Bioinformatic analysis of transcriptome data: application to helminth parasites

by

Ranjeeta Menon
Master of Science (Bioinformatics)
Annamalai University, India

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

Department of Chemistry and Biomolecular Sciences
Macquarie University
Sydney, Australia

April 2012

IN MEMORY OF MY FATHER

LATE MR. K. S. U. MENON

AND

DEDICATED TO MY FAMILY

MRS. BABY MENON (MOTHER), MR. RAJESH MENON (BROTHER), MR. MANOJ MENON (SPOUSE) AND MS. AKSHITA MENON (DAUGHTER)

DECLARATION

This thesis contains original work performed by me. A few aspects of this work have been carried out with help from collaborating researchers; these people have been acknowledged and their contributions recognised in the ackowledgement section with details of their assistance. This thesis contains no material that has been accepted for the award of any higher degree or diploma at any University or Institution and to the best of my knowledge, contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Ranjeeta Menon

April 2012

ACKNOWLEDGEMENTS

It is my pleasure to thank the following people who made this thesis possible through their continuous encouragement, prayers and support:

Professional

- My supervisor, Prof. Shoba Ranganathan, for her constant moral support and invaluable suggestions during this work. I thank her for giving me an opportunity to be a part of her group. I am very grateful for her patience, motivation, enthusiasm and intellectual support.

- Prof. Robin Gasser, my co-supervisor for the support throughout my PhD tenure and for the motivation. I also thank him for annotating my manuscripts with constructive criticism and for his valuable suggestions throughout my PhD.

- Prof. John Dalton and Dr. Mark W. Robinson, for giving an opportunity to collaborate in *Fasciola hepatica* project.

- Dr. Makedonka Mitreva, for giving an opportunity to collaborate in *Teladorsagia circumcincta* project.

- Prof. Alan Christoffels for his guidance at the starting of my carrier in Bioinformatics field at the National University of Singapore.

- Dr. Lesheng Kong for giving support and motivation during my PhD.

- Ms. Catherine Wong, Ms. Maria Hyland, Ms. Jane Yang, Dr. Chris McRae, Mr. Michael Baxter and Mr. Doan Lee, for administrative and IT support.

- Macquarie University, for the award of MQRES research scholarship for pursuing a PhD.

- The Department of Chemistry and Biomolecular Sciences, Faculty of Science and the Higher Degree Research Office at Macquarie University, Sydney, for having provided me with all the facilities for the successful completion of this research project.

- My colleagues, Mrs. Elsa Chacko, Dr. Javed Mohammed Khan, Mr. Gaurav Kumar, Mr. Gagan Garg, Dr. Varun Khanna, Dr. Jitendra Gaikwad, Mr. Mohammad Islam and Dr. Shivshanker Nagaraj.

Personal

- I am forever indebted to my husband, Manoj Menon whose endless patience, motivation, financial support and encouragement made it possible for me to reach this important milestone in my research career.

- I would like to thank my dear baby Akshita Menon for giving most wonderful times at her tender age of life that kept me going through the completion of my PhD.

- My heartiest thanks to my parents- in-laws, for their kindness and affection.

- Thanks also to my sister-in-law and my brother-in-law and family for their support.

- I am deeply thankful to my friends for their good company and for their moral support.

- Last, but by no means the least, my father, whose sudden demise in 2007 soon after my selection to pursue PhD motivated me throughout my PhD, his memory and love has been the guiding light and an intense source of inspiration right from the day he departed.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

Adl	Adult lethal
cDNA	complementary DNA
DNA	Deoxy ribonucleic acid
Emb	Embryonic lethal
ES	Excretory or secretory
EST	Expressed Sequence Tag
ESPs	Excretory/Secretory proteins
GWAS	Genome-wide association study
Let	Larval lethal
Lva	Larval arrest
mRNA	Messenger RNA
ORF	Open reading frames
RNA	Ribonucleic acid
RNAi	RNA Interference
NGS	Next-generation sequencing
SNP	Single nucleotide polymorphism
Ste	Maternal sterile
Stp	Sterile progeny
WGS	Whole genome sequencing

LIST OF FIGURES

LIST OF TABLES

LIST OF PUBLICATIONS INCLUDED IN THIS THESIS

The following publications are presented in their published form in this thesis and are referred to from this point onwards as listed in respective sections of the thesis, with my contributions to each paper:

1. Ranganathan S, **Menon R**, Gasser RB: Advanced in silico analysis of expressed sequence tag (EST) data for parasitic nematodes of major socio-economic importance--fundamental insights toward biotechnological outcomes. *Biotechnol Adv* 2009, 27(4):439-448.

   Contributions to: (i) concept: 40%; (ii) data gathering: 40%, (iii) data analysis: 60%, and (iv) writing: 50%.

2. Robinson MW, **Menon R**, Donnelly SM, Dalton JP, Ranganathan S: An integrated transcriptomics and proteomics analysis of the secretome of the helminth pathogen *Fasciola hepatica*: proteins associated with invasion and infection of the mammalian host. *Mol Cell Proteomics* 2009, 8(8):1891-1907.

   Contributions to: (i) concept: 40%; (ii) data gathering: 50%; (iii) data analysis: 50%; and (iv) writing: 45%.

3. **Menon R**, Mitreva M, Gasser R, Ranganathan S: An analysis of the transcriptome of *Teladorsagia circumcincta*: its biological and biotechnological implications. (Accepted – BMC Genomics).

   Contributions to: (i) concept: 50%; (ii) data gathering: 50%; (iii) data analysis: 100%; and (iv) writing: 90%.

4. **Menon R**, Garg G, Ranganathan S: TranSeqAnnotator for large-scale bioinformatic analysis of transcriptomic data (Accepted – BMC Bioinformatics)

   Contributions to: (i) concept: 80%; (ii) data gathering: 75%; (iii) data analysis: 80%; and (iv) writing: RM 90%.

5. **Menon R**, Ranganathan S: Comparison of adult and fourth larval stage transcriptomic data: *Teladorsagia circumcincta* (Under Preparation)

   Contributions to: (i) concept: 80%; (ii) data gathering: 100%; (iii) data analysis: 100%; and (iv) writing: 90%.

ABSTRACT

Parasitic nematodes of humans and other animals cause diseases of major global socio-economic importance, collectively known as neglected tropical diseases. These organisms are able to overcome the sophisticated host immune response mechanisms to colonize, mature and reproduce within the host. An in-depth understanding of parasite genomes, host-parasite relationships, the molecular biology of parasites and their functional annotation can help identify therapeutic molecular targets in helminths, from the discovery of novel genes for parasite control with minimum host side effects. With only a few nematode genomes completely sequenced, analysis is carried out with transcriptomic data, which requires a number of computational methods for their pre-processing, clustering, assembly and annotation to yield biologically relevant information. This thesis highlights improved bioinformatics approaches to analyse transcriptomic data from Expressed Sequence Tags (ESTs), and their application to parasitic nematodes. I first conducted a comprehensive review of the steps involved in transcriptome data analysis for the development of new semi-automated bioinformatic pipelines and their application to parasitic helminths. With the advent of Next-Generation sequencing technologies, my focus was to incorporate the assembly and annotation of short reads, and to concentrate on identifying molecules, especially excretory/secretory proteins (ESPs) involved in key biological processes or pathways that might serve as targets for new drugs or vaccines.

I carried out a preliminary analysis on *Fasciola hepatica*, a parasitic flatworm that causes the disease, fascioliasis, and also infects the liver of various mammals, including humans, leading to liver cancer. By integrating transcriptomic data with proteomic analysis emphasizing on proteases, I have been able to understand the complexities involved in the ability of a developing parasite to sustain itself within the mammalian host. The analysis revealed that a number of non-classically secreted proteins were identified by proteomics but not by bioinformatics, to be addressed in the design of a new analysis pipeline.

To benchmark current bioinformatics tools for transcriptome analysis for the new analysis pipeline, I then carried out large-scale analysis of the adult stage of *Teladorsagia circumcincta* (407357 raw ESTs), a parasitic nematode of sheep and goats.

Based on the benchmarking results, a robust transcriptome analysis pipeline (TranSeqAnnotator) has been developed, with contig generation from ESTs and short

reads, updated pathway analysis, non-classically secreted protein identification and extensive annotation. The pipeline accepts ESTs, quality values, protein sequences and short reads as input and provides as output, assembled contigs and singletons and their annotations including Gene Ontologies, secretory proteins, mapping to protein domains, motifs, metabolic pathways and interaction databases. ESPs are predicted by a combination of computational approaches to effectively identity proteins secreted by classical and non-classical pathways. The pipeline is available as web service and can be downloaded for local installation.

As part of evaluating the pipeline, I carried out an in-depth analysis of transcriptome from a nematode parasite, *Ascaris lumbricoides*. Results from the pipeline for the analysis of short read sequences of the fourth larval stage of *Teladorsagia circumcincta* (507,124 sequences) are compared to the adult stage EST annotations, as a large-scale application of TranSeqAnnotator.

Chapter 1: Introduction and literature survey

1.1 Overview

The study of genes and gene products helps in the understanding of the basic principles of biology at molecular and cellular levels. This ease the understanding how genes and their products interact to play a vital role that drives cellular mechanisms. Genes contain the instructions which code for a particular protein and are considered as the working subunits of DNA, which carries the instruction while the gene product, either RNA or protein are the biochemical material obtained as a result of expression of a gene. DNA are transcribed into corresponding RNA, referred as transcripts for carrying the instructions. The collections of these transcripts in a cell are termed transcriptome. Hence, a comprehensive analysis of transcriptome is considered important to comprehend an organism along with the genome and proteome analysis. In this thesis, new and updated bioinformatic approaches and tools are applied for the study of helminth transcriptome which helped to identify several novel and known genes with therapeutic experimental validations. At the start, a brief history of helminths, with special reference to parasitic nematode and their life cycle is presented. Following this, immune response and anti-parasite vaccine with the highlight in excretory/secretory products are described. The transcriptome of parasitic helminths/nematodes with the details on EST data analysis, their generation and annotation focussing on their potential role as the therapeutic targets for parasite control is provided. The framework of EST analysis pipelines and the current bioinformatic tools paves the way for the specific objectives for the thesis.

1.2 Brief history of Helminths

The worm-like parasitic helminths, a division of eukaryotic parasites are known to infect more than two billion people worldwide [1]. These complex multicellular organisms live and feed off living hosts to sustain themselves and perturb their host's nutrient absorption thus causing weakness and disease. The study of parasitic worms and their effect on their hosts is called Helminthology [2]. Contaminated water, soil, or food play a vital role in the spreading of these organisms based on the parasite species with more than one-third of human population considered to carry these organisms. The helminths prefer to restrict their selection of host with a preference to live in different locations of their host like intestine, liver, lumen, bile ducts, lungs or blood stream and they have mechanisms, obtained as a result of coevolution, with an emphasis on host-parasite relationship that

prevent their eradication by cellular immune responses by controlling the host's immune system. Recent studies of worm genomes reveal that the factors and receptors of worms show homology to molecules of the human immune system [1]. Worms have fully developed organs and complex tissues compared to bacteria and viruses [1, 3]. Infected individuals bare three life-cycle stages of heminth; infective larvae, adult worms and transmission stage parasites (eggs, immature larvae or microfilariae) [4]. These life-cycle stages are proved to be molecularly different both in proteomic [5] and DNA microarray [6] studies, and are believed to induce stage-specific immune responses.

1.3 Types of parasitic worms or helminths

Parasites that live inside the host body are called endoparasites, being the ectoparasites attached themselves to the outer parts of the host. Helminths are classified as Flukes (Trematodes), Tapeworms (Cestodes), Roundworms (Nematodes; including pinworms and hookworms) based on the external and internal morphology of egg, larval, and adult stages [2]. Worms that inhabit the interior of humans are the nematodes (roundworms), the platyhelminths (flatworms) which have distinct evolutionary histories and the annelida, which are segmented worms, with bodies divided into segments, or rings Figure 1.1 [7].

There are two classes of human flatworms called trematodes (flukes) and cestodes. Helminths fail to sustain human migration due to the genetic differences in the human host which have no association with human illness (e.g., *Trichuris suis*) [1]. A few other animal helminths cause pathology in humans like *Toxocara canis* (dog ascarid), *Dirofiliaria immitis* (dog heartworm), and *Trichinella spiralis*.

1.3.1 Flatworms or platyhelminthes

Platyhelminthes are flattened from the dorsal to ventral surfaces and include the classes Trematoda and Cestoda. Trematodes or flukes are leaf-shaped flatworms in their adult stage and are hermaphroditic other than blood flukes, which are bisexual. The bodies have ventral and oral suckers and animal is held by the sucker. Cestodes or tapeworms inhabit the intestinal lumen and are elongated, segmented, hermaphrodites in their adult stage.

Trematodes live mainly in the venous system (e.g., schistosome species), biliary system (e.g., *Clonorchis*), intestine (e.g. *Echinostoma*), gut (e.g., *Fasciolopsis*), or airway (e.g., *Paragonimus*). Cestodes include intestinal tapeworms like *Diphyllobothrium latum* (fish tapeworm), *Taenia saginata* (beef tapeworm), and *Taenia solium* (pig tapeworm).

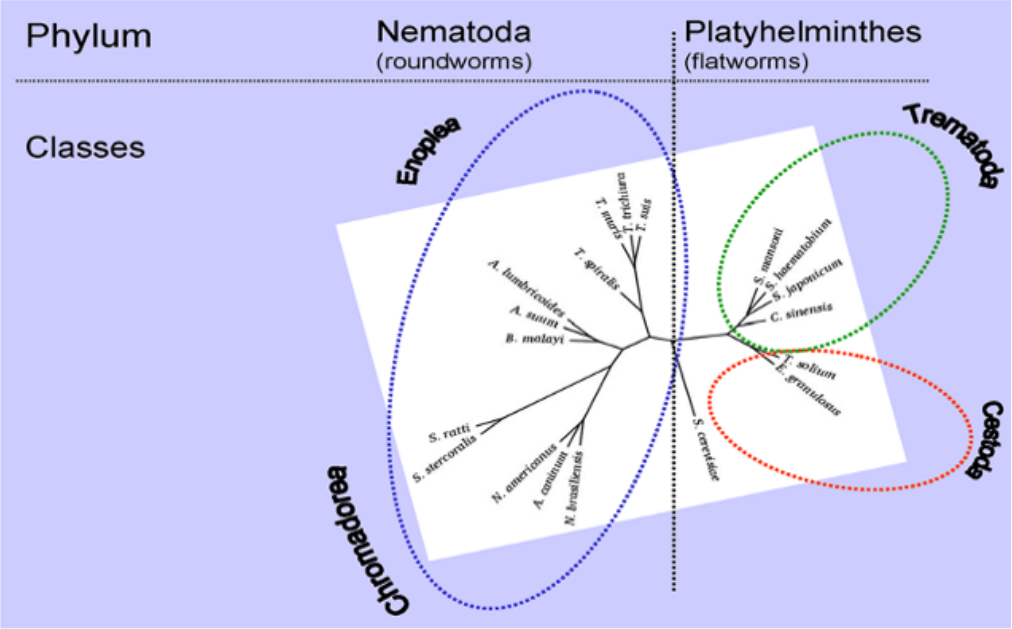


Figure 1.1: The classification of helminths – nematodes and platyhelminths
Adapted from Brindley et al. [7], with species relevant to this thesis added in blue.

1.3.2 Roundworms (nematodes)

Nematodes or roundworms (Phylum Nematoda from Greek (nema): "thread" +-ode "like") known to be the most abundant multicellular animals on earth, inhabit in the intestinal and extraintestinal sites and are bisexual, cylindrical worms in their adult and larval stages. They are known to be diverse in morphology, size (adults from less than a millimetre to over 6 metres), life cycles (from parthenogens to complex cycles of alternating sexual strategies), and ecology (including parasites of almost all other large multicellular organisms, plant and animal) [8]. Over a million nematode species on earth have been reported in recent studies though only 25,000 species have been described [9] as a result of their ability to adapt, small size, resistant cuticle, and simple body plan [10]. The phylum nematoda has a large variation in genome size ranging from 50-250Mb [11], due to high rate of large, spontaneous deletions [12]. They seem to have a well-defined digestive, nervous, excretory and reproductive systems, but lack a discrete circulatory or/and the respiratory system. Based on phylogenetic analysis of 53 small subunit ribosomal DNA sequences from a wide range of nematodes, evolutionary classifications of the nematodes are proposed (Table 1.1) [13]. The view of phylum nematoda shows three major branches, the Enoplia, Dorylaimia, and Chromadoria [8]. On the whole, there are five proposed clades: Clade I - Dorylaimia, Clade II - Enoplia, Clade III - Spirurina; Clade IV - Tylenchina; and Clade V – Rhabditina (Figure 1.2).

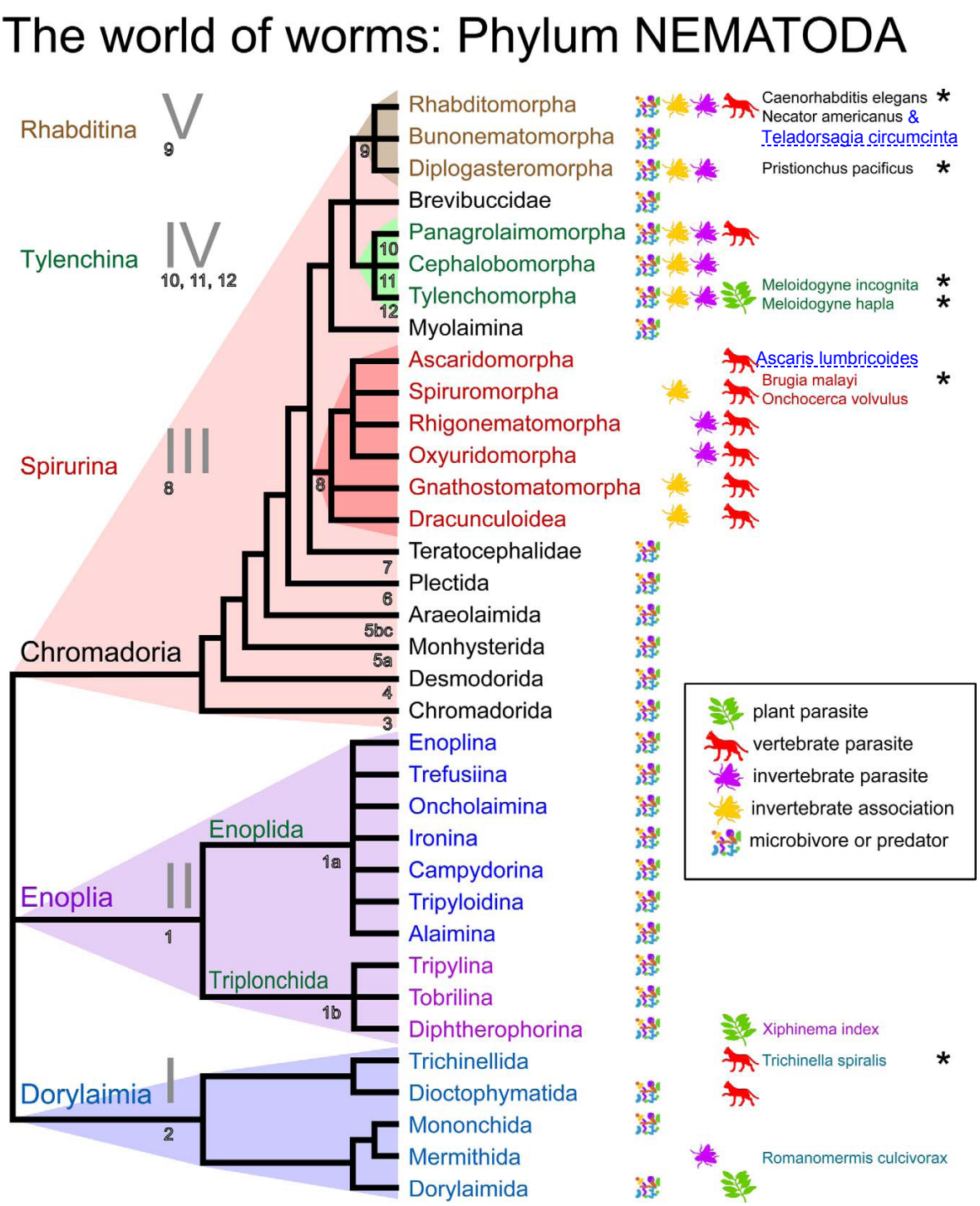


Figure 1.2: The classification of the phylum Nematoda.

Adapted from Blaxter's recent review [8], with species relevant to this thesis added in blue.

The diversity of nematodes is approached through comparative genomics, by sequencing of expressed genes of the target species, and transcriptome approach [8]. Studies report almost 60 transcriptome datasets generated for free-living, animal-parasitic, and plant-parasitic species [14]. The best known nematode is *Caenorhabditis elegans*, subject of the studies which granted the Nobel Prize for Physiology and Medicine in 2002 to Sydney

Brenner along with H. Robert Horvitz and John E. Sulston for their discoveries concerning "genetic regulation of organ development and programmed cell death" in that experimental model system. *C. elegans* is discussed in depth, later in this chapter.

Table 1.1 Taxonomy of the phylum Nematoda. Adapted from Blaxter's review [13]

Scientific classification
Lineage(full)
cellular organisms;
Eukaryota;
Fungi/Metazoa group;
Metazoa;
Eumetazoa;
Bilateria;
Pseudocoelomata;
Nematoda

Figure 1.3: *Trichuris trichiura* (roundworm)

This human parasitic roundworm causes trichuriasis when infecting human large intestine (http://www.latech.edu/ans/faculty-staff/liberatos-james-d/parasite-pictures.shtml).

1.4 Brief introduction to parasitic nematodes

Parasitic nematodes impose burden for significant public health and economy worldwide as they cause diseases and are known to infect humans, other animals and plants. Animal

and plant parasites can cause massive production or economic losses to farmers as well as to animal and plant industries [15], and are responsible for a range of "neglected tropical diseases", such as ancylostomatosis, necatoriasis, lymphatic filariasis, onchocerciasis, ascariasis and strongyloidiasis in humans [16-18]. Development of diagnostic tests and/or safe anti-parasitic compounds could be achieved with a better understanding of parasite genomes, host–parasite relationships and the molecular biology of parasites along with the functional annotation of parasite genomic sequence. An example of human parasitic nematode *Trichuris trichiura* is shown in Figure 1.3.

1.5 Life cycle of parasitic nematodes and their control

The parasitic nematode life cycle depends on the type of the host they feed on and the involvement of vector for transmission and also, differs in the presence or absence of free-living stages and intermediate hosts. Parasitic nematodes generally invade a new host by penetrating the host skin or mucous membrane by larvae or by the ingestion of mature infectious eggs or larvae. They tend to have multiple stages and alternate between host and regions in their host's body. In general there are six stages; an egg stage, four larval stages (L1, L2, L3, and L4 - each of which is followed by a molt or shedding of the skin, to facilitate growth) and an adult stage. The third larval stage (L3) is usually considered to be the infectious stage with respect to their host. Soil-transmitted intestinal parasitic nematodes adopt a direct life cycle, where the eggs or larvae exit the host in faecal material and then continue to develop in the soil, where they become infective. Parasitic nematode life cycles vary from free-living nematodes, inhabiting different hosts as well as diverse host organs, and their breeding system can be parthenogenetic or hermaphroditic [19]. They complete a new developmental stage inside each host showing different levels of host specificity at each stage. A summarized (generic) life cycle of parasitic nematodes (females and males) is illustrated in Figure 1.4.

Effector response directed against the parasite can be influenced with the life-cycle variation and antigen expression [4]. Life cycle patterns in general are phylogenetically conserved, and the variation in the type of hosts chosen by the parasite suggest the mechanisms that influence biogeographic variables [20]. Infections with parasitic worms are treated with anthelmintic drugs to both flatworms, such as flukes and tapeworms; and roundworms, such as nematodes. An alphabetic list of major parasitic nematodes of humans, other animals and plants are given in Table 1.2.

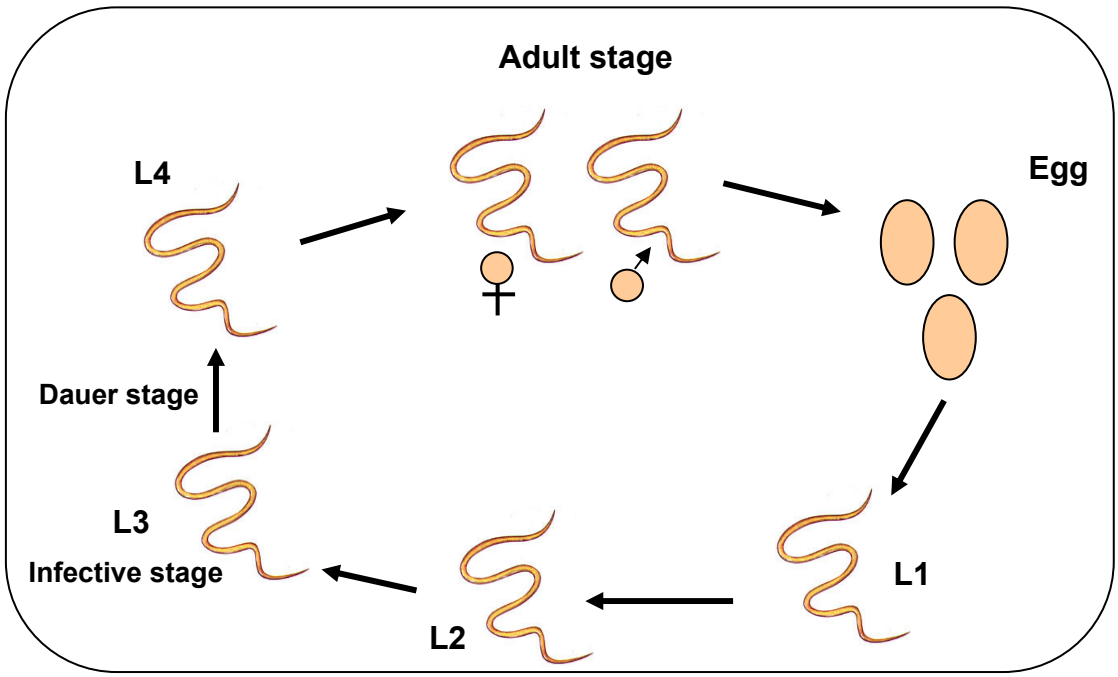


Figure 1.4: A generic life cycle for parasitic nematodes.

Table: 1.2 Nematodes parasitic in humans, other animals and plants (listed alphabetically for each host group) and their principal (definitive) hosts.

No.	Nematode Parasite	Common name or description	Principal [definitive] host or host group	Disease or common name of disease
Animal host				
1	*Ancylostoma duodenale*	Old World hookworm	Humans, cats and dogs	Hookworm disease
2	*Ancylostoma ceylanicum*	Hookworm	Human, dog and cat	Hookworm disease
3	*Ascaris lumbricoides*	Common roundworm	Human	Ascariasis
4	*Brugia malayi*	Filarial worm	Human	Lymphatic filariasis (elephantiasis)
5	*Necator americanus*	Hookworm	Human	Hookworm disease

No.	Nematode Parasite	Common name or description	Principal [definitive] host or host group	Disease or common name of disease
Animal host				
6	*Onchocerca volvulus*	NA	Human	Onchocerciasis (river blindness)
7	*Strongyloides stercoralis*	NA	Human	Strongyloidiasis
8	*Trichinella spiralis*	Trichina	Human, rats, canids, pigs	Trichinellosis
9	*Trichuris trichiura*	Whipworm	Human	Trichuriasis
10	*Wuchereria bancrofti*	Filarial worm	Human	Lymphatic filariasis (elephantiasis)
11	*Ancylostoma caninum*	Canine hookworm	Dog	Hookworm disease
12	*Ascaris suum*	Large common roundworm of pigs	Pig	Ascariasis
13	*Dirofilaria immitis*	Canine heartworm	Dog	Heartworm disease
14	*Dictyocaulus viviparus*	Lungworm	Cattle	Dictyocaulosis, parasitic bronchitis, husk
15	*Haemonchus contortus*	Barber's pole worm	Small ruminants (sheep, goat)	Haemonchosis
16	*Litomosoides sigmodontis*	Filarial worm	Rodent	Rodent filariasis
17	*Nippostrongylus brasiliensis*	Intestinal parasite of rats	Rodent	Nippostrongylosis
18	*Ostertagia ostertagi*	Brown stomach worm of cattle	Cattle and other bovids	Ostertagiasis
19	*Oesophagostomum dentatum*	Nodule worm of pigs	Pig	Oesophagostomiasis

No.	Nematode Parasite	Common name or description	Principal [definitive] host or host group	Disease or common name of disease
Animal host				
20	*Onchocerca ochengi*	NA	Cattle	Onchocerciasis
21	*Parastrongyloides trichosuri*	Intestinal parasite of Australian brush-tail(ed) possum	Possum	-
22	*Strongyloides ratti*	NA	Rat	Strongyloidiasis
23	*Teladorsagia circumcincta*	Brown stomach worm	Small ruminants (sheep, goat)	Ostertagiasis
24	*Toxocara canis*	Common roundworm of dogs	Dog	Toxocariasis (larval toxocariasis in intermediate host)
25	*Trichuris muris*	Murine whipworm	Rodent	Intestinal diseases
26	*Trichuris vulpis*	Canine whipworm	Canids	Trichuriasis
27	*Trichostrongylus vitrinus*	Black scour worm	Sheep and goat	Trichostrongylosis
Plant host				
1	*Globodera pallida*	White cyst nematode, Potato cyst nematode	Tomato, eggplant	Potato cyst disease
2	*Globodera rostochiensis*	Potato cyst nematode	Potato, tomato, eggplant	Potato cyst disease, yellowing or wilting of foliage
3	*Heterodera glycines*	Soybean cyst nematode	Soybean	Soybean cyst disease, yellow dwarf

No.	Nematode Parasite	Common name or description	Principal [definitive] host or host group	Disease or common name of disease
Plant host				
4	*Heterodera schachtii*	Sugar beet cyst nematode	Sugar beet, cabbage, cauliflower, brussel sprouts, mustard, radish, spinach, chard	Sugar beat cyst disease, stunted, wilt
5	*Meloidogyne arenaria*	Peanut root-knot nematode	Peanut, vegetables, grasses, fruit, ornamentals and tobacco	Root-knot disease, large galls on roots, pegs, pods and runners
6	*Meloidogyne chitwoodi*	Columbia root-knot nematode	Various plants, including potato, barley, wheat and alfalfa	Root-knot disease, nematode-induced blemish in tubers, stunting
7	*Meloidogyne hapla*	Northern root-knot nematode	>500 plant hosts, including clover, vegetables, alfalfa and ornamentals	root-knot disease, galls on roots; poor growth; shortened lifespan of the vine
8	*Meloidogyne incognita*	Cotton Root-knot Nematode	Cotton, tobacco, peanut and fibre crops	Root-knot disease; causes wilting of the infected plant leading to death, Blackshank of tobacco
9	*Meloidogyne javanica*	Root Knot Nematode	>770 species includ-ing peanut, sugar-cane and fibre crops	Root-knot disease
10	*Meloidogyne paranaensis*	Coffee root-knot nematode	Coffee	Coffee root-knot disease resulting in decline and dieback

No.	Nematode Parasite	Common name or description	Principal [definitive] host or host group	Disease or common name of disease
Plant host				
11	*Pratylenchus penetrans*	lesion nematode; meadow nematode	Apple, cherry, peach	Necrotic lesions and chlorosis
12	*Pratylenchus vulnus*	Walnut meadow nematode or Walnut root-lesion nematode	Apple, cherry, peach	Invades the cortex of roots, tubers, and results in necrotic lesions
13	*Radopholus similis*	Burrowing nematode	Banana, citrus	Migratory endoparasite; causes spreading decline
14	*Xiphinema index*	Dagger nematode	Grape	Root stunting and tip galling

1.6 Anthelmintics

The infections with parasitic worms are treated with anthelmintic drugs. They are known to form the basis for modern nematode control, which are classified on the basis of similar chemical structure and mode of action. The most important, extensively used and commercially available anthelmentics are classified into three chemical families; the benzimidazoles (BZs); levamisole and other imidazothiazoles (LEV); and the macrocyclic lactones (MLs), which includes piperazine, benzimidazoles, levamisole, pyrantel and morantel, paraherquamide, ivermectin (macrocylic lactones and milbemycins), emodepside (cyclodepsipeptides, PF1022A) [21]. Another class of drugs known to be recently discovered, the amino-acetonitrile derivatives (AADs) have been launched onto the small ruminant market in New Zealand and in the UK [22]. Immune heterogeneity is considered to arise from both the environment in which the immune response develops and from hosts and parasites in isolation. This further highlights the impact of anti-helminthic treatment [23]. Besides the economic effect on health system, anthelmintic resistance has been reported in livestock and is likely that this event occurs in parasites of humans and plants [24, 25]. A profound knowledge of host-parasite interactions and understanding of

candidate genes would help in developing intervention stratergies, including vaccine design.

1.6.1 Anthelmintic resistance

Anthelmintic resistance is defined as the ability of parasites to survive the drug dose that could kill them. Its nature is heritable and non-reversible [25]. A parasite will normaly be resistant to all anthelmintics within a class if it has been resistant to one anthelmintic in that drug class which is termed as side resistance [26]. Drug resistance has grown rapidly to the all three classes of drugs and is a major threat to livestock production in many parts of the world [27, 28]. The anthelmintic resistance development is facilitated by the large population size and inherent high genetic inconsistency, and compounded by the movement of their host species [29].

1.7 Immune system

Helminth parasites show varied challenge to immune system and their ability to modulate the immune system describes their permanence in the mammalian host [23, 30]. The immune system is generally divided into the innate and adaptive immune system. The innate immune system uses the pattern recognition of molecules foreign to the host to defend the host from infection in imprecise method with information of cells and mechanisms that defend the host.

The adaptive immune system employs T and B cells, recognizes specific antigens and confers long-lasting protective immunity. Cytokines that drive, modulate or suppress immune responses are the molecules made by cells of immune system which include effector T cells (Th1, Th2, Th17), regulatory T cells (Treg, Tr1, Th3), dendritic cells, and alternatively activated macrophages among others. Intestinal T-cell responsiveness to various mitogens and antigens are suppressed with worm infection [1].

1.7.1 Host immune response

Immune response developed during chronic infections often lead to pathologic changes which become the primary cause of disease. The cytotoxic effects of immune response is evaded with the newly developed mechanisms as part of long-term survival of helminth parasites within mammalian hosts [31]. Current chances of development of immunological disorders, such as autoimmunity and allergy, with the helminth infections are other major issues, with the highlight of the association of helminth infections and immune

responsiveness to allergens and/or allergic symptoms are considered as interesting areas of study [32]. The study of immune response paves way for the development of anti-parasite vaccines [33].

1.8 Anti-parasite vaccines

The control of parasitic infections with vaccines is another focus of study. Vaccines are considered environmentally friendly, being non toxic to dung and marine flora and fauna [34-36]. They are preferred over anthelmintics as they leave no chemical residues over food products. The commercially available anti-parasite vaccines include the vaccines for parasitic protozoa and use of live attenuated parasites [34, 37], and Bovilis Lungworm (Huskvac), formerly known as 30 Dictol, the nematode parasite vaccine for vaccinating calves against the lungworm *Dictyocaulus viviparous*. Presently, the nematode vaccine work concentrates on *H. contortus*. In the case of sheeps, exposing them to antigens like excreted or secreted by the parasite products (ESP) could help to determine the immunity that develops in sheep due to nematode infection, and can be studied with natural antigen candidates. Current vaccine discovery is based on the use of "hidden antigens" that reside on the gut surface of the parasites [38] and are known to be effective against blood-feeding parasites, the cattle tick, *Boophilus microplus* and in *H. contortus* [37, 39]. Vaccines should have a wide range of activities to contend against anthelminthics and also be able to give protection against at least the three major parasites of sheep, namely *H. contortus*, *T. circumcincta* and *Trichostrongylus* spp. [34].

1.9 Excretory/Secretory Products (ESP)

Parasitic helminths excrete or secrete a variety of molecules into their mammalian hosts. ESP represent a fraction of the full genomic complement, and are released by live parasites which can hinder the host immunity from initial recognition to end-stage effector mechanisms and are easily accessible to the immune system of the host as they are released from parasites into their surrounding environment. ES proteins play vital role as mediators and/or regulators of cellular and/or humoral immune responses, and also contribute to immune evasion strategies of the parasites with mechanisms like shedding of surface-bound ligands and cells, alteration of lymphocyte, macrophage and granulocyte functions and modulation of complement and other host inflammatory responses [40-44]. Molecular and genomic studies on large sets of secreted proteins from the plant-parasitic and human/animal-parasitic nematodes help to understand the secretions from the nematodes which are vital in invasion and establishment in the host [45]. Biochemical,

immunological, proteomic, transcriptomic analyses are carried out for all ESP irrespective of their physiological origin to determine the function of several secreted proteins involving cloning, recombinant expression and production of neutralising antibodies. Interactions between parasites and the host immune system is better understood with the combined study of proteomics and genomics [40]. The difficulties involved for a developing parasite to sustain itself within the host could be studied by integrating transcriptomic and proteomic analysis [46]. The N-terminal signal peptide cleavage and the absence of a transmembrane domain are the key factors in the prediction of excreted or secreted proteins. The study of EST datasets from parasitic helminths helps in predicting the putative ES proteins by direct sequence comparisons with known secreted proteins experimentally identified and characterised for a wide range of eukaryotic proteins [47-49]. Some secreted proteins like fibroblast growth factors (i.e. FGF-1 and FGF-2), interleukins (i.e. IL-1) and galectins proved experimentally to be secreted through a 'non-classical' secretory pathway and thus do not possess a N-terminal signal peptide [50, 51].

1.10 Transcriptome of parasitic helminths

Genome made up of deoxyribonucleic acid (DNA), the winding molecule that contains the instructions needed to build and maintain cells includes both genes and the non-coding sequences of the DNA/RNA. DNA is transcribed into corresponding molecules of ribonucleic acid (RNA), referred as transcripts to carry these instructions. The collection of the transcripts in a given cell is named transcriptome. The total protein complement of a genome is termed proteome, while the fraction of the genes expressed as mature mRNA from any genome at a given timepoint is termed transcriptome, which represents a small precentage of genetic code that is transcribed into RNA molecules, less than 5% of human genome [52]. Transcriptome study helps to determine what genes are active at various stages of development. The snapshot of the transcriptome from any cell, tissue or organ is obtained by making libraries of all expressed genes for an organ or developmental stage, a complementary DNA (cDNA) library. From either the 3' or 5' end of each cDNA clone, an adequate number of clones are sequenced resulting in the generation of expressed sequence tags (ESTs). ESTs can be used to produce probes to decide the presence or absence of similar transcripts in other tissues. This gives an outline of many genome projects with generation of new EST datasets and being deposited in public databases.

1.10.1 Expressed Sequence Tags (ESTs)

Expressed sequence tags or ESTs are short, unedited, randomly selected single-pass sequence reads of approximately 200-800 base pairs (bp) derived from complementary DNA (cDNA) libraries, providing a low-cost alternative to whole genome sequencing. ESTs represent a small region or a part of nucleotide sequence from a transcribed protein coding or non-coding messenger RNA (mRNA). ESTs play vital role in gene identification and verification of gene prediction as they represent the expressed region of a genome. So far almost 45 million ESTs have been generated from over 1400 different species of eukaryotes. With the advent of high-throughput sequencing, EST projects are used to either complement existing genome projects or serve as low-cost alternatives for purposes of gene discovery [53]. ESTs were used as the primary source for human gene discovery in early 1990s [54]. A detailed description of EST generation, through cDNA library construction and normalization applied to remove redundancies in EST dataset were later provided which arose exponential growth in the generation and accumulation of EST data in public databases [55].

Earlier, sequencing of organisms with large genomes used to be expensive and whole genome sequencing (WGS) was impractical hence ESTs were generated to complement genome sequencing as an alternative, earning the label of the 'poor man's genome' [56]. Presently ESTs enable gene discovery, complement genome annotation, aid gene structure identification, transcript profiling, guide single nucleotide polymorphism (SNP) characterization, phylogenetics and facilitate proteome analysis [56-58]. The recent release of "next-generation" sequencing techniques which generate millions of sequencing reads of 35-250 bp has revolutionize the depth of transcriptome data mainly for neglected organisms.

ESTs can be generated from a tissue, an organ or a cell of interest and also at different developmental stages. Generation of ESTs from mRNA is described briefly, with the different steps involved summarised in Figure 1.5.

Every gene encoded as DNA is transcribed into mRNA by the process called transcription, which in turn serves as a template for protein synthesis, a process called translation. mRNA are very unstable and transient outside of a cell and cannot be cloned directly. They represent copies from expressed genes, excluding the untranslated regions and the intronic segments present within many genes. Hence these sequences are reverse transcribed to

double-stranded cDNA using a specialized enzyme, the reverse transcriptase, resulting in cDNA, which are cloned to make libraries representing a set of transcribed genes of the original cell, tissue or organism. Later these cDNA clones are sequenced randomly from both directions in a single-pass run, with no validation or full-length sequencing, to obtain 5' and 3' ESTs.

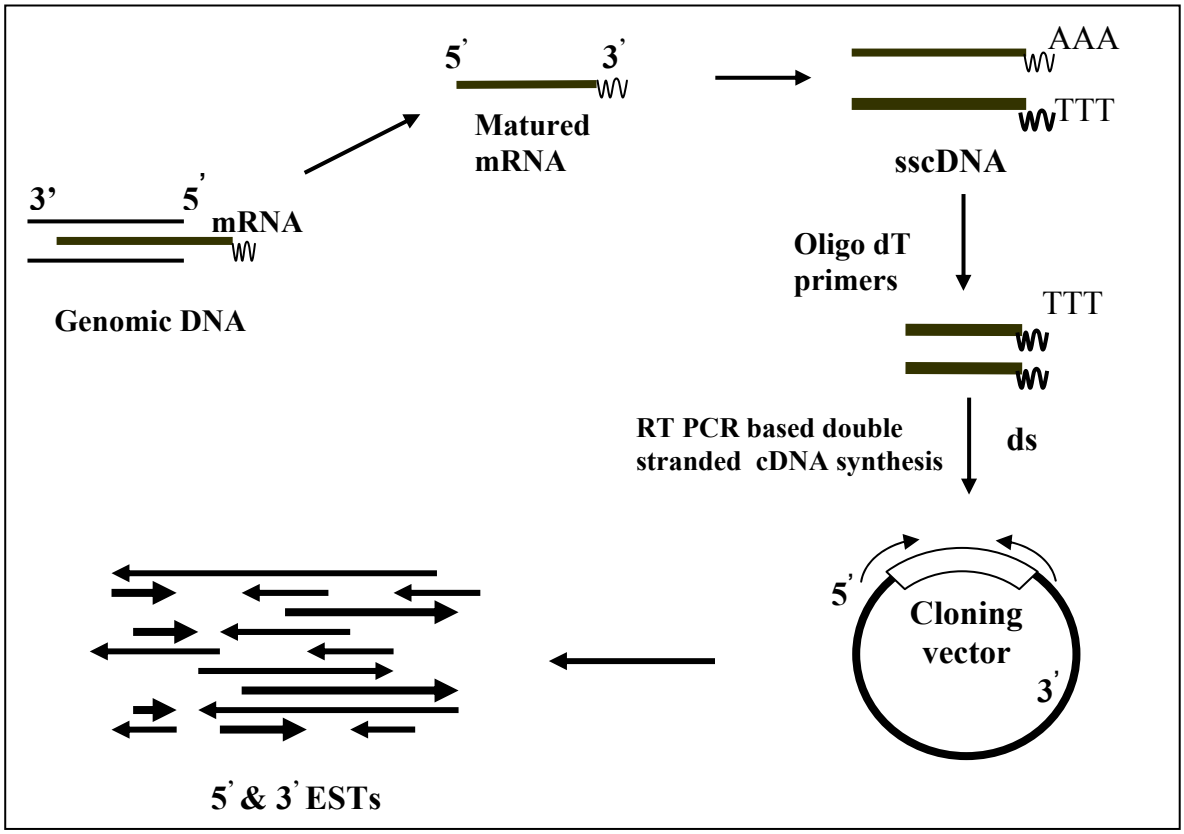


Figure 1.5: Illustration of steps involved in EST generation. Genomic DNA is transcribed to mRNA. The information on mRNA is copied onto cDNA which results in cDNA libraries. 5' and 3' ESTs are generated from such cDNA libraries.

1.10.2 Next Gen Sequencing

Next generation sequencing (NGS) technologies allow massive parallelised sequencing of genomes has substantially reduced the costs of generating large sequence datasets [59, 60]. During DNA synthesis (sequencing-by-synthesis), the entire genome is broken into small pieces followed by ligation of these pieces to designated adapters for random reading and hence called massively parallel sequencing [61]. Currently available NGS platforms include the 454 Life Sciences/Roche ([62]; www.454.com), Solexa/Illumina ([63];mwww.illumina.org) and SOLiD (Supported Oligonucleotide Ligation and Detection) ([64]; www.appliedbiosystems.com). Two other recently introduced technologies are, Polonator G.007 and the Helicos HeliScope platforms NGS, which

generates short sequences when compared to that generated by Sanger sequencing in terms of different base read lengths, different error rates, and different error profiles relative to Sanger sequencing data and to each other. 454/Roche platform uses the sequencing-by-synthesis approach to generate long reads (100-600 bp) and is used for de novo genomic and transcriptomic studies. The next introduced Illumina, formerly called Solexa sequencer generates billions of bases per run. It is considered best for re-sequencing projects, single nucleotide polymorphism (SNP), targeted sequencing, gene transcription studies and has different features compared to the 454 approach [63]. Recently introduced Illumina platform has similar chemistry to the 454 technology. Its 'two-base-calling' technique provides accurate base calling thus, helping the sequencing errors. Application of SOLiD system is similar to Illumina as they generate short read lengths. NGS would help to overcome the drawbacks of the studies conducted with genome-wide association study (GWAS) with the focus on sample size, limitation of arrays for certain genetic variations, and/or heterogeneity in phenotype [65]. Numerous studies have been carried out with the advent of NGS technology mainly in the areas of population genetics, systematic, molecular biology of pathogens including protozoa [66-68], bacteria [69], viruses [70], arthropods [71], nematodes [72, 73] and trematodes [74, 75]. The new genomic technologies helped in the study of genes or genomic regions involved in pathogenesis of human diseases which would lead to the advancement in genetic medicine and help in improvement of human health [76, 77]. The second complete individual genome (of James D. Watson) is marked as the first human genome sequenced using the new NGS technology [78]. The 454 technology was used for de novo sequencing of transcriptomes of trematodes like *Fasciola hepatica*, *Clonorchis sinensis* and *Opisthorchis viverrini* [74, 75]. The short read strategy of NGS has led to many challenges for bioinformatic analysis in data storage and management solution and creating informatic tools for analysis based on the sequence quality scoring, alignment, assembly, and data processing with their functions in the areas of alignment of reads to a reference sequence, *de novo* assembly, reference-based assembly, base-calling and/or genetic variation detection (such as SNV, indels), genome annotation, and utilities for data analysis [79, 80]. Recent studies with bioinformatic analysis and 454 approach of the transcriptome data at different development stages and sex of important parasitic nematodes like *T. colubriformis*, *H. contortus*, *N. americanus* and *O. dentatum* [81-84] are very similar to EST analysis, with specialized programs for clustering and assembly.

1.11 Study of nematode parasites using *Caenorhabditis elegans* data

C. elegans, the most remarkably studied free-living nematode, found in soil and compost heaps was the first multicellular organism for which a complete genome sequence was generated and remains the only metazoan (animal) for which the sequence of every nucleotide (i.e. 100 278 047 nt) has been determined at high confidence levels. The genome of *C. elegans* was estimated as ~97 Mb in size in 1998, with five autosomal and a single sex chromosome [85]. Due to its short lifecycle (~3 days) and easy to culture *in vitro* with simple anatomy with 959 somatic cells in the hermaphrodite and 1031 in the male, *C. elegans* is considered suitable for the study of genetic changes, genome, biology, physiology, biochemistry, as well as the localization and function of molecules [86].

C. elegans genome and proteome are compared to transcript data generated for parasitic nematodes to identify homologues and assign putative functions [87-94]. The *B. malayi* draft genome was annotated using the *C. elegans* genome [95]. The study of gene function in metazoan organisms has led to the detailed information of the functions of ~96% genes in *C. elegans*, with the introduction of the technique of RNAi interference [96-101]. RNA interference (RNAi), a mechanism that inhibits gene expression at the stage of translation or by hindering the transcription of specific genes is well established for most of the *C. elegans* genes. RNAi phenotypes, non-wild-type or loss-of-function *C. elegans* suggest the importance and function(s) of homologous genes in other parasitic nematodes where obligate parasitic life cycle, the lack of an effective *in vitro* culture system and/or an assay for RNAi make high-throughput screening have proved to be impractical [102]. Presently, wormbase [103] contains the detailed and curated information on ~19,000 *C. elegans* genes and association data, including the details with regard to cells and tissues, mutants and their phenotypes, transcription/expression profiles in different developmental stages, genetic and physical maps, information on gene-gene, and protein-protein interactions, SNPs and also, the peer-reviewed literature pertaining to *C. elegans*. Studies reveal that the *C. elegans* genes homologous in parasitic nematodes are considered to be essential for parasite survival and growth [102, 104]. As part of the therapeutic perspective, specific loss-of-function phenotypes such as Adl (adult lethal), Emb (embryonic lethal), Let (larval lethal), Lva (larval arrest), Stp (sterile progeny) and Ste (maternal sterile) are considered very important. Besides, the RNAi technique, the transgenesis of *C. elegans* is used for assessing and proving gene function [105, 106]. Microarray technology to study patterns of gene transcription during development and reproductive phase using *C. elegans* is another aspect of study other than the gene expression and localization [107-109]. The importance

of comparative investigation of biochemical and molecular pathways, linked to development in related nematodes using the free-living nematode, has proved to be important with the studies conducted, finding similarities of the characteristics of *C. elegans* and other parasitic nematodes. Amidst the argument of the advantages and disadvantages of using the *C. elegans* genome as a model for parasitic nematodes, the wealth of information for this organism will continue to hold its significance in future.

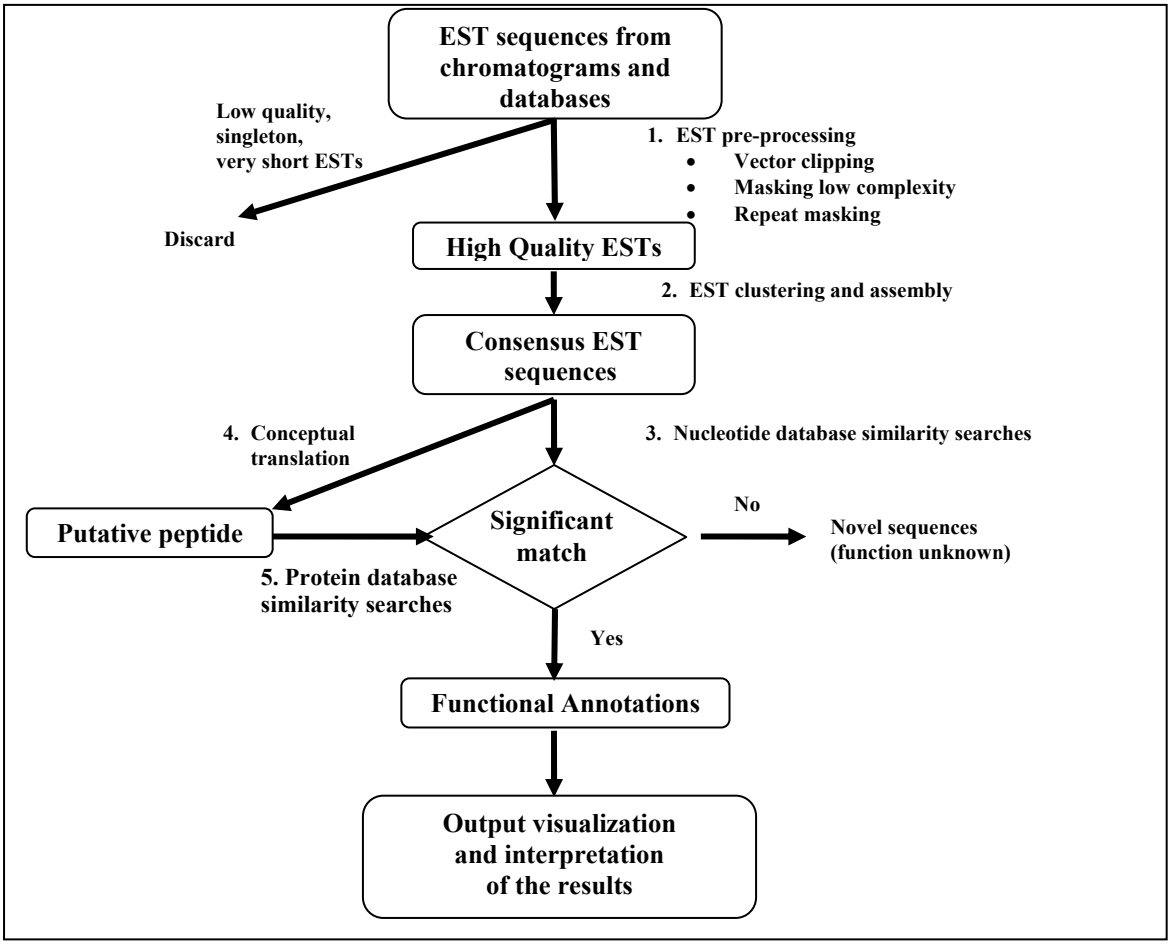


Figure 1.6: Generic steps involved in EST analysis. 1. Raw EST sequences are checked for vector contamination, low complexity and repeated regions, which are excised or masked. Low quality, singleton and very short sequences are also removed. 2. ESTs are then clustered and assembled to generate consensus sequences ('putative transcripts'). 3. DNA database similarity searches are carried out to assign, identify homologues and sign possible function. 4. Putative peptides are obtained by conceptual translation of consensus sequences. 5. Protein database similarity searches are performed to assign putative function(s). The analysis is extended to functional annotation followed by visualization and interpretation of results.

1.12 Overview of EST data analysis

The analysis of sequence data at cDNA and protein level has increased along with the number of ESTs in public databases. The analysis include different steps pre-processing, clustering, assembly and annotation using a myriad of tools to yield biological information. General steps in EST analysis are shown in Figure 1.6. The analysis of NGS data is very similar and therefore has not been explicitly described, to avoid repetition.

The analysis steps are briefly described in the following sections.

1.12.1 EST pre-processing

ESTs are initially screened for sequence repeats, contaminants and/or adaptor sequences [110-114] to generate high-quality ESTs. Informations from non-redundant vector databases like UniVec and EMVEC help to remove vector sequences using locally installed tools like BLAST [115], also avaliable as web interface [116] or Cross_Match. The popular tool for "cleaning" (validation and trimming) EST data, SeqClean has five programs along with the specific vector database for trimming.

Followed by sequence cleaning, the repetitive elements, such as LINEs (Long interspersed elements), SINEs (Short interspersed elements), LTRs (Long terminal repeat) and SSRs (Short simple repeats) are masked to avoid errors in assembly of sequences using tools like RepeatMasker or MaskerAid [117] to screen DNA sequences for low complexity DNA sequences and interspersed repeats. New method, RBR was introduced for library-less method for masking repeats in EST data for new organisms for which repeat libraries are not available [118]. Poly (A) tracks are trimmed to retain a few adenines (usually 6-10) to get high quality ESTs for clustering and assembly process as these are not encoded in the genomic sequence.

1.12.2 Transcript clustering and assembly

Transcript clustering is to collect overlapping reads from the same transcript of a single gene, into a unique cluster to reduce redundancy. Sequence assembly is to determine the sequence of a target transcript/gene which involves alignment and merging of DNA fragments to form long contiguous sequences (contigs) [110, 113]. The fragmented data resulted as part of clustering is indexed using gene sequence information [119]. Pair-wise sequence similarities are measured to cluster ESTs. Different steps for EST clustering are described in detail by Ptitsyn and Hide (2005) [120], which discusses two approaches for

EST clustering, “stringent” and “loose.” The stringent clustering method is conservative, generates shorter sequence consensi with low coverage of expressed genes and uses single-pass grouping of ESTs resulting in relatively accurate clusters while, loose clustering is “liberal” and it repeats low quality EST sequence alignments many times to generate less accurate but longer sequence consensi. Thus, there is a better coverage of expressed gene data and alternatively spliced transcripts, but there is a possibility of including paralogs in the clusters. StackPACK [121] is designed as for loose clustering, while TIGR Gene Indices [122] implement stringent clustering. Most broadly used programs for sequence clustering and assembly are Phrap and CAP3 [123] where, benchmarking analysis by Quackenbush and co-workers [124] showed CAP3 out-performed other similar programs, maintaining a high level of sensitivity to gene family members and producing high fidelity consensus sequences along with the handling of sequencing errors. Algorithms ‘overlap-layout-consensus’ [125] and ‘de-Bruijn-graph’ [126, 127] are used to assemble long-(generated by sanger sequencing, 454 platform) and short-reads (illumina and SOLiD platforms) respectively. CAP3, Phrap, TIGR assembler, the parallel contig assembly program (PCAP; [128]) and the mimicking intelligent read aseembly program (MIRA; MIRA: an automated genome and EST assembler) support long-read assembly and short-read assembly with short oligonucleotide analysis package (SOAP [129]), the assembly by short sequencing (ABYSS; [130]), the exact *de novo* assembler (EDENA [131]), Euler-SR [132] and velvet [127].

1.12.3 Conceptual translation of ESTs

To assign a predicted identity to each query sequence the contigs and singletons are compared to known sequence data in public database [110]. EST sequences are translated to identify the protein-coding regions or open reading frames (ORFs), from consensus EST sequences, to enhance the process of gene discovery and gene boundary predictions using OrfPredictor [133], ESTScan [134], DECODER [135]. After peptide prediction, protein analysis is carried out and known protein domains are inferred [110]. InterProScan [111], a tool for characterization of a protein family or a protein sequence, domain, functional site by compairing sequences with information available in databases PROSITE [136], Pfam [137], PRINTS [138], ProDom [139], SMART [140] or Gene Ontology [141]. Transmembrane doamin are predicted with program TMHMM [142], signal peptide motifs with SignalP [143] and non-classical secretions with SecretomeP [144].

1.12.4 EST Annotation

Annotation of both DNA or protein sequences are mainly assigning Gene Ontologies (GO), to establish pathway associations using KEGG pathways, mapping sequences to protein domains/motifs, identifying molecular interaction partners and finally to use comparative genomics approaches to assign function based on database similarity searches with known data.

1.12.5 Database similarity searches

Different flavours of the BLAST [145] programs from NCBI serve as universal tools for database similarity searches such as for comparing nucleotide sequence data with DNA (BLASTn) or cDNA or amino acid (BLASTx) sequences, or conceptually translated peptides with protein sequences (BLASTp) available in public databases [110]. GENBANK [146], EMBL[147], DDBJ [148] are the three on-line public databases used for sequence data analysis and annotation. PDB [149] represents the protein structure database and SWISS-PROT for protein sequence information.

1.12.6 Functional assignment using Gene Ontologies

Gene Ontologies (GO) have been used widely to predict gene function and classification and is categorised into three parts: i) biological process, referring to a biological objective to which the gene or gene product contributes; ii) molecular function defining the biochemical activity (including specific binding to ligands or structures) of a gene product; and iii) cellular component, to place in the cell where a gene product is active [141]. BLAST2GO [150], a sequence-based tool to assign GO terms, extracting them for each BLAST-match is obtained by mapping to extant annotation associations. InterPro [111], assigns GO terms to any sequence based on the known GO assignments of the functional domains/motifs identified in the sequence.

1.12.7 Pathway mapping

Biological pathways associated with EST sequences provide vital information on gene function, gene expression and regulation. The Kyoto Encyclopedia of Genes and Genomes database (KEGG) [151] is used to functionally classify EST data based on biochemical functionality. KOBAS [152], developed based on KO, is a web-based platform for automated annotation and pathway identification. KAAS [153] and PathFinder [154] are other recent publicly available tools.

1.12.8 Identification of interaction partners

Proteins usually work through number of interactions with other bio-molecules. Hence, insights into possible functions within the cell can be obtained with the study of each protein and its binding partners. Studying protein-protein interactions in parasites is important to understand the complex interplay between the cellular environments of the parasite and its host during the course of invasion and infection which will help in the development of new drug targets. Different interaction databases include IntAct [155], BIND [156], HPRD [157], DIP [158] and MINT [159]. User data can be matched with similarity search programs like BLAST and can be experimentally validated to get an idea about the role of each molecule in its specific cellular environment *in vivo*.

1.13 EST analysis pipelines and current bioinformatic tools

Automated analysis pipelines are necessary to store, organize and annotate ESTs with the advent of large-scale sequencing projects and generation of sequence data at protein and cDNA levels along with the development of new and updated bioinformatic tools. A pipeline would automatically clean, cluster, assemble and generate consensus sequences, conceptually translating these into possible protein products and assign putative function based on various DNA and protein similarity searches. The expansion of bioinformatics tools have led to the development of integrated pipelines or web-based programs such as ESTExplorer and EST2Secretome [160]. Current available bioinformatic tools and pipelines have reviewed the rules, principles, and methods for the analysis of EST data [110, 113]. An extensive review of analysis pipelines and tools is presented in Paper 1, at the end of this chapter. Another pipeline worth mentioning is EST-PAC [161], a web-based software package for EST annotation. EST-PAC provides automated EST annotation by: 1) searching local or remote biological databases for sequence similarities using BLAST services, 2) predicting protein coding sequence from EST data and, 3) annotating predicted protein sequences with functional domain predictions.

1.14 Objectives

In all these years, various studies have been conducted on the parasitic helminths of socio-economic importance which elucidated patterns of transcription. Major progress has been made in different aspects of study including the immunobiology, pathogenesis, epidemiology, diagnosis, treatments, etc. The transcriptome study and the integration of transcriptomic and proteomic analysis has helped to understand the parasite development and host-parasite interactions and disease along with the reason for the ability of a parasite

to sustain itself on a host which could lead to identifying novel intervention stratergies. ESTs generated for numerous organisms required the development of computational strategies to analyze and organize them. EST analysis had to undergo lot of steps and required different bioinformatic tools creating confusion in selection of the right tool. As a part of this, we conducted a review of the existing tools and the analysis pipelines and the future requirements. With the advent of NGS, new opportunities to explore the molecular biology of parasitic helminths have widened which requires serious evaluation and development of bioinformatic pipeline for assembly and analyses.

The overall objective of this thesis was to suggest improved bioinformatic approaches to analyse, validate and identify biomolecules, especially ES proteins responsible for causing disease through transcriptomic studies and to establish and integrate bioinformatic analysis pipeline with new tools.

Specific objectives are described below:

1. Review different methods for EST analysis, new semi-automated bioinformatic pipelines and their application to parasitic nematodes with the key focus of identifying pathways or molecules responsible for development of new drugs or vaccines (Paper 1).
2. Carry out an analysis on *Fasciola hepatica* by integrating transcriptomic and proteomic analysis to validate bioinformatics predictions with experimental proteomics data, as well as quantify the deficiencies of bioinformatics analysis (Paper 2).
3. Apply and benchmark new bioinformatics tools for the large-scale analysis of adult *Teladorsagia circumcincta* transcriptome data, as a step to develop a new and updated analysis pipeline (Paper 3).
4. Develop TranSeqAnnotator as a generic transcriptomic analysis pipeline with updated tools for assembly and annotation of EST and NGS data, with the tools selected from Paper 3 and its application to *Ascaris lumbricoides* data (Paper 4).
5. Apply TranSeqAnnotator to NGS transcriptome data for L4 stage of *Teladorsagia circumcincta* and compare the results with the transcriptome analysis of the adult *T. circumcincta*, reported in Paper 3, to understand the changes in the transcriptome with the parasite developmental stage (Paper 5).

The significance of the study, its novelty and future directions are presented in the concluding chapter.

Biotechnology Advances 27 (2009) 439–448

Contents lists available at ScienceDirect

Biotechnology Advances

journal homepage: www.elsevier.com/locate/biotechadv

Research review paper

Advanced *in silico* analysis of expressed sequence tag (EST) data for parasitic nematodes of major socio-economic importance — Fundamental insights toward biotechnological outcomes

Shoba Ranganathan [a,b,*], Ranjeeta Menon [a], Robin B. Gasser [c,*]

[a] Department of Chemistry and Biomolecular Sciences and Australian Research Council (ARC) Centre of Excellence in Bioinformatics, Macquarie University, Sydney, New South Wales 2109, Australia
[b] Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 119260
[c] Department of Veterinary Science, The University of Melbourne, Werribee, Victoria 3030, Australia

ARTICLE INFO

Article history:
Received 23 January 2009
Received in revised form 17 March 2009
Accepted 22 March 2009
Available online 2 April 2009

Keywords:
Helminths
Nematodes
Strongylida
Genomics
Transcriptomics
Bioinformatics
Drug targets

ABSTRACT

Parasitic nematodes infect humans, other animals and plants, and impose a significant public health and economic burden worldwide due to the diseases that they cause. A better understanding of parasite genomes, host–parasite relationships and the molecular biology of parasites themselves will enable the rational development of diagnostic tests and/or safe anti-parasitic compounds, following the functional annotation of parasite genomic sequences. With only a few completely sequenced nematode genomes, expressed sequence tag (EST) datasets provide a low-cost alternative ("poor man's genome") to whole genome sequences and a glimpse of the transcriptome of an organism. EST data require a number of computational methods for their pre-processing, clustering, assembly and annotation to yield biologically relevant information. In this article, we review the steps involved in EST data analysis, the development of new semi-automated bioinformatic pipelines and their application to parasitic nematodes of major socio-economic significance, focused on identifying molecules involved in key biological processes or pathways that might serve as targets for new drugs or vaccines.

Contents

1. Introduction

Parasitic worms of humans and other animals cause diseases of major socio-economic importance globally. In particular, parasitic flatworms (trematodes and cestodes) and roundworms (nematodes) have a long-term impact (directly and indirectly) on human health and cause substantial suffering, particularly in children. The World Health Organization (WHO) estimates that 2.9 billion people are

* Corresponding author. Ranganathan is to be contacted at Department of Chemistry and Biomolecular Sciences and Australian Research Council (ARC) Centre of Excellence in Bioinformatics, Macquarie University, Sydney, New South Wales 2109, Australia. Tel.: +61 2 98506262; fax: +61 2 98508313. Gasser, Tel.: +61 3 97312000; fax: +61 3 97312000.
E-mail addresses: shoba.ranganathan@mq.edu.au (S. Ranganathan), robinbg@unimelb.edu.au (R.B. Gasser).

0734-9750/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.biotechadv.2009.03.005

infected with nematodes (Hotez et al., 2007). Morbidity caused by nematodes is substantial and exceeds diabetes and lung cancer in disability adjusted life year (DALY) measurements (Engels and Savioli, 2006). Worldwide, the current financial losses caused by parasites to agriculture (i.e. domesticated animals and crops) have a major impact on farm profitability and exacerbate the global food shortage. In the agricultural setting, most parasites are controlled mainly through the use of chemotherapeutic agents (anthelmintics). This type of control can be costly and is not always effective in the absence of a management component. The excessive use of anthelmintics has resulted in serious problems with drug resistance in parasites. Furthermore, the use of such drugs can pose risks of residue problems in meat, milk and the environment. Given the increasingly stringent demands placed on maximum residue levels, the ongoing development of novel and improved control strategies (including non-chemical means) is crucial. Possibilities include the rational development of diagnostic tests and/or safe anti-parasitic compounds, based on a better understanding of parasite genomes, host–parasite relationships and the molecular biology of the parasites themselves. Hence, there are significant gains to be made by improving our knowledge of parasites, which will lead to such outcomes. Whole genome sequencing of key parasitic helminths can also provide an important foundation for a wide range of fundamental areas (including functional genomics, genetics, proteomics, systems biology, molecular biology, physiology, biochemistry, ecology, epidemiology, pathology, and many more), underpinning applied areas.

In 2004, a meeting funded by the Wellcome Trust was convened at the Sanger Institute (Cambridge, UK) to discuss whole genome sequencing of parasitic helminths (http://www.genome.gov/Pages/Research/Sequencing/SeqProposals/StrongylidaWormSeq.pdf). The global impact of "neglected" helminthic diseases, particularly through the subclinical and chronic infections that they cause, makes them key candidates for genome sequencing. The purpose of the meeting was to provide advice on the prioritization of helminth species of medical and veterinary importance, considering experimental models and scientific interest. The criteria for selecting a particular species included: clinical aspects; the availability of an animal model system; the comparative value of particular species; the size of the scientific community interested in or working on a species; and, the potential to enable research following genome sequencing. Several candidate organisms were identified at the meeting and nominated for sequencing (Mitreva et al., 2007; http://www.genome.gov/10002154). It was concluded that, at this stage, the scientific community should proceed with individual nominations while working together to defining packages (such as groups of related species) that might gain support for more ambitious plans from major funding agencies.

Although various projects on genomic and EST sequencing of metazoan organisms (cf. www.sanger.ac.uk/pathogens/; www.nematode.net/) have provided some information, limited progress has been made on socio-economically important parasitic helminths compared with human genomic sequencing programs. By contrast, the completion of the full genome sequence of the free-living nematode *C. elegans* (available from WormBase; http://www.wormbase.org) has provided an extremely valuable resource and a solid platform for comparative genome analyses for various nematode groups, particularly those representing clade V (proposed by Blaxter et al., 1998). *C. elegans* is also a powerful system for genetic and molecular investigations, since it has a rapid life cycle and is easy to maintain *in vitro*. The karyotype is $2n = 12$ (five pairs of autosomes and one pair of sex chromosomes), which appears to be consistent with a range of strongylid nematodes, and the genome of *C. elegans* contains ~20,000 genes. Strongylid nematodes, for example, are, based on molecular phylogenetic analysis, considered as relatively closely related to *C. elegans* (see Blaxter et al., 1998), as supported by findings from expressed sequence tag (EST) projects carried out on 40 nematode species other than *Caenorhabditis*, including 24 parasitic nematodes of mammals, 14 plant parasites and 2 free-living bacteriovores (see Mitreva et al., 2005a). Indeed, the strongylid datasets (Parkinson et al., 2004a) have the highest genetic similarity to *C. elegans* compared with the other taxa examined, based on the cumulative number of new genes found in each species. Both the distribution of homology matches and their relative scores support a close relationship between the Strongylida and *C. elegans* (see Parkinson et al., 2004a). Clusters of members of the Strongylida (represented by *Ancylostoma caninum* and *A. ceylanicum*; clade V) were evaluated relative to *Trichinella spiralis* (clade I), *Dirofilaria immitis* (clade III), *Strongyloides stercoralis* (clade IVA) and *Meloidogyne incognita* (clade IVB). For all of the non-strongylid species, 20.7 ± 9.6 of matches from database similarity searches were nematode-specific, and 6.3 ± 4.0 of matches were to non-nematode species. In contrast, clusters representing the strongylids were more skewed toward the nematode category, with only $3.3 \pm 0.01\%$ of matches being to non-nematodes, whereas 31.3 ± 0.06 had matches to nematodes. The ratio of nematode-only to non-nematode-only matches in *Ancylostoma* spp. (9.6 ± 5.6) differed highly significantly ($P<0.0001$, Student *t*-test) from other nematodes (3.3 ± 2.4) examined. Furthermore, *C. elegans* orthologues were identified for only 45% of *T. spiralis* clusters, compared with 60–65% for species of *Strongyloides* and *Ancylostoma* (Parkinson et al., 2004a). Further, the 15 most conserved gene products among *C. elegans*, *T. spiralis* and *S. stercoralis* were significant, while the top 15 *A. ceylanicum* scores were highly significant (Mitreva et al., 2004a,b; 2005a,b,c). Therefore, the extrapolation from the biology of a well-studied nematode, such as *C. elegans*, will be of great benefit when investigating members of the order Strongylida. Also, the chance that a cloned gene from a strongylid nematode has a homologue in *C. elegans* is high, with the exception of genes associated with host–parasite interactions. As there are no reliable culturing systems available for the propagation and maintenance of the entire life cycle of strongylid nematodes *in vitro*, *C. elegans* provides a powerful surrogate system to test the function of orthologous/homologous genes.

The future focus needs to be on the functional annotation of genes from large-scale EST and genomic sequence datasets for parasitic nematodes and on questions regarding the genetic basis of parasitism (Gasser and Newton, 2000; Newton et al., 2002). Functional characterization will require the application of genomic, proteomic and bioinformatic technologies. These will be greatly aided by the application of *C. elegans* data to whole genome sequences of key parasitic helminths. In spite of currently reported technological advances in genomics, transcriptomics and proteomics, surprisingly, progress on developing user-friendly and highly reliable and efficient bioinformatic pipelines, tailored specifically to parasitic helminths, has been slow.

Computational strategies to organise and analyse EST datasets are confounded by a bewildering number of analysis steps (Nagaraj et al., 2007a). An overwhelming number of tools is available for each step, with varying strengths/weaknesses for systematically extracting biologically relevant information from EST datasets. There has been some confusion about the choice of the right tools for individual steps of EST analysis and the subsequent annotation at the DNA or protein level. This confusion is compounded by the ability of some tools to handle high-throughput EST data, while others cannot. In this context, we considered that an appraisal of currently available EST analysis methods and platforms was necessary. We conducted such an evaluation and found that all available platforms terminated prior to downstream functional annotation, including gene ontologies (GOs), motif/pattern analysis and pathway mapping, necessitating the establishment of a comprehensive large-scale EST analysis pipeline (Nagaraj et al., 2007a). Given the rapidity with which enormous amounts of sequence data are currently being generated, there is an urgent need for advanced, high-throughput computational analyses of EST and genomic sequence datasets using automated platforms.

The purpose of the present article was to provide a background on the significance of EST analysis, to describe recent advances in the development of semi-automated bioinformatic pipelines and summarize their application to parasitic nematodes of major socio-economic significance, with a perspective toward identifying molecules involved in key biological processes or pathways which might serve as targets for new drugs or vaccines.

2. Critical appraisal of current bioinformatic tools and development of a semi-automated platform for EST analysis and annotation

Extensive computational strategies have been developed to organise and analyse both small and large EST datasets for gene discovery, and transcriptional and singe nucleotide polymorphism (SNP) analyses as well as the functional annotation of putative gene products. In a recent review (Nagaraj et al., 2007a), we provided an overview of the significance of ESTs in the genomic era, and their properties and applications. Methods adopted for each step of EST analysis by various research groups were compared. Challenges that lie ahead in organizing and analysing the ever-increasing EST data were identified, and the most suitable software tools for EST pre-processing, clustering and assembly as well as database matching and functional annotation were compiled.

Based on critical appraisal of the literature (Nagaraj et al., 2007a), we selected the best-suited programs for each step for EST analysis and incorporated them, to eventually design an semi-automated, user-friendly platform, called ESTExplorer, for pre-processing and assembly, followed by the annotation of relatively large amounts of EST data at both the DNA and protein levels (Nagaraj et al., 2007b). This user-friendly workflow system (Fig. 1) has been built on open-source technologies for EST data management and analysis. An appraisal of various EST analysis platforms revealed that they terminate prior to functional annotation (Nagaraj et al., 2007a,b). Some terminate at the assembly level, providing contigs and singletons as output, while others run nucleotide-based programs with limited annotation at the protein level. ESTExplorer (http://estexplorer.biolinfo.org) was designed to incorporate the best, currently available bioinformatic tools for pre-processing, clustering, gene ontology mapping, nucleotide functional annotation, detailed protein functional identification and metabolic pathway mapping. These programs are locally implemented over different dedicated processors on a multi-processor server, to ensure rapid and efficient computation.

ESTExplorer is a comprehensive workflow system for EST data management and analysis. Species-specific repeat masking and conceptual translation for 10 organisms (human, mouse, rat, rice, zebrafish, chicken, fly, dog, thale cress and the roundworm, *C. elegans*) are provided. This *in silico* system accepts a set of EST sequences as input, which can be analysed using programs selected by the user. Following pre-processing and assembly, datasets are annotated at the nucleotide and protein levels. The outputs include GOs, the 'functional' annotation of proteins, in terms of mapping to protein domains and metabolic pathways. This system can be applied specifically to annotate large EST datasets from nematodes and identify novel genes or gene products as potential drug targets.

The workflow in ESTExplorer can be divided into three phases, with Phase I being dedicated to EST sequence pre-processing and assembly. In Phases II and III, annotation is carried out at the nucleotide and protein levels. Input ESTs are first submitted for pre-processing (Phase I) and assembly (Input Option 1 in Fig. 1), followed by analyses in Phases II and III. While ESTExplorer has been designed primarily for large-scale EST assembly and analysis, ESTs assembled using other programs can also be submitted for annotation alone. Assembled ESTs can be submitted directly to Phases II and III for functional annotation, using Input Option 2 (Fig. 1).

Phase I employs three programs: SeqClean (http://www.tigr.org/tdb/tgi/software/), RepeatMasker (http://repeatmasker.org) and CAP3 (Huang and Madan, 1999), which are run sequentially to convert input EST sequences into high quality ESTs. SeqClean removes vector sequences (using the Univec database at NCBI; http://www.ncbi.nlm.nih.gov/), Poly-A tails, low quality segments at 5′ and 3′ cDNA ends and low complexity regions (using the DUST module) from input ESTs. All short ESTs (<100 bp) are eliminated. Output from SeqClean can be processed optionally, employing RepeatMasker, or can be assembled directly using CAP3. RepeatMasker employs cross-match and up-to-date repeat libraries for different species from RepBase, to carry out species-specific repeat masking. CAP3 either accepts repeat-masked, high quality EST sequences or the output from SeqClean, and performs clustering and assembly into large contiguous sequences, known as contigs (integrating multiple, related ESTs) and singletons.

In Phase II, nucleotide-level annotations of assembled EST contigs and singletons from Phase I or data uploaded by the user, are carried out. BLASTX (Altschul et al., 1997) and NCBI's non-redundant protein database, BLAST2GO (Conesa et al., 2005), which assigns functionality based on GOs (Ashburner et al., 2000), are the programs incorporated in this phase.

Protein-based annotation is carried out in Phase III, following the translation of individual EST sequences into amino acid sequences. ESTScan (Iseli et al., 1999) accepts contigs and singletons from CAP3

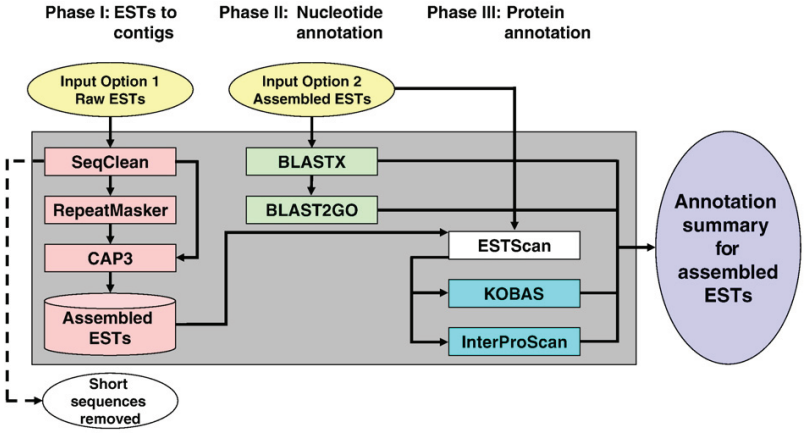


Fig. 1. ESTExplorer analysis and annotation workflow, showing Phase I (pre-processing and assembly), Phase II (nucleotide-level annotation) and Phase III (protein-level annotation).

and provides conceptual translations, using the genetic code from a related organism in a two-step process. Firstly, coding regions or open reading frames (ORFs) are detected and extracted. Secondly, these ORFs are translated into putative peptides. The peptide sequences from ESTScan can be analysed using InterProScan (Quevillon et al., 2005) and KOBAS (Wu et al., 2006) for processing. InterProScan matches protein sequences against InterPro (an integrated resource for protein families, domains and functional sites from member databases, such as PROSITE, PRINTS, PFAM, ProDom and SMART), providing details of domain/motif architecture for each sequence for the assignment of biological function. KOBAS maps protein sequences to biochemical pathways based on the Kyoto Encyclopaedia of Genes and Genomes, KEGG (Kanehisa et al., 2006).

It is usually difficult to collate results at the final output stage when a large dataset is subjected to analysis using a workflow containing several phases and multiple programs. To address this issue, ESTExplorer tracks each assembled sequence (contig/singleton) that has been annotated functionally. After an EST or contig dataset has been submitted to ESTExplorer, a status page is accessible to monitor the progress of the analysis in the program. As each selected step is completed, the status page is updated, and the output from that program becomes available. At the completion of processing, the outcome is summarized, with links to output files. While ESTExplorer can run on a fully automated mode, we have provided expert users with the option of modifying default parameter values as well as program selection options. For each assembled sequence, a detailed summary comprising GOs, the inference of protein function and mapping to protein domains and metabolic pathways, is displayed. A tutorial and a description of each program used are also available for reference.

3. Assessment of the efficiency and performance of ESTExplorer by comparison with manual approaches

Traditionally, biologists have relied on conventional database similarity searches, using BLAST (Altschul et al., 1997), to derive clues regarding biological function. Such an EST analysis was reported recently for the pine wood nematodes, *Bursaphelenchus xylophilus* and *B. mucronatus* (see Kikuchi et al., 2007). To evaluate the efficacy of ESTExplorer compared with standard database-derived functional annotation strategies, we used a well-defined dataset generated previously from the adult stage of a parasitic nematode, *Trichostrongylus vitrinus*, and analysed it manually (Nisbet and Gasser, 2004). Compared with this study, which was limited to annotation of ESTs *via* similarity searches (BLAST) for the assignment of putative function, we added annotations to the EST data and identified important genes for further investigation (Nagaraj et al., 2008a). Firstly, we included functional identifications, in terms of mapping to protein domains and metabolic pathways. Secondly, we categorised the *T. vitrinus* ESTs based on comparison with three databases, namely Wormpep (www.sanger.ac.uk/Projects/C_elegans/WORMBASE/current/wormpep.shtml), a 'parasitic nematode database' and a 'non-nematode database' (the last two were built 'in house'), and used the Java tool SimiTri (Parkinson and Blaxter, 2003) to visualize the data comparisons. Thirdly, we related ESTs to molecules in *C. elegans* which can be silenced by double-stranded RNA interference (RNAi) (www.wormbase.org). For comparative purposes, we repeated the earlier EST analysis (cf. Nisbet and Gasser, 2004) by running BLAST against the current databases, in order to minimize any database bias in the results and compared these results with the automated EST sequence annotations for *T. vitrinus* data. Previously, Nisbet and Gasser (2004) carried out the BLAST analysis of 1776 gender-enriched ESTs from *T. vitrinus* individually and categorised these molecules into different functional classes. This conventional analysis took ~16 weeks to perform compared with an enhanced analysis of the same dataset using ESTExplorer (with the exception of the secretome, SimiTri and RNAi phenotype analyses), which took less than 3 h to perform. A total of 301 ESTs representing 31 functional categories were defined and compared. Male-enriched ESTs of *T. vitrinus* encoded predominantly major sperm protein-like, protein kinases/phosphatases, transcription factors, nucleic acid synthesis and other categories, whereas female-enriched ESTs from this species represented vitellogenins, protein kinases/phosphatases and transcription factors and molecules involved in carbohydrate metabolism and modification and the ubiquitin–proteasome pathway.

We demonstrated that ESTExplorer provides a comprehensive "functional" annotation for EST datasets, which leads to an enhanced characterization and understanding of the molecules. When a gene has multiple predicted functions, it is possible to list them in advanced annotation protocols, such as GO and pathway mapping, which provides comprehensive information on the molecule, in order to underpin any molecular, biochemical and/or biological investigations. For instance, the EST TVm02_C07, which is homologous to a serine/threonine protein phosphatase, revealed the following GO terms: "chromatin modification, protein dephosphorylation, embryonic cleavage, cytokinesis, meiosis, oviposition, manganese ion binding, protein phosphatase type 1 activity, protein binding, mitosis, glycogen metabolic process, iron ion binding, mitochondrial outer membrane, nucleus", and has been predicted to be involved in multiple pathways, "long-term potentiation, regulation of actin cytoskeleton, focal adhesion and insulin signalling pathway". Other findings indicated that the *C. elegans* orthologue (a yeast glc seven-like phosphatase encoded by the gene F56C9.1; accession no. NP_498617.1) can be silenced in this free-living nematode, leading to progeny with Egl (egg laying deficit), Emb (embryonic lethal) and/or Sck (sick) phenotypes (Rual et al., 2004). Being able to achieve silencing in *C. elegans* showed that this gene is central to the development, reproduction and/or survival of this nematode. This information provided a basis for the detailed molecular characterization and transcriptional analysis of the full-length gene (*Tv-stp-1*) encoding this serine/threonine protein phosphatase (*Tv*-STP-1) from *T. vitrinus* (see Hu et al., 2007). The findings indicated that there is relative conservation in features and function of the serine/threonine protein phosphatase characterized for *T. vitrinus*, *O. dentatum* (see Boag et al., 2003) and *C. elegans*, likely to have significant implications for exploring molecular reproductive and developmental processes in strongylid nematodes of socio-economic importance.

4. Brief comparison of ESTExplorer with other currently available EST analysis pipelines

A number of EST analysis pipelines are currently available; these are listed in Table 1 in chronological order. The major processing steps, pre-processing, clustering and annotation at the nucleotide and protein levels have been presented for each pipeline. The specific features of ESTExplorer compared with those of other similar pipelines are listed.

Only nine analysis pipelines (Table 1, in bold font), *viz.* PipeOnline (Ayoubi et al., 2002), ESTAP (Mao et al., 2003), ESTAnnotator (Hotz-Wagenblatt et al., 2003), PartiGene (Parkinson et al., 2004b), openSputnik (Rudd, 2005), ParPEST (D'Agostino et al., 2005), ESTExplorer (Nagaraj et al., 2007b), ESTPass (Lee et al., 2007) and EST2uni (Forment et al., 2008), provide tools for comprehensive EST analysis. Of these, PipeOnline, ESTAP, ESTAnnotator, ParPEST, PartiGene and EST2uni rely solely on the annotation of BLAST hits for nucleotide and protein-level annotations, as do JUICE (Latorre et al., 2006) and annot8r (Schmid and Blaxter, 2008); openSputnik and ESTPass employ BLAST-derived annotations for nucleotide annotations.

TGICL (Pertea et al., 2003) and EGAssembler (Masoudi-Nejad et al., 2006) have no annotation phases and thus perform only pre-processing and clustering. Of these, TGICL requires the use of proprietary software (Paracel Transcript Assembler), while

Table 1
A list of currently available expressed sequence tag (EST) analysis pipelines.

Pipeline and website	Pre-processing programs	Clustering and assembly programs	DNA-level annotation	Protein-level annotation
PipeOnline (Ayoubi et al., 2002) http://bioinfo.okstate.edu/pipeonline/	Phred; Cross_Match	Phrap	*	*
EST Analysis Pipeline (ESTAP) (Mao et al., 2003) http://staff.vbi.vt.edu/estap/	Phred; Cross_Match	D2_cluster; CAP3	*	*
ESTAnnotator (Hotz-Wagenblatt et al., 2003) http://genome.dkfz-heidelberg.de/menu/biounit/dev.shtml#estannotator	Phred; RepMask	CAP3	*	*
ESTweb (Paquola et al., 2003) http://bioinfo.iq.usp.br/estweb/	Phred; Cross_Match	×	*	×
TGICL (Pertea et al., 2003) http://compbio.dfci.harvard.edu/tgi/	SeqClean; megaBLAST	CAP3; Paracel TranscriptAssembler	×	×
ESTIMA (Kumar et al., 2004) http://titan.biotec.uiuc.edu/ESTIMA/	×	BlastClust; CAP3	BLAST; Gene ontology (GO)	×
PartiGene (Parkinson et al., 2004b) http://www.nematodes.org/bioinformatics/PartiGene	Cross_Match	Phred	*	*
openSputnik (Rudd, 2005) http://sputnik.btk.fi	Cross_Match	HPT2 and CAP3	*	✓
ParPEST (D'Agostino et al., 2005) http://www.cab.unina.it/parpest/	SeqClean; Phrap; Repeatmasker	D2_cluster; PaCE; CAP3	*	*
EGAssembler (Masoudi-Nejad et al., 2006) http://egassembler.hgc.jp/	SeqClean	PHRAP; CAP3	×	×
JUICE (Latorre et al., 2006) http://genoma.unab.cl/juice_system/	×	CAP3	*	×
ESTExplorer (Nagaraj et al., 2007b) http://estexplorer.biolinfo.org	SeqClean; Repeatmasker	CAP3	BLAST, GO	✓
ESTPass (Lee et al., 2007) http://estpass.kobic.re.kr/	Cross_Match	D2_cluster; CAP3	*	✓
Annot8r (Schmid and Blaxter, 2008) http://www.nematodes.org/bioinformatics/annot8r	×	×	*	*
EST2uni (Forment et al., 2008) http://bioinf.comav.upv.es/est2uni/	Phred,SeqClean	CAP3	*	*
OREST (Waegele et al., 2008) http://mips.gsf.de/genre/proj/orest/index.html	WebTraceMiner	×	*	*

×: not available; ✓: available; *: BLAST-derived annotation; comprehensive analysis pipelines in bold-type.

EGAssembler provides a freely available webserver; these systems may be used to generate high quality EST contigs for annotation. Pre-processing is unavailable in ESTIMA (Kumar et al., 2004), JUICE and annot8r. Clustering is not incorporated in ESTweb (Paquola et al., 2003), annot8r and OREST (Waegele et al., 2008), so that relatively short input ESTs are directly matched with database sequences for annotation. These systems are thus unlikely to provide high quality EST sequences for annotation.

Thus, ESTExplorer provides a comprehensive analysis and annotation pipeline for ESTs. The use of multiple programs for nucleotide and protein annotations overcomes the limitations of BLAST-based annotations for these organisms, providing homologous sequences, GO classifications, KEGG pathways and protein domains and motifs.

5. Applications of the ESTExplorer pipeline to parasitic nematodes

Having assessed the performance of ESTExplorer (sections 3 and 4), we subsequently applied the pipeline to the detailed analyses of EST datasets for parasitic nematodes of major socio-economic importance, including *Dictyocaulus viviparus* (the bovine lungworm), *Haemonchus contortus* (the barber's pole worm of small ruminants), *A. caninum* (the hookworm of dogs) and *Ascaris suum* (the common roundworm of pigs) (Campbell et al., 2008; Datu et al., 2008; Huang et al., 2008; Ranganathan et al., 2007). These analyses allowed the limited information available in EST datasets to be transformed into biologically relevant information, enabling the formulation of hypotheses for future experimental work.

Lungworms of the genus *Dictyocaulus* (family Dictyocaulidae) are parasitic nematodes of major economic importance. They cause pathological effects and clinical disease in various ruminant hosts, particularly in young animals. *D. viviparus* is a major pathogen of cattle, particularly calves, with severe infections being fatal. In this study, we provided first insights into the transcriptome of the adult stage of *D. viviparus* through the analysis of ESTs determined conventionally. In brief, using ESTExplorer, we estimated that the dataset of 5271 ESTs was derived from 2258 genes, based on cluster and comparative genomic analyses. Of the 2258 representative ESTs, 1159 (51.3%) had homologues in *C. elegans*, 1174 (51.9%) in parasitic nematodes, 827 (36.6%) in organisms other than nematodes, and 863 (38%) had no significant matches to any sequence in the current databases. Of the homologues, 569 had observed 'non-wildtype' RNAi phenotypes *C. elegans*, including Emb (embryonic lethality), Ste (maternal sterility), Stp (sterility in progeny), Lva (larval arrest) and Gro (slow growth). We could functionally classify 776 (35%) sequences using the GOs and established pathway associations to 696 (31%) sequences in KEGG. In addition, we identified 85 putative secreted proteins which could represent potential anthelmintic or vaccine targets. In conclusion, the bioinformatic analyses of EST data for *D. viviparus* elucidated sets of relatively conserved and potentially novel genes. The characterization of the *D. viviparus* transcriptome also provides a foundation for whole genome sequencing and future comparative transcriptomic analyses.

In addition to lungworms, gastrointestinal nematodes of livestock cause substantial production losses due to poor productivity, failure to thrive and deaths (e.g., McLeod, 1995; Smith, 1997; Coles, 2001). The financial losses associated with these endoparasites are estimated (based on figures of the sales of anthelmintic drugs; unpublished data) at tens of billions of dollars per annum. A parasitic nematode of paramount importance in small ruminants (sheep and goats) is the blood-feeding, gastric nematode *H. contortus* (order Strongylida). This dioecious, haematophagous nematode, harboured in the abomasum, causes anaemia and associated complications, leading to death in severely affected animals (see Nikolaou and Gasser, 2006). The oviparous adult females release eggs *via* the faeces into the environment, in which first-stage larvae (L1s) develop and then hatch (within ~1 day, depending on conditions). The L1s develop into L2s and then moult to become infective third-stage larvae (L3s). The ensheathed L3s are ingested by the host with herbage, exsheath in the abomasum, and develop *via* the blood-feeding, fourth-stage larvae (L4s) to the adult females and males within ~3 weeks. Females become fully gravid between 3 and 4 weeks (Stoll, 1929) and each of them lays an average of 4500 eggs per day (Coyne and Smith, 1992). Hence, the development of *H. contortus* is complex and involves a number of timed processes, the timing of which can vary, depending on population, environmental and/or host factors (Rossanigo and Gruner, 1996). While there is considerable knowledge of the morphological changes which take place during the life cycle of *H. contortus*, very little is understood about the biochemical and molecular processes underlying the development and reproduction of this important parasite (Nikolaou and Gasser, 2006). Therefore, in a

recent study (Campbell et al., 2008), we took the first step toward exploring genes differentially transcribed between female and male *H. contortus* at the adult stage. We used ESTExplorer for the clustering and analysis of the currently available EST datasets for *H. contortus*, and selected representative subsets of molecules for transcriptional profiling by oligonucleotide microarray analysis and for the prediction of function and gene interactions based on comparative analyses with *C. elegans*.

Following the bioinformatic analysis of the *H. contortus* dataset using ESTExplorer, 1885 representative ESTs were selected, to which oligonucleotides (three per EST) were designed and spotted on to a microarray (Campbell et al., 2008). This microarray was hybridized with cyanine-dye labelled cRNA probes, synthesized from RNA from female or male adults of *H. contortus*. Differential hybridization was displayed for 301 of the 1885 rESTs (~16%). Of these, 165 (55%) had significantly greater signal intensities for female cRNA and 136 (45%) for male cRNA. Of these, 113 with increased signals in female or male *H. contortus* had homologues in *C. elegans* predicted to function in metabolism; information storage and processing; cellular processes and signalling; and, embryonic and/or larval development. Of the ESTs with no known homologues in *C. elegans*, 23 (~40%) had homologues in other nematodes, two had homologues in various other organisms and 30 (52%) had no homology to any sequence in current gene databases. A genetic interaction network was predicted for the *C. elegans* homologues/orthologues of the gender-enriched *H. contortus* genes, and a focused analysis of a subset revealed a tight network of molecules involved in amino acid, carbohydrate or lipid transport and metabolism; energy production and conversion; translation, ribosomal structure and biogenesis; and, importantly, those associated with meiosis and/or mitosis in the germline during oogenesis or spermatogenesis. This study (Campbell et al., 2008) provided a basis for the molecular, biochemical and functional exploration of selected molecules with differential transcription profiles in *H. contortus*, for further microarray analyses of transcription in different developmental stages of *H. contortus*, and for an extended functional analysis once the full genome sequence of this nematode is available.

More recently, we explored transcriptional changes in the early stage of development of the canine hookworm, *A. caninum*, using suppressive-subtractive hybridization (SSH)-based EST sequencing, followed by bioinformatic analysis with ESTExplorer. The L3s of *A. caninum* endure a period of apparent "arrested development" preceding transmission to a suitable host. Many of the mRNAs that are "up-regulated" at this stage of development are likely to encode proteins that act at the host–parasite interface and facilitate the transition from a free-living to a parasitic larva. The initial phase of the percutaneous infection of the canid host by *A. caninum* relates to the exsheathment and "activation" of the L3 stage, which can be mimicked *in vitro* by incubating L3s in culture medium containing dog serum. The mRNAs differentially transcribed between non-activated and activated L3s were identified by SSH. The analysis of these mRNAs on a customized oligonucleotide microarray, printed with the SSH-derived and publicly available *A. caninum* ESTs (non-subtracted), yielded a total of 602 differentially transcribed molecules, of which the most highly represented sequences ($n=27$) encoded products belonging to the pathogenesis-related protein (PRP) superfamily and different mechanistic classes of proteases. Comparison of these *A. caninum* mRNAs with those of *C. elegans* larvae exiting developmental arrest (or dauer) demonstrated large differences with respect to GO profiles. *C. elegans* L3s exiting developmental arrest were inferred to be linked to an up-regulated expression of collagens and other (mostly intracellular) molecules involved in growth and development. Such mRNAs are virtually absent from activated hookworm larvae, and, instead, are represented by an inordinately large number of mRNAs encoding extracellular proteins, suggesting that many of the activation-associated hookworm mRNAs are involved in host–parasite interactions. The near absence of mRNAs associated with reproduction, growth and development from activated hookworm L3s probably reflects their ability to further arrest (i.e. undergo hypobiosis) in tissues of the canid host, non-permissive hosts or in the external environment when conditions for transmission are unfavourable. Further investigations into the regulation of PRPs will provide a deeper understanding of the biological functions of these parasitism-associated genes. Although this information does not necessarily invalidate *C. elegans* dauer exit as a model for hookworm activation, it suggests a limitation of this free-living nematode as a model organism for the transition of hookworm larvae from a free-living to a parasitic state.

In the common roundworm of pigs, *A. suum*, transcription specific to the L3 stage was investigated also using an SSH-based sequencing approach, followed by detailed bioinformatic analyses (Huang et al., 2008). In brief, a cDNA archive enriched for molecules in the L3 of *A. suum* was constructed by SSH, and a subset of cDNAs from 3075 clones subjected to microarray analysis using probes derived from RNA from different developmental stages of *A. suum*. The cDNAs ($n=498$) shown by microarray analysis to be enriched in the L3 were sequenced and subjected to bioinformatic analyses using ESTExplorer. Using GO, 235 of these molecules were assigned to 'biological process' ($n=68$), 'cellular component' ($n=50$) or 'molecular function' ($n=117$). Of the 91 clusters assembled, 56 molecules (61.5%) had homologues/orthologues in the free-living nematodes *C. elegans* and *C. briggsae* and/or other organisms, whereas 35 (38.5%) had no significant similarity to any sequences available in current gene databases. Transcripts encoding protein kinases, protein phosphatases (and their precursors) and enolases were abundantly represented in the L3 of *A. suum*, as were molecules involved in cellular processes, such as ubiquitination/proteasome function, gene transcription, and protein–protein interactions and function. *In silico* analyses inferred the *C. elegans* orthologues/homologues ($n=50$) to be involved in apoptosis and insulin signalling (2%), ATP synthesis (2%), carbon metabolism (6%), fatty acid biosynthesis (2%), gap junction (2%), glucose metabolism (6%) or porphyrin metabolism (2%), although 34 (68%) of them could not be mapped to a specific metabolic pathway. Small numbers of these 50 molecules were predicted to be secreted (10%), anchored (2%) and/or transmembrane (12%) proteins. Functionally, 17 (34%) of them were predicted to be associated with (non-wild-type) RNAi phenotypes in *C. elegans*, the majority being Emb (embryonic lethality; 13 types; 58.8%), Lva (larval arrest; 23.5%) and Lvl (larval lethality; 47%). A genetic interaction network was predicted for these 17 *C. elegans* orthologues, revealing highly significant interactions for nine molecules associated with embryonic and larval development (66.9%), information storage and processing (5.1%), cellular processing and signalling (15.2%), metabolism (6.1%) and unknown function (6.7%). The roles of these molecules in development are suggested to be linked to those of their homologues/orthologues in *C. elegans* and some other organisms. The results of this study (Huang et al., 2008) provided a basis for future functional genomic studies to elucidate molecular aspects governing larval developmental processes in *A. suum* and/or the transition to parasitism.

6. Establishing a workflow system for high-throughput prediction of ES proteins from ESTs datasets

Molecules excreted or secreted by a cell or an organism, often referred to excretory/secretory (ES) products, play pivotal biological roles in parasitic helminths (e.g., Lightowlers and Rickard, 1988). ES proteins can represent $8\pm20\%$ of the proteome of an organism (Greenbaum et al., 2001) and can include functionally diverse classes of molecules, such as cytokines, chemokines, hormones, digestive enzymes, antibodies, extracellular proteinases, morphogens, toxins and antimicrobial peptides. Some of these proteins are known to be involved in vital biological processes, including cell adhesion, cell

migration, cell–cell communication, differentiation, proliferation, morphogenesis and immune responses (Maizels and Yazdanbakhsh, 2003). ES proteins can circulate throughout the body of an organism (e.g., in the extracellular space) or are localized to or released from the cell surface, making them readily accessible to drugs and/or the immune system. From a fundamental perspective, proteins secreted by parasites are of particular interest in relation to interactions with the host, because they are present or active at the interface between the parasite and host cells, and can regulate or mediate the host responses and/or cause disease (Kamoun, 2006; Muller et al., 2008). These latter characteristics make these molecules attractive targets for novel vaccines or therapeutics (Bonin-Debs et al., 2004).

ES proteins have long been the focus of biochemical and immunological studies of parasitic helminths; such worms secrete biologically active mediators which can modify or customize their niche within the host, in order to evade immune attack or to regulate or stimulate a particular host response (e.g., Lightowlers and Rickard, 1988; Hawdon et al., 1996; Maizels et al., 2001; Yatsuda et al., 2003). Parasitic nematodes are responsible for a range of neglected tropical diseases, such as ancylostomatosis, necatoriasis, lymphatic filariasis, onchocerciasis, ascariasis and strongyloidiasis in humans (Engels and Savioli, 2006; Hotez et al., 2007), and others can cause massive production or economic losses to farmers as well as to animal and plant industries (e.g., Koenning et al., 1999; McLeod, 1995; Smith, 1997; Coles, 2001; Sackett and Holmes, 2006).

There have been efforts to identify and characterize ES proteins in different parasitic nematodes in various studies. For instance, Robinson and Connolly (2005) used a proteomic approach to identify ES glycoproteins in *T. spiralis*, an enoplid nematode (or trichina) of musculature. In another study, Yatsuda et al. (2003) undertook an analysis of ES products from *H. contortus* (barber's pole worm); these authors identified a range of proteins but were only able (based on comparative analysis) to investigate known proteins, such as serine, metallo- and aspartyl-proteases and the microsomal peptidase H11, a vaccine candidate, previously recognised as a "hidden antigen" (Munn et al., 1997). The precise role of ES proteins from parasitic nematodes in mediating cellular processes is largely unknown due to the difficulty in experimentally assigning function to individual proteins (Robinson and Connolly, 2005). In this context, computational approaches applied to identify and annotate ES proteins have significantly complemented experimental studies of different cell types, tissues, organs and organisms. For example, in an early study, Grimmond et al. (2003) developed a computational strategy to identify and functionally classify secreted proteins in the mouse, based on the presence of a cleavable signal peptide (required for its entry into the secretory pathway), along with the lack of any transmembrane (TM) domain or intracellular localization signals in full-length molecules. This study was followed by the computational reconstruction of the secretome in human skeletal muscle from protein sequence data by Bortoluzzi et al. (2006). Also, Martinez et al. (2006) identified and annotated the secreted proteins involved in the early development of the kidney in the mouse from microarray 'expression' profiling, using computational strategies.

While EST data have been mined for many interesting functional molecules (e.g., Adams et al., 1993; Nagaraj et al., 2007a), predicting ES proteins from ESTs has been relatively uncommon. For example, Vanholme et al. (2006) identified putative secreted proteins from EST datasets for the plant parasitic nematode, *Heterodera schachtii*. Harcus et al. (2004) investigated the signal sequences inferred from the EST data for the parasitic nematode *Nippostrongylus brasiliensis*, and linked them to an "accelerated evolution" of secreted proteins in this parasite, compared with host or non-parasitic organisms. Also, Ranganathan et al. (2007) inferred ES proteins from EST data for the bovine lungworm, *D. viviparus*, while several mRNAs encoding novel secreted proteins without known homologues were identified from ESTs of *A. caninum* (see Datu et al., 2008). Nagaraj et al. (2008a) identified and classified putative secreted proteins from *T. vitrinus*, a parasitic nematode of ruminants and suggested some molecules as targets for novel anthelmintics. One of the suggested molecules, *Tv-stp-1*, was investigated further and its functionality predicted (Hu et al., 2007).

While single EST or protein datasets have been examined for the presence of secretory or ES proteins, large-scale analysis has not been conducted to date, due to the lack of effective high throughput, computational pipelines for analysis (Martinez et al., 2006). Based on recent work (Ranganathan et al., 2007; Datu et al., 2008; Nagaraj et al., 2008a), ESTExplorer has been modified to predict ES proteins with high confidence, and to provide extensive annotation, including GOs, pathway mapping, protein domain identification and the prediction of protein–protein interactions. The enhanced pipeline, called EST2Secretome (Nagaraj et al., 2008b), is a freely available web server that can directly process EST datasets or entire proteomes.

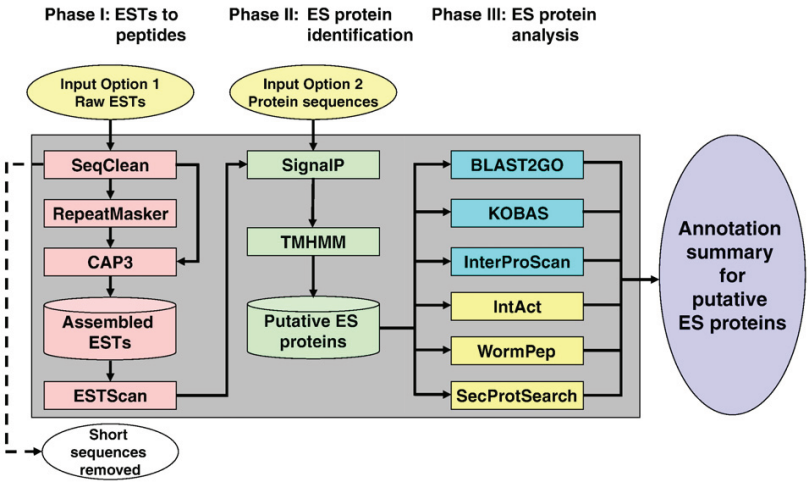


Fig. 2. EST2Secretome analysis and annotation workflow, showing Phase I (pre-processing, assembly and conceptual translation), Phase II (prediction of secreted proteins) and Phase III (secreted protein annotation).

EST2Secretome (http://EST2secretome.biolinfo.org/) is a comprehensive workflow system, comprising carefully selected computational tools to identify and annotate ES proteins inferred from ESTs. It provides a user-friendly interface and detailed online help to assist researchers in the analysis of EST datasets for ES proteins. The workflow can be divided into three phases (Fig. 2), with Phase I dedicated to pre-processing, assembly and conceptual translation, similar to that of ESTExplorer (see Nagaraj et al., 2008b). In Phase II, putative ES proteins are identified based on the presence of signal sequences and the absence of transmembrane helices. Phase III contains an annotation layer, comprising a suite of bioinformatic tools to annotate the ES proteins inferred in Phase II. ESTs can be submitted to Phase I for EST pre-processing, assembly and conceptual translation, followed by the identification of putative ES proteins in Phase II and annotation in Phase III. Alternatively, instead of EST data, protein sequences may be submitted directly to Phase II to identify putative ES proteins and to functionally annotate them in Phase III.

In Phase I, EST2Secretome shares programs SeqClean, RepeatMasker and CAP3 with ESTExplorer (Nagaraj et al., 2007b), based on the analysis presented earlier (Nagaraj et al., 2007a). The contig and singleton sequences generated *via* CAP3 are transferred to the program ESTScan for conceptual translation into proteins, using the genetic code from a related organism. Currently, EST2Secretome implements the genetic codes for 15 organisms, covering the most studied organisms including human, mouse, rat, pig, dog, chicken, rice, wheat, thale cress, zebrafish, fly, yeast and *C. elegans*.

In Phase II, putative ES proteins are inferred from the protein sequences generated in Phase I, using the two programs SignalP (Bendtsen et al., 2004) and TMHMM (Emanuelsson et al., 2007). SignalP first checks whether an N-terminal signal sequence is predicted based on both artificial neural network and hidden Markov model probability scores (SignalPNN and SignalP-HMM), using default parameters that can be modified by experienced users. Subsequently, all proteins with signal sequences are passed on to TMHMM, a hidden Markov model-based transmembrane helix prediction program, to "filter out" transmembrane proteins. The subset lacking transmembrane helices is selected as putative ES proteins for further annotation.

Phase III is the annotation layer, comprising a suite of six computational tools for the functional annotation of the predicted ES proteins, of which the first three (GO using BLAST2GO, InterProScan and pathway mapping employing KOBAS) are also implemented in ESTExplorer (Nagaraj et al., 2007b). The other three components are unique to EST2Secretome and incorporate protein BLAST searches against three different datasets derived from Wormpep (www.sanger.ac.uk/Projects/C_elegans/WORMBASE/current/wormpep.shtml) for the identification of nematode homologues, IntAct (Kerrien et al., 2007) for protein–protein interaction data and a non-redundant secreted protein database (SecProtSearch) derived from the literature, the secreted protein database, SPD (Chen et al., 2005) and the manually curated signal peptide database, SPdb (Choo et al., 2005). Mapping to Wormpep gives a list of homologous proteins in *C. elegans*, and is linked to WormBase (Bieri et al., 2007). Homologues from the IntAct database are determined using the concept of interlogs ("evolutionarily conserved interactions" identified based on conservation among homologous proteins in different species) and are linked to all molecular interaction partners of homologous proteins. EST2Secretome provides a link to the relevant interlog page at IntAct, containing all interaction partners. The interaction data culled from these interlogs can be extrapolated to predict protein interactions of the query sequence, for validation by RNAi, gene deletion or fluorescence-based interaction studies. The final module compares the query sequence to a specialised dataset of known secreted proteins (SecProtSearch), in order to identify orthologous secreted proteins, which would provide a second level of validation for the ES protein dataset. Phase III (Fig. 2) thus allows extensive characterization and validation of ES proteins predicted by EST2Secretome.

Once an EST (or protein) dataset has been submitted to EST2Secretome, a status page is accessible for the monitoring of the progress of analysis, at the program level. As each selected program is completed, the status page is updated and the output from the program is displayed. The outcome from each run is summarized, with links to output files from each selected program being listed. When a large dataset is analysed using a workflow, it is challenging to collate the results of the analysis from multiple steps. To address this issue, EST2Secretome provides a summary file for each ES protein, comprising the assembled contig/singleton sequence, the peptide sequence and all of the annotations (such as homologous proteins, protein domains, pathways and interaction partners). A detailed tutorial, frequently asked questions (FAQ) and sample EST and protein datasets are available online for the effective use of EST2Secretome. Thus, EST2Secretome extends the analysis and annotations capabilities of ESTExplorer by including specific tools relevant parasitic helminths, including the searching of WormPep, IntAct and secreted proteins databases, in order to enhance the annotation of ES proteins considered relevant in understanding parasite–host interactions.

7. Mining publicly available EST datasets for ES proteins of nematodes, as potential drug or vaccine targets

Recently, EST2Secretome was applied to 452134 ESTs, representing 39 economically important and disease-causing parasitic nematodes of humans, other animals and plants, to conduct a comprehensive analysis and detailed annotation of inferred ES proteins, also with specific reference to candidate molecules already being assessed as intervention targets (Nagaraj et al., 2008b). We predicted 4710 ES proteins from these ESTs. In total, 2632, 786 and 1292 ES proteins were predicted for animal-, human- and plant-parasitic nematodes. Subsequently, we systematically analysed ES proteins using computational methods. Of these 4710 proteins, 2490 (52.8%) had homologues in *C. elegans*, whereas 621 (13.8%) appeared to be novel, currently having no significant match to any molecules available in public databases. Of the *C. elegans* homologues, 267 were linked to strong 'loss-of-function' phenotypes by RNAi in this nematode. We could functionally annotate 1948 (41.3%) sequences using the GO terms, established pathway associations for 573 (12.2%) sequences using KEGG and predicted protein interaction partners for 1774 (37.6%) molecules. We also mapped 758 (16.1%) inferred peptides to protein domains, including the nematode-specific protein families "transthyretin-like" and "chromadorea ALT," presently considered as vaccine candidates against filariasis in humans (cf. Nagaraj et al., 2008b). In conclusion, this set of ES proteins provides an inventory of known and novel members of ES proteins, to support studies on the molecular biology of parasitic nematodes and their interactions with their hosts as well as for the development of novel drugs or vaccines for parasite intervention.

8. Conclusions and implications

We have described a systematic approach to the analysis of EST data from parasitic nematodes, applicable to a myriad EST datasets currently available. The platforms, ESTExplorer and EST2Secretome, are user-friendly and allow relevant analyses and annotations of data. These platforms can be logically extended for the analysis of a wide range of parasites, to identify novel molecules of nematodes implicated in host–parasite interactions. Our focus on N-terminal signal sequence-mediated, secreted protein identification is in accordance with other recent EST analyses of datasets for the root-knot nematode, *Meloidogyne chitwoodi* (see Roze et al., 2008) and the filarial nematode, *Brugia malayi* (see Hewitson et al., 2008). However, from combined transcriptomic and proteomic analyses of *B. malayi*,

Moreno and Geary (2008) reported the presence of several proteins that may be secreted through non-classical secretory pathways. Hence, additional tools for the inference of secreted proteins *in toto* in helminth parasites need to be developed and incorporated into EST2Secretome. New tools, such as SNP discovery (miraEST; Chevreux et al., 2004), digital gene expression profiling (GBA server; Wu et al., 2005) and the identification of molecular markers (GemProspector; Fredslund et al., 2006), could also be integrated. The accurate alignment of ESTs to their cognate or related genomic sequences is a prerequisite to the study of alternative splicing, gene discovery projects and comparative genomics. The incorporation of efficient tools (e.g., BLAT, Kent, 2002; MGAlign, Lee et al., 2003; GMAP, Wu and Watanabe, 2005) with the option of submitting EST data to ESTExplorer for alignment to a genome sequence of interest will enhance its annotation capabilities.

New generation sequencing technologies, such as 454-Roche (www.454.com; Margulies et al., 2005), ABI-SOLiD (www.appliedbiosystems.com; Pandey et al., 2008), Illumina-Solexa (www.illumina.com; Bentley et al., 2008) and Helicos (www.helicosbio.com; Harris et al., 2008), provide unique prospects for high-throughput transcriptomic and genomic explorations of parasitic helminths, yielding enormous amounts of data at substantially lower cost than incurred for conventional (Sanger) sequencing. This advance now requires substantial enhancements to be made to existing platforms for the efficient and accurate assembly and annotation of such large datasets at the DNA and protein levels, enriched with GO, protein domain and pathway information.

On the other hand, novel targets for anthelmintic drugs and/or vaccines have been proposed based on genomic–bioinformatic analyses. Importantly, numerous putative ES proteins containing nematode-specific domains and their absence from the respective host(s) are critical requirements to further select key subsets of molecules for characterization and pre-validation. Results from proteomic or biochemical studies will further enable the determination of threshold parameters for EST data analysis within an improved bioinformatic pipeline, as well as the validation of the protocols for the identification of ES proteins and other key molecular groups. The development of substantially enhanced high-throughput pipelines, capable of addressing detailed nucleotide and protein annotations as well as ES protein identification for parasitic nematodes, should accelerate biotechnology research, also by focussing the attention of the experimental biologist/parasitologist on key genes and gene products with potential as therapeutics or vaccines.

Acknowledgements

Original research was partially supported by grants from the Australian Research Council (ARC) (LP0667795 and DP0665230), Genetic Technologies Limited, and Meat and Livestock Australia.

References

Adams MD, Kerlavage AR, Fields C, Venter JC. 3,400 new expressed sequence tags identify diversity of transcripts in human brain. Nat Genet 1993;4:256–67.
Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.
Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25–9.
Ayoubi P, Jin X, Leite S, Liu X, Martajaja J, Abduraham A, et al. PipeOnline 2.0: automated EST processing and functional data sorting. Nucleic Acids Res 2002;30:4761–9.
Bendtsen JD, Nielsen H, von Heijne G, Brunak S. Improved prediction of signal peptides, SignalP 3.0. J Mol Biol 2004;340:783–95.
Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008;456:53–9.
Bieri T, Blasiar D, Ozersky P, Antoshechkin I, Bastiani C, Canaran P, et al. WormBase, new content and better access. Nucleic Acids Res 2007;35:D506–10.
Blaxter ML, De Ley P, Garey JR, Liu LX, Scheldeman P, Vierstraete A, et al. A molecular evolutionary framework for the phylum Nematoda. Nature 1998;392:71–5.
Boag PR, Ren P, Newton SE, Gasser RB. Molecular characterisation of a male-specific serine/threonine phosphatase from *Oesophagostomum dentatum* (Nematoda; Strongylida) and functional analysis of homologues in *Caenorhaboditis elegans*. Int J Parasitol 2003;33:313–25.
Bonin-Debs AL, Boche I, Gille H, Brinkmann U. Development of secreted proteins as biotherapeutic agents. Expert Opin Biol Ther 2004;4:551–8.
Bortoluzzi S, Scannapieco P, Cestaro A, Danieli GA, Schiaffino S. Computational reconstruction of the human skeletal muscle secretome. Proteins 2006;62:776–92.
Campbell BE, Nagaraj SH, Hu M, Zhong W, Sternberg PW, Ong EK, et al. Gender-enriched transcripts in *Haemonchus contortus* — predicted functions and genetic interactions based on comparative analyses with *Caenorhabditis elegans*. Int J Parasitol 2008;38:65–83.
Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WE, Wetter T, et al. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res 2004;14:1147–59.
Chen Y, Zhang Y, Yin Y, Gao G, Li S, Jiang Y, et al. SPD—a web-based secreted protein database. Nucleic Acids Res 2005;33:D169–73.
Choo KH, Tan TW, Ranganathan S. SPdb—a signal peptide database. BMC Bioinformatics 2005;6:249.
Coles GC. The future of veterinary parasitology. Vet Parasitol 2001;98:31–9.
Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 2005;21:3674–6.
Coyne MJ, Smith G. The mortality and fecundity of *Haemonchus contortus* in parasite-naive and parasite-exposed sheep following single experimental infections. Int J Parasitol 1992;22:315–25.
Datu BJ, Gasser RB, Nagaraj SH, Ong EK, O'Donoghue P, McInnes R, et al. Transcriptional changes in the hookworm, *Ancylostoma caninum*, during the transition from a free-living to a parasitic Larva. PLoS Negl Trop Dis 2008;2:e130.
D'Agostino N, Aversano M, Chiusano ML. ParPEST: a pipeline for EST data analysis based on parallel computing. BMC Bioinformatics 2005;6(Suppl 4):S9.
Emanuelsson O, Brunak S, von Heijne G, Nielsen H. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2007;2:953–71.
Engels D, Savioli L. Reconsidering the underestimated burden caused by neglected tropical diseases. Trends Parasitol 2006;22:363–6.
Forment J, Gilabert F, Robles A, Conejero V, Nuez F, Blanca JM. EST2uni: an open, parallel tool for automated EST analysis and database creation, with a data mining web interface and microarray expression data integration. BMC Bioinformatics 2008;9:5.
Fredslund J, Madsen LH, Hougaard BK, Sandal N, Stougaard J, Bertioli D, et al. GeMprospector—online design of cross-species genetic marker candidates in legumes and grasses. Nucleic Acids Res 2006;34:W670–5.
Gasser RB, Newton SE. Genomic and genetic research on bursate nematodes, significance, implications and prospects. Int J Parasitol 2000;30:509–34.
Greenbaum D, Luscombe NM, Jansen R, Qian J, Gerstein M. Interrelating different types of genomic data, from proteome to secretome, 'oming in on function. Genome Res 2001;11:1463–8.
Grimmond SM, Miranda KC, Yuan Z, Davis MJ, Hume DA, Yagi K, et al. The mouse secretome, functional classification of the proteins secreted into the extracellular environment. Genome Res 2003;13:1350–9.
Harcus YM, Parkinson J, Fernandez C, Daub J, Selkirk ME, Blaxter ML, et al. Signal sequence analysis of expressed sequence tags from the nematode *Nippostrongylus brasiliensis* and the evolution of secreted proteins in parasites. Genome Biol 2004;5:R39.
Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, et al. Single-molecule DNA sequencing of a viral genome. Science 2008;320:106–9.
Hawdon JM, Jones BF, Hoffman DR, Hotez PJ. Cloning and characterization of *Ancylostoma*-secreted protein. A novel protein associated with the transition to parasitism by infective hookworm larvae. J Biol Chem 1996;271:6672–8.
Hewitson JP, Harcus YM, Curwen RS, Dowle AA, Atmadja AK, Ashton PD, Wilson A, Maizels RM. The secretome of the filarial parasite, *Brugia malayi*, proteomic profile of adult excretory–secretory products. Mol Biochem Parasitol 2008;160:8-21.
Hotez PJ, Molyneux DH, Fenwick A, Kumaresan J, Sachs SE, Sachs JD, et al. Control of neglected tropical diseases. N Engl J Med 2007;357:1018–27.
Hotz-Wagenblatt A, Hankeln T, Ernst P, Glatting KH, Schmidt ER, Suhai S. ESTAnnotator: a tool for high throughput EST annotation. Nucleic Acids Res 2003;31:3716–9.
Hu M, Abs EL-Osta YG, Campbell BE, Boag PR, Nisbet AJ, Beveridge I, et al. *Trichostrongylus vitrinus* (Nematoda, Strongylida), molecular characterization and transcriptional analysis of *Tv-stp-1*, a serine/threonine phosphatase gene. Exp Parasitol 2007;117:22–34.
Huang X, Madan A. CAP3: a DNA sequence assembly program. Genome Res 1999;9:868–77.
Huang CQ, Gasser RB, Cantacessi C, Nisbet AJ, Zhong W, Sternberg PW, et al. Genomic–bioinformatic analysis of transcripts enriched in the third-stage larva of the parasitic nematode *Ascaris suum*. PLoS Negl Trop Dis 2008;2:e246.
Iseli C, Jongeneel CV, Bucher P. ESTScan, a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc Int Conf Intell Syst Mol Biol 1999:138–48.
Kamoun S. A catalogue of the effector secretome of plant pathogenic oomycetes. Annu Rev Phytopathol 2006;44:41–60.
Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, et al. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 2006;34:D354–7.
Kent WJ. BLAT—the BLAST-like alignment tool. Genome Res 2002;12:656–64.
Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, et al. IntAct—open source resource for molecular interaction data. Nucleic Acids Res 2007;35:D561–5.
Kikuchi T, Aikawa T, Kosaka H, Pritchard L, Ogura N, Jones JT. Expressed sequence tag (EST) analysis of the pine wood nematode *Bursaphelenchus xylophilus* and *B. mucronatus*. Mol Biochem Parasitol 2007;155:9-17.

Koenning SR, Overstreet C, Noling JW, Donald PA, Becker JO, Fortnum BA. Survey of crop losses in response to phytoparasitic nematodes in the United States for 1994. J Nematol 1999;31:587–618.
Kumar CG, LeDuc R, Gong G, Roinishivili L, Lewin HA, Liu L. ESTIMA, a tool for EST management in a multi-project environment. BMC Bioinformatics 2004;5:176.
Latorre M, Silva H, Saba J, Guziolowski C, Vizoso P, Martinez V, et al. JUICE: a data management system that facilitates the analysis of large volumes of information in an EST project workflow. BMC Bioinformatics 2006;7:513.
Lee BT, Tan TW, Ranganathan S. MGAlignIt. A web service for the alignment of mRNA/EST and genomic sequences. Nucleic Acids Res 2003;31:3533–6.
Lee B, Hong T, Byun SJ, Woo T, Choi YJ. ESTpass: a web-based server for processing and annotating expressed sequence tag (EST) sequences. Nucleic Acids Res 2007;35:W159–62.
Lightowlers MW, Rickard MD. Excretory–secretory products of helminth parasites, effects on host immune responses. Parasitology 1988;96:S123–66 Suppl.
Maizels RM, Yazdanbakhsh M. Immune regulation by helminth parasites, cellular and molecular mechanisms. Nat Rev Immunol 2003;3:733–44.
Maizels RM, Gomez-Escobar N, Gregory WF, Murray J, Zang X. Immune evasion genes from filarial nematodes. Int J Parasitol 2001;31:889–98.
Mao C, Cushman JC, May GD, Weller JW. ESTAP—an automated system for the analysis of EST data. Bioinformatics 2003;19:1720–2.
Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005;437:376–80.
Martinez G, Georgas K, Challen GA, Rumballe B, Davis MJ, Taylor D, et al. Definition and spatial annotation of the dynamic secretome during early kidney development. Dev Dyn 2006;235:1709–19.
Masoudi-Nejad A, Tonomura K, Kawashima S, Moriya Y, Suzuki M, Itoh M, et al. EGassembler: online bioinformatics service for large-scale processing, clustering and assembling ESTs and genomic DNA fragments. Nucleic Acids Res 2006;34:W459–62.
McLeod RS. Costs of major parasites to the Australian livestock industries. Int J Parasitol 1995;25:1363–7.
Mitreva M, Jasmer DP, Appleton J, Martin J, Dante M, Wylie T, et al. Gene discovery in the adenophorean nematode *Trichinella spiralis*, an analysis of transcription from three life cycle stages. Mol Biochem Parasitol 2004a;137:277–91.
Mitreva M, McCarter JP, Martin J, Dante M, Wylie T, Chiapelli B, et al. Comparative genomics of gene expression in the parasitic and free-living nematodes *Strongyloides stercoralis* and *Caenorhabditis elegans*. Genome Res 2004b;14:209–20.
Mitreva M, Blaxter ML, Bird DM, McCarter JP. Comparative genomics of nematodes. Trends Genet 2005a;21:573–81.
Mitreva M, McCarter JP, Arasu P, Hawdon J, Martin J, Dante M, et al. Investigating hookworm genomes by comparative analysis of two *Ancylostoma* species. BMC Genomics 2005b;6:58.
Mitreva M, Appleton J, McCarter JP, Jasmer DP. Expressed sequence tags from life cycle stages of *Trichinella spiralis*, application to biology and parasite control. Vet Parasitol 2005c;132:13–7.
Mitreva M, Zarlenga DS, McCarter JP, Jasmer DP. Parasitic nematodes — from genomes to control. Vet Parasitol 2007;148:31–42.
Moreno Y, Geary TG. Stage- and gender-specific proteomic analysis of *Brugia malayi* excretory–secretory products. PLoS Negl Trop Dis 2008;2:e326.
Muller O, Schreier PH, Uhrig JF. Identification and characterization of secreted and pathogenesis-related proteins in *Ustilago maydis*. Mol Genet Genomics 2008;279:27–39.
Munn EA, Smith TS, Smith H, James FM, Smith FC, Andrews SJ. Vaccination against *Haemonchus contortus* with denatured forms of the protective antigen H11. Parasite Immunol 1997;19:243–8.
Nagaraj SH, Gasser RB, Ranganathan S. A hitchhiker's guide to expressed sequence tag (EST) analysis. Brief Bioinform 2007a;8:6-21.
Nagaraj SH, Deshpande N, Gasser RB, Ranganathan S. ESTExplorer, an expressed sequence tag (EST) assembly and annotation platform. Nucleic Acids Res 2007b;35:W143–7.
Nagaraj SH, Gasser RB, Nisbet AJ, Ranganathan S. *In silico* analysis of expressed sequence tags from *Trichostrongylus vitrinus* (Nematoda), comparison of the automated ESTExplorer workflow platform with conventional database searches. BMC Bioinformatics 2008a;9(Suppl):S10.
Nagaraj SH, Gasser RB, Ranganathan S. Needles in the EST Haystack: large-scale identification and analysis of excretory–secretory (ES) proteins in parasitic nematodes using expressed sequence tags (ESTs). PLoS Negl Trop Dis 2008b;2:e301.
Newton SE, Boag PR, Gasser RB. Opportunities and prospects for investigating developmentally regulated and sex-specific genes and their expression in intestinal nematodes of humans. In, Holland, Kennedy CV, MW, editors. World Class Parasites, Vol.2. The Geohelminths, Ascaris, Trichuris and Hookworm. Kluwer Academic Publishers. Boston/Dordrecht/London, 2002. pp. 235–68.
Nikolaou S, Gasser RB. Prospects for exploring molecular developmental processes in *Haemonchus contortus*. Int J Parasitol 2006;36:859–68.
Nisbet AJ, Gasser RB. Profiling of gender-specific gene expression for *Trichostrongylus vitrinus* (Nematoda, Strongylida) by microarray analysis of expressed sequence tag libraries constructed by suppressive–subtractive hybridisation. Int J Parasitol 2004;34:633–43.
Pandey V, Nutter RC, Prediger E. Applied Biosystems SOLiD™ System: Ligation-Based Sequencing, Next Generation Genome Sequencing: Towards Personalized Medicine. Wiley; 2008. p. 29–41.
Paquola AC, Nishyiama Jr MY, Reis EM, da Silva AM, Verjovski-Almeida S. ESTWeb: bioinformatics services for EST sequencing projects. Bioinformatics 2003;19:1587–8.
Parkinson J, Blaxter M. SimiTri—visualizing similarity relationships for groups of sequences. Bioinformatics 2003;19:390–5.
Parkinson J, Mitreva M, Whitton C, Thomson M, Daub J, Martin J, et al. A transcriptomic analysis of the phylum Nematoda. Nat Genet 2004a;36:1259–67.
Parkinson J, Anthony A, Wasmuth J, Schmid R, Hedley A, Blaxter M. PartiGene—constructing partial genomes. Bioinformatics 2004b;20:1398–404.
Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, et al. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 2003;19:651–2.
Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al. InterProScan: protein domains identifier. Nucleic Acids Res 2005;33:W116–20.
Ranganathan S, Nagaraj SH, Hu M, Strube C, Schnieder T, Gasser RB. A transcriptomic analysis of the adult stage of the bovine lungworm, *Dictyocaulus viviparus*. BMC Genomics 2007;8:311.
Robinson MW, Connolly B. Proteomic analysis of the excretory–secretory proteins of the *Trichinella spiralis* L1 larva, a nematode parasite of skeletal muscle. Proteomics 2005;5:25–32.
Rossanigo CE, Gruner L. The length of strongylid nematode infective larvae as a reflection of developmental conditions in faeces and consequences on their viability. Parasitol Res 1996;82:304–11.
Roze E, Hanse B, Mitreva M, Vanholme B, Bakker J, Smant G. Mining the secretome of the root-knot nematode *Meloidogyne chitwoodi* for candidate parasitism genes. Mol Plant Pathol 2008;9:1-10.
Rual JF, Ceron J, Koreth J, Hao T, Nicot AS, Hirozane-Kishikawa T, et al. Toward improving *Caenorhabditis elegans* phenome mapping with an ORFeome-based RNAi library. Genome Res 2004;14:162-2168.
Rudd S. openSputnik—a database to ESTablish comparative plant genomics using unsaturated sequence collections. Nucleic Acids Res 2005;33:D622–7.
Sackett D, Holmes P. Assessing the Economic Cost of Endemic Disease on the Profitability Australian Beef Cattle and Sheep Producers. Sydney, Australia: Meat and Livestock Australia Limited1741910021; 2006.
Schmid R, Blaxter ML. annot8r: GO, EC and KEGG annotation of EST datasets. BMC Bioinformatics 2008;9:180.
Smith G. The economics of parasite control, obstacles to creating reliable models. Vet Parasitol 1997;72:37–44.
Stoll NR. Studies with the strongyloid nematode, *Haemonchus contortus*. Am J Hyg 1929;10:384–418.
Vanholme B, Mitreva M, Van Criekinge W, Logghe M, Bird D, McCarter JP, et al. Detection of putative secreted proteins in the plant-parasitic nematode *Heterodera schachtii*. Parasitol Res 2006;98:414–24.
Waegele B, Schmidt T, Mewes HW, Ruepp A. OREST: the online resource for EST analysis. Nucleic Acids Res 2008;36:W140–4.
Wu TD, Watanabe CK. GMAP, a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 2005;21:1859–75.
Wu X, Walker MG, Luo J, Wei L. GBA server, EST-based digital gene expression profiling. Nucleic Acids Res 2005;33:W673–6.
Wu J, Mao X, Cai T, Luo J, Wei L. KOBAS server: a web-based platform for automated annotation and pathway identification. Nucleic Acids Res 2006;34:W720–4.
Yatsuda AP, Krijgsveld J, Cornelissen AW, Heck AJ, de Vries E. Comprehensive analysis of the secreted proteins of the parasite *Haemonchus contortus* reveals extensive sequence variation and differential immune recognition. J Biol Chem 2003;278:16941–51.

Chapter 2: Methods and Applications

Methods and applications that were developed and used in this study are summarised in Table 2.1. The ensuing publications have also been listed and included in the relevant chapter.

Table 2.1: Methods, applications and publications

Methods/Applications	Chapter	Refer to Publication
Advanced *in silico* analysis of expressed sequence tag (EST) data for parasitic nematodes of major socio-economic importance--fundamental insights toward biotechnological outcomes.	1	1
An integrated transcriptomics and proteomics analysis of the secretome of the helminth pathogen *Fasciola hepatica*: proteins associated with invasion and infection of the mammalian host.	3	2
An analysis of the transcriptome of *Teladorsagia circumcincta*: its biological and biotechnological implications.	4	3
TranSeqAnnotator: Large-scale analysis of transcriptomic data.	5	4
Comparison of adult and fourth larval stage transcriptomic data: *Teladorsagia circumcincta*.	6	5

Chapter 3: An integrated transcriptomics and proteomics analysis of the secretome of the helminth pathogen, *Fasciola hepatica*: proteins associated with invasion and infection of the mammalian host

3.1 Summary

Fasciola hepatica, a parasitic flatworm of livestock, particularly sheep and cattle causes the disease, fascioliasis and has recently known to infect the liver of various mammals, including humans, leading to liver cancer. Identifying novel targets for anthelmintic compounds and potential antiparasite vaccine candidates is essential with a better perceptive of biochemical and molecular interactions between hosts and parasites.

We were able to integrate available transcriptomic data with proteomic analysis, emphasizing on proteases, to understand the complexities involved in the ability of a developing parasite to sustain itself within the mammalian host. The unannotated adult *F. hepatica* ESTs (14,031) from Wellcome Trust Sanger Centre were taken for the analysis of secretory proteins potentially involved in host-pathogen interactions, as only a few and mainly redundant *Fasciola* nucleotide sequences are available in Genbank. The significant findings of this analysis are the identification of novel proteases with important roles in host-parasite interplay, mainly cathepsin L and cathepsin B cysteine proteases and carboxypeptidase and trypsin-like serine proteases (members of two serine protease families). We could propose that a range of abundant and highly regulated antioxidants are released *via* a non-classical trans-tegumental pathway unlike the major proteases of *F. hepatica* secreted into the parasite gut via classical ER/Golgi pathway. Overall, the overview of the protein secretion by *F. hepatica* with transcriptomic and proteomic analysis approach represents a significant step in understanding the host-parasite interactions in fasciolosis.

An Integrated Transcriptomics and Proteomics Analysis of the Secretome of the Helminth Pathogen *Fasciola hepatica*

PROTEINS ASSOCIATED WITH INVASION AND INFECTION OF THE MAMMALIAN HOST*[S]

Mark W. Robinson‡§, Ranjeeta Menon¶, Sheila M. Donnelly‡‖, John P. Dalton‡‖‡‡, and Shoba Ranganathan¶****

To infect their mammalian hosts, *Fasciola hepatica* larvae must penetrate and traverse the intestinal wall of the duodenum, move through the peritoneum, and penetrate the liver. After migrating through and feeding on the liver, causing extensive tissue damage, the parasites move to their final niche in the bile ducts where they mature and produce eggs. Here we integrated a transcriptomics and proteomics approach to profile *Fasciola* secretory proteins that are involved in host-pathogen interactions and to correlate changes in their expression with the migration of the parasite. Prediction of *F. hepatica* secretory proteins from 14,031 expressed sequence tags (ESTs) available from the Wellcome Trust Sanger Centre using the semiautomated EST2Secretome pipeline showed that the major components of adult parasite secretions are proteolytic enzymes including cathepsin L, cathepsin B, and asparaginyl endopeptidase cysteine proteases as well as novel trypsin-like serine proteases and carboxypeptidases. Proteomics analysis of proteins secreted by infective larvae, immature flukes, and adult *F. hepatica* showed that these proteases are developmentally regulated and correlate with the passage of the parasite through host tissues and its encounters with different host macromolecules. Proteases such as FhCL3 and cathepsin B have specific functions in larvae activation and intestinal wall penetration, whereas FhCL1, FhCL2, and FhCL5 are required for liver penetration and tissue and blood feeding. Besides proteases, the parasites secrete an array of antioxidants that are also highly regulated according to their migration through host tissues. However, whereas the proteases of *F. hepatica* are secreted into the parasite gut via a classical endoplasmic reticulum/Golgi pathway, we speculate that the antioxidants, which all lack a signal sequence, are released via a non-classical trans-tegumental pathway. *Molecular & Cellular Proteomics 8:1891–1907, 2009.*

From the ‡Institute for the Biotechnology of Infectious Diseases, University of Technology Sydney, Level 6, Building 4, Corner of Thomas and Harris Street, Ultimo, Sydney, New South Wales 2007, Australia and ¶Department of Chemistry and Biomolecular Sciences and Australian Research Council Centre of Excellence in Bioinformatics, Macquarie University, Sydney, New South Wales 2109, Australia

Received, January 27, 2009, and in revised form, May 6, 2009

Published, MCP Papers in Press, May 14, 2009, DOI 10.1074/mcp.M900045-MCP200

Fasciola hepatica is a helminth (worm) parasite with a worldwide distribution. Although traditionally regarded as a parasite of livestock, particularly sheep and cattle, that results in a large economic loss to the agricultural community it has recently emerged as an important human infection in many regions of the world, including South America, Iran, Egypt, and mainland South-East Asia (1). Dormant larvae contained within cysts adhere to vegetation and emerge as infective juveniles (newly excysted juveniles (NEJs))[1] in the duodenum following ingestion by animals or humans. They infect their hosts by rapidly penetrating the intestinal wall and entering the peritoneal cavity where they break through the liver capsule. After 8–12 weeks of consistent burrowing, feeding, and growth within the liver parenchyma they move to their final destination within the bile ducts where they mature and produce enormous numbers of eggs (2). The two distinct clinical phases of fasciolosis are directly related to the migration of the parasites: acute fasciolosis, which manifests as fever, abdominal pain, weight loss, and hepatomegaly, is associated with liver tissue damage and inflammation caused by the migrating immature parasites, whereas chronic fasciolosis (usually subclinical) is coupled with the presence of the mature adult flukes in the bile ducts (1, 3).

These invasive helminth parasites undergo complex changes as they migrate within their definitive mammalian hosts. The developing parasites encounter different host tissues and macromolecules and have to contend with a continually changing physiological microenvironment (such as pH and oxygen availability) and a mounting humoral and cellular host immune response. Morphological and ultrastructural

[1] The abbreviations used are: NEJ, newly excysted juvenile; EST, expressed sequence tag; rEST, representative expressed sequence tag; BLAST, basic local alignment search tool; ER, endoplasmic reticulum; 2-DE, two-dimensional electrophoresis; 1-DE, one-dimensional electrophoresis; Bis-Tris, 2-[bis(2-hydroxyethyl)amino]-2-(hydroxymethyl)propane-1,3-diol; MSDB, mass spectrometry protein sequence database; emPAI, exponentially modified protein abundance index; E-64, *trans*-epoxysuccinyl-L-leucylamido(4-guanidino)butane; Z-Phe-Ala-CHN_2, carbobenzoxy-phenylalanyl-alanine-diazomethyl ketone; FaBP, fatty acid-binding protein; ABC, ATP-binding cassette; vpB1, vitelline protein B1; IL, interleukin; RNAi, RNA interference; 1-D, one-dimensional.

studies clearly show major alterations of the parasite surface and gastrodermis, the two host-parasite interfaces, as they migrate and grow (4). However, we are only beginning to understand the molecular and biochemical interactions that occur between host and parasite and how these adjust as the development of the parasite progresses. A deeper knowledge of such host-pathogen interplay should provide data on novel targets for anthelmintic compounds and potential antiparasite vaccine candidates.

Increasingly proteomics analysis is being used as a means to investigate the interaction of helminth parasites and their hosts particularly in cases where obtaining pathogen material is difficult (5–9). For some helminths these studies have been facilitated by the availability of large transcriptomics data sets (10–12). Unfortunately the *Fasciola* nucleotide sequences available in GenBank™ are relatively few and highly redundant (298 for *F. hepatica* and 142 for *Fasciola gigantica* as of January 15, 2009), and the adult *F. hepatica* ESTs currently available from the Wellcome Trust Sanger Centre (14,031 entries) are unannotated (therefore the current identification of peptides of interest requires manual BLAST analysis using specific query sequences). Accordingly in this study we used the semiautomated EST2Secretome pipeline to analyze all available *F. hepatica* ESTs for secretory proteins potentially involved in host-pathogen interactions (13). EST2Secretome was developed in our laboratory by optimizing signal peptide-mediated secreted protein prediction from our earlier predictions of parasitic nematode ESTs (14, 15). We integrated these transcriptomics data with a proteomics analysis of the molecules secreted by adult *F. hepatica* with a particular emphasis on proteolytic enzymes. Furthermore we also analyzed the somatic and secreted molecules of the infective NEJ parasites and compared these with the secretome of immature and adult parasites taken from liver tissues. In doing so we produced a comprehensive view of how *F. hepatica* differentially and developmentally express and secrete proteolytic enzymes and other molecules according to the specific challenges faced in the intestine, liver, and bile ducts. *F. hepatica* cathepsin Bs and cathepsin L (FhCL3) are stored as zymogens within the infective larvae ready to be *trans*-activated by specific asparaginyl endopeptidases and released to perform the highly specific function of host tissue invasion (intestinal epithelium and liver capsule). By contrast, cysteine proteases belonging to the phylogenetic clades FhCL1, FhCL2, and FhCL5 are expressed during the later stages within the liver and bile duct and function in tissue degradation and feeding alongside the cell lytic protein saposin and a newly described prolylcarboxypeptidase. Several novel developmentally regulated cathepsin L and cathepsin B cysteine proteases and members of two serine protease families, namely carboxypeptidase and trypsin-like serine proteases, which may also have important roles in host-parasite interplay, were also identified. Finally our observations led us to propose that whereas the major proteases of *F. hepatica* are secreted into the parasite gut via a classical ER/Golgi pathway, an array of abundant and highly regulated antioxidants are released via a non-classical trans-tegumental pathway.

EXPERIMENTAL PROCEDURES

Analysis of the F. hepatica ESTs Using EST2Secretome—Transcriptome analysis for secretory proteins from 14,031 *Fasciola* EST sequences, available from the Sanger Centre, UK, was performed using the semiautomated EST2Secretome pipeline recently reported by Nagaraj *et al.* (13). The EST2Secretome pipeline integrates a number of high quality programs for EST cleaning (SeqClean and RepeatMasker), assembly into contigs and singletons (CAP3), conceptual translation (ESTSCAN), and prediction of secreted proteins based on the presence of N-terminal secretory signal sequences (SignalP) and follows with the elimination of membrane proteins using transmembrane hidden Markov models (TMHMM). This approach has proven to be very reliable in identifying secretory proteins from a number of helminths including the bovine lungworm *Dictyocaulus viviparus* (16), gastrointestinal worm *Trichostrongylus vitrinus* (15), and the hookworm *Ancylostoma caninum* (17). The predicted secretory proteins are extensively annotated for functional protein families and motifs (InterProScan), gene ontologies (BLAST2GO) pathways KEGG orthology-based annotation system (KOBAS), protein-protein interactions (comparison with IntAct), mapping to *Caenorhabditis elegans* proteins (from WormPep), and subsequent correlation to RNAi phenotype data. The SignalP threshold value for secretory signal peptide prediction was set at 0.5 as determined for the large scale secretory protein prediction from helminth parasite ESTs (13).

Independently unannotated *Fasciola* ESTs that matched MS/MS data were used as queries for BLASTn (18) searches of all nucleotide sequences in the GenBank and European Molecular Biology Laboratory (EMBL) databases. Searches were performed using the National Center for Biotechnology Information (NCBI) server. Open reading frames were constructed from the *Fasciola* EST sequences (guided by their best BLASTn hits), and hypothetical proteins derived from their conceptual translation were submitted to the InterProScan algorithm (19) to detect conserved domains and motifs to help assign putative protein identifications.

Proteomics Analysis of Parasite Somatic and Secreted Proteins: Gel Electrophoresis and Mass Spectrometry—The profile of proteins secreted by adult *F. hepatica* has been recently studied using two-dimensional electrophoresis (2-DE) by us and others (8, 20, 21). However, because of the paucity of material that can be obtained for the infective NEJ parasites and 21-day-old liver stage parasites we used 1-DE in the present study. Nevertheless we validated this approach by subjecting adult parasite secretory proteins to 1-DE and comparing the data with those reported by Robinson *et al.* (8). Protein samples were analyzed by 1-DE using NuPAGE® Novex® 4–12% Bis-Tris gels (Invitrogen). NuPAGE lithium dodecyl sulfate sample buffer plus Sample Reducing Agent (Invitrogen) were added to the samples, and samples were heated at 95 °C for 5 min prior to electrophoresis. Gels were stained with colloidal Coomassie Blue G-250 (Sigma) and destained with 10% methanol (v/v) and 7% acetic acid (v/v). Following visualization, gel lanes were cut into six sections for analysis by mass spectrometry. Briefly the individual gel sections were cut into smaller pieces (~1 mm^2) and reduced and alkylated with 5 mM tributylphosphine and 20 mM acrylamide (Sigma) in 100 mM NH_4HCO_3 for 90 min. The excised sections were then in-gel digested with trypsin (Sigma proteomics grade), and the peptides were solubilized with 2% formic acid (Sigma) prior to analysis by nano-LC-ESI-MS/MS using a Tempo nano-LC system (Applied Biosystems) with a C_{18} column (Vydac) coupled to a QSTAR Elite QqTOF mass spectrometer running in information-dependent acquisition mode (Applied Biosystems). Peak list files generated by the Protein Pilot v1.0 soft-

ware (Applied Biosystems) using default parameters were exported to local MASCOT v2.1.0 (Matrix Science) or PEAKS (Bioinformatics Solutions Inc.) search engines for database searching.

Database Searches—MS/MS data were used to search 3,239,079 entries in the MSDB (September 8, 2006) using MASCOT whereas PEAKS software was used to search the 14,031 *F. hepatica* EST sequences from the Wellcome Trust Sanger Institute. The enzyme specificity was set to trypsin, propionamide (acrylamide) modification of cysteines was used as a fixed parameter, and oxidation of methionines was set as a variable protein modification. The mass tolerance was set at 100 ppm for precursor ions and 0.2 Da for fragment ions. Only one missed cleavage was allowed. For MASCOT searches, matches achieving a molecular weight search (MOWSE) score >70 with at least two high scoring individual peptides were considered to be significant (6, 7). For PEAKS searches of the *Fasciola* EST database, at least two high scoring (>60%) matching peptides were required. However, other criteria were considered in assigning a positive identification including concordance between the calculated theoretical molecular mass of the protein and the observed position of the polypeptide by 1-DE. To account for matches to multiple members of the *Fasciola* cathepsin family, we searched for peptides specific to individual enzymes or clades (Ref. 8; see "Results"). The proteomics data and transcriptome analysis were integrated to give a more complete view of gene expression/secretion in adult *F. hepatica*. This integrated data set provided a framework for comparison with the infective NEJs and 21-day-old immature flukes at the subproteome (*i.e.* secretome) level.

Quantitation of Fasciola Proteins by Exponentially Modified Protein Abundance Index (emPAI)—We analyzed the emPAI provided in the output of the MASCOT MS/MS ion search to estimate the relative expression of proteases and antioxidants identified in the developmental stages of *F. hepatica*. The emPAI used by MASCOT is a modification of the formula developed by Ishihama *et al.* (22) and gives a label-free relative quantification of the proteins in a mixture. The raw emPAI values obtained represent the transformed ratio of the number of experimentally observed peptides (composed of unique precursor ions, including different charge states of the same peptide, that match or exceed the threshold level for homology or identity) to the total number of peptides that can theoretically be detected within the operating mass range and retention time range of the mass spectrometer (calculated by MASCOT based on the mass of the protein, the average amino acid composition of the database searched, and the enzyme specificity). For this analysis, the raw emPAI values (averaged from three separate gels) for all *Fasciola* proteases or antioxidants identified were added to give a figure representing total expression for each within a particular developmental stage. The raw emPAI values for each individual protease or antioxidant were then converted to a percentage of this total to estimate their relative expression levels (see Figs. 2*A* and 3*A*). As the method used by MASCOT to calculate the number of observable peptides differs from that originally described by Ishihama *et al.* (22) and is not freely available, it was decided not to calculate emPAI values manually for those proteins identified using the PEAKS software to avoid introducing errors into the subsequent analysis. To account for redundancy, molecules potentially containing shared sequences were grouped together (*e.g.* all cathepsin B variants were classed as FhCB, all cathepsin L1 variants were classed as FhCL1, etc.).

Excystment of F. hepatica Metacercariae and Preparation of Somatic Larvae Proteins—The dormant cysts of *F. hepatica* metacercariae contain two layers, the outer of which can be contaminated with plant or other extraneous material. Here we describe a method for removing the outer cyst and adhered material so that somatic proteins of the dormant infective larvae can be analyzed. In addition, we describe a rapid method for activating the larvae and inducing them to emerge from the cysts so that *in vitro* secreted proteins can be isolated. Thus, *F. hepatica* metacercariae (Baldwin Aquatics Inc., Monmouth, OR) were vortexed for 10 s in 0.5% sodium hypochlorite and then incubated at room temperature for 20 min (this procedure dissolves the outer cyst layer). They were then washed three times in distilled water by centrifugation at 2000 × *g* for 2 min. The juvenile larvae, which were now only contained within a clear inner cyst layer, were used to prepare somatic protein extracts and secretory proteins. To prepare the somatic extracts the parasites were homogenized in RIPA buffer (50 mM Tris-HCl, pH 7.4, 150 mM NaCl, 1 mM PMSF, 1 mM EDTA, 1% Triton X-100, 1% sodium deoxycholate (Sigma), 0.1% SDS, 1× Complete Mini protease inhibitor mixture (Roche Applied Science)) and placed on ice for 30 min. The protein extract was centrifuged at 13,000 rpm for 10 min to remove insoluble debris, and the supernatant was stored at −20 °C until use.

To prepare secreted proteins the washed parasites were resuspended and incubated in excystment medium (0.5% sodium bicarbonate, 0.4% sodium chloride, 0.2% sodium taurocholate, 0.07% concentrated HCl, 0.006% L-cysteine) for up to 3 h at 37 °C in 5% CO_2. NEJs emerged from the cysts within 2–3 h and were transferred to prewarmed (37 °C) culture medium, RPMI 1640 medium (Invitrogen) containing 2 mM L-glutamine, 30 mM HEPES, 0.1% (w/v) glucose, and 2.5 μg/ml gentamycin, and cultured for 24 h.

To determine whether cysteine proteases were essential to cyst rupture, *F. hepatica* metacercariae (50 per treatment in triplicate) were either excysted with medium as described above or in medium lacking the 0.006% (w/v) L-cysteine or supplemented with the cysteine protease inhibitors *trans*-epoxysuccinyl-L-leucylamido(4-guanidino)butane (E-64, Sigma) or 1 mM carbobenzoxy-phenylalanyl-alanine-diazomethyl ketone (Z-Phe-Ala-CHN_2, Bachem, St. Helens, UK) to a final concentration of 1 mM. The numbers of excysted parasites were counted after 3-h incubation at 37 °C in 5% CO_2, and the data were analyzed using Student's *t* test. The experiments were repeated twice.

Preparation of Immature and Mature Adult F. hepatica Secretory Proteins—Immature *F. hepatica* flukes (21 days old) were recovered from the livers of female BALB/c mice (experimentally infected with 20 metacercariae), whereas adult parasites were recovered from the bile ducts of Merino sheep 16 weeks after an experimental infection with 200 metacercariae. Immature and adult parasites were washed in prewarmed (37 °C) PBS, pH 7.3, before transfer to culture medium (as described above) for 24 and 8 h, respectively. Secretory proteins were concentrated from the culture supernatants by precipitation with methanol/chloroform as described previously (6). Pellets were resuspended in 10 μl of RIPA buffer and stored at −20 °C prior to separation by electrophoresis.

RESULTS

Transcriptomics Profiling of Adult F. hepatica Secretory Proteins—Of the 14,031 adult *F. hepatica* raw EST sequences available, a total of 12,954 (92.3%) quality sequences were obtained (Table I). Cluster analysis of the 12,954 ESTs yielded 4236 representative ESTs (rESTs; 2749 contig and 1487 singleton sequences) of which 2960 (68.9%) had open reading frames. These preprocessed ESTs ranged from 60 to 2093 bp with a mean of 569 bp and S.D. of 268 bp. After clustering, the mean length of the contigs increased to 788 bp with S.D. of 358 bp. The G + C content of the coding sequences was 44.5%, which is similar to the figure of 43.5% reported for the adult bovine lungworm *D. viviparus* from EST analysis (16). This value is slightly higher than that reported for the related trematode *Schistosoma mansoni* (34%; Ref. 23) but consist-

TABLE I
Summary analysis of 14,031 adult F. hepatica ESTs using EST2Secretome

The contigs and singletons generated by preprocessing, overall rESTs, peptides from conceptual translation, and putative secretory proteins identified are shown.

F. hepatica ESTs	Numbers
Raw sequences obtained	14,031
Cleaned sequences	12,954 (92.3%)
Clusters of multiple sequences (contigs)	2,749 (19.6%)
Clusters of singletons	1,487 (10.6%)
Total rESTs	4,236 (30.2%)
Putative peptides	2,960 (68.9% rESTs)
Secreted proteins (SignalP cutoff, 0.5)	160 (5.4% peptides)

ent with those reported for nematodes (32–51%; Ref. 24), *C. elegans* (37%), and *Caenorhabditis briggsae* (38%) (25).

All rESTs were then subjected to analysis using our recently reported semiautomated bioinformatics platform (EST2Secretome; Ref. 13) to predict secretory proteins from the *F. hepatica* EST database. Using this approach, we identified a subset of parasite proteins likely to participate in the most significant interactions that occur between the adult stage of this parasite and its mammalian host. Thus, 173 *Fasciola* secretory proteins were predicted by the EST2Secretome pipeline with 160 true positives based on the EST2Secretome annotation mapping homology to proteins in a non-redundant secreted protein database (SecProtSearch) derived from the literature, the secreted protein database SPD (26), and the manually curated signal peptide database SPdb (27) as well the gene ontology annotations of subcellular localization of the top homologues identified by BLAST (92.5% accuracy) (Table I). Because ESTs are usually not full length and are often truncated, manual inspection of the final data set is required as it is possible that transmembrane sequences are erroneously identified as secreted proteins and thus elude the filtration step by TMHMM.

From the detailed annotations of the 160 adult secreted proteins the predominance of cathepsin L cysteine proteases is clearly evident as these are represented by a total of 66 (41.2%) proteins (Table II and supplemental Table 1). Robinson *et al.* (8) recently showed that the adult *F. hepatica* cathepsin L proteases separated into five distinct clades. Of these 66 adult rESTs encoding cathepsin Ls, 48 (72.5%) represented clade FhCL1 (38 subclade FhCL1A and 10 subclade FhCL1B), 11 (17%) encoded clade FhCL2, and two (3%) encoded clade FhCL5 cathepsins (8). Consistent with Robinson *et al.* (8) no cDNAs encoding clades FhCL3 or FhCL4 were detected in the adult ESTs because these proteases have been reported as specific for the infective NEJ (see below and Ref. 28). Interestingly five rESTs (7.5%) encoded cathepsin Ls that could not be placed into any of the five phylogenetic clades based on primary sequence alignment analysis.

The next most abundant secreted protein based on the number of ESTs identified was saposin-like protein 3 that has been reported as a secreted protein by Grams *et al.* (29) and suggested to play a role in red blood cell lysis (30). Other proteins of interest to our study include several novel cathepsin B cysteine endoproteases (designated cathepsins B4–B10 in the current study), four novel asparaginyl endopeptidases or legumains (designated legumains 4–7 in the current study), and a cysteine protease inhibitor, cystatin. Additionally three putative novel adult serine proteases were identified: a serine carboxypeptidase and two proteins with trypsin-like protease domains. Other ESTs encoded secreted vitelline protein B1 that is found in eggs produced by the adult parasite (31) (Table II).

A *Fasciola* protein-disulfide isomerase was also predicted that has previously been identified in the secretions of adult flukes (32). Protein-disulfide isomerases have roles in protein folding, and a *Fasciola* recombinant enzyme was shown to mediate the oxidative refolding of reduced RNase (32). A putative peptide with a number of cubulin domains was also predicted in the current analysis. Cubulin domains occur predominantly in extracellular proteins or plasma membrane-associated proteins with a range of functions including complement activation, tissue repair, cell signaling, and inflammation (33). Although the *Fasciola* peptide contains a predicted N-terminal transmembrane region, its molecular function remains unknown. We note that although orthologues are available for six proteins, including an uncharacterized secretory protein from *Clonorchis sinensis*, their function remains elusive. A total of 36 secreted proteins (21.9%) are novel, but no database matches exist at the present time (Table II).

Proteomics Profiling of Adult F. hepatica Secreted Proteins—We and others have previously characterized the major secretory proteins expressed by adult F. *hepatica* using 2-DE (8, 20, 21). To complement these earlier studies and to validate the use of 1-DE for proteomics analysis, we analyzed tryptic digests extracted from gel sections of adult *F. hepatica* secreted proteins (see Fig. 1) by mass spectrometry. Twenty-two different proteins secreted by adult *F. hepatica* were identified in this analysis: 19 matched to previously identified *Fasciola* cDNAs, and three corresponded to putative proteins encoded by novel ESTs identified by our present EST2Secretome analysis but were unidentified by Robinson *et al.* (8) (Figs. 2*B* and 3*B* and supplemental Tables 2 and 3).

In accordance with the transcriptomics predictions, cathepsin L proteases were highly represented in adult fluke secretions. Matches to 13 cathepsin L sequences were observed and included clades FhCL1, FhCL2, and FhCL5 enzymes (FhCL4 or FhCL3 enzymes were not detected) (8). These identifications were based on the presence of clade-specific peptide matches such as NSWGLSWGER (ion *m/z* 596.30, +2) and VTGYYTVHSGSEVELK (ion *m/z* 590.32, +3) that are clade 1-specific peptides and three peptides, DYYYVTEVK

TABLE II

Secretory proteins predicted from adult F. hepatica rESTs

Adult *F. hepatica* secretory proteins were predicted using the EST2Secretome pipeline (13) using the default SignalP threshold value of 0.5. Their putative functionality based on BLAST analysis and the presence of InterPro domains is shown. Supporting proteomics data from the three *Fasciola* life cycle stages and the RNAi phenotypes for their *C. elegans* homologues are also shown. Imm, 21-day-old immature fluke; Ad, adult fluke; Ste, sterile; Lvl, larval lethal; Lva, larval arrest; Emb, embryonic lethal; Gro, slow growth; Unc, locomotion abnormal; Daf, dauer formation abnormal; Pvl, protruding vulva; Dpy, dumpy; Slu, sluggish; Clr, clear; age, life span abnormal; Age, extended life span; Stp, sterile progeny; Rup, exploded through vulva; Ric, aldicarb-resistant; Cons, constipated; Mig, cell migration abnormal; Gom, gonad migration abnormal; Bmd, organism morphology abnormal. —, none.

Description (top BLAST hit)	Organism	InterPro domains	rESTs	Proteomics data	RNAi phenotypes of *C. elegans* homologues
Cathepsin L	*F. hepatica*	Peptidase C1A	66	NEJ, Imm, Ad	Emb, Gro, Unc
Novel proteins (no significant hits)	—	—	36	—	—
Saposin-like protein 3	*F. hepatica*	Saposin B	13	Imm, Ad	Emb, Gro, Unc, Ste, Pvl, Lva
Vitelline protein B1	*F. hepatica*	Trematode eggshell synthesis	9	—	Daf
Cathepsin B endoprotease	*F. hepatica*	Peptidase C1A	6	NEJ, Imm	—
Legumain	*Opisthorchis viverrini*	Peptidase C13, legumain	5	NEJ, Imm	Emb
Cystatin	*F. hepatica*	Proteinase inhibitor I25, cystatin	3	—	Gro
Unknown proteins	*C. sinensis*	—	3	NEJ, Ad[a]	Ste, Lvl, Lva, Emb
Uncharacterized proteins[b]	*S. japonicum*	New Pfam domain	2	—	—
Protein-disulfide isomerase	*F. hepatica*	Disulfide isomerase	2	Imm	Gro, Lva, Bmd, Dpy, Emb, Unc, Slu, Clr
SJCHGC01895 protein	*S. japonicum*	Peptidase S1 and S6	2	—	Emb
Cubulin	*Canis familiaris*	CUB	2	NEJ	—
Unnamed protein	*Tetraodon nigroviridis*	Deoxyribonuclease I	1	—	Emb, Lva
SJCHGC00967 protein	*S. japonicum*	Deoxyribonuclease II	1	—	age
Apoferritin-2	*S. japonicum*	Ferritin	1	Imm	bar-1(ga80)
Hypothetical protein	*Aedes aegypti*	Phospholipase D	1	—	—
SJCHGC06223 protein	*S. japonicum*	Peptidase S10, serine carboxypeptidase	1	—	—
Cyclophilin B	*Xenopus tropicalis*	Peptidyl-prolyl cis-trans isomerase	1	NEJ, Ad	—
Unnamed protein	*Kluyveromyces lactis*	—	1	—	—
Peptidoglycan recognition protein	*Argopecten irradians*	Amidase 2	1	—	Stp
Myoglobin 2	*P. westermani*	Globin, globin-like	1	—	Cons, Gom, Emb, Rup, Bmd, Mig, Gro
SJCHGC09717 protein	*S. japonicum*	—	1	—	—
Gag-pol polyprotein	*Strongylocentrotus purpuratus*	—	1	—	Age, Ric

[a] Identified in adult *F. hepatica* secretions following gel filtration (data not shown).

[b] New protein family assigned to this protein: PF11703.

(*m/z* 590.26, +2), VTGYYTVHSGDEIELK (*m/z* 604.32, +3), and LTHAVLAVGYGSQDGTDYWIVK (*m/z* 798.42, +3), characteristic of FhCL2 cathepsin Ls as well as the presence of peptides such as NSWGTWWGEDGYIR (*m/z* 863.90, +2) and FGLETESSYPYR (*m/z* 724.85, +2) that aid the identification of clade 5 cathepsin Ls (8). The three previously unidentified secreted proteins were saposin, a peptidyl-prolyl cis-trans isomerase, and a protein with homology to an uncharacterized *C. sinensis* secretory protein.

Several proteins that were predicted to be secreted by EST2Secretome were not identified by mass spectrometry of the adult parasite secretome both in this study and in the studies of Jefferies *et al.* (20), Morphew *et al.* (21), and Robinson *et al.* (8); these were cathepsin B, legumain, serine carboxypeptidase, two trypsin-like enzymes, and the eggshell vitelline protein B1. On the other hand, proteomics analysis detected four fatty acid-binding proteins (FaBP1, FaBP2, FaBP3, and Fh15) and two redox enzymes (peroxiredoxin and thioredoxin) that were not predicted by EST2Secretome analysis. A putative novel prolylcarboxypeptidase (also observed in 21-day-old immature parasites; see below) was also identified by mass spectrometry and yet was not predicted to be a secretory peptide. This is surprising because at least one adult fluke EST (Fhep06a01.q1k) encoding the N-terminal end of the enzyme contains a putative signal peptide. However, this EST was not retained as a singleton after repeat masking and truncation in the early stages of EST processing in EST2Secretome.

Identification of F. hepatica Dormant Larval Somatic Proteins and NEJ Secretory Proteins—By removing the outer cyst layer of the dormant metacercarial infective stage of the parasite we could extract somatic protein without contamination with extraneous proteins and analyze these by 1-DE. Approximately 14 protein bands could be visualized following Coo-

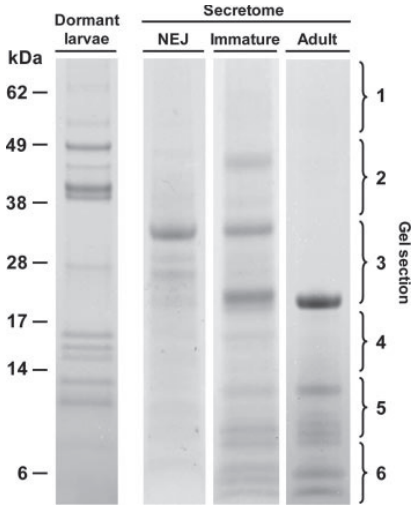


FIG. 1. **Analysis of *F. hepatica* somatic and secretory proteins by 1-DE.** Shown is the typical 1-D profile of somatic proteins expressed by dormant *F. hepatica* larvae as well as proteins secreted by *F. hepatica* NEJs (*NEJ*), 21-day-old immature flukes (*Immature*), and adult parasites (*Adult*). Proteins (10 μg) were separated using NuPAGE Novex 4–12% Bis-Tris gels (Invitrogen) and stained with colloidal Coomassie Blue G-250. Following trypsin digests, peptides were extracted from gel sections 1–6 and analysed by mass spectrometry.

massie Blue-stained SDS-PAGE (Fig. 1) that on inspection appeared very similar to the banding pattern of *Fasciola* larval somatic proteins reported by Tkalcevic *et al.* (34). Proteins in these bands were in-gel digested with trypsin and analyzed by nano-LC-ESI-MS/MS, and the resulting CID data were used for database searching. A total of 26 different *Fasciola* dormant larvae proteins were identified: 12 matched to previously identified *Fasciola* cDNAs (or cDNAs from related trematode species), and 14 corresponded to proteins encoded by novel ESTs (Figs. 2*B* and 3*B* and supplemental Tables 2 and 3). A further five peptides encoded by *Fasciola* ESTs were also identified, but these lacked conserved protein domains and could not be assigned putative functions based on BLAST searches.

Of the 26 positively matched *Fasciola* proteins six were proteases including two cathepsin L3 proteases, three cathepsin B endopeptidases, and an asparaginyl endopeptidase-like precursor (discussed in detail below). Others included structural proteins related to muscle function such as actin, myosin-regulatory light chain, and a troponin C homologue (supplemental Tables 2 and 3). Four metabolic enzymes were also identified: pyruvate carboxylase (gluconeogenesis), malate dehydrogenase (tricarboxylic acid cycle), and aldolase and enolase (glycolysis). Mass spectrometry data also matched to a putative cullin protein (roles in protein degradation and ubiquitination), a cyclophilin-like peptidyl-prolyl cis-trans isomerase, and a tetraspanin membrane protein. Other notable peptide matches included the *Fasciola* antioxidant enzyme peroxiredoxin, two histone proteins, a heat shock protein 70, dynein light chain, a ribosome production factor, ribosomal protein L30, and a protein bearing a conserved RNA-binding motif (supplemental Tables 2 and 3).

By using an artificial medium that replicated the surfactant and reducing conditions in the duodenum we could activate the dormant infective larvae and induce them to excyst. Proteins secreted by *F. hepatica* NEJs during *in vitro* culture were isolated from the medium, separated by 1-DE, and analyzed by nano-LC-ESI-MS/MS (Fig. 1). Matches to 29 different proteins, of which 10 were proteases including seven cathepsin L3 proteases, one cathepsin B, and two asparaginyl endopeptidase-like proteases, were obtained from MS/MS data from *Fasciola* NEJ secreted proteins.

Other protein matches to *Fasciola* cDNA sequences included enolase, three fatty acid-binding proteins (Fh2, Fh3, and Fh15), and peroxiredoxin (Fig. 3*B* and supplemental Tables 2 and 3). A further 16 different putative peptides were identified following searches of the *F. hepatica* EST database including the metabolic enzymes fructose-bisphosphate aldolase, phosphoenolpyruvate carboxykinase, glyceraldehyde-3-phosphate dehydrogenase, malate dehydrogenase, and an ATPase. Other matches included putative structural proteins such as calponin, spermadhesin, histones (H2A, H2B, and H4), redox enzymes (thioredoxin and peptidyl-prolyl cis-trans isomerase), and an uncharacterized protein with predicted transmembrane regions. Finally molecules with roles in protein turnover such as ubiquitin and a putative serpin were also identified.

Identification of Immature F. hepatica Secretory Proteins—Parasites were removed from the livers of infected mice and maintained *in vitro* for collection of secreted proteins. The overall complexity of the secretory proteins of these 21-day-old immature *F. hepatica* was greater compared with that of the NEJs or adult parasites (Fig. 1) and yielded a total of 45 different protein identifications (Figs. 2*B* and 3*B* and supplemental Tables 2 and 3). Of these, 34 were matched to previously identified *Fasciola* cDNAs, and 11 corresponded to putative proteins encoded by novel ESTs. Mass spectrometry data also matched to peptides encoded by a further two *F. hepatica* EST sequences that lacked conserved protein domains and could not be assigned putative functions based on BLAST searches.

Of the 45 positively matched proteins 22 were proteases and included 14 cathepsin Ls, four cathepsin Bs, three asparaginyl endopeptidases (legumains), and a newly discovered prolylcarboxypeptidase. The remaining 23 proteins secreted by the immature liver stage parasites included a GST Sigma class enzyme, four isoforms of Mu class GSTs (GST1, GST7, GST47, and GST 51), four fatty acid-binding proteins (FaBP1, FaBP2, FaBP3, and Fh15), two saposin-like proteins (SAP1 and SAP3), two enzymes of glycolysis (enolase and triosephosphate isomerase), and two enzymes involved in cell re-

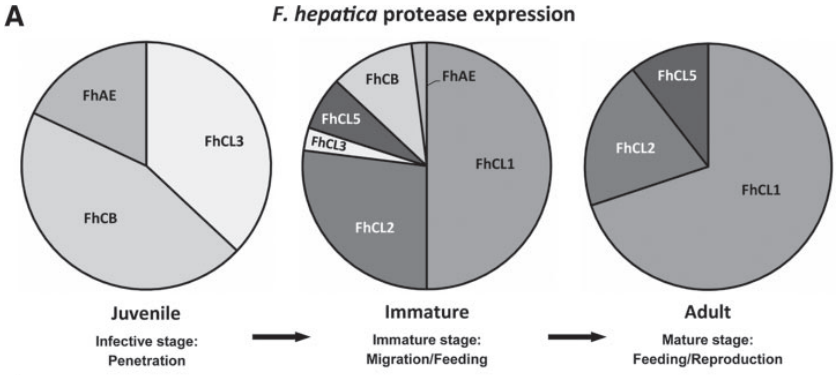
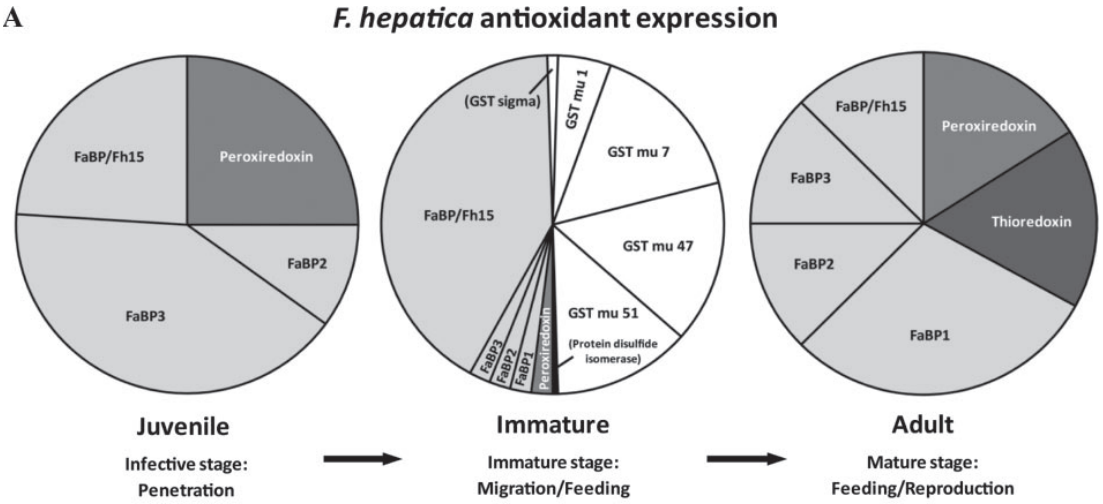


FIG. 2. **Expression of *F. hepatica* proteases.** *A*, the expression levels of the major proteolytic enzymes used by *F. hepatica* during its mammalian life cycle. The label-free emPAI (22) derived from *Fasciola* MS/MS CID data was used to determine the abundance of the various cathepsin Ls (*FhCL*), cathepsin Bs (*FhCB*), and asparaginyl endopeptidases (*FhAE*) that are expressed by juvenile flukes (dormant larvae) and secreted by the immature (liver stage) flukes and adult (bile duct stage) parasites. *B*, identification of *F. hepatica* proteases by nano-LC-ESI-MS/MS. Dormant larvae somatic proteins and those secreted by NEJs, 21-day old immature flukes and adult *F. hepatica* were separated by 1-DE and analyzed by nano-LC-ESI-MS/MS. Mass spectrometry data were submitted to either the MSDB or a custom-made database composed of all *F. hepatica* ESTs (14,031 entries) currently available from the Wellcome Trust Sanger Centre using MASCOT and PEAKS software, respectively. Adult *F. hepatica* ESTs that matched with MS/MS data were submitted as queries to BLASTn (18) or conceptually translated and submitted to InterProScan (19) to detect conserved domains and motifs. Footnotes are as follows. *1*, MASCOT scores are given for matches to *F. hepatica* cDNAs. *2*, number of unique peptide matches. *3*, raw emPAI values provided by the MASCOT search engine give an approximate quantification of the protein (22). These values were converted to a percentage of total protease expression within each *Fasciola* developmental stage (see "Experimental Procedures") and were used to generate the pie charts shown in *A*. *4*, section of the 1-D gel from which the peptides were identified. *5*, PEAKS scores (expressed as a percentage) are given for matches to peptides encoded by *F. hepatica* ESTs. *6*, single peptide-based identification (see supplemental Fig. 4).

B

Accession	*Fasciola* protein	Score[1]	Unique peptides[2]	% cover	Mr	emPAI[3]	Section[4]
Dormant larvae somatic							
A7UNB2	Cathepsin B	294	5	24	37.6	0.29	2
Q8I7B2	Procathepsin B2 precursor	111	3	9	38.0	0.09	2
A5X494	Cathepsin B3	223	1[6]	22	31.0	0.36	1-3
A8W7J0	Procathepsin L	235	7	21	35.6	0.30	2
Q9GRW6	Procathepsin L3	194	7	21	35.3	0.31	2
Q711M2	Legumain-like precursor 1	266	7	21	47.8	0.30	1-3
Fhep29h09.q1k	Putative legumain 2	76[5]	5	12	-	-	1
NEJ secretome							
P80527	Hemoglobinase-like protein 1	117	2	85	2.2	2.17	2,3
Fhep29h09.q1k	Putative legumain 2	99	12	12	-	-	1-5
Q9GRW4	Partial procathepsin L3	99[5]	11	23	35.1	-	2-5
Q9GRW6	Partial procathepsin L3	99[5]	8	23	35.4	-	2-5
Q95VA7	Cathepsin L	99[5]	5	27	37.4	-	2-5
B3TM67	Cathepsin L3	99[5]	9	26	35.1	-	2-5
B3TM68	Cathepsin L3	99[5]	9	39	35.3	-	2-5
A8W638	Metacercarial cathepsin L	99[5]	12	34	37.4	-	2-5
A8W7J0	Metacercarial procathepsin L	99[5]	13	38	35.7	-	2-5
HAN4015b05.q1kT3	Putative cathepsin B4	88[5]	3	11	-	-	3
Immature fluke secretome							
A7UNB2	Cathepsin B	113	2	7	37.6	0.09	2
Q8I7B2	Pro-cathepsin B2 precursor	99	2	6	38.0	0.18	5
Fhep45b05.q1k	Putative cathepsin B5	97[5]	6	13	-	-	2
HAN4006g01.q1kT3	Putative cathepsin B6	99[5]	6	9	-	-	2
Q711M2	Legumain-like precursor	412	31	16	47.8	0.70	2
A6Y9U8	Legumain-1	328	15	20	47.9	0.59	2
A6Y9U9	Legumain-2	176	3	10	48.3	0.22	2
Q8T5Z9	Cathepsin L	334	12	32	35.1	1.06	3
Q7JNQ9	Secreted cathepsin L1	447	16	37	36.7	1.82	3
Q6R018	Cathepsin L protein	292	13	24	36.6	1.00	3
Q24940	Cathepsin L-like proteinase	356	15	32	36.8	1.36	3
Q9GRW5	Cathepsin L1	282	12	27	35.1	0.88	3
Q7JNQ8	Secreted cathepsin L2	363	12	37	37.0	0.67	3
A5Z1V3	Secreted cathepsin L2	470	14	39	37.0	0.98	3
A5X483	Cathepsin L2	367	14	39	24.5	1.78	3
A3FMG6	Cathepsin L	228	9	30	36.9	0.19	3
Q95VA7	Cathepsin L	235	7	15	37.4	0.18	3
A8W638	Cathepsin L	274	6	25	37.3	0.29	3
Q9NGW3	Cathepsin L	204	7	11	36.9	0.54	3
Q9NB30	Cathepsin L	257	5	15	37.1	0.41	3
Q2HPD3	Cathepsin L1 proteinase	220	12	19	36.5	0.63	6
HAN4005b02.q1kT3	Putative prolylcarboxypeptidase	99[5]	10	33	-	-	
Adult secretome							
Q7JNQ9	Secreted cathepsin L1	552	20	39	36.7	1.82	1
A5Z1V3	Secreted cathepsin L2	444	12	34	37.0	1.16	1
Q9NB30	Cathepsin L	372	9	18	37.1	0.41	1
Q8T5Z9	Cathepsin L	481	17	40	35.1	1.25	1
Q9GRW5	Cathepsin L1	476	21	32	35.1	1.06	1
Q6R018	Cathepsin L protein	451	22	38	36.6	1.37	1
Q24940	Cathepsin L-like proteinase	444	19	33	36.8	1.58	1
A5X483	Cathepsin L2	401	16	45	24.5	2.16	1
Q24945	Cathepsin L-like protease	348	7	35	18.5	0.95	1
P81222	Cathepsin L-like proteinase	297	12	57	12.3	4.58	1
Q9NGW3	Cathepsin L	254	7	12	36.9	0.41	1
A5X484	Cathepsin L1	114	4	20	24.3	0.14	1
A8I598	Cathepsin L	77	2	7	28.1	-	1
Fhep30b01.q1k	Putative prolylcarboxypeptidase	92[5]	8	13	-	-	1

A



B

Accession	*Fasciola* Protein	Score	Unique peptides[1]	% cover	Mr	emPAI[2]	Section[3]
Dormant larvae somatic							
B0LT92	Peroxiredoxin	163	3	16	25.4	0.29	3
NEJ secretome							
Q7M4G1	Fatty acid-binding protein type 2	90	3	27	14.9	0.26	5
Q9U1G6	Fatty acid-binding protein type 3	217	3	31	14.6	1.03	5
Q7M4G0	Fatty acid-binding protein Fh15	104	2	20	14.7	0.61	5
Immature fluke secretome							
Q06A71	GST sigma class	138	4	23	24.5	0.14	3
P56598	GST mu class 26 kDa isozyme 1 (GST1)	303	9	38	25.7	0.63	3
P31671	GST mu class 26 kDa isozyme 7 (GST7)	361	14	34	25.3	2.05	3
P31670	GST mu class 26 kDa isozyme 47 (GST47)	381	15	44	25.3	2.04	3
P30112	GST mu class 26 kDa isozyme 51 (GST51)	390	16	39	25.3	1.70	3
B0LT92	Peroxiredoxin	237	5	21	24.5	0.29	3
O76945	Protein disulfide isomerase	199	6	14	55.1	0.06	3
Q9UAS2	Fatty acid-binding protein 1	77	4	16	14.6	0.23	5
Q7M4G1	Fatty acid-binding protein type 2	141	6	50	14.9	0.23	6
Q9U1G6	Fatty acid-binding protein type 3	141	3	28	14.6	0.23	5
Q7M4G0	Fatty acid-binding protein Fh15	349	14	59	14.7	5.55	5
Adult secretome							
B0LT92	Peroxiredoxin	211	5	21	25.4	0.29	1-3
Q9U1G7	Thioredoxin	83	2	22	11.6	0.30	2
Q9UAS2	Fatty acid-binding protein 1	107	3	16	14.6	0.52	2,3
Q7M4G1	Fatty acid-binding protein type 2	153	5	44	14.9	0.23	2,3
Q9U1G6	Fatty acid-binding protein type 3	81	2	18	14.6	0.23	2
Q7M4G0	Fatty acid-binding protein Fh15	208	7	51	14.7	0.23	2,3

FIG. 3. **Expression of *F. hepatica* antioxidants.** *A*, the expression levels of the major antioxidant molecules used by *F. hepatica* during its mammalian life cycle. The emPAI values (22) derived from *Fasciola* CID data were used to determine the abundance of the various antioxidants that are expressed and/or secreted by juvenile flukes (dormant larvae and NEJs) and secreted by the immature (liver stage) flukes and adult (bile duct stage) parasites. *B*, identification of *F. hepatica* antioxidants by nano-LC-ESI-MS/MS. Dormant larvae somatic proteins and those secreted by NEJs, 21-day-old immature flukes, and adult *F. hepatica* were separated by 1-DE and analyzed by nano-LC-ESI-MS/MS. MASCOT searches were performed against the MSDB. Footnotes are as follows. *1*, number of unique peptide matches. *2*, raw emPAI values provided by the MASCOT search engine give an approximate quantification of the protein (22). These values were converted to a percentage of total antioxidant expression within each *Fasciola* developmental stage (see "Experimental Procedures") and were used to generate the pie charts shown in *A*. *3*, section of the 1-D gel from which the peptides were identified.

dox homeostasis (peroxiredoxin and protein-disulfide isomerase). Other significant peptide matches included annexin, ferritin, ubiquitin, a 14-3-3 protein, a multicystatin, and a putative ATP-binding cassette (ABC) transporter protein.

DISCUSSION

The database of *F. hepatica* ESTs available from the Wellcome Trust Sanger Centre is now sufficiently large to allow a significant transcriptomics analysis of this helminth pathogen that until now has been lacking in this field. We used our newly developed EST2Secretome pipeline to analyze these data sets with the view to identifying molecules secreted by the adult trematode *F. hepatica* unlike earlier applications to nematode parasites that focused on the complete transcriptome. Furthermore we integrated these results with data generated from proteomics analysis performed here and in previous reports (8, 20, 21) to build a picture of how the developing parasite sustains itself within the mammalian host with particular emphasis on proteases as virulence and tissue-damaging factors.

EST2Secretome analysis of the adult *F. hepatica* ESTs identified 160 cDNAs encoding secreted proteins, 41% of which encoded cathepsin L cysteine proteases. The abundance of adult cathepsin L sequences was noted in a previous analysis of entries in the public databases (8, 35). When these sequences were subjected to a phylogenetic investigation it was demonstrated that they could be separated into five clusters or clades: FhCL1, FhCL2, FhCL3, FhCL4, and FhCL5. Because of the high level of conservation between the clades it was not possible to study temporal patterns of specific cathepsin gene expression by conventional means (such as RT-PCR). Analysis of the >14,000 adult parasite EST sequences in the present study confirmed the expression of three of these distinct clades (FhCL1, FhCL2, and FhCL5) in this fully mature stage of the parasite. It also supported the lack of expression of clades FhCL3 and FhCL4 in adult parasites, although proteases encoding these genes are expressed by NEJs (see below and Ref. 28). We also identified five novel cathepsin L proteases that although encoded by partial nucleotide sequences (contigs 1553, 2626, and 1886 and singletons Fhep55b05.q1k and Fhep18b10.q1k) may represent a new phylogenetic cluster.

The *F. hepatica* cathepsin L proteases, the largest family of proteases known in any helminth pathogen, arose by a series of gene duplications. Members underwent selective functional diversity brought about by specific alterations within the active site that ultimately produced a repertoire of proteases with overlapping but distinct substrate specificity (8). Our proteomics investigation was consistent with the EST2Secretome analysis by demonstrating the predominance of these proteases in the adult *F. hepatica* secretome (~80% of total protein secreted) and indicated a major role for the enzymes in supporting the existence of the parasites within the bile duct. Here the adult parasites are obligate blood feeders, and therefore the most obvious function for the cathepsin L enzymes is in the digestion of blood macromolecules. Recently Lowther *et al.* (36) showed that members of the most predominant cathepsin L clade, FhCL1, have evolved an active site with a strong preference for hydrophobic amino acids, such as Leu, Ala, Phe, and Val, that are particularly abundant in hemoglobin (42%) indicating a specific adaptation for digestion of this protein. Hemoglobin is the prime supplier of amino acids that the parasites use in the anabolism of egg proteins, a major task because they produce 30–50,000 eggs/day/worm (37). Members of the other major clade, FhCL2, have specific amino acid substitutions in the active site cleft that gives these proteases the ability to cleave substrates with Pro residues and facilitates the degradation of proline-rich interstitial collagen during the migration of the parasites through the tissues (38).

Consistent with the importance of assimilating nutrients from host red blood cells was the identification of ESTs encoding a saposin-like molecule as the next most abundant in the data set. This protein, which was termed FhSAP-2 because of its similarity to a saposin-like protein family and the amoebapore precursors of *Entamoeba histolytica*, has been implicated in the penetration and lysis of ingested host red blood cells not only in *Fasciola* (29, 30) but also in other helminths (39). Saposin-like molecules are likely secreted from granules within cells lining the parasite esophagus so that ingested red blood cells are lysed as soon as they enter the digestive track. Hydrolysis of released hemoglobin by the cathepsin L proteases then takes place in the acidic cecum (37).

Several *Fasciola* proteins predicted as components of the secretome by our EST2Secretome analysis were not identified by mass spectrometry in secretions of adult parasites maintained in culture (8, 20) or within the bile ducts (21). The most highly represented of these was the major eggshell protein vitelline protein B1 (vpB1), which is produced by mature vitelline cells to form the hard protective trematode eggshell (31). However, because this process occurs within the ootype and uterus of the adult fluke it is likely that vpB1 is retained within the eggshell and not secreted outside the parasite. Other examples are the transcripts encoding six novel cathepsin B cysteine proteases and five new asparaginyl endopeptidases. Although some isotypes of these proteases are secreted by the infective larvae and immature liver stage flukes (see below), they are absent from the secretome of adult *F. hepatica* or *F. gigantica* as previously reported using biochemical, proteomics, and immunoblotting methods (8, 40, 41). Considering these observations and the fact that cathepsin B and asparaginyl endopeptidases also function internally within cells (42, 43), it is likely that these newly discovered proteases have functions within the internal tissues of this complex multicellular parasite and are involved in generalized protein turnover, remodeling, and/or catabolism. Alternatively or additionally, because immunocytochemistry and *in situ* hybridization

have localized cysteine proteases and transcripts to the reproductive structures of other trematodes (44, 45) these proteases may be involved in the process of egg production. Other proteins identified by EST2Secretome analysis but not by proteomics, including the previously uncharacterized proteins such as deoxyribonucleases, carboxypeptidase, and trypsin-like peptidases that have homologues in *Schistosoma japonicum* and *C. sinensis*, may also be confined to internal tissues of the parasite or be expressed and secreted at levels below the detection capacity of the proteomics methods so far used (cDNAs of these were poorly represented in the transcriptome of the adult parasite).

Conversely a number of *F. hepatica* proteins that have previously been described as major components of the adult parasite secretions by standard protein and proteomics methods were not predicted by the EST2Secretome pipeline. These included the fatty acid-binding proteins FaBP1, FaBP2, FaBP3, and Fh15 (20, 21) and two redox enzymes, thioredoxin and peroxiredoxin (20, 46–48). Primary sequence analysis using SignalP (49) confirmed that these proteins lacked predicted N-terminal signal peptides for secretion via the classical ER/Golgi pathway, and hence they must be released by alternate mechanisms. Morphew *et al.* (21) suggested that release of *Fasciola* FaBPs was due to shedding of the tegument as part of a stress response during *in vitro* culture of parasites because they were not detected *in vivo*. However, morphological studies have shown that blebbing or shedding of the *F. hepatica* surface tegument and renewal by proteins synthesized by highly metabolic subtegumental cells is a continuous property that may be an immunodefensive strategy (4, 50). This non-classical mechanism for protein secretion may be compared with those reported for some leaderless eukaryotic proteins, for example the ER/Golgi-independent secretion of leaderless peptides including interleukin (IL)-1α, IL-1β, and fibroblast growth factor-2 (51). In the case of IL-1β, which is processed and activated by caspase-1, it has been proposed that both molecules are packaged together into plasma membrane blebs (by ABC transporters) and are rapidly released as microvesicles following phospholipase-mediated fusion with the plasma membrane (52–54). The identification of a phospholipase in the transcriptome of adult flukes (Table II) together with the presence of ABC transporters within the tegument (55) supports the possibility of a similar mechanism for the ER/Golgi-independent secretion pathway in *Fasciola*.

In summary, our EST2Secretome pipeline was successful in identifying the major secreted proteins of adult *F. hepatica*. Integration of this analysis with proteomics data is important for the study of helminth host-pathogen relationships to distinguish proteins that are secreted extracorporeally from those secreted within the internal tissues of the parasites. Additionally this integrated approach identified major helminth secreted proteins that may reach the exterior by novel or non-classical secretory pathways.

Proteomics Analysis of the Dormant and Infective Stage Larvae—Having collated the secretome data for adult *F. hepatica* we performed a comparative analysis of adult *F. hepatica* with the infective larvae that invade their host by penetrating the intestinal wall. The larvae of *F. hepatica* (<1 mm in length) are released by the intermediate snail host (*Galba truncatula*) and encyst on vegetation that is consumed by the host. For research purposes these are produced in an aquarium, and the larvae are allowed to encyst on cellophane. However, because under these conditions the encysted metacercariae can become contaminated with extraneous material we used a method of washing in 2% hypochlorite to remove this and the outer cyst wall so that a proteomics study of the dormant larvae could be performed. We also used a medium containing bile duct surfactants to activate the larvae and induce them to excyst so that their secretome could be analyzed. We found that the level of *Fasciola* excystment was significantly reduced when the excystment medium excluded the reducing agent L-cysteine (91% reduction, $p = 0.0008$; not shown) compared with medium containing L-cysteine. The requirement for reducing conditions implied a role for cysteine proteases in the excystment process because the activity of these thiol-dependent proteases is enhanced in the presence of reducing agents. We therefore proved this point by showing that the level of excystment was significantly reduced when the broad range cysteine protease inhibitor E-64 (98% reduction, $p = 0.0433$) or the cathepsin-specific inhibitor Z-Phe-Ala-CHN_2 (99% reduction, $p = 0.0205$) were added to the excystment medium. Proteomics analysis of soluble extracts of dormant larvae and the secretions of the NEJs identified 26 and 29 proteins, respectively (Figs. 2*B* and 3*B* and supplemental Tables 2 and 3). Significantly 23% of all the somatic larval proteins and 31% of the secreted proteins were cysteine proteases consisting of cathepsin L (37%), cathepsin Bs (45%), and asparaginyl endopeptidases (18%). Because the former two classes are potently inhibited by E-64 and Z-Phe-Ala-CHN_2 but the latter class is not, the results implicate cathepsin Ls and/or cathepsin B in the process of cyst rupture.

Phylogenetics studies showed that cDNA clones generated from transcripts of *Fasciola* larvae, designated clade FhCL3, encoded a cathepsin L protease that was specific to this infective stage (8, 35); the cDNA encoding this protease, FhCL3_nl22, was originally described by Harmsen *et al.* (56). Earlier studies by Tkalcevic *et al.* (34) using N-terminal sequencing to identify proteins expressed by larvae reported a sequence of a cathepsin L cysteine protease that we now know matches exactly with the predicted N terminus of FhCL3_nl22 (56). More recently, Cancela *et al.* (28) confirmed the restricted expression of FhCL3 and another protease, FhCL4, to infective larvae by isolating cDNAs encoding these enzymes from a *F. hepatica* larvae-specific cDNA library and subsequently by using RT-PCR expression analysis. No FhCL4 peptides were identified in the present analysis suggesting that this enzyme may be expressed at low levels,

performs specific intracellular functions, and is not secreted by *F. hepatica*. However, we confirmed the presence of FhCL3 in NEJ by identifying several FhCL3-specific peptides with matches to FhCL3_nl22. It is noteworthy that the mass of the dormant larvae somatic cathepsin L3, which was found in a protein band migrating at ~37 kDa, is consistent with that of an inactive cathepsin L precursor, or zymogen, which is supported by the presence of mass ion *m/z* 644.28, +2 that matched to a peptide (SNDVSWHEWK) found only within the N-terminal prosegment region of the protease. No peptides matching with cathepsin Ls were identified in protein bands in the region corresponding to the fully processed and active mature enzymes, *i.e.* mass ~24 kDa (Fig. 1, NEJ secretome gel *section 3*). By contrast, although a few cathepsin L3 peptides were present at ~37 kDa in the secretome of NEJ (including the FhCL3 prosegment peptides LGLNQFTDLTFEEFK (*m/z* 901.49, +2) and SNDVSWHEWK (*m/z* 644.28, +2)), the most robust peptide matches were obtained from gel sections at the molecular mass of ~24 kDa corresponding to the size of a fully activated mature enzyme (Fig. 1, NEJ secretome gel *section 3*). Collectively these observations show that the dormant larvae of *F. hepatica* express a specific cathepsin L protease, FhCL3, stored as an inactive zymogen that is rapidly secreted and activated following emergence from their cysts to become infective larvae.

Fasciola cathepsin B-like cysteine proteases were also identified in somatic extracts of larvae with peptides matching to proteins encoded by three different cDNAs that were previously designated cathepsin B1 (UniProt accession number A7UNB2),[2] cathepsin B2 (UniProt accession number Q8I7B2),[3] and cathepsin B3 (UniProt accession number A5X494; Ref. 28). All three cathepsin B proteases were identified in intense protein bands that migrated in reducing SDS-PAGE at molecular masses of ~36–50 kDa and ~50–70 kDa (Fig. 1, dormant larvae gel *sections 1* and *2*) suggesting that like FhCL3 these may represent stored zymogens. The presence of peptides FINIEHFK (*m/z* 524.28, +2) and QHLGLLEETPEER (*m/z* 775.89, +2) confirmed the existence of cathepsin B2 as an unprocessed zymogen whereas peptides QNLGVLEETPEDR (*m/z* 750.36, +2) and YSVSENDLPESFDAR (*m/z* 864.90, +2; which spans the juncture between the prosegment and the mature domain) indicate that cathepsins B1 and B3 are also stored as zymogens by *Fasciola* NEJs.

However, unlike the FhCL3, we also found some cathepsin Bs at ~20–36 kDa (Fig. 1, NEJ secretome gel *section 3*), which is consistent with the occurrence of processed active enzymes within the dormant larvae (which have a predicted molecular mass of 29.6 kDa; Ref. 57). Although each of the three cathepsin Bs were also found as fully active enzymes in the secretions of larvae, in the region of ~20–36 kDa, we also discovered a novel family member, encoded by a *Fasciola* EST HAN4015b05.q1kT3 (here designated cathepsin B4) (Fig. 2*B* and Table III).

Cysteine proteases are produced as inactive zymogens consisting of a prosegment and mature domain (58). The prosegment lies along the active site groove (in reverse to the direction of protein substrates) preventing unwanted hydrolysis during trafficking and storage of the protease within the cell. Removal of the prosegment from the mature domain exposes the active site of the hydrolase to entry of macromolecular protein substrates. Dalton *et al.* (59) proposed that this activation step is mediated by proteolytic clipping at a "protease-susceptible" region between the prosegment and mature enzyme and that this event may be performed by the same or another protease molecule. Dalton and Brindley (60) and Dalton *et al.* (59) proposed that in *F. hepatica* and *S. mansoni* an asparaginyl endopeptidase (otherwise known as legumain), which cleaves peptide bonds C-terminal to Asn residues, is involved in the *trans*-processing and activation of cathepsin L and B cysteine proteases. In support of this argument all members of these classes of cysteine protease in both helminth parasites possess an Asn residue in the vicinity of the prosegment/mature juncture (8, 60, 61). Furthermore, Sajid *et al.* (62) and Beckham *et al.* (57) provided experimental evidence by showing that a recombinant asparaginyl endopeptidase could *trans*-process and activate recombinant cathepsin B of *F. hepatica* and *S. mansoni*. In the present study, we identified two asparaginyl endopeptidases in the somatic proteins and secretome of *Fasciola* NEJs. One of the identified enzymes (legumain-like precursor, TrEMBL accession number Q711M2; termed legumain 1) matched exactly with an N-terminal sequence for asparaginyl endopeptidase obtained from NEJ somatic proteins by Tkalcevic *et al.* (34). Although this enzyme retains the His^{158} of the conserved catalytic dyad residues, the Cys^{200} is replaced by a serine residue, and therefore this enzyme may display an altered substrate specificity compared with the normal asparaginyl endopeptidases. However, the enzyme does contain a conserved asparagine residue (Asn^{323}) that is required for C-terminal processing to an active mature enzyme (63). The second legumain-like protease identified in the NEJ somatic protein extracts is encoded by a *Fasciola* EST (identifier Fhep29h09.q1k; termed legumain 2), but as this EST is incomplete at the 5′-end, it is not possible to determine whether the enzyme encoded by this sequence retains the conserved active site His^{158}/Cys^{200} dyad. Interestingly the enzyme encoded by the legumain 2 EST lacks the conserved Asn^{323} required for autoactivation. The overlapping primary sequence of the two enzymes exhibits 51% identity, and therefore individual enzymes were easily differentiated on the basis of their tryptic peptides. Notwithstanding the anomalies within the sequence of these two enzymes, the data suggest that the NEJ possesses the enzymatic machinery to rapidly process and activate the stored zymogens of cathepsins L and B. Recently Morita *et al.* (64) reported that purified bovine aspar-

[2] E. Ljunggren, M. Kozak, and L. Jedlina-Panasiuk, unpublished data.
[3] E. Khaznadji, M. Peloille, and N. Moire, unpublished data.

TABLE III

Database of the repertoire of proteases expressed by F. hepatica in the mammalian host

Somatic proteases expressed by dormant larvae and those secreted by NEJs, immature flukes, and adult *F. hepatica* identified by mass spectrometry are shown.

Accession no.	*Fasciola* protein	Dormant larvae	NEJ secretome	Immature secretome	Adult secretome
FhCL1A					
Q8T5Z9	Cathepsin L	−	−	+	+
Q7JNQ9	Cathepsin L1	−	−	+	+
Q6R018	Cathepsin L protein	−	−	+	+
Q2HPD3	Cathepsin L1 proteinase	−	−	+	+
Q24940	Cathepsin L-like proteinase	−	−	+	+
FhCL1B					
Q9GRW5	Cathepsin L1	−	−	+	+
FgCL1C					
Q8MUT6	Cathepsin L	−	−	−	−
FhCL2					
Q7JNQ8	Secreted cathepsin L2	−	−	+	+
A5Z1V3	Secreted cathepsin L2	−	−	+	+
A5X483	Cathepsin L2	−	−	+	+
A3FMG6	Cathepsin L	−	−	+	−
FhCL3					
Q95VA7	Cathepsin L	+	+	+	−
A8W638	Cathepsin L	+	+	+	−
A8W7J0	Metacercarial procathepsin L	−	+	−	−
Q9GRW4	Partial procathepsin L3	−	+	−	−
Q9GRW6	Partial procathepsin L3	−	+	−	−
B3TM67	Cathepsin L3	−	+	−	−
B3TM68	Cathepsin L3	−	+	−	−
FhCL4					
Fhep55b05.q1k	Putative cathepsin L	−	−	−	−
FhCL5					
Q9NGW3	Cathepsin L	−	−	+	+
Q9NB30	Cathepsin L	−	−	+	+
Cathepsin B					
A7UNB2	Cathepsin B1	+	−	+	−
Q8I7B2	Cathepsin B2	+	−	+	−
A5X494	Cathepsin B3	+	−	−	−
HAN4015b05.q1kT3	Putative cathepsin B4	−	+	−	−
Fhep45b05.q1k	Putative cathepsin B5	−	−	+	−
HAN4006g01.q1kT3	Putative cathepsin B6	−	−	+	+[a]
Fhep44e10.q1k	Putative cathepsin B7	−	−	−	+[a]
Fhep11b02.q1k	Putative cathepsin B8	−	−	−	+[a]
FhContig1639	Putative cathepsin B9	−	−	−	+[a]
FhContig2164	Putative cathepsin B10	−	−	−	+[a]
Asparaginyl endopeptidases					
Q711M2	Legumain-like precursor 1	+	+	+	−
Fhep29h09.q1k	Putative legumain 2	+	+	−	−
A6Y9U8	Legumain 3	−	−	+	+[a]
A6Y9U9	Legumain 4	−	−	+	+[a]
Fhep21f02.q1k	Putative legumain 5	−	−	−	+[a]
FhContig1272	Putative legumain 6	−	−	−	+[a]
FhContig2292	Putative legumain 7	−	−	−	+[a]
Prolylcarboxypeptidase (s28)					
Fhep30b01.q1k	Putative prolylcarboxypeptidase	−	−	+	+
Serine carboxypeptidase (s10)					
FhContig542	Putative serine carboxypeptidase	−	−	−	+[a]
Trypsin-like serine proteases					
FhContig492	Putative peptidase S1/S6	−	−	−	+[a]
FhContig2453	Putative trypsin-like Ser/Cys	−	−	−	+[a]

[a] Expression data only available at the transcript level.

aginyl endopeptidase can degrade fibronectin, a major component of the extracellular matrix. If they possess similar biochemical properties, a second tangible role for the asparaginyl endopeptidases secreted by the infective larvae could be to facilitate penetration of the host intestine.

Proteomics Analysis of the Migrating Liver Stage Parasite—The developmental stage of *F. hepatica* that migrates through the liver tissue is responsible for the clinical manifestations associated with acute animal and human fasciolosis. The migrating parasite causes extensive physical tissue destruction, tunneling, and hemorrhaging and induces immunologically related inflammatory damage (65). To gain insight into the variety of molecules produced by this parasite stage that may induce this pathology 21-day-old immature parasites were removed from the liver of infected mice and maintained in culture. A proteomics analysis of the medium revealed that these parasites secreted a greater range of proteins than the infective and adult stage flukes. In particular, they secreted the greatest range of cathepsin L cysteine proteases, including the FhCL3 that we identified in the NEJ (this was secreted solely as a mature active form of ~24 kDa). Members of cathepsin L clades FhCL1, FhCL2, and FhCL5 originally identified in the adult parasite were also secreted by the immature flukes (these could be distinguished on the basis of the clade-/subclade-specific peptide matches described above and reported previously by Robinson *et al.* (8)).

Fasciola cathepsins B1 and B2 were also secreted by 21-day-old immature flukes, whereas cathepsin B3 was not detected. However, mass spectrometry data matched with two novel putative cathepsin B enzymes that were discovered in the EST2Secretome of adult ESTs. The enzyme encoded by EST Fhep45b05.q1k (designated cathepsin B5) was identified based on matches with the sequence-specific peptides NIMYEIMK (*m/z* 521.25, +2), LLGWGVEDGEK (*m/z* 601.79, +2), and FYAISSYNVYGGEK (*m/z* 799.36, +2) whereas the other, EST HAN4006g01.q1kT3 (designated cathepsin B6), was differentiated due to specific matches with the peptides FSTPK (*m/z* 579.29, +1), HTTGALLGGHAIR (*m/z* 652.35, +2), and TSYNLLHNEETIMK (*m/z* 846.90, +2).

Three asparaginyl endopeptidases were identified within the secretory proteins of the immature 21-day-old flukes. These included the NEJ legumain-like precursor legumain 1 with the altered Cys^{200}/Ser^{200} active site residue but did not include the NEJ legumain 2. However, two asparaginyl endopeptidases (designated legumains 3 and 4) appeared to be specific to the liver migrating stage and, like their homologues that were originally identified in *F. gigantica* (41), retained the conserved catalytic His^{158}/Cys^{200} dyad and Asn^{323} that is required for C-terminal processing (63). Unlike the asparaginyl endopeptidases of NEJs, those secreted by the immature flukes all migrated at ~36 kDa suggesting that they had undergone some processing events to lead to fully active enzymes. It is possible that besides functioning in the *trans*-processing of cathepsins L and B the asparaginyl endopeptidases of the immature parasites participate in host tissue degradation.

Mass spectrometry data from the immature 21-day-old *F. hepatica* (and subsequently found in the adult parasite secretome) matched with several ESTs encoding a putative prolylcarboxypeptidase (Fig. 2*B* and Table III). The exopeptidase contains a predicted N-terminal signal peptide and a conserved catalytic triad (Ser^{161}, Asp^{413}, and His^{438}) that is characteristic of the s28 family of serine peptidases including prolylcarboxypeptidase. Although this is the first report of a putative prolylcarboxypeptidase from a trematode pathogen, orthologues have been identified in parasitic nematodes including the gastrointestinal worm *Haemonchus contortus* where the enzymes act as anticoagulants to facilitate a blood-feeding lifestyle (66). It is likely that the *Fasciola* prolylcarboxypeptidase performs a similar role, acting alongside the saposin-like proteins to prevent blood coagulation to ensure effective lysis of ingested host red blood cells. A role for the putative prolylcarboxypeptidase in *Fasciola* nutrition is supported by the absence of the enzyme in the NEJ, which does not possess a functional gut.

Relative Expression Levels of Fasciola Proteases and Relationship to Virulence and Tissue Invasion—For *F. hepatica* NEJs, the emPAI data indicated that cathepsin L3 and cathepsin B enzymes were expressed at similar high levels (accounting for 37 and 45% of total proteases, respectively) whereas asparaginyl endopeptidases represented 18% of total proteases detected. By using inhibitors of cathepsin-like proteases we showed that excystment of the infective larvae is dependent on the cathepsin L3 and cathepsin B proteases. Asparaginyl endopeptidases are likely to be essential for the rapid *trans*-processing and activation of the stored zymogens of cathepsins L and B, and hence it is probable that one of the first steps in activation of the dormant larvae is the switching on of the genes encoding these enzymes. A role for cathepsin L and B proteases in the penetration of the host by NEJ was recently demonstrated using RNA interference methods by McGonigle *et al.* (67). Knockdown of both cathepsin L and cathepsin B transcripts in NEJ parasites blocked the ability of the larvae to penetrate and traverse the intestinal wall of a rodent host. Therefore, the two types of cysteine proteases, working together, must be responsible for degrading the intestinal tissue macromolecules to enable the rapid penetration of the host intestine. Cysteine proteases are also essential for infectivity of the closely related schistosome parasites (*S. mansoni* and *Trichobilharzia regenti*) that enter their hosts via the skin (68) and for the cyst emergence and infectivity of larvae of the trematode *Paragonimus westermani* (69–71).

Correlating with the migration of the parasites into the liver tissue and the development of the gut apparatus (beginning at ~10 days; Ref. 36) the expression pattern of proteases becomes more complex. For the migrating parasite, the secretion of FhCL3 and cathepsin B becomes less important as these proteases account for only 3 and 2% of total proteases,

respectively. In contrast, members of the other cathepsin L clades become more highly represented in the secretome: FhCL1 (50%), FhCL2 (27%), and FhCL5 (7%) (Fig. 2*A*). This striking shift in protease expression indicates a requirement for a new set of proteases for liver migration and for the degradation of tissue and blood macromolecules. As discussed above FhCL1 proteases have a particular adaptation to hemoglobin digestion whereas FhCL2 proteases exhibit unique collagenase-like activities (FhCL5 appears to be more closely related to FhCL1; Ref. 72). Asparaginyl endopeptidase secretion remains relatively high, accounting for 11% of total proteases secreted by the immature flukes, that may be required for activation of the cathepsin L proteases and possibly for directly engaging in tissue degradation. The expression of the prolylcarboxypeptidase and saposin at this stage signals the beginning of blood feeding. Working together this profile of molecules is an effective tissue-destroying machinery essential to parasite migration and growth but, at the same time, responsible for the pathogenesis associated with acute animal and human fasciolosis.

After the parasite has entered the bile duct it completes its maturation and becomes an obligate blood feeder. Blood provides all the necessary nutrients (macro- and micromolecules) needed for production of an enormous numbers of eggs, the principle function of the adult parasite (37). The total protease secretion is assumed completely by the various cathepsin L proteases, clades FhCL1 (69%), FhCL2 (22%), and FhCL5 (9%), which can degrade macromolecules such as hemoglobin to small peptides that are absorbed by the parasite for further digestion to free amino acids within its tissues (36, 37). The emPAI values determined for adult parasite proteases are similar to expression data obtained using 2-DE and densitometry (67, 27, and 5%, respectively; Ref. 8) and validate the use of emPAI for estimating *Fasciola* secretory protein levels in the other developmental stages. Within the bile duct the parasite can exist for years (cattle) or even decades (sheep) causing relatively little pathology apart from anemia and bile duct hyperplasia.

F. hepatica Secretes a Battery of Antioxidant Molecules—It is worth highlighting that for all three developmental stages of *F. hepatica* investigated in this study an array of antioxidant molecules was secreted in an abundance second only to the proteases. These included peroxiredoxin, thioredoxin, protein-disulfide isomerase, four fatty acid-binding proteins (FaBP1–3 and Fh15), and five GSTs (Sigma and Mu classes) (20, 21, 32, 46–48, 73). These molecules have been implicated in fluke immune avoidance mechanisms; for example, secreted isoforms of *Fasciola* GSTs were shown to decrease the proliferative response of rat spleen cells and diminish nitric oxide production by macrophages (73) whereas peroxiredoxin plays a key role driving host Th2 immune responses via the recruitment and alternative activation of peritoneal macrophages (74, 75). *Fasciola* peroxiredoxin, thioredoxin, and protein-disulfide isomerase may also protect the parasites from harmful reactive oxygen species released by host immune cells (32, 47, 48, 76).

Analysis of the emPAI values for these various antioxidants indicates that, like the proteases, their secretion is highly regulated during the migration of the parasite (Fig. 3*A*). The profile of secretory antioxidants is similar for the NEJs and adult flukes; both secrete several FaBPs together with peroxiredoxin (although thioredoxin is also secreted by the adult worms). However, the diversity of the antioxidant molecules secreted by the immature liver stage parasites is notably different from those secreted by the other two stages with the dramatic production of GSTs, including Sigma class GST and four isoforms of Mu class GSTs (GST1, GST7, GST47, and GST 51); together these account for 50% of total antioxidants expressed at this stage. GST may provide a particular defense for the fluke against a mounting host cellular immune response as the immature parasite is in direct contact with the immune system as it migrates through the liver parenchyma. At this stage also the parasite has moved from an aerobic to an anaerobic environment, and therefore significant changes take place in its metabolism, particularly in the tegument, which is the primary host-parasite interface (77). Clearly the combined action of FaBPs, peroxiredoxin, and thioredoxin offers sufficient protection for the adult flukes residing within the immunologically privileged bile ducts.

It was also noted that each of the aforementioned antioxidant classes found in the secretome of *F. hepatica* does not possess a signal sequence for secretion (with the exception of protein-disulfide isomerase that also has a putative ER retention signal and may act as a chaperone to ensure proper folding of classically secreted *Fasciola* peptides). This suggests that they are secreted via non-classical mechanisms, possibly shedding or blebbing of the tegument (see above). Accordingly we propose that two distinct mechanisms for protein export operate for the most abundant proteins in *Fasciola*: (*a*) a classical secretion of proteases and other molecules via the ER/Golgi from specialized gastrodermal cells of the gut as reported previously (78) and (*b*) a non-classical secretion of leaderless *Fasciola* antioxidants via a trans-tegumental route. This has clear implications for the development of novel anti-*F. hepatica* chemotherapy and could explain the additive effects of two of the most potent flukicides, triclabendazole and clorsulon, which act upon the tegument and gastrodermal cells, respectively (79–81).

Conclusion—Our integrated transcriptomics and proteomics analysis has provided an in-depth overview of protein secretion by the animal and human pathogen *F. hepatica* that represents a significant step toward a comprehensive understanding of the host-parasite interactions in fasciolosis. Of vital importance now is the completion of transcriptome data sets for both the infective NEJ larvae and immature liver stage parasites so that a comparative analysis of secretory proteins can be performed. Nevertheless we have described the secretion of the major molecules of this parasite and correlated

their expression with critical steps in its migration and development within the mammalian host. This will allow a more strategic and rational approach to future anthelmintic or vaccine development programs. It is interesting that several of the major components of *Fasciola* secretions identified here are currently leading candidates for development as first generation antifluke vaccines (including cathepsins, peroxiredoxin, glutathione *S*-transferase, and fatty acid-binding proteins; Ref. 82) or are promising targets for novel flukicidal drugs (cathepsins; Ref. 83) demonstrating the value of our integrated approach for the future identification of new targets for therapeutic intervention.

Acknowledgments—We thank Matt Padula, Proteomics Technology Centre of Expertise, University of Technology Sydney, for assistance with the mass spectrometry and Matt Berriman of the Wellcome Trust Sanger Centre.

* The bioinformatics analysis was supported in part by Australian Research Council Grant LP0667795.

S The on-line version of this article (available at http://www.mcponline.org) contains supplemental material.

§ Supported by a University of Technology Sydney chancellor's postdoctoral fellowship. To whom correspondence should be addressed: Inst. for the Biotechnology of Infectious Diseases, University of Technology Sydney, Bldg. 4, Harris St., Ultimo, Sydney, New South Wales 2007, Australia. Tel.: 61-2-95144127; Fax: 61-2-95148206; E-mail: mark.robinson@uts.edu.au.

‖ Supported by the National Health and Medical Research Council of Australia Project Grant 352912.

** Both authors contributed equally to this work.

‡‡ Recipient of a New South Wales Government BioFirst Award 2004.

REFERENCES

1. Garcia, H. H., Moro, P. L., and Schantz, P. M. (2007) Zoonotic helminth infections of humans: echinococcosis, cysticercosis and fascioliasis. *Curr. Opin. Infect. Dis.* **20,** 489–494
2. Andrews, S. J. (1999) The life-cycle of *Fasciola hepatica*, in *Fasciolosis* (Dalton, J. P., ed) pp. 1–29, CAB International, Oxford
3. Mas-Coma, S., Bargues, M. D., and Esteban, J. G. (1999) Human fascioliosis, in *Fasciolosis* (Dalton, J. P., ed) pp. 411–434, CAB International, Oxford
4. Fairweather, I., Threadgold, L. T., and Hanna, R. E. (1999) Development of *Fasciola hepatica* in the mammalian host, in *Fasciolosis* (Dalton, J. P., ed) pp. 47–111, CAB International, Oxford
5. Hewitson, J. P., Harcus, Y. M., Curwen, R. S., Dowle, A. A., Atmadja, A. K., Ashton, P. D., Wilson, A., and Maizels, R. M. (2008) The secretome of the filarial parasite, *Brugia malayi*: proteomic profile of adult excretory-secretory products. *Mol. Biochem. Parasitol.* **160,** 8–21
6. Robinson, M. W., and Connolly, B. (2005) Proteomic analysis of the excretory-secretory proteins of the *Trichinella spiralis* L1 larva, a nematode parasite of skeletal muscle. *Proteomics* **5,** 4525–4532
7. Robinson, M. W, Greig, R., Beattie, K. A., Lamont, D. J., and Connolly, B. (2007) Comparative analysis of the excretory-secretory proteome of the muscle larva of *Trichinella pseudospiralis* and *Trichinella spiralis. Int. J. Parasitol.* **37,** 139–148
8. Robinson, M. W., Tort, J. F., Lowther, J., Donnelly, S. M., Wong, E., Xu, W., Stack, C. M., Padula, M., Herbert, B., and Dalton, J. P. (2008) Proteomics and phylogenetic analysis of the cathepsin L protease family of the helminth pathogen, *Fasciola hepatica*: expansion of a repertoire of virulence-associated factors. *Mol. Cell. Proteomics* **7,** 1111–1123
9. Knudsen, G. M., Medzihradszky, K. F., Lim, K. C., Hansell, E., and McKerrow, J. H. (2005) Proteomic analysis of *Schistosoma mansoni* cercarial secretions. *Mol. Cell. Proteomics* **4,** 1862–1875
10. Ojopi, E. P., Oliveira, P. S., Nunes, D. N., Paquola, A., DeMarco, R., Gregório, S. P., Aires, K. A., Menck, C. F., Leite, L. C., Verjovski-Almeida, S., and Dias-Neto, E. (2007) A quantitative view of the transcriptome of *Schistosoma mansoni* adult-worms using SAGE. *BMC Genomics* **8,** 186
11. Oliveira, G. (2007) The *Schistosoma mansoni* transcriptome: an update. *Exp. Parasitol.* **117,** 229–235
12. Laha, T., Pinlaor, P., Mulvenna, J., Sripa, B., Sripa, M., Smout, M. J., Gasser, R. B., Brindley, P. J., and Loukas, A. (2007) Gene discovery for the carcinogenic human liver fluke, *Opisthorchis viverrini*. *BMC Genomics* **8,** 189
13. Nagaraj, S. H., Gasser, R. B., and Ranganathan, S. (2008) Needles in the EST haystack: large-scale identification and analysis of excretory-secretory (ES) proteins in parasitic nematodes using expressed sequence tags (ESTs). *PLoS Negl. Trop. Dis.* **2,** e301
14. Nagaraj, S. H., Deshpande, N., Gasser, R. B., and Ranganathan, S. (2007) ESTExplorer: an expressed sequence tag (EST) assembly and annotation platform. *Nucleic Acids Res.* **35,** W143–147
15. Nagaraj, S. H., Gasser, R. B., Nisbet, A. J., and Ranganathan, S. (2008) In silico analysis of expressed sequence tags from *Trichostrongylus vitrinus* (Nematoda): comparison of the automated ESTExplorer workflow platform with conventional database searches. *BMC Bioinformatics* **9,** Suppl. 1, S10
16. Ranganathan, S., Nagaraj, S. H., Hu, M., Strube, C., Schnieder, T., and Gasser, R. B. (2007) A transcriptomic analysis of the adult stage of the bovine lungworm, *Dictyocaulus viviparus. BMC Genomics* **8,** 311
17. Datu, B. J., Gasser, R. B., Nagaraj, S. H., Ong, E. K., O'Donoghue, P., McInnes, R., Ranganathan, S., and Loukas, A. (2008) Transcriptional changes in the hookworm, *Ancylostoma caninum*, during the transition from a free-living to a parasitic larva. *PLoS Negl. Trop. Dis.* **2,** e130
18. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402
19. Zdobnov, E. M, and Apweiler, R. (2001) InterProScan: an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17,** 847–848
20. Jefferies, J. R., Campbell, A. M., van Rossum, A. J., Barrett, J., and Brophy, P. M. (2001) Proteomic analysis of *Fasciola hepatica* excretory-secretory products. *Proteomics* **1,** 1128–1132
21. Morphew, R. M., Wright, H. A., LaCourse, E. J., Woods, D. J., and Brophy, P. M. (2007) Comparative proteomics of excretory-secretory proteins released by the liver fluke *Fasciola hepatica* in sheep host bile and during *in vitro* culture ex host. *Mol. Cell. Proteomics* **6,** 963–972
22. Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., and Mann, M. (2005) Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol. Cell. Proteomics* **4,** 1265–1272
23. El-Sayed, N. M., Bartholomeu, D., Ivens, A., Johnston, D. A., and LoVerde, P. T. (2004) Advances in schistosome genomics. *Trends Parasitol.* **20,** 154–157
24. Mitreva, M., Zarlenga, D. S., McCarter, J. P., and Jasmer, D. P. (2007) Parasitic nematodes: from genomes to control. *Vet. Parasitol.* **148,** 31–42
25. Stein, L. D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M. R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., Coulson, A., D'Eustachio, P., Fitch, D. H., Fulton, L. A., Fulton, R. E., Griffiths-Jones, S., Harris, T. W., Hillier, L. W., Kamath, R., Kuwabara, P. E., Mardis, E. R., Marra, M. A., Miner, T. L., Minx, P., Mullikin, J. C., Plumb, R. W., Rogers, J., Schein, J. E., Sohrmann, M., Spieth, J., Stajich, J. E., Wei, C., Willey, D., Wilson, R. K., Durbin, R., and Waterston, R. H. (2003) The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol.* **1,** E45
26. Chen, Y., Zhang, Y., Yin, Y., Gao, G., Li, S., Jiang, Y., Gu, X., and Luo, J. (2005) SPD—a web-based secreted protein database. *Nucleic Acids Res.* **33,** D169–173
27. Choo, K. H., Tan, T. W., and Ranganathan, S. (2005) SPdb—a signal peptide database. *BMC Bioinformatics* **6,** 249
28. Cancela, M., Acosta, D., Rinaldi, G., Silva, E., Durán, R., Roche, L., Zaha, A., Carmona, C., and Tort, J. F. (2008) A distinctive repertoire of cathepsins is expressed by juvenile invasive *Fasciola hepatica. Biochimie* **90,**

1461–1475
29. Grams, R., Adisakwattana, P., Ritthisunthorn, N., Eursitthichai, V., Vichasri-Grams, S., and Viyanant, V. (2006) The saposin-like proteins 1, 2, and 3 of *Fasciola gigantica*. *Mol. Biochem. Parasitol.* **148,** 133–143
30. Espino, A. M., and Hillyer, G. V. (2003) Molecular cloning of a member of the *Fasciola hepatica* saposin-like protein family. *J. Parasitol.* **89,** 545–552
31. Robinson, M. W., Colhoun, L. M., Fairweather, I., Brennan, G. P., and Waite, J. H. (2001) Development of the vitellaria of the liver fluke, *Fasciola hepatica* in the rat host. *Parasitology* **123,** 509–518
32. Salazar-Calderón, M., Martín-Alonso, J. M., Castro, A. M., and Parra, F. (2003) Cloning, heterologous expression in *Escherichia coli* and characterization of a protein disulfide isomerase from *Fasciola hepatica*. *Mol. Biochem. Parasitol.* **126,** 15–23
33. Bork, P., and Beckmann, G. (1993) The CUB domain. A widespread module in developmentally regulated proteins. *J. Mol. Biol.* **231,** 539–545
34. Tkalcevic, J., Ashman, K., and Meeusen, E. (1995) Fasciola hepatica: rapid identification of newly excysted juvenile proteins. *Biochem. Biophys. Res. Commun.* **213,** 169–174
35. Irving, J. A., Spithill, T. W., Pike, R. N., Whisstock, J. C., and Smooker, P. M. (2003) The evolution of enzyme specificity in *Fasciola* spp. *J. Mol. Evol.* **57,** 1–15
36. Lowther, J., Robinson, M. W., Donnelly, S. M., Xu, W., Stack, C. M., Matthews, J. M., and Dalton, J. P. (2009) The importance of pH in regulating the function of *Fasciola hepatica* cathepsin L1 cysteine protease. *PLoS. Negl. Trop. Dis.* **3,** e369
37. Dalton, J. P., Caffrey, C. R., Sajid, M., Stack, C., Donnelly, S., Loukas, A., Don, T., McKerrow, J., Halton, D. W., and Brindley, P. J. (2006) Proteases in trematode biology, in *Parasitic Flatworms: Molecular Biology, Biochemistry, Immunology and Physiology* (Maule, A. G., and Marks, N. J., eds) pp. 348–368, CAB International, Oxford
38. Stack, C. M., Caffrey, C. R., Donnelly, S. M., Seshaadri, A., Lowther, J., Tort, J. F., Collins, P. R., Robinson, M. W., Xu, W., McKerrow, J. H., Craik, C. S., Geiger, S. R., Marion, R., Brinen, L. S., and Dalton, J. P. (2008) Structural and functional relationships in the virulence-associated cathepsin L proteases of the parasitic liver fluke, *Fasciola hepatica*. *J. Biol. Chem.* **283,** 9896–9908
39. Don, T. A., Oksov, Y., Lustigman, S., and Loukas, A. (2007) Saposin-like proteins from the intestine of the blood-feeding hookworm, *Ancylostoma caninum*. *Parasitology* **134,** 427–436
40. Law, R. H., Smooker, P. M., Irving, J. A., Piedrafita, D., Ponting, R., Kennedy, N. J., Whisstock, J. C., Pike, R. N., and Spithill, T. W. (2003) Cloning and expression of the major secreted cathepsin B-like protein from juvenile *Fasciola hepatica* and analysis of immunogenicity following liver fluke infection. *Infect. Immun.* **71,** 6921–6932
41. Adisakwattana, P., Viyanant, V., Chaicumpa, W., Vichasri-Grams, S., Hofmann, A., Korge, G., Sobhon, P., and Grams, R. (2007) Comparative molecular analysis of two asparaginyl endopeptidases and encoding genes from *Fasciola gigantica*. *Mol. Biochem. Parasitol.* **156,** 102–116
42. Turk, D., and Guncar, G. (2003) Lysosomal cysteine proteases (cathepsins): promising drug targets. *Acta Crystallogr. D Biol. Crystallogr.* **59,** 203–213
43. Chen, J. M., Dando, P. M., Stevens, R. A., Fortunato, M., and Barrett, A. J. (1998) Cloning and expression of mouse legumain, a lysosomal endopeptidase. *Biochem. J.* **335,** 111–117
44. Michel, A., Ghoneim, H., Resto, M., Klinkert, M. Q., and Kunz, W. (1995) Sequence, characterization and localization of a cysteine proteinase cathepsin L in *Schistosoma mansoni*. *Mol. Biochem. Parasitol.* **73,** 7–18
45. Meemon, K., Grams, R., Vichasri-Grams, S., Hofmann, A., Korge, G., Viyanant, V., Upatham, E. S., Habe, S., and Sobhon, P. (2004) Molecular cloning and analysis of stage and tissue-specific expression of cathepsin B encoding genes from *Fasciola gigantica*. *Mol. Biochem. Parasitol.* **136,** 1–10
46. Gourbal, B. E., Guillou, F., Mitta, G., Sibille, P., Thèron, A., Pointier, J. P., and Coustau, C. (2008) Excretory-secretory products of larval *Fasciola hepatica* investigated using a two-dimensional proteomic approach. *Mol. Biochem. Parasitol.* **161,** 63–66
47. Salazar-Calderón, M., Martín-Alonso, J. M., Ruiz de Eguino, A. D., and Parra, F. (2001) Heterologous expression and functional characterization of thioredoxin from *Fasciola hepatica*. *Parasitol Res.* **87,** 390–395
48. Sekiya, M., Mulcahy, G., Irwin, J. A., Stack, C. M., Donnelly, S. M., Xu, W., Collins, P., and Dalton, J. P. (2006) Biochemical characterisation of the recombinant peroxiredoxin (FhePrx) of the liver fluke, *Fasciola hepatica*. *FEBS Lett.* **580,** 5016–5022
49. Emanuelsson, O., Brunak, S., von Heijne, G., and Nielsen, H. (2007) Locating proteins in the cell using TargetP, SignalP, and related tools. *Nat. Protoc.* **2,** 953–971
50. Dalton, J. P., Skelly, P., and Halton, D. W. (2004) Role of the tegument and gut in nutrient uptake by parasitic platyhelminths. *Can. J. Zool.* **82,** 211–232
51. Keller, M., Rüegg, A., Werner, S., and Beer, H. D. (2008) Active caspase-1 is a regulator of unconventional protein secretion. *Cell* **132,** 818–831
52. MacKenzie, A., Wilson, H. L., Kiss-Toth, E., Dower, S. K., North, R. A., and Surprenant, A. (2001) Rapid secretion of interleukin-1beta by microvesicle shedding. *Immunity* **15,** 825–835
53. Andrei, C., Margiocco, P., Poggi, A., Lotti, L. V., Torrisi, M. R., and Rubartelli, A. (2004) Phospholipases C and A2 control lysosome-mediated IL-1 beta secretion: implications for inflammatory processes. *Proc. Natl. Acad. Sci. U.S.A.* **101,** 9745–9750
54. Mambula, S. S., Stevenson, M. A., Ogawa, K., and Calderwood, S. K. (2007) Mechanisms for Hsp70 secretion: crossing membranes without a leader. *Methods* **43,** 168–175
55. Kumkate, S., Chunchob, S., and Janvilisri, T. (2008) Expression of ATP-binding cassette multidrug transporters in the giant liver fluke *Fasciola gigantica* and their possible involvement in the transport of bile salts and anthelmintics. *Mol. Cell. Biochem.* **317,** 77–84
56. Harmsen, M. M., Cornelissen, J. B., Buijs, H. E., Boersma, W. J., Jeurissen, S. H., and van Milligen, F. J. (2004) Identification of a novel *Fasciola hepatica* cathepsin L protease containing protective epitopes within the propeptide. *Int. J. Parasitol.* **34,** 675–682
57. Beckham, S. A., Law, R. H., Smooker, P. M., Quinsey, N. S., Caffrey, C. R., McKerrow, J. H., Pike, R. N., and Spithill, T. W. (2006) Production and processing of a recombinant *Fasciola hepatica* cathepsin B-like enzyme (FhcatB1) reveals potential processing mechanisms in the parasite. *Biol. Chem.* **387,** 1053–1061
58. Coulombe, R., Grochulski, P., Sivaraman, J., Ménard, R., Mort, J. S., and Cygler, M. (1996) Structure of human procathepsin L reveals the molecular basis of inhibition by the prosegment. *EMBO J.* **15,** 5492–5503
59. Dalton, J. P., Brindley, P. J., Donnelly, S., and Robinson, M. W. (2009) The enigmatic asparaginyl endopeptidase of helminth parasites. *Trends Parasitol.* **25,** 59–61
60. Dalton, J. P., and Brindley, P. J. (1996) Schistosome asparaginyl endopeptidase Sm 32 in hemoglobin digestion. *Parasitol. Today* **12,** 125
61. Robinson, M. W., Dalton, J. P., and Donnelly, S. (2008) Helminth pathogen cathepsin proteases: it's a family affair. *Trends Biochem. Sci.* **33,** 601–608
62. Sajid, M., McKerrow, J. H., Hansell, E., Mathieu, M. A., Lucas, K. D., Hsieh, I., Greenbaum, D., Bogyo, M., Salter, J. P., Lim, K. C., Franklin, C., Kim, J. H., and Caffrey, C. R. (2003) Functional expression and characterization of *Schistosoma mansoni* cathepsin B and its trans-activation by an endogenous asparaginyl endopeptidase. *Mol. Biochem. Parasitol.* **131,** 65–75
63. Chen, J. M., Fortunato, M., and Barrett, A. J. (2000) Activation of human prolegumain by cleavage at a C-terminal asparagine residue. *Biochem. J.* **352,** 327–334
64. Morita, Y., Araki, H., Sugimoto, T., Takeuchi, K., Yamane, T., Maeda, T., Yamamoto, Y., Nishi, K., Asano, M., Shirahama-Noda, K., Nishimura, M., Uzu, T., Hara-Nishimura, I., Koya, D., Kashiwagi, A., and Ohkubo, I. (2007) Legumain/asparaginyl endopeptidase controls extracellular matrix remodeling through the degradation of fibronectin in mouse renal proximal tubular cells. *FEBS Lett.* **581,** 1417–1424
65. Behm, C. A., and Sangster, N. C. (1999) Pathology, pathophysiology and clinical aspects, in *Fasciolosis* (Dalton, J. P., ed) pp. 185–224, CAB International, Oxford
66. Geldhof, P., and Knox, D. (2008) The intestinal contortin structure in *Haemonchus contortus*: an immobilised anticoagulant? *Int. J. Parasitol.* **38,** 1579–1588
67. McGonigle, L., Mousley, A., Marks, N. J., Brennan, G. P., Dalton, J. P., Spithill, T. W., Day, T. A., and Maule, A. G. (2008) The silencing of cysteine proteases in *Fasciola hepatica* newly excysted juveniles using RNA interference reduces gut penetration. *Int. J. Parasitol.* **38,** 149–155
68. Kasny, M., Dalton, J. P., Mikes, L., and Horák, P. (2007) Comparison of cysteine and serine peptidase activities in *Trichobilharzia regenti* and *Schistosoma mansoni* cercariae. *Parasitology* **134,** 1599–1609

69. Ikeda, T. (2003) Involvement of cysteine proteinases in excystment of *Paragonimus ohirai* metacercariae induced by sodium cholate and A23187. *J. Helminthol.* **77,** 21–26
70. Chung, Y. B., Kim, T. S., and Yang, H. J. (2005) Early cysteine protease activity in excretory bladder triggers metacercaria excystment of *Paragonimus westermani*. *J. Parasitol.* **91,** 953–954
71. Na, B. K., Kim, S. H., Lee, E. G., Kim, T. S., Bae, Y. A., Kang, I., Yu, J. R., Sohn, W. M., Cho, S. Y., and Kong, Y. (2006) Critical roles for excretory-secretory cysteine proteases during tissue invasion of *Paragonimus westermani* newly excysted metacercariae. *Cell. Microbiol.* **8,** 1034–1046
72. Smooker, P. M., Whisstock, J. C., Irving, J. A., Siyaguna, S., Spithill, T. W., and Pike, R. N. (2000) A single amino acid substitution affects substrate specificity in cysteine proteinases from *Fasciola hepatica*. *Protein Sci.* **9,** 2567–2572
73. Cervi, L., Rossi, G., and Masih, D. T. (1999) Potential role for excretory-secretory forms of glutathione-S-transferase (GST) in *Fasciola hepatica*. *Parasitology* **119,** 627–633
74. Donnelly, S., O'Neill, S. M., Sekiya, M., Mulcahy, G., and Dalton, J. P. (2005) Thioredoxin peroxidase secreted by *Fasciola hepatica* induces the alternative activation of macrophages. *Infect. Immun.* **73,** 166–173
75. Donnelly, S., Stack, C. M., O'Neill, S. M., Sayed, A. A., Williams, D. L., and Dalton, J. P. (2008) Helminth 2-Cys peroxiredoxin drives Th2 responses through a mechanism involving alternatively activated macrophages. *FASEB J.* **22,** 4022–4032
76. McGonigle, S., Curley, G. P., and Dalton, J. P. (1997) Cloning of peroxiredoxin, a novel antioxidant enzyme, from the helminth parasite *Fasciola hepatica*. *Parasitology* **115,** 101–104
77. Tielens, G. M. (1999) Metabolism, in *Fasciolosis* (Dalton, J. P., ed) pp. 277–306, CAB International, Oxford
78. Collins, P. R., Stack, C. M., O'Neill, S. M., Doyle, S., Ryan, T., Brennan, G. P., Mousley, A., Stewart, M., Maule, A. G., Dalton, J. P., and Donnelly, S. (2004) Cathepsin L1, the major protease involved in liver fluke (*Fasciola hepatica*) virulence: propeptide cleavage sites and autoactivation of the zymogen secreted from gastrodermal cells. *J. Biol. Chem.* **279,** 17038–17046
79. Robinson, M. W., Trudgett, A., Hoey, E. M., and Fairweather, I. (2002) Triclabendazole-resistant *Fasciola hepatica*: β-tubulin and response to *in vitro* treatment with triclabendazole. *Parasitology* **124,** 325–338
80. Meaney, M., Allister, J., McKinstry, B., McLaughlin, K., Brennan, G. P., Forbes, A. B., and Fairweather, I. (2007) Fasciola hepatica: ultrastructural effects of a combination of triclabendazole and clorsulon against mature fluke. *Parasitol. Res.* **100,** 1091–1104
81. Meaney, M., Haughey, S., Brennan, G. P., and Fairweather, I. (2005) Ultrastructural observations on oral ingestion and trans-tegumental uptake of clorsulon by the liver fluke, *Fasciola hepatica*. *Parasitol. Res.* **95,** 201–212
82. McManus, D. P., and Dalton, J. P. (2006) Vaccines against the zoonotic trematodes *Schistosoma japonicum*, *Fasciola hepatica* and *Fasciola gigantica*. *Parasitology* **133,** S43–61
83. Alcala-Canto, Y., Ibarra-Velarde, F., Sumano-Lopez, H., Gracia-Mora, J., and Alberti-Navarro, A. (2007) Effect of a cysteine protease inhibitor on *Fasciola hepatica* (liver fluke) fecundity, egg viability, parasite burden, and size in experimentally infected sheep. *Parasitol. Res.* **100,** 461–465

Supplementary Table 1: Secreted proteins predicted from *Fasciola hepatica* ESTs and their homologues, functionality and RNAi phenotypes.

Number	EST sequence ID	Residue (aa)	Start	SP	Description (top NR hit)	E-value	Interproscan ID	InterPro Hit Description	Wormpep homologue (top hit)	*C. elegans* RNAi phenotype
1	FhContig155	115	M	20	SJCHGC09803 protein *[Schistosoma japonicum]*	6.76E-14	_	_	CAB04840.1 from WBGene00012057:chitinase status:Predicted TR:P92013	None
2	FhContig179	283	_	26	cubilin *[Canis familiaris]*	2.32E-06	IPR000859	CUB	CAA98557.1 from WBGene00013855:bone morphogenetic protein 1 like status:Partially_confirmed SW:Q20911	None
3	FhContig182	116	M	19	No significant BLAST hits	_	_	_	CAI46575.1 from WBGene00008767: Yeast mitochondrial carrier protein like status:Partially_confirmed TR:Q5FC59	life span abnormal(age)
4	FhContig245	189	M	15	apoferritin-2 *[Schistosoma japonicum]*	1.76E-51	IPR008331, IPR001519, IPR009078, IPR009040	Ferritin and Dps, Ferritin N-terminal, Ferritin/ribonucleotide reductase-like, Ferritin-like	AAK21364.1 from WBGene00001501: ftn-2 ferritin status:Confirmed TR:Q9TYS3	bar-1(ga80),C.elegans wild type, DR subclone of CB original (Tc1 pattern I)
5	FhContig302	124	M	21	No significant BLAST hits	_	_	_	CAB03153.2 from WBGene00010280: ATP-dependent RNA helicase like status:Partially_confirmed TR:P90897	None
6	FhContig305	84	M	20	SJCHGC09717 protein *[Schistosoma japonicum]*	3.85E-17	_	_	CAA90674.1from WBGene00006175 : str-124 status:Partially_confirmed TR:O18114	None
7	FhContig386	94	M	30	No significant BLAST hits	_	_	_	AAA97966.1 from WBGene00022603 : gck-2 serine\/threonine kinase status:Partially_confirmed TR:Q23290	None
8	FhContig490	327	_	23	conserved hypothetical protein *[Aedes aegypti]*	2.00E-63	IPR012411, IPR001736, IPR013582	Phospholipase D, chordopox, Phospholipase D/Transphosphatidylase, Phospholipase D/viral envelope	AAB71325.1 from WBGene00017316 : Partially_confirmed SW:O17405	None
9	FhContig492	278	_	23	SJCHGC01895 protein *[Schistosoma japonicum]*	3.61E-97	IPR001254, IPR001314, IPR009003	Peptidase S1 and S6, chymotrypsin/Hap, Peptidase S1A, chymotrypsin, Peptidase, trypsin-like serine and cysteine	AAA68746.2 from WBGene00006619: try-1 Plasminogen status:Confirmed TR:Q23528	embryonic lethal (Emb)
10	FhContig528	149	M	21	No significant BLAST hits	_	_	_	_	_
11	FhContig535	73	M	17	No significant BLAST hits	_	_	_	_	_
12	FhContig542	266	M	20	SJCHGC06223 protein *[Schistosoma japonicum]*	7.06E-63	IPR001563	Peptidase S10, serine carboxypeptidase	AAA68259.1 from WBGene00019605: Carboxypeptidase status:Partially_confirmed SW:Q09991	None
13	FhContig578	102	M	15	secreted saposin-like protein SAP-3 *[Fasciola gigantica]*	1.32E-38	IPR008139, IPR011001, IPR007856, IPR008138	Saposin B, Saposin-like, Saposin-like type B1, Saposin-like type B2	AAB37550.2 from WBGene00004995:locus:spp-10 7TM chemoreceptor, spp family status:Confirmed TR:Q4JFH6	None
14	FhContig595	102	M	15	secreted saposin-like protein SAP-3 *[Fasciola gigantica]*	1.32E-38	IPR008139, IPR011001, IPR007856, IPR008138	Saposin B, Saposin-like, Saposin-like type B1, Saposin-like type B2	AAB37550.2 from WBGene00004995:locus:spp-10 7TM chemoreceptor, spp family status:Confirmed TR:Q4JFH6	None

15	FhContig748	185	_	19	unnamed protein product *[Kluyveromyces lactis]*	3.27E-07	_	_	AAB37793.2 from WBGene00017815 : Partially_confirmed TR:P91279	cord commissures fail to reach target, Axon fasciculation abnormal,ventral cord patterning abnormal
16	FhContig788	203	_	15	peptidoglycan recognition protein [*Argopecten irradians]*	1.62E-35	IPR006619, IPR002502, IPR015510	Animal peptidoglycan recognition protein PGRP, N-acetylmuramoyl-L-alanine amidase family 2, peptidoglycan recognition proteins	CAA88310.1 from WBGene00008640: Zinc finger, C2H2 type status:Partially_confirmed TR:Q09534	sterile progeny (Stp), Pattern of transgene expression abnormal receptor mediated endocytosis defective,reduced brood size
17	FhContig1152	391	_	16	vitelline protein BI *[Fasciola gigantica]*	2.49E-21	IPR012615	Trematode Eggshell Synthesis	CAB03153.2 from WBGene00010280: ATP-dependent RNA helicase like status:Partially_confirmed TR:P90897	None
18	FhContig1258	161	_	36	No significant BLAST hits	_	_	_	CAE17836.2 from WBGene00009789 : status:Partially_confirmed TR:Q7YX10	_
19	FhContig1272	187	M	19	phosphatidylinositol glycan *[Bombyx mori]*	1.18E-68	IPR001096	Peptidase C13, legumain	CAA92977.2 from WBGene00011482 : haemoglobinase like status:Partially_confirmed SW:P49048	None
20	FhContig1352	310	_	22	unnamed protein product *[Tetraodon nigroviridis]*	6.19E-27	IPR008185	Deoxyribonuclease I	AAA81093.1 from WBGene00006522: abcx-1 status:Predicted TR:Q18901	embryonic lethal (Emb),larval arrest (Lva)
21	FhContig1378	222	_	21	cathepsin L2 *[Fasciola gigantica]*	1.68E-103	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb),slow growth (Gro),locomotion abnormal (Unc)
22	FhContig1420	64	M	22	No significant BLAST hits	_	_	_	CAB03918.1 from WBGene00007723: Partially_confirmed TR:O62074	None
23	FhContig1466	252	M	21	SJCHGC00967 protein *[Schistosoma japonicum]*	5.47E-38	IPR004947	Deoxyribonuclease II	CAA86412.1 from WBGene00003828: locus:nuc-1 status:Confirmed SW:Q17778	life span abnormal(age)
24	FhContig1553	363	M	20	SJCHGC06231 protein *[Schistosoma japonicum]*	2.94E-90	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb),slow growth (Gro),embryonic lethal (Emb),locomotion abnormal (Unc)
25	FhContig1563	90	M	20	unknown *[Clonorchis sinensis]*	2.53E-13	_	_	CAB61063.1 from WBGene00012697: Partially_confirmed TR:Q9NEI6	sterile(Ste),larval lethal(Lvl),larval arrest(Lva),embryonic lethal (Emb),maternal sterile(Ste)
26	FhContig1615	94	M	18	No significant BLAST hits	_	_	_	ABB88221.1 from WBGene00006600P:tph-1 Biopterin-dependent aromatic amino acid hydroxylase status:Confirmed TR:Q9XZD1	embryonic lethal(Emb)
27	FhContig1624	184	_	27	myoglobin 2 *[Paragonimus westermani]*	1.62E-38	IPR000971, IPR009050	Globin, Globin-like	R13F6.4d CE38629 WBGene00006498 locus:ten-1 status:Partially_confirmed	exploded through vulva(Rup),undeveloped,gonad development abnormal,constipated(Cons),gonad migration abnormal(Gom),embryonic lethal(Emb),reduced brood size,organism morphology abnormal,cell migration abnormal(Mig),slow growth(Gro)

28	FhContig1639	258	_	16	cathepsin B endopeptidase *[Schistosoma mansoni]*	2.01E-43	IPR013128, IPR000668, IPR012599, IPR000169	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase C1A, propeptide, Peptidase, cysteine peptidase active site	AAM51519.1 from WBGene00000786 : cpr-6 status:Confirmed TR:Q8MQC6	None
29	FhContig1733	188	_	28	No significant BLAST hits	_	_	_	CAB01662.1 from WBGene00011324: status:Confirmed TR:Q22058	None
30	FhContig1740	436	_	32	legumain *[Opisthorchis viverrini]*	4.81E-93	IPR001096	Peptidase C13, legumain	CAA99935.1from WBGene00012144:vacuolar processing enzyme like status:Confirmed TR:Q17945	embryonic lethal(Emb)
31	FhContig1766	84	M	24	SJCHGC09800 protein *[Schistosoma japonicum]*	1.73E-17	_	_	CAB04840.1 from WBGene00012057: chitinase status:Predicted TR:P92013	None
32	FhContig1795	215	_	21	PREDICTED: similar to gag-pol polyprotein *[Strongylocentrotus purpuratus]*	1.41E-27	_	_	AAB71008.3 from WBGene00019615 status:Partially_confirmed SW:O17237	extended life span (Age), Aldicarb resistant (Ric)
33	FhContig1886	155	_	24	secreted cathepsin L 1 *[Fasciola hepatica]*	1.01E-17	IPR013128, IPR000668	Peptidase C1A, papain, Peptidase C1A, papain C-terminal	AAB65956.2 from WBGene00007055:tag-196 cysteine protease and a protease inhibitor status:Partially_confirmed TR:O16454	None
34	FhContig1935	444	_	35	protein disulphide isomerase *[Fasciola hepatica]*	0	IPR005788, IPR000886, IPR005792, IPR013766, IPR012336, IPR006662	Disulphide isomerase, Endoplasmic reticulum targeting sequence, Protein disulphide isomerase, Thioredoxin domain, Thioredoxin-like fold, Thioredoxin-related	AAK39152.1 from WBGene00003963 locus:pdi-2 protein disulfide isomerase status:Confirmed SW:Q17770	slow growth(Gro),larval lethal,larval arrest(Lva),organism morphology abnormal, Dumpy(Dpy),embryonic lethal late emb,embryonic lethal(Emb),locomotion abnormal,sluggish(Slu),clear(Clr)
35	FhContig1954	244	_	19	No significant BLAST hits	_	_	_	_	_
36	FhContig1962	280	_	21	No significant BLAST hits	_	_	_	CAA21734.1 from WBGene00001045: dnj-27 DnaJ domain, Thioredoxin status:Partially_confirmed TR:Q9XWE1	None
37	FhContig2035	66	_	25	No significant BLAST hits	_	_	_	AAL06044.4 from WBGene00019832:Partially_confirmed SW:Q21653	reduced brood size
38	FhContig2164	260	_	15	cathepsin B *[Fasciola gigantica]*	1.96E-62	IPR013128, IPR000668, IPR012599, IPR000169	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site	C25B8.3c CE41106 WBGene00000786 locus:cpr-6 status:Confirmed	None
39	FhContig2169	326	M	15	Cathepsin L-like proteinase precursor *[Fasciola hepatica]*	0	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb),slow growth (Gro),locomotion abnormal (Unc)
40	FhContig2213	355	_	18	cathepsin L protein *[Fasciola hepatica]*	3.33E-147	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb),slow growth (Gro),locomotion abnormal (Unc)
41	FhContig2221	76	_	24	cathepsin L protein *[Fasciola hepatica]*	5.34E-11	_	_	CAA82576.1 from WBGene00011560 status:Confirmed TR:Q22290	None

42	FhContig2292	263	_	17	asparaginyl endopeptidase *[Schistosoma mansoni]*	4.64E-51	IPR001096, IPR001005	Peptidase C13, legumain,SANT, DNA-binding	CAA99935.1 from WBGene00012144 vacuolar processing enzyme like status:Confirmed TR:Q17945	embryonic lethal (Emb)
43	FhContig2337	97	_	26	Cathepsin L-like proteinase precursor *[Fasciola hepatica]*	3.55E-39	IPR013128, IPR013201	Peptidase C1A, papain, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb),slow growth (Gro),locomotion abnormal (Unc)
44	FhContig2374	180	_	36	No significant BLAST hits	_	_	_	_	_
45	FhContig2453	227	_	27	SJCHGC01895 protein *[Schistosoma japonicum]*	9.55E-41	IPR009003	Peptidase, trypsin-like serine and cysteine	CAA90302.2 from WBGene00000082 locus:adt-1 TSP type-1 repeats (13) status:Confirmed TR:Q8MYA8	None
46	FhContig2518	80	_	31	No significant BLAST hits	_	_	_	AAA83359.1 from WBGene00017989 locus:nol-10 status:Partially_confirmed TR:Q19974	slow growth (Gro),early larval arrest,embryonic lethal (Emb),egg laying abnormal (Egl),sterile progeny (Stp),larval arrest (Lva),maternal sterile (Ste)
47	FhContig2597	175	M	18	No significant BLAST hits	_	_	_	_	_
48	FhContig2623	342	_	31	secreted cathepsin L1 *[Fasciola hepatica]*	3.79E-177	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb),slow growth (Gro),locomotion abnormal (Unc)
49	FhContig2626	244	_	17	cathepsin L *[Fasciola gigantica]*	1.84E-112	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb),slow growth (Gro),locomotion abnormal (Unc)
50	FhContig2637	326	M	15	cathepsin L protein *[Fasciola hepatica]*	4.14E-178	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb),slow growth (Gro),locomotion abnormal (Unc)
51	FhContig2665	298	_	30	No significant BLAST hits	_	_	_	CAD27608.1 from WBGene00011629 nucleotide binding protein status:Partially_confirmed TR:Q8T3D2	None
52	FhContig2681	281	_	18	No significant BLAST hits	_	_	_	_	_
53	FhContig2687	466	M	16	No significant BLAST hits	_	_	_	CAB04815.1 from WBGene00006093 locus:str-27 7TM receptor status:Predicted TR:O45803	None
54	FhContig2690	169	_	24	No significant BLAST hits	_	_	_	AAK21364.1 from WBGene00001501: ftn-2 ferritin status:Confirmed TR:Q9TYS3	slow growth (Gro), Drug response abnormal
55	FhContig2710	274	_	19	cathepsin L1 proteinase *[Fasciola hepatica]*	1.48E-127	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	AAB65956.2 from WBGene00007055:tag-196 cysteine protease and a protease inhibitor status:Partially_confirmed TR:O16454	None
56	FhContig2720	333	_	22	cathepsin L *[Fasciola gigantica]*	6.88E-160	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
57	FhContig2721	67	M	19	No significant BLAST hits	_	_	_	_	_

58	FhContig2730	354	_	32	cathepsin B1 isotype 1 *[Schistosoma mansoni]*	3.13E-113	IPR002016, IPR013128, IPR000668, IPR012599, IPR000169	Haem peroxidase, plant/fungal/bacterial, Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase C1A, propeptide, Peptidase, cysteine peptidase active site	AAM51519.1 from WBGene00000786 : cpr-6 status:Confirmed TR:Q8MQC6	None
59	FhContig2734	222	_	20	asparaginyl endopeptidase *[Schistosoma mansoni]*	2.45E-70	IPR001096	Peptidase C13, legumain	CAA99935.1 from WBGene00012144 vacuolar processing enzyme like status:Confirmed TR:Q17945	embryonic lethal (Emb)
60	HAN4008b10.p1kT7	112	M	15	secreted saposin-like protein SAP-3 *[Fasciola gigantica]*	1.32E-38	IPR008139, IPR011001, IPR007856, IPR008138	Saposin B,Saposin-like,Saposin-like type B, 1,Saposin-like type B, 2	AAB37549.2 from WBGene00004995 locus:spp-10 7TM chemoreceptor, spp family status:Confirmed TR:Q18276	None
61	HAN4011a03.p1kT7	90	M	20	unknown *[Clonorchis sinensis]*	5.64E-13	_	_	CAB61063.1 from WBGene00012697: Partially_confirmed TR:Q9NEI6	sterile,larval lethal,larval arrest,embryonic lethal,maternal sterile
62	HAN4009b02.p1kT7	114	M	15	amoebapore-like protein *[Fasciola hepatica]*	2.39E-51	IPR008139, IPR011001, IPR007856, IPR008138	Saposin B,Saposin-like,Saposin-like type B, 1,Saposin-like type B, 2	AAM69087.1 from WBGene00004988 locus:spp-3 status:Confirmed TR:Q22336	None
63	HAN5015a11.q1kT3	83	M	20	unknown *[Clonorchis sinensis]*	3.99E-11	_	_	CAB61063.1 from WBGene00012697: Partially_confirmed TR:Q9NEI6	sterile (Ste),larval lethal (Lvl),larval arrest (Lva),embryonic lethal (Emb),maternal sterile (Ste)
64	HAN5012c07.q1kT3	224	_	25	amoebapore-like protein *[Fasciola hepatica]*	1.53E-43	IPR009050, IPR011001, IPR007856	Globin-like,Saposin-like,Saposin-like type B, 1	CAA16406.2 from WBGene00013075 status:Partially_confirmed TR:Q7K7P3	embryonic lethal (Emb),slow growth (Gro),locomotion abnormal (Unc),maternal sterile (Ste),thin sterile (Ste), Protruding vulva (Pvl),larval arrest (Lva)
65	HAN5008c12.q1kT3	110	M	15	secreted saposin-like protein SAP-3 *[Fasciola gigantica]*	1.33E-38	IPR008139, IPR011001, IPR007856, IPR008138	Saposin B,Saposin-like,Saposin-like type B, 1,Saposin-like type B, 2	AAB37550.2 from WBGene00004995:locus:spp-10 7TM chemoreceptor, spp family status:Confirmed TR:Q4JFH6	None
66	Fhep20e09.q1k	79	_	22	No significant BLAST hits	_	_	_	AAM69068.1 from WBGene00020732 status:Partially_confirmed TR:Q8MXJ1	None
67	Fhep21c03.q1k	127	M	21	No significant BLAST hits	_	_	_	CAB60424.2 from WBGene00012995 status:Partially_confirmed TR:Q9U286	None
68	Fhep12g01.q1k	60	_	19	No significant BLAST hits	_	_	_	CAB04651.3 from WBGene00011226 mitochondrial transporter protein status:Confirmed TR:O62347	None
69	Fhep13g05.q1k	136	M	18	No significant BLAST hits	_	_	_	_	_
70	Fhep21f02.q1k	360	_	26	legumain *[Opisthorchis viverrini]*	1.07E-68	IPR001096	Peptidase C13, legumain	CAA99935.1 from WBGene00012144 vacuolar processing enzyme like status:Confirmed TR:Q17945	embryonic lethal(Emb)
71	Fhep44e10.q1k	241	_	16	pro-cathepsin B2 *[Fasciola hepatica]*	1.94E-66	IPR013128, IPR000668, IPR012599, IPR000169	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase C1A, propeptide, Peptidase, cysteine peptidase active site	C25B8.3c CE41106 WBGene00000786 locus:cpr-6 status:Confirmed	None
72	Fhep05h03.q1k	73	M	17	No significant BLAST hits	_	_	_	_	_
73	Fhep11b02.q1k	211	_	18	cathepsin B1 isotype 1 *[Schistosoma mansoni]*	8.19E-25	IPR002016, IPR013128, IPR000668, IPR012599, IPR000169	Haem peroxidase, plant/fungal/bacterial, Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase C1A, propeptide, Peptidase, cysteine peptidase active site	AAA92327.1 from WBGene00000784 locus:cpr-4 cathepsin B-like cysteine proteinase 4 precursor status:Confirmed SW:P43508	None

74	Fhep25d07.q1k	399	M	16	No significant BLAST hits	_	_	_	_	_
75	Fhep19e04.q1k	295	_	20	secreted cathepsin L 1 [Fasciola hepatica]	3.76E-103	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
76	Fhep50d02.q1k	140	_	21	secreted cathepsin L 1 [Fasciola hepatica]	3.31E-61	IPR013128, IPR000668, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
77	Fhep20e12.q1k	141	_	21	cathepsin *[Fasciola gigantica]*	5.43E-64	IPR013128, IPR013201	Peptidase C1A, papain, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
78	Fhep55h04.q1k	213	_	24	cubilin *[Rattus norvegicus]*	0.000478271	IPR000859	CUB	CAB03030.2 from WBGene00009271 glutaminase status:Partially_confirmed SW:Q93650	None
79	Fhep05g06.q1k	228	_	25	protein disulphide isomerase *[Fasciola hepatica]*	7.21E-121	IPR013766, IPR012336, IPR006662	Thioredoxin domain,Thioredoxin-like fold,Thioredoxin-related	AAK39152.1 from WBGene00003963 locus:pdi-2 protein disulfide isomerase status:Confirmed SW:Q17770	slow growth (Gro),larval arrest (Lva),organism morphology abnormal (Bmd),locomotion abnormal (Unc),sluggish,larval lethal, Dumpy,embryonic lethal late emb (Emb),Clear(Clr)
80	Fhep05h02.q1k	121	_	24	cystatin *[Fasciola hepatica]*	2.35E-11	IPR000010	Proteinase inhibitor I25, cystatin	AAK31456.1 from WBGene00016120 status:Partially_confirmed TR:Q9BIA3	slow growth (Gro)
81	Fhep13h02.q1k	226	_	25	Cathepsin L-like proteinase precursor *[Fasciola hepatica]*	1.34E-111	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
82	Fhep14g11.q1k	265	_	32	No significant BLAST hits	_	_	_	AAC48151.1 from WBGene00000116 locus:alh-10 aldehyde dehydrogenase status:Partially_confirmed TR:Q18822	None
83	Fhep21a05.q1k	254	_	22	secreted cathepsin L1 *[Fasciola hepatica]*	4.63E-93	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
84	Fhep21h09.q1k	160	_	23	cathepsin L protein *[Fasciola hepatica]*	2.16E-39	IPR013128, IPR013201	Peptidase C1A, papain, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
85	Fhep23b08.q1k	78	_	27	No significant BLAST hits	_	_	_	CAA96680.3 from WBGene00011684 glycogenin like status:Predicted TR:Q22375	None
86	Fhep06c07.q1k	126	_	18	secreted saposin-like protein SAP-3 *[Fasciola gigantica]*	4.62E-31	IPR008139, IPR011001, IPR007856	Saposin B,Saposin-like, Saposin-like type B, 1	AAB37550.2 from WBGene00004995:locus:spp-10 7TM chemoreceptor, spp family status:Confirmed TR:Q4JFH6	None
87	Fhep06d06.q1k	295	_	20	vitelline protein B1 *[Fasciola hepatica]*	1.02E-116	IPR012615	Trematode Eggshell Synthesis	AAC26317.2 from WBGene00020360 locus:math-41 status:Confirmed TR:O76640	None
88	Fhep06d12.q1k	316	_	24	amoebapore-like protein *[Fasciola hepatica]*	7.52E-44	IPR011001, IPR007856	Saposin-like,Saposin-like type B, 1	AAM69087.1 from WBGene00004988 locus:spp-3 status:Confirmed TR:Q22336	None

89	Fhep06g03.q1k	283	_	19	cathepsin L-like protease *[Fasciola hepatica]*	1.01E-118	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
90	Fhep07a05.q1k	370	_	23	secreted cathepsin L2 *[Fasciola hepatica]*	2.61E-142	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
91	Fhep07f12.q1k	227	_	20	Cathepsin L-like proteinase precursor *[Fasciola hepatica]*	4.96E-114	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
92	Fhep08a08.q1k	299	_	30	cathepsin *[Fasciola gigantica]*	9.25E-150	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
93	Fhep08c01.q1k	309	_	16	Cathepsin L-like proteinase precursor *[Fasciola hepatica]*	6.59E-138	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
94	Fhep08c02.q1k	314	_	36	cathepsin B1 isotype 1 *[Schistosoma mansoni]*	2.53E-76	IPR013128, IPR000668, IPR012599, IPR000169	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase C1A, propeptide, Peptidase, cysteine peptidase active site	AAM51519.1 from WBGene00000786 : cpr-6 status:Confirmed TR:Q8MQC6	None
95	Fhep08h11.q1k	218	_	19	secreted cathepsin L1 *[Fasciola hepatica]*	3.31E-88	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
96	Fhep10c09.q1k	173	_	24	No significant BLAST hits	_	_	_	AAN60532.1 from WBGene00016006 locus:flna-1 status:Partially_confirmed TR:Q8IG48	molt defect (Mlt),egg laying defective(Egl_D),exploded through vulva(Rup),locomotion abnormal (Unc),slow growth(Gro),
97	Fhep10h01.q1k	199	_	25	unknown *[Schistosoma japonicum]*	3.34E-06	IPR008139, IPR011001, IPR008138	Saposin B,Saposin-like,Saposin-like type B, 2	AAX88825.1 from WBGene00004999 locus:spp-14 status:Confirmed TR:Q2XN18	None
98	Fhep12d04.q1k	289	_	18	cathepsin L2 *[Fasciola gigantica]*	1.08E-131	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
99	Fhep12h11.q1k	408	_	25	No significant BLAST hits	_	_	_	CAB76739.1 from WBGene00013762 status:Partially_confirmed TR:Q9NAL5	None

100	Fhep15b08.q1k	258	_	33	cathepsin L-like protease *[Fasciola hepatica]*	9.87E-83	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
101	Fhep15b10.q1k	290	_	19	cathepsin L-like protease *[Fasciola hepatica]*	6.65E-97	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
102	Fhep15c08.q1k	153	_	17	cathepsin L-like protease *[Fasciola hepatica]*	7.14E-85	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
103	Fhep15d08.q1k	252	_	17	cathepsin L-like protease *[Fasciola hepatica]*	1.09E-94	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
104	Fhep17a08.q1k	140	_	26	cystatin [Fasciola hepatica]	3.59E-07	_	_	AAK31456.1 from WBGene00016120 status:Partially_confirmed TR:Q9BIA3	slow growth (Gro)
105	Fhep18b10.q1k	149	_	17	cathepsin L-like protease *[Fasciola hepatica]*	3.08E-59	IPR013128, IPR000668, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
106	Fhep18d10.q1k	402	_	17	cathepsin L-like protease *[Fasciola hepatica]*	2.60E-106	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
107	Fhep18e10.q1k	277	_	34	vitelline protein BI [Fasciola gigantica]	1.19E-92	IPR012615	Trematode Eggshell Synthesis	CAB05785.1 from WBGene00006856 locus:usp-14 Queuine trna-ribosyltransferase status:Confirmed SW:Q17361	None
108	Fhep18f12.q1k	237	_	21	cathepsin L-like protease *[Fasciola hepatica]*	3.14E-130	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
109	Fhep20a06.q1k	154	_	15	cathepsin L1 proteinase *[Fasciola hepatica]*	4.19E-69	IPR013128, IPR000668, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
110	Fhep20c05.q1k	153	_	22	cathepsin L protein *[Fasciola hepatica]*	6.48E-78	IPR013128, IPR000668, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
111	Fhep20d03.q1k	102	_	15	secreted saposin-like protein SAP-3 [Fasciola gigantica]	5.03E-38	IPR008139, IPR011001, IPR007856, IPR008138	Saposin B,Saposin-like,Saposin-like type B, 1,Saposin-like type B, 2	AAB37550.2 from WBGene00004995:locus:spp-10 7TM chemoreceptor, spp family status:Confirmed TR:Q4JFH6	None

112	Fhep20h10.q1k	101	_	14	secreted saposin-like protein SAP-3 [Fasciola gigantica]	2.51E-37	IPR008139, IPR011001, IPR007856, IPR008138	Saposin B,Saposin-like,Saposin-like type B, 1,Saposin-like type B, 2	AAB37550.2 from WBGene00004995:locus:spp-10 7TM chemoreceptor, spp family status:Confirmed TR:Q4JFH6	None
113	Fhep25b01.q1k	306	_	24	cathepsin L protein *[Fasciola hepatica]*	5.87E-131	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
114	Fhep25h05.q1k	270	_	16	vitelline protein B1 *[Fasciola hepatica]*	6.07E-110	IPR012615	Trematode Eggshell Synthesis	CAA93499.1 from WBGene00000695 locus:col-121 collagen status:Partially_confirmed TR:Q20880	organism stress response abnormal, Dauer formation abnormal (Daf)
115	Fhep29e10.q1k_1	161	_	19	peptidylprolyl isomerase B (cyclophilin B) *[Xenopus tropicalis]*	1.65E-47	IPR002130	Peptidyl-prolyl cis-trans isomerase, cyclophilin-type	AAA91355.1 from WBGene00000882 locus:cyn-6 status:Confirmed SW:P52014	None
116	Fhep30a01.q1k	283	_	35	cathepsin L-like protease *[Fasciola hepatica]*	1.33E-94	IPR013128, IPR000668, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
117	Fhep30e02.q1k	314	_	32	cathepsin L-like protease *[Fasciola hepatica]*	5.22E-114	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
118	Fhep31b09.q1k	249	_	19	cathepsin L-like protease *[Fasciola hepatica]*	4.43E-109	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
119	Fhep31b10.q1k	306	_	37	vitelline protein B1 *[Fasciola hepatica]*	2.11E-88	IPR012615	Trematode Eggshell Synthesis	CAB05785.1 from WBGene00006856 locus:usp-14 Queuine trna-ribosyltransferase status:Confirmed SW:Q17361	None
120	Fhep31c09.q1k	126	_	32	No significant BLAST hits	_	_	_	_	_
121	Fhep31c11.q1k	269	_	22	cathepsin L-like protease *[Fasciola hepatica]*	2.28E-117	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
122	Fhep31d03.q1k	179	_	24	No significant BLAST hits	_	_	_	AAN60532.1 from WBGene00016006 locus:flna-1 status:Partially_confirmed TR:Q8IG48	molt defect (Mlt),egg laying defective(Egl_D),exploded through vulva(Rup),locomotion abnormal (Unc),slow growth(Gro),
123	Fhep31g11.q1k	335	_	21	cathepsin L-like protease *[Fasciola hepatica]*	1.17E-98	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
124	Fhep31h08.q1k	193	_	20	cathepsin L-like protease *[Fasciola hepatica]*	1.24E-103	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)

125	Fhep32c04.q1k	240	_	23	Cathepsin L-like proteinase precursor *[Fasciola hepatica]*	1.45E-114	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
126	Fhep32c08.q1k	212	_	26	No significant BLAST hits	_	_	_	_	_
127	Fhep32d02.q1k	300	M	15	cathepsin L2 *[Fasciola gigantica]*	1.57E-80	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
128	Fhep32g02.q1k	187	_	17	Cathepsin L-like proteinase precursor *[Fasciola hepatica]*	1.33E-104	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
129	Fhep33e01.q1k	197	_	19	SJCHGC00969 protein *[Schistosoma japonicum]*	7.48E-11	IPR008139, IPR011001, IPR008138	Saposin B,Saposin-like,Saposin-like type B, 2	AAX88825.1 from WBGene00004999 locus:spp-14 status:Confirmed TR:Q2XN18	None
130	Fhep34a04.q1k	161	M	20	cystatin *[Fasciola hepatica]*	1.54E-53	IPR000010	Proteinase inhibitor I25, cystatin	CAA21703.2 from WBGene00013224 status:Confirmed TR:Q95Q15	None
131	Fhep36b02.q1k	128	_	31	secreted saposin-like protein SAP-3 *[Fasciola gigantica]*	1.33E-38	IPR008139, IPR011001, IPR007856, IPR008138	Saposin B,Saposin-like,Saposin-like type B, 1,Saposin-like type B, 2	AAB37549.2 from WBGene00004995 locus:spp-10 7TM chemoreceptor, spp family status:Confirmed TR:Q18276	None
132	Fhep36b03.q1k	223	_	20	cathepsin L *[Fasciola hepatica]*	2.99E-108	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
133	Fhep36e03.q1k	244	_	29	Cathepsin L-like proteinase precursor *[Fasciola hepatica]*	2.34E-91	IPR013128, IPR000668, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
134	Fhep37a08.q1k	227	_	13	secreted cathepsin L2 *[Fasciola hepatica]*	2.32E-87	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
135	Fhep37d08.q1k	214	_	21	cathepsin L protein *[Fasciola hepatica]*	3.98E-99	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
136	Fhep37e07.q1k	299	_	21	vitelline protein B1 *[Fasciola hepatica]*	4.36E-99	IPR012615	Trematode Eggshell Synthesis	_	_
137	Fhep37g04.q1k	73	M	17	No significant BLAST hits	_	_	_	_	_
138	Fhep37h05.q1k	209	_	17	cathepsin L *[Fasciola hepatica]*	1.53E-100	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb),slow growth (Gro),embryonic lethal (Emb),locomotion abnormal (Unc)

139	Fhep38a09.q1k_1	194	_	19	No significant BLAST hits	_	_	_	_	_
140	Fhep39b03.q1k	291	_	19	cathepsin L-like protease *[Fasciola hepatica]*	5.67E-96	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
141	Fhep39e10.q1k	266	_	20	secreted cathepsin L1 *[Fasciola hepatica]*	5.56E-84	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
142	Fhep39f01.q1k	172	_	21	cathepsin L protein *[Fasciola hepatica]*	4.66E-84	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
143	Fhep40a09.q1k	288	_	30	vitelline protein B1 *[Fasciola hepatica]*	1.12E-96	IPR012615	Trematode Eggshell Synthesis	AAB70342.2 from WBGene00001596 locus:gld-2 status:Confirmed TR:O17087	reduced brood size,exploded through vulva (Rup),sterile (Ste), Protruding vulva (Pvl),embryonic lethal (Emb),maternal sterile (Ste),slow growth(Gro), Pattern of transgene expression abnormal receptor mediated endocytosis defective,tumorous germline (Tum), Passage through meiosis abnormal early emb (Emb),embryo osmotic integrity abnormal early emb (Emb)
144	Fhep40e01.q1k	234	_	22	Cathepsin L-like proteinase *[Fasciola hepatica]*	7.98E-86	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
145	Fhep40h11.q1k	314	_	21	cathepsin L1 proteinase *[Fasciola hepatica]*	1.16E-137	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
146	Fhep42d08.q1k	116	_	21	cathepsin *[Fasciola gigantica]*	4.76E-60	IPR013128, IPR013201	Peptidase C1A, papain, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
147	Fhep43f03.q1k	209	_	22	Cathepsin L-like protease *[Fasciola hepatica]*	3.62E-110	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
148	Fhep45e01.q1k	213	_	21	vitelline protein B1 *[Fasciola hepatica]*	1.19E-71	IPR012615	Trematode Eggshell Synthesis	CAB05785.1 from WBGene00006856 locus:usp-14 Queuine trna-ribosyltransferase status:Confirmed SW:Q17361	None
149	Fhep48b11.q1k	316	_	30	Cathepsin L-like proteinase precursor *[Fasciola hepatica]*	1.17E-121	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)

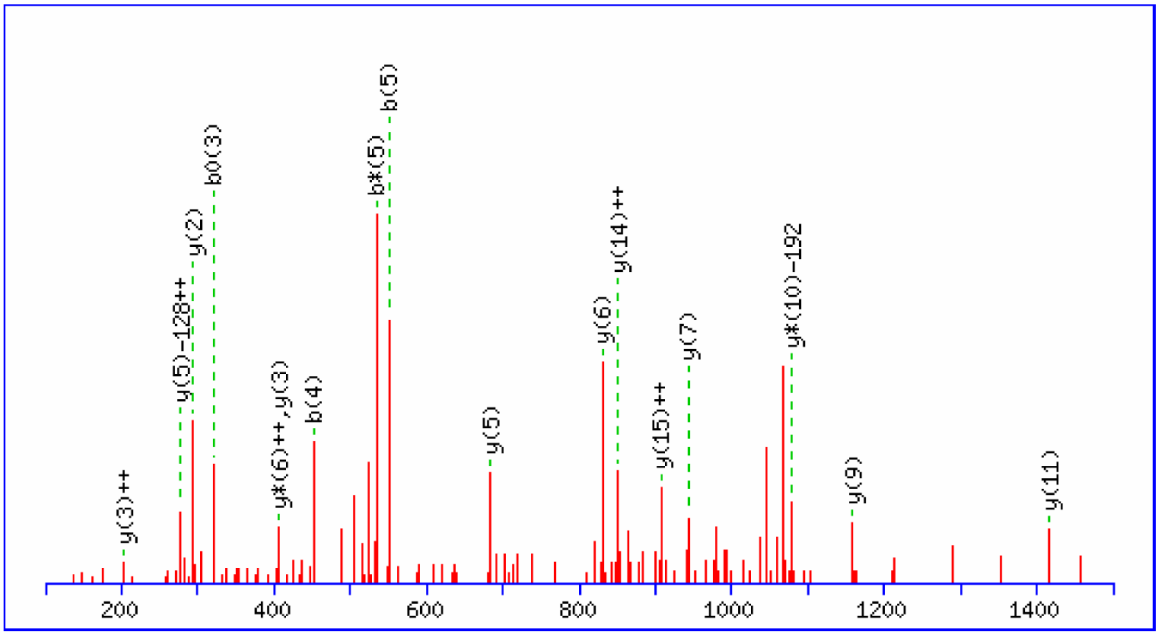
150	Fhep49h03.q1k	296	M	15	cathepsin L2 *[Fasciola gigantica]*	2.82E-82	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
151	Fhep50a05.q1k	361	_	21	cathepsin L-like protease *[Fasciola hepatica]*	1.29E-138	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
152	Fhep50f12.q1k	279	_	29	vitelline protein B1 *[Fasciola hepatica]*	1.42E-117	IPR012615	Trematode Eggshell Synthesis	AAC26317.2 from WBGene00020360 locus:math-41 status:Confirmed TR:O76640	None
153	Fhep51b06.q1k	341	_	31	cathepsin L protein *[Fasciola hepatica]*	1.77E-134	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	_
154	Fhep51e06.q1k_1	189	M	29	No significant BLAST hits	_	_	_	AAM48557.1 from WBGene00021448 status:Partially_confirmed TR:Q8MXE3	None
155	Fhep51h01.q1k	239	_	18	Cathepsin L-like proteinase precursor *[Fasciola hepatica]*	2.37E-133	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
156	Fhep51h12.q1k	83	_	24	Cathepsin L-like proteinase precursor *[Fasciola hepatica]*	7.22E-37	IPR013128, IPR013201	Peptidase C1A, papain, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
157	Fhep53e12.q1k	125	_	21	cathepsin *[Fasciola gigantica]*	1.87E-56	IPR013128, IPR013201	Peptidase C1A, papain, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
158	Fhep55b05.q1k	257	_	17	cathepsin L *[Fasciola gigantica]*	8.79E-108	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb),slow growth (Gro),embryonic lethal (Emb),locomotion abnormal (Unc)
159	Fhep55b12.q1k	268	_	17	cathepsin L *[Fasciola gigantica]*	1.42E-111	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)
160	Fhep55g04.q1k	258	_	24	secreted cathepsin L1 *[Fasciola hepatica]*	4.66E-141	IPR013128, IPR000668, IPR000169, IPR013201	Peptidase C1A, papain, Peptidase C1A, papain C-terminal, Peptidase, cysteine peptidase active site, Proteinase inhibitor I29, cathepsin propeptide	CAB07275.1 from WBGene00000776: cpl-1 cathepsin-like protease status:Confirmed TR:O45734	embryonic lethal (Emb), slow growth (Gro),locomotion abnormal (Unc)

Accession	Protein	Score	Unique peptides[1]	% cover	Mr	emPAI[2]	Section[3]
Dormant larvae somatic							
A7UNB2	*F. hepatica* cathepsin B	294	7	24	37.6	0.29	2
Q8I7B2	*F. hepatica* procathepsin B2 precursor	111	3	9	38.0	0.09	2
A5X494	*F. hepatica* cathepsin B3	223	5	22	31.0	0.36	1-3
A8W7J0	*F. hepatica* procathepsin L	235	7	21	35.6	0.30	2
Q9GRW6	*F. hepatica* procathepsin L3	194	7	21	35.3	0.31	2
Q711M2	*F. hepatica* legumain-like precursor	266	7	21	47.8	0.30	1-3
Q71TT8	*S. japonicum* actin	193	6	21	41.7	0.16	2
Q27655	*F. hepatica* enolase	169	3	9	46.2	0.15	2
Q5DF69	*S. japonicum* SJCHGC06318 protein	98	4	15	41.8	0.08	2
B0LT92	*F. gigantica* thioredoxin peroxidase	163	3	16	25.4	0.29	3
Q2YHG1	*C. sinensis* myosin regulatory light chain	93	3	15	23.5	0.14	3
Q86ET3	*S. japonicum* Clone ZZD1444 mRNA sequence	76	2	14	17.8	0.19	4
NEJ secretome							
Q27655	*F. hepatica* enolase	871	10	35	46.2	1.17	1,2,4,5
P80527	*F. hepatica* Hemoglobinase-like protein 1	117	2	85	2.2	2.17	2,3
P91883	*F. hepatica* thioredoxin peroxidase	160	3	8	21.6	0.63	4,5
Q7M4G1	*F. hepatica* fatty acid-binding protein type 2	90	3	27	14.9	0.26	5
Q9U1G6	*F. hepatica* fatty acid-binding protein type 3	217	3	31	14.6	1.03	5
Q7M4G0	*F. hepatica* fatty acid-binding protein Fh15	104	2	20	14.7	0.61	5
Q9GRW4	*F. hepatica* partial procathepsin L3 (pFH64)	99	11	23	35.1	-	2-5
Q9GRW6	*F. hepatica* partial procathepsin L3 (pFH22)	99	8	23	35.4	-	2-5
Q95VA7	*F. gigantica* cathepsin L	99	5	27	37.4	-	2-5
B3TM67	*F. hepatica* cathepsin L3 (CL3-uy5)	99	9	26	35.1	-	2-5
B3TM68	*F. hepatica* cathepsin L3 (CL3-uy9)	99	9	39	35.3	-	2-5
A8W638	*F. hepatica* metacercariae cathepsin L	99	12	34	37.4	-	2-5
A8W7J0	*F. hepatica* metacercariae procathepsin L	99	13	38	35.7	-	2-5
Immature fluke secretome							
Q27655	*F. hepatica* enolase	602	10	30	46.2	0.62	1
Q5FX78	*F. gigantica* 14-3-3 protein	210	3	17	28.7	0.39	2
A7UNB2	*F. hepatica* cathepsin B	113	2	7	37.6	0.09	2
Q8I7B2	*F. hepatica* pro-cathepsin B2 precursor	99	2	6	38.0	0.18	5
Q711M2	*F. gigantica* legumain-like precursor	412	31	16	47.8	0.70	2
A6Y9U8	*F. gigantica* legumain-1	328	15	20	47.9	0.59	2
A6Y9U9	*F. gigantica* legumain-2	176	3	10	48.3	0.22	2
Q06A71	*F. hepatica* GST sigma class	138	4	23	24.5	0.14	3
P56598	*F. hepatica* GST mu class 26 kDa isozyme 1 (GST1)	303	9	38	25.7	0.63	3
P31671	*F. hepatica* GST mu class 26 kDa isozyme 7 (GST7)	361	14	34	25.3	2.05	3
P31670	*F. hepatica* GST mu class 26 kDa isozyme 47 (GST47)	381	15	44	25.3	2.04	3
P30112	*F. hepatica* GST mu class 26 kDa isozyme 51 (GST51)	390	16	39	25.3	1.70	3
B0LT92	*F. gigantica* thioredoxin peroxidase	237	5	21	24.5	0.29	3
O76945	*F. hepatica* protein disulphide isomerase	199	6	14	55.1	0.06	3
Q8T5Z9	*F. hepatica* cathepsin L	334	12	32	35.1	1.06	3
Q7JNQ9	*F. hepatica* secreted cathepsin L1	447	16	37	36.7	1.82	3
Q6R018	*F. hepatica* cathepsin L protein	292	13	24	36.6	1.00	3
Q24940	*F. hepatica* cathepsin L-like proteinase	356	15	32	36.8	1.36	3
Q9GRW5	*F. hepatica* cathepsin L1	282	12	27	35.1	0.88	3
Q7JNQ8	*F. hepatica* secreted cathepsin L2	363	12	37	37.0	0.67	3
A5Z1V3	*F. hepatica* secreted cathepsin L2	470	14	39	37.0	0.98	3
A5X483	*F. hepatica* cathepsin L2	367	14	39	24.5	1.78	3
A3FMG6	*F. hepatica* cathepsin L	228	9	30	36.9	0.19	3
Q95VA7	*F. gigantica* cathepsin L	235	7	15	37.4	0.18	3
A8W638	*F. hepatica* cathepsin L	274	6	25	37.3	0.29	3
Q9NGW3	*F. hepatica* cathepsin L	204	7	11	36.9	0.54	3
Q9NB30	*F. hepatica* cathepsin L	257	5	15	37.1	0.41	3
Q711N7	*F. hepatica* putative cys1 protein	197	4	7	79.2	0.13	5
Q9UAS2	*F. gigantica* fatty acid-binding protein 1	77	4	16	14.6	0.23	5
Q7M4G1	*F. hepatica* fatty acid-binding protein type 2	141	6	50	14.9	0.23	6
Q9U1G6	*F. hepatica* fatty acid-binding protein type 3	141	3	28	14.6	0.23	5
Q7M4G0	*F. hepatica* fatty acid-binding protein Fh15	349	14	59	14.7	5.55	5
Q2HPD3	*F. hepatica* cathepsin L1 proteinase	220	12	19	36.5	0.63	6
Q4KSL8	*F. gigantica* saposin-1	94	4	38	11.1	0.31	6
Adult secretome							
Q7JNQ9	*F. hepatica* secreted cathepsin L1	552	20	39	36.7	1.82	1
A5Z1V3	*F. hepatica* secreted cathepsin L2	444	12	34	37.0	1.16	1
Q9NB30	*F. hepatica* cathepsin L	372	9	18	37.1	0.41	1
Q8T5Z9	*F. hepatica* cathepsin L	481	17	40	35.1	1.25	1
Q9GRW5	*F. hepatica* cathepsin L1	476	21	32	35.1	1.06	1
Q6R018	*F. hepatica* cathepsin L protein	451	22	38	36.6	1.37	1
Q24940	*F. hepatica* cathepsin L-like proteinase precursor	444	19	33	36.8	1.58	1
A5X483	*F. hepatica* cathepsin L2	401	16	45	24.5	2.16	1
Q24945	*F. hepatica* cathepsin L-like protease	348	7	35	18.5	0.95	1
P81222	*F. hepatica* cathepsin L-like cysteine proteinase	297	12	57	12.3	4.58	1
Q9NGW3	*F. hepatica* cathepsin L	254	7	12	36.9	0.41	1
A5X484	*F. hepatica* cathepsin L1	114	4	20	24.3	0.14	1
A8I598	*F. hepatica* cathepsin L	77	2	7	28.1	-	1
B0LT92	*F. gigantica* thioredoxin peroxidase	211	5	21	25.4	0.29	1-3
Q9U1G7	*F. hepatica* thioredoxin	83	2	22	11.6	0.30	2
Q9UAS2	*F. gigantica* fatty acid-binding protein 1	107	3	16	14.6	0.52	2,3
Q7M4G1	*F. hepatica* fatty acid-binding protein type 2	153	5	44	14.9	0.23	2,3
Q9U1G6	*F. hepatica* fatty acid-binding protein type 3	81	2	18	14.6	0.23	2
Q7M4G0	*F. hepatica* fatty acid-binding protein Fh15	208	7	51	14.7	0.23	2,3

EST identifier	Score[1]	Unique peptides[2]	% cover	Top BLASTn hit	BLAST e	InterPro domains	Section[3]
Dormant larvae somatic							
Fhep12g03.q1k	99	7	7	*S. japonicum* SJCHGC07506 protein	5e-38	Pyruvate carboxylase; IPR005481	1
Fhep13a09.q1k	98	6	21	*F. hepatica* heat shock protein 70	0.0	Heat shock protein 70; IPR013126	1
Fhep22a04.q1k	97	7	23	*S. mansoni* actin	0.0	Actin; IPR004000	1
Fhep10a09.q1k	93	3	9	*S. japonicum* SJCHGC01242 protein	3e-53	Ribosomal protein L30; IPR000231	1
Fhep14c04.q1k	90	3	5	*S. japonicum* SJCHGC04174 protein	2e-35	PUA domain protein, IPR002478	1
Fhep12e03.q1k	90	4	10	*S. japonicum* SJCHGC00933 protein	2e-36	Cullin; IPR001373	1
Fhep29h09.q1k	76	5	12	*F. gigantica* legumain-1	1e-141	Legumain; IPR001096	1
Fhep13d07.q1k	99	9	26	*O. nigricans* aldolase	1e-92	Aldolase; IPR000741	2
Fhep27d08.q1k	99	8	33	*S. mansoni* actin	0.0	Actin; IPR004000	2
Fhep42h09.q1k	97	2	8	None	-	None	2
HAN3004-1f04.q1k	84	4	33	*F. hepatica* enolase trans-spliced	0.0	Enolase; IPR000941	2
Fhep19g06.q1k	79	3	5	None	-	None	2
Fhep40d04.q1k	98	6	17	*C. sinensis* mitochondrial malate dehydrogenase	2e-73	Malate dehydrogenase; IPR001236, IPR001252	3
Fhep22e12.q1k	95	4	15	*F. hepatica* thiol-specific antioxidant protein	0.0	Peroxiredoxin; IPR000866, IPR012335	3
Fhep11f10.q1k	91	3	6	None	-	Ribosome production factor; IPR007109	4
Fhep13f02.q1k	91	3	13	None	-	None	4
Fhep31e02.q1k	91	2	8	*S. mansoni* 23 kDa integral membrane protein	1e-28	Tetraspanin; IPR000301, IPR008952	4
Fhep48c07.q1k	90	2	9	*H. sapiens* clone RP11-215A20	-	None	4
HAN5021d04.p1kT7	79	5	11	None	1e-78	Histone H2A; IPR002119	4
Fhep25c09.q1k	76	4	12	*N. fischeri* peptidyl-prolyl cis-trans isomerase	1e-66	Cyclophilin; IPR002130	4
Fhep10a01.q1k	99	9	15	*A. mellifera* similar to germinal histone H4	6e-76	Histone H4; IPR001951	3,5
Fhep40g04.q1k	90	3	7	*N. tabacum* glycine-rich protein	0.017	None	5
Fhep49a02.q1k	70	2	12	*S. mansoni* Sm10 protein	5e-41	Dynein light chain; IPR001372	5
NEJ secretome							
Fhep29h09.q1k	99	12	12	*F. gigantica* legumain-1	1e-141	Peptidase C13, legumain; IPR001096	1-5
HAN4009g10.q1kT3	99	7	35	*F. hepatica* enolase	0.0	Enolase; IPR000941	1,2,4
HAN4004c01.q1kT3	99	7	30	*S. japonicum* fructose 1,6-bisphosphate aldolase	5e-97	Fructose-bisphosphate aldolase; IPR000741	2,5
Fhep20a08.q1k	99	6	26	*S. herbacea* tonoplast intrinsic protein gamma	2e-07	Protease inhibitor I4, serpin; IRP000215	2
HAN4005a07.q1kT3	99	3	8	*M. smithii* ATCC 35061 genomic sequence	0.22	Spermadhesin, CUB domain; IPR000859	1,2
Fhep07g02.q1k	95	3	8	*R. microplus* phosphoenolpyruvate carboxykinase	2e-47	Phosphoenolpyruvate carboxykinase; IPR008209	1,2
Fhep15c10.q1k	93	6	26	*F. hepatica* GAPDH	1e-144	GAPDH; IPR000173	2
HAN4015b05.q1kT3	88	3	11	*F. hepatica* cathepsin B (cat-B3)	2e-108	Peptidase C1A, cathepsin B; IPR015643	3
HAN4004e01.q1kT3	99	13	23	*F. gigantica* thioredoxin peroxidase	0.0	Thiol-specific antioxidant; IPR000866	4,5
HAN4005g04.q1kT3	99	15	22	*F. hepatica* thiol-specific antioxidant protein	0.0	Thiol-specific antioxidant; IPR000866	4
Fhep12b09.q1k	80	4	13	*C. sinensis* mitochondrial malate dehydrogenase	8e-98	Lactate/Malate dehydrogenase; IPR001236	4
Fhep10a01.q1k	99	14	19	*A. assectella* histone H4	7e-76	Histone H4: IPR001951	4-6
HAN5013f06.q1kT3	99	5	17	*T. asiatica* clone TaHC3-G5	4e-76	Histone H2B; IPR000558	4,5
HAN3004-1f09.q1k	99	6	24	*F. hepatica* fatty acid binding protein	0.0	Cytosolic fatty-acid binding; IPR000463	5
HAN4018d09.q1kT3	98	7	18	*F. hepatica* thioredoxin	0.0	Thioredoxin; IPR005746	5
HAN5021d04.p1kT7	97	4	19	*A. aegypti* histone h2a	1e-78	Histone H2A; IPR002119	5
Fhep46d02.q1k	96	3	15	*F. hepatica* fatty acid binding protein homolog	0.0	Cytosolic fatty-acid binding; IPR000463	5
HAN3004-1a05.q1k	94	2	11	*S. japonicum* SJCHGC09424 protein	1e-149	ATPase, F1/V1/A1 complex; IPR000194	5
Fhep27d03.q1k	90	2	8	*S. mansoni* FM fimbrin	2e-34	Calponin-like actin-binding; IPR001715	5
Fhep20f11.q1k	76	9	23	*F. hepatica* fatty acid binding protein homolog	2e-107	Cytosolic fatty-acid binding; IPR000463	5
HAN5016d06.q1kT3	99	6	27	*S. japonicum* SJCHGC00176 protein	2e-106	Ubiquitin; IPR000626	6
HAN4019c12.p1kT7	99	4	10	*S. japonicum* reinfection related protein 338	2e-21	Transmembrane regions	6
HAN5005h03.q1kaT3	98	2	13	*A. clavatus* peptidyl-prolyl cis-trans isomerase	7e-69	Peptidyl-prolyl cis-trans isomerase; IPR002130	6
Immature fluke secretome							
HAN4005b02.q1kT3	99	10	33	*S. japonicum* SJCHGC02147 protein	1e-34	Prolylcarboxypeptidase, IPR008758	1
HAN4018f01.q1kT3	97	5	45	*S. japonicum* SJCHGC00176 protein	2e-94	Ubiquitin, IPR000626	2
HAN5019h01.q1kT3	98	7	18	*F. gigantica* SAP-3	6e-179	Saposin, IPR008139, IPR007856, IPR008138	3
HAN5012c07.q1kT3	96	5	15	*F. hepatica* amoebapore-like protein	0.0	Saposin, IPR008139, IPR007856, IPR008138	5
Fhep45b05.q1k	97	6	13	*F. hepatica* cathepsin B3	1e-68	Cathepsin B, IPR015643	2
HAN4006g01.q1kT3	99	6	9	*T. regenti* cathepsin B1 isotype 2	1e-35	Cathepsin B, IPR015643	2
Fhep27f02.q1k	99	5	11	*Micromonas sp*. light-harvesting protein	7e-08	None	2
Fhep15h08.q1k	98	7	23	*S. herbacea* tonoplast intrinsic protein gamma	2e-08	None	2
HAN5010a10.q1kT3	98	7	30	*S. japonicum* SJCHGC03972 protein	0.003	Annexin, IPR001464	2
Fhep22a04.q1k	98	3	10	*S. mansoni* actin	0.0	Actin/actin-like, IPR004000	2
Fhep26g11.q1k	96	7	27	*Stylochus* sp. triosephosphate isomerase	6e-53	Triosephosphate isomerase, IPR000652	3
HAN4014h01.q1kT3	99	10	30	*S. japonicum* SJCHGC06812 protein	1e-41	Ferritin, IPR001519	4
Fhep08a09.q1k	95	3	9	*S. japonicum* clone SJCHGC05231	3e-51	ABC transporter-related, IPR003439	4
Adult secretome							
Fhep30b01.q1k	92	8	13	*G. domesticus* Gly d 13 allergen	2e-10	Prolylcarboxypeptidase, IPR008758	1
HAN5007d12.q1kaT3	99	20	36	*F. gigantica* SAP-3	0.0	Saposin, IPR008139, IPR007856, IPR008138	1
Fhep55d10.q1k	99	9	17	*A. clavatus* peptidyl-prolyl cis-trans isomerase	2e-64	Peptidyl-prolyl cis-trans isomerase, IPR002130	1

MS/MS data for cathepsin B3 identified in *F. hepatica* dormant larvae.

Accession	*m/z*	*z*	*Error (ppm)*	*Peptide*	*Score*
A5X494	1076.94	2	-10.70	SSYNVGEQETDIMMEIMK	54.2



#	b	b^{++}	b*	b^{*++}	b^{0}	b^{0++}	Seq.	y	y^{++}	y*	y^{*++}	y^{0}	y^{0++}	#
1	88.0393	44.5233			70.0287	35.5180	S							18
2	175.0713	88.0393			157.0608	79.0340	S	2065.8763	1033.4418	2048.8497	1024.9285	2047.8657	1024.4365	17
3	338.1347	169.5710			**320.1241**	160.5657	Y	1978.8443	989.9258	1961.8177	981.4125	1960.8337	980.9205	16
4	**452.1776**	226.5924	435.1510	218.0792	434.1670	217.5871	N	1815.7809	**908.3941**	1798.7544	899.8808	1797.7704	899.3888	15
5	**551.2460**	276.1266	**534.2195**	267.6134	533.2354	267.1214	V	1701.7380	**851.3726**	1684.7115	842.8594	1683.7274	842.3674	14
6	608.2675	304.6374	591.2409	296.1241	590.2569	295.6321	G	1602.6696	801.8384	1585.6430	793.3252	1584.6590	792.8331	13
7	737.3101	369.1587	720.2835	360.6454	719.2995	360.1534	E	1545.6481	773.3277	1528.6216	764.8144	1527.6376	764.3224	12
8	865.3686	433.1880	848.3421	424.6747	847.3581	424.1827	Q	**1416.6055**	708.8064	1399.5790	700.2931	1398.5950	699.8011	11
9	994.4112	497.7093	977.3847	489.1960	976.4007	488.7040	E	1288.5470	644.7771	1271.5204	636.2638	1270.5364	635.7718	10
10	1095.4589	548.2331	1078.4324	539.7198	1077.4483	539.2278	T	**1159.5044**	580.2558	1142.4778	571.7425	1141.4938	571.2505	9
11	1210.4859	605.7466	1193.4593	597.2333	1192.4753	596.7413	D	1058.4567	529.7320	1041.4301	521.2187	1040.4461	520.7267	8
12	1323.5699	662.2886	1306.5434	653.7753	1305.5594	653.2833	I	**943.4297**	472.2185	926.4032	463.7052	925.4192	463.2132	7
13	1470.6053	735.8063	1453.5788	727.2930	1452.5948	726.8010	M	**830.3457**	415.6765	813.3191	**407.1632**	812.3351	406.6712	6
14	1617.6407	809.3240	1600.6142	800.8107	1599.6302	800.3187	M	**683.3103**	342.1588	666.2837	333.6455	665.2997	333.1535	5
15	1746.6833	873.8453	1729.6568	865.3320	1728.6728	864.8400	E	536.2749	268.6411	519.2483	260.1278	518.2643	259.6358	4
16	1859.7674	930.3873	1842.7408	921.8741	1841.7568	921.3820	I	**407.2323**	**204.1198**	390.2057	195.6065			3
17	2006.8028	1003.9050	1989.7762	995.3918	1988.7922	994.8997	M	**294.1482**	147.5777	277.1217	139.0645			2
18							K	147.1128	74.0600	130.0863	65.5468			1

3.2 Conclusions

As a result of this study, several major components of *Fasciola hepatica* secretions identified are currently the leading candidates for development as first generation antifluke vaccines (including cathepsins, peroxiredoxin, glutathione S-transferase, and fatty acid-binding proteins) or are potential targets for novel flukicidal drugs (cathepsins) signifying the value of our integrated approach for the future identification of new targets for therapeutic intervention. As a step for future studies in anthelmintic or vaccine development, we correlated the secretions of major molecules of this parasite and their expression with the important steps in the migration and development of the parasite within the mammalian host. Further, the study of the transcriptome data for the infective NEJ larvae and immature liver stage parasites will help in the comparative analysis of secretory proteins.

At the same time, a number of non-classically secreted proteins were identified by proteomic analysis, which were not detected by the bioinformatics pipeline that we used, which was dependent on the presence of the classical N-terminal secretory signal, using SignalP [143], to identify secreted proteins, which we have addressed in Chapter 4.

Chapter 4: An analysis of the transcriptome of *Teladorsagia circumcincta*: its biological and biotechnological implications

4.1 Summary

Teladorsagia circumcincta classified in the order Strongylid, Superfamily Trichostrongyloidea and Family Trichostrongylidae is an economically important parasitic nematode of small ruminants (including sheep and goats) in temperate climatic regions of the world. This study was focussed to define the transcriptome of the adult stage of *T. circumcincta* (407,357 ESTs) and to surmise the main pathways associated to molecules known to be expressed in this nematode.

In order to address the bioinformatics prediction of non-classically secreted proteins, we have included SecretomeP [144] prediction. We also compared the pathway analysis program, KOBAS [152], used in Paper 3, with KAAS [153], a newly released program from the authors of the KEGG database [151]. The detailed annotation which includes pathway mapping of predicted proteins (including 112 excreted/secreted [ES] and 226 transmembrane peptides), domain analysis and GO annotation along with secretory signal peptides using a combination of bioinformatics tools. The mechanisms of resistance and anthelmintic actions are yet to be understood completely. As an alternative method to control parasites an insight into the molecular biology is essential to evade major issues related with the anthelmintic resistance. Development of a vaccine against this parasite is essential as sheep developed acquired immunity against this parasite. We report the first comprehensive analysis of the transcriptome from the adult stage of *T. circumcincta*, with an emphasis on characterization of molecules inferred to be ES proteins as possible immunogens and vaccine candidates.

An analysis of the transcriptome of *Teladorsagia circumcincta:* its biological and biotechnological implications

Ranjeeta Menon[1], Robin B. Gasser[2], Makedonka Mitreva[3,4] and Shoba Ranganathan[1, 5,*]

[1]Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, New South Wales 2109, Australia

[2]Department of Veterinary Science, The University of Melbourne, 250 Princes Highway, Werribee, Victoria 3030, Australia

[3] The Genome Institute, Washington University School of Medicine, 444 Forest Park Boulevard, St. Louis, MO 63108

[4] Department of Genetics, Washington University School of Medicine, 444 Forest Park Boulevard, St. Louis, MO 63108

[5]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, Singapore 117597

*Corresponding author

E-mail:

RM: ranjeetamaoj@gmail.com;

RGB: robinbg@unimelb.edu.au;

MM: mmitreva@watson.wustl.edu;

SR: shoba.ranganathan@mq.edu.au

Abstract

Background

Teladorsagia circumcincta (order Strongylida) is an economically important parasitic nematode of small ruminants (including sheep and goats) in temperate climatic regions of the world. Improved insights into the molecular biology of this parasite could underpin alternative methods required to control this and related parasites, in order to circumvent major problems associated with anthelmintic resistance. The aims of the present study were to define the transcriptome of the adult stage of *T. circumcincta* and to infer the main pathways linked to molecules known to be expressed in this nematode. Since sheep develop acquired immunity against *T. circumcincta*, there is some potential for the development of a vaccine against this parasite. Hence, we infer excretory/secretory molecules for *T. circumcincta* as possible immunogens and vaccine candidates.

Results

A total of 407,357 ESTs were assembled yielding 39,852 putative gene sequences. Conceptual translation predicted 24,013 proteins, which were then subjected to detailed annotation which included pathway mapping of predicted proteins (including 112 excreted/secreted [ES] and 226 transmembrane peptides), domain analysis and GO annotation was carried out using InterProScan along with BLAST2GO. Further analysis was carried out for secretory signal peptides using SignalP and non-classical sec pathway using SecretomeP tools.

For ES proteins, key pathways, including Fc epsilon RI, T cell receptor, and chemokine signalling as well as leukocyte transendothelial migration were inferred to be linked to immune responses, along with other pathways related to neurodegenerative diseases and infectious diseases, which warrant detailed future studies. KAAS could identify new and updated pathways like phagosome and protein processing in endoplasmic reticulum. Domain analysis for the assembled dataset revealed families of serine, cysteine and proteinase inhibitors which might represent targets for parasite intervention. InterProScan could identify GO terms pertaining to the extracellular region. Some of the important domain families identified included the SCP-like extracellular proteins which belong to the pathogenesis-related proteins (PRPs) superfamily along with C-type lectin, saposin-like proteins. The ‘extracellular region’ that corresponds to allergen

V5/Tpx-1 related, considered important in parasite-host interactions, was also identified.

Six cysteine motif (SXC1) proteins, transthyretin proteins, C-type lectins, activation-associated secreted proteins (ASPs), which could represent potential candidates for developing novel anthelmintics or vaccines were few other important findings. Of these, SXC1, protein kinase domain-containing protein, trypsin family protein, trypsin-like protease family member (TRY-1), putative major allergen and putative lipid binding protein were identified which have not been reported in the published *T. circumcincta* proteomics analysis.

Detailed analysis of 6,058 raw EST sequences from dbEST revealed 315 putatively secreted proteins. Amongst them, C-type single domain activation associated secreted protein ASP3 precursor, activation-associated secreted proteins (ASP-like protein), cathepsin B-like cysteine protease, cathepsin L cysteine protease, cysteine protease, TransThyretin-Related and Venom-Allergen-like proteins were the key findings.

Conclusions

We have annotated a large dataset ESTs of *T. circumcincta* and undertaken detailed comparative bioinformatics analyses. The results provide a comprehensive insight into the molecular biology of this parasite and disease manifestation which provides potential focal point for future research. We identified a number of pathways responsible for immune response. This type of large-scale computational scanning could be coupled with proteomic and metabolomic studies of this parasite leading to novel therapeutic intervention and disease control strategies. We have also successfully affirmed the use of bioinformatics tools, for the study of ESTs, which could now serve as a benchmark for the development of new computational EST analysis pipelines.

Introduction

Parasitic nematodes have a free-living state with their growth and survival controlled by the surrounding environment, especially by factors such as temperature and moisture.

Teladorsagia circumcincta is a key parasite that affect small ruminants in many countries around the world. Its lifecycle is direct and is similar to a number of gastrointestinal strongylid nematodes [1]. In brief, eggs released in faeces develop, and first-stage larvae (L1s) hatch usually within a day. L1s develop through to infective third-stage larvae (L3s) within about a week. L3s on pasture are ingested by the ruminant host, within which they exsheath in the rumenoreticulum and then pass to the abomasum to enter gastric glands and moult to fourth-stage larvae (L4). After this histotrophic phase, these larvae develop to adult female and male worms which reproduce.

T. circumcincta can be a major cause of economic loss due to poor productivity of ruminants, such as sheep and goats, failure to thrive and deaths, mainly in lambs [2, 3]. Together with other trichostrongylid nematodes, this parasite is usually controlled using a combination of anthelmintic treatment and management strategies. The emergence of resistance in trichostrongylids to the three main classes of anthelmintic drugs, including benzimidazoles (white drenches), imidazothiazoles/tetrahydropyrimidines (yellow/pink drenches) and macrocyclic lactones (clear drenches) compromises effective control. Improved insights into the molecular biology of these parasites have the potential to support the development of alternative methods of parasite control, in order to circumvent these resistance problems. Vaccination is considered by some researchers [4] to be a possible alternative approach to anthelmintic treatment, but attempts to develop a practical, commercial vaccine have been unsuccessful to date, likely because of a lack of detailed understanding of the immuno-molecular biology of the parasites, host-parasite interactions and disease. In spite of the economic significance of *T. circumcincta*, particularly in lambs, our understanding of the spectrum of antigens and immunogens involved in immune responses is still limited [5-7]. Nonetheless, there is evidence that excretory/secretory (ES) molecules are intimately involved in inducing and/or modulating the host's immune response [8], and it has been proposed that some of them are immunogens which could serve as potential vaccine targets [9, 10].

Antigenic or immunogenic molecules can be studied using a range of immunochemical or proteomic approaches [11], and transcriptomic studies can

strengthen such investigations by providing annotated datasets to allow the identification and classification of such key molecules. For instance, transcriptomic study of *T. circumcincta* has identified a number of components, including N-type and C-type single domain, activation-associated secreted proteins (ASPs) [5]. Preliminary evidence showed that the proteins inferred to represent the secretome in *T. circumcincta* larvae were associated with specific antibody responses in sheep against this parasite. These proteins might be incorporated into a vaccine for immunizing sheep to combat the Teladorsagiosis disease [12]. Importantly, N-type and C-type single domain activation-associated secreted proteins (ASPs) and *T. circumcincta* apyrase-1 (Tci-APY-1) in excretory/secretory products of L4s of *T. circumcincta*, identified also in transcriptomic studies [5, 13], have been demonstrated to be targets for early, specific IgA responses in infected sheep [5]. In addition, it has been reported that Tci-MIF-1, a macrophage migration inhibitory factor (MIF)-like molecule with tautomerase activity, might influence both host immune responses and nematode physiology [14]. Therefore, a detailed exploration of the transcriptome of *T. circumcincta* will provide a vital insight into the molecular biology of this parasite and should also provide a basis for studying parasite-host interactions and disease as well as parasite development and reproduction, with a view towards establishing new methods of prevention, treatment or control. Extending previous studies of strongylid nematodes [15-18], we report the first comprehensive analysis of the transcriptome from the adult stage of *T. circumcincta*, with an emphasis on characterization of molecules inferred to be ES proteins.

Materials and methods

The ESTs (NCBI EST database accession numbers SRR328404 and SRR328405) was generated by LS454 RNAseq sequencing of *T. circumcincta* 2284716780 fragment cDNA library using 454 GS FLX Titanium instrument. The dataset was initially assembled and annotated using different tools. Initially, all ESTs were pre-processed (using SeqClean [19] and RepeatMasker (Smit AFA & Green P)), for the removal of low-quality regions and consensus sequence generation using the Contig Assembly Program CAP3 which was followed by assembly [20]. This step was followed by ESTScan [21] translation of the

contiguous sequences (contigs) into peptides, which were then characterized *via* InterProScan [22] domain/motifs. Gene ontologies were inferred using BLAST2GO (V 2.3.5) [23], from Gene Ontology (MySQL-DB-data release go_200903) and InterProScan. Peptides predicted were also compared, using BLASTP, with data in the non-redundant protein sequence database from National Centre for Biotechnology Information (NCBI). The peptides were mapped to respective pathways in *C. elegans* using KOBAS [24] (KEGG [25] Orthology-Based Annotation System, KOBAS-1.1.0). The results were compared with pathway mapping using KAAS [26]. Similarity searches were done for protein databases for 'parasitic nematodes' and 'non-nematodes' generated in-house. Homologues/orthologues were identified *via* comparisons against WormBase using BLASTX. In addition, data for *C. elegans*, including RNA interference (RNAi), gene ontology, pathway and domain analyses were used for functional annotation.

The program SimiTri [27] was used for the comparison of inferred amino acid sequence data for *T. circumcincta* with those available for *C. elegans*, parasitic nematode and other organisms in public databases. SimiTri provides a two-dimensional display of relative similarity relationships among three different datasets. ES proteins were predicted using SignalP [28] to infer the presence of secretory signal peptides and signal anchors in predicted proteins. SecretomeP [29] was also used to predict proteins involved in a non-classical secretory pathway. Transmembrane proteins were predicted using TMHMM [30], a hidden Markov model-based program. Predicted proteins lacking transmembrane domains were subjected to further annotation using data available in Wormpep [31].

Results

cDNA analysis

From a total of 407,357 raw ESTs representing *T. circumcincta*, we obtained 366,897 high quality ESTs (Table 1), which ranged from 100-415 bp in length (mean: 206 bp; standard deviation: 43 bp). After clustering and assembly, the mean length of contigs increased to 360 bp (standard deviation: 173 bp). The G+C content of the coding sequence was 42%, consistent with other strongylid

nematodes [15, 32]. The assembly of the 366,897 ESTs yielded 39,852 representative sequences (22,382 contigs and 17,470 singletons; Table 1), of which 24,013 (60.3%) had open reading frames (ORFs). Similarity searches of these representative sequences identified 19,540 (49%) homologues in *C. elegans*, 32,476 (81.5%) in other parasitic nematodes and 13,064 (32.78%) in organisms other than nematodes.

Of the 6,628 (16.63 %) well-characterized molecules known to be associated with various biological processes (Additional File 1). Similarly, a comparative analysis of all 39,852 rESTs was also carried out using data from various nematodes (such as *Haemonchus contortus*, *Necator americanus*, *Nippostrongylus brasiliensis*, *Ostertagia ostertagi*, *Oesophagostomum dentatum*, *Ancylostoma caninum*, *Dictyocaulus viviparus*) [32];Mitreva et al., 2006) to explore gene conservation within clade V (Additional File 2). The analysis showed that 13,531 ESTs (33.95%) had significant sequence similarity to molecules from the members of clade V at an e-values cut-off of 1e-05.

6156 of them were mapped to 234 KEGG pathways of the homologues identified in *C. elegans*. *Oxidative phosphorylation* (n=357) and *Peptidases* (n=277 peptidases) were the highest represented according to the number of peptides mapped. Other groups of molecules were mapped to metabolic pathways such as *glycine, serine and threonine metabolism* (n=93), *insulin signaling pathway* (n=68), *signal transduction mechanisms* (n=54), *N-glycan biosynthesis* (n=33), *galactose metabolism* (n=31), *GnRH signaling pathway* (n=13), *aminosugars metabolism* (n=11), *linoleic acid metabolism* (n=5), *immune and complement and coagulation cascades* (n=4). A list of the KEGG pathways and the corresponding rESTs is provided as supplementary information (Additional File 3).

Peptides/Proteins

Of the 39,852 rESTs, 24,013 were inferred to have open reading frame (ORFs). 6,470 sequences mapped to 309 KEGG pathways, with the top 30 ‘highly represented’ pathways categorized by the number of peptides mapped, presented in Table 2. The main KEGG pathways represented were the *peptidases* (n=254)

and *ribosomal protein assembly pathway* (n=220). Other highly represented pathways by the peptides include *oxidative phosphorylation* (n=187) and *chaperones and folding catalysts* (n=144). Peptides were mapped to several pathways, including *purine metabolism* and *glycolysis/gluconeogenesis*. We have also compared our results by mapping the sequences using KAAS where 2,897 sequences were characterized as belonging to 257 pathways, with 30 'highly represented' pathways, categorized according by the number of peptides mapped, are presented in Table 3. The main KAAS pathways represented were *Huntington's disease* (n=91) and *oxidative phosphorylation* (n=84). Other highly represented pathways include the *ribosomal protein assembly pathway* (n=80), *ubiquitin mediated proteolysis* (n=33) and *glycolysis/gluconeogenesis* (n=29).

Peptides were also mapped to several other pathways, including *purine metabolism* and *pyrimidine metabolism*, pathways in *cancer, cysteine and methionine metabolism*, *glycolipid metabolism* and *glutathione metabolism*. Among the highly represented pathways, both KEGG and KAAS identified *oxidative phosphorylation*, *purine metabolism*, *glycolysis/gluconeogenesis* and *ribosomal protein assembly* pathways. We could identify GO terms using InterProScan for 24,013 proteins with 3,801 being assigned as involved in biological process (BP), 5,220 as associated with molecular function (MF) and 1,862 as part of the cellular component (CC) (Additional File 4). The analysis revealed that *oxidation reduction* (GO:0055114) and *metabolic process* (GO:0008152) were the most common GO categories representing biological processes. The highest represented GO terms in molecular function were *binding* (GO: 0005488) *and oxidoreductase activity* (GO:0016491). Whereas in cellular component, the highly represented GO terms were *ribosome* (GO:0005840) and *membrane* (GO:0016020). With 138 protein entries, the *protein kinase-like domain* family of proteins was the most represented, followed by *SCP-like extracellular domain* family, with 126 protein entries. Other highly represented group of domains are the *NAD(P)-binding domain*, *allergen V5/Tpx-1 related domain* and *transthyretin-like domain* (Table 4).

Secretome

We inferred 112 excreted/secreted proteins from the present data set of 39,852 rESTs (Additional File 5). Six Transthyretin proteins followed by three saposin-like

protein1 from *A. caninum*, three SXC1 (Six Cysteine Motif) proteins of *O. ostertagi*, two C-type single domain activation associated secreted protein ASP3 precursor from *O. ostertagi* were identified. Two C-type lectin-1 proteins represented in *Heligmosomoides polygyrus* and FMRFamide-like prepropeptide from *Oesophagostomum dentatum* one each of globin-like protein and putative L3 ES proteins of *O. ostertagi*, the bovine parasite which is closely related to *T. circumcincta* [33] were also identified. Neuropeptides or neuropeptide precursor molecules were represented among the annotated ES dataset.

Upon detailed annotations of the 112 adult secreted proteins, few novel proteins such as SXC1, protein kinase domain containing protein, trypsin family protein, TRYpsin-like protease family member (try-1), putative lipid binding protein were also identified. These novel proteins were not reported in the *T. circumcincta* proteomics analysis [12, 34] (Additional File 6). Subsequent detailed annotation of 226 transmembrane proteins helped in the identification of SXC1 (Six Cysteine Motif) proteins of *O. ostertagi*, putative L3 ES protein (*O. ostertagi*), putative major allergen (*Brugia malayi*). The details of these proteins are listed in Additional File 7.

We were able to functionally assign GO terms to 112 putative ES proteins with 50 being assigned as involved in biological process (BP), 81 as associated with molecular function (MF). The GO annotation summary with biological process, cellular component and molecular function details is provided in Figure 1. *Oxidation reduction* (GO:0055114) and *transmembrane transport* (GO:0055085) were the most common GO categories representing biological processes. The highest represented GO terms in molecular function were *binding* (GO: 0005488) and *catalytic activity* (GO: 0003824), known for their role in the identification of vaccine candidates or drug discovery. Additional File 8 gives a list of GO mappings consigned to ES protein data is provided in. 63 KEGG pathways showed mapping to 90 sequences with the top 30 'highly represented' pathways, categorized according to the number of putative ES proteins mapped, are presented in Table 5. *Protein kinases* (n=3) and *oxidative phosphorylation* (n=3) were the main KEGG pathways that mapped to the ES protein sequences.

Few other highly represented pathways by the ES proteins include the *glycerophospholipid metabolism* (n=3), *long-term depression* (n=3), *glycolysis/gluconeogenesis* (n=2). Several pathways including *purine metabolism*,

protein folding and associated processing, *MAPK signaling pathway*, *linoleic acid metabolism*, *GnRH signaling pathway* and *glutathione metabolism* were mapped by ES protein sequences. The list of KEGG pathways for ES proteins is available from Additional File 9.

55 KEGG pathways contained 85 sequences using KAAS with the top 30 'highly represented' pathways, categorized by the number of peptides mapped, are presented in Table 6. *Glycerophospholipid metabolism* (n=3) and *oxidative phosphorylation* (n=3) were the main KEGG pathways that mapped to the sequences. Few other highly represented pathways by ES proteins included *long-term depression* (n=3) and *Wnt signaling pathway* (n=2). ES proteins were mapped to several pathways such as *MAPK signaling pathway*, *linoleic acid metabolism*, *GnRH signaling pathway*, *glutathione metabolism* and *TGF-β signaling pathway*. The KEGG pathways with the corresponding ES proteins are provided in Additional File 10.

Table 7 gives the top 20 representative protein families with *metridin-like ShK toxin* as the highly represented family of proteins, comprising of 14 ES protein entries. Followed by *transthyretin-like* family of proteins, comprising 11 ES protein entries. *C-type lectin*, *saposin-like domain* and *SCP-like extracellular domain superfamily of the pathogenesis-related proteins* (PRPs) [35, 36] were the few other well-represented domain families in the present datasets. SecretomeP identified 615 sequences as non-classical secreted proteins at a cut-off value of 0.9. The detailed annotation of 615 secreted proteins revealed 62 KEGG pathways mapped by 105 sequences (Additional File 11) with the top highly represented pathways presented in Table 8.

Translation factors and *oxidative phosphorylation* were the main KEGG pathways that mapped to the sequences. *Protein kinases, peptidases, chaperones* and *folding catalysts* are among other well represented pathways by ES proteins. The analysis of 6,058 raw EST sequences from dbEST with an overlap of 20.3% with the cDNA resulted in 745 contigs and 1,696 singletons, where 2,242 had ORFs.

We could identify 315 putatively secreted proteins and 183 transmembrane proteins. An in-depth analysis of secreted proteins, identified 11 *C-type single domain activation associated secreted protein* (ASP3) precursors (*O. ostertagi*),

ten *ancyclostoma-secreted protein-like proteins* (*O. ostertagi*), five *cathepsin B-like cysteine proteases* (*O. ostertagi*), one *cathepsin L cysteine protease* (*H. contortus*), three *cysteine proteases*, four *precursor transthyretin like protein 1* (*O. ostertagi*), six *putative L3 ES proteins* (*O. ostertagi*), five *saposin-like protein 1* (*A. caninum*), three *secreted cathepsin F* (*T. circumcincta*), two *SXC1 proteins* (*O. ostertagi*), three *TransThyretin-related proteins*, two *venom-allergen-like proteins*.

Discussion

In the absence of a genomic sequence for *T. circumcincta*, 407,357 raw EST sequences were analysed to obtain quality ESTs with a sequencing success of 90.06% which is consistent with previous studies [15, 34, 37]. To infer the proteome for *T. circumcincta*, all rESTs were then subjected to analyses against three databases containing protein sequences. Data were compared with protein sequences available for (i) *C. elegans* (from WORMPEP v.182 Wombase([http://wormbase.org/)), (ii) parasitic nematodes (available protein sequences and peptides from conceptually translated ESTs) and (iii) organisms other than nematodes (from NCBI non-redundant protein database) [38]. Three-way comparison of *T. circumcincta* rESTs with homologues from *C. elegans*, WORMPEP and parasitic nematodes have been figuratively presented (Figure 2) using SimiTri.

Some of the proteins predicted to be parasite- or nematode-specific were identified by similarity searches of rESTs and these proteins in parasitic nematodes were either absent from or very different from the corresponding molecules in their host(s).

Comparative analysis was carried out to identify homologues in *C. elegans*, the best characterized nematode in relation to its genome, genetics, biology, physiology, biochemistry as well as the localization and functions of molecules Wormbase [39]. This study showed that 7,537 of them were mapped to key biological pathways including *oxidative phosphorylation*, *peptidases* and the *ribosomal protein assembly pathway*. *Oxidative phosphorylation* relates to genes that encode NADH dehydrogenases, succinate dehydrogenases, cytochrome c oxidases, cytochrome c reductases, ATPases and ATP synthases (complexes I-V)

[40]. Several *peptidases* are known to play a vital role in the moulting process [41], these include metallo-peptidases that might be candidates for chemotherapeutic interventions [42-45]. The *ribosomal protein assembly pathway* is composed of genes that encode various proteins of the ribosomal subunits. These proteins are closely related functionally and need to interact with each other physically to form a large protein complex known as the ribosome [40]. Other pathways represented include the *carbon fixation pathways*. Several enzymes in nematodes map to KEGG carbon fixation pathways [http://www.genome.jp/kegg-bin/show_pathway?category=Nematodes&mapno=00720], which refer to normal energy pathways such as *glycolysis, gluconeogenesis* (which is actually carbon fixing) and *tricarboxylic acid cycle*.

The pathways identified using KOBAS such as *TGF-β signaling pathway* and *insulin signaling pathway* trigger an ‘‘alternative’’ developmental pathway and regulate the transition of environmental stress on *C. elegans* in the first larval stage of its life cycle [46, 47]. The disruption of both insulin-like and DAF-7 transforming growth factor (TGF)-β signalling pathways causes developmental arrest [48, 49]. Abundant levels of transcription of GTP-CH transcripts in some parasitic species could be associated with production of serotonin to regulate these processes, in a way that is similar to that of *C. elegans*, if a TGF-β pathway does indeed regulate developmental events in parasitic nematodes [34]. These areas are of great interest and deserve detailed investigation, particularly given that molecules representing the TGF-β pathway have been described for a number of parasitic nematodes such as *B. pahangi*, *B. malayi* and *P. trichosuri* [50-52].

Proteins expected to play critical roles in host-parasite interactions including immune responses are predicted to be involved in antigen processing and presentation or complement and coagulation cascades.

Nematode enzymes mapped to known human disease pathways such as *Huntington’s disease, Alzheimers disease, Parkinson's disease* and *Vibrio cholerae infection*. The neurological disorder pathways are known to describe the morbidity and depression associated with helminthic infections. The *Vibrio cholera*

infection pathway supports this parasite being similar to gastrointestinal strongylid nematodes.

Clearly, much more work is required to establish the functional roles of such proteins in the parasite and/or the host and also to identify essential proteins required in each pathway, even though they are not well represented. Some of the proteins are inferred to be excreted/secreted from the nematode. These include *serine proteinase inhibitors* and *cathepsin B-like cysteine proteases* which are proposed to interfere with the immune system at the antigen processing and presentation stages, thereby, to interrupt the cytokine network and to down-regulate inflammation [53]. Families of proteins considered as important targets for parasite invention and control were also identified represented by *serine, cysteine* as well as *proteinase inhibitors* which are also supported by domain analysis [54-56]. The proteinase inhibitors might protect the parasite against digestion by endogenous or host-derived proteinases [53].

Of the 39,852 rESTs, 24,013 were inferred to have open reading frame (ORFs). The most represented domain family of proteins were the *protein kinase-like* and the *SCP-like extracellular domains*, followed by *NAD(P)-binding domain, allergen V5/Tpx-1 related domain* and *transthyretin-like domain*. Analysis of several protein and protein domains present in *C. elegans* [57] revealed that protein kinases comprise the second largest family of protein domains in worms. Protein kinases are required for the existence of multicellular organisms and are likely to be involved in the complex signal transduction pathways including cell–substratum and cell–cell adhesion, transmembrane signaling in response to humoral factors and cell survival or programmed cell death. Other protein kinases provide signals that regulate metazoan-specific transcription factors, particularly those containing Zn-finger domains [58].

SCP/TAPS family members belong to the *cysteine-rich secretory protein* (CRISP) and have been identified in various eukaryotes. They also seem to have some biological roles linked with the member proteins within this superfamily [59].

The *sperm-coating protein (SCP)-like* extracellular proteins, also called SCP/Tpx-1/Ag5/PR-1/Sc7, play major biological roles in the host–pathogen

interplay [60] along with other groups of proteins [61] . $NADP^+$ plays a vital role in developmental process and also acts as a reducing agent in anabolism along with NAD^+, a coenzyme involved in key pathways like *glucose metabolism* and *fatty acid synthesis* [62]. In *Strongyloidae*, the *allergen V5/Tpx-1 related domain* is considered as one of the most abundant InterPro domain that may be important in parasitism [32]. It symbolizes various members such as the *ancylostoma-secreted or activation-associated proteins* (ASPs) that belong to the pathogenesis-related protein (PRP) superfamily [35]. The *transthyretin-like domain*, an abundant nematode-specific motif [63] was recently identified as being abundantly transcribed in the transcriptome of *B. malayi* [64]. *Lectins* are carbohydrate binding proteins and the CLec fold constitutes a general ligand (including protein)-binding motif [65].

The vertebrate immune cell signalling and trafficking, activation of innate immunity in both vertebrates and invertebrates and venom-induced haemostasis, have the involvement of C-type lectins [66]. *Metridin-like ShK toxin domains* are highly represented in the Strongylida [32]. Though the specific function of these proteins are not known, they are assumed to be involved in defense or digestion [67]. *WD40 repeats* (also known as WD or beta-transducin repeats) are involved in signal transduction and transcription regulation along with cell-cycle control and apoptosis [68, 69].

Heat shock proteins, such as HSP-20 are reported to be present in the parasitic nematode, *H. contortus* (barber's pole worm) which afflicts small ruminant species and in the adult stage of *A. caninum* and other nematodes including the bovine lungworm *Dictyocaulus viviparus* and the common roundworm of canids *Toxocara canis*. The expression of this molecule was shown not to be controlled by heat shock treatment [70].

'EF-hand' domains are involved in protein-protein interactions regulated by various specialized systems (e.g., Golgi system, voltage dependent calcium channels and calcium transporters) [71]. The maturation of the nervous system and the formation of ciliated sensory neurons require both EF-hand and WD40 proteins in *C. elegans* [72, 73]. *Major sperm proteins (MSPs)*, a large protein family, are known to be largely involved in nematode sperm motility [74, 75]. MSPs

(expressed in recombinant form) have been proposed as vaccine candidates [76]. The entire list of domains and their details are given in Additional File 12. The protein sequences were assigned functionality based on BLASTP against the NR database (Additional File 13). Different classes of proteases are assigned based on the catalytic mechanisms and are named based on their active catalytic centre residues (*aspartic, serine and cysteine proteases*) or after their dependence on co-factors for activity (*metalloproteases*). Of the four classes of proteases aspartic proteases are considered to be the most conserved group.

Cysteine proteases are most likely involved in tissue penetration and feeding [77]. *Cysteine, aspartic and metallo-proteases* represented in *N. americanus*, are known to function in a multi-enzyme cascade to digest haemoglobin and other serum proteins [78, 79]. *SCP (sperm coating protein)-1 superfamily* members include insect venom allergens, plant pathogenesis family-1 (PR-1) proteins and VAL proteins beside mammalian cysteine-rich sperm proteins (CRISPs). No rational function for this protein family has been demonstrated despite the sequence similarity [8]. *Astacin-like metalloproteases* are vital for establishment of the parasite in the host. *MTP-1* and *the astacin-like MTP* secreted by infective larvae of hookworms, are primarily reported in *A. caninum* [80] [81, 82]. The enzyme *guanosine-50-triphosphate (GTP)-cyclohydrolase* may be involved in larval development [35]. In parasitic nematodes, *astacin-like* molecules are considered to be involved with moulting, tissue penetration and immunomodulation besides feeding [34, 80]. They are also anticipated to be vaccine candidates against parasitic nematodes [82, 83].

Pathway analysis using KOBAS [24] mapped a total of 6,470 sequences to 309 KEGG pathways. The results were compared by mapping the sequences using KAAS [26], where a total of 2,897 sequences were mapped to 257 KEGG pathways. The perceptive of such mapping in biological pathways will help in identifying vital proteins required in each pathway.

Functionally varied classes of molecules such as digestive enzymes, extracellular proteinases, chemokines, morphogens, cytokines, toxins, hormones, antibodies, antimicrobial peptides included in secretome constitute the entire set of secreted proteins, representing up to 30% of the proteome of an organism [84]. *SXC1 (Six Cysteine Motif)* proteins of *O. ostertagi*, *transthyretin* proteins, *saposin-*

like protein 1, *C-type lectin-1*, *globin-like protein*, *Na-ASP-2,* a PR-1 protein from *N. americanus*, *ASP-3* from *O. ostertagi*, *neuropeptides* and *cytochrome P450s* were also identified from the 112 excreted/secreted proteins inferred from the data set of 39,852 rESTs.

The *SXC domain*, also termed nematode-six cysteine, NC6 [85], was identified in surface coat proteins of the parasitic ascarid *T. canis* [86, 87] along with zinc metalloproteases and tyrosinases of *C. elegans*. SXC domains have also been identified in other helminths such as *Ascaris, Brugia*, *Trichuris muris* and *Necator* [88]. The function of the motif is not known but it is suggested that it is involved in protein-protein interactions, particularly those associated with nematode surfaces [89] or that it acts as a signalling ligand [90]. In general, SXC motif containing proteins have a putative secretory signal peptide and are therefore extracellular. The *transthyretin-like* (ttl) gene family, also known as ‘‘family 2’’ [91], has been classified as nematode-specific based on the genome-wide study of *C. elegans*. These are the largest conserved nematode-specific gene families, coding for a group of proteins with significant sequence similarity to transthyretins (TTR) and transthyretin-related proteins (TRP) [92]. Transthyretin-like protein families are potential vaccine candidates against human filariasis [93].

As part of transcriptomic analysis of some members of the phylum Nematoda more than 4,000 nematode-specific protein families encoded by nematode-restricted genes were defined with TTL family representing one of the largest [32]. TTL protein domain was represented 185 times in all nematodes studied. This included 18 ttl genes in *O. ostertagi* as a result of protein domain search using the NEMBASE database [92]. The TTL family shows characteristics comparable with those of neuropeptides, i.e., a large protein family with secretion signals and different expression patterns between the members of the family and are likely to play a role in the nervous system of the nematodes [94]. SAPLIPs (saposin-like proteins) are a diverse family of lipid interacting proteins [95] that have six conserved cysteine residues forming three disulfide bridges [95-98]. The majority of Ac-slp-1 is expressed in the L3 and adult worm, although it is detected in RNA from all developmental stages of *A. caninum*.

While the Ac-slp-1 and slp-2 mRNAs are expressed in the intestines of multiple developmental stages of *A. caninum*, suggesting multiple functions in parasite biology, both Ac-SLP-1 and SLP-2 are localized to the intestines and could play a role in parasite feeding. The SLP-1 protein could also interact with host cells [99]. Worm carbohydrates may be masked from host immune cells by parasite C-TLs. Nematode C-TLs may also have roles unconnected with immune evasion [8]. Antigen uptake and presentation, cell adhesion, apoptosis and T cell polarization are the few immune processes in which C-type lectins and galectins are involved [66]. CTLs are perhaps the most prominent in the mammalian immune system. *Heligmosomoides polygyrus*, the natural parasites of mice, are the most widely-studied amongst the parasitic nematodes. Immunological interactions with the host are presumed to be mediated by the new C-type lectins from these rodent parasites which are preferentially expressed by the mature adult stages [100].

Craig *et al*. [101] were able to identify a homologue of a globin-like ES protein from *O. ostertagi* in L4 and adult *T. circumcincta* protein. Adult ES proteins in *O. ostertagi* identified a homologue of an ASP and a vitellogenin [92], which were not identified in *T. circumcincta* ES proteins [101]. However, we have successfully identified a *globin-like protein* and *Na-ASP-2* - a PR-1 protein from *Necator americanus*) [102] and *ASP-3* from *O. ostertagi* [103]. ASPs are the members of a group of nematode-specific molecules [5]. Proteins in this family have been identified in a wide range of organisms [35], including human hookworm [104], filarial nematodes [105, 106], trichostrongylids such as *H. contortus* [107, 108], *schistosomes* [59, 109, 110] as well as free-living *C. elegans* [111]. It has been suggested that ASPs are key to the transition of nematodes from free-living to the parasitic state [112]. It has also been suggested that they exhibit homology to a diverse, yet evolutionarily-related, group of secreted proteins classified as the SCP/Tpx-1/Ag5/PR-1/Sc7 family [5].

Na-ASP-2 has recently been shown to induce neutrophil chemotaxis *in vitro* and *in vivo* [113], but it remains uncertain if this is a widespread property of VAL homologues [8]. The role of nematode ASPs as valid vaccine candidates has also been investigated [114]. ASPs have been suggested to have the role of allergens [34]. They also have a role in modulation of the host immune response [115], in maintenance of the parasites at their host niche [116, 117]and in maintenance

and/or exit from arrested development [118]. ASPs are highly represented in EST datasets derived from parasitic stages of *T. circumcincta* and are abundant in the L4 ES proteins of this nematode [34]. Neuropeptide-like proteins have shown to be present in *O. ostertagi* [119]. These intercellular signaling molecules and particularly the FMRFamide-related peptides (FaRPs), have been most widely studied in *Ascaris suum* where they are present throughout the nervous system [34]. Cysteine-rich proteins were highly represented in *T. circumcincta* L4-specific dataset and were suggested to have a role in establishment and immune evasion [113].

Members of the astacin family have a wide range of functions [120] including immunomodulation [121], growth-factor processing, pattern formation in embryos [122], digestion, tissue penetration [80, 123] and hatching [124]. Nematode AST-like metalloproteinases play role in stimulating innate and adaptive immune responses early in infection [83]. Cytochrome P450s, the candidate drug-resistance genes, were also identified. These could affect the expression of the functional group 'xenobiotic degradation and metabolism' [6]. We have attempted to integrate the transcriptomics data with the proteomics analysis from previous reports to understand the role of ES proteins in host-parasite interaction (Additional File 6). Kyoto Encyclopedia of Genes and Genomes database (KEGG) was searched with KOBAS and KAAS to categorize functionality by assigning secreted protein sequences to biological pathways. *Fc epsilon RI signaling pathway*, *T cell receptor signaling pathway*, *leukocyte transendothelial migration* and *chemokine signaling pathway* represent the immune system related pathways which could play a critical role in understanding the immune responses.

We were also able to identify pathways related to neurodegenerative diseases and infectious diseases. Figure 3 shows the pathways represented using the ipath tool [125]. Identification of the role of such proteins as potential players in pathway analysis will help in our understanding of nematode biology in the context of parasite-host interplay. However, they are thought to be involved in immune responses in either the host or the parasite, which can be the focus of future studies. Of the pathways identified using KAAS, the protein family comprising serine, cysteine and metallo-proteinases and proteinase inhibitors in the EST datasets could form the basis of *in vitro* and *in vivo* studies. The parasite might be

protected against digestive degradation by blocking endogenous proteinases within the host, with proteinase inhibitors. Tissue migration and other interactions with host cells may be facilitated by the function of these enzymes, by mediating or changing proteolytic functions [53]. Several studies have considered these enzymes as important therapeutic targets for parasite control [54-56, 93]. Results from the pathway analysis carried out using KOBAS were compared with the results obtained using KAAS. The identification of domain/motif or region in a protein sequence characteristic for a particular protein family helps in the annotation by the assignment of protein function. We also searched the InterPro member databases [126] using Interproscan. Amongst the InterPro domains identified, the *Metridin-like ShK* and *transthyretin-like domains* were amongst the most represented, followed by *C-type lectin*, *saposin-like* and *SCP-like* extracellular domains. The *Metridin-like ShK* domain has already been shown to be highly represented in Strongylida and is often present in metallopeptidases [127, 128]. The results showed that the most common molecules associated with the extracellular region correspond to *allergen V5/Tpx-1 related protein*. Additional File 14 contains the domain details of ES proteins. Overall, KOBAS and KAAS provided similar results.

Homologues RNAi phenotypes were identified by the comparison of 112 predicted ES proteins with the free-living nematode *C. elegans* and the associated RNAi phenotypes were studied to understand the function(s) and importance of homologous genes in other nematodes (of animals).

From these, 133 *C. elegans* homologues were retrieved with RNAi phenotypes (Additional File 15): *Emb* (embryonic lethal, including pleiotropic defects severe early emb), *Lva* (larval arrest), *Gro* (slow growth). *Stp* (sterile progeny), *Lvl* (larval lethal) and *Ste* (maternal sterile). In the current dataset, we have selected RNAi phenotypes essential for nematode survival or growth as well as those representing potential drug and/or vaccine targets [129, 130]. Lethality can be considered as the most attractive RNAi phenotype applicable to all developmental stages that are less susceptible to available drugs as a result of interference with a vital process. Other attractive phenotypes include sterility that would lead to death. RNAi phenotypes help in understanding the concerns regarding genetic redundancy [131].

Authors' contributions

MM generated and pre-processed the data. RM carried out the analysis, computational studies and drafted the manuscript. RM, SR and RBG participated in the design of the study and interpretation of data. SR, MM and RBG conceived the project and SR finalized the manuscript. All authors have read and approved the final manuscript.

Acknowledgements

RM gratefully acknowledges the award of a Macquarie University Research Excellence Scholarship. Open access application charges were borne by Macquarie University. The data generation and research at Washington University School of Medicine was supported by grant from NHGRI and NIAID.

References

1. Abubucker S, Zarlenga DS, Martin J, Yin Y, Wang Z, McCarter JP, Gasbarree L, Wilson RK, Mitreva M: **The transcriptomes of the cattle parasitic nematode Ostertagia ostartagi**. *Veterinary parasitology* 2009, **162**(1-2):89-99.
2. O'Connor LJ, Walkden-Brown SW, Kahn LP: **Ecology of the free-living stages of major trichostrongylid parasites of sheep**. *Vet Parasitol* 2006, **142**(1-2):1-15.
3. Gibson TE, Everett G: **Effect of different levels of intake of Ostertagia circumcincta larvae on the faecal egg counts and weight gain of lambs**. *J Comp Pathol* 1976, **86**(2):269-274.
4. McNeilly TN, Devaney E, Matthews JB: **Teladorsagia circumcincta in the sheep abomasum: defining the role of dendritic cells in T cell regulation and protective immunity**. *Parasite Immunol* 2009, **31**(7):347-356.
5. Nisbet AJ, Smith SK, Armstrong S, Meikle LI, Wildblood LA, Beynon RJ, Matthews JB: **Teladorsagia circumcincta: activation-associated secreted proteins in excretory/secretory products of fourth stage larvae are targets of early IgA responses in infected sheep**. *Exp Parasitol* 2010, **125**(4):329-337.
6. Dicker AJ, Nath M, Yaga R, Nisbet AJ, Lainson FA, Gilleard JS, Skuce PJ: **Teladorsagia circumcincta: the transcriptomic response of a multi-drug-resistant isolate to ivermectin exposure in vitro**. *Exp Parasitol* 2011, **127**(2):351-356.

7. Hein WR, Pernthaner A, Piedrafita D, Meeusen EN: **Immune mechanisms of resistance to gastrointestinal nematode infections in sheep**. *Parasite Immunol* 2010, **32**(8):541-548.
8. Hewitson JP, Grainger JR, Maizels RM: **Helminth immunoregulation: the role of parasite secreted proteins in modulating host immunity**. *Mol Biochem Parasitol* 2009, **167**(1):1-11.
9. Hotez PJ, Bethony JM, Diemert DJ, Pearson M, Loukas A: **Developing vaccines to combat hookworm infection and intestinal schistosomiasis**. *Nat Rev Microbiol* 2010, **8**(11):814-826.
10. De Vries E, Bakker N, Krijgsveld J, Knox DP, Heck AJ, Yatsuda AP: **An AC-5 cathepsin B-like protease purified from Haemonchus contortus excretory secretory products shows protective antigen potential for lambs**. *Vet Res* 2009, **40**(4):41.
11. Greub G, Kebbi-Beghdadi C, Bertelli C, Collyn F, Riederer BM, Yersin C, Croxatto A, Raoult D: **High throughput sequencing and proteomics to identify immunogenic proteins of a new pathogen: the dirty genome approach**. *PLoS One* 2009, **4**(12):e8423.
12. Smith SK, Nisbet AJ, Meikle LI, Inglis NF, Sales J, Beynon RJ, Matthews JB: **Proteomic analysis of excretory/secretory products released by Teladorsagia circumcincta larvae early post-infection**. *Parasite Immunol* 2009, **31**(1):10-19.
13. Nisbet AJ, Zarlenga DS, Knox DP, Meikle LI, Wildblood LA, Matthews JB: **A calcium-activated apyrase from Teladorsagia circumcincta: an excretory/secretory antigen capable of modulating host immune responses?** *Parasite Immunol* 2011, **33**(4):236-243.
14. Nisbet AJ, Bell NE, McNeilly TN, Knox DP, Maizels RM, Meikle LI, Wildblood LA, Matthews JB: **A macrophage migration inhibitory factor-like tautomerase from Teladorsagia circumcincta (Nematoda: Strongylida)**. *Parasite Immunol* 2010, **32**(7):503-511.
15. Ranganathan S, Nagaraj SH, Hu M, Strube C, Schnieder T, Gasser RB: **A transcriptomic analysis of the adult stage of the bovine lungworm, Dictyocaulus viviparus**. *BMC Genomics* 2007, **8**:311.
16. Nagaraj SH, Gasser RB, Nisbet AJ, Ranganathan S: **In silico analysis of expressed sequence tags from Trichostrongylus vitrinus (Nematoda): comparison of the automated ESTExplorer workflow platform with conventional database searches**. *BMC Bioinformatics* 2008, **9 Suppl 1**:S10.
17. Nisbet AJ, Gasser RB: **Profiling of gender-specific gene expression for Trichostrongylus vitrinus (Nematoda: Strongylida) by microarray analysis of expressed sequence tag libraries constructed by suppressive-subtractive hybridisation**. *Int J Parasitol* 2004, **34**(5):633-643.
18. Cantacessi C, Mitreva M, Campbell BE, Hall RS, Young ND, Jex AR, Ranganathan S, Gasser RB: **First transcriptomic analysis of the economically important parasitic nematode, Trichostrongylus colubriformis, using a next-generation sequencing approach**. *Infect Genet Evol* 2010, **10**(8):1199-1207.
19. Chen YA, Lin CC, Wang CD, Wu HB, Hwang PI: **An optimized procedure greatly improves EST vector contamination removal**. *BMC Genomics* 2007, **8**:416.
20. Huang X, Madan A: **CAP3: A DNA sequence assembly program**. *Genome Res* 1999, **9**(9):868-877.

21. Iseli C, Jongeneel CV, Bucher P: **ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences**. *Proc Int Conf Intell Syst Mol Biol* 1999:138-148.
22. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R: **InterProScan: protein domains identifier**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W116-120.
23. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research**. *Bioinformatics* 2005, **21**(18):3674-3676.
24. Wu J, Mao X, Cai T, Luo J, Wei L: **KOBAS server: a web-based platform for automated annotation and pathway identification**. *Nucleic Acids Res* 2006, **34**(Web Server issue):W720-724.
25. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG**. *Nucleic Acids Res* 2006, **34**(Database issue):D354-357.
26. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M: **KAAS: an automatic genome annotation and pathway reconstruction server**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W182-185.
27. Parkinson J, Blaxter M: **SimiTri--visualizing similarity relationships for groups of sequences**. *Bioinformatics* 2003, **19**(3):390-395.
28. Bendtsen JD, Nielsen H, von Heijne G, Brunak S: **Improved prediction of signal peptides: SignalP 3.0**. *J Mol Biol* 2004, **340**(4):783-795.
29. Bendtsen JD, Jensen LJ, Blom N, Von Heijne G, Brunak S: **Feature-based prediction of non-classical and leaderless protein secretion**. *Protein Eng Des Sel* 2004, **17**(4):349-356.
30. Emanuelsson O, Brunak S, von Heijne G, Nielsen H: **Locating proteins in the cell using TargetP, SignalP and related tools**. *Nat Protoc* 2007, **2**(4):953-971.
31. Bieri T, Blasiar D, Ozersky P, Antoshechkin I, Bastiani C, Canaran P, Chan J, Chen N, Chen WJ, Davis P *et al*: **WormBase: new content and better access**. *Nucleic Acids Res* 2007, **35**(Database issue):D506-510.
32. Parkinson J, Mitreva M, Whitton C, Thomson M, Daub J, Martin J, Schmid R, Hall N, Barrell B, Waterston RH *et al*: **A transcriptomic analysis of the phylum Nematoda**. *Nat Genet* 2004, **36**(12):1259-1267.
33. Vercauteren I, Geldhof P, Peelaers I, Claerebout E, Berx G, Vercruysse J: **Identification of excretory-secretory products of larval and adult Ostertagia ostertagi by immunoscreening of cDNA libraries**. *Mol Biochem Parasitol* 2003, **126**(2):201-208.
34. Nisbet AJ, Redmond DL, Matthews JB, Watkins C, Yaga R, Jones JT, Nath M, Knox DP: **Stage-specific gene expression in Teladorsagia circumcincta (Nematoda: Strongylida) infective larvae and early parasitic stages**. *Int J Parasitol* 2008, **38**(7):829-838.
35. Henriksen A, King TP, Mirza O, Monsalve RI, Meno K, Ipsen H, Larsen JN, Gajhede M, Spangfort MD: **Major venom allergen of yellow jackets, Ves v 5: structural characterization of a pathogenesis-related protein superfamily**. *Proteins* 2001, **45**(4):438-448.
36. Lu G, Villalba M, Coscia MR, Hoffman DR, King TP: **Sequence analysis and antigenic cross-reactivity of a venom allergen, antigen 5, from hornets, wasps, and yellow jackets**. *J Immunol* 1993, **150**(7):2823-2830.
37. Cottee PA, Nisbet AJ, Abs El-Osta YG, Webster TL, Gasser RB: **Construction of gender-enriched cDNA archives for adult**

Oesophagostomum dentatum by suppressive-subtractive hybridization and a microarray analysis of expressed sequence tags. *Parasitology* 2006, **132**(Pt 5):691-708.
38. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank**. *Nucleic Acids Res* 2011, **39**(Database issue):D32-37.
39. http://wormbase.org
40. Huang R, Wallqvist A, Covell DG: **Comprehensive analysis of pathway or functionally related gene expression in the National Cancer Institute's anticancer screen**. *Genomics* 2006, **87**(3):315-328.
41. Craig H, Isaac RE, Brooks DR: **Unravelling the moulting degradome: new opportunities for chemotherapy?** *Trends Parasitol* 2007, **23**(6):248-253.
42. Bennuru S, Semnani R, Meng Z, Ribeiro JM, Veenstra TD, Nutman TB: **Brugia malayi excreted/secreted proteins at the host/parasite interface: stage- and gender-specific proteomic profiling**. *PLoS Negl Trop Dis* 2009, **3**(4):e410.
43. Hong X, Bouvier J, Wong MM, Yamagata GY, McKerrow JH: **Brugia pahangi: identification and characterization of an aminopeptidase associated with larval molting**. *Exp Parasitol* 1993, **76**(2):127-133.
44. Rhoads ML, Fetterer RH, Urban JF, Jr.: **Secretion of an aminopeptidase during transition of third- to fourth-stage larvae of Ascaris suum**. *J Parasitol* 1997, **83**(5):780-784.
45. Rhoads ML, Fetterer RH, Urban JF, Jr.: **Effect of protease class-specific inhibitors on in vitro development of the third- to fourth-stage larvae of Ascaris suum**. *J Parasitol* 1998, **84**(4):686-690.
46. Patterson GI, Padgett RW: **TGF beta-related pathways. Roles in Caenorhabditis elegans development**. *Trends Genet* 2000, **16**(1):27-33.
47. Beall MJ, Pearce EJ: **Transforming growth factor-beta and insulin-like signalling pathways in parasitic helminths**. *Int J Parasitol* 2002, **32**(4):399-404.
48. Ren P, Lim CS, Johnsen R, Albert PS, Pilgrim D, Riddle DL: **Control of C. elegans larval development by neuronal expression of a TGF-beta homolog**. *Science* 1996, **274**(5291):1389-1391.
49. Sze JY, Victor M, Loer C, Shi Y, Ruvkun G: **Food and metabolic signalling defects in a Caenorhabditis elegans serotonin-synthesis mutant**. *Nature* 2000, **403**(6769):560-564.
50. Gomez-Escobar N, van den Biggelaar A, Maizels R: **A member of the TGF-beta receptor gene family in the parasitic nematode Brugia pahangi**. *Gene* 1997, **199**(1-2):101-109.
51. Gomez-Escobar N, Gregory WF, Maizels RM: **Identification of tgh-2, a filarial nematode homolog of Caenorhabditis elegans daf-7 and human transforming growth factor beta, expressed in microfilarial and adult stages of Brugia malayi**. *Infect Immun* 2000, **68**(11):6402-6410.
52. Crook M, Thompson FJ, Grant WN, Viney ME: **daf-7 and the development of Strongyloides ratti and Parastrongyloides trichosuri**. *Mol Biochem Parasitol* 2005, **139**(2):213-223.
53. Knox DP: **Proteinase inhibitors and helminth parasite infection**. *Parasite Immunol* 2007, **29**(2):57-71.
54. Karanu FN, Rurangirwa FR, McGuire TC, Jasmer DP: **Haemonchus contortus: identification of proteases with diverse characteristics in adult worm excretory-secretory products**. *Exp Parasitol* 1993, **77**(3):362-371.

55. Kovaleva ES, Masler EP, Skantar AM, Chitwood DJ: **Novel matrix metalloproteinase from the cyst nematodes Heterodera glycines and Globodera rostochiensis**. *Mol Biochem Parasitol* 2004, **136**(1):109-112.
56. Yatsuda AP, Bakker N, Krijgsveld J, Knox DP, Heck AJ, de Vries E: **Identification of secreted cysteine proteases from the parasitic nematode Haemonchus contortus detected by biotinylated inhibitors**. *Infect Immun* 2006, **74**(3):1989-1993.
57. Clarke ND, Berg JM: **Zinc fingers in Caenorhabditis elegans: finding families and probing pathways**. *Science* 1998, **282**(5396):2018-2022.
58. Plowman GD, Sudarsanam S, Bingham J, Whyte D, Hunter T: **The protein kinases of Caenorhabditis elegans: a model for signal transduction in multicellular organisms**. *Proc Natl Acad Sci U S A* 1999, **96**(24):13603-13610.
59. Chalmers IW, McArdle AJ, Coulson RM, Wagner MA, Schmid R, Hirai H, Hoffmann KF: **Developmentally regulated expression, alternative splicing and distinct sub-groupings in members of the Schistosoma mansoni venom allergen-like (SmVAL) gene family**. *BMC Genomics* 2008, **9**:89.
60. Vermeire JJ, Cho Y, Lolis E, Bucala R, Cappello M: **Orthologs of macrophage migration inhibitory factor from parasitic nematodes**. *Trends Parasitol* 2008, **24**(8):355-363.
61. Cantacessi C, Campbell BE, Visser A, Geldhof P, Nolan MJ, Nisbet AJ, Matthews JB, Loukas A, Hofmann A, Otranto D *et al*: **A portrait of the "SCP/TAPS" proteins of eukaryotes--developing a framework for fundamental research and biotechnological outcomes**. *Biotechnol Adv* 2009, **27**(4):376-388.
62. Cantacessi C, Campbell BE, Young ND, Jex AR, Hall RS, Presidente PJ, Zawadzki JL, Zhong W, Aleman-Meza B, Loukas A *et al*: **Differences in transcription between free-living and CO2-activated third-stage larvae of Haemonchus contortus**. *BMC Genomics* 2010, **11**:266.
63. McCarter JP, Mitreva MD, Martin J, Dante M, Wylie T, Rao U, Pape D, Bowers Y, Theising B, Murphy CV *et al*: **Analysis and functional classification of transcripts from the nematode Meloidogyne incognita**. *Genome Biol* 2003, **4**(4):R26.
64. Hewitson JP, Harcus YM, Curwen RS, Dowle AA, Atmadja AK, Ashton PD, Wilson A, Maizels RM: **The secretome of the filarial parasite, Brugia malayi: proteomic profile of adult excretory-secretory products**. *Mol Biochem Parasitol* 2008, **160**(1):8-21.
65. McMahon SA, Miller JL, Lawton JA, Kerkow DE, Hodes A, Marti-Renom MA, Doulatov S, Narayanan E, Sali A, Miller JF *et al*: **The C-type lectin fold as an evolutionary solution for massive sequence variation**. *Nat Struct Mol Biol* 2005, **12**(10):886-892.
66. Loukas A, Maizels RM: **Helminth C-type lectins and host-parasite interactions**. *Parasitol Today* 2000, **16**(8):333-339.
67. McElwee JJ, Schuster E, Blanc E, Thomas JH, Gems D: **Shared transcriptional signature in Caenorhabditis elegans Dauer larvae and long-lived daf-2 mutants implicates detoxification system in longevity assurance**. *J Biol Chem* 2004, **279**(43):44533-44543.
68. Wolf DA, Jackson PK: **Cell cycle: oiling the gears of anaphase**. *Curr Biol* 1998, **8**(18):R636-639.
69. Leipe DD, Koonin EV, Aravind L: **STAND, a class of P-loop NTPases including animal and plant regulators of programmed cell death:**

multiple, complex domain architectures, unusual phyletic patterns, and evolution by horizontal gene transfer. *J Mol Biol* 2004, **343**(1):1-28.
70. Hartman D, Cottee PA, Savin KW, Bhave M, Presidente PJ, Fulton L, Walkiewicz M, Newton SE: **Haemonchus contortus: molecular characterisation of a small heat shock protein**. *Exp Parasitol* 2003, **104**(3-4):96-103.
71. Nagamune K, Moreno SN, Chini EN, Sibley LD: **Calcium regulation and signaling in apicomplexan parasites**. *Subcell Biochem* 2008, **47**:70-81.
72. Fujiwara RT, Cancado GG, Freitas PA, Santiago HC, Massara CL, Dos Santos Carvalho O, Correa-Oliveira R, Geiger SM, Bethony J: **Necator americanus infection: a possible cause of altered dendritic cell differentiation and eosinophil profile in chronically infected individuals**. *PLoS Negl Trop Dis* 2009, **3**(3):e399.
73. Estevez AO, Cowie RH, Gardner KL, Estevez M: **Both insulin and calcium channel signaling are required for developmental regulation of serotonin synthesis in the chemosensory ADF neurons of Caenorhabditis elegans**. *Dev Biol* 2006, **298**(1):32-44.
74. Roberts TM, Stewart M: **Acting like actin. The dynamics of the nematode major sperm protein (msp) cytoskeleton indicate a push-pull mechanism for amoeboid cell motility**. *J Cell Biol* 2000, **149**(1):7-12.
75. Buttery SM, Ekman GC, Seavy M, Stewart M, Roberts TM: **Dissection of the Ascaris sperm motility machinery identifies key proteins involved in major sperm protein-based amoeboid locomotion**. *Mol Biol Cell* 2003, **14**(12):5082-5088.
76. Strube C, Buschbaum S, Schnieder T: **Molecular characterization and real-time PCR transcriptional analysis of Dictyocaulus viviparus major sperm proteins**. *Parasitol Res* 2009, **104**(3):543-551.
77. Williamson AL, Brindley PJ, Knox DP, Hotez PJ, Loukas A: **Digestive proteases of blood-feeding nematodes**. *Trends Parasitol* 2003, **19**(9):417-423.
78. Williamson AL, Lecchi P, Turk BE, Choe Y, Hotez PJ, McKerrow JH, Cantley LC, Sajid M, Craik CS, Loukas A: **A multi-enzyme cascade of hemoglobin proteolysis in the intestine of blood-feeding hookworms**. *J Biol Chem* 2004, **279**(34):35950-35957.
79. Ranjit N, Zhan B, Stenzel DJ, Mulvenna J, Fujiwara R, Hotez PJ, Loukas A: **A family of cathepsin B cysteine proteases expressed in the gut of the human hookworm, Necator americanus**. *Mol Biochem Parasitol* 2008, **160**(2):90-99.
80. Hotez P, Haggerty J, Hawdon J, Milstone L, Gamble HR, Schad G, Richards F: **Metalloproteases of infective Ancylostoma hookworm larvae and their possible functions in tissue invasion and ecdysis**. *Infect Immun* 1990, **58**(12):3883-3892.
81. Williamson AL, Lustigman S, Oksov Y, Deumic V, Plieskatt J, Mendez S, Zhan B, Bottazzi ME, Hotez PJ, Loukas A: **Ancylostoma caninum MTP-1, an astacin-like metalloprotease secreted by infective hookworm larvae, is involved in tissue migration**. *Infect Immun* 2006, **74**(2):961-967.
82. Hotez PJ, Ashcom J, Zhan B, Bethony J, Loukas A, Hawdon J, Wang Y, Jin Q, Jones KC, Dobardzic A *et al*: **Effect of vaccination with a recombinant fusion protein encoding an astacinlike metalloprotease (MTP-1) secreted by host-stimulated Ancylostoma caninum third-stage infective larvae**. *J Parasitol* 2003, **89**(4):853-855.

83. Borchert N, Becker-Pauly C, Wagner A, Fischer P, Stocker W, Brattig NW: **Identification and characterization of onchoastacin, an astacin-like metalloproteinase from the filaria Onchocerca volvulus**. *Microbes Infect* 2007, **9**(4):498-506.
84. Skach WR: **The expanding role of the ER translocon in membrane protein folding**. *J Cell Biol* 2007, **179**(7):1333-1335.
85. Gems D, Ferguson CJ, Robertson BD, Nieves R, Page AP, Blaxter ML, Maizels RM: **An abundant, trans-spliced mRNA from Toxocara canis infective larvae encodes a 26-kDa protein with homology to phosphatidylethanolamine-binding proteins**. *J Biol Chem* 1995, **270**(31):18517-18522.
86. Maizels RM, Tetteh KK, Loukas A: **Toxocara canis: genes expressed by the arrested infective larval stage of a parasitic nematode**. *Int J Parasitol* 2000, **30**(4):495-508.
87. Loukas A, Hintz M, Linder D, Mullin NP, Parkinson J, Tetteh KK, Maizels RM: **A family of secreted mucins from the parasitic nematode Toxocara canis bears diverse mucin domains but shares similar flanking six-cysteine repeat motifs**. *J Biol Chem* 2000, **275**(50):39600-39607.
88. Daub J, Loukas A, Pritchard DI, Blaxter M: **A survey of genes expressed in adults of the human hookworm, Necator americanus**. *Parasitology* 2000, **120 (Pt 2)**:171-184.
89. De Maere V, Vercauteren I, Saverwyns H, Claerebout E, Berx G, Vercruysse J: **Identification of potential protective antigens of Ostertagia ostertagi with local antibody probes**. *Parasitology* 2002, **125**(Pt 4):383-391.
90. Blaxter M: **Caenorhabditis elegans is a nematode**. *Science* 1998, **282**(5396):2041-2046.
91. Sonnhammer EL, Durbin R: **Analysis of protein domain families in Caenorhabditis elegans**. *Genomics* 1997, **46**(2):200-216.
92. Saverwyns H, Visser A, Van Durme J, Power D, Morgado I, Kennedy MW, Knox DP, Schymkowitz J, Rousseau F, Gevaert K *et al*: **Analysis of the transthyretin-like (TTL) gene family in Ostertagia ostertagi--comparison with other strongylid nematodes and Caenorhabditis elegans**. *Int J Parasitol* 2008, **38**(13):1545-1556.
93. Nagaraj SH, Gasser RB, Ranganathan S: **Needles in the EST haystack: large-scale identification and analysis of excretory-secretory (ES) proteins in parasitic nematodes using expressed sequence tags (ESTs)**. *PLoS Negl Trop Dis* 2008, **2**(9):e301.
94. Jacob J, Vanholme B, Haegeman A, Gheysen G: **Four transthyretin-like genes of the migratory plant-parasitic nematode Radopholus similis: members of an extensive nematode-specific family**. *Gene* 2007, **402**(1-2):9-19.
95. Bruhn H: **A short guided tour through functional and structural features of saposin-like proteins**. *Biochem J* 2005, **389**(Pt 2):249-257.
96. Leippe M, Andra J, Nickel R, Tannich E, Muller-Eberhard HJ: **Amoebapores, a family of membranolytic peptides from cytoplasmic granules of Entamoeba histolytica: isolation, primary structure, and pore formation in bacterial cytoplasmic membranes**. *Mol Microbiol* 1994, **14**(5):895-904.
97. Andersson M, Gunne H, Agerberth B, Boman A, Bergman T, Sillard R, Jornvall H, Mutt V, Olsson B, Wigzell H *et al*: **NK-lysin, a novel effector**

peptide of cytotoxic T and NK cells. Structure and cDNA cloning of the porcine form, induction by interleukin 2, antibacterial and antitumour activity. *Embo J* 1995, **14**(8):1615-1625.
98. Zhai Y, Saier MH, Jr.: **The amoebapore superfamily**. *Biochim Biophys Acta* 2000, **1469**(2):87-99.
99. Don TA, Oksov Y, Lustigman S, Loukas A: **Saposin-like proteins from the intestine of the blood-feeding hookworm, Ancylostoma caninum**. *Parasitology* 2007, **134**(Pt 3):427-436.
100. Harcus Y, Nicoll G, Murray J, Filbey K, Gomez-Escobar N, Maizels RM: **C-type lectins from the nematode parasites Heligmosomoides polygyrus and Nippostrongylus brasiliensis**. *Parasitol Int* 2009, **58**(4):461-470.
101. Craig H, Wastling JM, Knox DP: **A preliminary proteomic survey of the in vitro excretory/secretory products of fourth-stage larval and adult Teladorsagia circumcincta**. *Parasitology* 2006, **132**(Pt 4):535-543.
102. Hotez PJ, Zhan B, Bethony JM, Loukas A, Williamson A, Goud GN, Hawdon JM, Dobardzic A, Dobardzic R, Ghosh K *et al*: **Progress in the development of a recombinant vaccine for human hookworm disease: the Human Hookworm Vaccine Initiative**. *Int J Parasitol* 2003, **33**(11):1245-1258.
103. Visser A, Van Zeveren AM, Meyvis Y, Peelaers I, Van den Broeck W, Gevaert K, Vercruysse J, Claerebout E, Geldhof P: **Gender-enriched transcription of activation associated secreted proteins in Ostertagia ostertagi**. *Int J Parasitol* 2008, **38**(3-4):455-465.
104. Bin Z, Hawdon J, Qiang S, Hainan R, Huiqing Q, Wei H, Shu-Hua X, Tiehua L, Xing G, Zheng F *et al*: **Ancylostoma secreted protein 1 (ASP-1) homologues in human hookworms**. *Mol Biochem Parasitol* 1999, **98**(1):143-149.
105. Moreno Y, Geary TG: **Stage- and gender-specific proteomic analysis of Brugia malayi excretory-secretory products**. *PLoS Negl Trop Dis* 2008, **2**(10):e326.
106. Murray J, Gregory WF, Gomez-Escobar N, Atmadja AK, Maizels RM: **Expression and immune recognition of Brugia malayi VAL-1, a homologue of vespid venom allergens and Ancylostoma secreted proteins**. *Mol Biochem Parasitol* 2001, **118**(1):89-96.
107. Schallig HD, van Leeuwen MA, Verstrepen BE, Cornelissen AW: **Molecular characterization and expression of two putative protective excretory secretory proteins of Haemonchus contortus**. *Mol Biochem Parasitol* 1997, **88**(1-2):203-213.
108. Yatsuda AP, Krijgsveld J, Cornelissen AW, Heck AJ, de Vries E: **Comprehensive analysis of the secreted proteins of the parasite Haemonchus contortus reveals extensive sequence variation and differential immune recognition**. *J Biol Chem* 2003, **278**(19):16941-16951.
109. Cass CL, Johnson JR, Califf LL, Xu T, Hernandez HJ, Stadecker MJ, Yates JR, 3rd, Williams DL: **Proteomic analysis of Schistosoma mansoni egg secretions**. *Mol Biochem Parasitol* 2007, **155**(2):84-93.
110. Curwen RS, Ashton PD, Sundaralingam S, Wilson RA: **Identification of novel proteases and immunomodulators in the secretions of schistosome cercariae that facilitate host entry**. *Mol Cell Proteomics* 2006, **5**(5):835-844.

111. Maizels RM, Gomez-Escobar N, Gregory WF, Murray J, Zang X: **Immune evasion genes from filarial nematodes**. *Int J Parasitol* 2001, **31**(9):889-898.
112. Hawdon JM, Jones BF, Hoffman DR, Hotez PJ: **Cloning and characterization of Ancylostoma-secreted protein. A novel protein associated with the transition to parasitism by infective hookworm larvae**. *J Biol Chem* 1996, **271**(12):6672-6678.
113. Bower MA, Constant SL, Mendez S: **Necator americanus: the Na-ASP-2 protein secreted by the infective larvae induces neutrophil recruitment in vivo and in vitro**. *Exp Parasitol* 2008, **118**(4):569-575.
114. Ghosh K, Hotez PJ: **Antibody-dependent reductions in mouse hookworm burden after vaccination with Ancylostoma caninum secreted protein 1**. *J Infect Dis* 1999, **180**(5):1674-1681.
115. Asojo OA, Goud G, Dhar K, Loukas A, Zhan B, Deumic V, Liu S, Borgstahl GE, Hotez PJ: **X-ray structure of Na-ASP-2, a pathogenesis-related-1 protein from the nematode parasite, Necator americanus, and a vaccine antigen for human hookworm infection**. *J Mol Biol* 2005, **346**(3):801-814.
116. Zhan B, Liu Y, Badamchian M, Williamson A, Feng J, Loukas A, Hawdon JM, Hotez PJ: **Molecular characterisation of the Ancylostoma-secreted protein family from the adult stage of Ancylostoma caninum**. *Int J Parasitol* 2003, **33**(9):897-907.
117. Tawe W, Pearlman E, Unnasch TR, Lustigman S: **Angiogenic activity of Onchocerca volvulus recombinant proteins similar to vespid venom antigen 5**. *Mol Biochem Parasitol* 2000, **109**(2):91-99.
118. Wang J, Kim SK: **Global analysis of dauer gene expression in Caenorhabditis elegans**. *Development* 2003, **130**(8):1621-1634.
119. Moore J, Tetley L, Devaney E: **Identification of abundant mRNAs from the third stage larvae of the parasitic nematode, ostertagia ostertagi**. *Biochem J* 2000, **347 Pt 3**:763-770.
120. Zhan B, Hotez PJ, Wang Y, Hawdon JM: **A developmentally regulated metalloprotease secreted by host-stimulated Ancylostoma caninum third-stage infective larvae is a member of the astacin family of proteases**. *Mol Biochem Parasitol* 2002, **120**(2):291-296.
121. Culley FJ, Brown A, Conroy DM, Sabroe I, Pritchard DI, Williams TJ: **Eotaxin is specifically cleaved by hookworm metalloproteases preventing its action in vitro and in vivo**. *J Immunol* 2000, **165**(11):6447-6453.
122. Blelloch R, Kimble J: **Control of organ shape by a secreted metalloprotease in the nematode Caenorhabditis elegans**. *Nature* 1999, **399**(6736):586-590.
123. Geldhof P, Claerebout E, Knox DP, Jagneessens J, Vercruysse J: **Proteinases released in vitro by the parasitic stages of the bovine abomasal nematode Ostertagia ostertagi**. *Parasitology* 2000, **121 Pt 6**:639-647.
124. Hawdon JM, Jones BF, Perregaux MA, Hotez PJ: **Ancylostoma caninum: metalloprotease release coincides with activation of infective larvae in vitro**. *Exp Parasitol* 1995, **80**(2):205-211.
125. Letunic I, Yamada T, Kanehisa M, Bork P: **iPath: interactive exploration of biochemical pathways and networks**. *Trends Biochem Sci* 2008, **33**(3):101-103.

126. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R *et al*: **New developments in the InterPro database**. *Nucleic Acids Res* 2007, **35**(Database issue):D224-228.
127. Rawlings ND, Barrett AJ: **Evolutionary families of metallopeptidases**. *Methods Enzymol* 1995, **248**:183-228.
128. Mohrlen F, Hutter H, Zwilling R: **The astacin protein family in Caenorhabditis elegans**. *Eur J Biochem* 2003, **270**(24):4909-4920.
129. Geldhof P, Visser A, Clark D, Saunders G, Britton C, Gilleard J, Berriman M, Knox D: **RNA interference in parasitic helminths: current situation, potential pitfalls and future prospects**. *Parasitology* 2007, **134**(Pt 5):609-619.
130. Kumar S, Chaudhary K, Foster JM, Novelli JF, Zhang Y, Wang S, Spiro D, Ghedin E, Carlow CK: **Mining predicted essential genes of Brugia malayi for nematode drug targets**. *PLoS One* 2007, **2**(11):e1189.
131. Foster JM, Zhang Y, Kumar S, Carlow CK: **Mining nematode genome data for novel drug targets**. *Trends Parasitol* 2005, **21**(3):101-104.

Figures

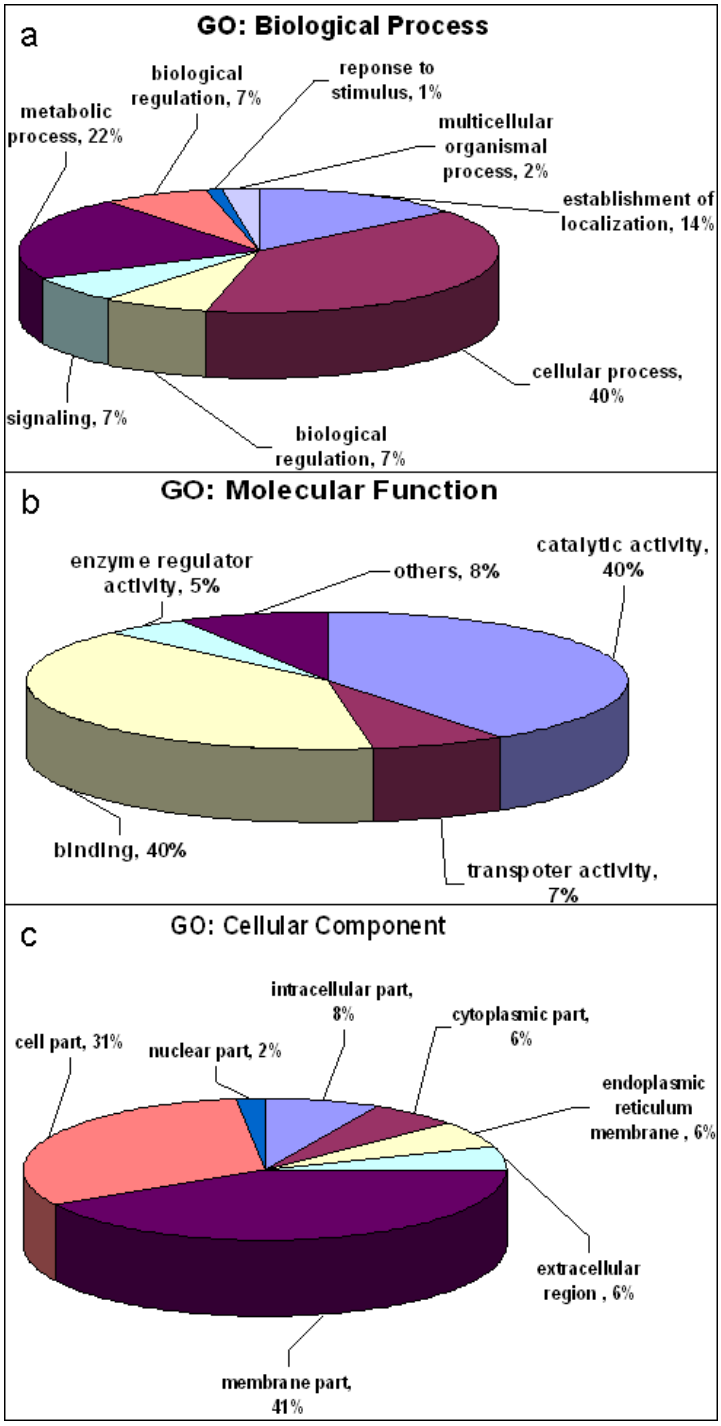


Figure 1: Putative excretory-secretory proteins and Gene Ontology (GO) terms identified. Percentages show the annotated categories a. Cellular Component, b. Molecular Function and c. Biological Process.

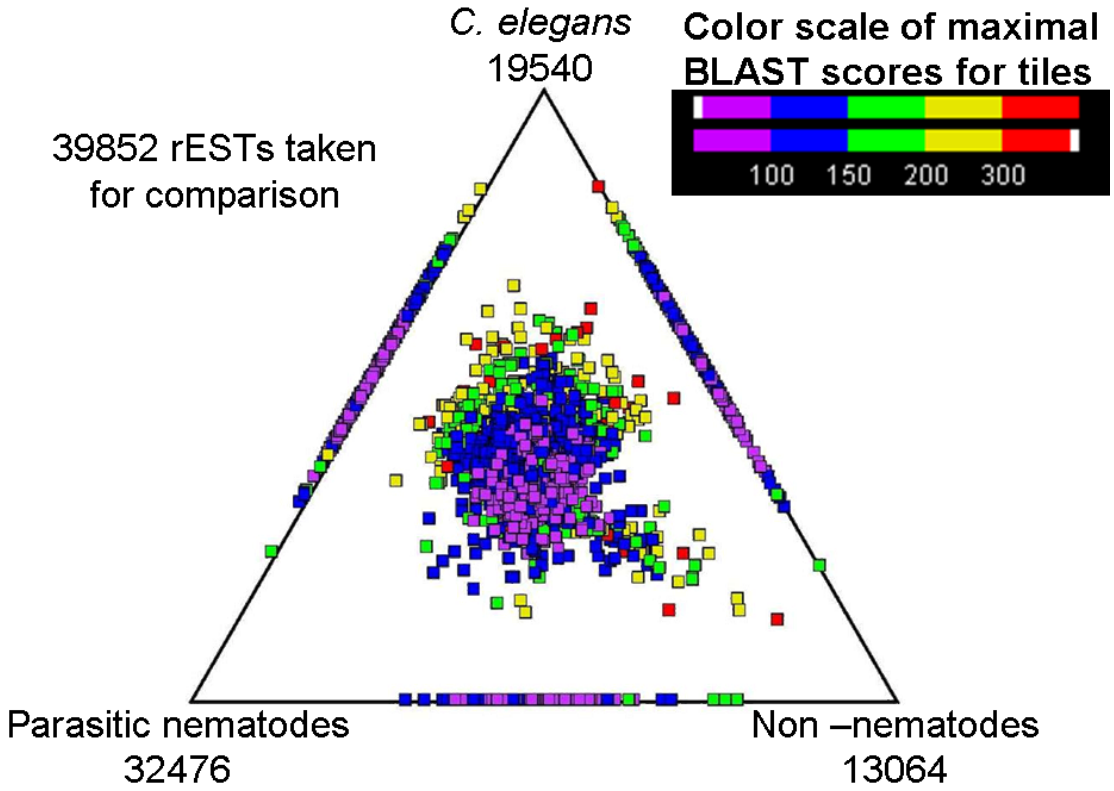


Figure 2: Comparison of *T. circumcincta* rESTs with *C. elegans*, other parasitic nematodes and organisms other than nematodes, from SimiTri analysis. The numbers at each vertex indicate rESTs matching that specific database.

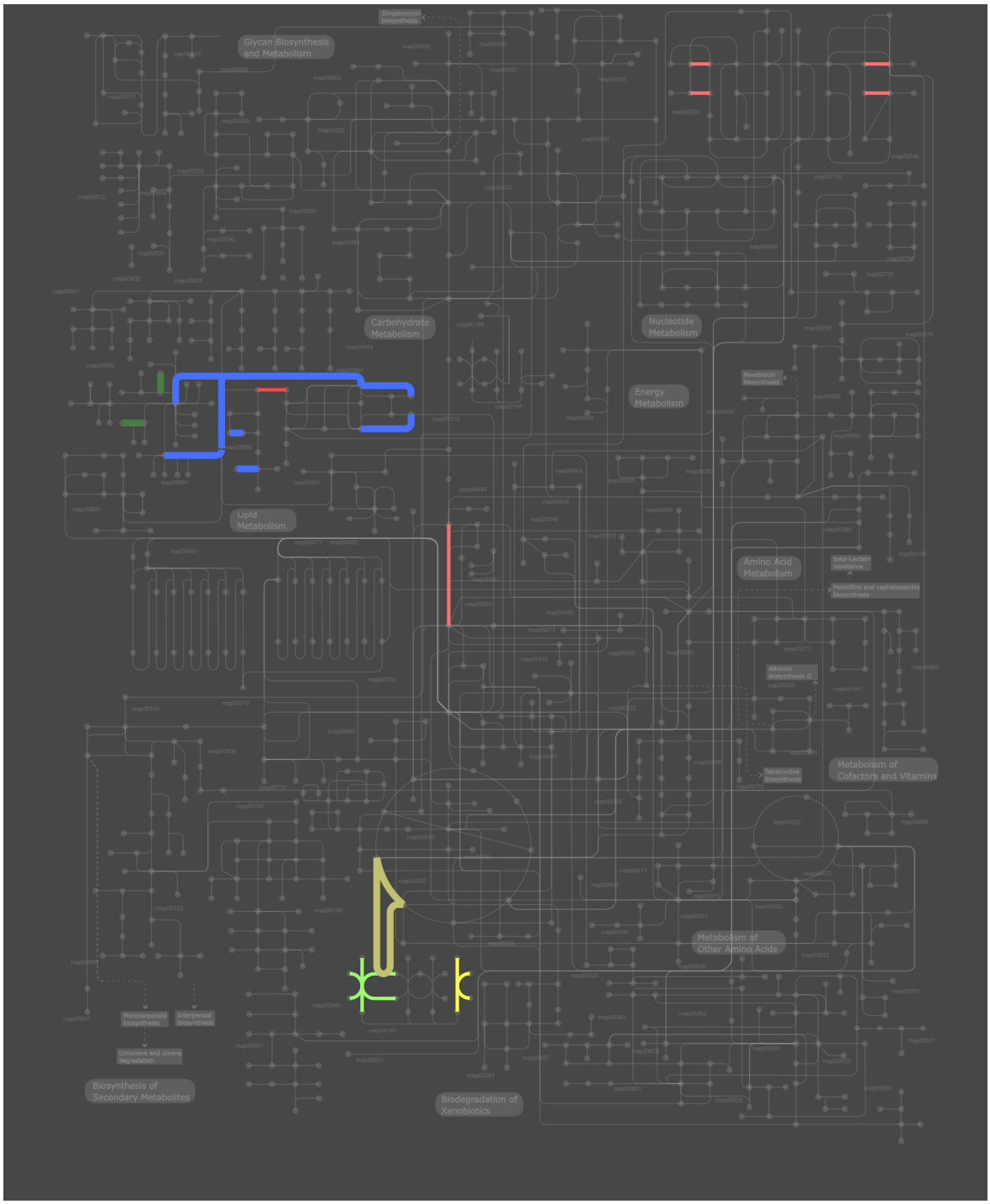


Figure 3: Biological pathways mapped using iPath tool for putative excretory-secretory proteins. The highlighted areas represent the pathways identified in the whole pathway.

Tables

Table 1: Preliminary analysis of the 407357 *T. circumcincta* ESTs .The contigs and singletons generated by preprocessing, overall representative ESTs (rESTs), peptides from conceptual translation and putative excretory-secretory (E/S) proteins identified are shown.

T. circumcincta ESTs	**Numbers (percentage)**
Raw sequences obtained	407357
Cleaned sequences	366897 (90.06)
Clusters of multiple sequences (contigs)	22382 (5.4)
Clusters of singletons	17470 (4.2)
Total rESTs	39852 (9.7)
Putative peptides	24013 (60.25 % rESTs)
E/S proteins (cut-off: 0.5)	112

Table 2: Top 30 metabolic pathways mapped by Kyoto Encyclopedia of Genes and Genomes in *T. circumcincta* protein sequences

KEGG PATHWAY	SEQUENCE COUNT
Peptidases	254
Ribosome	220
Oxidative phosphorylation	187
Other enzymes	168
Chaperones and folding catalysts	144
Cytoskeleton proteins	109
Protein kinases	108
Purine metabolism	102
Translation factors	96
Ubiquitin enzymes	90
Proteasome	89
Starch and sucrose metabolism	86
Pyruvate metabolism	86
Glycolysis / Gluconeogenesis	83
Fatty acid metabolism	83
Lysine degradation	78
Valine, leucine and isoleucine degradation	76
Tryptophan metabolism	72
Aminoacyl-tRNA biosynthesis	69
Insulin signaling pathway	68
GTP-binding proteins	68
Citrate cycle (TCA cycle)	68
Regulation of actin cytoskeleton	65
Propanoate metabolism	64
Cell cycle	64
Carbon fixation	64
Focal adhesion	62
Ubiquitin mediated proteolysis	60
Fructose and mannose metabolism	60
Butanoate metabolism	59

Table 3: Top 30 metabolic pathways mapped by KAAS in *T. circumcincta* protein sequences

KEGG PATHWAY	PROTEINS
Huntington's disease	91
Oxidative phosphorylation	84
Ribosome	80
Spliceosome	79
Alzheimer's disease	72
Parkinson's disease	70
Purine metabolism	56
Pyrimidine metabolism	51
Cell cycle	34
Ubiquitin mediated proteolysis	33
Proteasome	33
Lysosome	33
Endocytosis	32
Cell cycle - yeast	31
Peroxisome	30
Glycolysis / Gluconeogenesis	29
Pathways in cancer	28
Aminoacyl-tRNA biosynthesis	28
DNA replication	26
Valine, leucine and isoleucine degradation	25
Regulation of actin cytoskeleton	25
Citrate cycle (TCA cycle)	25
Vibrio cholerae infection	23
Fatty acid metabolism	23
Amino sugar and nucleotide sugar metabolism	23
RNA degradation	22
Nucleotide excision repair	21
Lysine degradation	21
RNA polymerase	20
Meiosis - yeast	20

Table 4: Top 30 domain description for the protein sequences

Description	InterProscan ID	Protein sequences
Protein kinase-like domain	IPR011009	138
SCP-like extracellular	IPR014044	126
NAD(P)-binding domain	IPR016040	96
Allergen V5/Tpx-1 related	IPR001283	95
Transthyretin-like	IPR001534	88
C-type lectin fold	IPR016187	85
C-type lectin	IPR001304	78
C-type lectin-like	IPR016186	71
Nucleotide-binding, alpha-beta plait	IPR012677	71
Serine/threonine-protein kinase-like domain	IPR017442	69
Metridin-like ShK toxin	IPR003582	67
RNA recognition motif, RNP-1	IPR000504	64
Peptidase C1A, papain	IPR013128	59
Thioredoxin-like fold	IPR012336	57
WD40 repeat, subgroup	IPR019781	56
WD40 repeat-like-containing domain	IPR011046	56
WD40/YVTN repeat-like-containing domain	IPR015943	54
Thioredoxin fold	IPR012335	53
Pyridoxal phosphate-dependent transferase, major domain	IPR015424	52
Heat shock protein Hsp20	IPR002068	51
Protein-tyrosine phosphatase, receptor/non-receptor type	IPR000242	50
EF-hand-like domain	IPR011992	49
Peptidase A1	IPR001461	48
Tyrosine-protein kinase	IPR020685	47
Peptidase C1A, papain C-terminal	IPR000668	47
Peptidase aspartic	IPR021109	45
Short-chain dehydrogenase/reductase SDR	IPR002198	45

Table 5: Top 30 selected metabolic pathways in excretory-secretory proteins mapped using KEGG database

KEGG PATHWAY	ES Proteins
Protein kinases	3
Oxidative phosphorylation	3
Long-term depression	3
Glycerophospholipid metabolism	3
Arachidonic acid metabolism	3
VEGF signaling pathway	2
Purine metabolism	2
Protein folding and associated processing	2
Peptidases	2
MAPK signaling pathway	2
Linoleic acid metabolism	2
GnRH signaling pathway	2
Glycolysis / Gluconeogenesis	2
Glutathione metabolism	2
Fc epsilon RI signaling pathway	2
Ether lipid metabolism	2
Cytoskeleton proteins	2
CAM ligands	2
alpha-Linolenic acid metabolism	2
Wnt signalling pathway	1
Urea cycle and metabolism of amino groups	1
Ubiquitin mediated proteolysis	1
Ubiquitin enzymes	1
Tyrosine metabolism	1
Type II diabetes mellitus	1
Translation factors	1
Transcription factors	1
Tight junction	1
TGF-beta signaling pathway	1
Signal transduction mechanisms	1

Other enzymes	4

Table 6: Pathway Analysis of secreted proteins using KAAS

KEGG PATHWAY	ES Proteins
Glycerophospholipid metabolism	3
Oxidative phosphorylation	3
Vascular smooth muscle contraction	3
Long-term depression	3
Arachidonic acid metabolism	3
Alzheimer's disease	3
Wnt signaling pathway	2
VEGF signaling pathway	2
Tight junction	2
TGF-beta signaling pathway	2
Parkinson's disease	2
Oocyte meiosis	2
Meiosis - yeast	2
MAPK signaling pathway	2
Lysosome	2
Linoleic acid metabolism	2
Huntington's disease	2
GnRH signaling pathway	2
Glutathione metabolism	2
Fc epsilon RI signaling pathway	2
Ether lipid metabolism	2
Cell cycle - yeast	2
Axon guidance	2
alpha-Linolenic acid metabolism	2
Pyruvate metabolism	1
Glycolysis / Gluconeogenesis	1
Carbon fixation in photosynthetic organisms	1
Citrate cycle	1
Vibrio cholerae infection	1
Ubiquitin mediated proteolysis	1

Table 7: Top 20 protein families of known function found in excretory-secretory proteins

Description	ES sequences	Type	Interproscan ID
Metridin-like ShK toxin	14	Domain	IPR003582
Transthyretin-like	11	Family	IPR001534
SCP-like extracellular	7	Domain	IPR014044
Saposin-like	7	Domain	IPR011001
C-type lectin	7	Domain	IPR001304
C-type lectin fold	6	Domain	IPR016187
C-type lectin-like	6	Domain	IPR016186
Proteinase inhibitor I2, Kunitz metazoa	5	Domain	IPR002223
Protein kinase-like domain	4	Domain	IPR011009
Major facilitator superfamily, general substrate transporter	4	Domain	IPR016196
Destabilase	3	Family	IPR008597
Allergen V5/Tpx-1 related	3	Family	IPR001283
Tyrosine-protein kinase	3	Region	IPR020685
Phospholipase A2	2	Family	IPR016090
Thioredoxin-like fold	2	Domain	IPR012336
Thioredoxin fold	2	Domain	IPR012335
Globin	2	Domain	IPR012292
Serine/cysteine peptidase, trypsin-like	2	Domain	IPR009003
Saposin B	2	Domain	IPR008139
Protein of unknown function DUF148	2	Domain	IPR003677

Table 8: Top 30 Pathway analysis of secreted proteins obtained from SecretomeP

KEGG PATHWAY	ES Proteins
Translation factors	6
Oxidative phosphorylation	4
Cell cycle	4
Regulation of actin cytoskeleton	3
Protein kinases	3
Progesterone-mediated oocyte maturation	3
Peptidases	3
DNA polymerase	3
Chaperones and folding catalysts	3
Ubiquitin mediated proteolysis	2
Ubiquitin enzymes	2
Transcription factors	2
Tight junction	2
RNA polymerase	2
Ribosome	2
Reductive carboxylate cycle (CO2 fixation)	2
Pyruvate metabolism	2
Pores ion channels	2
mTOR signaling pathway	2
MAPK signaling pathway	2
Glutathione metabolism	2
General function prediction only	2
Gap junction	2
Fatty acid metabolism	2
Fatty acid biosynthesis	2
Cytoskeleton proteins	2
Citrate cycle (TCA cycle)	2
Cell cycle - yeast	2
Arginine and proline metabolism	2
Other enzymes	2

Additional Files

Additional File 1: Comparison of rESTs from *Teladorsagia circumcincta* with *C. elegans*.

Additional File 2: *Teladorsagia circumcincta* homologues in Clade V of the phylum Nematoda, comprising *Haemonchus contortus*, *Necator americanus*, *Nippostrongylus brasiliensis*, *Ostertagia ostertagi*, *Pristionchus pacificus*, *Ancylostoma caninum*, *Ancylostoma ceylanicum* and *Dictyocaulus viviparus*.

Additional File 3: Metabolic pathways in *Teladorsagia circumcincta* mapped by Kyoto Encyclopedia of Genes and Genomes.

Additional File 4: GO Annotation for proteins.

Additional File 5: Secreted proteins predicted from rESTs from *Teladorsagia circumcincta* and their homologues and RNAi phenotypes.

Additional File 6: Secretory proteins predicted from *Teladorsagia circumcincta* rESTs - comparison with proteomic data.

Additional File 7: Transmembrane proteins predicted from rESTs from *Teladorsagia circumcincta* and their homologues and RNAi phenotypes.

Additional File 8: GO Annotation for secreted proteins.

Additional File 9: Pathway Analysis of secreted proteins using KOBAS.

Additional File 10: Pathway Analysis of secreted proteins using KAAS.

Additional File 11: Pathway analysis of secreted proteins obtained from SecretomeP.

Additional File 12: Protein Domain Analysis in *Teladorsagia circumcincta*.

Additional File 13: Top NR description of *Teladorsagia circumcincta* protein sequences.

Additional File 14: InterProScan analysis :representative protein domains/families.

Additional File 15: Comparison of secreted proteins from *Teladorsagia circumcincta* with *C. elegans*.

4.2 Conclusions

The study could identify number of pathways for immune response and was also successful in benchmarking different bioinformatics tools for EST analysis which could help in the development of robust computational EST analysis pipelines. The incorporation of SecretomeP increased the number of ES proteins predicted while the benchmarking results between KOBAS and KAAS clearly indicating that both programs are equally proficient in predicting biochemical pathways. The results of this large-scale analysis provide a step towards future research with the focus on disease and molecular biology of the parasite and could also be integrated with proteomic and metabolomic studies for identifying novel disease control strategies.

The bioinformatics improvements by way of predicting non-classically secreted proteins and complete functional annotation for all putative peptides (not ES proteins alone as offered by EST2Secretome [162]) have now been incorporated into a generic transcriptome analysis pipeline, described in the next chapter.

Chapter 5: TranSeqAnnotator: Large-scale analysis of transcriptomic data

5.1 Summary

A user-friendly robust transcriptome analysis pipeline was developed extending the existing ESTExplorer [160] and EST2Secretome [162] pipelines with advanced features and updated tools based on the review as discussed in paper 1, and the results of papers 3 and 4.

The new pipeline incorporates the results of the benchmarking carried out in Chapter 4 (Paper 3). Furthermore, a detailed evaluation of the performance of this pipeline was carried out for a large-scale NGS dataset. The pipelines developed previously processed only ESTs and failed to address the non-classically secreted proteins. TranSeqAnnotator accepts short read fasta sequences along with EST sequences for pre-processing, assembly, conceptual translation and blast against NR followed by annotation with identification of putative ES proteins based on the presence of signal sequences and similarity searches using Blast. Non-classically secreted proteins are identified using the tool SecretomeP along with the absence of transmembrane helices. The pipeline also looks for the combined annotation, comprising a suite of bioinformatic tools to annotate the protein sequence and ES proteins inferred in different phases with GO ontology from Interproscan and pathway analysis with KOBAS for all sequences. The hardware issues like CPU memory and storage capacity are addressed in the new pipeline. Programs for EST/short read pre-processing, assembly and annotation at DNA, protein, and secretome levels were carefully selected and also, applied the *Ascaris lumbricoides* dataset as an example of the pipeline workflow.

TranSeqAnnotator: Large-scale analysis of transcriptomic data

Ranjeeta Menon[1], Gagan Garg[1], Robin B. Gasser[2] and Shoba Ranganathan[1, 3,*]

[1]*Department of Chemistry and Biomolecular Sciences and ARC Centre of Excellence, Macquarie University, Sydney, NSW 2109, Australia;* [2]*Department of Veterinary Sciences, The University of Melbourne, Werribee, VIC 3030, Australia;* [3]*Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117597.*

* To whom correspondence

Email addresses:

RM: ranjeeta.menon@mq.edu.au

GG: gagan.garg@mq.edu.au

RBG: robinbg@unimelb.edu.au

SR: shoba.ranganathan@mq.edu.au

ABSTRACT

Background

The transcriptome of an organism can be studied with the analysis of expressed sequence tag (EST) data sets that offers a rapid and cost effective approach with several new and updated bioinformatics approaches and tools for assembly and annotation. The comprehensive analyses comprehend an organism along with the genome and proteome analysis. With the advent of large-scale sequencing projects and generation of sequence data at protein and cDNA levels, automated analysis pipeline is necessary to store, organize and annotate ESTs.

Results

TranSeqAnnotator is a workflow for large-scale analysis of transcriptomic data with the most appropriate bioinformatics tools for data management and analysis. The pipeline automatically cleans, clusters, assembles and generates consensus sequences, conceptually translates these into possible protein products and assigns putative function based on various DNA and protein similarity searches. Excretory/secretory (ES) proteins inferred from ESTs/short reads are also identified. The TranSeqAnnotator accepts FASTA format raw and quality ESTs along with protein and short read sequences and are analysed with user selected programs. After pre-processing and assembly, the dataset is annotated at the nucleotide, protein and ES protein levels.

Conclusion

TranSeqAnnotator has been developed in a Linux cluster, to perform an exhaustive and reliable analysis and provide detailed annotation. TranSeqAnnotator outputs gene ontologies, protein functional identifications in terms of mapping to protein domains and metabolic pathways. The pipeline is applied to annotate large EST datasets to identify several novel and known genes with therapeutic experimental validations and could serve

as potential targets for parasite intervention. TransSeqAnnotator is freely available for the scientific community at http://estexplorer.biolinfo.org/TranSeqAnnotator/.

BACKGROUND

Expressed sequence tags or ESTs, derived from complementary DNA (cDNA) libraries provide a low-cost transcriptomic alternative to whole genome sequencing as these are short, unedited, randomly selected single-pass sequence reads of approximately 200-800 base pairs (bp) which represent a small region or a part of nucleotide sequence from a transcribed protein coding or non-coding messenger mRNA. They play vital role in gene identification and verification of gene prediction as they represent the expressed region of a genome. The analysis of EST data can facilitate gene discovery, help in gene structure identification, complement genome annotation, establish the viability of alternative transcripts, direct single nucleotide polymorphism (SNP) characterization and facilitate proteomic exploration [1-3]. They were used as the primary source for human gene discovery in early 1990s [4]. Besides ESTs, millions of sequencing reads of 35-250 bp are generated with the advent of "next-generation" sequencing (NGS) which further help in the study of transcriptome data mainly for neglected organisms and also, understanding different isoforms of an organism at different stages of development. Studies using experimental proteomic approach have shown the identification of proteins in ESP with transcriptome assembly [5]. Many challenges are faced in the areas of bioinformatics analysis in data storage and management solution and developing informatics tools for analysis with the focus on sequence quality scoring, alignment, assembly, and data processing with the advent of short read strategy of NGS [6, 7]. A comprehensive analysis pipeline is required to store, organize and annotate ESTs with several computational tools for pre-processing, clustering, assembly into contiguous segments known as contigs and annotation to yield biological information. The web resources available were reviewed for large-scale EST dataset at each step including clustering, assembly, consensus generation

and tools for DNA, protein and ES annotation [8]. A number of analysis steps and tools confounded computational strategies to organize and analyse transcriptomic dataset [9] which is compounded by the ability of some tools to handle high-throughput EST data. An evaluation revealed that all available platforms terminated prior to downstream functional annotation, including gene ontologies (GOs), motif/pattern analysis and pathway mapping. Hence, the establishment of a comprehensive large-scale transcriptomic analysis pipeline [9] was required to be developed to keep up with the rapidity with which enormous amounts of sequence data are currently being generated. An urgent need for advanced, high-throughput computational analyses of EST and genomic sequence datasets using automated platforms is highlighted. EST data are been applied to study of functional biomolecules [9, 10] but, predicting ES proteins, from ESTs have been uncommon. Excretory/Secretory (ES) products are the molecules excreted or secreted by a cell or an organism that can circulate throughout the body of an organism (e.g., in the extracellular space) or are localized to or released from the cell surface, making them readily accessible to drugs and/or the immune system. ES products cover 8±20% of the proteome of an organism [11] and include molecules of varied functionality, including chemokines, digestive enzymes cytokines, hormones, toxins, antibodies, morphogens, extracellular proteinases and antimicrobial peptides. They are known to be involved in vital biological processes, including cell adhesion, cell migration, cell–cell communication, differentiation, proliferation, morphogenesis and immune responses [12]. Biochemical and immunological studies of parasitic helminths were focussed on ES proteins. Worms secrete biologically active mediators which can transform or customize their niche within the host [13-15] to regulate or to elude immune attack or stimulate a particular host response.

Some platforms terminate at the assembly level, providing contigs and singletons [16] (referred to as rESTs) while other platforms exclusively run nucleotide-based programs with limited annotation at the protein level [17-20]. Based on the benchmarking results, a

robust transcriptome analysis pipeline (TranSeqAnnotator) is constructed with contig generation from ESTs and short reads, updated pathway analysis, non-classically secreted protein identification and extensive annotation with an option to select specific analysis phases by users (detailed below). Proteins secreted by classical and non-classical pathways are identified by a combination of computational approaches to predict ESPs. The pipeline accepts ESTs, quality values, protein sequences and short reads as input and provides as output, assembled rESTs and their annotations including gene ontologies, secretory proteins, mapping to protein domains, motifs, metabolic pathways and interaction databases. TranSeqAnnotator (TSA) is available as web service and can be downloaded for local installation.

Implementation

TranSeqAnnotator workflow has three phases with Phase I (a) for EST or (b) short read fasta sequence pre-processing, assembly, conceptual translation and blast against NR, Phase II for the identification of putative ES proteins, from classically and non-classically secreted proteins and the elimination of transmembrane proteins and Phase III for the combined annotation of the protein sequence and ES proteins involving a carefully selected suite of bioinformatic tools, based on a large-scale transcriptome analysis [21] (Figure 1). TranSeqAnnotator currently implements the genetic codes for 15 organisms, covering the most studied organisms, including human, rat, pig, dog, chicken, rice, wheat, thale cress (*Arabidopsis thaliana*), zebrafish, yeast and a free-living roundworm (*Caenorhabditis elegans*).

Phase I accept ESTs and short reads as well as quality values in the case of ESTs as input for pre-processing and assembly (Figure 1).

The sequence cleaning step uses seqclean [22] and seqtrim [23] with ESTs alone and with ESTs and quality sequences respectively followed by masking the repeats using RepeatMasker [24] which is optional. The Phase I (b) accepts short reads and pre-processing is carried out using seqclean. The masked sequences are then passed on for clustering and assembly with iAssembler (http://bioinfo.bti.cornell.edu/tool/iAssembler/) which incorporates MIRA [25] and CAP3 assemblers for ESTs and short reads. For conceptual translation into proteins, the program ESTScan [26] applies the genetic code from the nearest organism to the contig and singleton sequences generated by CAP3 or iAssembler.

In Phase II, the protein sequences generated in Phase I, using TMHMM [27] and putative ES proteins identified using SecretomeP [28] are annotated (Figure 1). Firstly, the signal sequence is checked with SignalP while, SecretomeP looks for non-classically secreted proteins and the hidden Markov model probability scores (SignalPNN and SignalP-HMM), using default parameters that can be modified by experienced users. Subsequently, all proteins with signal sequences are passed on to TMHMM, a hidden Markov model-based transmembrane helix prediction program, to ‘‘filter out’’ of transmembrane proteins. ES proteins, the subset lacking transmembrane helices are further annotated. Phase III, the annotation level for protein sequences or ES proteins comprises a suite of computational tools InterProScan [29] for domain analysis and Gene Ontology, pathway mapping using KOBAS (KEGG Orthology-Based Annotation System) [30. 31]. Also, protein BLAST is employed to search databases derived from Wormpep [32] for locating nematode homologues and a list of homologous proteins in *C. elegans*, archived in WormBase as well as interaction databases like IntAct [33], BioGrid [34] and DIP [35] which give information on molecular interaction data and experimentally verified protein-protein interactions.

TSA accepts a dataset submitted by the user and optional programs can be selected as required (Figure 2). The progress of the analysis is monitored on the status page which is updated after each selected process is completed and the output of each program is available along with a summarized output. Some of these tools are provided in the ESTExplorer [36] and EST2Secretome [37] pipeline but, the analysis of large-scale EST dataset and short read sequences with updated bioinformatics tools is incorporated with TranSeqAnnotator as part of the benchmarking with the large-scale analysis of *Teladorsagia circumcincta* dataset (unpublished work). Also, the program SecretomeP showed the identification of important proteins which the previous pipelines failed to identify with SignalP. The identification of both classically and non-classically secreted proteins with secretomeP is the highlight of the robust analysis pipeline as our earlier analysis on *Fasciola hepatica* [38].

SOFTWARE/HARDWARE ENVIRONMENT

TranSeqAnnotator is developed using PERL v5.10.0 which links the different bioinformatics programs and MySQL as backend for data management and analysis. The front end is developed using PHP and the processes are run based on CPU availability. Each input sequence submitted by the user is tagged with a request ID to trace the process. The pipeline runs on a 16-node Linux cluster (2.4GHz, Intel(R) Xeon (R) CPU, 16 Processors, 32 GB RAM) running on ubuntu server operating system. The output files for viewing and downloading are provided as final results which are available for a week.

Results and Discussion

Application of TranSeqAnnotator

Ascaris lumbricoides, the soil-transmitted helminths or geohelminths is the largest common intestinal nematode parasites of human that causes the disease ascariasis [39]. It infects an estimated 1.2 billion people worldwide, but is usually asymptomatic [40]. 1822 *A. lumbricoides* EST sequences from dbEST [41], were analysed using the

TranSeqAnnotator. The dataset is from the adult male whole body *Ascaris lumbricoides* cDNA clone. The phase I of pre-processing (SeqClean and RepeatMasker) aligned/clustered using CAP3 followed by assembly, was carried out which yielded 236 contigs and 658 singletons. These rESTs were mapped to the non-redundant (NR) dataset using BLAST, for nucleotide level annotation. Using a translational matrix, ESTScan conceptually translates these high quality rESTs, which are then transferred to Phase II of TSA, for the prediction of ES proteins, by sequentially running SecretomeP (with a threshold value for the NN-score of 0.9) and TMHMM programs. The cluster dataset, translated peptide sequences and ES proteins were annotated with biochemical pathways, employing KOBAS, domain/family motif and GeneOntology using InterProScan. The query sequences were compared using BLASTP against Wormpep [32] and against the IntAct database (version 1.7.0) to extract all interaction partners. The 894 rESTS were conceptually translated to yield 510 peptide sequences. The GO terms were identified for these putative protein sequences using InterProScan, with 108 peptide sequences assigned biological process (BP), 156 associated with molecular function (MF) and 83 as part of a cellular component (CC) (Additional File 1). The analysis revealed that *translation* (GO:0006412) and *oxidation-reduction process* (GO:0055114) were the highly represented GO categories signifying biological processes. The major number of GO terms in molecular function was *structural constituent of ribosome* (GO:0003735), *oxidoreductase activity* (GO:0016491) and *ATP binding* (GO:0005524) whereas in cellular component, the highly represented GO terms were *ribosome* (GO:0005840) and *extracellular space* (GO:0005615).

A total of 239 peptide sequences were mapped to 113 KEGG pathways using KOBAS. The main KEGG pathways mapped included *ribosomal protein assembly pathway* (n=34) and *cytoskeleton proteins* (n=19). Other well represented pathways include *tight junction*

(n=14), *regulation of actin cytoskeleton* (n=12), *focal adhesion* (n=12), *valine, leucine and isoleucine degradation* (n=8) and *propanoate metabolism* (n=7). Peptides were mapped to several pathways, including *glycolysis/gluconeogenesis, MAPK signaling pathway* and *ubiquitin mediated proteolysis* (Additional File 2).

Domain mapping by Interproscan provides details as to the family, fold and functional domains present in the putative peptides. The most represented was the *collagen triple helix repeat of proteins*, comprising 14 protein entries, followed by *C-type lectin fold* and *transthyretin-like family*, with nine protein entries each. Other highly represented domains are the *actin-like* and *C-type lectin* (Additional File 3).

A total of 32 were predicted by SecretomeP. Of these, 6 are classically secreted peptides; with N-terminal signal sequences while 26 are non-classical, supporting the use of SecretomeP vs. SignalP alone, which can only predict classically secreted proteins. Of these 32, six proteins with transmembrane helices, predicted by TMHMM were eliminated, resulting in 26 excreted/secreted proteins inferred from the present dataset of 894 rESTs. We could identify cecropin (including the cecropin-P1, cecropin-P2, cecropin-P3), cathepsin L from *Ascaris suum* and cathepsin L-like protease from *Strongylus vulgaris*, chymotrypsin/elastase isoinhibitor 1 from *Ascaris suum*, C-type lectin protein 160 from *Ascaris suum* and C-type lectin domain-containing protein 160 from *Ascaris suum*. Gelsolin from *Ascaris suum* and GelSoliN-Like family member (gsnl-1) from *Caenorhabditis elegans* were also identified (Additional File 4). Cecropins, represent a large family of antibacterial and toxic peptides are known to execute host defence functions mainly against micro-organisms [42, 43] and are found in insects [44]. Ascaris cecropins (P1–P4) were identified as antimicrobial peptides that were positively inducible by bacterial injection. Ascaris cecropins synthesized chemically were bactericidal against a wide range of microbes, i.e. Gram-positive (*Staphylococcus aureus*, *Bacillus subtilis* and

Micrococcus luteus) and Gram-negative (*Pseudomonas aeruginosa*, *Salmonella typhimurium*, *Serratia marcescens* and *Esherichia coli*) bacteria, and were weakly but detectably active against yeasts (*Saccharomyces cerevisiae* and *Candida albicans*) [45]. A large family of proteins that binds carbohydrate moieties in a Ca2+-dependent manner are represented by C-type lectins (CTLs) which act as a pathogen recognition molecule or an antibacterial protein in immune responses to protect the worm itself against microbial infection [46-49]. They also play vital role in immune homeostasis by endogenous 'self' ligand recognition [50], and they themselves have a bactericidal activation [51]. Studies have shown that *A. suum* C-type lectin-1(As-CTL-) shows high similarity to *Toxocara canis* C-type lectin (Tc-CTLs) and are exposed to attack by host immune responses. Hence, to avoid protective immune responses in infected animals during tissue migration *A. suum* larvae might interfere with host inflammation processes by As-CTL-1 [52]. The Gelsolin family belongs to a group of actin binding proteins are known to be involved in cell structure, motility, apoptosis, amyloidosis and cancer. Gelsolin-like protein-1 (GSNL-1) from *C. elegans* is a new member of the gelsolin family of actin regulatory proteins which provide new insight into functional diversity and evolution of gelsolin-related proteins [53, 54]. We were able to functionally assign GO terms to 26 putative ES proteins with proteolysis (GO:0006508) the most common GO category representing biological processes, cysteine-type peptidase activity (GO:0008234) in molecular function and extracellular region (GO:0005576) in cellular component. Protein processing in endoplasmic reticulum, phagosome, lysosome, antigen processing and presentation, rheumatoid arthritis represented the sequences mapped to KEGG pathways using KOBAS. The TranSeqAnnotaor methodology was benchmarked using the large-scale dataset of *Teladorsagia circumcincta* (unpublished work) and applied for the annotation of *A. lumbricoides*.

FUTURE DIRECTIONS

TranSeqAnnotator currently supports nucleotide, short reads, protein and ES level annotation. Our aim is to extend the pipeline with updating the masking the repeats with repeatless libraries to annotate newly sequenced organisms and also to carry out annotations for different datasets like RNA-seq, microarray datasets.

AUTHORS' CONTRIBUTIONS

RM carried out the analysis, computational studies and drafted the manuscript. RM, GG, SR and RBG participated in the design of the study and interpretation of data. SR and RBG conceived the project and finalized the manuscript. All authors have read and approved the final manuscript.

Conflict of interest: none declared.

ACKNOWLEDGEMENTS

We are grateful to Macquarie University for the award of postgraduate research scholarships. Funding to pay the Open Access publication charges for this article was provided by Macquarie University.

REFERENCES

1. Rudd S: **Expressed sequence tags: alternative or complement to whole genome sequences?** *Trends Plant Sci* 2003, **8**(7):321-329.

2. Dong Q, Kroiss L, Oakley FD, Wang BB, Brendel V: **Comparative EST analyses in plant systems**. *Methods Enzymol* 2005, **395**:400-418.

3. Jongeneel CV: **Searching the expressed sequence tag (EST) databases: panning for genes**. *Brief Bioinform* 2000, **1**(1):76-92.

4. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF *et al*: **Complementary DNA sequencing: expressed sequence tags and human genome project**. *Science* 1991, **252**(5013):1651-1656.

5. Moreno Y, Gros PP, Tam M, Segura M, Valanparambil R, Geary TG, Stevenson MM: **Proteomic analysis of excretory-secretory products of Heligmosomoides polygyrus assessed with next-generation sequencing transcriptomic information**. *PLoS neglected tropical diseases* 2011, **5**(10):e1370.

6. Wold B, Myers RM: **Sequence census methods for functional genomics**. *Nat Methods* 2008, **5**(1):19-21.

7. Yang MQ, Athey BD, Arabnia HR, Sung AH, Liu Q, Yang JY, Mao J, Deng Y: **High-throughput next-generation sequencing technologies foster new cutting-edge computing techniques in bioinformatics**. *BMC genomics* 2009, **10 Suppl 1**:I1.

8. Ranganathan S, Menon R, Gasser RB: **Advanced in silico analysis of expressed sequence tag (EST) data for parasitic nematodes of major socio-economic importance--fundamental insights toward biotechnological outcomes**. *Biotechnol Adv* 2009, **27**(4):439-448.

9. Nagaraj SH, Gasser RB, Ranganathan S: **A hitchhiker's guide to expressed sequence tag (EST) analysis**. *Brief Bioinform* 2007, **8**(1):6-21.

10. Adams MD, Kerlavage AR, Fields C, Venter JC: **3,400 new expressed sequence tags identify diversity of transcripts in human brain**. *Nat Genet* 1993, **4**(3):256-267.

11. Greenbaum D, Luscombe NM, Jansen R, Qian J, Gerstein M: **Interrelating different types of genomic data, from proteome to secretome: 'oming in on function**. *Genome Res* 2001, **11**(9):1463-1468.

12. Maizels RM, Yazdanbakhsh M: **Immune regulation by helminth parasites: cellular and molecular mechanisms**. *Nat Rev Immunol* 2003, **3**(9):733-744.

13. Lightowlers MW, Rickard MD: **Excretory-secretory products of helminth parasites: effects on host immune responses**. *Parasitology* 1988, **96 Suppl**:S123-166.

14. Hawdon JM, Jones BF, Hoffman DR, Hotez PJ: **Cloning and characterization of Ancylostoma-secreted protein. A novel protein associated with the transition to parasitism by infective hookworm larvae**. *J Biol Chem* 1996, **271**(12):6672-6678.

15. Maizels RM, Gomez-Escobar N, Gregory WF, Murray J, Zang X: **Immune evasion genes from filarial nematodes**. *Int J Parasitol* 2001, **31**(9):889-898.

16. Masoudi-Nejad A, Tonomura K, Kawashima S, Moriya Y, Suzuki M, Itoh M, Kanehisa M, Endo T, Goto S: **EGassembler: online bioinformatics service for large-scale processing, clustering and assembling ESTs and genomic DNA fragments**. *Nucleic Acids Res* 2006, **34**(Web Server issue):W459-462.

17. D'Agostino N, Aversano M, Chiusano ML: **ParPEST: a pipeline for EST data analysis based on parallel computing**. *BMC Bioinformatics* 2005, **6 Suppl 4**:S9.

18. Latorre M, Silva H, Saba J, Guziolowski C, Vizoso P, Martinez V, Maldonado J, Morales A, Caroca R, Cambiazo V *et al*: **JUICE: a data management system that facilitates the analysis of large volumes of information in an EST project workflow**. *BMC Bioinformatics* 2006, **7**:513.

19. Paquola AC, Nishyiama MY, Jr., Reis EM, da Silva AM, Verjovski-Almeida S: **ESTWeb: bioinformatics services for EST sequencing projects**. *Bioinformatics* 2003, **19**(12):1587-1588.

20. Hotz-Wagenblatt A, Hankeln T, Ernst P, Glatting KH, Schmidt ER, Suhai S: **ESTAnnotator: A tool for high throughput EST annotation**. *Nucleic Acids Res* 2003, **31**(13):3716-3719.

21. Menon R, Gasser RB, Miterva M, Ranganathan S: An analysis of the transcriptome of Teladorsagia circumcincta: its biological and biotechnological implications. *BMC Genomics* 2012, in press.

22. Chen YA, Lin CC, Wang CD, Wu HB, Hwang PI: **An optimized procedure greatly improves EST vector contamination removal**. *BMC Genomics* 2007, **8**:416.

23. Falgueras J, Lara AJ, Fernandez-Pozo N, Canton FR, Perez-Trabado G, Claros MG: **SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read**. *BMC Bioinformatics* 2010, **11**:38.

24. RepeatMasker. http://www.repeatmasker.org

25. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WE, Wetter T, Suhai S: **Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs**. *Genome Res* 2004, **14**(6):1147-1159.

26. Iseli C, Jongeneel CV, Bucher P: **ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences**. *Proc Int Conf Intell Syst Mol Biol* 1999:138-148.

27. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes**. *J Mol Biol* 2001, **305**(3):567-580.

28. Bendtsen JD, Jensen LJ, Blom N, Von Heijne G, Brunak S: **Feature-based prediction of non-classical and leaderless protein secretion**. *Protein Eng Des Sel* 2004, **17**(4):349-356.

29. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L *et al*: **InterPro: the integrative protein signature database**. *Nucleic Acids Res* 2009, **37**(Database issue):D211-215.

30. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li CY, Wei L: **KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases**. *Nucleic acids research* 2011, **39**(Web Server issue):W316-322.

31. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG**. *Nucleic acids research* 2006, **34**(Database issue):D354-357.

32. Bieri T, Blasiar D, Ozersky P, Antoshechkin I, Bastiani C, Canaran P, Chan J, Chen N, Chen WJ, Davis P *et al*: **WormBase: new content and better access**. *Nucleic Acids Res* 2007, **35**(Database issue):D506-510.

33. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, Feuermann M, Ghanbarian AT, Kerrien S, Khadake J *et al*: **The IntAct molecular interaction database in 2010**. *Nucleic Acids Res* 2010, **38**(Database issue):D525-531.

34. Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, Livstone M, Oughtred R, Lackner DH, Bahler J, Wood V *et al*: **The BioGRID Interaction Database: 2008 update**. *Nucleic Acids Res* 2008, **36**(Database issue):D637-640.

35. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update**. *Nucleic acids research* 2004, **32**(Database issue):D449-451.

36. Nagaraj SH, Deshpande N, Gasser RB, Ranganathan S: **ESTExplorer: an expressed sequence tag (EST) assembly and annotation platform**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W143-147.

37. Nagaraj SH, Gasser RB, Ranganathan S: **Needles in the EST haystack: large-scale identification and analysis of excretory-secretory (ES) proteins in parasitic nematodes using expressed sequence tags (ESTs)**. *PLoS Negl Trop Dis* 2008, **2**(9):e301.

38. Robinson MW, Menon R, Donnelly SM, Dalton JP, Ranganathan S: **An integrated transcriptomics and proteomics analysis of the secretome of the helminth pathogen Fasciola hepatica: proteins associated with invasion and infection of the mammalian host**. *Mol Cell Proteomics* 2009, **8**(8):1891-1907.

39. Dold C, Holland CV: **Ascaris and ascariasis**. *Microbes Infect* 2011, **13**(7):632-637.

40. Holland CV: **Predisposition to ascariasis: patterns, mechanisms and implications**. *Parasitology* 2009, **136**(12):1537-1547.

41. Boguski MS, Lowe TM, Tolstoshev CM: **dbEST--database for "expressed sequence tags"**. *Nat Genet* 1993, **4**(4):332-333.

42. Tamang DG, Saier MH, Jr.: **The cecropin superfamily of toxic peptides**. *J Mol Microbiol Biotechnol* 2006, **11**(1-2):94-103.

43. Bulet P, Stocklin R: **Insect antimicrobial peptides: structures, properties and gene regulation**. *Protein Pept Lett* 2005, **12**(1):3-11.

44. Steiner H, Hultmark D, Engstrom A, Bennich H, Boman HG: **Sequence and specificity of two antibacterial proteins involved in insect immunity. Nature 292: 246-248. 1981**. *J Immunol* 2009, **182**(11):6635-6637.

45. Pillai A, Ueno S, Zhang H, Lee JM, Kato Y: **Cecropin P1 and novel nematode cecropins: a bacteria-inducible antimicrobial peptide family in the nematode Ascaris suum**. *Biochem J* 2005, **390**(Pt 1):207-214.

46. O'Rourke D, Baban D, Demidova M, Mott R, Hodgkin J: **Genomic clusters, putative pathogen recognition molecules, and antimicrobial genes are induced by infection of C. elegans with M. nematophilum**. *Genome Res* 2006, **16**(8):1005-1016.

47. Schulenburg H, Hoeppner MP, Weiner J, 3rd, Bornberg-Bauer E: **Specificity of the innate immune system and diversity of C-type lectin domain (CTLD) proteins in the nematode Caenorhabditis elegans**. *Immunobiology* 2008, **213**(3-4):237-250.

48. Drickamer K: **Two distinct classes of carbohydrate-recognition domains in animal lectins**. *J Biol Chem* 1988, **263**(20):9557-9560.

49. Drickamer K: **Ca(2+)-dependent sugar recognition by animal lectins**. *Biochem Soc Trans* 1996, **24**(1):146-150.

50. Garcia-Vallejo JJ, van Kooyk Y: **Endogenous ligands for C-type lectin receptors: the true regulators of immune homeostasis**. *Immunol Rev* 2009, **230**(1):22-37.

51. Cash HL, Whitham CV, Behrendt CL, Hooper LV: **Symbiotic bacteria direct expression of an intestinal bactericidal lectin**. *Science* 2006, **313**(5790):1126-1130.

52. Yoshida A, Nagayasu E, Horii Y, Maruyama H: **A novel C-type lectin identified by EST analysis in tissue migratory larvae of Ascaris suum**. *Parasitol Res* 2011.

53. Liu Z, Klaavuniemi T, Ono S: **Distinct roles of four gelsolin-like domains of Caenorhabditis elegans gelsolin-like protein-1 in actin filament severing, barbed end capping, and phosphoinositide binding**. *Biochemistry* 2010, **49**(20):4349-4360.

54. Klaavuniemi T, Yamashiro S, Ono S: **Caenorhabditis elegans gelsolin-like protein 1 is a novel actin filament-severing protein with four gelsolin-like repeats**. *J Biol Chem* 2008, **283**(38):26071-26080.

Figures

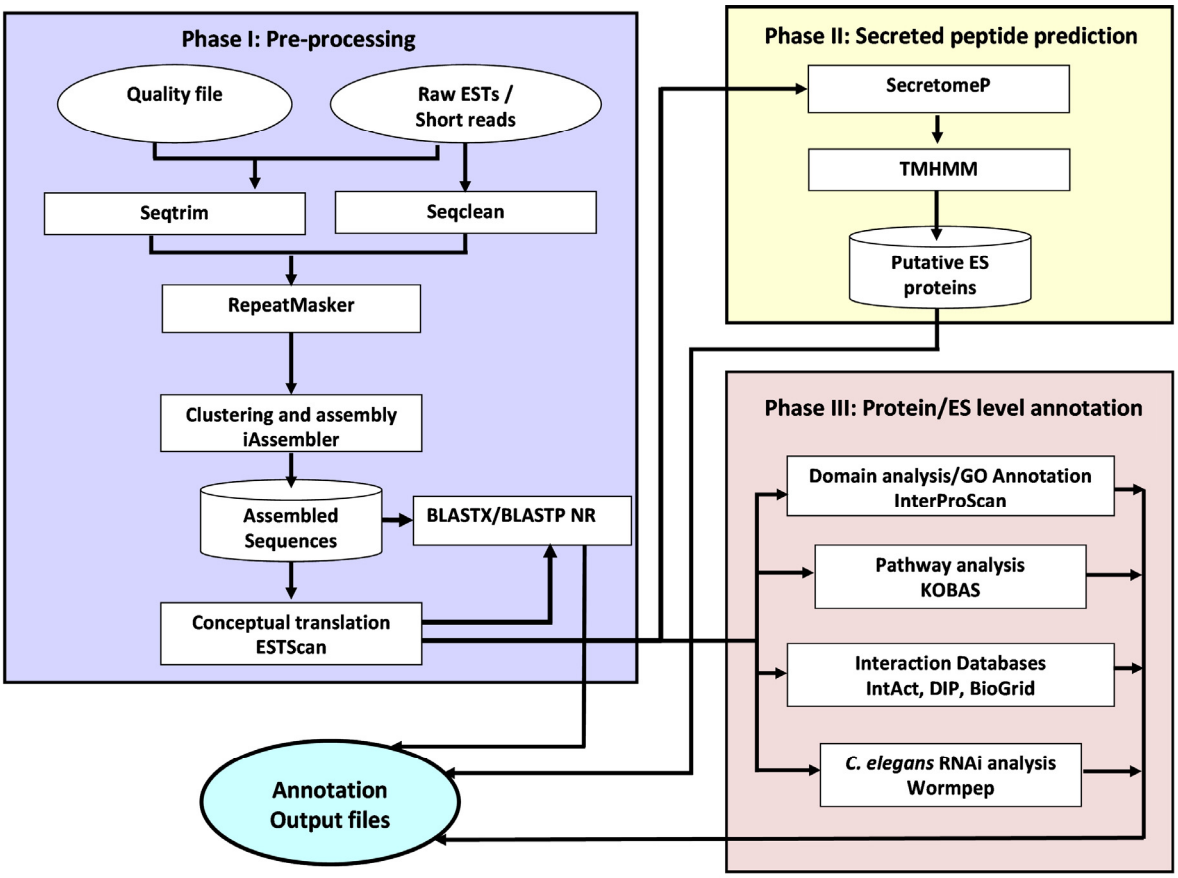


Figure 1: Schematic diagram of TranSeqAnnotator workflow.

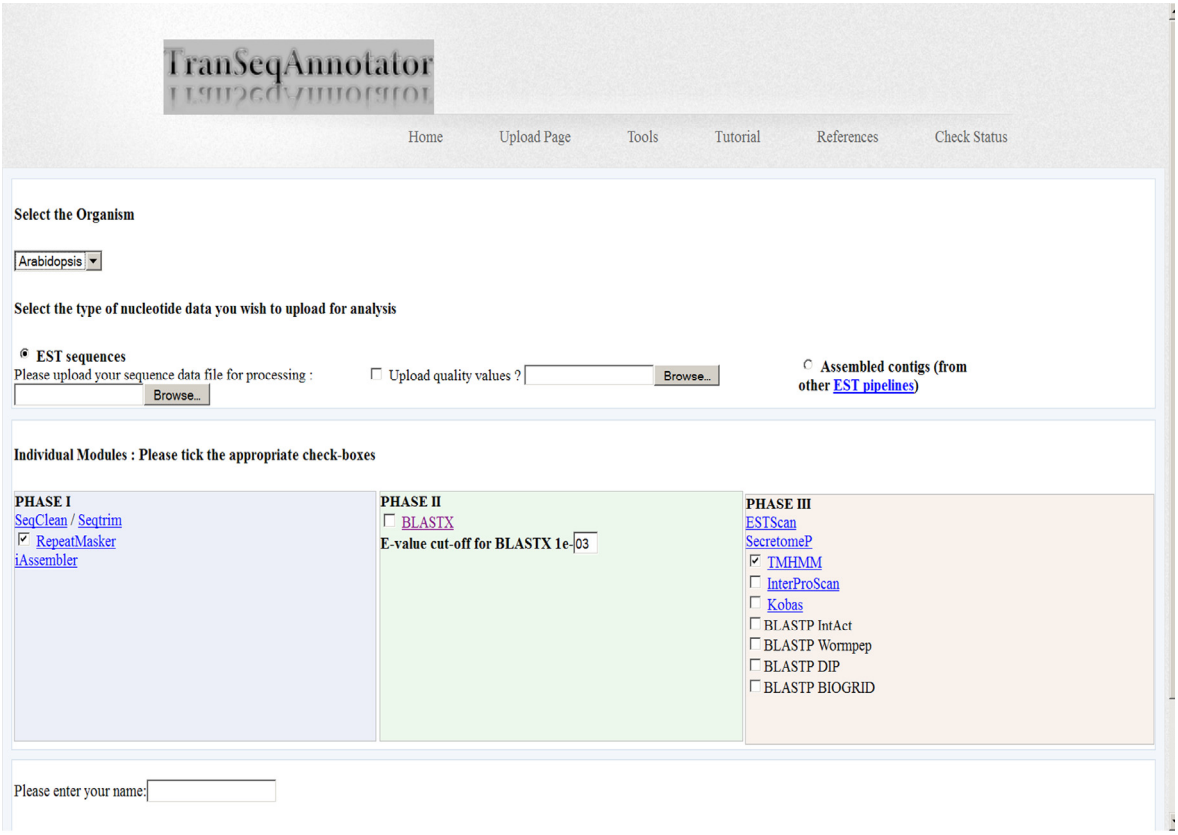


Figure 2: TransSeqAnnotator data submirssion page

Additional files

Additional file 1 (*.xls)

Title: GO annotation for putative peptides

Description: Gene Ontology annotations from Interproscan reported.

Additional file 2 (*.xls)

Title: KEGG Pathway analysis of proteins (E-value threshold of 1E-05)

Description: Database matches reported.

Additional file 3 (*.xls)

Title: Domain description for the protein sequences

Description: Interproscan domains reported.

Additional file 4 (*.xls)

Title: Top BLAST hits for secreted proteins

Description: Non-redundant database matches reported.

5.2 Conclusions

The TranSeqAnnotator combines ESTExplorer and EST2Secretome with additional features for handling next-generation sequencing (NGS) data, extended ES protein prediction and comprehensive annotation. As a few programs overlap with existing pipelines, the design and code is easily adaptable and additional modules can be added for specific requirement. ESTExplorer provides EST analysis for all EST applications and EST2Secretome for ES protein detection. The ES proteins with N-terminal signal sequences alone were addressed in EST2Secretome pipeline but, the analysis discussed in chapter 3 highlighted the importance of non-classically secreted proteins. Hence TranSeqAnnotator uses the new tool, SecretomeP. With the advent of NGS, the existing pipelines were modified to accept short read sequence data as input along with EST data, followed by protein and ES analysis. The pre-processing of both FASTA and qual files using seqtrim program and also, clustering and assembly using iAssembler are the other highlights of the pipeline. TranSeqAnnotator has better annotation capabilities such as identification of interaction partners with different interaction databases to gain better insights of cellular function based on binding partners. Users also have the option to select different optional programs as required. TranSeqAnnotator has been developed for varied applications as a multi-purpose, high-throughput transcriptome analysis pipeline.

Chapter 6: Comparison of adult and fourth larval stage transcriptomic data: *Teladorsagia circumcincta*

6.1 Summary

Teladorsagia circumcincta (Strongylida, Trichostrongylidae), a pathogenic parasitic nematode cause disease (teladorsagiosis) that leads to anorexia, diarrhoea and death in lambs with the major impact in the areas of lamb productivity with a reduction in weight gain. They inhabit in the abomasum of small ruminants and are endemic in temperate regions of the world. The study was focussed to define and compare the transcriptome of the fourth larval stage (L4) of *T. circumcincta* (507,124 ESTs) against the adult stage (presented in Chapter 4) with the focus on main pathways associated to molecules known to be expressed in this nematode. The early host responses are considered important in preventing the development of worms in infected/immune sheeps hence, the study was concentrated on different stages of the nematode. The detailed annotation included pathway mapping of predicted proteins (including 930 excreted/secreted proteins), domain analysis and GO annotation along with secretory signal peptides. Understanding of resistance and anthelmintic actions is still not clear hence, an alternative method is important. Studies have shown vaccination as a suitable option for control of sheep which developed acquired immunity against this parasite. The study of parasite excretory/secretory products (ESP) which have the prospective to invoke protective immunity against *T. circumcincta* is main the highlight of the project. We report the an imminent aspect into the molecular biology of this parasite and disease manifestation highlights the future research, proteomic and metabolomic studies of this parasite and could lead to novel intervention and control strategies.

Comparison of adult and fourth larval stage transcriptomic data: *Teladorsagia circumcincta*

Ranjeeta Menon[1], Shoba Ranganathan[1, 2,]*

[1]Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, New South Wales 2109, Australia

[2]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, Singapore 117597

*Corresponding author

E-mail: Shoba Ranganathan*- shoba.ranganathan@mq.edu.au

Abstract

Background

Teladorsagia circumcincta (order Strongylida) is an economically important parasitic nematode of sheep and goats in temperate regions of the world. The aims of the present study were to define and compare the transcriptomic analysis of the fourth and adult stage of *T. circumcincta* and to predict the main pathways that are linked in this nematode. Though the *T. circumcincta* are known to induce intense changes in areas related to parasite ES in infected sheep and goats, the nature of the proteins with in vitro excretory/secretory products (ES) of the fourth-stage larva (L4) and adult *Teladorsagia circumcincta* are undefined. We focussed on these stages because it has been shown previously that the early host responses (i.e., within 5 days of challenge) are important in preventing the establishment of worms in previously infected, immune sheep. We are investigating parasite excretory/secretory products for molecules, which have potential to invoke protective immunity against *T. circumcincta* as the sheep develop acquired immunity to *T. circumcincta* so vaccination is a valid option for control.

Results

A total of 507,124 cDNA representing 30493 genes were assembled. The key pathways identified for 930 ES proteins, include RNA transport, protein processing in endoplasmic reticulum, RNA degradation and lysosome showed the highest representation. Other important pathways include detailed protein analysis were purine metabolism, protein processing in endoplasmic reticulum, Ribosome and Oxidative phosphorylation. Pathways other than purine metabolism and Oxidative phosphorylation were unique to L4 larval stage. A novel, immunogenic cathepsin F, the most abundant molecule in excretory/secretory products released in vitro by *T. circumcincta* was identified in L4 stage while, Tci-CF-1, a cathepsin F family is reported in both adult and fourth larval stage. Other important identifications in both adult and fourth stage larval include the cathepsin B-like cysteine protease, leading vaccine candidate. Another most abundant protein identified include C-type lectin family members in *T. circumcincta* larvae exposed to a naïve host environment, legumain of *Haemonchus contortus* were not identified in adult *Teladorsagia circumcincta* analysis.

Conclusions

We were successful in comparing the results for the large dataset of fourth larval stage of *T. circumcincta* ESTs against the adult stage. The results provide an imminent aspect into the molecular biology of this parasite and disease manifestation which provides potential focus for future research, proteomic and metabolomic studies of this parasite and could lead to novel intervention and control strategies.

Background

Teladorsagia circumcincta (Strongylida, Trichostrongylidae), a pathogenic parasitic nematode that inhabit in the abomasum of small ruminants are endemic in temperate regions of the world. *T. circumcincta* primarily causes disease (teladorsagiosis) that leads to anorexia, diarrhoea and death in lambs with the major impact of its effect on lamb productivity with a reduction in weight gain [1]. Parasitic lifestyle transition occurs when ensheathed infective third stage larvae (L3) are ingested by the host while grazing. Then, L3s exsheath (XL3) and develop into fourth stage larvae (L4), immature and mature adult worms. Transcriptomic and proteomic studies have been conducted for the identification of genes and proteins specifically expressed during the early parasitic phase of *T. circumcincta* [2, 3]. Anthelmintic resistance in *T. circumcincta* against all three available anthelmintic drug classes are common with advanced studies conducted [4, 5]. Recently, the faecal Egg Count Reduction Test (FECRT) showed the resistance of a *Teladorsagia circumcincta* isolate against LEV (39-58%), IVM (88-92%) and doramectin (85%) [6]. A combination of anthelmintic treatments and pasture management are usually used to control the gastrointestinal parasitic nematodes of sheep [7].

The spreading of anthelmintic resistance has reduced the efficacy of the drug against the control of gastro-intestinal nematodes which is mainly based on anthelminthic treatments. Recently studies have reported the efficacy of the drug against inhibited fourth-stage larvae of Australian origin as 99.8% for *Teladorsagia* spp. [8] and also, the first molecule from the aminoacetonitrile derivative class of anthelmintics, monepantel has been developed for the use in sheep [9, 10]. The first published evidence on the efficacy of monepantel against natural infections of *T. circumcincta* parasite and its stages was the report that the untreated control sheep were additionally infected with developing fourth-stage *Teladorsagia* spp. larvae at necropsy [11]. Host immunomodulatory factors are released by numerous parasitic helminths in their excretory and secretory (ES) products [12-16]. Studies on several parasitic nematode species have shown nematode excretory/secretory (ES) products as a major source of antigen [17-20] and studies were mainly conducted on ES products to understand their effects on the host tissues or on characterization of proteases released in vitro. The ES material from L3, L4 and adult *T. circumcincta* has been studied for stage-specific protease activity at a range of pH optima [21]. Parasites within the host with various functions such as penetration of host tissue barriers, feeding and evasion or modulation of the host immune response are carried out with the proteins excreted or secreted by parasites (ES) [22]. ES proteins have the prospective to identify

novel targets to control infection with a particular focus on vaccination and also provide an insight into the dynamics of the host-parasite relationship. ES components have been evaluated as vaccine candidates in a number of ruminant nematode parasites with promising results [20, 23-25].

RESULTS AND DISCUSSIONS

From 507,124 L4 454 sequences downloaded from University of Liverpool (http://worm1.liv.ac.uk/file_summary.html) after pre-processing, assembly and conceptual translation a total of 2811 protein sequences were obtained which mapped to 266 KEGG pathways. The top 30 (highly represented) pathways are shown in Table 1. Purine metabolism (n=67), protein processing in endoplasmic reticulum (n=60), ribosome (n=53) and oxidative phosphorylation (n=52) had the highest representation for the sequences mapped to KEGG pathways. Protein processing in endoplasmic reticulum was represented only in L4 stage. Some of the other pathways that were well represented by the peptides unique to L4 stage include Huntington's disease (n=51), spliceosome (n=50), pathways in cancer (n=28). Peptides were also mapped to several other pathways glycolysis / Gluconeogenesis, insulin signalling pathway in both L4 and adult *T. circumcincta*. A complete listing of the KEGG mappings to all the pathways is available as supplementary data (Additional File 1).

The detailed annotation of proteins identified Secreted Cathepsin F, excretory/secretory antigen, cytochrome c oxidase, cytochrome P450 which is known to play vital role in host-parasite interaction. Cathepsin B family is considered as the commonest type of nematode protease [26]. A novel, immunogenic cathepsin F secreted by L4 *T. circumcincta* already identified in previous studies is the most abundant molecule in excretory/secretory products released in vitro by *T. circumcincta* [27]. *T. circumcincta* cathepsin F-1, Tci-CF-1, of cathepsin F class have shown identity to ESTs present in the *Ostertagia ostertagi* and *Haemonchus contortus* expressed sequence tag databases and to hypothetical proteins identified in the genomes of *Caenorhabditis elegans* and *Caenorhabditis briggsae.* Tci-cf-1 represents the only cathepsin-encoding EST sequence derived from both L4 and adult stage cDNA libraries. The identification of a cathepsin F (Tci-CF-1), the most abundant molecule in ES products derived from *T. circumcincta* L4 is considered as a key finding in the study.

The identification of antigens with protective capacities has always remained a challenge for the development of an effective vaccine**.** ES antigens released through specialized excretory/secretory organs or expressed at the parasite surface are often stage specific. ES products are considered important as the candidate protective antigens due to their highly immunogenic nature in infections and they are the targets for immune effector mechanisms [17, 28]. Cytochrome P450s, the candidate resistance genes are believed to be involved in drug efflux or metabolism [29, 30]. In response to drug exposure the changes in the parasite's respiration rate causes changes in expression, cytochrome *c* oxidases are part of the mitochondrial electron transport chain [31]. Other well identified proteins include cathepsin B-like cysteine protease, astacin-like metalloprotease, heat shock protein, metalloprotease, putative L3 ES protein, SCP-Like extracellular protein, NADH dehydrogenase, NADH Ubiquinone Oxidoreductase family member. Astacin-like metalloproteases are vital for establishment of the parasite in the host and has been identified in the excretory/secretory (ES) products of *T. circumcincta* larvae [3]. Studies show the penetration of the host intestinal mucosa and the penetration of mucosal capillaries helped by metalloproteinase and have specific roles to play in parasite maintenance within the abomasum [21, 32].

Based on the active catalytic centre residues (aspartic, serine and cysteine proteases) or on the dependence on co-factors for activity (metalloproteases) different classes of proteases are assigned. Aspartic proteases are considered to be the most conserved group while, cysteine proteases are involved in tissue penetration and feeding [33]. Heat shock proteins are known to play vital role in response to stimulus and developmental process. HSPs are reported to be present in the parasitic nematodes *A. caninum, H. contortus, D. viviparous, T. canis.* Studies have shown stress-induced heat shock protein are included in the immune-exposed larvae [34] where the host immune system produce oxidative free radicals to kill pathogens. Other significant identification were the *C. elegans* EGg Laying defective family member (egl-21) homologue, which facilitates acetylcholinerelease at neuromuscular junctions [35] and a *C. elegans* C-type LECtin family member (clec-1) homologue. C-type lectin family members (clec-200) the most abundant in *T. circumcincta* larvae exposed to a naïve host environment were identified. Galectin shows high expression in sheeps infected with *H. contortus*. They are secreted into the lumen and detected in mucus. These have also been reported in *T. circumcincta* third larval stage studies [36]. The fatty acid and retinol binding protein were reported in adult hookworm ES products.

Nematode lipid binding proteins are known for their functions in transport and metabolism of fatty acids, sterols, or retinol, used in the metabolic and developmental processes of embryogenesis, scavenging, growth and cellular differentiation, glycoprotein synthesis [37]. We identified APY-1, APYrase family member, a *C. elegans* homologue of the apyrase. Studies have shown constitutive activation of unfolded protein response with the depletion of APY-1. It was first found in the saliva of blood-sucking insects and it is reported in closely related enzymes in human [38, 39]. FaRPs are believed to interact with different peptide receptors as a result of different effects seen on nematode somatic musculature. Non-specific receptor interactions are also reported with structurally related molecules [40]. The entire list of domain details in Additional File 2.

We could functionally assign GO terms with 4424 being assigned biological process (BP), 5744 molecular function (MF) and 1928 cellular component (CC) GO terms. A summary of GO annotation by biological process, molecular function and cellular component with metabolic process (GO:0008152), oxidation reduction (GO:0055114) and proteolysis (GO:0006508) were the most common GO categories representing biological processes. The largest number of GO terms in molecular function was ATP binding (GO:0005524), oxidoreductase activity (GO:0016491) and catalytic activity (GO:0003824). The cellular component represent membrane (GO:0016020) and cytoplasm (GO:0005737) which are significant from the viewpoint of identifying novel drug or vaccine candidates. The entire list of GO annotation is shown in Additional File 3.

930 putatively secreted proteins were identified from the present data set of 30493 rESTs. We identified two legumain of *Haemonchus contortus*, C-type Lectin from *Caenorhabditis elegans* and *Necator americanus*, putative L3 ES protein of *Ostertagia ostertagi*, astacin-like metalloprotease from *Haemonchus contortus*, serine/threonine protein phosphatise, kettin from *Caenorhabditis elegans*, Zinc finger from *Caenorhabditis elegans*, *Brugia malayi* (Additional File 4).

We could functionally assign GO terms to 930 putative ES proteins with 62 being assigned biological process (BP), 82 molecular function (MF) and 34 cellular component (CC) GO terms. A summary of GO annotation by biological process, molecular function and cellular component with proteolysis (GO:0006508), transmembrane transport (GO:0055085) and regulation of transcription (GO:0006355) were the most common GO categories

representing biological processes. The largest number of GO terms in molecular function was ATP binding (GO:0005524), protein binding (GO:0005515) and binding (GO: 0005488). The cellular component represent membrane (GO:0016020) and intracellular (GO:0005622) which are significant from the viewpoint of identifying novel drug or vaccine candidates (Additional File 5).

A total of 94 sequences were mapped to 75 KEGG pathways. RNA transport (n=3), Protein processing in endoplasmic reticulum (n=3), RNA degradation (n=3) and Lysosome (n=3) had the highest representation for the sequences mapped to KEGG pathways (Table 2). Some of the other well represented pathways by ES proteins included the Glycine, serine and threonine metabolism, Lysine degradation, Tight junction, Antigen processing and presentation, Insulin signaling pathway. The entire lists of pathway analysis in Additional File 6.

Discussion

Proteins inferred most represented domain family of proteins were the secreted cathepsin F of *Teladorsagia circumcincta*, ryanodine receptor of *Caenorhabditis elegans*, metalloprotease I of *Ostertagia ostertagi*, protein disulphide isomerase of *Ostertagia ostertagi*, glutamate dehydrogenase ryanodine receptor encoded by the unc-68 gene are conserved in nematodes [41]. 60 kDa protein similar to a protein disulphide isomerase (PDI) were identified in both adult and L4 *T. circumcincta* proteomic analysis also identified in the ES from parasitic stages of *O. ostertagi* [28, 42].These ubiquitous enzyme play vital role during generation of many secretory and outer membrane proteins as a catalyst of protein folding [43]. Proteomic analysis of adult ES identified enzymes, a protein at ~20 kDa with sequence identity to nucleoside diphosphate kinase and a putative glutamate dehydrogenase from *H. Contortus* (Additional File 2).

We could identify proteins which were homologues of heat shock-protein, cytochrome p450, Cytochrome-c oxidase, EGg Laying defective family member, NADH dehydrogenase which were also identified in the adult and third larval stage [36] . The peptidyl-glycine alpha-amidating monooxygenase were identified only in third and fourth larval stage.

The analysis of 930 sequences identified legumain of *H. contortus.* The identification of 11 kDa excretory-secretory protein with homology to *Trichostrongylus colubriformis*, a

parasitic gastrointestinal nematode of sheep, goats, cattle and deer shows the possibility of producing immunomodulatory factors by *T. circumcincta*, that may aid the survival of the parasites observed in adult *Teladorsagia circumcincta* [44]. Followed by the identification of 15/24 kDa proteomic studies have shown that vaccination with recombinant 15/24 kDa ES proteins of *H. contortus* did confer protection in 9-month-old sheep, but not in 3-month-old lambs [3, 45]. The studies with adult *Teladorsagia circumcincta* show 15/24 kDa with homologue to *Brugia malayi* and *Ascaris suum*. Ancylostoma secreted protein 1 precursor from *Necator americanus* were identified while in adult *Teladorsagia circumcincta* ASP-3 from *O. ostertagi* (ASP3) precursors (*Ostertagia ostertagi*), ancyclostoma-secreted protein-like proteins (*Ostertagia ostertagi*) was identified (Additional File 4).

ASPs are group of secreted proteins classified as the SCP/Tpx-1/Ag5/PR-1/Sc7 family are members of nematode-specific molecules [7]. A proteomics approach was used to show ASPs as the most abundant protein in adult *O. ostertagi* ES [46] and also, an important vaccine candidate [7, 47]. ASPs have been identified in L4 *T. circumcincta* ES and in it would seem that ASPs are abundant in *T. circumcincta* L4 ES and are also highly represented in EST datasets of different parasitic stages of this nematode [48]. *T. circumcincta* expressed sequence tag (EST) has shown significant sequence identity to astacin (AST)-like metalloprotease from *Ostertagia ostertagi* in previous studies. An AST-like metalloproteinase and a cathepsin have been identified in 3rd larval stage [3]. The study of the fourth larval stage has identified astacin-like metalloprotease of *Haemonchus contortus*. Nematode cathepsins [49] and ASTs [50] are known to digest the host proteins. The nematode AST-like metalloproteinases are known to play vital role in adaptive immune responses early in infection and in stimulating innate [51]. Cathepsin B-like cysteine protease was identified in fourth stage and adult stage larval. The cytokine network and down-regulate inflammation are interrupted with the anticipation of cathepsin B-like cysteine protease to interfere with the immune system [52]. Cathepsin B-like cysteine protease of *Haemonchus contortus* identified in fourth stage larvae are considered leading vaccine candidate. C-type lectin, carbohydrate binding proteins identified in both adult and fourth stages are known to be involved in immune processes, such as antigen uptake and presentation, cell adhesion, apoptosis and T cell polarisation form a part of the superfamily of the pathogenesis-related proteins (PRPs) [53, 54]. They are known to be abundant in *T. canis* [55, 56]. Galectins which were identified in adult stage was not identified in fourth stage. Legumain or asparaginyl proteinases

identified as the novel class of cysteine proteinase were first reported in animals in the human blood fluke, *Schistosoma mansoni* [57]. Legumains involved in the activation of, cathepsin B-like proteinases, cysteine proteases are considered to help degrade the bloodmeal in blood-feeding helminths such as schistosomes, hookworms and other nematode species [58]. The legumain of *Haemonchus contortus* were not identified in adult *Teladorsagia circumcincta* analysis.

Myosin heavy chain of *Haemonchus contortus* are known as antigenic proteins and are used in the vaccine studies [59]. Nuclear Hormone Receptor is a transcription factor belonging to the superfamily of steroid nuclear hormone receptors. They are known to affect in areas of sex determination, molting, developmental timing, diapause, and life span. Studies have been reported that the three members of the NR4A1/Nur77/NGFIB orphan nuclear hormone receptor subfamily (NR4A1, NR4A2, and NR4A3) are known to be down-regulated in the HeLaHF revertant of cervical cancer [60, 61]. The future focus of NR studies of *C. elegans* is in the area of ion balance, stress response and immunity. Serine, cysteine as well as proteinase inhibitors identified in both adult and fourth stage represent the protein families are considered as important targets for parasite invention and control. Parasites are protected against digestion by endogenous or host-derived proteinases with the help of proteinase inhibitors [52, 62-64].

Conclusions

Large-scale analysis with fourth larval stage and comparing with the adult stage of *Teladorsagia circumcincta* identified important excretory/secretory products that could be step towards identifying novel intervention and control strategies. The fourth stage analysis identified cathepsin F, the most abundant molecule in excretory/secretory products released in vitro by *T. Circumcincta.* The proteins that represent the important targets for parasite invention and control, Serine, cysteine as well as proteinase inhibitors, the cathepsin B-like cysteine protease, the leading vaccine candidate and C-type lectin family members in *T. circumcincta* larvae exposed to a naïve host environment were identified in both adult and fourth stage. The proteins identified in adult like Galectins were not reported in fouth larval stage. Similarly, legumain of *Haemonchus contortus,* the novel class of cysteine proteinase was identified in fourth larval stage were not identified in adult *Teladorsagia circumcincta* analysis. These results present a step towards future research on disease manifestation and molecular biology of the parasite with the focus on studies integrating proteomics and metabolomic studies.

METHODS

The short reads were were pre-processed using SeqClean, RepeatMasker, for the removal of flanking vector and adapter sequences aligned/clustered using the iassembler [65]. rESTs were then conceptually translated into peptides using ESTScan [66], which were further characterized via InterProScan (domain/motifs) and Gene ontologies were inferred using InterProScan [67]. Peptides predicted from rEST were also compared, using BLASTP, with the non-redundant protein sequence database from National Centre for Biotechnology Information (NCBI). The peptides were mapped to respective pathways in *C. elegans* using KOBAS [68] . SecretomeP [69] was used to identify the non-classical sec pathway. All proteins with signal sequences are passed on to TMHMM [70], a hidden Markov model-based transmembrane helix prediction program, to ‘‘filter out’’ transmembrane proteins. The subset lacking transmembrane helices is selected as ES proteins for further annotation. The predicted proteins are searched against the data sets derived from Wormpep Wombase [http://wormbase.org/] for locating nematode homologues. Lists of homologous proteins in *C. elegans*, linked to WormBase are obtained by mapping against Wormpep. Results are further validated by the gene ontology and pathway analysis. InterProScan was used to analyse the protein domain/motif.

Abbreviations

EST: Expressed sequence tag; ESPs: excretory/secretory proteins; KEGG: Kyoto Encyclopedia of Genes and Genomes database; KOBAS (KEGG Orthology Based Annotation System)

Authors' contributions

RM carried out the analysis, computational studies and drafted the manuscript. RM and SR participated in the design of the study and interpretation of data. SR conceived the project and finalized the manuscript. All authors have read and approved the final manuscript.

Acknowledgments

RM gratefully acknowledges the award of a Macquarie University Research Excellence Scholarship. The data was generated by The University of Liverpool, United Kingdom. Open access application charges were borne by Macquarie University.

References

1. Gibson TE, Everett G: **Effect of different levels of intake of Ostertagia circumcincta larvae on the faecal egg counts and weight gain of lambs**. *Journal of comparative pathology* 1976, **86**(2):269-274.
2. Nisbet AJ, Redmond DL, Matthews JB, Watkins C, Yaga R, Jones JT, Nath M, Knox DP: **Stage-specific gene expression in Teladorsagia circumcincta (Nematoda: Strongylida) infective larvae and early parasitic stages**. *Int J Parasitol* 2008, **38**(7):829-838.
3. Smith SK, Nisbet AJ, Meikle LI, Inglis NF, Sales J, Beynon RJ, Matthews JB: **Proteomic analysis of excretory/secretory products released by Teladorsagia circumcincta larvae early post-infection**. *Parasite immunology* 2009, **31**(1):10-19.
4. Bartley DJ, Jackson F, Jackson E, Sargison N: **Characterisation of two triple resistant field isolates of Teladorsagia from Scottish lowland sheep farms**. *Veterinary parasitology* 2004, **123**(3-4):189-199.
5. Wrigley J, McArthur M, McKenna PB, Mariadass B: **Resistance to a triple combination of broad-spectrum anthelmintics in naturally-acquired Ostertagia circumcincta infections in sheep**. *New Zealand veterinary journal* 2006, **54**(1):47-49.
6. Martinez-Valladares M, Famularo MR, Fernandez-Pato N, Cordero-Perez C, Castanon-Ordonez L, Rojo-Vazquez FA: **Characterization of a multidrug resistant Teladorsagia circumcincta isolate from Spain**. *Parasitology research* 2011.
7. Nisbet AJ, Smith SK, Armstrong S, Meikle LI, Wildblood LA, Beynon RJ, Matthews JB: **Teladorsagia circumcincta: activation-associated secreted proteins in excretory/secretory products of fourth stage larvae are targets of early IgA responses in infected sheep**. *Experimental parasitology* 2010, **125**(4):329-337.
8. Stein PA, Rolfe PF, Hosking BC: **The control of inhibited fourth-stage larvae of Haemonchus contortus and Teladorsagia spp. in sheep in Australia with monepantel**. *Veterinary parasitology* 2010, **169**(3-4):358-361.
9. Kaminsky R, Ducray P, Jung M, Clover R, Rufener L, Bouvier J, Weber SS, Wenger A, Wieland-Berghausen S, Goebel T *et al*: **A new class of anthelmintics effective against drug-resistant nematodes**. *Nature* 2008, **452**(7184):176-180.

10. Kaminsky R, Gauvry N, Schorderet Weber S, Skripsky T, Bouvier J, Wenger A, Schroeder F, Desaules Y, Hotz R, Goebel T *et al*: **Identification of the amino-acetonitrile derivative monepantel (AAD 1566) as a new anthelmintic drug development candidate**. *Parasitology research* 2008, **103**(4):931-939.
11. Ramage C, Bartley DJ, Jackson F, Cody R, Hosking BC: **The efficacy of monepantel against naturally acquired inhibited and developing fourth-stage larvae of Teladorsagia circumcincta in sheep in the United Kingdom**. *Veterinary parasitology* 2011.
12. Maizels RM, Yazdanbakhsh M: **Immune regulation by helminth parasites: cellular and molecular mechanisms**. *Nature reviews Immunology* 2003, **3**(9):733-744.
13. Vermeire JJ, Cho Y, Lolis E, Bucala R, Cappello M: **Orthologs of macrophage migration inhibitory factor from parasitic nematodes**. *Trends in parasitology* 2008, **24**(8):355-363.
14. Johnston MJ, MacDonald JA, McKay DM: **Parasitic helminths: a pharmacopeia of anti-inflammatory molecules**. *Parasitology* 2009, **136**(2):125-147.
15. McSorley HJ, Grainger JR, Harcus Y, Murray J, Nisbet AJ, Knox DP, Maizels RM: **daf-7-related TGF-beta homologues from Trichostrongyloid nematodes show contrasting life-cycle expression patterns**. *Parasitology* 2010, **137**(1):159-171.
16. Nisbet AJ, Bell NE, McNeilly TN, Knox DP, Maizels RM, Meikle LI, Wildblood LA, Matthews JB: **A macrophage migration inhibitory factor-like tautomerase from Teladorsagia circumcincta (Nematoda: Strongylida)**. *Parasite immunology* 2010, **32**(7):503-511.
17. Lightowlers MW, Rickard MD: **Excretory-secretory products of helminth parasites: effects on host immune responses**. *Parasitology* 1988, **96 Suppl**:S123-166.
18. Hawdon JM, Narasimhan S, Hotez PJ: **Ancylostoma secreted protein 2: cloning and characterization of a second member of a family of nematode secreted proteins from Ancylostoma caninum**. *Molecular and biochemical parasitology* 1999, **99**(2):149-165.
19. Maizels RM, Gomez-Escobar N, Gregory WF, Murray J, Zang X: **Immune evasion genes from filarial nematodes**. *International journal for parasitology* 2001, **31**(9):889-898.
20. Yatsuda AP, Krijgsveld J, Cornelissen AW, Heck AJ, de Vries E: **Comprehensive analysis of the secreted proteins of the parasite Haemonchus contortus reveals**

extensive sequence variation and differential immune recognition. *The Journal of biological chemistry* 2003, **278**(19):16941-16951.

21. Young CJ, McKeand JB, Knox DP: **Proteinases released in vitro by the parasitic stages of Teladorsagia circumcincta, an ovine abomasal nematode**. *Parasitology* 1995, **110 (Pt 4)**:465-471.
22. Knox DP: **Development of vaccines against gastrointestinal nematodes**. *Parasitology* 2000, **120 Suppl**:S43-61.
23. Schallig HD, Van Leeuwen MA: **Protective immunity to the blood-feeding nematode Haemonchus contortus induced by vaccination with parasite low molecular weight antigens**. *Parasitology* 1997, **114 (Pt 3)**:293-299.
24. Tsuji N, Suzuki K, Kasuga-Aoki H, Isobe T, Arakawa T, Matsumoto Y: **Mice intranasally immunized with a recombinant 16-kilodalton antigen from roundworm Ascaris parasites are protected against larval migration of Ascaris suum**. *Infection and immunity* 2003, **71**(9):5314-5323.
25. Vercauteren I, Geldhof P, Vercruysse J, Peelaers I, van den Broeck W, Gevaert K, Claerebout E: **Vaccination with an Ostertagia ostertagi polyprotein allergen protects calves against homologous challenge infection**. *Infection and immunity* 2004, **72**(5):2995-3001.
26. Tort J, Brindley PJ, Knox D, Wolfe KH, Dalton JP: **Proteinases and associated genes of parasitic helminths**. *Advances in parasitology* 1999, **43**:161-266.
27. Redmond DL, Smith SK, Halliday A, Smith WD, Jackson F, Knox DP, Matthews JB: **An immunogenic cathepsin F secreted by the parasitic stages of Teladorsagia circumcincta**. *International journal for parasitology* 2006, **36**(3):277-286.
28. Vercauteren I, Geldhof P, Peelaers I, Claerebout E, Berx G, Vercruysse J: **Identification of excretory-secretory products of larval and adult Ostertagia ostertagi by immunoscreening of cDNA libraries**. *Molecular and biochemical parasitology* 2003, **126**(2):201-208.
29. Kotze AC: **Cytochrome P450 monooxygenase activity in Haemonchus contortus (Nematoda)**. *International journal for parasitology* 1997, **27**(1):33-40.
30. Prichard RK, Roulet A: **ABC transporters and beta-tubulin in macrocyclic lactone resistance: prospects for marker development**. *Parasitology* 2007, **134**(Pt 8):1123-1132.
31. Dicker AJ, Nath M, Yaga R, Nisbet AJ, Lainson FA, Gilleard JS, Skuce PJ: **Teladorsagia circumcincta: the transcriptomic response of a multi-drug-**

resistant isolate to ivermectin exposure in vitro. *Experimental parasitology* 2011, **127**(2):351-356.

32. Hotez PJ, Trang NL, McKerrow JH, Cerami A: **Isolation and characterization of a proteolytic enzyme from the adult hookworm Ancylostoma caninum**. *The Journal of biological chemistry* 1985, **260**(12):7343-7348.
33. Williamson AL, Brindley PJ, Knox DP, Hotez PJ, Loukas A: **Digestive proteases of blood-feeding nematodes**. *Trends in parasitology* 2003, **19**(9):417-423.
34. Ding L, Candido EP: **Association of several small heat-shock proteins with reproductive tissues in the nematode Caenorhabditis elegans**. *The Biochemical journal* 2000, **351**(Pt 1):13-17.
35. Jacob TC, Kaplan JM: **The EGL-21 carboxypeptidase E facilitates acetylcholine release at Caenorhabditis elegans neuromuscular junctions**. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 2003, **23**(6):2122-2130.
36. Halliday AM, Lainson FA, Yaga R, Inglis NF, Bridgett S, Nath M, Knox DP: **Transcriptional changes in Teladorsagia circumcincta upon encountering host tissue of differing immune status**. *Parasitology* 2012, **139**(3):387-405.
37. Basavaraju SV, Zhan B, Kennedy MW, Liu Y, Hawdon J, Hotez PJ: **Ac-FAR-1, a 20 kDa fatty acid- and retinol-binding protein secreted by adult Ancylostoma caninum hookworms: gene transcription pattern, ligand binding properties and structural characterisation**. *Molecular and biochemical parasitology* 2003, **126**(1):63-71.
38. Smith TM, Hicks-Berger CA, Kim S, Kirley TL: **Cloning, expression, and characterization of a soluble calcium-activated nucleotidase, a human enzyme belonging to a new family of extracellular nucleotidases**. *Archives of biochemistry and biophysics* 2002, **406**(1):105-115.
39. Uccelletti D, Pascoli A, Farina F, Alberti A, Mancini P, Hirschberg CB, Palleschi C: **APY-1, a novel Caenorhabditis elegans apyrase involved in unfolded protein response signalling and stress responses**. *Molecular biology of the cell* 2008, **19**(4):1337-1345.
40. Maule AG, Bowman JW, Thompson DP, Marks NJ, Friedman AR, Geary TG: **FMRFamide-related peptides (FaRPs) in nematodes: occurrence and neuromuscular physiology**. *Parasitology* 1996, **113 Suppl**:S119-135.

41. Maryon EB, Coronado R, Anderson P: **unc-68 encodes a ryanodine receptor involved in regulating C. elegans body-wall muscle contraction**. *The Journal of cell biology* 1996, **134**(4):885-893.

42. Craig H, Wastling JM, Knox DP: **A preliminary proteomic survey of the in vitro excretory/secretory products of fourth-stage larval and adult Teladorsagia circumcincta**. *Parasitology* 2006, **132**(Pt 4):535-543.

43. Pirneskoski A, Klappa P, Lobell M, Williamson RA, Byrne L, Alanen HI, Salo KE, Kivirikko KI, Freedman RB, Ruddock LW: **Molecular characterization of the principal substrate binding site of the ubiquitous folding catalyst protein disulfide isomerase**. *The Journal of biological chemistry* 2004, **279**(11):10374-10381.

44. Riffkin MC, Seow HF, Wood PR, Brown LE, Jackson DC, Scheerlinck JP: **Trichostrongylus colubriformis extract upregulates TNF-alpha receptor expression and enhances TNF-alpha sensitivity of L929 cells**. *Immunology and cell biology* 2000, **78**(6):575-579.

45. Vervelde L, Van Leeuwen MA, Kruidenier M, Kooyman FN, Huntley JF, Van Die I, Cornelissen AW: **Protection studies with recombinant excretory/secretory proteins of Haemonchus contortus**. *Parasite immunology* 2002, **24**(4):189-201.

46. Geldhof P, Vercauteren I, Gevaert K, Staes A, Knox DP, Vandekerckhove J, Vercruysse J, Claerebout E: **Activation-associated secreted proteins are the most abundant antigens in a host protective fraction from Ostertagia ostertagi**. *Molecular and biochemical parasitology* 2003, **128**(1):111-114.

47. Goud GN, Zhan B, Ghosh K, Loukas A, Hawdon J, Dobardzic A, Deumic V, Liu S, Dobardzic R, Zook BC *et al*: **Cloning, yeast expression, isolation, and vaccine testing of recombinant Ancylostoma-secreted protein (ASP)-1 and ASP-2 from Ancylostoma ceylanicum**. *The Journal of infectious diseases* 2004, **189**(5):919-929.

48. Nisbet AJ, Redmond DL, Matthews JB, Watkins C, Yaga R, Jones JT, Nath M, Knox DP: **Stage-specific gene expression in Teladorsagia circumcincta (Nematoda: Strongylida) infective larvae and early parasitic stages**. *International journal for parasitology* 2008, **38**(7):829-838.

49. Williamson AL, Brindley PJ, Loukas A: **Hookworm cathepsin D aspartic proteases: contributing roles in the host-specific degradation of serum proteins and skin macromolecules**. *Parasitology* 2003, **126**(Pt 2):179-185.

50. Feng J, Zhan B, Liu Y, Liu S, Williamson A, Goud G, Loukas A, Hotez P: **Molecular cloning and characterization of Ac-MTP-2, an astacin-like metalloprotease released by adult Ancylostoma caninum**. *Molecular and biochemical parasitology* 2007, **152**(2):132-138.

51. Borchert N, Becker-Pauly C, Wagner A, Fischer P, Stocker W, Brattig NW: **Identification and characterization of onchoastacin, an astacin-like metalloproteinase from the filaria Onchocerca volvulus**. *Microbes and infection / Institut Pasteur* 2007, **9**(4):498-506.

52. Knox DP: **Proteinase inhibitors and helminth parasite infection**. *Parasite immunology* 2007, **29**(2):57-71.

53. Loukas A, Maizels RM: **Helminth C-type lectins and host-parasite interactions**. *Parasitol Today* 2000, **16**(8):333-339.

54. Hewitson JP, Grainger JR, Maizels RM: **Helminth immunoregulation: the role of parasite secreted proteins in modulating host immunity**. *Molecular and biochemical parasitology* 2009, **167**(1):1-11.

55. Loukas A, Doedens A, Hintz M, Maizels RM: **Identification of a new C-type lectin, TES-70, secreted by infective larvae of Toxocara canis, which binds to host ligands**. *Parasitology* 2000, **121 Pt 5**:545-554.

56. Loukas A, Mullin NP, Tetteh KK, Moens L, Maizels RM: **A novel C-type lectin secreted by a tissue-dwelling parasitic nematode**. *Current biology : CB* 1999, **9**(15):825-828.

57. Dalton JP, Hola-Jamriska L, Brindley PJ: **Asparaginyl endopeptidase activity in adult Schistosoma mansoni**. *Parasitology* 1995, **111 (Pt 5)**:575-580.

58. Oliver EM, Skuce PJ, McNair CM, Knox DP: **Identification and characterization of an asparaginyl proteinase (legumain) from the parasitic nematode, Haemonchus contortus**. *Parasitology* 2006, **133**(Pt 2):237-244.

59. Liu F, Cui SJ, Hu W, Feng Z, Wang ZQ, Han ZG: **Excretory/secretory proteome of the adult developmental stage of human blood fluke, Schistosoma japonicum**. *Molecular & cellular proteomics : MCP* 2009, **8**(6):1236-1251.

60. Ke N, Claassen G, Yu DH, Albers A, Fan W, Tan P, Grifman M, Hu X, Defife K, Nguy V *et al*: **Nuclear hormone receptor NR4A2 is involved in cell transformation and apoptosis**. *Cancer research* 2004, **64**(22):8208-8212.

61. Antebi A: **Nuclear hormone receptors in C. elegans**. *WormBook : the online review of C elegans biology* 2006:1-13.

62. Karanu FN, Rurangirwa FR, McGuire TC, Jasmer DP: **Haemonchus contortus: identification of proteases with diverse characteristics in adult worm excretory-secretory products**. *Experimental parasitology* 1993, **77**(3):362-371.

63. Kovaleva ES, Masler EP, Skantar AM, Chitwood DJ: **Novel matrix metalloproteinase from the cyst nematodes Heterodera glycines and Globodera rostochiensis**. *Molecular and biochemical parasitology* 2004, **136**(1):109-112.

64. Yatsuda AP, Bakker N, Krijgsveld J, Knox DP, Heck AJ, de Vries E: **Identification of secreted cysteine proteases from the parasitic nematode Haemonchus contortus detected by biotinylated inhibitors**. *Infection and immunity* 2006, **74**(3):1989-1993.

65. Zheng Y, Zhao L, Gao J, Fei Z: **iAssembler: a package for de novo assembly of Roche-454/Sanger transcriptome sequences**. *BMC bioinformatics* 2011, **12**:453.

66. Iseli C, Jongeneel CV, Bucher P: **ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences**. *Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB International Conference on Intelligent Systems for Molecular Biology* 1999:138-148.

67. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R: **InterProScan: protein domains identifier**. *Nucleic acids research* 2005, **33**(Web Server issue):W116-120.

68. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li CY, Wei L: **KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases**. *Nucleic Acids Res* 2011, **39**(Web Server issue):W316-322.

69. Bendtsen JD, Jensen LJ, Blom N, Von Heijne G, Brunak S: **Feature-based prediction of non-classical and leaderless protein secretion**. *Protein Eng Des Sel* 2004, **17**(4):349-356.

70. Emanuelsson O, Brunak S, von Heijne G, Nielsen H: **Locating proteins in the cell using TargetP, SignalP and related tools**. *Nature protocols* 2007, **2**(4):953-971.

Tables and Captions

Table 1: Top 30 KEGG pathways in L4 *T. circumcincta* protein sequences.

KEGG PATHWAY	No. of sequences
Purine metabolism	67
Protein processing in endoplasmic reticulum	60
Ribosome	53
Oxidative phosphorylation	52
Huntington's disease	51
Spliceosome	50
RNA transport	46
Lysosome	44
Alzheimer's disease	43
Parkinson's disease	40
Valine, leucine and isoleucine degradation	35
Peroxisome	34
Ubiquitin mediated proteolysis	29
Proteasome	29
Endocytosis	29
Ribosome biogenesis in eukaryotes	28
Pathways in cancer	28
Aminoacyl-tRNA biosynthesis	27
Phagosome	27
Glycolysis / Gluconeogenesis	26
HTLV-I infection	26
Fatty acid metabolism	24
Insulin signaling pathway	24
Citrate cycle (TCA cycle)	23
Regulation of actin cytoskeleton	23
Amino sugar and nucleotide sugar metabolism	22
Propanoate metabolism	22
Focal adhesion	22
mRNA surveillance pathway	21
Wnt signaling pathway	21

Table 2: Top 30 metabolic pathways mapped by Kyoto Encyclopedia of Genes and Genomes in L4 *T. circumcincta* ES proteins

KEGG PATHWAY	No. of sequences
RNA transport	3
Protein processing in endoplasmic reticulum	3
RNA degradation	3
Lysosome	3
Citrate cycle	2
Nitrogen metabolism	2
Glycine, serine and threonine metabolism	2
Lysine degradation	2
Ribosome biogenesis in eukaryotes	2
Tight junction	2
Antigen processing and presentation	2
Insulin signaling pathway	2
Alzheimer's disease	2
Huntington's disease	2
Viral myocarditis	2
Glycolysis / Gluconeogenesis	1
Galactose metabolism	1
Starch and sucrose metabolism	1
Pyruvate metabolism	1
Propanoate metabolism	1
Inositol phosphate metabolism	1
Oxidative phosphorylation	1
Methane metabolism	1
Fatty acid biosynthesis	1
Purine metabolism	1
Pyrimidine metabolism	1
Alanine, aspartate and glutamate metabolism	1
Cysteine and methionine metabolism	1
Arginine and proline metabolism	1
Tryptophan metabolism	1

List of additional files (available on CD attached)

Additional File 1: Pathway Analysis of *Teladorsagia circumcincta* L4 proteins

Additional File 2: Top BLAST hits of *Teladorsagia circumcincta* L4 protein sequences

Additional File 3: GO Annotation for *Teladorsagia circumcincta* L4 proteins

Additional file 4: Top BLAST hits of *Teladorsagia circumcincta* L4 ES sequences

Additional File 5: GO Annotation for *Teladorsagia circumcincta* L4 ES proteins

Additional File 6: Pathway Analysis of *Teladorsagia circumcincta* L4 secreted proteins

6.2 Conclusions

Large-scale analysis helped in comparison of adult and fourth larval stage *Teladorsagia circumcincta*. The most abundant molecule in excretory/secretory products released *in vitro* by *T. circumcincta,* cathepsin F was identified in fourth larval stage. Other important identifications in both adult and fourth larval stage include the cathepsin B-like cysteine protease, the leading vaccine candidate and C-type lectin family members in *T. circumcincta* larvae exposed to a naïve host environment. The proteins that represent the important targets for parasite invention and control, serine-, cysteine-proteases as well as proteinase inhibitors were identified in both adult and fourth larval stage. Galectins which were identified in adult stage was not identified in fourth stage while legumain of *Haemonchus contortus,* the novel class of cysteine proteinase was identified in fourth larval stage were not identified in adult *Teladorsagia circumcincta* analysis. Our results provide a step for future research on disease manifestation and molecular biology of the parasite and also future studies integrating proteomics and metabolomic studies for identifying novel intervention and control strategies.

Chapter 7: Conclusions and future directions

7.1 Summary

ESTs accredited as one of the richest sources for discoveries in modern biology with a wide range of application in genomics, transcriptomics and proteomics, with a range of computational methods and algorithms for the analysis.

A review of different tools and pipelines in EST and NGS data analysis was necessary for the development of new analysis pipeline with new and updated bioinformatic tools. With the introduction of Next-generation sequencing, short read sequences are generated extensively for different organisms including the new ones.

Analysis of *Fasciola hepatica* for secretory proteins involved in host-pathogen interactions was carried out with the existing pipeline which identified novel proteases with important roles in host-parasite interplay, mainly cathepsin L and cathepsin B cysteine proteases and carboxypeptidase and trypsin-like serine proteases, the members of two serine protease families. Abundant and vastly regulated antioxidants released via a non-classical trans-tegumental pathway were identified which revealed the requirement to integrate the bioinformatic tool, SecretomeP, for the identification of non-classically secreted proteins. The *F. hepatica* analysis with the focus on protein secretion using the transcriptomic and proteomic approach represents a significant step in understanding the host-parasite interactions in fasciolosis.

The benchmarking of the tools were carried out with the large-scale analysis of *Teladorsagia circumcincta* dataset (407,357 ESTs) with an emphasis on description of molecules inferred to be ES proteins as possible immunogens and vaccine candidates. The large-scale analysis paved the way for future research with the concentration of study on disease and molecular biology of the parasite along with studies to identify novel disease control stratergies.

A robust analysis pipeline, TranSeqAnnotator was developed for large-scale analysis with different phases including the ESTs, short reads, proteins and ES sequence analysis. The annotation carried out for all the phases and a summary of independent analysis is

provided. The efficiency of TranSeqAnnotator is analysed using the *Ascaris lumbricoides* dataset.

Studies on ES proteins across the clade in phylum Nematoda have proposed them as potential therapeutic targets [41, 163, 164]. Also, our study on *F. hepatica* by integrating transcriptomic and proteomic studies have led to the identification of more novel proteases. These findings made us expand our study on large-scale dataset of *T. circumcincta*. The large-scale NGS data analysis helped to benchmark the new and updated bioinformatic tools with vast amounts of EST data including the short read sequences and accurately predict and analyse ES proteins. Understanding the role of ES proteins in helminths and also, the development of novel anthelmintics are the highlights of the study. We have enabled experimental biologists to carry out a detailed and directed functional annotation in understanding the complexities of host-parasite interactions in a cell with the development of a robust analysis pipeline.

7.2 Significance and contributions

The study and analysis of EST dataset is reviewed based on the currently available methods and pipelines and written in the form of a review article. Based on this review, we have developed TranSeqAnnotator, a comprehensive workflow system for EST and NGS data management and analysis. The *in silico* approach in the identification of several important proteases during EST analysis of parasitic nematode *F. hepatica* supported by experimental validation and hence represented as an integrated approach of transcriptomics and proteomics. The large-scale analysis from *Teladorsagia circumcincta* helped in the development of TranSeqAnnotator to facilitate better understanding of parasite biology and for the development of novel drugs or vaccines for parasite intervention and control.

7.3 Future directions

In this section, we present the extention of the research work carried out in this thesis. The TranSeqAnnotator could be developed with additional features addressing the Next-generation sequencing, the newly developed technology. The pipeline can be extended for microarray analysis, new tools could be incorporated for SNP discovery like autoSNP [165, 166], and SNPServer [167], and to construct a relational database to enable the efficient mining of the identified polymorphisms. Other areas to explore would be the gene expression data and identification of biomarkers. Include short read aligner tool like bowtie [168]. Recent advances in sequencing technology now produces 200-500 million reads per

lane with Illumina HiSeq leading to an increasing accumulation of transcriptomic (RNAseq) data. The main application of RNA-Seq is in the discovery of new transcripts, transcript boundaries, splice junctions. Also, helps to learn about expression outside of annotated exons, alternative splicing along with which genes are up and down regulated in different cell lines [169]. Study of alternative splicing, gene discovery projects and comparative genomics is carried out with the initial study of the alignment of EST sequences to their associated or related genomic sequences. We here have reported several novel candidates for parasite control and intervention with the focus on several ES proteins which could help in the identification of novel therapeutic targets. The experimental verification would help in characterization of the molecules in detail.

References

1. Weinstock JV, Elliott DE: **Helminths and the IBD hygiene hypothesis.** *Inflammatory Bowel Diseases* 2009, **15:**128-133.
2. Castro GA: **Helminths: Structure, Classification, Growth, and Development.** In: Baron S, editor. *Medical Microbiology*. 4th edition. Galveston (TX): University of Texas Medical Branch at Galveston; 1996. Chapter 86.
3. Robinson MW, Hutchinson AT, Donnelly S, Dalton JP: **Worm secretory molecules are causing alarm.** *Trends in Parasitology* 2010, **26:**371-372.
4. Bourke CD, Maizels RM, Mutapi F: **Acquired immune heterogeneity and its sources in human helminth infection.** *Parasitology*, **138:**139-159.
5. Moreno Y, Geary TG: **Stage- and gender-specific proteomic analysis of *Brugia malayi* excretory-secretory products.** *PLoS Neglected Tropical Diseases* 2008, **2:**e326.
6. Jolly ER, Chin CS, Miller S, Bahgat MM, Lim KC, DeRisi J, McKerrow JH: **Gene expression patterns during adaptation of a helminth parasite to different environmental niches.** *Genome Biology* 2007, **8:**R65.
7. Brindley PJ, Mitreva M, Ghedin E, Lustigman S: **Helminth genomics: The implications for human health.** *PLoS Neglected Tropical Diseases* 2009, **3:**e538.
8. Blaxter M: **Nematodes: the worm and its relatives.** *PLoS Biology* 2011, **9:**e1001050.
9. Blaxter ML: **Nematoda: genes, genomes and the evolution of parasitism.** *Advances in Parasitology* 2003, **54:**101-195.
10. Coghlan A: **Nematode genome evolution.** *WormBook: the online review of C elegans biology* 2005**:**1-15.
11. Leroy S, Duperray C, Morand S: **Flow cytometry for parasite nematode genome size measurement.** *Molecular and Biochemical Parasitology* 2003, **128:**91-93.
12. Witherspoon DJ, Robertson HM: **Neutral evolution of ten types of mariner transposons in the genomes of *Caenorhabditis elegans* and *Caenorhabditis briggsae*.** *Journal of Molecular Evolution* 2003, **56:**751-769.
13. Blaxter ML, De Ley P, Garey JR, Liu LX, Scheldeman P, Vierstraete A, Vanfleteren JR, Mackey LY, Dorris M, Frisse LM, et al: **A molecular evolutionary framework for the phylum Nematoda.** *Nature* 1998, **392:**71-75.

14. Wasmuth J, Schmid R, Hedley A, Blaxter M: **On the extent and origins of genic novelty in the phylum Nematoda.** *PLoS Neglected Tropical Diseases* 2008, **2:**e258.
15. Koenning SR, Overstreet C, Noling JW, Donald PA, Becker JO, Fortnum BA: **Survey of crop losses in response to phytoparasitic nematodes in the United States for 1994.** *Journal of Nematology* 1999, **31:**587-618.
16. Engels D, Savioli L: **Reconsidering the underestimated burden caused by neglected tropical diseases.** *Trends in Parasitology* 2006, **22:**363-366.
17. Hotez PJ, Molyneux DH, Fenwick A, Kumaresan J, Sachs SE, Sachs JD, Savioli L: **Control of neglected tropical diseases.** *The New England Journal of Medicine* 2007, **357:**1018-1027.
18. Yamey G, Hotez P: **Neglected tropical diseases.** *British Medical Journal* 2007, **335:**269-270.
19. Blouin MS, Liu J, Berry RE: **Life cycle variation and the genetic structure of nematode populations.** *Heredity* 1999, **83 (Pt 3):**253-259.
20. Poulin R, Krasnov BR, Mouillot D, Thieltges DW: **The comparative ecology and biogeography of parasites.** *Philosophical Transactions of the Royal Society of London B Biological Science*, **366:**2379-2390.
21. Holden-Dye L, Walker RJ: **Anthelmintic drugs.** *WormBook : the online review of C elegans biology* 2007**:**1-13.
22. Kaminsky R, Ducray P, Jung M, Clover R, Rufener L, Bouvier J, Weber SS, Wenger A, Wieland-Berghausen S, Goebel T, et al: **A new class of anthelmintics effective against drug-resistant nematodes.** *Nature* 2008, **452:**176-180.
23. Bourke CD, Maizels RM, Mutapi F: **Acquired immune heterogeneity and its sources in human helminth infection.** *Parasitology* 2011, **138:**139-159.
24. Geary TG, Sangster NC, Thompson DP: **Frontiers in anthelmintic pharmacology.** *Veterinary Parasitology* 1999, **84:**275-295.
25. Wolstenholme AJ, Fairweather I, Prichard R, von Samson-Himmelstjerna G, Sangster NC: **Drug resistance in veterinary helminths.** *Trends in Parasitology* 2004, **20:**469-476.
26. Kaplan RM, Vidyashankar AN, Howell SB, Neiss JM, Williamson LH, Terrill TH: **A novel approach for combining the use of *in vitro* and *in vivo* data to measure and detect emerging moxidectin resistance in gastrointestinal nematodes of goats.** *International Journal for Parasitology* 2007, **37:**795-804.

27. Lespine A, Alvinerie M, Vercruysse J, Prichard RK, Geldhof P: **ABC transporter modulation: a strategy to enhance the activity of macrocyclic lactone anthelmintics.** *Trends in Parasitology* 2008, **24:**293-298.
28. Sutherland IA, Bailey J, Shaw RJ: **The production costs of anthelmintic resistance in sheep managed within a monthly preventive drench program.** *Veterinary Parasitology* 2010, **171:**300-304.
29. Redman E, Packard E, Grillo V, Smith J, Jackson F, Gilleard JS: **Microsatellite analysis reveals marked genetic differentiation between *Haemonchus contortus* laboratory isolates and provides a rapid system of genetic fingerprinting.** *International Journal for Parasitology* 2008, **38:**111-122.
30. Behnke JM, Barnard CJ, Wakelin D: **Understanding chronic nematode infections: evolutionary considerations, current hypotheses and the way forward.** *International Journal for Parasitology* 1992, **22:**861-907.
31. Pearce EJ, Sher A: **Mechanisms of immune evasion in schistosomiasis.** *Contributions to Microbiology and Immunology* 1987, **8:**219-232.
32. MacDonald AS, Araujo MI, Pearce EJ: **Immunology of parasitic helminth infections.** *Infection and Immunity* 2002, **70:**427-433.
33. Hedeler C, Paton NW, Behnke JM, Bradley JE, Hamshere MG, Else KJ: **A classification of tasks for the systematic study of immune response using functional genomics data.** *Parasitology* 2006, **132:**157-167.
34. Dalton JP, Mulcahy G: **Parasite vaccines--a reality?** *Veterinary Parasitology* 2001, **98:**149-167.
35. Vercruysse J, Knox DP, Schetters TP, Willadsen P: **Veterinary parasitic vaccines: pitfalls and future directions.** *Trends in Parasitology* 2004, **20:**488-492.
36. Omura S: **Ivermectin: 25 years and still going strong.** *International Journal of Antimicrobial Agents* 2008, **31:**91-98.
37. Smith WD, Zarlenga DS: **Developments and hurdles in generating vaccines for controlling helminth parasites of grazing ruminants.** *Veterinary Parasitology* 2006, **139:**347-359.
38. Knox DP, Smith WD: **Vaccination against gastrointestinal nematode parasites of ruminants using gut-expressed antigens.** *Veterinary Parasitology* 2001, **100:**21-32.

39. Willadsen P, Bird P, Cobon GS, Hungerford J: **Commercialisation of a recombinant vaccine against Boophilus microplus.** *Parasitology* 1995, **110 Suppl:**S43-50.

40. Hewitson JP, Grainger JR, Maizels RM: **Helminth immunoregulation: the role of parasite secreted proteins in modulating host immunity.** *Molecular and Biochemical Parasitology* 2009, **167:**1-11.

41. Lightowlers MW, Rickard MD: **Excretory-secretory products of helminth parasites: effects on host immune responses.** *Parasitology* 1988, **96 Suppl:**S123-166.

42. Dalton JP, Brindley PJ, Knox DP, Brady CP, Hotez PJ, Donnelly S, O'Neill SM, Mulcahy G, Loukas A: **Helminth vaccines: from mining genomic information for vaccine targets to systems used for protein expression.** *International Journal for Parasitology* 2003, **33:**621-640.

43. Hartmann S, Lucius R: **Modulation of host immune responses by nematode cystatins.** *International Journal for Parasitology* 2003, **33:**1291-1302.

44. Pearson MS, Ranjit N, Loukas A: **Blunting the knife: development of vaccines targeting digestive proteases of blood-feeding helminth parasites.** *Biological Chemistry* 2010, **391:**901-911.

45. Bird DM, Opperman CH: **The secret(ion) life of worms.** *Genome Biology* 2009, **10:**205.

46. Robinson MW, Menon R, Donnelly SM, Dalton JP, Ranganathan S: **An integrated transcriptomics and proteomics analysis of the secretome of the helminth pathogen *Fasciola hepatica*: proteins associated with invasion and infection of the mammalian host.** *Molecular and Cellular Proteomics* 2009, **8:**1891-1907.

47. Choo KH, Tan TW, Ranganathan S: **A comprehensive assessment of N-terminal signal peptides prediction methods.** *BMC Bioinformatics* 2009, **10 Suppl 15:**S2.

48. Kall L: **Prediction of transmembrane topology and signal peptide given a protein's amino acid sequence.** *Methods in Molecular Biology* 2010, **673:**53-62.

49. Frank K, Sippl MJ: **High-performance signal peptide prediction based on sequence alignment techniques.** *Bioinformatics* 2008, **24:**2172-2176.

50. Nickel W: **The mystery of nonclassical protein secretion. A current view on cargo proteins and potential export routes.** *European Journal of Biochemistry* 2003, **270:**2109-2119.

51. Nickel W: **Unconventional secretory routes: direct protein export across the plasma membrane of mammalian cells.** *Traffic* 2005, **6:**607-614.

52. Frith MC, Pheasant M, Mattick JS: **The amazing complexity of the human transcriptome.** *European Journal of Human Genetics* 2005, **13:**894-897.
53. Parkinson J, Blaxter M: **Expressed sequence tags: an overview.** *Methods in Molecular Biology* 2009, **533:**1-12.
54. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, et al.: **Complementary DNA sequencing: expressed sequence tags and human genome project.** *Science* 1991, **252:**1651-1656.
55. Bonaldo MF, Lennon G, Soares MB: **Normalization and subtraction: two approaches to facilitate gene discovery.** *Genome Research* 1996, **6:**791-806.
56. Rudd S: **Expressed sequence tags: alternative or complement to whole genome sequences?** *Trends in Plant Science* 2003, **8:**321-329.
57. Dong Q, Kroiss L, Oakley FD, Wang BB, Brendel V: **Comparative EST analyses in plant systems.** *Methods in Enzymology* 2005, **395:**400-418.
58. Jongeneel CV: **Searching the expressed sequence tag (EST) databases: panning for genes.** *Briefings in Bioinformatics* 2000, **1:**76-92.
59. Morozova O, Marra MA: **Applications of next-generation sequencing technologies in functional genomics.** *Genomics* 2008, **92:**255-264.
60. Metzker ML: **Sequencing technologies - the next generation.** *Nature Reviews Genetics* 2010, **11:**31-46.
61. Zhang J, Chiodini R, Badr A, Zhang G: **The impact of next-generation sequencing on genomics.** *Journal of Genetics and Genomics* 2011, **38:**95-109.
62. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005, **437:**376-380.
63. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al: **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature* 2008, **456:**53-59.
64. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, Dimeo J, Efcavitch JW, et al: **Single-molecule DNA sequencing of a viral genome.** *Science* 2008, **320:**106-109.
65. Daly AK: **Genome-wide association studies in pharmacogenomics.** *Nature Reviews Genetics* 2010, **11:**241-246.

66. Chen XS, Collins LJ, Biggs PJ, Penny D: **High throughput genome-wide survey of small RNAs from the parasitic protists *Giardia intestinalis* and *Trichomonas vaginalis*.** *Genome Biology and Evolution* 2009, **1:**165-175.

67. Franzen O, Jerlstrom-Hultqvist J, Castro E, Sherwood E, Ankarklev J, Reiner DS, Palm D, Andersson JO, Andersson B, Svard SG: **Draft genome sequencing of *Giardia intestinalis* assemblage B isolate GS: is human giardiasis caused by two different species?** *PLoS Pathogens* 2009, **5:**e1000560.

68. Otto TD, Wilinski D, Assefa S, Keane TM, Sarry LR, Bohme U, Lemieux J, Barrell B, Pain A, Berriman M, et al: **New insights into the blood-stage transcriptome of *Plasmodium falciparum* using RNA-Seq.** *Molecular Microbiology* 2010, **76:**12-24.

69. Tettelin H: **The bacterial pan-genome and reverse vaccinology.** *Genome Dynamics* 2009, **6:**35-47.

70. Kuroda M, Katano H, Nakajima N, Tobiume M, Ainai A, Sekizuka T, Hasegawa H, Tashiro M, Sasaki Y, Arakawa Y, et al: **Characterization of quasispecies of pandemic 2009 influenza A virus (A/H1N1/2009) by de novo sequencing using a next-generation DNA sequencer.** *PloS ONE* 2010, **5:**e10256.

71. Wang XW, Luan JB, Li JM, Bao YY, Zhang CX, Liu SS: ***De novo* characterization of a whitefly transcriptome and analysis of its gene expression during development.** *BMC Genomics* 2010, **11:**400.

72. Wang Z, Abubucker S, Martin J, Wilson RK, Hawdon J, Mitreva M: **Characterizing *Ancylostoma caninum* transcriptome and exploring nematode parasitic adaptation.** *BMC Genomics* 2010, **11:**307.

73. Jex AR, Hu M, Littlewood DT, Waeschenbach A, Gasser RB: **Using 454 technology for long-PCR based sequencing of the complete mitochondrial genome from single *Haemonchus contortus* (Nematoda).** *BMC Genomics* 2008, **9:**11.

74. Young ND, Campbell BE, Hall RS, Jex AR, Cantacessi C, Laha T, Sohn WM, Sripa B, Loukas A, Brindley PJ, Gasser RB: **Unlocking the transcriptomes of two carcinogenic parasites, *Clonorchis sinensis* and *Opisthorchis viverrini*.** *PLoS Neglected Tropical Diseases* 2010, **4:**e719.

75. Young ND, Hall RS, Jex AR, Cantacessi C, Gasser RB: **Elucidating the transcriptome of *Fasciola hepatica* - a key to fundamental and biotechnological discoveries for a neglected parasite.** *Biotechnology Advances* 2010, **28:**222-231.

76. Novelli G, Predazzi IM, Mango R, Romeo F, Mehta JL: **Role of genomics in cardiovascular medicine.** *World Journal of Cardiology* 2010, **2:**428-436.

77. Gonzalez-Angulo AM, Hennessy BT, Mills GB: **Future of personalized medicine in oncology: a systems biology approach.** *Journal of Clinical Oncology* 2010, **28:**2777-2783.

78. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, et al: **The complete genome of an individual by massively parallel DNA sequencing.** *Nature* 2008, **452:**872-876.

79. Wold B, Myers RM: **Sequence census methods for functional genomics.** *Nature Methods* 2008, **5:**19-21.

80. Yang MQ, Athey BD, Arabnia HR, Sung AH, Liu Q, Yang JY, Mao J, Deng Y: **High-throughput next-generation sequencing technologies foster new cutting-edge computing techniques in bioinformatics.** *BMC Genomics* 2009, **10 Suppl 1:**I1.

81. Cantacessi C, Jex AR, Hall RS, Young ND, Campbell BE, Joachim A, Nolan MJ, Abubucker S, Sternberg PW, Ranganathan S, et al: **A practical, bioinformatic workflow system for large data sets generated by next-generation sequencing.** *Nucleic Acids Research* 2010, **38:**e171.

82. Cantacessi C, Campbell BE, Young ND, Jex AR, Hall RS, Presidente PJ, Zawadzki JL, Zhong W, Aleman-Meza B, Loukas A, et al: **Differences in transcription between free-living and CO2-activated third-stage larvae of *Haemonchus contortus*.** *BMC Genomics* 2010, **11:**266.

83. Cantacessi C, Mitreva M, Jex AR, Young ND, Campbell BE, Hall RS, Doyle MA, Ralph SA, Rabelo EM, Ranganathan S, et al: **Massively parallel sequencing and analysis of the *Necator americanus* transcriptome.** *PLoS Neglected Tropical Diseases* 2010, **4:**e684.

84. Cantacessi C, Mitreva M, Campbell BE, Hall RS, Young ND, Jex AR, Ranganathan S, Gasser RB: **First transcriptomic analysis of the economically important parasitic nematode, *Trichostrongylus colubriformis*, using a next-generation sequencing approach.** *Infection, Genetics and Evolution* 2010, **10:**1199-1207.

85. **Genome sequence of the nematode *C. elegans*: a platform for investigating biology.** *Science* 1998, **282:**2012-2018.

86. Brenner S: **The genetics of *Caenorhabditis elegans*.** *Genetics* 1974, **77:**71-94.

87. Cottee PA, Nisbet AJ, Abs El-Osta YG, Webster TL, Gasser RB: **Construction of gender-enriched cDNA archives for adult *Oesophagostomum dentatum* by suppressive-subtractive hybridization and a microarray analysis of expressed sequence tags.** *Parasitology* 2006, **132:**691-708.

88. Mitreva M, Appleton J, McCarter JP, Jasmer DP: **Expressed sequence tags from life cycle stages of *Trichinella spiralis*: application to biology and parasite control.** *Veterinary Parasitology* 2005, **132:**13-17.

89. Li BW, Rush AC, Crosby SD, Warren WC, Williams SA, Mitreva M, Weil GJ: **Profiling of gender-regulated gene transcripts in the filarial nematode *Brugia malayi* by cDNA oligonucleotide array analysis.** *Molecular and Biochemical Parasitology* 2005, **143:**49-57.

90. Yin Y, Martin J, McCarter JP, Clifton SW, Wilson RK, Mitreva M: **Identification and analysis of genes expressed in the adult filarial parasitic nematode *Dirofilaria immitis*.** *International Journal for Parasitology* 2006, **36:**829-839.

91. Ranganathan S, Nagaraj SH, Hu M, Strube C, Schnieder T, Gasser RB: **A transcriptomic analysis of the adult stage of the bovine lungworm, *Dictyocaulus viviparus*.** *BMC Genomics* 2007, **8:**311.

92. McCarter JP, Mitreva MD, Martin J, Dante M, Wylie T, Rao U, Pape D, Bowers Y, Theising B, Murphy CV, et al: **Analysis and functional classification of transcripts from the nematode *Meloidogyne incognita*.** *Genome Biology* 2003, **4:**R26.

93. Jacob J, Mitreva M, Vanholme B, Gheysen G: **Exploring the transcriptome of the burrowing nematode *Radopholus similis*.** *Molecular Genetics and Genomics* 2008, **280:**1-17.

94. Thompson FJ, Mitreva M, Barker GL, Martin J, Waterston RH, McCarter JP, Viney ME: **An expressed sequence tag analysis of the life-cycle of the parasitic nematode *Strongyloides ratti*.** *Molecular and Biochemical Parasitology* 2005, **142:**32-46.

95. Ghedin E, Wang S, Spiro D, Caler E, Zhao Q, Crabtree J, Allen JE, Delcher AL, Guiliano DB, Miranda-Saavedra D, et al: **Draft genome of the filarial nematode parasite *Brugia malayi*.** *Science* 2007, **317:**1756-1760.

96. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC: **Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*.** *Nature* 1998, **391:**806-811.

97. Barstead R: **Genome-wide RNAi.** *Current opinion in Chemical Biology* 2001, **5:**63-66.

98. Kamath RS, Ahringer J: **Genome-wide RNAi screening in *Caenorhabditis elegans*.** *Methods* 2003, **30:**313-321.

99. Simmer F, Moorman C, van der Linden AM, Kuijk E, van den Berghe PV, Kamath RS, Fraser AG, Ahringer J, Plasterk RH: **Genome-wide RNAi of *C. elegans* using the hypersensitive rrf-3 strain reveals novel gene functions.** *PLoS Biology* 2003, **1:**E12.

100. Sugimoto A: **High-throughput RNAi in *Caenorhabditis elegans*: genome-wide screens and functional genomics.** *Differentiation* 2004, **72:**81-91.

101. Sonnichsen B, Koski LB, Walsh A, Marschall P, Neumann B, Brehm M, Alleaume AM, Artelt J, Bettencourt P, Cassin E, et al: **Full-genome RNAi profiling of early embryogenesis in *Caenorhabditis elegans*.** *Nature* 2005, **434:**462-469.

102. Geldhof P, Visser A, Clark D, Saunders G, Britton C, Gilleard J, Berriman M, Knox D: **RNA interference in parasitic helminths: current situation, potential pitfalls and future prospects.** *Parasitology* 2007, **134:**609-619.

103. Yook K, Harris TW, Bieri T, Cabunoc A, Chan J, Chen WJ, Davis P, de la Cruz N, Duong A, Fang R, et al: **WormBase 2012: more genomes, more data, new website.** *Nucleic Acids Research* 2012, **40:**D735-741.

104. Kumar S, Chaudhary K, Foster JM, Novelli JF, Zhang Y, Wang S, Spiro D, Ghedin E, Carlow CK: **Mining predicted essential genes of *Brugia malayi* for nematode drug targets.** *PloS ONE* 2007, **2:**e1189.

105. Stinchcomb DT, Shaw JE, Carr SH, Hirsh D: **Extrachromosomal DNA transformation of *Caenorhabditis elegans*.** *Molecular and Cellular Biology* 1985, **5:**3484-3496.

106. Fire A: **Integrative transformation of *Caenorhabditis elegans*.** *The EMBO Journal* 1986, **5:**2673-2680.

107. Reinke V, Smith HE, Nance J, Wang J, Van Doren C, Begley R, Jones SJ, Davis EB, Scherer S, Ward S, Kim SK: **A global profile of germline gene expression in *C. elegans*.** *Molecular Cell* 2000, **6:**605-616.

108. Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, Eizinger A, Wylie BN, Davidson GS: **A gene expression map for *Caenorhabditis elegans*.** *Science* 2001, **293:**2087-2092.

109. Jiang M, Ryu J, Kiraly M, Duke K, Reinke V, Kim SK: **Genome-wide analysis of developmental and sex-regulated gene expression profiles in *Caenorhabditis***

***elegans*.** *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98:**218-223.

110. Nagaraj SH, Gasser RB, Ranganathan S: **A hitchhiker's guide to expressed sequence tag (EST) analysis.** *Briefings in Bioinformatics* 2007, **8:**6-21.

111. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, et al: **InterPro: the integrative protein signature database.** *Nucleic Acids Research* 2009, **37:**D211-215.

112. Soderlund C, Johnson E, Bomhoff M, Descour A: **PAVE: program for assembling and viewing ESTs.** *BMC Genomics* 2009, **10:**400.

113. Ranganathan S, Menon R, Gasser RB: **Advanced *in silico* analysis of expressed sequence tag (EST) data for parasitic nematodes of major socio-economic importance--fundamental insights toward biotechnological outcomes.** *Biotechnology Advances* 2009, **27:**439-448.

114. Falgueras J, Lara AJ, Fernandez-Pozo N, Canton FR, Perez-Trabado G, Claros MG: **SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read.** *BMC Bioinformatics* 2010, **11:**38.

115. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Research* 1997, **25:**3389-3402.

116. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL: **NCBI BLAST: a better web interface.** *Nucleic Acids Research* 2008, **36:**W5-9.

117. Bedell JA, Korf I, Gish W: **MaskerAid: a performance enhancement to RepeatMasker.** *Bioinformatics* 2000, **16:**1040-1041.

118. Malde K, Schneeberger K, Coward E, Jonassen I: **RBR: library-less repeat detection for ESTs.** *Bioinformatics* 2006, **22:**2232-2236.

119. Burke J, Davison D, Hide W: **d2_cluster: a validated method for clustering EST and full-length cDNAsequences.** *Genome Research* 1999, **9:**1135-1142.

120. Ptitsyn A, Hide W: **CLU: a new algorithm for EST clustering.** *BMC Bioinformatics* 2005, **6 Suppl 2:**S3.

121. Miller RT, Christoffels AG, Gopalakrishnan C, Burke J, Ptitsyn AA, Broveak TR, Hide WA: **A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base.** *Genome Research* 1999, **9:**1143-1155.

122. Lee Y, Tsai J, Sunkara S, Karamycheva S, Pertea G, Sultana R, Antonescu V, Chan A, Cheung F, Quackenbush J: **The TIGR Gene Indices: clustering and**

assembling EST and known genes and integration with eukaryotic genomes. *Nucleic Acids Research* 2005, **33:**D71-74.

123. Huang X, Madan A: **CAP3: A DNA sequence assembly program.** *Genome Research* 1999, **9:**868-877.

124. Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J: **An optimized protocol for analysis of EST sequences.** *Nucleic Acids Research* 2000, **28:**3657-3665.

125. Myers EW: **Toward simplifying and accurately formulating fragment assembly.** *Journal of Computational Biology* 1995, **2:**275-290.

126. Idury RM, Waterman MS: **A new algorithm for DNA sequence assembly.** *Journal of Computational Biology* 1995, **2:**291-306.

127. Zerbino DR, Birney E: **Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs.** *Genome Research* 2008, **18:**821-829.

128. Huang X, Wang J, Aluru S, Yang SP, Hillier L: **PCAP: a whole-genome assembly program.** *Genome Research* 2003, **13:**2164-2170.

129. Li R, Li Y, Kristiansen K, Wang J: **SOAP: short oligonucleotide alignment program.** *Bioinformatics* 2008, **24:**713-714.

130. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I: **ABySS: a parallel assembler for short read sequence data.** *Genome Research* 2009, **19:**1117-1123.

131. Hernandez D, Francois P, Farinelli L, Osteras M, Schrenzel J: **De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer.** *Genome Research* 2008, **18:**802-809.

132. Pevzner PA, Tang H, Waterman MS: **An Eulerian path approach to DNA fragment assembly.** *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98:**9748-9753.

133. Min XJ, Butler G, Storms R, Tsang A: **OrfPredictor: predicting protein-coding regions in EST-derived sequences.** *Nucleic Acids Research* 2005, **33:**W677-680.

134. Iseli C, Jongeneel CV, Bucher P: **ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences.** *Proceedings / International Conference on Intelligent Systems for Molecular Biology* 1999**:**138-148.

135. Fukunishi Y, Hayashizaki Y: **Amino acid translation program for full-length cDNA sequences with frameshift errors.** *Physiological Genomics* 2001, **5:**81-87.

136. Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K, Bairoch A: **The PROSITE database, its status in 2002.** *Nucleic Acids Research* 2002, **30:**235-238.

137. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, et al: **The Pfam protein families database.** *Nucleic Acids Research* 2010, **38:**D211-222.

138. Attwood TK, Croning MD, Flower DR, Lewis AP, Mabey JE, Scordis P, Selley JN, Wright W: **PRINTS-S: the database formerly known as PRINTS.** *Nucleic Acids Research* 2000, **28:**225-227.

139. Corpet F, Gouzy J, Kahn D: **Recent improvements of the ProDom database of protein domain families.** *Nucleic Acids Research* 1999, **27:**263-267.

140. Schultz J, Copley RR, Doerks T, Ponting CP, Bork P: **SMART: a web-based tool for the study of genetically mobile domains.** *Nucleic Acids Research* 2000, **28:**231-234.

141. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nature Genetics* 2000, **25:**25-29.

142. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.** *Journal of Molecular Biology* 2001, **305:**567-580.

143. Nielsen H, Engelbrecht J, Brunak S, von Heijne G: **A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites.** *International Journal of Neural Systems* 1997, **8:**581-599.

144. Bendtsen JD, Jensen LJ, Blom N, Von Heijne G, Brunak S: **Feature-based prediction of non-classical and leaderless protein secretion.** *Protein engineering, design & selection* 2004, **17:**349-356.

145. Mount DW: **Using the Basic Local Alignment Search Tool (BLAST).** *CSH Protocols* 2007, **2007:**pdb top17.

146. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Research* 2011, **39:**D32-37.

147. Kulikova T, Aldebert P, Althorpe N, Baker W, Bates K, Browne P, van den Broek A, Cochrane G, Duggan K, Eberhardt R, et al: **The EMBL Nucleotide Sequence Database.** *Nucleic Acids Research* 2004, **32:**D27-30.

148. Tateno Y, Imanishi T, Miyazaki S, Fukami-Kobayashi K, Saitou N, Sugawara H, Gojobori T: **DNA Data Bank of Japan (DDBJ) for genome scale research in life science.** *Nucleic Acids Research* 2002, **30:**27-30.

149. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Research* 2000, **28:**235-242.

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

151. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG.** *Nucleic Acids Research* 2006, **34:**D354-357.

152. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li CY, Wei L: **KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases.** *Nucleic Acids Research* 2011, **39:**W316-322.

153. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M: **KAAS: an automatic genome annotation and pathway reconstruction server.** *Nucleic Acids Research* 2007, **35:**W182-185.

154. Goesmann A, Haubrock M, Meyer F, Kalinowski J, Giegerich R: **PathFinder: reconstruction and dynamic visualization of metabolic pathways.** *Bioinformatics* 2002, **18:**124-129.

155. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, Feuermann M, Ghanbarian AT, Kerrien S, Khadake J, et al: **The IntAct molecular interaction database in 2010.** *Nucleic Acids Research* 2010, **38:**D525-531.

156. Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, Bobechko B, Boutilier K, Burgess E, et al: **The Biomolecular Interaction Network Database and related tools 2005 update.** *Nucleic Acids Research* 2005, **33:**D418-424.

157. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TK, Gronborg M, et al: **Development of human protein reference database as an initial platform for approaching systems biology in humans.** *Genome Research* 2003, **13:**2363-2371.

158. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update.** *Nucleic Acids Research* 2004, **32:**D449-451.

159. Ceol A, Chatr Aryamontri A, Licata L, Peluso D, Briganti L, Perfetto L, Castagnoli L, Cesareni G: **MINT, the molecular interaction database: 2009 update.** *Nucleic Acids Research* 2010, **38:**D532-539.

160. Nagaraj SH, Deshpande N, Gasser RB, Ranganathan S: **ESTExplorer: an expressed sequence tag (EST) assembly and annotation platform.** *Nucleic Acids Research* 2007, **35:**W143-147.

161. Strahm Y, Powell D, Lefevre C: **EST-PAC a web package for EST annotation and protein sequence prediction.** *Source code for Biology and Medicine* 2006, **1:**2.

162. Nagaraj SH, Gasser RB, Ranganathan S: **Needles in the EST haystack: large-scale identification and analysis of excretory-secretory (ES) proteins in parasitic nematodes using expressed sequence tags (ESTs).** *PLoS Neglected Tropical Diseases* 2008, **2:**e301.

163. Yatsuda AP, Krijgsveld J, Cornelissen AW, Heck AJ, de Vries E: **Comprehensive analysis of the secreted proteins of the parasite *Haemonchus contortus* reveals extensive sequence variation and differential immune recognition.** *The Journal of Biological Chemistry* 2003, **278:**16941-16951.

164. Hawdon JM, Narasimhan S, Hotez PJ: **Ancylostoma secreted protein 2: cloning and characterization of a second member of a family of nematode secreted proteins from *Ancylostoma caninum*.** *Molecular and Biochemical Parasitology* 1999, **99:**149-165.

165. Barker G, Batley J, H OS, Edwards KJ, Edwards D: **Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP.** *Bioinformatics* 2003, **19:**421-422.

166. Batley J, Barker G, O'Sullivan H, Edwards KJ, Edwards D: **Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data.** *Plant Physiology* 2003, **132:**84-91.

167. Savage D, Batley J, Erwin T, Logan E, Love CG, Lim GA, Mongin E, Barker G, Spangenberg GC, Edwards D: **SNPServer: a real-time SNP discovery tool.** *Nucleic Acids Research* 2005, **33:**W493-495.

168. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biology* 2009, **10:**R25.

169. Haas BJ, Zody MC: **Advancing RNA-Seq analysis.** *Nature Biotechnology* 2010, **28:**421-423.