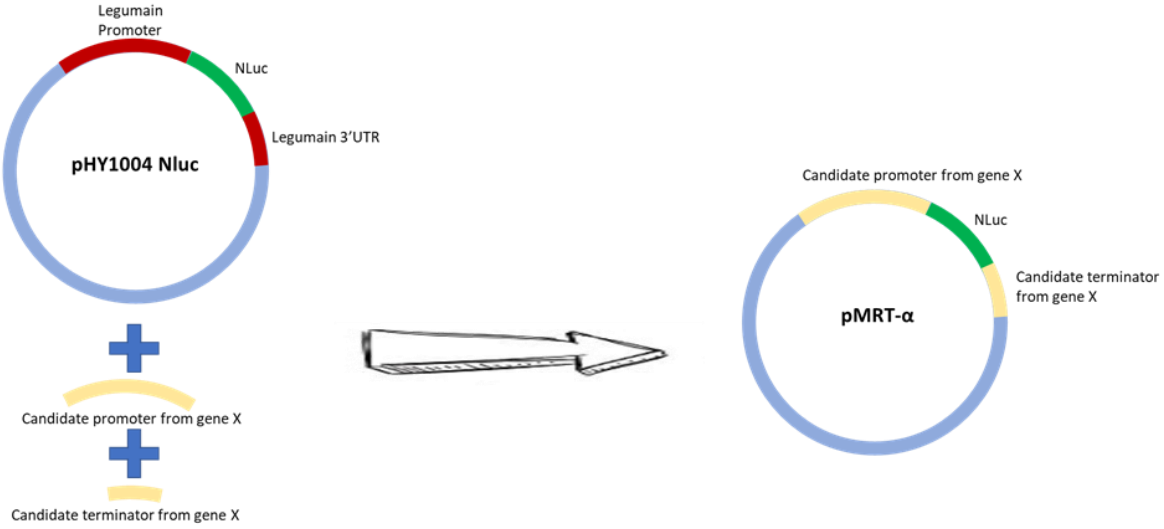

Legumain Promoter
NLuc
Legumain 3'UTR
pHY1004 Nluc
Candidate promoter from gene X
Candidate terminator from gene X
Candidate promoter from gene X
NLuc
Candidate terminator from gene X
pMRT-α

Figure 8 Illustration of the vector constructed for cloning of genes.

The software Benchling (https://www.benchling.com/) was used to create both promoters and terminators for the genes. For gene 60SRPL32, the promoters of the two fragments 1000 and 1295, and one terminator were made. For the PC1A gene, there were promoters consisting of three fragments: 643, 1263, and 2207, and terminator. Further primers and oligo bridges needed to build these vectors in the lab were also created in Benchling and ordered from Thermo Fischer.

When designing primers, it is important to follow some essential rules and guidelines focusing on considerations such as primer length, melting temperature, and specificity, amongst other important factors to succeed. The length of the primer should be around 18-24 nucleotides (Addgene, n.d.). Shorter primers could lack specificity, while longer primers may have reduced annealing efficiency (ThermoFisher, 2019). The melting temperature (T_m) is the temperature at which the DNA duplex formed by the primers and the target DNA separates, and should be 50-60 °C for each primer, and the coupled primers should have a range of ±5 °C (Javed, 2022). The percentage of guanine (G) and cytosine (C) in primers is known as the GC content. Adenine (A) and thymine (T) form two hydrogen bonds, while GC base pairs form three hydrogen bonds when primers anneal to their target sequence. Since three hydrogen bonds are stronger than two, it takes more energy to separate G and C than A and C (Bera & Schaefer, 2005). The single DNA strands (ssDNA) will bind more firmly to the primer with a higher GC content, raising the T_m. Higher GC content can result in mismatches and the production of primer-dimers, which is the hybridization of two primers with each other, which could be problematic during PCR (Polymerase Chain Reaction) (Naz & Fatima, 2013). Therefore, aiming for a GC content of 40–60% to keep the stability and specificity (Kumar & Kaur, 2014). The start and the end of the primer should also contain 1-2 G/C pairs.

2.1.7 Sanger sequencing

The "chain termination method," often known as Sanger sequencing, is a widely used technique for determining the precise order of nucleotides (A, T, C, and G) in a DNA molecule (*Sanger Sequencing Steps and Method.*, n.d.). Sanger sequencing was used to confirm genomic sequences and is done automatically way by using a sequencing machine. All sequences in this study were sent to a company that performed this work for us. Each technique has three fundamental steps that are listed below in Figure 9.

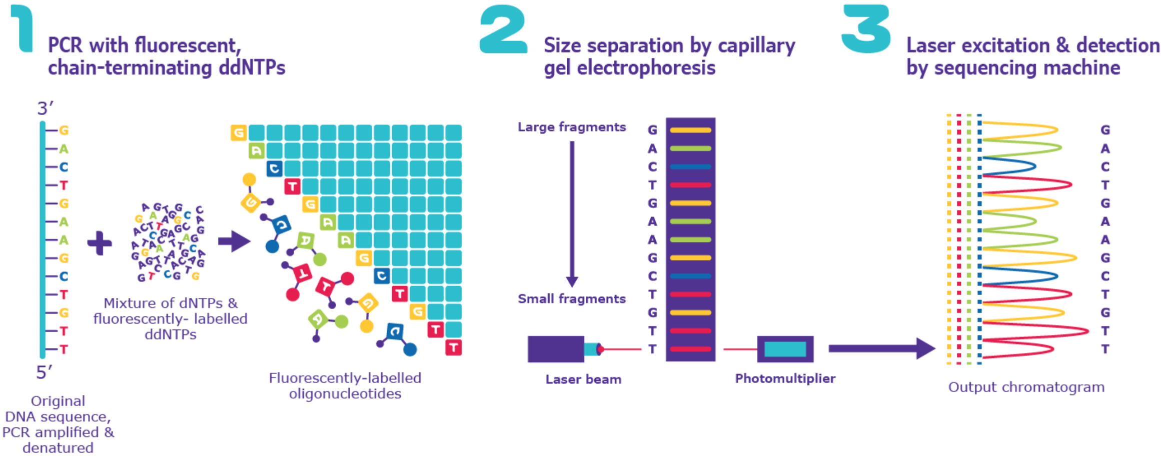

1
PCR with fluorescent, chain-terminating ddNTPs
3′
G A C T G A A G C T G T T
5′
Original DNA sequence, PCR amplified & denatured
Mixture of dNTPs & fluorescently- labelled ddNTPs
Fluorescently-labelled oligonucleotides
2
Size separation by capillary gel electrophoresis
Large fragments
Small fragments
G A C T G A A G C T G T T
Laser beam
Photomultiplier
3
Laser excitation & detection by sequencing machine
G A C T G A A G C T G T T
Output chromatogram

Figure 9 Method of Sanger sequencing in steps (*Sanger Sequencing Steps and Method.*, n.d.)

To check that the sequences being evaluated were correct, Benchling was used to align the sequences. Sequences could be worked on further if they had no or few mismatches.

2.2 Molecular part

2.2.1 Isolate promoters and terminators

Table 1 presents the genes used for the experimental part, the fragment version, forward primer (FP) and reverse primers (RP) employed, and the corresponding sizes for each fragment in the base pair.

Table 1 List of promoters and terminators for the genes PC1A and 60SRPL32. The primers used for each version are listed below with the forward primer (FP) and reverse primer (RP), and the size in base pair.

Gene	Version	FP	RP	Size, bp
PC1A	V643	134	130	643
PC1A	V1263	129	130	1263
PC1A	V2207	133	130	2207
PC1A	Terminator	135	136	500
60SRPL32	V1000	13 7	132	1000
60SRPL32	V1295	138	132	1295
60SRPL32	Terminator	139	131	500

Table 2 illustrates the PCR recipe, including all reagents utilized and their respective quantities in µl. The template that was used is genomic DNA from *Blastocystis* ST7.

Table 2 The PCR mix recipe uses Platinum buffer and polymerase. The total volume for each reaction is 50 µl. *genomic DNA from *Blastocystis* ST7

Reagents	Amount (µl)
5x Platinum Buffer	10
10 µM FP	2.5
10 µM RP	2.5
10 µM dNTP's	1
Template*	1
DMSO	1.5
Platinum Polymerase	0.25
NFW	31.25
Total volume	50

The PCR program was set with the settings shown in Table 3. Each step has an optimized temperature and duration, where the steps of denaturation, annealing, and extension were repeated in 25-35 cycles. This depends on the length of the fragments.

Table 3 Program used for the PCR reaction with the temperature used for each step and the duration of each step. *Steps 2-4 were repeated 25-35 times depending on the length of the fragments.

PCR Program steps	Temperature (°C)	Time
Initial denaturation	98	10 sec
Denaturation*	98	10 sec
Annealing*	72	30 sec
Extension*	72	2 min
Final extension	60	10 min
Hold	12	∞

2.2.2 Agarose gel electrophoresis

Agarose gel was prepared with a ratio of 1 g agarose per 100 ml 1 x TAE buffer in reagent bottles. GelRed (10000X) was added for bands to visualize in the gel. In a 500 ml reagent bottle of agarose gel, 5 µl of GelRed was added. This was heated until the agarose was completely dissolved and poured into the cast until set. The gel ran for approximately 40 minutes at 110V (Volt) before confirming bands in UV light.

2.2.3 Cloning of genetic constructs

For assembly and transformation, NEBuilder HiFi Assembly was used. The calculation of the assembly mix is to first find the mass (ng) with a pmol of 0.05.

$$\text{moles of dsDNA} \times \left(\text{length (bp) of dsDNA } \times \frac{617.96\,\frac{\text{g}}{\text{mol}}}{\text{bp}} \right) + 36.04\,\frac{\text{g}}{\text{mol}}$$

Where 617.96 g/mol/bp is the average molecular weight of a base pair and 36.04 g/mol are the two -OH and two -H added back to the ends. These are both constant values.

Then the volume (µl) was determined from the DNA concentration (ng/µl) and mass (ng). DNA concentration used is 30-50 ng.

$$\frac{\text{Mass (g)}}{[\text{DNA}](\frac{\text{ng}}{\mu\text{l}})}$$

This is calculated for the insert (promoter/terminator) and the backbone (Φ and τ). All the calculated volumes are summed together (DNA, backbone, and ssOB mix), and the total amount will be the same for the NEB buffer. When the assembly is performed for more than 4 fragments, the recommendations of 1-hour incubation, which was utilized.

For transformation, competent *E. coli* cells were used. Assembly product was mixed with the competent *E. coli* cells, chilled on ice before heat shocked at 42 °C and put on ice again. Then 150 µl was spread and incubated on agar plates with ampicillin. The remaining product was spun down and spread on a separate plate to grow plates with both high and low concentrations. Plates were incubated overnight at 37 °C.

After overnight incubation, 20-30 colonies were picked on the agar plates, half from each transformation plate, for each construct assembly and screened by colony-PCR (cPCR) using DreamTaq Green Master mix (5X) and with the recipe and settings shown in Table 4.

Table 4 Colony PCR recipe with a total volume of 15 µl. The colony is taken from an agar plate put in a PCR tube and resuspended with a pipette tip.

Reagents	Amount (µl)
5x DreamTaq Buffer	1.5
10 µM FP	0.2
10 µM RP	0.2
10 µM dNTP's	0.15
Colony	-
DreamTaq Polymerase	0.15
NFW	12.8
Total volume	15

The program used for colony PCR is slightly different from the PCR of the fragments. The parameters for each stage are shown in Table 5, along with the temperature and time.

Table 5 Overview of colony PCR program with temperature and duration of each step involved. *Steps 2-4 were repeated 28 times.

PCR Program steps	Temperature (°C)	Time
Initial denaturation	95	5 min
Denaturation*	95	30 sec
Annealing*	55	30 sec

Extension*	72	45 sec
Final extension	72	5 min
Hold	12	∞

To confirm the correct sizes of constructs and quality, gel electrophoresis was used on cPCR products. GelRed was mixed with 1% agarose gel and run at 100V for approximately 40 minutes. Generuler DNA Ladder Mix was used to verify the length of the bands.

2.2.4 Plasmid purification miniprep

To confirm the presence of the colony, gel electrophoresis was performed on the product obtained from the cPCR. Subsequently, purification of the colony was carried out before its verification using Sanger sequencing for a second confirmation. Purification was performed by using a kit from Qiagen (QIAGEN®, 2021). The initial step involved the preparation of cells by selecting a colony from an agar plate and transferring it to 5 ml of LB (Lysogeny Broth) containing the antibiotic ampicillin. The mixture was then subjected to overnight shaking at 300 rpm and maintained at a temperature of 37 °C. Following the overnight incubation, the cell culture was harvested by centrifugation at maximum speed (approximately 13 rpm) for 1 minute. The resulting pellet was subsequently dried on the laboratory bench and then resuspended in 300 µl of Buffer P1. Afterward, Buffer P2 was added by gently inverting the tube 4-6 times, and the mixture was allowed to incubate at room temperature for 5 minutes. Buffer P3 was added and mixed immediately by inverting and then incubated on ice for 5 min for a white fluffy consistency. Subsequently, the mixture was centrifuged at maximum speed for 10 minutes to obtain a clear supernatant. A column with a mixture was then put into the tube and 1000 µl QBT was added and flowed through by gravity. After that, the supernatant was added and entered through the column by gravity flow before washed with 2000 µl Buffer QC and repeating the washing. DNA was eluted with 800 µl Buffer QF.

2.2.5 Nanodrop protocol

To measure the concentration of the DNA samples, NanoDrop (NanoDrop One, Thermo Scientific) was used. To prepare the instrument, the cleaning of the pedestal was done thoroughly by applying UltraPure water and then gently wiping it off with filter paper (lint-free, VWR). Blank was set with the buffer used.

2.2.6 Plasmid purification maxiprep

The experimental protocol employed in this study follows the same principle as the mini prep, to obtain higher concentrations achieved by utilizing larger volumes in the maxiprep. The kit employed was Zymopure (ZymoResearch, 2022). To accommodate the increased volumes and enhance efficiency, a vacuum manifold was used.

The initial step involves preparing the cells through cell cultivation. A small culture of 5 ml LB mixed with ampicillin was prepared in test tubes. A clone was then selected from an agar plate using a pipette tip and transferred into the broth. Following an incubation period of 12-14 hours at 37 °C and 250 rpm, 200 µl of the small culture containing the grown bacteria was transferred to an Erlenmeyer flask containing prepared LB with ampicillin (100 ml). Subsequently, the large culture was incubated overnight in a shaker at 37 °C at 250 rpm. To harvest the cells, centrifugation was performed at 7000 rpm for 10 minutes.

2.2.7 Storage of samples

To ensure the prolonged preservation of bacterial cultures, glycerol stock was utilized as a secure storage method to sustain the viability and shelf life of bacterial cells. The inclusion of glycerol plays a crucial role in stabilizing the frozen bacteria and prevents the degradation of their cell membranes, thereby ensuring their long-term storage (Schaudien et al., 2007). Glycerol stocks can be safely maintained at -80 °C for numerous years. To prepare the glycerol stock, 50% sterile glycerol was used. Single colonies of confirmed clones were selected for each stock and incubated in 800 µl of LB overnight at 37 °C. The following day, the bacterial culture was mixed with 800 µl of glycerol, inverted 5-10 times to ensure thorough mixing, and subsequently stored at -80 °C. Ensuring the growth of the stock is crucial, and this can be achieved by transferring a sterile loopful of the stock onto an agar plate for incubation.

2.2.8 Cell strain and culture

Culturing cells of *Blastocystis* can be a challenging task due to the organism's anaerobic nature, and they can easily perish if their specific requirements are not met. However, a commonly followed procedure involves reducing the cells in Iscove’s modified Dulbecco’s medium (IMDM) with 10% horse serum at 37 °C under anaerobic conditions. Alternatively, calf serum can also be used as a supplement. To maintain the culture in an anaerobic environment, oxoic jars equipped with sachets to generate anaerobic gas are employed. This creates an atmosphere with less than 1% oxygen and an appropriate level of carbon dioxide, which is ideal for the growth of anaerobic organisms (Ho et al., 1993; Zhang et al., 2012).

Once the *Blastocystis* cells have reached the log phase of growth, they are harvested through centrifugation at 1000 g for 10 minutes at room temperature. Before transfection, it is essential to wash the cells with a pre-reduced incomplete cytomix buffer and then resuspend them in a pre-reduced complete cytomix. To assess the electropore-forming rate, the cells are stained with 5 µg/ml of the cell-impermeable dye propidium iodide (PI). This staining method helps determine the efficiency of electroporation. By utilizing a cytomix buffer, chemically permeabilized cells can be effectively maintained with an ionic composition like that found in intracellular cells.

2.2.9 Cell transfection

Although not many *Blastocystis* transformations have been carried out, a successful method developed and made public by Kevin Tan (University of Singapore) is noteworthy (Li et al., 2019). The transfection was employed with *Blastocystis* isolate ST7. The method was not pursued in this study for unfortunate reasons, but a description serves to provide an understanding of its potential application and relevance.

Blastocystis cells mixed with 25 µg of plasmid DNA in a 0.4 cm transfection cuvette to introduce plasmid DNA into the *Blastocystis* cells, is a technique called electroporation can be employed using a Bio-Rad Gene Pulser system. Electroporation involves the application of an electric pulse to create temporary pores in the cell membrane, allowing the plasmid DNA to enter the *Blastocystis* cells. By subjecting the mixture to a single pulse of electricity, the plasmid DNA was delivered into the *Blastocystis* cells, facilitating genetic modification or analysis.

The optimal settings are determined from the article to be 370 V and 30 ms (time constant). With these settings, an electropore-forming rate of 94.3% and a survival rate of 9.4% was found. As *Blastocystis* cells are large, a recommendation of low voltage is more likely to be successful. The optimal number of cells used for each transfection is 2 x 10^7, and if increased, there could be an increased probability of arcing. Using 10^8 cells, 370 V and 20 ms showed results with ideal transfection efficiency (Li et al., 2019). If there are more cells, the electroporation time should be decreased. The ideal transfection program for two different cell counts is shown in Table 6, from the study developed by Tan (2019).

Table 6 Program for an optimal cell transfection with 2x10^7 and 2x10^8 cells developed by Kevin Tan (Li et al., 2019).

Cells for transfection	Plasmid DNA	Voltage, V	The time constant, ms
2 x 10^7	25 µg	370	30
10^8	100 µg	370	20

After the transfection procedure, cells can be enumerated by using a hemocytometer under fluorescence microscopy.

The percentage of positive PI (propidium iodide) cells post-transfection needs to be calculated, and to determine the survival rate, similar batches of cells without PI staining are also subjected to electroporation and stained with PI after 12 hours' incubation to allow membrane sealing. The PI-negative cells are to be counted and the survival rates are to be calculated. The hypothesis is that cells survive electrotransfection more in a cytomix rather than in phosphate-buffered saline (PBS), or culture medium as these media contain ions at concentrations that are harmful to the cells. Both transfection and cell survival showed a higher rate in cytomix (lowest in PBS) at 500 µF efficiency from the successful transfection. After electroporation, cells need to be transferred to a fresh pre-reduced culture medium and kept at 37 °C anaerobically for 12-16 h.

2.2.10 Nanoluc luciferase assay

To conduct a Nanoluc luciferase assay, the first step involves preparing a cell lysate by collecting and processing the cells. The following method can be employed for this procedure. Initially, the cells need to be gathered and prepared for lysis. To accomplish this, 1 x Luciferase Cell Culture Lysis reagent can be utilized, consisting of 25 mM Trisphosphate (pH 7.8), 2 mM DTT, 2 mM 1,2-diaminocyclohexane-N,N,N',N'-tetraacetic acid, 10% glycerol, and 1% Triton X-100. The cells go through lysis in this solution to release their contents.

After cell lysis, the resulting cell lysate is combined with an equal volume of Nano-Glo Luciferase Assay reagents. This mixture facilitated the detection and quantification of luciferase activity, a key component of the experiment. Subsequently, the relative luminescence units (RLU), indicative of luciferase activity, can be measured using the Hidex Sense multimode microplate reader (Hidex). This reader is employed for its ability to provide accurate and reliable readings.

By following these steps, one can collect the cells, lyse them, and measure the resulting luminescence, which enables the evaluation of the luciferase activity in the experiment.

2.2.11 Immunofluorescent microscopy

To experiment with immunofluorescent microscopy, *Blastocystis* cells can be subjected to the following procedure. Initially, the cells need to be washed and resuspended in PBS. Subsequently, the cells can be fixed using a solution of 4% paraformaldehyde in PBS, at room temperature, for a time of 30 minutes. This fixation process is aimed at preserving the cellular structure and components.

2.2.12 Permeabilization and Blocking

Following the fixation, cells are permeabilized using 1% NP-40 to facilitate antibody penetration into the cells. To prevent non-specific binding, cells can then be blocked with 3% bovine serum albumin before antibody staining. This blocking step minimizes background signals and enhances the specificity of the subsequent antibody labeling.

2.2.13 Antibody Staining

The monoclonal anti-Ty55 antibody can be employed at a dilution of 1:100 to specifically label eGFP-Ty. Following primary antibody incubation, cells need to be subjected to staining with a secondary antibody, goat-anti-mouse IgG H&L, at a dilution of 1:2000. This secondary antibody can facilitate the detection and visualization of the primary antibody binding to the target of interest.

2.2.14 Microscopy and Imaging

Subsequently, the prepared cells can be placed onto slides and examined using an inverted microscope. The setup is important for precise observation and capturing of cellular images. Through the procedure, *Blastocystis* cells can be described to be appropriately prepared, fixed, stained with specific antibodies, and subsequently imaged under a microscope, enabling the visualization and analysis of the eGFP-Ty labeling in the cells.

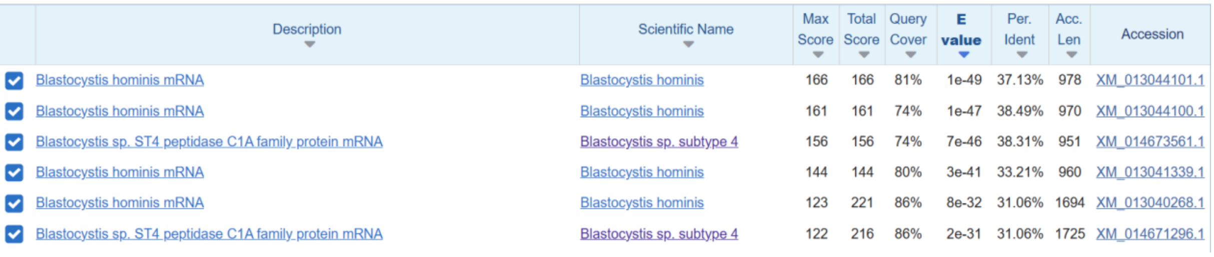
3 Results and discussion

This section presents the results of the investigation into proven virulence factors within *G. intestinalis*, *E. histolytica,* and *C. pavrum,* as well as a discussion of the findings. The results include the organisms mentioned as well as *Blastocystis* subtype 7 and subtype 4. Notably, the first protein presented (cathepsin B) will encompass the detail of the search outcomes, whereas the remaining protein's results are referred to in Appendix 1 and 2. Table 7 below presents an overview of the virulence factors examined in this study. Each factor is accompanied by a distinct label, the organism in which the virulence factor is identified, and the corresponding protein accession code found in NCBI.

Table 7 Overview of virulence factors originating from *G. intestinalis*, *E. histolytica*, and *C. pavrum*, including details such as the specific virulence factor, corresponding protein accession code on NCBI, and associated label in this study.

Virulence factor	Organism	Protein accession	Label
cathepsin B	*Giardia intestinalis*	KAE8304515.1	GiP1
cysteine proteinase 2	*Entamoeba histolytica*	XP_650642.1	EhP2
cryptopain-1	*Cryptosporidium pavrum*	ABA40395.1	CpP3
cysteine proteinase 5	*Entamoeba histolytica*	CAA62835.1	EhP4

The first protein, cathepsin B, the known virulence factor in *Giardia intestinalis* is searched for in the *Blastocystis* genome, and outcomes are shown in Figure 10 below.



	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Blastocystis hominis mRNA	Blastocystis hominis	166	166	81%	1e-49	37.13%	978	XM_013044101.1
☑	Blastocystis hominis mRNA	Blastocystis hominis	161	161	74%	1e-47	38.49%	970	XM_013044100.1
☑	Blastocystis sp. ST4 peptidase C1A family protein mRNA	Blastocystis sp. subtype 4	156	156	74%	7e-46	38.31%	951	XM_014673561.1
☑	Blastocystis hominis mRNA	Blastocystis hominis	144	144	80%	3e-41	33.21%	960	XM_013041339.1
☑	Blastocystis hominis mRNA	Blastocystis hominis	123	221	86%	8e-32	31.06%	1694	XM_013040268.1
☑	Blastocystis sp. ST4 peptidase C1A family protein mRNA	Blastocystis sp. subtype 4	122	216	86%	2e-31	31.06%	1725	XM_014671296.1

Figure 10 Search results of cathepsin B from *G. intestinalis* in *Blastocystis* genome. Outcomes are found by using translated nucleotide BLAST search (tBLASTn) for both *Blastocystis* ST7 and ST4. The query cover and e-value are both within the threshold.

Blastocystis ST7 had a query cover of 81% and an e-value of 1e^{-49}, which aligns with the established threshold. *Blastocystis* ST4 had a query cover of 74%, which is still noted as a high score, and an e-value of 7e^{-46} which fills the requirements. A graphic summary of the search outcomes from cathepsin B in *Blastocystis* (Figure 10) is shown in Figure 12 below. Pink bars denote solid matches with an alignment score between 80-200.

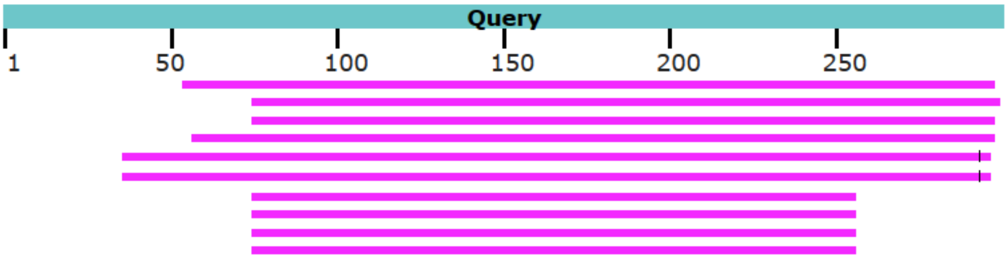
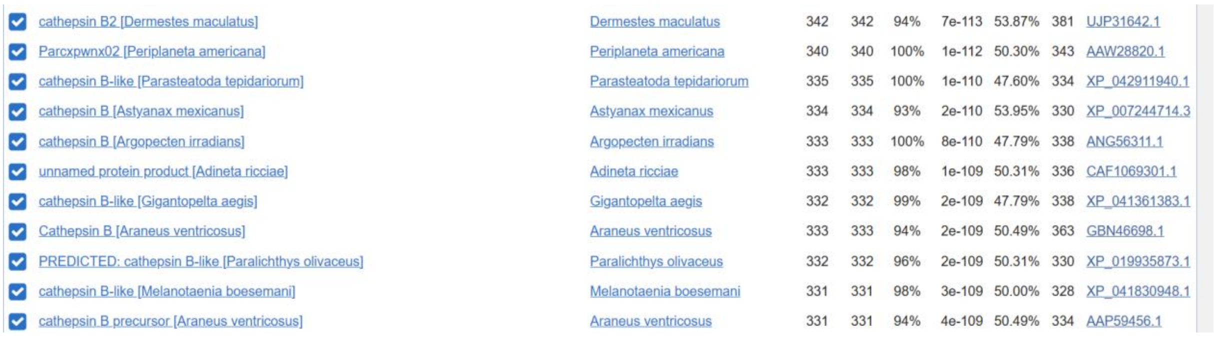


Figure 11 Graphic summary of results from tBLASTn search of cathepsin B from GiP1 in *Blastocystis* (Figure 10). Pink bars indicate solid matches.

Blastocystis ST7 displays a result linked to an uncharacterized protein. In the NCBI page from the search, the region of the uncharacterized protein is noted as "Cathepsin B group; composed of cathepsin B and similar proteins, including tubulointerstitial nephritis antigen (TIN-Ag)". To ascertain the nature of *Blastocystis* ST7, its protein sequence also underwent investigation through reverse protein BLAST (BLASTp), with the exclusion of searches within the *Blastocystis* genome. The findings reveal a predominant presence of cathepsin B across various organisms. The resultant findings are illustrated in Figure 12.



☑	cathepsin B2 [Dermestes maculatus]	Dermestes maculatus	342	342	94%	7e-113	53.87%	381	UJP31642.1
☑	Parcxpwnx02 [Periplaneta americana]	Periplaneta americana	340	340	100%	1e-112	50.30%	343	AAW28820.1
☑	cathepsin B-like [Parasteatoda tepidariorum]	Parasteatoda tepidariorum	335	335	100%	1e-110	47.60%	334	XP_042911940.1
☑	cathepsin B [Astyanax mexicanus]	Astyanax mexicanus	334	334	93%	2e-110	53.95%	330	XP_007244714.3
☑	cathepsin B [Argopecten irradians]	Argopecten irradians	333	333	100%	8e-110	47.79%	338	ANG56311.1
☑	unnamed protein product [Adineta ricciae]	Adineta ricciae	333	333	98%	1e-109	50.31%	336	CAF1069301.1
☑	cathepsin B-like [Gigantopelta aegis]	Gigantopelta aegis	332	332	99%	2e-109	47.79%	338	XP_041361383.1
☑	Cathepsin B [Araneus ventricosus]	Araneus ventricosus	333	333	94%	2e-109	50.49%	363	GBN46698.1
☑	PREDICTED: cathepsin B-like [Paralichthys olivaceus]	Paralichthys olivaceus	332	332	96%	2e-109	50.31%	330	XP_019935873.1
☑	cathepsin B-like [Melanotaenia boesemani]	Melanotaenia boesemani	331	331	98%	3e-109	50.00%	328	XP_041830948.1
☑	cathepsin B precursor [Araneus ventricosus]	Araneus ventricosus	331	331	94%	4e-109	50.49%	334	AAP59456.1

Figure 12 The reverse BLASTp search results for the protein discovered in *Blastocystis* ST7 strongly indicate its identity as cathepsin B. A significant number of hits were found for the cathepsin B protein, enhancing the reliability of the discovery that the identified protein *Blastocystis* ST7 is indeed cathepsin B.

The graphic summary shown in Figure 13 presents the distribution of sequences from Figure 12, comparing the sequence of protein found in *Blastocystis* ST7 of known proteins in BLAST. The alignment scores of the sequences are denoted by red (red ≥ 200), indicating a high degree of similarity.

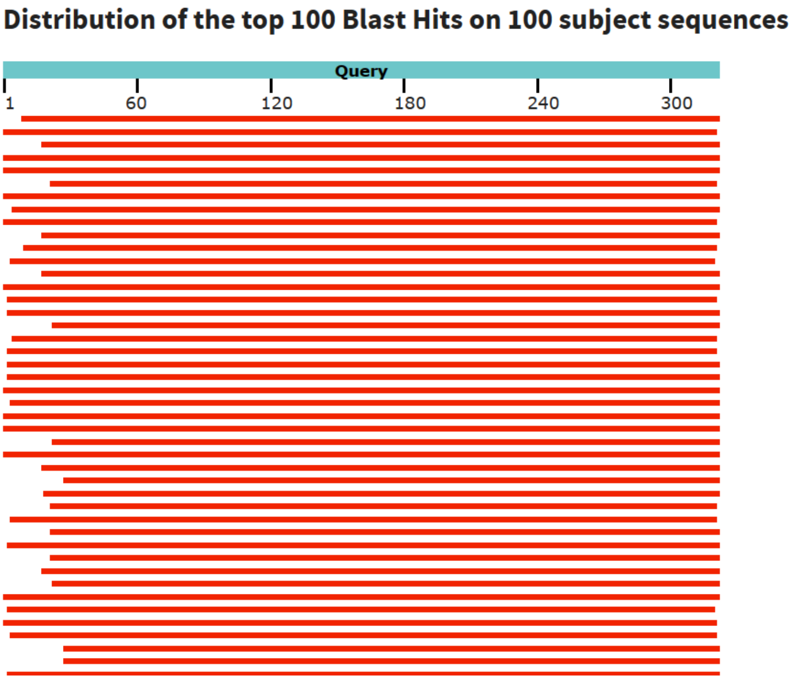


Figure 13 Illustration of the distribution resulting from the reverse BLASTp search for the protein identified in *Blastocystis* ST7 (excluding *Blastocystis* genome itself). This search mirrors the one conducted in Figure 12. The red color of the lines signifies strong alignment scores, indicative of favorable sequence matches.

Based on the results above, it is highly probable that the uncharacterized protein *Blastocystis* ST7 corresponds to the cathepsin B protein. In the case of *Blastocystis* ST4 results, the protein corresponds to the peptidase C1A family protein, and just like ST7, it is noted as Cathepsin B in NCBI. However, minor differences exist in the protein sequence.

The table below (Table 8) shows all outcomes from BLAST searches of the known virulence factors in *Blastocystis* ST7 and ST4 with the protein accession, the protein name in NCBI, and which virulence factor the protein found is retrieved using. As you can see, the virulence factors from EhP2 and CpP3 resulted in the finding of protein BsP2 ST7 and BsP2 ST4. Similarly, in EhP4, the protein BsP2 ST4 was identified in ST4.

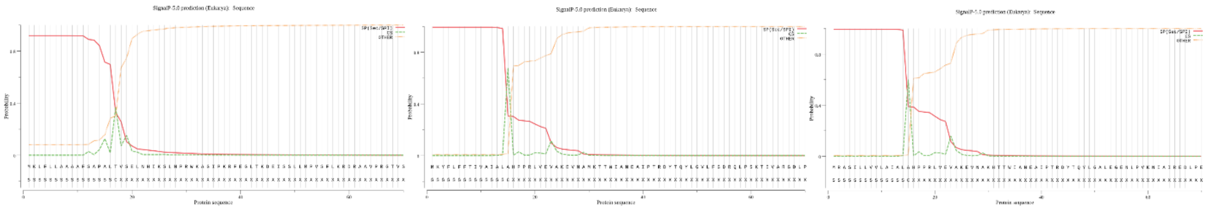
Table 8 Summary of the identified outcomes regarding the known virulence factor cathepsin B (GiP1) from *G. intestinalis*, cysteine proteinase 2 (EhP2) and 5 (EhP4) from *E. histolytica*, and cryptopain-1 (CpP3) from *C. pavrum*.

Label	Organism	Protein accession	NCBI protein	Retrieved using
BsP1 ST7	*Blastocystis* spp. ST7	XP_012899555.1	uncharacterized protein	GiP1
BsP1 ST4	*Blastocystis* spp. ST4	XP_014529047.1	peptidase C1A family protein	GiP1
BsP2 ST7	*Blastocystis* spp. ST7	XP_012897923.1	uncharacterized protein	EhP2, CpP3
BsP2 ST4	*Blastocystis* spp. ST4	XP_014529703.1	peptidase C1A domain-containing protein	EhP2, CpP3, EhP4
BsP4 ST7	*Blastocystis* spp. ST7	XP_012894811.1	uncharacterized protein	EhP4

All other search results from the other proteins can be found in Appendix 1. A detailed outcome of the BLAST analysis is accessible in Appendix 2, providing essential data such as e-values, query coverage, and protein lengths. The query coverage for most proteins exceeds 70%, except for a single instance of *Blastocystis* ST4 protein identified from EhP4, which exhibited a query coverage of 69%. This result falls just below the threshold. On the other hand, all protein results maintain e-values below the predetermined threshold.

3.1 Signal peptide

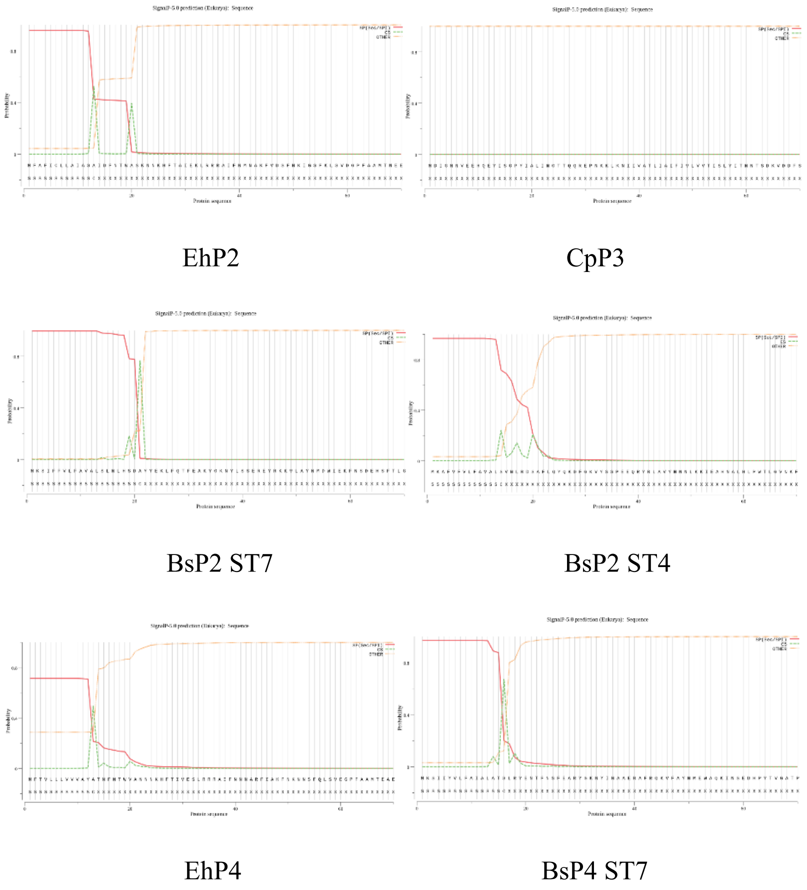
Signal peptides were found by using the online database SignalP 5.0 (Petersen et al., 2011). This is used to determine the presence of signal peptides and to find if the protein is likely to be targeted for secretion or membrane insertion, thereby providing crucial insights into its functional role within the cell (Owji et al., 2018). GiP1 was found to have a cleavage site located between positions 17 and 18 (ALT-VS) with a high likelihood of 0.9171. Meanwhile, BsP1 ST7 possesses a signal peptide with a cleavage site positioned between 15 and 16 (ALA-HP) and displays a probability of 0.9926. In the case of BsP1 ST4, it shares the same cleavage site position with a probability as high as 0.9942. Subsequently, these signal peptides were excised from the sequences to yield matured proteins. Results are shown in Figure 14.



GiP1 BsP1 ST7 BsP1 ST4

Figure 14 Figure of signal peptide analysis for protein Cathepsin B in *Giardia intestinalis*, *Blastocystis* ST7 and *Blastocystis* ST4. Results retrieved from the server of SignalP 5.0.

EhP2 has a signal peptide in between positions 13 and 14 (ASA-ID) with a likelihood of 0.9563, whereas CpP3 lacked any signal peptides. BsP2 ST7 has a likelihood of a signal peptide of 0.9946 with a cleavage site between positions 21 and 22 (SDA-YY). BsP2 ST4 has a signal peptide with a likelihood of 0.9681 with a cleavage site between positions 14 and 15 (ALS-VN). EhP4 exhibited a signal peptide likelihood of 0.7123, featuring a cleavage site between positions 13 and 14 (AYA-TN). In contrast, BsP4 ST7 displayed a signal peptide likelihood of 0.9691, with a cleavage site positioned between positions 16 and 17 (ATS-LR). Results from the signal peptide results are shown in Figure 15.



EhP2

CpP3

BsP2 ST7

BsP2 ST4

EhP4

BsP4 ST7

Figure 15 Figure of signal peptide analysis from virulence factors cysteine proteinase 2 (EhP2), cryptopain-1 (CpP3) and cysteine proteinase 5 (EhP4) and proteins of *Blastocystis* ST7 (BsP2 ST7 and BsP4 ST7) and *Blastocystis* ST4 (BsP2 ST4).

Apart from CpP3, all proteins contain signal peptides within their sequences. These peptides contain valuable information making them particularly significant for disease diagnosis and immunization research (Owji et al., 2018). Proteins lacking these signal peptides are usually enclosed in the cell's cytoplasm (Purves, 2003). In some cases, the signal peptide is cleaved off during translation or it is embedded into the endoplasmic reticulum membrane creating a transmembrane segment that anchors the protein to the membrane (Kapp et al., 2009).

3.2 Motifs and domains

The SMART database enables the identification of motifs and domains within a sequence of interest. The outcomes of the protein search for all virulence factors and protein results of *Blastocystis* ST7 and ST4, in this database are described in Figure 16 below.

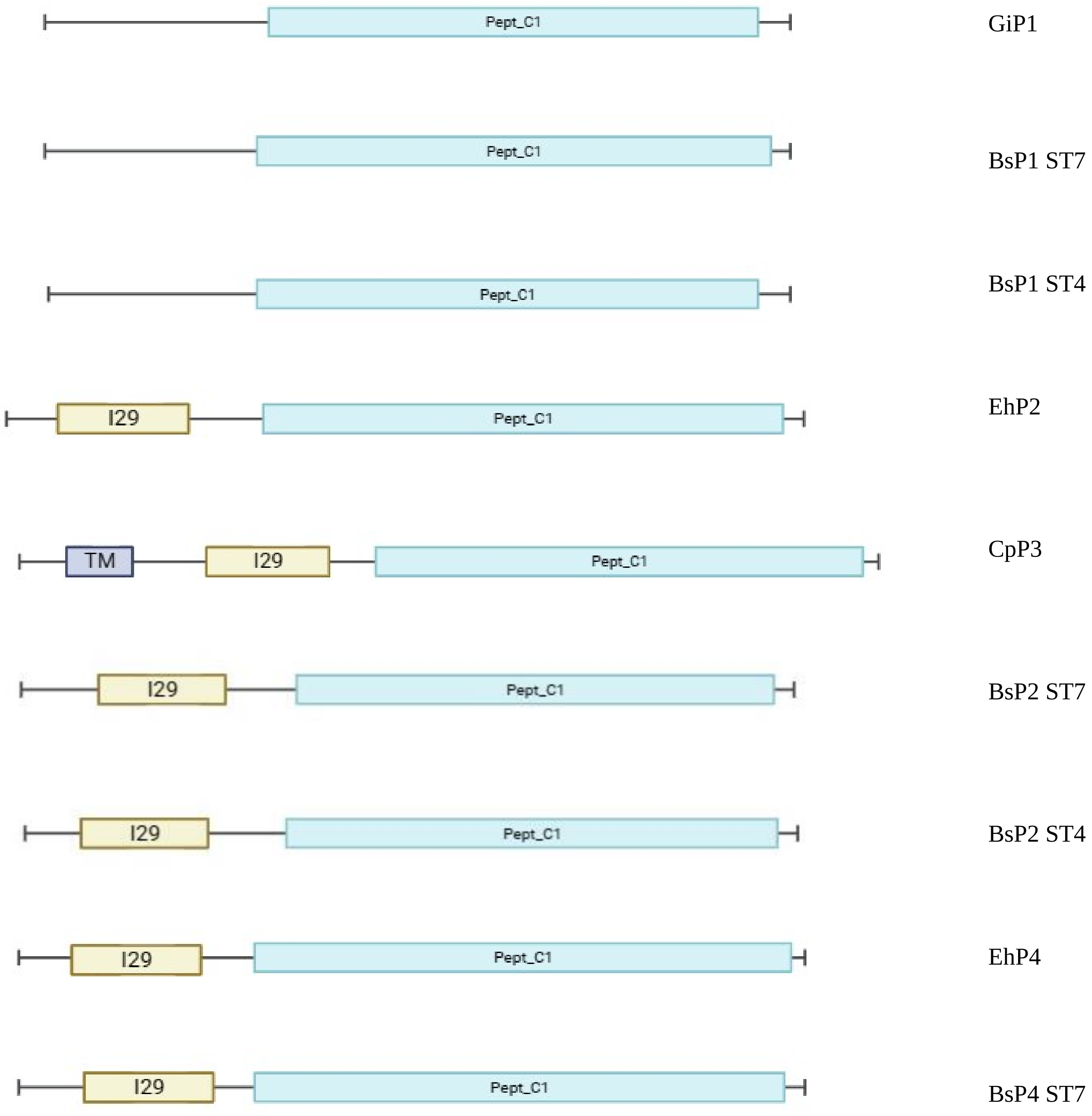
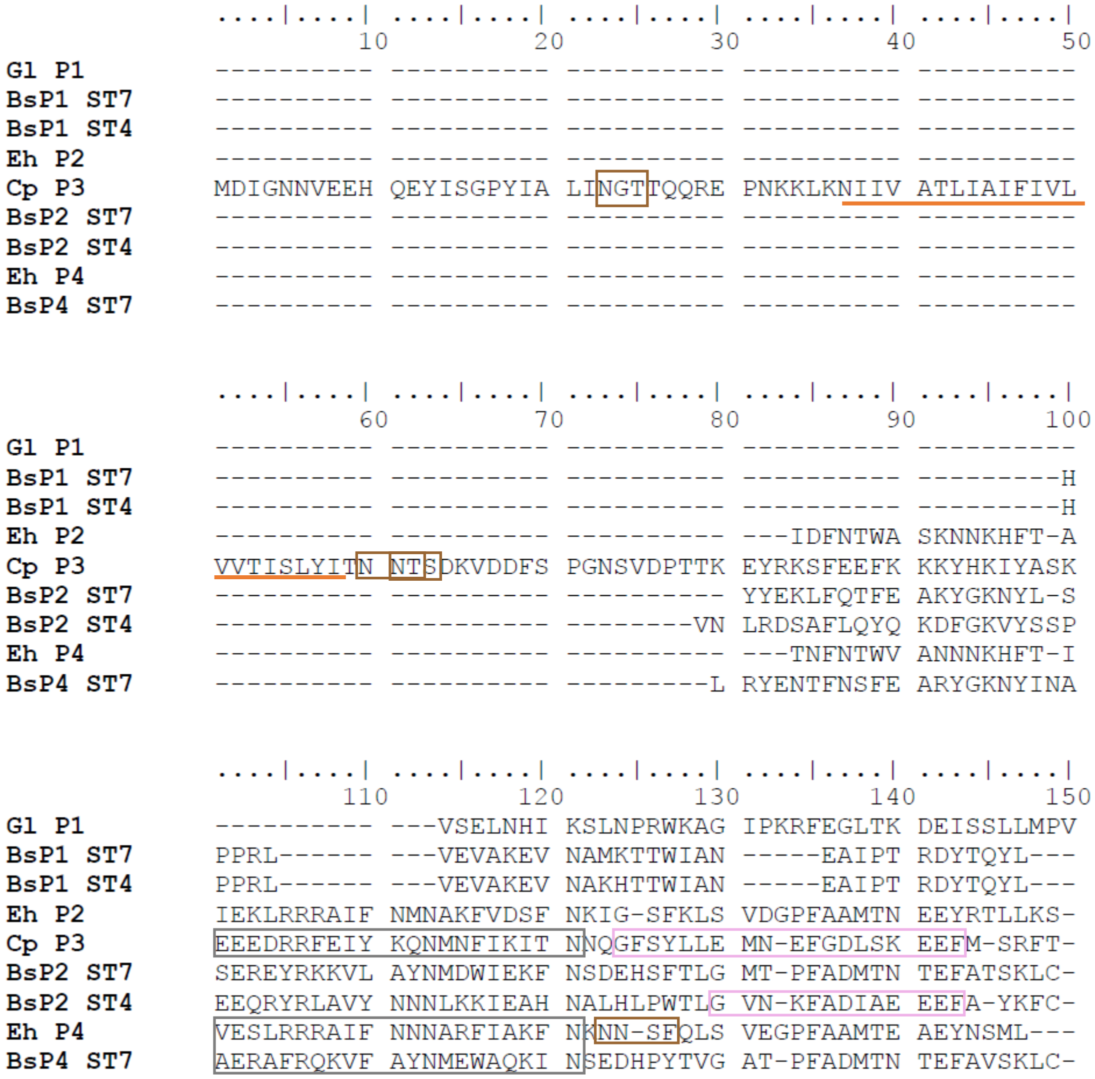


Figure 16 Illustration of SMART search of virulence factors from *G. intestinalis* (GiP1), *E. histolytica* (EhP2, EhP4), and *C. pavrum* (CpP3). As well as proteins from *Blastocystis* ST7 and ST4). All proteins identified encompass the peptidase C1 domain family (Pept_C1). Notably, *G. intestinalis* and its corresponding results for *Blastocystis* ST7 and ST4 proteins lack the inhibitor I29 region (indicated by the yellow box), a feature present in all other virulence factors and the *Blastocystis* proteins. Created with BioRender.com

The findings from the SMART analysis indicate that the cysteine peptidase family C1 domain is shared among all the organisms. All investigated proteins feature a region with a notation indicating possible catalytic inactivity in the SMART database. Although the known virulence factors are known to be causing virulence based on research, this might not necessarily apply to the proteins under analysis. The protein lengths remain consistent with a variation of ±20 amino acids, with a few exceptions; *C. pavrum* was notably longer because it possesses a transmembrane region close to the amino terminus. If this transmembrane domain were cut, it would lead to greater similarity in the lengths of all sequences, as well as align the positions of the inhibitor I29 and Pept_C1 domains. EhP2 also had a longer sequence originally, a decision to shorten the protein sequence for the SMART search was made by cutting away the first part which consisted of low complexity regions and to make the Pept_C1 align with the other proteins compared.

3.3 Sequence alignment

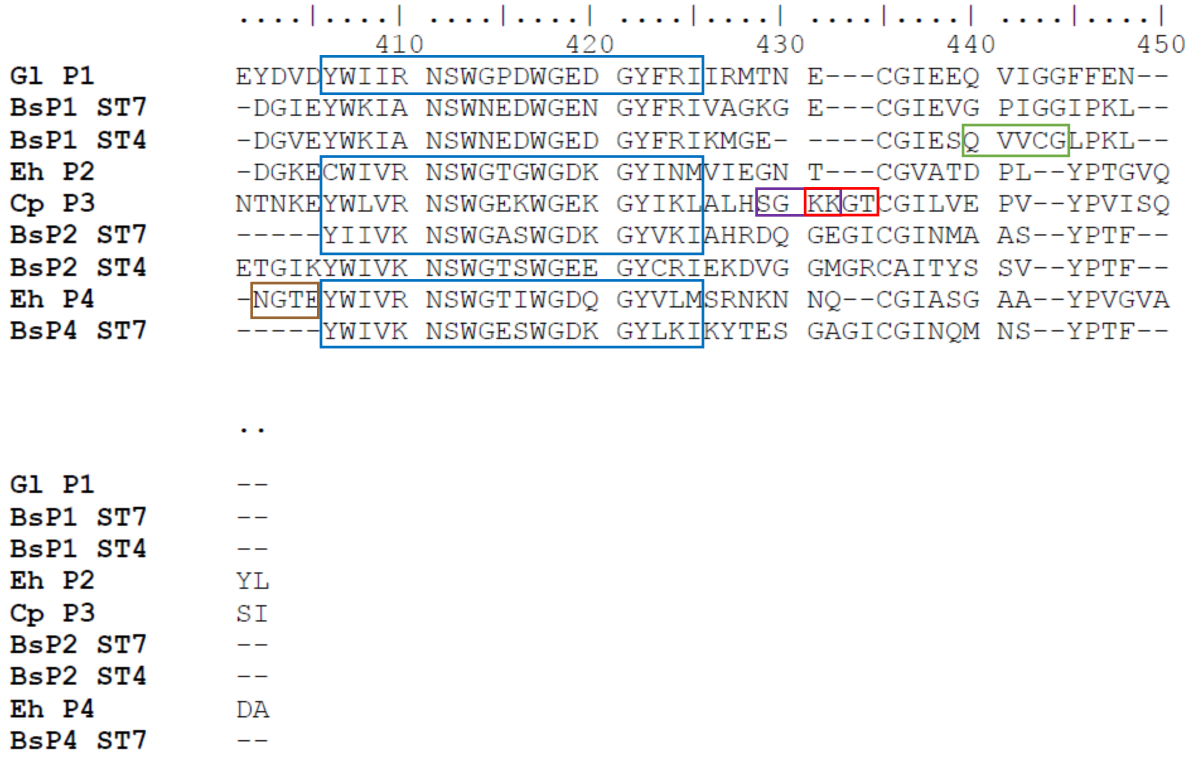
A sequence alignment is conducted to compare the known virulence factors of *G. intestinalis*, *E. histolytica*, and *C. pavrum* with the proteins present in *Blastocystis* ST7 and ST4 (Figure 17). Active sites and motifs were found by a search in ScanProsite, detailed results of one search conducted in *Blastocystis* ST7 (BsP1 ST7) are shown in Appendix 1. Motifs that have a high probability of occurrence were applied in the scan, but just a selection of the outcomes are shown and discussed here. The figure displays color codes corresponding to the following sites: active sites (blue), QxVxG motif (green), CxxC motif (black), occluding loop (yellow line), RGD motif (orange box), cAMP- and cGMP-dependent protein kinase phosphorylation site (red), transmembrane domain (orange line), ERFNIN motif (grey), GNFD motif (pink), N-glycosylation site (brown) and amidation site (purple).



```
               ....|....| ....|....| ....|....| ....|....| ....|....|
                        10         20         30         40         50
Gl  P1         ---------- ---------- ---------- ---------- ----------
BsP1 ST7       ---------- ---------- ---------- ---------- ----------
BsP1 ST4       ---------- ---------- ---------- ---------- ----------
Eh  P2         ---------- ---------- ---------- ---------- ----------
Cp  P3         MDIGNNVEEH QEYISGPYIA LINGTTQQRE PNKKLKNIIV ATLIAIFIVL
BsP2 ST7       ---------- ---------- ---------- ---------- ----------
BsP2 ST4       ---------- ---------- ---------- ---------- ----------
Eh  P4         ---------- ---------- ---------- ---------- ----------
BsP4 ST7       ---------- ---------- ---------- ---------- ----------


               ....|....| ....|....| ....|....| ....|....| ....|....|
                        60         70         80         90        100
Gl  P1         ---------- ---------- ---------- ---------- ----------
BsP1 ST7       ---------- ---------- ---------- ---------- ---------H
BsP1 ST4       ---------- ---------- ---------- ---------- ---------H
Eh  P2         ---------- ---------- ---------- ---IDFNTWA SKNNKHFT-A
Cp  P3         VVTISLYITN NTSDKVDDFS PGNSVDPTTK EYRKSFEEFK KKYHKIYASK
BsP2 ST7       ---------- ---------- ---------- YYEKLFQTFE AKYGKNYL-S
BsP2 ST4       ---------- ---------- --------VN LRDSAFLQYQ KDFGKVYSSP
Eh  P4         ---------- ---------- ---------- ---TNFNTWV ANNNKHFT-I
BsP4 ST7       ---------- ---------- ---------L RYENTFNSFE ARYGKNYINA


               ....|....| ....|....| ....|....| ....|....| ....|....|
                       110        120        130        140        150
Gl  P1         ---------- ---VSELNHI KSLNPRWKAG IPKRFEGLTK DEISSLLMPV
BsP1 ST7       PPRL------ ---VEVAKEV NAMKTTWIAN -----EAIPT RDYTQYL---
BsP1 ST4       PPRL------ ---VEVAKEV NAKHTTWIAN -----EAIPT RDYTQYL---
Eh  P2         IEKLRRRAIF NMNAKFVDSF NKIG-SFKLS VDGPFAAMTN EEYRTLLKS-
Cp  P3         EEEDRRFEIY KQNMNFIKIT NNQGFSYLLE MN-EFGDLSK EEFM-SRFT-
BsP2 ST7       SEREYRKKVL AYNMDWIEKF NSDEHSFTLG MT-PFADMTN TEFATSKLC-
BsP2 ST4       EEQRYRLAVY NNNLKKIEAH NALHLPWTLG VN-KFADIAE EEFA-YKFC-
Eh  P4         VESLRRRAIF NNNARFIAKF NKNN-SFQLS VEGPFAAMTE AEYNSML---
BsP4 ST7       AERAFRQKVF AYNMEWAQKI NSEDHPYTVG AT-PFADMTN TEFAVSKLC-
```

```
                ....|....| ....|....| ....|....| ....|....| ....|....|
                        160        170        180        190        200
Gl P1           SFLKRDRAAV PRGTVSATQA ---------P DSFDFREEYP HC--IPEVVD
BsP1 ST7        GVLFGDRQLP SKTIVARGDL ---------P ESFDPVEKWP ECPSLKEIRD
BsP1 ST4        GALEGE-NLP VKNIAIREDL ---------P ESFDAAEQWP ECPSLKEIRD
Eh P2           ---KRTTEEN GQVKYLNIQA ---------P ESVD----WR KEGKVTPIRD
Cp P3           GYRKDLDDNE GRFKASRVSA IEFEEDFAIP DSVN----WV EAGCVNPIRN
BsP2 ST7        GCMKKPLNHK -QARVLNNMA ---------V ESID----WR EKGAVTPVKN
BsP2 ST4        GCAKDPKSRA GRVTPIYGDA ---------P ERID----WR EKGAVTPVKD
Eh P4           KPFVIDKQHE EIVYDSRGDV ---------P ESVD----WR AKGKVPAIRD
BsP4 ST7        GCMLKPKMTK P-ATPIMEPA ---------A EAVD----WR EKGAVTPVKN

                ....|....| ....|....| ....|....| ....|....| ....|....|
                        210        220        230        240        250
Gl P1           QGGCGSCWAF SSVASVGDRR CFAGLDKK-- -AVKYSPQYV VSC----DRG
BsP1 ST7        QSVCGSCWAF GAAEAATDRL CIASKGKI-- -QDRLSEQDL LTCC---DSC
BsP1 ST4        QSVCGSCWAF GAAEAATDRL CIASKGKV-- -QDRLSSEDL LTCC---DIC
Eh P2           QAQCGSCYTF GSLAALEGRL LIEKGGDA-- NTLDLSEEHM VQCTR--DNG
Cp P3           QKNCGSCWAF SAVAALEGAI CAQTNQGL-- --PSLSEQQL VDCSK--KNG
BsP2 ST7        QGSCGSCWAF SATGALEGGN FVATGKLV-- ---SLSEQQL VDC----DTE
BsP2 ST4        QAACGSCWSF STTGTVEGAY FIHSGKLV-- ---SLSEQQL VDCAREPKYQ
Eh P4           QASCGSCYSF ASVAAIEGRL LVAGSKKFTV DDLDLSEQQL VDCSV--SVG
BsP4 ST7        QASCGSCWAF SATGAMEGRN FVANGELI-- ---SLSEQQL VDC----DHQ
```

```
                ....|....| ....|....| ....|....| ....|....| ....|....|
                        260        270        280        290        300
Gl P1           DMACDGGWLP SVWRF-LTKT GTTTDECVP- ------YQ-- ----------
BsP1 ST7         GFGCDGGWLD MAWRW-FQST GVTTGGEYGS KDWCNAYSFP KCEHHAEGKY
BsP1 ST4         GFGCQGGFPS SAWNW-FHTV GVTTGGEYGS KDWCNAYAFA KCEHHSTGKY
Eh P2           NNGCNGGLGS NVYDY-IIEH GVAKESDYP- ------YT-- ----------
Cp P3           NFGCSGGTMG LAFQYAIKNK YLCTNDDYP- ------YY-- ----------
BsP2 ST7         DAGCGGGFMD TAFEY-VMKK GLCTEEDYP- ------YH-- ----------
BsP2 ST4         AEGCGGGWPW SVMDY-VSDH GLCTEEDYP- ------YH-- ----------
Eh P4           NKGCNGGSLL LSFRY-VKLN GIMQEKDYP- ------YV-- ----------
BsP4 ST7         SSGCGGGLMT YAFEY-AKKK GMCKEEDYP- ------YH-- ----------

                ....|....| ....|....| ....|....| ....|....| ....|....|
                        310        320        330        340        350
Gl P1           --SGSTGARG TCPTKCADG- -----SDLPI YKATKAVDYG LDCDLIMKAL
BsP1 ST7         PPCGESQETP ECVKQCQEGY PVEYEKDKHF FGEAYYVQGG IDA--IKTEL
BsP1 ST4         PPCGETQDTP ECVTECQEGY PVSYENDKHF FEEGYSVRGI EA---IKTEL
Eh P2           ------GSDS TCKTNVK--- -----SFAKI TGYTKVPRNN EAE--LKAAL
Cp P3           ------AEEN ICRESLCENY -----VEVPV KAYRYVFPRN VNS--LKSAL
BsP2 ST7         ------AKDE DCKDDQCT-- -----SVISI TGYEDVPAND GVA--LKQAL
BsP2 ST4         ------AKDE DCKDDKCT-- -----VAVQS VGKVQLPQGD EES--LANAV
Eh P4           ------AAEE TCTYDKKK-- -----VAVKI TGQKLVRPGS EKA--LMRAA
BsP4 ST7         ------AVDE DCKDDKCT-- -----PVVFP KGYEEVPRFD GAA--LKQAV

                ....|....| ....|....| ....|....| ....|....| ....|....|
                        360        370        380        390        400
Gl P1           ATGGPLQTAF TVYS-DFMYY EGGVYQHTYG RVE---GGHA VEMVGYGTD-
BsP1 ST7         MTNGPLEVSF FVYE-DFLTY KSGIYQHVAG KYL---GGHA VKLVGWGVE-
BsP1 ST4         MTNGPMEVAF TVYE-DFMTY KSGIYQHVTG SRL---GGHA VKLVGWGVE-
Eh P2           SQ-GLVDVSI DASSAKFQLY KSGAYTDTKC KNNYFALNHE VCAVGYGVV-
Cp P3           ARYGPISVAI QADQTPFQFY KSGVFD-APC GTR---VNHG VVLVGYDIDE
BsP2 ST7         TK-APVSVAI QADSFVFQMY TGGVLDSDMC GTS---LNHG VLAVGYAKE-
BsP2 ST4         AL-TPVSIVL DASA--MQLY KEGII--TKC SES---INHA VLAVGYGVEE
Eh P4           AE-GPVAAAI DASGVKFQLY KSGIYNSKEC SST--QLNHG VAVVGYGTQ-
BsP4 ST7         SQ-GPVSVAV EADSIVFQMY TGGVIDSSAC GTS---LNHG VLAVGYGAD-
```



```
                ....|....| ....|....| ....|....| ....|....| ....|....|
                        410        420        430        440        450
Gl  P1          EYDVDYWIIR NSWGPDWGED GYFRIIRMTN E---CGIEEQ VIGGFFEN--
BsP1  ST7       -DGIEYWKIA NSWNEDWGEN GYFRIVAGKG E---CGIEVG PIGGIPKL--
BsP1  ST4       -DGVEYWKIA NSWNEDWGED GYFRIKMGE- ----CGIESQ VVCGLPKL--
Eh  P2          -DGKECWIVR NSWGTGWGDK GYINMVIEGN T---CGVATD PL--YPTGVQ
Cp  P3          NTNKEYWLVR NSWGEKWGEK GYIKLALHSG KKGTCGILVE PV--YPVISQ
BsP2  ST7       -----YIIVK NSWGASWGDK GYVKIAHRDQ GEGICGINMA AS--YPTF--
BsP2  ST4       ETGIKYWIVK NSWGTSWGEE GYCRIEKDVG GMGRCAITYS SV--YPTF--
Eh  P4          -NGTEYWIVR NSWGTIWGDQ GYVLMSRNKN NQ--CGIASG AA--YPVGVA
BsP4  ST7       -----YWIVK NSWGESWGDK GYLKIKYTES GAGICGINQM NS--YPTF--


                ..

Gl  P1          --
BsP1  ST7       --
BsP1  ST4       --
Eh  P2          YL
Cp  P3          SI
BsP2  ST7       --
BsP2  ST4       --
Eh  P4          DA
BsP4  ST7       --
```

Figure 17 Sequence alignment for all virulence factors from *G. intestinalis*, *E. histolytica*, and *C. pavrum* together with the protein outcomes of *Blastocystis* ST7 and ST4. Active sites of cysteine, histidine, and asparagine are noted in blue. The length of active sites differs; the enclosed sequence is only approximate. The annotated boxes in the alignment are the RGD motif in red, QxVxG motif in green, CxxC motif in black, ERFNIN motif in grey, GNFD motif in pink, KRTT motif in red, N-glycosylation sites in brown and amidation site in purple. The yellow line presents a possible occluding loop (only in *Blastocystis*), and the orange line is a transmembrane region only found in CpP3. Created in BioEdit.

The sequence alignment of the proteins is robust and encompasses numerous conserved regions, indicating significant similarities among them and implying potential functional activity. However, distinct motifs and sites are also observed with further details provided in the following sections.

3.3.1 Active sites

Active sites vary in length for the proteins, but a blue box highlights the area of the approximate site as all are in the same positions. All proteins share active sites of Cys, His, and Asn forming a catalytic triad. However, this triad is absent in all proteins of *Blastocystis* ST4 and one *Blastocystis* ST7; BsP1 ST7. In these cases, they instead exhibit a catalytic dyad comprised of the active site residues Cys and His. Even though the proteins mentioned don't have the Asn site, the same area is conserved and shows a similar amino acid sequence compared to the other proteins.

Active sites play a crucial role in catalyzing biochemical reactions (Robinson, 2015). In this case of *Blastocystis*, it could be possible that *Blastocystis* ST7 displays a higher enzymatic activity compared to ST4 due to the presence of a complete catalytic triad. This suggests that biological processes associated with these active sites might affect the organism's possibility of causing virulence. The presence of a catalytic dyad in ST4 implies a possibly different mechanism or it could have reduced efficiency in enzymatic reactions. This variation in active site configuration between different subtypes of *Blastocystis* might result in diverse metabolic capabilities, affecting their ability to interact with the host or possibly other microorganisms in the environment. Some studies have already found results supporting *Blastocystis* ST7 being related to pathogenicity, while ST4 is not (Deng et al., 2022; Deng & Tan, 2022).

3.3.2 QxVxG motif

The motif QxVxG is unique to BsP1 ST7 and ST4, with ST7 including this motif at two positions and ST4 at three positions. This is marked in green (Figure 17).

QxVxG motifs have been recognized as crucial elements for binding and inhibiting cysteine protease activity (Cuesta-Astroz et al., 2014). These motifs are directly involved in protein-protein interactions and are typically seen in cystatins (Dutt et al., 2010). Cystatins play a role in the regulation of protease activity and can have an important role in the parasite's survival within the host (Khatri et al., 2020). The discovery and its potential link with cystatins are noteworthy for further investigation in studying the pathogenicity of *Blastocystis*.

3.3.3 CxxC motif

The CxxC motif is important for specific protein functions and is enclosed by black boxes in the sequence alignment (Figure 17). The motif is characterized by two cysteines separated by two other residues, and from research known to be utilized by numerous redox proteins for disulfide bond formation, isomerization, reduction, and other redox functions (Fomenko & Gladyshev, 2003). This motif is present in all proteins in the active site of Cys with pattern CGSC. Researchers have systematically explored genetic sequences to identify patterns equivalent to the CxxC motif. They discovered that this motif plays an important role in modulating protein function. Altering this motif significantly impacts the protein's properties and interactions, acting as a molecular switch that influences their behavior in various biochemical processes (Quan et al., 2007).

3.3.4 Occluding loop

In both BsP1 ST7 and BsP1 ST4, a possible occluding loop was found. This data was found based on a comparison with the occluding loop found in human cathepsin B (Liu, 2019). The site is distinguished by the amino acid pattern HH and is not found in any of the other analyzed proteins.

CEHHVNGSRPPCTGEGDTPKC	human cathepsin B
CEHHAEGKYPPCGESQETPEC	BsP1 ST7
CEHHSTGKYPPCGETQDTPEC	BsP1 ST4

These sequence alignments of human cathepsin B compared with BsP1 ST7 and ST4 reveal similarities. In addition to an established role in exopeptidase activity, the occluding loop of cathepsin B serves to restrict access of macromolecules to the active site, effectively functioning as an exopeptidase. Although the residual endopeptidase activity is not strictly necessary intracellularly, it enables the enzyme to function extracellularly as an endopeptidase, which is poorly inhibited by cystatins (Illy et al., 1997). This may have pathological consequences for *Blastocystis* for degradation of its host tissues for survival and proliferation. The resistance of cathepsin B to inhibition by cystatins in the extracellular environment could enhance the virulence of these parasites, allowing them to evade host immune responses. Further research into the regulation of cathepsin B activity and its interactions with host factors is important for understanding the molecular basis of parasite pathogenicity and for developing targeted therapeutic approaches to fight parasitic infections.

3.3.5 RGD motif

The orange box signifies the presence of the RGD motif, which are proteins that contain the amino acids Arg-Gly-Asp (RGD). The motif is found at one position before the Cys active site in BsP1 ST7 and EhP4, while in GiP1 it is found further downstream of the Cys active site. This motif is not found in BsP1 ST4. The RGD serves as an attachment site for numerous adhesive extracellular matrix, blood, and cell surface proteins, along with integrins acting as their receptors, constituting a vital recognition system for cell adhesion (Yamada et al., 2023). This motif, crucial for integrin-mediated cell attachment, plays a significant role in regulating processes like cell migration, growth, differentiation, and apoptosis (Ruoslahti, 1996). RGD peptides and their mimics serve as valuable tools to explore integrin functions across diverse biological systems. The recognized RGD motif found in *Giardia* and *Blastocystis* ST7 could hold promise for pharmaceutical development. Employing strategies such as exploring the motif's structure could be a possible method in drug design. However, additional *in vitro* studies are essential to confirm these assumptions.

3.3.6 cAMP- and cGMP-dependent protein kinase phosphorylation site

Both virulence factors of *E. histolytica* as well as *C. pavrum* are shown to be the only proteins containing this site that function as a target for cAMP- and cGMP-dependent protein kinase phosphorylation. The phosphorylation site is highlighted in red in Figure 17, located before the active site of Cys (KRtT) in EhP2, whereas in EhP4, it is positioned after the Cys active site (KKfT). In *C. pavrum* CpP3, the phosphorylation site (KKgT) is found at the end of the sequence. Moreover, these sites regulate the rate of cAMP production and degradation, making them responsive to a wide range of extracellular and intracellular signals (Caretta & Mucignat-Caretta, 2011). The function of these phosphorylation sites is possible to affect the regulation of a variety of cell functions, from metabolism to ion channel activation, cell growth and differentiation, gene expression, and apoptosis (Asaoka, 2012; K. V. Chin et al., 2002). Protein Kinase A (PKA), a key player in cAMP signaling, regulates diverse cellular processes such as metabolism, cell growth, and gene expression. Its involvement in infection-related pathways underscores its significant impact on various pathological processes due to its influence on intracellular functions driven by cAMP regulation (Haidar et al., 2017). This could be relevant for further studies but was not found present in the proteins investigated in *Blastocystis* ST7 or ST4.

3.3.7 Transmembrane domain

The only instance of a transmembrane domain detected by the PRED-TM algorithm is in *C. pavrum*, indicated by the orange line in the sequence alignment (Figure 17). See Appendix 1 for the results of the search. Integral membrane proteins typically feature a transmembrane region composed of hydrophobic α-helices clustered together (Wayne Albers, 2012). This specific arrangement may result from a folding mechanism where stable transmembrane helices align without undergoing topological changes (Popot, 1993). These regions can play various crucial roles, like catalyzing enzymes, facilitating membrane transport, acting as receptors for signaling molecules like hormones and growth factors, and participating in energy transfer for ATP synthesis (Ramasarma, 1996). It remains to be explored whether this region contributes to the enzyme's processing and localization within the parasite. This could be interesting to investigate other potential proteins in *Blastocystis* as it is not found in the proteins investigated in this study.

3.3.8 ERFNIN motif

The motif ERFNIN, represented in grey (Figure 17), serves as the pro-domain of cysteine cathepsins and features a highly preserved amino acid sequence represented by ERFNIN (Aich & Biswas, 2018). This motif was identified in CpP3, EhP4, and BsP4 ST7. This area also appears to be conserved in all proteins except for the analysis of GiP1 with *Blastocystis* proteins. Although EhP2, BsP2 ST7, and BsP2 ST4 possess a comparable conserved region, they lack one or more amino acids necessary to fully establish the presence of this motif. Studies on the zymogen structure of cathepsin L and K unveiled that the arginine (R) residue within the ERFNIN motif acts as a central element in a salt-bridge and hydrogen-bond network, thereby stabilizing the scaffold of the pro-domain (Aich & Biswas, 2018). This is interesting for further investigation in *Blastocystis*, where it could have an impact on the activation and functioning of cysteine proteases. Exploring the presence of the ERFNIN motif in *Blastocystis* might bring notice to their enzymatic activities and regulatory mechanisms. Additionally, understanding the role of the conserved arginine (R) residue within the motif could provide valuable insights into the structural stability and functional importance.

3.3.9 GNDF motif

The GNFD motif, enclosed in pink in Figure 17, is found in virulence factor cryptopain-1 from *C. pavrum* and BsP2 ST4. The region is conserved for both *E. histolytica* proteins (EhP2 and EhP4) as well as both *Blastocystis* ST7 proteins (BsP2 ST7 and BsP4 ST7). This motif has been proven to ensure proper folding and stability in numerous papain family proteases (Kumar et al., 2004). Understanding the role of this conserved motif in *Blastocystis* proteins may provide important awareness of their functional properties and potential implications in host-pathogen interactions. Investigative the impact of the GNFD motif on the folding and stability of *Blastocystis* proteins could unlock paths for studying their role in the pathogenicity mechanisms of this organism.

Both the motifs ERFNIN and GNFD, are found to be conserved in cathepsin L sub-family papain family proteases, as the mediator of prodomain inhibitory activity (Pandey et al., 2009). This implies that these factors might affect the enzymatic activity and virulence of a pathogen, and affect the functioning of a potential host negatively.

3.3.10 N-glycosylation sites

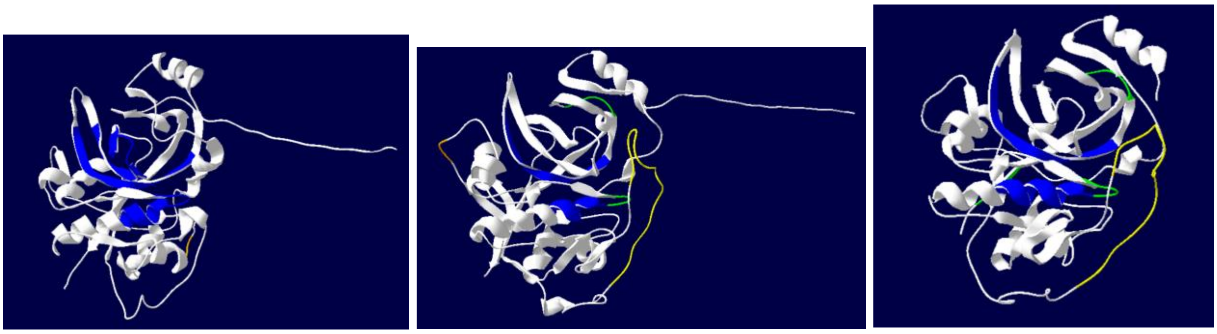
N-glycosylation sites are presented in brown in Figure 17, with the sites found at four locations for cryptopain-1 (CpP3) and at two locations for cysteine proteinase 5 (EhP4). There was none found in any of the *Blastocystis* proteins investigated. EhP4 contains an Asn-X-(Ser/Thr) recognition sequence within the prosequence, which may be posttranslationally modified by glycosylation (Jacobs et al., 1998). Concerning *E. histolytica* pathogenicity, EhP4 appears to be of special importance, since it is the only cysteine protease known so far that is present on the amoeba’s surface (Jacobs et al., 1998). However, the mechanism of surface association remains to be determined, since the molecule does not contain any known surface attachment part (Bruchhaus et al., 2003). Later research has revealed that N-glycosylation could affect cellular functions such as secretion, cytoskeleton organization, proliferation, and apoptosis (Kukuruzinska & Lennon, 1998). This could have been a relevant investigation into a potential treatment, but it was not found in the analyzed *Blastocystis* proteins.

3.3.11 Amidation sites

There is only one amidation site found in *C. pavrum* (sGKK) in Figure 17. The site is not found in any of the proteins in *Blastocystis*. Amidation has a big impact on how well peptides attach to their G-protein-coupled receptors (Kumar et al., 2014). It usually happens at the C-terminal, and the connection with the receptors makes the attachment stronger and improves how well they work together (Shahmiri & Mechler, 2020). Research done on a human parasite (*Schistosoma mansoni*) suggests this site is a target for new chemotherapeutics (Mair et al., 2004).

3.4 3D structure

The 3D models of all proteins are retrieved from Alphafold and Phyre 2 in this study. Consequently, the proteins encompass their signal peptides at the N-terminus. Figure 18 shows the 3D models of results retrieved from cathepsin B (*Giardia*) in *Blastocystis* ST7 and ST4.



GiP1 BsP1 ST7 BsP1 ST4

Figure 18 Illustration of GiP1, BsP1 ST7, and BsP1 ST4 in 3D view of the molecule. The molecule model is a prediction taken from Alphafold and Phyre 2 and edited in Swiss-PDB viewer. Blue regions are active sites, green regions are QxVxG motifs, and the orange region presents an RGD motif. Both *Blastocystis* ST7 and ST4 include a possible occluding loop (yellow).

The 3D figures show structural similarities of virulence factor catB from *G. intestinalis* with both *Blastocystis* subtypes. The active sites of Cys and His are positioned alike, but both *Blastocystis* ST7 and ST4 only form a catalytic dyad of Cys and His, while GiP1 forms a triad with an additional Asn active site. The occluding loop (yellow) in both *Blastocystis* ST7 and

ST4 are similarly positioned. This loop could influence the binding of substrates at the active site and is very interesting for further investigation to find out if this could act as an inhibitor for *Blastocystis*. An occluding loop regulates the entry of substrates, facilitating the chemical reaction, and controlling the release of products in enzymes (Cavallo-Medved et al., 2011; Redzynia et al., 2008). This structural feature is vital for the proper functioning and regulation of various enzymatic processes in biological systems (Illy et al., 1997). This could lead to significant pathological effects.

Figure 19 presents models of cysteine proteinase 2 (A: EhP2) and cryptopain-1 (B: CpP3) in a 3D view with the results for *Blastocystis* ST7 (C: BsP2 ST7) and ST4 (D: BsP2 ST4), as well as cysteine proteinase 5 (E: EhP4) and the compared protein of *Blastocystis* ST7 (F: BsP4 ST7). These are all put together as some of the virulence factors had the same protein results in *Blastocystis*. The colors are coded similarly to the sequence alignment.

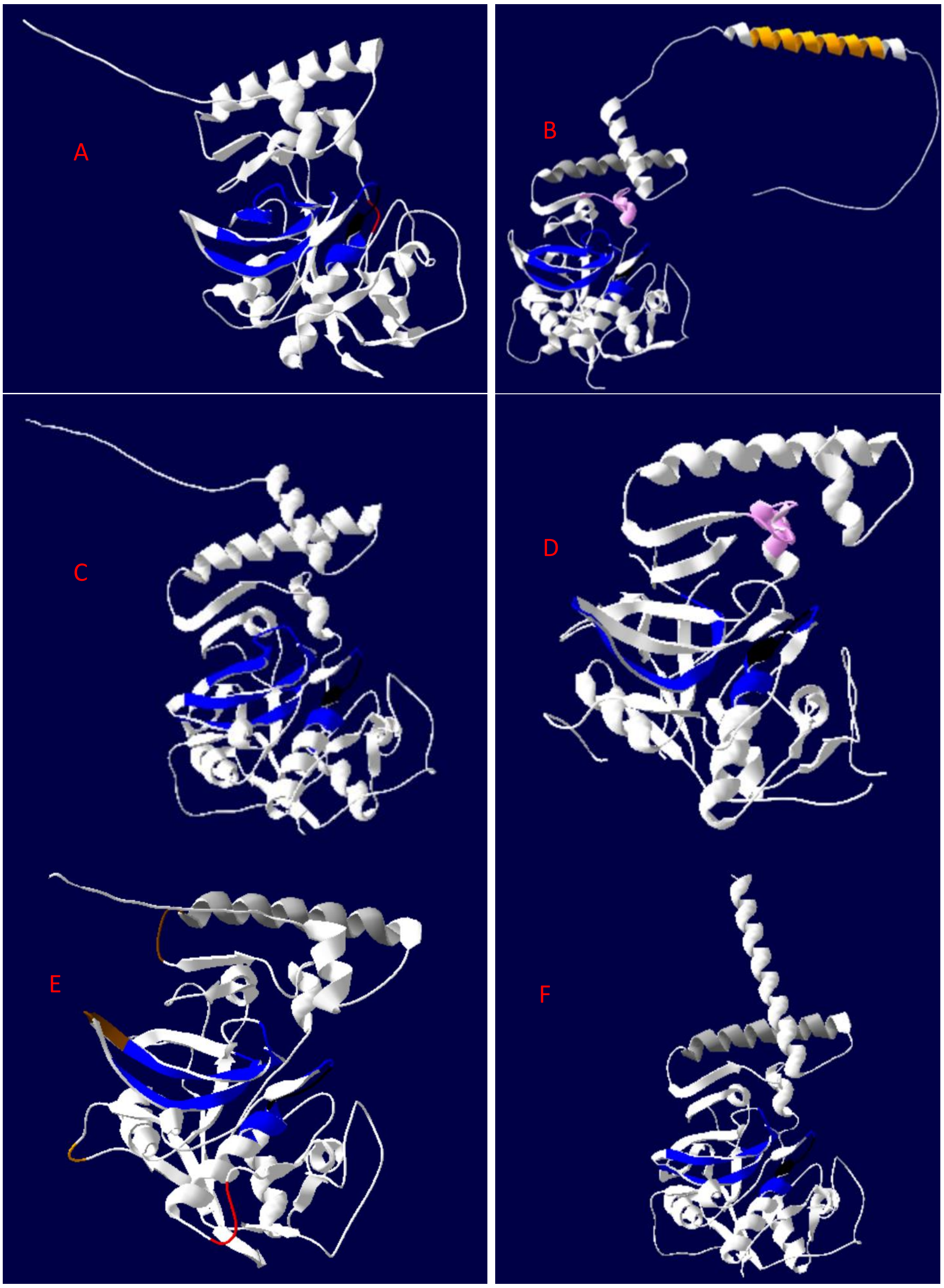


Figure 19 Illustration of 3D representation of (A) EhP2, (B) CpP3, (C) BsP2 ST7 and (D) BsP2 ST4, (E) EhP4, and (F) BsP4 ST7, as a prediction from Alphafold and Phyre 2. Positions highlighted are the active sites (blue), CxxC motif (black), ERFNIN motif (grey), GNDF motif (pink), N-glycosylation site (brown), cAMP- and cGMP-dependent protein kinase phosphorylation sites (red) and RGD motif (orange). CpP3 also has a possible transmembrane domain highlighted in orange at N-terminus and an amidation site shown in purple. Models are edited in Swiss-PDB viewer.

All proteins show structural similarities. Active sites are alike consisting of the catalytic triad, except for *Blastocystis* ST4 which includes a catalytic dyad (lacking Asn active site). The CxxC motif is represented in black and is similarly positioned. The proteins CpP3, EhP4, and BsP4 ST7 include a coiled ERFNIN motif on the outside of the molecule. The GNFD motif is positioned in a way that is like the structures of molecules CpP3 and BsP2 ST4. The molecules lacking both ERFNIN and GNFD motifs share a similar structure even though the motifs are not present.

CpP3 has the exclusive domain of the transmembrane, as well as an amidation site shown as a purple string positioned behind the active site of CpP3. An integral part of a protein spans a biological membrane's phospholipid bilayer, like a cell's outer membrane. Amino acids in these sections interact with the membrane's fatty acyl groups, securing the protein in the membrane. The transmembrane region serves the essential function of anchoring a protein within a biological membrane, such as the cell's plasma membrane (Alberts B, 2002). This region spans the hydrophobic lipid bilayer, and its amino acids interact with the hydrophobic fatty acyl groups of the membrane's phospholipids, thereby firmly embedding the protein within the membrane structure. This anchoring is crucial for the protein to carry out its specific tasks, which might include transporting molecules, receiving signals, or facilitating various cellular processes that involve interactions between the cell's interior and exterior environments. The N-terminal for all proteins is somewhat structurally different, and this is also where the signal peptide is located, in the prepeptide.

3.5 Phylogenetic tree

To build the phylogenetic tree, organisms were identified using BLAST and subsequently assembled on phylogeny.fr. Detailed protein information corresponding to the phylogenetic trees can be found in Appendix 2.

3.5.1 Cathepsin B

The majority of the results of catB in *G. intestinalis* were marine organisms. *Giardia* is found in the habitat of fresh water, which makes it interesting to observe such a prevalence of marine organisms associated with catB in this context. A specific search for the protein in BLAST was also performed for fungi to determine its evolutionary distance. Organisms are organized into colors presenting which kingdom it belongs to: protists are purple, chromists are green, fungi are pink, animal is peach, and plant is yellow (Figure 20).

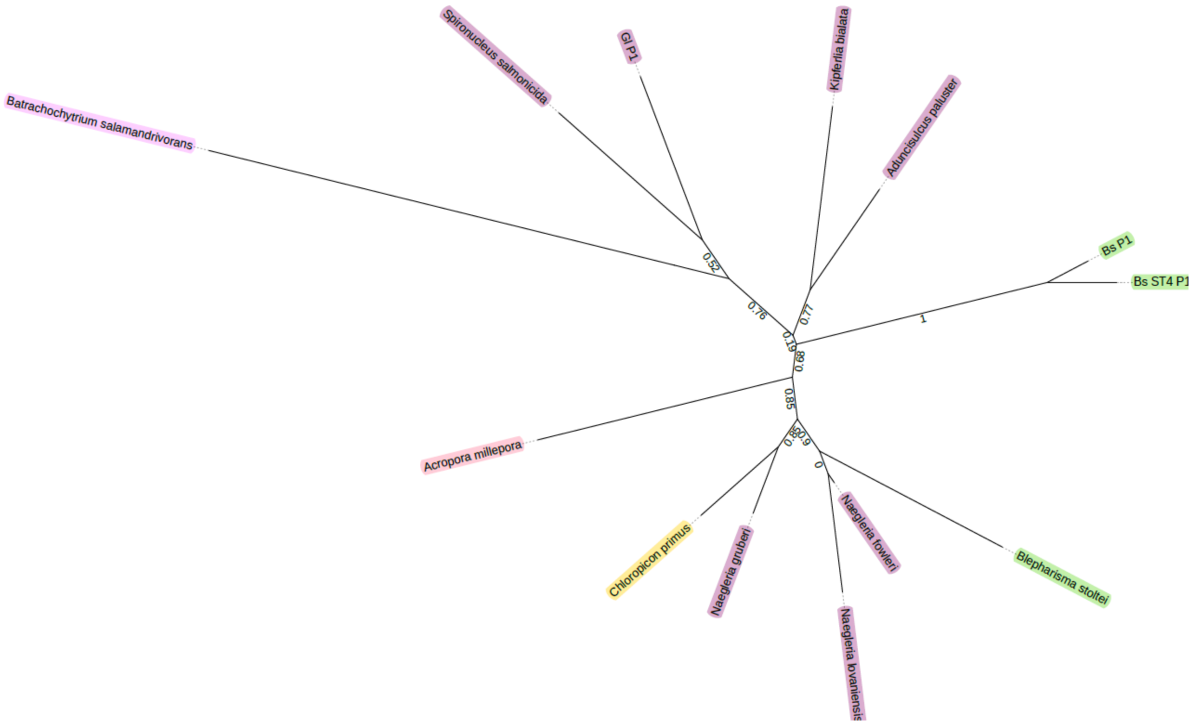


Figure 20 The phylogenetic tree displays the results obtained from the virulence factor cathepsin B in *G. intestinalis*. Organisms are categorized by color, with protists represented in purple, chromists in green, fungi in pink, animals in peach, and plants in yellow. Bootstrap values, denoting branch support, are indicated on the branches. The tree was constructed using phylogeny.fr and edited in iTOL and Adobe Acrobat.

The branch lengths and bootstrap values within the phylogenetic tree symbolize the inferred genetic distances or the level of support for each branching point within the tree. In this search, the branch connecting GiP1 and *Spironucleus salmonicida* exhibits a support value of 0.52. On the other hand, its association with chromalveolates, such as *Blastocystis* ST7 and

ST4, is less distinct, as evidenced by a branch length of 0.19, indicating a relatively distant relationship. Upon closer examination of BsP1 ST7 and BsP1 ST4, their proximity is upheld by a strong support value at their shared node with a support value of 1. When analyzing all branches within the tree, it becomes apparent that yeast shares a common ancestral lineage with GiP1, but this relationship is notably distant, as indicated by the substantial branch length that separates them.

3.5.2 Cysteine proteinase 2

The phylogenetic tree constructed for cysteine proteinase 2 shown in Figure 21 illustrates a wide range of organisms from protists, animals, fungi, plants, chromists, and bacteria. A specific search was performed for fungi to get a better overview of where the protein is evolutionary from. The best hit was *Clostridia* bacterium. The colors present the kingdom they belong to; protists (purple), chromists (green), fungi (pink), animals (peach), bacteria (brown), and plants (yellow). The branch support values indicate the level of confidence in these relationships.

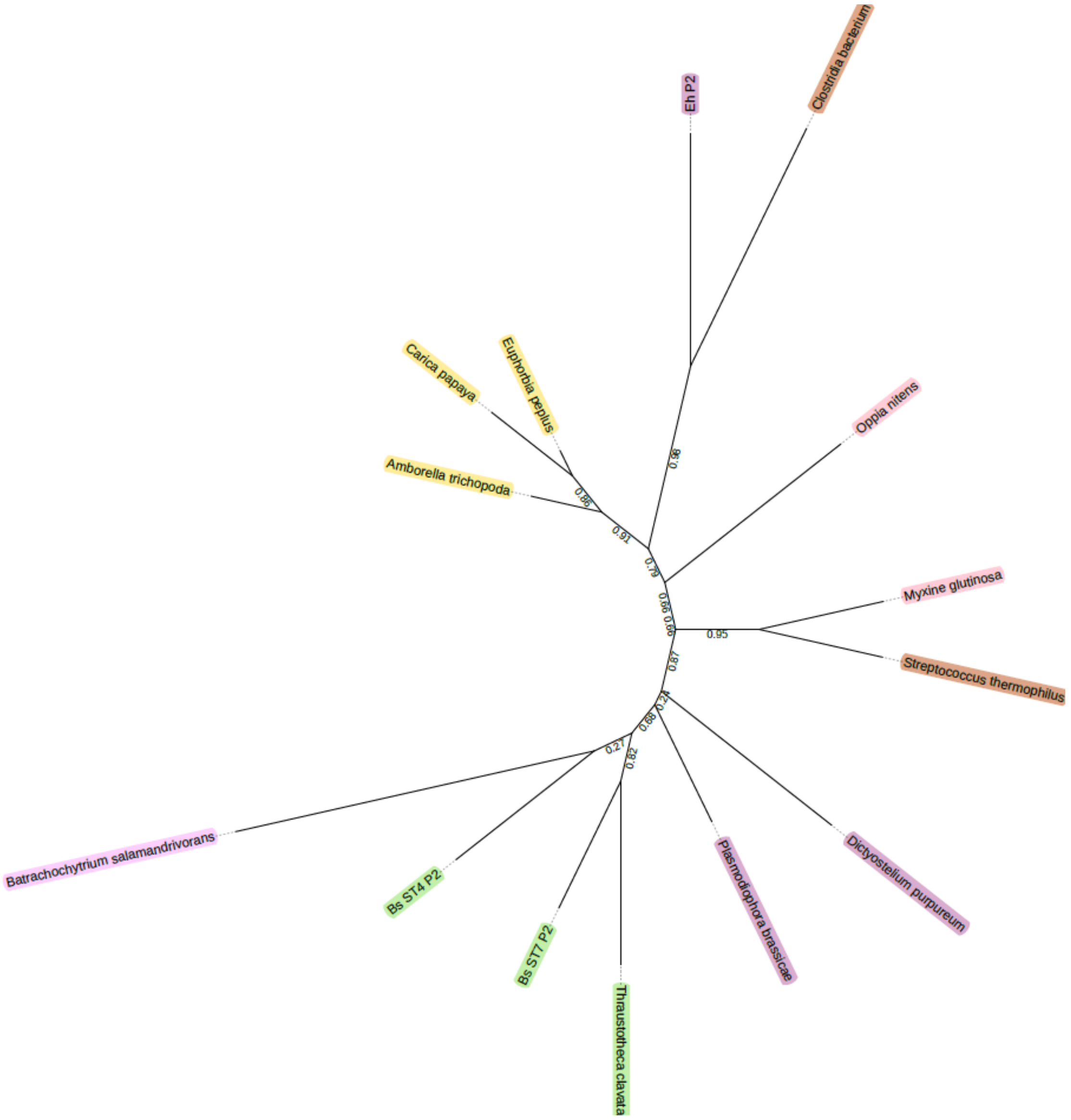


Figure 21 The phylogenetic tree of protein cysteine proteinase 2 from *E. histolytica* was constructed using phylogeny.fr and further edited in iTOL and Adobe Acrobat to assign specific colors to different groups. In this representation, yellow represents plant species, purple represents protists, green represents chromists, pink represents fungi, peach represents animals, and brown represents bacteria. The bootstrap values are indicated on the branches of the tree.

From the search of protein cysteine proteinase 2 in *E. histolytica*, the major result showed to be from plants. *Entamoeba* and plants are evolutionarily distant from each other, but they both belong to the domain of *Eukaryotes*, and cysteine proteinases are found in both plants and microorganisms (Kędzior et al., 2016). The evolutionary relationship of cysteine proteinases found in plants to those in other organisms is based on the shared ancestry of the genes that code for these enzymes. Genes encoding cysteine proteinases could have originated early in the evolution of life, and as organisms diverged and evolved, it is possible that these genes underwent various changes and adaptations to suit the specific functions and requirements of different species. If plants obtained cysteine proteinase genes from prokaryotes, it does not necessarily mean that all other lineages would retain these genes. Evolutionary pressures, genetic drift, and other factors can lead to the loss of genes in different lineages over millions of years, resulting in the diversity of gene presence and absence that we observe in the biological world today (Star & Spencer, 2013). Certain cysteine proteases from plants and animals are homologous to each other (Hughes, 1994).

EhP2 shares a common ancestor with the cluster of plants (*Euphorbia peplus*, *Carica papaya*, *Amborella trichopoda)*, and *Clostridia* bacterium. These organisms are part of the same branch within the phylogenetic tree. The branch that includes EhP2 and the above-mentioned organisms is sister to the animal species (*Oppia nitens* and *Myxine glutinosa*) but is not as closely related. The branch lengths in the Newick format provide information about the genetic divergence or evolutionary distance between the organisms. In this context, EhP2 and *Clostridia* bacterium have a branch length of 0.98, indicating a certain level of genetic divergence from its closest relatives within the group. It is also distinct from *Oppia nitens*, and EhP2 has a certain level of genetic divergence from its closest relatives within this group. In contrast, *Blastocystis* ST7 and ST4 exhibit a more distant relationship with EhCP2 from *E. histolytica*, along with yeast.

3.5.3 Cryptopain-1

From the search of cryptopain-1 in *C. pavrum*, the major hits were found to be in protists and some in plants. Specific searches in fungi and animals were done to see the evolutionary distances to these groups. The colors organize the kingdom they belong to; protists are purple, chromists are green, fungi are pink, the animal is peach, and the plant is yellow (Figure 22). The support values of the branches signify the degree of confidence in these relationships in bootstrap values.

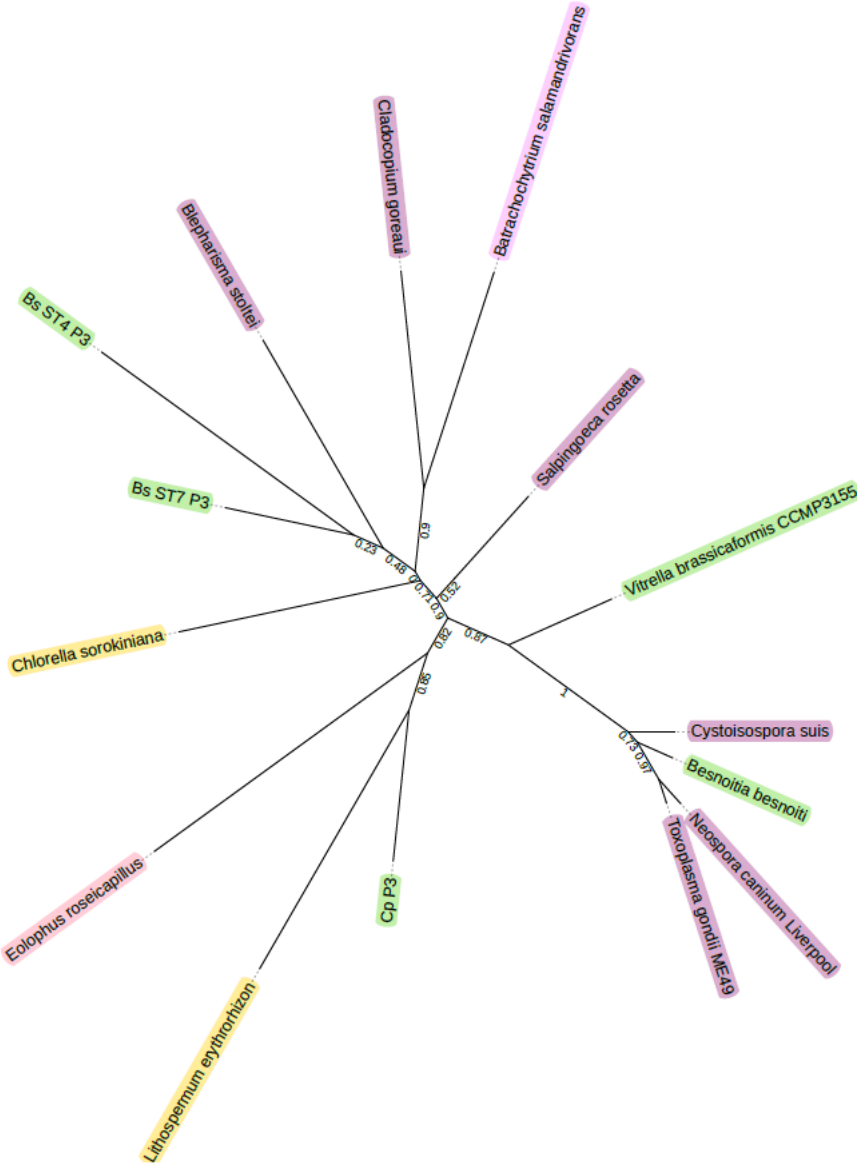


Figure 22 The phylogenetic tree of protein cryptopain-1 from *C. pavrum* was generated through a BLAST search encompassing protists (purple), chromists (green), plants (yellow), animals (peach), and fungi (pink). The tree was constructed using phylogeny.fr and subsequently edited using iTOL and Adobe Acrobat. Bootstrap values indicating branch support are displayed on the tree branches.

CpP3 appears to be more closely related to the plant (*Lithospermum erythrorhizon*) and the animal (*Eolophus roseicapillus*), than to the other organisms in the tree. BsP2 ST4 and BsP2

ST7 share a common ancestor with a branch length of 0.23. This common ancestor is further connected to the protist *Blepharisma stoltei* with a branch length of 0.48. The entire subtree, including BsP2 ST4, BsP2 ST7, and *Blepharisma stoltei*, is connected to the rest of the tree with a branch length of 0.06238. This indicates a distant relationship with cryptopain-1.

3.5.4 Cysteine proteinase 5

From the search for protein cysteine proteinase 5 in E. *histolytica*, many results from animals were found. A specific search in plants and fungi was performed to analyze the distance for different groups. The colors show which kingdom the organisms belong to where protists are purple, chromists are green, fungi are pink, the animal is peach, and the plant is yellow (Figure 23).

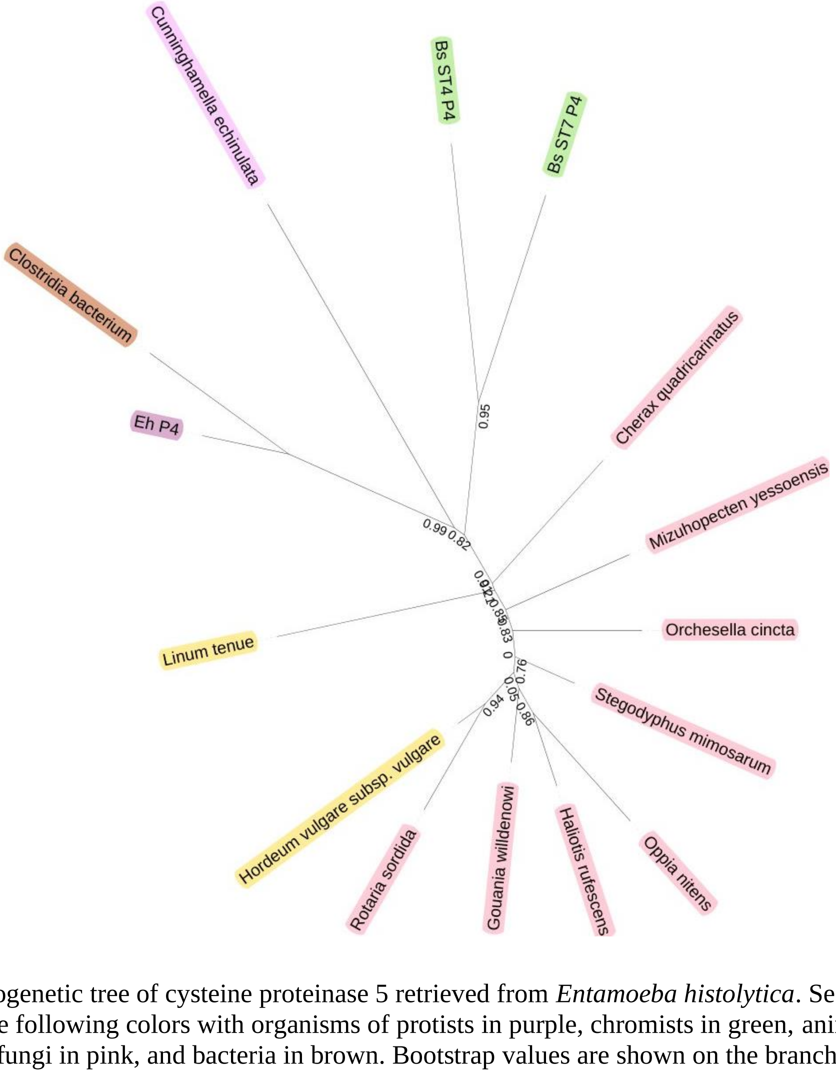


Figure 23 Phylogenetic tree of cysteine proteinase 5 retrieved from *Entamoeba histolytica*. Search results are organized in the following colors with organisms of protists in purple, chromists in green, animals in peach, plants in yellow, fungi in pink, and bacteria in brown. Bootstrap values are shown on the branches. Construction of the tree is done in phylogeny.fr and edited in iTOL and Adobe Acrobat.

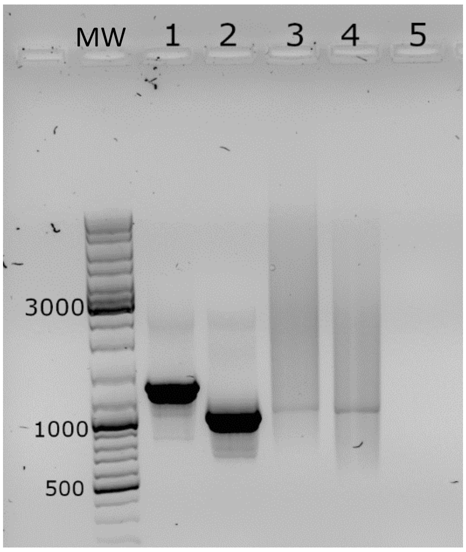
The outcome of the phylogenetic tree shows that the *Clostridia* bacterium is closely related to EhP4 with a support value of 0.99. The fungi *Cunninghamella echinulate* shows a common ancestor with EhP4 but is relatively distant as the branch is extensive. The tree also shows a closer relationship with *Blastocystis* spp. with a support value of 0.84.

The phylogenetic analysis revealed a distinct separation of most of the virulence factors from the other proteins, suggesting a significant evolutionary divergence. This separation likely stems from extensive adaptation to specific organisms over a considerable period, rendering these factors uniquely suited for inducing virulence in their respective hosts. The limited scope of our search, which excluded the host organisms themselves, may contribute to the observed distinctiveness of the results, reinforcing the idea that these virulence factors are highly organism-specific in their function and evolution.

4 Results laboratory work

4.1 Isolation and PCR of promoters and terminators

During the process of amplifying the specific promoters and terminators through PCR, gel electrophoresis was utilized to confirm the correct sizes of the fragments. Since each primer was designed and created for a specific fragment, it was necessary to verify its effectiveness with individual promoters and terminators before moving forward with the cloning process into the vector. As illustrated in Figure 24, it shows that PC1A did not yield the correct size of 2207 bp (well 3 and 4), whereas the 60SRPL32 promoters were successfully prepared with a correct size of 1295 bp (well 1) and 1000 bp (well 2). All primers used through the experiments can be found in Appendix 3.



MW: Molecular Weight

1: 60SRPL32 V1295

2: 60SRPL32 V1000

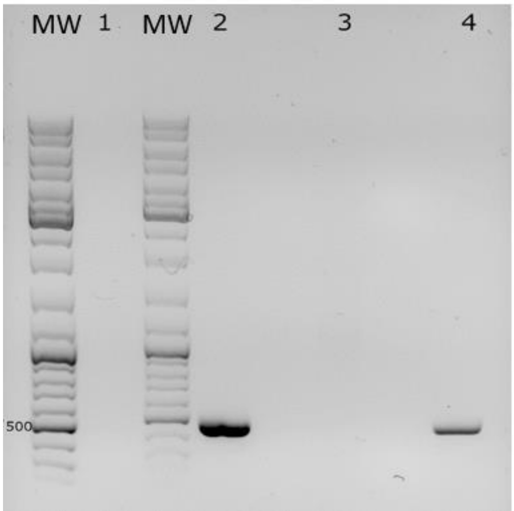
3: PC1A V2207

4: PC1A V2207 «diluted»

5: neg control

Figure 24 Gel electrophoresis of PCR. The gel is made of 1% agarose gel with GelRed staining. Lane with molecular weight (MW) shows the DNA ladder used and estimates the size of the samples. DNA ladder mix was used, and wells are numbered by gene and promoter (1-5). Numbers 1 and 2 show successful size, and the primers can therefore be used further. Number 3 and 4 is not successful size. Negative control shows no bands meaning no contamination. Each well was loaded with 5 µl of sample and 3 µl for the DNA ladder mix.

Results of gel electrophoresis of successful terminators of both genes PC1A and 60SRPL32 are shown in Figure 25 below. Neither of the promoters showed any visible band in this attempt.



1: 60S Ribosomal Protein L32 V1295

2: 60S Ribosomal Protein L32 Terminator

3: PC1A V2207

4: PC1A Terminator

Figure 25 Gel electrophoresis of PCR with promoters and terminators for both 60SRPL32 and PC1A. The terminators for both genes exhibited the expected size of 500 bp, indicating successful amplification. However, in lane 1 and 3, the promoters did not show any bands, suggesting potential issues during the experimental procedure. To assess the gel's integrity and size markers, a DNA ladder was loaded into two separate lanes for testing purposes. In each well, 5 µl of the sample and 3 µl of the DNA ladder mix were loaded.

This is only a summary of many attempts to isolate all promoters and terminators. Because of not enough genomic DNA, PC1A promoter V2207 was not successfully isolated. Except for that, all versions of promoters and terminators of both genes PC1A and 60SRPL32 were successfully isolated through PCR, gel wash, and purification.

4.2 Colony PCR

To verify the correct size of the insert within the vector, a colony PCR approach was employed. Bacterial colonies grown on agar plates were collected and mixed with the PCR mixture. The resulting PCR products were then subjected to gel electrophoresis to confirm their sizes. A total of 19 colonies were picked from both the high-concentration and low-concentration samples. Figure 26 shows the results of gel electrophoresis after colony PCR of genes PC1A and 60SRPL32. This is just a brief outline; all promoters and terminators were successfully inserted into the vector.

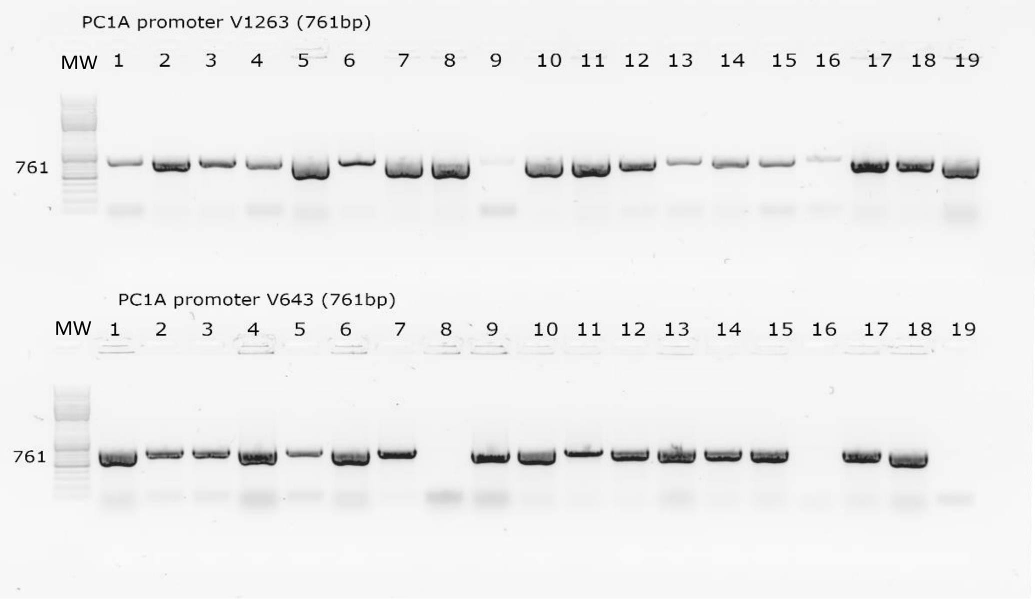

PC1A promoter V1263 (761bp)
MW 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
761
PC1A promoter V643 (761bp)
MW 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
761

Figure 26 Gel electrophoresis of Peptidase C1A versions V1263 and V643. Samples from colonies grown on an agar plate and resuspended in PCR reaction mix. Both samples are expected to be 761bp. Molecular weight (MW) with DNA ladder estimating the size of the samples. In each well it was loaded 5 µl of sample and 3 µl for the DNA ladder mix.

Unfortunately, the completion of the lab work was hindered due to an unpredictable circumstance. This unexpected situation prevented the successful completion of the planned experiments and analysis. To complete the study, it is necessary to finalize the cloning process for all promoters and terminators related to genes PC1A and 60SRPL32 into the appropriate vectors. Subsequently, the transformation into *Blastocystis* cells must proceed, followed by the application of the Nanoluc luciferase assay and antibody staining methods outlined in the procedures section. Using microscopy, the expression levels of the genes of interest should be assessed, along with any potential impact on biological pathways.

5 Conclusion

The discovery of possible virulence factors inside *Blastocystis* has generated interesting challenges concerning the pathogenic potential of the organism. By comparing the previously identified virulence factors in the pathogens of interest; *Giardia intestinalis*, *Entamoeba histolytica*, and *Cryptosporidium pavrum*, it appears that *Blastocystis* subtypes ST7 and ST4 may share similar mechanisms and functions, suggesting the possibility of pathogenicity. This study highlights the importance of researching *Blastocystis* and its many subtypes to solve the puzzles surrounding its pathogenic nature. In addition, this thesis provides material and data pointing out the need to resolve the fundamental question of whether *Blastocystis* causes pathogenicity in hosts.

Furthermore, the similarities revealed between *Blastocystis* and known virulence factors give a good platform for future research. As *Blastocystis* exhibits extensive subtype variation, some subtypes could cause illness while others do not (Wawrzyniak et al., 2012). Consequently, individuals might undergo treatment for *Blastocystis* without any adverse effects, owing to this diversity among its subtypes (Scanlan et al., 2015). Determining whether *Blastocystis* plays a role in disease is not only important for identifying suitable treatments but also has the potential to enhance the economic aspects of the healthcare system. This uncertainty arises from its prevalence in both individuals without health issues and those experiencing intestinal symptoms, such as diarrhea and IBS (Scanlan et al., 2014). In the United States alone, IBS imposes a substantial economic burden on the healthcare system due to extensive resource usage, leading to direct medical expenses and indirect (workplace) of a total of $30 billion annually (Leong et al., 2003). In Europe, IBS is identified as the main cause of hospitalizations and emergency room visits (Tack et al., 2019). As *Blastocystis* has been correlated with IBS symptoms, understanding the diversity of the *Blastocystis* subtypes and their relation to symptoms will help decrease these costs (Hulisz, 2004). This will also help to minimize the demand on the healthcare system, especially since the exact involvement of *Blastocystis* in digestive disorders in people is uncertain (Boorom et al., 2008). Furthermore, the research might pave the way for new treatment procedures and prevention methods.

Despite encountering challenges in the molecular part of this project, the completion of the planned experiments and analysis has the potential to reveal fundamental insights into *Blastocystis* and its genetic composition and cellular processes. Understanding its interactions with the host immune system and the mechanisms underlying its pathogenicity in various subtypes is essential to improve our knowledge of the biology of the parasite, its function in health and illness, and potential new strategies for parasite management. Nonetheless, the reliability of our findings needs to be further strengthened through additional research *in vivo* research.

In summary, this study is an important step toward understanding the complexities of *Blastocystis* pathogenicity. The findings presented lay a framework for future research, providing a look into the complex world of parasitic illnesses. As scientists learn more about *Blastocystis* and its subtypes, we get closer to finding appropriate treatments that could improve global health outcomes and our capacity to tackle parasitic illnesses efficiently.

References

Abrahamsen, M. S., Templeton, T. J., Enomoto, S., Abrahante, J. E., Zhu, G., Lancto, C. A., Deng, M., Liu, C., Widmer, G., Tzipori, S., Buck, G. A., Xu, P., Bankier, A. T., Dear, P. H., Konfortov, B. A., Spriggs, H. F., Iyer, L., Anantharaman, V., Aravind, L., & Kapur, V. (2004). Complete Genome Sequence of the Apicomplexan, *Cryptosporidium parvum*. *Science*, *304*(5669), 441-445. https://doi.org/10.1126/science.1094786

Addgene. (n.d.). *Primer Design for PCR*. https://www.addgene.org/protocols/primer-design/

Adeyemo, F. E., Singh, G., Reddy, P., Bux, F., & Stenström, T. A. (2019). Efficiency of chlorine and UV in the inactivation of *Cryptosporidium* and *Giardia* in wastewater. *PloS one*, *14*(5), e0216040. https://doi.org/10.1371/journal.pone.0216040

Aich, P., & Biswas, S. (2018). Highly Conserved Arg Residue of ERFNIN Motif of Pro-Domain is Important for pH-Induced Zymogen Activation Process in Cysteine Cathepsins K and L. *Cell Biochem Biophys*, *76*(1-2), 219-229. https://doi.org/10.1007/s12013-017-0838-x

Ajjampur, S. S., & Tan, K. S. (2016). Pathogenic mechanisms in *Blastocystis* spp. - Interpreting results from *in vitro* and *in vivo* studies. *Parasitol Int*, *65*(6 Pt B), 772-779. https://doi.org/10.1016/j.parint.2016.05.007

Alberts B, J. A., Lewis J, et al. (2002). *Membrane Proteins*. 4th edition. New York: Garland Science. https://www.ncbi.nlm.nih.gov/books/NBK26878/

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

Allain, T., Fekete, E., & Buret, A. G. (2019). *Giardia* Cysteine Proteases: The Teeth behind the Smile. *Trends in Parasitology*, *35*(8), 636-648. https://doi.org/10.1016/j.pt.2019.06.003

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, *215*(3), 403-410. https://doi.org/10.1016/s0022-2836(05)80360-2

Andersen, L. O. B., & Stensvold, C. R. (2016). *Blastocystis* in Health and Disease: Are We Moving from a Clinical to a Public Health Perspective? *Journal of Clinical Microbiology*, *54*(3), 524-528. https://doi.org/doi:10.1128/jcm.02520-15

Ankri, S., Stolarsky, T., Bracha, R., Padilla-Vaca, F., & Mirelman, D. (1999). Antisense Inhibition of Expression of Cysteine Proteinases Affects *Entamoeba histolytica*-Induced Formation of Liver Abscess in Hamsters. *Infection and immunity*, *67*(1), 421-422. https://doi.org/10.1128/iai.67.1.421-422.1999

Argüello-García, R., Carrero, J. C., & Ortega-Pierres, M. G. (2023). Extracellular Cysteine Proteases of Key Intestinal Protozoan Pathogens-Factors Linked to Virulence and Pathogenicity. *International Journal of Molecular Sciences*, *24*(16), 12850. https://www.mdpi.com/1422-0067/24/16/12850

Argüello-García, R., & Ortega-Pierres, M. G. (2021). *Giardia duodenalis* Virulence — "To Be, or Not To Be". *Current Tropical Medicine Reports*, *8*(4), 246-256. https://doi.org/10.1007/s40475-021-00248-z

Asaoka, Y. (2012). Chapter thirteen - Phosphorylation of Gli by cAMP-Dependent Protein Kinase. In G. Litwack (Ed.), *Vitamins & Hormones* (Vol. 88, pp. 293-307). Academic Press. https://doi.org/10.1016/B978-0-12-394622-5.00013-4

Barrett, A. J. (2000). Proteases. *Current protocols in protein science*, *21*(1), 21.21.21-21.21.12. https://doi.org/10.1002/0471140864.ps2101s21

Bera, P. P., & Schaefer, H. F., 3rd. (2005). (G-H)*-C and G-(C-H)* radicals derived from the guanine.cytosine base pair cause DNA subunit lesions. *Proc Natl Acad Sci U S A*, *102*(19), 6698-6703. https://doi.org/10.1073/pnas.0408644102

Bergtrom, G. (2022). Protein Domains, Motifs, and Folds in Protein Structure. In *Basic Cell and Molecular Biology 5e: What We Know and How We Find Out*. https://dc.uwm.edu/biosci_facbooks_bergtrom/14

Betanzos, A., Bañuelos, C., & Orozco, E. (2019). Host Invasion by Pathogenic Amoebae: Epithelial Disruption by Parasite Proteins. *Genes (Basel)*, *10*(8). https://doi.org/10.3390/genes10080618

Beyhan, Y. E., Yilmaz, H., Cengiz, Z. T., & Ekici, A. (2015). Clinical significance and prevalence of *Blastocystis hominis* in Van, Turkey. *Saudi Medical Journal*, *36*(9), 1118-1121. https://doi.org/10.15537/smj.2015.9.12444

Boorom, K. F., Smith, H., Nimri, L., Viscogliosi, E., Spanakos, G., Parkar, U., Li, L.-H., Zhou, X.-N., Ok, Ü. Z., Leelayoova, S., & Jones, M. S. (2008). Oh my aching gut: irritable bowel syndrome, *Blastocystis*, and asymptomatic infection. *Parasites & Vectors*, *1*(1), 40. https://doi.org/10.1186/1756-3305-1-40

Bouzid, M. (2014). WATERBORNE PARASITES | Detection of Food- and Waterborne Parasites: Conventional Methods and Recent Developments. In C. A. Batt & M. L. Tortorello (Eds.), *Encyclopedia of Food Microbiology (Second Edition)* (pp. 773-781). Academic Press. https://doi.org/10.1016/B978-0-12-384730-0.00355-4

Bouzid, M., Hunter, P. R., Chalmers, R. M., & Tyler, K. M. (2013). *Cryptosporidium* Pathogenicity and Virulence. *Clinical microbiology reviews*, *26*(1), 115-134. https://doi.org/10.1128/cmr.00076-12

Britannica, T. E. o. E. (2022). Cysteine. *Encyclopedia Britannica.* . https://www.britannica.com/science/cysteine

Brock, T. D., Madigan, M. T., Martinko, J. M., & Parker, J. (2003). *Brock biology of microorganisms* (10th ed.). Upper Saddle River (N.J.) : Prentice-Hall. http://lib.ugent.be/catalog/rug01:000745286

Bruchhaus, I., Loftus, B. J., Hall, N., & Tannich, E. (2003). The Intestinal Protozoan Parasite *Entamoeba histolytica* Contains 20 Cysteine Protease Genes, of Which Only a Small Subset Is Expressed during *In Vitro* Cultivation. *Eukaryotic Cell*, *2*(3), 501-509. https://doi.org/10.1128/ec.2.3.501-509.2003

Caretta, A., & Mucignat-Caretta, C. (2011). Protein kinase a in cancer. *Cancers (Basel)*, *3*(1), 913-926. https://doi.org/10.3390/cancers3010913

Casadevall, A., & Pirofski, L.-a. (2009). Virulence factors and their mechanisms of action: the view from a damage–response framework. *Journal of Water and Health*, *7*(S1), S2-S18. https://doi.org/10.2166/wh.2009.036

Cavallo-Medved, D., Moin, K., & Sloane, B. (2011). Cathepsin B: Basis Sequence: Mouse. *AFCS Nat Mol Pages*, *2011*. https://pubmed.ncbi.nlm.nih.gov/28781583/

Chin, A. C., Teoh, D. A., Scott, K. G.-E., Meddings, J. B., Macnaughton, W. K., & Buret, A. G. (2002). Strain-Dependent Induction of Enterocyte Apoptosis by *Giardia lamblia* Disrupts Epithelial Barrier Function in a Caspase-3-Dependent Manner. *Infection and immunity*, *70*(7), 3673-3680. https://doi.org/10.1128/iai.70.7.3673-3680.2002

Chin, K. V., Yang, W. L., Ravatn, R., Kita, T., Reitman, E., Vettori, D., Cvijic, M. E., Shin, M., & Iacono, L. (2002). Reinventing the wheel of cyclic AMP: novel mechanisms of cAMP signaling. *Ann N Y Acad Sci*, *968*, 49-64. https://doi.org/10.1111/j.1749-6632.2002.tb04326.x

Chou, A., & Austin, R. L. (2023). *Entamoeba histolytica* Infection. In *StatPearls*. StatPearls Publishing.

Cotton, J. A., Bhargava, A., Ferraz, J. G., Yates, R. M., Beck, P. L., & Buret, A. G. (2014). *Giardia duodenalis* Cathepsin B Proteases Degrade Intestinal Epithelial Interleukin-8 and Attenuate Interleukin-8-Induced Neutrophil Chemotaxis. *Infection and immunity*, *82*(7), 2772-2787. https://doi.org/10.1128/iai.01771-14

Coulombe, R., Grochulski, P., Sivaraman, J., Ménard, R., Mort, J. S., & Cygler, M. (1996). Structure of human procathepsin L reveals the molecular basis of inhibition by the prosegment. *Embo j*, *15*(20), 5492-5503. https://pubmed.ncbi.nlm.nih.gov/8896443/

Cuellar, P., Hernández-Nava, E., García-Rivera, G., Chávez-Munguía, B., Schnoor, M., Betanzos, A., & Orozco, E. (2017). *Entamoeba histolytica* EhCP112 Dislocates and Degrades Claudin-1 and Claudin-2 at Tight Junctions of the Intestinal Epithelium. *Front Cell Infect Microbiol*, *7*, 372. https://doi.org/10.3389/fcimb.2017.00372

Cuesta-Astroz, Y., Scholte, L. L. S., Pais, F. S.-M., Oliveira, G., & Nahum, L. A. (2014). Evolutionary analysis of the cystatin family in three *Schistosoma* species [Original Research]. *Frontiers in Genetics*, *5*. https://doi.org/10.3389/fgene.2014.00206

Cummings, R. D., & van Die, I. (2015). *Parasitic Infections* (3rd ed.). Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY). https://www.ncbi.nlm.nih.gov/books/NBK453068

Current, W. L., & Garcia, L. S. (1991). Cryptosporidiosis. *Clin Microbiol Rev*, *4*(3), 325-358. https://doi.org/10.1128/cmr.4.3.325

Deng, L., Lee, J. W. J., & Tan, K. S. W. (2022). Infection with pathogenic *Blastocystis* ST7 is associated with decreased bacterial diversity and altered gut microbiome profiles in diarrheal patients. *Parasit Vectors*, *15*(1), 312. https://doi.org/10.1186/s13071-022-05435-z

Deng, L., & Tan, K. S. W. (2022). Interactions between *Blastocystis* subtype ST4 and gut microbiota *in vitro*. *Parasites & Vectors*, *15*(1), 80. https://doi.org/10.1186/s13071-022-05194-x

Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J. F., Guindon, S., Lefort, V., Lescot, M., Claverie, J. M., & Gascuel, O. (2008). Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res*, *36*(Web Server issue), W465-469. https://doi.org/10.1093/nar/gkn180

Dixon, B. R. (2021). *Giardia duodenalis* in humans and animals - Transmission and disease. *Res Vet Sci*, *135*, 283-289. https://doi.org/10.1016/j.rvsc.2020.09.034

Drag, M. (2013). Chapter 478 - OTU1 Peptidase. In N. D. Rawlings & G. Salvesen (Eds.), *Handbook of Proteolytic Enzymes (Third Edition)* (pp. 2121-2123). Academic Press. https://doi.org/10.1016/B978-0-12-382219-2.00477-4

Dutt, S., Singh, V., Marla, S., & Kumar, A. (2010). In silico Analysis of Sequential, Structural and Functional Diversity of Wheat Cystatins and Its Implication in Plant Defense. *Genomics, proteomics & bioinformatics*, *8*, 42-56. https://doi.org/10.1016/S1672-0229(10)60005-8

Emery, S. J., Mirzaei, M., Vuong, D., Pascovici, D., Chick, J. M., Lacey, E., & Haynes, P. A. (2016). Induction of virulence factors in *Giardia duodenalis* independent of host attachment. *Scientific Reports*, *6*(1), 20765. https://doi.org/10.1038/srep20765

Espinosa-Cantellano, M., & Martínez-Palomo, A. (2000). Pathogenesis of intestinal amebiasis: from molecules to disease. *Clin Microbiol Rev*, *13*(2), 318-331. https://doi.org/10.1128/cmr.13.2.318

Examining your blast results. Fungal Diversity Survey. https://fundis.org/component/sppagebuilder/41-examining-your-blast-results

Faheem, M., Martins-de-Sa, D., Vidal, J. F. D., Álvares, A. C. M., Brandão-Neto, J., Bird, L. E., Tully, M. D., von Delft, F., Souto, B. M., Quirino, B. F., Freitas, S. M., & Barbosa, J. A. R. G. (2016). Functional and structural characterization of a novel putative cysteine protease cell wall-modifying multi-domain enzyme selected from a microbial metagenome. *Scientific Reports*, *6*(1), 38031. https://doi.org/10.1038/srep38031

Fairlie, D. P., Tyndall, J. D., Reid, R. C., Wong, A. K., Abbenante, G., Scanlon, M. J., March, D. R., Bergman, D. A., Chai, C. L., & Burkett, B. A. (2000). Conformational selection of inhibitors and substrates by proteolytic enzymes: implications for drug design and polypeptide processing. *J Med Chem*, *43*(7), 1271-1281. https://doi.org/10.1021/jm990315t

Faust, D. M., & Guillen, N. (2012). Virulence and virulence factors in *Entamoeba histolytica,* the agent of human amoebiasis. *Microbes and Infection*, *14*(15), 1428-1441. https://doi.org/10.1016/j.micinf.2012.05.013

Fayer, R., Morgan, U., & Upton, S. J. (2000). Epidemiology of *Cryptosporidium*: transmission, detection and identification. *Int J Parasitol*, *30*(12-13), 1305-1322. https://doi.org/10.1016/s0020-7519(00)00135-1

Figaj, D., Ambroziak, P., Przepiora, T., & Skorko-Glonek, J. (2019). The Role of Proteases in the Virulence of Plant Pathogenic Bacteria. *Int J Mol Sci*, *20*(3). https://doi.org/10.3390/ijms20030672

Fomenko, D. E., & Gladyshev, V. N. (2003). Identity and functions of CxxC-derived motifs. *Biochemistry*, *42*(38), 11214-11225. https://doi.org/10.1021/bi034459s

Gerace, E., Lo Presti, V. D. M., & Biondo, C. (2019). *Cryptosporidium* Infection: Epidemiology, Pathogenesis, and Differential Diagnosis. *Eur J Microbiol Immunol (Bp)*, *9*(4), 119-123. https://doi.org/10.1556/1886.2019.00019

Guerrant, R. L. (1997). Cryptosporidiosis: an emerging, highly infectious threat. *Emerg Infect Dis*, *3*(1), 51-57. www.ncbi.nlm.nih.gov/pmc/articles/PMC2627589/

Haidar, M., Ramdani, G., Kennedy, E. J., & Langsley, G. (2017). PKA and Apicomplexan Parasite Diseases. *Horm Metab Res*, *49*(4), 296-300. https://doi.org/10.1055/s-0042-118459

Hall, T. A. (1999). *BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids. Symp. Ser. 41:95-98.* In

He, C., Nora, G. P., Schneider, E. L., Kerr, I. D., Hansell, E., Hirata, K., Gonzalez, D., Sajid, M., Boyd, S. E., Hruz, P., Cobo, E. R., Le, C., Liu, W. T., Eckmann, L., Dorrestein, P. C., Houpt, E. R., Brinen, L. S., Craik, C. S., Roush, W. R., McKerrow, J., … Reed, S. L. (2010). A novel *Entamoeba histolytica* cysteine proteinase, EhCP4, is key for invasive amebiasis and a therapeutic target. . *The Journal of biological chemistry*, *285(24)*, 18516–18527. https://doi.org/10.1074/jbc.M109.086181

Hellberg, A., Nickel, R., Lotter, H., Tannich, E., & Bruchhaus, I. (2001). Overexpression of cysteine proteinase 2 in *Entamoeba histolytica* or *Entamoeba dispar* increases amoeba-induced monolayer destruction *in vitro* but does not augment amoebic liver abscess formation in gerbils. *Cellular Microbiology*, *3*(1), 13-20. https://doi.org/10.1046/j.1462-5822.2001.00086.x

Ho, L. C., Singh, M., Suresh, G., Ng, G. C., & Yap, E. H. (1993). Axenic culture of *Blastocystis hominis* in Iscove's modified Dulbecco's medium. *Parasitology Research*, *79*(7), 614-616. https://doi.org/10.1007/BF00932249

Hou, Y., Mortimer, L., & Chadee, K. (2010). *Entamoeba histolytica* cysteine proteinase 5 binds integrin on colonic cells and stimulates NFkappaB-mediated pro-inflammatory responses. *J Biol Chem*, *285*(46), 35497-35504. https://doi.org/10.1074/jbc.M109.066035

Hughes, A. L. (1994). Evolution of cysteine proteinases in eukaryotes. *Mol Phylogenet Evol*, *3*(4), 310-321. https://doi.org/10.1006/mpev.1994.1038

Hulisz, D. (2004). The burden of illness of irritable bowel syndrome: current challenges and hope for the future. *J Manag Care Pharm*, *10*(4), 299-309. https://doi.org/10.18553/jmcp.2004.10.4.299

Illy, C., Quraishi, O., Wang, J., Purisima, E., Vernet, T., & Mort, J. S. (1997). Role of the Occluding Loop in Cathepsin B Activity *. *Journal of Biological Chemistry*, *272*(2), 1197-1202. https://doi.org/10.1074/jbc.272.2.1197

Irmer, H., Tillack, M., Biller, L., Handal, G., Leippe, M., Roeder, T., Tannich, E., & Bruchhaus, I. (2009). Major cysteine peptidases of *Entamoeba histolytica* are required for aggregation and digestion of erythrocytes but are dispensable for phagocytosis and cytopathogenicity. *Molecular microbiology*, *72*(3), 658-667. https://doi.org/10.1111/j.1365-2958.2009.06672.x

Javed, K., Ebertz, A. . (2022). Primer Design Guide – The Top 5 Factors to Consider For Optimum Performance. *Eurofins*. https://the-dna-universe.com/2022/09/05/primer-design-guide-the-top-5-factors-to-consider-for-optimum-performance/

Jenkins, M. B., Eaglesham, B. S., Anthony, L. C., Kachlany, S. C., Bowman, D. D., & Ghiorse, W. C. (2010). Significance of wall structure, macromolecular composition, and surface polymers to the survival and transport of *Cryptosporidium parvum* oocysts. *Appl Environ Microbiol*, *76*(6), 1926-1934. https://doi.org/10.1128/aem.02295-09

Jiménez, P., Muñoz, M., & Ramírez, J. D. (2022). An update on the distribution of *Blastocystis* subtypes in the Americas. *Heliyon*, *8*(12), e12592. https://doi.org/10.1016/j.heliyon.2022.e12592

Junger, W. G. (2008). Purinergic regulation of neutrophil chemotaxis. *Cell Mol Life Sci*, *65*(16), 2528-2540. https://doi.org/10.1007/s00018-008-8095-1

Kantor, M., Abrantes, A., Estevez, A., Schiller, A., Torrent, J., Gascon, J., Hernandez, R., & Ochner, C. (2018). *Entamoeba Histolytica*: Updates in Clinical Manifestation, Pathogenesis, and Vaccine

Development. *Can J Gastroenterol Hepatol*, *2018*, 4601420. https://doi.org/10.1155/2018/4601420

Kapp, K., Schrempf, S., Lemberg, M., & Dobberstein, B. (2009). Post-Targeting Functions of Signal Peptides. https://www.ncbi.nlm.nih.gov/books/NBK6322/

Karki, G. (2017). *Entamoeba histolytica*: Morphology, life cycle, Pathogenesis, clinical manifestation, lab diagnosis and Treatment. https://www.onlinebiologynotes.com/entamoeba-histolytica-morphology-life-cycle-pathogenesis-clinical-manifestation-lab-diagnosis-treatment/

Kędzior, M., Seredyński, R., & Gutowicz, J. (2016). Microbial inhibitors of cysteine proteases. *Med Microbiol Immunol*, *205*(4), 275-296. https://doi.org/10.1007/s00430-016-0454-1

Kermasha, S., & Eskin, M. N. A. (2021). Chapter Nine - Selected industrial enzymes. In S. Kermasha & M. N. A. Eskin (Eds.), *Enzymes* (pp. 259-305). Academic Press. https://doi.org/10.1016/B978-0-12-800217-9.00009-5

Khalil, I. A., Troeger, C., Rao, P. C., Blacker, B. F., Brown, A., Brewer, T. G., Colombara, D. V., De Hostos, E. L., Engmann, C., Guerrant, R. L., Haque, R., Houpt, E. R., Kang, G., Korpe, P. S., Kotloff, K. L., Lima, A. A. M., Petri, W. A., Jr., Platts-Mills, J. A., Shoultz, D. A., . . . Mokdad, A. H. (2018). Morbidity, mortality, and long-term consequences associated with diarrhoea from *Cryptosporidium* infection in children younger than 5 years: a meta-analyses study. *Lancet Glob Health*, *6*(7), e758-e768. https://doi.org/10.1016/s2214-109x(18)30283-3

Khatri, V., Chauhan, N., & Kalyanasundaram, R. (2020). Parasite Cystatin: Immunomodulatory Molecule with Therapeutic Activity against Immune Mediated Disorders. *Pathogens*, *9*(6). https://doi.org/10.3390/pathogens9060431

Kissoon-Singh, V., Moreau, F., Trusevych, E., & Chadee, K. (2013). *Entamoeba histolytica* Exacerbates Epithelial Tight Junction Permeability and Proinflammatory Responses in Muc2−/− Mice. *The American Journal of Pathology*, *182*(3), 852-865. https://doi.org/10.1016/j.ajpath.2012.11.035

Kukuruzinska, M. A., & Lennon, K. (1998). Protein N-glycosylation: molecular genetics and functional significance. *Crit Rev Oral Biol Med*, *9*(4), 415-448. https://doi.org/10.1177/10454411980090040301

Kumar, A., Dasaradhi, P. V. N., Chauhan, V. S., & Malhotra, P. (2004). Exploring the role of putative active site amino acids and pro-region motif of recombinant falcipain-2: a principal hemoglobinase of *Plasmodium falciparum*. *Biochemical and Biophysical Research Communications*, *317*(1), 38-45. https://doi.org/10.1016/j.bbrc.2004.02.177

Kumar, A., & Kaur, J. (2014). Primer Based Approach for PCR Amplification of High GC Content Gene: *Mycobacterium* Gene as a Model. *Molecular Biology International*, *2014*, 937308. https://doi.org/10.1155/2014/937308

Kumar, D., Eipper, B. A., & Mains, R. E. (2014). Amidation☆. In *Reference Module in Biomedical Sciences*. Elsevier. https://doi.org/10.1016/B978-0-12-801238-3.04040-X

Lechner, A. M., Assfalg-Machleidt, I., Zahler, S., Stoeckelhuber, M., Machleidt, W., Jochum, M., & Nägler, D. K. (2006). RGD-dependent binding of procathepsin X to integrin alphavbeta3 mediates cell-adhesive properties. *Journal of Biological Chemistry*, *281*(51), 39588-39597. https://doi.org/10.1074/jbc.M513439200

Leitão, J. H. (2020). Microbial Virulence Factors. *Int J Mol Sci*, *21*(15). https://doi.org/10.3390/ijms21155320

Leitch, G. J., & He, Q. (2012). Cryptosporidiosis-an overview. *J Biomed Res*, *25*(1), 1-16. https://doi.org/10.1016/s1674-8301(11)60001-8

Lendner, M., & Daugschies, A. (2014). *Cryptosporidium* infections: molecular advances. *Parasitology*, *141*(11), 1511-1532. https://doi.org/10.1017/s0031182014000237

Leong, S. A., Barghout, V., Birnbaum, H. G., Thibeault, C. E., Ben-Hamadi, R., Frech, F., & Ofman, J. J. (2003). The Economic Consequences of Irritable Bowel Syndrome: A US Employer Perspective. *Archives of Internal Medicine*, *163*(8), 929-935. https://doi.org/10.1001/archinte.163.8.929

Lepczyńska, M., Białkowska, J., Dzika, E., Piskorz-Ogórek, K., & Korycińska, J. (2017). Blastocystis: how do specific diets and human gut microbiota affect its development and pathogenicity? *Eur J Clin Microbiol Infect Dis*, *36*(9), 1531-1540. https://doi.org/10.1007/s10096-017-2965-0

Letunic I, & Bork., P. *SMART: recent updates, new developments and status in 2020 Nucleic Acids Res 2020;* https://doi.org/10.1093/nar/gkaa937

Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, *49*(W1), W293-W296. https://doi.org/10.1093/nar/gkab301

Li, F.-J., Tsaousis, A. D., Purton, T., Chow, V. T. K., He, C. Y., & Tan, K. S. W. (2019). Successful Genetic Transfection of the Colonic Protistan Parasite *Blastocystis* for Reliable Expression of Ectopic Genes. *Scientific Reports*, *9*(1), 3159. https://doi.org/10.1038/s41598-019-39094-5

Libretexts. (2021). Protein domains, motifs, and folds in protein structure. *Biology LibreTexts.* . https://bio.libretexts.org/Bookshelves/Cell_and_Molecular_Biology/Book%3A_Basic_Cell_and_Molecular_Biology_(Bergtrom)/03%3A_Details_of_Protein_Structure/3.06%3A_Protein_Domains_Motifs_and_Folds_in_Protein_Structure

Lim, M. X., Png, C. W., Tay, C. Y. B., Teo, J. D. W., Jiao, H., Lehming, N., Tan, K. S. W., & Zhang, Y. (2014). Differential Regulation of Proinflammatory Cytokine Expression by Mitogen-Activated Protein Kinases in Macrophages in Response to Intestinal Parasite Infection. *Infection and immunity*, *82*(11), 4789-4801. https://doi.org/10.1128/iai.02279-14

Liu, J. (2019). *Characterization of secreted Giardia intestinalis cysteine proteases* (Publication Number 1763) [Doctoral thesis, comprehensive summary, Acta Universitatis Upsaliensis]. DiVA. Uppsala. http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-372997

Liu, J., Ma'ayeh, S., Peirasmaki, D., Lundström-Stadelmann, B., Hellman, L., & Svärd, S. G. . (2018). Secreted *Giardia intestinalis* cysteine proteases disrupt intestinal epithelial cell junctional complexes and degrade chemokines. . *Virulence*, *9(1)*, 879–894. https://doi.org/10.1080/21505594.2018.1451284

López-Otín, C., & Bond, J. S. (2008). Proteases: multifunctional enzymes in life and disease. *J Biol Chem*, *283*(45), 30433-30437. https://doi.org/10.1074/jbc.R800035200

Løge, I. (2012). *Giardia lamblia assosiert med kronisk fatigue og irritabel tarm*. https://nhi.no/for-helsepersonell/fra-vitenskapen/giardia-lamblia-assosiert-med-kronisk-fatigue-og-irritabel-tarm/

Mair, G. R., Niciu, M. J., Stewart, M. T., Brennan, G., Omar, H., Halton, D. W., Mains, R., Eipper, B. A., Maule, A. G., & Day, T. A. (2004). A functionally atypical amidating enzyme from the human parasite *Schistosoma mansoni*. *Faseb j*, *18*(1), 114-121. https://doi.org/10.1096/fj.03-0429com

Marquay Markiewicz, J., Syan, S., Hon, C. C., Weber, C., Faust, D., & Guillen, N. (2011). A proteomic and cellular analysis of uropods in the pathogen *Entamoeba histolytica*. *PLoS Negl Trop Dis*, *5*(4), e1002. https://doi.org/10.1371/journal.pntd.0001002

Martín-Escolano, R., Ng, G. C., Tan, K. S. W., Stensvold, C. R., Gentekaki, E., & Tsaousis, A. D. (2023). Resistance of *Blastocystis* to chlorine and hydrogen peroxide. *Parasitol Res*, *122*(1), 167-176. https://doi.org/10.1007/s00436-022-07713-2

McGinnis, S., & Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*, *32*(suppl_2), W20-W25. https://doi.org/10.1093/nar/gkh435

Meléndez-López, S. G., Herdman, S., Hirata, K., Choi, M.-H., Choe, Y., Craik, C., Caffrey, C. R., Hansell, E., Chávez-Munguía, B., Chen, Y. T., Roush, W. R., McKerrow, J., Eckmann, L., Guo, J., Stanley, S. L., & Reed, S. L. (2007). Use of Recombinant *Entamoeba histolytica* Cysteine Proteinase 1 To Identify a Potent Inhibitor of Amebic Invasion in a Human Colonic Model. *Eukaryotic Cell*, *6*(7), 1130-1136. https://doi.org/10.1128/ec.00094-07

Melo, G. B. d., Bosqui, L. R., Costa, I. N. d., Paula, F. M. d., & Gryschek, R. C. B. (2021). Current status of research regarding *Blastocystis* sp., an enigmatic protist, in Brazil [10.6061/clinics/2021/e2489]. *Clinics*, *76*. https://doi.org/10.6061/clinics/2021/e2489

Mirza, H., & Tan, K. S. (2009). *Blastocystis* exhibits inter- and intra-subtype variation in cysteine protease activity. *Parasitol Res*, *104*(2), 355-361. https://doi.org/10.1007/s00436-008-1203-1

Mirza, H., Teo, J. D. W., Upcroft, J., & Tan, K. S. W. (2011). A Rapid, High-Throughput Viability Assay for *Blastocystis* spp. Reveals Metronidazole Resistance and Extensive Subtype-Dependent Variations in Drug Susceptibilities. *Antimicrobial Agents and Chemotherapy*, *55*(2), 637-648. https://doi.org/10.1128/aac.00900-10

Mort, J. S., & Buttle, D. J. (1997). Cathepsin B. *The International Journal of Biochemistry & Cell Biology*, *29*(5), 715-720. https://doi.org/10.1016/S1357-2725(96)00152-5

Mótyán, J. A., Tóth, F., & Tőzsér, J. (2013). Research Applications of Proteolytic Enzymes in Molecular Biology. *Biomolecules*, *3*(4), 923-942. https://www.mdpi.com/2218-273X/3/4/923

Na, B. K., Kang, J. M., Cheun, H. I., Cho, S. H., Moon, S. U., Kim, T. S., & Sohn, W. M. (2009). Cryptopain-1, a cysteine protease of *Cryptosporidium parvum*, does not require the pro-domain for folding. *Parasitology*, *136*(2), 149-157. https://doi.org/10.1017/s0031182008005350

Nash, A., Dalziel, R., & Fitzgerald, J. (2015). Mechanisms of Cell and Tissue Damage. *Mims' Pathogenesis of Infectious Disease*, *171-231*. https://doi.org/10.1016/B978-0-12-397188-3.00008-1

Naz, S., & Fatima, A. (2013). Amplification of GC-rich DNA for high-throughput family-based genetic studies. *Mol Biotechnol*, *53*(3), 345-350. https://doi.org/10.1007/s12033-012-9559-y

Noël, C., Dufernez, F., Gerbod, D., Edgcomb, V. P., Delgado-Viscogliosi, P., Ho, L. C., Singh, M., Wintjens, R., Sogin, M. L., Capron, M., Pierce, R., Zenner, L., & Viscogliosi, E. (2005). Molecular phylogenies of *Blastocystis* isolates from different hosts: implications for genetic diversity, identification of species, and zoonosis. *J Clin Microbiol*, *43*(1), 348-355. https://doi.org/10.1128/jcm.43.1.348-355.2005

Nourrisson, C., Wawrzyniak, I., Cian, A., Livrelli, V., Viscogliosi, E., Delbac, F., & Poirier, P. (2016). On *Blastocystis* secreted cysteine proteases: a legumain-activated cathepsin B increases paracellular permeability of intestinal Caco-2 cell monolayers. *Parasitology*, *143*(13), 1713-1722. https://doi.org/10.1017/S0031182016001396

Novinec, M., & Lenarčič, B. (2013). Papain-like peptidases: structure, function, and evolution. *BioMolecular Concepts*, *4*(3), 287-308. https://doi.org/10.1515/bmc-2012-0054

Ochieng, J., & Chaudhuri, G. (2010). Cystatin superfamily. *J Health Care Poor Underserved*, *21*(1 Suppl), 51-70. https://doi.org/10.1353/hpu.0.0257

Owji, H., Nezafat, N., Negahdaripour, M., Hajiebrahimi, A., & Ghasemi, Y. (2018). A comprehensive review of signal peptides: Structure, roles, and applications. *European Journal of Cell Biology*, *97*(6), 422-441. https://doi.org/10.1016/j.ejcb.2018.06.003

Padilla-Vaca, F., & Anaya-Velazquez, F. (2010). Insights Into *Entamoeba histolytica* Virulence Modulation. *Infectious Disorders - Drug Targets*, *10*(4), 242-250. https://doi.org/10.2174/187152610791591638

Pandey, K., Barkan, D., Sali, A., & Rosenthal, P. (2009). Regulatory Elements within the Prodomain of Falcipain-2, a Cysteine Protease of the Malaria Parasite *Plasmodium falciparum*. *PloS one*, *4*, e5694. https://doi.org/10.1371/journal.pone.0005694

Parija, S. C., & Jeremiah, S. (2013). *Blastocystis*: Taxonomy, biology and virulence. *Trop Parasitol*, *3*(1), 17-25. https://pubmed.ncbi.nlm.nih.gov/23961437/

Peirasmaki, D., Ma'ayeh, S. Y., Xu, F., Ferella, M., Campos, S., Liu, J., & Svärd, S. G. (2020). High Cysteine Membrane Proteins (HCMPs) Are Up-Regulated During *Giardia*-Host Cell Interactions [Original Research]. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.00913

Pertuz Belloso, S., Ostoa Saloma, P., Benitez, I., Soldevila, G., Olivos, A., & García-Zepeda, E. (2004). *Entamoeba histolytica* cysteine protease 2 (EhCP2) modulates leucocyte migration by proteolytic cleavage of chemokines. *Parasite Immunology*, *26*(5), 237-241. https://doi.org/10.1111/j.0141-9838.2004.00706.x

Petersen, T. N., Brunak, S., von Heijne, G., & Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, *8*(10), 785-786. https://doi.org/10.1038/nmeth.1701

Petri, W. A. (2005). Treatment of Giardiasis. *Curr Treat Options Gastroenterol*, *8*(1), 13-17. https://doi.org/10.1007/s11938-005-0047-3

Popot, J. L. (1993). Integral membrane protein structure: transmembrane α-helices as autonomous folding domains: Current opinion in structural biology. *Curr Opin Struct Biol*, *3*(4), 532-540. https://doi.org/10.1016/0959-440x(93)90079-z

Purves, W. K., Sadava, D., Orians, G.H., and Heller, H.C. (2003). *Chemical signals in proteins direct them to their cellular destinations* (7th ed.). Sunderland, MA: Sinauer Associates.

Puthia, M. K., Lu, J., & Tan, K. S. W. (2008). *Blastocystis ratti* Contains Cysteine Proteases That Mediate Interleukin-8 Response from Human Intestinal Epithelial Cells in an NF-kB-Dependent Manner. *Eukaryotic Cell*, *7*(3), 435-443. https://doi.org/doi:10.1128/ec.00371-07

Puthia, M. K., Vaithilingam, A., Lu, J., & Tan, K. S. W. (2005). Degradation of human secretory immunoglobulin A by *Blastocystis*. *Parasitology Research*, *97*(5), 386-389. https://doi.org/10.1007/s00436-005-1461-0

QIAGEN®. (2021). *Plasmid Purification Handbook*. https://www.qiagen.com/us/resources/resourcedetail?id=22df6325-9579-4aa0-819c-788f73d81a09&lang=en

Quan, S., Schneider, I., Pan, J., Von Hacht, A., & Bardwell, J. C. A. (2007). The CXXC motif is more than a redox rheostat. *J Biol Chem*, *282*(39), 28823-28833. https://doi.org/10.1074/jbc.M705291200

Que, X., Brinen, L. S., Perkins, P., Herdman, S., Hirata, K., Torian, B. E., Rubin, H., McKerrow, J. H., & Reed, S. L. (2002). Cysteine proteinases from distinct cellular compartments are recruited to phagocytic vesicles by *Entamoeba histolytica*. *Molecular and Biochemical Parasitology*, *119*(1), 23-32. https://doi.org/10.1016/S0166-6851(01)00387-5

Que, X., Kim, S. H., Sajid, M., Eckmann, L., Dinarello, C. A., McKerrow, J. H., & Reed, S. L. (2003). A surface amebic cysteine proteinase inactivates interleukin-18. *Infect Immun*, *71*(3), 1274-1280. https://doi.org/10.1128/iai.71.3.1274-1280.2003

Que, X., & Reed, S. L. (2000). Cysteine Proteinases and the Pathogenesis of Amebiasis. *Clinical microbiology reviews*, *13*(2), 196-206. https://doi.org/10.1128/cmr.13.2.196

Rajah Salim, H., Suresh Kumar, G., Vellayan, S., Mak, J. W., Khairul Anuar, A., Init, I., Vennila, G. D., Saminathan, R., & Ramakrishnan, K. (1999). *Blastocystis* in animal handlers. *Parasitol Res*, *85*(12), 1032-1033. https://doi.org/10.1007/s004360050677

Ramasarma, T. (1996). Transmembrane domains participate in functions of integral membrane proteins. *Indian J Biochem Biophys*, *33*(1), 20-29. https://pubmed.ncbi.nlm.nih.gov/8744829/

Ranasinghe, S. L., & McManus, D. P. (2017). Protease Inhibitors of Parasitic Flukes: Emerging Roles in Parasite Survival and Immune Defence. *Trends in Parasitology*, *33*(5), 400-413. https://doi.org/10.1016/j.pt.2016.12.013

Rawlings, N. D., Barrett, A.J., Thomas, P.D., Huang, X., Bateman, A. & Finn, R.D. . (2018). The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. https://doi.org/10.1093/nar/gkx1134

Razzaq, A., Shamsi, S., Ali, A., Ali, Q., Sajjad, M., Malik, A., & Ashraf, M. (2019). Microbial Proteases Applications. *Front Bioeng Biotechnol*, *7*, 110. https://doi.org/10.3389/fbioe.2019.00110

Redzynia, I., Ljunggren, A., Abrahamson, M., Mort, J. S., Krupa, J. C., Jaskolski, M., & Bujacz, G. (2008). Displacement of the Occluding Loop by the Parasite Protein, Chagasin, Results in Efficient Inhibition of Human Cathepsin B *. *Journal of Biological Chemistry*, *283*(33), 22815-22825. https://doi.org/10.1074/jbc.M802064200

Renko, M., Požgan, U., Majera, D., & Turk, D. (2010). Stefin A displaces the occluding loop of cathepsin B only by as much as required to bind to the active site cleft. *The FEBS Journal*, *277*(20), 4338-4345. https://doi.org/10.1111/j.1742-4658.2010.07824.x

Robertson, S. (2019). Giardiasis Intestinal Infection. *News-Medical*. https://www.news-medical.net/health/Giardiasis-Intestinal-Infection.aspx

Robinson, P. K. (2015). Enzymes: principles and biotechnological applications. *Essays Biochem*, *59*, 1-41. https://doi.org/10.1042/bse0590001

Rossle, N. F., & Latif, B. (2013). Cryptosporidiosis as threatening health problem: A review. *Asian Pac J Trop Biomed*, *3*(11), 916-924. https://doi.org/10.1016/s2221-1691(13)60179-3

Rumsey, P., & Waseem, M. (2023). *Giardia Lamblia* Enteritis. In *StatPearls*. StatPearls Publishing.

Ruoslahti, E. (1996). RGD and other recognition sequences for integrins. *Annu Rev Cell Dev Biol*, *12*, 697-715. https://doi.org/10.1146/annurev.cellbio.12.1.697

Sanger Sequencing Steps and Method. (n.d.). https://www.sigmaaldrich.com/NO/en/technical-documents/protocol/genomics/sequencing/sanger-sequencing

Sanvictores, T., & Farci, F. (2023). Biochemistry, Primary Protein Structure. In *StatPearls*. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK564343/

Scanlan, P. D., Stensvold, C. R., & Cotter, P. D. (2015). Development and Application of a *Blastocystis* Subtype-Specific PCR Assay Reveals that Mixed-Subtype Infections Are Common in a Healthy Human Population. *Appl Environ Microbiol*, *81*(12), 4071-4076. https://doi.org/10.1128/aem.00520-15

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

Schaeffer, R. D., Kinch, L. N., Liao, Y., & Grishin, N. V. (2016). Classification of proteins with shared motifs and internal repeats in the ECOD database. *Protein Sci*, *25*(7), 1188-1203. https://doi.org/10.1002/pro.2893

Schaudien, D., Baumgärtner, W., & Herden, C. (2007). High preservation of DNA standards diluted in 50% glycerol. *Diagn Mol Pathol*, *16*(3), 153-157. https://doi.org/10.1097/PDM.0b013e31803c558a

Shahmiri, M., & Mechler, A. (2020). The role of C-terminal amidation in the mechanism of action of the antimicrobial peptide aurein 1.2. *The EuroBiotech Journal*, *4*(1), 25-31. https://doi.org/10.2478/ebtj-2020-0004

Sharma, A. K., Dhasmana, N., Dubey, N., Kumar, N., Gangwal, A., Gupta, M., & Singh, Y. (2017). Bacterial Virulence Factors: Secreted for Survival. *Indian Journal of Microbiology*, *57*(1), 1-10. https://doi.org/10.1007/s12088-016-0625-1

Sienzel, D. J., Boreham, P. F. L., & McDougall, R. (1991). Ultrastructure of *Blastocystis hominis* in human stool samples. *International Journal for Parasitology*, *21*(7), 807-812. https://doi.org/10.1016/0020-7519(91)90149-2

Siqueira-Neto, J. L., Debnath, A., McCall, L.-I., Bernatchez, J. A., Ndao, M., Reed, S. L., & Rosenthal, P. J. (2018). Cysteine proteases in protozoan parasites. *PLOS Neglected Tropical Diseases*, *12*(8), e0006512. https://doi.org/10.1371/journal.pntd.0006512

Smith, H. V., Nichols, R. A., & Grimason, A. M. (2005). *Cryptosporidium* excystation and invasion: getting to the guts of the matter. *Trends Parasitol*, *21*(3), 133-142. https://doi.org/10.1016/j.pt.2005.01.007

Soghra, V., Zahra, R., Iman, P., Asad, M., & Jahangir, A. (2022). The Prevalence of *Blastocystis* sp. and Its Relationship with Gastrointestinal Disorders and Risk factors. *Iranian Journal of Parasitology*, *17*(1). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9375720/

Star, B., & Spencer, H. G. (2013). Effects of genetic drift and gene flow on the selective maintenance of genetic variation. *Genetics*, *194*(1), 235-244. https://doi.org/10.1534/genetics.113.149781

Stensvold, C. R., Berg, R. P. K. D., Maloney, J. G., Molokin, A., & Santin, M. (2023). Molecular characterization of *Blastocystis* and *Entamoeba* of muskoxen and sheep in Greenland. *International Journal for Parasitology*, *53*(11), 673-685. https://doi.org/10.1016/j.ijpara.2023.05.005

Stenzel, D. J., & Boreham, P. F. (1996). *Blastocystis hominis* revisited. *Clinical microbiology reviews*, *9*(4), 563-584. https://doi.org/10.1128/cmr.9.4.563

Tack, J., Stanghellini, V., Mearin, F., Yiannakou, Y., Layer, P., Coffin, B., Simren, M., Mackinnon, J., Wiseman, G., Marciniak, A., Bouchoucha, Ducrotte, Macaigne, Mion, Schneider, Sidanier, L., Tardy, Andresen, Z., Bischoff, . . . on behalf of the, I.-C. S. g. (2019). Economic burden of moderate to severe irritable bowel syndrome with constipation in six European countries. *BMC Gastroenterology*, *19*(1), 69. https://doi.org/10.1186/s12876-019-0985-1

Tamura K, S. G., and Kumar S. (2021). *Molecular Biology and Evolution*. https://academic.oup.com/mbe/article/38/7/3022/6248099?login=true

Tan, K. S., Mirza, H., Teo, J. D., Wu, B., & Macary, P. A. (2010). Current Views on the Clinical Relevance of *Blastocystis* spp. *Curr Infect Dis Rep*, *12*(1), 28-35. https://doi.org/10.1007/s11908-009-0073-8

Tan, K. S. W. (2008). New Insights on Classification, Identification, and Clinical Relevance of *Blastocystis* spp. *Clinical microbiology reviews*, *21*(4), 639-665. https://doi.org/10.1128/cmr.00022-08

ThermoFisher. (2019). *PCR Primer Design Tips*. https://www.thermofisher.com/blog/behindthebench/pcr-primer-design-tips/

Thibeaux, R., Avé, P., Bernier, M., Morcelet, M., Frileux, P., Guillén, N., & Labruyère, E. (2014). The parasite *Entamoeba histolytica* exploits the activities of human matrix metalloproteinases to invade colonic tissue. *Nat Commun*, *5*, 5142. https://doi.org/10.1038/ncomms6142

Tyndall, J. D. A., Nall, T., & Fairlie, D. P. (2005). Proteases Universally Recognize Beta Strands In Their Active Sites. *Chemical Reviews*, *105*(3), 973-1000. https://doi.org/10.1021/cr040669e

van der Velden, V. H., & Hulsmann, A. R. (1999). Peptidases: structure, function and modulation of peptide-mediated effects in the human lung. *Clin Exp Allergy*, *29*(4), 445-456. https://doi.org/10.1046/j.1365-2222.1999.00462.x

Venczel, L. V., Arrowood, M., Hurd, M., & Sobsey, M. D. (1997). Inactivation of *Cryptosporidium parvum* oocysts and *Clostridium perfringens* spores by a mixed-oxidant disinfectant and by free chlorine. *Appl Environ Microbiol*, *63*(4), 1598-1601. https://doi.org/10.1128/aem.63.4.1598-1601.1997

Vivancos, V., González-Alvarez, I., Bermejo, M., & Gonzalez-Alvarez, M. (2018). Giardiasis: Characteristics, Pathogenesis and New Insights About Treatment. *Curr Top Med Chem*, *18*(15), 1287-1303. https://doi.org/10.2174/1568026618666181002095314

Vorster, B. J., Schlüter, U., Du Plessis, M., Van Wyk, S., Makgopa, M. E., Ncube, I., Quain, M. D., Kunert, K., & Foyer, C. H. (2013). The Cysteine Protease–Cysteine Protease Inhibitor System Explored in Soybean Nodule Development. *Agronomy*, *3*(3), 550-570. https://www.mdpi.com/2073-4395/3/3/550

Wang, Y., Zhang, H., Zhong, H., & Xue, Z. (2021). Protein domain identification methods and online resources. *Comput Struct Biotechnol J*, *19*, 1145-1153. https://doi.org/10.1016/j.csbj.2021.01.041

Wawrzyniak, I., Poirier, P., Viscogliosi, E., Dionigia, M., Texier, C., Delbac, F., & Alaoui, H. E. (2013). *Blastocystis*, an unrecognized parasite: an overview of pathogenesis and diagnosis. *Ther Adv Infect Dis*, *1*(5), 167-178. https://doi.org/10.1177/2049936113504754

Wawrzyniak, I., Texier, C., Poirier, P., Viscogliosi, E., Tan, K. S. W., Delbac, F., & El Alaoui, H. (2012). Characterization of two cysteine proteases secreted by *Blastocystis* ST7, a human intestinal parasite. *Parasitology International*, *61*(3), 437-442. https://doi.org/10.1016/j.parint.2012.02.007

Wayne Albers, R. R. W. (2012). Chapter 2 - Cell Membrane Structures and Functions. In S. T. Brady, G. J. Siegel, R. W. Albers, & D. L. Price (Eds.), *Basic Neurochemistry (Eighth Edition)* (pp. 26-39). Academic Press. https://doi.org/10.1016/B978-0-12-374947-5.00002-X

Wlodawer, A. (2002). Structure-based design of AIDS drugs and the development of resistance. *Vox Sang*, *83 Suppl 1*, 23-26. https://doi.org/10.1111/j.1423-0410.2002.tb05261.x

Wu, Z., Mirza, H., & Tan, K. S. W. (2014). Intra-Subtype Variation in Enteroadhesion Accounts for Differences in Epithelial Barrier Disruption and Is Associated with Metronidazole Resistance in *Blastocystis* Subtype-7. *PLOS Neglected Tropical Diseases*, *8*(5), e2885. https://doi.org/10.1371/journal.pntd.0002885

Xiong, J. (2006). Protein Motifs and Domain Prediction. In *Essential Bioinformatics* (pp. 85-94). Cambridge University Press. https://doi.org/10.1017/CBO9780511806087.008

Yamada, Y., Onda, T., Wada, Y., Hamada, K., Kikkawa, Y., & Nomizu, M. (2023). Structure–Activity Relationships of RGD-Containing Peptides in Integrin αvβ5-Mediated Cell Adhesion. *ACS Omega*, *8*(5), 4687-4693. https://doi.org/10.1021/acsomega.2c06540

Yang, N., Matthew, M. A., & Yao, C. (2023). Roles of Cysteine Proteases in Biology and Pathogenesis of Parasites. *Microorganisms*, *11*(6), 1397. https://www.mdpi.com/2076-2607/11/6/1397

Yason, J. A., Ajjampur, S. S. R., & Tan, K. S. W. (2016). *Blastocystis* Isolate B Exhibits Multiple Modes of Resistance against Antimicrobial Peptide LL-37. *Infection and immunity*, *84*(8), 2220-2232. https://doi.org/10.1128/iai.00339-16

Yason, J. A., Koh, K., & Tan, K. S. W. (2018). Viability Screen of LOPAC(1280) Reveals Phosphorylation Inhibitor Auranofin as a Potent Inhibitor of *Blastocystis* Subtype 1, 4, and 7 Isolates. *Antimicrob Agents Chemother*, *62*(8). https://doi.org/10.1128/aac.00208-18

Yoshikawa, H., Morimoto, K., Wu, Z., Singh, M., & Hashimoto, T. (2004). Problems in speciation in the genus *Blastocystis*. *Trends in Parasitology*, *20*(6), 251-255. https://doi.org/10.1016/j.pt.2004.03.010

Zachary, J. F. (2017). Mechanisms of Microbial Infections. https://doi.org/10.1016/B978-0-323-35775-3.00004-7

Zajaczkowski, P., Lee, R., Fletcher-Lartey, S. M., Alexander, K., Mahimbo, A., Stark, D., & Ellis, J. T. (2021). The controversies surrounding *Giardia intestinalis* assemblages A and B. *Curr Res Parasitol Vector Borne Dis*, *1*, 100055. https://doi.org/10.1016/j.crpvbd.2021.100055

Zhang, X., Qiao, J., Wu, X., Da, R., Zhao, L., & Wei, Z. (2012). *In vitro* culture of *Blastocystis hominis* in three liquid media and its usefulness in the diagnosis of blastocystosis. *International Journal of Infectious Diseases*, *16*(1), e23-e28. https://doi.org/10.1016/j.ijid.2011.09.012

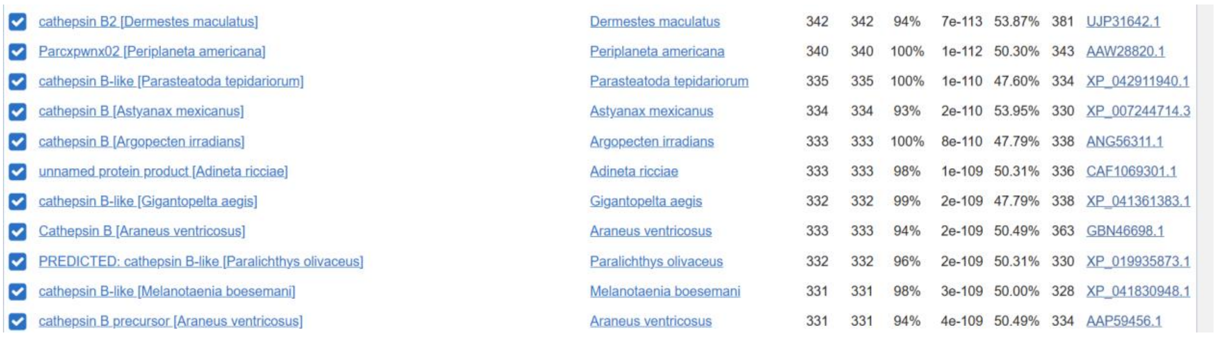
Zhang, Z., Wang, L., Seydel, K. B., Li, E., Ankri, S., Mirelman, D., & Stanley, S. L., Jr. (2000). *Entamoeba histolytica* cysteine proteinases with interleukin-1 beta converting enzyme (ICE) activity cause intestinal inflammation and tissue damage in amoebiasis. *Mol Microbiol*, *37*(3), 542-548. https://doi.org/10.1046/j.1365-2958.2000.02037.x

Zierdt, C. H. (1991). *Blastocystis hominis*--past and future. *Clinical microbiology reviews*, *4*(1), 61-79. https://doi.org/10.1128/cmr.4.1.61

ZymoResearch. (2022). *ZymoPURE™ II Plasmid Maxiprep Kit*. https://files.zymoresearch.com/protocols/_d4202_d4203_zymopure_ii_plasmid_maxiprep.pdf

Appendix 1

Cathepsin B



☑	cathepsin B2 [Dermestes maculatus]	Dermestes maculatus	342	342	94%	7e-113	53.87%	381	UJP31642.1
☑	Parcxpwnx02 [Periplaneta americana]	Periplaneta americana	340	340	100%	1e-112	50.30%	343	AAW28820.1
☑	cathepsin B-like [Parasteatoda tepidariorum]	Parasteatoda tepidariorum	335	335	100%	1e-110	47.60%	334	XP_042911940.1
☑	cathepsin B [Astyanax mexicanus]	Astyanax mexicanus	334	334	93%	2e-110	53.95%	330	XP_007244714.3
☑	cathepsin B [Argopecten irradians]	Argopecten irradians	333	333	100%	8e-110	47.79%	338	ANG56311.1
☑	unnamed protein product [Adineta ricciae]	Adineta ricciae	333	333	98%	1e-109	50.31%	336	CAF1069301.1
☑	cathepsin B-like [Gigantopelta aegis]	Gigantopelta aegis	332	332	99%	2e-109	47.79%	338	XP_041361383.1
☑	Cathepsin B [Araneus ventricosus]	Araneus ventricosus	333	333	94%	2e-109	50.49%	363	GBN46698.1
☑	PREDICTED: cathepsin B-like [Paralichthys olivaceus]	Paralichthys olivaceus	332	332	96%	2e-109	50.31%	330	XP_019935873.1
☑	cathepsin B-like [Melanotaenia boesemani]	Melanotaenia boesemani	331	331	98%	3e-109	50.00%	328	XP_041830948.1
☑	cathepsin B precursor [Araneus ventricosus]	Araneus ventricosus	331	331	94%	4e-109	50.49%	334	AAP59456.1

Figure 27 Search results of uncharacterized protein from *Blastocystis* ST7 in BLAST, by the exclusion of *Blastocystis* and *Giardia* genome. Results show a majority of cathepsin B.

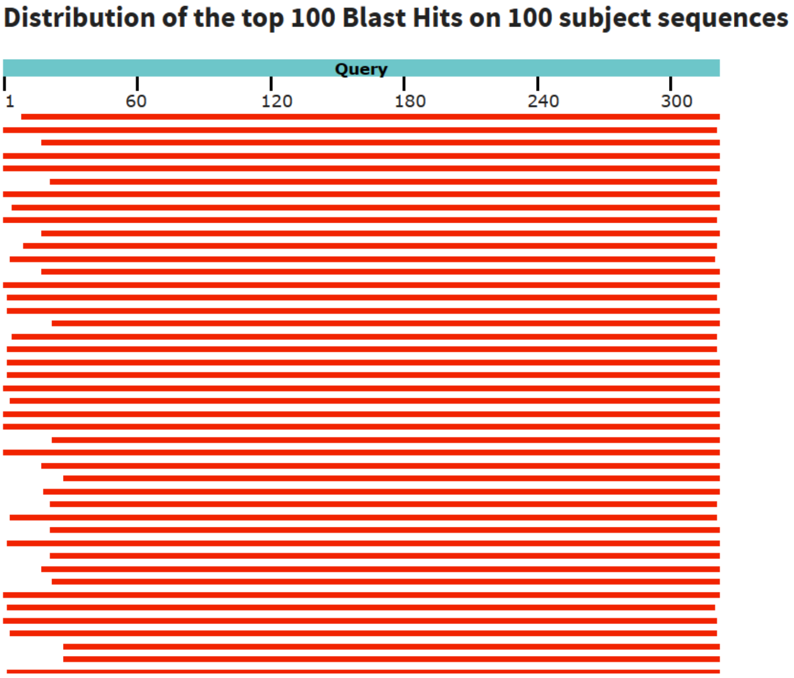
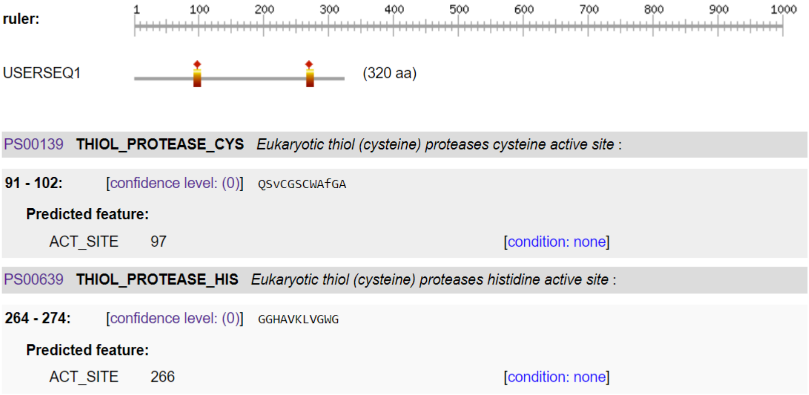


Figure 28 Distribution of search results from Figure 27 in a graphic summary. Red bars, denoting strong alignment scores, indicative of favorable sequence matches.



PS00139 **THIOL_PROTEASE_CYS** *Eukaryotic thiol (cysteine) proteases cysteine active site :*

91 - 102: [confidence level: (0)] `QSvCGSCWAfGA`

Predicted feature:

ACT_SITE 97 [condition: none]

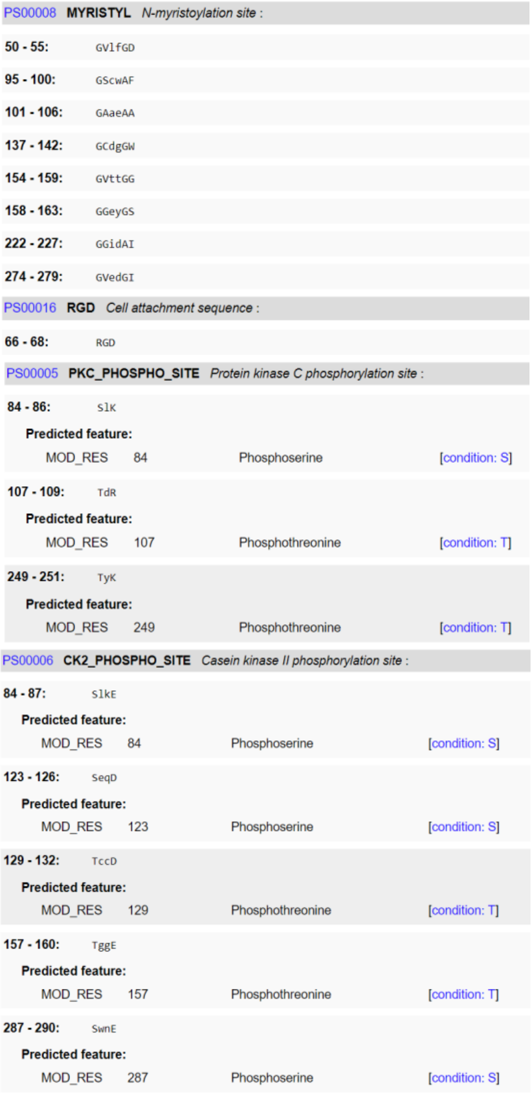
PS00639 **THIOL_PROTEASE_HIS** *Eukaryotic thiol (cysteine) proteases histidine active site :*

264 - 274: [confidence level: (0)] `GGHAVKLVGWG`

Predicted feature:

ACT_SITE 266 [condition: none]

Figure 29 Results of active sites found in GiP1 (cathepsin B) in *Blastocystis* ST7 performed in ScanProsite. The confidence level of 0 is described as a reliable cut-off in the database. This was performed for all proteins to determine the active sites and their positions, only showing results from BsP1 ST7.



PS00008 **MYRISTYL** *N-myristoylation site :*

50 - 55: `GVlfGD`

95 - 100: `GScwAF`

101 - 106: `GAaeAA`

137 - 142: `GCdgGW`

154 - 159: `GVttGG`

158 - 163: `GGeyGS`

222 - 227: `GGidAI`

274 - 279: `GVedGI`

PS00016 **RGD** *Cell attachment sequence :*

66 - 68: `RGD`

PS00005 **PKC_PHOSPHO_SITE** *Protein kinase C phosphorylation site :*

84 - 86: `SlK`

Predicted feature:

MOD_RES 84 Phosphoserine [condition: S]

107 - 109: `TdR`

Predicted feature:

MOD_RES 107 Phosphothreonine [condition: T]

249 - 251: `TyK`

Predicted feature:

MOD_RES 249 Phosphothreonine [condition: T]

PS00006 **CK2_PHOSPHO_SITE** *Casein kinase II phosphorylation site :*

84 - 87: `SlkE`

Predicted feature:

MOD_RES 84 Phosphoserine [condition: S]

123 - 126: `SeqD`

Predicted feature:

MOD_RES 123 Phosphoserine [condition: S]

129 - 132: `TccD`

Predicted feature:

MOD_RES 129 Phosphothreonine [condition: T]

157 - 160: `TggE`

Predicted feature:

MOD_RES 157 Phosphothreonine [condition: T]

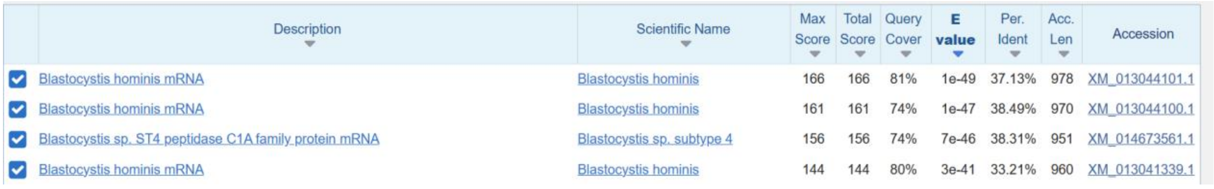
287 - 290: `SwnE`

Predicted feature:

MOD_RES 287 Phosphoserine [condition: S]

Figure 30 Results from ScanProsite of motifs found in GiP1 (cathepsin B) in *Blastocystis* ST7 including motifs that have a high probability of occurrence. Only some of the motifs are put into further analysis.

Cysteine proteinase 2



	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Blastocystis hominis mRNA	Blastocystis hominis	166	166	81%	1e-49	37.13%	978	XM_013044101.1
☑	Blastocystis hominis mRNA	Blastocystis hominis	161	161	74%	1e-47	38.49%	970	XM_013044100.1
☑	Blastocystis sp. ST4 peptidase C1A family protein mRNA	Blastocystis sp. subtype 4	156	156	74%	7e-46	38.31%	951	XM_014673561.1
☑	Blastocystis hominis mRNA	Blastocystis hominis	144	144	80%	3e-41	33.21%	960	XM_013041339.1

Figure 31 Results from a search of cysteine proteinase 2 (*E. histolytica*) in *Blastocystis* genome performed in translated nucleotide BLAST search (tBLASTn). They all include a query cover over 70% and the e-value is within the threshold.

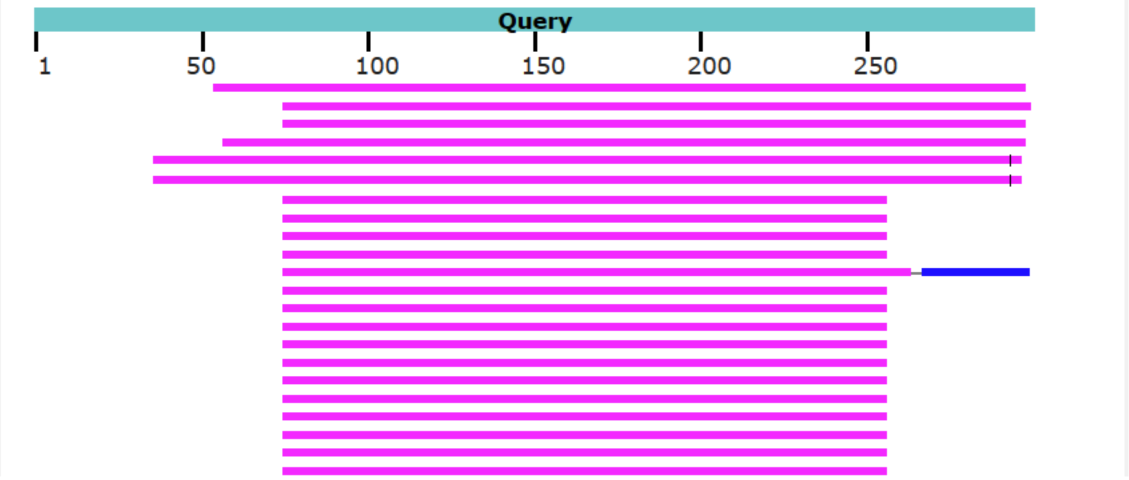
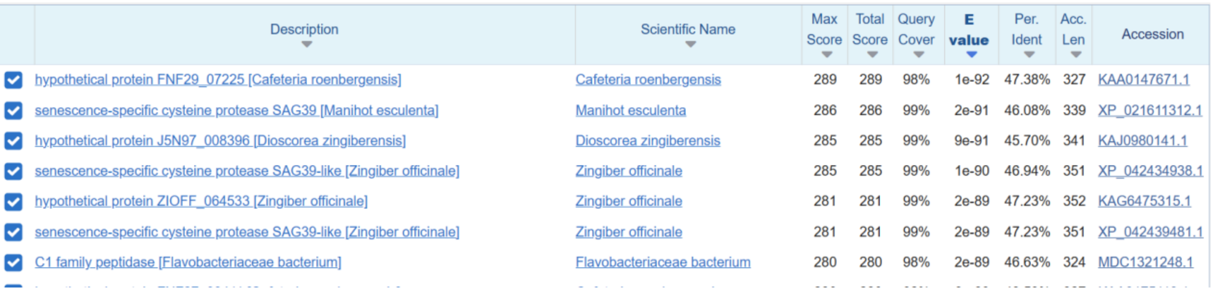


Figure 32 Graphic summary of results from Figure 31, of the search results of cysteine proteinase 2 from *E. histolytica* in *Blastocystis* genome in tBLASTn. Pink bars denoting solid matches.



	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	hypothetical protein FNF29_07225 [Cafeteria roenbergensis]	Cafeteria roenbergensis	289	289	98%	1e-92	47.38%	327	KAA0147671.1
☑	senescence-specific cysteine protease SAG39 [Manihot esculenta]	Manihot esculenta	286	286	99%	2e-91	46.08%	339	XP_021611312.1
☑	hypothetical protein J5N97_008396 [Dioscorea zingiberensis]	Dioscorea zingiberensis	285	285	99%	9e-91	45.70%	341	KAJ0980141.1
☑	senescence-specific cysteine protease SAG39-like [Zingiber officinale]	Zingiber officinale	285	285	99%	1e-90	46.94%	351	XP_042434938.1
☑	hypothetical protein ZIOFF_064533 [Zingiber officinale]	Zingiber officinale	281	281	99%	2e-89	47.23%	352	KAG6475315.1
☑	senescence-specific cysteine protease SAG39-like [Zingiber officinale]	Zingiber officinale	281	281	99%	2e-89	47.23%	351	XP_042439481.1
☑	C1 family peptidase [Flavobacteriaceae bacterium]	Flavobacteriaceae bacterium	280	280	98%	2e-89	46.63%	324	MDC1321248.1

Figure 33 Results of search in BLASTp of protein found in *Blastocystis* ST7, and excluding *E. histolytica* and *Blastocystis* from the search. Many results show cysteine protease.

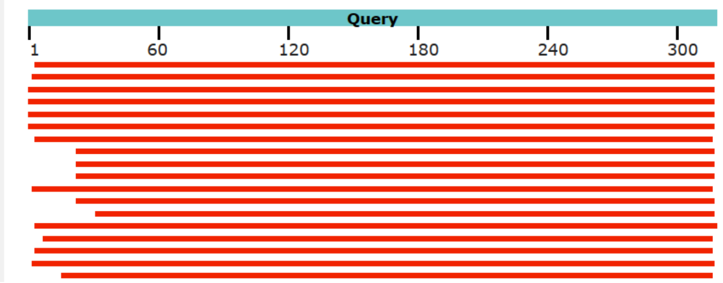
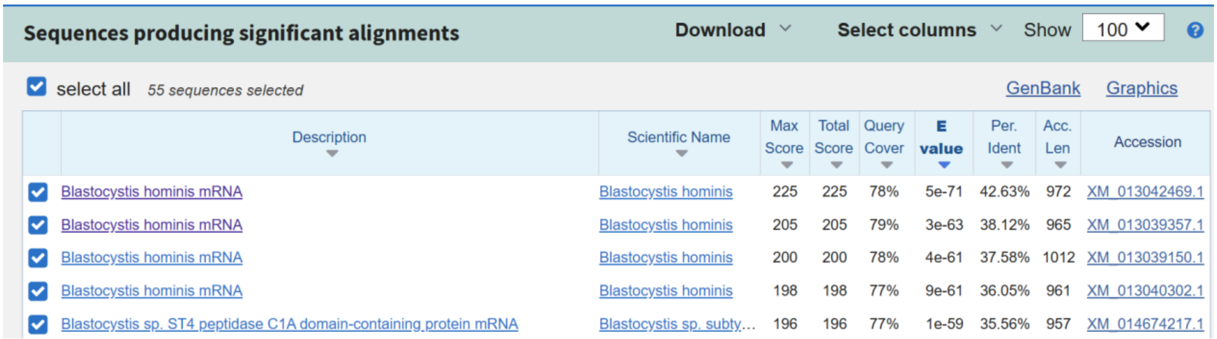


Figure 34 Graphical summary of reverse protein BLASTp results from Figure 33, of the search of cysteine proteinase 2 and excluding results from *E. histolytica* and *Blastocystis* spp.

Cryptopain-1



☑	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Blastocystis hominis mRNA	Blastocystis hominis	225	225	78%	5e-71	42.63%	972	XM_013042469.1
☑	Blastocystis hominis mRNA	Blastocystis hominis	205	205	79%	3e-63	38.12%	965	XM_013039357.1
☑	Blastocystis hominis mRNA	Blastocystis hominis	200	200	78%	4e-61	37.58%	1012	XM_013039150.1
☑	Blastocystis hominis mRNA	Blastocystis hominis	198	198	77%	9e-61	36.05%	961	XM_013040302.1
☑	Blastocystis sp. ST4 peptidase C1A domain-containing protein mRNA	Blastocystis sp. subty…	196	196	77%	1e-59	35.56%	957	XM_014674217.1

Figure 35 Results of the search for cryptopain-1 in *Blastocystis* genome performed in translated nucleotide BLAST search (tBLASTn). They all include a query cover over 70% and the e-value is within the threshold.

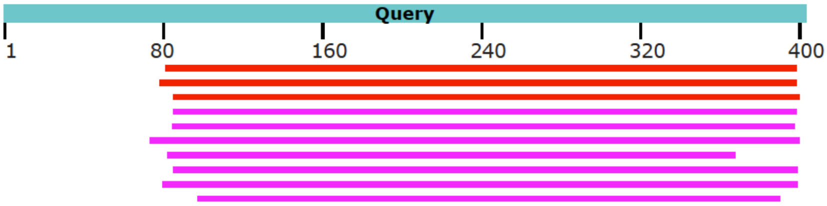
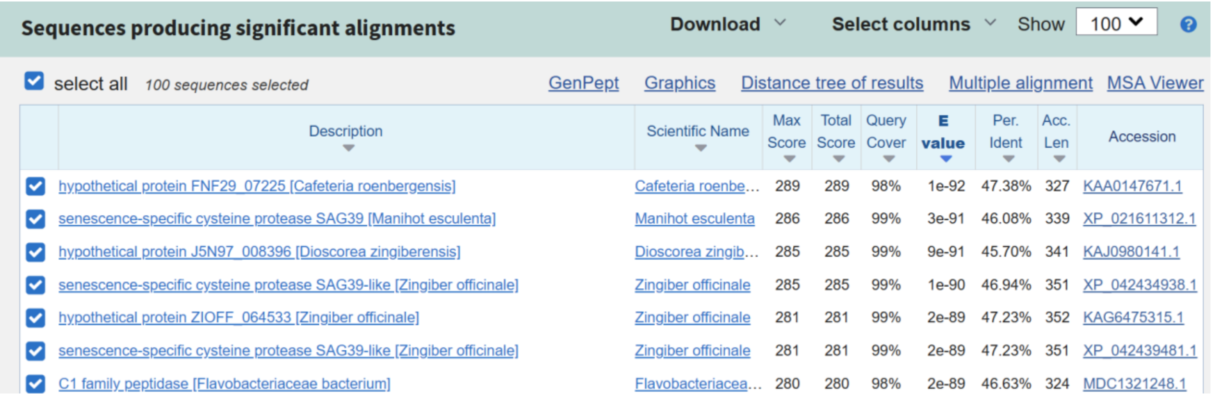


Figure 36 Graphic summary of results from Figure 35, of the search of cryptopain-1 from *C. pavrum* in genome *Blastocystis* in tBLASTn. Red and pink bars indicate some robust results and some solid matches.



☑	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	hypothetical protein FNF29_07225 [Cafeteria roenbergensis]	Cafeteria roenbe…	289	289	98%	1e-92	47.38%	327	KAA0147671.1
☑	senescence-specific cysteine protease SAG39 [Manihot esculenta]	Manihot esculenta	286	286	99%	3e-91	46.08%	339	XP_021611312.1
☑	hypothetical protein J5N97_008396 [Dioscorea zingiberensis]	Dioscorea zingib…	285	285	99%	9e-91	45.70%	341	KAJ0980141.1
☑	senescence-specific cysteine protease SAG39-like [Zingiber officinale]	Zingiber officinale	285	285	99%	1e-90	46.94%	351	XP_042434938.1
☑	hypothetical protein ZIOFF_064533 [Zingiber officinale]	Zingiber officinale	281	281	99%	2e-89	47.23%	352	KAG6475315.1
☑	senescence-specific cysteine protease SAG39-like [Zingiber officinale]	Zingiber officinale	281	281	99%	2e-89	47.23%	351	XP_042439481.1
☑	C1 family peptidase [Flavobacteriaceae bacterium]	Flavobacteriacea…	280	280	98%	2e-89	46.63%	324	MDC1321248.1

Figure 37 Results of protein search in reverse protein BLAST of *Blastocystis* ST7, excluding *Cryptosporidium* and *Blastocystis* from the search. Results show the protein to be cysteine protease.

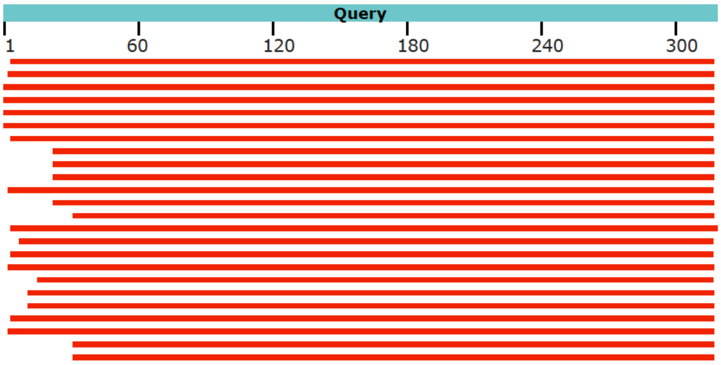
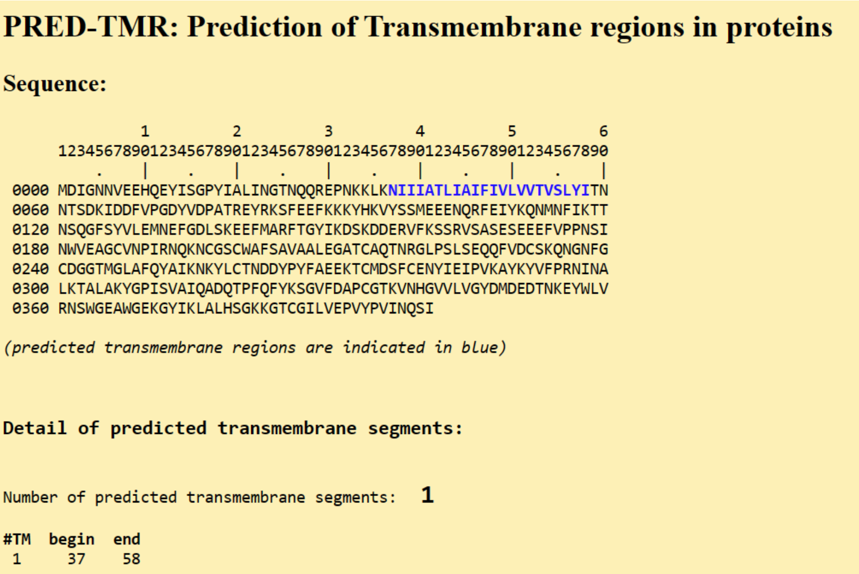


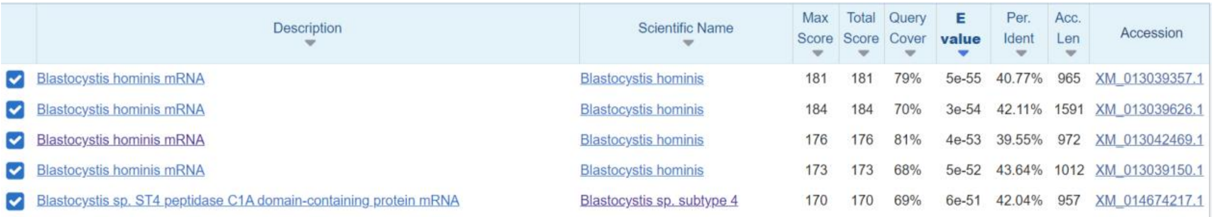
Figure 38 Graphic summary of results from Figure 37, of the reverse protein BLAST search of cryptopain-1. The search excludes results from *Cryptosporidium* and *Blastocystis* spp.



```
PRED-TMR: Prediction of Transmembrane regions in proteins

Sequence:

                 1         2         3         4         5         6
      123456789012345678901234567890123456789012345678901234567890
           .    |    .    |    .    |    .    |    .    |    .    |
 0000 MDIGNNVEEHQEYISGPYIALINGTNQQREPNKKLKNIIIATLIAIFIVLVVTVSLYITN
 0060 NTSDKIDDFVPGDYVDPATREYRKSFEEFKKKYHKVYSSMEEENQRFEIYKQNMNFIKTT
 0120 NSQGFSYVLEMNEFGDLSKEEFMARFTGYIKDSKDDERVFKSSRVSASESEEEFVPPNSI
 0180 NWVEAGCVNPIRNQKNCGSCWAFSAVAALEGATCAQTNRGLPSLSEQQFVDCSKQNGNFG
 0240 CDGGTMGLAFQYAIKNKYLCTNDDYPYFAEEKTCMDSFCENYIEIPVKAYKYVFPRNINA
 0300 LKTALAKYGPISVAIQADQTPFQFYKSGVFDAPCGTKVNHGVVLVGYDMDEDTNKEYWLV
 0360 RNSWGEAWGEKGYIKLALHSGKKGTCGILVEPVYPVINQSI

(predicted transmembrane regions are indicated in blue)


Detail of predicted transmembrane segments:


Number of predicted transmembrane segments:  1

#TM  begin  end
 1     37    58
```

Figure 39 Results of search in PRED-TMR for cryptopain-1 of *C. pavrum*. Results from *Blastocystis* are not included as there were no predictions in either ST7 or ST4.

Cysteine proteinase 5



	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Blastocystis hominis mRNA	Blastocystis hominis	181	181	79%	5e-55	40.77%	965	XM_013039357.1
☑	Blastocystis hominis mRNA	Blastocystis hominis	184	184	70%	3e-54	42.11%	1591	XM_013039626.1
☑	Blastocystis hominis mRNA	Blastocystis hominis	176	176	81%	4e-53	39.55%	972	XM_013042469.1
☑	Blastocystis hominis mRNA	Blastocystis hominis	173	173	68%	5e-52	43.64%	1012	XM_013039150.1
☑	Blastocystis sp. ST4 peptidase C1A domain-containing protein mRNA	Blastocystis sp. subtype 4	170	170	69%	6e-51	42.04%	957	XM_014674217.1

Figure 40 Results of translated nucleotide BLAST search (tBLASTn) of cysteine proteinase 5 from *E. histolytica* found in *Blastocystis* genome. *Blastocystis* ST7 has a query cover over 70%, while *Blastocystis* ST4 is under the threshold with 69% in query cover.

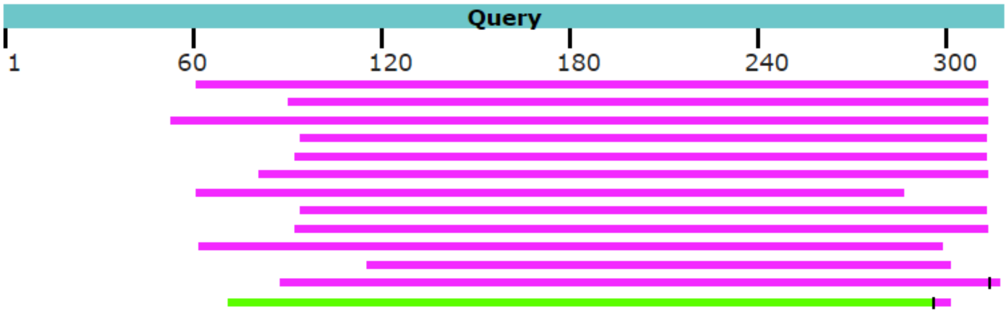
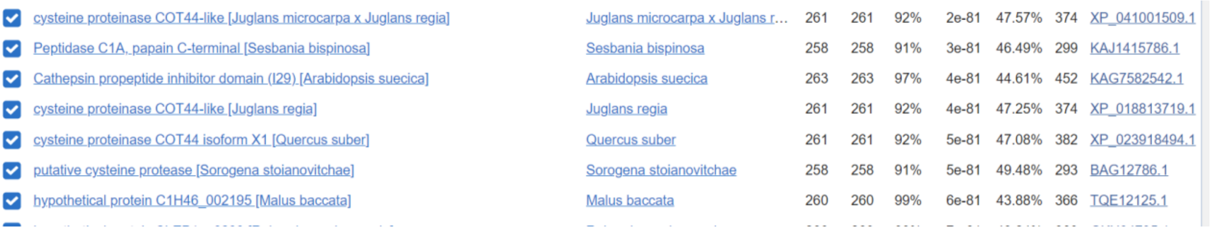


Figure 41 Graphic summary of results from Figure 40, of the tBLASTn search results of cysteine proteinase 5 in *Blastocystis* spp.



☑	cysteine proteinase COT44-like [Juglans microcarpa x Juglans regia]	Juglans microcarpa x Juglans r...	261	261	92%	2e-81	47.57%	374	XP_041001509.1
☑	Peptidase C1A, papain C-terminal [Sesbania bispinosa]	Sesbania bispinosa	258	258	91%	3e-81	46.49%	299	KAJ1415786.1
☑	Cathepsin propeptide inhibitor domain (I29) [Arabidopsis suecica]	Arabidopsis suecica	263	263	97%	4e-81	44.61%	452	KAG7582542.1
☑	cysteine proteinase COT44-like [Juglans regia]	Juglans regia	261	261	92%	4e-81	47.25%	374	XP_018813719.1
☑	cysteine proteinase COT44 isoform X1 [Quercus suber]	Quercus suber	261	261	92%	5e-81	47.08%	382	XP_023918494.1
☑	putative cysteine protease [Sorogena stoianovitchae]	Sorogena stoianovitchae	258	258	91%	5e-81	49.48%	293	BAG12786.1
☑	hypothetical protein C1H46_002195 [Malus baccata]	Malus baccata	260	260	99%	6e-81	43.88%	366	TQE12125.1

Figure 42 Results of search in reverse BLASTp of cysteine protease 5 found in *Blastocystis* ST7, excluding search in *Entamoeba* and *Blastocystis*. Some results show cysteine protease.

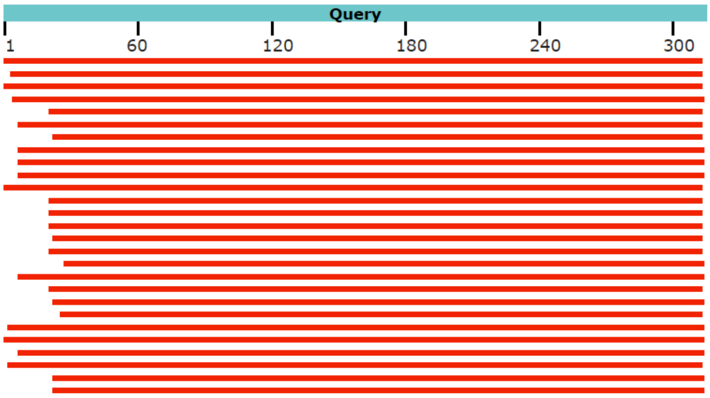


Figure 43 Graphic summary of results from Figure 42, of the reverse protein BLAST search of cysteine proteinase 5. The search excludes results from *Entamoeba* and *Blastocystis* spp.

Appendix 2 – Excel files

Cathepsin B

Table 9 Results of search performed in BLAST for cathepsin B in *Blastocystis* ST7 and ST4, as well as results of organisms used to construct a phylogenetic tree of the protein.

Organism	Gene ID	Protein ID	Protein	Length protein	Query cover	e-value
G.intestinalis	GL50803_16779	KAE8304515.1	Cathepsin B	298		
Blastocystis spp. ST7	XM_013044101.1	XP_012899555.1	uncharacterized protein	320	81 %	1e^-49
Blastocystis spp. ST4	XM_014673561.1	XP_014529047.1	peptidase C1A family protein	316	74 %	7e^-46

Phylogeny tree

Organism	Kingdom	Domain	Protein ID	Protein	Length protein
G.intestinalis	Protist	Eukaryota	KAE8304515.1	Cathepsin B	298
Spironucleus salmonicida	Protist	Eukaryota	KAH0570486.1	Cathepsin B	286
Kipferlia bialata	Protist	Eukaryota	AHW50664.1	peptidase C1A, partial	334
Aduncisulcus paluster	Protist	Eukaryota	GKT34524.1	peptidase C1A, partial	438
Naegleria fowleri	Protist	Eukaryota	AHW50664.1	cathepsin B-like protein	313
Naegleria lovaniensis	Protist	Eukaryota	XP_044544572.1	uncharacterized protein C9374_010021	316
Naegleria gruberi	Protist	Eukaryota	EFC44879.1	Cathepsin B	321
Acropora millepora	Animal	Eukaryota	XP_029203614.2	cathepsin B-like CP3	325
Blastocystis spp. ST7	Chromista	Eukaryota	XP_012899555.1	uncharacterized protein	320
Blastocystis spp. ST4	Chromista	Eukaryota	XP_014529047.1	peptidase C1A family protein	316
Blepharisma stoltei	Chromista	Eukaryota	CAG9313979.1	unnamed protein product	299
Chloropicon primus	Plant	Eukaryota	QDZ22777.1	cathepsin B cysteine protease	319
Batrachochytrium salamandrivorans	Fungi	Eukaryota	KAH9256964.1	hypothetical protein BASA81_004785	482

Cysteine proteinase 2

Table 10 Results of search performed in BLAST for cysteine proteinase 2 in *Blastocystis* spp., as well as results of organisms used to construct a phylogenetic tree of the protein.

Organism	Gene ID	Protein ID	Protein	Length protein	Query cover	e-value
E. histolytica	EHI_033710	XP_650642.1	cysteine proteinase 2	315		
Blastocystis spp. ST7	XM_013042469.1	XP_012897923.1	uncharacterized protein	316	97 %	1e^-60
Blastocystis spp. ST4	XM_014674217.1	XP_014529703.1	peptidase C1A domain-containing protein	318	98 %	2e^-48

Phylogeny tree					
Organism	**Kingdom**	**Domain**	**Protein ID**	**Protein**	**Length**
E. histolytica	Protist	Eukaryota	XP_650642.1	cysteine proteinase 2	315
Dictyostelium purpureum	Protist	Eukaryota	XP_003290609.1	hypothetical protein DICPUDRAFT_92519	333
Plasmodiophora brassicae	Protist	Eukaryota	CEO98669.1	hypothetical protein PBRA_006783	336
Blastocystis spp. ST7	Chromista	Eukaryota	XP_012897923.1	uncharacterized protein	316
Blastocystis spp. ST4	Chromista	Eukaryota	XP_014529703.1	peptidase C1A domain-containing protein	318
Thraustotheca clavata	Chromista	Eukaryota	AIG55389.1	secreted protein	520
Euphorbia peplus	Plant	Eukaryota	WCJ32692.1	Cysteine proteinases superfamily protein	341
Carica papaya	Plant	Eukaryota	XP_021887163.1	senescence-specific cysteine protease SAG39-like	339
Amborella trichopoda	Plant	Eukaryota	ERN19263.1	hypothetical protein AMTR_s00061p00215230	344
Clostridia bacterium	Bacteria	Bacteria	MBO5344712.1	MAG: hypothetical protein J6A51_02335	317
Streptococcus thermophilus	Bacteria	Bacteria	MCE2196779.1	hypothetical protein GQ599_09565	322
Batrachochytrium salamandrivorans	Fungi	Eukaryota	KAH9261048.1	hypothetical protein BASA81_000752	369
Myxine glutinosa	Animal	Eukaryota	AAF19631.1	cysteine proteinase precursor, partial	324
Oppia nitens	Animal	Eukaryota	XP_054166525.1	procathepsin L-like	335

Cryptopain-1

Table 11 Results of search performed in BLAST for cryptopain-1 in *Blastocystis* spp., as well as results of organisms used to construct phylogenetic tree of the protein.

Organism	**Sequence ID**	**Protein ID**	**Protein**	**Length protein**	**Query cover**	**e-value**
Cryptosporidium parvum Iowa II	XM_627814.1	XP_627814.1	cryptopain	401		
Blastocystis spp. ST7	XM_013042469.1	XP_012897923.1	uncharacterized	316	78 %	5e^-71
Blastocystis spp. ST4	XM_014674217.1	XP_014529703.1	peptidase C1A domain-containing protein	318	77 %	1e^-59
Phylogeny tree						
Organism	**Kingdom**	**Domain**	**Protein ID**	**Protein**	**Length**	
C. pavrum	Chromista	Eukaryota	ABA40395.1	cryptopain-1	401	
Blastocystis spp. *ST7*	Chromista	Eukaryota	XP_012897923.1	uncharacterized	316	
Blastocystis spp. *ST4*	Chromista	Eukaryota	XP_014529703.1	peptidase C1A domain-containing protein	318	
Besnoitia besnoiti	Chromista	Eukaryota	XP_029215951.1	cathepsin CPL	430	
Vitrella brassicaformis CCMP3155	Chromista	Eukaryota	CEM03624.1	unnamed protein product	385	
Cystoisospora suis	Protist	Eukaryota	PHJ22756.1	cathepsin cpl	471	
Blepharisma stoltei	Protist	Eukaryota	CAG9326208.1	unnamed protein product	350	
Cladocopium goreaui	Protist	Eukaryota	CAI3978150.1	unnamed protein product	462	
Toxoplasma gondii ME49	Protist	Eukaryota	XP_002371694.1	cathepsin CPL	422	
Neospora caninum Liverpool	Protist	Eukaryota	CEL64542.1	TPA: Cathepsin L, related	423	
Salpingoeca rosetta	Protist	Eukaryota	XP_004998235.1	cysteine proteinase	448	
Batrachochytrium salamandrivorans	Fungi	Eukaryota	KAH9256963.1	hypothetical protein BASA81_004784	484	
Lithospermum erythrorhizon	Plant	Eukaryota	KAG9153891.1	hypothetical protein Leryth_005995	359	
Chlorella sorokiniana	Plant	Eukaryota	PRW18306.1	cysteine ase RD21a-like	467	
Eolophus roseicapillus	Animal	Eukaryota	NXD73214.1	CATS protein, partial	330	

Cysteine proteinase 5

Table 12 Results of search performed in BLAST for cysteine proteinase 5 in *Blastocystis* spp., as well as results of organisms used to construct a phylogenetic tree of the protein.

Organism	**Gene ID**	**Protein ID**	**Protein**	**Length protein**	**Query cover**	**e-value**
Entamoeba histolytica		CAA62835.1	cysteine proteinase	318		
Blastocystis spp. ST7	XM_013039357.1	XP_012894811.1	uncharacterized protein	313	79 %	5e^-55
Blastocystis spp. ST4	XM_014674217.1	XP_014529703.1	peptidase C1A domain-containing protein	318	69 %	6e^-51

Phylogeny tree

Organism	**Kingdom**	**Domain**	**Protein ID**	**Protein**	**Length**
Entamoeba histolytica	Protist	Eukaryote	CAA62835.1	cysteine proteinase	318
Oppia nitens	Animal	Eukaryote	XP_054166525.1	procathepsin L-like	335
Cherax quadricarinatus	Animal	Eukaryote	XP_053637670.1	digestive cysteine proteinase 1-like	349
Stegodyphus mimosarum	Animal	Eukaryote	KFM79807.1	Cathepsin L, partial	384
Orchesella cincta	Animal	Eukaryote	ODN04735.1	Cathepsin L	330
Gouania willdenowi	Animal	Eukaryote	XP_028293264.1	cathepsin L1-like	328
Mizuhopecten yessoensis	Animal	Eukaryote	XP_021341415.1	cathepsin L1-like	346
Haliotis rufescens	Animal	Eukaryote	XP_046336476.2	procathepsin L-like	327
Rotaria sordida	Animal	Eukaryote	CAF3796554.1	unnamed protein product	336
Cunninghamella echinulat	Fungi	Eukaryote	KAI9303884.1	hypothetical protein BJ944DRAFT_241048	232
Clostridia bacterium	Bacteria	Bacteria	MBO5344712.1	MAG: hypothetical protein J6A51_02335	317
Blastocystis spp. ST7	Chromista	Eukaryote	XP_012894811.1	uncharacterized protein	313
Blastocystis spp. ST4	Chromista	Eukaryote	XP_014529703.1	peptidase C1A domain-containing protein	318
Linum tenue	Plant	Eukaryote	CAI0384080.1	unnamed protein product	344
Hordeum vulgare subsp. vulgare	Plant	Eukaryote	BAK02675.1	predicted protein	333

Appendix 3 – List of primers and oligo bridges

Table 13 Overview of the primers used for the promoters and terminators for the genes Peptidase C1A and 60S Ribosomal Protein L32. This was used for the experimental part of the project.

Primer nr.	Sequence 5'-3'	Description
MRT-129	GTAGGGCTCTGCTGCCCG	Forward primer for PC1A promoter V1263
MRT-130	CGCTCACAAGATTATTATGAATAGAAACGTTG	Reverse primer for PC1A promoters
MRT-131	CAGACACAGAAGCCGCCTCAG	Reverse primer for 60SRPL32 terminator
MRT-132	TTTATTGATCGGAGTGATTGACAATAAATCTGTAGAGA	Reverse primer for 60SRPL32 promoters
MRT-133	TTGCTGTAGCCGCGGATG	Forward primer for PC1A promoter V2207
MRT-134	GTAGCGATTGGGCGAAGGCT	Forward primer for PC1A promoter V643
MRT-135	ATAGAATCTCATGGCAAAGTATTATAATAAAGAAAAG	Forward primer for PC1A terminator
MRT-136	TAGAAAATCAGCCGTTCCTATTTATATTATTCAC	Reverse primer for 60SRPL32 terminator
MRT-137	TTCTCTAACGCTTCTGCGTGTTCTG	Forward primer for 60SRPL32 promoter V1000
MRT-138	CTGCGGTTGAGAATGACAAAAATAGAAAC	Forward primer for 60SRPL32 promoter V1295
MRT-139	TTTTGTTTGTTGGAATCAAATTCGAACGCTC	Forward primer for 60SRPL32 terminator

Table 14 Overview of oligo bridges with a numbering system and sequence with direction 5'-3'.

Oligo nr.	Sequence 5'-3'
ssOB_70	cacgacgttgtaaaacgacggccagtgccaCTGCGGTTGAGAATGACAAAAATAGAAACT
ssOB_71	GATTTATTGTCAATCACTCCGATCAATAAAATGGTCTTCACACTCGAAGATTTCGTTGGG
ssOB_72	cacgacgttgtaaaacgacggccagtgccaTTCTCTAACGCTTCTGCGTGTTCTGTGTGT
ssOB_73	TGGCGGCTGTGCGAACGCATTCTGGCGTAATTTTGTTTGTTGGAATCAAATTCGAACGCT
ssOB_74	GGATGGAAGCTGAGGCGGCTTCTGTGTCTGaattcgtaatcatggtcatagctgtttcct
ssOB_75	cacgacgttgtaaaacgacggccagtgccaGTAGCGATTGGGCGAAGGCTGAGATGTAGG
ssOB_76	ACGTTTCTATTCATAATAATCTTGTGAGCGATGGTCTTCACACTCGAAGATTTCGTTGGG
ssOB_77	cacgacgttgtaaaacgacggccagtgccaGTAGGGCTCTGCTGCCCGGATGAATCATGC
ssOB_78	cacgacgttgtaaaacgacggccagtgccaTTGCTGTAGCCGCGGATGGGGACGAATCAG

Table 15 Scheme of oligo bridges used for the different vectors.

Plasmid name	Oligo bridge nr.
pMRT-ϕ_IH_60SRPL32_V1295	ssOB_70 ssOB_71
pMRT-ϕ_IH_60SRPL32_V1000	ssOB_71 ssOB_72
pMRT-τ_IH_60SRPL32	ssOB_73 ssOB_74
pMRT-ϕ_IH_PC1A_V643	ssOB_75 ssOB_76
pMRT-ϕ_IH_PC1A_V1263	ssOB_76 ssOB_77
pMRT-ϕ_IH_PC1A_V2207	ssOB_76 ssOB_78
pMRT-τ_IH_PC1A	ssOB_79 ssOB_80