JIDC | THE JOURNAL OF INFECTION IN DEVELOPING COUNTRIES

Original Article

Functional annotation and identification of novel drug targets from uncharacterized proteome of *Trichuris trichiura*

Kanchan Rauthan[1], Saranya Joshi[1 #], Lokesh Kumar[1], Sudhir Kumar[1,2]

[1] Department of Biotechnology, H.N.B. Garhwal University, Srinagar (Garhwal), Uttarakhand, India
[2] Special Centre for Molecular Medicine, Jawaharlal Nehru University, New Delhi, India
[#] Present Address: Department of Biotechnology, School of Applied and Life Sciences, Uttaranchal University, Dehradun, Uttarakhand, India

Abstract

Introduction: *Trichuris trichiura*, a soil-transmitted helminth, resides in the large intestine of humans, causing an asymptomatic disease known as trichuriasis. This global health concern is particularly prevalent in low- or middle-income countries, representing a significant burden on public health as one of the most neglected tropical diseases. The diminishing effects of currently available anthelmintic drugs, attributed to escalating drug resistance, warrants an urgent need for alternative and more potent vaccines or drugs. A substantial portion of the proteins in the *T. trichiura* genome are uncharacterized and their annotation might offer insight into the parasite's invasion, interaction, and survival mechanisms inside the host. Hence, this study is aimed to provide functional annotations for the uncharacterized proteins identified in the proteome of *T. trichiura*.

Methodology: The uncharacterized proteome of *T. trichiura* was subjected to physiological parameter computation, localization analysis, domain identification, homology, and druggability analysis. The programs used were evaluated using the Receiver Operating Characteristic (ROC) analysis.

Results: Functional annotation was assigned to 165 out of the 1726 uncharacterized proteins. Out of these, 85 proteins were found to be non-homologous with the human host and considered to be potential novel drug targets. Two proteins were identified as essential proteins in the DEG database.

Conclusions: Our study identified 165 new proteins from the uncharacterized proteome of the *T. trichiura* and several novel targets that can be further analyzed for drug designing and vaccine-related studies.

Key words: Trichuris; annotation; uncharacterized proteins; drug targets.

J Infect Dev Ctries 2025; 19(6):948-961. doi:10.3855/jidc.19924

(Received 27 January 2024 – Accepted 26 November 2024)

Introduction

Soil-transmitted helminths (STH) are gastrointestinal parasites that infect about 1.5 billion people., It primarily affects the poorest and most deprived populations, with limited access to adequate water, sanitation, and hygiene, especially in Sub-Saharan Africa (SSA), East Asia, China, India, and South America [1]. *Trichuris trichiura* is one of the most prevalent soil-transmitted helminths worldwide, followed by *Ascaris lumbricoides* and hookworms (*Ancylostoma duodenale* and *Necator americanus*). *T. trichiura,* also known as a whipworm, is a gastrointestinal parasitic nematode responsible for causing trichuriasis in humans, particularly in preschool and school-age children. The transmission of this parasite occurs through the ingestion of embryonated eggs present in soil contaminated with feces, and an estimated 604-795 million people are infected worldwide [2]. The Institute for Health Metrics and Evaluation (IHME) estimated that in 2019, trichuriasis contributed to a global burden of disease of 236,000 disability-adjusted life years (DALYs) with 464.6 million people infected with *T. trichiura* [3]. Trichuriasis can appear as a severe ailment with symptoms such as loss of appetite, recurrent abdominal pain, chronic bloody diarrhea, severe anemia, weight loss, rectal prolapse, malnutrition, and severe stunting. It is typically associated with poverty and inadequate environmental sanitation leading to negligence, and thus categorized as a neglected tropical disease.

Anthelminthic drugs like benzimidazole and albendazole, are part of the primary treatment in combination with other drugs and are administered as preventive therapy in Mass Drug Administration (MDA) programs. Despite mass treatment efforts, reinfection with *T. trichiura* can occur due to the partial or short-term protection of these drugs, emphasizing the urgency for more effective drugs with long-term

protection [4].

Despite contributing significantly to the global disease burden and high reinfection rate, *T. trichiura* remains only partially explored. Its genome is 75.2 Mb, housing 9650 genes and an equal number of proteins, with a GC content of 42%. It comprises 4439 contigs and 4156 scaffolds (70602-N50 and 265-L50 scaffolds) [5]. Approximately 18% of the proteome remains uncharacterized which could offer insights into the identification and development of multiple therapeutic targets. Our study employed a subtractive genomic approach to functionally annotate the uncharacterized proteins of *T. trichiura*, leading to the identification of 173 proteins having functional domains. Through sequential characterization and domain analysis, specific proteins have been identified as potential targets for drugs or vaccines. These findings could contribute to a deeper understanding of the parasite's pathogenesis, survival, and invasion mechanisms.

Methodology

Sequence Retrieval

The proteome of *T. trichiura* was downloaded from the UNIPROT database in FASTA format using Proteome ID UP000030665 (data retrieved on 11 January 2021). It contains 9625 proteins of which 1806 proteins are listed as uncharacterized. This set of proteins was analyzed using the CD-HIT program [6] to remove the redundant sequences having an identity of more than 90%. This process yielded 1726 non-redundant sequences of uncharacterized proteins, which were then selected for further analysis.

Identification of Functional Domain

Protein domains are the conserved structural and functional units of proteins that provide information about their role in biological activities. *T. trichiura* uncharacterized proteins were submitted to various web servers such as InterProScan, MOTIF Search, SMART, NCBI CDART, and HMMER for conserved domain and protein family identification.

InterProScan performs the identification and classification of protein sequences into different protein families based on the domains present. Apart from domain analysis, the presence of transmembrane helices and signal peptides in a protein sequence is also predicted, leading to the categorization of the protein as a component of biological, molecular, or cellular processes [7]. MOTIF, HMMER [8], and SMART (Simple Modular Architecture Research Tool) [9] web servers were used for the identification of domain architecture. NCBI-Conserved Domain Architecture Retrieval Tool (CDART) uses the RPS-BLAST algorithm against the NCBI entrez protein database to find protein similarities across significant evolutionary distances using sensitive protein domain profiles [10].

Non-homology analysis against Human proteome

Annotated uncharacterized proteins were subjected to BLASTp with an e value of 10^{-3} against the human proteome (taxid: 9606) database to identify non-homologous proteins. The proteins with no similarity or hit were non-homologous and selected for further analysis.

Sequence Characterization

Physico-chemical properties

Expasy's Protparam tool calculated the physiochemical properties of 173 functionally annotated uncharacterized proteins. Physicochemical properties such as molecular weight, extinction coefficient, isoelectric point, grand average of hydropathicity, etc. are predicted by this server using the amino acid sequence of the protein. An instability index of less than 46 represents the protein as stable while more than 127 denotes it as unstable [11].

Sub-cellular localization

FUEL-mLoc web-based server was used to predict protein subcellular localization. It uses an elastic-net (EN) based multi-label classifier with ProSeq and ProSeq-GO database to interpret the results [12].

Prediction of Signal Peptide and Transmembrane helices

Web applications like SignalP (v5.0) [13], Outcyte [14], and Target P (v2.0) [15] were used to predict the involvement of protein in classical and non-classical secretion pathway. SignalP 5.0 uses a deep neural network-based method in combination with conditional random field classification to classify the identified signal peptides into 4 types, i.e., Sec/SPI, Sec/SPII, Tat/SPI, and, others [13]. However, few proteins lack signal sequence at their N-terminal and translocate through non-classical secretion pathways. For such proteins, Outcyte was used which predicted the protein secretion in two steps: first, it filtered out the proteins with N-terminal signal, and then classified proteins as unconventional protein secretions (UPS) or intracellular proteins. TMHMM (v2.0) server was used to identify transmembrane proteins [16].

Protein-Protein Interaction

STRING database shows the functional association of a protein using various techniques and evidence that

is divided into seven categories or "channels." Three of these channels rely on predictions using genomic context information. Another channel looks at co-expression, meaning whether genes are turned on or off together. The fifth channel involves information found through text mining, while the sixth one is based on actual biochemical or genetic experiments. The last channel focuses on knowledge gathered from previously curated pathways and protein complexes in databases [17]. So, to search for the interaction partners of non-homologous uncharacterized proteins, the String database (v11.0) was used, and proteins with confidence score > 1 were identified**.**

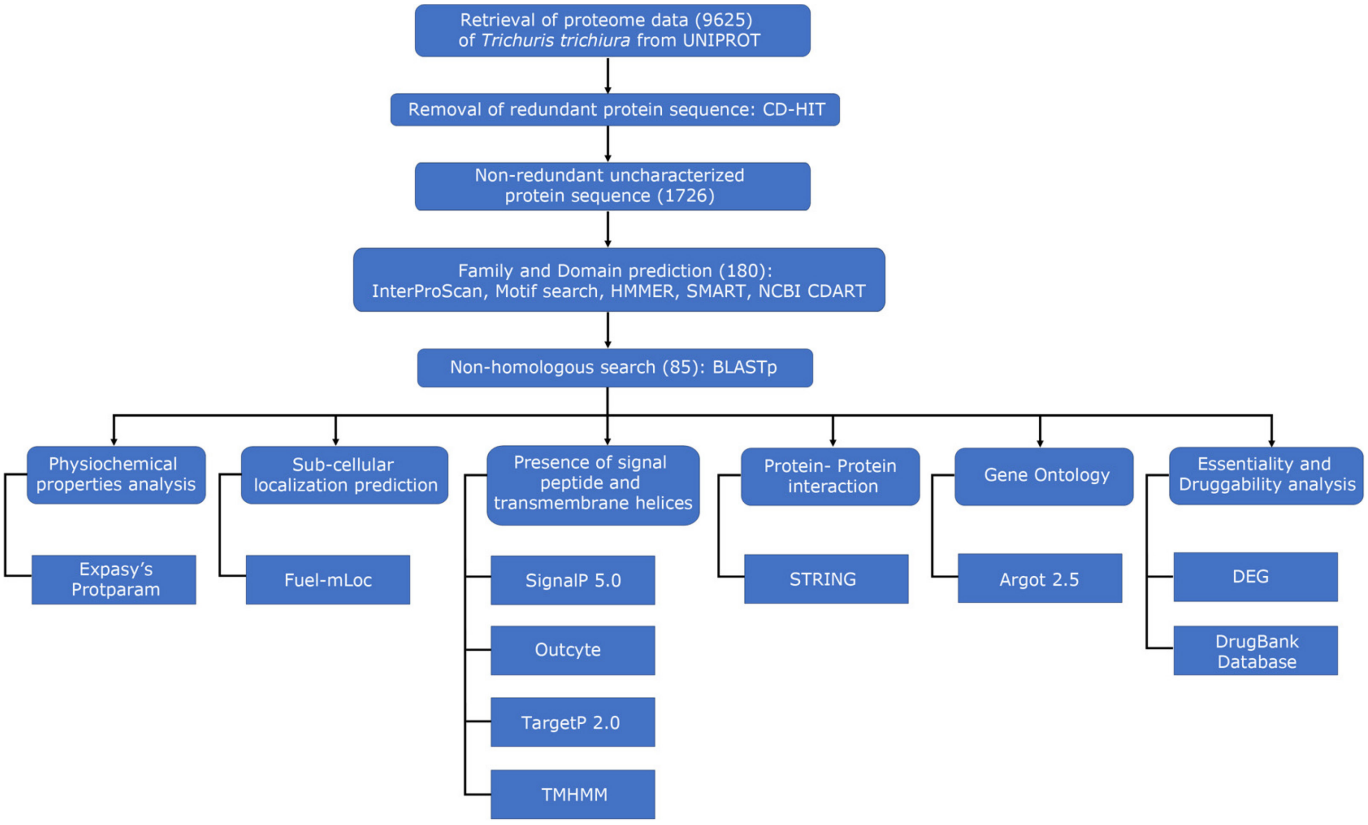
Gene Ontology

Gene Ontology categorizes the characteristics of genes and proteins into three hierarchical ontologies, specifying their molecular function (biochemical activity), biological process (pathway), and cellular component (localization). Here, GO prediction was conducted using the Argot 2.5 (Annotation Retrieval of Gene Ontology Terms) server which employed BLAST and HMMER algorithms to search for similarities against reference databases (UniProtKB and Pfam). It then annotates the sequences with GO terms retrieved from the UniProtKB-GOA database [18].

Essentiality and Druggability Analysis

To find out the essential genes of *T. trichiura*, the non-homologous, uncharacterized proteins were subjected to BLAST against the Database of Essential Genes (DEG) with an e value of 10^{-5}. DEG compiles the experimentally verified essential genes across the three domains of life. Query proteins that had similarities to genes or proteins in DEG were considered as possible essential genes or proteins [19]. Additionally, these proteins were further searched using BLASTp against the DrugBank database with an e value 10^{-5} to pinpoint potential druggable candidates. Proteins without a match in the DrugBank database were categorized as novel targets, while those with a match were considered druggable proteins. The complete methodology is depicted as a flowchart in Figure 1.

Figure 1. Sequential methodology.



Each block represents a distinct methodological phase, with directional arrows indicating the progression and interdependencies among steps. The complete proteome dataset of T. trichiura was retrieved from Uniprot, subsequently filtered to remove redundancy among proteins, and subjected to domain search. The proteins where domains were identified were then screened for their homology to the human host. Non-homologous proteins were identified and further analyzed for biophysical parameter determination, essentiality, and Druggability analysis. Tools and webservers used at each step are mentioned.

Performance assessment

The accuracy, sensitivity, and specificity of various webservers, used for function prediction or domain analysis, were validated through Receiver Operating Characteristic (ROC) analysis [20]. Approximately 100 proteins with known functions (acting as representative proteins) were employed to assess the accuracy of the domain and function prediction tools. The diagnostic efficacy was evaluated using a six-level classification that employs binary numerals '0' or '1' to classify predictions as true positive ('1') or true negative ('0'). Additionally, integers '2', '3', '4', and '5' were utilized to assess prediction efficacy. The ROC web server generated a curve between sensitivity and specificity, with the area under the curve representing effective accuracy measures ranging from 0 to 1.

Results

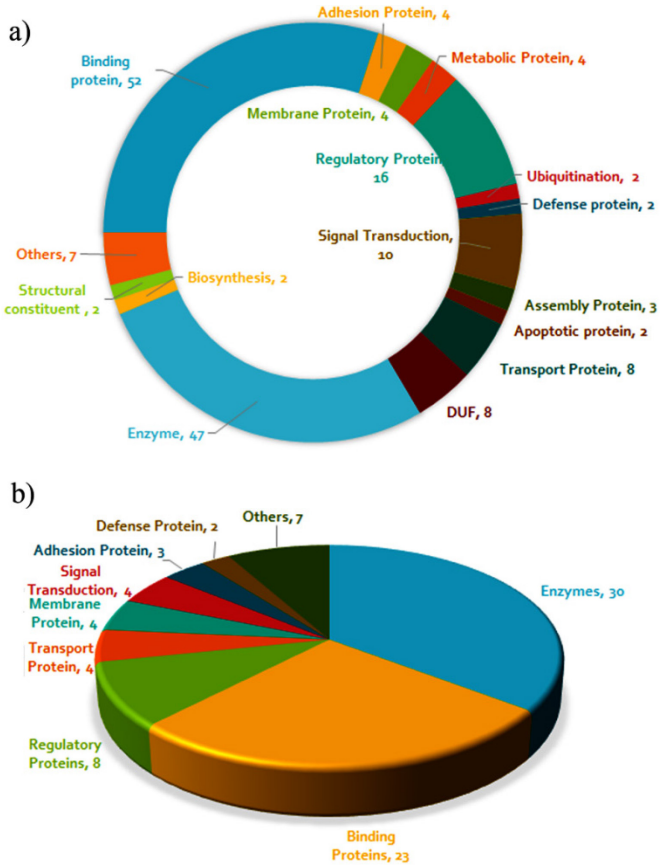
Functional Annotation of proteins

A combination of bioinformatics tools, including InterProScan, MOTIF Search, SMART, NCBI CDART, and HMMER, were utilized collectively to predict and identify conserved domains, family, and protein superfamily in 1726 uncharacterized proteins. A high level of functional annotation was achieved by assessing only the consistent functions annotated across three or more different programs which resulted in the identification of 173 uncharacterized proteins. Among these, eight proteins had conserved domains of unknown function (DUF). Subsequently, the remaining 165 proteins were functionally annotated and categorized into distinct functional classes, i.e. enzymes, binding proteins, regulatory proteins, transport proteins, membrane proteins, adhesion proteins, signal transduction, and defense proteins (Figure 2a).

Identification of non-homologous uncharacterized proteins

A non-homology search compared these 165 functionally characterized proteins of *T. trichiura* against the human proteome (taxid: 9606) using BLASTp with an e-value < 0.001. Among these, 85 proteins were identified as non-homologous to the human host, while the remaining 80 proteins, which exhibited homology to human proteins, were excluded from the study. These 85 proteins included enzymes (35%), binding proteins (27%), regulatory proteins (9%), membrane proteins (4%), signal transduction (4%), transport proteins (4%), adhesion proteins (3%), defenses proteins (2%) and proteins involved in other functions (8%) (Figure 2b), Table 1).

Figure 2. Assignment of probable functions to the proteins. **a)** Chart shows the probable functions of 173 uncharacterized proteins; **b)** Predicted function of 85 non-homologous uncharacterized proteins.



Physico-chemical properties and Sub-cellular localization

The sequences of 173 functionally characterized proteins were systematically examined to assess the physicochemical characteristics and cellular distribution. Special emphasis was placed on the non-homologous proteins, which were considered to be potential drug targets.

The theoretical isoelectric point (pI) for non-homologous uncharacterized proteins ranged from 4.24 to 10.2 and the molecular weights of these proteins ranged from 6.87 kDa to 285.117 kDa.

Most of these proteins (63) were identified as unstable, while 22 proteins had an instability index < 40 and were considered stable proteins. The extinction coefficient of these proteins varied between 1615 and 410405 M^{-1} cm^{-1}. The Grand Average of Hydropathy (GRAVY) value is crucial in discerning whether a protein is globular or membranous. Proteins with a GRAVY score below 0 are identified as hydrophilic, while those with a score above 0 are considered

hydrophobic. In this study, 65 (37.8%) proteins are hydrophilic while the rest 20 (11.6%) proteins have > 0 GRAVY score (Figure 3).

Table 1. List of functionally characterized 85 non-homologous proteins of *T. trichiura.*

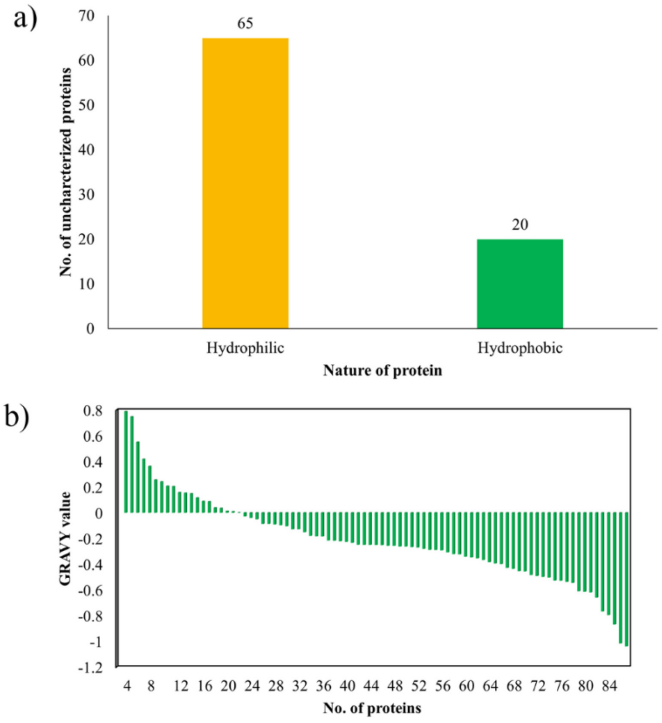
S. No.	Accession Id	Identified protein
1	A0A077Z4J9	Lysosome-associated membrane glycoprotein (Lamp)
2	A0A077Z499	KH domain
3	A0A077ZLH9	KH domain
4	A0A077Z4I5	Knot1 Protein
5	A0A077ZK07	Aspartic peptidase
6	A0A077ZHT5	BTB/POZ Protein
7	A0A077ZJP0	Ribonuclease H-like Protein
8	A0A077Z3X5	4'-phosphopantetheinyl transferase protein
9	A0A077Z4P0	Transient receptor potential cation channel
10	A0A077ZBU8	Reverse transcriptase
11	A0A077ZNU9	Reverse transcriptase
12	A0A077ZFH2	Nucleoporin Nup188
13	A0A077ZLE0	Ribonuclease H-like Protein
14	A0A077ZHK1	YdcK Beta solenoid repeat Protein
15	A0A077ZR80	Parallel Beta helix repeat Protein
16	A0A077ZFV0	SH3 Protein
17	A0A077ZLG5	Ribonuclease H-like Protein
18	A0A077ZFX3	Nucleotidyltransferase
19	A0A077ZFT8	Ribonuclease H-like Protein
20	A0A077YY48	SH2 Protein
21	A0A077ZM21	Aspartic peptidase
22	A0A077Z791	BTB/POZ Protein
23	A0A077ZG87	Regulatory subunit (RIIα) of Protein Kinase A
24	A0A077ZCZ7	Rhodopsin
25	A0A077ZMF2	Exodeoxyribonuclease I
26	A0A077ZJI5	Aspartic peptidase
27	A0A077ZHK4	TraI, Relaxase
28	A0A077ZHS0	Ribonuclease H-like Protein
29	A0A077Z447	Ionotropic glutamate receptor
30	A0A077Z971	Cadherin Protein
31	A0A077ZBK2	β-propeller folds containing Protein
32	A0A077Z4X6	Reverse transcriptase
33	A0A077Z2Q9	F-box Protein
34	A0A077ZCB7	SH2 Protein
35	A0A077Z0U1	Galectin Protein
36	A0A077YXF7	Mitogen-activated Protein Kinase
37	A0A077Z4T8	Immunoglobulin-likeProtein
38	A0A077ZEN3	Organic solute transporter
39	A0A077Z9S8	Tetracycline repressor Protein
40	A0A077ZAM4	Cyclic nucleotide-binding Protein
41	A0A077Z8H7	PNN Protein
42	A0A077ZDL0	Orc3 Protein
43	A0A077ZEG2	MFS transporter Protein
44	A0A077Z5E3	Transmembrane Protein 126
45	A0A077Z280	NRF Protein
46	A0A077ZG60	MapZ Protein
47	A0A077ZQN1	Reverse transcriptase
48	A0A077ZCR1	Aspartic peptidase
49	A0A077ZJW0	Transthyretin-like Protein
50	A0A077ZIJ5	Zinc finger, CCHC-type Protein
51	A0A077ZEZ6	GWT1 Protein
52	A0A077YWD8	Knot1 Protein
53	A0A077ZAP5	Double-stranded RNA binding Protein
54	A0A077ZE44	Lipocalin / cytosolic fatty-acid binding Protein
55	A0A077Z0H8	Ribonuclease H-like Protein
56	A0A077Z8J5	Ribonuclease H-like Protein
57	A0A077YVW8	Quiver Protein
58	A0A077ZQN6	Reverse transcriptase
59	A0A077ZIA6	AT hook Protein
60	A0A077ZH10	DotD Protein
61	A0A077ZJ13	Reverse transcriptase
62	A0A077Z4K5	Transmembrane Protein 127
63	A0A077ZL16	Phage tail tube Protein
64	A0A077ZEG9	tRNA synthetase
65	A0A077ZFN5	Quiver Protein
66	A0A077ZC43	DNA-directed RNA polymerase III
67	A0A077ZKR7	SH2 Protein
68	A0A077ZIG5	RMI2 Protein
69	A0A077YVJ2	Low-density lipoprotein (LDL) receptor class A protein
70	A0A077ZAW2	SAM Protein
71	A0A077ZKM0	HAD-like Protein
72	A0A077Z733	Rhabdovirus nucleoprotein
73	A0A077Z5E5	Globin
74	A0A077ZDU0	TAF1B/Rrn7 transcription initiation factor
75	A0A077YZD3	WD40 Protein
76	A0A077ZLK6	Aspartic peptidase
77	A0A077ZC12	Globin
78	A0A077ZHE0	Ribonuclease H-like Protein
79	A0A077Z8W2	HDc Protein
80	A0A077Z8F1	EGF Protein
81	A0A077ZG91	Quiver Protein
82	A0A077Z677	Nucleotidyltransferase
83	A0A077ZKT8	Ribonuclease H-like Protein
84	A0A077Z0E8	SH2 Protein
85	A0A077ZMI6	KH Protein

The prediction of sub-cellular localization, secretion or signaling capabilities, and the presence of transmembrane helices for the 85 uncharacterized proteins was conducted utilizing diverse tools, including Fuel-mLoc, SignalP 5.0, TargetP, Outcyte, and TMHMM 2.0. The proteins were predicted to be localized in cytoplasm (28), extracellular location (20), cell membrane (16), and nucleus (12) (Figure 4).

Proteins spanning the membrane are known for their roles in signaling pathways and facilitating the transport of nutrients across various biological environments both within and outside the cell. Around 17 of these proteins were identified to have transmembrane helices, as predicted by TMHMM (A0A077Z4J9, A0A077Z4P0, A0A077ZR80, A0A077ZCZ7, A0A077Z447, A0A077ZBK2, A0A077YXF7, A0A077Z4T8, A0A077ZEN3, A0A077ZEG2, A0A077Z5E3, A0A077ZJW0, A0A077ZEZ6, A0A077YVW8, A0A077Z4K5, A0A077Z8F1, and A0A077ZG91).

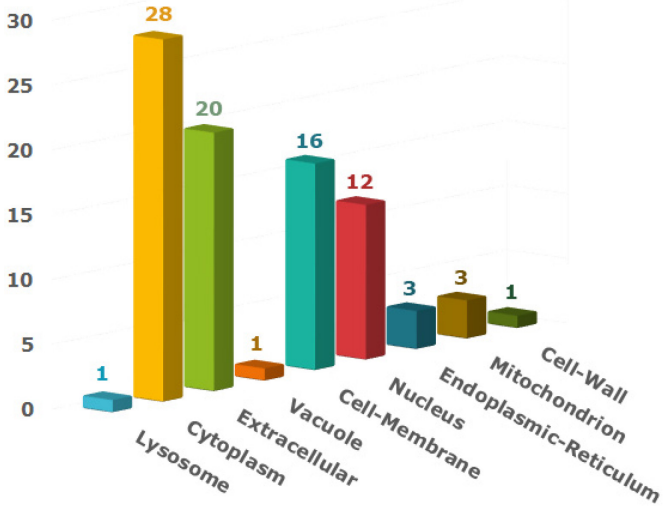
A total of 8 proteins were identified, which possess typical secretory signal peptides that undergo cleavage by Signal Peptidase I as predicted by SignalP 5.0 server,

Figure 3. Grand average of hydropathy (GRAVY) index (GI) of uncharacterized proteins.



GI score below 0 signifies the prediction of hydrophilic (globular) proteins, whereas scores exceeding 0 are indicative of hydrophobic (membrane) proteins. A total of 65 proteins were classified as hydrophilic, while 20 were identified as hydrophobic.

Figure 4. Sub-cellular localization of uncharacterized proteins that are non-homologous to humans.



hence involved in classical secretion. Only 1 protein, A0A077ZAP5, contained mitochondrial transfer peptide, while 23 proteins are involved in non-classical secretion as predicted by the Outcyte webserver.

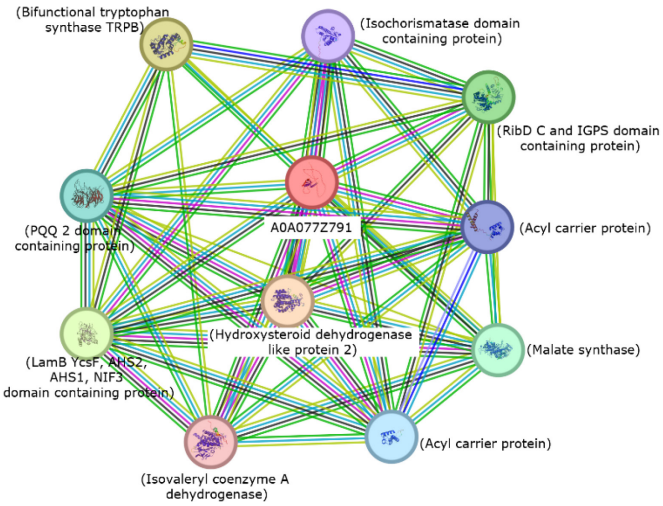
Protein-protein interactions

A protein-protein interaction analysis was conducted to uncover the uncharacterized proteins' interaction partners. Out of the 85 proteins queried on the String database, 42 proteins were identified with a confidence score greater than 1, while 35 proteins showed no interactional partners. Notably, protein with the accession ID A0A077Z791 exhibited the highest confidence score of 5.125 and was associated with 41 interaction partners, as detailed in Figure 5.

Gene Ontology analysis

In this study, the gene ontology (GO) terms of 85 functionally annotated uncharacterized proteins were

Figure 5. String analysis A0A077Z791 protein.

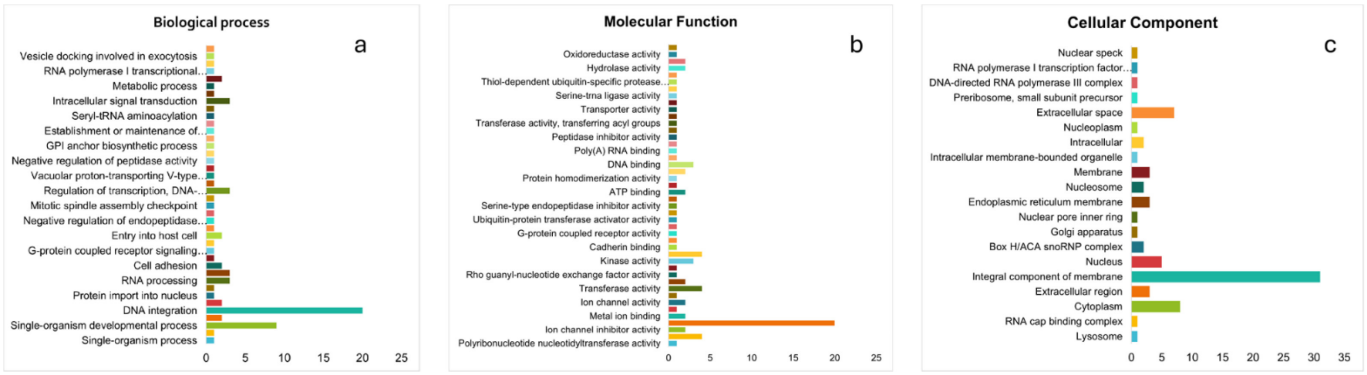


Interaction partners of A0A077Z791 (red colour) with other proteins having known or predicted structure.

examined to ascertain their associations with any of the following GO categories: Biological Process (BP), Cellular Components (CC), and Molecular Functions (MF). Among 85 proteins, 78 have been associated with biological processes with 40 GO terms. Around 20 proteins were involved in DNA integration followed by single organism developmental processes, signal transduction, and RNA processing, as shown in Figure 6 (a).

A total of 80 proteins were associated with molecular function using 44 different GO keywords. Twenty proteins were found to be nucleic-acid binding proteins, and 4 proteins were identified as RNA-binding proteins (Figure 6b). In terms of cellular components, 20 distinct GO terms for 73 proteins have been identified, in which the highest number of proteins (31)

Figure 6. Gene Ontology Classification.



Classification of the uncharacterized proteins based on gene ontology data predicted by Argot2.0 webserver to identify functional categories: **a)** Biological processes, **b)** Molecular function and **c)** Cellular components.

were associated with integral components of membrane proteins (Figure 6c).

Essentiality and Druggability analysis

The DEG database was utilized to identify fundamental genes necessary for pathogen survival. This database utilizes the fundamental genes identified via both *in vitro* and *in vivo* experiments and help in detection of genes with necessary cellular function. Proteins A0A077ZLE0 and A0A077Z971 were found to be essential for *T. trichiura*, as these proteins show similarity against the DEG database with DEG accession number DEG20280006, DEG20020065 respectively. For the druggability analysis, the non-homologous uncharacterized proteins were queried against the targets available in DrugBank database. None of the 85 proteins exhibited similarities with the Drug Bank database (E-value 0.00001) and may be considered novel drug targets.

ROC analysis

Receiver Operating Characteristic (ROC) analysis was done by subjecting the 100 *T. trichiura* proteins with known functions (acting as representative proteins) to the methodology adopted in this study and the results were scored. The average accuracy and ROC area for the utilized database(s) were determined to be 94.6% and 0.97, respectively.

Discussion

The uncharacterized proteome makes a large fraction of total known proteins of *T. trichiura*. A detailed in silico analysis was carried out to functionally annotate about 165 proteins whose functions ranged from enzymes to transport proteins and regulatory proteins.

Enzymes

A0A077ZK07, A0A077ZM21, A0A077ZJI5, A0A077ZCR1, and A0A077ZLK6 were identified as aspartic peptidases, which are widely prevalent in nematodes, parasites, plants, fungi, and viruses and are characterized by two catalytic aspartyl residues in their conserved Asp-Thr/Ser-Gly motif [21]. Aspartic peptidases are endopeptidases that play a key role in nematode parasitism due to their involvement in host-parasite interactions, thus considered as potential candidates for antiparasitic drugs and protective immunization [22]. In other parasitic nematodes, e.g. in *Ascaris suum* and *Clonorchis sinensis*, aspartic peptidases support parasites by facilitating nutrient assimilation, evasion of host immune responses, and the overall persistence of the parasite within the host organism [23].

A0A077ZJP0, A0A077ZLE0, A0A077ZLG5, A0A077ZFT8, A0A077ZHS0, A0A077Z0H8, A0A077Z8J5, A0A077ZHE0, and A0A077ZKT8 belong to retroviral integrase (Rve) superfamily also known as the Ribonuclease H-like (RNHL) superfamily which contains numerous enzymes involved in replication, homologous recombination, DNA repair, transposition, and RNA interference [24]. A0A077Z3X5 is identified as a 4'-phosphopantetheinyl transferase (PPTases) domain superfamily member. PPTases catalyze posttranslational modifications by transferring the 4′-phosphopantetheine prosthetic group from CoA to conserved serine residues in carrier proteins. This facilitates the transformation of inactive "apo" forms of fatty acid synthases (FASs), polyketide synthases (PKSs), and non-ribosomal peptide synthases (NRPSs) to active "holo" forms [25].

A0A077ZBU8, A0A077ZNU9, A0A077Z4X6, A0A077ZQN6, A0A077ZJ13, and A0A077ZQN1 were identified as reverse transcriptase enzymes; an RNA-dependent DNA polymerase well-known for converting RNA into complementary DNA (cDNA). Reverse transcriptase functions as retrotransposon RTs (with a role in gene regulation) and telomerase (involved in the maintenance of chromosomal ends, cellular aging, and immortalization) [26].

A0A077ZFX3 and A0A077Z677 are Nucleotidyltransferase superfamily (NTS) enzymes that transfer nucleotides to acceptor hydroxyl groups attached to proteins, nucleic acids, and small molecules. This class of enzymes is characterized by the presence of αβαβαβ topology and three or four conserved motifs of carboxylates (the DDE or DEDD motifs) in the active site. The NTase fold superfamily plays an important role in DNA replication and repair, transcription, RNA processing, viral replication, telomere maintenance, etc. [27].

A0A077ZMF2 is identified as exodeoxyribonuclease I or exonuclease I. Nucleases are essential to DNA replication; they remove RNA primers and have 5'-3' and 3'-5' proofread activity. The activity of nucleases is vital to several processes occurring in DNA, including DNA metabolism, recombination and repair, topoisomerization, site-specific recombination, and RNA splicing [28].

A0A077ZBK2 is identified as a protein containing β-propeller folds which function as hydrolase, lyase, isomerase, signaling protein, structural protein, and ligand-binding protein [29]. A0A077YXF7 is identified as a mitogen-activated protein (MAP) kinase that

regulates cellular processes such as proliferation, stress responses, apoptosis, and immune defense [30]. A0A077ZEG9 is recognized as a protein containing tRNA synthetase class II core domain. Researchers have identified and characterized the lysyl-tRNA synthetases from *Loa loa* and *Schistosoma mansoni*, demonstrating their inhibition by cladosporin, thereby highlighting the targeting of protein synthesis pathways as a novel approach for drug development [31]. A0A077ZC43 is found to be an RPC4 subunit of DNA-directed RNA polymerase III synthesizing a wide variety of small essential RNAs, encompassing tRNAs, 5S rRNA, and some snRNAs [30]. A0A077ZKM0 is a member of haloacid dehalogenase (HAD)-like superfamily composed of potential phosphatases, ATPases, phosphonatases, and phosphomutases, all playing substantial roles in hydrolytic enzyme activities. A0A077Z8W2 is a member of HDc superfamily featuring doublet of His-Asp residues (HXnHDXnD motif) that coordinates an active site metallocentre and performs functions such as immune response, nucleic acid metabolism, inflammation, virulence, stress response, and small molecule activation [32].

Binding Proteins

A0A077Z499, A0A077ZLH9, and A0A077ZMI6 are RNA binding proteins with K Homology (KH) type I domain, potentially involved in transcriptional and translational regulation [33]. A0A077ZFV0 is identified as SH3 (Src homology 3) domain containing protein which is a small noncatalytic domain of 50 to 60 amino acids. In addition to proliferating cells and differentiation, SH3 domains regulate cell survival, cell migration and cytoskeletal rearrangements, protein trafficking, degradation, protein-protein interactions, morphogenesis, signal transduction, and immunity [34]. A0A077YY48, A0A077ZCB7, A0A077ZKR7 and A0A077Z0E8 belong to the SH2 domain superfamily. The SH2 (Src Homology 2) domain is a small non-catalytic protein module (100 amino acids) involved in several pathways that regulate important physiological activities [35]. A0A077Z8F1 contains an epidermal growth factor-like (EGFL) domain and is thought to be involved in the interaction between proteins [36].

A0A077Z971 is identified as a protein containing the cadherin domain responsible for facilitating cadherin-mediated cell-cell interactions. Generally, cadherins play an important role in cell adhesion, cell migration, tissue organization, morphogenesis, and cell signaling [37]. A0A077Z2Q9 is identified as an F-box domain containing protein. These domains are 50 amino acid long adapter peptides that are involved in the SKP1-CUL1-F-box protein (SCF)-mediated ubiquitination in the protein degradation pathway [38]. A0A077Z0U1 is a protein containing galectin domain used by parasites for adhesion, cytolysis, invasion, resistance to complement lysis, and perhaps encystment to gain host entry and evade the host's immune response. The galectin protein from *Haemonchus contortus* was identified among uncharacterized proteins which potentially inhibits T cell proliferation and alter MHC-II expression on monocytes, thereby manipulating the host immune response [39]. This highlights the potential of galectin as a target for therapeutic strategies that are aimed at disrupting parasitic immune evasion [40]. A0A077ZAM4 is a cyclic nucleotide-binding domain protein, an important signaling module that controls various cellular processes in both prokaryotes and eukaryotes. It responds to accumulating second messengers, such as cyclic AMP (cAMP) and cyclic GMP (cGMP). Due to its role in cellular signaling, it could be a potential target for pathogen-specific treatments or drug development [41].

A0A077ZDL0 is a third subunit of hexameric origin recognition complex (ORC) that controls the initiation of DNA replication during the cell cycle in all eukaryotic species. In addition to DNA replication, ORC subunits regulate neuronal maturation in non-proliferating cells (adrenal cortical cells, cardiac myocytes, and neurons) [42]. A0A077ZIJ5 is a zinc finger, CCHC-type superfamily member containing a ZnF motif characterized by conserved cysteine and histidine amino acid residues. This motif contributes to the recognition and binding of single-stranded nucleic acids, particularly single-stranded RNAs, thus enhancing RNA functions and processes such as transcription, biogenesis, splicing, translation, and degradation [43]. A0A077ZIA6 is a protein containing AT-hook motif. The AT-hook represents a constituent of the high mobility group non-histone chromosomal protein HMG-I(Y). AT-hook proteins have crucial functions in both chromatin structure and act as cofactors for transcription factors [44].

A0A077ZE44 is identified as a member of Lipocalin / cytosolic fatty-acid binding protein family that works in retinol transport, invertebrate cryptic coloration, olfaction, pheromone transport, and prostaglandin synthesis. A0A077ZIG5 is an RMI2 domain containing protein that helps maintain genome stability [45]. A0A077ZAW2 is recognized as a protein containing the sterile alpha motif (SAM) domain

involved in critical processes such as signal transduction and mediating cell-cell communication. A0A077Z5E5 and A0A077ZC12 are identified as globin proteins harboring a prosthetic heme group surrounded by 8 helices. In addition to the reversible binding of oxygen for transport and storage, they also serve as cytoprotectants against reactive oxygen species and NO scavengers within oxygen-dependent metabolic pathways [46]. A0A077ZDU0 is identified as a TAF1B/Rrn7 transcription initiation factor of RNA polymerase I. RRN7 is a 60 kDa protein that plays a crucial role in transcription initiation within RNA polymerase I. A0A077YZD3 belongs to a WD40-repeat-containing domain superfamily. These repeats comprise a Glycine-Histidine (GH) and Tryptophan-Aspartate (WD) motif of roughly 40–60 amino acids. WD40 repeat containing proteins play a role in signal transduction, transcription, vesicle trafficking, cytoskeletal assembly, apoptosis, cell cycle regulation, and chromatin modification, which help in homeostasis and normal body functioning. Due to their widespread distribution and functional diversity, WD-repeat proteins are suitable targets for pharmacological studies [47].

Regulatory Proteins

A0A077ZHT5 and A0A077Z791 belong to SKP1/BTB/POZ domain superfamily involved in a variety of functional roles, such as protein ubiquitination/degradation, transcription repression, tetramerization and gating of ion channels and protein-protein interaction. A0A077ZG87 is identified as a regulatory subunit (RIIα) of Protein Kinase A (PKA). A0A077Z9S8 contains tetracycline repressor (TetR) domain, a homodimeric (24.3 kDa monomer) α-helical protein that regulates the transcription of genes whose byproducts are implicated in osmotic stress, antibiotic production, and multidrug resistance [48]. A0A077ZAP5 is recognized as a protein containing a double-stranded RNA binding motif.

A0A077YVW8, A0A077ZFN5, and A0A077ZG91 are sleepless/quiver (SSS) proteins that are highly conserved, small (~15 kDa), glycosylphosphatidylinositol (GPI)–anchored membrane proteins crucial for the regulation of sleep, in insects and worms [49].

Membrane Proteins

A0A077Z4J9 is identified as lysosome-associated membrane glycoprotein (LAMPs), a highly glycosylated lysosome-specific, integral membrane protein involved in cholesterol transport. A0A077Z5E3 is recognized as a transmembrane protein 126 (TMEM126) involved in mitochondrial function. A0A077Z4K5 is a transmembrane protein 127 (TMEM 127), potentially a tumor suppressor while A0A077YVJ2 is a low-density lipoprotein (LDL) receptor class A protein. In *salmonella*, TMEM127 is known to damage MHC Class II Molecules by ubiquitination [50].

Signal transduction/Transducer

A0A077Z4P0 is a transient receptor potential (TRP) channel representing a substantial and versatile family of cation-conducting channels that are pivotal in mediating host-parasite interactions, critical for parasite life cycle and pathogenesis [51]. A0A077ZCZ7 is a member of Rhodopsin-like G protein-coupled receptors (GPCRs). A0A077ZCZ7 is a Rhodopsin-like G protein-coupled receptor (GPCR). A0A077Z447 is an ionotropic glutamate receptor (iGluR), a ligand-gated ion channel specific to invertebrates, which may disrupt a parasite's life cycle [52]. A0A077ZL16 is identified as a phage tail tube protein (TTP). The GPCRs and ionotropic glutamate receptors have previously been explored as drug targets for anti-parasitic interventions. Sriram *et. al*. reported that about 134 human GPCRs had been developed as FDA-approved drugs, while a lot more are in the process [53]. Ionotropic glutamate receptors are associated with memory in *C. elegans* and are potential drug targets [51].

Transport Protein

A0A077ZEN3 is identified as a subunit of heteromeric organic solute transporter (OST) protein. OSTs are involved in the transport of cholesterol, drugs, and other important molecules. It has been found that OSTs are crucial for the survivability of trematodes, as excessive bile acids can cause the death of the worms [54]. A0A077ZEG2 is a member of the major facilitator superfamily proteins, which play an important role in cellular growth, metabolism, and homeostasis and are involved in several physiological processes, such as development, neurotransmission, and signaling. A0A077Z280 was identified as a protein containing the NRF domain. A0A077ZH10 is identified as a DotD protein, an outer membrane lipoprotein of the Dot/Icm Type IV Secretion System (T4SS) playing a crucial role in the outer membrane targeting of DotH, another protein in the T4BSS core complex [55].

Adhesion Protein

A0A077ZHK1 is a YdcK β-solenoid repeat-containing protein and is identified as a virulence factor

or adhesin in bacteria and fungi. A0A077ZR80 is identified as a parallel β-helix repeat-containing protein, which appears to be a single cooperative repetitive unit that functions as a toxin, virulence factor, adhesin, or surface protein involved in the pathogenesis of bacteria and fungi [56]. A0A077Z8H7 harbored a PNN-interacting serine/arginine-rich 140 kDa phosphoprotein domain, which is known to transcribe genes in epithelial cells and works in the establishment and preservation of adhesion in corneal epithelium [57]. These proteins are biomarkers in various human cancers [58].

Defense Proteins

A0A077Z4I5 and A0A077YWD8 contain a knottin, scorpion toxin-like domain, which is a disulfide bridge-containing peptide (DBP) found in unrelated protein families, e.g., toxins and antimicrobials produced by plants, arthropods, mollusks, and nematodes. These molecules exhibit a distinctive knotted topology characterized by three disulfide bridges; one penetrates through a macrocycle formed by the other two disulfides, creating interconnected peptide bonds within the structure. Cysteine knot peptides have exceptional stability and well-defined scaffolds, making them attractive candidates for pharmaceutical use as lead molecules or molecular frameworks. Cysteine knot proteins have found applications in developing new protease inhibitors, integrin-binding medicines, anti-HIV medications, growth factor mimics, and multi-epitope vaccination development [59,60].

Others

A0A077ZFH2 is recognized as a Nucleoporin 188 protein, a component of nuclear pore complexes (NPCs) that controls the passage of membrane proteins in the nuclear envelope. It also facilitates microtubule formation at centrosomes, thus maintaining the integrity of nuclear membranes and chromosome alignment respectively [61]. A0A077ZHK4 is identified as the TraI domain of the type IV secretion system (T4SS). A0A077Z4T8 is identified as an immunoglobulin-like domain superfamily protein (IgSF), which plays an important role in cell-to-cell recognition, cell surface receptors, and muscle structure [62]. A0A077ZG60 is identified as a transmembrane mid-cell anchored protein Z (MapZ), that regulates cell division by controlling the FtsZ (Filamenting temperature-sensitive mutant Z) ring constriction in prokaryotes [63]. A0A077ZJW0 is recognized as a secretory TTR-52 protein belonging to the transthyretin-like protein family, a subfamily of the transthyretin-related protein family (TRPs) [64]. A0A077ZEZ6 is recognized as a PIG-W/GWT1 superfamily transmembrane protein necessary for facilitating the attachment of cell surface proteins to the lipid bilayer [65]. A0A077Z733 shows similarity with a rhabdovirus nucleoprotein (RNP), which plays a crucial role in virus assembly and serves as a template for transcription and replication for the viral polymerase [66].

Novel Drug targets

Because of their distinct, non-homologous origin, the proteins identified in this study confer encouraging opportunities for developing therapeutics. Most of these proteins, including enzymes, binding proteins, and regulatory proteins, are essential for the survival and pathogenicity of the organism. Targeting these proteins via specific inhibitors that disrupt essential biological pathways within the parasite will aid in reducing its virulence or leading to its eradication without affecting the human host.

Additionally, proteins, such as aspartic peptidases, membrane proteins, and cell adhesion proteins, which are involved in host-parasite interactions, play critical roles in the lifecycle of parasitic nematodes. Many studies have explored the possibility of multiepitope vaccines utilizing novel B and T cell epitopes of cellular proteins of parasites [67-69]. Such studies have been successful in identifying uncharacterized proteins as valid vaccine candidates in *Plasmodium falciparum* [70], *Leishmania donovani* [71], *Mycobacterium orygis* [72], and *Neisseria gonorrhoeae* [73] to list a few. The proteins identified here could also be used in a similar manner for epitope detection and vaccine development, potentially reducing infection severity or preventing infection altogether. Simultaneously, future studies might also focus on structural and functional assays of these proteins to validate their roles as drug targets, paving the way for targeted therapeutic approaches.

Conclusions

The rapid advancement of technologies enables the generation of extensive genomic or transcriptomic data for numerous organisms in a short timeframe. Despite this progress, a significant portion of the data remains uncharacterized, potentially participating in various biological processes. Therefore, the functional annotation of this uncharacterized data or proteins is essential. While experimental elucidation of functional assignments for uncharacterized proteins is possible, it is often expensive, time-consuming, and, in many

cases, technically unfeasible.

This study employs diverse strategies to functionally annotate 1726 uncharacterized proteins from *T. trichiura*. Based on the identified domains and physiochemical characterization, the functions of 173 uncharacterized proteins of *T. trichiura* were predicted. Subsequent non-homologous analysis identified 85 proteins as non-homologous in the human proteome database. These 85 proteins were further categorized into nine functional classes, including enzymes, binding and regulatory proteins. Druggability and essentiality analysis of these 85 proteins led to the identification of all proteins as novel targets while 2 proteins were recognized as essential for *T. trichiura.* In addition, proteins predicted to function as tRNA synthetase, 4'-phosphopantetheinyl transferase, SH2 protein, reverse transcriptase enzyme, Knot1 protein, aspartic peptidases, galectin protein, and cyclic nucleotide-binding protein have been identified as potential therapeutic targets. On the other hand, proteins characterized as transient receptor potential (TRP) channels, rhodopsin-like G protein–coupled receptors, and ionotropic glutamate receptors could serve as potential vaccine targets. Notably, Dot D and type IV secretion system proteins, typically exclusive to bacterial cells, were identified in the proteome of *T. trichiura*. These results hint at a potential endosymbiotic relationship between the organism and bacteria or the association of this parasite with gut microbiota for survival, necessitating further investigations for confirmation. This study represents a significant step toward the drug design process for *T. trichiura*, with validation anticipated through laboratory experiments. There are still many proteins in this parasite which remain uncharacterized and need further exploration. Moreover, the methodology employed in this study holds promise for characterizing hypothetical or uncharacterized proteins in other organisms.

Acknowledgements

Kanchan Rauthan, Saranya Joshi, Lokesh Kumar acknowledge the non-net fellowship of UGC.

Availability of data and materials

The datasets utilized and examined in the present investigation can be provided upon request.

Corresponding author

Sudhir Kumar, PhD
Special Centre for Molecular Medicine,
Jawaharlal Nehru University, New Delhi, India-110067
Tel: +91-11-26739223
Email: sudhir.k@mail.jnu.ac.in; sudhir.1685@gmail.com

Conflict of interests

No conflict of interests is declared.

References

1. Mascarini-Serra L (2011) Prevention of soil-transmitted helminth infection. J Glob Infect Dis 3: 175-82. doi: 10.4103/0974-777X.81696.
2. Centre for disease control (CDC) (2023) Parasites - Trichuriasis (also known as Whipworm Infection). Available: https://www.cdc.gov/parasites/whipworm/index.html. Accessed: 20 September 2023.
3. Pullan RL, Smith JLJasrasaria R, Brooker SJ (2010) Global numbers of infection and disease burden of soil transmitted helminth infections in 2010. Parasit Vectors 7: 37. doi: 10.1186/1756-3305-7-37.
4. Masaku J, Njomo DW, Njoka A, Okoyo C, Mutungi FM, Njenga SM (2020) Soil-transmitted helminths and schistosomiasis among pre-school age children in a rural setting of Busia County, Western Kenya: a cross-sectional study of prevalence, and associated exposures. BMC Public Health 20: 356. doi: 10.1186/s12889-020-08485-z.
5. Foth BJ, Tsai IJ, Reid AJ, Bancroft AJ, Nichol S, Tracey A, Holroyd N, Cotton JA, Stanley EJ, Zarowiecki M, Liu JZ, Huckvale T, Cooper PJ, Grencis RK, Berriman M (2014) Whipworm genome and dual-species transcriptome analyses provide molecular insights into an intimate host-parasite interaction. Nat Genet 46: 693-700. doi: 10.1038/ng.3010.
6. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658-9. doi: 10.1093/bioinformatics/btl158.
7. Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C, McMenamin C, Nuka G, Pesseat S, Sangrador-Vegas A, Scheremetjew M, Rato C, Yong SY, Bateman A, Punta M, Attwood TK, Sigrist CJ, Redaschi N, Rivoire C, Xenarios I, Kahn D, Guyot D, Bork P, Letunic I, Gough J, Oates M, Haft D, Huang H, Natale DA, Wu CH, Orengo C, Sillitoe I, Mi H, Thomas PD, Finn RD (2015) The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res 43: D213-D21. doi: 10.1093/nar/gku1243.
8. Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39: W29-W37. doi: 10.1093/nar/gkr367.
9. Letunic I, Khedkar S, Bork P (2021) SMART: recent updates, new developments and status in 2020. Nucleic Acids Res 49: D458-D60. doi: 10.1093/nar/gkaa93710.

10. Geer LY, Domrachev M, Lipman DJ, Bryant SH (2002) CDART: protein homology by domain architecture. Genome Res 12: 1619-23. doi: 10.1101/gr.278202.
11. Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, Williams KL, Appel RD, Hochstrasser DF (1999) Protein identification and analysis tools in the ExPASy server. Methods Mol Biol 112: 531-52. doi: 10.1385/1-59259-584-7:531.
12. Wan S, Mak MW, Kung SY (2017) FUEL-mLoc: feature-unified prediction and explanation of multi-localization of cellular proteins in multiple organisms. Bioinformatics 33: 749-750. doi: 10.1093/bioinformatics/btw717.
13. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H (2019) SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol 37: 420-423. doi: 10.1038/s41587-019-0036-z.
14. Zhao L, Poschmann G, Waldera-Lupa D, Rafiee N, Kollmann M, Stühler K (2019) OutCyte: a novel tool for predicting unconventional protein secretion. Sci Rep 9: 19448. doi: 10.1038/s41598-019-55351-z.
15. Armenteros JJA, Salvatore M, Emanuelsson O, Winther O, von Heijne G, Elofsson A, Nielsen H. Detecting sequence signals in targeting peptides using deep learning. Life Sci Alliance 2: e201900429. doi: 10.26508/lsa.201900429.
16. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305: 567-580. doi: 10.1006/jmbi.2000.4315.
17. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, Mering CV (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 47: D607–D613. doi: 10.1093/nar/gky1131.18.
18. Fontana P, Cestaro A, Velasco R, Formentin E, Toppo S (2009) Rapid annotation of anonymous sequences from genome projects using semantic similarities and a weighting scheme in gene ontology. PLoS One 4: e4619. doi: 10.1371/journal.pone.0004619.
19. Zhang R, Ou HY, Zhang CT (2004) DEG: a database of essential genes. Nucleic Acids Res 32: D271-D272. doi: 10.1093/nar/gkh024.
20. Eng J (2014) ROC analysis: web-based calculator for ROC curves. Baltimore: Johns Hopkins University [updated 2014 March 19]. Available: http://www.jrocfit.org. Accessed: 13 June 2023.
21. Davies DR (1990) The structure and function of the aspartic proteinases. Annu Rev Biophys Biophys Chem 19: 189-215. doi: 10.1146/annurev.bb.19.060190.001201.
22. Santos LO, Garcia-Gomes AS, Catanho M, Sodre CL, Santos AL, Branquinha MH, d'Avila-Levy CM (2013) Aspartic peptidases of human pathogenic trypanosomatids: perspectives and trends for chemotherapy. Curr Med Chem 20: 3116-3133. doi: 10.2174/0929867311320250007.
23. Malagón D, Benítez R, Kasny M, Adroher FJ (2013) Peptidases in parasitic nematodes. A review. In: Parasites: ecology, diseases and management. Eds Gilmar S. Erzinger. pp 61-102.
24. Majorek KA, Dunin-Horkawicz S, Steczkiewicz K, Muszewska A, Nowotny M, Ginalski K, Bujnicki JM (2014) The RNase H-like superfamily: new members, comparative structural analysis and evolutionary classification. Nucleic Acids Res 42: 4160-4179. doi: 10.1093/nar/gkt1414.
25. Lambalot RH, Gehring AM, Flugel RS, Zuber P, LaCelle M, Marahiel MA, Reid R, Khosla C, Walsh CT (1996) A new enzyme superfamily - the phosphopantetheinyl transferases. Chem Biol 3: 923-936. doi: 10.1016/s1074-5521(96)90181-7.
26. Coffin JM, Fan H (2016) The discovery of reverse transcriptase. Annu Rev Virol 3: 29-51. doi: 10.1146/annurev-virology-110615-035556.
27. Kuchta K, Knizewski L, Wyrwicz LS, Rychlewski L, Ginalski K (2009) Comprehensive classification of nucleotidyltransferase fold proteins: identification of novel families and their representatives in human. Nucleic Acids Res 37: 7701-7714. doi: 10.1093/nar/gkp854.
28. Yang W (2011) Nucleases: diversity of structure, function and mechanism. Q Rev Biophys 44: 1-93. doi: 10.1017/S0033583510000181.
29. Andersen OM, Dagil R, Kragelund BB (2013) New horizons for lipoprotein receptors: communication by beta-propellers. J Lipid Res 54: 2763-2774. doi: 10.1194/jlr.M039545.
30. Saravanan P, Venkatesan SK, Mohan CG, Patra S, Dubey VK (2010) Mitogen-activated protein kinase 4 of Leishmania parasite as a therapeutic target. Eur J Med Chem 45: 5662-5670. doi: 10.1016/j.ejmech.2010.09.020.
31. Sharma A, Sharma M, Yogavel M, Sharma A (2016) Protein translation enzyme lysyl-tRNA synthetase presents a new target for drug development against causative agents of Loiasis and Schistosomiasis. PLoS Negl Trop Dis 10: e0005084. doi: 10.1371/journal.pntd.0005084
32. Langton M, Sun S, Ueda C, Markey M, Chen J, Paddy I, Jiang P, Chin N, Milne A, Pandelia ME (2020) The HD-domain metalloprotein superfamily: an apparent common protein scaffold with diverse chemistries. Catalysts 10: 1191. doi: 10.3390/catal10101191.
33. Valverde R, Edwards L, Regan L (2008) Structure and function of KH domains. FEBS J 275: 2712-2726. doi: 10.1111/j.1742-4658.2008.06411.x.
34. Kurochkina N, Guha U (2013) SH3 domains: modules of protein-protein interactions. Biophys Rev 5: 29-39. doi: 10.1007/s12551-012-0081-z.
35. Waksman G, Kumaran S, Lubman O (2004) SH2 domains: role, structure and implications for molecular medicine. Expert Rev Mol Med 6: 1-18. doi: 10.1017/S1462399404007331.
36. Song IJ, Ikram M, Subhan F, Choi DJ, Lee JR, Kim HS, Lim YT, Yoon S (2015) Molecular characterization and expression analysis of mouse epidermal growth factor-like domain 8. Int J Mol Med 36: 541-550. doi: 10.3892/ijmm.2015.2252.
37. Maitre JL, Heisenberg CP (2013) Three functions of cadherins in cell adhesion. Curr Biol 23: R626-R633. doi: 10.1016/j.cub.2013.06.019.
38. Kipreos ET, Pagano M (2000) The F-box protein family. Genome Biol 1: reviews3002.1. doi: 10.1186/gb-2000-1-5-reviews3002.
39. Wang W, Wang S, Zhang H, Yuan C, Yan R, Song X, Xu L, Li X (2014) Galectin Hco-gal-m from Haemonchus contortus modulates goat monocytes and T cell function in different patterns. Parasit Vectors 7: 342. doi : 10.1186/1756-3305-7-342
40. Vasta GR (2020) Galectins in host-pathogen interactions: structural, functional and evolutionary aspects. Adv Exp Med Biol 1204: 169-196. doi: 10.1007/978-981-15-1580-4_7.
41. Mohanty S, Kennedy EJ, Herberg FW, Hui R, Taylor SS, Langsley G, Kannan N (2015) Structural and evolutionary divergence of cyclic nucleotide binding domains in eukaryotic

pathogens: Implications for drug design. Biochim Biophys Acta 1854: 1575-1585. doi: 10.1016/j.bbapap.2015.03.01242.
42. Sasaki T, Gilbert DM (2007) The many faces of the origin recognition complex. Curr Opin Cell Biol 19: 337-343. doi: 10.1016/j.ceb.2007.04.007.
43. Wang Y, Yu Y, Pang Y, Yu H, Zhang W, Zhao X, Yu J (2021) The distinct roles of zinc finger CCHC-type (ZCCHC) superfamily proteins in the regulation of RNA metabolism. RNA Biol 18: 2107-2126. doi: 10.1080/15476286.2021.1909320.
44. Garabedian A, Jeanne Dit Fouque K, Chapagain PP, Leng F, Fernandez-Lima F (2022) AT-hook peptides bind the major and minor groove of AT-rich DNA duplexes. Nucleic Acids Res 50: 2431-2439. doi: 10.1093/nar/gkac115.
45. Velkova M, Silva N, Dello Stritto MR, Schleiffer A, Barraud P, Hartl M, Jantsch V (2021) Caenorhabditis elegans RMI2 functional homolog-2 (RMIF-2) and RMI1 (RMH-1) have both overlapping and distinct meiotic functions within the BTR complex. PLoS Genet 17: e1009663. doi: 10.1371/journal.pgen.1009663.
46. Wajcman H, Kiger L, Marden MC (2009) Structure and function evolution in the superfamily of globins. C R Biol 332: 273-282. doi: 10.1016/j.crvi.2008.07.026.
47. Jain BP, Pandey S (2018) WD40 repeat proteins: signalling scaffold with diverse functions. Protein J 37: 391-406. doi: 10.1007/s10930-018-9785-7.
48. Ramos JL, Martínez-Bueno M, Molina-Henares AJ, Terán W, Watanabe K, Zhang X, Gallegos MT, Brennan R, Tobes R (2005) The TetR family of transcriptional repressors. Microbiol Mol Biol Rev 69: 326-356. doi: 10.1128/MMBR.69.2.326-356.2005.
49. Koh K, Joiner WJ, Wu MN, Yue Z, Smith CJ, Sehgal A (2008) Identification of SLEEPLESS, a sleep-promoting factor. Science 321: 372-376. doi: 10.1126/science.1155942.
50. Alix E, Godlee C, Cerny O, Blundell S, Tocci R, Matthews S, Liu M, Pruneda JN, Swatek KN, Komander D, Sleap T, Holden DW (2020) The tumour suppressor TMEM127 Is a Nedd4-Family E3 Ligase Adaptor required by Salmonella SteD to ubiquitinate and degrade MHC Class II molecules. Cell Host Microbe 28: 54-68. doi: 10.1016/j.chom.2020.04.024
51. Wolstenholme AJ, Williamson SM, Reaves BJ (2011) TRP channels in parasites. Adv Exp Med Biol 704: 359-71. doi: 10.1007/978-94-007-0265-3_20.
52. Brockie PJ, Maricq AV (2006) Ionotropic glutamate receptors: genetics, behavior and electrophysiology. WormBook 19: 1-16. doi: 10.1895/wormbook.1.61.1.
53. Sriram K, Insel PA (2018) G protein-coupled receptors as targets for approved drugs: how many targets and how many drugs? Mol Pharmacology 93: 251. doi:10.1124/mol.117.111062
54. Lu Y, Yoo WG, Dai F, Lee JY, Pak JH, Sohn WM, Hong SJ (2018) Characterization of a novel organic solute transporter homologue from Clonorchis sinensis. PLoS Negl Trop Dis 12: e0006459. doi: 10.1371/journal.pntd.0006459.
55. Nakano N, Kubori T, Kinoshita M, Imada K, Nagai H (2010) Crystal structure of Legionella DotD: insights into the relationship between type IVB and type II/III secretion systems. PLoS Pathog 6: e1001129. doi: 10.1371/journal.ppat.1001129.
56. Jenkins J, Pickersgill R (2001) The architecture of parallel beta-helices and related folds. Prog Biophys Mol Biol 77: 111-175. doi: 10.1016/s0079-6107(01)00013-x.
57. Zimowska G, Shi J, Munguba G, Jackson MR, Alpatov R, Simmons MN, Shi Y, Sugrue SP (2003) Pinin/DRS/memA interacts with SRp75, SRm300 and SRrp130 in corneal epithelial cells. Invest Ophthalmol Vis Sci 44: 4715-4723. doi: 10.1167/iovs.03-0240.
58. Wang R, Qin Z, Luo H, Pan M, Liu M, Yang P, Shi T (2022) Prognostic value of PNN in prostate cancer and its correlation with therapeutic significance. Front in Genetics 13: 1056224. doi: 10.3389/fgene.2022.1056224
59. Pai PP, Mondal S (2017) Intriguing cystine-knot miniproteins in drug design and therapeutics. In: Toxins and Drug Disc. Cruz LJ, Luo S, Gopalakrishnakone P (Eds.). Dordrecht: Springer Netherlands. doi: 10.1007/978-94-007-6726-3_25-1.
60. Santibanez-Lopez CE, Possani LD (2015) Overview of the knottin scorpion toxin-like peptides in scorpion venoms: insights on their classification and evolution. Toxicon 107: 317-326. doi: 10.1016/j.toxicon.2015.06.029.
61. Theerthagiri G, Eisenhardt N, Schwarz H, Antonin W (2010) The nucleoporin Nup188 controls passage of membrane proteins across the nuclear pore complex. J Cell Biol 189: 1129-1142. doi: 10.1083/jcb.200912045.
62. Teichmann SA, Chothia C (2000) Immunoglobulin superfamily proteins in Caenorhabditis elegans. J Mol Biol 296: 1367-1383. doi: 10.1006/jmbi.1999.3497.
63. Fleurie A, Lesterlin C, Manuse S, Zhao C, Cluzel C, Lavergne JP, Franz-Wachtel M, Macek B, Combet C, Kuru E, VanNieuwenhze MS, Brun YV, Sherratt D, Grangeasse C (2014) MapZ marks the division sites and positions FtsZ rings in Streptococcus pneumoniae. Nature 516: 259-262. doi: 10.1038/nature13966.
64. Wang X, Li W, Zhao D, Liu B, Shi Y, Chen B, Yang H, Guo P, Geng X, Shang Z, Peden E, Kage-Nakadai E, Mitani S, Xue D (2010) Caenorhabditis elegans transthyretin-like protein TTR-52 mediates recognition of apoptotic cells by the CED-1 phagocyte receptor. Nat Cell Biol 12: 655-664. doi: 10.1038/ncb2068.
65. Eisenhaber B, Sinha S, Wong WC, Eisenhaber F (2018) Function of a membrane-embedded domain evolutionarily multiplied in the GPI lipid anchor pathway proteins PIG-B, PIG-M, PIG-U, PIG-W, PIG-V, and PIG-Z. Cell Cycle 17: 874-880. doi: 10.1080/15384101.2018.1456294.
66. Luo M, Green TJ, Zhang X, Tsao J, Qiu S (2007) Conserved characteristics of the rhabdovirus nucleoprotein. Virus Res 129: 246-251. doi: 10.1016/j.virusres.2007.07.011.
67. Shey RA, Ghogomu SM, Esoh KK, Nebangwa ND, Shintouo CM, Nongley NF, Asa BF, Ngale FN, Vanhamme L, Souopgui, J (2019) In-silico design of a multi-epitope vaccine candidate against onchocerciasis and related filarial diseases. Sci Rep 9: 4409. doi: 10.1038/s41598-019-40833-x.
68. Saha S, Vashishtha S, Kundu B, Ghosh M (2022) In-silico design of an immunoinformatics based multi-epitope vaccine against Leishmania donovani. BMC Bioinform 23: 319. doi: 10.1186/s12859-022-04816-6.
69. Atapour A, Vosough P, Jafari S, Sarab GA (2022) A multi-epitope vaccine designed against blood-stage of malaria: an immunoinformatic and structural approach. Sci Rep 12: 11683. doi: 10.1038/s41598-022-15956-3.
70. Aguttu C, Okech BA, Mukisa A, Lubega GW (2021) Screening and characterization of hypothetical proteins of Plasmodium falciparum as novel vaccine candidates in the fight against malaria using reverse vaccinology. J Genet Eng Biotechnol 19: 103. doi: 10.1186/s43141-021-00199-y.

71. Khan MAA, Ami JQ, Faisal K, Chowdhury R, Ghosh P, Hossain F, Wahed AAE, Mondal D (2020) An immunoinformatic approach driven by experimental proteomics: in silico design of a subunit candidate vaccine targeting secretory proteins of Leishmania donovani amastigotes. Parasit Vectors 13: 196. doi: 10.1186/s13071-020-04064-8.
72. Mukherjee A, Dandapat P, Haque MZ, Mandal S, Jana PS, Samanta S, Pal S, Das AK, Nanda PK, Bandopadhyay S, Guha C (2023) Computational analysis of hypothetical proteins from Mycobacterium orygis identifies proteins with therapeutic and diagnostic potentials. Anim Gene 29: 200154. doi: 10.1016/j.angen.2023.200154
73. Kant R, Khan MS, Chopra M, Saluja D (2024) Artificial intelligence-driven reverse vaccinology for Neisseria gonorrhoeae vaccine: Prioritizing epitope-based candidates. Front Mol Biosci 11: 1442158. doi: 10.3389/fmolb.2024.1442158.