Functional annotation of selected *Theileria parva* hypothetical proteins without known sequence descriptions and pathway associations

Submitted in partial fulfillment of the requirements of the degree

Magister Scientiae

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Veterinary Science

Department of Veterinary Tropical Diseases

By

Bongiwe P. Mahlobo (15406360)

April 2017

ABSTRACT

Cattle theileriosis is infamous for hampering the economic development of south, central and east African countries due to the exorbitant numbers of cattle mortalities. The disease is caused by *Theileria parva*, a tick-transmitted hemoprotozoan parasite belonging to the phylum Apicomplexa. Infection of cattle with the cattle-derived *T. parva* isolates is responsible for the East Coast fever while infections by buffalo-derived isolates result in the Corridor disease. A transcriptome study comparing two *T. parva* isolates, representing cattle- and buffalo-derived parasites, identified several differentially expressed transcripts, of which 54.4% encode hypothetical proteins (HPs). These proteins are believed to be crucial in understanding the disease syndromes caused by *T. parva* infections; hence, the purpose of this study was to annotate their function. The 309 proteins analysed in this study exclude HPs that were assigned sequence descriptions and had pathway associations with initial screenings using Blast2GO and KEGG pathway analyses. For function prediction, an integrated bioinformatics approach was employed which facilitated sequence comparison, protein family classification, domains discovery, sub-cellular localisation, protein-protein interactions and identification of virulence factors. Overall, 277 (90%) HPs were successfully annotated for function with 224 of these being possible virulent proteins. Enzymes, membrane-associated proteins, transcription factors and secreted proteins, were some of the protein families detected among the HPs. Secretome analysis revealed 57 HPs containing signal peptides, suggesting possible interactions with the host. Thus, among the HPs investigated, there are proteins that could have various functions significant to the pathogenesis of cattle theileriosis including attachment of the pathogen to the host surfaces, disruption of host signal pathways, colonisation of the host cell, immunosuppression, host cell phenotype modulation and proliferation. Sub-cellular localisation revealed three HPs that did not have homologs to any of the vertebrate host proteins, which can be investigated as possible therapeutic targets. The findings of this study will facilitate a better understanding of the mechanism of pathogenesis associated with cattle theileriosis and identification of novel targets to improve disease control strategies. Thus, HPs with predicted biological roles of interest should be further explored experimentally to confirm their roles in cattle theileriosis.

DECLARATION

I, Bongiwe Priscah Mahlobo, declare that this is my own unaided work hereby submitted for the Master's degree at the University of Pretoria. I also confirm that this work has not been submitted before for any degree or examination at any other university.

_____ _____

Bongiwe Priscah Mahlobo **Date**

ACKNOWLEDGEMENTS

I wish to thank God for giving me the strength to complete this study.

I am also extremely grateful to my supervisors for their guidance, expertise and mentorship throughout the duration of the project. Your support and encouragement indispensably made it possible for me to finish this project.

I would also like to extend further acknowledgement to the National Research Foundation of South Africa and AgriSeta for funding my study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF OUTPUTS

Manuscript in preparation:

Bongiwe P. Mahlobo, Fortunate Mokoena, Paul T. Matjila and Kgomotso P. Sibeko-Matjila. *In silico* functional prediction and characterisation of selected *T. parva* hypothetical proteins.

Conference Proceedings:

Bongiwe P. Mahlobo, Fortunate Mokoena, Paul T. Matjila and Kgomotso P. Sibeko-Matjila. *In silico* functional prediction and characterisation of selected *T. parva* hypothetical protein (Poster presentation). 45th PARSA Conference, Lagoon Beach Hotel, Cape Town, South Africa, 28-31 August 2016.

Bongiwe P. Mahlobo, Fortunate Mokoena, Paul T. Matjila and Kgomotso P. Sibeko-Matjila. *In silico* functional prediction and characterisation of selected *T. parva* hypothetical proteins (Poster presentation).4th Genomics Research Institute Symposium, University of Pretoria, South Africa, 18 November 2016.

CHAPTER 1

1.1. BACKGROUND

Tick-borne diseases (TBDs) affecting cattle have devastating socio-economic impacts on many developing continents, with a more emphasized crippling economic impact seen in countries across East Africa [1]. The common TBDs of livestock in Africa include babesiosis, cowdriosis, anaplasmosis and theileriosis, and they are responsible for high mortality and morbidity in cattle. Control of TBDs is usually through the use of acaricides, which target the tick vectors [1]. In the African region, *Theileria parva*, the etiological agent of cattle theileriosis, is transmitted by *Rhipicephalus appendiculatus* and *R. zambeziensis* ticks [2] and cause a lympho-proliferative disease, East Coast fever (ECF). East Coast fever is a virulent TBD infamous for hampering the economic development of south, central and east African countries due to exorbitant numbers of cattle mortalities [3]. At least 5.5 million deaths resulted from ECF following its introduction to South Africa (SA) and the eradication process of the disease cost the country approximately R100 million [4]. In 1989 alone, the cost of ECF was estimated to be at $186 million [5] in 11 affected African countries. *Theileria parva* infections transform bovine lymphocytes, thus, inducing uncontrolled proliferation of the affected cells. Infected animals can recover from ECF [6], however, the recovery period is usually prolonged and recovered animals may become carriers of the parasite [7]. In certain instances, the animals which recover from ECF remain frail and may take a while before they are productive again [8].

East Coast fever was introduced to SA in 1902, as a result of cattle importation from the eastern coast of Africa. The importation was followed by a prompt spread of the disease within the low-lying areas of SA. Indeed, by 1912, ECF had invaded all areas in SA, where it was able to maintain itself. The epidemic spread of ECF led to government intervention; as a result, extremely stringent measures, such as dipping, spraying and quarantine, were put in place to eradicate ECF [4]. However, a few years following the eradication of ECF, it became apparent that another form of theileriosis, Corridor disease, still existed. Corridor disease has similar manifestations as ECF, but it only occurs in the presence of *T. parva*-infected buffalo and kills the affected animals more rapidly than the ECF. To distinguish between the two disease syndromes, the causative agents were designated *T. p. Lawrencei* and *T.p. parva* for Corridor disease and ECF, respectively [9]. This nomenclature was later revised to cattle-

derived and buffalo-derived *T. parva* for the etiological agents of ECF and Corridor disease, respectively [10].

There are clinical differences that distinguish Corridor disease and ECF; however, the causative agents bear similar morphological and serological characteristics, making it impossible to perform differential diagnosis using traditional methods such as microscopy and serology-based tools. Molecular characterisation studies have demonstrated that the buffalo-derived *T. parva* isolates are more genetically diverse compared to their counterparts [11]; these findings have been recently confirmed by genome sequence comparative studies [12, 13]. Since the cattle-derived and buffalo-derived *T. parva* parasites are not distinguishable by morphology, serology or genetics; it is believed that transcriptome profiles may allow differentiation of these parasites. Consequently, Sibeko K.P. (Department of Veterinary Tropical Diseases, University of Pretoria, personal communication) has analysed the transcriptome profiles of two *T. parva* isolates, *T. parva* Muguga and *T. parva* 7014, respectively, representing cattle-derived and buffalo-derived *T. parva* isolates. The study identified 3969 transcripts and expression of 1089 of these, was differentially regulated. Analysis of the differentially expressed transcripts (DETs) revealed that 593 encode some hypothetical proteins (HPs). Although the functions of the *T. parva* HPs are unknown, it is suspected that they may be vital for explaining the different disease syndromes resulting from the *T. parva* infections.

1.2. LITERATURE REVIEW

Cattle theileriosis is a disease of cattle resulting from infections with *Theileria* species, mainly *T. parva* and *T. annulata*. Disease syndromes caused by *T. parva* include the ECF, January disease and Corridor disease, occurring in the eastern, central and southern Africa, while *T. annulata*, occurring in North Africa, the southern Europe and Asia, causes a form of cattle theileriosis known as the Tropical theileriosis in these regions. However, *T. parva* is the more malignant of the two species. Collectively, *Theileria* parasites are responsible for losses amounting to hundreds of million dollars yearly in Asia and Sub-Saharan Africa [14].

1.2.1. *THEILERIA* PARASITES

The genus *Theileria*, of the Theileriidae family, comprises tick-transmitted protozoan parasites known to cause a number of disease syndromes in domestic and wild ruminants, including cattle, sheep, Asian water buffalo, African buffalo and goats [14, 15]. This genus has been included in the Apicomplexa sub-phylum, which is believed to have originated about 930 million years ago [16], as its members possess an apical complex [14]. *Theileria* species are also classified in the class Sporozoa together with apicomplexan species of human parasites such as *Plasmodium* and *Toxoplasma* [17]. The *Theileria* species are also phylogenetically closely related to *Babesia* species, the tick-borne protozoan parasites which infect the erythrocytes of mammals, including domestic livestock [17].

Various diseases caused by the *Theileria* species are listed in **Table 1.1**. *Theileria parva* and *T.annulata* are the most virulent *Theileria* parasites of cattle [18] while *T. lestoquardi* is the most pathogenic species in small ruminants [15]. Other known *Theileria* species comprise the mildly pathogenic *T. mutans, T. tauroiragi* [19] and *T. velifera* which is non-pathogenic [20] (**Table 1.1**). Cattle theileriosis can lead to reduced production and high mortality in animals, thus the disease is of economic importance [10]. Consequently, the threat of *Theileria* infections has led to restricted movements of cattle between countries. Cattle theileriosis is most severe in recently introduced animals, thus a major constraint in the importation of new breeds into endemic areas.

Table 1.1. *Theileria* species, their main vector ticks, known geographical distribution and diseases they cause.

***Theileria* species**	**Major vector tick**	**Known distribution**	**Disease**	**Reference(s)**
T. parva	*Rhipicephalus appendiculatus* *R. zambesiensis* *R. duttoni*	Eastern, Central and Southern Africa	ECF, Corridor disease, January Disease	[2, 22, 23]
T. annulata	*Hyalomma anatolicum* and other *Hyalomma* spp.	Southern Europe Western, Southern and Eastern Asia Northern Africa	Tropical/ Mediterranean theileriosis	[14]
T. mutans	*Amblyomma variegatum* and other *Amblyomma* spp.	Western, Eastern Central and Southern Africa Caribbean islands	Benign theileriosis	[19, 24]
T. sergenti	*Haemaphysalis* spp.	Japan, Korea, Southern Europe, Asia and Australia	Benign theileriosis	[14]
T. buffeli	*Haemaphysalis* spp.	Europe, Asia, Australia, Eastern Africa	Benign theileriosis	[14, 25]
T. lestoquardi	*Hyalomma* spp.	Asia and Northern Africa	Malignant theileriosis	[13]
T. ovis	*Hyalomma* spp.	Asia	Benign theileriosis	[25]

1.2.2. DISEASE SYNDROMES RESULTING FROM *THEILERIA PARVA* INFECTIONS

Theileria parasites infect the erythrocytes and lymphocytes of cattle and African buffalo, causing ECF, Corridor disease and January disease (also known as the Zimbabwean theileriosis) [26]. *Theileria parva* has long been present in African buffalo (*Syncerus caffer*) [27] and the parasite still circulates within the buffalo population in SA. Thus, *T. parva* infections can affect both the cattle and game farming industries in this country.

1.2.2.1. East Coast fever (ECF)

East Coast fever is observed when *T. parva* is transmitted from infected cattle to susceptible cattle through tick bites. The transmission of *T. parva* by several tick species has been experimentally shown, however; it is widely accepted that *R. appendiculatus* is the primary vector [22, 29]. At the schizont stage, *T. parva* transforms the bovine host lymphocyte malignantly [29], thus, contributing significantly to a number of mortalities and pathogenesis related to this parasite. The clinical signs of ECF include; pulmonary oedema [33], high fever, respiratory distress, peripheral lymphadenopathy and anorexia [34].

An estimated 5.5 million cattle died from ECF in SA during the first epidemic in the 1900s and the cost of eradication was enormous [4]. Although immense efforts were taken to eradicate *T. parva* in SA in 1954 [30], the persistence of the tick-vector undermined these efforts [31]. *Rhipicephalus appendiculatus* is still wide spread in the country and the ecto-parasite still persists in the buffalo population. Thus, the presence of a highly susceptible cattle population requires rigorous regulations to prevent transmission of the protozoan parasite [32].

1.2.2.2. Corridor Disease

After the eradication of ECF, Corridor disease became the most important form of theileriosis in SA and it is still a serious threat in the country. The disease derives its name from the location where it was first identified, a corridor between Hluhluwe and Umfolozi Game Reserves in KwaZulu-Natal (SA) [35]. Corridor disease is a fatal cattle disease which is caused by buffalo-derived *T. parva* parasites, formerly designated *T. p. Lawrencei* [36]. Contrary to ECF, this does not require the presence of buffalo for transmission, Corridor

disease is observed when the parasite is transmitted from an infected buffalo to susceptible cattle. Thus, it is not surprising that Corridor disease is commonly observed in cattle grazing pastures shared with *T. parva*-infected buffalo. It is believed that the buffalo-derived parasite cannot adapt to cattle, as most of the cattle infected with these parasites die before the parasite can mature to the tick-infective stage, the piroplasm [35]; hence, the disease is considered self-limiting. The course of the disease is normally shorter than that of ECF. Symptoms and clinical signs include; oedema, lachrymation, corneal opacity, dyspnoea and generalised lymphadenopathy [7]. In SA, Corridor disease is endemic to the North Eastern Limpopo, Northern KwaZulu-Natal and Northern Mpumalanga Provinces [36]. The presence of this disease has also been noted in eastern and central Africa [26].

1.2.2.3. January disease

January disease, also referred as the Zimbabwean theileriosis, is a rapidly fatal cattle disease caused by a cattle-associated *T. parva*, previously known as *T. parva* bovis [36]. The disease was first recognized and distinguished from the classical ECF in 1936. The name January disease is influenced by the seasonal occurrence (between December and March) of the disease, which is well in accordance with the rampant activity of the tick vector; *R. Appendiculatus* in Zimbabwe [37]. January disease is found mainly in Zimbabwe; there it causes a significant number of cattle deaths per year. The clinical features, pathogenesis and pathology of this January disease remain largely similar in presentation and manifestation as the other two *T. parva* infections, ECF and Corridor disease.

1.2.3. LIFE CYCLE OF *THEILERIA PARVA*

The life cycle of the *T. parva* (**Figure 1.1**) is extremely complex in both the mammalian host and tick vector [38]. During a blood meal, an infected tick injects sporozoites into the host's bloodstream. The sporozoites then travel through the bloodstream of the host, where they infect the lymphocytes [39]. Inside the lymphocytes, the sporozoites mature into schizonts [40]. The infected lymphocytes grow larger and start to divide; each enlarged lymphocyte is called a lymphoblast. As each lymphoblast divides, the schizonts inside also divides (clonal expansion), ensuring infection of each of the two daughter cells produced by the dividing

lymphoblast. The infected lymphocytes rapidly expand and spread throughout the lymphoid system of the animal, resulting into an extensive destruction of the host cells. Some schizonts inside the lymphoblast undergo merogony, to form merozoites [41]. Later, the merozoites rupture, and are subsequently, released into the host bloodstream where they invade erythrocytes. In the erythrocytes, the parasites mature into piroplasms that can be ingested by the tick during a blood meal [41].

As ticks feed on the *T. parva*-infected animals, they ingest erythrocytes comprising the piroplasms. Once inside the tick gut, the piroplasms differentiate into male and female macrogametes, which later undergo syngamy (sexual reproduction) to form zygotes [39] (**Figure 1.1**). The zygotes differentiate into kinetes, which then move to the salivary gland and enter the digestive cells [40, 43]. Here, the parasites form sporoblasts, which later give rise to sporozoites. The sporozoites are introduced into a mammalian host during a blood meal, along with tick saliva, initiating a new cycle of parasite development [39]. Generally, for transmission to occur, the infected tick has to attach to the host for some days to enable the sporozoites to mature and be emitted in the saliva of the feeding tick.

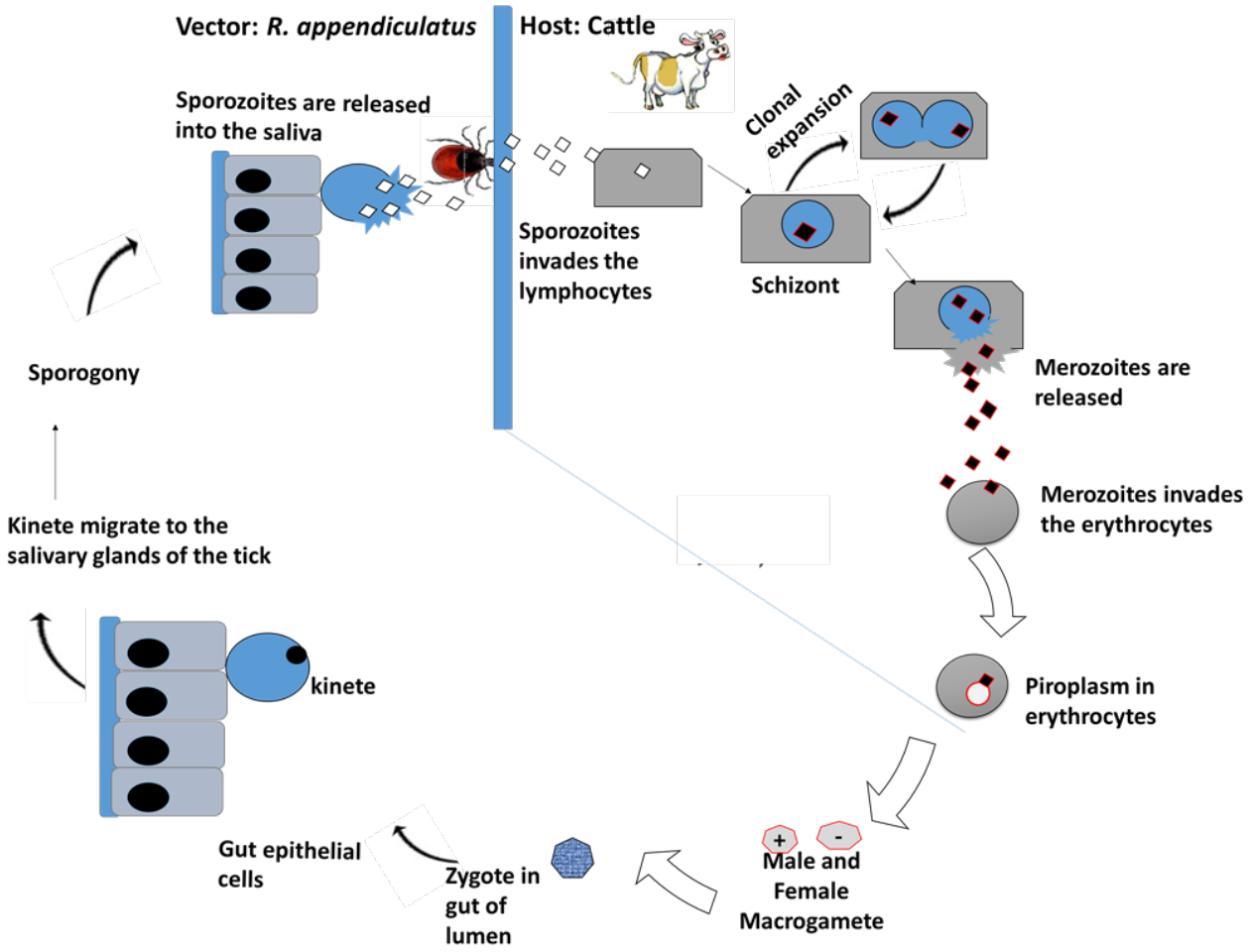


Figure 1.1.The life cycle of *T. parva* in the vertebrate host and tick vector.

1.2.4. CONTROL AND TREATMENT OF *THEILERIA PARVA* INFECTIONS

To control Corridor disease in South Africa, the government has enforced quarantine measures aimed at preventing contact between buffalo infected with *T. parva* and susceptible cattle through the use of fences [36]. Tick control measures are also implemented such as the spraying of cattle with acaricide substances [43] and dipping. In other regions of Africa still most affected by ECF, the infection and treatment method (ITM) is used as a vaccination strategy; ITM requires simultaneous infection with the sporozoite-stage of the parasite and treatment with a tetracycline [44]. The Muguga cocktail vaccine, used for ITM, consists of the three strains of *T. parva* including Muguga, Serengeti-transformed, and Kiambu 5 and is employed extensively in East Africa [12]. The use of local strains is highly recommended to avoid the risk of introducing new parasite strains. Hence, in Zambia, ITM uses the local stocks, namely Katete and Chitongo, to control ECF [3]. Vaccination and chemotherapy are

not allowed in SA to prevent the development of a carrier state. DNA-based recombinant vaccines are being investigated to improve on the limitation experienced with the ITM.

1.2.5. WHOLE-GENOME SEQUENCING OF *THEILERIA PARVA*

Genome-wide research started a decade ago, when the whole genome of the *Haemophilus influenzae* strain Rd KW20 was sequenced successfully [45]. Whole-genomes of several organisms have since been sequenced. Gardner *et al.*[29] reported sequencing of the haploid nuclear genome of *T. parva*. The *T. parva* whole-genome project is available in the DNA Data Bank (accession number: AAGK00000000). According to Gardner *et al.*[29], the *T. parva* nuclear genome consists of four chromosomes and is approximately 8.3 x 106 base pairs in size. In comparison to another apicomplexan parasite, *Plasmodium falciparum* (**Table1.2**), the chromosomes of the *T. parva* parasite display limited variance of gene synteny and its plastid-like genome represents the first example, where all the apicoplast genes are encoded on one DNA strand [29]. Similar to *P. falciparum*, *T. parva* chromosomes comprise a single remarkably A+T rich region (>97%; **Table 1.2**) that is roughly 3000bp in length [46]. Compared to *T. annulata*, the *T. parva* genome is smaller in size but has more protein encoding genes, 4035 compared to 3807; however, this number is 20% less than the protein encoding genes identified in *P. falciparum* (**Table 1.2**).

Table 1.2. Comparison of the *T. parva* nuclear genome coding characteristics with other sequenced apicomplexans [15, 29, 47, 48]

Features	**Parasitic apicomplexan**		
	T. parva	*P. falciparum*	*T. annaluta*
Size (bp)	8,308,027	22,853,764	9,100,000
Number of chromosomes	4	14	8
Number of protein encoding genes	4035	5268	3807
Number of hypothetical proteins	2498	3208	925
Apicoplast genome size	39.5	35	NA
Total A+T content (%)	>97%	~84%	NA

Attributed to the successful and rapid genome sequencing projects, indeed, two other *Theileria* parasites; namely, *T. orientalis* and *T. annulata* have been sequenced [49]. The *T. orientalisis* genome is approximately 9.0 Mb in size; 8% larger than that of *T. parva*, and *T. annulata*. The number of predicted protein-coding genes found in *T. orientalis* is however, almost similar to that identified in *T. parva* [15].

The apicoplast and mitochondrial genomes of *T. parva* have also been sequenced [46]. The apicoplast plays a critical role in parasite metabolism such as isoprenoid biosynthesis, they are the sites of type II fatty acid and heme biosynthesis [50]. In contrast to *P. falciparum*, all genes of the apicoplast genome of *T. parva* are transcribed in the same direction [29]. Indeed, 44 protein encoding genes are reportedly encoded in the *T. parva* apicoplast genome [46], of these, (26/44) 59%, are common to those of *P. falciparum*. Comparing *T. parva* to *T. annulata*, approximately 100 AT (Apicoplast-Targeted) proteins were identified in both species and 40% of these proteins were HPs. It is assumed that some of these proteins may execute some apicoplast-associated roles, possibly involved in host-pathogen interaction.

Recently, Hayashida *et al*.[12] performed a genome sequence comparative study, where nine strains of *T. parva* including seven cattle-derived and two buffalo-derived strains, were investigated. The genomes of the nine strains were sequenced by the next-generation sequencing (NGS) technology. This technology uses the same principle as the Sanger sequencing technology, however, it has a higher throughput, thus; allows for the sequencing of many pieces of DNA at the same time.

In Hayashida *et al*.[12] study, phylogenetic analysis and recombination detection were performed to identify the relationship between the nine strains of *T. parva*; the genome sequence comparisons were done against the Muguga reference genome sequence. Single nucleotide polymorphisms (SNPs), were shown to be more abundant in the buffalo-derived strains [12]. Finally, this study developed the high-density SNPs map for genotyping or linkage analysis of the *T. parva* parasites. Essentially, SNPs-based genotyping can now be used to distinguish field and vaccine strains of *T. parva*, where ITM is used for immunisation against cattle theileriosis [12]. The genome sequence of *T. parva* provides information that can facilitate research on parasite biology, assist in vaccine development by the detection of schizont antigens and extend comparative apicomplexan genomics [29].

1.2.6. ANALYSIS OF *THEILERIA PARVA* TRANSCRIPTOME

Genome-wide transcription data is a significant tool towards understanding the biology of an organism in a systems context. Several methods are used for powerful and high-throughput analysis of organism's transcriptome profile including microarrays based on hybridisation, serial analysis of gene expression (SAGE) and massively parallel signature sequencing (MPSS; [51]). The massively parallel signature sequencing is an improvement from the SAGE, using novel sequencing and amplification technologies, and indeed, covering more reads [51]. One of the biggest advantage of MPSS, is its ability to detect transcripts that are expressed at very low levels [52], while its major limitation is the detection of transcripts based on the presence of the DpnII restriction site. Therefore, MPSS will not detect transcripts lacking this site.

Bovine lymphocytes that have been infected with *T. parva* can be propagated *in vitro* just like tumors, allowing RNA derived from the schizont stage of the parasite to be obtained in sufficient amounts for analysis. This enabled Bishop *et al.*[51] to analyze *T. parva* transcriptome using MPSS and to annotate signatures derived from the schizont stage RNA using the genome sequence [29]. The analysis of the transcriptome of *T. parva* revealed that most of its genes are transcriptionally active in the schizonts stage. Recently, the transcriptome profiles of two *T. parva* isolates, one representing ECF and the other one representing Corridor disease, have also been studied using NGS (Sibeko KP, Department of Veterinary Tropical Diseases, University of Pretoria, unpublished data). About 98.4% (3969) of the 4035 predicted coding genes were mapped from the two transcriptomes using NGS compared to the 73% detected using MPSS [51]. This highlighted the limitation of the MPSS approach which depends primarily on the presence of the DnpII restriction site for mapping of protein coding genes. Moreover, in the NGS study, the differentially expressed transcripts (DETs, n = 1089) were identified between the cattle-derived and buffalo-derived isolates investigated; however, bulk of these (~57%) were genes coding for HPs.

1.2.7. HYPOTHETICAL PROTEINS

Due to a number of limitations such as the time required for experimental methods and high cost, complete annotation of many genomes has not yet been achieved. Consequently, increasing amounts of the so-called "hypothetical proteins", novel proteins with unknown functions, have been discovered in recent years [53–56]. Computer-based methods, which generally search various databases using different algorithms to extrapolate functions of proteins is a good alternative to the time-consuming laboratory-based approaches. Computational analysis is thought to be an effective means of accomplishing structural and functional annotation of HPs [57].

Functional annotation of HPs can help our understanding of their precise molecular function; identification of proteins vital for the organism's survival can also enable the development of novel vaccines or drug targets that can improve the control of pathogen infections [58, 59]. Furthermore, functional annotation of HPs may assist in identifying novel protein cascades or pathways, hence, complete our understanding of biological importance of other novel proteins [60, 61].

Previous studies have successfully annotated the roles of HPs using the following bioinformatics web tools: PFAM (http://pfam.sanger.ac.uk/), COGs (http://www.ncbi.nlm.nih.gov/COG/), INTERPROSCAN, BLAST (http://www.ncbi.nlm.nih.gov/ BLAST/) and KEGG (www.genome.jp/kegg/kegg4.html). Al-Khafaji *et al*.[62] used these tools to analyse HPs from *Mycobacterium tuberculosis* for function prediction, while Kumar *et al*.[61] performed sequence analysis of 43 HPs from *Candida dubliniensis*. To date, bioinformatics web tools have had a wide application in different organisms and have proven to be powerful tools in elucidating biological information.

1.2.8. COMPUTATIONAL BIOLOGY

High-throughput techniques for DNA sequencing and analysis of gene expression have caused an exponential growth in the quantity of genomic data which is stored in the NCBI (National Centre for Biotechnology Information) [63]. There is a need for tools that can allow analysis of high-throughput DNA and RNA sequence data, because of the exponential growth

in the sequence data. This led to the emergence of the field of computational biology, also referred to as bioinformatics. Computational biology refers to the use of computer science, statistics, and mathematics for solving problems in biology. Computational biology tools have also allowed annotation of function for proteins which had no known functions for many years, including HPs, leading to the discovery of proteins with significant biological functions and applications, for drug and vaccine development.

1.2.8.1. Biological features exploited in *in silico* methods for protein function prediction

Several approaches to predict protein function based on sequence similarity, structure and interactions to known protein have been developed [64–67]. It is also interesting to note that clustering patterns from phylogenetic profiles [46, 69] have also been applicable for protein function prediction. Genes with the same expression patterns may have related functions [69–71] hence, gene expression patterns data is also useful for the prediction of protein function [72]. Also, the presence of conserved motifs in the protein sequence can be used for function annotation [73]. Although several approaches used for the prediction of proteins functionality rely on finding similarity in the sequence and structure between a protein of unknown function and annotated proteins, in some cases, homologous proteins may perform different functions [74]. Binding clefts, binding pockets and domains present in proteins can also be used as annotation features for protein function prediction [75].

I. Sequence-based approaches

Sequence analysis is the first step used to determine the level of homology between proteins, subsequently; function can be putatively assigned [76]. Sequence-based approaches used for function prediction are applicable for a larger dataset than structure-based approaches. This is obviously because sequence information is available for the majority of proteins and also because most of function information is stored in sequence databases.

Sequence alignment:

This process organizes RNA, DNA, or protein sequences to detect similar regions that may possibly be important for structural, functional, or evolutionary relationships between sequences. Pair-wise sequence alignments produces statistically significant estimates with strong confidence when homologs have been identified (**Table 1.3**) [77]. On the other hand, multiple sequence alignments (MSA) are commonly used to identify conserved regions across a set (two or more) of sequences [78]. They are much more informative compared to pair-wise alignments, they provide additional information about the conserved regions within the protein [79] (**Table 1.3**). Multiple sequence alignment programs use specific algorithms to deduce the different interrelations between sequences and thus eliminating a biasness that may result from using a single program. Examples of sequence alignment tools include the Tree-based Consistency Objectives Function for alignment Evaluation (T-COFFEE), a MSA program that produces accurate results in a modest speed compared to other MSA programmes and Clustal Omega: another MSA programme that aligns protein sequences and distinctly delivers more reliable alignments than the other most widely used programmes [80].

Sequence alignment has a high sensitivity and is more reliable in discovering evolutionary relationships between genes or proteins [82, 83]; hence, they are mostly used in bioinformatics analyses that involve comparison of homologous protein sequences [80].

Table 1.3. Comparison of pair-wise alignment and multiple sequence alignments

Pair-wise alignment	**Multiple sequence alignment**
Categorised as either a local or global method	Is generally a global multiple sequence alignment
Compares two biological sequences of either DNA , RNA or protein	Compares three or more biological sequences of either DNA , RNA or protein
A comparatively simple algorithm is used	Complex algorithms are used
Finds conserved regions between two sequence	Detects regions of conservation in a protein family
Does similarity searches in a database	Phylogenetic analysis

Homology analysis:

Homolog refers to sequences that typically share a common evolutionary ancestor and tend to perform similar functions. It had been generally accepted in the bioinformatics community that protein sequences should at least share 30% sequence identity over their entire lengths to be considered homologous. However, this 30% criterion is known to miss many homologs that can be easily detected as some homologous proteins can share less than 20% sequence identity [77]. Homologous sequences can be detected by tools that employ scoring matrices to calculate scores and/or analyse sequences, such as BLAST and PSI-BLAST [83] both of which use the BLOSUM62 matrix. These similarity search tools calculate local sequence alignments which detect the regions of similarity between two sequences. The inference of functional similarity on the basis of global similarity and conserved active site residues is much more reliable than inferences that are solely based on local similarity [77]. When inferring homology, expectation values (E-value), sequence identity and coverage are provided in order to retrieve reliable and significant results. In addition, it is also very crucial to determine if functional domains are part of the alignment [79] as this can drastically improve predictions and lessen errors caused by merely copying the function of the reference protein to the query protein [77]. Below is a brief discussion of the sequence similarity search programmes employed in the study:

- BLAST is a most widely used tool for sequence-similarity searches; it searches against the whole database for optimum local alignments with the query and generates outputs of various hits that match the query sequence. BLAST employs a scoring matrix known as (BLOcks SUbstitution Matrix) BLOSUM62, which is designed to identify protein similarities.

- PSI-BLAST (Position Specific iterated-BLAST) [83] is another program used for sequence similarity search. The program achieves this by constructing sequence profiles using multiple sequence alignments and then calculating a Position Specific Scoring Matrix (PSSM) for residue conservation. A PSSM is generated by searching through sequence profiles and then the probability of a particular residue is calculated at each position. PSI-BLAST iteration runs faster and much more sensitive than the standard BLAST, it also provides more re-iterations. The multiple iterations provided by PSI-BLAST helps in refining the results and also detects protein family members that are

distantly related [83]. This method is based on the theory that functionally important residues are conserved in all sequences that are identical [83].

Orthology analysis:

Orthology prediction is also one of the most popular platforms used for inferring functional similarity. Orthology is a term used to refer to proteins that have originated from a common ancestor separated by speciation, these group of proteins tend to retain a common function over evolutionary time, thus, making orthologs identification a potent tool for functional annotation [84]. The concept of orthology was initially introduced to the field of molecular systematics [85] to differentiate between two kinds of homologs: orthologs (defined above) and paralogs, which are homologous sequences that have diverged after a duplication event [86]. The latter tend to attain novel functions and thus are not suitable for functional assignment [79]. Several previous studies have successfully applied orthology for functional characterisation [87–89].

EuPathDB, is a Eukaryotic Pathogen Database (EuPathDB;http://eupathdb.org) that can be used for identification of orthologs, resources about 11 other databases which support functional genomic and eukaryotic pathogen genomic data, phylogenomics and isolate data. These resources are similar on the basis of the infrastructure used to build them and they also provide a search strategy system that allows complex analyses of the underlying data. PiroplasmaDB (http://piroplasmadb.org) [90] is a programme recently added to the EuPathDB family of databases, which allows the analysis of *Theileria* and *Babesia* parasites. PiroplasmaDB uses the search strategy system [91] to conduct searches; it involves sequential addition of searches to produce refined results [92].

Domains detection:

Domains are defined as basic functional units found in protein sequences [87]. These functional units and the nature of their interactions contribute to the overall function of a protein [94]. A single domain can either function independently or in cooperation with other domains or contribute to the function of a multidomain protein [94]. Therefore, studying protein domains may provide significant clues about the functional role of uncharacterised proteins [95, 96]. Typically, simple proteins comprise one or two domains, while larger

proteins may contain multiple domains necessary for complex cellular functions [97]. Below is a brief description of domain prediction tools employed in the study:

- Simple Modular Architecture Research Tool (SMART): Is a Hidden Markov Models-based online database used for the annotation and identification of domains in proteins by searching for sequences with similar domains through comparing the query sequences with annotated sequences in the database based on domain profiles and architecture. It constructs MSA of protein domain families and then identifies regions in the sequence such as signal peptides, coiled coil and trans-membrane [98]. Additionally, SMART extensively annotates protein domains on the basis of functional class, functionally critical residues, phyletic distributions and tertiary structures.

- InterProScan: Is a tool used for functional analysis of proteins by classifying them into their respective families and predicting domains and other important regions that are present in the protein for domain identification [99]. It analyses a target protein sequence against a combination of protein signature recognition methods of the InterPro databases; ProDom, PROSITE, Pfam, SMART, and PRINTS [99]. InterProScan can be utilised as a system for simple retrieval of the underlying data. Furthermore, this tool was created to be easily accessible and as an extensible system with a powerful internal architecture [99].

- PROSITE: Is another tool largely used to identify and annotate conserved regions in protein sequences, covering domains, protein families, and motifs [100]. PROSITE identifies these regions by using two types of descriptors, generalized and patterns profiles. Generalized profiles describe protein families and modular protein domains, whereas, the pattern profile describes sequence motifs which corresponds to structurally and/or functionally critical residues [101]. The PROSITE descriptors are associated with ProRules (annotation rules) [102], these ProRules provides information about structurally and functionally important amino acid residues thus increasing the discriminatory power of the descriptors [100].

- National Center for Biotechnology Information (NCBI) -Conserved Domain Database (CDD): Is a resource for conserved domain annotation of protein sequences [103]. This

tool contains over 23 500 PSI-BLAST-derived Position Specific Score Matrices representing domains taken from the SMART, Pfam, and from domain alignments derived from COGs. CDD involves the use of a protein's 3D structure to refine domain models and also provide understandings into structure-sequence-function relationships.

- Proteins Families Database (Pfam): Is a HMM-based database widely used for protein domain and families analysis. Pfam is also extensively used in structural biology for the identification of novel targets for structure determination; it employs HMMER3, the improved version of the most popular profile hidden Markov model package (HMMER2) [104]. The HMMER3 software is more sensitive and much faster compared to the former version. Pfam has11 912 families known as Pfam-A, they are found in about three quarters of characterised proteins. In order to further intensify the coverage of the database, the Pfam-A family collection was increased with a set of families (Pfam-B) which are generated automatically [104]. Pfam-B is a derivative of the Automatic Domain Decomposition Algorithm (ADDA) domain collection [105] which automatically identifies protein domains solely from protein sequence alignments [104].

Subcellular localisation:

The functional role of a protein is usually associated to its subcellular localisation and because of this; the prediction of subcellular localisation from protein sequences provides useful clues for assigning cellular protein functions [107, 108]. Several protein subcellular localisation prediction tools/databases have been developed and these databases integrate different data from proteomics-based experiments [109, 110], microscopy-based high-throughput localisation studies [111, 112] and cDNA tagging projects [112]. Thus, it is always important to consult multiple prediction tools and databases in order to get a complete view of the protein subcellular localisation. Below is a brief discussion of the subcellular localisation prediction tools/databases employed in the current study:

- WoLF PSORT: Is the latest version of the PSORT II tool used for the prediction of protein subcellular localisation [113]. WoLF PSORT operates based on PSORT principles and converts the amino acid sequences of a protein into numerical localisation features; this is achieved based on amino acid composition, functional motifs and sorting

signals. Ultimately, a simple k-nearest neighbor classifier is used for predicting the specific subcellular localisation of a protein.

- YLoc: Is a protein subcellular localisation prediction programme, which can attain over 90% prediction accuracy [114]. YLoc does not only predict the subcellular localisation but it also provides reasons as to why a particular prediction was made and identifies which biological features of the protein sequence were considered the prediction. These features include motifs or localisation signals relevant to protein sorting [114]. Additionally, a confidence estimate is provided which helps researchers to rate a prediction as reliable or not. YLoc can also reliably localize proteins targeted to multiple compartments with high accuracy [114].

- TargetP 1.1: Is another programme that predicts the subcellular location of eukaryotic proteins [115]. TargetP assigns the subcellular location based on the occurrence of N-terminal sequence motifs: the mitochondrial targeting peptide (mTP), chloroplast transit peptide (cTP) or secretory pathway signal peptide (SP). Moreover, it also predicts cleavage sites for the predicted pre-sequences/sequence motifs. To achieve this, it employs ChloroP and SignalP for SP and cTP cleavage site predictions, respectively.

Trans-membrane helixes/domains:

Trans-membrane domains are regions found in membrane proteins which transverse in and out, looping through the membrane [116]. Functions of membrane proteins include enzymatic processes, anchoring of other proteins, ion channel activity or transport of other molecules across the membrane and receptor signalling [40]. Moreover, membrane proteins are considered as putative vaccine targets [117], thus, prediction of membrane proteins is crucial for the identification of putative drug targets.

- Trans-membrane helixes Hidden Markov Model (TMHMM) is a Hidden Markov Model (HMM)-based trans-membrane helixes prediction tool. This tool successfully predicts about 97%-98% of trans-membrane helices [118]. The trans-membrane helices Hidden Markov Model tool discriminates between soluble and membrane proteins with both sensitivity and specificity levels better than 99%; however, the accuracy of the tool decreases when a signal peptide has been detected [118].

- Hidden Markov Mode Topology of Proteins (HMMTOP) is another Hidden Markov Model-based trans-membrane topology prediction tool which predicts both the topology of the trans-membrane proteins and the localisation of helical trans-membrane segments [119]. To augment the prediction power as well as help in the interpretation of results, HMMTOP allows the submission of additional data about segment localisation [119].

Protein-protein interaction (PPI) network:

Non-covalent interactions between protein's residue side chains are the basis for protein assembly, protein folding, and protein-protein interaction (PPI) [120]. These contacts lead to various interactions amongst proteins. It has been revealed that, proteins typically (over 80%) operate in complexes rather than alone when performing their functions [122, 123]. Protein-protein interaction approaches are based on the hypothesis that interacting proteins most likely have similar functions [123–126]. This theory has previously been supported by studies which indicated that 70%-80% of the interacting partners perform at least one similar function [127]. Therefore, the role of HPs can be predicted based on their interaction with a known/characterised protein [128]. Furthermore, data generated from the protein-protein interaction predictions can also be used to predict the druggability of a particular molecule [128]. STRING database [129] is the most popular tool used for the prediction of protein-protein interactions, and it currently covers more than 2031 organisms [130]. STRING (http://string-db.org) incorporates interaction data from different sources such as co-expression (conserved), genomic context and high-throughput experiments. The database accept queries for single and/or multiple protein sequences as inputs and the search can be restricted to one specific organism or clades of organisms [130]. In addition, STRING interactions are provided with a confidence score, and additional information such as protein 3D-structures and domains.

Gene Ontology:

Functional annotation methods frequently allow the use of Gene Ontology (GO) hierarchical vocabulary to unravel proteins functional properties [132, 133]. Gene Ontology terms are catagorised into three domains and each of them describes the different features of gene and/or protein function: Molecular Function (MF), Biological Process (BP) and Cellular Component (CC) [132]. Blast2GO (B2G) is a GO annotation and a statistics platform for high throughput functional annotation and genomic data analysis [133]. Briefly, it employs

BLAST [134] to search for sequences similar to query sequences and then extract the GO terms from each attained sequence by mapping to the existent annotation associations. Finally, an annotation rule assigns GO terms to the input sequence. The annotation and the functional analysis are viewed in a graph format displaying the GO relationships and the most significant areas are color-highlighted [133].

Physiochemical properties:

A proteins' physiochemical property is known to influence its structure as well as its function, hence, knowledge of these properties provides a suitable platform towards the characterisation and prediction of its functional role [135–137]. ExPASy's ProtParam tool: This program computes several physiochemical properties that can simply be inferred from the protein sequence. The input sequence can either be in a raw sequence format or as a Swiss-Prot/TrEMBL accession number [138]. Physiochemical properties calculated by the tool include; the extinction coefficient (EC), amino acid composition, molecular weight (MW), atomic composition, isoelectric point (pl), instability index (II), grand average of hydropathicity (GRAVY) index, aliphatic index and estimated half-life [138].

II. Structure-based analysis

Prediction of the three-dimensional (3D) structure of an unknown protein based on its amino acid sequence using computational methods allows for its biochemical and biophysical functions to be realized [140, 141]. Generally, the prediction of a protein's function from its structure is carried out when sequence-based approaches have shown some limitations, as the protein structure is more conserved than the sequence [141, 142]. The first step towards the inference of protein function from its 3D-structure is usually to use global structure comparison wherein the query protein is compared to structures in the Protein Data Bank in order to detect structural neighbors which are likely to perform similar functions [142]. There are many different methods that are widely employed for protein 3D structures prediction. One method is the threading-based method employed when the identical residues fall below a "safe" region for homology modeling technique; the method uses the structure–sequence alignment strategy and fold assignments procedures for 3D-structure prediction [143]. Another method is the *ab initio* method also known as *de novo* protein modeling method, which is employed when no suitable template for modeling is detected [144]. This method is purely based on physiochemical properties and model structures from scratch rather than relying on previously solved structures as templates [146, 147].

Finally, homology modeling also called comparative modeling or template-based modeling (TBM), is the most widely used method because of its simple implementation [147]. Homology modeling involves a process of modeling a 3D structure of an unknown protein using a known protein structure which has been determined experimentally [using Nuclear magnetic resonance (NMR) /X-ray crystallography] as a template [148, 149]. This method is highly reliant on the sequence similarity with limited errors [144]. It takes into account the biological concept which is based on the fact that during evolution, the structure of a protein is more stable and better conserved compared to the associated sequence [141, 148]. As long as the length of the two sequences and the percentage of identical residues fall in the region marked as "safe" as is shown below in **Figure 1.2**, the two sequences are guaranteed to adopt a similar structure [147].

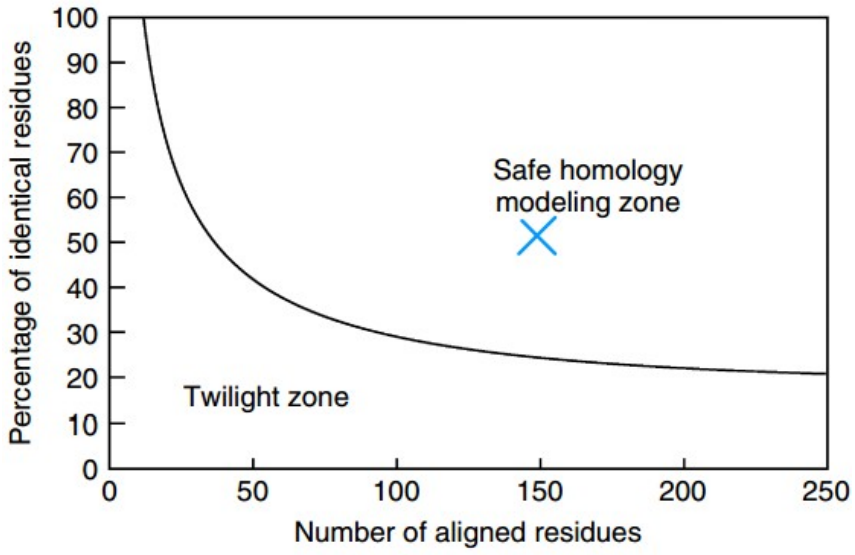


Figure 1.2. The diagram illustrates two zones of sequence alignments, Safe homology modeling and twilight zone.
Two sequences often have similar structures if their sequence length and % sequence identity fall into the safe zone (marked with a cross, **X**). This figure was adapted from Krieger *et al*.[147].

Homology modeling is used in molecular biology, for drug design/discovery [149] and functional annotation [150]. It also provides starting models for solving structures from X-ray crystallography, NMR and electron microscopy [152, 153]. The prediction of a proteins' 3D-structure using computational methods allows researchers to unravel and discover the function of a protein [139]. Therefore, the process of determining the 3D structure of a protein is vital. There are four main steps involved in protein structure homology modeling (**Figure 1.3**); these steps can be reiterated until suitable models are built: Template(s) identification; Template-target alignment; Model construction and Model refinement and verification.

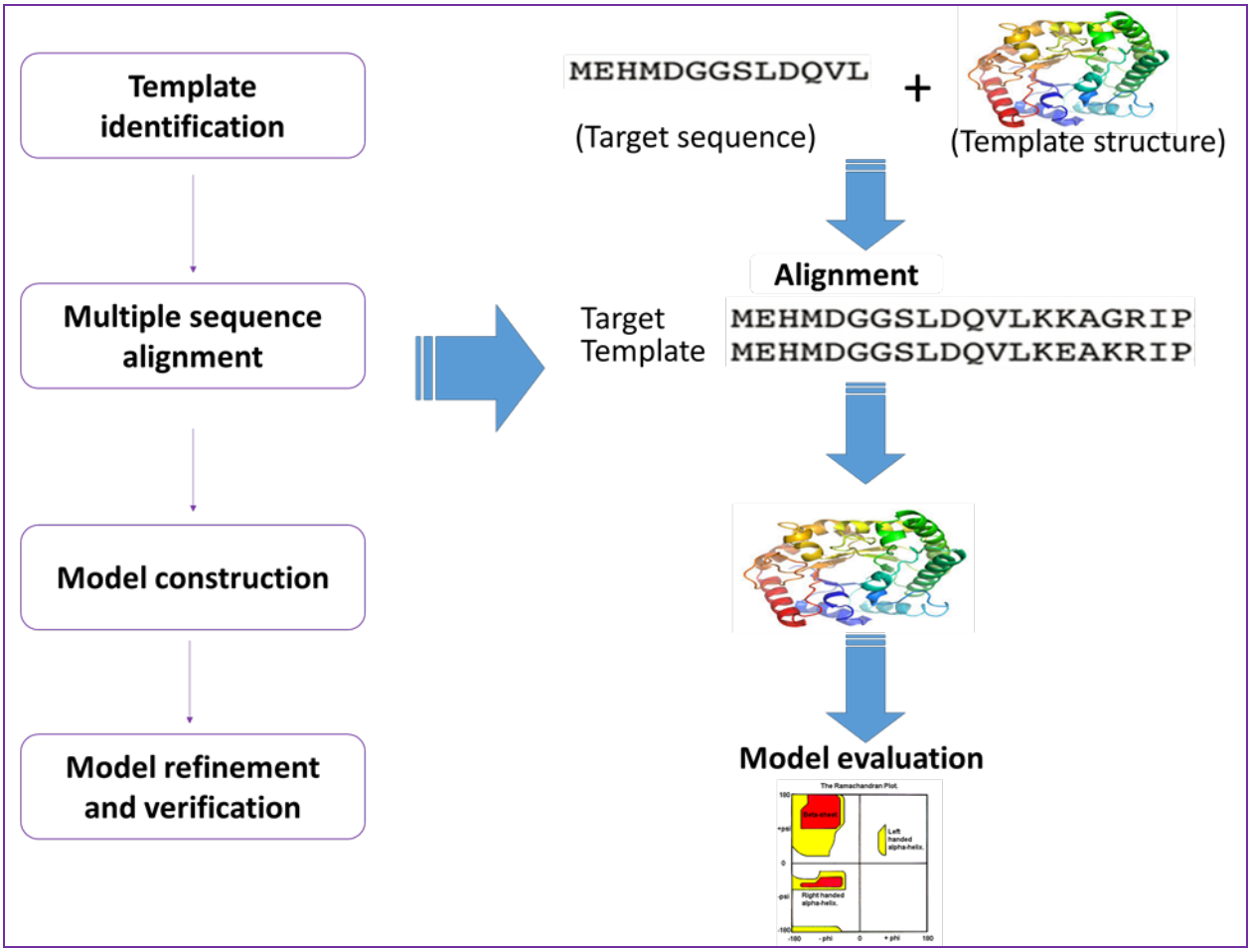


Figure 1.3. Steps involved in protein structure homology modeling.
A template structure is identified based on sequence identity. Subsequently alignment of target sequence with the template sequence is performed. A model is built for the target-template alignment file and the information from the template structure and lastly, the prediction of the model errors or validation of the structure is obtained.

i. Template(s) identification

This is the initial and most important step in homology modeling, in which sequence-similarity search programs compares the target (query) sequence of unknown structure to all known protein structures available in the PDB (**Figure 1.3**) [148, 154]. Fast Alignment (FASTA) and BLAST are the two most widely used servers for template detection. The homology detection prediction (HHpred) interactive server [154] is another popular and efficient tool used for the identification of template(s). The sequence identity shared between a template and the target protein is also important and a foundation for template identification. There are other factors that need to be considered as well, such as, the experimental accuracy and/or reliability of the template structure to be used for model construction.

ii. Template-target/multiple sequence alignment

Multiple sequence alignments (MSA) are crucial in most bioinformatics analyses for comparing homologous sequences [80]. As previously stated, sequence alignment is more reliable compared to just a simple sequence-similarity search. There are numerous programmes that are used for sequence alignment, for example ClustalW, DIALIGN-T, MAFFT, MUSCLE, T-COFFEE, PROBCONS; all the six programmes are said to be the most popular MSA programmes [155] and employ different approaches and/or scoring matrixes for the analysis of sequences. Relying on just one programme may be unreliable; thus, it is advisable to use at least two tools. Most MSA programs employ progressive alignment systems to attain accurate results. However, because of the commonly occurring sequence features such as deletions and insertions, the accuracy of most MSA programs usually get affected.

iii. Model construction

After template-target/multiple sequence alignment, the next step of homology modeling is model construction. In this step, the structure of the query protein is modeled based on the structure of the selected template.

iv. Model refinement and verification

The assessment of the accuracy and reliability of a model is the final step in homology modeling and is based on template selection and alignment accuracy [153]. Most model assessment programs are based on physical energy function and statistically effective energy function [156]. Each and every program use a different approach, hence using more than one programme helps decrease the occurrence of errors in models. Several model assessment methods are available including the ProSA-web (Protein Structure Analysis) (https://prosa.services.came.sbg.ac.at; [157]) ; Ramachandran plots are also generated using tools such as PROCHECK (http://www.ebi.ac.uk/thorntonsrv/software/PROCHECK/; [158]) for model refinement; the superimposition and visualisation of constructed models is usually carried out by using tools such as the MetaMQAP (https://genesilico.pl/toolkit/; [159]).

III.3D-structure prediction programme used in the current study:

SWISS-MODEL(http://swissmodel.expasy.org/;[160]) was used in the current study and is one of the most popular and broadly used web server for the construction of a protein 3D-structure from its amino acid sequence using a homology modeling process. Model quality estimates provided by SWISSMODEL are based on a QMEAN (local composite scoring function) [161]) potential, which is presented as a Z-score. The latter relates the attained scores to values calculated from X-ray high-resolution structures [161] while the overall performance and credibility of SWISS-MODEL is continuously assessed by the CAMEO project (Continuous Automated Model Evaluation). In addition, another quality estimate is given, which combines the GMQE estimate attained from target-template alignment with the QMEAN scoring function. The GMQE estimate is presented by numbers from zero to one; where in high values are indicative of high reliability.

1.3 PROBLEM STATEMENT/HYPOTHESIS

The persistence of *T. parva* in buffalo, in the presence of the tick vector led to the occurrence of a new form of cattle theileriosis, Corridor disease, which was observed after the eradication of ECF in SA. It has since become necessary to differentiate the causative agents of ECF and Corridor disease in SA. This is in order to characterise the *T. parva* parasites circulating in the buffalo population to establish if there are parasites that can adapt in cattle and to eventually cause ECF. Although there has been many studies undertaken, to date, there are no molecular markers identified or available for this purpose. From the recent investigation of transcriptome profiles of two *T. parva* isolates including the cattle-derived and buffalo-derived isolates, 26.28% of genes detected were identified to code for HPs while 54.45% (593) of these formed part of the differentially expressed transcripts (DETs) (KP Sibeko, Department of Veterinary Tropical Diseases, University of Pretoria, unpublished data). The fact that some of the HPs are encoded by DETs, opens the possibility that the encoded proteins are crucial for the disease stage of the parasite (*viz.* the schizont) and in understanding the different disease syndromes resulting from infection with cattle-derived and buffalo-derived parasite isolates. Therefore, it is of importance to functionally characterise these HPs.

Thus, this study hypothesized that these HPs are likely to play a significant role in the disease outcome resulting from infections by different isolates of *T. parva* (cattle-derived and buffalo-derived parasite isolates) and their function may possibly assist in explaining the different disease syndromes resulting from *T. parva* infection.

1.3.1. PROJECT OBJECTIVES

The main aim of the study was to functionally characterise *T. parva* HPs encoded by some of the DETs, previously detected from transcriptome analysis of cattle-derived and buffalo-derived parasite isolates, using various bioinformatics tools.

Specific objectives

i. To classify selected HPs into canonical functional families.
ii. To annotate the functions of these HPs based on sequence homology and orthology.
iii. To infer function of these HPs from the 3D-structures predictions using homology modeling.
iv. To predict function from the physiochemical properties of the HPs.
v. To determine the subcellular localisation of the HPs, for identification of possible putative therapeutic vaccine candidates and drug targets.
vi. To identify secreted proteins for potential disease biomarkers and protein therapeutics.
vii. To detect virulence factors.

CHAPTER 2

2.0 MATERIALS AND METHODS

This study reports a step-wise two-phase analysis for annotating functions of 309 *T. parva* HPs. The computational framework adopted for the functional annotation of the selected HPs is divided into two categories, primary and secondary analyses (**Figure 2.1)**. The Receiver Operating Characteristic (ROC) statistical analysis was employed to evaluate the accuracy of the various bioinformatics tools used in the analysis of the eukaryotic proteins. The bioinformatics tools and databases that were used for sequence-based function annotation are listed below in **Table 2.1**.

Table 1.1. Bioinformatics tools and databases used in the current study for the analysis of protein amino acid sequences.

Software	URL	Remark
Classification into canonical functional families		
Blast2GO	https://www.blast2go.com/	Platform for high-quality protein function prediction and functional analysis of genomic datasets
Sequence comparison		
BLAST: Basic Local Alignment Search Tool	http://www.ncbi.nlm.nih.gov/BLAST/	BLASTp is used for finding similar sequences in protein databases
PSI-BLAST: Position-Specific Iterated- BLAST	blast.ncbi.nlm.nih.gov/Blast.cgi	A tool for the detection of similar sequences
PiroplasmaDB	piroplasmadb.org/	Used for Ortholog identification
Physicochemical characterisation		
ExPASy – ProtParam tool	http://web.expasy.org/protparam/	Computes various physiochemical properties
Virulence factors analysis		
VirulentPred	bioinfo.icgeb.res.in/virulent	Virulent protein prediction tool
MP3	metagenomics.iiserb.ac.in/mp3/	A tool for identification of virulent proteins
VICMpred	www.imtech.res.in/raghava/vicmpred/	Used for prediction of virulence factors
Sub-cellular localisation prediction		
WoLF PSORT	http://www.genscript.com/wolf-psort.html	Subcellular localisation prediction
TargetP	http://www.cbs.dtu.dk/services/TargetP/	Predicts subcellular location
Yloc	abi.inf.uni-tuebingen.de/Services/YLoc/webloc.cgi	A web server for predicting subcellular localisation
MITOPROT	https://ihg.gsf.de/ihg/mitoprot.html	Allows detection of Mitochondrial Targeting Signals
NLS Mapper	nls-mapper.iab.keio.ac.jp/	Used for Nuclear Localisation Signals
Identification of secreted proteins		
SignalP	http:www.cbs.dtu.dk/services/SignalP/	Predicts the presence and location of signal peptide
SecretomeP	http:www.cbs.dtu.dk/services/SecretomeP/	Allows prediction of signal peptides
Detection of membrane proteins		
TMHMM	http://www.cbs.dtu.dk/services/TMHMM/	Predicts membrane topology

HMMTOP	http://www.enzim.hu/hmmtop	Predicts the presence of trans-membrane helices
Detection of GPI- Anchored proteins		
PredGPI	gpcr.biocomp.unibo.it/predgpi/pred.htm	Allows prediction of GPI-anchored proteins
Multiple sequence alignments construction		
T-COFFEE: Tree-based Consistency Objectives Function for Alignment Evaluation	tcoffee.crg.cat/	Is a multiple sequence alignment package
CLUSTAL OMEGA	www.ebi.ac.uk/Tools/msa/clustalo/	A multiple sequence alignment program
Domain identification		
SMART: Simple Modular Architecture Research Tool	http://smart.embl.de/	Allows identification and annotation of protein domains
InterProScan	http://www.ebi.ac.uk/InterProScan/	Searches InterProScan for motif and domain discovery
NCBI-CDD	http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml	Predicts conserved domains
PROSITE	http://www.prosite.expasy.org/	Predicts domains, families and other known functional sites
Pfam	http://pfam.sanger.ac.uk/	Is a database with a large collection of protein families
Protein-protein interaction network		
STRING: Search Tool for the Retrieval of Interacting Genes/Proteins	http://string-db.org	Detects known and predicted Protein-Protein interactions
3D-structure determination		
SWISS-MODEL	https://swissmodel.expasy.org/	Used for 3D-structure prediction

2.1. EXPERIMENTAL DESIGN

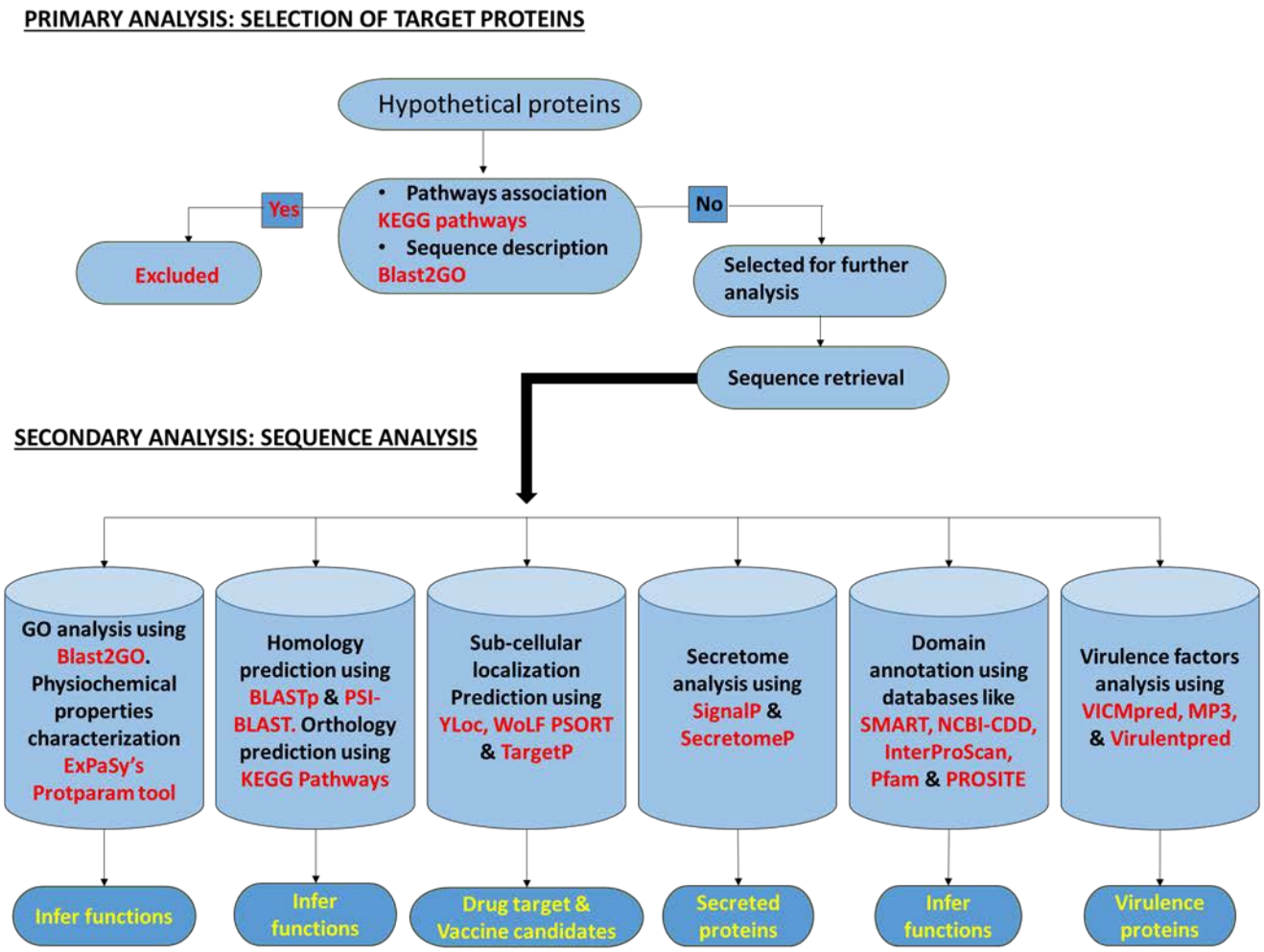


Figure 2.1. The computational framework adopted for the functional annotation of 309 *T. parva* hypothetical proteins.
The analysis was composed of two categories; the primary and secondary analyses. The **primary analysis** include selection of target proteins while **secondary analysis** consist of GO annotation, physicochemical properties characterisation; sequence-similarity search for the identification of homologs and orthologs; prediction of sub-cellular localisation for the identification vaccine candidates and drug targets; prediction of domains, detection of virulence factors and secretome analysis. The latter phase also included ROC analysis, which was carried out to assess the performance of the bioinformatics tools used in the study.

2.2 EXPERIMENTAL PROCEDURES

2.2.1. Selection and retrieval of sequences

In a previous study, the transcriptome profiles of buffalo-derived and cattle-derived *T. parva* strains were investigated, resulting in the identification of 1089 DETs (KP Sibeko-Matjila, Department of Veterinary Tropical Diseases, University of Pretoria, unpublished data). Sequence similarity analysis using an nr database showed that 593 of these transcripts encode HPs. For selection of the proteins of interest, the HPs were analyzed on Blast2GO [133] and KEGG (Kyoto Encyclopedia of Genes and Genomes) [162], in order to exclude proteins with sequence descriptions and pathway association. The sequences of retained HPs which could not be assigned sequence descriptions and are not associated with any metabolic pathways (n = 309) were retrieved in FASTA format from the Universal Protein Resource (UniProt) [163] database to allow for further analysis. Unless specified, note that all the tools used in subsequent analysis employed input files of amino acid sequences in a FASTA format.

2.2.2. Classification of proteins into canonical functional families

Classification of proteins into functional families is usually helpful in better understanding the functional roles of a large set of unknown proteins [27]. Furthermore, it provides important clues about activity, structure, and metabolic role of uncharacterised proteins [30]. Blast2GO [27], a bioinformatics platform for high-quality protein function prediction and functional analysis of genomic datasets, has a suitable framework to allow for classification of proteins into their canonical functional families.

Thus, sequences of each of the selected HPs were used as input onto Blast2GO and a search for the gene ontology (GO) terms was then performed using default parameters; E-value-Hit-Filter of 1.0E-6, annotation cut-off of 55, GO weight of 5 and cut-off of zero. Blast2GO produced a GO distribution of the query proteins based on the biological processes (BP) they are involved in, their cellular compartment (CC) and molecular functions (MF). This information was then used to classify the HPs into their appropriate functional families.

2.2.3. Characterisation of the physicochemical properties

The ExPASy's ProtParam tool, used in the computation of protein's physical and chemical parameters [138], was used for identification of the Isoelectric point, aliphatic index, instability index, grand average of hydropathicity (GRAVY) and extinction coefficient of each protein. Sequences of the HPs were used as input to the ExPASy's ProtParam server, which produced an output giving a list of all the properties studied for each query protein.

2.2.4. Sub-cellular localisation prediction

The localisation of a protein in a cell is very vital for the prediction of protein function, genome annotation and drug discovery [117, 164]. Membrane-associated proteins may act as possible vaccine candidates, while cytoplasmic proteins have previously been considered as potent drug targets [117]. WoLF PSORT [113], YLoc [114] and TargetP [115] were used for predicting the sub-cellular localisation of the HPs. Using WolF PSORT, the HPs' amino acid sequences were submitted for the analysis. The server generated a list of all the possible sites with scores and nearest neighbors; the sub-cellular site with the highest score was considered to be the probable site of the target protein. Sub-cellular localisation prediction by YLoc requires the use of a single raw protein sequence or multiple sequences for analysis. The output of YLoc prediction is displayed in a table showing the predicted location, the probability of the predicted location and the YLoc confidence score which indicates if the prediction is reliable or not. The confidence score ranges between zero and one, where, a large value indicates a high confidence that the specific prediction is reliable.

Analysis using TargetP was performed using default parameters. The possible location of each protein was accompanied by a Reliability class (RC), which indicates the prediction accuracy of TargetP [115]. The RC is represented by numbers between one and five, where one is indicative of a strongest/reliable prediction and five a less reliable prediction; thus, the lower the RC value, the more reliable the prediction and *vise versa*. If the protein sequence does not contain any trans-membrane domains, TargetP localisations prediction output will be "other"(-), meaning that the protein could either be nuclear, cytoplasmic, or peroxisomal. Unfortunately, the current version of TargetP is unable to differentiate between the N-terminal sorting signals of these three localisations to provide specific predictions.

2.2.4.1. Confirmation of the nuclear, mitochondrial and membrane-associated proteins

Proteins associated with the mitochondria usually have an N-terminal pre-sequence known as the Mitochondrial-targeting signal (MTS) that transport them to the mitochondria. This pre-sequence is composed of charged and hydrophobic amino acid residues forming an amphipathic helix that is responsible for the transportation of the protein to the mitochondrion [165]. Similarly, for proteins to be targeted to the nucleus they require a nuclear localisation signal (NLS; ~6 to 20 amino acids long), which is usually rich in lysine and arginine. Accordingly, to predict proteins likely to be localized in the mitochondria and cell nucleus, the 'localisation-specific signals' were detected using MitoProt [166] and cNLS Mapper [167], respectively.

Mitochondrial proteins: To predict MTS, a raw amino acid sequence of each HP was used as an input to MitoProt. MitoProt calculates the N-terminal region of a protein that supports MTS and searches for the occurrence of two close aspartic acid (Asp) or glutamic acid (Glu) within a determined distance. Together, these factors are used to determine the probability of the protein to be exported to the mitochondria.

Nuclear proteins: For detection of the NLS, the sequence of each HP was used as input to cNLS Mapper, wherein a cut-off score of 5.0 was applied. Higher scores display stronger NLS activities. cNLS mapper outputs the query sequence with the probable NLS marked in red.

Membrane-associated proteins: The presence of trans-membrane domains/helices and a glycosyl-phosphatidylinositol (GPI) anchor were detected to predict localisation on the cell membrane. A membrane protein constitutes at least one membrane-spanning domain; these could either be α helices or multiple β strands [168]. These specific protein sequence features are used by several bioinformatics tools to allow for the identification of membrane associated proteins, for example TMHMM [118] and HMMTOP [119] used in this study. Both tools predict trans-membrane helices based on a hidden Markov model.

GPI-Anchor detection: Most GPI-anchored proteins migrate to the plasma membrane, though some have been shown to reside in different compartments [169]. GPI-anchored proteins were predicted using PredGPI. The amino acid sequence of each HP was used for the analysis. For every query protein, the prediction system provided the likelihood of the

presence of the GPI-Anchor as a *Specificity* index. When this index is above 99.9%, the presence of the GPI-anchor is "highly probable"; between 99.9% and 99.5%, it is said to be "probable", and when it ranges from 99.5% to 99.0%, the presence of the GPI-anchor is considered "lowly probable".

2.2.5. Identification of secreted proteins

In eukaryotes, secreted proteins play crucial biological regulatory roles, and are also considered to be potential disease biomarkers and protein therapeutics [170]; additionally, knowledge of their functional roles is key to genome annotation. Secreted proteins constitute signal peptides responsible for their transport to the classical endoplasmic reticulum (ER)/Golgi-dependent secretion pathway [171]. These signal peptides comprise short chains of amino acids which are frequently cleaved by signal peptidases after export to the specific subcellular compartments [172].

For identification of the secreted proteins, sequences of each of the HPs were used for analysis in two programmes; SecretomeP [173]and SignalP 4.1 [174]. Using SecretomeP, a threshold of 0.6 was applied to filter redundant results. Usually a value above 0.5 suggests possible secretion, however, a threshold of 0.6 is recommended for eukaryotic sequences. SecretomeP is trained to predict non-classical secretion, yet it also provides high score to proteins secreted through the classical secretory pathway. For SignalP, a cut-off value of 0.450 was applied, and the scores generated by SignalP for each input sequence were used to filter redundant results. Secreted protein without signal peptide are often secreted outside of the plasma membrane via the non-classical secretory pathways [175] and these can also be predicted using the above stated subcellular prediction tools, YLoc and TargetP as well as SecretomeP.

2.2.6. Sequence comparisons

2.2.6.1. Detection of homologs of the hypothetical proteins:

Homologous proteins are likely to have a common biochemical function in the course of evolution [176], therefore, this information may allow functional annotation of uncharacterised proteins. For the identification of known functional homologs, each HP sequence was used as a query on the National Center for Biotechnology Information (NCBI) through BLASTp [134] and PSI-BLAST [83]. PSI-BLAST provides more iterations; hence, it can detect distantly related proteins that may have been missed by BLASTp [83]. Hits obtained were filtered based on the E-value (<0.005), sequence identity (≥25%) and sequence coverage (≥50%). Normally, proteins that meet the criteria are considered to be close homologs of the protein of interest, while those with low identities (≤25%) are considered to be distant homologues. In this study, proteins with <25% sequence identity and ≤50% coverage were considered non-homologous.

For confirmation of homology and to provide more support to putative functional annotation, a sequence similarity search was performed on BLASTp using homologs of the *T. parva* HPs as query sequences. This was done in order to see if the identified "putative homologs" would retrieve the corresponding HPs as its homologs. Subsequently, the "putative homolog" and its corresponding HPs were analysed to check if they possess common domains. This step was followed by the submission of the homolog's accession number or sequence description to a database known as HomoloGene wherein the family of each homolog was obtained. However, most of the proteins accession number or sequence description gave an error. As a result, some protein families were not obtained and a pair-wise sequence alignment of the HPS and their homolog was then carried out.

A pair-wise sequence alignment of the query sequences (HP) and their detected homologs was performed using CLUSTAL OMEGA [80] and T-COFFEE [177]. Each of the programs uses a specific algorithm or scoring matrices to deduce the different interrelations between sequences thus eliminating biasness that may result from using one program. In both programs, the input file consisted of the protein sequences of both the target and its homolog. Ultimately, the functional information generated from the homolog was then transferred to the HP or used to assign a probable function of the HP.

For proteins that resulted in good alignments with their homologs but had different domains, a pair-wise sequence alignment of solely the residues found within the domain region was carried out for reliable functional annotations. This was done in order to identify the level of conservation between the residues spanning the domain region across the two proteins (the HP and its corresponding homolog), as these residues may be important for protein function.

2.2.6.2. Detection of orthologs of the hypothetical proteins

Orthology identification has been broadly used for function assignment of uncharacterised proteins [87–89]. Orthologs are proteins that have originated from a common ancestor separated by speciation. These groups of proteins tend to retain a common function over evolutionary time, thus, making their identification a potent tool for functional annotation [84]. To identify known functional orthologs, each HP was used as a query on PiroplasmaDB [90]. Proteins with a sequence identity of ≥25%, coverage of ≥50% and an E-value of <0.005 were considered to be orthologous to the HP.

2.2.7. Predictions of sequence domains

Domains are structurally and/or functionally important regions in proteins; they are responsible for a particular function or interaction, which contributes to the overall function of a protein. Identification of such regions may provide significant clues about the functional role of uncharacterised proteins. The HPs were analysed for presence of domains using tools such as SMART [98], InterProScan [99], NCBI-CDD [103] PROSITE [100] and Pfam [104]. The results were provided as a graphic display which indicated the domain present in the query sequence plus its length.

2.2.8. Detection of virulence factors

The identification of virulence factors (VFs) is required to better understand the mechanism of pathogenesis and identify possible therapeutic targets for combating diseases [65]. For identification of the VFs, VirulentPred [178], VICMpred [179] and MP3 [180] were used. VirulentPred and VICMpred are both SVM (Support Vector Machines)-based methods used to predict VFs from protein sequences with an accuracy level of 70.75% and 81.8%, respectively; while MP3 uses an integrated SVM-HMM (Hidden Markov Model) approach.

Using VirulentPred, the search was performed with the cascade SVM module as a prediction approach for the input sequence and a threshold value of 0.4 was considered. For VICMpred, the patterns based module was used as the prediction method of the HP query sequence. VICMpred produced a list of the functional classes; metabolism molecule, cellular process, information and storage, and virulence factors. Each functional class is assigned a score and the one with the highest score was taken as the possible functional class of the target protein. Using the MP3 software, the SVM based approach was also employed as a prediction approach and a threshold of -0.2 was considered and the outcome presented as “pathogenic” or “non-pathogenic”.

2.2.9. Protein-protein interaction network

Proteins usually interact with one another in order to perform a common function. Thus, the prediction of interaction partners can be used to annotate functions of novel proteins. Here, the STRING database [129] was used to predict protein interaction partners of the *T. parva* HPs. This database incorporates interaction data from different sources such as co-expression, genomic context and high-throughput experiments to predict the possible functional partners of the protein of interest.

From the analysis, each protein-protein interaction was annotated with a score; these scores are indicators of confidence of how likely the database judges an interaction to be true. The scores are used to see if the predictions are of high confidence or not. Scores rank from 0 to 1, wherein 1 signifies the highest possible prediction confidence. The output of the analysis is presented as a diagram that indicates the association between the predicted functional partners.

2.2.10. Prediction of 3D-structures of the HPs

SWISS-MODEL web tool was employed in the current study for the construction of the HPs 3D-structures from its amino acid sequence. SWISS-MODEL provided model quality estimates which are presented as a Z-score. Another quality estimate was given, which combines the GMQE estimate attained from target-template alignment with the QMEAN

scoring function. The GMQE estimate is presented by numbers from zero and one, wherein high values are indicative of high reliability. These quality estimates were used to select the suitable template for the construction of the 3D-structures.

2.2.11. Performance assessment

The ROC statistical analysis was used to evaluate the accuracy of the prediction methods employed in this study. Hundred (100) proteins sequences with known function were selected for this analysis and the diagnostics efficiency was assessed at six levels. The true positive was classified as 1 and the true negative prediction as 0. The adopted confidence ratings were defined as 1, 2, 3, 4, 5 and 6. The web-based calculator for ROC curves was used for ROC analysis, this online software calculates the ROC based on the submitted data and the results are presented as sensitivity, accuracy, specificity and the ROC area.

CHAPTER 3

3.0. RESULTS

3.1. SELECTION OF HPS

Initially, 593 HPs were selected for this study; however, after the HPs were analysed using Blast2GO and KEGG to identify proteins with sequence descriptions and pathway association, only 309 HPs were retained for further analysis as these had no sequence descriptions and/or metabolic pathway association.

3.2. PREDICTED FUNCTIONAL FAMILIES

By using a variety of available bioinformatics tools, 309 HP sequences were extensively analysed in order to predict their possible functions. Based on the GO terms obtained from annotation with Blast2GO, 43 HPs (**Figure 3.1 and Table 3.1**) were characterised into seven canonical functional families and the majority of these were binding proteins (n = 21) followed by proteins with transcription factor (n = 11) and catalytic (n = 5) activities.

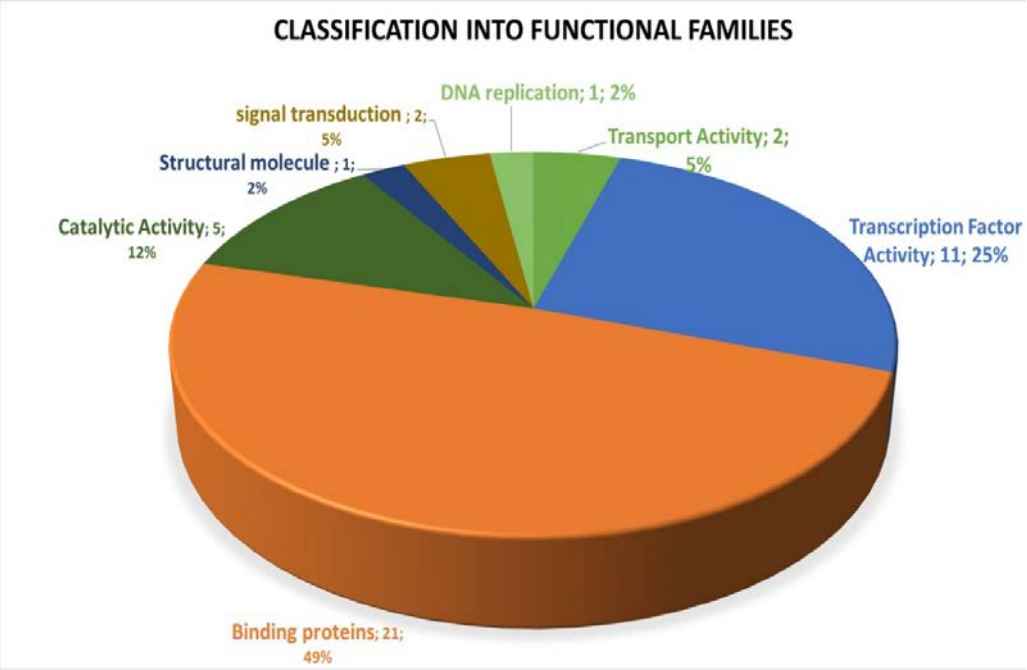


Figure 3.1.Classification of 43, of the 309 *T. parva* HPs selected for investigation, into their canonical functional families from gene ontology analysis.

Table 3.1. Hypothetical proteins (n = 43) classified into their functional families.

Sequence. No.	HP gene name (protein length)	*GO Term	Domain(s)	Homology		Probable function(s)
				Higher eukaryotes	Lower eukaryotes	
1.	TP03_0620 (452)	C:mediator complex; F:RNA polymerase II transcription cofactor activity; P:regulation of transcription from RNA polymerase II promoter	Smc (Chromosome segregation ATPase [Cell cycle control, cell division, chromosome partitioning (309-390); Trans-membrane (415–437); Coiled coil (360–380)	None	*Toxoplasma_gondii_ME49* : TGME49_244210; RNA polymerase-associated protein RTF1 (25%)	RNA Polymerase II-associated protein
2.	TP03_0742 (273)	C:cytoplasm; C:integral component of membrane; F:protein transporter activity; F:transmembrane transporter activity; P:regulation of transcription, DNA-templated; P:intracellular protein transport; P:transmembrane transport	Lectin_N (Hepatic lectin, N-terminal domain;18-100); CALCOCO1 (Calcium binding and coiled-coil domain (24-102); AIP3 (Actin interacting protein 3;26-124); Exonuc_VII_L (Exonuclease VII, large subunit;26-104); Tropomyosin_1(70-139); Mplasa_alph_rch(68-142); Smc (65-142); ATG16 (67-137)	None	*Toxoplasma_gondii_ME49* : TGME49_286440; Malic Enzyme (37%)	Protein transporter
3.	TP02_0860 (246)	C:endoplasmic reticulum membrane; P:GPI anchor biosynthetic process; C:integral component of membrane	GPI biosynthesis protein family Pig-F (51-234); DUF261 (87-131)	None	*Theileria equi_strain_WA*:BEWA_035950, conserved hypothetical protein (34%)	GPI biosynthesis protein family Pig-F
4.	TP01_0817 (409)	C:integral component of membrane	TauE (11-148, 203-397); Trans-membrane (7-32, 52-74, 99-119, 131– 152, 389-408)	None	*Neospora caninum*: NCLIV_012630, Os03g0726500 Protein, related (34%)	Os03g0726500 protein, related
5.	TP04_0869 (283)	C:intracellular; P:intracellular signal transduction	C1_1 (232-264); C1_2 (233-263); Zn_ribbon_17 (234-277); Double Zinc Ribbon (234-254); Protein kinase C-like, phorbol	None	*Theileria equi_strain_WA*: BEWA_013490, DEAD box ATP-dependent RNA	RNA helicase family member protein

			ester/diacylglycerol-binding domain (220 -277)		helicase family member protein (31%)	
6.	TP02_0150 (385)	C:mRNA cleavage and polyadenylation specificity factor complex; P:mRNA polyadenylation; P:mRNA cleavage	Cleavage and polyadenylation factor 2 C-terminal (294-376)	None	*Babesia bigemina*: BBBOND_0300290, Brix domain containing protein, putative (34%)	Cleavage and polyadenylation factor
7.	TP04_0069 (228)	C:NELF complex; P:negative regulation of transcription elongation from RNA polymerase II promoter	Negative elongation factor A (30-149)	None	*Toxoplasma_gondii_ME49* : TGME49_246070, SAG-related sequence SRS56 (38%)	SAG-related sequence SRS56
8.	TP01_0391 (680)	C:nucleus; F:DNA binding; P:transcription, DNA-templated	Transcription factor IIS, N-terminal (191 - 304); Med26, Mediator of RNA polymerase II transcription subunit 26 (227-289)	None	*Theileria equi_strain_WA:* BEWA_030570, hypothetical protein (45%)	Mediator of RNA polymerase II transcription subunit 26 (Med26)
9.	TP02_0812 (576)	F:calcium ion binding	EF-hand domain pair (476 -563)	None	*Neospora_caninum_Liver pool*: NCLIV_017560, Calcium-dependent protein kinase 2, related (28%)	Calcium-dependent protein
10.	TP01_0564 (162)	F:calcium ion binding	Peptidase_M54(Peptidase family M54;18-92); EF-hand domain pair (11-160)	None	*Neospora_caninum_Liver pool*: NCLIV_063600, Calmodulin, putative (38%)	Calmodulin
11.	TP03_0475 (459)	F:double-stranded RNA binding	TONB_DEPENDENT_REC_1 (1-53)	None	*Theileria equi_strain_WA*:BEWA_005200, elongation Factor ts, putative (26%)	Elongation Factor
12.	TP02_0267 (413)	F:GTP binding	Trans-membrane(210-228, 248-	None	*Neospora_caninum_Liver pool:* NCLIV_031590,	GTP binding

		C:integral component of membrane	266, 272-291, 300-323, 367-388); Cytoplasmic_domain(229-247, 292-299, 389-413); Non_cytoplasmic_domain(1-209, 267-271, 324-366); DUF2207 (159-293); DUF2981 (53-118)		conserved hypothetical protein (27%)	
13.	TP02_0880 (187)	F:GTP binding; P:small GTPase mediated signal transduction; F:nucleotide binding; C:intracellular, integral component of membrane	Trans-membrane, helical (104 – 125)); Cytoplasmic_domain (1-103); Non_cytoplasmic_domain (126-187)	None	*Babesia bigemina:* BBBOND_0300940, hypothetical protein, conserved (36%)	Guanylate-binding protein,
14.	TP02_0512 (1336)	F:GTPase activity; F:GTP binding; P:metabolic process	Guanylate-binding protein, N-terminal (121 -222); P-loop containing nucleoside triphosphate hydrolase (104 -163, 197-228, 276-430); Trans-membrane region (7-29, 1195-1215, 1222-1241, 1289-1308)	None	*Toxoplasma_gondii_ME49* :TGME49_304990, guanylate-binding protein, N-terminal domain-containing protein (29%)	Guanylate-binding protein,P-loop_NTPase domain-containing protein
15.	TP02_0916 (1070)	F:GTPase activity; F:GTP binding; P:metabolic process	Guanylate-binding protein, N-terminal(225-353); SieB (21-63); MMR_HSR1(246-321); Flavi_NS4B(282-355); ATPase (646-720); RNA_lig_T4_1(655-732); Coiled coil (658-697)	None	*Theileria equi_strain_WA* :BEWA_034930, conserved hypothetical protein 38%)	Guanylate-binding protein,
16.	TP03_0055 (557)	F:hydrogen ion transmembrane transporter activity; P:ATP synthesis coupled proton transport; C:Mitochondrionl proton-transporting ATP synthase complex, coupling factor F(o)	Coiled coil domain (305 – 325);ATP5H, ATP synthase D chain, Mitochondrial (25-95); HAND (494-550)	None	*Theileria equi_strain_WA* :BEWA_016560, Conserved hypothetical protein (53%)	ATP5H, ATP synthase D chain, Mitochondrial

17.	TP03_0329 (317)	F:iron-sulfur cluster binding	Hanta_G2 (35-131); Coil (121-162)	None	*Neospora_caninum_Liver pool:* NCLIV_043230, putative DNA double-strand break repair rad50 ATPase (37%)	DNA double-strand break repair rad50 ATPase
18.	TP01_0890 (187)	F:metal ion binding	Zinc finger, CCCH-type (61-93)	None	*Toxoplasma_gondii_ME49* : TGME49_294785, zinc finger (CCCH type) motif-containing protein (34%)	Zinc finger (CCCH type) motif-containing protein
19.	TP04_0121 (114)	F:metal ion binding; F:zinc ion binding; F:DNA binding	None	None	*Theileria equi_strain_WA*: BEWA_003200 hypothetical protein (44%)	Metal ion, DNA binding
20.	TP04_0353 (477)	F:methyltransferase activity; P:methylation	Fungal tRNA ligase phosphodiesterase domain(133-184); LigT_Pease(138-165); TetR_C_10 (Tetracycline repressor, C-terminal all-alpha domain ,403-466)	None	*Theileria equi_strain_WA:* BEWA_006230, hypothetical protein (29%)	Methyltransferase
21.	TP04_0715 (212)	F:protein binding	GYF domain (21-78); Coiled coil(162 -196)	None	*Toxoplasma_gondii_ME49* : TGME49_224610, GYF domain-containing protein (45%)	GYF domain-containing protein
22.	TP01_0305 (78)	F:RNA polymerase II transcription cofactor activity; C:mediator complex; P:regulation of transcription from RNA polymerase II promoter	None	None	*Theileria equi_strain_WA* :BEWA_031410 hypothetical protein (58%)	Transcription regulator
23.	TP03_0169 (239)	F:transcription factor activity, sequence-specific DNA binding; F:sequence-specific DNA binding; P:regulation of transcription, DNA-templated	Tropomyosin_1 (173-234); TPR_MLP1_2 (175-215); DivIC (189-218); bZIP_1 (Basic leucine zipper 1, 189-223); RasGAP_C (187-227); DNA_topoisoIV (117-	None	*Theileria equi_strain_WA* :BEWA_049610, hypothetical protein (31%)	Regulation of transcription

			221)			
24.	TP02_0687 (323)	F:transferase activity, transferring alkyl or aryl (other than methyl) groups; P:metabolic process	Cytoplasmic_domain (233-323); Non_cytoplasmic_domain (33-208); Trans-membrane (209-232)	None	*Theileria equi_strain_WA:* BEWA_014360, signal peptide containing protein (40%)	Transferase
25.	TP03_0658 (165)	F:zinc ion binding	Zinc finger, RanBP2-type (133 - 162)	None	*Neospora_caninum_LivFerpool* :NCLIV_042280,zinc-finger-Ran binding domain-containing protein, related (40%)	Zinc-finger-Ran binding domain-containing protein, related
26.	TP03_0132 (112)	F:zinc ion binding	None	None	*Babesia bigemina:*BBBOND_0207010,hypothetical protein, conserved (27%)	Zinc ion binding protein
27.	TP04_0254 (220)	P:cell redox homeostasis; C:cell; C:integral component of membrane	Thioredoxin-like fold (50-136)	None	*Toxoplasma_gondii_ME49* : TGME49_202440, hypothetical protein (29%)	Thioredoxin domain containing protein
28.	TP02_0776 (2126)	P:DNA replication; C:origin recognition complex; C:nucleus	ORC5_C (origin recognition complex, 1851-2050);Salp15 (1910-1984); Coil (1031-1051, 1261-1281); Trans-membrane (2081-2099); Cytoplasmic_domain (2100 – 2126); Non_cytoplasmic_domain (1-2080) termin_org_DnaJ (1635-1899)	None	*Theileria equi_strain_WA* :BEWA_026970, hypothetical Protein (33%)	ATP-dependent DNA-replication protein
29.	TP03_0729 (379)	P:regulation of transcription, DNA-templated; F:transcription factor activity, sequence-specific DNA binding	CBFB_NFYA (CCAAT-binding transcription factor, 21-45)	*Theileria equi_strain_WA:* BEWA_013670, DNA primase large subunit, putative	None	Transcription factor

				(27%)		
30.	TP03_0557 (447)	P:transcription initiation from RNA polymerase II promoter	Ribonuclease R winged-helix domain (191-217); TFIIE beta subunit core domain(193-253)	*Neospora_caninum_ Liverpool*: NCLIV_021590, hypothetical Protein (33%)	None	Initiation of transcription
31.	TP01_0891 (150)	C:mediator complex; F:RNA polymerase II transcription cofactor activity; P:regulation of transcription from RNA polymerase II promoter	Mediator complex, subunit Med7 (7-132)	None	*Toxoplasma_gondii_ME49 :* TGME49_306280, mediator Complex subunit MED7 (MED7) (26%)	Mediator of RNA polymerase II transcription subunit 7 (Med7)
32.	TP01_0636 (151)	F:nucleic acid binding; F:nucleotide binding	RRM1_RRT5 (RNA recognition motif,85-147); DUF1764 (11-99)	None	*Theileria equi_strain_WA:* BEWA_045820, hypothetical protein (40%)	RNA recognition motif-containing protein
33.	TP02_0651 (228)	C:ribosome; F:structural constituent of ribosome; P:translation	Ribosomal protein L27(72-150)	None	*Babesia bigemina*: BBBOND_0103370, hypothetical protein, conserved (71%)	Ribosomal protein L27
34.	TP02_0213 (364)	F:DNA binding	DUF529 (148-213);Coil (254 - 278); Non_cytoplasmic_domain (23-364); Atg31 (267-325)	None	*Theileria equi_strain_WA:* BEWA_053980, signal peptide containing protein (26%)	DNA binding
35.	TP03_0523 (238)	F:metal ion binding	Zinc finger, RING/FYVE/PHD-type (44- 104)	None	*Toxoplasma_gondii_ME49* : TGME49_237870, FYVE zinc finger domain-containing protein (28%)	FYVE zinc finger domain-containing protein
36.	TP03_0323 (178)	P:protein folding; F:Chaperone binding	None	*Saccharomyces Cereviae*, Tetratricopeptide repeat protein,	None	Chaperone binding

				putative (35%)		
37.	TP03_0034 (201)	C:nucleus	DUF3545, Protein of unknown function (147-181)	None	*Babesia bigemina*: BBBOND_0311100, hypothetical protein, conserved (35%)	**NFP**
38.	TP03_0018 (98)	P:transcription, DNA-templated; F:DNA-directed RNA polymerase activity; F:DNA binding	None	None	*Theileria equi_strain_WA:* BEWA_017160, Conserved hypothetical protein (44%)	DNA binding
39.	TP03_0825 (446)	P:regulation of transcription, DNA-templated; F:transcription factor activity, sequence-specific DNA binding	AP2 domain(12-45)	None	*Theileria equi_strain_WA:* BEWA_051710, conserved hypothetical protein (46%)	Regulation of transcription
40.	TP02_0428 (243)	F:DNA binding	None	None	*Theileria equi_strain_WA* :BEWA_021670,hypothetical protein, (293) (31%)	DNA binding
41.	TP04_0200 (197)	C:integral component of membrane	None	None	*Theileria equi_strain_WA* :BEWA_004700 , hypothetical protein (25%)	**NFP**
42.	TP01_1123 (180)	P:cellular metabolic process	None	None	None	**NFP**
43.	TP02_0910 (181)	F:GTP binding; F:GTPase activity; P:metabolic process	DNA replication terminus site-binding protein (Ter protein;101-138); WHEP-TRS domain (111-137)	None	None	GTP binding

NFP = **No function predicted**
GO terms: **C = Cellular Compartment, BP = Biological Process, MF = Molecular Function**

3.3. PHYSIOCHEMICAL PROPERTIES CHARACTERISATION

The ExPASy's ProtParam server was utilised to predict the theoretical physiochemical properties of the 309 HPs, including the isoelectric point (pI), extinction coeffient (EC), instability index (II), aliphatic index (AI) and grand average of hydropathicity (GRAVY) index. A summary of results obtained from this analysis is provided below in **Table 3.2**. Isoelectric point (pI) refers to the pH at which a particular molecule carries no net electrical charge. One hundred sixty-two (162) HPs were shown to have pI <7 indicating negatively charged proteins, while, 146 HPs had pI>7, indicating positively charged proteins. Only one HP (TP02_0583) had a pI value of 7.00, which indicates that the HP is a neutral protein. Experimental protein stability was estimated by the instability index (II). According to literature [138], a protein whose II is <40 is predicted to be stable, while proteins with an II >40 are classified as unstable. This information is important for storing proteins in the suitable solvent [137]. The II values of the investigated HPs ranged from 14.01 to 80.24, whereby 164 (53%) HPs had an II >40, while 145 (47%) had an II <40, indicating unstable and stable proteins, respectively.

Another protein property that was investigated was the GRAVY index; where a negative GRAVY index infers a soluble and hydrophilic protein, while a positive GRAVY index suggests insoluble and hydrophobic proteins. GRAVY indeces obtained in this study ranged from -0.708 to -1.930; 279 (90%) of HPs had negative GRAVY values, indicating that the majority of HPs investigated were hydrophilic (globular) and soluble. The remainder (n = 30) displayed positive GRAVY values suggesting that these HPs were hydrophobic (membrane) and insoluble proteins.

The EC was also determined to measure how much light each of the HPs absorbs at a given wavelength with respect to the concentration of cysteine (Cys), tryptophan (Trp) and tyrosine (Tyr). A high extinction coefficient indicates a high Cys, Trp and Tyr concentration in the protein. Extinction coefficient values of the HPs at 280 nm ranged from 1490 to 212060, and revealed that the majority (n = 301) of the HPs had high concentrations of Cys, Trp, and Tyr. The estimate of this coefficient is useful in protein purification.

Finally, an aliphatic index of the proteins was investigated; this is the relative volume occupied by aliphatic side chain and is used to predict or measure the thermo-stability of proteins [138]. A high aliphatic index indicates that, a particular protein is thermo-stable over a wide temperature range [181]. The aliphatic index of the HPs involved in this study ranged from 29.62 to 143.53,

which indicated that a large number (n = 300) of the HPs were thermo-stable. The details of the predicted properties for all the 309 HPs are provided in **Appendix A.1.**

Table 3.2. Predicted physicochemical properties of the 309 *T. parva* HPs determined by Expasy's ProtParam.

Physiochemical properties		**Number of HPs**
Isoelectric point (pI)	*Positive*	146
	Negative	162
	Neutral	1
Extinction co-efficient (M-1 cm-1)	*High Cys, Trp & Tyr concentration*	301
	Low Cys, Trp & Tyr concentration	8
Aliphatic Index (II)	*Thermo-stable*	300
	Non-thermo-stable	9
Instability Index(II)	*Stable*	145
	Unstable	164
Grand Average of Hydropathicity (GRAVY)	*Hydrophilic & soluble*	279
	Hydrophobic & insoluble	30

3.4. SUBCELLULAR LOCALISATION

The information obtained from the protein sub-cellular localisation prediction can be used to infer protein cellular functions and to find novel vaccine or drug targets [61]. The subcellular localisation of the *T. parva* HPs was predicted using three online servers; WoLF PSORT, YLoc, and TargetP. The majority of the HPs (n = 111; 36%) were predicted to localize in the cytoplasm, followed by those localized in the mitochondrion (n = 26; 8%), nucleus (n = 42; 13%) and on the plasma membrane (n = 5; 2%) (**Figure 3.2; Appendix A.2**). Some proteins can move across locations and localize to multiple subcellular compartments [182]; in this study, a set of such proteins (n = 36; 12%) was also detected. This analysis also revealed that a significant number of HPs (n = 34; 13%) were secreted proteins.

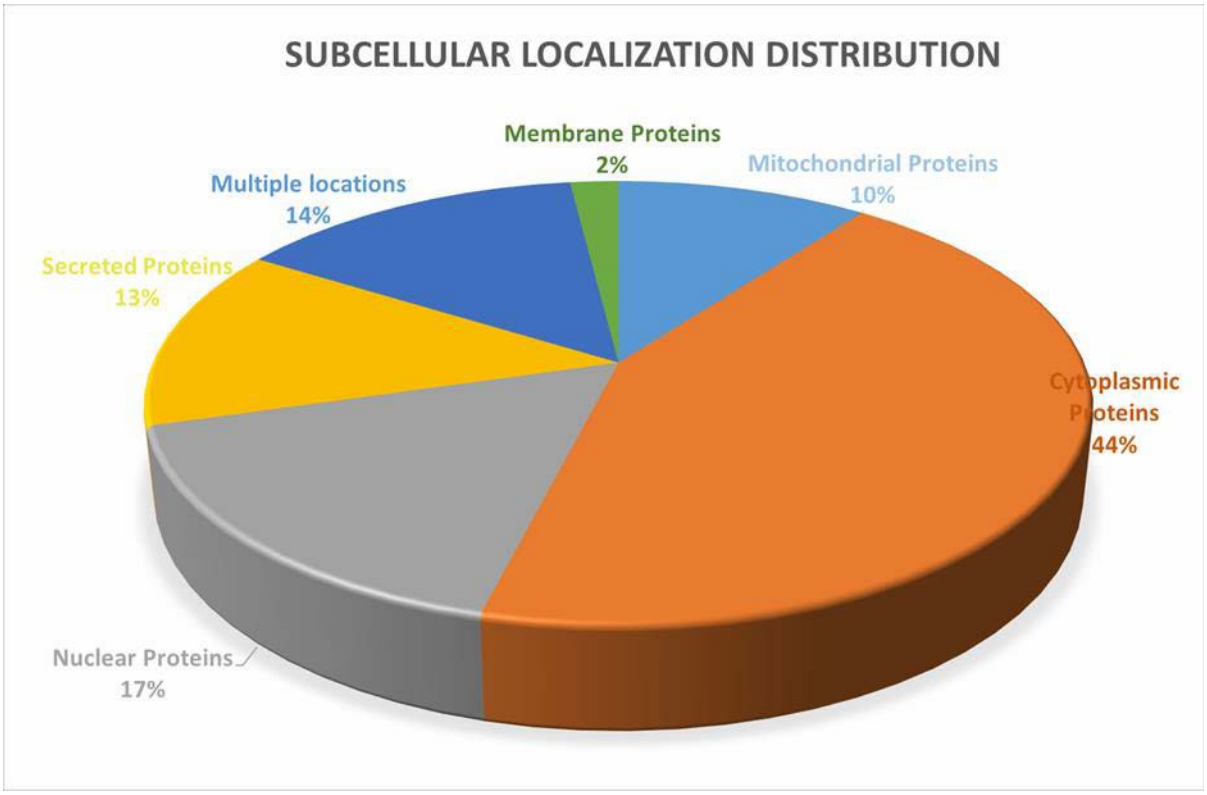


Figure 3.2. Subcellular localisation analysis of the 309 *T. parva* hypothetical proteins.

3.4.1. Confirmation of the nuclear, mitochondrial and membrane-associated proteins

3.4.1.1. *Nuclear proteins*

Proteins are usually imported into the nucleus by a nuclear localisation signal (NLS); thus, the protein amino acid sequences were further analysed using the cNLS Mapper prediction program to detect presence of these signals. Proteins predicted to localise in the nucleus by at least two of the used subcellular localisation tools (n = 42), were further analysed for the detection of NLS and the analysis revealed the presence of NLS in 14 HPs.

3.4.1.2. *Mitochondrial proteins*

The importation of a protein into the mitochondrion is facilitated by MTS, to fully characterise the protein localisation and subsequently annotate its functional role within the cell, the presence of these MTS was detected in all the predicted mitochondrial proteins (n = 26). Using MitoProt server, 16 HPs showed the presence of the MTS peptides.

3.4.1.3. *Membrane-associated proteins*

The presence of trans-membrane domains/helices and a glycosyl-phosphatidylinositol (GPI) anchor were detected to further establish localisation on the cell membrane. The HMMTOP and TMHMM servers were used in order to specifically predict the likelihood of a protein to be a membrane protein. Of the 309 HPs, 195 showed the presence of trans-membrane helices, suggesting membrane-associated proteins (**Appendix A.2)**. GPI-anchored proteins were predicted using the PredGPI; most of these proteins migrate to the plasma membrane, though some proteins have been shown to reside in different compartments [169]. Only, 6 proteins of the studied HPs were identified to be GPI-anchored (**Table 3.3**); these proteins had specificity values higher than 99.9% which indicates a very high probability of being GPI-anchored.

Table 3.3. Six proteins predicted to be GPI-anchored and the position of the anchor on the protein.

HP gene names (protein length)	GPI-anchor position
TP01_0004 (468)	439-468
TP01_0679 (543)	514-543
TP02_0512 (1336)	1316-1336
TP03_0038 (376)	351-376
TP03_0136 (189)	166-189
TP03_0564 (187)	161-187

3.4.1.4. *Detection of secreted proteins/secretome analysis*

Compared to other known intracellular protozoa, *T. parva* invades different types of cells to evade the host immune system. Hence, *T. parva* parasites have multiple secreted proteins that can manipulate host cell signalling pathways to promote parasite adhesion, recognition, and invasion [183]. Ninety-one (91) secreted proteins were detected; 57 containing N-terminal signal peptides according to the secretome analysis (**Appendix A.2)**, signifying that they are secreted by the classical pathways and 34 lacking the N-terminal signal peptide based on the subcellular localisation analysis, suggesting that they are possibly secreted by the non-classical secretion pathway (**Appendix A.2)**.

3.5. SEQUENCE COMPARISONS

3.5.1. Detection of homologs of hypothetical protein sequences

The BLASTp and PSI-BLAST searches performed against the non-redundant database showed that 294 of the 309 *T. parva* HPs had homology with proteins present in *Homo sapiens*, *Mus musculus*, *Plamodium falciparum*, *Bos taurus*, *Babesia bovis*, *Toxoplasma gondii*, *Neospora caninum* and *Sacchoromyces cerevisiae*. Most of the HPs (n = 212) investigated had homologs with proteins from the parasitic *T. equi*, and the non-parasitic *H. sapiens* and *M. muscus* eukaryotes (**Figure 3.3)**.

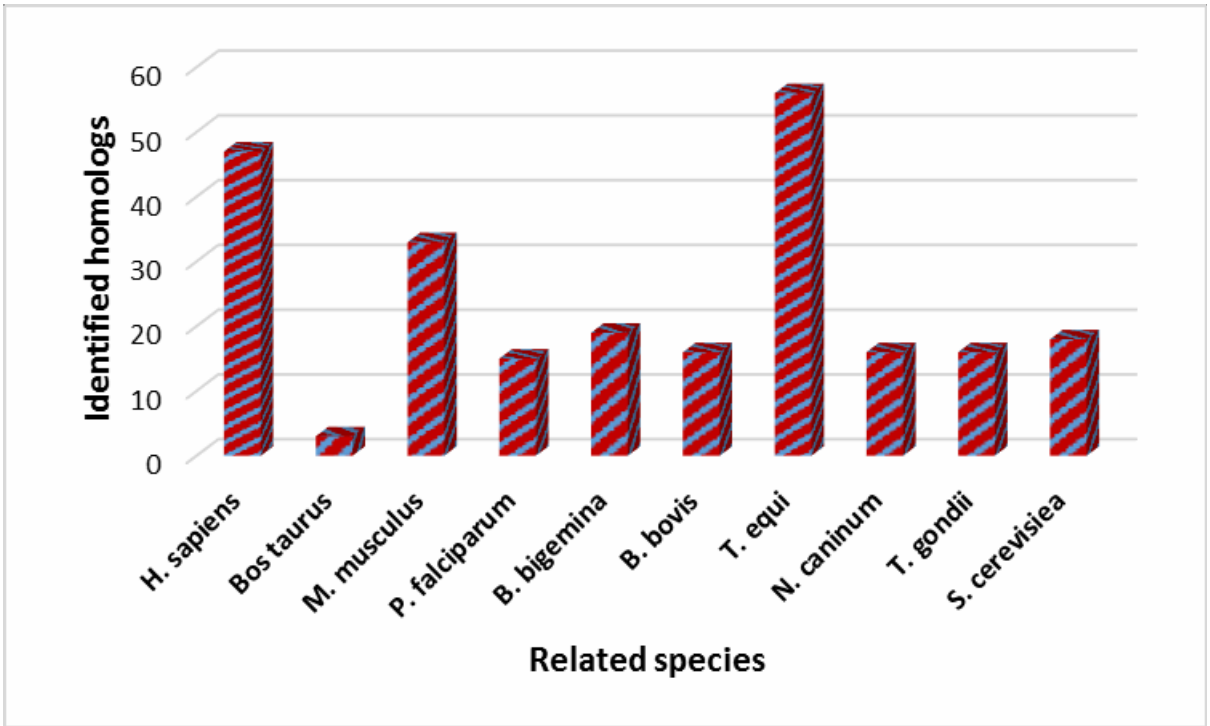


Figure 3.3. Hypothetical proteins with homology to proteins detected in other different related species.

3.5.1.1. *Homology confirmation*

Sequence similarity search (Blast) analysis:

To provide more support for annotations obtained from the sequence homology analysis, homologs of each of the *T. parva* HP were used as query sequences on Blast analysis, to check if they could retrieve back the corresponding HP sequences. Using this approach, 72% (212/294) of analyzed homologs retrieved back their respective HPs; the remaining HPs were all considered to be non-homologous. Hypothetical protein sequences which were retrieved by their respective homologs were subjected to pair-wise sequence alignment and domain analysis for further confirmation.

Pair-wise sequence alignment analysis:

Sequence alignments of both the HPs and their homologs were constructed using T-COFFEE and Clustal Omega, and most (n = 134) of the *T. parva* HP sequences had good alignments with corresponding homologs from both tools. The level of sequence conservation in T-COFFEE is indicated by colours where red indicates high sequence conservation and blue displays low sequence conservation. Clustal Omega outputs show the conservation level by consensus symbols, asterisk (***)** and colon (**:**) and a dot/period (**.**), wherein an asterisk (*****) shows fully conserved residues. Conservation between groups that share strong similar properties is indicated by a colon (**:**), while, conservation between groups that share a weak similarity is indicated by a dot/period (**.**); an example is shown in **Figure 3.4**.

(A)

```
TP01_0061       MIRLRPHLRSLQSHVSSGQIPTREYLESFEPKSLRRLKVFNQTPPEDIRKTFFEKLTSHY   60
BEWA_034260     MLRSSACRRALQSYLSTGQIATREYLESFEPKSLRRLKVFDETPPEDLQRSFFEKLTANY   60
                *:*     *:***::*:*** ********************::*****::::******:.*

TP01_0061       SNPRSILSALF---------------------------LSTWQL--------------IF   79
BEWA_034260     SNPRSILKLYRSYIAVDDYPCYSWLVRCLCQLGNSFSFNSFWSPIDRQAIVKLPLFKYLI   120
                *******.                                * *.              ::

TP01_0061       ------FQLFLPRHVPRMIFALSALEYRSWPLLTHLLELVQKHLKDWRISTISVMIGCLV   133
BEWA_034260     YDLIERKHYLHPRHVPRVIFALTALDYRCWPLLSELLSLVRAHLQDWRLPTLSCMIGCLV   180
                       : : ******:****:**:**.****:.**.**: **:***: *:* ******

TP01_0061       KLGLDNIPN-NSNLLPNVTSLLLNEVESRDPSYSSPFDWAILSYSLVLTNRYEWPNYENA   192
BEWA_034260     QLGISNTNFPDEDQSTNIISLLLNEVDRRDPSDSTPFDWALLSYSLVLTNRYEWPTRDNA   240
                :**:.*    :.:   *: *******: **** *:*****:**************. :**

TP01_0061       SLPVFLNRACKGLNLENLPLSGWTQYYLYISLYCTDVEKPRNEIQIKESVPPEIQQSLHT   252
BEWA_034260     SLPRYLKRACEGLTLENLPLSGWMQYYLYLTLYCTDVEKPSNEKDIKNSIPYEFQESLHT   300
                *** :*:***:**.********* *****::********* ** :**:*:* *:*:****

TP01_0061       RWLNEILIHSQPQGSEMLQKDVDTVLEGLGNTEGLINCSVGRGDDEQHCLFTGHLFPNRK   312
BEWA_034260     RWLNDILIHAQPQGSEMLQKDVDTVLEGLGITEGLINCSVGREDDEQHCLFCGHLFTHRK   360
                ****:****:****************** ********** ******* **** .**

TP01_0061       LCFEYNYLMPMPDGTPVESGLVSFRRRLFNKFGYNTVVIHRHQWEKLTILEKEEQLLQIL   372
BEWA_034260     LCFEYNYLMPKPDGMPVESGLISFRRRMFNKFGYNTAVIHKHQWDSLSLIEKEEQLLRIL   420
                ********** *** ******:*****:********.***:***:.*:::*******:**

TP01_0061       SKFPIIESNLYESKPEAKTFE-GSNYKYLKHLRPKITTWPPDKITL 417
BEWA_034260     STFPAVEPEVYDSKPAPKTFPEDADLKYMKHLRPKITTWPPEKFTV 466
                *.** :* ::*:***  ***   :: **:************:*:*:
```

(B)

```
TP01_0061         1 MIRLRPHLRSLQSHVSSGQIPTREYLESFEPKSLRRLKVFNQTPPEDIRKTFFEKLTSHYSNPRSILSA   69
BEWA_034260       1 MLRSSACRRALQSYLSTGQIATREYLESFEPKSLRRLKVFDETPPEDLQRSFFEKLTANYSNPRSILKL   69

cons              1 *:*  .  *:***::*:***.*******************::*****::::******::********.   69

TP01_0061        70 ---------------------------LFLSTWQLIFFQ-------------------LFLPRHVPRM   91
BEWA_034260      70 YRSYIAVDDYPCYSWLVRCLCQLGNSFSFNSFWSPIDRQAIVKLPLFKYLIYDLIERKHYLHPRHVPRV  138

cons             70                          * * *. *  *                  : ******:  138

TP01_0061        92 IFALSALEYRSWPLLTHLLELVQKHLKDWRISTISVMIGCLVKLGLDNI--PNNSNLLPNVTSLLLNEV  158
BEWA_034260     139 IFALTALDYRCWPLLSELLSLVRAHLQDWRLPTLSCMIGCLVQLGISNTNFPD-EDQSTNIISLLLNEV  206

cons            139 ****:**:**.****:.**.**: **:***:.*:* ******:**:.*   *: .:  .*: *******  207

TP01_0061       159 ESRDPSYSSPFDWAILSYSLVLTNRYEWPNYENASLPVFLNRACKGLNLENLPLSGWTQYYLYISLYCT  227
BEWA_034260     207 DRRDPSDSTPFDWALLSYSLVLTNRYEWPTRDNASLPRYLKRACEGLTLENLPLSGWMQYYLYLTLYCT  275

cons            208 : **** *:*****:***************. :***** :*:***:**.********* *****::****  276

TP01_0061       228 DVEKPRNEIQIKESVPPEIQQSLHTRWLNEILIHSQPQGSEMLQKDVDTVLEGLGNTEGLINCSVGRGD  296
BEWA_034260     276 DVEKPSNEKDIKNSIPYEFQESLHTRWLNDILIHAQPQGSEMLQKDVDTVLEGLGITEGLINCSVGRED  344

cons            277 ***** ** :**:*:* *:*:********:****:********************** ********** *  345

TP01_0061       297 DEQHCLFTGHLFPNRKLCFEYNYLMPMPDGTPVESGLVSFRRRLFNKFGYNTVVIHRHQWEKLTILEKE  365
BEWA_034260     345 DEQHCLFCGHLFTHRKLCFEYNYLMPKPDGMPVESGLISFRRRMFNKFGYNTAVIHKHQWDSLSLIEKE  413

cons            346 ******* ****.:************ *** ******:*****:********.***:***:.*:::*** 414

TP01_0061       366 EQLLQILSKFPIIESNLYESKPEAKTF-EGSNYKYLKHLRPKITTWPPDKITL  417
BEWA_034260     414 EQLLRILSTFPAVEPEVYDSKPAPKTFPEDADLKYMKHLRPKITTWPPEKFTV  466

cons            415 ****:***.** :*.::*:*** .*** *.:: **:************:*:*:  467
```

Figure 3.4.Alignment of the sequence of the *T. parva* HPs TP01_0061 and their corresponding homolog, BEWA_034260, from *T. equi* using Clustal Omega (A) and T-Coffee (B).

Domain analysis:

Domain predictions were carried out to check if the *T. parva* HPs contained domains corresponding to those of their respective homologs. Of the 212 HPs analysed, only 29 HPs had domains corresponding to those of their respective homologs. Examples of these proteins are shown in **Figure 3.5 and 3.6.**

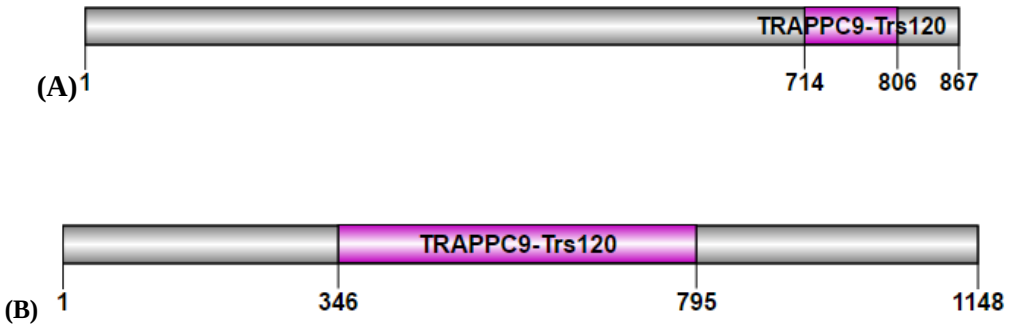


Figure 3.5.A schematic representation of the *T. parva* HPTP01_0306 (A) containing the domain TRAPPC9-Trs120 shared with the *Homo sapiens* homolog, trafficking protein XP_011515631.1 (B).

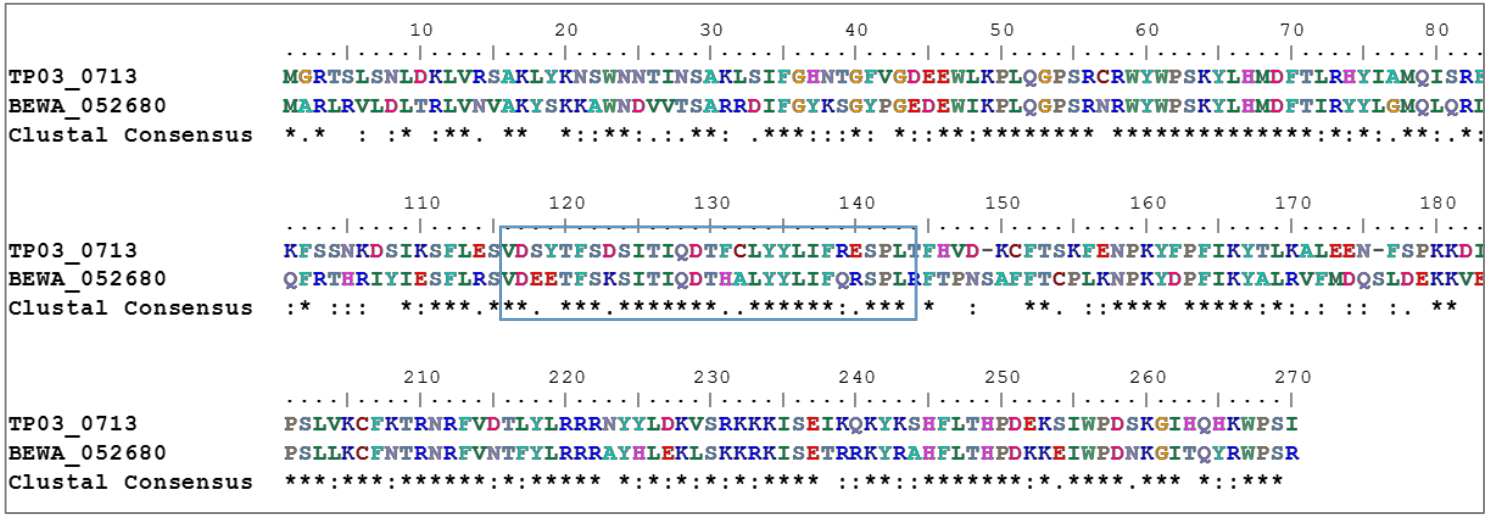
Figure 3.6.(A) A schematic representation of the *T. parva* HP TP01_0061 showing a RAP domain shared with the *T. equi* homolog, BEWA_034260 (B).

Function prediction based on domain and sequence alignments analysis:

The homology shared between the HPs and their corresponding functionally characterised homologs, identified from related organisms, permitted transfer of functional information from the known functional homologs to the HPs. Based on corresponding functional domains and good sequence alignments, 29 HPs were assigned relevant functions with a high level of confidence (**Table 3.4 and Figures 3.4-3.7.**).

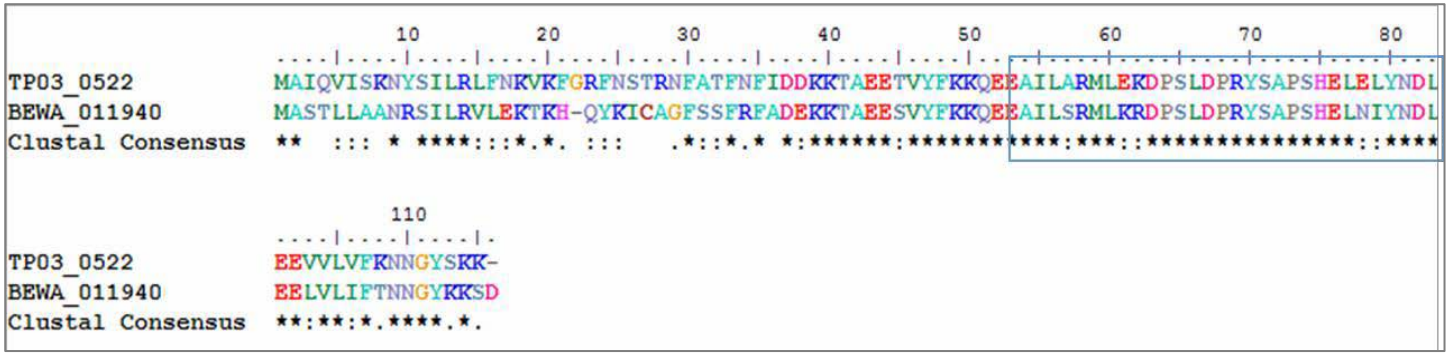
However, not all HPs shared common domains with their homologs; 142 HPs that had good alignments with respective homologs fell under this category. Therefore, to allow for more reliable functional annotations, a pair-wise sequence alignment analysis was performed to evaluate the conservation of the amino acid residues found within the domain region. Eight HPs displayed high levels of sequence conservation against their corresponding homolog; upon investigation of the residues spanning the domain region. Consequently, these HPs were also assigned functions with high confidence (proteins in the bottom eight in **Table 3.4**). Representative alignments are shown in **Figure 3.7**.

(A)



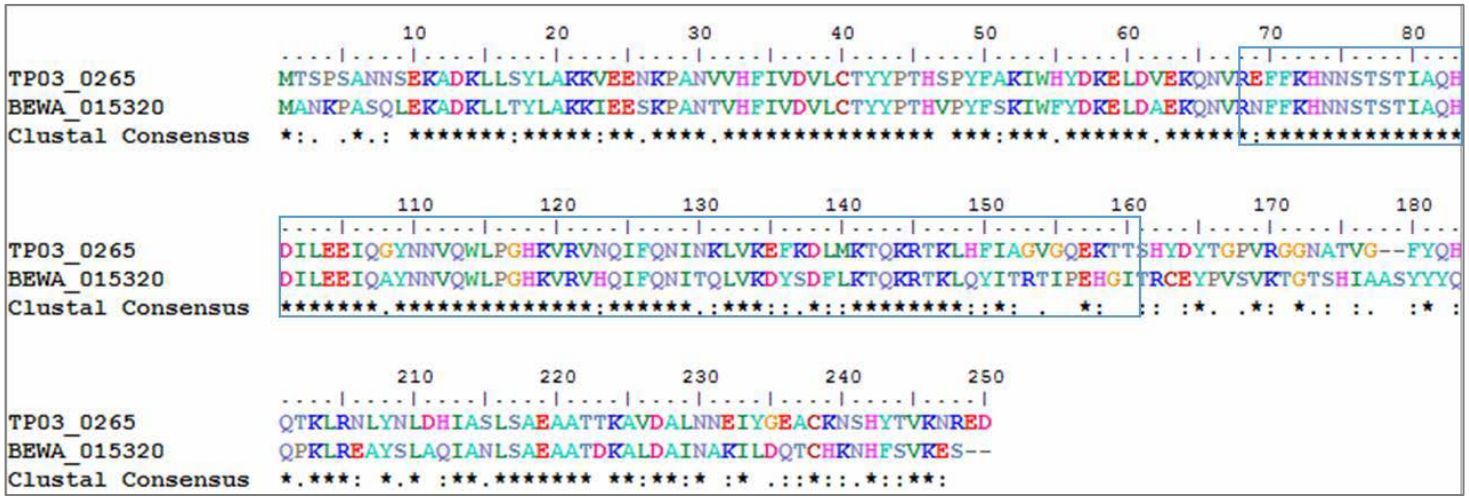
TP03_0713 (length: 268)
Identities: 72%

(B)



TP03_0522 (length: 115)
Identities: 75%

(C)



TP03_0265 (length: 248)
Identities: 75.5%

Figure 3.7. Diagrams A, B and C, depict a representation of a pair-wise sequence alignment analysis output showing the comparison of the sequences spanning the domain regions of the *T. parva* HPs against their homologs from related species.
Domain regions are marked in blue boxes and the identities are also presented below each alignment.

Table 3.4. A list of predicted functions assigned to *T. parva* HPs (n = 37) with a high level of confidence.

Sequence. No	*HP gene names* (protein length)	Probable function(s)
Good alignments and similar domains (n = 29)		
1.	TP01_0061 (417)	RAP domain-containing protein
2.	TP01_0306 (867)	Trafficking protein particle complex subunit 9 isoform X5
3.	TP01_0669 (427)	Acetyltransferase, GNAT family protein, related
4.	TP01_0676 (254)	SF-assemblin/beta giardin protein
5.	TP01_1011 (356)	Signal peptide-containing protein
6.	TP02_0357 (201)	RRP7 domain containing protein
7.	TP02_0419 (342)	FMLP-related receptor II
8.	TP02_0591 (1066)	Trans-membrane domain-containing protein
9.	TP02_0592 (1057)	Coiled coil and Trans-membranedomain-containing protein
10.	TP03_0008 (911)	Proteoglycan 4 isoform E preproprotein
11.	TP03_0125 (177)	Trans-membrane containing domain
12.	TP03_0305 (820)	Zinc finger protein 28 isoform X3
13.	TP03_0483 (344)	Trans-membrane protein
14.	TP03_0537 (427)	Phosphatidate phosphatase APP1
15.	TP03_0642 (248)	UPF0667 protein C1orf55 homolog
16.	TP03_0680 (181)	Signal peptide-containing protein
17.	TP03_0761 (185)	Csm1p
18.	TP03_0832 (155)	Trans-membrane domain-containing protein
19.	TP03_0856 (527)	Signal peptide-containing protein
20.	TP04_0283 (206)	Signal peptide and GMGPP repeat at C-terminus- containing protein
21.	TP04_0399 (357)	Membrane protein
22.	TP04_0455 (556)	Membrane protein
23.	TP04_0532 (206)	Calcium-binding protein 4 isoform X3
24.	TP04_0786 (381)	Trans-membrane domain-containing protein
25.	TP04_0910 (276)	Membrane protein
26.	TP01_0679 (543)	membrane protein, putative
27.	TP04_0905 (676)	Lyncein
28.	TP04_0729 (218)	Signal peptide containing-protein
29.	TP04_0188 (199)	MICOS complex subunit Mic19 isoform X2
	Different domains but good pairwise alignments of domain regions (n = 8)	
30.	TP01_0027 (127)	Peptidase_A25 and Cytokin_check_N domain-containing protein
31.	TP01_0144 (1218)	Trans-membrane domain-containing protein
32.	TP02_0302 (681)	Membrane protein
33.	TP03_0265 (248)	Homocysteine S-methyltransferase
34.	TP03_0522 (115)	ATP-dependent RNA helicase
35.	TP03_0525 (243)	Flavinator of succinate dehydrogenase
36.	TP03_0779 (300)	BNIP2 domain-contating protein
37.	TP04_0503 (245)	Coiled coil and V-ATPase_G_2 domain-containing protein

The remaining group of homologous HPs (n = 134), although their sequences showed very low conservation with the corresponding homologs, this group of HPs was still assigned probable functions based on the functional information of the annotated homologs, however, with a low level of confidence (**Table 3.5 and Appendix A.3**).

Table 3.5.List of predicted functions assigned to the *T. parva* HPs (n = 134) with a low level of confidence.

Sequence. No	HP gene names (protein length)	Probable function(s)
1.	TP01_0026 (722)	Elongator subunit ELP4
2.	TP01_0034 (625)	Gasdermin-C isoform X4
3.	TP01_0090 (424)	Protein kinase, cAMP-dependent, regulatory, type I, alpha (tissue specific extinguisher 1), isoform CRA_b
4.	TP01_0135 (215)	SAG-related sequence SRS36C (SAG5A)
5.	TP01_0143 (829)	NEDD4-like E3 ubiquitin-protein ligase WWP2
6.	TP01_0255 (155)	mCG116971, partial
7.	TP01_0346 (453)	Cadherin-18 isoform X5
8.	TP01_0392 (257)	Signal peptide-containing protein
9.	TP01_0493 (234)	Retinoblastoma-binding protein 5 isoform X5
10.	TP01_0537 (183)	Glycos_transf_4 domain containing protein
11.	TP01_0554 (520)	Trans-membrane protein
12.	TP01_0555 (146)	mCG14668, isoform CRA_c
13.	TP01_0563 (393)	Trans-membrane protein
14.	TP01_0626 (178)	APG5 domain-containing protein
15.	TP01_0671 (455)	Putative U2 small Nucleus ribonucleoprotein
16.	TP01_0864 (373)	14-3-3-like protein, related
17.	TP01_0912 (145)	mCG121048
18.	TP01_0931 (274)	Myosin light chain
19.	TP01_0953 (264)	PhyH domain-containing protein
20.	TP01_0985 (139)	mCG20969, isoform CRA_a, partial
21.	TP01_0993 (1124)	mCG132223
22.	TP01_1001 (727)	RIKEN cDNA 1300018J18 gene
23.	TP01_1003 (171)	hCG2024596
24.	TP01_1026 (421)	Pcf11p
25.	TP01_1034 (204)	Coiled coil-domain containing protein
26.	TP01_1064 (348)	Annexin A9
27.	TP01_1118 (650)	PL48 domain-containing protein
28.	TP01_1146 (117)	Ecm19p
29.	TP01_1179 (562)	DNA binding protein
30.	TP01_1197 (286)	Tetratricopeptide repeat protein 38 isoform X3
31.	TP02_0009 (336)	RNA recognition motif domain-containing protein
32.	TP02_0024 (570)	V-type proton ATPase subunit d 2
33.	TP02_0043 (583)	Calcyphosin-2 isoform X10
34.	TP02_0065 (272)	Dpy-19-like 4 , isoform CRA_a

35.	TP02_0092 (102)	X-linked lymphocyte regulated gene
36.	TP02_0133 (601)	T-cell activation Rho GTPase-activating protein isoform a
37.	TP02_0147 (167)	Chromosome 12 open reading frame 60
38.	TP02_0273 (355)	SWI/SNF complex-related protein
39.	TP02_0296 (132)	Soluble nsf attachment protein SNAP,
40.	TP02_0308 (287)	Coiled coil domain-containing protein
41.	TP02_0336 (212)	Coiled coil domain-containing protein
42.	TP02_0410 (596)	Unc-13 homolog D, isoform CRA_c, partial
43.	TP02_0420 (1743)	Dedicator of cytokinesis protein 9 isoform X22
44.	TP02_0424 (580)	GRB2-associated binding protein 1
45.	TP02_0427 (175)	Mitochondrionl carrier-like protein
46.	TP02_0534 (298)	Myozenin-3 isoform X1
47.	TP02_0575 (179)	Cytidine deaminase
48.	TP02_0582 (482)	Coiled coil domain-containing protein
49.	TP02_0585 (183)	Adamts15 protein, partial
50.	TP02_0586 (251)	GATA zinc finger domain containing 1, isoform CRA_b
51.	TP02_0679 (123)	Mu opioid receptor splice variant MOR-1H
52.	TP02_0682 (160)	PPR_2 domain-containing protein
53.	TP02_0695 (145)	Bli1p
54.	TP02_0705 (272)	Adrenergic receptor alpha 2B
55.	TP02_0711 (126)	Membrane protein
56.	TP02_0754 (85)	Non-cytoplasmic domain-containing protein
57.	TP02_0773 (172)	Aquaporin 3
58.	TP02_0849 (224)	Serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 6, isoform CRA_a
59.	TP02_0859 (556)	Coiled coil domain-containing protein
60.	TP02_0863 (222)	Muc3 protein, partial
61.	TP02_0871 (163)	Myosin phosphatase-Rho interacting protein
62.	TP02_0907 (766)	Ephexin-1 isoform 1
63.	TP02_0965 (149)	Striatin-3
64.	TP03_0024 (89)	Death domain-associated protein
65.	TP03_0028 (544)	Coiled coil containing protein
66.	TP03_0038 (376)	hCG1749747, isoform CRA_b
67.	TP03_0045 (477)	TTN protein, partial
68.	TP03_0051 (184)	Coiled coil domain-containing protein
69.	TP03_0095 (229)	Selenoprotein S isoform 2
70.	TP03_0098 (158)	Protein CASC4 isoform X1
71.	TP03_0125 (177)	Bifunctional anthranilate synthase/indole-3-glycerol-phosphate synthase
72.	TP03_0136 (189)	Adenosine deaminase-like protein isoform X5
73.	TP03_0138 (351)	Bifunctional anthranilate synthase/indole-3-glycerol-phosphate synthase
74.	TP03_0148 (1433)	PERQ amino acid-rich with GYF domain-containing protein 1
75.	TP03_0177 (79)	Ubiquitous TPR motif protein UTY
76.	TP03_0193 (346)	Zinc finger and SCAN domain-containing protein 22 isoform c
77.	TP03_0194 (368)	Zinc finger protein 420 isoform X5
78.	TP03_0234 (252)	Mas-related G-protein coupled receptor member H
79.	TP03_0246 (310)	Phosphotransferase enzyme family protein
80.	TP03_0256 (114)	MTCP1 domain-containing protein

81.	TP03_0271 (275)	KIAA0564 protein, partial
82.	TP03_0335 (200)	Oxoglutarate/malate translocator protein
83.	TP03_0336 (1003)	Uso1 / p115 like vesicle tethering protein
84.	TP03_0378 (121)	UPF0686 protein C11orf1 homolog isoform X3
85.	TP03_0381 (506)	Target of myb1-like protein 2
86.	TP03_0388 (195)	MYT1L protein
87.	TP03_0389 (321)	PLXNB2 protein, partial
88.	TP03_0463 (524)	Dennd5b protein
89.	TP03_0471 (538)	Vba1p
90.	TP03_0484 (1029)	Acetyl-Coenzyme A acyltransferase 2 (Mitochondrionl 3-oxoacyl-Coenzyme A thiolase), isoform CRA_e
91.	TP03_0556 (657)	Membrane protein
92.	TP03_0581 (307)	HaeIII restriction endonuclease
93.	TP03_0597 (1509)	Liprin-alpha-1 isoform X5
94.	TP03_0605 (146)	mCG1028421, isoform CRA_b, partial
95.	TP03_0641 (299)	Adenovirus endoprotease domain-containing protein
96.	TP03_0647 (670)	Hda3p
97.	TP03_0657 (1081)	Bud3p
98.	TP03_0681 (373)	Bmp domain-containing protein
99.	TP03_0725 (290)	Plm2p
100.	TP03_0759 (195)	Pleckstrin homology domain containing, family Q member 1, isoform CRA_d, partial
101.	TP03_0768 (346)	Hexokinase 2
102.	TP03_0820 (255)	Interactor protein for cytohesin exchange factors 1 isoform X4
103.	TP03_0827 (885)	Trans-membrane domain-containing protein
104.	TP03_0850 (397)	hCG2041388, partial
105.	TP03_0896 (524)	Dennd5a protein
106.	TP03_0898 (223)	Zfp827 protein, partial
107.	TP03_0901 (522)	DENND5B protein
108.	TP04_0068 (233)	Cir1p
109.	TP04_0073 (379)	FANCI_S1-cap domain-containing protein
110.	TP04_0081 (935)	Protein MAM3
111.	TP04_0082 (207)	Large ribosomal subunit processing factor
112.	TP04_0087 (356)	Alkaline ceramidase 1
113.	TP04_0114 (198)	Tumor protein D53
114.	TP04_0127 (540)	Topoisomerase (DNA) III beta, isoform CRA_b, partial
115.	TP04_0128 (587)	DNA primase large subunit
116.	TP04_0144 (370)	Pantothenate kinase
117.	TP04_0171 (306)	General transcription factor IIH polypeptide 5 GTF2H5
118.	TP04_0172 (912)	Sushi domain-containing protein 4 isoform b precursor
119.	TP04_0190 (74)	Alternative protein RBM15B
120.	TP04_0240 (327)	Protein L-Myc isoform 3 (41%)
121.	TP04_0252 (649)	C15orf42 protein, partial
122.	TP04_0259 (137)	hCG1658138, partial
123.	TP04_0275 (2405)	CsoS2_M domain-containing protein
124.	TP04_0405 (718)	SCL-interrupting locus protein isoform X6
125.	TP04_0414 (768)	Sld5 and Coiled coil domain-containing protein

126.	TP04_0579 (106)	Epididymal secretory protein E3-beta precursor
127.	TP04_0633 (193)	Hepatic leukemia factor isoform X2
128.	TP04_0654 (591)	Merlin isoform 1
129.	TP04_0693 (221)	Peptide chain release factor 1
130.	TP04_0708 (349)	3-methyl-2-oxobutanoate hydroxymethyltransferase
131.	TP04_0833 (147)	mCG1029182
132.	TP04_0834 (398)	Sfh1p
133.	TP04_0896 (489)	Ran binding protein
134.	TP05_0020 (133)	Interferon beta-2 precursor

3.5.2. Orthology analysis

Using PiroplasmaDB, a genome database for the genus Piroplasma, 22 HPs were shown to be orthologous to *T. annulata* proteins belonging to the SVSP family (*Theileria*- specific sub-telomeric proteins) and the remainder was found to be orthologous to uncharacterised proteins. Characteristic features of the SVSP family include; a putative signal peptide responsible for secretion, QP-rich region, followed by the FAINT domain [184]. As such, all 22 HPs identified as belonging to the SVSP family contained some of the characteristics features of the SVSPs family except the QP-rich region. Eighteen (18) contained signal peptides, FAINT domain and NLS, two had a signal peptide only, while another two had FAINT domain and NLS only. The list of predicted SVSP proteins is provided in **Table 3.6**.

Table 3.6. The twenty-two (22) *T. parva* HPs annotated as members of the SVSP family.

Sequence. No	HP gene name (protein length)	Domain	Signal peptide	Nuclear Localisation Sequence	Orthology	
					T. annulata	Sequence identity
1.	TP01_0004 (468)	DUF529(336-395); DUF1640 (64-104); DUF342 (63-107); DUF4407 (55-116)	Present	2 NLS	tan:TA02740 (525)	44%
2.	TP01_0007 (510)	DUF529 (333-387, 439-489)	Present	1 NLS	tan:TA17540 (576)	41%
3.	TP01_0008 (444)	DNA translocase FtsK (152-286); Non_cytoplasmic_domain (22-444)	Present	1 NLS	tan:TA02740 (525)	39%
4.	TP01_0009 (498)	DUF529 (305-370, 409-471)	Present	1 NLS	tan:TA17540 (576)	42%
5.	TP02_0004 (579)	DUF529 (407-463); Non_cytoplasmic_domain (21-579)	Present	2 NLS	tan:TA17120 (743)	35%
6.	TP02_0005 (570)	DUF529 (480-541)	Present	2 NLS	tan:TA17540 (576)	41%
7.	TP02_0006 (595)	DUF529 (410-464, 507-549	Present	3 NLS	tan:TA17540 (576)	44%
8.	TP02_0007 (515)	DUF529 (425-475)	Present	3 NLS	tan:TA17540 (576)	43%
9.	TP03_0678 (210)	Trans-membrane region (13-35); coiled coil (82-132)	Absent	Not present	tan:TA17695 (210)	41%
10.	TP03_0839 (260)	None	Absent	Not present	tan:TA18455 (298)	40%
11.	TP03_0870 (563)	DUF529 (375-429, 478-521)	Present	1 NLS	tan:TA17540 (576)	44%
12.	TP03_0873 (510)	DUF529 (429-477)	Present	1 NLS	tan:TA17540 (576)	36%
13.	TP03_0875 (509)	DUF529 (419-482)	Present	1 NLS	tan:TA02740 (525)	43%
14.	TP03_0877 (341)	DUF529 (158-213, 260 -301); Coil (162-182)	Absent	1 NLS	tan:TA17540 (576)	42%
15.	TP03_0880 (609)	DNA translocase FtsK (36-154); DUF529 (519-582)	Present	3 NLS	tan:TA02740 (525)	40%
16.	TP03_0881 (568)	DNA translocase FtsK (85-228); DUF529 (380-438, 485-537)	Absent	2 NLS	tan:TA17540 (576)	39%
17.	TP03_0882 (607)	DUF1421 (146-343); DUF529 (520-577)	Present	2 NLS	tan:TA17540 (576)	37%
18.	TP03_0883 (537)	DUF529 (458-505); DNA translocase FtsK (114-294); Non_cytoplasmic_domain (22-537)	Present	1 NLS	tan:TA02740 (525)	39%
19.	TP03_0885 (481)	DUF529 (390-440); DNA translocase FtsK (65-235); DUF4195 (136-175)	Present	3 NLS	tan:TA17540 (576)	40%
20.	TP03_0893 (568)	DUF529 (377-438, 483-531); DNA translocase FtsK (93-304)	Present	2 NLS	tan:TA17540 (576)	42%
21.	TP04_0003 (561)	DNA translocase FtsK (101-356); Neuroendocrine-specific golgi protein P55 (NESP55; 232-363); Non_cytoplasmic_domain (21-561)	Present	1 NLS	tan:TA17540 (576)	38%
22.	TP04_0006 (522)	DUF529 (109-220); DNA translocase FtsK (72-185)	Present	1 NLS	tan:TA02740 (525)	40%

3.6. PREDICTION OF SEQUENCE DOMAINS

Besides confirmation of sequence homology, domains were also considered as the basis to assign functional roles of the HPs. Thus, in addition to the 29 HPs detected with domains corresponding to respective homologs, from the analysis of 212 HPs that had good sequence alignments, possible functions were assigned to 26 HPs (that had no homologs) based on this analysis (**Table 3.7**). Overall, it was observed that most of the HPs had a common domain known as the non-cytoplasmic domain (extracellular; n = 106). On the contrary, 65 HPs were shown to possess a cytoplasmic (intracellular) domain **(Figure 3.8)**. Also, a large number of proteins (n = 79) containing trans-membrane domain associated with membrane proteins was observed, followed by HPs containing the FAINT domain (n = 73). The list of all predicted domains (n = 244) and respective HPs are shown in **Appendix A.4.**

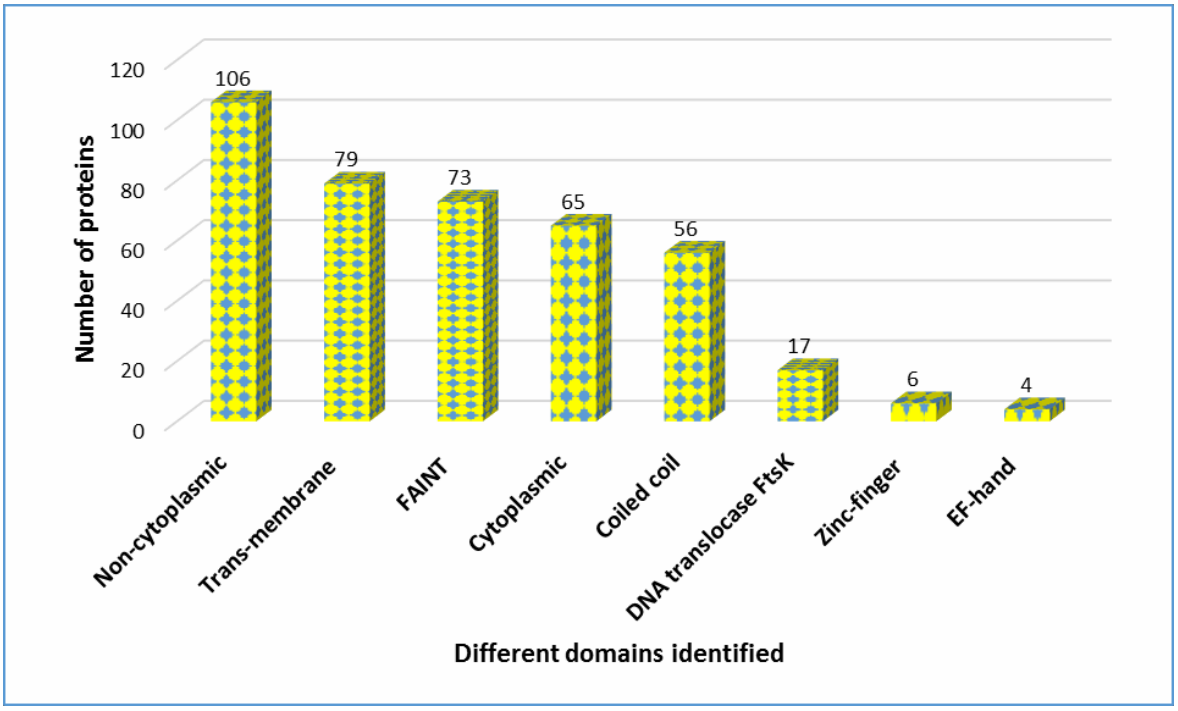


Figure 3.8. A distribution of the major domain classes identified in hypothetical protein sequences.

Table 3.7. Protein functions annotated based on domains possessed by *T. parva* HPs.

Sequence. No	HP gene names (protein length)	Probable function(s)
	Domains predicted by more than one analysis tool (n = 9)	
1.	TP01_0138 (131)	Ribosomal_L32p protein
2.	TP02_0528 (552)	Trans-membrane domain-containing protein
3.	TP02_0555 (102)	Trans-membrane domain-containing protein
4.	TP02_0897 (1499)	Trans-membrane domain-containing protein
5.	TP03_0042 (120)	Coiled coil domain-containing protein
6.	TP03_0060 (137)	Coiled coil domain-containing protein
7.	TP03_0094 (176)	Protein filamin
8.	TP03_0268 (1154)	Sexual stage antigen-associated protein
9.	TP05_0039 (126)	Trans-membrane domain-containing protein
	Domains predicted by a single analysis tool (n = 17)	
10.	TP01_0457 (163)	Colicin_immun domain-containing protein
11.	TP01_0699 (76)	Cytomegalovirus early phosphoprotein P34
12.	TP01_0736 (589)	Exonuclease VII
13.	TP01_0847 (642)	Trans-membrane and cytoplasmic domain-containing protein
14.	TP02_0026 (166)	Immunity protein
15.	TP02_0280 (144)	ATP binding, AAA superfamily
16.	TP02_0363 (889)	PI-PLC-X domain-containing protein
17.	TP02_0459 (104)	Trafficking protein Mon1
18.	TP02_0569 (438)	Phosphopantetheine
19.	TP02_0752 (476)	Transcriptional regulatory/DNA binding protein
20.	TP03_0606 (163)	Putative stress-responsive Nucleus envelope protein
21.	TP03_0819 (112)	Sin-like protein
22.	TP03_0899 (524)	Regulation of transcription
23.	TP04_0192 (110)	FAST kinase-like protein
24.	TP04_0327 (507)	Mediator of homo- and hetero-oligomerisation
25.	TP04_0422 (129)	Non-cytoplasmic domain-containing protein
26.	TP04_0576 (447)	Aminotransferase I & II

3.7. Detection Of Virulence Factors

Various bioinformatics tools such as VICMpred, VirulentPred and MP3 were used to predict virulence factors among the 309 HPs. Data generated from the analysis indicated that 224 of the 309 HPs investigated had virulence factors (**Table 3.8**), which were detected by at least two of the three computational tools used.

Table 3.8. List of *T. parva* HPs with virulence factors (n = 224).

Sequence. No	HP gene names (protein length)	Virulent proteins			Consensus
		VICMpred	Mp3	VirulentPred	
1.	TP01_0305 (78)	Cellular process	Pathogenic	Virulent	Virulent
2.	TP01_0391 (680)	Cellular process	Pathogenic	Virulent	Virulent
3.	TP01_0636 (151)	Cellular process	Pathogenic	Virulent	Virulent
4.	TP01_0817 (409)	Cellular process	Pathogenic	Virulent	Virulent
5.	TP01_0891 (150)	Cellular process	Pathogenic	Virulent	Virulent
6.	TP01_1123 (180)	Cellular process	Pathogenic	Virulent	Virulent
7.	TP02_0150 (385)	Metabolism molecule	Pathogenic	Virulent	Virulent
8.	TP02_0213 (364)	Cellular process	Pathogenic	Virulent	Virulent
9.	TP02_0267 (413)	Cellular process	Pathogenic	Virulent	Virulent
10.	TP02_0428 (243)	Metabolism Molecule	Pathogenic	Virulent	Virulent
11.	TP02_0651 (228)	Cellular process	Pathogenic	Virulent	Virulent
12.	TP02_0776(2126)	Virulence factors	Pathogenic	Virulent	Virulent
13.	TP02_0812 (576)	Cellular process	Pathogenic	Virulent	Virulent
14.	TP02_0860 (246)	Metabolism molecule	Pathogenic	Virulent	Virulent
15.	TP02_0880 (187)	Information and storage	Pathogenic	Virulent	Virulent
16.	TP02_0916 (1070)	Virulence factors	Pathogenic	Virulent	Virulent
17.	TP03_0034 (201)	Cellular process	Pathogenic	Virulent	Virulent
18.	TP03_0055 (557)	Cellular process	Pathogenic	Virulent	Virulent
19.	TP03_0132 (112)	Cellular process	Pathogenic	Virulent	Virulent
20.	TP03_0169 (239)	Information and storage	Pathogenic	Virulent	Virulent
21.	TP03_0323 (178)	Virulence factors	Pathogenic	Virulent	Virulent
22.	TP03_0329 (317)	Virulence factors	Pathogenic	Virulent	Virulent
23.	TP03_0475 (459)	Cellular process	Pathogenic	Virulent	Virulent
24.	TP03_0523 (238)	Cellular process	Pathogenic	Virulent	Virulent
25.	TP03_0620 (452)	Cellular process	Pathogenic	Virulent	Virulent
26.	TP03_0729 (379)	Cellular process	Pathogenic	Virulent	Virulent
27.	TP03_0742 (273)	Cellular process	Pathogenic	Virulent	Virulent
28.	TP03_0825 (446)	Metabolism Molecule	Pathogenic	Virulent	Virulent
29.	TP04_0254 (220)	Cellular process	Pathogenic	Virulent	Virulent
30.	TP04_0353 (477)	Virulence factors	Pathogenic	Virulent	Virulent
31.	TP04_0715 (212)	Cellular process	Pathogenic	Virulent	Virulent
32.	TP04_0869 (283)	Metabolism molecule	Pathogenic	Virulent	Virulent
33.	TP01_0004 (468)	Metabolism Molecule	Pathogenic	Virulent	Virulent
34.	TP01_0008 (444)	Virulence factors	Non-Pathogenic	Virulent	Virulent
35.	TP01_0027 (127)	Metabolism Molecule	Pathogenic	Virulent	Virulent
36.	TP01_0034 (625)	Cellular Process	Pathogenic	Virulent	Virulent
37.	TP01_0038 (458)	Cellular Process	Pathogenic	Virulent	Virulent
38.	TP01_0061 (417)	Metabolism Molecule	Pathogenic	Virulent	Virulent
39.	TP01_0124 (345)	Virulence factors	Pathogenic	Virulent	Virulent
40.	TP01_0090 (424)	Information and Storage	Pathogenic	Virulent	Virulent
41.	TP01_0135 (215)	Cellular Process	Pathogenic	Virulent	Virulent
42.	TP01_0143 (829)	Cellular Process	Pathogenic	Virulent	Virulent
43.	TP01_0144 (1218)	Metabolism Molecule	Pathogenic	Virulent	Virulent
44.	TP01_0270 (151)	Information and storage	Pathogenic	Virulent	Virulent
45.	TP01_0306 (867)	Virulence factors	Pathogenic	Virulent	Virulent
46.	TP01_0307 (584)	Metabolism Molecule	Pathogenic	Virulent	Virulent
47.	TP01_0346 (453)	Cellular Process	Pathogenic	Virulent	Virulent
48.	TP01_0392 (257)	Cellular Process	Pathogenic	Virulent	Virulent

49.	TP01_0457 (163)	Cellular Process	Pathogenic	Virulent	Virulent
50.	TP01_0554 (520)	Metabolism Molecule	Pathogenic	Virulent	Virulent
51.	TP01_0555 (146)	Information and storage	Pathogenic	Virulent	Virulent
52.	TP01_0559 (240)	Information and storage	Pathogenic	Virulent	Virulent
53.	TP01_0563 (393)	Cellular Process	Pathogenic	Virulent	Virulent
54.	TP01_0626 (178)	Virulence factors	Pathogenic	Virulent	Virulent
55.	TP01_0629 (110)	Cellular Process	Pathogenic	Virulent	Virulent
56.	TP01_0669 (427)	Cellular Process	Pathogenic	Virulent	Virulent
57.	TP01_0671 (455)	Metabolism Molecule	Pathogenic	Virulent	Virulent
58.	TP01_0676 (254)	Cellular Process	Pathogenic	Virulent	Virulent
59.	TP01_0679 (543)	Metabolism Molecule	Pathogenic	Virulent	Virulent
60.	TP01_0736 (589)	Cellular Process	Pathogenic	Virulent	Virulent
61.	TP01_0759 (944)	Virulence factors	Pathogenic	Virulent	Virulent
62.	TP01_0847 (642)	Virulence factors	Pathogenic	Virulent	Virulent
63.	TP01_0864 (373)	Cellular Process	Pathogenic	Virulent	Virulent
64.	TP01_0912 (145)	Cellular Process	Pathogenic	Virulent	Virulent
65.	TP01_0917 (405)	Cellular Process	Pathogenic	Virulent	Virulent
66.	TP01_0931 (274)	Information and storage	Pathogenic	Virulent	Virulent
67.	TP01_0953 (264)	Virulence factors	Pathogenic	Virulent	Virulent
68.	TP01_0993 (1124)	Metabolism Molecule	Pathogenic	Virulent	Virulent
69.	TP01_1001 (727)	Cellular Process	Pathogenic	Virulent	Virulent
70.	TP01_1003 (171)	Cellular Process	Pathogenic	Virulent	Virulent
71.	TP01_1011 (356)	Cellular Process	Pathogenic	Virulent	Virulent
72.	TP01_1026 (421)	Information and storage	Pathogenic	Virulent	Virulent
73.	TP01_1034 (204)	Cellular Process	Pathogenic	Virulent	Virulent
74.	TP01_1064 (348)	Virulence factors	Pathogenic	Virulent	Virulent
75.	TP01_1118 (650)	Information and storage	Pathogenic	Virulent	Virulent
76.	TP01_1179 (562)	Virulence factors	Pathogenic	Virulent	Virulent
77.	TP01_1197 (286)	Cellular Process	Pathogenic	Virulent	Virulent
78.	TP01_1208 (197)	Cellular Process	Pathogenic	Virulent	Virulent
79.	TP01_1228 (60)	Cellular Process	Pathogenic	Virulent	Virulent
80.	TP02_0004 (579)	Cellular Process	Pathogenic	Virulent	Virulent
81.	TP02_0006 (595)	Cellular Process	Pathogenic	Virulent	Virulent
82.	TP02_0009 (336)	Cellular Process	Pathogenic	Virulent	Virulent
83.	TP02_0024 (570)	Metabolism Molecule	Pathogenic	Virulent	Virulent
84.	TP02_0026 (166)	Cellular Process	Pathogenic	Virulent	Virulent
85.	TP02_0092 (102)	Cellular Process	Pathogenic	Virulent	Virulent
86.	TP02_0109 (99)	Cellular Process	Pathogenic	Virulent	Virulent
87.	TP02_0133 (601)	Cellular Process	Pathogenic	Virulent	Virulent
88.	TP02_0147 (167)	Cellular Process	Pathogenic	Virulent	Virulent
89.	TP02_0273 (355)	Cellular Process	Pathogenic	Virulent	Virulent
90.	TP02_0280 (144)	Metabolism Molecule	Pathogenic	Virulent	Virulent
91.	TP02_0296 (132)	Cellular Process	Pathogenic	Virulent	Virulent
92.	TP02_0302 (681)	Cellular Process	Pathogenic	Virulent	Virulent
93.	TP02_0308 (287)	Cellular Process	Pathogenic	Virulent	Virulent
94.	TP02_0336 (212)	Cellular Process	Pathogenic	Virulent	Virulent
95.	TP02_0357 (201)	Cellular Process	Pathogenic	Virulent	Virulent
96.	TP02_0363 (889)	Cellular Process	Pathogenic	Virulent	Virulent
97.	TP02_0410 (596)	Cellular Process	Pathogenic	Virulent	Virulent
98.	TP02_0419 (342)	Cellular Process	Pathogenic	Virulent	Virulent
99.	TP02_0420 (1743)	Metabolism Molecule	Pathogenic	Virulent	Virulent
100.	TP02_0424 (580)	Cellular Process	Pathogenic	Virulent	Virulent
101.	TP02_0426 (89)	Cellular Process	Pathogenic	Virulent	Virulent
102.	TP02_0427 (175)	Metabolism Molecule	Pathogenic	Virulent	Virulent
103.	TP02_0432 (402)	Cellular Process	Pathogenic	Virulent	Virulent
104.	TP02_0447 (411)	Information and storage	Pathogenic	Virulent	Virulent

105.	TP02_0459 (104)	Metabolism Molecule	Pathogenic	Virulent	Virulent
106.	TP02_0528 (552)	Metabolism Molecule	Pathogenic	Virulent	Virulent
107.	TP02_0534 (298)	Cellular Process	Pathogenic	Virulent	Virulent
108.	TP02_0569 (438)	Cellular Process	Pathogenic	Virulent	Virulent
109.	TP02_0582 (482)	Virulence factors	Pathogenic	Virulent	Virulent
110.	TP02_0583 (840)	Virulence factors	Non-Pathogenic	Virulent	Virulent
111.	TP02_0592 (1057)	Metabolism Molecule	Pathogenic	Virulent	Virulent
112.	TP02_0614 (80)	Information and storage	Pathogenic	Virulent	Virulent
113.	TP02_0644 (246)	Cellular Process	Pathogenic	Virulent	Virulent
114.	TP02_0695 (145)	Metabolism Molecule	Pathogenic	Virulent	Virulent
115.	TP02_0752 (476)	Cellular Process	Pathogenic	Virulent	Virulent
116.	TP02_0859 (556)	Information and storage	Pathogenic	Virulent	Virulent
117.	TP02_0871 (163)	Cellular Process	Pathogenic	Virulent	Virulent
118.	TP02_0897 (1499)	Cellular Process	Pathogenic	Virulent	Virulent
119.	TP02_0907 (766)	Virulence factors	Pathogenic	Virulent	Virulent
120.	TP02_0965 (149)	Cellular Process	Pathogenic	Virulent	Virulent
121.	TP03_0008 (911)	Virulence factors	Pathogenic	Virulent	Virulent
122.	TP03_0024 (89)	Cellular Process	Pathogenic	Virulent	Virulent
123.	TP03_0028 (544)	Cellular Process	Pathogenic	Virulent	Virulent
124.	TP03_0038 (376)	Cellular Process	Pathogenic	Virulent	Virulent
125.	TP03_0045 (477)	Metabolism Molecule	Pathogenic	Virulent	Virulent
126.	TP03_0051 (184)	Cellular Process	Pathogenic	Virulent	Virulent
127.	TP03_0094 (176)	Cellular Process	Pathogenic	Virulent	Virulent
128.	TP03_0095 (229)	Virulence factors	Pathogenic	Virulent	Virulent
129.	TP03_0098 (158)	Cellular Process	Pathogenic	Virulent	Virulent
130.	TP03_0107 (913)	Information and storage	Pathogenic	Virulent	Virulent
131.	TP03_0119 (198)	Cellular Process	Pathogenic	Virulent	Virulent
132.	TP03_0123 (481)	Information and storage	Pathogenic	Virulent	Virulent
133.	TP03_0125 (177)	Cellular Process	Pathogenic	Virulent	Virulent
134.	TP03_0138 (351)	Virulence factors	Pathogenic	Virulent	Virulent
135.	TP03_0148 (1433)	Virulence factors	Pathogenic	Virulent	Virulent
136.	TP03_0194 (368)	Cellular Process	Pathogenic	Virulent	Virulent
137.	TP03_0234 (252)	Metabolism Molecule	Pathogenic	Virulent	Virulent
138.	TP03_0246 (310)	Cellular Process	Pathogenic	Virulent	Virulent
139.	TP03_0255 (139)	Metabolism Molecule	Pathogenic	Virulent	Virulent
140.	TP03_0256 (114)	Information and storage	Pathogenic	Virulent	Virulent
141.	TP03_0305 (820)	Cellular Process	Pathogenic	Virulent	Virulent
142.	TP03_0336 (1003)	Information and storage	Pathogenic	Virulent	Virulent
143.	TP03_0378 (121)	Cellular Process	Pathogenic	Virulent	Virulent
144.	TP03_0380 (736)	Virulence factors	Pathogenic	Virulent	Virulent
145.	TP03_0381 (506)	Metabolism Molecule	Pathogenic	Virulent	Virulent
146.	TP03_0388 (195)	Virulence factors	Non-Pathogenic	Virulent	Virulent
147.	TP03_0389 (321)	Virulence factors	Non-Pathogenic	Virulent	Virulent
148.	TP03_0437 (90)	Cellular Process	Pathogenic	Virulent	Virulent
149.	TP03_0463 (524)	Cellular Process	Pathogenic	Virulent	Virulent
150.	TP03_0472 (125)	Metabolism Molecule	Pathogenic	Virulent	Virulent
151.	TP03_0483 (344)	Cellular Process	Pathogenic	Virulent	Virulent
152.	TP03_0484 (1029)	Cellular Process	Pathogenic	Virulent	Virulent
153.	TP03_0525 (243)	Cellular Process	Pathogenic	Virulent	Virulent
154.	TP03_0537 (427)	Virulence factors	Pathogenic	Virulent	Virulent
155.	TP03_0556 (657)	Virulence factors	Pathogenic	Virulent	Virulent
156.	TP03_0564 (187)	Cellular Process	Pathogenic	Virulent	Virulent
157.	TP03_0597 (1509)	Virulence factors	Pathogenic	Virulent	Virulent
158.	TP03_0605 (146)	Cellular Process	Pathogenic	Virulent	Virulent
159.	TP03_0641 (299)	Cellular Process	Pathogenic	Virulent	Virulent
160.	TP03_0642 (248)	Virulence factors	Pathogenic	Virulent	Virulent

161.	TP03_0647 (670)	Cellular Process	Pathogenic	Virulent	Virulent
162.	TP03_0657 (1081)	Cellular Process	Pathogenic	Virulent	Virulent
163.	TP03_0680 (181)	Cellular Process	Pathogenic	Virulent	Virulent
164.	TP03_0681 (373)	Cellular Process	Pathogenic	Virulent	Virulent
165.	TP03_0725(290)	Metabolism Molecule	Pathogenic	Virulent	Virulent
166.	TP03_0738 (666)	Metabolism Molecule	Pathogenic	Virulent	Virulent
167.	TP03_0754 (819)	Cellular Process	Pathogenic	Virulent	Virulent
168.	TP03_0759 (195)	Cellular Process	Pathogenic	Virulent	Virulent
169.	TP03_0761 (185)	Cellular Process	Pathogenic	Virulent	Virulent
170.	TP03_0768 (346)	Cellular Process	Pathogenic	Virulent	Virulent
171.	TP03_0779 (300)	Cellular Process	Pathogenic	Virulent	Virulent
172.	TP03_0812 (64)	Cellular Process	Pathogenic	Virulent	Virulent
173.	TP03_0819 (112)	Metabolism Molecule	Pathogenic	Virulent	Virulent
174.	TP03_0827 (885)	Virulence factors	Pathogenic	Virulent	Virulent
175.	TP03_0839 (260)	Virulence factors	Pathogenic	Virulent	Virulent
176.	TP03_0850 (397)	Information and storage	Pathogenic	Virulent	Virulent
177.	TP03_0856 (527)	Virulence factors	Pathogenic	Virulent	Virulent
178.	TP03_0870 (563)	Cellular Process	Pathogenic	Virulent	Virulent
179.	TP03_0877 (341)	Cellular Process	Pathogenic	Virulent	Virulent
180.	TP03_0883 (537)	Cellular Process	Pathogenic	Virulent	Virulent
181.	TP03_0893 (568)	Cellular Process	Pathogenic	Virulent	Virulent
182.	TP03_0896 (524)	Cellular Process	Pathogenic	Virulent	Virulent
183.	TP03_0898 (223)	Metabolism Molecule	Pathogenic	Virulent	Virulent
184.	TP03_0899 (524)	Cellular Process	Pathogenic	Virulent	Virulent
185.	TP03_0900 (521)	Cellular Process	Pathogenic	Virulent	Virulent
186.	TP03_0901 (522)	Virulence factors	Non-Pathogenic	Virulent	Virulent
187.	TP03_0903 (310)	Information and storage	Pathogenic	Virulent	Virulent
188.	TP04_0003 (561)	Cellular Process	Pathogenic	Virulent	Virulent
189.	TP04_0006 (522)	Cellular Process	Pathogenic	Virulent	Virulent
190.	TP04_0068 (233)	Metabolism Molecule	Pathogenic	Virulent	Virulent
191.	TP04_0073 (379)	Virulence factors	Pathogenic	Virulent	Virulent
192.	TP04_0081 (935)	Virulence factors	Pathogenic	Virulent	Virulent
193.	TP04_0082 (207)	Cellular Process	Pathogenic	Virulent	Virulent
194.	TP04_0087 (356)	Metabolism Molecule	Pathogenic	Virulent	Virulent
195.	TP04_0114 (198)	Cellular Process	Pathogenic	Virulent	Virulent
196.	TP04_0127 (540)	Cellular Process	Pathogenic	Virulent	Virulent
197.	TP04_0144 (370)	Virulence factors	Pathogenic	Virulent	Virulent
198.	TP04_0171 (306)	Cellular Process	Pathogenic	Virulent	Virulent
199.	TP04_0172 (912)	Metabolism Molecule	Pathogenic	Virulent	Virulent
200.	TP04_0181 (67)	Cellular Process	Pathogenic	Virulent	Virulent
201.	TP04_0188 (199)	Cellular Process	Pathogenic	Virulent	Virulent
202.	TP04_0210 (384)	Information and storage	Pathogenic	Virulent	Virulent
203.	TP04_0223 (48)	Cellular Process	Pathogenic	Virulent	Virulent
204.	TP04_0232 (884)	Cellular Process	Pathogenic	Virulent	Virulent
205.	TP04_0245 (816)	Cellular Process	Pathogenic	Virulent	Virulent
206.	TP04_0259 (137)	Cellular Process	Pathogenic	Virulent	Virulent
207.	TP04_0283 (206)	Metabolism Molecule	Pathogenic	Virulent	Virulent
208.	TP04_0405 (718)	Cellular Process	Pathogenic	Virulent	Virulent
209.	TP04_0414 (768)	Metabolism Molecule	Pathogenic	Virulent	Virulent
210.	TP04_0422 (129)	Cellular Process	Pathogenic	Virulent	Virulent
211.	TP04_0505 (74)	Metabolism Molecule	Pathogenic	Virulent	Virulent
212.	TP04_0532 (206)	Metabolism Molecule	Pathogenic	Virulent	Virulent
213.	TP04_0576 (447)	Virulence factors	Pathogenic	Virulent	Virulent
214.	TP04_0633 (193)	Metabolism Molecule	Pathogenic	Virulent	Virulent
215.	TP04_0638 (64)	Cellular Process	Pathogenic	Virulent	Virulent
216.	TP04_0654 (591)	Metabolism Molecule	Pathogenic	Virulent	Virulent

217.	TP04_0693 (221)	Cellular Process	Pathogenic	Virulent	Virulent
218.	TP04_0708 (349)	Cellular Process	Pathogenic	Virulent	Virulent
219.	TP04_0786 (381)	Cellular Process	Pathogenic	Virulent	Virulent
220.	TP04_0833 (147)	Metabolism Molecule	Pathogenic	Virulent	Virulent
221.	TP04_0910 (276)	Cellular Process	Pathogenic	Virulent	Virulent
222.	TP05_0020 (133)	Metabolism Molecule	Pathogenic	Virulent	Virulent
223.	TP03_0713 (268)	Information and storage	Pathogenic	Virulent	Virulent
224.	TP03_0648 (98)	Metabolism	Pathogenic	Virulent	Virulent

3.8. PROTEIN-PROTEIN NETWORK PREDICTIONS

Prediction of interaction partners can be used to assign functions of uncharacterised proteins, hence, in this study, the STRING database was utilised, which revealed three *T. parva* HPs to have interaction partners, including TP04_0172, TP03_0738 and TP01_0124. However, their functional interacting partners are also uncharacterised proteins **(Figure 3.9)**.

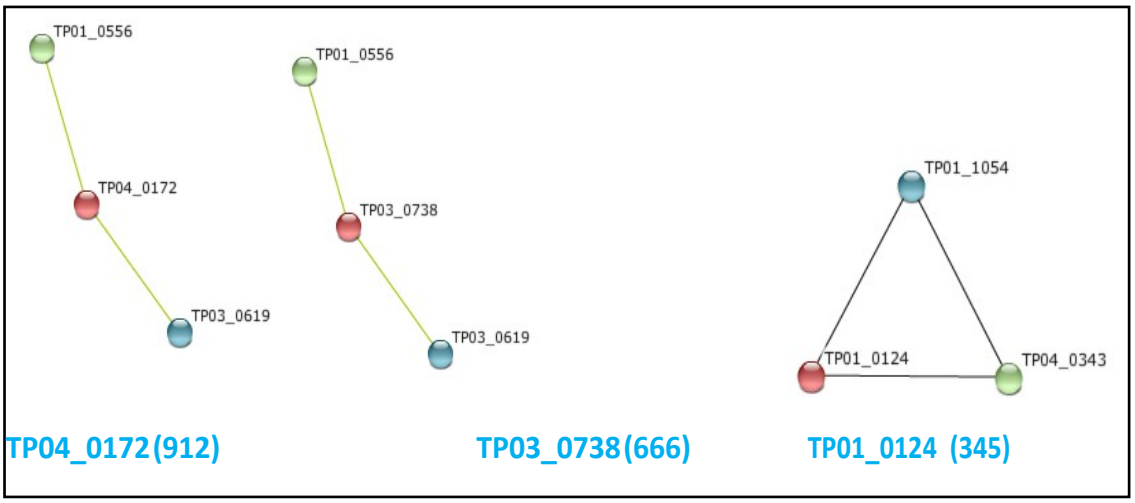


Figure 3.9. The display of functional partners predicted by STRING for the three *T. parva* HPs. The target protein is represented by a red circle and functional partners are presented as blue and green circles.

3.9. PREDICTION OF 3D-STRUCTURES

SWISS-MODEL was used to predict 3D-structures of 32 HPs that could not be assigned functions based on the sequence-based analysis. However, results attained from this analysis were redundant because of lack of structural similarities with annotated proteins in available structural databases.

3.10. PERFORMANCE ASSESSMENT

The accuracy of the bioinformatics pipeline used in this study was found to be 99% (**Table 3.9**); suggesting that the results of functional annotation of the HPs from the study are reliable and could further be utilised for future experimental research.

Table 3.9. List of accuracy, sensitivity, specificity and ROC area of various bioinformatics tools used for the prediction of functions of *T. parva* HPs obtained after ROC analysis

	Software name	Accuracy of prediction	Sensitivity	Specificity	ROC Area
1.	BLAST	96%	96%	N/A	NA
2.	SMART	98%	100%	N/A	N/A
3.	INTERPROSCAN	100%	100%	N/A	N/A
4.	Pfam	99%	99%	100%	N/A
5.	NCBI-CDD	100%	100%	N/A	N/A
6.	PROSITE	98%	100%	94.1%	0.97
7.	VICMPred	100%	100%	100%	1
8.	VirulentPred	100%	100%	100%	1
9.	MP3	100%	100%	100%	1
	AVERAGE	99%	99.4%	-	-

CHAPTER 4

4.0 DISCUSSION

The main aim of the study was to assign probable functions of HPs from the *T. parva* using protein sequence and structural information. The process of function annotation has been unable to keep stride with the high throughput genome sequencing projects. This causes a challenge for both computational and experimental researchers. To annotate function of uncharacterised proteins following experimental procedures is a laborious, time-consuming and a costly undertaking. On the other hand, computational approaches can significantly speed up the process of functional annotation and they are a much cheaper; hence, this was the favorable approach to achieve the objectives of this study.

Consequently, protein sequences of 309 HPs were extensively analysed by using a variety of available computational tools and 90% (n = 277) of the investigated HPs were successfully assigned probable functions. Overall, the transfer of annotation was achieved using sequence homology and orthology, and subcellular localisation, which revealed proteins that may be critical to the parasite biochemical and molecular processes, and its pathogenecity and persistence in the target host. Furthermore, potential drug targets were also detected among the HPs investigated; these polypeptides may be exploited in the development of novel disease control strategies.

Classification into functional families

Generally, analysis of canonical protein families revealed that the majority of HPs were binding proteins, mainly DNA-binding and RNA-binding, followed by proteins with transcription factor activity, enzymes, transporters and signal transduction proteins.

Binding proteins

DNA-binding and RNA-binding proteins play a significant role in several regulatory and cellular processes including translation, recombination and transcription; apparently, they are indispensable in the propagation and survival of the parasite within the bovine host [184, 185]. *Theileria* DNA binding proteins have been reported to translocate to the host cell nucleus [187], suggesting possible contact with the host. Evidently, eight of the binding

proteins were found to possess a trans-membrane domain, which indicates membrane association. Membrane-associated proteins are involved in several important processes such as cell wall degradation, metal ion binding, trans-membrane transport, and more important, they are associated with several virulence factors [185, 187]; accordingly they are implicated in attachment or invasion of the host cell. Membrane proteins are also good drug targets [182]. Thus, these HPs could have a significant role in the disease initiation and can also be considered in the development of disease control strategies, particularly as targets for chemotherapeutic drugs.

Of the 21 HPs identified as binding proteins, 43% (n = 9) were DNA-binding proteins and 55% (n = 5) of these have transcription factor activity; moreover, one of the DNA-binding HPs (TP03_0825) was identified as a member of the family of proteins known as apicomplexan AP2 (ApiAP2) transcription factors. Although very little is known about transcription factor-based gene regulation in apicomplexan parasite, the impression is that it is well coordinated considering the complexity of the life cycle of these parasites. Bioinformatics searches for DNA-binding proteins from the genome sequence of *Plasmodium* resulted in the identification of a family of regulatory proteins ApiAP2 protein family [189]; members of this family were also detected in published genome sequences of other apicomplexan parasites and some may be species-specific (reviewed by [190]). Interestingly, the 26 members of the ApiAP2 protein family identified in *Plasmodium* were initially annotated as HPs [191]. Plant homologs regulate genes involved in pathogenesis, stress response pathways, plant development, depending on the number of AP2 domains they possess [190]. In apicomplexan parasites, it is suggested that ApiAP2 proteins regulate timed and well coordinated gene expression governing multi-stage parasite development and interhost transition (athropod/mammalian host) required for successful propagation [190] and the ApiAP2 domain is conserved across the Apicomplexan genera [192]. Interhost transition is critical for establishment of infection and transmission of the parasite to a new host.

Transcription factors

Prior to the discovery of the lineage-specific ApiAP2 protein family, comparative genomics of apicomplexans, including *Theileria*, revealed the presence of specific transcription factors with DNA-binding domains that are closely related to those found in major families of

transcription factors from other eukaryotes [189]. Hence in this study, 11 proteins with transcription factor activity, containing various domains, were also identified among the HPs analyzed. Domains detected among these HPs included N-terminal IIS, AP2/ERF and CBFB_NFYA domains. As discussed above, transcription factors have an affinity to bind specific DNA sequences, this enables the control of the rate of genetic information transcription from DNA to mRNA [193]. Generally, transcription factors can enter the nucleus of the cell; as DNA binding proteins they can bind host promoter sequences, and activate transcription of host genes to promote pathogen infection [183, 184]. Hence, HPs with transcription factor activity may possibly play an essential role in the persistence of *T. parva* infection in the host. One of the HPs (TP03_0620) in this group had a homolog in another apicomplexan parasite, *viz*. *Toxoplasma gondii* strain ME49, an RNA polymerase-associated protein, further confirming the predicted function of this protein.

Enzymes

Theileria parva is an obligate parasite, thus it is dependent on the host for most of its nutritional needs [195]. Enzymes may possibly facilitate the parasite's survival within the host by carrying out several cellular processes making it viable for the course of infection in the host. Five types of enzymes were identified including transferase, methyltransferase, and GTPases from HPs the investigated. Transferases are a group of enzymes that facilitate the transfer of functional groups (for example, a methyl group; methyltransferase) from one molecule to another. These enzymes are indispensable in the biosynthesis of lipoproteins, of which some are known to play a direct role in virulence-associated functional roles such as invasion, colonisation, evasion of host defence and immunomodulation [76, 77]. TP04_0353 was identified to encode a methyltransferase, an enzyme known to be involved in the methylation process [198]. Post-translational modifications are important in Apicomplexan parasites for many processes, such as invasion into and egress from a host cell (prenylation, phosphorylation and palmitoylation), regulation of gene expression (acetylation and methylation), cell cycle regulation (ubiquitination) and many more [199]. Based on their essential roles, enzymes responsible for some of the post-translational modification processes have been investigated in the development of antiparasitic drugs [199] with methyltransferases being among these.

Proteins encoded by TP02_0512 and TP02_0916 were also shown to contain guanylate-binding domain, hence were predicted to have GTPase activity. GTPases are a family of enzymes that bind and hydrolyze guanosine triphosphate (GTP); they play a significant role in signal transduction, protein biosynthesis and translocation of proteins through membranes [200, 201]. To gain entrance into host cell tissues, the parasite manipulate the host signalling networks and host signalling molecules, such as chemokines; cytokines or interferons are then secreted in response to the infection [202].

Transporters

Transporter proteins play a critical role in metabolism processes, such as the uptake of nutrients and excretion of waste products. Thus it makes sense that these proteins have also been identified to be significant in the survival of intracellular pathogens and may also be involved in the pathogen's virulence [203]. In view of this, two transporters encoded by TP03_0742 and TP03_0055, were annotated as calcium-binding proteins because of the domains they possess, the calcium-binding and EF-hand domains, respectively. Calcium-binding proteins are known to contribute in calcium cell signalling pathways by binding to Ca^{2+}, the calcium ion plays a vital role in several cellular processes such as gene transcription and cell proliferation [204]. *Theileria parva* survives immune attack by transforming the host lymphocytes, inducing uncontrolled/continuous proliferation [21, 22]. This uncontrolled proliferation causes the infection to spread through the infected animal. Therefore, the role of the two *T. parva* HPs, TP03_0742 and TP03_0055, as modulators of transformation of infected cells, might be worth investigating. The localisation of these HPs revealed that TP03_0055 is a mitochondrial protein and TP03_0742 is involved in secretory pathway although it lacked a signal peptide, suggesting that it is possibly secreted via the non-classical secretory pathway. The secretion of the parasite protein suggests contact with the host; hence it is one of the characteristics used to identify modulators of transformation of the infected cells.

Signal transduction proteins

The tick-transmitted hemoprotozoan parasite *T. parva* induces transformation of the infected lymphocytes by disturbing the host cell signalling pathways that control survival and proliferation, resulting in continuous proliferation and clonal expansion of the infected cells.

This process induces a lympho-proliferative disorder known as East Coast fever (ECF) which mainly results in the death of the infested animal [205]. Also among other things, the process of *T. parva* sporozoites entry into host cell requires the stimulation of the signal transduction pathways in both the bovine host cell and the parasite [33]. In this study, two proteins encoded by TP04_0869 and TP02_0880 HPs were found to be involved in signal transduction, according to GO analysis. Many eukaryotic proteins involved in signal transduction usually have (Src homology 2) SH2 domains [207]; some have catalytic domains for kinase activity while others have phosphatase [208], and others have domains with effector and stimulator functions [208]. Protein TP04_0869 was shown to contain a catalytic domain, Protein kinase C-like domain, responsible for its kinase activity. A number of diseases have been linked to a dysfunction in the signal transduction networks and thus, understanding the key components of these networks as well as their associations to diseases could potentially provide a useful tool towards the detection of novel drug targets.

Characterisation of the physiochemical properties

In an effort to annotate the HPs, the physiochemical properties of these proteins were also investigated. These properties are useful for understanding a protein charge stability (pI) and indicating a protein's probability of being hydrophilic and soluble or likelihood of being hydrophobic and insoluble (GRAVY) [138]. Information obtained from the GRAVY analysis may also be used in localising these HPs; hydrophobic proteins are likely to be membrane proteins while hydrophilic proteins are likely to be globular proteins. For example, TP03_0381 and TP04_0200 were predicted to localize on the plasma membrane by the subcellular localisation prediction tools used in the study; this prediction was further supported by their positive GRAVY indices obtained from the physiochemical properties characterisation. However, conflicting results were also obtained; three proteins (TP03_0045, TP04_0654 and TP02_0267) shown to be membrane-associated by the subcellular localisation analysis had negative GRAVY indices. This discrepancy has previously been observed in a study conducted by Jaiswal *et al.*[209], where they were characterising physiochemical properties of chickpea membrane proteins. Similarly, they observed that the membrane proteins that showed a negative GRAVY index possessed trans-membrane domains while, those with positive GRAVY scores did not contain trans-membrane domains.

Most integral membrane proteins have been reported to be hydrophobic in nature, whereas the outer membrane ones are said to be hydrophilic [210, 211] causing the discovered ambiguity in the GRAVY indices.

Sequence homology analysis

Some (n = 4) important homologs of the *T. parva* HPs were identified in different eukaryotic species; one homolog of interest was detected in the *N. caninum*, and it belongs to the acetyltransferase family protein, known to be involved in virulence-associated functional roles such as invasion, colonisation, evasion of host defence and immunomodulation [76, 77]. This homolog shares a sequence identity of 35% with the corresponding HP (TP01_0669). Similar to methyltransferases, acetyltransferases are also considered to be highly interesting antiparasitic drugs [199].

Another homolog of interest was identified in *Mus musculus*, a Mas-related G-protein coupled receptor which shares 50% sequence identity with the corresponding HP (TP03_0234). G-protein-coupled receptors (GPCRs) represents one of the most vital classes of targets for drug discovery and malfunction of these receptors results in a wide variety of serious diseases, for instance congestive heart failure and cancer [212]; the latter displays a disease phenotype similar to that resulting from infection with *T. parva* [213].

One more important homolog is a tetratricopeptide repeat protein, which was detected in two species, *Homo sapiens* and *Saccharomyces cereviae*. This homolog respectively shares a relative sequence identity of 38% and 35%, with the corresponding HPs; TP01_1197 and TP03_0323. Tetratricopeptide repeat proteins have been reported to have virulence-associated functions, such as adhesion to host cells, targeting virulence factors into host cells, as well as hindering phagolysosomal development [214, 215]. Moreover, they are also involved in many regulatory and metabolic processes.

Otherwise, for other HPs, low sequence identities between homologs and HPs were also detected (25%-29%), suggesting that they could be distant homologs of *T. parva*. Nonetheless, alignment carried out using two MSA programs, T-COFFEE and Clustal Omega indicated significant sequence conservation between the target sequences and their homologs, thus the HPs may possibly perform similar functions. Noteworthy is that both programs use different methods for sequence alignment, yet most results attained from them were consistent.

A good correlation between data from the GO terms and homology analyses was observed in only nine HPs. One example is proteins TP01_0564 and TP02_0812, the GO term analysis revealed that these two are calcium ion-binding proteins, this molecular function was further supported by the calmodulin (calcium-binding protein) homolog identified from *Neospora caninum* (NCLIV_017560 and NCLIV_063600). The proteins also contained an EF-hand domain pair which is usually found in a large family of calcium-binding proteins. Another example is protein TP02_0512 which was shown to be a GTP-binding protein/GTPase based on the GO term analysis, this protein was also found to be homologous to a guanylate-binding protein (forms part of the GTPases family) identified from *Toxoplasma gondii* (TGME49_304990). The protein (TP02_0512) was also predicted to possess a guanylate-binding protein, N-terminal domain. Generally, this information further confirmed the predicted homology-based probable functions. Furthermore, two (TP01_1123 and TP02_0910) of the 43 HPs with GO terms were shown to be non-homologous to proteins of related species as well as the host proteome; these require further analysis as they may potentially represent drug targets and/or vaccine candidates.

Subcellular localisation predictions

The information obtained from the protein sub-cellular localisation prediction can be used to infer protein cellular functions and to find novel vaccine or drug targets [61]. Therefore, in an effort to annotate the HPs, the proteins were subjected to subcellular localisation analysis which revealed localisation of *T. parva* HPs investigated to the cytoplasm, mitochondria, and nucleus and on the membrane. About 25 HPs were identified as cytoplasmic proteins, which appeared to be unique to the *T. parva*, as the general sequence-similarity (Blast) search involving closely related Apicomplexa could not reveal any homologs; these proteins can be suitable candidates for drug development. Proteins found to localize in the cytoplasm are known to be ideal putative drug targets while those associated with the membrane are considered as potential vaccine targets [216]. Thus, pathogen proteins expressed in the cytoplasm and are non-homologous to the host proteome may possibly be facilitated as drugs to target the parasite's system without having any effect on the bovine host [217]. Cytoplasmic proteins make the best drug targets because the cytoplasm hosts several pathogenic, signalling and metabolic processes that are targets for various diseases [217]. Thus, targeting proteins that are in the cytoplasm may possibly be a valuable platform towards the discovery of drugs. Importation of proteins into the nucleus is usually facilitated

by a nuclear localisation signal (NLS). Although analysis based on the NLS detection revealed the presence of NLS in some HPs (n = 14), there are several studies that have provided evidence that some proteins may still be imported to the nucleus even if they do not contain NLS [218–220]. As a result, nine nuclear proteins did not contain the localisation signal but were however predicted to localize in the nucleus by at least two of the subcellular localisation tools used. These proteins could be interacting with other nuclear proteins that comprise a functional NLS for their import to the nucleus. In the host nucleus, parasite proteins have an opportunity to interact and bind to host DNA, and thus activate transcription of host genes to promote pathogen infection [221]. This suggests a high possibility that these proteins may play a role in the transformation of the lymphocyte by *Theileria* parasites [221].

Nuclear proteins have also been revealed as putative targets for cancer drug therapy due to involvement in the modifications of the chromatin structure, sub-nuclear compartments and protein-protein interactions [222]. The disease phenotype resulting from infection with *T. parva* is also similar to those of some cancers [207, 217], in that proteins released by *T. parva* infect the bovine host leukocytes, inducing phenotypes such as hyperproliferation, dissemination, and immortalisation [217, 218]. As suggested by Tretina and co-workers [225], since pathogen-host interactions observed in theileriosis and cancer biology are profusely comparable, cancer treatment methods should be considered in the discovery of chemotherapeutic targets against *Theileria* infections [213].

Eight percent (8%) of the 309 HPs were shown to be localized in the mitochondrion. Mitochondrial proteins are known to be involved in several complex biochemical processes [226] and they have also been identified as possible anti-cancer targets [221, 222]. The mitochondrion is crucial for normal organ and cell function; because of its principal role in energy production, it plays the main role in metabolic homeostasis. Additionally, it also plays important roles in innate immune response, apoptosis, control of cytosolic calcium ion levels, and metabolic cell signalling [215-217]. It is therefore expected that mitochondrial dysfunction is associated with many human diseases (over 100), including cancer [218-221]; it may be tempting, but not unfounded, to make an inference that the same may apply to theileriosis cases.

Some proteins move across locations and localize to multiple subcellular compartments [182]. This protein movement from one location to another allows the protein to perform multiple distinct functions. Proteins with multiple compartments were also detected in this study, and included TP02_0428, TP02_0812, TP03_0024 and TP04_0188.

Identification of GPI- anchored proteins

GPI-anchored proteins are known to be involved in immunomodulation, signalling processes, and host-pathogen response. Furthermore, GPI-anchored proteins can also act as surface antigens, membrane receptors, adhesion molecules and enzymes [236]. Six proteins encoded by HPs were identified to be GPI-anchored. GPI-anchored proteins possess both a C-terminal GPI-anchor and an N-terminal trans-membrane helix [237]. Consequently, all the six GPI-anchored proteins were also shown to comprise of trans-membrane helixes. Xue *et al.*[238] identified a GPI-anchored protein known as gp34, which is expressed specifically by schizonts of both *T. parva* and *T. annulata* [238]. Findings of their study suggested that the GPI-anchored protein may possibly contribute to the imperative parasite-host interactions during host cell division as this protein did not have homologs in non-transforming *Theileria* species. Comparably, two of the GPI-anchored proteins (TP01_0004 and TP03_0564) predicted in the current study were also non-homologous to the non-transforming *Theileria* species as well as all the related species. Worth noting is that the parasite-host interactions are very complex and involve a wide range of proteins that are unknown [238]; therefore, predicted GPI-anchored proteins detected from the current study could potentially also be responsible or contribute to the parasites' interaction with the host cell.

Identification of trans-membrane proteins

Roughly 30% of all proteins of the newly sequenced genomes represent membrane proteins [212]. Membrane proteins have important roles in several physiological functions, such as energy transduction, ion regulation and molecular recognition [212]. Regardless of the experimental difficulties of studying membrane proteins specifically in purification, crystallisation and expression, these proteins represent more than 60% of drug targets hence, understanding them is an essential key towards the discovery of novel drugs [233, 234]. Membrane-associated proteins also function as transporters and receptors; these key functions explain their suitability as vaccine candidates because of their potential to be easily

recognized by the immune system [235, 236]. However, a good vaccine candidate should also not share homology with the host proteins in order to prevent the occurrence of a possible autoimmune response [242]. Using HMMTOP and TMHMM servers, 63% of the HPs investigated in this study showed the presence of trans-membrane helices suggesting membrane-associated proteins. Unfortunately, all proteins identified from this analysis were found to be homologous to the host proteome and therefore are unsuitable to be considered as vaccine candidates.

Identification of secreted proteins

Theileria parva is an intracellular parasite that invades different types of cells to evade the host immune system. Such parasites have complex life cycles that include various developmental stages, and typically, have multiple secreted proteins that can manipulate host cell signalling pathways to promote parasite adhesion, recognition, and invasion; thus, may possibly be putative therapeutic targets [183]. In *T. parva*, proteins targeted for introduction into the host cell requires the presence of a signal peptide as they have to go through the schizont membrane [221]. Consequently, in the current study, candidates for drug development (n = 32; non-homologous to the host proteome) were identified based on the presence of the signal peptide. *Theileria parva*-secreted proteins are likely to affect the host cell environment and influence the disease phenotype. Thus, identification of *T. parva*-secreted proteins is important to better understand the pathogenesis of cattle theileriosis. Secreted proteins do not only make suitable drug targets but vaccine candidates too; a large number of promising vaccine candidates are known to either be membrane or secreted proteins as they stimulate protective immune response [241].

Detection of orthologs of some of the *T. parva* hypothetical proteins

Orthology detection is also one of the most popular platforms used for inferring functional similarity. Orthology refers to proteins/genes that have originated from a common ancestor but separated by speciation, these tend to retain a common function over evolutionary time, thus, making orthologs identification a potent tool for functional annotation [84]. In the current study, 22 HPs were shown to be orthologous to subtelomeric variable secreted proteins (SVSP), present in *T. annulata*. The majority of the SVSP-encoded proteins are expressed at the RNA level by the macroschizont stage of both *T. parva* and *T. annulata* and

contribute to the secretome [223, 243]. Moreover, they have also been suggested to play a role in the transformation of the host cell. The SVSP family has also been demonstrated to be similar to TashAT genes, another *T. annulata* 17-member protein family; both protein families contain a FAINT domain and a signal peptide. The ~70 amino acid long (FAINT) domain is frequently associated with proteins of transforming *Theileria* species, specifically *T. parva* and *T. annulata.* However, the function of this domain is unknown [243], although signal peptides predicted in these proteins suggests possible secretion into the cytoplasm of the host cell during infection. Nonetheless, *T. parva* proteins secreted in the host cell cytoplasm represent an important tool for discovering the mysteries of the *Theilerial* intracellular life as well as a better understanding of its pathogenicity mechanisms.

Contrary to the general features of the SVSP proteins, the amino acid sequence of TP03_0839 gene product did not contain any typical characteristics; however, this protein was shown to share a significant sequence identity (40%) with an SVSP protein present in *T. annulata* (TA18455). Proteins that share sequence identity above 30% have previously been reported [77] to have the likelihood of performing similar functions. Hence, based on the sequence identity observed between the two proteins, protein TP03_0839 may possibly be part of the SVSP family. Also, protein TP03_0678 contained only the FAINT domain; because of the high level of sequence identity (41%) observed between the protein and its corresponding ortholog (TA17695) it was concluded that this protein is possibly an SVSP.

Discrepancies were also observed from the analysis of TP03_0882 gene product. Previous analysis of this protein showed that it has a signal peptide (from residue 1-21) at the N-terminal region, a large C-terminal region containing two FAINT domains (from residue 146-343 and 520-577) two NLSs and a QP-rich domain [223]. However, the QP-rich domain was not detected in this protein in the current study. This inconsistency could be attributed to the fact that, in the previous study, the QP-rich region was detected using the yeast-2-hybrid system which is a well-established genetic *in vivo* method and here, an *in silico* approach was employed which is based on the amino acid sequence of a protein. Noteworthy, is that the lack of the QP-rich domain has previously been observed in other SVSPs and this may signify that this region/domain is not important for protein function as it is predicted to be unstructured [183, 217] and its function is unknown [244].

Members of the SVSP family are also known to contain functional NLSs suggesting that these parasite proteins are secreted into the nucleus and may possibly contribute to host cell phenotypic changes [223]. Consequently, the NLS was detected in 90% of the 22 SVSPs

identified in this study. It is well known that both *T. parva* and *T. annulata* invade and transform host lymphocytes, and recently it has been reported that these parasites transform the host cell by using families of secreted proteins [245] and the SVSP family being one example.

Identification of domains in sequences of the *T. parva* hypothetical proteins

A domain is a conserved part and a functional unit of a protein sequence and each has a different structure as well as function. Among domains detected from the *T. parva* HPs investigated, a zinc finger domain was identified in one of the zinc ion binding proteins (TP03_0658). A zinc finger domain is a protein domain that has multiple finger-like structures that are responsible for contact with their target molecules. Zinc finger domain-containing proteins have very diverse functions within a cell that include regulation of apoptosis and gene transcription, among others [246]. Eleven of these were also discovered to be cytoplasmic and non-homologous to the host proteome; and are listed in **Appendix A.2**. Apoptosis is a mechanism that is responsible for maintenance of cellular homeostasis in tissues, together with the immune system [247]. It is also known to contribute to the pathogenesis of a number of diseases [248]. Although the transformation of the host-lymphocytes by *T. parva* is said to involve the manipulation of the host cell signal transduction pathways that regulate apoptosis [187], the parasite proteins/factors implicated in this host phenotype are unknown [186, 207]. Thus, proteins such as TP03_0658 can further be investigated for possible involvement in the regulation of apoptosis in cattle theileriosis.

Another domain which was detected is the cytoplasmic (intracellular) domain, which has been shown to influence ligand binding of epidermal growth factor and to determine signal specificity and cellular routing characteristics [249]. Furthermore, it also known to be essential in cell adhesion for trans-membrane proteins [250].

Pain *et al.*[243], identified the presence of ~900 copies of FAINT domains in the genomes of *T. annulata* and *T. parva* [251] and Hayashida *et al.*[13] identified 686 in 137 predicted proteins of *T. orientalis*. Consistent with this, a number of HP sequences (n = 73) contained this domain in the current study. FAINT domain-containing proteins in *T. parva* and *T. annulata* usually do not contain any other domains except a signal peptide as they are said to be secretory proteins, which is congruent with export to the host cell [186, 237]. Similarly,

most of the FAINT domain-containing proteins predicted in the study were found to consist of signal peptides, suggesting possible export to the host cell.

The FAINT domains were formerly known as DUF529 and there is still a huge amount of protein domains (more than 20%) assigned as “domains of unknown function” (DUFs), of which 1,500 are found in eukaryotes [252]. Using a simple subtractive procedure on transferred annotations, Goodacre and co-workers [252] investigated functions that might be associated with ten specific bacterial DUFs. Although functional annotation was not the main focus of their study, they were able to collect probable functional attributes for the proteins containing DUFs, which indicated membrane association, based on subcellular location. This partially explains why the function of these domains still remains unknown, given the struggle of studying membrane proteins [252]. Membrane-associated proteins are highly important because of their role in various diseases and they are also considered as major therapeutic targets [253]. Thus the DUFs may represent interesting targets for future research and may have essential functions.

Virulence factors detection

Various bioinformatics tools such as VICMpred, VirulentPred and MP3 were used to predict virulence factors in this group of 309 HPs. Worth mentioning is that, currently, the majority of virulence prediction tools that are freely available are trained for bacterial datasets, including tools used in this study. Nevertheless, it should also be noted that other studies investigating non-bacterial sequences have previously employed bacterial-trained tools and successfully identified virulence factors in these organisms. For instance, Adebayo *et al.* [254] employed VirulentPred for the annotation of virulence factors in Schistosomes, where 72% of the protein entries were shown to have virulence factors [254]. Coincidentally, in the current study, virulence factors were detected in 72% of the investigated proteins; virulence factors have been reported to allow the pathogen to survive within the host and therefore enhancing its ability to cause a disease [116]. Furthermore, these virulence proteins are also known to be responsible for attachment of the pathogen to the host surfaces, evasion of host defence mechanisms, colonisation of the host cell, immunosuppression, and in intracellular pathogens, they mediate the entry and exit in host cells [199, 248, 249]. It has previously been hypothesized that the virulence proteins hydrophobic residues support their integration

into the membrane and stimulate binding of proteins to targets in the hosts[187, 250]. Consequently, the HPs predicted to be virulence protein may be essential in the establishment of the disease within a susceptible host and may thus provide potent pathogen-specific therapeutic targets to control *T. parva* infection. It is no surprise that so many HPs with virulence factors were detected since the majority of HPs were selected from genes with up-regulated expression in the cattle-derived parasite isolate of *T. parva* (*T. parva* Muguga), which was maintained in bovine lymphoblasts, known to cause the fatal bovine disease, ECF.

CONCLUSION

The hypothesis for this study was that HPs encoded by DETs, previously identified from a transcriptome profiling study, are likely to play a significant role in the disease outcome resulting from infections by different isolates of *T. parva*. Using *in silico* approaches, probable functions were successfully assigned to 277(90%) of the 309 HPs analysed, including enzymes, membrane-associated proteins, transcription factors, secreted proteins, and proteins with virulence factors, among others. Accordingly and in support of the hypothesis, these proteins could have various functions significant to the pathogenesis of cattle theileriosis including the attachment of the pathogen to the host surfaces, disruption of the host signal pathways, colonisation of the host cell, immunosuppression, host cell phenotype modulation and proliferation. Membrane proteins may also serve as antigenic proteins; thus, can be explored for the development of vaccines against cattle theileriosis caused by *T. parva*. This study also facilitated the identification of putative therapeutic targets, which did not have homologs to any of the vertebrate host proteins that may be considered for control of the *T. parva* infections. The HPs with predicted biological roles of interest should be further explored experimentally to confirm their roles in cattle theileriosis.

STUDY LIMITATIONS

Although *in silico* prediction methods have advantages already mentioned earlier, these methods are also inevitably prone to error. Most researchers are tempted to assign a single functional role to a protein while many proteins are multifunctional; this may result in the annotation of incorrect or incomplete information [258]. Also, *in silico* methods depend largely on the information in the sequence of the query protein and the availability of analogous information. Thus, another major limitation is with regard to the pair-wise sequence similarity search, whereby the best hit can also be a protein of unknown function or may be incorrectly predicted, which can lead to inaccurate annotation of the HP [258]. This highlights the importance of validating the predicted biological roles assigned by employing *in silico* analysis, by *in vitro* and/or *in vivo* experiments [259].

In the case of sequence similarity, where a best hit is also an uncharacterised protein, the biological role can also be determined by its native 3D structure [144]. During evolution, protein structures are often better conserved than the sequence [141]; thus, protein 3D structures can be used to infer function, where sequence-based approaches show limitations [140]. Consequently, SWISS-MODEL was used to predict the 3D structures of 32 HPs in order to use structural similarities as another platform to annotate their function. However, this approach was unsuccessful because of lack of structural similarities with annotated proteins in available structural databases. Protein pairs require at least ≥30% sequence identity to share structural similarities [260]. Using multiple databases that employ different algorithms may have resulted in a favorable output. The results obtained may also suggest proteins unique to the parasite and these can be exploited in drug discovery or identification of novel vaccine candidates, of course taking into consideration the relevant criteria.

In another attempt to overcome the limitations experienced with sequence similarity analysis, the current study further attempted to use protein-protein network predictions to assign functions to some of the HPs. Prediction of interaction partners has previously been used to assign functions of uncharacterised proteins [124, 185, 254, 255]. Consequently, STRING database was utilised to predict putative protein-protein interaction partners of HPs under investigation. STRING database revealed that only three proteins has interaction partners including TP01_0124, TP03_0738, and TP04_0172; each of these had two functional

partners, however, all the identified partners were also uncharacterised proteins. As a result, this information also could not be used to infer functions of the target proteins.

The receiver operating characteristics (ROC) analysis was conducted in order to assess the performance of the prediction tools used in the current study. This analysis is based on true-positive rate against the corresponding false-positive rate. It plays an essential role in high-throughput projects such as structural genomics and genome sequencing [263]. Therefore, the evaluation of the credibility of classifiers has become fundamental to ensure data quality. Receiver operating characteristics analysis is the most popular of the many performance assessment methods available [264]. This is because in contrast to simple performance indices; it provides a visual and a numerical summary of the classifiers performance. It is also becoming a predominant technique in the bioinformatics community. However, like any other evaluation metric, conducting ROC analysis accurately requires knowing its characteristics and limitations [263]. This analysis is conceptually simple; however, there are some common misconceptions and difficulties that may be experienced when using it in research. Receiver operating characteristics analysis is predominantly aimed for binary classifiers, where datasets can be safely identified [265]. If there are unknown cases, ROC analysis can be applied only to a manually curated subset, and this may be questionable whether or not a subset is adequate to represent the variability of an entire database. Hence, it has been recommended that, having the same number of positive and negative examples may be favorable when evaluating binary classifiers [266]. However, this condition is almost impossible in bioinformatics because there are usually fewer positives than negatives [263]. It is recommended that performance assessment should also involve other comparison measures into the evaluation process other than only relying on one method.

These few examples highlight the limitations of *in silico* methods, showing that not all uncharacterised proteins can be annotated through this approach; hence, experimental work is necessary. Nonetheless, there is information that is obtained from *in silico* analysis which can be useful in *in vitro* and *in vivo* experimental studies, such as subcellular localisation, protein domains and physiochemical properties.

FUTURE RECOMMENDATIONS

Among important proteins identified from annotation of the selected *T. parva* HPs were non-homologous cytoplasmic proteins and virulent proteins which represent an important set of proteins that could be exploited for future drug design. The former is important in drug design since drug development largely focuses on pathogen proteins that are non-homologous to the host proteome as the effect of the drug should be solely on the parasite and not any other aspect of the host biology. On the other hand, virulence proteins are known to be involved in the establishment of the pathogen inside the host.

Secreted proteins were also detected; these are also known to manipulate host cell signalling pathways to promote parasite adhesion, recognition, and host invasion [171]. Proteins identified in this study can be further investigated experimentally to determine whether they fulfill these roles in *T. parva* infections as these are important processes in pathogenesis and disease progression. Secreted proteins can also be investigated for possible drug targets. Generally, this study recommends the recombinant expression and experimental biological characterisation of the newly annotated secreted proteins. The roles of these secreted and virulent proteins in the *T. parva* parasite life cycle need to be defined and maybe manipulated for control of *T. parva* infections.

REFERENCES

1. Muhanguzi D, Picozzi K, Hatendorf J, Thrusfield M, Welburn SC, Kabasa JD, *et al*. Prevalence and spatial distribution of *Theileria parva* in cattle under crop-livestock farming systems in Tororo District, Eastern Uganda. Parasitology Vectors. 2014;7: 91. doi:10.1186/1756-3305-7-91

2. Waladde SM, Young AS, Ochieng SA, Mwaura SN, Mwakima FN. Transmission of *Theileria parva* to cattle by *Rhipicephalus appendiculatus* adults fed as nymphae *in vitro* on infected blood through an artificial membrane. Parasitology. 1993; 249–56. do:org/10.1017/S0031182000079221

3. Muleya W, Namangala B, Simuunza M, Nakao R, Inoue N, Kimura T, *et al*. Population genetic analysis and sub-structuring of *Theileria parva* in the northern and eastern parts of Zambia. Parasit Vectors. BioMed Central; 2012;5: 255. doi:10.1186/1756-3305-5-255

4. Anonymous. The eradication of East Coast fever in South Africa. J S Afr Vet Assoc. 1981;52: 71–73.

5. Mukhebi A, Perry B, Kruska R. Estimated economics of theileriosis control in Africa. Prev Vet Med. 1992; 12(1-2): 73-85. doi:org/10.1016/0167-5877(92)90070-V

6. Marcotty T, Brandt J, Billiouw M, Chaka G, Losson B, Berkvens D. Immunisation against *Theileria parva* in eastern Zambia: influence of maternal antibodies and demonstration of the carrier status. Vet Parasitol. 2002;110: 45–56.doi:10.1016/s0304-4017(02)00314-x

7. Potgieter FT, Stoltsz WH, Blouin EF, Roos J a. Corridor disease in South Africa: a review of the current status. J S Afr Vet Assoc. 1988;59: 155–160.

8. Lawrence JA, Perry BD, Williamson SM. East Coast Fever. Infect Dis Livest. 2004; 448–467. doi:10.1007/0-306-48380-7_1323

9. Uilenberg G. *Theilerial* species of domestic livestock. Adv Control Theileriosis. 1981; Available: http://link.springer.com/chapter/10.1007/978-94-009-8346-5_2

10. Perry BD, Young AS. The naming game: the changing fortunes of East Coast fever and *Theileria parva*. Vet Rec. BMJ Publishing Group Limited; 1993;133: 613–6. doi:10.1136/VR.133.25-26.613

11. Collins NE, Allsopp BA. *Theileria parva* ribosomal internal transcribed spacer sequences exhibit extensive polymorphism and mosaic evolution: application to the characterization of parasites from cattle and buffalo. Parasitology. 1999; 541–51.

12. Hayashida K, Abe T, Wier W, Nakao R, Ito K, Kajino K, *et al*. Whole-genome sequencing of *Theileria parva* strains provides insight into parasite migration and diversification in the African continent. Oxford University Press; 2013; 20(3):209-20. doi: 10.1093/dnares/dst003

13. Hayashida K, Hara Y, Abe T, Yamasaki C, Toyoda A, Kosuge T, *et al*. Comparative Genome Analysis of Three Eukaryotic Parasites with Differing Abilities To Transform Leukocytes Reveals Key Mediators of *Theileria*-Induced Leukocyte Transformation. MBio. American Society for Microbiology; 2012;3: e00204-12-e00204-12. doi:10.1128/mBio.00204-12

14. Bishop R, Musoke A, Morzaria S, Gardner M, Nene V. *Theileria*: intracellular protozoan parasites of wild and domestic ruminants transmitted by ixodid ticks. Parasitology. 2004;129: S271– S283. doi:10.1017/S0031182003004748

15. Hayashida K, Hara Y, Abe T, Yamasaki C, Toyoda A, Kosuge T, *et al*. Comparative genome analysis of three eukaryotic parasites with differing abilities to transform leukocytes reveals key mediators of *Theileria*-induced leukocyte transformation. MBio. 2012;3: e00204-12. doi:10.1128/mBio.00204-12

16. Escalante AA, Ayala FJ. Evolutionary origin of *Plasmodium* and other Apicomplexa based on rRNA genes. Proc Natl Acad Sci U S A. National Academy of Sciences; 1995; 92: 5793–7.

17. Allsopp MT, Cavalier-Smith T, De Waal DT, Allsopp BA. Phylogeny and evolution of the piroplasms. Parasitology. 1994; 147–52. doi:10.1016/j.2007.04.002

18. Morrison WI, McKeever DJ. Current status of vaccine development against *Theileria* parasites. Parasitology. 2006;133 Suppl: S169-87. doi:10.1017/S0031182006001867

19. Devos AJ, Roos JA. The isolation of *Theileria taurotragi* in South Africa. Onderstepoort J vet Res. 1981; 48: 149–153.

20. Berger J. *Theileria velifera* demonstrated in cattle in the Eastern Cape Province of the Republic of South Africa. J S Afr Vet Assoc. 1979;50: 45–46. ISSN: 1019-9128

21. Olwoch JM, Reyers B, Engelbrecht FA, Erasmus BFN. Climate change and the tick-borne disease, Theileriosis (East Coast fever) in sub-Saharan Africa. J Arid Environ. 2008;72: 108–120. doi:10.1016/j.jaridenv.2007.04.003

22. Lawrence JA, Perry BD, Williamson SM. Zimbabwe theileriosis. Infect Dis Livest. 2004; 472–474.

23. Splitter EJ. *Theileria mutans* associated with bovine anaplasmosis in the United States. J Am Vet Med Assoc. 1950;117: 134–5. Available: http://www.ncbi.nlm.nih.gov/pubmed/15428394

24. Cossio-Bayugar R, Pillars R, Schlater J, Holman PJ. *Theileria buffeli* infection of a Michigan cow confirmed by small subunit ribosomal RNA gene analysis. Vet Parasitol. 2002; 105: 105–110. doi: org/10.1016/S0304-4017(02)00003-1

25. Zakkyeh T, Mohammad Ali O, Nasibeh HV, Mohammad Reza YE, Farhang B, Fatemeh M. First molecular detection of *Theileria ovis* in *Rhipicephalus sanguineus* tick in Iran. Asian Pac J Trop Med. 2012;5: 29–32. doi:10.1016/S1995-7645(11)60240-X

26. Katzer F, Ngugi D, Oura C, Bishop RP, Taracha ELN, Walker AR, *et al*. Extensive Genotypic Diversity in a Recombining Population of the Apicomplexan Parasite *Theileria parva*. Infect Immun. American Society for Microbiology; 2006;74: 5456–5464. doi:10.1128/IAI.00472-06

27. Young A. The epidemiology of theileriosis in East Africa. Adv Control Theileriosis. 1981; 14: 38-55. doi:10.1007/978-94-009-8346-5_3

28. Lawrence JA, Norval RAI, Uilenberg G. *Rhipicephalus zambeziensis* as a vector of bovine theileriae. Trop Anim Health Prod. Kluwer Academic Publishers-Human Sciences Press; 1983;15: 39–42. doi:10.1007/BF02250760

29. Gardner MJ. Genome Sequence of *Theileria parva*, a Bovine Pathogen That Transforms Lymphocytes. Science (80-). 2005;309: 134–137. doi:10.1126/science.1110439

30. Toye P, Musoke A, Naessens J. Role of the polymorphic immunodominant molecule in entry of *Theileria parva* sporozoites into bovine lymphocytes. Infect Immun. American Society for Microbiology (ASM); 2014;82: 1786–92. doi:10.1128/IAI.01029-13

31. Fry LM, Schneider DA, Frevert CW, Nelson DD, Morrison WI, Knowles DP. East Coast Fever Caused by *Theileria parva* Is Characterized by Macrophage Activation Associated with Vasculitis and Respiratory Failure. PLoS One. Public Library of Science; 2016;11: e0156004. doi:10.1371/journal.pone.0156004

32. Lawrence JA. History of bovine theileriosis in southern Africa; 1992:1–39. doi:19920511973

33. Collins NE, Allsopp MTEP, Allsopp BA. Molecular diagnosis of theileriosis and heartwater in bovines in Africa. Trans R Soc Trop Med Hyg. No longer published by Elsevier; 2002;96: S217–S224. doi:10.1016/S0035-9203(02)90079-9

34. Stoltsz H. *Theileria parva* infections. 2013. 21: 3797–3809. doi:10.1000/1522-2683

35. Neitz WO, Canham AS, Kluge EB. Corridor disease: a fatal form of bovine theileriosis encountered in Zululand. 1955; 26(2): 79–87.

36. Mbizeni S, Potgieter FT, Troskie C, Mans BJ, Penzhorn BL, Latif AA. Field and laboratory studies on Corridor disease (*Theileria parva* infection) in cattle population at the livestock/game interface of uPhongolo-Mkuze area, South Africa. Ticks Tick Borne Dis. 2013;4: 227–234. doi:10.1016/j.ttbdis.2012.11.005

37. Matson BA. Theileriosis in Rhodesia: 1. A study of diagnostic specimens over two seasons. J S Afr Vet Assoc. AOSIS; 1967;38: 93–102. ISSN:1079-9128

38. Norval RAI, Perry BD, Young AS. The epidemiology of theileriosis in Africa. Academic Press Limited; 1992; ISBN 0-12-521740-1

39. Walker AR, Katzer F, Ngugi D, McKeever D. Cloned *Theileria parv* produces lesser infections in ticks compared to uncloned *T. parva* & despite similar infections in cattle : research communication. Onderstepoort J Vet Res. 2006;73. doi:10.4102/ojvr.v73i2.163

40. Goeyse I De, Jansen F, Madder M, Hayashida K, Berkvens D, Dobbelaere D, *et al*. Veterinary Parasitology Transfection of live , tick derived sporozoites of the protozoan Apicomplexan parasite *Theileria parva*. Vet Parasitol. Elsevier B.V.; 2015;208: 238–241. doi:10.1016/j.vetpar.2015.01.013

41. Gauer M, Mackenstedt U, Mehlhorn H, Schein E, Zapf F, Njenga E, *et al*. DNA measurements and ploidy determination of developmental stages in the life cycles of *Theileria annulata* and *T. parva*. Parasitol Res. Springer-Verlag; 1995;81: 565–574. doi:10.1007/BF00932023

42. Konnai S, Imamura S, Nakajima C, Witola WH, Yamada S, Simuunza M, *et al*. Acquisition and transmission of *Theileria parva* by vector tick, *Rhipicephalus appendiculatus*. Acta Trop. 2006;99: 34–41. doi:10.1016/j.actatropica.2006.06.008

43. Schetters TPM, Arts G, Niessen R, Schaap D. Development of a new score to estimate clinical East Coast Fever in experimentally infected cattle. Vet Parasitol. 2010;167: 255–259. doi:10.1016/j.vetpar.2009.09.027

44. Brown CG, Radley DE, Burridge MJ, Cunningham MP. The use of tetracyclines on the chemotherapy of experimental East Coast Fever (*Theileria parva* infection of cattle). Tropenmed Parasitol. 1977;28: 513–20.

45. Sivashankari S, Shanmughavel P. Functional annotation of hypothetical proteins – A review. Bioinformation. 2006;1: 335–338. doi:10.6026/97320630001335

46. Kairo A, Fairlamb AH, Gobright E, Nene V. A 7.1 kb linear DNA molecule of *Theileria parva* has scrambled rDNA sequences and open reading frames for mitochondrially encoded proteins. EMBO J. 1994;13 (4): 898–905.

47. Arisue N, Hashimoto T, Mitsui H, Palacpac NMQ, Kaneko A, Kawai S, *et al*. The *Plasmodium* Apicoplast Genome: Conserved Structure and Close Relationship of P. ovale to Rodent Malaria Parasites. Mol Biol Evol. 2012;29: 2095–2099. doi:10.1093/molbev/mss082

48. Brocchieri L. Low-Complexity Regions in *Plasmodium* Proteins: In Search of a Function. Genome Res. 2001. 11: 195-197. doi:10.1101/gr.176401.

49. Hikosaka K, Watanabe Y i., Tsuji N, Kita K, Kishine H, Arisue N, *et al*. Divergence of the Mitochondrial Genome Structure in the Apicomplexan Parasites, *Babesia* and *Theileria*. Mol Biol Evol. 2010;27: 1107–1116. doi:10.1093/molbev/msp320

50. Waller RF, McFadden GI. The apicoplast: a review of the derived plastid of apicomplexan parasites. Curr Issues Mol Biol. 2005;7: 57–79.

51. Bishop R. Analysis of the transcriptome of the protozoan *Theileria parva* using MPSS reveals that the majority of genes are transcriptionally active in the schizont stage. Nucleic Acids Res. 2005;33: 5503–5511. doi:10.1093/nar/gki818

52. Brenner S. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol. Nat Biotechnol. 2000;18: 1021–1021. doi: 10.1038/76469

53. Roberts RJ. Identifying Protein Function—A Call for Community Action. PLoS Biol. Public Library of Science; 2004;2: e42.doi:10.1371/journal.pbio.0020042

54. Mehmood T, Martens H, Sæbø S, Warringer J, Snipen L. Mining for genotype-phenotype relations in Saccharomyces using partial least squares. BMC Bioinformatics. BioMed Central; 2011;12: 318. doi:10.1186/1471-2105-12-318

55. Benso A, Di Carlo S, ur Rehman H, Politano G, Savino A, Suravajhala P. A combined approach for genome wide protein function annotation/prediction. Proteome Sci. BioMed Central; 2013;11: S1. doi:10.1186/1477-5956-11-S1-S1

56. Galperin MY. Conserved "Hypothetical" Proteins: New Hints and New Puzzles. Comp Funct Genomics. 2001;2: 14–18.doi:10.1002/cfg.66

57. Paul S, Saha M, Bhoumik NC, Talukdar SN. *In silico* Structural and Functional Annotation of *Mycoplasma genitalium* Hypothetical Protein MG_377. Int J Bioautomation. Academic Publishing House ; 2015;19: 15–24.

58. Minion FC, Lefkowitz EJ, Madsen ML, Cleary BJ, Swartzell SM, Mahairas GG. The genome sequence of *Mycoplasma hyopneumoniae* strain 232, the agent of swine mycoplasmosis. J Bacteriol. American Society for Microbiology (ASM); 2004;186: 7123–33. doi:10.1128/JB.186.21.7123-7133

59. Lubec G, Afjehi-Sadat L, Yang J-W, John JPP. Searching for hypothetical proteins: theory and practice based upon original data and literature. Prog Neurobiol. 77: 90–127. doi:10.1016/j.pneurobio.2005.10.001

60. Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, Linial M, *et al*. Protein function annotation by homology-based inference. Genome Biol. 2009;10: 207. doi:10.1186/gb-2009-10-2-207

61. Kumar K, Prakash A, Tasleem M, Islam A, Ahmad F, Hassan MI. Functional annotation of putative hypothetical proteins from *Candida dubliniensis*. Gene. Elsevier B.V.; 2014;543: 93–100. doi:10.1016/j.gene.2014.03.060

62. Al-Khafaji ZM. *In Silico* Investigation of Rv Hypothetical Proteins of Virulent Strain *Mycobacterium tuberculosis* H37Rv. Indian Journal of Pharmaceutical & Biological Research. 2013; 1 (4): 81-88. doi: 123456789/157258

63. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, *et al*. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2006;34: D173-80. doi:10.1093/nar/gkj158

64. Sael L, Chitale M, Kihara D. Structure- and sequence-based function prediction for non-homologous proteins. J Struct Funct Genomics. 2012;13: 111–23.doi:10.1007/s10969-012-9126-6

65. Wang Z, Cao R, Cheng J, Martin D, Berriman M, Barton G, *et al*. Three-Level Prediction of Protein Function by Combining Profile-Sequence Search, Profile-Profile Search, and Domain Co-Occurrence Networks. BMC Bioinforma 2013 143. BioMed Central; 2013;14: 178. doi:10.1186/1471-2105-14-S3-S3

66. Laskowski RA, Watson JD, Thornton JM. ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res. 2005;33: W89–W93.doi:10.1093/nar/gki414

67. Glazer DS, Radmer RJ, Altman RB. Improving Structure-Based Function Prediction Using Molecular Dynamics. Structure. 2009;17: 919–929. doi:10.1016/j.str.2009.05.010

68. Kotaru AR, Joshi RC. Classification of Phylogenetic Profiles for Protein Function Prediction: An SVM Approach. Springer Berlin Heidelberg; 2009. pp. 510–520. doi:10.1007/978-3-642-03547-0_49

69. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. National Academy of Sciences; 1998; 95: 14863–8.

70. Zhou X, Kao M-CJ, Wong WH. Transitive functional annotation by shortest-path analysis of gene

expression data. Proc Natl Acad Sci. National Academy of Sciences; 2002;99: 12783–12788. doi:10.1073/pnas.192159399

71. Beer MA, Tavazoie S. Predicting Gene Expression from Sequence. Cell. 2004;117: 185–198. doi:10.1016/S0092-8674(04)00304-6

72. Zhao X-M, Wang Y, Chen L, Aihara K, Chien C, Bartel P, *et al.* Gene function prediction using labeled and unlabeled data. BMC Bioinformatics. BioMed Central; 2008;9: 57.doi:10.1186/1471-2105-9-57

73. Salavati R, Najafabadi HS. Sequence-based functional annotation: what if most of the genes are unique to a genome? Trends Parasitol. 2010;26: 225–229. doi:10.1016/j.pt.2010.02.001

74. Whisstock JC, Lesk AM. Prediction of protein function from protein sequence and structure. Q Rev Biophys. 2003;36: 307–340.doi:10.1017/S0033583503003901

75. Watson JD, Laskowski RA, Thornton JM. Predicting protein function from sequence and structural data. Curr Opin Struct Biol. 2005;15: 275–284. doi:10.1016/j.sbi.2005.04.003

76. Skolnick J, Fetrow JS. From genes to protein structure and function: Novel applications of computational approaches in the genomic era. Trends in Biotechnology. 2000. doi:10.1016/S0167-7799(99)01398-0

77. Pearson WR. An introduction to sequence similaritysearching. Curr Protoc Bioinformatics. NIH Public Access; 2013;Chapter 3: Unit3.1. doi:10.1002/0471250953.bi0301s42

78. Nagar A, Hahsler M, Altschul S, Gish W, Miller W, Myers E, *et al*. Fast discovery and visualization of conserved regions in DNA sequences using quasi-alignment. BMC Bioinforma 2013 1411. BioMed Central; 2013;14: 403–410.doi:10.1186/1471-2105-14-S11-S2

79. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A. 1988;85: 2444–8.

80. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, *et al*. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7: 539. doi:10.1038/msb.2011.75

81. Lindahl EA. Identification of related proteins on family, superfamily and fold level. J Mol Biol . 2000; 295: 613–25. doi: org/10.1006/jmbi.1999.3377

82. Sauder JM, Arthur JW, Dunbrack RL. Large-scale comparison of protein sequence alignment algorithms with structure alignments. Proteins. 2000;40: 6–22. doi: 10.1002/(SICI)1097-0134(20000701)40:1<6::AID-PROT30>3.0.CO;2-7

83. Altschul S. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25: 3389–3402.doi:10.1093/nar/25.17.3389

84. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13: 2178–89.doi:10.1101/gr.1224503

85. Fitch WM. Distinguishing Homologous from Analogous Proteins. Syst Zool. 1970;19: 99. doi:10.2307/2412448

86. Koonin E V, Galperin MY. Evolutionary Concept in Genetics and Genomics. Kluwer Academic. 2003; 25-49.doi.org/10.1007/978-1-4757-3783-7_3

87. Tatusov RL, Natale DA, Garkavtsev I V, Tatusova TA, Shankavaram UT, Rao BS, *et al*. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 2001; 29: 22–8. doi: org/10.1093/nar/29.1.22

88. Mushegian AR, Garey JR, Martin J, Liu LX. Large-Scale Taxonomic Profiling of Eukaryotic Model Organisms: A Comparison of Orthologous Proteins Encoded by the Human, Fly, Nematode, and Yeast Genomes. Genome Res. 1998;8: 590–598.doi:10.1101/gr.8.6.590

89. Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, *et al*. Comparative genomics of the eukaryotes. Science. 2000;287: 2204–15. doi: 10.1126/science.287.5461.2204

90. Aurrecoechea C, Brestelli J, Brunk BP, Fischer S, Gajria B, Gao X, *et al*. EuPathDB: a portal to eukaryotic pathogen databases. Nucleic Acids Res. 2010;38:D415-9. doi:10.1093/nar/gkp941

91. Fischer S, Aurrecoechea C, Brunk BP, Gao X, Harb OS, Kraemer ET, *et al*. The Strategies WDK: a graphical search interface and web development kit for functional genomics databases. Database (Oxford). Oxford University Press; 2011;2011: bar027. doi:10.1093/database/bar027

92. Aurrecoechea C, Barreto A, Brestelli J, Brunk BP, Caler E V, Fischer S, *et al*. AmoebaDB and MicrosporidiaDB: functional genomic resources for *Amoebozoa* and *Microsporidia* species. Nucleic Acids Res. Oxford University Press; 2011;39:D612-9. doi:10.1093/nar/gkq1006

93. Zhang S, Chen H, Liu K, Sun Z. Inferring protein function by domain context similarities in protein-protein interaction networks. BMC Bioinformatics. BioMed Central; 2009;10: 395. doi:10.1186/1471-2105-10-395

94. Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA. Structure, function and evolution of multidomain proteins. Curr Opin Struct Biol. 2004;14: 208–16. doi:10.1016/j.sbi.2004.03.011

95. Reid AJ, Yeats C, Orengo CA. Methods of remote homology detection can be combined to increase coverage by 10% in the midnight zone. Bioinformatics. 2007;23: 2353–60. doi:10.1093/bioinformatics/btm355

96. Apic G1, Gough J TS. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J Mol Biol. 2001;310: 311–25. doi: org/10.1006/jmbi.2001.4776

97. Traut T, Traut T. Multidomain Proteins. eLS. Chichester, UK: John Wiley & Sons, Ltd; 2014. doi:10.1002/9780470015902.a0005053

98. Letunic I, Doerks T, Bork P. SMART 7: recent updates to the protein domain annotation resource. Nucleic Acids Res. 2012;40: D302-5. doi:10.1093/nar/gkr931

99. Zdobnov EM, Apweiler R. InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics. 2001;17: 847–8. doi:org/10.1093/bioinformatics/17.9.847

100. Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, *et al*. New and continuing developments at PROSITE. Nucleic Acids Res. Oxford University Press; 2013;41: D344-7. doi:10.1093/nar/gks1067

101. Sigrist CJA, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, *et al*. PROSITE: A documented database using patterns and profiles as motif descriptors. Brief Bioinform. Oxford University Press; 2002;3: 265–274. doi:10.1093/bib/3.3.265

102. Sigrist CJA, De Castro E, Langendijk-Genevaux PS, Le Saux V, Bairoch A, Hulo N. ProRule: a new database containing functional and structural information on PROSITE profiles. Bioinformatics. 2005;21: 4060–6. doi:10.1093/bioinformatics/bti614

103. Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res. 2002;30: 281–3. https://doi.org/10.1093/nar/30.1.281

104. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, *et al*. The Pfam protein families database. Nucleic Acids Res. 2010;38: D211-22. doi:10.1093/nar/gkp985

105. Heger A, Wilton CA, Sivakumar A, Holm L. ADDA: a domain database with global coverage of the protein universe. Nucleic Acids Res. Oxford University Press; 2005;33: D188-91. doi:10.1093/nar/gki096

106. Yu C-S, Chen Y-C, Lu C-H, Hwang J-K. Prediction of protein subcellular localization. Proteins. 2006;64: 643–51. doi:10.1002/prot.21018

107. Binder JX, Pletscher-Frankild S, Tsafou K, Stolte C, O'Donoghue SI, Schneider R, *et al*. COMPARTMENTS: unification and visualization of protein subcellular localization evidence. Database (Oxford). Oxford University Press; 2014;2014: bau012. doi:10.1093/database/bau012

108. Bell AW, Ward MA, Blackstock WP, Freeman HN, Choudhary JS, Lewis AP, *et al*. Proteomics characterization of abundant Golgi membrane proteins. J Biol Chem. 2001;276: 5152–65. doi:10.1074/jbc.M006143200

109. Andersen JS, Lyon CE, Fox AH, Leung AKL, Lam YW, Steen H, *et al*. Directed proteomic analysis of the human nucleolus. Curr Biol. 2002; 12: 1–11. doi: org/10.1016/S0960-9822(01)00650-9

110. Herold N, Will CL, Wolf E, Kastner B, Urlaub H, Lührmann R. Conservation of the protein composition and electron microscopy structure of *Drosophila melanogaster* and human spliceosomal complexes. Mol Cell Biol. American Society for Microbiology (ASM); 2009;29: 281–301. doi:10.1128/MCB.01415-08

111. Kumar A, Agarwal S, Heyman JA, Matson S, Heidtman M, Piccirillo S, *et al*. Subcellular localization of the yeast proteome. Genes Dev. Cold Spring Harbor Laboratory Press; 2002;16: 707–19. doi:10.1101/gad.970902

112. Simpson JC, Wellenreuther R, Poustka A, Pepperkok R, Wiemann S. Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. EMBO Rep. European Molecular Biology Organization; 2000;1: 287–92. doi:10.1093/embo-reports/kvd058

113. Horton P, Park K-J, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, *et al*. WoLF PSORT: protein localization predictor. Nucleic Acids Res. 2007;35: W585-7. doi:10.1093/nar/gkm259

114. Briesemeister S, Rahnenführer J, Kohlbacher O. YLoc--an interpretable web server for predicting subcellular localization. Nucleic Acids Res. 2010;38: W497-502. doi:10.1093/nar/gkq477

115. Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence. J Mol Biol. 2000;300: 1005–1016. doi:10.1006/jmbi.2000.3903

116. Chaudhuri R, Ramachandran S. Prediction of virulence factors using bioinformatics approaches. Methods in molecular biology (Clifton, NJ). 2014. pp. 389–400. doi:10.1007/978-1-4939-1115-8_22

117. Damte D, Suh J-W, Lee S-J, Yohannes SB, Hossain MA, Park S-C. Putative drug and vaccine target protein identification using comparative genomic analysis of KEGG annotated metabolic pathways of *Mycoplasma hyopneumoniae*. Genomics. 2013;102: 47–56. doi:10.1016/j.ygeno.2013.04.011

118. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001;305: 567–80. doi:10.1006/jmbi.2000.4315

119. Tusnady GE, Simon I. The HMMTOP transmembrane topology prediction server. Bioinformatics. 2001;17: 849–850. doi:10.1093/bioinformatics/17.9.849

120. Ofran Y. Analysing six types of protein-protein interfaces. J Mol Biol. 2003;325: 377–87. doi.org/10.1016/S0022-2836(02)01223-8

121. Yanagida M. Functional proteomics; current achievements. J Chromatogr B. 2002;771: 89–106. doi:10.1016/S1570-0232(02)00074-0

122. Berggård T, Linse S, James P. Methods for the detection and analysis of protein–protein interactions. Proteomics. WILEY-VCH Verlag; 2007;7: 2833–2842. doi:10.1002/pmic.200700131

123. Deng M, Zhang K, Mehta S, Chen T, Sun F. Prediction of Protein Function Using Protein–Protein Interaction Data. Mary Ann Liebert, Inc. 2004; 10(6): 947-960. doi:org/101089/106652703322756168.

124. Letovsky S, Kasif S. Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics. 2003; 19(1): i197-204. doi:org/10.1093/bioinformatics/btg1026

125. Karaoz U, Murali TM, Letovsky S, Zheng Y, Ding C, Cantor CR, *et al*. Whole-genome annotation by using evidence integration in functional-linkage networks. Proc Natl Acad Sci. National Academy of Sciences; 2004;101: 2888–2893. doi:10.1073/pnas.0307326101

126. Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics. 2005;21 Suppl 1: i302-10. doi:10.1093/bioinformatics/bti1054

127. Zhang S, Chen H, Liu K, Sun Z, Fields S, Aebersold R, *et al*. Inferring protein function by domain context similarities in protein-protein interaction networks. BMC Bioinformatics. BioMed Central; 2009;10: 395. doi:10.1186/1471-2105-10-395

128. Rao VS, Srinivas K, Sujini GN, Kumar GNS, Rao VS, Srinivas K, *et al*. Protein-protein interaction detection: methods and analysis. Int J Proteomics. Hindawi Publishing Corporation; 2014;2014: 147648. doi:10.1155/2014/147648

129. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, *et al*. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. Oxford University Press; 2011;39: D561-8. doi:10.1093/nar/gkq973

130. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, *et al*. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. Oxford University Press; 2015;43: D447-52. doi:10.1093/nar/gku1003

131. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, *et al*. Gene Ontology: tool for the unification of biology. Nat Genet. Nature Publishing Group; 2000;25: 25–29. doi:10.1038/75556

132. Kourmpetis YA, van Dijk AD, Ter Braak CJ. Gene Ontology consistent protein function prediction: the FALCON algorithm applied to six eukaryotic genomes. Algorithms Mol Biol. 2013;8: 10. doi:10.1186/1748-7188-8-10

133. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talin M, Robles M. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21: 3674–3676. doi:10.1093/bioinformatics/bti610

134. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215: 403–10. doi:10.1016/S0022-2836(05)80360-2

135. Xu Y, Ding J, Wu L-Y. iSulf-Cys: Prediction of S-sulfenylation Sites in Proteins with Physicochemical Properties of Amino Acids. Liu B, editor. PLoS One. Public Library of Science; 2016;11: e0154237. doi:10.1371/journal.pone.0154237

136. Sivakumar K, Balaji S. *In silico* characterization of antifreeze proteins using computational tools and servers. J Chem Sci. 2007; 119: 571–579. doi:org/10.1007/s12039-007-0072-y

137. Grasso EJ, Sottile AE, Coronel CE. Structural Prediction and In Silico Physicochemical Characterization for Mouse Caltrin I and Bovine Caltrin Proteins. Bioinform Biol Insights. Libertas Academica; 2016;10: 225–236. doi:10.4137/BBI.S38191

138. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res. Oxford University Press; 2003;31: 3784–8. doi: org/10.1093/nar/gkg563

139. Xu D, Xu Y, Uberbacher E. Computational Tools For Protein Modeling. Curr Protein Pept Sci. 2000;1: 1–21. doi:10.2174/1389203003381469

140. Bordoli L, Kiefer F, Arnold K, Benkert P, Battey J, Schwede T. Protein structure homology modeling using SWISS-MODEL workspace. Nat Protoc. Nature Publishing Group; 2009;4: 1–13. doi:10.1038/nprot.2008.197

141. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J.1986; 5(4):823–6. PMCID:PMC1166865

142. Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, Linial M, *et al*. Protein function annotation by homology-based inference. Genome Biol. 2009;10: 207. doi:10.1186/gb-2009-10-2-207

143. Wu S, Zhang Y. LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Res. Oxford University Press; 2007;35: 3375–82.doi:10.1093/nar/gkm251

144. Gupta CL, Akhtar S, Bajpai P. *In silico* protein modeling: possibilities and limitations. EXCLI J. Leibniz Research Centre for Working Environment and Human Factors; 2014;13: 513–5. ISSN 1611-2156

145. K Mishra P, Sonkar SC, Raj SR, Saluja UC and D. Functional Analysis of Hypothetical Proteins of *Chlamydia Trachomatis*: A Bioinformatics Based Approach for Prioritizing the Targets. J Comput Sci Syst Biol. OMICS International; 2014;7. doi:10.4172/jcsb.1000132

146. Wu S, Skolnick J, Zhang Y. *Ab initio* modeling of small proteins by iterative TASSER simulations. BMC Biol. BioMed Central; 2007;5: 17.doi:10.1186/1741-7007-5-17

147. Krieger E, Nabuurs SB, Vriend G. Homology Modeling. John Wiley & Sons, Inc.; 2005; 509–523. doi:10.1002/0471721204.ch25

148. Sánchez R, Sali A. Evaluation of comparative protein structure modeling by MODELLER-3. Proteins. 1997; l 1: 50–8.

149. Zhou Y, Johnson ME. Comparative molecular modeling analysis of-5-amidinoindole and benzamidine binding to thrombin and trypsin: specific H-bond formation contributes to high 5-amidinoindole potency and selectivity for thrombin and factor Xa. J Mol Recognit. 12: 235–41. doi:10.1002/(SICI)1099-1352(199907/08)12:4<235::AID-JMR460>3.0.CO;2-X

150. Ceulemans H1 RR. Fast fitting of atomic structures to low-resolution electron density maps by surface overlap maximization. Journal of Molecular Biology; 2004; 338 (4): 783-793. doi: org/10.1016/j.jmb.2004.02.066

151. Teichmann SA1, Chothia C GM. Advances in structural genomics. Curr Opin Struct Biol . 1999; 9(3): 390-399. doi: org/10.1016/S0959-440X(99)80053-0

152. Capener CE, Shrivastava IH, Ranatunga KM, Forrest LR, Smith GR, Sansom MS. Homology modeling and molecular dynamics simulation studies of an inward rectifier potassium channel. Biophys J. The Biophysical Society; 2000;78: 2929–42. doi:10.1016/S0006-3495(00)76833-0

153. Vyas VK, Ukawala RD, Ghate M, Chintha C. Homology modeling a fast tool for drug discovery: current perspectives. Indian J Pharm Sci. Medknow Publications; 2012;74: 1–17. doi:10.4103/0250-474X.102537

154. Söding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. Oxford University Press; 2005;33: W244-8. doi:10.1093/nar/gki408

155. Golubchik T, Wise MJ, Easteal S, Jermiin LS. Mind the gaps: evidence of bias in estimates of multiple sequence alignments. Mol Biol Evol. Oxford University Press; 2007;24: 2433–42. doi:10.1093/molbev/msm176

156. Sippl MJ. Recognition of errors in three-dimensional structures of proteins. Proteins Struct Funct Genet. 1993;17: 355–362.doi:10.1002/prot.340170404

157. Wiederstein M, Sippl MJ. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res. 2007;35: W407–W410. doi:10.1093/nar/gkm290

158. Laskowski RA, MacArthur MW, Moss DS, Thornton JM, IUCr. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr. International Union of Crystallography; 1993;26: 283–291.doi:10.1107/S0021889892009944

159. Pawlowski M, Gajda MJ, Matlak R, Bujnicki JM, Lazaridis T, Karplus M, *et al*. MetaMQAP: A meta-server for the quality assessment of protein models. BMC Bioinformatics. BioMed Central; 2008;9: 403. doi:10.1186/1471-2105-9-403

160. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, *et al*. SWISS-MODEL: modeling

protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res. 2014;42: W252-8. doi:10.1093/nar/gku340

161. Benkert P, Künzli M, Schwede T. QMEAN server for protein model quality estimation. Nucleic Acids Res. 2009;37: W510-4. doi:10.1093/nar/gkp322

162. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28: 27–30. doi: org/10.1093/nar/28.1.27

163. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, *et al*. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. 2004;32: D115-9. doi:10.1093/nar/gkh131

164. Su E, Chiu H-S, Lo A, Hwang J-K, Sung T-Y, Hsu W-L, *et al*. Protein subcellular localization prediction based on compartment-specific features and structure conservation. BMC Bioinformatics. BioMed Central; 2007;8: 330. doi:10.1186/1471-2105-8-330

165. Savojardo C, Martelli PL, Fariselli P, Casadio R. TPpred2: improving the prediction of mitochondrial targeting peptide cleavage sites by exploiting sequence motifs. Bioinformatics. 2014;30: 2973–4. doi:10.1093/bioinformatics/btu411

166. Claros. MitoProt, a Macintosh application for studying mitochondrial proteins. Comput Appl Biosci; 1995. 6(4): 122-1129. PMID: 8521054

167. Kosugi S, Hasebe M, Tomita M, Yanagawa H. Systematic identification of cell cycle-dependent yeast nucleocytoplasmic shuttling proteins by prediction of composite motifs. Proc Natl Acad Sci USA. 2009;106: 10171–6. doi:10.1073/pnas.0900604106

168. Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J. Membrane Proteins. 2000; 25: 397-399. doi: 10.1001/1552-36361

169. Chatterjee S, Mayor S. The GPI-anchor and protein sorting. Cell Mol Life Sci. 2001;58: 1969–1987. doi:10.1007/PL00000831

170. Luo J, Yu L, Guo Y, Li M. Functional classification of secreted proteins by position specific scoring matrix and auto covariance. Chemom Intell Lab Syst. 2012;110: 163–167. doi:10.1016/j.chemolab.2011.11.008

171. Watanabe Costa R, da Silveira JF, Bahia D. Interactions between *Trypanosoma cruzi* Secreted Proteins and Host Cell Signaling Pathways. Front Microbiol. Frontiers Media SA; 2016;7: 388. doi:10.3389/fmicb.2016.00388

172. Tjalsma H, Bolhuis A, Jongbloed JD, Bron S, van Dijl JM. Signal peptide-dependent protein transport in *Bacillus subtilis*: a genome-based survey of the secretome. Microbiol Mol Biol Rev. 2000;64: 515–47. doi: 10.1128/9781555817992

173. Bendtsen JD, Kiemer L, Fausbøll A, Brunak S. Non-classical protein secretion in bacteria. BMC Microbiol. 2005;5: 58. doi:10.1186/1471-2180-5-58

174. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8: 785–6. doi:10.1038/nmeth.1701

175. J. Simpson R, Mathivanan S. Extracellular Microvesicles: The Need for Internationally Recognised Nomenclature and Stringent Purification Criteria. J Proteomics Bioinform. OMICS International; 2012;5. doi:10.4172/jpb.10000e10

176. Taylor WR, Orengo CA. Protein structure alignment. J Mol Biol. 1989; 208: 1–22. doi:org/10.1016/0022-2836(89)90084-3

177. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol. 2000;302: 205–17. doi:10.1006/jmbi.2000.4042

178. Garg A, Gupta D. VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. BMC Bioinformatics. BioMed Central; 2008; 9: 62. doi:10.1186/1471-2105-9-62

179. Saha S, Raghava GPS. VICMpred: An SVM-based Method for the Prediction of Functional Proteins of Gram-negative Bacteria Using Amino Acid Patterns and Composition. Genomics Proteomics Bioinformatics. 2006;4: 42–47. doi:10.1016/S1672-0229(06)60015-6

180. Gupta A, Kapil R, Dhakan DB, Sharma VK. MP3: a software tool for the prediction of pathogenic proteins in genomic and metagenomic data. PLoS One. Public Library of Science; 2014;9: e93907. doi:10.1371/journal.pone.0093907

181. Ikai A. Thermostability and aliphatic index of globular proteins. J Biochem. 1980;88: 1895–8. doi:org/10.1093/oxfordjournals.jbchem.a133168

182. Murphy RF. Communicating subcellular distributions. Cytometry A. NIH Public Access; 2010;77: 686–92. doi:10.1002/cyto.a.20933

183. Burleigh BA, Woolsey AM. Cell signalling and *Trypanosoma cruzi* invasion. Cell Microbiol. 2002;4: 701–11. doi: 10.1046/j.1462-5822.2002.00226.x

184. Prilusky J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O, Beckmann JS, *et al*. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics. 2005;21: 3435–8.doi:10.1093/bioinformatics/bti537

185. Ariyachet C, Solis N V, Liu Y, Prasadarao N V, Filler SG, McBride AE. SR-like RNA-binding protein Slr1 affects *Candida albicans* filamentation and virulence. Infect Immun. American Society for Microbiology (ASM); 2013;81: 1267–76. doi:10.1128/IAI.00864-12

186. Naqvi AAT, Shahbaaz M, Ahmad F, Hassan MI. Identification of functional candidates amongst hypothetical proteins of *Treponema pallidum* ssp. pallidum. PLoS One. Public Library of Science; 2015;10: e0124177.doi:10.1371/journal.pone.0124177

187. Shiels B, Langsley G, Weir W, Pain A, McKellar S, Dobbelaere D. Alteration of host cell phenotype by *Theileria annulata* and *Theileria parva*: Mining for manipulators in the parasite genomes [Internet]. International Journal for Parasitology. 2006. pp. 9–21. doi:10.1016/j.ijpara.2005.09.002

188. Blanco MT, Sacristán B, Lucio L, Blanco J, Pérez-Giraldo C, Cándido Gómez-García A. La hidrofobicidad de la superficie celular como indicador de otros factores de virulencia en *Candidaalbicans*. Rev Iberoam Micol. 2010;27: 195–199. doi:10.1016/j.riam.2010.09.001

189. Balaji S, Babu MM, Iyer LM, Aravind L. Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. Nucleic Acids Res. Oxford University Press; 2005;33: 3994–4006. doi:10.1093/nar/gki709

190. Painter HJ, Campbell TL, Llinás M. The Apicomplexan AP2 family: integral factors regulating *Plasmodium* development. Mol Biochem Parasitol. NIH Public Access; 2011;176: 1–7. doi:10.1016/j.molbiopara.2010.11.014

191. De Silva EK, Gehrke AR, Olszewski K, Leon I, Chahal JS, Bulyk ML, *et al.* Specific DNA-binding by Apicomplexan AP2 transcription factors. Proc Natl Acad Sci. 2008;105: 8393–8398. doi:10.1073/pnas.0801993105

192. Pieszko M, Weir W, Goodhead I, Kinnaird J, Shiels B, Dyer M, *et al.* ApiAP2 Factors as Candidate Regulators of Stochastic Commitment to Merozoite Production in *Theileria annulata*. Dinglasan RR, editor. PLoS Negl Trop Dis. Public Library of Science; 2015;9: e0003933. doi:10.1371/journal.pntd.0003933

193. Latchman DS. Transcription factors: An overview. Int J Biochem Cell Biol. Pergamon; 1997;29: 1305–1312. doi:10.1016/S1357-2725(97)00085-X

194. Anantharaman V, Iyer LM, Aravind L. Comparative genomics of protists: new insights into the evolution of eukaryotic signal transduction and gene regulation. Annu Rev Microbiol. 2007;61: 453–475. doi:10.1146/annurev.micro.61.080706.093309

195. Peeling RW, Hook EW. The pathogenesis of syphilis: the Great Mimicker, revisited. J Pathol. 2006; 208: 224–32. doi:10.1002/path.1903

196. Hutchings MI, Palmer T, Harrington DJ, Sutcliffe IC. Lipoprotein biogenesis in Gram-positive bacteria: knowing when to hold 'em, knowing when to fold 'em. Trends Microbiol. 2009;17: 13–21. doi:10.1016/j.tim.2008.10.001

197. Khandavilli S, Homer KA, Yuste J, Basavanna S, Mitchell T, Brown JS. Maturation of *Streptococcus pneumoniae* lipoproteins by a type II signal peptidase is required for ABC transporter function and full virulence. Mol Microbiol. 2008;67: 541–57. doi:10.1111/j.1365-2958.2007.06065

198. Lin WJ, Gary JD, Yang MC, Clarke S, Herschman HR. The mammalian immediate-early TIS21 protein and the leukemia-associated BTG1 protein interact with a protein-arginine N-methyltransferase. J Biol Chem. 1996; 271: 15034–44. doi: 10.1074/jbc.271.25.15034

199. Jortzik E, Kehr S, Becker K. Post-translational Modifications in Apicomplexan Parasites. Progress in Parasitology. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. pp. 93–120. doi:10.1007/978-3-642-21396-0_6

200. Scheffzek K, Ahmadian MR. GTPase activating proteins: structural and functional insights 18 years after discovery. Cell Mol Life Sci. 2005;62: 3014–38. doi:10.1007/s00018-005-5136

201. Leipe DD, Wolf YI, Koonin E V, Aravind L. Classification and evolution of P-loop GTPases and related ATPases. J Mol Biol. 2002;317: 41–72. doi:10.1006/jmbi.2001.5378

202. Schmid-Hempel P. Immune defence, parasite evasion strategies and their relevance for "macroscopic phenomena" such as virulence. doi:10.1098/rstb.2008.0157

203. Freeman ZN, Dorus S, Waterfield NR, Groisman E, Mouslim C, Groisman E, *et al*. The KdpD/KdpE Two-Component System: Integrating K+ Homeostasis and Virulence. Chitnis CE, editor. PLoS Pathog. Public Library of Science; 2013;9: e1003201. doi:10.1371/journal.ppat.1003201

204. Bootman MD, Lipp P. Calcium Signalling and Regulation of Cell Function. 2012; 5(3): 221-225. doi: 10.1002/9780470015902.a0001265.pub3

205. Dobbelaere DA, Fernandez PC, Heussler VT. *Theileria parva*: taking control of host cell proliferation and survival mechanisms. Cell Microbiol. 2000;2: 91–9. doi: 10.1046/j.1462-5822.2000.00045.x

206. Heussler VT, Machado J, Fernandez PC, Botteron C, Chen CG, Pearse MJ, *et al*. The intracellular parasite *Theileria parva* protects infected T cells from apoptosis. Proc Natl Acad Sci U S A. National Academy of Sciences; 1999; 96: 7312–7. doi: 10.1073/pnas.96.13.7312

207. Cohen GB, Ren R, Baltimore D. Modular binding domains in signal transduction proteins. Cell. Cell Press; 1995;80: 237–248. doi:10.1016/0092-8674(95)90406-9

208. Bork P, Schultz J, Ponting CP. Cytoplasmic signalling domains: the next generation. Trends Biochem Sci. 1997;22: 296–8. doi: org/10.1016/S0968-0004(97)01084-0

209. Jaiswal DK, Ray D, Subba P, Mishra P, Gayali S, Datta A, *et al*. Proteomic analysis reveals the diversity and complexity of membrane proteins in chickpea (Cicer arietinum L.). Proteome Sci. 2012;10: 59. doi:10.1186/1477-5956-10-59

210. Nouwens AS, Cordwell SJ, Larsen MR, Molloy MP, Gillings M, Willcox MDP, *et al*. Complementing genomics with proteomics: The membrane subproteome of*Pseudomonas aeruginosa* PAO1. Electrophoresis. Wiley Subscription Services, Inc., A Wiley Company; 2000; 21: 3797–3809. doi:10.1002/1522-2683

211. Santoni V, Molloy M, Rabilloud T. Membrane proteins and proteomics: Un amour impossible? Electrophoresis. Wiley Subscription Services, Inc., A Wiley Company; 2000;21: 1054–1070. doi:10.1002/1522-2683

212. Arinaminpathy Y, Khurana E, Engelman DM, Gerstein MB. Computational analysis of membrane proteins: the largest class of drug targets. Drug Discov Today. 2009; 14: 1130–5. doi:10.1016/j.drudis.2009.08.006

213. Tretina K, Gotia HT, Mann DJ, Silva JC. *Theileria*-transformed bovine leukocytes have cancer hallmarks. Trends Parasitol. 2015; 31: 306–14. doi:10.1016/j.pt.2015.04.001

214. Chakraborty S, Monfett M, Maier TM, Benach JL, Frank DW, Thanassi DG. Type IV pili in *Francisella tularensis*: Roles of pilF and pilT in fiber assembly, host cell adherence, and virulence. Infect Immun. 2008;76: 2852–2861.doi:10.1128/IAI.01726-07

215. Bröms JE, Edqvist PJ, Forsberg A, Francis MS. Tetratricopeptide repeats are essential for PcrH chaperone function in Pseudomonas aeruginosa type III secretion. FEMS Microbiol Lett. 2006;256: 57–66.doi:10.1111/j.1574-6968.2005.00099

216. Vetrivel U, Subramanian G, Dorairaj S. A novel *in silico* approach to identify potential therapeutic targets in human bacterial pathogens. Hugo J. 2011;5: 25–34. doi:10.1007/s11568-011-9152-7

217. Rajendran L, Knölker H-J, Simons K. Subcellular targeting strategies for drug design and delivery. Nat Rev Drug Discov. Nature Publishing Group; 2010;9: 29–42. doi:10.1038/nrd2897

218. Leslie DM, Zhang W, Timney BL, Chait BT, Rout MP, Wozniak RW, *et al*. Characterization of karyopherin cargoes reveals unique mechanisms of Kap121p-mediated nuclear import. Mol Cell Biol. American Society for Microbiology (ASM); 2004;24: 8487–503. doi:10.1128/MCB.24.19.8487-8503.2004

219. Jans DA, Briggs LJ, Gustin SE, Jans P, Ford S, Young IG. The cytokine interleukin-5 (IL-5) effects cotransport of its receptor subunits to the nucleus *in vitro*. FEBS Lett. 1997; 410: 368–72. doi: 10.1016/S0014-5793(97)00622-4

220. Shiota C, Coffey J, Grimsby J, Grippo JF, Magnuson MA. Nuclear import of hepatic glucokinase depends upon glucokinase regulatory protein, whereas export is due to a nuclear export signal sequence in glucokinase. J Biol Chem. 1999; 274: 37125–30. doi: 10.1074/jbc.274.52.37125

221. Ravindran S, Boothroyd JC. Secretion of Proteins into Host Cells by Apicomplexan Parasites. Traffic. Blackwell Publishing Ltd; 2008;9: 647–656. doi:10.1111/j.1600-0854.2008.00723

222. Yao Y-L, Yang W-M. Nuclear proteins: promising targets for cancer drugs. Curr Cancer Drug Targets. 2005;5: 595–610. doi: https://doi.org/10.2174/156800905774932815

223. Schmuckli-Maurer J, Casanova C, Schmied S, Affentranger S, Parvanova I, Kang'a S, *et al*. Expression analysis of the *Theileria parva* subtelomere-encoded variable secreted protein gene family. PLoS One. Public Library of Science; 2009;4: e4839. doi:10.1371/journal.pone.0004839

224. Sivakumar T, Hayashida K, Sugimoto C, Yokoyama N. Evolution and genetic diversity of *Theileria*. Infect Genet Evol. 2014; 27: 250–263.doi:10.1016/j.meegid.2014.07.013

225. Tretina K, Gotia HT, Mann DJ, Silva JC, Adl SM, *et al*. *Theileria*-transformed bovine leukocytes have cancer hallmarks. Trends Parasitol. Elsevier; 2015; 31: 306–314. doi:10.1016/j.pt.2015.04.001

226. Guda C, Guda P, Fahy E, Subramaniam S. MITOPRED: a web server for the prediction of mitochondrial proteins. Nucleic Acids Res. Oxford University Press; 2004; 32: W372-4. doi:10.1093/nar/gkh374

227. Wen S, Zhu D, Huang P. Targeting cancer cell mitochondria as a therapeutic approach. Future Med Chem. NIH Public Access; 2013; 5: 53–67.doi:10.4155/fmc.12.190

228. Barbosa IA, Machado NG, Skildum AJ, Scott PM, Oliveira PJ. Mitochondrial remodeling in cancer metabolism and survival: potential for new therapies. Biochim Biophys Acta. 2012; 1826: 238–54. doi:10.1016/j.bbcan.2012.04.005

229. Cheng Z, Ristow M. Mitochondria and metabolic homeostasis. Antioxid Redox Signal. 2013; 19: 240–2. doi:10.1089/ars.2013.5255

230. Suen D-F, Norris KL, Youle RJ. Mitochondrial dynamics and apoptosis. Genes Dev. 2008; 22: 1577–90. doi:10.1101/gad.1658508

231. Tait SWG, Green DR. Mitochondria and cell signalling. J Cell Sci. 2012; 125: 807–15. doi:10.1242/jcs.099234

232. Van Houten B, Woshner V, Santos JH. Role of mitochondrial DNA in toxic responses to oxidative stress. DNA Repair (Amst). 2006; 5: 145–52. doi:10.1016/j.dnarep.2005.03.002

233. Fukasawa Y, Tsuji J, Fu S-C, Tomii K, Horton P, Imai K. MitoFates: improved prediction of mitochondrial targeting sequences and their cleavage sites. Mol Cell Proteomics. American Society for Biochemistry and Molecular Biology; 2015;14: 1113–26. doi:10.1074/mcp.M114.043083

234. Hsu C-C, Tseng L-M, Lee H-C. Role of mitochondrial dysfunction in cancer progression. Exp Biol Med. 2016;241: 1281–1295.doi:10.1177/1535370216641787

235. Boland ML, Chourasia AH, Macleod KF. Mitochondrial Dysfunction in Cancer. Front Oncol. 2013; 3. doi:10.3389/fonc.2013.00292

236. Pierleoni A, Martelli P, Casadio R. PredGPI: a GPI-anchor predictor. BMC Bioinformatics. 2008; 9: 392. doi:10.1186/1471-2105-9-392

237. Kupzig S, Korolchuk V, Rollason R, Sugden A, Wilde A, Banting G. Bst-2/HM1.24 Is a Raft-Associated Apical Membrane Protein with an Unusual Topology. Traffic. Blackwell Publishing Ltd; 2003; 4: 694–709.doi:10.1034/j.1600-0854.2003.00129

238. Xue G, von Schubert C, Hermann P, Peyer M, Maushagen R, Schmuckli-Maurer J, *et al*. Characterisation of gp34, a GPI-anchored protein expressed by schizonts of *Theileria parva* and *T. annulata*. Mol Biochem Parasitol. 2010; 172: 113–120. doi:10.1016/j.molbiopara.2010.03.018

239. Terstappen GC, Reggiani A. *In silico* research in drug discovery. Trends Pharmacol Sci. 2001; 22: 23–6. doi: org/10.1016/S0165-6147(00)01584-4

240. Davey J. G-protein-coupled receptors: new approaches to maximise the impact of GPCRS in drug discovery. Expert Opin Ther Targets. 2004; 8: 165–70. doi:10.1517/14728222.8.2.165

241. Rinaudo CD, Telford JL, Rappuoli R, Seib KL. Vaccinology in the genome era. J Clin Invest. American Society for Clinical Investigation; 2009; 119: 2515–25. doi:10.1172/JCI38330

242. Chaudhuri R, Kulshreshtha D, Raghunandanan MV, Ramachandran S. Integrative immunoinformatics for Mycobacterial diseases in R platform. Syst Synth Biol. 2014; 8: 27–39. doi:10.1007/s11693-014-9135-9

243. Pain A, Renauld H, Berriman M, Murphy L, Yeats CA, Weir W, *et al*. Genome of the host-cell transforming parasite *Theileria annulata* compared with *T. parva*. Science. American Association for the Advancement of Science; 2005; 309: 131–3.doi:10.1126/science.1110418

244. Schmuckli-Maurer J, Casanova C, Schmied S, Affentranger S, Parvanova I, Kang'a S, *et al*. Expression Analysis of the *Theileria parva* Subtelomere-Encoded Variable Secreted Protein Gene Family. Rodrigues MM, editor. PLoS One. 2009; 4: e4839. doi:10.1371/journal.pone.0004839

245. Reid AJ. Large, rapidly evolving gene families are at the forefront of host-parasite interactions in Apicomplexa. Parasitology. Cambridge University Press; 2015; S57-70. doi:10.1017/S0031182014001528

246. Laity JH, Lee BM, Wright PE. Zinc finger proteins: new insights into structural and functional diversity. Curr Opin Struct Biol. 2001; 11: 39–46. https://doi.org/10.1016/S0959-440X(00)00167-6

247. Osborne BA. Apoptosis and the maintenance of homoeostasis in the immune system. Curr Opin Immunol . 1996;8: 245–54. doi.org/10.1016/S0952-7915(96)80063-X

248. Evan G. Cancer--a matter of life and cell death. Int J cancer. 1997; 71: 709–11. doi: 10.1002/(SICI)1097-0215(19970529)71:5<709::AID-IJC2>3.0.CO;2-V

249. Riedel H, Dull TJ, Honegger AM, Schlessinger J, Ullrich A. Cytoplasmic domains determine signal specificity, cellular routing characteristics and influence ligand binding of epidermal growth factor and insulin receptors. EMBO J. 1989; 8: 2943–54. PMCID:PMC 401363

250. Chang DD, Wong C, Smith H, Liu J. ICAP-1, a novel beta1 integrin cytoplasmic domain-associated protein, binds to a conserved and functionally important NPXY sequence motif of beta1 integrin. J Cell

Biol. The Rockefeller University Press; 1997; 138: doi:1149–57. 10.1083/jcb.138.5.1149

251. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, *et al*. The Pfam protein families database. Nucleic Acids Res. Oxford University Press; 2004; 32: D138-41. doi:10.1093/nar/gkh121

252. Goodacre NF, Gerloff DL, Uetz P. Protein domains of unknown function are essential in bacteria. MBio. American Society for Microbiology; 2014;5: e00744-13. doi:10.1128/mBio.00744-13

253. Snider J, Kittanakom S, Damjanovic D, Curak J, Wong V, Stagljar I. Detecting interactions with membrane proteins using a membrane two-hybrid assay in yeast. Nat Protoc. Nature Publishing Group; 2010;5: 1281–1293. doi:10.1038/nprot.2010.83

254. Adebayo S. Annotation of virulence factors in schistosomes for the development of a SchistoVir database. J Comput Biol Bioinforma Res. 2013; 5: 6–14. doi:10.5897/JCBBR12.013

255. Sousa CP de. Pathogenicity mechanisms of prokaryotic cells: an evolutionary view. Brazilian J Infect Dis. The Brazilian Journal of Infectious Diseases and Contexto Publishing; 2003; 7: 23–31. doi:10.1590/S1413-86702003000100004

256. Cui W, Chen L, Huang T, Gao Q, Jiang M, Zhang N, *et al*. Computationally identifying virulence factors based on KEGG pathways. Mol Biosyst. The Royal Society of Chemistry; 2013; 9: 1447. doi:10.1039/c3mb70024k

257. Katsir L, Schilmiller AL, Staswick PE, He SY, Howe GA. COI1 is a critical component of a receptor for jasmonate and the bacterial virulence factor coronatine. Proc Natl Acad Sci. 2008; 105: 7100–7105. doi:10.1073/pnas.0802332105

258. Wu CH, Huang H, Yeh LSL, Barker WC. Protein family classification and functional annotation. Comput Biol Chem. 2003; 27: 37–47. doi:10.1016/S1476-9271(02)00098-1

259. Saeidnia S, Manayi A, Abdollahi M. The Pros and Cons of the *In-silico* Pharmaco-toxicology in Drug Discovery and Development. Int J Pharmacol. 2013; 9: 176–181. doi:10.3923/ijp.2013.176.181

260. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins. 1991; 9: 56–68. doi:10.1002/prot.340090107

261. Shahbaaz M, Hassan MI, Ahmad F. Functional annotation of conserved hypothetical proteins from *Haemophilus influenzae* Rd KW20. PLoS One. 2013; 8. doi:10.1371/journal.pone.0084263

262. McDermott J, Bumgarner R, Samudrala R. Functional annotation from predicted protein interaction networks. Bioinformatics. 2005; doi:10.1093/bioinformatics/bti514

263. Sonego P. ROC analysis : applications to the classification of biological sequences and 3D structures. 2008;9: 198–209. doi:10.1093/bib/bbm064

264. Zweig H. Receiver-Operating Clinical Medicine (ROC) Plots : A Fundamental Evaluation Tool in. 1993; 39: 561–577. doi:10.1002/p.312290112

265. Wootton JC. Evaluating the effectiveness of sequence analysis algorithms using measures of relevant information. 1997; 21: 191–202. doi:org/10.1016/S0097-8485(97)00001-6

266. Duda RO, Hart PE, Stork DG. Pattern Classification, 2nd edn. John Wiley and Sons, New York (2001).

APPENDIX

Appendix A: Supplementary data for Chapter 3 (Results)

Appendix B: Ethical Clearance Certificate

Appendix A. 1. The details of the analysis of physicochemical properties of 309 *T. parva* hypothetical proteins determined using Expasy's ProtParam tool

Sequence. No	HP gene name(protein length)	Molecular weight (Da)	Isoelectric point (PI)	Extinction coefficient (EC;M-1 cm-1)	Aliphatic Index (AI)	Instability Index(II)		GRAVY
						Computed	Classification	
1.	TP05_0039 (126)	15901.0	9.70	51005	97.54	33.53	Stable	0.170
2.	TP05_0020 (133)	15585.6	9.00	29465	143.53	17.48	Stable	0.831
3.	TP04_0910 (276)	31960.9	7.69	13410	68.55	37.23	Stable	-1.034
4.	TP04_0905 (676)	80376.8	5.18	96610	80.81	38.64	Stable	-0.583
5.	TP04_0903 (170)	19559.7	4.58	13075	69.35	44.43	Unstable	-0.744
6.	TP04_0896 (489)	56392.6	5.08	23380	97.08	29.98	Stable	-0.647
7.	TP04_0869 (283)	32295.4	5.02	24910	79.86	48.74	Unstable	-0.278
8.	TP04_0834 (398)	46263.7	8.90	49850	108.22	32.78	Stable	-0.206
9.	TP04_0833 (147)	17356.1	9.88	8940	88.84	25.04	Stable	-0.327
10.	TP04_0786 (381)	43218.1	4.71	40465	60.66	46.00	Unstable	-1.025
11.	TP04_0729 (218)	25531.3	5.35	15930	99.77	69.26	Unstable	-0.417
12.	TP04_0715 (212)	24566.9	4.56	18910	58.87	51.07	Unstable	-0.606
13.	TP04_0708 (349)	42120.0	6.51	49405	103.78	44.63	Unstable	-0.066
14.	TP04_0693 (221)	25434.7	8.47	21430	71.40	62.62	Unstable	-0.677
15.	TP04_0654 (591)	68837.8	6.41	72240	105.52	32.59	Stable	-0.059
16.	TP04_0638 (64)	7466.2	6.26	8480	53.44	66.33	Unstable	-0.988
17.	TP04_0633 (193)	21844.0	9.76	17420	88.39	56.85	Unstable	-0.101
18.	TP04_0579 (106)	12083.4	9.75	7115	99.25	26.90	Stable	-0.013
19.	TP04_0576 (447)	51523.1	8.82	32780	106.11	38.04	Stable	-0.160
20.	TP04_0532 (206)	23286.3	5.80	13980	86.99	36.07	Stable	-0.307
21.	TP04_0505 (74)	8866.3	8.66	8940	107.97	33.57	Stable	-0.366
22.	TP04_0503 (245)	28705.2	10.32	9970	64.41	46.48	Stable	-1.263
23.	TP04_0455 (556)	63577.9	5.85	77810	114.96	48.56	Unstable	0.483
24.	TP04_0422 (129)	14666.0	6.25	7700	120.85	57.17	Unstable	0.154
25.	TP04_0414 (768)	86777.8	5.45	51230	108.68	31.92	Stable	-0.182
26.	TP04_0405 (718)	80949.8	5.52	42750	77.84	30.83	Stable	-0.523
27.	TP04_0399 (357)	39665.9	6.07	12045	82.44	26.66	Stable	-0.496
28.	TP04_0353 (477)	54620.5	5.70	39810	80.19	33.48	Stable	-0.423
29.	TP04_0327 (507)	58310.1	8.83	34645	79.19	33.69	Stable	-0.485
30.	TP04_0283 (206)	23179.8	9.75	47440	80.97	28.95	Stable	-0.100
31.	TP04_0275 (2405)	272538.0	6.10	164785	78.20	35.63	Stable	-0.598
32.	TP04_0259 (137)	15599.7	4.50	10430	115.26	40.82	Unstable	0.137

33.	TP04_0254 (220)	25859.0	8.61	49850	105.41	54.41	Unstable	-0.065
34.	TP04_0252 (649)	75705.1	5.20	45730	93.85	42.72	Unstable	-0.419
35.	TP04_0245 (816)	90854.8	5.64	67270	84.00	36.68	Stable	-0.479
36.	TP04_0240 (327)	37006.4	10.17	4970	81.07	33.91	Stable	-0.619
37.	TP04_0237 (164)	18543.6	10.23	8940	70.73	60.09	Unstable	-0.868
38.	TP04_0232 (884)	102790.4	7.87	102680	96.12	35.99	Stable	-0.219
39.	TP04_0223 (48)	5305.1	6.04	6990	99.58	30.37	Stable	-0.090
40.	TP04_0210 (384)	44508.1	6.25	50100	75.86	48.87	Unstable	-0.832
41.	TP04_0200 (197)	22728.8	9.22	30830	113.25	35.62	Stable	0.550
42.	TP04_0192 (110)	13066.7	4.90	15930	77.91	53.88	Unstable	-0.465
43.	TP04_0190 (74)	8937.4	6.89	19940	102.43	34.75	Stable	0.246
44.	TP04_0188 (199)	23448.6	8.75	26360	85.23	51.50	Unstable	-0.440
45.	TP04_0181 (67)	7748.2	5.43	1490	30.45	72.83	Unstable	-1.815
46.	TP04_0172 (912)	105272.6	5.76	128885	78.93	37.22	Stable	-0.438
47.	TP04_0171 (306)	35183.5	9.20	18005	-0.123	42.20	Unstable	-0.123
48.	TP04_0144 (370)	42485.1	6.72	26820	67.16	43.68	Unstable	-0.901
49.	TP04_0128 (587)	67327.0	5.14	63260	89.23	40.05	Unstable	-0.464
50.	TP04_0127 (540)	62261.1	8.07	48250	90.07	37.15	Stable	-0.448
51.	TP04_0121 (114)	13316.4	11.51	24980	80.35	36.82	Stable	-0.594
52.	TP04_0114 (198)	22844.5	5.58	24075	76.31	30.14	Stable	-0.752
53.	TP04_0087 (356)	40835.6	8.62	20985	96.57	26.85	Stable	-0.352
54.	TP04_0082 (207)	23711.6	4.44	14565	99.28	18.15	Stable	-0.215
55.	TP04_0081 (935)	108092.1	6.39	83660	106.46	28.82	Stable	-0.116
56.	TP04_0077 (55)	6400.4	5.64	5960	109.82	41.34	Unstable	-0.322
57.	TP04_0073 (379)	42811.1	9.27	19410	76.97	35.59	Stable	-0.806
58.	TP04_0069 (228)	26966.8	5.75	31650	90.22	29.00	Stable	-0.534
59.	TP04_0068 (233)	26731.0	5.66	9970	106.61	37.79	Stable	-0.100
60.	TP04_0006 (522)	61656.6	5.62	81530	56.34	53.61	Unstable	-1.119
61.	TP04_0003 (561)	66150.3	4.88	82895	59.71	61.29	Unstable	-1.063
62.	TP03_0903 (310)	35706.6	6.09	44350	92.39	33.38	Stable	-0.241
63.	TP03_0901 (522)	62387.3	9.15	75070	92.85	35.04	Stable	-0.311
64.	TP03_0900 (521)	61770.1	8.91	86445	85.37	33.22	Stable	-0.400
65.	TP03_0899 (524)	62330.9	9.17	67620	85.08	31.90	Stable	-0.431
66.	TP03_0898 (223)	25313.3	7.71	15150	99.24	36.44	Stable	0.260
67.	TP03_0896 (524)	62639.5	9.27	77590	91.22	35.02	Stable	-0.377
68.	TP03_0893 (568)	67075.5	4.86	85500	61.73	60.39	Unstable	-1.044
69.	TP03_0885 (481)	57573.0	4.91	72465	60.52	54.11	Unstable	-1.099
70.	TP03_0883 (537)	63369.1	4.84	79915	60.54	52.96	Unstable	-0.972
71.	TP03_0882 (607)	72202.4	4.98	116790	53.57	69.21	Unstable	-1.203
72.	TP03_0881 (568)	67922.8	6.53	88605	57.96	62.36	Unstable	-1.233

73.	TP03_0880 (609)	72231.6	4.96	88480	49.89	69.78	Unstable	-1.218
74.	TP03_0877 (341)	39868.3	9.48	40590	113.43	25.83	Stable	-0.221
75.	TP03_0875 (509)	60703.4	4.85	74165	46.31	80.24	Unstable	-1.366
76.	TP03_0873 (510)	60874.3	4.94	82520	50.94	64.39	Unstable	-1.222
77.	TP03_0870 (563)	66072.6	4.81	92950	50.66	63.19	Unstable	-1.171
78.	TP03_0856 (527)	61582.0	8.92	66740	94.35	30.06	Stable	-0.365
79.	TP03_0850 (397)	45162.3	4.92	46300	69.19	36.17	Stable	-0.540
80.	TP03_0839 (260)	29658.3	9.12	32890	66.58	35.88	Stable	-0.714
81.	TP03_0832 (155)	18175.9	9.07	39420	84.19	47.51	Unstable	-0.053
82.	TP03_0827 (885)	99704.4	9.45	113275	61.48	32.16	Stable	-0.863
83.	TP03_0825 (446)	50893.0	5.14	35340	83.16	36.89	Stable	-0.528
84.	TP03_0820 (255)	30215.1	6.27	27850	79.80	30.24	Stable	-0.849
85.	TP03_0819 (112)	12948.3	5.86	12950	64.38	27.87	Stable	-1.066
86.	TP03_0812 (64)	7389.5	9.99	5960	71.56	36.21	Stable	-0.166
87.	TP03_0779 (300)	34822.1	4.71	26485	66.20	43.88	Unstable	-0.744
88.	TP03_0768 (346)	38783.7	5.73	19620	98.61	34.59	Stable	-0.167
89.	TP03_0761 (185)	21463.0	4.99	33920	70.05	30.24	Stable	-0.601
90.	TP03_0759 (195)	22124.6	5.19	10345	101.38	31.64	Stable	0.161
91.	TP03_0754 (819)	94473.2	4.95	77060	108.24	29.36	Stable	0.065
92.	TP03_0742 (273)	31253.5	8.82	14440	80.00	30.06	Stable	-0.538
93.	TP03_0738 (666)	77637.8	8.49	105505	80.98	40.27	Unstable	-0.317
94.	TP03_0729 (379)	42813.7	5.70	27195	67.15	37.03	Stable	-0.713
95.	TP03_0725 (268)	32082.9	9.74	57870	75.26	45.37	Unstable	-0.671
96.	TP03_0713 (268)	32082.9	9.74	57870	75.26	45.37	Unstable	-0.671
97.	TP03_0681 (373)	43283.0	5.44	35785	78.93	48.28	Unstable	-0.812
98.	TP03_0680 (181)	20790.5	5.02	11460	80.77	39.30	Stable	-0.551
99.	TP03_0678 (210)	22754.8	5.35	6990	74.71	34.27	Stable	-0.658
100.	TP03_0658 (165)	19763.0	10.17	13200	43.15	78.63	Unstable	-1.461
101.	TP03_0657 (1081)	123939.7	5.43	125015	73.14	35.05	Stable	-0.553
102.	TP03_0647 (670)	76410.8	5.11	32320	83.34	48.88	Unstable	-0.785
103.	TP03_0642 (248)	28790.0	9.20	20900	78.87	53.59	Unstable	-0.814
104.	TP03_0641 (299)	34437.7	4.45	33350	77.53	42.90	Unstable	-0.767
105.	TP03_0620 (452)	50618.4	5.20	20400	94.18	46.54	Unstable	-0.290
106.	TP03_0606 (163)	18756.8	8.71	10430	90.25	40.69	Unstable	-0.214
107.	TP03_0605 (146)	17218.2	7.08	35660	53.49	37.40	Stable	-0.999
108.	TP03_0597 (1509)	173402.7	6.65	84580	87.36	43.84	Unstable	- 0.708
109.	TP03_0581 (307)	35621.4	4.38	24870	70.78	70.23	Unstable	-0.745

110.	TP03_0564 (187)	21133.1	9.54	28420	122.46	31.32	Stable	0.419
111.	TP03_0557 (447)	52092.5	8.96	48360	98.95	31.80	Stable	-0.345
112.	TP03_0556 (657)	75819.1	9.84	60740	99.91	26.43	Stable	-0.267
113.	TP03_0537 (427)	48614.8	8.39	50880	92.60	32.63	Stable	-0.259
114.	TP03_0525 (243)	27891.5	9.48	16765	90.62	42.49	Unstable	-0.242
115.	TP03_0523 (238)	27797.2	9.24	16765	100.38	44.19	Unstable	-0.365
116.	TP03_0522 (115)	13273.3	8.80	7450	87.39	27.53	Stable	-0.327
117.	TP03_0484 (1029)	123555.6	9.85	183930	99.04	38.87	Stable	-0.337
118.	TP03_0483 (344)	39903.0	8.83	39545	79.27	49.56	Unstable	-0.672
119.	TP03_0475 (459)	51830.9	5.67	29340	99.50	35.96	Stable	-0.162
120.	TP03_0472 (125)	14131.9	9.50	5960	67.84	14.20	Stable	-0.991
121.	TP03_0471 (538)	62868.3	5.08	40145	95.43	42.81	Unstable	-0.456
122.	TP03_0463 (524)	62175.9	8.54	63610	93.65	42.35	Unstable	-0.195
123.	TP03_0437 (90)	10252.9	8.85	2980	100.67	39.40	Stable	-0.051
124.	TP03_0389 (321)	37038.7	5.82	24660	90.78	54.50	Unstable	-0.427
125.	TP03_0388 (195)	23335.2	9.86	17420	79.44	36.46	Stable	-1.045
126.	TP03_0381 (506)	59505.2	5.41	49740	124.74	32.15	Stable	0.105
127.	TP03_0380 (736)	84728.8	4.95	69400	118.83	35.90	Stable	0.069
128.	TP03_0378 (121)	13686.4	9.55	13075	57.19	26.62	Stable	-0.759
129.	TP03_0336 (1003)	116628.7	7.49	106760	87.83	35.29	Stable	-0.437
130.	TP03_0335 (200)	23853.5	6.09	26150	85.70	54.87	Unstable	-0.487
131.	TP03_0329 (317)	35387.1	5.68	23950	63.44	41.37	Unstable	-0.968
132.	TP03_0323 (178)	20049.6	4.95	14565	90.90	29.01	Stable	-0.379
133.	TP03_0310 (127)	14873.5	9.23	24200	93.62	37.20	Stable	0.031
134.	TP03_0305 (820)	94806.7	6.36	74315	82.43	26.20	Stable	-0.471
135.	TP03_0271 (275)	31016.4	5.52	22350	95.85	33.71	Unstable	-0.217
136.	TP03_0268 (1154)	132451.5	5.78	142450	84.32	46.28	Unstable	-0.374
137.	TP03_0265 (248)	28411.9	7.74	30370	75.48	41.78	Unstable	-0.629
138.	TP03_0256 (114)	12915.3	5.10	9315	49.56	42.76	Unstable	-0.690
139.	TP03_0255 (139)	16015.3	5.29	26025	105.90	43.46	Unstable	0.059
140.	TP03_0246 (310)	35036.2	5.11	15025	102.81	39.43	Stable	-0.267
141.	TP03_0234 (252)	27440.1	5.11	1615	82.34	39.42	Stable	-0.632
142.	TP03_0194 (368)	43060.9	6.57	69830	68.56	32.83	Stable	-0.784
143.	TP03_0193 (346)	40604.2	5.77	63830	71.76	41.04	Unstable	-0.745
144.	TP03_0177 (79)	9245.8	10.22	None	85.06	44.17	Unstable	-0.729
145.	TP03_0169 (239)	28193.2	5.43	35410	87.66	60.70	Unstable	-0.777
146.	TP03_0148	163498.3	6.29	199970	74.54	32.46	Stable	-0.612

	(1433)							
147.	TP03_0138 (351)	40078.6	7.74	19035	91.03	37.25	Stable	-0.584
148.	TP03_0136 (189)	20583.1	8.48	9970	98.62	14.01	Stable	-0.220
149.	TP03_0132 (112)	13103.6	4.87	16055	78.30	30.99	Stable	-0.522
150.	TP03_0125 (177)	20266.2	8.91	37400	122.03	35.04	Stable	0.846
151.	TP03_0123 (481)	56773.3	5.72	61450	92.56	35.46	Stable	-0.465
152.	TP03_0119 (198)	23129.6	7.68	14815	115.61	44.25	Unstable	-0.162
153.	TP03_0107 (913)	103149.5	5.05	91680	69.24	37.43	Stable	-0.714
154.	TP03_0098 (158)	18628.4	10.19	21555	87.59	29.87	Stable	-0.487
155.	TP03_0095 (229)	26607.2	6.33	23045	83.41	35.55	Stable	-0.683
156.	TP03_0094 (176)	19464.9	5.68	20065	104.60	36.62	Stable	-0.223
157.	TP03_0060 (137)	16171.6	9.54	9970	67.59	28.49	Stable	-1.174
158.	TP03_0055 (557)	65086.1	5.32	120000	66.16	43.44	Unstable	-0.797
159.	TP03_0051 (184)	20643.0	4.92	4595	90.71	51.50	Unstable	-0.534
160.	TP03_0045 (477)	54251.0	8.62	24870	87.44	40.55	Unstable	-0.340
161.	TP03_0042 (120)	14199.0	4.77	6990	104.00	52.85	Unstable	-0.832
162.	TP03_0038 (376)	43104.4	4.73	23380	92.74	51.18	Unstable	-0.632
163.	TP03_0034 (201)	22987.2	5.15	15470	62.54	38.71	Stable	-1.170
164.	TP03_0028 (544)	62440.2	4.90	43780	90.13	34.31	Stable	-0.386
165.	TP03_0024 (89)	9947.4	3.74	1490	59.10	37.66	Stable	-1.338
166.	TP03_0018 (98)	11444.4	11.44	1490	74.59	71.04	Unstable	-1.196
167.	TP03_0008 (911)	104106.8	6.75	100300	75.30	31.47	Stable	-0.507
168.	TP02_0965 (149)	17121.7	4.73	5960	53.62	40.87	Unstable	-1.069
169.	TP02_0916 (1070)	123077.1	4.43	83715	71.35	36.19	Stable	-0.766
170.	TP02_0910 (181), TP02_0914	19625.3	5.72	6335	82.98	25.30	Stable	-0.375
171.	TP02_0907 (766)	89717.7	5.56	100510	79.70	36.25	Stable	-0.567
172.	TP02_0897 (1499)	171336.1	6.59	194820	94.18	40.60	Unstable	0.058
173.	TP02_0880 (187)	21433.6	9.91	19940	96.95	42.19	Unstable	-0.225
174.	TP02_0874 (54)	6179.0	10.44	None	61.30	49.27	Unstable	-1.243
175.	TP02_0871 (163)	18846.6	6.98	2980	104.60	51.38	Unstable	-0.504
176.	TP02_0863 (222)	24812.4	3.58	10430	67.03	73.89	Unstable	-0.893
177.	TP02_0860 (246)	27822.2	9.70	19285	129.84	35.51	Stable	0.821
178.	TP02_0859 (556)	62601.4	5.49	52830	73.71	27.13	Stable	-0.766
179.	TP02_0849 (224)	26526.6	9.41	43320	104.87	29.14	Stable	0.651
180.	TP02_0812 (576)	64936.0	6.17	34380	98.42	29.63	Stable	-0.180
181.	TP02_0776	245981.2	8.69	212060	106.10	31.00	Stable	-0.141

	(2126)							
182.	TP02_0773 (172)	20497.9	6.91	15025	109.42	48.06	Unstable	-0.228
183.	TP02_0754 (85)	9481.2	8.53	6990	96.12	35.61	Stable	0.428
184.	TP02_0752 (476)	55586.6	5.89	51285	102.73	36.04	Stable	-0.151
185.	TP02_0711 (126)	14892.1	9.36	8940	92.78	31.49	Stable	-0.025
186.	TP02_0705 (272)	30440.6	5.74	34380	80.99	24.61	Stable	-0.099
187.	TP02_0695 (145)	16879.1	6.30	22835	92.76	43.00	Unstable	-0.568
188.	TP02_0687 (323)	37204.4	9.22	30745	107.43	20.70	Stable	-0.082
189.	TP02_0682 (160)	19393.7	9.96	31065	99.25	46.52	Unstable	-0.416
190.	TP02_0679 (123)	13771.6	10.26	3105	117.15	39.26	Stable	0.667
191.	TP02_0651 (228)	26349.3	10.14	33015	81.97	35.23	Stable	-0.535
192.	TP02_0644 (246)	28883.4	5.72	37820	64.59	35.62	Stable	-0.776
193.	TP02_0614 (80)	9559.8	6.55	15470	77.88	56.52	Unstable	-0.655
194.	TP02_0592 (1057)	122402.2	8.65	133510	87.28	30.15	Stable	-0.097
195.	TP02_0591 (1066)	122010.1	8.56	144510	90.21	37.41	Stable	-0.004
196.	TP02_0589 (798)	92619.1	9.06	72295	112.94	34.72	Stable	-0.040
197.	TP02_0586 (251)	28663.0	7.63	18700	99.76	28.84	Stable	-0.310
198.	TP02_0585 (183)	21186.4	6.24	12950	82.51	36.70	Stable	-0.458
199.	TP02_0583 (840)	98207.4	7.00	68190	106.18	40.15	Unstable	-0.170
200.	TP02_0582 (482)	54305.0	5.33	15275	70.56	42.40	Unstable	-0.674
201.	TP02_0575 (179)	20261.3	9.30	10430	107.82	36.67	Stable	-0.032
202.	TP02_0569 (438)	48871.3	7.20	24660	67.72	28.54	Stable	-0.597
203.	TP02_0555 (102)	11339.8	4.48	13410	92.65	36.67	Stable	-0.014
204.	TP02_0534 (298)	34689.6	5.52	36495	67.75	45.19	Unstable	-0.789
205.	TP02_0528 (552)	63819.8	8.41	91720	122.97	33.15	Stable	0.603
206.	TP02_0526 (547)	63895.3	5.20	37250	92.82	41.26	Unstable	-0.574
207.	TP02_0512 (1336)	155289.8	8.66	89970	89.52	35.30	Stable	-0.341
208.	TP02_0459 (104)	11813.5	7.62	12170	79.62	48.15	Unstable	-0.010
209.	TP02_0447 (411)	47335.8	4.81	42290	80.80	42.32	Unstable	-0.549
210.	TP02_0432 (402)	45836.6	9.09	26360	60.02	43.14	Unstable	-0.719
211.	TP02_0428 (243)	27445.4	9.52	17420	86.58	31.56	Stable	-0.466
212.	TP02_0427 (175)	20130.2	9.30	26930	48.57	26.20	Stable	-0.679
213.	TP02_0426 (89)	9799.2	7.78	6085	81.12	46.79	Unstable	0.081
214.	TP02_0424 (580)	63996.9	7.20	38850	70.03	44.48	Unstable	-0.684
215.	TP02_0420 (1743)	199489.3	6.14	141880	102.14	35.05	Stable	-0.129
216.	TP02_0419 (342)	37964.5	5.66	37610	123.60	34.29	Stable	0.680

217.	TP02_0410 (596)	68225.1	5.95	78645	101.49	34.49	Stable	-0.131
218.	TP02_0363 (889)	101955.3	5.30	73120	67.75	47.13	Unstable	-0.834
219.	TP02_0357 (201)	23493.9	9.07	14900	85.87	40.89	Unstable	-0.505
220.	TP02_0336 (212)	23961.0	8.52	14440	96.89	37.93	Stable	-0.390
221.	TP02_0308 (287)	33755.8	5.87	46300	84.46	43.51	Unstable	-0.497
222.	TP02_0302 (681)	79613.9	7.28	98475	93.89	36.00	Stable	-0.337
223.	TP02_0296 (132)	14864.8	5.14	6085	90.23	41.05	Unstable	-0.429
224.	TP02_0280 (144)	16852.5	9.78	17420	108.96	61.92	Unstable	-0.153
225.	TP02_0273 (355)	40798.3	6.99	32695	92.51	34.08	Stable	-0.271
226.	TP02_0267 (413)	47502.5	8.74	46300	88.96	45.10	Unstable	-0.059
227.	TP02_0213 (364)	41765.9	4.66	36245	82.94	47.74	Unstable	-0.602
228.	TP02_0192 (122)	14488.7	10.06	18575	64.67	50.99	Unstable	-0.720
229.	TP02_0150 (385)	43821.4	5.52	43820	82.21	46.19	Unstable	-0.559
230.	TP02_0147 (167)	19258.1	9.08	5960	89.16	31.75	Stable	-0.734
231.	TP02_0133 (601)	69208.0	5.37	38280	97.12	36.95	Stable	-0.279
232.	TP02_0109 (99)	11238.7	6.52	7700	95.56	35.36	Stable	-0.228
233.	TP02_0092 (102)	11709.0	9.19	5500	52.55	63.32	Unstable	-1.342
234.	TP02_0065 (272)	31126.1	5.39	14440	69.85	57.55	Unstable	-0.798
235.	TP02_0043 (583)	68040.9	4.85	42290	86.78	38.29	Stable	-0.578
236.	TP02_0026 (166)	19221.3	8.20	17420	106.27	34.93	Stable	-0.254
237.	TP02_0024 (570)	65673.8	8.46	69955	93.61	39.24	Stable	-0.192
238.	TP02_0009 (336)	39157.6	5.44	46315	59.64	60.08	Unstable	-1.022
239.	TP02_0007 (515)	60841.4	5.02	79540	53.69	65.20	Unstable	-1.135
240.	TP02_0006 (595)	69824.0	5.31	92950	53.03	61.16	Unstable	-1.234
241.	TP02_0005 (570)	67875.7	5.12	99035	51.07	65.50	Unstable	-1.211
242.	TP02_0004 (579)	68589.7	5.07	97670	50.62	68.36	Unstable	-1.211
243.	TP01_1228 (60)	6705.9	9.52	3105	116.83	16.83	Stable	0.032
244.	TP01_1208 (197)	22694.9	5.60	15930	103.35	38.27	Stable	-0.222
245.	TP01_1197 (286)	33344.9	9.27	35090	99.13	29.89	Stable	-0.056
246.	TP01_1179 (562)	63323.8	8.76	35300	66.71	42.64	Unstable	-0.713
247.	TP01_1146 (117)	13533.8	6.27	8480	107.44	56.37	Unstable	-0.259
248.	TP01_1123 (180)	20904.6	7.69	18910	83.89	31.31	Stable	-0.285
249.	TP01_1118 (650)	75259.1	5.24	59960	68.34	40.73	Unstable	-1.084
250.	TP01_1064 (348)	38390.2	5.34	11920	69.68	41.63	Unstable	-0.516
251.	TP01_1034 (204)	23839.4	9.14	21890	90.29	39.03	Stable	-0.640
252.	TP01_1026 (421)	49529.4	8.33	70710	99.24	36.54	Stable	-0.298
253.	TP01_1011 (356)	39808.6	3.47	11920	79.07	52.92	Unstable	-0.514
254.	TP01_1003 (171)	19427.5	6.19	10430	79.06	39.75	Stable	-0.474
255.	TP01_1001 (727)	83999.4	5.27	48820	95.71	45.53	Unstable	-0.293
256.	TP01_0993	129862.1	8.53	143940	93.28	31.73	Stable	0.029

	(1124)							
257.	TP01_0985 (139)	16163.6	8.62	17670	100.94	22.14	Stable	-0.214
258.	TP01_0953 (264)	30288.9	9.27	30745	91.21	37.88	Stable	-0.385
259.	TP01_0931 (274)	31846.8	6.97	26360	87.81	38.88	Stable	-0.764
260.	TP01_0917 (405)	45561.4	5.91	21890	73.88	50.40	Unstable	-0.833
261.	TP01_0912 (145)	16808.9	7.47	17420	96.07	34.38	Stable	0.034
262.	TP01_0891 (150)	17466.9	4.91	19940	65.00	48.85	Unstable	-0.549
263.	TP01_0890 (187)	21755.3	8.88	26065	78.72	46.35	Unstable	-0.304
264.	TP01_0878 (115)	12918.8	5.59	0	124.52	47.42	Unstable	-0.220
265.	TP01_0864 (373)	43306.0	9.28	38850	107.88	49.54	Unstable	-0.195
266.	TP01_0847 (642)	73616.5	5.87	67730	84.08	48.05	Unstable	-0.409
267.	TP01_0817 (409)	45743.8	8.82	30830	115.31	39.80	Stable	0.487
268.	TP01_0759 (944)	108531.5	9.26	85370	103.10	34.62	Stable	0.046
269.	TP01_0736 (589)	66027.2	4.39	20400	44.45	52.32	Unstable	-1.454
270.	TP01_0699 (76)	8655.4	5.38	2980	51.45	55.22	Unstable	-1.346
271.	TP01_0679 (543)	61292.1	6.63	57300	78.84	33.52	Stable	-0.444
272.	TP01_0676 (254)	29363.4	5.25	9970	95.98	50.47	Unstable	-0.583
273.	TP01_0671 (455)	52720.0	8.96	54250	110.86	26.64	Stable	-0.104
274.	TP01_0669 (427)	50380.4	4.96	65015	97.68	53.02	Unstable	-0.206
275.	TP01_0636 (151)	17626.5	9.92	7450	85.89	43.01	Unstable	-0.868
276.	TP01_0629 (110)	12311.3	9.72	8480	91.18	45.97	Unstable	-0.231
277.	TP01_0626 (178)	20518.5	10.25	15930	89.21	61.34	Unstable	-0.618
278.	TP01_0564 (162)	18336.8	5.68	6085	99.20	44.89	Unstable	-0.219
279.	TP01_0563 (393)	44019.1	8.73	33920	117.40	25.96	Stable	0.574
280.	TP01_0559 (240)	28635.7	10.04	37360	29.62	62.64	Unstable	-1.930
281.	TP01_0555 (146)	16810.6	11.55	0	73.36	46.13	Unstable	-1.062
282.	TP01_0554 (520)	59773.9	5.09	36330	100.87	29.39	Stable	-0.274
283.	TP01_0537 (183)	21002.5	5.36	8940	111.69	37.24	Stable	0.061
284.	TP01_0493 (234)	26870.8	4.03	16180	77.39	71.96	Unstable	-0.384
285.	TP01_0457 (163)	18077.4	9.22	12950	78.34	50.86	Unstable	-0.292
286.	TP01_0392 (257)	29104.9	5.23	16390	59.96	41.63	Unstable	-1.112
287.	TP01_0391 (680)	76638.4	8.29	71210	75.34	30.32	Stable	-0.688
288.	TP01_0346 (453)	53055.1	7.65	41050	135.43	30.04	Stable	0.150
289.	TP01_0307 (584)	65702.3	9.11	61155	89.85	29.23	Stable	-0.195
290.	TP01_0306 (867)	99221.0	7.57	74245	105.57	31.52	Stable	-0.048
291.	TP01_0305 (78)	9235.6	7.78	20400	102.31	33.05	Stable	-0.224
292.	TP01_0270 (151)	17791.5	10.35	12045	100.66	51.98	Unstable	-0.584
293.	TP01_0255 (155)	18043.0	9.89	17085	82.97	56.18	Unstable	-0.802
294.	TP01_0144 (1218)	142086.7	8.44	181855	87.15	35.12	Stable	-0.396

295.	TP01_0143 (829)	93068.4	6.47	57190	80.60	35.39	Stable	-0.339
296.	TP01_0138 (131)	15631.4	11.01	14440	74.43	44.82	Unstable	-0.592
297.	TP01_0135 (215)	24027.8	8.94	35410	58.60	33.53	Stable	-0.692
298.	TP01_0124 (345)	39573.7	4.75	30830	75.16	34.23	Stable	-0.385
299.	TP01_0090 (424)	49141.3	5.23	42010	99.27	43.26	Unstable	-0.209
300.	TP01_0061 (417)	48723.1	8.38	75205	96.26	49.34	Unstable	-0.329
301.	TP01_0038 (458)	51680.9	4.63	27850	73.58	53.21	Unstable	-0.814
302.	TP01_0034 (625)	72483.8	8.86	69680	93.98	40.49	Unstable	-0.225
303.	TP01_0027 (127)	14805.6	8.95	18910	73.70	55.33	Unstable	-0.794
304.	TP01_0026 (722)	84721.3	9.33	83660	110.76	41.63	Unstable	0.070
305.	TP01_0009 (498)	58375.6	6.04	70140	72.69	53.66	Unstable	-0.830
306.	TP01_0008 (444)	51980.3	4.47	57775	61.69	64.32	Unstable	-1.049
307.	TP01_0007 (510)	60190.1	5.04	88855	59.76	64.36	Unstable	-1.004
308.	TP01_0004 (468)	54244.9	5.30	59140	75.56	73.10	Unstable	-0.791
309.	TP03_0648 (98)	11353.1	9.51	31050	118.27	17.53	Stable	-0.219

Appendix A.2. Subcellular localization, Signal peptide and Trans-membrane helices predictions of the 309 *T. parva* HPs.

Sequence. No	HP gene names (protein length)	Subcellular localization prediction				Signal peptides		Trans-membrane helices	
		Wolf PSORT	TargetP	YLoc	Consensus	SignalP	SecretomeP	HMMTOP	TMHMM
	TP01_0004 (468)	Endoplasmic Reticulum	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	1	1
2.	TP01_0007 (510)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	0	0
3.	TP01_0008 (444)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	0	0
4.	TP01_0009 (498)	Cytoplasm	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	0	0
5.	TP01_0026 (722)	Plasma membrane	-	Cytoplasm	Cytoplasm	No	No	3	0
6.	TP01_0027 (127)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
7.	TP01_0034 (625)	Mitochondrion	Mitochondrion	Cytoplasm	Mitochondrion	No	No	0	0
8.	TP01_0038 (458)	Nucleus	-	Cytoplasm	Nucleus	No	No	0	0
9.	TP01_0061 (417)	Mitochondrion	Mitochondrion	Mitochondrion	Mitochondrion	No	No	1	0
10.	TP01_0090 (424)	Nucleus	-	Cytoplasm	Cytoplasm	No	No	0	0
11.	TP01_0124 (345)	Nucleus	-	Cytoplasm	Cytoplasm	No	No	0	0
12.	TP01_0135 (215)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
13.	TP01_0138 (131)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	0	0
14.	TP01_0143 (829)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	No	Yes	1	1
15.	TP01_0144 (1218)	Endoplasmic Reticulum	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	1	1
16.	TP01_0255 (155)	Mitochondrion	-	Nucleus	Multiple locations	No	No	0	0
17.	TP01_0270 (151)	Mitochondrion	Mitochondrion	Secreted pathway	Mitochondrion	No	No	1	0
18.	TP01_0305 (78)	Cytoplasm	-	Secreted pathway	Cytoplasm	No	No	1	0
19.	TP01_0306 (867)	Plasma membrane	-	Cytoplasm	Cytoplasm	No	No	0	0
20.	TP01_0307 (584)	Nucleus	-	Cytoplasm	Nucleus	No	No	0	0
21.	TP01_0346 (453)	Mitochondrion	-	Secreted pathway	Multiple locations	No	No	2	0
22.	TP01_0391 (680)	Nucleus	-	Cytoplasm	Nucleus	No	No	1	0
23.	TP01_0392 (257)	Extracellular	Secretory pathway	Nucleus	Multiple locations	Yes	Yes	1	1
24.	TP01_0457 (163)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	1	0
25.	TP01_0493 (234)	Plasma membrane	Secretory pathway	Secreted pathway	Secretory pathway	No	No	3	0
26.	TP01_0537 (183)	Cytoplasm	-	Secreted pathway	Cytoplasm	No	No	1	0
27.	TP01_0554 (520)	Endoplasmic Reticulum	Secretory pathway	Secreted pathway	Secretory pathway	No	Yes	1	0
28.	TP01_0555 (146)	Nucleus	Mitochondrion	Mitochondrion	Mitochondrion	No	No	0	0
29.	TP01_0559 (240)	Cytoplasm	-	Nucleus	Nucleus	No	No	0	0
30.	TP01_0563 (393)	Plasma membrane	Secretory pathway	Secreted pathway	Secretory pathway	No	Yes	7	8
31.	TP01_0564 (162)	Cytoplasm	-	Secreted pathway	Cytoplasm	No	No	0	0

32.	TP01_0626 (178)	Mitochondrion	Mitochondrion	Mitochondrion	Mitochondrion	No	No	0	0
33.	TP01_0629 (110)	Extracellular	-	Secreted pathway	Multiple locations	No	No	0	0
34.	TP01_0636 (151)	Nucleus	-	Nucleus	Nucleus	No	No	0	0
35.	TP01_0669 (427)	Cytoplasm	-	Secreted pathway	Cytoplasm	No	No	2	1
36.	TP01_0671 (455)	Mitochondrion	Mitochondrion	Mitochondrion	Mitochondrion	No	No	0	0
37.	TP01_0676 (254)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
38.	TP01_0679 (543)	Microsome	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	1	1
39.	TP01_0699 (76)	Nucleus	-	Nucleus	Nucleus	No	No	0	0
40.	TP01_0736 (589)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	1	0
41.	TP01_0759 (944)	Plasma Membrane	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	9	8
42.	TP01_0817 (409)	Plasma membrane	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	10	9
43.	TP01_0847 (642)	Extracellular	-	Cytoplasm	Cytoplasm	No	No	1	0
44.	TP01_0864 (373)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	1	0
45.	TP01_0878 (115)	Cytoplasm	-	Secreted pathway	Cytoplasm	No	No	0	0
46.	TP01_0890 (187)	Extracellular	-	Cytoplasm	Cytoplasm	No	No	0	0
47.	TP01_0891 (150)	Cytoplasm	-	Secreted pathway	Cytoplasm	No	No	0	0
48.	TP01_0912 (145)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	0	0
49.	TP01_0917 (405)	Cytoplasm	-	Nucleus	Nucleus	No	No	0	0
50.	TP01_0931 (274)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
51.	TP01_0953 (264)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
52.	TP01_0985 (139)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
53.	TP01_0993 (1124)	Plasma membrane	Secretory pathway	Secreted pathway	Secretory pathway	No	Yes	13	14
54.	TP01_1001 (727)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
55.	TP01_1003 (171)	Nucleus	-	Cytoplasm	Cytoplasm	No	No	0	0
56.	TP01_1011 (356)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	3	0
57.	TP01_1026 (421)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
58.	TP01_1034 (204)	Mitochondrion	Mitochondrion	Mitochondrion	Mitochondrion	No	No	0	0
59.	TP01_1064 (348)	Mitochondrion	Mitochondrion	Nucleus	Mitochondrion	No	No	0	0
60.	TP01_1118 (650)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
61.	TP01_1123 (180)	Mitochondrion	Mitochondrion	Mitochondrion	Mitochondrion	No	No	1	0
62.	TP01_1146 (117)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	1	0
63.	TP01_1179 (562)	Nucleus	-	Nucleus	Nucleus	No	No	0	0
64.	TP01_1197 (286)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	1	0
65.	TP01_1208 (197)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
66.	TP01_1228 (60)	Extracellular	-	Secreted pathway	Multiple locations	No	No	0	0
67.	TP02_0004 (579)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	1	0
68.	TP02_0005 (570)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	0	0
69.	TP02_0006 (595)	Mitochondrion	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	0	0
70.	TP02_0007 (515)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	0	0
71.	TP02_0009 (336)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	0	0
72.	TP02_0024 (570)	Nucleus	-	Cytoplasm	Nucleus	No	No	0	0
73.	TP02_0026 (166)	Extracellular	-	Secreted pathway	Multiple locations	No	No	0	0
74.	TP02_0043 (583)	Cytoplasm	-	Nucleus	Nucleus	No	No	0	0

75.	TP02_0065 (272)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	1	1
76.	TP02_0092 (102)	Nucleus	-	Nucleus	Nucleus	No	No	0	0
77.	TP02_0109 (99)	Extracellular	-	Secreted pathway	Multiple locations	No	No	0	0
78.	TP02_0133 (601)	Nucleus	-	Cytoplasm	Nucleus	No	No	0	0
79.	TP02_0147 (167)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
80.	TP02_0150 (385)	Nucleus	-	Nucleus	Nucleus	No	No	0	0
81.	TP02_0192 (122)	Mitochondrion	Mitochondrion	Mitochondrion	Mitochondrion	No	No	0	0
82.	TP02_0213 (364)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	0	0
83.	TP02_0267 (413)	Plasma membrane	-	Nucleus	Plasma membrane	No	No	4	5
84.	TP02_0273 (355)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	No	Yes	0	1
85.	TP02_0280 (144)	Mitochondrion	Mitochondrion	Nucleus	Mitochondrion	No	No	0	0
86.	TP02_0296 (132)	Nucleus	-	Cytoplasm	Cytoplasm	No	No	0	0
87.	TP02_0302 (681)	Mitochondrion	Mitochondrion	Cytoplasm	Mitochondrion	No	No	0	0
88.	TP02_0308 (287)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
89.	TP02_0336 (212)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
90.	TP02_0357 (201)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
91.	TP02_0363 (889)	Mitochondrion	Mitochondrion	Nucleus	Mitochondrion	No	No	0	0
92.	TP02_0410 (596)	Nucleus	-	Cytoplasm	Cytoplasm	No	No	0	0
93.	TP02_0419 (342)	Plasma membrane	Secretory pathway	Secreted pathway	Secretory pathway	No	Yes	6	6
94.	TP02_0420 (1743)	Plasma membrane	-	Cytoplasm	Cytoplasm	No	No	0	0
95.	TP02_0424 (580)	Nucleus	-	Nucleus	Nucleus	No	No	1	0
96.	TP02_0426 (89)	Extracellular	-	Secreted pathway	Multiple locations	No	No	0	0
97.	TP02_0427 (175)	Mitochondrion	-	Mitochondrion	Mitochondrion	No	No	1	1
98.	TP02_0428 (243)	Nucleus	-	Mitochondrion	Multiple locations	No	No	0	0
99.	TP02_0432 (402)	Cytoplasm	-	Nucleus	Nucleus	No	No	0	0
100.	TP02_0447 (411)	Nucleus	-	Cytoplasm	Cytoplasm	No	No	0	0
101.	TP02_0459 (104)	Extracellular	-	Secreted pathway	Multiple locations	No	No	0	4
102.	TP02_0512 (1336)	Plasma membrane	Secretory pathway	Secreted pathway	Secreted pathway	Yes	Yes	3	4
103.	TP02_0526 (547)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
104.	TP02_0528 (552)	Plasma membrane	-	Cytoplasm	Multiple locations	No	No	12	9
105.	TP02_0534 (298)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
106.	TP02_0555 (102)	Plasma membrane	-	Cytoplasm	Cytoplasm	No	No	1	1
107.	TP02_0569 (438)	Nucleus	-	Cytoplasm	Cytoplasm	No	No	0	0
108.	TP02_0575 (179)	Plasma membrane	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	2	0
109.	TP02_0582 (482)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
110.	TP02_0583 (840)	Mitochondrion	Mitochondrion	Mitochondrion	Mitochondrion	No	No	0	0
111.	TP02_0585 (183)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
112.	TP02_0586 (251)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0

113.	TP02_0589 (798)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
114.	TP02_0591 (1066)	Plasma membrane	Secretory pathway	Secreted pathway	Secreted pathway	No	No	10	12
115.	TP02_0592 (1057)	Plasma membrane	Secretory pathway	Secreted pathway	Secretory pathway	No	Yes	9	10
116.	TP02_0614 (80)	Cytoplasm	-	Secreted pathway	Cytoplasm	No	No	0	0
117.	TP02_0644 (246)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
118.	TP02_0651 (228)	Mitochondrion	-	Secreted pathway	Multiple locations	No	No	0	0
119.	TP02_0679 (123)	Plasma membrane	Secretory pathway	Secreted pathway	Secreted pathway	No	No	2	2
120.	TP02_0682 (160)	Mitochondrion	Mitochondrion	Mitochondrion	Mitochondrion	No	No	0	0
121.	TP02_0687 (323)	Microxisomal	Secretory pathway	Cytoplasm	Multiple locations	No	No	2	0
122.	TP02_0695 (145)	Nucleus	-	Cytoplasm	Cytoplasm	No	No	0	0
123.	TP02_0705 (272)	Extracellular	Secretory pathway	Secreted pathway	Secreted pathway	No	Yes	1	1
124.	TP02_0711 (126)	Mitochondrion	Secretory pathway	Secreted pathway	Secreted pathway	No	Yes	1	1
125.	TP02_0752 (476)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
126.	TP02_0754 (85)	Extracellular	Secretory pathway	Secreted pathway	Secreted pathway	Yes	Yes	1	0
127.	TP02_0773 (172)	Cytoplasm	-	Nucleus	Cytoplasm	No	No	2	0
128.	TP02_0776(2126)	Plasma membrane	-	Secreted pathway	Multiple locations	No	No	0	0
129.	TP02_0812 (576)	Extracellular	Secretory pathway	Mitochondrion	Multiple locations	No	Yes	1	1
130.	TP02_0849 (224)	Extracellular	Secretory pathway	Secreted pathway	Secreted pathway	No	Yes	6	4
131.	TP02_0859 (556)	Nucleus	-	Cytoplasm	Nucleus	No	No	0	0
132.	TP02_0860 (246)	Plasma membrane	Secretory pathway	Secreted pathway	Secretory pathway	No	Yes	6	0
133.	TP02_0863 (222)	Extracellular	-	Cytoplasm	Cytoplasm	No	No	0	0
134.	TP02_0871 (163)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
135.	TP02_0874 (54)	Nucleus	-	Mitochondrion	Multiple locations	No	No	0	0
136.	TP02_0880 (187)	Mitochondrion	Mitochondrion	Mitochondrion	Mitochondrion	No	No	2	1
137.	TP02_0897 (1499)	Plasma membrane	Secretory pathway	Secreted pathway	Secretory pathway	No	No	12	12
138.	TP02_0907 (766)	Endoplasm Reticulum	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	0	0
139.	TP02_0910,TP02_0914 (181)	Nucleus	-	Secreted pathway	Multiple locations	No	No	0	0
140.	TP02_0916 (1070)	Nucleus	-	Cytoplasm	Nucleus	No	No	0	1
141.	TP02_0965 (149)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
142.	TP03_0008 (911)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	2	1
143.	TP03_0018 (98)	Nucleus	-	Nucleus	Nucleus	No	No	0	0
144.	TP03_0024 (89)	Extracellular	-	Nucleus	Multiple locations	No	No	0	0
145.	TP03_0028 (544)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
146.	TP03_0034 (201)	Nucleus	-	Nucleus	Nucleus	No	No	0	0
147.	TP03_0038 (376)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	2	0
148.	TP03_0042 (120)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
149.	TP03_0045 (477)	Plasma	-	Nucleus	Plasma membrane	No	No	4	4

		membrane							
150.	TP03_0051 (184)	Cytoplasm	-	Nucleus	Cytoplasm	No	No	0	0
151.	TP03_0055 (557)	Mitochondrion	Mitochondrion	Mitochondrion	Mitochondrion	No	No	0	0
152.	TP03_0060 (137)	Extracellular	-	Nucleus	Multiple locations	No	Yes	0	0
153.	TP03_0094 (176)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
154.	TP03_0095 (229)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
155.	TP03_0098 (158)	Mitochondrion	-	Secreted pathway	Multiple locations	No	No	1	1
156.	TP03_0107 (913)	Nucleus	-	Cytoplasm	Nucleus	No	No	0	0
157.	TP03_0119 (198)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
158.	TP03_0123 (481)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	0	0
159.	TP03_0125 (177)	Plasma membrane	Secretory pathway	Secreted pathway	Secretory pathway	No	Yes	4	4
160.	TP03_0132 (112)	Mitochondrion	-	Secreted pathway	Multiple locations	No	No	0	0
161.	TP03_0136 (189)	Extracellular	Secretory pathway	Secreted pathway	Secreted pathway	Yes	Yes	1	0
162.	TP03_0138 (351)	Nucleus	-	Nucleus	Nucleus	No	No	0	0
163.	TP03_0148 (1433)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	No	No	1	1
164.	TP03_0169 (239)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
165.	TP03_0177 (79)	Nucleus	-	Nucleus	Nucleus	No	No	0	0
166.	TP03_0193 (346)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
167.	TP03_0194 (368)	Cytoplasm	-	Nucleus	Nucleus	No	No	0	0
168.	TP03_0234 (252)	Mitochondrion	Secretory pathway	Nucleus	Multiple locations	No	No	1	0
169.	TP03_0246 (310)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
170.	TP03_0255 (139)	Cytoplasm	-	Secreted pathway	Cytoplasm	No	No	0	0
171.	TP03_0256 (114)	Extracellular	-	Cytoplasm	Cytoplasm	No	No	0	0
172.	TP03_0265 (248)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
173.	TP03_0268 (1154)	Plasma membrane	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	0	0
174.	TP03_0271 (275)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
175.	TP03_0305 (820)	Mitochondrion	-	Secreted pathway	Multiple locations	No	No	0	0
176.	TP03_0310 (127)	Extracellular	Secretory pathway	Mitochondrion	Multiple locations	No	Yes	1	0
177.	TP03_0323 (178)	Cytoplasm	-	Secreted pathway	Cytoplasm	No	No	0	0
178.	TP03_0329 (317)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
179.	TP03_0335 (200)	Cytoplasm	Secretory pathway	Secreted pathway	Secretory pathway	No	No	0	0
180.	TP03_0336 (1003)	Endoplasmic Reticulum	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	1	1
181.	TP03_0378 (121)	Cytoplasm	-	Nucleus	Nucleus	No	No	1	1
182.	TP03_0380 (736)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
183.	TP03_0381 (506)	Plasma membrane	-	Secreted pathway	Multiple locations	No	No	4	0
184.	TP03_0388 (195)	Nucleus	Mitochondrion	Nucleus	Nucleus	No	No	0	0
185.	TP03_0389 (321)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0

186.	TP03_0437 (90)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	1	0
187.	TP03_0463 (524)	Endoplasmic Reticulum	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	1	0
188.	TP03_0471 (538)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
189.	TP03_0472 (125)	Nucleus	-	Nucleus	Nucleus	No	No	0	0
190.	TP03_0475 (459)	Nucleus	-	Cytoplasm	Nucleus	No	No	0	0
191.	TP03_0483 (344)	Mitochondrion	Mitochondrion	Mitochondrion	Mitochondrion	No	No	1	1
192.	TP03_0484 (1029)	Nucleus	-	Nucleus	Nucleus	No	No	0	0
193.	TP03_0522 (115)	Mitochondrion	Mitochondrion	Mitochondrion	Mitochondrion	No	No	0	0
194.	TP03_0523 (238)	Cytoplasm	-	Mitochondrion	Cytoplasm	No	No	0	0
195.	TP03_0525 (243)	Cytoplasm	-	Secreted pathway	Cytoplasm	No	No	0	0
196.	TP03_0537 (427)	Endoplasmic Reticulum	Secretory pathway	Secreted pathway	Secretory pathway	No	No	1	1
197.	TP03_0556 (657)	Nucleus	-	Nucleus	Nucleus	No	No	0	0
198.	TP03_0557 (447)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
199.	TP03_0564 (187)	Cytoplasm	Secretory pathway	Secreted pathway	Secretory pathway	No	No	4	3
200.	TP03_0581 (307)	Endoplasmic Reticulum	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	1	0
201.	TP03_0597 (1509)	Mitochondrion	Mitochondrion	Cytoplasm	Mitochondrion	No	No	0	0
202.	TP03_0605 (146)	Nucleus	-	Nucleus	Nucleus	No	No	0	0
203.	TP03_0606 (163)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	1	0
204.	TP03_0620 (452)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	1	1
205.	TP03_0641 (299)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
206.	TP03_0642 (248)	Nucleus	-	Cytoplasm	Cytoplasm	No	No	0	0
207.	TP03_0647 (670)	Nucleus	-	Cytoplasm	Nucleus	No	No	0	0
208.	TP03_0648 (98)	Cytoplasm	-	Secreted pathway	Cytoplasm	No	No	0	0
209.	TP03_0657 (1081)	Plasma membrane	-	Cytoplasm	Multiple locations	No	No	3	0
210.	TP03_0658 (165)	Nucleus	Mitochondrion	Nucleus	Nucleus	No	No	0	0
211.	TP03_0678 (210)	Cytoplasm	Secretory pathway	Nucleus	Multiple locations	No	Yes	1	1
212.	TP03_0680 (181)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	1	1
213.	TP03_0681 (373)	Nucleus	-	Nucleus	Nucleus	No	No	1	0
214.	TP03_0713 (268)	Nucleus	Mitochondrion	Nucleus	Nucleus	No	No	0	0
215.	TP03_0725 (290)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	1	1
216.	TP03_0729 (379)	Cytoplasm	-	Nucleus	Cytoplasm	No	No	0	0
217.	TP03_0738 (666)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
218.	TP03_0742 (273)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	No	Yes	0	0
219.	TP03_0754 (819)	Plasma membrane	-	Cytoplasm	Multiple locations	No	No	4	0
220.	TP03_0759 (195)	Cytoplasm	-	Secreted pathway	Cytoplasm	No	No	0	0

221.	TP03_0761 (185)	Mitochondrion	Secretory pathway	Secreted pathway	Secretory pathway	No	No	1	0
222.	TP03_0768 (346)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
223.	TP03_0779 (300)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	1	0
224.	TP03_0812 (64)	Mitochondrion	-	Secreted pathway	Multiple locations	No	No	1	0
225.	TP03_0819 (112)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
226.	TP03_0820 (255)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
227.	TP03_0825 (446)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
228.	TP03_0827 (885)	Mitochondrion	Secretory pathway	Cytoplasm	Multiple locations	No	Yes	1	0
229.	TP03_0832 (155)	Cytoplasm	-	Secreted pathway	Cytoplasm	No	No	2	0
230.	TP03_0839 (260)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
231.	TP03_0850 (397)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	No	Yes	1	0
232.	TP03_0856 (527)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	1	0
233.	TP03_0870 (563)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	0	0
234.	TP03_0873 (510)	Endoplasmic Reticulum	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	0	0
235.	TP03_0875 (509)	Endoplasmic Reticulum	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	0	0
236.	TP03_0877 (341)	Mitochondrion	-	Secreted pathway	Multiple locations	No	No	0	0
237.	TP03_0880 (609)	Endoplasmic Reticulum	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	0	0
238.	TP03_0881 (568)	Endoplasmic Reticulum	Secretory pathway	Secreted pathway	Secretory pathway	No	No	0	0
239.	TP03_0882 (607)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	0	0
240.	TP03_0883 (537)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	0	0
241.	TP03_0885 (481)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	0	0
242.	TP03_0893 (568)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	0	0
243.	TP03_0896 (524)	Endoplasmic Reticulum	Secretory pathway	Secreted pathway	Secretory pathway	No	Yes	1	1
244.	TP03_0898 (223)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	1	0
245.	TP03_0899 (524)	Endoplasmic Reticulum	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	0	0
246.	TP03_0900 (521)	Endoplasmic Reticulum	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	0	0
247.	TP03_0901 (522)	Endoplasmic Reticulum	Secretory pathway	Secreted pathway	Secretory pathway	No	Yes	1	1
248.	TP03_0903 (310)	Mitochondrion	Mitochondrion	Mitochondrion	Mitochondrion	No	No	0	0
249.	TP04_0003 (561)	Endoplasmic Reticulum	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	0	0
250.	TP04_0006 (522)	Endoplasmic Reticulum	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	0	0

251.	TP04_0068 (233)	Microsome	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	1	1
252.	TP04_0069 (228)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
253.	TP04_0073 (379)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
254.	TP04_0077 (55)	Cytoplasm	-	Secreted pathway	Cytoplasm	No	No	0	0
255.	TP04_0081 (935)	Plasma membrane	-	Cytoplasm	Multiple locations	No	No	7	7
256.	TP04_0082 (207)	Cytoskeleton	-	Secreted pathway	Multiple locations	No	No	1	0
257.	TP04_0087 (356)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
258.	TP04_0114 (198)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
259.	TP04_0121 (114)	Mitochondrion	-	Mitochondrion	Mitochondrion	No	No	0	0
260.	TP04_0127 (540)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
261.	TP04_0128 (587)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
262.	TP04_0144 (370)	Mitochondrion	-	Nucleus	Nucleus	No	No	0	0
263.	TP04_0171 (306)	Mitochondrion	Mitochondrion	Mitochondrion	Mitochondrion	No	No	2	1
264.	TP04_0172 (912)	Plasma Membrane	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	3	1
265.	TP04_0181 (67)	Nucleus	-	Nucleus	Nucleus	No	No	0	0
266.	TP04_0188 (199)	Nucleus	-	Secreted pathway	Multiple locations	No	No	0	0
267.	TP04_0190 (74)	Cytoplasm	Mitochondrion	Secreted pathway	Multiple locations	No	No	1	0
268.	TP04_0192 (110)	Cytoplasm	-	Secreted pathway	Cytoplasm	No	No	0	0
269.	TP04_0200 (197)	Plasma membrane	-	Mitochondrion	Plasma membrane	No	No	0	0
270.	TP04_0210 (384)	Nucleus	-	Nucleus	Nucleus	No	No	0	0
271.	TP04_0223 (48)	Cytoplasm	-	Secreted pathway	Cytoplasm	No	No	0	0
272.	TP04_0232 (884)	Nucleus	-	Cytoplasm	Cytoplasm	No	No	0	0
273.	TP04_0237 (164)	Nucleus	Mitochondrion	Mitochondrion	Mitochondrion	No	No	0	0
274.	TP04_0240 (327)	Cytoplasm	Secretory pathway	Secreted pathway	Secretory pathway	No	No	0	0
275.	TP04_0245 (816)	Nucleus	-	Cytoplasm	Cytoplasm	No	No	0	0
276.	TP04_0252 (649)	Endoplasmic Reticulum	Secretory pathway	Secreted pathway	Secretory pathway	No	Yes	1	0
277.	TP04_0254 (220)	Cytoplasm	-	Secreted pathway	Cytoplasm	No	No	0	0
278.	TP04_0259 (137)	Cytoplasm	Secretory pathway	Secreted pathway	Secreted pathway	No	No	0	0
279.	TP04_0275 (2405)	Nucleus	-	Nucleus	Nucleus	No	No	0	0
280.	TP04_0283 (206)	Plasma membrane	-	Secreted pathway	Multiple locations	No	No	2	2
281.	TP04_0327 (507)	Cytoplasm	-	Nucleus	Nucleus	No	No	0	0
282.	TP04_0353 (477)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
283.	TP04_0399 (357)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	2	2
284.	TP04_0405 (718)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
285.	TP04_0414 (768)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0

286.	TP04_0422 (129)	Extracellular	Secretory pathway	Secreted pathway	Secreted pathway	Yes	Yes	1	0
287.	TP04_0455 (556)	Plasma membrane	-	Cytoplasm	Plasma membrane	No	No	11	10
288.	TP04_0503 (245)	Nucleus	-	Nucleus	Nucleus	No	No	0	0
289.	TP04_0505 (74)	Cytoplasm	-	Secreted pathway	Cytoplasm	No	No	1	1
290.	TP04_0532 (206)	Mitochondrion	-	Cytoplasm	Cytoplasm	No	No	0	0
291.	TP04_0576 (447)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
292.	TP04_0579 (106)	Cytoplasm	-	Mitochondrion	Cytoplasm	No	No	0	0
293.	TP04_0633 (193)	Nucleus	Mitochondrion	Mitochondrion	Mitochondrion	No	No	3	3
294.	TP04_0638 (64)	Nucleus	-	Cytoplasm	Cytoplasm	No	No	0	0
295.	TP04_0654 (591)	Plasma membrane	-	Cytoplasm	Plasma membrane	No	No	7	6
296.	TP04_0693 (221)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
297.	TP04_0708 (349)	Plasma membrane	Secretory pathway	Secreted pathway	Secreted pathway	No	Yes	4	4
298.	TP04_0715 (212)	Nucleus	-	Cytoplasm	Nucleus	No	No	0	0
299.	TP04_0729 (218)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	2	0
300.	TP04_0786 (381)	Nucleus	-	Nucleus	Nucleus	No	No	1	0
301.	TP04_0833 (147)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	1	1
302.	TP04_0834 (398)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
303.	TP04_0869 (283)	Extracellular	-	Cytoplasm	Cytoplasm	No	No	0	0
304.	TP04_0896 (489)	Endoplasmic Reticulum	Secretory pathway	Secreted pathway	Secretory pathway	No	Yes	1	1
305.	TP04_0903 (170)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
306.	TP04_0905 (676)	Cytoplasm	-	Cytoplasm	Cytoplasm	No	No	0	0
307.	TP04_0910 (276)	Extracellular	Secretory pathway	Secreted pathway	Secretory pathway	Yes	Yes	1	0
308.	TP05_0020 (133)	Plasma membrane	Secretory pathway	Secreted pathway	Secretory pathway	No	Yes	4	4
309.	TP05_0039 (126)	Cytoplasm	Secretory pathway	Secreted pathway	Secretory pathway	No	No	3	2

Appendix A.3. Examples of the 134 HPs which were assigned relevant functions with low level of confidence.

(A)

```
TP01_1026       ------------------------------------------------------------  0
NP_010514.3     MDHDTEVIVKDFNSILEELTFNSRPIITTLTKLAEENISCAQYFVDAIESRIEKCMPKQK  60


TP01_1026       ------MNGLTKFSNIPSTKWLHFNFITSIYNLYTKNNTHPNSNYVRELPRLDHKSAFKP  54
NP_010514.3     LYAFYALDSICKNVGSPYTIY----FSRNLFNLYKRTYLLVDN-----TTRTKLINMFKL  111
                      ::.: *    * * :    *  .::***.:.    :.       * .  . **


TP01_1026       YEKEENEENDSISHLPGRLICHL---------NKHRE---SLLR---SDLYYGNITKWAY-  99
NP_010514.3     WLNP---NDTGLPLFEGSALEKIEQFLIKASALHQKNLQAMLPTPTVPLLLRDIDKLTCL  168
                : :    :: .:  : *  : ::          *::   ::*      *   :* * :


TP01_1026       -------------YGLFSVVIYNWNSIHKREKLALMN--SIVK------------IIRSI  132
NP_010514.3     TSERLKNQPNDEKLKMKLLVLSQLKQELKREKLTLNALKQVQMQLRQVFSQDQQVLQERM  228
                             :  :*: : :.  *****:*    .:              : . :


TP01_1026       SYNSSENNP---------NSLNLDNYKQSLIVFLKFILLSDSNTLSNINNTSFVSDNCIN  183
NP_010514.3     RYHELQQQQQQQQQQQQQQQQQQQQYHETKDMVGSYTQNSNS-AIPLFGNNS---D-TTN  283
                 *.. :::          :. : ::*:::  :. .:   *:* ::  : *.*   *   *


TP01_1026       NWKYVYNCIHNCVNNLNSFDSVGNGTFTELVNECNYVRDFQKLFSVLYWVKSLKIIYSQL  243
NP_010514.3     QQNSLSSSLFGNISGVESFQEIEK------------KKSLNKINNLYASLKAEGLIYTPP  331
                : : : ..:.  :. ::**:.: :             :.::*: .:   :*:  :**:


TP01_1026       ELLSSAVLRRLK--SVYDEDCDLRDIINISYYTSMIDPKLLSDGFISDCLFSR-------  294
NP_010514.3     KESIVTLYKKLNGHSNYSLDSHEKQLMK-----N-LPKIPLLNDILSDCKAYFATVNIDV  385
                :    :: ::*:  * *. *.. :::::     . :    * : ::***


TP01_1026       FNNRYHS--------LNVIDCYHLLHFLKT-------YLNYSLGNE-----IQYTFLRLA  334
NP_010514.3     LNNPSLQLSEQTLLQENPIVQNNLIHLLYRSKPNKCSVCGKRFGNSESEKLLQNEHLDWH  445
                :**   .         * *   .*:*:*              :**.     :*  .*


TP01_1026       FNILVNIH----------------------------CYINNSGELDIKEL--------EE  358
NP_010514.3     FRINTRIKGSQNTANTGISNSNLNTTTTRKNIQSRNWYLSDSQWAAFKDDEITSTKHKND  505
                *.* ..*:                             *:.:*    :*:         ::


TP01_1026       LSNKIFNYKVDLTCLNQKNQPNE-----------DKRYYKLRVISNLLQEVSQDLGVLTP  407
NP_010514.3     YTDPHANKNIDKSALNIHADENDEGSVDNTLGSDRSNELEIRGKYVVVPETSQDMAFKCP  565
                 ::   * ::* :.** : : *:           ..  ::*    :: *.***:..  *


TP01_1026       FCIKVYNKVVNFLN----------------------------------------------  421
NP_010514.3     ICKETVTGVYDEESGEWVWKNTIEVNGKYFHSTCYHETSQNSSKSNSGKVGLDDLKKLVT  625
                :* :. . * :  .


TP01_1026       -      421
NP_010514.3     K      626
```

(B)

```
TP01_1034          MKICFRVGKNNSLAVLRKLRIYGSRYKIEKIQKLCENFEKACWRSVRVAEESAYMRKYHF60
BBBOND_0100600     MWN-RARVKTSSLQALAQLSRLSRRYKIEKVQRLSANFERACWRSVRTAEERAYHQKYAF59
                   *       *..** .* :*   . ******:*:*. ***:*******.*** ** :** *

TP01_1034          LKHCMNVINSYRQNKLYTPSAYD----SDPTRFKLS------QYILNKYD--LDNAKACL108
BBBOND_0100600     LRHCVDTINFRKADDRSRRVPLPNTAAALASQFTGSNGPQLQKRVQSHIDELTAGLPSCL119
                   *:**::.**  : :.            :  ::*. *      : : .: *       :**

TP01_1034          DTANKILNLVTNVNDINDDPSSNITKETEENNDLIRQLDVTSLIRGMDNQTLNQDVEKLK168
BBBOND_0100600     SEADASLV---------------TADGSHAA----QIVARCRAHGMTTPECERAVCALS159
                   . *:  *                 * : ..      *: . .  :** .   :: *  *.

TP01_1034          NIITKLEEKVQLLDKYGKD---PHVIGYSKSQLESMIKL-----204
BBBOND_0100600     EIA-RLRTLHARLATYTTHSEGVEDRCYTKRELELVVSGESGCQ202
                   :*  :*.     * .* ..    .   *:* :** ::.
```

(C)

```
TP02_0682          MICKAISLNFFRGFSTLTTPHVTRRRPVWKIKQTYYHRLWNLLTSKKWDEFHQTLQLMRE   60
BEWA_024710        --------MIFRSFTT------VAHKPVWRLKQTFYHRLWHALTGKKWDEFDNLLRRMRD   46
                           :**.*:*      . ::***::***:*****. **.******.: *: **:

TP02_0682          KGLNNDEVTYTLKAHYFILNPYVAVE----------------------NNLINSYFELEE   98
BEWA_024710        QGLNHDEVTYTLKAHYYILNPRTPVENTYLVIEEMKKALMHPSIIRMNENIINSYFELED   106
                   :***.**********:**** . **                        :*:********:

TP02_0682          IGCEPPKLLWQNFTKMIFQTSIRLNRVRRHNLKRQLLLKDPEDVMKLTDKVRL-SLLQIL   157
BEWA_024710        ISCEPPKSQWQNFTKLIWQTALKLNRQRRHDLKQQLLLKDPNDVMKITQNDVQIMALDEF   166
                   *.*****  ******:*:**:::*** ***:**:*******:****:*::      *: :

TP02_0682          QR-L-------------------------------------------------------   160
BEWA_024710        NQAMITPALSVDEIYDEPISVNAEKYRELPVKKLDIQSQHNLEECNYPEIGDRRDLLEA   225
                   :: :
```

Figure 4.10. A representation of a pair-wise sequence alignment analysis output showing the comparison of the sequences of the *T. parva* hypothetical proteins against their homolog from related species.

(A) Protein sequence alignment of the *T. parva* TP01_1026 with the homolog Pcf11p sequence of *Saccharomyces cereviae*. **(B)** Protein sequence alignment of the *T. parva* TP01_1034 with the homolog BBBOND_0100600 sequence from *Babesia bigemina*. **(C)** Finally, protein sequence alignment of *T. parva* TP02_0682 with the *T. equi* homolog BEWA_024710 sequence.

Appendix A.4. The list of all predicted domains and their respective *T. parva* HPs (n = 244).

Sequence. No	**HP ID**(protein length)	**Pfam**	**InterProScan**	**NCBI-CDD**	**PROSITE**	**SMART**
1.	TP03_0648(98)	Mitotic-spindle organizing gamma-tubulin ring associated (MOZART1; 31-73)	Mitotic-spindle organizing protein 1 (31-73)	Mitotic-spindle organizing gamma-tubulin ring associated (MOZART1 (31-73)	None	None
2.	TP01_1034 (204)	KIX_2(11-57); HEPN domain (74-120); BLOC1_2 (120-186); YlbD_coat (126-180)	Coil (157-184)	None	None	coiled coil (155-184)
3.	TP01_0993 (1124)	None	Trans-membrane (41-59, 66 – 83, 89-107, 114 – 133, 172–193); Cytoplasmic_domain (60-65, 108-113, 161-171); Non_cytoplasmic_domain (1-40, 84–88, 134–138, 194–204,)	Metallothio_6(561-627); OpgE (36-195); DUF4383 (41-129)	None	Trans-membrane region (7-24, 39-61, 66-85, 89-107, 114-132, 142-164,)
4.	TP01_0985 (139)	None	Nucleic acid-binding, OB-fold (19-117)	None	None	None
5.	TP01_0953 (264)	DUF1479 (251-279); PhyH (372-444)	None	None	None	None
6.	TP01_0891 (150)	Med (77-132)	Mediator complex, subunit Med7 (7-132)	Med7 (8-143)	None	None
7.	TP01_0890 (187)	None	Zinc finger, CCCH-type (61-93)	None	ZF_C3H1 (65-92)	None
8.	TP01_0736 (589)	Exonuc_VII_L (204-330)	Coil (201-382, 512–532, 547-567); Non_cytoplasmic_domain (20-589)	ZapB (224-280); Bap31 (208-265); Smc (203-411); SMC_prok_A (209-351); APG6 (203-283); MAEBL (190-588)	None	Coiled coil (200-389)
9.	TP01_0699 (76)	HSV U79 / HCMV P34 (28-59)	None	None	None	None
10.	TP01_0679 (543)	None	Trans-membrane (517-538); Non_cytoplasmic_domain (22-516); Cytoplasmic_domain (539-	None	None	Trans-membrane region (519-541)

			543)			
11.	TP01_0671 (455)	CAMSAP_CKK (25-86)	None	None	None	None
12.	TP01_0669 (427)	None	Acyl-CoA N-acyltransferase (40-119; 241 – 314; 378 – 425)	None	None	None
13.	TP01_0636 (151)	RRM_1 (106-144); DUF1764 (11-99)	None	DUF1764 (6-99); RRM1_RRT5 (85-147)	None	None
14.	TP01_0564 (162)	Peptidase family M54 (18-92)	EF-hand domain pair (11-160)	None	None	None
15.	TP01_0555 (146)	DUF4837 (31-83)	None	None	None	None
16.	TP01_0537 (183)	Glycos_transf_4 (115-161)	None	None	None	None
17.	TP01_0391 (680)	Med26 (226-288)	Transcription factor IIS, N-terminal (191 - 304)	Med26 (, 226-288); TFIIS_I (209-288); FSII (218-318)	TFIIS_N(202-292)	TFS2N (224-288)
18.	TP01_0306 (867)	TRAPPC9-Trs120 (743-835)	None	None	None	None
19.	TP01_0144 (1218)	None	Non_cytoplasmic_domain (28–1218)	None	None	Trans-membrane region(24-46)
20.	TP01_0138 (131)	Ribosomal_L32p (89-122); DUF3615 (26-68)	None	Ribosomal_L32p (89-112)	None	None
21.	TP01_0061 (417)	RAP (331-368)	RAP domain (313-371)	RAP(331-371)	RAP (313–371)	None
22.	TP01_0027 (127)	Peptidase_A25 (43-107); Cytokin_check_N (54-80)	None	None	None	None
23.	TP02_0907 (766)	Pec_lyase (679-739)	Non_cytoplasmic_domain (20-766)	None	None	None
24.	TP02_0880 (187)	None	Trans-membrane (104-125); Cytoplasmic_domain (1-103); Non_cytoplasmic_domain (126-187)	None	None	Trans-membrane region (103-125)
25.	TP02_0863 (222)	None	Coil (137 - 157)	None	None	None
26.	TP02_0860 (246)	HPS3_Mid (Hermansky-Pudlak syndrome 3, middle region, 39-142); GPI biosynthesis protein family Pig-F (51-234);DUF261 (87-131)	Trans-membrane (9-28, 48-70, 96-113, 125– 145, 184-207, 219-241); Cytoplasmic_domain (1-8, 71-95, 146-183, 242-246); Non_cytoplasmic_domain (29-47, 114-124, 208-218)	None	None	Trans-membrane region (7-29, 44-66, 96-113, 123-145, 186-208, 218-240)
27.	TP02_0859 (556)	BLOC1S3 (247-360); V-type ATPase 116kDa subunit family (431-514)	Coil (437-471)	None	None	Coiled coil (435-475)
28.	TP02_0812 (576)	None	EF-hand domain pair (476 - 563)	None	EF_HAND_2 (477-503, 536-	None

					571)	
29.	TP02_0776 (2126)	ORC5_C (1851-2050);Salp15 (1910-1984)	Coil (1031-1051, 1261-1281); Trans-membrane (2081-2099); Cytoplasmic_domain (2100 – 2126); Non_cytoplasmic_domain (1-2080)	termin_org_DnaJ (1635-1899)	None	None
30.	TP02_0773 (172)	None	Cytoplasmic_domain(103-172); Non_cytoplasmic_domain (1-84); Trans-membrane (85-102)	None	None	None
31.	TP02_0754 (85)	DUF261 (4-61)	Non_cytoplasmic_domain (28-85)	None	None	None
32.	TP02_0687 (323)	None	Cytoplasmic_domain (233-323); Non_cytoplasmic_domain (33-208); Trans-membrane (209-232)	None	None	None
33.	TP02_0682 (160)	PPR_2 (42-71)	None	None	None	None
34.	TP02_0651 (228)	Ribosomal_L27 (64-145)	Ribosomal protein L27 (70-226)	Ribosomal_L27(64-145)	None	None
35.	TP02_0586 (251)	RNA pseudouridylate synthase (19-116)	Coil (153-173)	None	None	None
36.	TP02_0575 (179)	None	Trans-membrane (127-145); Cytoplasmic_domain(1-126); Non_cytoplasmic_domain (146-179)	None	None	None
37.	TP02_0534 (298)	AP2 (202-242)	None	None	None	None
38.	TP02_0528 (552)	None	Trans-membrane (25-45, 66-85, 91-115, 458 -479, 530-548); Cytoplasmic_domain (46-65, 382-387); Non_cytoplasmic_domain (1-24, 184 – 194)	None	None	Trans-membrane region (25-47, 62-84, 89-111, 395-417, 523-545)
39.	TP02_0512 (1336)	Guanylate-binding protein (121-222, 275-385); MMR_HSR1 (142-162)	Guanylate-binding protein, N-terminal (121 -222); P-loop containing nucleoside	Guanylate-binding protein (137-365)	None	Trans-membrane region (7-29, 1195-1215, 1222-1241, 1289-1308)

			triphosphate hydrolase (104 -163, 197-228, 276-430)			
40.	TP02_0459 (104)	Trafficking protein Mon1 (1-75)	None	None	None	None
41.	TP02_0427 (175)	None	Trans-membrane (144-161); Cytoplasmic_domain (1-143); Non_cytoplasmic_domain (162-175)	None	None	Trans-membrane region (142-161)
42.	TP02_0424 (580)	Gly-zipper_OmpA(388-425)	None	None	None	None
43.	TP02_0420 (1743)	None	Coil(289-309, 1656-1676)		None	None
44.	TP02_0419 (342)	Thiamine pyrophosphokinase C terminal (39-75)	Trans-membrane (40-63, 86-105, 112-132, 138-163, 184-203, 255 – 278)	None	None	Trans-membrane region (40-62, 8-106, 110-132, 142-164, 177-199, 254 - 276)
45.	TP02_0363 (889)	PI-PLC-X (483-532)	None	None	None	None
46.	TP02_0357 (201)	Ribosomal RNA-processing protein 7 (132-197); DUF3939 (127-156); DUF5093 (140-178)	Coil (149-169)	RRP7_like (135-201)	None	None
47.	TP02_0302 (681)	None	Cytoplasmic_domain (625-681); Non_cytoplasmic_domain (1-606); Trans-membrane (607-624)	None	None	None
48.	TP02_0280 (144)	ATPase_2 (33-69); AAA_14 (37-115)	None	None	None	None
49.	TP02_0273 (355)	DUF2321 (85-208)	Coil (171-195)	None	None	Coiled coil (168-199)
50.	TP02_0267 (413)	DUF2207 (159-293)	Trans-membrane(210-228, 248-266, 272-291, 300-323, 367-388); Cytoplasmic_domain(229-247, 292-299, 389-413); Non_cytoplasmic_domain(1-209, 267-271, 324-366)	DUF2981 (53-118)	WD_REPEATS_1 (356-370)	Trans-membrane region (210-232, 245-262, 272-291, 304-326, 366-388)
51.	TP02_0213 (364)	Atg31 (267-325); FAINT DUF529 (27-98, 148-213)	Protein of unknown function DUF529 (,148 -213); Coil (254 - 278); Non_cytoplasmic_domain (23-364)	FAINT DUF529 (147-214)	None	Coiled coil (252-281)
52.	TP02_0150 (385)	Cleavage and polyadenylation factor 2 C-terminal (294-376)	None	Cleavage and polyadenylation factor 2 C-terminal (224-378)	None	None

53.	TP02_0007 (515)	DUF529 (425-475)	Non_cytoplasmic_domain (22-515)	DNA translocase FtsK (72-188); DUF1421 (74-185)	None	None
54.	TP02_0006 (595)	Mononeg_mRNAcap (122-165); DUF529 (410-464, 507-549)	Non_cytoplasmic_domain (21-595)	None	None	None
55.	TP02_0005 (570)	DUF529 (480-541)	Non_cytoplasmic_domain (22-570)	Polyadenylate binding protein (101-188); DNA translocase FtsK (76-172); RCR superfamily (98-180); DUF1421 (73-183)	None	None
56.	TP04_0006 (522)	None	Non_cytoplasmic_domain (21-522)	Neuroendocrine-specific golgi protein P55 (156-312); DNA translocase FtsK (72-185); DUF1421 (109-220)	None	None
57.	TP04_0910 (276)	DUF1443(9-30); UPF0239(13-53); Coiled-coil domain-containing protein 55 (DUF2040, 163-249)	Domain of unknown function DUF2040 (163-249)	DUF2040(144-257)	None	None
58.	TP04_0833 (147)	DNA_Packaging_2(76-118); ING (Inhibitor of growth proteins N-terminal histone-binding ,63-113)	Trans-membrane (125-144); Cytoplasmic_domain (145-147); Non_cytoplasmic_domain (1-124)	None	None	Trans-membrane region (125-144)
59.	TP04_0786 (381)	None	Trans-membrane (178-201); Cytoplasmic_domain (1-177); Non_cytoplasmic_domain (202-381)	None	None	Trans-membrane region (178-200)
60.	TP04_0729 (218)	None	Non_cytoplasmic_domain (18-218)	None	None	None
61.	TP04_0715 (212)	GYF (23-68)	GYF domain (21-78)	GYF domain (24-54)	GYF (21-78)	Coiled coil(162 -196)
62.	TP04_0708 (349)	PepSY-associated TM helix(22-79); DUF2207 (35-144)	Coil(315-335); Trans-membrane(12-35, 62-83, 104-127, 168-189); cytoplasmic_domain(1-11, 84-103, 190-349); non_cytoplasmic_domain(36-61, 128-167)	None	None	Coiled coil (312-340); Trans-membrane region (13-35, 61-83, 104-126, 167-189)
63.	TP04_0505 (74)	Peptidase_M48_N (7-44); P12 (22-45); SieB (25-66); DUF1049 (24-63)	Trans-membrane (24-41); Coil (50 - 70); non_cytoplasmic_domain (1-23); cytoplasmic_domain (42-74)	None	None	Trans-membrane region (24-41)

64.	TP04_0422 (129)	None	Non_cytoplasmic_domain (25-129)	None	None	None
65.	TP04_0414 (768)	Sld5 (142-238)	None	None	None	coiled coil (85-113)
66.	TP04_0353 (477)	LigT_Pease(138-165); tRNA_lig_CPD(290-313); TetR_C_10(403-466)	None	None	None	None
67.	TP04_0327 (507)	Nore1-SARAH (49-65)	None	None	None	None
68.	TP04_0275 (2405)	None	None	CsoS2_M (2070-2239)	None	None
69.	TP04_0259 (137)	NAF (96-129)	None	None	None	None
70.	TP04_0240 (327)	None	None	None	None	None
71.	TP04_0237 (164)	Protein of unknown function (DUF2423, 1-20)	None	None	None	None
72.	TP04_0200 (197)	None	Trans-membrane (21-39, 51-68, 80-98, 104-121, 128-149, 169-189); Cytoplasmic_domain (1-20, 69-79, 122-127, 190-197); Non_cytoplasmic_domain (40-50, 99-103, 150-168)	None	None	Trans-membrane region (21-39, 49-71, 106-124, 128-150, 163-185)
73.	TP04_0192 (110)	FAST_1(FAST kinase-like protein, subdomain 1; 28-55)	None	None	None	None
74.	TP04_0188 (199)	DUF529 (22-92)	None	None	None	None
75.	TP04_0128 (587)	DUF529 (412-481)	Protein of unknown function DUF529 (412 -481)	None	None	None
76.	TP04_0081 (935)	None	Trans-membrane (30-49, 56– 75, 81–101, 113 –134, 154–174, 181–202, 246–265, 295-319); Cytoplasmic_domain (50-55, 102–112, 175– 180, 266-294); Non_cytoplasmic_domain (1-29, 76-80, 135–153, 203–245, 320-935)	None	None	Trans-membrane region (36-58, 78-100, 113-135, 150-172, 177-199, 242-264, 295-317)
77.	TP04_0073 (379)	FANCI_S1-cap (8-35)	None	None	None	None
78.	TP04_0069 (228)	None	Negative elongation factor A (30-149)	None		None
79.	TP03_0581 (307)	RE_HaeIII (122-161)	Non_cytoplasmic_domain (23-307)	None	LECTIN_LEGUME_BETA (296-302)	None

80.	TP03_0557 (447)	Ribonuclease R winged-helix domain (191-217); TFIIE beta subunit core domain(193-253)	None	None	None	None
81.	TP03_0537 (427)	DUF2183 (253-317)	Domain of unknown function DUF2183 (252 - 319)	App1 (148-321); DUF2183 (251-315)	None	Trans-membrane region (12-34)
82.	TP03_0523 (238)	FYVE (52-93); IBR (52-90); C1_1 (54-93); zf-B_box (54-89); RMP (8-121)	Zinc finger, RING/FYVE/PHD-type (44-104); Zinc finger, FYVE/PHD-type (44- 104)	FYVE_RUFY4 (58-89)	None	None
83.	TP03_0522 (115)	None	None	ATP-dependent RNA helicase HrpB; Provisional (53-88)	None	None
84.	TP03_0475 (459)	None	None	None	TONB_DEPENDENT_REC_1 (1-53)	None
85.	TP03_0899 (524)	Med14 (369-443)	None	None	None	None
86.	TP03_0898 (223)	None	Non_cytoplasmic_domain(19-223)	None	None	None
87.	TP03_0437 (90)	Myc_target_1 (35-88)	Non_cytoplasmic_domain (22-90)	None	SERPIN (2-12)	None
88.	TP03_0381 (506)	HECW_N (389-449)	Cytoplasmic_domain(1-111, 174-192, 252-506); Non_cytoplasmic_domain(136-154, 213-231); Trans-membrane(112-135, 155-173)	None	None	None
89.	TP03_0335 (200)	None	Non_cytoplasmic_domain(20-200)	None	DNA_LIGASE_A1 (118- 126)	None
90.	TP03_0329 (317)	Hanta_G2 (35-131)	Coil (121-162)	None	None	None
91.	TP03_0265 (248)	Homocysteine S-methyltransferase (68-161)	None	None	None	None
92.	TP03_0194 (368)	FAINT (121-171); UPF0515 (175-227)	None	Herpes_LMP1 (Herpesvirus latent membrane protein 1, 10-65)	None	None
93.	TP03_0169 (239)	Tropomyosin_1 (173-234); TPR_MLP1_2 (175-215); Spc24 (182-230); Mto2_bdg (197-218); MCC-bdg_PDZ (194-224); DivIC (189-218); bZIP_1 (Basic leucine zipper 1, 189-223); RasGAP_C	Coil (185-219)	DNA_topoisoIV (117-221)	None	BRLZ (Coiled-coil domain-containing protein, 178-229); coiled coil (185-223)

		(187-227); DUF904 (187-223)				
94.	TP03_0148 (1433)	Domain of unknown function, DUF529 (69-141, 326-360, 564-637, 683-747, 1107-1153)	Protein of unknown function DUF529 (69 – 141, 564-637)	Essential cell division protein FtsN (1325-1424); DUF3682 (1340-1431)	None	Trans-membrane region (20-37)
95.	TP03_0138 (351)	FAM76 (165-245); DUF641 (200-246)	Coil (191-211, 213-240)	NT-C2 (11-130); Smc (186-262)	None	Coiled coil (188-245)
96.	TP03_0136 (189)	PDCD9 (38-101)	Trans-membrane (167-188); Cytoplasmic_domain (189-189); Non_cytoplasmic_domain (19-166)	None	None	None
97.	TP03_0125 (177)	Sec3-IP2_bind(150-168)	Cytoplasmic_domain, 1–20, 147- 177, 75-80; Non_cytoplasmic_domain, 42–52, 104-122; Trans-membrane (123-146, 81-103, 53-74, 21 - 41)	None		Trans-membrane region (21-43, 53-75, 82-104, 124-146)
98.	TP03_0123 (481)	None	Non_cytoplasmic_domain (17-481)	None	None	None
99.	TP03_0055 (557)	HAND (494-550); ATP5H, ATP synthase D chain, mitochondrial (25-95)	Coils (305-325)	None	None	Coiled coil (296-323)
100.	TP03_0051 (184)	CENP-H (4-61); Taxilin (7-62); Phage_GP20 (12-42); DivIC (17-54); DUF972 (14-58); DUF4795 (15-62); UPF0242 (17-82)	Coil (20-47)	UPF0242 (7-80)	None	Coiled coil (11-53)
101.	TP03_0045 (477)	None	Trans-membrane (318-342, 348-372, 411-433, 445-466); Cytoplasmic_domain (1-317, 373-410, 467-477); Non_cytoplasmic_domain (343-347, 434- 444)	7tm_7 (307-470)	None	Trans-membrane region (318-340, 350-372, 411-433, 448-470)
102.	TP03_0042 (120)	Phage_min_cap2 (49-114); Z1 (51-108); Arc_PepC_II (58-110); Goodbye (60-103); COQ7 (66-103); Atrophin-1(74-116); PET117 (77-107); Minor_tail_Z (82-115)	Coil (63-118)	None	None	Coiled coil (62-118)
103.	TP03_0034 (201)	DUF3545 (147-181)	None	None	None	None
104.	TP03_0885 (481)	DUF4195 (136-175); DUF529 (390-440)	Non_cytoplasmic_domain (22-481)	DNA translocase FtsK (65-235); PAT1 (65-237)	None	None

105.	TP03_0880 (609)	DUF529 (519-582)	Non_cytoplasmic_domain (22-609)	DNA translocase FtsK (36-154); PAT1 (57-191)	None	None
106.	TP03_0870 (563)	DUF529 (375-429, 478-521)	Non_cytoplasmic_domain (22-563)	PAT1 (122-362); DNA translocase FtsK (120-225)	None	None
107.	TP03_0832 (155)	None	Trans-membrane (20-41); Cytoplasmic_domain (42-155); Non_cytoplasmic_domain (1-19)	None	None	None
108.	TP03_0825 (446)	AP2 (12-44)	None	None	None	None
109.	TP03_0768 (346)	Dor1 (24-273)	None	None	None	None
110.	TP03_0761 (185)	DUF4294(54-87)	Non_cytoplasmic_domain (19 – 185)	None	None	None
111.	TP03_0754 (819)	RSN1_TM (493-556)	Trans-membrane (503 -525, 597-618); Cytoplasmic_domain (526-596); Non_cytoplasmic_domain (1-502, 619-819)	None	None	None
112.	TP03_0742 (273)	Uso1_p115_C (PF04871, 73-142);Lectin_N(57-137); CALCOCO1(Calcium-binding and coiled-coiled domain-containg protein 1, 61-139); AIP3 (Actin-interacting protein (Bud6/Aip3) ,63-160)	Uso1/p115-like vesicle tethering protein, C-terminal (73-142); Paired amphipathic helix (201 - 271)	Tropomyosin_1(70-139); Mplasa_alph_rch(68-142); Smc (65-142); ATG16 (67-137)	None	coiled coil (65-150)
113.	TP03_0725 (290)	Syndecan domain;259-281); SCIMP(262-287)	Trans-membrane (259 – 280, 281 – 290); Non_cytoplasmic_domain (1 - 258)	None		Trans-membrane (258 – 280)
114.	TP03_0657 (1081)	None	None	CirA (323-503)	PROKAR_LIPO PROTEIN (1-20)	None
115.	TP03_0642 (248)	Telomere_Sde2(55-167)	Sde2 N-terminal domain (55-167)	Telomere_Sde2(126-167)	None	None
116.	TP03_0641 (299)	Adenovirus endoprotease(26-81)	Coils (75–95)	None	None	
117.	TP03_0620 (452)	None	Trans-membrane(415-437); Coil(360-380); Cytoplasmic_domain(438-	Chromosome segregation ATPase (309-390)	None	Trans-membrane region (415-437)

			452); Non_cytoplasmic_domain(1-414)			
118.	TP05_0020 (133)	Ribosomal_L23 (33-54)	Trans-membrane (47-65, 86-104, 110-132); Cytoplasmic_domain (66-85); Non_cytoplasmic_domain (14-46, 66-85)	None	None	Trans-membrane region (2-19, 43-65, 85-104, 109-131)
119.	TP01_0004 (468)	HAP1 N-terminal conserved region (38-106); Allexivirus 40kDa protein (52-137); Outer membrane protein (OmpH-like, 56-121); AATF-Che1 (69-122); IncA (71-104); BST2 (72-105); :CLZ (74-105);Afadin- and alpha -actinin-Binding (75-105); ABC_tran_CTD (77-106); DUF529(336-395); DUF1640 (64-104); DUF342 (63-107); DUF4407 (55-116)	Coil (76-103); Trans-membrane (443-467); Non_cytoplasmic_domain (22-442); Cytoplasmic_domain (468-468)	None	None	Coiled coil (70-108); Trans-membrane region (445-467)
120.	TP01_0007 (510)	DUF529 (333-387, 439-489)	Non_cytoplasmic_domain (22-510)	DUF1421 (150-281); DNA translocase FtsK (151-312); beta-CASP ribonuclease, RNase J family (379-454)	None	None
121.	TP01_0008 (444)	None	Non_cytoplasmic_domain (22-444)	DNA translocase FtsK (152-286)	None	None
122.	TP01_0009 (498)	DUF529 (305-370, 409-471); Transcription factor IIA, alpha/beta subunit (51-279)	None	DUF1421 (37-173); DNA translocase FtsK (66-166)	None	None
123.	TP01_0026 (722)	None	Cytoplasmic_domain (210-388); Non_cytoplasmic_domain (1-187, 407- 722); Trans-membrane (188-209, 407-722)	None	None	None
124.	TP01_0034 (625)	Rep_Org_C (105-191); Cas_Csd1 (116-184); HCaRG (351-421)	None	None	None	None
125.	TP01_0090 (424)	COG5 (71-159)	None	None	None	Coiled coil (109-131)
126.	TP01_0143 (829)	Hamartin (220-333); DUF4488 (619-667)	Trans-membrane (7-27); Cytoplasmic_domain (1-6);	Herpes_BLLF1 (163-335); PHA0325	None	Trans-membrane region (7-29)

			Non_cytoplasmic_domain (28-829)	BDLF3; Provisional (197-319)		
127.	TP01_0346 (453)	None	Trans-membrane (382-405); Cytoplasmic_domain (406-453); Non_cytoplasmic_domain (1-381)	None	None	None
128.	TP01_0392 (257)	DUF3810 (15-195); Nop14-like family (53-123)	Coil (168-188); Non_cytoplasmic_domain (32-257)	None	None	Trans-membrane region (12-34)
129.	TP01_0457 (163)	Colicin_immun (13-47)	None	None	None	None
130.	TP01_0493 (234)	None	Non_cytoplasmic_domain (1-19);Trans-membrane (20-42); Cytoplasmic_domain (43-234); Coil (186-209)	None	None	coiled coil (186-229)
131.	TP01_0554 (520)	None	Trans-membrane (12-30); Non_cytoplasmic_domain (1-11); Cytoplasmic_domain (31-520)	None	None	None
132.	TP01_0563 (393)	None	Trans-membrane (6-25, 32-53, 84 –112, 310-331, 352-375); Cytoplasmic_domain (26-31, 113-140, 195– 270, 332-351); Non_cytoplasmic_domain (54-83, 291 – 309, 376-393)	None	None	Trans-membrane region (4-26, 33-55, 85-107, 309-331, 352-374)
133.	TP01_0626 (178)	APG5 (117-161)	None	None	None	None
134.	TP01_0676 (254)	SF-assemblin(20-246); Cluap1(43-129); DUF1465(74-132); CLZ (84-129); T3SchapCesA(94-128)	Coil (92-126)	SF-assemblin (20-248); SMC_prok_B (11-149)	None	coiled coil(92-131)
135.	TP01_0759 (944)	Cadherin-like (45-170)	Trans-membrane (388-410, 417-440, 465-484, 851-872, 892-913); Cytoplasmic_domain (411-416, 485-503, 914-944); Non_cytoplasmic_domain (31-387, 873-891)	None	PROKAR_LIPOPROTEIN (PS51257, 1-30)	Trans-membrane region (388-410, 469-486, 499-521, 541-63, 797-819, 830-852, 859-881, 891-913)

136.	TP01_0817 (409)	TauE (11-148, 203-397)	Trans-membrane (7-32, 52-74, 99-119, 131– 152, 389-408); Cytoplasmic_domain (1-6, 75-98, 383-388); Non_cytoplasmic_domain (33-51, 120-130, 318-359)	TauE (297-397, 11-66); Uncharacterized membrane protein YfcA (285-397)	None	Trans-membrane region (7-29, 44-66, 99-121, 297-319, 360-382)
137.	TP01_0847 (642)	None	Cytoplasmic_domain (1-11); Non_cytoplasmic_domain (33-642); Transm-embrane (12-32)	None	None	None
138.	TP01_0912 (145)	DnaB_2 (43-76)	Non_cytoplasmic_domain (23-145)	None	None	None
139.	TP01_1001 (727)	SepSecS (347-382)	None	None	None	None
140.	TP01_1011 (356)	None	Non_cytoplasmic_domain (18-356)	None	None	None
141.	TP01_1026 (421)	Receptor serine/threonine kinase(116-146)	Coil (356–376)	Valyl-tRNA synthetase-like protein (103-198)	None	None
142.	TP01_1118 (650)	PL48 (175-236)	None	None	None	None
143.	TP01_1146 (117)	Pcc1 (74-115)	Non_cytoplasmic_domain(24-117)	None	None	None
144.	TP01_1179 (562)	None	Trans-membrane (536-557); Cytoplasmic_domain (1-535); Non_cytoplasmic_domain (558-562)	None	ASP_PROTEASE (386-397)	None
145.	TP01_1197 (286)	Toprim_N (180-205); DUF1401 (20-56)	Non_cytoplasmic_domain (23-286)	None	None	None
146.	TP01_1208 (197)	None	Coil (157-177)	None	None	Coiled coil (148-180)
147.	TP02_0004 (579)	Mononeg_mRNAcap (122-165); DUF529 (407-463)	Non_cytoplasmic_domain (21-579)	None	None	None
148.	TP02_0009 (336)	Tash protein PEST motif (257-273, 309-325)	Non_cytoplasmic_domain(22-336)	DNA translocase FtsK (100-199); Topoisomerase II-associated protein PAT1 (65-266)	None	None
149.	TP02_0024 (570)	Leucine-rich repeat domain, L domain-like (31-112, 329-532)	None	None	None	None
150.	TP02_0026 (166)	Imm42 (3-90)	None	None	None	None
151.	TP02_0043 (583)	MBA1-like protein (37-73)	None	None	None	None
152.	TP02_0065 (272)	Pox_polyA_pol (142-171)	Non_cytoplasmic_domain	None	None	Trans-membrane region

			(24-272)			(7-26)
153.	TP02_0133 (601)	FAST_1 (43-96); ArsA_ATPase (270-364); YpjP (41-106)	None	None	None	None
154.	TP02_0147 (167)	DUF1395 (25-142); DUF1474 (44-95)	None	None	None	None
155.	TP02_0308 (287)	None	Coil (191-211)	None	None	None
156.	TP02_0336 (212)	DUF2321 (85-208)	Coil (171-195)	None	None	Coiled coil (168-199)
157.	TP02_0410 (596)	None	Non_cytoplasmic_domain (1-298); Cytoplasmic_domain (324–596); Trans-membrane (299-323)	Pantothenate kinase; Provisional, 211-261)	None	None
158.	TP02_0428 (243)	Sigma_reg_C (79-154)	None	None	None	None
159.	TP02_0526 (547)	DUF2387 (215-260)	None	None	None	None
160.	TP02_0555 (102)	None	Trans-membrane (29-62); Cytoplasmic_domain (63-102); Non_cytoplasmic_domain (1-28)	None	None	Trans-membrane region (29-51)
161.	TP02_0569 (438)	None	None	None	PHOSPHOPANTETHEINE (138-153)	None
162.	TP02_0582 (482)	DUF2666 (120-185)	None	None		Coiled coil (220- 250)
163.	TP02_0591 (1066)	None	Trans-membrane (12-31, 43–61, 73-106, 118- 140) Cytoplasmic_domain (1-11, 62-72, 141-151, 255-273, 321-331, 774-1066); Non_cytoplasmic_domain (32-42, 107-117, 174 – 233, 295– 299, 356-753)	Toxin_16 (549-573)	EGF_1 (550–561)	Trans-membrane region (15-32, 39-61, 84-106, 119-141, 156-173, 234-256, 301-323, 330-352, 754-776, 869-891)
164.	TP02_0592 (1057)	Zip (53-145)	Trans-membrane (6 -24, 31–50, 56–77, 89–107, 127–145, 212–239, 289–310, 322–341, 748–769, 874-895); Cytoplasmic_domain (25-30, 78-88, 146-211, 311-321, 770-873); Non_cytoplasmic_domain (1-5, 51–55, 108 -126)	None	None	Coiled coil (818-845); Trans-membraneregion (5-24, 31-49, 59-81, 88-107, 127-146, 228-250, 289-311, 318-340, 748-770, 874-896)

165.	TP02_0679 (123)	None	Trans-membrane (25-47, 90-111); Cytoplasmic_domain (1-24, 112-123); Non_cytoplasmic_domain (48 – 89)	None	None	Trans-membrane region (25-47, 90-112)
166.	TP02_0695 (145)	DUF3342 (15-127)	Coil (17-37)	None	None	coiled coil (8-39)
167.	TP02_0705 (272)	DUF4330 (7-50)	Trans-membrane (6-27); Cytoplasmic_domain (28-272); Non_cytoplasmic_domain (1-5)	None	None	Trans-membrane region (5-27)
168.	TP02_0711 (126)	PBC domain (86-122); DUF2633 (7-24)	Trans-membrane (14-34); Cytoplasmic_domain (35-126); NOn_cytoplasmic_domain (1-13)	None	None	Trans-membrane region (15-37)
169.	TP02_0752 (476)	helix-turn-helix _16(296-319); DUF4116 (88-120)	None	None	None	None
170.	TP02_0849 (224)	None	Cytoplasmic_domain(75- 93, 161-171, 217-224); Non_cytoplasmic_domain (31-53, 119 -129, 193-197); Trans-membrane(54-74, 94-118, 130-160)	None	None	Trans-membrane region(7-29, 53-75, 132-154, 174-193)
171.	TP02_0871 (163)	DUF2046 (7-74); DUF16 (13-77); DUF1147 (76-104) Atg14 (7-100); Rootletin (8-68); SHE3 (11-134); Fez1 (11-110); Septum formation initiator (19-52); TPR_MLP1_2 (30-75); ZPR1 zinc-finger domain (32-124)	Coiled coil (11-52)	Atg14 (13-110)	None	Coiled coil (6-65)
172.	TP02_0897 (1499)	None	Trans-membrane (234-252, 272-291); Cytoplasmic_domain (1,013-1,020, 292–954); Non_cytoplasmic_domain(1045 – 1499; 981-991)	Metallothionein (818-889)	None	Trans-membrane region(58-80, 15-37, 84-106, 111-130, 145-167, 186-205, 237-259, 272-294, 951-973, 994-1011, 1021-1040, 1061-1080)
173.	TP02_0910 (181), TP02_0914	WHEP-TRS (111-137); DUF4398 (85-147)	None	None	None	Coiled coil (113-140)
174.	TP02_0916	GBP (225-353); SieB (21-63);	P-loop containing nucleoside	GBP(225-492)	Aldoketo_reduct	Coiled coil (658-697)

	(1070)	MMR_HSR1(246-321); Flavi_NS4B(282-355); ATPase (646-720); RNA_lig_T4_1(655-732)	triphosphate hydrolase(204-360, 394-524); Guanylate-binding protein, N-terminal (225-353)		ase_3(706 – 721)	
175.	TP03_0008 (911)	Syndecan (3-35)	Non_cytoplasmic_domain (20-911)	104 kDa microneme/rhoptry antigen (171-290); rypan_PARP (196-264)	None	Trans-membrane region (5-27)
176.	TP03_0024 (89)	NARP1 (6-75); Flu_B_NS1 (23-84)	Coil (53-73)	None	None	None
177.	TP03_0028 (544)	Human Cytomegalovirus UL139 protein (391-423)	Coil(393-416)	None	None	Coiled coil (392-420)
178.	TP03_0038 (376)	Mlh1_C (305-375)	Coil (67-87, 315 -335); Non_cytoplasmic_domain (22-355); Cytoplasmic_domain (376-376); Trans-membrane (356-375)	None	None	Coiled coil (305-336)
179.	TP03_0060 (137)	None	Coil (48 - 68); Non_cytoplasmic_domain (15-137)	None	None	coiled coil (28-56)
180.	TP03_0094 (176)	Filamin (17-110)	Immunoglobulin-like fold (16-110)	Filamin (17-110)	None	None
181.	TP03_0095 (229)	Mnd1 (27-141); Osmo_CC (76-111); TelA (78-170); SKA2 (79-118); Spectrin (81-131); DivIC (83-128); DUF4407 (39-187); DUF342 (71-149); DUF1877 (74-132); DUF5082 (77-126)	Coil (79-113)	None	None	Coiled coil (74-117)
182.	TP03_0098 (158)	NAD_Gly3P_dh_N (38-86)	Trans-membrane (130-150); Cytoplasmic_domain (1-129); Non_cytoplasmic_domain (151-158)	None	None	Trans-membrane region (130-152)
183.	TP03_0119 (198)	DUF1098(47-95); Mononeg_mRNAcap(121-188)	None	None	None	None
184.	TP03_0193 (346)	DUF529 (94-152)	DUF529 (96- 150)	None	None	None
185.	TP03_0234 (252)	BAR_3 (59-104)	Non_cytoplasmic_domain (47-252)	None	None	None
186.	TP03_0256 (114)	MTCP1 (30-65)	None	None	None	None
187.	TP03_0268 (1154)	s48_45 (252-297; 1030-1140)	6-Cysteine domain (1028 - 1154)	Sexual stage antigen (1019-1088)	6_CYS (1028-1154)	s48_45 (1029-1093)

188.	TP03_0271(275)	tRNA_anti-codon (94-131)	Nucleic acid-binding, OB-fold (61-144)	RPA2_DBD_D(94-143); RFA2 (49-181)	None	None
189.	TP03_0305 (820)	CHY zinc finger (738-789)	None	Rubredoxin_like (741-790)	None	None
190.	TP03_0336 (1003)	Uso1_p115_head (843-928)	Coil (235–255)	None	None	Trans-membrane region (7-25)
191.	TP03_0378 (121)	SPATA3(43-116)	Cytoplasmic_domain(1-101); Non_cytoplasmic_domain (121-121); Trans-membrane (102-120)	None	None	Trans-membrane region(102-120)
192.	TP03_0388 (195)	DUF4602 (83-180)	None	DUF4602(85-195)	None	None
193.	TP03_0389 (321)	Rop (179-208)	None	None	None	None
194.	TP03_0463 (524)	None	Trans-membrane (6-25); Cytoplasmic_domain (26-524); Non_cytoplasmic_domain (1-5)	KAR9 (362-465)	None	None
195.	TP03_0483 (344)	None	Trans-membrane (76-94); Coil (283–313); Cytoplasmic_domain (1-75); Non_cytoplasmic_domain (95-344)	None	None	Trans-membrane region (76-95); Coiled coil (280-324)
196.	TP03_0525 (243)	Sdh5 (138-190)	Flavinator of succinate dehydrogenase (138 -208)	None	None	None
197.	TP03_0564 (187)	None	Trans-membrane (32-50, 62-84, 104-125, 132-149, 169-186); Non_cytoplasmic_domain (1-31, 85-103, 150-168); Cytoplasmic_domain (126-131, 51-61)	None		Trans-membrane region (104-126, 130-149, 169-189)
198.	TP03_0597 (1509)	HALZ (901-932)	Coil (774-822, 845-938, 946– 980, 1016-1050, 1072-1099, 1344-1364, 1408-1428)	DUF342 (780-863); Smc (819-1177); chromosome segregation protein;, 756-1242); SMC_prok_A (793-1091); SCP-1 (764-1414)	None	Coiled coil (773- 938, 1008-1101, 1341-1434)
199.	TP03_0606 (163)	Putative stress-responsive nuclear envelope protein (137-157)	Non_cytoplasmic_domain(17-163)	None	None	None
200.	TP03_0647 (670)	CENP-F_leu_zip (235-340); CC2-LZ (234-314); DUF904 (385-418)	Coils (220 – 310, 322 – 349, 371 - 405)	CENP-_leu_zip (235-340); UBAN(polyubiquitin binding domain of NEMO(233-314); SCP-1 (	None	Spc7 (kinetochore protein, 249-380)

				Synaptonemal complex protein 1, 173-413)		
201.	TP03_0658 (165)	zf-RanBP (134-160)	Zinc finger, RanBP2-type (133 -162)	None	ZF_RANBP2_2 (133-162);ZF_RANBP2_1 (137-156)	ZnF_RBZ (135-159)
202.	TP03_0678 (210)	SurA_N_3 (1-108); DUF1013 (77-125)	Trans-membrane (12-39); Coil (95-132); Cytoplasmic_domain (1-11); Non_cytoplasmic_domain (40-210)	RIB43A (81-146)	None	Trans-membrane region (13-35); coiled coil (82-132)
203.	TP03_0680 (181)	Uncharacterised protein family (28 -109)	Non_cytoplasmic_domain(19 -181)	None	None	None
204.	TP03_0681 (373)	Bmp (158-246)	None	None	None	None
205.	TP03_0729 (379)	CBFB_NFYA (21-45)	None	None	None	None
206.	TP03_0738 (666)	Domain of unknown function, DUF529, 137-213; 254-330); Hydrolase_4 (Serine aminopeptidase, S33, 483-661)	Protein of unknown function DUF529 (137 – 211; 254 - 328); Serine aminopeptidase, S33 (492-651); Alpha/Beta hydrolase fold (355 – 430; 480 – 650)	None	None	None
207.	TP03_0779 (300)	BNIP2 (63-155)	None	None	None	None
208.	TP03_0819 (112)	DUF1604 (56-95); Sin-like protein conserved region(40-91); DUF241 Arabidopsis protein of unknown function (12-64)	DUF1604 (56- 95)	DUF1604 (56-98)	None	None
209.	TP03_0827 (885)	None	Trans-membrane (12-33); Coil (86-113, 404-424); Non_cytoplasmic_domain (34-885); Cytoplasmic_domain (1-11)	None	None	Trans-membrane region (13-32); coiled coil (84-116)
210.	TP03_0850 (397)	None	Non_cytoplasmic_domain (26-397)	None	None	Trans-membrane region (5-24)
211.	TP03_0856 (527)	FHA domain (409-453)	Non_cytoplasmic_domain (30-527)	None	None	None
212.	TP03_0873 (510)	DUF529 (429-477)	Non_cytoplasmic_domain (22-510)	NESP55 (246-303); DNA translocase FtsK(88-138)	None	None
213.	TP03_0875 (509)	Presenilin (312-452); DUF529 (419-482)	Non_cytoplasmic_domain (22-509)	DNA translocase FtsK(36-154); PAT1 (57-191)	None	None

214.	TP03_0877 (341)	Synapsin (162-232); DUF529 (158-213, 260 -301)	Coil (162-182)	None	None	None
215.	TP03_0881 (568)	DUF529 (380-438, 485-537)	Protein of unknown function DUF529 (485 -537)	DNA translocase FtsK (85-228)	None	None
216.	TP03_0882 (607)	Translocated intimin receptor (Tir) C-terminus (344-393); Domain of unknown function, DUF529 (520-577)	Non_cytoplasmic_domain (22-607)	Chitin synthesis regulation, resistance to Congo red; RCR (254-349); DNA translocase FtsK (166-363); DUF1421 (146-343)	None	None
217.	TP03_0883 (537)	DUF529 (458-505)	Non_cytoplasmic_domain (22-537)	DNA translocase FtsK (114-294)	None	None
218.	TP03_0893 (568)	DUF529 (377-438, 483-531)	DUF529 (483-531)	DUF529 (465-529); DNA translocase FtsK (93-304)	None	None
219.	TP03_0896 (524)	None	Trans-membrane (6-25); Cytoplasmic_domain (26-524); Non_cytoplasmic_domain (1-5)	None	None	Trans-membrane region (5-27)
220.	TP03_0901 (522)	YwpF (385-453)	Trans-membrane (6-25); Cytoplasmic_domain (26-522)	None	None	Trans-membrane region (5-27)
221.	TP04_0003 (561)	None	Non_cytoplasmic_domain (21-561)	DNA translocase FtsK (101-356); Neuroendocrine-specific golgi protein P55 (NESP55)(232-363)	None	None
222.	TP04_0068 (233)	None	Trans-membrane (196-216); Cytoplasmic_domain (217-233); Non_cytoplasmic_domain (22-195)	None	None	Trans-membraneregion (192-214)
223.	TP04_0171 (306)	None	Coil (261-281); Trans-membrane (68-95); Cytoplasmic_domain (96-306); Coil (261-281)	None	TONB_DEPENDENT_REC_1 (1-60)	Trans-membrane region (71-93); coiled coil (252-273)
224.	TP04_0172 (912)	DUF529 (46-121, 244-306, 356-430); DUF900 (636-683); Hydrolase_4 (Serine aminopeptidase, S33, 573-826);	Alpha/Beta hydrolase fold (28-121 , 573-876) ; Serine aminopeptidase, S33 (579 -824);	Hydrolase; alpha/beta fold family protein (586-723); 104 kDa microneme/rhoptry antigen (238-412)	LIPASE_SER (657- 666)	Trans-membrane region (874-896)

		alpha/beta hydrolase fold (575-675)	Protein of unknown function DUF529 (244 -306, 356-430); Trans-membrane (874-897)			
225.	TP04_0190 (74)	DUF529 (27-56)	None	None	None	None
226.	TP04_0223 (48)	Predicted SPOUT methyltransferase (3-29)	None	None	PTS_HPR_SER (15-30)	None
227.	TP04_0252 (649)	None	Non_cytoplasmic_domain (1-33); Cytoplasmic_domain (55-649); Trans-membrane (34-54)	None	None	None
228.	TP04_0254 (220)	Thioredoxin (69-135)	Thioredoxin-like fold (50-136)	TRX_family (57-131)	None	None
229.	TP04_0283 (206)	Peptidase_A8 (53-96); DUF2244 (57-110)	Trans-membrane (58-78, 84 – 103); Cytoplasmic_domain (1-57, 104-206); Non_cytoplasmic_domain (79-83)	None	None	Trans-membrane region (56-78, 82-104)
230.	TP04_0399 (357)	None	Trans-membrane (327-348); Cytoplasmic_domain (349-357); Non_cytoplasmic_domain (20-326)	Na_trans_assoc (235-285)	None	Trans-membrane region (4-21, 327-349)
231.	TP04_0405 (718)	Calcipressin (82-130)	None	None	None	None
232.	TP04_0455 (556)	None	Cytoplasmic_domain (1-20; 85-95; 145-155, 221-254, 391-401, 496-556); Trans-membrane (21-44, 64 – 84, 96–113, 119 –144, 156–179, 474-495); Non_cytoplasmic_domain (279–367, 420-473)	RnfA (56-173)	None	Trans-membrane region (23-45, 60-82, 89-111, 121-143, 473-495)
233.	TP04_0503 (245)	V-ATPase_G_2 (156-222)	Coil (192-215)	None	None	Coiled coil (187-220)
234.	TP04_0532 (206)	EF-hand_7(125-185)	EF-hand domain pair (42-187)	None	None	
235.	TP04_0576 (447)	Aminotran_1_2 (206-261)	None	None	None	None
236.	TP04_0633 (193)	None	Trans-membrane (60-78; 98-115, 127-149); Cytoplasmic_domain (1-59;	None	None	Trans-membrane region (60-78, 98-120, 127-149)

			116-126); Non_cytoplasmic_domain (79-97; 150-193)			
237.	TP04_0638 (64)	DUF4268 (16-58)	None	None	None	None
238.	TP04_0654 (591)	Baculo_PEP_C (265-358); Major Facilitator Superfamily (361-505)	Coil (309-354); Trans-membrane (24–47, 67-84, 118-139, 370-390, 410-430, 442-461, 514-532)	Putative minor structural protein (282-402); TPR/MLP1/MLP2-like protein (256-368); helix-rich Mycoplasma protein (157-368); Uncharacterized coiled-coil protein, contains DUF342 domain (177-365); chromosome segregation protein (163-371)	None	Coiled coil (307-359); Trans-membrane region (24-46, 67-84, 118-140, 369-391, 411-433, 440-462); Spc7 kinetochore protein (270-375)
239.	TP04_0834 (398)	Acetyltransf_13 (50-91)	Coil (352 - 372)	None	None	None
240.	TP04_0869 (283)	C1_1 (232-264); C1_2 (233-263); Zn_ribbon_17 (234-277); Double Zinc Ribbon (234-254)	Protein kinase C-like, phorbol ester/diacylglycerol-binding domain (220 -277)	None	ZF_DAG_PE_2(220-277)	None
241.	TP04_0896 (489)	Opioid growth factor receptor (OGFr) conserved region (237-317)	Trans-membrane (12-33); Cytoplasmic_domain (1-11); Non_cytoplasmic_domain (34-489)	None	None	Trans-membrane region (9-31)
242.	TP04_0905 (676)	DUF1762 (124-177)	Coil (49-76)	None	None	coiled coil (48-76)
243.	TP05_0039 (126)	None	Trans-membrane (21-41, 61-79); Cytoplasmic_domain (1-20, 80-126); Non_cytoplasmic_domain (42-60)	None	None	Trans-membrane region (29-51, 61-80)
244.	TP03_0713 (268)	DUF1951 (116-141)	None	None	None	None

Appendix B: Ethical Clearance Certificate

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Animal Ethics Committee

PROJECT TITLE	***In silico* characterization and functional prediction of selected *Theileria parva* hypothetical proteins**
PROJECT NUMBER	**V024-16**
RESEARCHER/PRINCIPAL INVESTIGATOR	**BP Mahlobo**

STUDENT NUMBER (where applicable)	**UP-15406360**
DISSERTATION/THESIS SUBMITTED FOR	**MSc**

ANIMAL SPECIES	**n/a**	
NUMBER OF ANIMALS	**n/a**	
Approval period to use animals for research/testing purposes		**April 2016-April 2017**
SUPERVISOR	**Dr. K Sibeko-Matjila**	

KINDLY NOTE:

Should there be a change in the species or number of animal/s required, or the experimental procedure/s - please submit an amendment form to the UP Animal Ethics Committee for approval before commencing with the experiment

APPROVED	Date 18 April 2016
CHAIRMAN: UP Animal Ethics Committee	Signature