Elsevier required licence: © <2019>.
This manuscript version is made available under the CC-BY-NC-ND 4.0 license

http://creativecommons.org/licenses/by-nc-nd/4.0/

The definitive publisher version is available online at

https://www.sciencedirect.com/science/article/pii/S030440171830387X?via%3Dihub

Annotating the ‘hypothetical’ in hypothetical proteins: *in-silico* analysis of uncharacterised proteins for the Apicomplexan parasite, *Neospora caninum*

Larissa Calarco[1*], John Ellis[1]

[1] School of Life Sciences, University of Technology Sydney,
PO Box 123, Broadway,
NSW 2007, Australia
Larissa.M.Calarco@student.uts.edu.au
John.Ellis@uts.edu.au

***Corresponding Author**: Larissa Calarco, University of Technology Sydney, School of Life Sciences, Sydney, Australia, P: +61 2 9514 4161, F: +61 2 9514 8206,
E: Larissa.M.Calarco@student.uts.edu.au

Abstract

Neospora caninum is a parasite of veterinary and economic importance, affecting beef and dairy cattle industries globally. While this species has been recognised as a serious cause of disease in cattle and dogs for over 30 years, treatment and control options are still not available. Furthermore, whilst vaccination was identified as the most economic control strategy, vaccine discovery programs require new leads to investigate as vaccines.

The current lack of gene annotation available for *N. caninum*, especially compared to the closely related model organism, *Toxoplasma gondii*, considerably hinders vaccine related research. Moreover, due to the high degree of similarity between the two organisms, a significant amount of gene annotation available for *N. caninum* stems from sequence homology between the species. However, there is a plethora of literature identifying conserved virulence factors between members of the Apicomplexa, which suggests that key players are contributing to successful parasite invasion, motility, and host cell attachment.

In this study, bioinformatic approaches classified 125 uncharacterised proteins within the *N. caninum* genome, as transmembrane proteins with signal peptide sequences. Functional annotation assigned enriched gene ontologies for cell-adhesion, ATP binding, protein serine/threonine phosphatase complex, immune system process, antigen binding, and proteolysis. Additionally, 32 of these proteins were also identified as adhesins, or having adhesin-like properties, which were further characterised through the discovery of domains and and gene ontology, to reveal their potential functional significance as virulence factors for *N. caninum*. This study identifies a new, small subset of proteins within *N. caninum*, that may be involved in host-cell interaction, parasite adhesion, and invasion, thereby implicating them as potential targets to exploit in the development of control options against the disease.

Key words: hypothetical protein, *in*-silico, annotation, virulence factors, adhesin, transmembrane protein

Introduction

The Apicomplexa represent a phylum of diverse, ubiquitous, and successful parasites that are responsible for a range of medical, economical, and veterinary diseases. The increasing significance and relevance of this group of parasites has sparked a plethora of research elucidating parasite biology, host cell interaction, and diversity between and within species.

Shared amongst Apicomplexans is the presence of specialised secretory organelles that form part of the unique apical complex (Carruthers and Sibley, 1997; Gubbels and Duraisingh, 2012; English et al., 2015). The release of effector molecules from these secretory organelles provides a catalyst for the execution of crucial processes, which promote parasite motility, host cell attachment, and subsequent invasion (Carruthers and Sibley, 1997; Sibley, 2004; English et al., 2015). Invasion begins with attachment to the host cell via the apical complex, resulting in the organised secretion of proteins from rhoptries and adhesive micronemes (Sam-Yellowe, 1996; Carruthers and Sibley, 1997; Carruthers et al., 1999; Sibley, 2004). This is followed by creation of the protective parasitophorous vacuole (PV), where subsequently the parasite is able to grow, replicate, and disrupt host cell signalling and defence mechanisms (Sibley, 2004; Plattner and Soldati-Favre, 2008; Luder et al., 2009; Pelle et al., 2015; Clough and Frickel, 2017).

A protein’s structure determines its function, and membrane proteins are vital to a plethora of cellular processes, including cellular attachment, invasion, molecule transport and signalling, thereby representing a category of biologically significance proteins (Reynolds et al., 2008). Conversely, proteins that are transported to secretory organelles generally contain an N-terminal signal sequence (Chen et al., 2008), where the mechanisms for coordinated parasite egress and invasion, rely on signal transduction (Gubbels and Duraisingh, 2012). Effector molecules that function to facilitate parasite invasion and direct modulation of host cell signalling in apicomplexans are constantly being identified, many of which contain such important structural features.

For example, microneme (MICs), rhoptry (ROPs) and dense granule proteins (GRAs) are classified as excretory/secretory antigens (ESA), representing a group of proteins instrumental in parasite invasion, intracellular survival, and successful replication (Decoster et al., 1988; Cesbron-Delauw and Capron, 1993; Cesbron-Delauw et al., 1996; Hoppe et al., 2000; Nam, 2009; Sheiner et al., 2010). Many of these secreted proteins commonly possess a signal peptide, and/or transmembrane domains, conducive to their function (Ngo et al., 2004; Nam, 2009; Sheiner et al., 2010; Cabrera et al., 2012; Huynh et al., 2014). The MIC family of proteins can also be organised based on their adhesive motifs, which are predicted to mediate parasite motility, invasion, and attachment (Sibley et al., 1998; Tomley and Soldati, 2001). These commonly include epidermal growth factor (EGF), von Willebrand Factor A (vWF), and thrombospondin type 1 (TSP-1) (Lawler and Hynes, 1986; Bork and Rohde, 1991; Tordai et al., 1999; Tomley and Soldati, 2001; Chen et al., 2008).

N. caninum is a cyst forming protozoan parasite of veterinary and economic importance, that affects beef and dairy cattle industries globally (Dubey, 1999, 2003). While neosporosis as a disease has been recognised for over 30 years, the development of treatment and control options is severely lacking, but becoming increasingly vital (Reichel and Ellis, 2002). The current extent of genome annotation for *N. caninum* however, presents a hindrance to the crucial identification of key contributors to pathogenicity. Many proteins are termed ‘hypothetical’ or ‘unnamed’ due to either their unknown function, or lack of sequence homology to recognised proteins (Galperin and Koonin, 2004). Furthermore, recent studies focusing on improving and expanding the available gene structure and annotations for *N.*

caninum are yet to be integrated into popular online databases, such as NCBI or ToxoDB (Gajria et al., 2008) reference resources.

While it is logical to assume that essential protein-coding genes implicated in parasite virulence have been identified, the sheer number of unclassified or hypothetical regions cannot be neglected or deemed unimportant, prompting this study. Current vaccine candidates for parasites within this phylum involve either surface or secreted antigens that appear to be fundamental to parasite invasion, the mechanisms and contributors of which appear to be mostly conserved (Kim and Weiss, 2004; Hemphill, 2015). Consequently, the identification and investigation of uncharacterised proteins through sequence homology, structure, and known hallmarks of parasite virulence, has the power to perpetuate the discovery of targets for vaccine development. This study aimed to exploit bioinformatic techniques to identify previously uncharacterised proteins of biological and functional significance, based on protein sequence topology, structure, and discerning features.

Methods

A range of tools were used for the functional annotation of hypothetical proteins in this study, are detailed in Table 1.

PLACE TABLE 1 HERE

Sequence retrieval

The Entrez GeneIDs of proteins classified as 'hypothetical' or 'unnamed' (collectively referred to as 'uncharacterised' from here on in), were extracted from NCBI (GenBank assembly accession #GCA_000208865.2: https://www.ncbi.nlm.nih.gov/genome/proteins/248?genome_assembly_id=28617), and uploaded to UniProtKB. The sequences for all uncharacterised proteins that also had no gene names or annotations in UniProtKB were then extracted in FASTA format.

Protein topology prediction and annotation

The final list of uncharacterised protein sequences was submitted to the Philius Prediction Server for individual classification by protein type (http://www.yeastrc.org/philius/runPhilius.do) (Reynolds et al., 2008). This software categorises proteins as globular (G), globular with signal peptides (G+SP), transmembrane

(TM), or transmembrane with signal peptides (TM+SP). Sequences were subsequently obtained for proteins classified as TM+SP in FASTA format, and submitted to Blast2GO (version 5), for functional annotation (Conesa et al., 2005). The gene ontology (GO) annotation workflow available in Blast2GO incorporates BLAST analysis, GO, and InterProScan (https://www.ebi.ac.uk/interpro/) (Quevillon et al., 2005; Finn et al., 2017).

Integrating the InterProScan database allowed identification of homologous superfamilies, domains, and repeats present within each query protein sequence. It also incorporates the transmembrane topology predictor Phobius (Kall et al., 2004), and signal peptide predictor SignalP (Petersen et al., 2011). This allowed confirmation of protein sequence classification by Philius to be corroborated by these tools. Proteins that were not identified as TM+SP by at least two of these tools were discarded.

In an attempt to further assign biological function to the remaining unannotated proteins in this list, the protein sequences were uploaded to the SECLAF webserver (https://pitgroup.org/seclaf/), to identify enriched or over represented gene ontologies in this protein callset. This server uses deep neural networks for the hierarchical classification of biological sequences (Szalkai and Grolmusz, 2018b, a).

Identification and annotation of adhesion-like proteins

All TM+SP uncharacterised protein sequences were analysed by MAAP, a malarial adhesins proteins predictor (http://maap.igib.res.in/) (Ansari et al., 2008). This predictor is based on Support Vector Machines, where a default threshold of $P_{\text{maap}} = 0$ was used, characterising any protein sequences above this threshold as adhesin or adhesin-like. The identified adhesin proteins were cross-referenced with their predicted Philius protein classification, resulting in a final list of uncharacterised proteins, identified as adhesin or adhesin-like transmembrane proteins, containing signal peptides. The original Blast2GO results were retrieved for proteins in this callset for further analysis. The bioinformatics workflow is summarised in Figure 1.

PLACE FIGURE 1 HERE

The list of TM+SP proteins were also uploaded to the ExPASy PROSITE database of protein domains, families, and functional sites (de Castro et al., 2006; Sigrist et al., 2013). This involved identifying sequence patterns, sites, and profiles, and also calculating the amino acid composition of each sequence. The protein browser available in ToxoDB, PBrowse, was subsequently used to identify any orthologous sites across each protein, through BLASTP. Lastly, the 32 proteins were searched against the Database of Essential Genes (DEG; http://www.essentialgene.org/) using default BLAST parameters, which consolidates currently available genomic elements considered essential and indispensable for the survival of an organism.

Evidence for expression of proteins using RNA-seq data

To obtain experimental evidence supporting the expression of the proteins in the final callset, the *de novo* transcriptome assembled using RNA-seq data generated from NC-Liverpool tachyzoites as per a previous study (Calarco et al., 2018) was exploited. Each protein sequence was subjected to a BLAST analysis against the NC-Liverpool transcriptome, using the command-line NCBI BLAST tool (version 2.7.1), where the most confident transcriptome contig hits (low e-value and high bit score) for each protein were retained. For any proteins not returning a result, data integrated into ToxoDB from Reid *et al.* (2012), generated from the transcriptomes of days three and four NC-Liverpool tachyzoites, was used to determine mRNA expression levels.

Sequence variation within the final protein callset

Calarco *et al.* (2018) compared RNA-seq data from tachyzoites of the NC-Liverpool and NC-Nowra isolates. By employing a variant analysis pipeline, sequence variants located within functionally significant genes or regions that differed between the two isolates were identified and reported. The 32 adhesin-like transmembrane proteins with signal peptides presented in this study were investigated for SNPs and whether they were located in a genome hotspot.

Results

Topology prediction and annotation for all uncharacterised proteins

There were 4008 "hypothetical" proteins, and 256 unnamed" proteins extracted from NCBI, from a total of 6936 *N. caninum in-silico* predicted proteins. These proteins are listed in Supplementary File S1, detailing their chromosome location, GeneIDs, locus tags, and lengths. Once the GeneIDs were uploaded to UniProt for sequence retrieval, 981 proteins were removed as they were assigned predicted protein descriptions and annotations based on the data available in UniProt (Supplementary File S2). This includes annotations assigned based on sequence similarity, or experimental evidence at the protein and transcript level. The details of the remaining proteins, whose sequences were retrieved from UniProt in FASTA format, are provided in Supplementary File S3. This process is summarised in Supplementary File S4.

Philius identified more than half of the uncharacterised proteins as globular, with no transmembrane domains or signal peptide sequences (Figure 2). There were however 147 proteins predicted to be TM+SP proteins by Philius.

PLACE FIGURE 2 HERE

Of the 147 TM+SP proteins identified by Philius, the topologies of 20 were not corroborated by either Phobius or SignalP following Blast2GO analysis. There were also an additional two proteins with no topology features predicted by Phobius or SignalP (F0V8W3 and F0VLJ1). All of these proteins, which also had relatively low confidence scores in Philius, were discarded from functional annotation. The Blast2GO results for all 147 TM+SP proteins are provided in Supplementary File S5.

The Blast2GO results identified enriched gene ontologies, featured protein families, and domains occurring across the remaining 125 TM+SP proteins under investigation. This workflow also aimed to assign gene descriptions to each protein based on sequence similarity to closely related organisms. Of the 125 TM+SP proteins, 43 of these were assigned sequence descriptions (Supplementary File S6). The remaining proteins however, were still only classified as 'putative transmembrane protein' or 'hypothetical protein'.

The homologous protein superfamilies represented multiple times in this callset of TM+SP proteins, included growth factor receptor cysteine-rich domain superfamily (IPR009030), which includes proteins involved in signal transduction by receptor tyrosine kinases (Ward et al., 1995; Garrett et al., 1998; Cho and Leahy, 2002), major facilitator transporter superfamily (IPR036259), consisting of membrane transport proteins (Pao et al., 1998; Walmsley et al., 1998), and vWF A-like domain superfamily (IPR036465), where such proteins participate in cell adhesion, signal transduction, membrane transport, and immune defence mechanisms (Colombatti et al., 1993).

The main GOs related to molecular function included nucleic acid binding (GO:0003676), DNA binding (GO:0003677), ATP binding (GO:0005524), and serine-type endopeptidase activity (GO:0004252). Conversely, the most represented GOs pertaining to biological function were proteolysis (GO:0006508) and regulation of apoptotic process (GO:0042981). As expected, most of the protein sequences were assigned cellular component GOs for 'integral component of membrane' (GO:0016021) and 'membrane' (GO:0016020).

While the Blast2GO analysis returned only minimal GOs for all 125 proteins, the SECLAF webserver provided a more thorough and extensive list of enriched gene ontology protein function prediction. The most represented GOs that were associated with almost all proteins in this callset, included cell junction (GO:0030054), protein serine/threonine phosphatase complex (GO:0008287), cell tip of elongated cells (GO:0051286), and binding (GO:000584). Other GOs of functional interest associated with many of these TM+SP proteins included signal transduction (GO:0007165), immune system process (GO:0002376), anchoring junction (GO:0070161), adhesion of symbiont to host (GO:0044406), and interaction with symbiont (GO:0051702).

Prediction of adhesin-like proteins and their classification

Of the 3283 uncharacterised proteins investigated, 654 (20%) were identified as having adhesin properties by MAAP (Supplementary File S7). Supplementary File S8 contains a small subset of these proteins that were investigated through InterProScan sequence analysis, to justify the applicability and efficacy of this malarial adhesins predictor for *N. caninum*.

Figure 3 is a pie chart presenting the percentage of adhesin proteins, and their predicted protein classification according to Philius. A total of 32 uncharacterised proteins (~1%) were identified as adhesin-like transmembrane proteins, with signal peptides.

PLACE FIGURE 3 HERE

Annotation of adhesin TM+SP proteins

The Blast2GO analysis assigned gene descriptions for 20 proteins, based on sequence similarity. This included proteins identified as MIC2 (F0VIM1), subtilisin SUB2 (F0VNN6), septin (F0VML7), and *T. gondii* family A protein. There were however 12 proteins that remained described as ‘hypothetical’ or simply ‘putative transmembrane protein’, due to a lack of sequence homology with related species. Additionally, two hypothetical proteins that were not assigned descriptions, had between 34-38% identity with *T. gondii* GRA11 (F0V9X3 and F0V9Z2). The featured domains identified within this final protein callset included vWF type A domain, (IPR002035), CARD or caspase recruitment domain (IPR001315), subtilisin SUB1-like catalytic domain (IPR034204), and peptidase S8 domain (IPR036852).

The only represented GOs from Bast2GO included ‘serine-type endopeptidase activity’ (GO:0004252) and ‘protein binding’ (GO:005515) for molecular function, and ‘proteolysis’ (GO:0006508) and ‘regulation of apoptotic process’ (GO:0042981) for biological process. Again however, the SECLAF webserver assigned further functionally relevant GOs to the 32 proteins, where those over-represented included locomotion (GO:0040011), cell adhesion (GO:0098602), antigen binding (GO:0003823), cofactor transmembrane transporter activity (GO:0051184), and structural molecule activity (GO:0005198).

The 32 adhesin-like transmembrane proteins with signal peptides identified in this study are listed in Table 2, along with their Blast2GO descriptions, gene ontologies, and InterProScan features.

PLACE TABLE 2 HERE

Represented across the 32 adhesin proteins with transmembrane domains and signal peptides, were serine-rich (PS50324) and alanine-rich regions (PS50310), as determined by PROSITE. Further to this, after calculating the amino acid composition for each protein, serine was the most abundant amino acid in 14 of the 32 protein sequences, followed by alanine in 12 protein sequences. In relation to sequence similarity, four proteins contained regions with BLASTP hits to gel-forming secreted mucin-19 from mice. An additional four proteins had sequence similarity to serine-rich adhesin for platelets segments, with BLAST hits to various *Staphyloccocus* and *Streptococcus* species. Other notable hits included adhesin-like cell wall proteins from *Candida albicans* (23% PID), and endochitenase from *Aspergillus fumigatus* (24-32% PID). The amino acid composition and BLASTP results for each of the 32 proteins are presented in Supplementary File S9.

Three proteins in the final callset returned hits to genes in the Database of Essential Genes. Protein F0VIM1 (MIC2) returned BLAST hits to thrombospondin, integrin subunit alpha 1, collagen alpha 1, and ADAM (disintegrin and metalloproteinase) metalloendopeptidase genes in both humans and mice. Genes returned for protein sequence F0VNN6 (SUB2) included membrane-bound transcription factor peptidase and proprotein covertase subtilisin/kexin genes from human and mice, and protein F0VM28 (vWF type A domain containing protein), aligned to huge dynein-related AAA-type ATPase (midasin) from *Saccharomyces cerevisiae*, mediating ATP-dependent remodelling of 60S subunits and subsequent export from nucleoplasm to cytoplasm.

Evidence of protein expression provided by RNA-seq data

All but three of the final 32 proteins (F0VIG7, F0VEH5, and F0VK21) had high confidence BLAST hits to contigs in the NC-Liverpool transcriptome published in Calarco *et al.* (2018), with percentage identities (PID) > 80%. Additionally, some of the proteins had BLAST hits to the same NC-Liverpool transcriptome contig, indicating that they may be paralogous genes within *N. caninum*. Of the three remaining aforementioned proteins, transcript expression was recorded for each of these based on RNA-seq data generated from either day 3 and 4 tachyzoites, published by Reid *et al*. (2012). Supplementary File S10 contains a list of the final 32 proteins, along with the recorded FPKM (Fragments Per Kilobase Million) and percentiles from Reid *et al.* (2012), based on the transcriptomes of days three and four tachyzoites.

Sequence variation within 32 adhesin-like TM+SP proteins

Calarco *et al.* (2018) identified subtilisin SUB2 protease as present in a SNP hotspot, based on the number of sequence variants identified within the gene sequence, when comparing the NC-Liverpool and NC-Nowra isolates. Protein F0VNN6 in this study was annotated as SUB2, and predicted to contain adhesin-like properties, a signal peptide, and transmembrane domains. The only other protein in this final callset which was previously found to contain SNPs, was F0VNG1.

Discussion

Identifying and characterising key players of the parasite invasion process, and elucidating how they could represent treatment, control, and vaccine targets is an important step for any vaccine discovery program. Host-modulating effectors currently of interest include parasite surface antigens, and proteins secreted from the unique Apicomplexan secretory organelles: the rhoptries, micronemes, and dense granules (Carruthers and Sibley, 1997; Sibley, 2004; Gubbels and Duraisingh, 2012). To exploit the current understanding of parasite virulence in the context of *N. caninum*, this study employed various bioinformatic tools to identify a small subset of biologically important proteins, potentially associated with parasite adhesion, invasion, and host cell interactions. The reasoning behind this approach stems from the disturbing lack of genomic annotation available for *N. caninum*, and the sizeable existence of unidentified, theoretically important proteins that await characterisation.

With a focus on non-model organisms lacking sequence annotation for many predicted proteins, there is currently a plethora of research dedicated to the description of hypothetical proteins, through *in-silico* analysis of available sequencing data. This can subsequently result in the identification of functionally significant proteins involved in essential processes, pertinent to the organism under investigation. For example, the *in-silico* analysis of hypothetical proteins in the *Plasmodium falciparum* proteome by Oladele *et al.* (2011), resulted in the classification of several sequences as potential biomarkers of malaria. Another study exploited bioinformatics tools to identify potential new drug targets, from a set of hypothetical proteins classified in a previous immunoproteomics study for *Leishmania* spp. (Chavez-Fumagalli et al., 2017). The *in-silico* workflow elucidated the cellular localisation, biological function, and structure of these proteins, thereby presenting a method for the functional annotation and elucidation of potential drug candidates against Leishmaniasis. In *Trypanosoma cruzi*, all proteins with predicted transmembrane regions were computationally analysed for potential biological function (Silber and Pereira, 2012). A total of 54 proteins were found to be involved in signal-transduction processes through sequence annotation, which again could represent putative drug targets. Lastly, a comprehensive bioinformatics study on the hypothetical protein dataset for *Leishmania donovani*, assigned putative functions, GO terms, and protein domains to a previously uncharacterised set of proteins (Ravooru et al., 2014). The association of these proteins to specific biological pathways and classification as essential genes, demonstrated the advantage of robust computational strategies for the identification of molecules as potential therapeutic targets against such diseases.

The fact that 4264 of 6936 genes in the published *N. caninum* genome are uncharacterised and described as ‘hypothetical’ proteins, presents a major and concerning hindrance to the study of potential virulence factors. Compounding this problem is the lack of consistency and consensus between popular online databases containing genomic data and gene annotation. For example during sequence retrieval, a total of 981 of these proteins had annotation information in UniProt that was not present in NCBI or integrated into ToxoDB. These included functionally important proteins such as apical membrane antigen AMA1 (NCLIV_065490), rhoptry proteins including ROP5-ROP8, rhoptry neck protein RON2 (NCLIV_064620), MIC8 (NCLIV_062770), and multiple GRA proteins including GRA6, GRA7, GRA10, and GRA14.

Based on sequence similarity, Blast2GO assigned descriptions for 43 of the 125 predicted TM+SP proteins (Supplementary Files S5 and S6). Protein F0VQ63 was described as rhomboid-like protease ROM6, belonging to a large family of intramembrane-cleaving serine proteases that are ubiquitous in almost all organisms (Urban and Dickey, 2011). In Apicomplexans, rhomboid proteases are involved in the shedding of adhesins from the cell surface during parasite motility and host-cell invasion, and hence play an important role in host-parasite interactions (Santos et al., 2012; Sibley, 2013). Another TM+SP protein was described as a cytoadherence-linked asexual protein (Clag; F0V7G1), which is thought to be essential for the adhesion and survival of *P. falciparum* in vivo and is regarded as a major determinant of the parasite's virulence (Ocampo et al., 2005). Additionally, protein F0VPV9 was annotated as lectin C-type domain protein, which are integral membrane proteins that have been shown to play a role in the recognition of glycosylated parasite antigens (Vazquez-Mendoza et al., 2013). These proteins have been implicated in processes such as cell adhesion, platelet activation, and pathogen recognition in various pathogenic organisms (Weis et al., 1998; Kilpatrick, 2002; Kerrigan and Brown, 2009).

The processes of parasite invasion are facilitated by organised, sequential protein secretion from specialised apical organelles, to release adhesins for cell attachment and protein transport to the PV membrane (Carruthers and Sibley, 1997; Bradley and Sibley, 2007). Studies have implicated various apicomplexan surface proteins in host cell recognition, where such proteins can be identified by the presence of conserved domains found across a wide range of organisms (Templeton et al., 2004). This class of proteins usually contains adhesion domains, the structural patterns of which can be exploited by an adhesin predictor such as MAAP. An assessment of MAAP indicated it was applicable to the dataset from *N. caninum*, based on the sequence annotation of proteins classified as adhesins in this study (Supplementary File S8). Many of the proteins contained adhesion domains, implicating them in cell adhesion and host cell recognition. Additionally, enriched GOs assigned to the adhesin-like proteins characterising the final callset, such as ‘cell adhesion’, ‘cell-cell adhesion’, and ‘antigen binding’, provided further reassurance and confidence in the use of this tool for *N. caninum* proteins. This was also supported by the BLASTP results, where some of these protein sequences contained ‘serine-rich adhesin for platelets’ segments present in bacterial species, or ‘adhesin-like cell wall protein’ segments found in some fungi. Further investigation of the 654 adhesin-like proteins identified in this study, may reveal further key players involved in crucial parasite adhesion and invasion mechanisms conducive to their success.

While the annotation of many *N. caninum* proteins remains incomplete or insufficient, it is expected that important information can be gained through sequence homology searches with closely related species, especially those part of the Apicomplexa phylum. Arguably, one of the most significant proteins identified and described in the final callset through sequence similarity, was MIC2. Huynh and Carruthers (2006) demonstrated that reduced MIC2 expression led to ineffective host-cell attachment and parasite invasion, as well as reduced gliding motility. This implicated the MIC2 complex as a major determinant of virulence in *Toxoplasma* infection, and identified the potential for MIC2-deficient parasites as an effective live attenuated vaccine against the disease. However, although MIC2 was previously described for *N. caninum* by Lovett *et al.* (2000), it still remains annotated as a hypothetical protein in the NCBI, UniProt, and ToxoDB reference databases.

Another protein described in the final callset was SUB2 (F0VNN6). The success of host cell invasion by Apicomplexan species is contingent on the secretion of proteins from specialised apical organelles (Carruthers et al., 1999). Much of the research concerning these important secretory organelles and their protein contents implicates proteolytic processing as central to the maturation of these crucial proteins (Sam-Yellowe, 1996; Miller et al., 2003). Studies have shown that serine proteinase inhibitors obstruct host cell invasion, implicating subtilisin-like serine proteinases as biologically important in Apicomplexans (Conseil et al., 1999; Blackman, 2000; Miller et al., 2003). The MIC2 and SUB2 proteins identified here also returned BLAST hits to protein-coding genes within the eukaryotic Database of Essential Genes, further cementing their functional significance within the *N. caninum* proteome. The annotation of SUB2 in this study, as well as the previous identification of the SUB2 gene as a SNP hotspot (Calarco et al., 2018), suggests that this protein could represent a potential virulence factor of *N. caninum* that warrants future investigation.

What was still surprising despite the efforts and measures taken in this study, was the lack of descriptions and sequence features present for 12 of the 32 final proteins, such as domains, repeats, and gene ontology. However, the annotation workflow implemented in this study identified functionally significant proteins such as MIC2 and SUB2, within the *N. caninum* proteome. This therefore suggests that the remaining proteins identified represent previously uncharacterised, but biologically important proteins, based on their sequence topology and predicted adhesin-like properties. Such proteins may potentially be involved in adhesion, invasion, and secretion processes that are responsible for the success of such parasite species, despite the lack of sequence features assigned.

Most of the adhesin TM+SP proteins were rich in serine, alanine, and threonine (Supplementary File S9), which likely reflects the training dataset incorporated into MAAP. Furthermore, based on the integrated protein browser in ToxoDB (PBrowse), many of these protein sequences were found to contain mucin-like segments, identified through BLAST analysis (Supplementary File S9). In *Cryptosporidium parvum* for example, there are more than 30 unique mucin-like surface proteins, which are also characterised by serine- and threonine- rich repeats in their extracellular regions, and hence proposed to facilitate adhesion between the parasite and host cell surface (Barnes et al., 1998; Ward and Cevallos, 1998; Cevallos et al., 2000a; Cevallos et al., 2000b; Winter et al., 2000; Templeton et al., 2004). These *C. parvum* mucins, which include gp900 and gp40/gp15, are described as highly immunogenic, and hence potentially important vaccine candidates (Barnes et al., 1998; Cevallos et al., 2000b; O'Connor et al., 2007; O'Connor et al., 2009; Chatterjee et al., 2010). This class of mucin-like proteins is also shared with *T. gondii*, however these proteins are not present in *Plasmodium* and *Theileria* species, and therefore may represent adaptations of coccidians to harsh environments, and immune system evasion mechanisms (Templeton et al., 2004). As the antigens shown thus far to be crucial for the attachment and invasion of *Cryptosporidium* species into host cells, are all mucin-like glycoproteins (O'Connor et al., 2009), the identification of similar proteins in this study hence bolsters their potential significance as part of the *N. caninum* proteome.

Conclusion

This study characterised previously unannotated proteins of the *N. caninum* proteome, using *in-silico* tools. The workflow implemented resulted in the identification of 125 proteins with predicted transmembrane domains and signal peptide sequences. Further analysis of these proteins classified 32 as having adhesin features, which suggests they may be part of crucial parasite mechanisms of cell invasion, adhesion, and motility. Such processes are conserved in the Apicomplexan phylum, the key contributors of which may represent virulence factors to target in the development of therapeutic drugs or vaccines against the disease.

The relevance and value of the bioinformatics tools exploited in this study, was supported by the biologically significant annotations collated. Enriched gene ontologies for the prioritised proteins included proteolysis, cell adhesion, protein serine/threonine phosphatase complex, and integral component of membrane. The *in-silico* approach described is especially useful for non-model organisms or those in the early stages of genomic and proteomic exploration, which may lack sufficient or robust characterisation of functionally significant proteins.

Acknowledgments

This study was completed by LC in partial fulfilment of the PhD degree at UTS.

Competing interests

The authors declare they have no competing interests.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Ansari, F.A., Kumar, N., Bala Subramanyam, M., Gnanamani, M., Ramachandran, S., 2008. MAAP: malarial adhesins and adhesin-like proteins predictor. Proteins 70, 659-666.

Barnes, D.A., Bonnin, A., Huang, J.X., Gousset, L., Wu, J., Gut, J., Doyle, P., Dubremetz, J.F., Ward, H., Petersen, C., 1998. A novel multi-domain mucin-like glycoprotein of Cryptosporidium parvum mediates invasion. Mol Biochem Parasitol 96, 93-110.

Blackman, M.J., 2000. Proteases involved in erythrocyte invasion by the malaria parasite: function and potential as chemotherapeutic targets. Curr Drug Targets 1, 59-83.

Bork, P., Rohde, K., 1991. More von Willebrand factor type A domains? Sequence similarities with malaria thrombospondin-related anonymous protein, dihydropyridine-sensitive calcium channel and inter-alpha-trypsin inhibitor. Biochem J 279 (Pt 3), 908-910.

Bradley, P.J., Sibley, L.D., 2007. Rhoptries: an arsenal of secreted virulence factors. Current opinion in microbiology 10, 582-587.

Cabrera, A., Herrmann, S., Warszta, D., Santos, J.M., John Peter, A.T., Kono, M., Debrouver, S., Jacobs, T., Spielmann, T., Ungermann, C., Soldati-Favre, D., Gilberger, T.W., 2012. Dissection of minimal sequence requirements for rhoptry membrane targeting in the malaria parasite. Traffic 13, 1335-1350.

Calarco, L., Barratt, J., Ellis, J., 2018. Genome Wide Identification of Mutational Hotspots in the Apicomplexan Parasite Neospora caninum and the Implications for Virulence. Genome Biol Evol.

Carruthers, V.B., Giddings, O.K., Sibley, L.D., 1999. Secretion of micronemal proteins is associated with toxoplasma invasion of host cells. Cell Microbiol 1, 225-235.

Carruthers, V.B., Sibley, L.D., 1997. Sequential protein secretion from three distinct organelles of Toxoplasma gondii accompanies invasion of human fibroblasts. Eur J Cell Biol 73, 114-123.

Cesbron-Delauw, M.F., Capron, A., 1993. Excreted/secreted antigens of Toxoplasma gondii--their origin and role in the host-parasite interaction. Res Immunol 144, 41-44.

Cesbron-Delauw, M.F., Lecordier, L., Mercier, C., 1996. Role of secretory dense granule organelles in the pathogenesis of toxoplasmosis. Curr Top Microbiol Immunol 219, 59-65.

Cevallos, A.M., Bhat, N., Verdon, R., Hamer, D.H., Stein, B., Tzipori, S., Pereira, M.E., Keusch, G.T., Ward, H.D., 2000a. Mediation of Cryptosporidium parvum infection in vitro by mucin-like glycoproteins defined by a neutralizing monoclonal antibody. Infection and immunity 68, 5167-5175.

Cevallos, A.M., Zhang, X., Waldor, M.K., Jaison, S., Zhou, X., Tzipori, S., Neutra, M.R., Ward, H.D., 2000b. Molecular cloning and expression of a gene encoding Cryptosporidium parvum glycoproteins gp40 and gp15. Infection and immunity 68, 4108-4116.

Chatterjee, A., Banerjee, S., Steffen, M., O'Connor, R.M., Ward, H.D., Robbins, P.W., Samuelson, J., 2010. Evidence for mucin-like glycoproteins that tether sporozoites of Cryptosporidium parvum to the inner surface of the oocyst wall. Eukaryot Cell 9, 84-96.

Chavez-Fumagalli, M.A., Schneider, M.S., Lage, D.P., Machado-de-Avila, R.A., Coelho, E.A., 2017. An in silico functional annotation and screening of potential drug targets derived from Leishmania spp. hypothetical proteins identified by immunoproteomics. Exp Parasitol 176, 66-74.

Chen, Z., Harb, O.S., Roos, D.S., 2008. In silico identification of specialized secretory-organelle proteins in apicomplexan parasites and in vivo validation in Toxoplasma gondii. PLoS One 3, e3611.

Cho, H.S., Leahy, D.J., 2002. Structure of the extracellular region of HER3 reveals an interdomain tether. Science 297, 1330-1333.

Clough, B., Frickel, E.M., 2017. The Toxoplasma Parasitophorous Vacuole: An Evolving Host-Parasite Frontier. Trends Parasitol 33, 473-488.

Colombatti, A., Bonaldo, P., Doliana, R., 1993. Type A modules: interacting domains found in several non-fibrillar collagens and in other extracellular matrix proteins. Matrix 13, 297-306.

Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M., Robles, M., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674-3676.

Conseil, V., Soete, M., Dubremetz, J.F., 1999. Serine protease inhibitors block invasion of host cells by Toxoplasma gondii. Antimicrob Agents Chemother 43, 1358-1361.

de Castro, E., Sigrist, C.J., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P.S., Gasteiger, E., Bairoch, A., Hulo, N., 2006. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res 34, W362-365.

Decoster, A., Darcy, F., Capron, A., 1988. Recognition of Toxoplasma gondii excreted and secreted antigens by human sera from acquired and congenital toxoplasmosis: identification of markers of acute and chronic infection. Clin Exp Immunol 73, 376-382.

Dubey, J.P., 1999. Neosporosis in cattle: biology and economic impact. Journal of the American Veterinary Medical Association 214, 1160-1163.

Dubey, J.P., 2003. Review of Neospora caninum and neosporosis in animals. Korean J Parasitol 41, 1-16.

English, E.D., Adomako-Ankomah, Y., Boyle, J.P., 2015. Secreted effectors in Toxoplasma gondii and related species: determinants of host range and pathogenesis? Parasite Immunol 37, 127-140.

Finn, R.D., Attwood, T.K., Babbitt, P.C., Bateman, A., Bork, P., Bridge, A.J., Chang, H.Y., Dosztanyi, Z., El-Gebali, S., Fraser, M., Gough, J., Haft, D., Holliday, G.L., Huang, H., Huang, X., Letunic, I., Lopez, R., Lu, S., Marchler-Bauer, A., Mi, H., Mistry, J., Natale, D.A., Necci, M., Nuka, G., Orengo, C.A., Park, Y., Pesseat, S., Piovesan, D., Potter, S.C., Rawlings, N.D., Redaschi, N., Richardson, L., Rivoire, C., Sangrador-Vegas, A., Sigrist, C., Sillitoe, I., Smithers, B., Squizzato, S., Sutton, G., Thanki, N., Thomas, P.D., Tosatto, S.C., Wu, C.H., Xenarios, I., Yeh, L.S., Young, S.Y., Mitchell, A.L., 2017. InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res 45, D190-D199.

Gajria, B., Bahl, A., Brestelli, J., Dommer, J., Fischer, S., Gao, X., Heiges, M., Iodice, J., Kissinger, J.C., Mackey, A.J., Pinney, D.F., Roos, D.S., Stoeckert, C.J., Jr., Wang, H., Brunk, B.P., 2008. ToxoDB: an integrated Toxoplasma gondii database resource. Nucleic Acids Res 36, D553-556.

Galperin, M.Y., Koonin, E.V., 2004. 'Conserved hypothetical' proteins: prioritization of targets for experimental study. Nucleic Acids Res 32, 5452-5463.

Garrett, T.P., McKern, N.M., Lou, M., Frenkel, M.J., Bentley, J.D., Lovrecz, G.O., Elleman, T.C., Cosgrove, L.J., Ward, C.W., 1998. Crystal structure of the first three domains of the type-1 insulin-like growth factor receptor. Nature 394, 395-399.

Gubbels, M.J., Duraisingh, M.T., 2012. Evolution of apicomplexan secretory organelles. Int J Parasitol 42, 1071-1081.

Hemphill, A., 2015. Vaccines and drugs against Neospora caninum, an important apicomplexan causing abortion in cattle and other farm animals. Reports in Parasitology 2015:4, 31-41.

Hoppe, H.C., Ngo, H.M., Yang, M., Joiner, K.A., 2000. Targeting to rhoptry organelles of Toxoplasma gondii involves evolutionarily conserved mechanisms. Nat Cell Biol 2, 449-456.

Huynh, M.H., Boulanger, M.J., Carruthers, V.B., 2014. A conserved apicomplexan microneme protein contributes to Toxoplasma gondii invasion and virulence. Infection and immunity 82, 4358-4368.

Huynh, M.H., Carruthers, V.B., 2006. Toxoplasma MIC2 is a major determinant of invasion and virulence. PLoS Pathog 2, e84.

Kall, L., Krogh, A., Sonnhammer, E.L., 2004. A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338, 1027-1036.

Kerrigan, A.M., Brown, G.D., 2009. C-type lectins and phagocytosis. Immunobiology 214, 562-575.

Kilpatrick, D.C., 2002. Animal lectins: a historical introduction and overview. Biochim Biophys Acta 1572, 187-197.

Kim, K., Weiss, L.M., 2004. Toxoplasma gondii: the model apicomplexan. Int J Parasitol 34, 423-432.

Lawler, J., Hynes, R.O., 1986. The structure of human thrombospondin, an adhesive glycoprotein with multiple calcium-binding sites and homologies with several different proteins. J Cell Biol 103, 1635-1648.

Lovett, J.L., Howe, D.K., Sibley, L.D., 2000. Molecular characterization of a thrombospondin-related anonymous protein homologue in Neospora caninum. Mol Biochem Parasitol 107, 33-43.

Luder, C.G., Stanway, R.R., Chaussepied, M., Langsley, G., Heussler, V.T., 2009. Intracellular survival of apicomplexan parasites and host cell modification. Int J Parasitol 39, 163-173.

Miller, S.A., Thathy, V., Ajioka, J.W., Blackman, M.J., Kim, K., 2003. TgSUB2 is a Toxoplasma gondii rhoptry organelle processing proteinase. Mol Microbiol 49, 883-894.

Nam, H.W., 2009. GRA proteins of Toxoplasma gondii: maintenance of host-parasite interactions across the parasitophorous vacuolar membrane. Korean J Parasitol 47 Suppl, S29-37.

Ngo, H.M., Yang, M., Joiner, K.A., 2004. Are rhoptries in Apicomplexan parasites secretory granules or secretory lysosomal granules? Mol Microbiol 52, 1531-1541.

O'Connor, R.M., Burns, P.B., Ha-Ngoc, T., Scarpato, K., Khan, W., Kang, G., Ward, H., 2009. Polymorphic mucin antigens CpMuc4 and CpMuc5 are integral to Cryptosporidium parvum infection in vitro. Eukaryot Cell 8, 461-469.

O'Connor, R.M., Wanyiri, J.W., Cevallos, A.M., Priest, J.W., Ward, H.D., 2007. Cryptosporidium parvum glycoprotein gp40 localizes to the sporozoite surface by association with gp15. Mol Biochem Parasitol 156, 80-83.

Ocampo, M., Rodriguez, L.E., Curtidor, H., Puentes, A., Vera, R., Valbuena, J.J., Lopez, R., Garcia, J.E., Ramirez, L.E., Torres, E., Cortes, J., Tovar, D., Lopez, Y., Patarroyo, M.A., Patarroyo, M.E., 2005. Identifying Plasmodium falciparum cytoadherence-linked asexual protein 3 (CLAG 3) sequences that specifically bind to C32 cells and erythrocytes. Protein Sci 14, 504-513.

Oladele, T.O., Sadiku, J.S., Bewaji, C.O., 2011. *In silico* characterisation of some hypothetical proteins in the proteome of *Plasmodium falciparum*. Centrepoint J. 17, 129-139.

Pao, S.S., Paulsen, I.T., Saier, M.H., Jr., 1998. Major facilitator superfamily. Microbiol Mol Biol Rev 62, 1-34.

Pelle, K.G., Jiang, R.H., Mantel, P.Y., Xiao, Y.P., Hjelmqvist, D., Gallego-Lopez, G.M., A, O.T.L., Kang, B.H., Allred, D.R., Marti, M., 2015. Shared elements of host-targeting pathways among apicomplexan parasites of differing lifestyles. Cell Microbiol 17, 1618-1639.

Petersen, T.N., Brunak, S., von Heijne, G., Nielsen, H., 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8, 785-786.

Plattner, F., Soldati-Favre, D., 2008. Hijacking of host cellular functions by the Apicomplexa. Annu Rev Microbiol 62, 471-487.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., Lopez, R., 2005. InterProScan: protein domains identifier. Nucleic Acids Res 33, W116-120.

Ravooru, N., Ganji, S., Sathyanarayanan, N., Nagendra, H.G., 2014. Insilico analysis of hypothetical proteins unveils putative metabolic pathways and essential genes in Leishmania donovani. Front Genet 5, 291.

Reichel, M.P., Ellis, J.T., 2002. Control options for Neospora caninum infections in cattle--current state of knowledge. New Zealand veterinary journal 50, 86-92.

Reid, A.J., Vermont, S.J., Cotton, J.A., Harris, D., Hill-Cawthorne, G.A., Konen-Waisman, S., Latham, S.M., Mourier, T., Norton, R., Quail, M.A., Sanders, M., Shanmugam, D., Sohal, A., Wasmuth, J.D., Brunk, B., Grigg, M.E., Howard, J.C., Parkinson, J., Roos, D.S., Trees, A.J., Berriman, M., Pain, A., Wastling, J.M., 2012. Comparative genomics of the apicomplexan parasites Toxoplasma gondii and Neospora caninum: Coccidia differing in host range and transmission strategy. PLoS Pathog 8, e1002567.

Reynolds, S.M., Kall, L., Riffle, M.E., Bilmes, J.A., Noble, W.S., 2008. Transmembrane topology and signal peptide prediction using dynamic bayesian networks. PLoS Comput Biol 4, e1000213.

Sam-Yellowe, T.Y., 1996. Rhoptry organelles of the apicomplexa: Their role in host cell invasion and intracellular survival. Parasitol Today 12, 308-316.

Santos, J.M., Graindorge, A., Soldati-Favre, D., 2012. New insights into parasite rhomboid proteases. Mol Biochem Parasitol 182, 27-36.

Sheiner, L., Santos, J.M., Klages, N., Parussini, F., Jemmely, N., Friedrich, N., Ward, G.E., Soldati-Favre, D., 2010. Toxoplasma gondii transmembrane microneme proteins and their modular design. Mol Microbiol 77, 912-929.

Sibley, L.D., 2004. Intracellular parasite invasion strategies. Science 304, 248-253.

Sibley, L.D., 2013. The roles of intramembrane proteases in protozoan parasites. Biochim Biophys Acta 1828, 2908-2915.

Sibley, L.D., Hakansson, S., Carruthers, V.B., 1998. Gliding motility: an efficient mechanism for cell penetration. Curr Biol 8, R12-14.

Sigrist, C.J., de Castro, E., Cerutti, L., Cuche, B.A., Hulo, N., Bridge, A., Bougueleret, L., Xenarios, I., 2013. New and continuing developments at PROSITE. Nucleic Acids Res 41, D344-347.

Silber, A.M., Pereira, C.A., 2012. Assignment of putative functions to membrane "hypothetical proteins" from the Trypanosoma cruzi genome. J Membr Biol 245, 125-129.

Szalkai, B., Grolmusz, V., 2018a. Near perfect protein multi-label classification with deep neural networks. Methods 132, 50-56.

Szalkai, B., Grolmusz, V., 2018b. SECLAF: a webserver and deep neural network design tool for hierarchical biological sequence classification. Bioinformatics 34, 2487-2489.

Templeton, T.J., Iyer, L.M., Anantharaman, V., Enomoto, S., Abrahante, J.E., Subramanian, G.M., Hoffman, S.L., Abrahamsen, M.S., Aravind, L., 2004. Comparative analysis of apicomplexa and genomic diversity in eukaryotes. Genome Res 14, 1686-1695.

Tomley, F.M., Soldati, D.S., 2001. Mix and match modules: structure and function of microneme proteins in apicomplexan parasites. Trends Parasitol 17, 81-88.

Tordai, H., Banyai, L., Patthy, L., 1999. The PAN module: the N-terminal domains of plasminogen and hepatocyte growth factor are homologous with the apple domains of the prekallikrein family and with a novel domain found in numerous nematode proteins. FEBS Lett 461, 63-67.

Urban, S., Dickey, S.W., 2011. The rhomboid protease family: a decade of progress on function and mechanism. Genome Biol 12, 231.

Vazquez-Mendoza, A., Carrero, J.C., Rodriguez-Sosa, M., 2013. Parasitic infections: a role for C-type lectins receptors. Biomed Res Int 2013, 456352.

Walmsley, A.R., Barrett, M.P., Bringaud, F., Gould, G.W., 1998. Sugar transporters from bacteria, parasites and mammals: structure-activity relationships. Trends Biochem Sci 23, 476-481.

Ward, C.W., Hoyne, P.A., Flegg, R.H., 1995. Insulin and epidermal growth factor receptors contain the cysteine repeat motif found in the tumor necrosis factor receptor. Proteins 22, 141-153.

Ward, H., Cevallos, A.M., 1998. Cryptosporidium: molecular basis of host-parasite interaction. Adv Parasitol 40, 151-185.

Weis, W.I., Taylor, M.E., Drickamer, K., 1998. The C-type lectin superfamily in the immune system. Immunol Rev 163, 19-34.

Winter, G., Gooley, A.A., Williams, K.L., Slade, M.B., 2000. Characterization of a major sporozoite surface glycoprotein of Cryptosporidum parvum. Funct Integr Genomics 1, 207-217.

Zhang, R., Ou, H.Y., Zhang, C.T., 2004. DEG: a database of essential genes. Nucleic Acids Res 32, D271-272.

List of Tables

Table 1. Summary of the tools used in the annotation of uncharacterised *N. caninum* proteins.

Tool/Database	Description	Reference
Philius	Prediction of transmembrane topology and signal peptides	(Reynolds et al., 2008)
Blast2GO 5 PRO	Bioinformatics platform for the functional annotation and analysis of datasets	(Conesa et al., 2005)
InterProScan (v68.0)	Scans sequences against InterPro protein signature databases to identify protein families, domains, and repeats	(Finn et al., 2017)
MAAP	Prediction of adhesins and adhesin-like proteins	(Ansari et al., 2008)
ExPASy PROSITE	Database of protein families, functional sites, and sequence patterns	(de Castro et al., 2006; Sigrist et al., 2013)
ToxoDB PBrowse (v2.48)	Interactive and integrated protein browser	(Gajria et al., 2008)
SECLAF	Webserver that uses deep neural networks for the hierarchical classification of protein sequences	(Szalkai and Grolmusz, 2018b)
Database of Essential Genes (DEG)	Contains records of currently available essential genomic elements among bacteria, archaea, and eukaryotes	(Zhang et al., 2004)

Table 2. Annotations for the final 32 proteins, including their gene descriptions, InterProScan results, and gene ontologies.

Protein	Description	GO	InterPro IDs
F0VIM1	Microneme protein MIC2	F: GO:005515, protein binding	**IPR002035**, von Willebrand factor, type A domain **IPR036465,** von Willebrand factor A-like domain superfamily **IPR036383**, Thrombospondin type-1 (TSP-1) repeat superfamily **IPR000884,** TSP-1 repeat
F0VIG7	Putative transmembrane protein	C: GO:0016021, integral component of membrane	
F0V9X3	Putative transmembrane protein	C: GO:0016021, integral component of membrane	
F0VII0	Hypothetical protein	C: GO:0016021, integral component of membrane	
F0VEH5	Hypothetical protein		
F0VNG1	Transmembrane protein	C: GO:0016021, integral component of membrane	
F0VPX6	Transmembrane protein	C: GO:0016021, integral component of membrane	
F0VFU2	Hypothetical protein	C: GO:0016021, integral component of membrane, GO:0016020, membrane	
F0V9Z2	Putative transmembrane protein	C: GO:0016021, integral component of membrane P: GO:0042981, regulation of apoptotic process	**IPR001315**, CARD (caspase recruitment) domain
F0VM28	Hypothetical protein	-	-
F0VML7	Septin	C: GO:0016021, integral component of membrane, GO:0016020, membrane	CATH Superfamily **G3DSA:3.40.50.300**, P-loop containing nucleotide triphosphate hydrolases

F0VK21	Putative transmembrane protein	C: GO:0016021, integral component of membrane, GO:0016020, membrane	
F0VQ97	Hypothetical protein	C: GO:0016021, integral component of membrane, GO:0016020, membrane	
F0VQ28	Putative transmembrane protein	C: GO:0016021, integral component of membrane, GO:0016020, membrane	
F0VE57	*T. gondii* family A protein	C: GO:0016020, membrane	
F0VE64	*T. gondii* family A protein	C: GO:0016020, membrane	
F0VNN6	Subtilisin SUB2	F: GO:0004252, serine-type endopeptidase activity	**IPR015500**, Peptidase S8, subtilisin-related protein family
		P: GO:0006508, proteolysis	**IPR036852**, Peptidase S8/S53 domain superfamily
		C: GO:0016021, integral component of membrane	**IPR000209**, Peptidase S8/S53 domain **IPR034204**, Subtilisin SUB1-like catalytic domain
F0VRI3	*T. gondii* family A protein	C: GO:0016021, integral component of membrane	
F0VRI6	*T. gondii* family A protein	C: GO:0016021, integral component of membrane	**PS51257**, prokaryotic membrane lipoprotein lipid attachment site profile
F0VRI7	*T. gondii* family A protein	C: GO:0016021, integral component of membrane	
F0VRI8	*T. gondii* family A protein	C: GO:0016021, integral component of membrane, GO:0016020, membrane	**PS51257**, prokaryotic membrane lipoprotein lipid attachment site profile
F0VRI9	*T. gondii* family A protein	C: GO:0016021, integral component of membrane, GO:0016020, membrane	**PS51257**, prokaryotic membrane lipoprotein lipid attachment site profile
F0VRJ2	T. *gondii* family A protein	C: GO:0016021, integral component of membrane, GO:0016020, membrane	**PS51257**, prokaryotic membrane lipoprotein lipid attachment site profile
F0VRJ3	*T. gondii* family A protein	C: GO:0016021, integral component of membrane, GO:0016020, membrane	**PS51257**, prokaryotic membrane lipoprotein lipid attachment site profile

F0VRJ6	*T. gondii* family A protein	C: GO:0016021, integral component of membrane, GO:0016020, membrane	**PS51257**, prokaryotic membrane lipoprotein lipid attachment site profile
F0VRJ8	*T. gondii* family A protein	C: GO:0016020, membrane	**PS51257**, prokaryotic membrane lipoprotein lipid attachment site profile
F0VRK0	*T. gondii* family A protein	C: GO:0016020, membrane	**PS51257**, prokaryotic membrane lipoprotein lipid attachment site profile
F0VRL7	*T. gondii* family A protein	C: GO:0016021, integral component of membrane	
F0VRL8	*T. gondii* family A protein	C: GO:0016021, integral component of membrane	
F0VRL9	*T. gondii* family A protein	C: GO:0016021, integral component of membrane	
F0VRM0	*T. gondii* family A protein	C: GO:0016021, integral component of membrane	
F0VRM4	*T. gondii* family A protein	C: GO:0016021, integral component of membrane	

List of Figures

Figure 1. A summary of the workflow used to annotate uncharacterised proteins for *N. caninum*. After retrieving the sequences for uncharacterised proteins from online reference databases (i.e. those described as "hypothetical" or "unnamed"), Philius was used to identify those with transmembrane domains and signal peptides (TM+SP proteins). After further defining this set of proteins by whether they had adhesin-like properties using the MAAP predictor, various tools were used to annotate the sequences and identify features such as domains, repeats, sequence similarity, and gene ontology.

Figure 2. Classification of all 3283 uncharacterised proteins under investigation by Philius. The Philius protein topology predictor classifies protein sequences as either globular, globular with a signal peptide, containing transmembrane domains, or containing both transmembrane domains and a signal peptide sequence. As presented, over half of the protein sequences submitted to Philius were classified simply as globular proteins, where the 4% identified as transmembrane proteins with signal peptides were selected for further annotation.

Figure 3. Pie of a pie chart presenting the predicted topology for uncharacterised proteins predicted to be adhesins. A total of 654 proteins were identified as having adhesin-like properties, which were further classified by Philius. The 1% in the second pie chart represents the 32 predicted adhesin-like transmembrane proteins with signal peptides, which were subjected to further sequence annotation.

*** Black and white to be used for Figures 1-3 in print.**

Figure 1

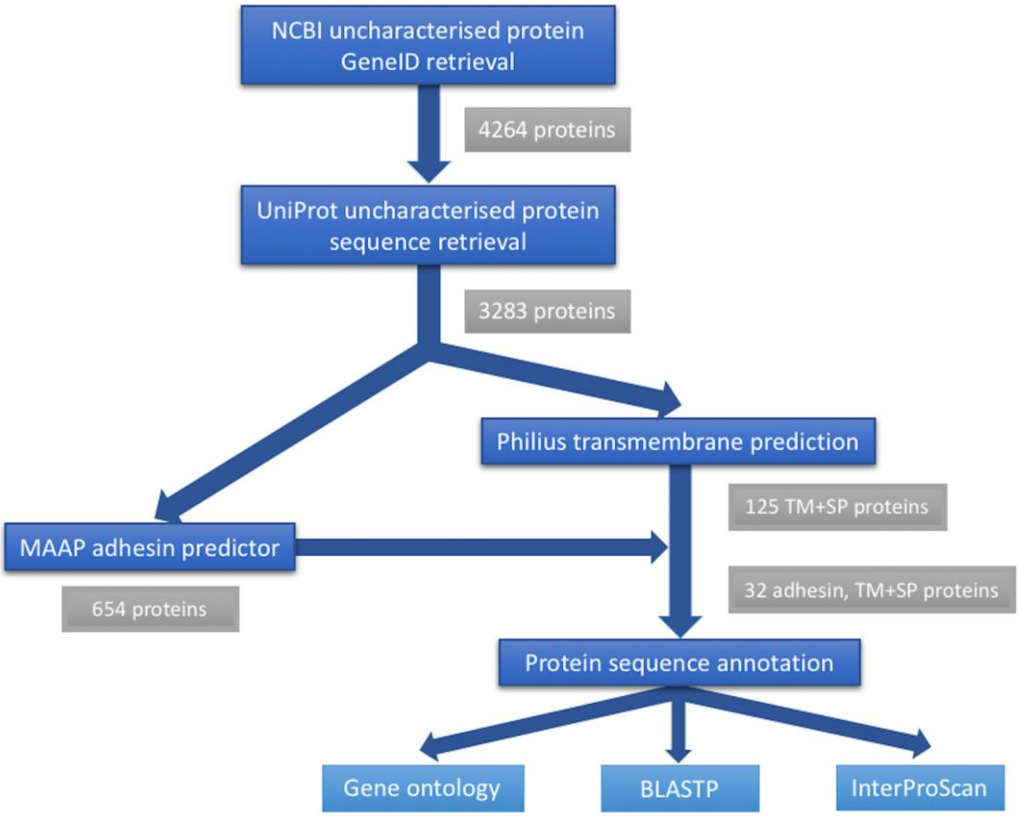

NCBI uncharacterised protein GeneID retrieval
4264 proteins
UniProt uncharacterised protein sequence retrieval
3283 proteins
Philius transmembrane prediction
125 TM+SP proteins
MAAP adhesin predictor
654 proteins
32 adhesin, TM+SP proteins
Protein sequence annotation
Gene ontology
BLASTP
InterProScan

Figure 2

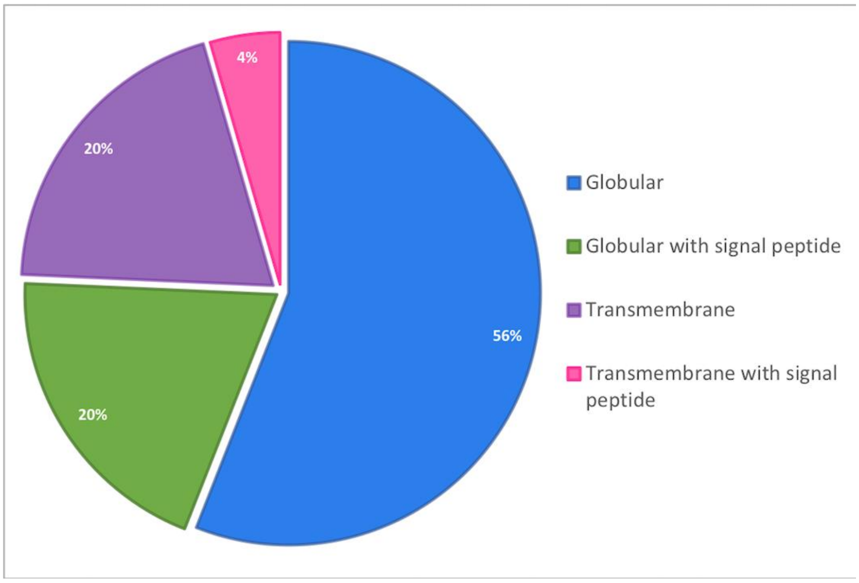
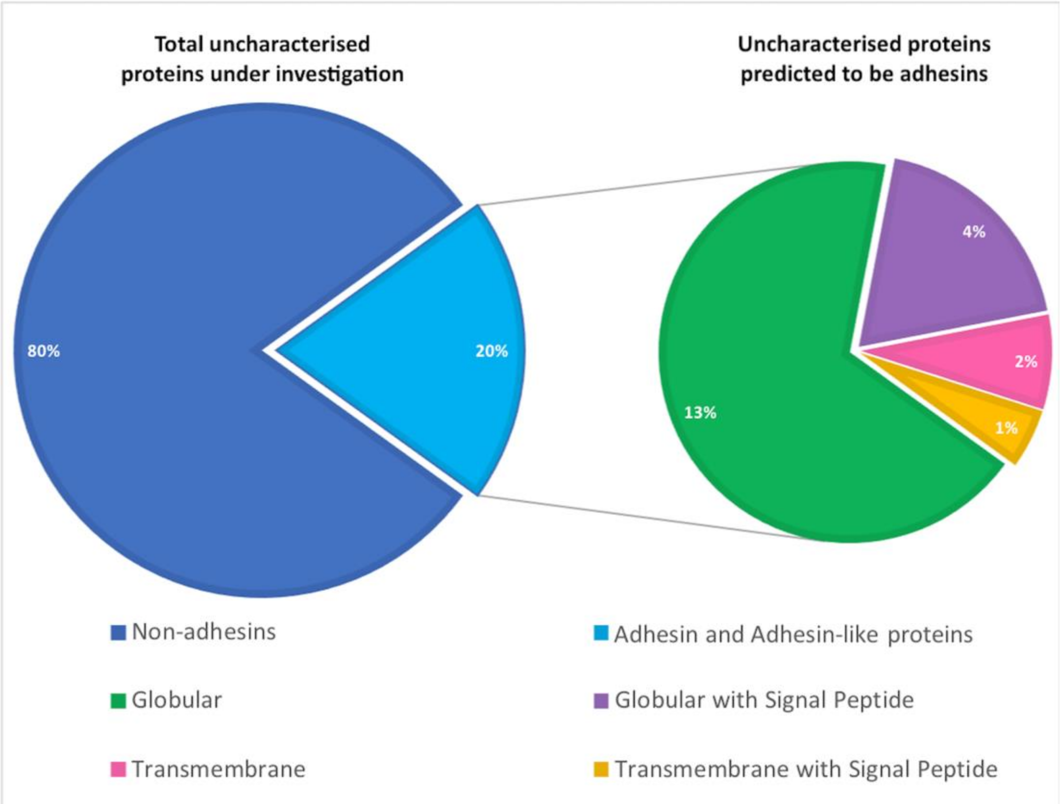


Figure 3


Total uncharacterised proteins under investigation
Uncharacterised proteins predicted to be adhesins
Non-adhesins
Adhesin and Adhesin-like proteins
Globular
Globular with Signal Peptide
Transmembrane
Transmembrane with Signal Peptide