In silico identification of novel therapeutic targets from secretome analysis of parasites

by

Gagan Garg

Master of Applied Science (Molecular Biotechnology),

The University of Sydney, Australia

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

January 2013

DEDICATED TO MY FAMILY

MRS. RAMA GARG (MUMMY), MR. SUNIL GARG (PAPA) AND MR. LOVISH GARG (BHAI)

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.

Gagan Garg

January 2013

ACKNOWLEDGEMENTS

It is my pleasure to thank the following people who made this thesis possible through their continuous encouragement, motivation 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.

- Dr. Antonio Marcilla for giving an opportunity to collaborate on the *Strongyloides stercoralis* and *Echinostoma caproni* projects.

- Dr. Steve Peterson for giving an opportunity to work on the *Strongyloides ratti* data.

- Prof. Robin Gasser, my adjunct supervisor and his group members for the support during the lab visits and motivation.

- My co-supervisor, Professor Helena Nevalainen for her support throughout my PhD tenure.

- Ms. Catherine Wong, Ms. Maria Hyland, Ms. Anne Micallef, Ms. Michelle Kang, Ms. Jane Yang, Ms. Meredith McGregor, Dr. Chris McRae, Mr. Michael Baxter, Mr. Doan Lee and Mr. Suresh Mulavineth for administrative and IT support.

- Macquarie University, for the award of Australian Postgraduate Award for pursuing Ph.D. and PGRF for attending Genome Informatics meeting and Comparative Genomics course at Cold Spring Harbor Laboratory, New York, USA.

- 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 and funds for the successful completion of this research project.

- Past and current colleagues: Mr. Mohammad Islam, Mrs. Elsa Chacko, Dr. Javed Mohammed Khan, Dr. Gaurav Kumar, Dr. Ranjeeta Menon, Dr. Varun Khanna, Dr. Jitendra Gaikwad, Mr. Rajesh Ganeshan and Ms. Sowmya Gopichandran.

Personal

- I am deeply indebted to all my following friends and relatives for making these years a memorable experience with their good company and for their moral support during tough times: Parvinder Singh, Falguni Patel, Prince Modi, Balwan and Family, Ajit Garg and Family, Aditya Bansal and Family and all other friends and relatives in Australia and abroad.

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

- Special thanks to my good friends in India: Paras Bhardwaj, Pardeep Nahata and Parveen Kikan for their support.

- Last, but not the least, my parents, who motivated me throughout my Ph.D., their love has been the guiding light and a source of inspiration for me.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

Adl	Adult lethal
BLAST	Basic local alignment search tool
BIND	Biomolecular interaction network database
BioGRID	The biological general repository for interaction datasets
cDNA	complementary DNA
DIP	The database of interacting proteins
DNA	Deoxy ribonucleic acid
DOE	Department of Energy
Emb	Embryonic lethal
EMBL	European molecular biology laboratory
ES	Excretory/Secretory
EST	Expressed Sequence Tag
ESPs	Excretory/Secretory proteins
E-value	Expectation value
FDA	US Food and Drug Administration
GO	Gene ontology
GWAS	Genome-wide association study
HMM	Hidden markov model
HPRD	Human protein reference database
JGI	Joint Genome Institute
KAAS	KEGG automatic annotation server
KEGG	Kyoto encyclopedia of genes and genomes
KOBAS	KEGG orthology based annotation system
Let	Larval lethal
Lva	Larval arrest
mRNA	Messenger RNA
MINT	Molecular interaction database

MS	Mass spectrometry
NGS	Next generation sequencing
NN	Neural network
OLC	Overlap layout consensus
ORF	Open reading frames
PCR	Polymerase chain reaction
PDTD	Potential drug target database
PDB	Protein Data Bank
RNA	Ribonucleic acid
RNAi	RNA Interference
NGS	Next generation sequencing
NTD	Neglected tropical diseases
SNP	Single nucleotide polymorphism
SOAP	Short oligonucleotide analysis package
Ste	Maternal sterile
Stp	Sterile progeny
TM	Transmembrane
TTD	Therapeutic target database
WGS	Whole genome sequencing

LIST OF TABLES

LIST OF FIGURES

LIST OF PUBLICATIONS INCLUDED IN THIS THESIS

The following papers are presented 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, **Garg G** (2009): **Secretome: clues into pathogen infection and clinical applications.** *Genome Medicine*, **1:**113

   Contributions to: (i) concept: 40%; (ii) data gathering: 70% and (iii) writing: 50%.

2. **Garg G**, Ranganathan S (2013) ***In silico* secretome analysis of *Echinococcus multilocularis* and *Echinococcus granulosus* using expressed sequence tags** (under submisson)

   Contributions to: (i) concept: 60%; (ii) data gathering: 80%; (iii) data analysis: 70%; and (iv) writing: 70%.

3. **Garg G**, Ranganathan S (2011): ***In silico* secretome analysis approach using next generation sequencing transcriptomic data.** *BMC Genomics*, 12 Suppl 3, S14

   Contributions to: (i) concept: 70%; (ii) data gathering: 100%; (iii) data analysis: 100%; and (iv) writing: 70%.

4. **Garg G**, Ranganathan S: **Helminth secretome database (HSD): a collection of helminth excretory/secretory proteins predicted from expressed sequence tags (ESTs).** *BMC Genomics*, 13 Suppl 8, S8

   Contributions to: (i) concept: 70%; (ii) data gathering: 100%; (iii) data analysis: 100%; and (iv) writing: 70%.

5. Marcilla A, **Garg G**, Bernal D, Ranganathan S et al. (2012) **Transcriptome analysis of *Strongyloides stercoralis* L3i larvae identifies targets for intervention in a neglected disease**. *PLoS Neglected Tropical Diseases*, 6:e1513

   Contributions to: (i) concept: 30%; (ii) data gathering: 40%; (iii) data analysis: 50%; and (iv) writing: 30%.

6. **Garg G**, Bernal D, Trelis M, Forment J et al. **Transcriptome characterization of the model organism *Echinostoma caproni*** (under submisson)

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

7. **Garg G**, Ranganathan S (2013): **High-throughput functional annotation and data mining of fungal genomes to identify therapeutic targets**, in: Gupta, V.K.; Tuohy, M.G.; Ayyachamy, M.; Turner, K.M.; O'Donovan, A. (Eds.): *Laboratory Protocols in Fungal Biology: Current Methods in Fungal Biology*, Springer, USA, pp. 569-574.

   Contributions to: (i) concept: 70%; (ii) data gathering: 100%; (iii) data analysis: 100%; and (iv) writing: 70%.

ABSTRACT

The secretome of an organism is defined as the subset of proteins secreted by its cell, usually known as excretory/secretory (ES) proteins. These proteins play an important role in producing clinical infections inside the host organism during parasite attack. ES proteins are the choice of new therapeutic solutions for different clinical infections, especially in the case of parasitic and fungal infections because these proteins are present at the host-parasite interface and act as immunoregulators to host immune recognition for parasite survival inside the host organism.

An in-depth study of ES proteins will lead to better understanding of host-parasite relationships and the molecular biology of parasites while their functional annotation can help identify therapeutic molecular targets to control parasitic infections with minimum host side effects. This thesis focusses on the role of the secretome in the molecular interplay between parasites and their respective hosts. As the involvement of ES proteins as a key factor in parasitism, my focus was to predict, annotate and analyse parasitic excretory/secretory (ES) proteins for the prediction of novel therapeutic solutions against parasitic infections using transcriptomic data. To achieve these goals, I have carried out an initial review of the different approaches for secretome analysis, compiling recent secretome data available for parasites and application of bioinformatic tools to parasites.

Based on this review, we carried out a preliminary analysis on *Echinococcus multilocularis* and *Echinococcus granulosus* by integrating SecretomeP, the most widely used program for non-classical ES protein prediction, into an existing pipeline, EST2Secretome. However in case of parasites, SecretomeP is not able to completely predict non-classical secretory proteins, as shown in the other parasitic secretome studies. To address these issues, we developed a new secretome analysis approach, which use sequence similarity search against experimentally identified ES proteins collected from literature along with computational prediction.The updated approach was applied to 454 transcriptomic data of *Strongyloides ratti*, which is a gastrointestinal nematode that infects rats and used as a model to study human stronglyloidasis. By integrating different computational tools together we were able to study ES proteins comprehensively in *S. ratti* transcriptome. The analysis revealed the involvement of *S. ratti* ES proteins in pathways such as purine metabolism and glutathione metabolism, which are important for parasite survival inside the host.

Our updated computational approach was applied to analyse largest publicly available helminth transcriptome data from dbEST. From 870,223 ESTs for 78 helminth species, predicted ES proteins along with annotation results were compiled as a database (Helminth Secretome Database). This unique resource is a collection of 18,992 helminth ES proteins and freely available to scientific community.

The bioinformatics approach developed has been applied to novel transcriptome datasets of two parasitic organisms. Our primary computational application was first applied to the analysis of transcriptome data from the infective third larval stage (L3i) of *Strongyloides stercoralis*. This dataset is the first transcriptome of L3i of *S. stercoralis,* using 454 sequencing coupled with a semi-automated bioinformatic analyses. Along with ES proteins we carried out functional annotation of all putative proteins translated from 11,250 contiguous sequences, of which most were novel. Secondly, we studied the transcriptome of the adult stage of *Echinostoma caproni* generated using 454 sequencing. Our bioinformatic workflow was employed to predict 3,415 putative ES proteins and potential therapeutic targets.

With the advent of next-generation sequencing technologies, massive sequencing datasets have been generated. Despite this increase in sequence data, several proteins remain unannotated, as hypothetical proteins. To fill this gap we have extended our bioinformatics workflow with other relevant computational tools for the annotation of hypothetical proteins and the prediction and analysis of secreted proteins as therapeutic targets. This protocol was applied to pathogenic fungi, *Cryptococcus gattii* and *Cryptococcus neoformans var. grubii*, causative agents of disease (cryptococcosis) in healthy, immunocompetent and immunosuppressive humans.

In conclusion, an updated computational secretome analysis approach using transcriptomic or proteomic data was developed to significantly reduce the time taken for secretome analysis, improved annotations at the protein function levels and identified key ES proteins inferred to be involved in parasite-host interactions. Several novel candidates for parasite intervention were discovered during these analyses.

Chapter 1: Introduction

1.1 Overview

The study of genes and gene products helps in the understanding of the basic principles of biology at the biochemical level. For the complete understanding of fundamental principles of biology at molecular, cellular and organismal levels, we must identify and catalogue the function of all the genes and gene products of an organism.

Briefly, the genome is the master blueprint for the total set of an organism's genes. Genes contain the information to code for gene products and are considered as the instructional subunits of deoxyribonucleic acid (DNA), while the gene products, the intermediate ribonucleic acid (RNA) and the proteins, are the biochemical materials obtained as a result of the expression of a gene. The transcriptome is the fraction of genes expressed as mature messenger RNA (mRNA) from any genome at a given timepoint. The transcriptome is thus analogous to a snapshot of a time period in the cell's life and is related to the proteins present in a cell at that instance. The genome is static while the transcriptome is dynamic and variable, depending on the exact conditions of the cellular environment. The proteome is the total set of proteins expressed by a genome, cell, tissue or organism. The secretome is a subset of proteome which constitutes the complete set of proteins that are secreted by a cell. It constitutes upto 30% of the total proteome [1].

A diverse array of living organisms can be found in the biosphere of our planet. Most of them are free living, but there are also subsets of organisms, known as parasites that colonise and spend a significant part of their life cycle in a suitable 'host' organism by utilizing the host's cellular machinery. Proteins secreted by the parasites customize the host environment and provide protection against the host immune system [2]. These proteins, usually known as excretory/secretory (ES) proteins, comprise the secretome of parasites and help us in understanding the basic principles of biology involved in parasitic infections, as these are present at the host-parasite interface and regulate the host immune system Thus, the study of ES proteins leads us to understand the molecular basis of parasitism and provides valuable clues for developing novel therapeutic strategies against parasitic infections. In the case of parasites, transcriptomics has been used extensively for ES protein prediction and for identifying novel therapeutic targets against parasitic infections.

In this thesis, new and updated computational approaches and tools are applied for the secretome analysis of parasites, using transcriptomic or proteomic data. Improved and faster methods to analyse large amounts of transcriptomic data are suggested and more importantly, several novel therapeutic targets have been identified, with experimental validations already underway for a subset.

At the outset, the transcriptome of parasites and different transcriptomic data generation methods are described. This is followed by the secretome, its importance and experimental analysis approaches are described. We then present the bioinformatics approach for secretome analysis with basic steps involved in the secretome analysis using transcriptomic data is presented. A brief introduction to helminth parasites and other pathogenic organisms (pathogenic fungi) covered in this thesis is then provided. From the requirements for a generic framework of computational secretome analysis, the objectives for the thesis have been set out.

1.2 Transcriptome of parasites

Every cell of an organism has a genome that contains all the biological information needed to build and maintain life. The genome is made up of DNA (or RNA in the case of some viruses). It includes both coding and non-coding sequences in multicellular organisms, with nearly every cell containing the same genome, and thus having the same genes. DNA is transcribed into complementary molecules of RNA, referred to as transcripts, to carry the genetic instructions. Not every gene is transcriptionally active in every cell. Therefore, different cells show different patterns of gene expression, leading to entirely different transcripts. This variation in gene expression differentiates various types of cells. A transcriptome represents the small percentage of genes that is transcribed into RNA. In the case of humans, this is estimated to be less than 5% [3]. 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 as a complementary DNA (cDNA) library. Sequencing from the 3’ or 5' end of each cDNA generates expressed sequence tags (ESTs). ESTs can be used to produce probes to decide the presence or absence of similar transcripts in other tissues. This feature has been used to charcterize newly sequenced genomes. For parasites, transcriptomics has proved to be an effective tool to determine which genes are active in these organisms at various stages of development and essential for parasitism. Various

organisms like *Strongyloides ratti* [4] and *Teladorsagia circumcinta* [5], whose genomes are underway, have been studied by generating transcriptome datasets provide a low-cost alternative to whole genome sequencing.

1.2.1 Expressed Sequence Tags (ESTs)

Expressed sequence tags or ESTs are short, unedited, randomly selected single-pass sequence reads of approximately 100-800 base pairs (bp) length, derived from complementary DNA (cDNA) clones [6, 7] using Sanger sequencing. An EST represents a small region or a part of nucleotide sequence from a transcribed protein coding or non-coding messenger RNA (mRNA). ESTs can be generated at a reasonably low cost from either the 5' or 3' end of a cDNA clone, to get an insight into transcriptionally active regions of the genome of the original tissue used for the construction of the cDNA bank. ESTs were the primary source for human gene discovery in the seminal contribution by Adams *et al.* [8]. ESTs have become an invaluable resource for gene discovery, genome annotation, alternative splicing, single nucleotide polymorphism (SNP) discovery, molecular markers for population analysis and expression analysis in animal, plant, and microbial species [9-11]. In parasitology, ESTs have been used for gene discovery, therapeutic target identification and verification of predicted genes [12-14]. Over the years there has been an exponential growth in the generation and accumulation of EST data in public databases for myriad organisms, due to significant advancements and decreasing costs in high throughput sequencing technologies. For instance, the Expressed Sequence Tags database (dbEST) (release 130101) currently holds over 74 million ESTs, derived from more than 1400 organisms, representing the largest publicly available resources for ESTs [15].

According to the central dogma of molecular biology [16], the DNA codes for the production of messenger RNA (mRNA) during transcription. This mRNA serves as a template for protein synthesis in the cytoplasm, *via* the process of translation. The mRNA represents copies from expressed genes. As RNA cannot be cloned directly, these sequences are reverse transcribed to double-stranded cDNA using a specialized enzyme, the reverse transcriptase. The cDNA is then cloned to make libraries representing a set of transcribed genes of the original cell, tissue or organism. These cDNA clones are then sequenced randomly from both directions in a single-pass run, with no validation or full-length sequencing, to obtain 5' and 3' ESTs. The resultant set of ESTs is redundant, as the cDNA template used can be of partial or full length

ESTs can be generated from any tissue, organ or cell of interest and also at different developmental stages, such as foetal or adult stages. The generation of ESTs from mRNA is described briefly in the following sections, with the different steps summarised in Figure 1.1.

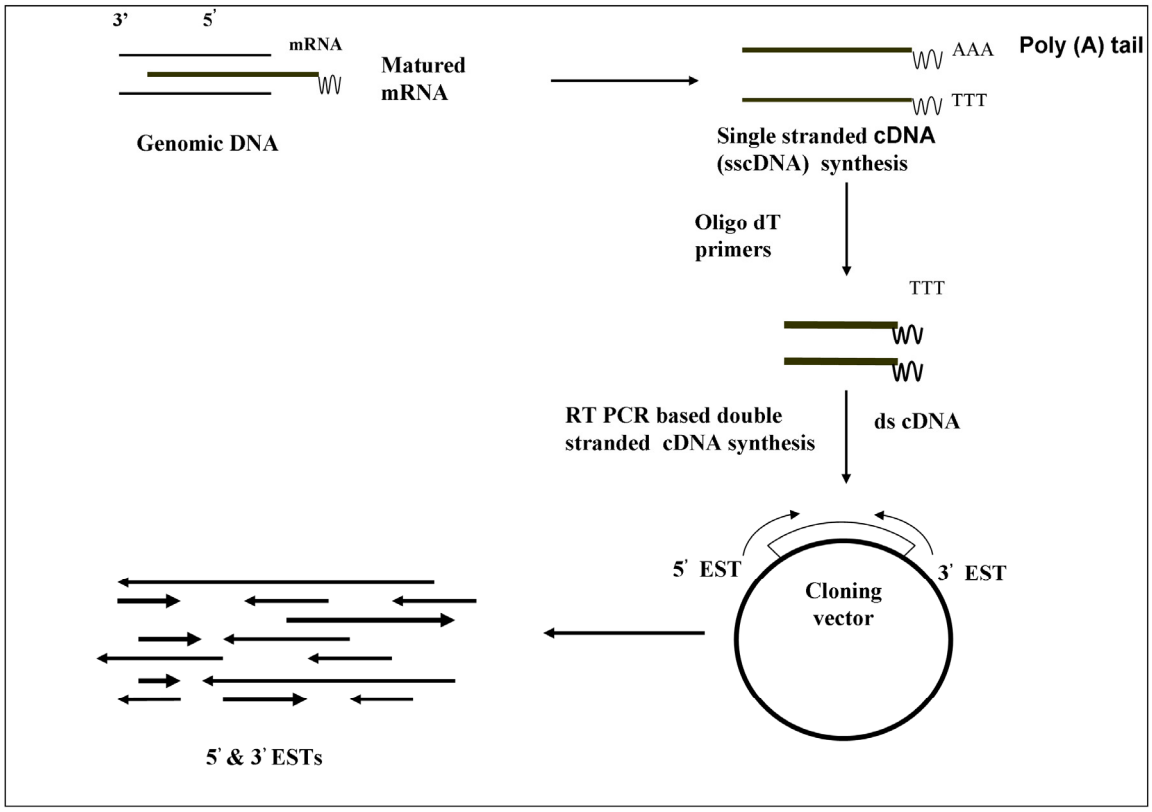


Figure 1.1. Illustration of the steps involved in EST generation. Genomic DNA is transcribed to mRNA. The information on the mRNA is copied onto cDNA which results in cDNA libraries. 5' and 3' ESTs are generated from such cDNA libraries.

1.2.2 Transcriptome data from next-generation sequencing (NGS) platforms

Back in 2005, high-throughput sequencing (HTS) or next-generation sequencing (NGS) platforms, based on sequencing by synthesis were introduced. This technology revolutionised the field of transcriptomics by providing a cost-effective means by executing millions of sequencing reactions in parallel, and producing nucleotide data at ultrahigh rates [17].

The main NGS platforms include the 454 system (http://www.454.com), based on pyrosequencing technology [18], Illumina Solexa system (http://www.illumina.com), which detects fluorescence signals [19] and ABI SOLiD (www.appliedbiosystems.com),

based on sequencing by ligation [20]. Other recently introduced technologies are Polonator G.007 [21], Helicos HeliScope platform [22], Pacific Biosciences Pacbio RS [23] and Ion Personal Genome Machine (PGM) by Ion Torrent [24]. Although read lengths generated by NGS platforms are much shorter (averaging 100–230 bp and 300–400 bp for 454FLX and 454Titanium, respectively) than with capillary sequencing, these platforms generate sufficient data to re-sequence complete bacterial genomes in a single run [25-27]. The various NGS methods have been reviewed by Mardis [28]. The most commonly used NGS techniques are described briefly in the following sections.

1.2.2.1 454/Roche pyrosequencing

The basis of this technology is pyrophosphate detection [29]. In this sequencing system, DNA fragments are ligated to beads by means of specific adapters. To obtain sufficient light signal intensity for detection in the sequencing-by-synthesis reaction step, emulsion PCR is carried out for amplification. Once the PCR amplification cycles are complete, each bead with its fragment is placed at the top end of an optical fiber that has the other end facing a sensitive CCD camera, which enables the positional detection of emitted light. In the last step, polymerase enzyme and primer are added to the beads so that the synthesis of the complementary strand can start: the incorporation of a base by the polymerase enzyme in the growing chain releases a pyrophosphate group, which can be detected as emitted light.

A limitation of the 454 sequencing platform is that base calling cannot properly interpret long stretches (>6) of the same nucleotide (homopolymer DNA segments); for this reason homopolymer segments are prone to base insertion and deletion errors during base calling. The 454/Roche platform is used for *de novo* genomic and transcriptomic studies. In the case of helminths, many transcriptome projects involving secretome analysis have been carried out using this technology [30, 31]. 454 cDNA sequencing datasets of several nematode species like *Teladorsagia circumcinta* are available from nematode.net [32].

1.2.2.2 Illumina Genome Analyser (Solexa)

This technology is the most widely available NGS technology. In this platform, the amplified sequencing features are generated by bridge PCR [33, 34] and after immobilization in the array, all the molecules are sequenced in parallel by means of sequencing by synthesis. During the sequencing process, each nucleotide is recorded through imaging techniques, and is then converted into base calls. The Illumina sequencer

is able to sequence reads up to 100 bp with relatively low error rates (~ 0.1% compared to 1% for 454/Roche Platform). Due to the ability to generate millions of reads per run, it is considered the best for re-sequencing, single nucleotide polymorphism (SNP), targeted sequencing and gene transcription studies. It has different features compared to the 454 approach [35]. In the case of helminths, Illumina sequencing was used for the *Fasciola gigantica* transcriptome [36]. Recently sequenced helminth genomes like *Ascaris suum* [37] and *Schistosoma herbatorium* [38] also used Illumina sequencing.

1.2.2.3 ABI SOLiD

The sequencing process used by ABI SOLiD is very similar to the Ilumina workflow with some differences. First of all, the clonal sequencing features are generated by emulsion PCR, instead of bridge PCR. Secondly, the SOLiD system uses a di-base sequencing technique in which two nucleotides are read (*via* sequencing by ligation) simultaneously at every step of the sequencing process, while the Illumina system reads the DNA sequences directly. The latest ABI SOLiD 4 machines are able to generate up to 1 billion 50 bp paired-end reads per run for a total of 100 GB of data with a throughput of around 5 GB per day. A variety of genomic studies have been carried out using this system [39].

With the advent of high-throughput sequencing, a new method termed RNA-Seq (RNA sequencing) for both mapping and quantifying transcriptomes has been developed, using deep-sequencing technologies such as 454/Roche and Illumina [40]. In general, a population of RNA (total or fractionated, such as poly(A)+) is converted to a library of cDNA fragments with adaptors attached to one or both ends. Each molecule, with or without amplification, is then sequenced in a high-throughput manner to obtain short sequences from one end (single-end sequencing) or both ends (pair-end sequencing).The reads are typically 30–400 bp, depending on the DNA-sequencing technology used.

The short read strategy of NGS has led to many challenges for bioinformatic analysis in data storage and management solutions, leading to the creation of 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 [41, 42].

1.3 Secretome and its importance in clinical infections

When a parasite invades a host organism, it must secrete biologically active substances that either mimic the host environment or protect it from the host organism's immune system. Therefore, molecules secreted by the parasite, often referred to ES proteins, are believed to be one of the most important class of molecules in parasites and are of interest to a number of laboratories investigating the basic biology of parasites, their interaction with host systems and the possibility of employing ES products in antiparasitic vaccines. ES proteins constitute many functionally diverse classes of molecules, such as cytokines, chemokines, hormones, digestive enzymes, antibodies, extracellular proteinases, morphogens, toxins and antimicrobial peptides. Many of these proteins are known to be associated with vital biological processes, including cell adhesion, cell migration, cell-cell communication and the regulation of the immune response [43]. ES proteins circulate throughout the body (in the extracellular space) or are localized on the cell surface, making them readily accessible to various drug delivery mechanisms and/or the immune system. These characteristics make them attractive to be considered as targets for novel therapeutic targets, and they are currently the focus of major drug discovery research programmes [44].

1.4 Approaches to secretome analysis

With the advances in proteomics and sequencing technologies, numerous studies have been carried out on the secretome, ranging from different parasites to human cancer [45-49]. Mostly, these approaches can be divided into three categories, namely genome sequence analysis, proteomic approaches and bioinformatics approach using in the main, transcriptomic or proteomic data.

1.4.1 Genome sequence analysis

This approach is based on a combination of computational prediction and transcript profiling. The computational prediction of secreted proteins is based on the presence of N-terminal signal peptides, which are considered as signatures for classically secreted proteins. According to the signal hypothesis [50], the majority of secreted proteins have an N-terminal signal peptide sequence, which targets proteins to the endoplasmic reticulum (ER) lumen *via* the *sec*-dependent protein translocation complex. This approach has been used extensively to study significant secretome genes (genes encoding secretory proteins) in human cancers [51, 52].

The genome-based approach allows researchers to scan the whole genome quickly for potentially secreted proteins; however it has the following limitations. Firstly, the basis of this analysis is that an organism's genome sequence has to be available. In case of pathogenic organisms, the genomes of several pathogens such as *Brugia malayi* [53], *Ascaris suum* [37], *Schistosoma mansoni* [54], *Schistosoma japonicum* [55], *Schistosoma herbatorium* [38] and *Vibrio cholerae* [56] are now available but there are several other organisms such as *Ascaris lumbricoides* and *Wuchereria bancrofti* which are still awaiting whole genome sequencing. Secondly, this approach is based on the prediction of N-terminal signal peptides; however, many secretory proteins lack the N-terminal signal peptides and are known as non-classical secretory proteins [57], which are not predicted by this method. Thirdly, secreted proteins are regulated at the post-transcriptional level. Therefore the real level of expression of secreted proteins does not always correlate with mRNA expression [58, 59].

1.4.2 Proteomic approaches

Proteomic techniques are currently the main resource for secretome studies. With the massive development in the area of mass spectrometry, proteomic approaches has been vastly used for the secretome analysis of pathogens. There are two main technologies available: gel-based technology and gel-free mass spectrometry (MS) based technology.

1.4.2.1 Gel-based proteomic analysis

This method is based on two-dimensional gel electrophoresis (2-DE) combined with MS. This method allows the separation of complex mixtures of intact proteins at high resolution. The protein mixtures are first separated according to their charge in the first round using isoelectric focusing and then in second round, according to size using sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE). After gel seperation, proteins are analyzed by peptide mass fingerprinting after in-gel tryptic digestion. This approach has been widely used in pathogen secretome studies like secretome studies of *Helicobacter pylori* [60].

Two dimensional gel electrophoresis (2-DE) currently remains the most efficient method for the separation of complex protein mixtures but this technique has a number of limitations: poor reproducibility between gels, low sensitivity to detect proteins at low concentrations and hydrophobic membrane proteins, limited sample capacity and low

linear range of visualization procedures, time- consuming, labor intensive and low efficiency in protein detection due to its limited accountability to automation.

To counteract some of the problems of standard 2-DE procedure, a modified method, differential in-gel electrophoresis (DIGE) was developed [61]. This method use three spectrally distinct, charge and mass-matched fluorescent dyes (Cy2, Cy3 or Cy5), which can primarily combine covalently with lysine. Protein samples are differently labelled by these dyes before electrophoresis and then mixed and separated on one single gel. By allowing two protein samples to run on single gel, it reduces the experimental variations. Fluorescent labelling enhances the linear dynamic range and sensitivity for DIGE [62]. A differential secretome analysis study based on SW480 human colon carcinoma cells [63] demonstrated that DIGE is more relable and powerful than traditional 2-DE.

Although DIGE proved to be more powerful technique than traditional 2-DE, it still has following shortcomings: low throughput and difficulties in the identification of proteins with extreme isoelectric points or molecular weight.

1.4.2.2 Gel-free proteomic analysis

To overcome the drawbacks of gel-based approaches (mentioned in previous section), gel-free MS-based or shotgun proteomics have been introduced. In these newly developed techniques, instead of using, complex mixtures of proteins are first digested into peptides or peptide fragments, then separated by one or several steps of capillary chromatography and finally analyzed by tandem MS (MS/MS). Multidimensional protein identification technology (MudPIT) [64] is one of the most typical approaches in gel free techniques. In MudPIT, strong cation exchange (SCX) and reverse-phase (RP) liquid chromatographies (LC) are combined with automated MS/MS. MudPIT is powerful in the analysis of membrane or low abundance proteins which are not detected by gel based methods [65, 66]. MudPIT has now become the popular technique in secretome analysis especially human cancers [67, 68]. However, MudPIT is not a quantitative proteomic approach so it is not good for differential proteome analysis [69].

Quantitative proteomics is carried out using molecular labelling methods [70, 71]. In these methods, proteins from different samples are first labelled with different stable isotopes or chemicals, then mixed, separated and identified by LC coupled MS/MS. Different stable isotope labelling approaches are isotope-coded affinity tag (ICAT) [72], Stable isotope

labelling by amino acids in cell culture (SILAC) [73] and Isobaric tag for relative and absolute quantization (iTRAQ) [74]. The secretome analysis of *Leishmania donovani* [75] adopted liquid chromatography coupled with automated MS/MS. Matrix-Assisted Laser Desorption/Ionization-Time of Flight (MALDI-TOF) mass spectrometry, a popular tool for the analysis of complex molecules, was used to analyze the secretome of HepG2 cells infected with the dengue virus [76]. Definitely shotgun proteomics is far more powerful method than its gel based counter parts but it has poor capability to detect peptides in complex proteins mixture. Another exciting approach is the Surface enhanced laser desorption/ionization- Time of flight- Mass spectrometry (SELDI-TOF-MS). This approach is a variation of MALDI and uses protein chip arrays, having chromatographic features. A recent study by Michalski *et al*. [77] shows that only 16% of peptides had been targeted for MS/MS in standard LC runs of cell lysate.

While proteomic data is the ideal method to obtain sectrome data, proteomics does not detect all the ES proteins predicted from even a limited expressed sequence tag (EST) dataset (e.g. *Fasciola hepatica* study [78]). Also, the access to high-end machines, sample preparation and data analysis are bottlenecks to obtaining large-scale ES protein information for parasites. Hence, we have selected transcriptomic data as an alternative data resource.

1.4.3 Bioinformatics approach for secretome analysis using transcriptomic data

With the boom in sequencing technologies, sequencing data has been generated worldwide on a massive scale especially in the area of genomics and transcriptomics. Transcriptomic data provides the snapshot of actively expressed genes in a cell at any given time. In the case of parasites especially parasitic helminths, this data has been extensively used for ES protein prediction and annotation using bioinformatics analysis systems such as EST2Secretome [79].

EST2Secretome is a web server for the prediction of secretory proteins from ESTs by preprocessing, assembly and conceptual translation into protein sequences, followed by identification an N-terminal secretory signal peptide using SignalP [80]. The ES proteins predicted were further checked for the presence of transmembrane domains using TMHMM [81]. Proteins, which are predicted to contain a signal peptide by SignalP but no transmembrane domain by TMHMM, are considered as ES proteins by EST2Secretome. These ES proteins are further functionally annotated in terms of gene ontologies, protein

domains and families, metabolic pathways and by identifying homologues from a well-studied model organism such as *Caenorhabditis elegans* [82]. EST2Secretome was applied to approximately 0.5 million EST sequences from parasitic nematodes to predict ES proteins, some of which are already being trialed as vaccine candidates and as targets for therapeutic intervention. The advantage of using this bioinformatic approach is that the authors were able to rapidly produce reliable secretome data in real time, providing valuable clues for experimental validation.

A study on *Fasciola hepatica* [78] by our group estimated the accuracy of EST-based predictions of ES proteins when assessed with proteomic data. EST2Secretome pipeline is restricted to using EST data as input and predicts only classical secretory proteins whereas many proteins are found to be secreted by non-classical secretory pathways [57]. The limitations of the current secretome analysis system paved the way for the development of an updated computational approach for the comprehensive prediction and annotation of ES proteins using Sanger dideoxy sequencing or NGS transcriptomic data, which is the core aim of this thesis.

1.5 Steps for *in silico* secretome analysis using transcriptomic data

The analysis includes different steps (pre-processing, clustering, assembly, ES protein prediction and annotation) using several tools to yield biological information. The steps involved in *in silico* secretome analysis using transcriptomic data are shown in Figure 1.2.

The analysis steps are briefly described in the following sections.

1.5.1 Pre-processing of raw data

Raw cDNA sequence reads from NGS platforms or classic Sanger sequencing (ESTs) are initially screened for vector contaminants and low quality and low complexity sequences to generate high-quality data [83]. Information from non-redundant (NR) vector databases like UniVec [84] and EMVEC [85] are used with tools like such as Seqclean [86] to remove vector sequences. Other tools used for cleaning of raw data are Seqtrim [87] and

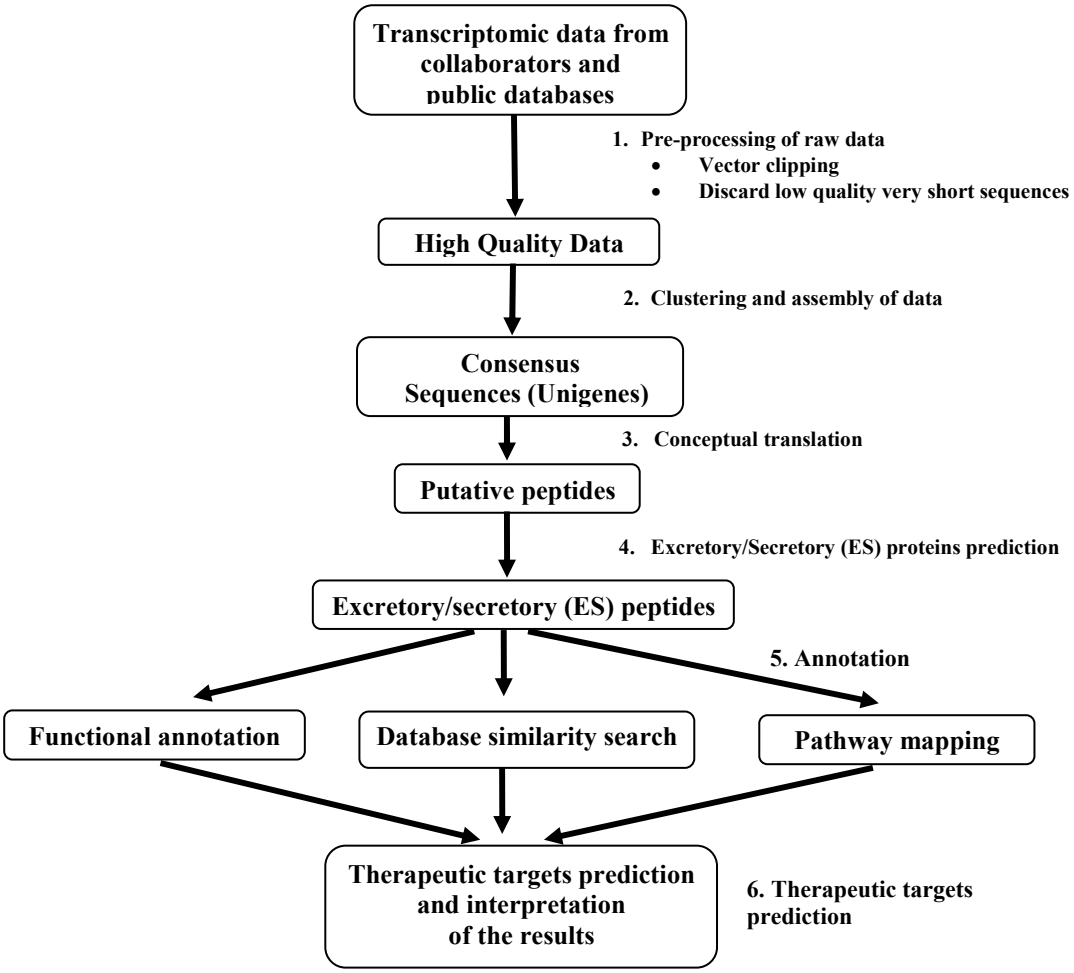


Figure 1.2. Generic steps involved in *in silico* secretome analysis using transcriptomic data. 1. Raw transcriptome data sequences are checked for vector contamination and low quality and very short sequences are removed. 2. High quality data are then clustered and assembled to generate consensus sequences ("unigenes"). 3. Putative peptides are obtained by conceptual translation of consensus sequences. 4. ES proteins are predicted from putative peptides 5. ES protein database similarity searches, functional annotation and pathway mapping are performed to assign putative function(s). 6. The analysis is extended to therapeutic target prediction.

Lucy [88]. These tools use locally installed software such as BLAST [89, 90] to search for vector sequences against NR vector databases. SeqClean is one of the most popular tools used for preprocessing of raw data. This tool carries out trimming and validation of ESTs or other DNA sequences by screening for various contaminants, low quality and low-complexity sequences automatically using specific vector database like Univec [84] for trimming. Poly (A) tracks are trimmed to retain a few adenine nucleotides (usually 6-10) to

get high quality data for the next step of clustering and assembly into concensus sequences. SeqClean has been successfully deployed in many transcriptome projects [91-93].

1.5.2 Clustering and assembly into unigenes

Once the raw data has been cleaned, the next step is the clustering and assembly. This is achieved by using sequence assembly tools like CAP3 [94], MIRA [95] and Newbler (also known as the GS *de novo* assembler) [96]. The aim of clustering is to collect overlapping reads from the same transcript of a single gene into a single cluster. This process can be extremely time-consuming due to the intrinsic need for all pairs of raw ESTs to be tested for overlaps. Pair-wise sequence similarities are measured to cluster ESTs. The idea of EST clustering has been implemented by TIGR [97], UniGene [98], IMAGEne [99] and MERCK [100].

After the clustering is complete, one or more consensus assemblies for each cluster are produced. Assembly aids in determining the sequence of a target transcript/gene, by alignmment and merging of DNA fragments to form long contiguous sequences (contigs) [83]. This process results in longer and more interpretable coding sequences than their individual ESTs. With sufficient coverage, miscall and indel errors present in individual ESTs can be removed by assembly [101]. Overall, the process reduces redundancy and increases the overall quality of the derived sequence. Assembly results in contigs and single sequences composed of only one sequence known as singletons. Contigs and singletons are together known as unigenes (or rESTs). The different steps for EST clustering are described in detail by Ptitsyn and Hide [102], presenting 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. A comparision study based on Phrap [103], CAP3 [94] and TIGR Assembler [104] based on an 118,000 rat EST dataset was carried by Liang *et al.* [105]. This study showed that CAP3 out-performed other similar programs, maintaining a high level of sensitivity to gene family members and producing high fidelity consensus sequences while handling sequencing errors. CAP3 is based on ‘overlap-layout-consensus (OLC) approach’. The OLC approach uses an overlap graph, which involves overlap discovery by all-against-all, pair-wise read comparison followed by construction

and manipulation of the overlap graph [106, 107]. CAP3 has been implemented successfully in many EST projects.

Newbler [96], distributed by 454 Life Sciences is another example of OLC assembler. Newbler implements OLC twice. The first-phase OLC generates unitigs from reads. Unitigs are mini-assemblies that are, ideally, uncontested by overlaps to reads in other unitigs [108]. The unitigs serve as preliminary, high-confidence, conservative contigs that seed the rest of the assembly pipeline. The second-phase OLC generates larger contigs from the unitigs. This phase joins unitigs into a contig layout based on pair-wise overlaps between unitigs. CAP3 [94] has been used as a prime tool for different 454 transcriptome data projects [109-111]. The MIRA EST Assembler [95] is another good example, which has been reliably used in different EST projects [30, 31]. It works by reconstructing the mRNA transcripts from raw data while performing SNP detection.

A recent study by Kumar and Blaxter [112] compares different *de novo* assemblers for 454 transcriptome data. This study suggests the use of multiple assemblers for the assembly, to reduce the errors. Recently, a new Perl package iAssembler [113], employing MIRA and CAP3 has been developed to identify and correct two common types of transcriptome assembly errors: 1) ESTs from different transcripts (mainly alternatively spliced transcripts or paralogs) are incorrectly assembled into the same contig; and 2) ESTs from the same transcripts fail to be assembled together. iAssembler can be used to assemble ESTs generated using the traditional Sanger method and/or the Roche-454 massive parallel pyrosequencing technology.

Another category of assemblers are those based on the *de Bruijn* graph approach [114]. This approach is most widely applied to the assembly of short reads from the Solexa and SOLiD platforms. This approach relies on *k-mer* graphs, whose attributes make it attractive for vast quantities of short reads. Various *de Bruijn* graph-based assemblers are SOAP [115], ABYSS [116], Euler [117], Velvet [118] and AllPaths [119].

Unigenes obtained from this step can be translated into putative peptides or compared to nucleotide or protein databases using BLASTN or BLASTX respectively from the BLAST [89, 90] suite of programs. Along with individual sequence assembly programs, many EST pipelines like ESTExplorer [91], EST-PAC [120], EST2uni [121], ESTPiper [122] and

many more has been developed. The main aim of these pipelines is to automate the processing of ESTs, mainly the preprocessing and assembly steps.

1.5.3 Conceptual translation of unigenes

Transcriptomic data can be correlated with protein-centric annotations by accurate and robust polypeptide translations, since polypeptides are better templates for identifying domains and motifs, protein localization and to assign gene ontologies. The first step in translating unigenes is in identifying the protein-coding regions or open reading frames (ORFs) from unigenes. Various tools have been created for this purpose such as ESTScan [123], DECODER [124] and OrfPredictor [125]. ESTScan, which can detect and extract coding regions from low quality or partial cDNAs while correcting for frame shift errors and provide conceptual translations, has been used in many transcriptome projects [91-92, 109]. This tool has been used for different *in silico* secretome studies carried out as part of this thesis. The putative peptides obtained from this step can be compared to protein databases and assigned a functional annotation.

1.5.4 Excretory/secretory (ES) protein prediction

With the generation of large volumes of transcriptomic data, especially in the case of parasites and role of ES proteins during parasitic infections, transcriptomic data has been used for ES protein prediction using different processing steps explained above and variety of computational tools like SignalP [80], SecretomeP [126], TargetP [127], PSORT [128], Sigcleave [129], Phobius [130] and TMHMM [81]. These tools can also be used for ES proteome prediction from putative proteins, generated as a result of various genome projects. All these tools work on the basis of machine learning prediction models like hidden Markov models (HMM) and neural networks (NN). These prediction models are derived from training and test datasets. These prediction models are trained on generic and best datasets available for the model. However, as there is huge diversity among the proteins across different species, it is hard to achieve 100% accuracy in ES protein prediction using these tools individually.

To achieve reliable ES protein prediction, these tools are used together, such as a combination of using CJ-SPHMM, TMHMM and PSORT suggested by Chen *et al.* [131]. Many helminth transcriptomic studies [78, 79] have used a combination of TMHMM (for prediction of transmembrane domains) and SignalP (for prediction of classical secretory proteins) for prediction of ES protein prediction. Classical secretory proteins are those,

which are secreted only through conventional secretion pathways, using N-terminal signal peptide signatures [50]. In addition to classical secretory proteins, there are now many proteins which are found to be secreted by non-classical secretory pathways [57, 78]. These non-classical secretory proteins are usually predicted using SecretomeP [126]. In addition to using these above mentioned prediction tools, a homology-based search can be performed against experimentally verified ES proteins of respective species collected from literature or protein databases like Uniprot [132] to generate a comprehensive list of ES proteins.

1.5.5 Database similarity searches

Running a homology-based search of unigenes, putative or ES proteins against known nucleotide and protein databases like GenBank [133] and the NR databases [134] using different flavours of the BLAST [89, 90] programs available from NCBI is usually the first step towards functional annotation. BLASTN can be used to search unigenes against the nucleotide sequence databases. BLASTX can be used to search unigenes against the protein sequence database by translating unigenes into protein products in all six reading frames. BLASTP can be used to search putative proteins against the protein sequence database. Other alignment programs used for sequence similarity search are BLAT [135], GMAP [136] and MGALIGN [137].

Unigenes are searched against known protein databases like the NR protein database and genomic data of related species. In case of nematodes, *Caenorhaditis elegans* data [138, 139] is usually considered, as it is the best studied free-living nematode. The different protein sequence databases are Swiss-Prot [140], Entrez protein database [141], TrEMBL [140] and UniProtKB [132]. PDB [142], which represents the protein structure database, also contains sequence information of proteins.

Although each protein sequence is annotated individually using these tools, it is important to study proteins as a part of larger protein complexes and pathways within a cell as proteins work by interacting with other proteins. In case of parasites, protein interactions are important as a complex interplay exists between the cellular environments of the parasite and its host during the course of invasion and infection.

Information on protein interactions is available in different interaction databases, which include IntAct [143], BOND (formally known as BIND) [144-146], HPRD [147], DIP

[148], BioGRID [149] and MINT [150]. Putative or ES proteins can be matched against these databases, by using BLAST [89, 90], by setting a reasonably permissible E-value threshold value (1e-05 or less) to report significant matches. The results can be validated by experimental assays.

1.5.6 Functional annotation

Besides database similarity search, there are other tools used to provide functional annotation to predicted proteins from transcriptomic data. Gene Ontology (GO) [151] has been used widely to classify gene function, as the knowledge of gene and protein roles in cells accumulates. It provides a dynamic vocabulary and hierarchy that unifies descriptions of biological, cellular and molecular functions across genomes. GO is categorised into three parts: biological process, molecular function and cellular component. Biological process refers to a biological objective to which the gene or gene product contributes. Molecular function is defined as the biochemical activity (including specific binding to ligands or structures) of a gene product, while cellular component refers to the place in the cell where a gene product is active [151]. Various tools like BLAST2GO [152], Goblet [153], OntoBlast [154] and GeneTools [155] are used to assign GO terms to EST data. There are many tools developed within the Gene Ontology Consortium and other independent labs to help users to assign GO terms to diverse set of data. The complete list is available at Gene Ontology project website [156].

Another strategy for annotation is mapping putative proteins to well characterized protein domains, motifs and signatures as well as GO assignments of the functional domains/motifs identified. This characterization is achieved by using InterPro [157-159] database, which integrates data from 13 member databases including PROSITE [160] and Pfam [161]. A complete list of InterPro member database resources is provided in Table 1.1. InterPro is queried using an integrated tool, InterProScan [162]. While some of the databases may appear to cover that same functional description, the scientific basis for each collection is different and therefore, for comprehensive annotation, all InterPro databases are usually checked for matches.

Table 1.1. **List of InterPro member databases**

Database	Description	Reference
PROSITE	Database of protein families and domains, containing biologically significant sites, patterns and profiles	[160]
PRINTS	Collection of protein fingerprints	[163]
PFAM	Collection of multiple sequence alignments and hidden Markov models covering many common protein domains.	[161]
PRODOM	Database of an automatic compilation of homologous domains	[164, 165]
PANTHER	Collection of protein families	[166]
SMART	Resource for the identification and annotation of protein domains and the analysis of protein domain architectures.	[167, 168]
SUPERFAMILY	Library of profile hidden markov models that represent all proteins of known structure.	[169]
TIGRFAM	Collection of protein families, featuring curated multiple sequence alignments, hidden Markov models (HMMs) and annotation	[170]
CATH-Gene3D	Describes protein families and domain architectures in complete genomes.	[171]
HAMAP	Collection of manually created profiles by expert curators they identify proteins that are part of well-conserved bacterial, archaeal and plastid-encoded proteins families or subfamilies.	[172]
PIRSF	Network with multiple levels of sequence diversity from superfamilies to subfamilies that reflects evolutionary relationship	[173]

1.5.7 Pathway mapping

Mapping of conceptually translated proteins from transcriptomic data to biological pathways provide vital information on gene function, gene expression and regulation. The Kyoto Encyclopedia of Genes and Genomes (KEGG) [174], provides a reference knowledgebase for linking genomes to biological systems. The KEGG PATHWAY database contains pathway maps for molecular systems in both normal and perturbed state. Other metabolic or enzyme databases are BioCyc [175], BRENDA [176] and Reactome [177]. The Kyoto Encyclopedia of Genes and Genomes database (KEGG) [178, 179] is used to functionally classify transcriptomic data based on biochemical functionality. A complete list of KEGG resources is provided in Table 1.2. Different pathway mapping tools are available like KOBAS [180], KAAS [181] and Pathfinder [182]. KOBAS (KEGG Orthology Based Annotation System) is a web server for automated annotation and pathway identification based on KEGG orthology (KO). KAAS (KEGG Automatic Annotation Server) provides functional annotation of genes by performing BLAST comparisons against the manually curated KEGG GENES database.

Table 1.2. List of KEGG resources.

Database/content	URL
KEGG PATHWAY	http://www.genome.jp/kegg/pathway.html
KEGG GENES	http://www.genome.jp/kegg/genes.html
KEGG LIGAND	http://www.genome.jp/kegg/ligand.html
KEGG BRITE	http://www.genome.jp/kegg/brite.html
KEGG DRUG	http://www.genome.jp/kegg/drug/
KEGG GLYCAN	http://www.genome.jp/kegg/glycan/
KEGG REACTION	http://www.genome.jp/kegg/reaction/
KEGG EXPRESSION	http://www.genome.jp/kegg/expression/
KEGG ANNOTATION	http://www.genome.jp/tools/kaas/
KEGG DISEASE	http://www.genome.jp/kegg/disease/
KEGG ORTHOLOGY	http://www.genome.jp/kegg/ko.html

1.5.8 Therapeutic target prediction

ES proteins circulating throughout the body of an organism (e.g. in the extracellular space) are localized to or released from the cell surface, making them readily accessible to drugs and/or the immune system. These proteins are also responsible for regulating the host immune system for parasite survival inside the host. These characteristics make these molecules extremely attractive targets for novel vaccines and therapeutics, which are currently the focus of major drug discovery research programs [43, 44]. In order to check the potential of parasite ES proteins as therapeutic targets, we took the subtractive genomics approach [183], which involves subtraction between host and parasite gene products for enlisting essential parasite specific proteins. This step is usually achieved using BLAST [89, 90] to check the sequence similarity between two protein datasets.

These parasite specific proteins are further tested for essentiality for parasite survival. In case of parasitic nematodes, this is achieved by mapping to *C. elegans* lethal RNAi phenotype mapping. After a protein is found to be essential for parasite survival, it is further checked by homology matching against a set of known drug targets. Relevant databases containing information on drug targets are PDTD [184], Drugbank [185], TDR Targets Database [186] and the therapeutic targets database (TTD) [187]. Among these databases, DrugBank is a unique bioinformatics and cheminformatics resource that contains detailed drug data with their target information (i.e. sequence, structure, and pathway). The database contains 6711 drug entries, which includes 1447 FDA-approved small molecule drugs, 131 FDA-approved biotech (protein/peptide) drugs, 85 nutraceuticals and 5080 experimental drugs. Therefore, in this thesis, we used DrugBank [185] as the main database for therapeutic targets prediction.

1.6 Genome/putative proteome data

Using NGS, the study of genomics has been increasingly dominated by the growth of raw sequencing datasets. Recently many new helminth genomes like *Ascaris suum* [37], *Trichinella spiralis* [188], *Schistosoma haemato*bium [38] and *Clonorchis sinensis* [189] have been sequenced. The assembled genome data along with putative proteins are freely available to the scientific community. This data can be functionally annotated in terms of pathway mapping, protein domain mapping, GO mapping, ES protein prediction and drug targets mapping using different computational tools. Fungal pathogen genomes are available from two major genome sequencing centers, the Broad Institute [190] and the Department of Energy (DOE) Joint Genome Institute (JGI) [191]. Few fungal pathogens

were sequenced at The Institute for Genomic Research (TIGR) [192] and the Genome Sequencing Center at Washington University (WU-GSC) [193].

1.7 Introduction to helminths

Helminths are parasitic worms that cause a wide variety of infectious diseases, *viz.* Filariasis and cysticercosis. These complex multicellular organisms live and feed off living hosts to sustain themselves and perturb their host's nutrient absorption thus causing morbidity, weakness and disease. The helminths are separated according to their general external shape and the host organ they inhabit. Many helminths are free-living organisms in aquatic and terrestrial environments whereas others occur as parasites in most animals and some plants. Parasitic helminths are an almost universal feature of vertebrate animals; most species have worms in them somewhere. 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. Helminths develop through egg, larval and adult stages. Knowledge of the different stages in relation to their growth and development is the basis for understanding the epidemiology and pathogenesis of helminthiasis. Most common symptoms results from helminth infections include diarrhea, foul breath, headache, nausea and abdominal pain. Various stages of helminth species have been studied by generating transcriptomic sequencing data [78, 93]. Recent studies of worm genomes reveal that the factors and receptors of worms show greater homology to molecules of the human immune system [194]. Worms have fully developed organs and complex tissues compared to bacteria and viruses. Infected individuals bare three life-cycle stages of heminth; infective larvae, adult worms and transmission stage parasites (eggs, immature larvae or microfilariae). These life-cycle stages are proved to be molecularly different both in proteomic [195] and DNA microarray [196] studies and are believed to induce stage-specific immune responses.

1.8 Types of helminths

The helminths are invertebrates characterized by elongated, flat or round bodies. In medical terms, helminths are classified as flatworms or platyhelminths, which include flukes (trematodes) and tapeworms (cestodes) and roundworms (nematodes). Other classification is based on host organ, where they reside like lung flukes and intestinal roundworms.

1.8.1 Nematodes

Nematodes or roundworms are known to be the most abundant eukaryotic animals on earth (Figure 1.3). Adult and larval nematodes are bisexual and cyclindrical in shape, which inhabit in the intestinal and extraintestinal sites. Nematodes (roundworms) have long thin unsegmented tube-like bodies with anterior mouths and longitudinal digestive tracts. They have a fluid-filled internal body cavity (pseudocoelum) which acts as a hydrostatic skeleton providing rigidity (so-called 'tubes under pressure'). Worms use longitudinal muscles to produce a sideways thrashing motion. Adult worms form separate sexes with well-developed reproductive systems. 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) [197]. Over a million nematode species on earth have been reported in recent studies though only 25,000 species have been described [198] as a result of their ability to adapt, small size, resistant cuticle, and simple body plan [199]. The phylum nematoda has a large variation in genome size ranging from 50-250Mb [200], due to high rate of large, spontaneous deletions [201]. We studied 454 transcriptome datasets of *Strongyloides ratti* (Chapter 4) and *Strongyloides stercoralis* (Chapter 6) as part of this thesis.

Scientific classification of Nematodes

Lineage(full)
cellular organisms;
Eukaryota;
Fungi/Metazoa group;
Metazoa;
Eumetazoa;
Bilateria;
Coelomata;
Protostomia;
Ecdysozoa;
Nematoda

Figure 1.3. Taxonomy of the phylum Nematoda. Adapted from Blaxter [202].

1.8.1.1 Strongyloides ratti

Strongyloides ratti is a parasitic nematode of rats. It can undergo two types of development ouside the host, homogonic (direct larval development) and heterogonic (free-living adults

and sexual reproduction). Direct larval development is similar to the development of the dauer stage of free-living nematodes like *C. elegans.* In the life cycle of *Strongyloides ratti,* there are two adult phases: one parasitic and one free-living (Figure 1.4). The adult parasitic generation occurs in the mucosa of the small intestine and is only female, which reproduce by parthenogenesis [203]. Eggs produced by these females are passed out of the host in faeces. These eggs develop into different ways: some develop directly into infective larvae while others develop into free-living adult males and females. These adult stages reproduced by sexual reproduction and their progeny develop, in turn, into infective larvae [204]. Thus, the free-living phase of the life cycle produces infective third stage larvae (iL3s) either directly or *via* a facultative free-living adult generation. Different life stages of *Strongyloides ratti* are shown in Figure 1.4.

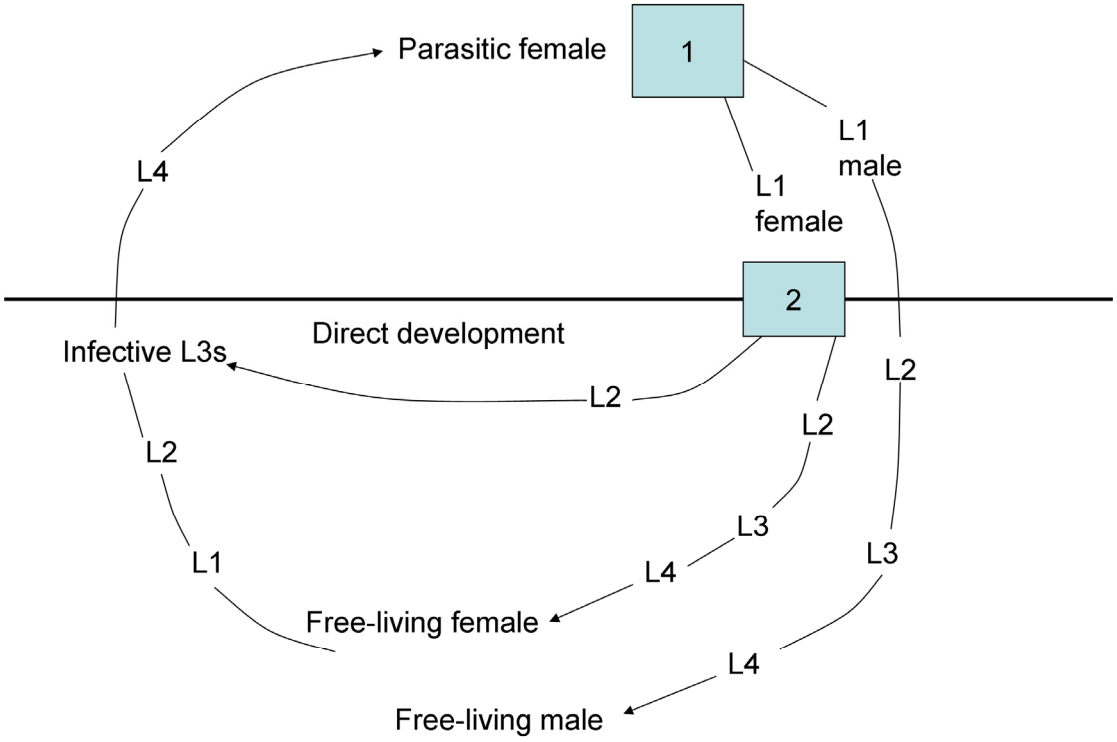


Figure 1.4. Different life stages of *Strongyloides ratti*. Adapted from Viney and Lok [205].

1.8.1.2 Strongyloides stercoralis

Strongyloides stercoralis is an intestinal nematode, which is one of two species of *Strongyloides* which infects humans [205]. Strongyloidiasis caused by *Strongyloides stercoralis* is a soiltransmitted helminthiasis distributed worldwide, affecting more than 100 million people [206, 207]. Recently, it was classified as one of the most neglected tropical diseases (NTD) [208]. Chronic infections in endemic areas may be maintained

asymptomatically for decades through the autoinfective cycle with the filariform larvae L3 [209, 210]. The *S. stercoralis* life cycle encompasses both free-living and parasitic stages. Adult female worms parasitizing the human small intestine lay eggs in the intestinal mucosa that hatch into rhabditiform larvae, which are shed in the stool. In the environment, under warm moist conditions that often characterize the tropical and subtropical areas where *S. stercoralis* is endemic, rhabditiform larvae can either molt into infective filariform larvae or develop through succeeding rhabditiform stages into free-living adults.

1.8.2 Trematodes

Trematodes (flukes) are flatworms (Figure 1.5) small flat leaf-like bodies with oral and ventral suckers and a blind sac-like gut. They do not have a body cavity (acoelomate) and are dorsoventrally flattened with bilateral symmetry. They exhibit elaborate gliding or creeping motion over substrates using compact 3-D arrays of muscles. Most species are hermaphroditic (individuals with male and female reproductive systems) although some blood flukes form separate male and female adults. Trematodes live mainly in the venous system (e.g., schistosome species), biliary system (e.g., Clonorchis), gut (e.g., Fasciolopsis), or airway (e.g., Paragonimus). We studied 454 transcriptome dataset of *Echinostoma caproni* (Chapter 7) as a part of this thesis.

Scientific classification of Trematodes

Lineage(full)
cellular organisms;
Eukaryota;
Fungi/Metazoa group;
Metazoa;
Eumetazoa;
Bilateria;
Acoelomata;
Platyhelminthes;
Trematoda

Figure 1.5. Taxonomy of the phylum Trematoda. Adapted from Blaxter [202].

1.8.2.1 Echinostoma caproni

Echinostoma caproni is an intestinal trematode which infects wide range of hosts, although its compatibility differs considerably between species based on worm survival and development [211].

Many animals may serve as definitive hosts for various echinostome species, including aquatic birds, carnivores, rodents and humans. Unembryonated eggs are passed in feces and develop in the water before hatching and penetrate into the first intermediate host, a snail. The intramolluscan stages include a sporocyst, one or two generations of rediae, and cercariae. The cercariae may encyst as metacercariae within the same first intermediate host or leave the host and penetrate into new second intermediate host. Depending on the species, several animals may serve as the second intermediate host, including other snails, bivalves or fish. The definitive host becomes infected after eating infected second intermediate hosts. The full life cycle flow chart is shown in Figure 1.6.

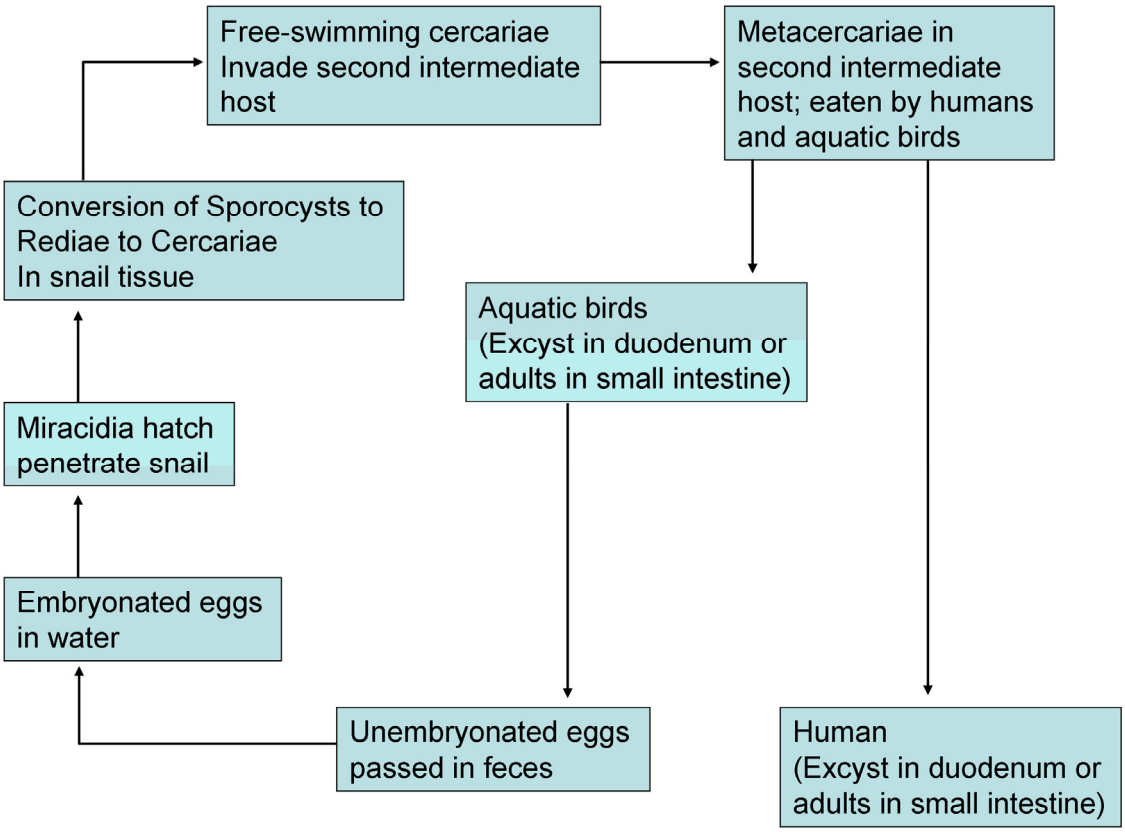


Figure 1.6. Life cycle of *Echinostoma caproni*. Adapted from the Divison of Parasitic diseases, Centers for Disease Control and Prevention, USA [212].

1.8.3 Cestodes

Cestodes (tapeworms, Figure 1.7) have long flat ribbon-like bodies with a single anterior holdfast organ (scolex) and numerous segments. They do not have a gut and all nutrients are taken up through the tegument. They do not have a body cavity (acoelomate) and are flattened to facilitate perfusion to all tissues. All tapeworms are hermaphroditic and each segment contains both male and female organs. Cestodes include the intestinal tapeworms like *Diphyllobothrium latum* (fish tapeworm), *Taenia saginata* (beef tapeworm), and *Taenia solium* (pig tapeworm). We studied EST datasets of *Echinococcus multilocularis* and *Echinococcus granulosus* (preliminary analysis section, Chapter 3) available from the Sanger Institute [213] as a part of this thesis.

Scientific classification of Cestodes

Lineage(full)
cellular organisms;
Eukaryota;
Fungi/Metazoa group;
Metazoa;
Eumetazoa;
Bilateria;
Acoelomata;
Platyhelminthes;
Cestoda

Figure 1.7. Taxonomy of the phylum Cestoda. Adapted from Blaxter [202]

1.8.3.1 Echinococcus granulosus

Echinococcus granulosus or Dog Tapeworm is a cyclophyllid cestode that reside in the small intestine of canids as an adult. It causes hydatid disease in livestock and humans. From definitive hosts eggs are passed in the feces. After ingestion by a suitable intermediate host (sheep, goat or cattle), the egg hatches the larva that penetrates the intestinal wall and migrates through the circulatory system into various organs, especially the liver and lungs. In these organs, the larva develops into a cyst that enlarges gradually, producing protoscolices and daughter cysts that fill the cyst interior. The definitive host becomes infected by ingesting the cyst-containing organs of the infected intermediate host. The full life cycle flow chart is shown in Figure 1.8.

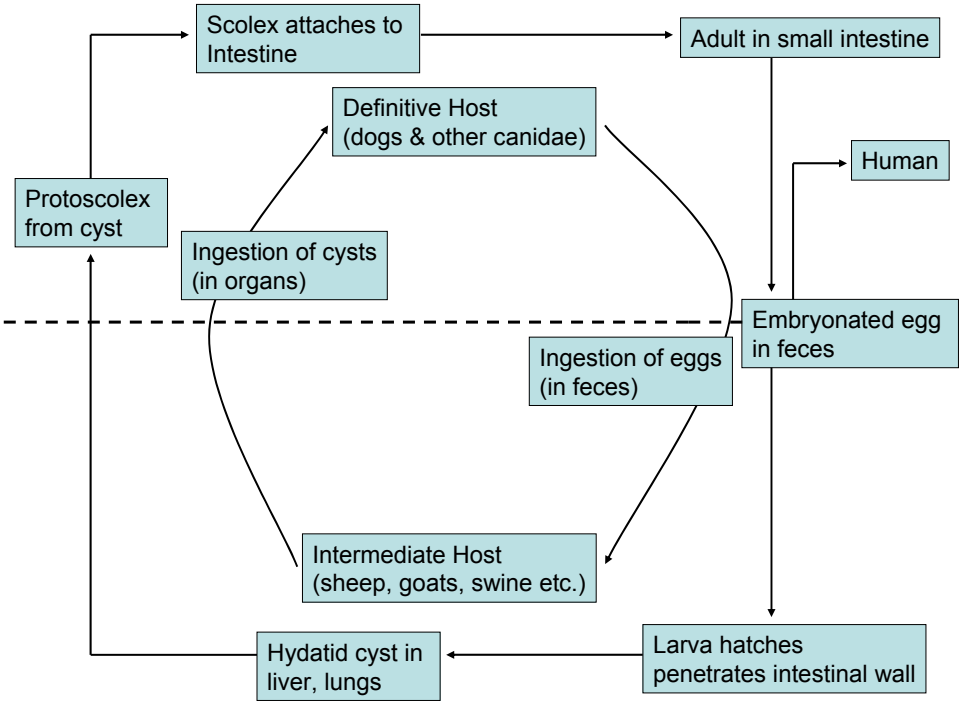


Figure 1.8. Life cycle of *Echinococcus granulosus.* Adapted from the Divison of Parasitic diseases, Centers for Disease Control and Prevention, USA [214].

1.8.3.2 Echinococcus multilocularis

Echinococcus multilocularis is a cyclophyllid tapeworm, which produces the disease known as echinococcosis in certain terrestrial mammals including humans.The life cycle of *Echinococcus multilocularis* is similar to one in *Echinococcus granulosus* but larval growth (in the liver) remains indefinitely in the proliferative stage, resulting in invasion of the surrounding tissues. Humans become infected by ingesting eggs, with resulting release of larva in the intestine and the development of cysts in various organs.

1.9 Other organisms: pathogenic fungi

Apart from helminths, there are other pathogenic organisms like bacteria, fungi or virus, which can cause infectious diseases to human, plants and animals. To extent the scope of computational approach for secretome analysis, developed as a part of this thesis, we have applied our secretome analysis approach with modifications according to the data to proteome data of pathogenic fungi (*Cryptococcus neoformans* forms: *Cryptococcus gattii* and *Cryptococcus grubii*) available from the Broad Institute [215, 216]. Computational prediction of ES fungal proteins has been deployed previously in the form of databases like

FunSecKB [217] and Fungal Secretome Database [218]. This application shows the general applicability of our approach to pathogenic organisms other than helminths.

1.9.1 *Cryptococcus neoformans*

Cryptococcus neoformans is an encapsulated yeast that can live in both plants and animals. *C. neoformans* has three varieties: *C. neoformans v. neoformans*, *C. neoformans v. grubii* or *Cryptococcus grubii* and *C. neoformans v. gattii* or *Cryptococcus gattii.C. grubii* and *v. neoformans* have a worldwide distribution and are often found in soil which has been contaminated by bird excrement. In the past, *C. gattii* has often been associated with *Eucalyptus* trees in tropical and subtropical climates, causing disease in immunocompetent hosts at low incidences [219, 220] but over the past decade, *C. gattii* has emerged as a primary pathogen in northwestern North America, including both Canada and the United States [221-226]. The genome sequence of *C. neoformans v. neoformans* was published in 2005 [227]. With the availability of new genome sequences of different varieties of *C. neoformans* and other pathogens**,** it is now possible to get insights in molecular epidemiology of pathogens. The compuataional approaches used in this thesis will help to improve understanding of population dynamics during an outbreak by understanding ES proteins, and may lead to novel methods for the rapid identification, treatment, and diagnosis of emerging infections through novel therapeutic targets.

1.10 Objectives

The overall objective of this thesis was to suggest improved bioinformatic approaches to analyse, validate and identify ES proteins responsible for parasites through transcriptomic and proteomic studies. These studies led to the establishment and integration of bioinformatic analysis protocols with new tools.

As a part of this thesis, a series of related studies have been conducted on the secretome analysis of parasites of socio-economic importance, elucidating novel therapeutic targets and the involvement of ES proteins in key biological processes important for parasite survival inside the host. The secretome studies and the integration of transcriptomic and proteomic analysis have helped to understand the parasite development inside the host and host-parasite interactions, which could lead to identifying novel intervention stratergies.

In the case of parasitic helminths, transcriptomics has been used extensively to understand the molecular basis of parasitism and for developing novel therapeutic strategies against parasitic infections, based on predicted ES proteins. Different experimental and computational methods are available for secretome analysis and therefore, these methods need to be studied by considering advantages and disadvantages for each method. In view of this, we conducted a review of the different methods and computational tools in secretome analysis for setting up the scene and rationale for the development of new analysis pipeline with new and updated bioinformatic tools. With the advent of NGS, new opportunities to explore the molecular biology of parasites have widened and large scale transcriptomic data were generated from disease causing parasitic helminths like *S. stercoralis, S. ratti* and *E. caproni* in our collaboration labs. These datasets are good resources for secretome analysis, which require serious evaluation of different computational tools and the development of bioinformatic pipeline for the analysis.

The suite of computational tools for the different phases of functional annotation (pathway mapping, protein domain mapping, GO mapping, ES protein prediction and drug targets mapping) can also be used to annotate newly sequenced pathogen genomes. This extension of our secretome analysis approach has been applied to pathogenic organisms other than helminths*, viz.* pathogenic fungi.

The studies outlined above represent the specific aims of this thesis, which are described in the following sections and have been addressed in detail in seven publications presented in this thesis:

Specific objectives are described below:

1. Review the different methods for secretome analysis, identify recent secretome data for parasites and evaluate bioinformatic tools for their application to parasites with the key focus of identifying novel therapeutic targets (Chapter 1).

2. Carry out a preliminary analysis on representative parasite transcriptomic datasets: *Echinococcus multilocularis* and *Echinococcus granulosus*; by modifying the existing bioinformatics analysis pipelines to address the current requirements of NGS transcriptomic data as well as for the successful prediction ES proteins produced by classical and non-classical secretion pathways. (Chapter 3).

3. Develop an updated secretome analysis protocol to analyse transcriptomic data arising from both EST and NGS technologies, based on the findings of the preliminary analysis in Chapter 3 and apply it to representative NGS data: 454 *Strongyloides ratti* (Chapter 4)

4. Implement the secretome analysis protocol, developed in Chapter 4, for the large-scale analysis of helminth EST data available from dbEST and make the data publicly available for novel therapeutic applications, *via* the Helminth Secretome Database (Chapter 5).

5. Apply our secretome analysis approach, developed in Chapters 4 and 5, and other relevant computational methods in collaboration with parasitologists, to analyse the transciptome of important parasitic helminths:
    5.1 The infective third larval (L3i) stage of *Strongyloides stercoralis.* This organism is responsible for causing the human disease, strongyloidiasis, affecting annually more than 100 million people worldwide. (Chapter 6)

    5.2 The first large-scale transcriptome analysis of the adult stage of *Echinostoma caproni.* This organism belongs to the family, Trematoda: Echinostomatidae, affecting more than 40 million people worldwide each year. Also, this family of helminths serves as an ideal model for the study of several aspects of the biology of intestinal helminths. (Chapter 7)

6. Extend our generic computational approach with other relevant organism/family-specific datasets to analyse the proteomic data from pathogens other than parasitic helminths, i.e. pathogenic fungi (Chapter 8)

The significance of the study, its novelty and future directions have been summarized in the concluding chapter (Chapter 9).

Pages 31-37 of this thesis have been removed as they contain published material. Please refer to the following citation for details of the article contained in these pages:

Ranganathan, S., Garg, G. (2009). Secretome: clues into pathogen infection and clinical applications. Genome Medicine, 1, Article number 113. https://doi.org/10.1186/gm113

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	Thesis Publication
In silico secretome analysis of *Echinococcus multilocularis* and *Echinococcus granulosus* using expressed sequence tags	3	2
In silico secretome analysis approach using next generation sequencing transcriptomic data	4	3
Helminth Secretome Database (HSD): a collection of helminth excretory/secretory proteins predicted from expressed sequence tags (ESTs)	5	4
Transcriptome analysis of *Strongyloides stercoralis* L3i larvae identifies targets for intervention in a neglected disease	6	5
Transcriptome characterization of the model organism, *Echinostoma caproni*	7	6
High-throughput functional annotation and data mining of fungal genomes to identify therapeutic targets	8	7

Chapter 3: *In silico* secretome analysis of *Echinococcus multilocularis* and *Echinococcus granulosus* using expressed sequence tags

3.1 Preliminary data analysis

Transcriptomic data is the representation of actively expressed genes in a cell at any given time. In case of parasitic helminths; transcriptomics has been used extensively to understand the molecular basis of parasitism and for developing novel therapeutic strategies against parasitic infections by prediction of ES proteins (as detailed in Chapter 1). Transcriptomic studies lead to the prediction of ES protein prediction, for identifying novel therapeutic targets, especially for neglected organisms such as parasites.

The only bioinformatics pipleline available for ES protein prediction and annotation, EST2Secretome [79] can only predict classically secreted proteins whereas it is well established that non-classically secreted proteins are especially important for parasitic organisms [57, 78].

We carried out preliminary secretome analysis of *Echinococcus multilocularis* (EM) and *Echinococcus granulosus* (EG) ESTs to achieve the goals of this thesis, addressing specifically secretion *via* classical as well as non-classical pathways. To this end, we integrated the bioinformatic tool, SecretomeP [126] for the identification of non-classically and classically secreted proteins into the existing EST2Secretome pipeline.

We further extended the ES protein analysis section in EST2Secretome to therapeutic target prediction by carrying out a BLAST search against DrugBank (detailed in Section 1.5.8 Therapeutic target prediction) and the results are presented in Publication 2.

In silico secretome analysis of *Echinococcus multilocularis* and *Echinococcus granulosus* using expressed sequence tags

Gagan Garg[1] and Shoba Ranganathan[1,2*]

[1] Dept. of Chemistry and Biomolecular Sciences, Macquarie University, Sydney NSW 2109, Australia.

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

*Corresponding author

Email addresses:

GG: gagan.garg@mq.edu.au

SR: shoba.ranganathan@mq.edu.au

Abstract

Background

The secretome of a parasite play an important role during parasitic infections. Parasites secrete or excrete a variety of molecules into their hosts through classical and non-classical secretory pathways, that modify or customize their niche within the host, in order to survive immune attack. Identifying secretory proteins involved during echinococcosis will lead to the discovery of potential therapeutic targets to control this disease.

Results

We developed a computational approach for prediction and annotation of excretory/secretory (ES) proteins using expressed sequence tags (ESTs) data. For the prediction of non-classically secreted proteins, we have used an improved computational strategy. We have analysed *Echinoccocus granulosus* (EG) and *Echinococcus multilocularis* (EM) EST data available from the Sanger Institute. From 23,684 ESTs for EG and EM, we derived 6173 unigenes (contigs and singletons), which were translated into 4,435 proteins. Based on our improved ES protein prediction approach, we identified 2,630 ES proteins. We were able to annotate 1943 ES proteins with protein domains and families and 743 ES proteins with pathways. In addition, we have identified five representative ES proteins, which have no homologues in the host organism but are homologous to lethal RNAi phenotypes in *C. elegans*, as potential therapeutic targets.

Conclusion

We report a preliminary computational approach using freely available computational tools for the secretome analysis of EST data. This approach has been used to analyse *Echinococcus granulosus* and *Echinococcus multilocularis* EST data for *in silico* excretory/secretory protein prediction and analysis, providing a foundation for developing new therapeutic solutions for echinococcosis disease.

Background

The secretome is defined as the complete set of proteins secreted by a cell. The secretome of a parasite plays an important role in parasitic infections and host-parasite interactions. During infection, excretory/secretory (ES) proteins are present at the host-parasite interface and help parasites to proliferate inside the host [1].

In case of parasitic helminths including tapeworms [2], transcriptomics has been used extensively to understand the molecular basis of parasitism. These data have been reliably used for the prediction of ES proteins [3, 4]. However, none of transcriptomic studies have extensively covered ESP prediction, especially the non-classically ES proteins. A study on *Fasciola hepatica* [5] from our group revealed that parasites produce ES proteins through classical as well as non-classical secretory pathways, to modify or customize their niche within the host, in order to suppress the host's immune system. Identification of these non-classical ES proteins along with classical ES proteins is thus essential to understanding the molecular mechanism of parasitic infections.

Here, we discuss a bioinformatics approach for secretome analysis, which is based on improvements in our existing pipeline, EST2Secretome [3]. We have addressed the issue of predicting non-classically secreted proteins, as well as compared our results with drug targets, available in the DrugBank [6] and applied our new approach to example datasets of *Echinococcus multilocularis* (EM) and *Echinococcus granulosus* (EG) expressed sequence tag (ESTs). These organisms are cyclophyllid cestodes (tapeworms), responsible for causing the life threatening disease, hepatic echinococcosis (HD) or hydatid disease in mammals, including humans. *E. granulosus* is a small tapeworm of length of 2-7 mm. There are ten distinct genetic types (G1-10) within *E. granulosus*, with differing geographical distribution. Cystic echinococcosis (CE) occurs as the result of infection by the larval stages of *E. granulosus*. CE is the most common form of HD, with a worldwide distribution. *E. multilocularis* is a small cestode, 1.2-4.5 mm long. Hepatic alveolar echinococcosis (AE) results from infection by the larval forms of *E. multilocularis* (EM).

EM develops in the liver and is characterized by an alveolar structure, made up by several vesicles surrounded by large granulomas. EM is able to elicit a strong cellular immune response in the host liver [7]. Secretory products of these parasitic helminths are involved in their parasitic activity. These products have been characterized before in these two organisms [8, 9]. Identifying secretory proteins involved in hydatid disease will help identify the mechanism of hydratid disease and also, provide leads for the discovery of new therapeutic solutions to control this disease.

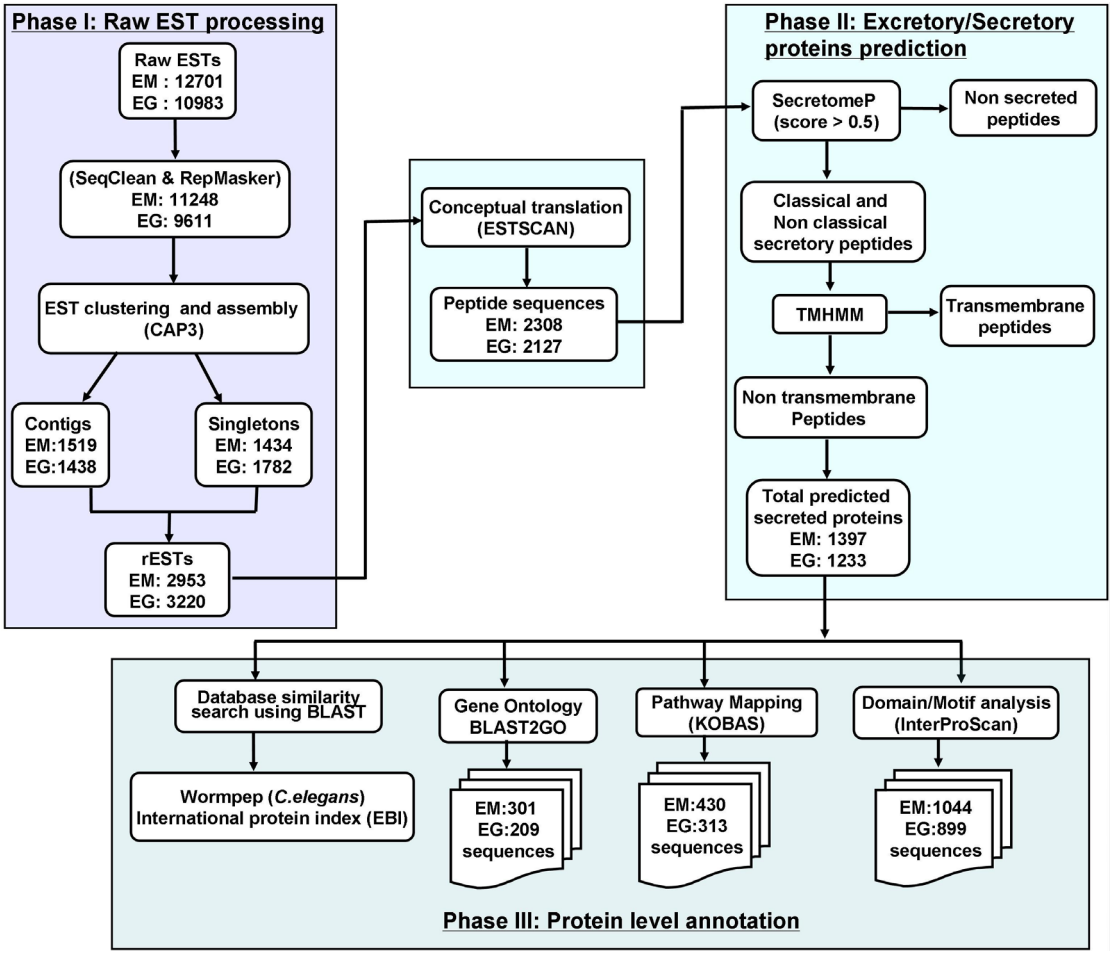


Figure 1: Bioinformatics workflow for secretome analysis.

Bioinformatics workflow comprising Phase I (pre-processing and assembly), II (prediction of excretory/secretory proteins) and III (Protein-level annotation)

Results

Our computational approach, incorporating three key components, was constructed. The different components of the workflow system (Figure 1) are linked using Perl, Python and bash shell scripts.

Extraction and assembly of EG and EM EST datasets

Initially a total of 12701 ESTs for EM and 10983 ESTs for EG were downloaded and stored in different directories on our local Linux machine. As set out in the workflow (Figure 1), raw ESTs were pre-processed using Seqclean [10] and RepeatMasker [11] for removing very short or vector sequences and to mask sequence repeats, respectively. 20,859 (88.07%) processed ESTs were passed to CAP3 [12] for *de novo* assembly. The assembly results in 1,519 contigs and 1,434 singetons (2,953 unigenes) for EM and 1,438 contigs and 1,782 singletons (3,220 unigenes) for EG. These unigenes were conceptually translated into 2,308 proteins for EM and 2,127 proteins for EG using ESTScan [13] (shown in Figure 1).

Prediction of excretory/secretory (ES) proteins

ESP prediction is carried out in Phase II of the workflow (Figure 1). Firstly, 2,308 putative proteins generated from EM contigs were passed to SecretomeP [14] and 1,553 were predicted as secreted both using classical and non-classical secretory pathways. These 1,553 secreted proteins were then input to TMHMM [15] and 156 proteins were identified as transmembrane proteins. These transmembrane proteins were removed from the set of secreted proteins, resulting in 1,397 proteins predicted as ES proteins from the computational prediction pipeline for EM.

In case of EG, 1,379 proteins were predicted as secreted by SecretomeP. O these, 146 were predicted as transmembrane proteins by TMHMM. Finally, 1,233 proteins were considered as ES proteins for EG.

Annotation of EG and EM ES proteins

ES proteins of EG and EM were annotated based on protein families and domains using Interproscan [16] and mapped to biochemical pathways using KOBAS [17]. Out of 1,397 ES proteins predicted for EM, we were able to annotate 1044 (74.7%) proteins with protein domains and families and 430 ES proteins were mapped to KEGG pathways [18]. For EG, 899 (72.9%) ES proteins were annotated with protein domains and families and 313 ES proteins were mapped to KEGG pathways (Figures 2 and 3). 301 and proteins were mapped to gene ontology terms (GO) [19] for EM and EG respectively. 616 (44 %) EM ES proteins were found homologous to *Caenorhabditis elegans* proteins [20], whereas in case of EG 467 (37.8 %) were found homologous to *C. elegans*.proteins.

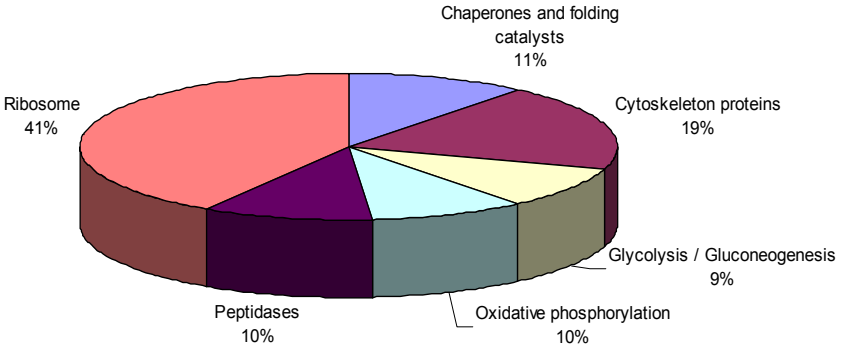


Figure 2: Major pathways found in EG ES proteins.

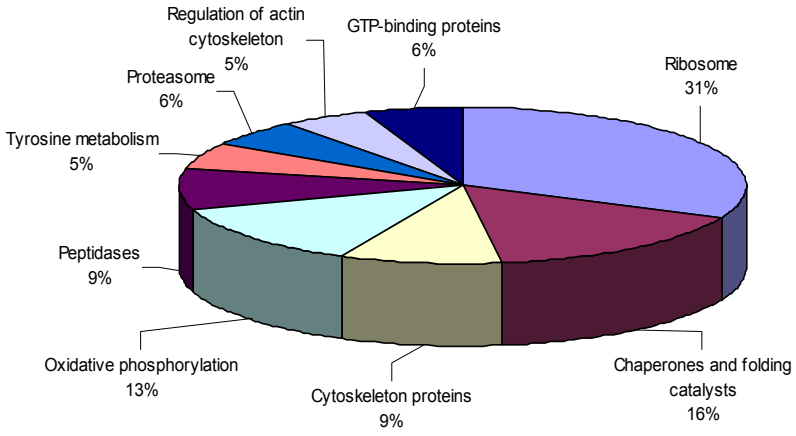


Figure 3: Major pathways found in EG ES proteins.

We also checked our set of predicted ES proteins against known drug targets in DrugBank and 283 and 204 ES proteins were found homologous to known drug targets for EM and EG, respectively (Figure 1).

EM and EG ES proteins as therapeutic targets

Based on annotation and mapping of EM and EG ES proteins with known drug targets, we found 5 ES proteins (4 for EG and 1 for EM), which have lethal RNAi phenotypes present in *C. elegans* and represent potential therapeutic targets (Table 1).

Discussion

Using the EG and EM EST data, we have carried a preliminary study for prediction and analysis of ES proteins with the help of computational tools. For this study, we have selected programs that are freely available under academic licence. All the programs used in our approach are available with free academic licence, which can be easily installed on UNIX based operating system.

Biological implications of the results

Millions of people globally suffer from echnicoccosis, caused by the cyclophilid cestodes, EG and EM. Here, we have analysed EG and EM EST data available from the Sanger Institute for the prediction and analysis of ES proteins. The recently published *in vitro* cultured *E. granulosus* (EG) ES protein study identified 32 proteins [21], which is approximately 40-fold less than our predicted ES proteins from EG. Most of the targets predicted from *Echinococcus* ES protein data here show high similarity to *Schistosoma mansoni* and *Schistosoma japonicum*, trematodes responsible for causing schistosomasis, which is the second most socioeconomically devastating parasitic disease after malaria. The genomes of *S. mansoni* [22] and *S. japonicum* [23] have been sequenced and there is literature evidence that these drug targets are used in schistosoma studies. EG_Contig 25 maps to tubulin β-1 chain (in Drugbank), which is a target for albendazole. This drug is

used for the treatment of hydatid disease, supporting the effectiveness of our approach in identifying therapeutic targets [6]

EG_Contig 60 maps to SJCHGC04801 protein, in *S. japonicum*. This protein belongs to the Ras family, several of whose members are involved in the GTPase pathway. This protein is annotated as involved in the GTPase-mediated signal transduction biological process. The GTPase pathway is involved in many diseases [24].

EG_Contig 107 maps to annexin in *S. mansoni*. Annexin was used as a potential biomarker in schistosoma transcriptomic studies [25] and also observed at the host-parasite interface in *Echinococcus granulosus* infection [26]. Therefore, annexin might serve as a potential therapeutic agent for echinococcosis.

EG_Contig 376 maps to calmodulin. Calmodulin is involved in calcium ion binding. Calmodulin is also tested as a drug target in case of *Schistosoma spp ,* however no activity was observed against *S. mansoni,* although the calmodulin domain was recently predicted to be a drug target for *S. mansoni* [27].

EM_Contig 252 maps to a hypothetical protein, with no annotation. There are many hypothetical proteins provided as drug targets in the TDR Targets Database [28]. We have predicted these five peptide sequences as potential targets of therapeutic value but further validation is required as all the facts described above are based on Schistosoma studies, and not on actual *Echnicoccus* spp as there is huge variation in the biochemistry of *Echnicoccus* spp. in different hosts [29].

Conclusions

We predicted and analysed *Echinococcus granulosus* and *Echinococcus multilocularis* ES proteins, from transcriptomic (EST) data, addressing both classical and non-classical secretory pathways. This approach is effective in analysing large-scale EST data for organisms whose genomes have not been sequenced. This preliminary study will contribute to the development of more comprehensive *in silico* secretome analysis

computational approach, applicable to parasites as well as major host and model organisms.

Methods

Expressed sequence tags (ESTs) data sets

For this study, EST datasets for EM and EG were downloaded from the Wellcome Trust Sanger Institute, UK website [30] and analyzed locally

Components of the computational approach

Our approach to predict and annotate ES proteins is divided into three phases, shown in Figure 1, corresponds approximately to those in EST2Secretome [3]. EST2Secretome was developed with the aim to predict and annotate classical ES proteins from ESTs. In the study, we have extended the approach to reliably predict non-classical and classical secreted proteins, followed by detailed functional annotation and therapeutic target prediction.

Phase I: Preprocessing and assembly of raw data

FASTA files for each organism were cleaned to remove short and vector sequences using Seqclean and Univec [31]. Cleaned sequences were passed to Repeatmasker to mask repeats. The output from Repeatmasker was assembled using CAP3. Unigenes (contigs and singletons; also known as rESTs) generated as a result of assembly were conceptually translated into putative proteins using ESTScan.

Phase II: Prediction of excretory/secretory (ES) proteins

ES proteins were predicted using a combination of two tools, SecretomeP and TMHMM. SecretomeP is used for the prediction of both classical and non-classical secretory proteins. incorporating SignalP [32], to predict classical secretory proteins while TMHMM is used

to identify transmembrane proteins. In case of EST2Secretome, only SignalP was deployed for the prediction of ES proteins, thereby identifying only classically secreted proteins.

Firstly, the proteins generated from ESTScan were passed to SecretomeP for prediction of classical and non-classical secreted proteins. All the proteins, with a neural network (NN) score greater than 0.5 (default NN score) were considered as secretory and then passed to TMHMM for prediction of transmembrane proteins. Proteins which are predicted to have no transmembrane helices by TMHMM are considered finally as ES proteins.

Phase III: Annotation and comparative analysis of ES proteins

All the predicted ES proteins are annotated using a number of tools. We used Interproscan for protein domain and family classification, KOBAS is used for mapping ES proteins to KEGG pathways and BLAST2GO [33] was used for gene ontology mapping. ES proteins are searched for sequence similarity against the Wormpep database [20] for proteins similar to *C. elegans*. ES proteins are also searched for sequence similarity against known drug targets available from DrugBank for the prediction of therapeutic targets for echinococcosis disease using BLAST [34].

References

1. Ranganathan S, Garg G: **Secretome: clues into pathogen infection and clinical applications**. *Genome Med* 2009, **1**(11):113.
2. Yang D, Fu Y, Wu X, Xie Y, Nie H, Chen L, Nong X, Gu X, Wang S, Peng X *et al*: **Annotation of the transcriptome from *Taenia pisiformis* and its comparative analysis with three Taeniidae species**. *PLoS One*, **7**(4):e32283.
3. 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.
4. 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.** *Biotechnol Adv* 2010, **28:**222-231.
5. 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.
6. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, et al: **DrugBank 3.0: a comprehensive resource for 'omics' research on drugs.** *Nucleic Acids Res* 2011, **39:**D1035-1041.
7. Nunnari G, Pinzone MR, Gruttadauria S, Celesia BM, Madeddu G, Malaguarnera G, Pavone P, Cappellani A, Cacopardo B: **Hepatic echinococcosis: clinical and therapeutic aspects**. *World J Gastroenterol*, **18**(13):1448-1458.
8. Carmena D, Martinez J, Benito A, Guisantes JA: **Characterization of excretory-secretory products from protoscoleces of *Echinococcus granulosus* and evaluation of their potential for immunodiagnosis of human cystic echinococcosis**. *Parasitology* 2004, **129**(Pt 3):371-378.
9. Walker M, Baz A, Dematteis S, Stettler M, Gottstein B, Schaller J, Hemphill A: **Isolation and characterization of a secretory component of *Echinococcus multilocularis* metacestodes potentially involved in modulating the host-parasite interface**. *Infect Immun* 2004, **72**(1):527-536.
10. **Seqclean** Available at: http://www.tigr.org/
11. Smit, AFA, Hubley, R & Green, P. **RepeatMasker Open-3.0.** 1996-2010 Available at: http://www.repeatmasker.org

12. Huang X, Madan A: **CAP3: A DNA sequence assembly program.** *Genome Res* 1999, **9:**868-877.
13. 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.
14. 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:**349-356.
15. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes**. J Mol Biol 2001, 305:567-580.
16. Zdobnov EM, Apweiler R: **InterProScan--an integration platform for the signature-recognition methods in InterPro.** *Bioinformatics* 2001, **17:**847-848.
17. 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:**W720-724.
18. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic Acids Res* 2000, **28**(1):27-30.
19. 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**. *Nat Genet* 2000, **25**(1):25-29.
20. **Wormpep database**,http://www.sanger.ac.uk/Projects/C_elegans/WORMBASE/current/wormpep.shtml
21. Virginio VG, Monteiro KM, Drumond F, de Carvalho MO, Vargas DM, Zaha A, Ferreira HB: **Excretory/secretory products from in vitro-cultured *Echinococcus granulosus* protoscoleces**. *Mol Biochem Parasitol*, **183**(1):15-22.
22. Berriman M, Haas BJ, LoVerde PT, Wilson RA, Dillon GP, Cerqueira GC, Mashiyama ST, Al-Lazikani B, Andrade LF, Ashton PD *et al*: **The genome of the blood fluke Schistosoma mansoni**. *Nature* 2009, **460**(7253):352-358.
23. **The *Schistosoma japonicum* genome reveals features of host-parasite interplay**. *Nature* 2009, **460**(7253):345-351.
24. Lu Q, Longo FM, Zhou H, Massa SM, Chen YH: **Signaling through Rho GTPase pathway as viable drug target.** *Curr Med Chem* 2009, **16:**1355-1365.

25. Gobert GN, Tran MH, Moertel L, Mulvenna J, Jones MK, McManus DP, Loukas A: **Transcriptional changes in *Schistosoma mansoni* during early schistosomula development and in the presence of erythrocytes.** *PLoS Negl Trop Dis*, **4:**e600.
26. Diaz A, Ibarguren S, Breijo M, Willis AC, Sim RB: **Host-derived annexin II at the host-parasite interface of the *Echinococcus granulosus* hydatid cyst.** *Mol Biochem Parasitol* 2000, **110:**171-176.
27. Caffrey CR, Rohwer A, Oellien F, Marhofer RJ, Braschi S, Oliveira G, McKerrow JH, Selzer PM: **A comparative chemogenomics strategy to predict potential drug targets in the metazoan pathogen, *Schistosoma mansoni*.** *PLoS One* 2009, **4:**e4413.
28. Magarinos MP, Carmona SJ, Crowther GJ, Ralph SA, Roos DS, Shanmugam D, Van Voorhis WC, Aguero F: **TDR Targets: a chemogenomics resource for neglected diseases**. *Nucleic Acids Res*, **40**(Database issue):D1118-1127.
29. McManus DP: **Reflections on the biochemistry of Echinococcus: past, present and future.** *Parasitology* 2009, **136:**1643-1652.
30. Helminth data from Sanger Insitute Available at: http://www.**sanger**.ac.uk/resources/downloads/**helminths**/
31. **Univec** Available at : http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html
32. Bendtsen JD, Nielsen H, von Heijne G, Brunak S: **Improved prediction of signal peptides: SignalP 3.0.** *J Mol Biol* 2004, **340:**783-795.
33. 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.
34. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.

Table 1. Representative therapeutic target set of EG and EM secretory proteins, with lethal RNAi phenotypes and homologous to known drug targets

No.	Cluster ID (Sequence Length)	Homology (non-redundanant protein database search with BLASTP)	RNAi phenotype	Gene Ontology (BLAST2GO)	Protein domain analysis (Interproscan)	Drug target mapping (DrugBank)
1	EG_Contig25 (368)	Tubulin beta chain (*Echnicoccus multilocularis*, E-value = 0.0)	Embryonic lethal Slow growth Pronuclear migration defective early emb	GO:0005525:GTP binding GO:0007018:microtubule-based movement GO:0051258:protein polymerization GO:0003924:GTPase activity	PF00091: Tubulin	Tubulin beta-1 chain
2	EG_Contig 60 (236)	SJCHGC04801 protein (*Schistosoma japonicum*, E-value = 4e-84)	Protruding vulva Larval arrest Reduced brood size	GO:0005525:GTP binding GO:0007264:GTPase mediated signal transduction	PF00071: Ras family	Ras-related protein SEC4
3	EG_Contig 107 (176)	Annexin (*Schistosoma mansoni*,,E-value = 5e-30)	Embryonic lethal Reduced brood size Germ cell morphology variant	GO:0005544: calcium-dependent phospholipid binding	PF00191: Annexin	Methylaspartate ammonia-lyase
4	EG_Contig 376 (192)	Calmodulin (*Schistosoma mansoni*, E-value = 2e-74)	Slow growth Embryonic lethal Meiotic spindle defective	GO:0005509 :calcium ion binding	PF00036: EF hand	Calmodulin
5	EM_Contig 252 (93)	Hypothetical protein (*Schistosoma mansoni*, E-value = 9e-20)	Early larval lethal Larval arrest Reduced brood size	GO:0006898: receptor-mediated endocytosis	PTHR12868: NADH-ubiquinone oxidoreductase b22 subunit	NADH dehydrogenase [ubiquinone] 1 beta subcomplex subunit 9

3.2 Conclusions

Based on preliminary data analysis, we predicted and analysed *E. granulosus* and *E. multilocularis* ESPs, from transcriptomic (EST) data, addressing both classical and non-classical secretory pathways. The study results in the prediction of five therapeutic targets against echinococcois and also validates our approach for the extension of the available pipeline for secretome analysis using transcriptomic data. This approach is also effective in analysing large-scale EST data for organisms whose genomes have not been sequenced.

This preliminary study contributes to the development of more comprehensive *in silico* secretome analysis computational approach (Chapter 4), applicable to parasites as well as major host and model organisms.

Chapter 4: *In silico* secretome analysis approach using next generation sequencing transcriptomic data

4.1 Summary

Following the successful extension of the EST2Secretome analysis pipeline [79] in Chapter 3, to include the prediction of non-classical ES proteins and to identify therapeutic drug targets from EST datasets, the bioinformatics pipleline still needs to address the use of the currently available short read transcriptomic data from NGS technologies.

Here in this chapter, we developed a semi-automated computational approach for prediction and annotation of ES proteins using transcriptomic data from NGS platforms, using computational tools that are freely available under academic licence. For the prediction of non-classically secreted proteins, we have further improved the computational strategy, by including a homology matching to a dataset of 1080 experimentally determined parasitic helminth ES proteins, compiled from the literature on *Ancylostoma caninum, B. malayi, Clonorchis sinesis, F. hepatica, S. mansoni, S. japonicum, and T. circumcinta* . The protocol was tested on the 454 short reads of the parasitic nematode, *Strongyloides ratti*, a common gastrointestinal parasite of the rat, which is used as a model to study the disease, strongyloidiasis in humans.

The details of the protocol and the test results are presented in Publication 3, with Additional Files provided on CD.

Garg and Ranganathan *BMC Genomics* 2011, **12**(Suppl 3):S14
http://www.biomedcentral.com/1471-2164/12/S3/S14

BMC Genomics

PROCEEDINGS Open Access

In silico secretome analysis approach for next generation sequencing transcriptomic data

Gagan Garg[1], Shoba Ranganathan[1,2*]

From Asia Pacific Bioinformatics Network (APBioNet) Tenth International Conference on Bioinformatics – First ISCB Asia Joint Conference 2011 (InCoB/ISCB-Asia 2011)
Kuala Lumpur, Malaysia. 30 November - 2 December 2011

Abstract

Background: Excretory/secretory proteins (ESPs) play a major role in parasitic infection as they are present at the host-parasite interface and regulate host immune system. In case of parasitic helminths, transcriptomics has been used extensively to understand the molecular basis of parasitism and for developing novel therapeutic strategies against parasitic infections. However, none of transcriptomic studies have extensively covered ES protein prediction for identifying novel therapeutic targets, especially as parasites adopt non-classical secretion pathways.

Results: We developed a semi-automated computational approach for prediction and annotation of ES proteins using transcriptomic data from next generation sequencing platforms. For the prediction of non-classically secreted proteins, we have used an improved computational strategy, together with homology matching to a dataset of experimentally determined parasitic helminth ES proteins. We applied this protocol to analyse 454 short reads of parasitic nematode, *Strongyloides ratti*. From 296231 reads, we derived 28901 contigs, which were translated into 20877 proteins. Based on our improved ES protein prediction pipeline, we identified 2572 ES proteins, of which 407 (1.9%) proteins have classical N-terminal signal peptides, 923 (4.4%) were computationally identified as non-classically secreted while 1516 (7.26%) were identified by homology to experimentally identified parasitic helminth ES proteins. Out of 2572 ES proteins, 2310 (89.8%) ES proteins had homologues in the free-living nematode *Caenorhabditis elegans* and 2220 (86.3%) in parasitic nematodes. We could functionally annotate 1591 (61.8%) ES proteins with protein families and domains and establish pathway associations for 691 (26.8%) proteins. In addition, we have identified 19 representative ES proteins, which have no homologues in the host organism but homologous to lethal RNAi phenotypes in *C. elegans*, as potential therapeutic targets.

Conclusion: We report a comprehensive approach using freely available computational tools for the secretome analysis of NGS data. This approach has been applied to *S. ratti* 454 transcriptomic data for *in silico* excretory/secretory proteins prediction and analysis, providing a foundation for developing new therapeutic solutions for parasitic infections.

Background

The secretome of an organism is defined as the subset of proteins secreted by the cell [1]. This subset of proteins is usually known as excretory/secretory (ES) proteins [2], plays an important role in producing clinical infections in the host organism. ES proteins are the choice of new therapeutic solutions for different clinical infections, especially in the case of parasitic infections [3,4] because these proteins are present at the host-parasite interface and act as immunoregulators to host immune recognition for parasite survival inside the host organism [5].

Transcriptomic data is the representation of actively expressed genes in a cell at any given time. Earlier transcriptomic studies were based on generation of expressed sequence tags (ESTs) generated at different

* Correspondence: shoba.ranganathan@mq.edu.au
[1]Dept. of Chemistry and Biomolecular Sciences, Macquarie University, Sydney NSW 2109, Australia
Full list of author information is available at the end of the article

stages of an organism using traditional Sanger sequencing. These studies were restricted to the analysis of a few thousand ESTs at a time. Recent technological improvements in cDNA sequencing, using next generation sequencing (NGS) platforms, are able to generate millions of reads, to record the transcript profile of an organism at a given developmental stage. The read length generated through NGS is quiet short (50-400 bases) as compared to traditional Sanger sequencing (800-1000 bases). Thus, the assembly of shorter reads is challenging in terms of computational power and resources needed. These reads are assembled into long consensus sequences (clusters) known as contigs using assemblers such as ABySS [6], Velvet [7] and MIRA [8], which have been reviewed in a recent study [9]. ABySS and Velvet provide good results for genome assembly, while MIRA is very well tested for handling *de novo* transcriptome assembly [10]. Since the genomes of only a very few parasitic nematodes are currently available, *de novo* assemblers such as MIRA are the only option for NGS data from these neglected organisms.

Recently, NGS platforms have been used to generate large amounts of transcriptomic data for different organisms, including several helminth parasites like *Fasciola gigantica* [11], *Fasciola hepatica*[12], *Trichostrongylus colubriformis*[13], *Oesophagostomum dentatum*[14], *Haemonchus contortus* [15], *Dictyocaulus viviparus*[16], *Necator americanus* [17], *Clonorchis sinensis*[18], *Opisthorchis viverrini* [18] and *Teladorsagia circumcinta* [*19*]. Here, NGS data has been assembled with CAP3 alone [14,16] or with MIRA followed by CAP3 [12,18], based on combinations of assemblers performing better in a recent study [10]. However, none of these studies have extensively covered ES protein prediction and further analysis, for identifying therapeutic targets.

ES proteins were once considered to be secreted only through conventional secretion pathways, using N-terminal signal peptide signatures, but there are now many proteins which are found to be secreted by non-classical secretory pathways [20]. Usually non-classical secretory proteins are predicted through SecretomeP [21], which is the most widely used tool for non-classical secretory proteins. However in case of parasites, SecretomeP is not able to completely predict non-classical secretory proteins, as shown in the study of *Brugia malayi*[*22*]. Hence, a novel approach to identifying non-classically secreted proteins is required for comprehensive secretome analysis.

Transcriptomic data has been used extensively for the prediction of ES proteins in parasitic helminth studies [23]. EST2Secretome, a computational prediction and annotation pipeline for ES proteins from our group, was designed to handle ESTs from Sanger sequencing and currently has the following limitations: (i) assembly of short reads, (ii) prediction of non-classical secretory proteins and (iii) pathway mapping using KOBAS [24,25], which contains pathways that are not regularly updated.

In the present study, we have developed an updated computational approach for the prediction and annotation of ES proteins using NGS transcriptomic data overcoming the limitations of the earlier EST2Secretome pipeline. We have developed a robust assembly protocol for NGS data. In order to identify non-classically secreted proteins that are missed by SecretomeP, we have also compiled a dataset of experimentally determined ES proteins of parasitic helminths for homology-based prediction (details in the Methods section). Additionally, we have replaced KOBAS with KAAS [26], for efficient and up-to-date pathway identification.

We applied our approach to ~0.3M 454 transcriptomic reads for a parasitic nematode, *Strongyloides ratti*, which is a gastro intestinal nematode that infects rats, comprehensively reviewed by Viney [27] and is a Clade IV parasite [28]. Genome data is available only for the free living nematodes, *C. elegans*[*29*] and *C. briggsae*[*30*] from Clade V, which is adjacent to Clade IV and for a parasite, *Brugia malayi*[*31*] from Clade III, which is not similar to Clade IV parasites, whereas limited transcriptomic and proteomic data from experimental studies are available for several helminth parasites. As such, a BLASTX against a reference organism, as proposed recently [32] will not provide comprehensive annotation results, unless the fully annotated proteome of a very similar organism is available.

In adult phase, *S. ratti* is present in both parasitic (females only) and free living forms (male and female) [27]. Eggs produced by parasitic females develop into free living males, free living females and parasitic females by different larval stages. Our dataset is derived from the adult nematode, which includes parasitic and free living forms (sequencing details in the Methods section). The NGS data has been clustered and translated into proteins and ES proteins predicted using a series of computational tools, augmented by homology matching to our in-house dataset of experimentally determined parasitic helminth ES proteins. Predicted ES proteins have been annotated functionally in terms of protein families, domains and biochemical pathways. ES proteins have also been compared with proteomic data of the host (rat) and other nematodes, with an emphasis on the best characterized nematode, *C. elegans*. Such annotation techniques have enabled us to identify 19 novel targets, matching to lethal RNAi phenotypes in *C. elegans*, which could be considered in the development of future therapeutic strategies.

Methods

cDNA sequencing data sets

For this study, *S. ratti* cDNA sequencing data from the University of Liverpool [33] is used. cDNA libraries were prepared from adult helminths, comprising a mixture of parasitic females, free-living males and free-living females. Sequencing was performed using 454-FLX platform (Roche diagnostics). The pyrosequencing procedure used to prepare this dataset is described elsewhere [34].

Components of computational approach

Our approach to predict and annotate ES proteins is divided into three phases, shown in Figure 1, corresponding approximately to those in EST2Secretome [23]. EST2Secretome was developed with the aim to predict and annotate ES proteins from ESTs (generated mainly using Sanger sequencing) mainly from parasitic nematodes. Now with the use of NGS, the input sequence data has changed considerably in terms of read length and number; necessitating modifications to tackle NGS data as well reliably predict non-classical protein secretion and use updated annotation tools.

Phase I: extraction and assembly of data

FASTA and associated quality files were extracted from SFF file along with clipping of sequence adapters using the sff_extract software [35]. Extracted data from sff files is first assembled using the MIRA [8] (V3.2.0rc1) assembler using quality information. MIRA is our preferred assembler as it is an open source tool which is considered reliable for data from different NGS platforms [8] and it

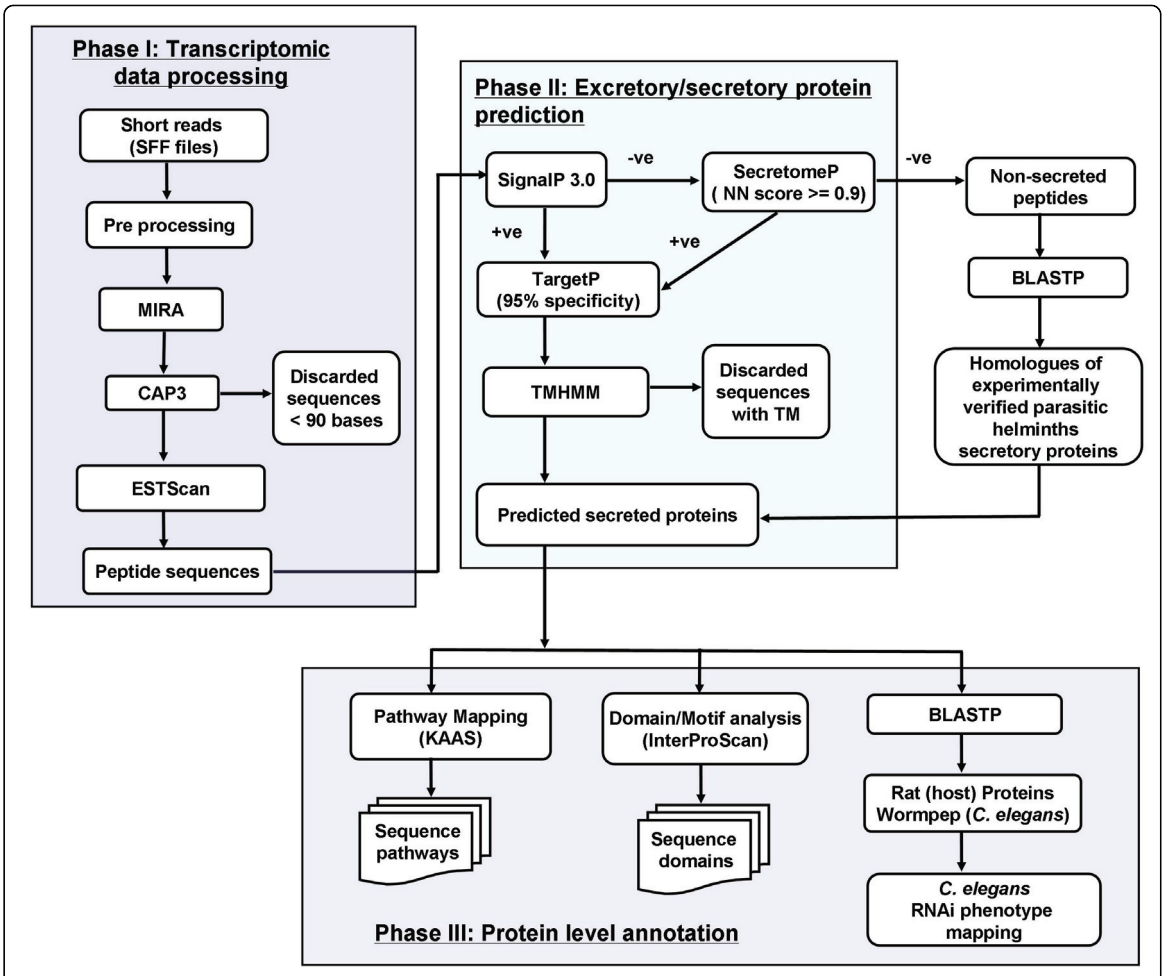


Figure 1 Bioinformatics workflow for secretome analysis. Bioinformatics workflow comprising Phase I (pre-processing and assembly), II (prediction of excretory/secretory proteins) and III (Protein-level annotation) were augmented by homologue identification from nematodes as well as parasitic nematodes, using specialized databases.

has been very well tested in other parasitic helminth transcriptomic studies [12,18]. For this dataset, we have used MIRA, ABYSS and Velvet, compared with Newbler (data not shown), MIRA giving the longest contigs. Contigs generated by MIRA are further passed to the Contig Assembly Program (CAP3) [36], to extend the MIRA assembly. This is in accord with an earlier study which suggests that serial assembly from two assemblers can improve the quality of the assembly [10]. Second order contigs generated using CAP3 are combined with MIRA contigs, to be conceptually translated into putative proteins using ESTScan [37].

Phase II: prediction of excretory secretory proteins

ES proteins were predicted using a combination of four tools, SecretomeP [21], SignalP [38], TargetP [39] and TMHMM [40]. SignalP is used for the prediction of classical secretory proteins, while SecretomeP predicts non-classical secretory proteins. TargetP is for the prediction of mitochondrial proteins and TMHMM identifies transmembrane proteins. Firstly, the proteins generated from ESTScan are passed to SignalP for prediction of classical secreted proteins. All the proteins, which are predicted as non-secretory (proteins having D score and signal peptide probability less than 0.5) are then passed to SecretomeP for prediction of non-classical secretory proteins. Proteins which obtain neural network (NN) score of greater than or equal to 0.9 are considered as non-classical secretory proteins. All the classical and non-classical secretory proteins are merged together and then scanned by TargetP. Proteins predicted as mitochondrial proteins by TargetP are omitted out from the set of predicted ES proteins and passed to TMHMM. Finally the proteins which are predicted to have no transmembrane helices are considered as ES proteins.

In addition to standard computational approaches for the prediction of ES proteins, we compiled a list of 1080 ES protein sequences of parasitic helminths (*Brugia malayi*, *Teladorsagia circumcinta*, *Schistosoma mansoni*, *Ancylostoma caninum*, *Schistosoma japonicum*, *Clonorchis sinesis* and *Fasciola hepatica*) from the literature [22,41-49]. A homology-based search with BLASTP [50] is used to further extract ES proteins from proteins which are predicted to be non-secretory by SecretomeP.

The results from computational tools are combined with those from BLAST searches, for functional annotation and analysis in Phase III.

Phase III: annotation and comparative analysis of ES proteins

All the predicted ES proteins are annotated using a number of tools. We used Interproscan [51] for protein domain and family classification. KAAS [26] is used for mapping ES proteins to KEGG pathways and to KEGG BRITE objects [52-54]. ES proteins are searched for sequence similarity against the Wormpep database (WS224) [55] for proteins similar to *C. elegans*. ES proteins are also searched for sequence similarity against rat (host) proteins and parasitic nematodes using BLASTP algorithm, to identify parasite-specific proteins. Comparative analysis of similarity of ES proteins with rat, parasitic nematodes and *C. elegans* proteins are analyzed using Simitri [56]. Proteins not homologous to the host (rat) proteome are further screened for RNAi phenotypes in *C. elegans*.

Hardware specifications

All the programs used in this study were installed on a 16 CPU Linux cluster (2.4 GHz, Intel(R)Xeon(R) E5530, 32 RAM) running on ubuntu server operating system. The computer intensive steps are sequence assembly (MIRA, CAP3) and protein functional annotation mapping (Interproscan). All other programs will run efficiently on current desktop systems.

Results

A semi-automated computational approach, incorporating three key components, was constructed. The different components of the workflow system (Figure 1) are linked using Perl, Python and bash shell scripts. This approach was applied to *S. ratti* 454 transcriptomic dataset to show its efficacy and utility.

Extraction and assembly of *S. ratti* data sets

Initially 296231 short reads (69488625 bases) were extracted from the sff file with 234±62 bases (average length ± standard deviation), and a GC content of 39.7%. The *de novo* assembly from MIRA results in 33222 contigs, which were passed to CAP3 to get a more robust assembly, with a minimum sequence overlap length of 40 bases and an identity threshold of 90%. Using CAP3, we are able to achieve a maximum contig length of 3620 bases as compared to maximum contig length of 2607 bases by Newbler [34]. The CAP3 assembly results in 3056 second order contigs and 25845 MIRA contigs (not assembled further by CAP3). The difference in results using MIRA+CAP3 and Newbler are shown in Table 1. We consider 25765 (99.6%) contigs with a minimum length of 90 bases, discarding sequences yielding peptides <30 amino acids, for further secretory protein prediction and analyses. A total of 3056 second order contigs and 25765 contigs were conceptually translated into 20877 proteins by ESTScan.

Prediction of ES proteins

ES protein prediction is carried out in Phase II of the pipeline (Figure 1). Firstly, 407 (1.9%) proteins were predicted as classical secreted proteins using SignalP. The remaining 20470 (98.05%) proteins, which were

Table 1 Comparison of results from different NGS assemblers

Assembler	No. of second order contigs	No. of contigs	Largest contig	Average length	N50*	N90*	Number of bases
MIRA [8] + CAP3 [29]	3056	25845	3620	402.36	406	253	11628536
Newbler [26]		25127	2607	407.11	409	252	10229510

*N50 refers to the length of the shortest contig such that the sum of contigs of equal length or longer is at least 50% of the total assembly size. While N90 refers to the length of the shortest contig such that the sum of contigs of equal length or longer is at least 90% of the total assembly size.

predicted as non secretory by SignalP were processed by SecretomeP for prediction of non-classical secretory proteins. A total of 923 (4.4%) proteins were predicted as non-classical secretory proteins using SecretomeP. The classical and non-classical secretory proteins (1330, 6.3%) from these two programs were analyzed by TargetP for mitochondrial proteins. Only 18 proteins were predicted as mitochondrial proteins using TargetP at 95% specificity. These 18 proteins were removed from the set of 1330 secreted proteins while 1312 secretory proteins were passed to TMHMM for the prediction of transmembrane proteins. 256 proteins, predicted as transmembrane proteins having one or more transmembrane helices, were removed from the secretory protein dataset. A total of 1056 (5.05%) proteins were finally predicted as ES proteins from the computational prediction pipeline.

Proteins that were considered non-secretory by SecretomeP were matched to our in-house dataset of 1080 non redundant experimentally determined parasitic helminth proteins, using the BLASTP similarity search. We found an additional 1516 (7.26%) proteins similar to known ES proteins by this homology search approach. Thus, for annotation and analyses in Phase III, we compiled a total of 2572 ES proteins, which is 12.3% of our putative proteins. This dataset is a more comprehensive collection of ES proteins of *S. ratti*, compared to those reported by other *S. ratti* secretome studies [57,58].

Annotation of *S. ratti* ES proteins

ES proteins are annotated based on protein families and domains using Interproscan and mapped to biochemical pathways using KAAS. Out of 2572 ES proteins predicted, we were able to annotate 1591 (61.8%) proteins with protein domains and families. The most represented Interpro terms are shown in Table 2 (complete results available from Additional file 1). We established pathway associations to 691 (26.8%) ES proteins. Among the most represented pathways are metabolic pathways, which are important for parasite survival inside the host. Predicted ES proteins are associated with important biological molecules, like enzymes, peptidases and protein kinases. The most represented KEGG BRITE objects and KEGG pathways are shown in Table 3 (full annotation available from Additional file 2) and Table 4 (full annotation available from Additional file 3).

Comparative analysis of *S. ratti* ES proteins with other organisms

2310 (89.8%) *S. ratti* ES proteins had homologues in the free-living nematode, *C. elegans*. 2220 (86.3%) ES proteins had homologues in parasitic nematodes. As *S. ratti* infects rats, we checked the similarity of ES proteins with the rat proteome. Similarity of *S. ratti* ES proteins to *C. elegans*, parasitic nematodes and rat proteins is shown using Simitri in Figure 2. We found 537 (20.8%) ES proteins had no homologues present in rat and are therefore preferred targets for parasite intervention strategies. 142 ES proteins are novel in the *S. ratti* dataset, with no known homologues to the host or any other nematode. 233 (9%) ES proteins, which are not present in the host (rat), have homologues present in *C. elegans*. Of these, 19 ES proteins (predicted from second order contigs from CAP3 assembly), which have lethal RNAi phenotypes present in *C. elegans*, (complete RNAi phenotype mapping available from Additional file 4) and represent potential therapeutic targets (Additional file 5).

Discussion

We demonstrated the utility of our new computational approach for the comprehensive prediction and analysis

Table 2 Top 15 most represented protein domains found in ES proteins using Interproscan

InterPro description	InterPro code	Number of ES proteins (%)
Protein Kinase like domain	IPR011009	126 (4.90)
Protein kinase, catalytic domain	IPR000719	114 (4.43)
Serine/threonine-protein kinase like domain	IPR017442	99 (3.85)
Serine/threonine-protein kinase domain	IPR002290	64 (2.49)
Serine/threonine-protein kinase active site	IPR008271	52 (2.02)
WD40 repeat like domain	IPR011046	40 (1.55)
WD40 repeat subgroup	IPR019781	39 (1.52)
WD40/YVTN repeat like domain	IPR015943	39 (1.52)
WD40 repeat	IPR001680	39 (1.52)
WD40 repeat domain	IPR017986	38 (1.47)
Tyrosine-protein kinase catalytic domain	IPR020635	37 (1.44)
WD40 repeat 2	IPR019782	37 (1.44)
Helicase C	IPR001650	35 (1.36)
NAD(P)-binding domain	IPR016040	29 (1.13)
Immunoglobulin-like fold	IPR013783	28 (1.09)

Table 3 Top 15 most represented KEGG pathways found in ES proteins predicted by KAAS

Pathway name	Number of ES proteins represented (%)
Metabolic pathways	109 (4.24)
Protein processing in endoplasmic reticulum	57 (2.22)
Ubiquitin mediated proteolysis	44 (1.71)
Wnt signalling pathway	29 (1.13)
Glycolysis / Gluconeogenesis	28 (1.08)
Spliceosome	28 (1.08)
Glutathione metabolism	26 (1.01)
Circadian rhythm - mammal	22 (0.85)
TGF- beta signalling pathway	22 (0.85)
RNA transport	20 (0.77)
Endocytosis	20 (0.77)
Purine metabolism	19 (0.74)
Phagosome	19 (0.74)
Proteasome	18 (0.70)
Drug metabolism	17 (0.66)

of ES proteins from transcriptomic data generated by NGS. The protocol will be implemented in a web server, in the future, after extensive testing of different assembly programs, and considering the choice of specific assemblers, based on the transcriptomic dataset, as proposed by Kumar and Blaxter [10]. For this study, we have selected programs that are freely available under academic licence. All the programs used in our approach are available with free academic licence, which can be easily installed on Linux platforms. Our use of MIRA followed by CAP3 for assembly of NGS data is simpler than the assembler combinations proposed by Kumar and Blaxter [10] and also used by studies on *Fasciola hepatica* [*12*], *Clonorchis sinensis* [*18*] and *Opisthorchis viverrini*[*18*] to generate second order contigs by CAP3 from contigs generated by MIRA which have open reading frames. The whole assembly for the current dataset was performed in approximately 3 hours CPU time using both MIRA and CAP3, whereas the use of CAP3 alone was not possible due to memory overflow with the current dataset, using hardware specified in the methods section. Although all the studies discussed here are more comprehensive in terms of transcriptome coverage (more than 0.5M 454 reads were generated), which is higher as compared to our current dataset of ~0.3M, none of them have comprehensively studied ES proteins. For example, the 454 transcriptomic study on *Fasciola hepatica* [*12*] reported only 1812 ES proteins (only 4%) from 44597 putative protein sequences generated from ESTScan, followed by ES protein predictions based on signal peptide identification by SignalP.

Table 4 Top 15 most represented KEGG BRITE objects found in ES proteins predicted by KAAS

BRITE object	Number of ES proteins represented (%)
Enzymes	282 (10.96)
Spliceosome	49 (1.90)
Chaperons and folding catalysts	44 (1.71)
Peptidases	44 (1.71)
Protein kinases	43 (1.67)
Ubiquitin system	37 (1.44)
Chromosome	34 (1.32)
Cytoskeleton proteins	27 (1.05)
DNA repair and recombination proteins	21 (0.82)
GTP-binding proteins	19 (0.74)
Proteasome	18 (0.70)
Transcription factors	17 (0.66)
Ribosome biogenesis	16 (0.62)
Translation factors	11 (0.43)
DNA replication proteins	9 (0.35)

Biological implications of the results

Millions of people globally suffer from Strongyloidiasis, caused by the parasitic nematode, *Strongyloides stercoralis*. *S. ratti* is a common gastro-intestinal parasite of the rat, which is used as a model to study Strongyloidiasis. Here, we have analysed *S. ratti* transcriptomic data from parasitic females, free-living males and free-living females for the prediction and analysis of ES proteins. Of the dataset of 2572 ES proteins 2310 (89.8%) had homologues in the free-living nematode, *C. elegans*, which is similar to earlier reported findings in Strongyloides EST analysis studies [59]. Many predicted ES proteins map to protein kinase domains as shown in Table 2, which are reported to be essential for parasitic activity in parasitic nematodes [60]. Protein kinases play a central role in signal transduction and hence are considered as drugabble targets. Another representative Interpro protein domains among *S. ratti* ES proteins were WD40 repeat domains (7.5%), which are associated with signalling transduction pathways [61]. These domains were also found among the top 20 most represented Interpro protein domains of *O. dentatum* putative proteins [14]. ES proteins also map to ribosomal protein interpro domains such as IPR000589 (Ribosomal protein S15), which is associated with ageing in *S. ratti* [62]. All the most representative KEGG pathways mapped to ES proteins shown in table 3 are required for parasite survival inside the host, as the secretome of a parasite is representative of its genome in the host environment. Major ES proteins map to enzymes, which are essential for metabolic pathways functioning and also very well reflected in our protein domain mapping. Other KEGG

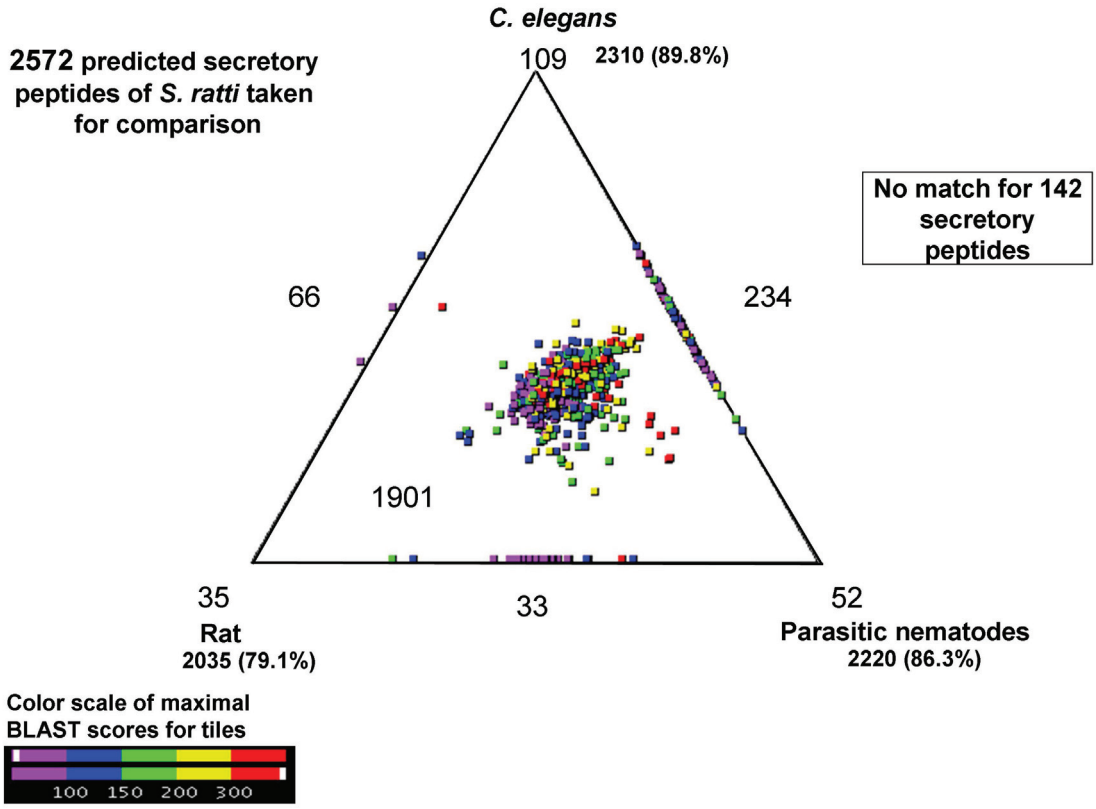


Figure 2 Comparison of *S. ratti* ES proteins with *C. elegans*, parasitic nematodes and rat (host) protein sequence databases using SimiTri. The numbers at each vertex indicate the number of proteins matching only that specific database. The numbers on the edges indicate the number of proteins matching the two databases linked by that edge. The number within the triangle indicates the number of S. ratti ES proteins with matches to all three databases.

pathways like purine metabolism and glutathione metabolism found in this study were also found in other parasitic nematodes excretory/secretory proteins analysis [23]. 22 (0.85%) ES proteins were mapped to the circadian rhythm – mammal pathway in *C. elegans*. This pathway is unexpected in the case of ES proteins of nematodes, however three proteins S-phase kinase-associated protein 1 (KO3094), cullin 1 (KO3347) and F-box and WD-40 domain protein 1/11 (KO3362) which were found in our ES proteins are common to Ubiquitin mediated proteolysis in *C. elegans*. The common components of several pathways have led to this unexpected result. KEGG BRITE objects (representative objects shown in Table 4) reflect the presence of essential proteins such as protein kinases, peptidases and proteasome among ES proteins for *S. ratti* survival inside the host organism. 44 (1.71%) ES proteins map to chaperones, which are responsible for host immune system modulation, such as the recently characterised *S. ratti* heat shock protein 10 [63]. Along with well known protein families found in ES proteins, we found some protein categories such as chromosome, DNA replication proteins and DNA repair and recombination proteins which are expected to be localized in the nucleus but found in *S. ratti* ES proteins. This pattern of exporting nuclear proteins to the secretome of a parasitic nematode was also observed in *Meloidogyne incognita* [64]. 66 secreted proteins were identified with putative nuclear localization such as DNA and RNA binding proteins including helicases in *M. incognita*, of which we observed the presence of helicase C domain in 35 (1.36%) *S. ratti* ES proteins. Contig 1289 and Contig 428 map to the metalloproteinase precursor in *S. stercoralis* [65], this is also well characterized protein in *Trichinella spirallis*[66]. Expresssion of an *S. stercoralis* metalloproteinase homologue was also found in the recent

transcript analysis of another intestinal nematode, *Strongyloides venezuelensis* [67]. Many of these potential therapeutic targets map to hypothetical proteins present in *C. elegans*, *C. briggsae* and *B. malayi* and having lethal phenotypes according to *C. elegans* RNAi phenotype mapping and could be considered as parasitism central genes [68] of *S. ratti*. Many of the putative proteins from *S. ratti* could be examined further after the publication of *S. ratti* genome, which is expected soon [69].

Methodological limitations

Integrated approaches similar to the one discussed in this paper have been applied to several socio-economically important parasites. These approaches are based on data available on the reference organism of that taxonomic order where limited data is available for the subject organism. For example, *C. elegans* is the most studied organism among nematodes. *C. elegans* data was used to create the translation matrix used by ESTScan, to translate potential coding regions in the assembled contigs into protein sequences. These translated coding regions were then used for ES proteins prediction. The use of a reference organism data for the translation matrix instead of using actual organism information may lead to false positives in peptides prediction as well as in ES protein prediction. Another limiting factor is that we are looking into the annotation of protein function in terms of primary sequence alone, rather than the 3D structure. Therefore, all the therapeutic targets predicted in this study are preliminary predictions which need to be further validated by additional computation analysis such as structural modelling and by experimental assays.

Conclusions

In this paper we demonstrate how different computational tools can be used together to extract the useful information of ES proteins from transcriptomic data. All the programs used in our approach are open source tools that are freely available for academic purposes. With the advent of NGS technologies, while there is a massive increase in sequence data, this data is extremely fragmented and of no use for information extraction as output from the sequencer. Our methodology will help in rapid assembly, fast annotation and reliable prediction of ES proteins. The approach is a generalized method which can be applied to any organism, although its main application is for neglected organisms whose genomes are not yet sequenced, with limited functional knowledge. Although we have used 454 transcriptomic data in this study but this methodology can be applied to transcriptomic data from other NGS platforms with slight modifications in terms of pre-processing, as data output formats obtained from different NGS platforms are different. Thus, this system will help us to carry out secretome studies for other parasitic organisms in future.

Additional material

Additional file 1: Protein domain mapping of *S. ratti* ES proteins. Represented Interpro domains found in *S. ratti* ES proteins using Interproscan (sheet1). Protein domains mapping of *S. ratti* excretory/secretory proteins (sheet2).

Additional file 2: KEGG pathways mapping of *S. ratti* ES proteins. Represented KEGG pathways found in ES proteins predicted by KAAS (Table S2).

Additional file 3: KEGG BRITE objects mapping of *S. ratti* ES proteins. Represented KEGG BRITE objects found in ES proteins predicted by KAAS (Table S3).

Additional file 4: RNAi Phenotype mapping of *S. ratti* ES proteins. RNAi Phenotype mapping of *S. ratti* ES proteins against known *C. elegans* known phenotypes (sheet1).

Additional file 5: Representative therapeutic targets set of *S. ratti* ES proteins. Representative therapeutic targets set of *S. ratti* ES proteins, homologous to *C. elegans* proteins with lethal RNAi phenotype and with no homologue in the host, rat.

List of abbreviations used

BRITE: Biomolecular Relations in Information Transmission and Expression; KEGG: Kyoto Encyclopedia of Genes and Genomes; KAAS: KEGG automatic annotation server.

Acknowledgements

We would like to thank Dr. Steve Paterson for providing the *Strongyloides ratti* cDNA sequencing data. We are thankful to Prof. Minoru Kanehisa for providing us the stand alone copy of KAAS program. GG would like to acknowledge Macquarie University for an Australian Post-graduate Award scholarship.

This article has been published as part of *BMC Genomics* Volume 12 Supplement 3, 2011: Tenth International Conference on Bioinformatics – First ISCB Asia Joint Conference 2011 (InCoB/ISCB-Asia 2011): Computational Biology. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2164/12?issue=S3.

Author details

[1]Dept. of Chemistry and Biomolecular Sciences, Macquarie University, Sydney NSW 2109, Australia. [2]Dept. of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, Singapore 117597.

Authors' contributions

SR directed the study. GG did the analysis. SR and GG contributed to writing the manuscript.

Competing interests

The authors declare that they have no competing interests.

Published: 30 November 2011

References

1. Skach WR: **The expanding role of the ER translocon in membrane protein folding.** *J Cell Biol* 2007, **179**:1333-5.
2. Tjalsma H, Bolhuis A, Jongbloed JD, Bron S, van Dijl JM: **Signal peptide-dependent protein transport in Bacillus subtilis: a genome-based survey of the secretome.** *Microbiol Mol Biol Rev* 2000, **64**:515-547.
3. Bonin-Debs AL, Boche I, Gille H, Brinkmann U: **Development of secreted proteins as biotherapeutic agents.** *Expert Opin Biol Ther* 2004, **4**:551-8.
4. Hotez PJ, Zhan B, Bethony JM, Loukas A, Williamson A, Goud GN, Hawdon JM, Dobardzic A, Dobardzic R, Ghosh K, Bottazzi ME, Mendez S,

Zook B, Wang Y, Liu S, Essiet-Gibson I, Chung-Debose S, Xiao S, Knox D, Meagher M, Inan M, Correa-Oliveira R, Vilk P, Shepherd HR, Brandt W, Russell PK: **Progress in the development of a recombinant vaccine for human hookworm disease: the Human Hookworm Vaccine Initiative.** *Int J Parasitol* 2003, **33**:1245-1258.

5. Hewitson JP, Grainger JR, Maizels RM: **Helminth immunoregulation: the role of parasite secreted proteins in modulating host immunity.** *Mol Biochem Parasitol* 2009, **167**:1-11.
6. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I: **ABySS: a parallel assembler for short read sequence data.** *Genome Res* 2009, **19**:1117-1123.
7. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18**:821-829.
8. 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**:1147-1159.
9. Miller JR, Koren S, Sutton G: **Assembly algorithms for next-generation sequencing data.** *Genomics* 2010, **95**:315-327.
10. Kumar S, Blaxter ML: **Comparing de novo assemblers for 454 transcriptome data.** *BMC Genomics* 2010, **11**:571.
11. Young ND, Jex AR, Cantacessi C, Hall RS, Campbell BE, Spithill TW, Tangkawattana S, Tangkawattana P, Laha T, Gasser RB: **A portrait of the transcriptome of the neglected trematode, Fasciola gigantica–biological and biotechnological implications.** *PLoS Negl Trop Dis* 2011, **5**:e1004.
12. 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.** *Biotechnol Adv* 2010, **28**:222-231.
13. 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**:1199-1207.
14. 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 Res* 2010, **38**:e171.
15. 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.
16. Cantacessi C, Gasser RB, Strube C, Schnieder T, Jex AR, Hall RS, Campbell BE, Young ND, Ranganathan S, Sternberg PW, Mitreva M: **Deep insights into Dictyocaulus viviparus transcriptomes provides unique prospects for new drug targets and disease intervention.** *Biotechnol Adv* 2011, **29**:261-271.
17. 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 Negl Trop Dis* 2010, **4**:e684.
18. 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 Negl Trop Dis* 2010, **4**:e719.
19. 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**:351-356.
20. Nickel W: **The mystery of nonclassical protein secretion. A current view on cargo proteins and potential export routes.** *Eur J Biochem* 2003, **270**:2109-2119.
21. 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**:349-356.
22. 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**:e410.
23. 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**:e301.
24. Mao X, Cai T, Olyarchuk JG, Wei L: **Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary.** *Bioinformatics* 2005, **21**:3787-3793.
25. 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**:W720-724.
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**:W182-185.
27. Viney ME: **The biology and genomics of Strongyloides.** *Med Microbiol Immunol* 2006, **195**:49-54.
28. Aguinaldo AM, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA: **Evidence for a clade of nematodes, arthropods and other moulting animals.** *Nature* 1997, **387**:489-493.
29. The C. elegans Sequencing Consortium: **Genome sequence of the nematode C. elegans: a platform for investigating biology.** *Science* 1998, **282**:2012-2018.
30. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A, *et al*: **The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics.** *PLoS Biol* 2003, **1**:E45.
31. 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.
32. Surget-Groba Y, Montoya-Burgos JI: **Optimization of de novo transcriptome assembly from next-generation sequencing data.** *Genome Res* 2010, **20**:1432-1440.
33. **Strongyloides ratti cDNA sequencing data.** , Available at: http://worm1.liv.ac.uk/file_summary.html.
34. Mello LV, O'Meara H, Rigden DJ, Paterson S: **Identification of novel aspartic proteases from Strongyloides ratti and characterisation of their evolutionary relationships, stage-specific expression and molecular structure.** *BMC Genomics* 2009, **10**:611.
35. **Sff_extract software.** , Available at: http://bioinf.comav.upv.es/sff_extract.
36. Huang X, Madan A: **CAP3: a DNA sequence assembly program.** *Genome Res* 1999, **9**:868-877.
37. 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.
38. Bendtsen JD, Nielsen H, von Heijne G, Brunak S: **Improved prediction of signal peptides: SignalP 3.0.** *J Mol Biol* 2004, **340**:783-795.
39. Emanuelsson O, Nielsen H, Brunak S, von Heijne G: **Predicting subcellular localization of proteins based on their N-terminal amino acid sequence.** *J Mol Biol* 2000, **300**:1005-1016.
40. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.** *J Mol Biol* 2001, **305**:567-580.
41. 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**:535-543.
42. Gourbal BE, Guillou F, Mitta G, Sibille P, Theron A, Pointier JP, Coustau C: **Excretory-secretory products of larval Fasciola hepatica investigated using a two-dimensional proteomic approach.** *Mol Biochem Parasitol* 2008, **161**:63-66.
43. Ju JW, Joo HN, Lee MR, Cho SH, Cheun HI, Kim JY, Lee YH, Lee KJ, Sohn WM, Kim DM, *et al*: **Identification of a serodiagnostic antigen, legumain, by immunoproteomic analysis of excretory-secretory products of Clonorchis sinensis adult worms.** *Proteomics* 2009, **9**:3066-3078.
44. Knudsen GM, Medzihradszky KF, Lim KC, Hansell E, McKerrow JH: **Proteomic analysis of Schistosoma mansoni cercarial secretions.** *Mol Cell Proteomics* 2005, **4**:1862-1875.
45. 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.** *Mol Cell Proteomics* 2009, **8**:1236-1251.
46. Moreno Y, Geary TG: **Stage- and gender-specific proteomic analysis of Brugia malayi excretory-secretory products.** *PLoS Negl Trop Dis* 2008, **2**:e326.
47. 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**:10-19.
48. Mulvenna J, Hamilton B, Nagaraj SH, Smyth D, Loukas A, Gorman JJ: **Proteomics analysis of the excretory/secretory component of the blood-feeding stage of the hookworm, Ancylostoma caninum.** *Mol Cell Proteomics* 2009, **8**:109-121.
49. Knudsen GM, Medzihradszky KF, Lim KC, Hansell E, McKerrow JH: **Proteomic analysis of Schistosoma mansoni cercarial secretions.** *Mol Cell Proteomics* 2005, **4**:1862-1875.
50. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403-410.
51. Zdobnov EM, Apweiler R: **InterProScan–an integration platform for the signature-recognition methods in InterPro.** *Bioinformatics* 2001, **17**:847-848.
52. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: **KEGG for representation and analysis of molecular networks involving diseases and drugs.** *Nucleic Acids Res* 2010, **38**:D355-360.
53. 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**:D354-357.
54. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 2000, **28**:27-30.
55. **Wormpep database.** http://www.sanger.ac.uk/Projects/C_elegans/WORMBASE/current/wormpep.shtml, release wormpep224, date Mar 04, 2011.
56. Parkinson J, Blaxter M: **SimiTri–visualizing similarity relationships for groups of sequences.** *Bioinformatics* 2003, **19(3)**:390-395.
57. Abe T, Nawa Y, Yoshimura K: **Protease resistant interleukin-3 stimulating components in excretory and secretory products from adult worms of Strongyloides ratti.** *J Helminthol* 1992, **66**:155-158.
58. Tazir Y, Steisslinger V, Soblik H, Younis AE, Beckmann S, Grevelding CG, Steen H, Brattig NW, Erttmann KD: **Molecular and functional characterisation of the heat shock protein 10 of Strongyloides ratti.** *Mol Biochem Parasitol* 2009, **168**:149-157.
59. Mitreva M, McCarter JP, Martin J, Dante M, Wylie T, Chiapelli B, Pape D, Clifton SW, Nutman TB, Waterston RH: **Comparative genomics of gene expression in the parasitic and free-living nematodes Strongyloides stercoralis and Caenorhabditis elegans.** *Genome Res* 2004, **14**:209-220.
60. Liotta F, Siekierka JJ: **Apicomplexa, trypanosoma and parasitic nematode protein kinases as antiparasitic therapeutic targets.** *Curr Opin Investig Drugs* 2010, **11**:147-156.
61. Jiang J, Struhl G: **Regulation of the Hedgehog and Wingless signalling pathways by the F-box/WD40-repeat protein Slimb.** *Nature* 1998, **391**:493-496.
62. Thompson FJ, Barker GL, Nolan T, Gems D, Viney ME: **Transcript profiles of long- and short-lived adults implicate protein synthesis in evolved differences in ageing in the nematode Strongyloides ratti.** *Mech Ageing Dev* 2009, **130**:167-172.
63. Tazir Y, Steisslinger V, Soblik H, Younis AE, Beckmann S, Grevelding CG, Steen H, Brattig NW, Erttmann KD: **Molecular and functional characterisation of the heat shock protein 10 of Strongyloides ratti.** *Mol Biochem Parasitol* 2009, **168**:149-157.
64. Bellafiore S, Shen Z, Rosso MN, Abad P, Shih P, Briggs SP: **Direct identification of the Meloidogyne incognita secretome reveals proteins with host cell reprogramming potential.** *PLoS Pathog* 2008, **4**:e1000192.
65. Gomez Gallego S, Loukas A, Slade RW, Neva FA, Varatharajalu R, Nutman TB, Brindley PJ: **Identification of an astacin-like metallo-proteinase transcript from the infective larvae of Strongyloides stercoralis.** *Parasitol Int* 2005, **54**:123-133.
66. Lun HM, Mak CH, Ko RC: **Characterization and cloning of metallo-proteinase in the excretory/secretory products of the infective-stage larva of Trichinella spiralis.** *Parasitol Res* 2003, **90**:27-37.
67. Yoshida A, Nagayasu E, Nishimaki A, Sawaguchi A, Yanagawa S, Maruyama H: **Transcripts analysis of infective larvae of an intestinal nematode, Strongyloides venezuelensis.** *Parasitol Int* 2011, **60**:75-83.
68. Thompson FJ, Barker GL, Hughes L, Viney ME: **Genes important in the parasitic life of the nematode Strongyloides ratti.** *Mol Biochem Parasitol* 2008, **158**:112-119.
69. **Welcome Trust Sanger Institute: five year plan for helminth sequencing.** , Available at: http://www.sanger.ac.uk/Projects/Helminths/.

doi:10.1186/1471-2164-12-S3-S14
Cite this article as: Garg and Ranganathan: *In silico* secretome analysis approach for next generation sequencing transcriptomic data. *BMC Genomics* 2011 **12**(Suppl 3):S14.

4.2 Conclusions

As a result of this study, of the 2572 ES proteins identified, 2310 (89.8%) ES proteins had homologues in the free-living nematode *C. elegans* and 2220 (86.3%) in parasitic nematodes. We could functionally annotate 1591 (61.8%) ES proteins with protein families and domains and establish pathway associations for 691 (26.8%) proteins. KEGG BRITE objects mapping reflect the presence of essential proteins such as protein kinases, peptidases and proteasome among ES proteins for *S. ratti* survival inside the host organism. Of the ES proteins identified, 19 have no homologues in the host or other mammalian organisms and are also homologues to lethal RNAi phetypes in *C. elegans* and therefore represent therapeutic targets, for strongyloidiasis for further examination using experimental assays.

With the advent of NGS technologies, there is a massive increase in sequence data and our methodology is timely for the rapid assembly, fast annotation and reliable prediction of ES proteins. The approach is a generalized methodology which serves as the basis for the development of the Helminth Secretome Database (HSD, Chapter 5) and has been applied to novel transcriptome datasets (Chapters 6 and 7) as well as adapted for complete genome/putative proteome annotation (Chapter 8).

Chapter 5: Helminth secretome database (HSD): a collection of helminth excretory/secretory proteins predicted from expressed sequence tags (ESTs)

5.1 Summary

Helminths are important socio-economic organisms, responsible for causing major parasitic infections in humans, plants and other animals (detailed in Sections 1.7 Introduction to helminths and 1.8 Types of helminths). These infections affect billions of people worldwide and leads to the loss of billions of dollars due to damage of crops and livestock every year. Another important factor to study these helmintic infections is due to their role in autoimmune diseases. Although a few studies have reported ES proteins from helminths only one large scale bioinformatics study has been carried out till date [79], for which the results are not freely available.

In this study, we focussed on comprehensively analysing the EST data available from dbEST, for 78 helminth species (64 nematodes, 7 trematodes and 7 cestodes), ranging from parasitic to free living organisms, for predicting ES proteins, followed by their annotation and therapeutic target prediction, adapting the protocol developed in Chapter 4. Experimentally determined ES proteins from helminth parasites from the literature have been used in HSD to validate the ES proteins predicted computationally. We have analysed our in house 1485 experimental ES data (collected from literature) with our computational pipeline and 461 (31%) ES proteins out of 1485 experimental ES proteins were determined by our computational pipeline. This result is comparable with the value of 25.2% for *Brugia malayi* L3 secreted proteins identified by SecretomeP [46]. The predictive ability of SecretomeP for parasites has been discussed in Chapter 4, publication 3 (p.58.)

The analysis results are presented in a free Helminth Secretome Database (HSD) at http://estexplorer.biolinfo.org/hsd. HSD is a repository for ES proteins predicted using classical and non-classical secretory pathways from EST data. A BLAST server is also implemented at HSD, for researchers to check the sequence similarity of other proteins against HSD predicted helminth ES proteins. The details of the database development and the results for all the helminth species analyzed are presented in Publication 4, with the Additional Files available on CD.

Garg and Ranganathan *BMC Genomics* 2012, **13**(Suppl 7):S8
http://www.biomedcentral.com/1471-2164/13/S7/S8

BMC Genomics

PROCEEDINGS Open Access

Helminth secretome database (HSD): a collection of helminth excretory/secretory proteins predicted from expressed sequence tags (ESTs)

Gagan Garg[1], Shoba Ranganathan[1,2*]

From Asia Pacific Bioinformatics Network (APBioNet) Eleventh International Conference on Bioinformatics (InCoB2012)
Bangkok, Thailand. 3-5 October 2012

Abstract

Background: Helminths are important socio-economic organisms, responsible for causing major parasitic infections in humans, other animals and plants. These infections impose a significant public health and economic burden globally. Exceptionally, some helminth organisms like *Caenorhabditis elegans* are free-living in nature and serve as model organisms for studying parasitic infections. Excretory/secretory proteins play an important role in parasitic helminth infections which make these proteins attractive targets for therapeutic use. In the case of helminths, large volume of expressed sequence tags (ESTs) has been generated to understand parasitism at molecular level and for predicting excretory/secretory proteins for developing novel strategies to tackle parasitic infections. However, mostly predicted ES proteins are not available for further analysis and there is no repository available for such predicted ES proteins. Furthermore, predictions have, in the main, focussed on classical secretory pathways while it is well established that helminth parasites also utilise non-classical secretory pathways.

Results: We developed a free Helminth Secretome Database (HSD), which serves as a repository for ES proteins predicted using classical and non-classical secretory pathways, from EST data for 78 helminth species (64 nematodes, 7 trematodes and 7 cestodes) ranging from parasitic to free-living organisms. Approximately 0.9 million ESTs compiled from the largest EST database, dbEST were cleaned, assembled and analysed by different computational tools in our bioinformatics pipeline and predicted ES proteins were submitted to HSD.

Conclusion: We report the large-scale prediction and analysis of classically and non-classically secreted ES proteins from diverse helminth organisms. All the Unigenes (contigs and singletons) and excretory/secretory protein datasets generated from this analysis are freely available. A BLAST server is available at http://estexplorer.biolinfo.org/hsd, for checking the sequence similarity of new protein sequences against predicted helminth ES proteins.

Background

According to the World Health Organization, over two billion people are suffering from human helmintasis and many more are at risk worldwide, especially in developing nations [1]. Helmintasis also results in the economic loss of billions of dollars due to damage of crops and livestock every year [2,3]. Besides their role in causing diseases, helminths also provide some protection against autoimmune diseases [4]. Free-living helminths such as *Caenorhabditis elegans* (the most studied helminth till date) serve as models to understand parasitism [5]. In the case of parasitic organisms, excretory/secretory (ES) proteins play an important role during the parasitic infection as these proteins are responsible for the regulation of the host's immune system for parasite survival inside the host. Such important roles played by ES proteins make these proteins

* Correspondence: shoba.ranganathan@mq.edu.au
[1]Dept. of Chemistry and Biomolecular Sciences and ARC Centre of Excellence in Bioinformatics, Macquarie University, Sydney NSW 2109, Australia
Full list of author information is available at the end of the article

attractive targets for the development of therapeutic strategies [6].

With rapid advances in sequencing technologies, sequencing data has been generated on large scale especially in the area of genomics and transcriptomics. Although short reads generated using 454 Roche pyrosequencing is the major sequencing technique used these days for generating transcriptomic data, expressed sequence tags (ESTs) remain the largest resource of helminthic transcriptomic data, with data available for several helminths. dbEST [7], the largest global repository of ESTs, recorded 71,276,166 entries (as on December 1, 2011, release 120111). EST data has been widely used for ES protein prediction in different transcriptomic studies [8,9] but most of the studies do not cover ES proteins comprehensively, especially non-classically secreted ones [10]. Also, it must be noted here that although the helminth proteome is directly affected by the developmental stage-specific expression and indirectly by change/decrease of 3'UTRs with their developmental stages, the data is so sparse in dbEST for some organisms that all available EST data from different stages are pooled together for the data analysis reported here. These mixed datasets have been used before for other nematode transcriptome studies like *S. ratti* studies [11,12]. We have used such a composite *S. ratti* dataset [12] in our previous secretome analysis [13].

In this study, we compiled ESTs for each helminth organism, covering nematodes, trematodes and cestodes and predicted ES proteins encoded by them, followed by functional annotation and therapeutic target analysis. Our earlier large-scale helminth secretome analysis was carried out using EST2Secretome [14] but the study only considered the classically secreted proteins, based on N-terminal secretory signals and covered only parasitic nematodes. Also, the ES protein sequences predicted as a part of this earlier study were not provided to the scientific community. We believe such predicted ES proteins are a valuable resource for understanding host-parasite interactions and for the development of new therapeutic strategies against helminth infections, for further validation using wet lab assays.

Recently we proposed a new bioinformatics workflow [13] for the prediction of classically and non-classically secreted proteins using 454 transcriptomic data of parasitic nematode, *Strongyloides ratti*. In the present study, we applied our workflow with minor modifications to accommodate EST datasets of 78 different helminth species available from dbEST, including those also available from Nematode.net [15], the largest provider of nematode ESTs.

The data were cleaned, assembled into Unigenes (contigs and singletons), which were then translated into proteins. From these putative proteins, ES proteins were predicted using a series of computational tools, which were further verified by sequence similarity to our in-house experimentally-determined parasitic helminth ES protein dataset (detailed in **Materials and methods**). Predicted ES proteins were functionally annotated in terms of similarity to other known proteins, biochemical pathways, protein families and domains. ES proteins were also searched for homologues in human, *C. elegans, Schistosoma mansoni* and *Schistosoma japonicum*. The analysis results are made available to the scientific community *via* the Helminth Secretome Database (HSD) [16] web portal All the Unigenes and ES protein sequence datasets can be browsed in FASTA format and are available for download. A BLAST web service is also provided for researchers to check the similarity of their protein sequences with our predicted ES datasets.

Materials and methods

Expressed sequence tags (ESTs) data sets

For this study, EST datasets for different helminth species were downloaded from NCBI dbEST [7] and analysed locally.

Bioinformatics approach components

Our bioinformatics approach has three phases as shown in Figure 1, similar to one tested on the *S. ratti* transcriptomic data [13] where we have used MIRA and CAP3 for reliable *de novo* transcriptome assembly, with these tools now combined by a Perl wrapper in iAssembler [17] for the robust assembly of both 454 and Sanger EST datasets. We have implemented our computational approach to the large helminth EST data from dbEST.

Phase I: Preprocessing and assembly of raw EST data

Each organism raw EST data were cleaned to remove short and vector sequences using Seqclean [18] and Univec [19] as a vector database. Seqclean is used to trim and validate ESTs for screening of vector contaminants, low quality and low complexity sequences. Cleaned sequences were assembled using iAssembler (version 1.3.1) [17]. The assembly was carried out using a minimum percent identity for sequence clustering and assembly of 95% contigs and singletons, collectively referred to as Unigenes. ESTScan [20] was used to conceptually translate Unigenes into putative proteins.

Phase II: Prediction and validation of excretory/secretory (ES) proteins

Prediction of ES proteins was carried out using a pipeline of four tools; SignalP [21], SecretomeP [22], TargetP [23] and TMHMM [24] followed by validation with experimentally determined helminth ES proteins as shown in the bioinformatic workflow (Figure 1). This approach of computational prediction of ES proteins has been successfully applied earlier to *Stronglyloides ratti* [13]. SignalP (version 3.0) was used for predicting classically secreted proteins applying options of organism category of eukaryotes and truncation of protein sequence at 70 amino

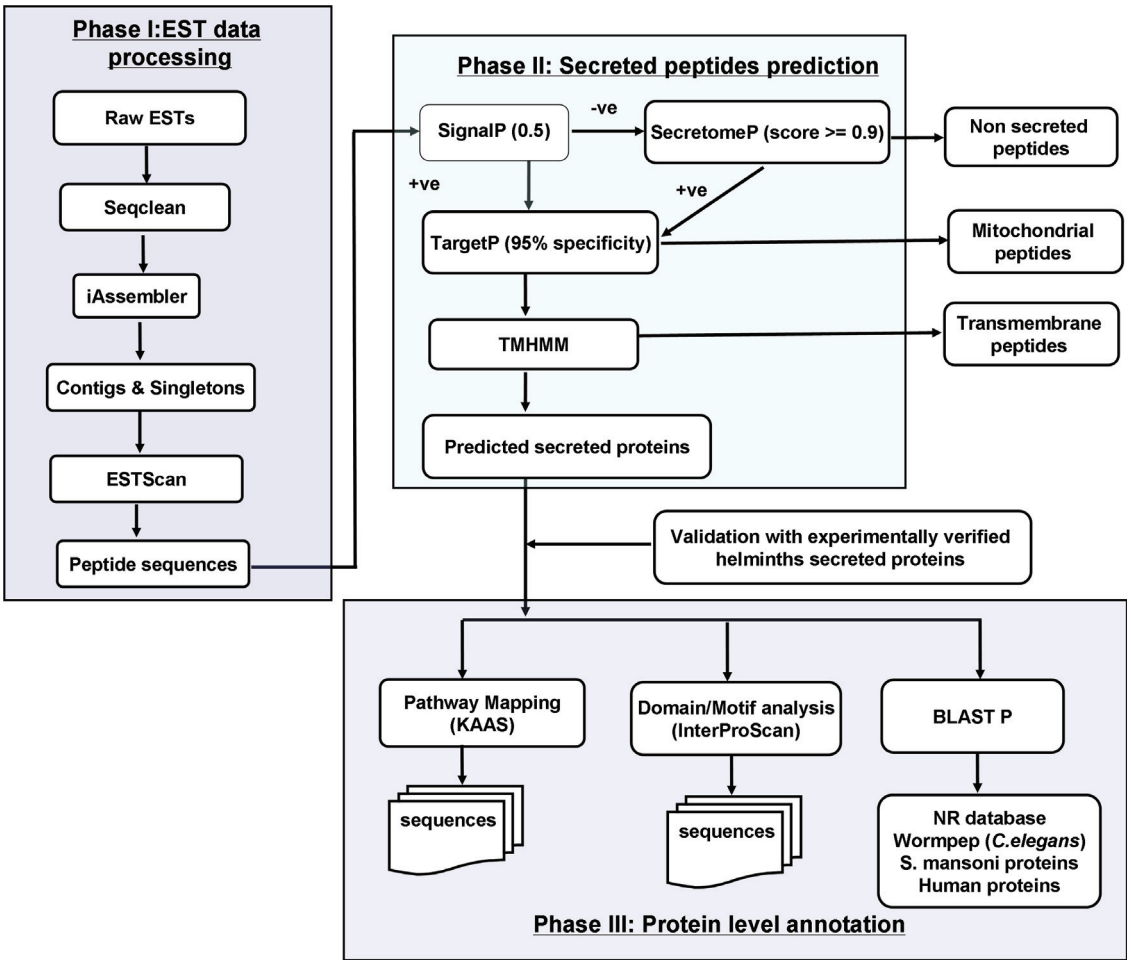


Figure 1 Secretome analysis workflow based on EST data. Secretome analysis workflow comprising Phase I (pre-processing and assembly of raw data), II (excretory/secretory (ES) protein prediction) and III (Protein-level functional annotation) based on homologue identification against different databases.

acids. SecretomeP (version 1.0) was used for predicting non-classically secreted proteins using default options. TargetP (version 1.1) was used for the prediction of mitochondrial proteins with a prediction cut-off of 0.78 for mitochondrial protein prediction and 0.73 for other locations. TMHMM (version 2.0) was used for the prediction of transmembrane proteins with default options. Firstly, putative proteins generated from ESTScan were analyzed by SignalP for predicting classically secreted proteins. Proteins were considered secreted, if the D-score and the signal peptide probability computed by SignalP are greater than 0.5. The remaining proteins were then input to SecretomeP for non-classical secretory protein prediction. Proteins were considered as secreted, if the neural network (NN) score from SecretomeP is greater than or equal to 0.9. The combined set of classical and non-classical secretory proteins is then passed to TargetP, to check for mitochondrial proteins. Mitochondrial proteins predicted by TargetP were then removed and the remaining predicted ES proteins analyzed by TMHMM. ES proteins with no transmembrane segments are considered for further analysis.

For the validation of computationally predicted ES proteins, we checked their sequence similarity against our compiled set of 1485 experimentally derived ES proteins of parasitic helminths (*Ancylostoma caninum, Brugia malayi, Clonorchis sinesis, Fasciola hepatica, Schistosoma mansoni, Schistosoma japonicum, Strongyloides ratti* and *Teladorsagia circumcinta)* compiled from literature [25-35] using BLAST [36].

Phase III: ES proteins annotation

Predicted ES proteins from phase II, were annotated for protein domain and family classification using Interproscan [37] including gene ontology (GO) terms option. KAAS [38], provide functional annotation by BLAST comparisons against the manually curated KEGG databases. This tool was used for KEGG pathways BRITE objects mapping [39,40]. ES proteins were independently also searched for homology matching against NCBI's non-redundant protein database and Wormpep (*C. elegans* proteins) [41] using BLAST [36]. ES proteins were also checked for homology matching against human proteins. BLAST was used with permissive (E-value: 1e-05), moderate (1e-15) and/or stringent (1e-30) search strategies. These tools provide fast annotation of large volumes of ES proteins and also reliably used before in other helminth transcriptomic studies [13,14].

Hardware and Software specifications

The Helminth Secretome database (HSD) is developed using MySQL 5 relational database [42]. The user-friendly interface is developed using PHP [43] for BLAST service and data management. The data is served using the Apache web server [44]. Open source tools used for this study were installed on a ubuntu server operating system based 16-CPU Linux cluster (2.4 GHz, Intel(R) Xeon(R) E5530, 32 RAM). Sequence assembly using iAssembler and protein functional annotation mapping using Interproscan are the most computationally intensive steps.

Results

Our recently developed bioinformatics workflow applied to 454 transcriptomic dataset of *S. ratti* was modified slightly to be applicable to EST data. The different components of the workflow were linked by Perl, Python and bash shell scripts (Figure 1).

Preprocessing and assembly of EST datasets

Initially a total of 870,223 ESTs ranging from 59 to 80,905 ESTs for different helminth species were downloaded and stored in different directories on our Linux server. According to the workflow (Figure 1), raw ESTs were cleaned first using Seqclean for removing very short or vector sequences. 846,741 (97.3%) cleaned ESTs were passed to iAssembler for *de novo* assembly. iAssembler is a standalone Perl package to assemble ESTs using iterative cycles of MIRA assemblies followed by CAP3 assembly. The tool gives much higher accuracy in EST assembly than other existing assemblers by employing an iterative assembly strategy and automated error corrections of mis-assemblies [17]. This strategy of using MIRA+CAP3 for *de novo* transcriptome assembly has been successfully implemented earlier for other helminth organisms [13] and therefore, using iAssembler is not only equivalent to these two programs but eliminates an extra step by incorporating the running of both programs in a single step. The assembly results in 303,657 Unigenes, comprising 103,791 contigs and 199,866 singletons. 245,814 proteins were obtained by conceptual translation of Unigenes using ESTScan (Table 1). Statistics of the EST analysis reported here, are provided in Additional file 1: Table S1.

ES protein prediction

Firstly, 18,287 (7.44%) proteins were predicted as classically secreted proteins out of 245,814 total putative proteins using SignalP. The remaining 227,527 (92.56%) putative proteins, predicted to be non-secretory by SignalP, were then scanned by SecretomeP for predicting non-classical secretory proteins. SecretomeP predicted a total of 9,244 (3.76%) non-classically secreted proteins. Combining the results from these two programs yielded a total of 27,531 (11.2%) classical and non-classical proteins which wer then checked by TargetP for identifying mitochondrial proteins. TargetP predicted only 0.17% proteins as mitochondrial, at 95% specificity. The remaining 27,116 proteins after removing 415 mitochondrial proteins were analysed by TMHMM for the prediction of transmembrane proteins. A total of 18,992 (7.72%) proteins were predicted finally as ES proteins after removing 8,126 proteins, which were predicted by TMHMM as transmembrane proteins with at least one transmembrane helix. This number is four fold higher than earlier reported (4710 ES proteins) in the secretome analysis of 39 parasitic nematodes [14].

All ES proteins that were predicted computationally were searched for sequence similarity against our non-redundant dataset of 1,485 experimentally determined ES proteins of various parasitic helminth organisms using BLASTP. We found 4,260 (22.43%) computationally predicted ES proteins homologous to known ES proteins. To the best of our knowledge, the HSD dataset is the most comprehensive collection of ES proteins of helminth organisms. It will serve as a rich source for developing new treatment strategies against parasitic infections and to study the molecular mechanisms of helminth organisms.

Annotation of ES proteins

ES proteins predicted in Phase II were mapped to known protein families and domains using Interproscan. These proteins were also mapped to biochemical pathways using KAAS. Of the 18,992 ES proteins predicted, we could annotate a total of 7,802 (41.08%) proteins with 2,340 different protein domains and families. ES proteins were annotated with Gene Ontology (GO) terms (2,893 for Biological Process, 4,558 for Molecular Function and 1,588 for Cellular Component) based on Interproscan annotations (species wide annotation available from Additional file 2: Table S2). Table 2 contains the most represented

Table 1 Summary of EST data analysis

Total number of species	78
Number of Nematode species	64
Number of Trematode species	7
Number of Cestode species	7
Total number of expressed sequence tags (ESTs) analysed	870,223
Total number of Unigenes (contigs + singletons)	303,657
Total number of putative peptides	245,814
Total number of excretory/secretory (ES) proteins predicted	18,992
Total number of ES proteins with annotation	11,390
Total number of ES proteins verified with experimentally derived helminth ES proteins	4,260

Interpro terms (complete results in Additional file 3: Table S3). Pathway associations were established for 5,893 (31.02%) ES proteins. Maximum number of ES proteins belongs to ***metabolism and human diseases***, making these proteins important in parasitic infections (Table 3). The predicted ES protein dataset comprises important biological molecules, including enzymes, the spliceosome and the ribosome. Table 4 contains the most represented KEGG BRITE objects among the different helminth species (full results available in Additional file 4: Table S4).

Comparative analysis of ES proteins with well-studied organisms

All computationally predicted ES proteins were searched for homology matching against the proteomes of *C. elegans* (Wormpep), *S. mansoni, S. japonicum* and human (Table 5) using BLASTP at an E-value of 1e-05. We also checked for homologues at more stringent E-values (1e-15, 1e -30) (complete results in Additional files 5, 6 and 7). Along with the similarity of our helminth ES protein dataset with other organisms, we checked these proteins for interacting partners based on data obtained from IntAct [45], BioGRID [46] and DIP [47] using BLASTP (interaction results in Additional file 8: Table S8).

Our dataset comprises a fairly high number (23, 30%) of parasitic helminth organisms infecting humans so ES proteins were checked for homology matching against the human proteome (Table 5). We found 13,756 (72.4%) ES proteins had no sequence similarity against human proteins and could be preferred targets for parasitic infections. These human dissimilar ES proteins were further searched for sequence similarity against known drug targets available from DrugBank [48]. Of these, 39 ES proteins from human parasitic helminth organisms were found similar to 27 known drug targets and represent potential therapeutic targets. These 27 drug targets are targeted by 75 small drug molecules, out of which 14 are clinically approved drugs. These therapeutic targets are also available from HSD.

Helminth Secretome database (HSD) data

All the ES proteins and Unigenes generated from this study can be viewed from the HSD data page for each organism. Along with proteins and Unigenes, users have

Table 2 Top 15 most represented domains found in ES proteins using Interproscan

InterPro description	InterPro code	Number of ES proteins (%)
Peptidase C1A, papain	IPR013128	305 (1.60%)
Transthyretin-like	IPR001534	298 (1.57%)
Peptidase C1A, papain C-terminal	IPR000668	276 (1.45%)
CAP domain	IPR014044	267 (1.40%)
Peptidase, cysteine peptidase active site	IPR000169	226 (1.19%)
Allergen V5/Tpx-1-related	IPR001283	204 (1.07%)
Thioredoxin-like fold	IPR012336	190 (1.00%)
C-type lectin fold	IPR016187	170 (0.89%)
Peptidase C1A, cathepsin B	IPR015643	137 (0.72%)
C-type lectin	IPR001304	135 (0.71%)
Metridin-like ShK toxin	IPR003582	127 (0.67%)
Domain of unknown function DUF148	IPR003677	127 (0.67%)
Saposin B	IPR008139	121 (0.64%)
Saposin-like	IPR011001	120 (0.63%)
Glycoside hydrolase, superfamily	IPR017853	120 (0.63%)

Table 3 KEGG pathways inferred from predicted ES proteins

Parent KEGG pathway	No. of ESPs	Top KEGG pathway in the category
Metabolism:		
Carbohydrate metabolism	296	Citrate cycle (TCA cycle)
Lipid metabolism	221	Fatty acid metabolism
Amino acid metabolism	217	Valine, leucine and isoleucine degradation
Energy metabolism	188	Oxidative phosphorylation
Glycan biosynthesis and metabolism	167	N-Glycan biosynthesis
Nucleotide metabolism	137	Purine metabolism
Xenobiotics Biodegradation and Metabolism	104	Metabolism of xenobiotics by cytochrome P450, Drug metabolism - other enzymes
Metabolism of Cofactors and Vitamins	95	Riboflavin metabolism
Metabolism of other amino acids	70	Glutathione metabolism
Biosynthesis of other Secondary Metabolites	38	Isoquinoline alkaloid biosynthesis
Metabolism of Terpenoids and Polyketides	29	Terpenoid backbone biosynthesis, Limonene and pinene degradation
Genetic Information processing:		
Folding, sorting and degradation	446	Protein processing in endoplasmic reticulum
Translation	334	RNA transport
Transcription	176	Spliceosome
Replication and repair	72	Nucleotide excision repair
Environmental information processing:		
Signal transduction	243	MAPK signaling pathway
Signalling, molecules and interaction	23	Cell adhesion molecules (CAMs)
Membrane transport	6	ABC transporters
Cellular processes:		
Transport and catabolism:	436	Lysosome
Cell Growth and Death	208	Cell cycle
Cell communication	130	Tight junction
Cell Motility	35	Regulation of actin cytoskeleton
Organismal systems:		
Immune system	291	Antigen processing and presentation
Nervous System	186	Glutamatergic synapse
Endocrine system	172	Insulin signaling pathway
Digestive System	80	Pancreatic secretion
Circulatory System	52	Cardiac muscle contraction
Excretory System	51	Proximal tubule bicarbonate reclamation
Development	47	Axon guidance
Environmental Adaptation	30	Circadian rhythm - mammal
Sensory System	15	Phototransduction
Human Diseases:		
Infectious Diseases	522	HTLV-I infection
Neurodegenerative Diseases	417	Alzheimer's disease
Cancers	241	Pathways in cancer (overview)
Cardiovascular Diseases	55	Hypertrophic cardiomyopathy (HCM), Arrhythmogenic right ventricular cardiomyopathy (ARVC)
Immune Diseases	44	Rheumatoid arthritis
Endocrine and Metabolic Diseases	19	Type II diabetes mellitus

the choice to view protein domain mapping and pathway mapping results. For ES proteins found homologous to known proteins, we provide annotation in the form of sequence identifiers along with percent identity and E-value for BLAST search, e.g. {Acantortus_UN0312; similar to gi|256096002|emb|CAR63732.1| hypothetical

Table 4 Top 15 putative functions inferred from predicted ES proteins

BRITE object	No. of species represented (%)
Peptidases	61
Spliceosome	50
Ribosome	49
Transcription Machinery	47
Protein kinases	38
Transfer RNA biogenesis	38
Chaperones and folding catalysts	34
Cytoskeleton proteins	34
Transcription factors	33
Ubiquitin system	26
Translation factors	25
Glycosyltransferases	24
DNA replication proteins	20
Amino acid related enzymes	19
Transporters	18

protein [Angiostrongylus cantonensis] (Evalue:2e-26, identity:50.00) unverified}. Each annotated ES protein is also tagged as verified or unverified based on the presence or absence of sequence similarity to experimentally determined parasitic helminth ES proteins (Phase II, Figure 1).

Helminth Secretome database (HSD) BLAST server

We have set up a BLAST server to run sequence similarity searches against our predicted ES protein datasets (Figure 2). All ES proteins are divided into three datasets (Nematode ES proteins, Cestode ES proteins and Trematode ES proteins) based on the organism. Users can also query our dataset of experimentally determined helminth ES proteins compiled from literature. The input data uploaded can be either nucleotide or protein sequences in FASTA format. A text box is also provided to paste the sequences directly into the BLAST query submission page. The results from the BLAST search are displayed in HTML format.

Table 5 Sequence homology inferred between predicted ES proteins in major helminth organism classes and other well-studied protein datasets at an E-value of 1e-05, using BLASTP

Dataset	Nematode hits	Trematode hits	Cestode hits
C. elegans proteins (Wormpep)	8457	345	280
S. mansoni proteins	3440	598	419
Human proteins	4539	408	326
NR protein database	10116	652	497
S. japonicum proteins	3456	612	416

Discussion

Here, we demonstrates the utility of our computational approach, integrating various open source tools, for the prediction and analysis of ES proteins using EST data available from dbEST. All software used in this study are freely available under academic licence. These tools can be installed on different flavours of UNIX based operating system. With the advent of next-generation sequencing (NGS) technologies, there are many transcriptomic studies completed especially for individual helminth species with good coverage but we have focussed on the coverage of a large number of helminth organisms for secretome analysis. The earlier analysis from our group using the EST2Secretome pipeline has now been extended to cover non-classical secretory proteins, with validation against experimentally known excretory/secretory proteins. We plan to carry out further prediction of ES proteins using more comprehensive helminth transcriptomic datasets from NGS platforms and provide the results through HSD.

Biological implications of this study

Several billion people worldwide are afflicted by infections caused by parasitic helminths. Infections from parasitic helminths, especially from nematodes also results in heavy economic losses worth billions of dollars due to agricultural crop and livestock infection each year. In this study, we have predicted and analysed ES proteins from the largest freely available EST data of several helminth organisms from dbEST.

Many predicted ES proteins map to peptidase domains and families (944,5%) which are reported to be involved in virulence activity (Table 2) and recently, cysteine peptidase expression was studied in a helminth pathogen, *Fasciola hepatica* [49]. Peptidases are well studied in *F. hepatica* for their role in migration and maturation of the parasite within its mammalian host [10]. Another representative Interpro protein domain among the helminth ES proteins is the transthyretin-like domain (1.57%). Transthyretin-like proteins were reported as novel proteins in the *B. malayi* secretome [50]. The most represented functional class among the helminth ES proteins are enzymes, essential for the function of metabolic pathways. Protein kinases, which play a key role in signal transduction, are also present in 38 species of this analysis.

Among the most representative KEGG pathways found in ES proteins are metabolic pathways (8.2%, as shown in Table 3). The top energy metabolism pathway, Oxidative phosphorylation and the top nucleotide metabolism pathway, purine metabolism, found in our pathway analysis were also reported in other helminth transcriptomic studies [13,51]. The second most represented KEGG pathway category among helminth ES proteins are

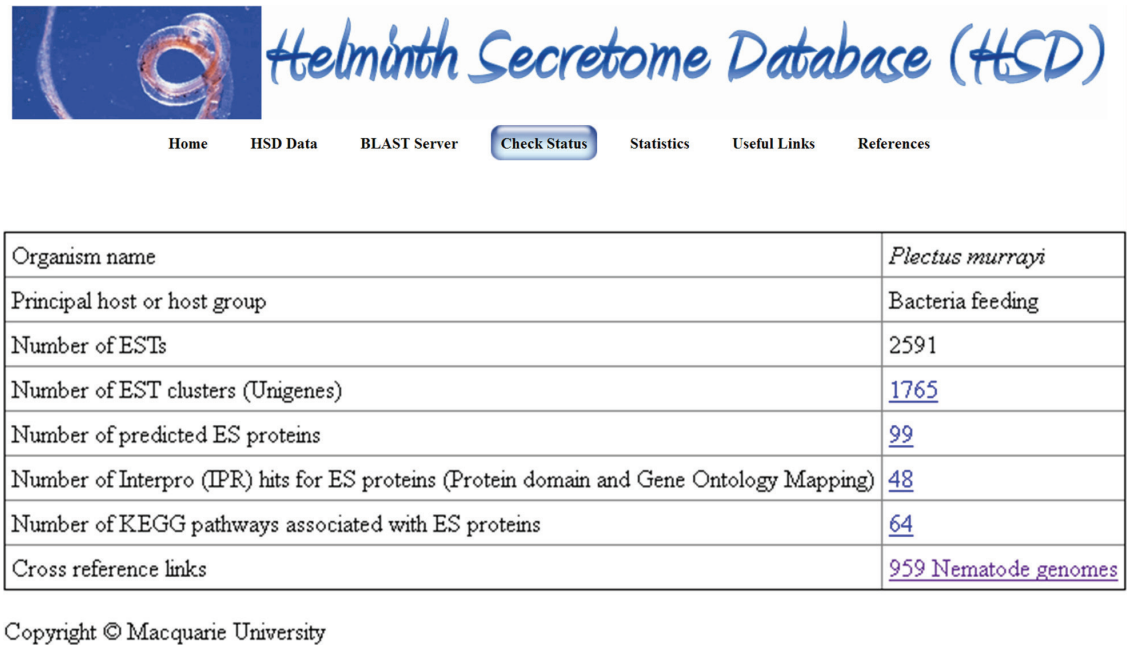


Figure 2 Screen shot of Helminth Secretome Database (HSD) species page. Helminth Secretome Database (HSD) species page of *Plectus murrayi*, a bacterial feeding nematode. Users can view Unigenes, ES proteins, protein domain and gene ontology and pathway mapping results from this page.

human diseases (6.83%). Association of helminth infections mainly by trematodes with cancers has been recently reviewed [52]. Carcinogenic parasitic trematodes like *Opisthorchis viverrini, Clonorchis sinensis* and *Schistosoma haematobium* were studied in different transcriptomics or genomics studies [53,54].

Representation of ES proteins with immune diseases leads us towards hygiene hypothesis [55]. It is well known that helminth ES proteins modulate the host immune system during the infection for helminth survival inside the host [56]. It is also suggested by regulating the host immune system; helminth species reduce the host susceptibility to allergic and autoimmune diseases [4]. A number of studies are currently underway to test the association of helminth infection with allergic diseases [57]. KEGG pathways contain disease pathways from which we note top neurodegenerative disorder as Alzheimer's disease and top endocrine and metabolic disease as Type II diabetes mellitus (Table 3) in our current ES proteins, which were also found in other helminth transcriptomic studies [13,51]. It is well studied that helminth infection is also associated with diabetes [58,59]. It was hypothesized that helmintic infections may attenuate the development of cardiovascular diseases like atherosclerosis [60]. With the properties of helminth ES proteins for host immune system modulation and involvement of helminth infections in many other disorders, these ES proteins demand further investigation for the development of novel therapeutic strategies. In our attempt to investigate predicted helminth ES proteins as drug targets, we found 27 targets using Drug Bank. Ten *O. viverrini* ES proteins were found similar to β-galactosidase which is used for the development of diagnostic tool for human helminthiasis [61]. *S. stercoralis* ES protein (Sstercoralis_UN2092) was found similar to Cathepsin F. A cathepsin F cysteine protease of *O. viverrini* (human liver fluke) has been characterized [62] and could be a potential therapeutic target as in helminth parasites as this protein is involved in excystation, tissue invasion, catabolism of host proteins for nutrition and immunoevasion [63,64]. We found heme as a potential drug molecule for helminth infection targeting fumarate reductase flavoprotein subunit. This target can be further investigated as helminths lack the heme synthesis pathway [65].

In the present study we have predicted ES proteins from helminth EST data available from dbEST followed by functional annotation of ES proteins in terms of protein domains, pathways and gene ontology and also 39 ES proteins from human parasitic helminth organisms were found similar to known drug targets but it is noteworthy to mention that only few of the targets are validated in helminth organisms. Nearly 40% of predicted ES proteins remain unannotated, which needs to be

further investigated using genomic and functional characterization studies.

Limitations of the current methodology

Integrated computational approaches, similar to those used in this paper, have been applied to other transcriptomic studies [8][13]. These approaches depend on the availability of data for a reference organism from the same taxonomic order. Annotation of the subject organism is based on sequence similarity against proteins present in non-redundant protein database from NCBI and proteins available for well helminth organisms like *C. elegans* (Wormpep), *S. mansoni* and *S. japonicum*. Availability of secretome experimental data is another limiting factor for validation of computationally predicted ES proteins. In the current study, experimentally derived ES proteins from 8 species are used to validate computational predicted ES protein data from 78 species using BLAST. Current validation percentage (22.43%) of computational predicted ES proteins can be further improved by availability of more experimental data. Another limiting factor is that we are predicting functionality based on primary sequence annotation alone, whereas protein function is actually determined by its three dimensional (3D) structure. Therefore, these preliminary predictions of therapeutic targets from this study needs to be further validated using wet-lab assays.

Conclusion

Our bioinformatics approach made possible the large scale prediction and analysis of ES proteins. As a result of our analysis we develop a unique resource HSD (Helminth Secretome Database) of ES proteins for the parasitology/infectious diseases/pharmacy communities. Our approach can be used on new large-scale transcriptomic data sets from NGS platforms, for rapid prediction and annotation of ES proteins. The approach can be applied to any organism but its main application is for neglected organisms with limited knowledge.

Additional material

Additional File 1: Summary of large scale helminth EST analysis. Statistics of excretory/secretory proteins and Unigenes across different helminth species (Table S1)

Additional File 2: Gene Ontology distribution of helminth ES proteins. Statistics of Gene Ontology distribution across different helminth species (Table S2)

Additional File 3: Helminth ES protein domain mapping. Represented Interpro domains found in helminth ES proteins. (Table S3)

Additional File 4: KEGG BRITE objects mapping of helminth ES proteins. Represented KEGG BRITE objects found in ES proteins predicted by KAAS (Table S4)

Additional File 5: Comparison of putative helminth ES proteins with *C. elegans* (Wormpep) and *S. mansoni* proteins. Statistics of sequence similarity results of helminth ES proteins with *C. elegans* (Wormpep) and *S. mansoni* proteins using BLASTP across different helminth species (Table S5)

Additional File 6: Comparison of putative helminth ES proteins with NR database proteins. Statistics of sequence similarity results of helminth ES proteins with NR database proteins using BLASTP across different helminth species (Table S6)

Additional File 7: Comparison of putative helminth ES proteins with *S. japonicum*, human proteins. Statistics of sequence similarity results of helminth ES proteins with *S. japonicum*, human proteins using BLASTP across different helminth species (Table S7)

Additional File 8: Comparison of putative helminth ES proteins with interaction databases proteins. Statistics of sequence similarity results of helminth ES proteins with interaction databases proteins using BLASTP across different helminth species (Table S8)

List of abbreviations used

BRITE: Biomolecular Relations in Information Transmission and Expression; KEGG: Kyoto Encyclopedia of Genes and Genomes; KAAS: KEGG automatic annotation server.

Acknowledgements

We would like to thank Mr. Mohammad T. Islam for helping us to set up the Helminth Secretome Database website. GG would like to acknowledge Macquarie University for the grant of Australian Postgraduate Award scholarship and Post Graduate Research Fund.
This article has been published as part of *BMC Genomics* Volume 13 Supplement 7, 2012: Eleventh International Conference on Bioinformatics (InCoB2012): Computational Biology. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/13/S7.

Author details

[1]Dept. of Chemistry and Biomolecular Sciences and ARC Centre of Excellence in Bioinformatics, Macquarie University, Sydney NSW 2109, Australia. [2]Dept. of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, Singapore 117597.

Authors' contributions

SR directed the study. GG developed the database and carried out the analysis. SR and GG contributed to writing the manuscript.

Competing interests

The authors declare that they have no competing interests.

Published: 13 December 2012

References

1. **Soil-transmitted helminths.** World Health Organization; [http://www.who.int/intestinal_worms/en/].
2. Torgerson PR: **Economic effects of echinococcosis.** *Acta Trop* 2003, **85**:113-118.
3. **Bibliography of Estimated Crop Losses in the United States Due to Plant-parasitic Nematodes.** *J Nematol* 1987, **19**:6-12.
4. Wilson MS, Maizels RM: **Regulation of allergy and autoimmunity in helminth infection.** *Clin Rev Allergy Immunol* 2004, **26**:35-50.
5. Geary TG, Thompson DP: **Caenorhabditis elegans: how good a model for veterinary parasites?** *Vet Parasitol* 2001, **101**:371-386.
6. Ranganathan S, Garg G: **Secretome: clues into pathogen infection and clinical applications.** *Genome Med* 2009, **1**:113.
7. Boguski MS, Lowe TM, Tolstoshev CM: **dbEST–database for "expressed sequence tags".** *Nat Genet* 1993, **4**:332-333.
8. Cantacessi C, Young ND, Nejsum P, Jex AR, Campbell BE, Hall RS, Thamsborg SM, Scheerlinck JP, Gasser RB: **The transcriptome of Trichuris suis–first molecular insights into a parasite with curative properties for key immune diseases of humans.** *PLoS One* 2011, **6**:e23590.
9. 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. *Biotechnol Adv* 2010, **28**:222-231.
10. 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**:1891-1907.
11. Evans H, Mello LV, Fang Y, Wit E, Thompson FJ, Viney ME, Paterson S: **Microarray analysis of gender- and parasite-specific gene transcription in Strongyloides ratti.** *Int J Parasitol* 2008, **38(11)**:1329-1341.
12. Mello LV, O'Meara H, Rigden DJ, Paterson S: **Identification of novel aspartic proteases from Strongyloides ratti and characterisation of their evolutionary relationships, stage-specific expression and molecular structure.** *BMC Genomics* 2009, **10**:611.
13. Garg G, Ranganathan S: **In silico secretome analysis approach for next generation sequencing transcriptomic data.** *BMC Genomics* 2011, **12(Suppl 3)**:S14.
14. 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**:e301.
15. Martin J, Abubucker S, Heizer E, Taylor CM, Mitreva M: **Nematode.net update 2011: addition of data sets and tools featuring next-generation sequencing data.** *Nucleic Acids Res* 2012, **40**:D720-728.
16. **Helminth Secretome Database (HSD).** [http://estexplorer.biolinfo.org/hsd].
17. 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.
18. **Seqclean.** [http://compbio.dfci.harvard.edu/tgi/software/].
19. **Univec.** [http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html].
20. 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.
21. Bendtsen JD, Nielsen H, von Heijne G, Brunak S: **Improved prediction of signal peptides: SignalP 3.0.** *J Mol Biol* 2004, **340**:783-795.
22. 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**:349-356.
23. Emanuelsson O, Nielsen H, Brunak S, von Heijne G: **Predicting subcellular localization of proteins based on their N-terminal amino acid sequence.** *J Mol Biol* 2000, **300**:1005-1016.
24. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.** *J Mol Biol* 2001, **305**:567-580.
25. Mulvenna J, Hamilton B, Nagaraj SH, Smyth D, Loukas A, Gorman JJ: **Proteomics analysis of the excretory/secretory component of the blood-feeding stage of the hookworm, Ancylostoma caninum.** *Mol Cell Proteomics* 2009, **8**:109-121.
26. 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**:e410.
27. Moreno Y, Geary TG: **Stage- and gender-specific proteomic analysis of Brugia malayi excretory-secretory products.** *PLoS Negl Trop Dis* 2008, **2**:e326.
28. Ju JW, Joo HN, Lee MR, Cho SH, Cheun HI, Kim JY, Lee YH, Lee KJ, Sohn WM, Kim DM, *et al*: **Identification of a serodiagnostic antigen, legumain, by immunoproteomic analysis of excretory-secretory products of Clonorchis sinensis adult worms.** *Proteomics* 2009, **9**:3066-3078.
29. Gourbal BE, Guillou F, Mitta G, Sibille P, Theron A, Pointier JP, Coustau C: **Excretory-secretory products of larval Fasciola hepatica investigated using a two-dimensional proteomic approach.** *Mol Biochem Parasitol* 2008, **161**:63-66.
30. Knudsen GM, Medzihradszky KF, Lim KC, Hansell E, McKerrow JH: **Proteomic analysis of Schistosoma mansoni cercarial secretions.** *Mol Cell Proteomics* 2005, **4**:1862-1875.
31. Knudsen GM, Medzihradszky KF, Lim KC, Hansell E, McKerrow JH: **Proteomic analysis of Schistosoma mansoni cercarial secretions.** *Mol Cell Proteomics* 2005, **4**:1862-1875.
32. 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.** *Mol Cell Proteomics* 2009, **8**:1236-1251.
33. Soblik H, Younis AE, Mitreva M, Renard BY, Kirchner M, Geisinger F, Steen H, Brattig NW: **Life cycle stage-resolved proteomic analysis of the excretome/secretome from Strongyloides ratti–identification of stage-specific proteases.** *Mol Cell Proteomics* 2011, **10, M111 010157.**
34. 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**:535-543.
35. 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**:10-19.
36. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403-4.
37. Zdobnov EM, Apweiler R: **InterProScan–an integration platform for the signature-recognition methods in InterPro.** *Bioinformatics* 2001, **17**:847-848.
38. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M: **KAAS: an automatic genome annotation and pathway reconstruction server.** *Nucleic Acids Res* 2007, **35**:W182-185.
39. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: **KEGG for representation and analysis of molecular networks involving diseases and drugs.** *Nucleic Acids Res* 2010, **38**:D355-360.
40. 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**:D354-357.
41. **Wormpep database.** [http://www.sanger.ac.uk/Projects/C_elegans/WORMBASE/current/wormpep.shtml].
42. **MySQL 5 relational database.** [http://www.mysql.com/].
43. **PHP.** [http://www.php.net/].
44. **Apache webserver.** [http://www.apache.org/].
45. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, *et al*: **IntAct–open source resource for molecular interaction data.** *Nucleic Acids Res* 2007, **35**:D561-565.
46. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, Nixon J, Van Auken K, Wang X, Shi X, *et al*: **The BioGRID Interaction Database: 2011 update.** *Nucleic Acids Res* 2011, **39**:D698-704.
47. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D: **DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions.** *Nucleic Acids Res* 2002, **30**:303-305.
48. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, *et al*: **DrugBank 3.0: a comprehensive resource for 'omics' research on drugs.** *Nucleic Acids Res* 2011, **39**:D1035-1041.
49. McVeigh P, Maule AG, Dalton JP, Robinson MW: **Fasciola hepatica virulence-associated cysteine peptidases: a systems biology perspective.** *Microbes Infect* 2012.
50. 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.
51. Young ND, Jex AR, Cantacessi C, Hall RS, Campbell BE, Spithill TW, Tangkawattana S, Tangkawattana P, Laha T, Gasser RB: **A portrait of the transcriptome of the neglected trematode, Fasciola gigantica–biological and biotechnological implications.** *PLoS Negl Trop Dis* 2011, **5**:e1004.
52. Fried B, Reddy A, Mayer D: **Helminths in human carcinogenesis.** *Cancer Lett* 2011, **305**:239-249.
53. 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 Negl Trop Dis* 2010, **4**:e719.
54. Young ND, Jex AR, Li B, Liu S, Yang L, Xiong Z, Li Y, Cantacessi C, Hall RS, Xu X, *et al*: **Whole-genome sequence of Schistosoma haematobium.** *Nat Genet* 2012, **44**:221-225.
55. Strachan DP: **Hay fever, hygiene, and household size.** *BMJ* 1989, **299**:1259-1260.
56. 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.
57. Flohr C, Quinnell RJ, Britton J: **Do helminth parasites protect against atopy and allergic disease?** *Clin Exp Allergy* 2009, **39**:20-32.

58. Liu Q, Sundar K, Mishra PK, Mousavi G, Liu Z, Gaydo A, Alem F, Lagunoff D, Bleich D, Gause WC: **Helminth infection can reduce insulitis and type 1 diabetes through CD25- and IL-10-independent mechanisms.** *Infect Immun* 2009, **77**:5347-5358.
59. Saunders KA, Raine T, Cooke A, Lawrence CE: **Inhibition of autoimmune type 1 diabetes by gastrointestinal helminth infection.** *Infect Immun* 2007, **75**:397-407.
60. Magen E, Borkow G, Bentwich Z, Mishal J, Scharf S: **Can worms defend our hearts? Chronic helminthic infections may attenuate the development of cardiovascular diseases.** *Med Hypotheses* 2005, **64**:904-909.
61. Sugane K, Sun SH: **Detection of anti-helminth antibody by microenzyme-linked immunosorbent assay using recombinant antigen and anti-beta-galactosidase monoclonal antibody.** *J Immunol Methods* 1994, **168**:55-60.
62. Pinlaor P, Kaewpitoon N, Laha T, Sripa B, Kaewkes S, Morales ME, Mann VH, Parriott SK, Suttiprapa S, Robinson MW, *et al*: **Cathepsin F cysteine protease of the human liver fluke, Opisthorchis viverrini.** *PLoS Negl Trop Dis* 2009, **3**:e398.
63. 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**:35950-35957.
64. Perrigoue JG, Marshall FA, Artis D: **On the hunt for helminths: innate immune cells in the recognition and response to helminth parasites.** *Cell Microbiol* 2008, **10**:1757-17.
65. Rao AU, Carta LK, Lesuisse E, Hamza I: **Lack of heme synthesis in a free-living eukaryote.** *Proc Natl Acad Sci USA* 2005, **102**:4270-4275.

doi:10.1186/1471-2164-13-S7-S8
Cite this article as: Garg and Ranganathan: **Helminth secretome database (HSD): a collection of helminth excretory/secretory proteins predicted from expressed sequence tags (ESTs).** *BMC Genomics* 2012 **13** (Suppl 7):S8.

5.2 Conclusions

Here, we analysed the largest EST dataset of helminths covering the maximum number of species comprehensively, for the prediction and annotation of classical and non-classical ES proteins. The predicted ES proteins from this large-scale transcriptomic analysis of helminths are available to scientific community as a searchable database. Using our computational approach we were able to predict 18992 ES proteins, out of which 11390 ES proteins were annotated using database similarality searches. 4260 (22.4%) of annotated ES proteins were verified with experimentally determined helminth ES proteins. Furthermore, 39 ES proteins from human parasitic helminth organisms were found to match with 27 known drug targets, targeted by 14 clinically approved drugs. These potential drug targets are also available from HSD.

The study could identify number of pathways associated to diseases like Alzheimer's disease, type II diabetes mellitus and cancer. The results of this large-scale analysis provide a step towards future research with the focus on disease and molecular biology of helminths and could also be integrated with proteomic and metabolomic studies for identifying novel disease control strategies. The robust secretome analysis pipeline for this large scale analysis of EST data has been applied to novel parasitic nematode transcriptome datasets (*Strongloides stercoralis*, Chapter 6 and *Echinostoma caproni*, Chapter 7). With minor modifications to the specific datasets selected for homology matching, this pipeline can be adapted to other pathogenic organisms as well as described in Chapter 8.

Chapter 6: Transcriptome analysis of *Strongyloides stercoralis* L3i larvae identifies targets for intervention in a neglected disease

6.1 Summary

Strongyloides stercoralis is an important nematode parasite of humans, considered as the causative agent of strongyloidiasis, affecting more than 100 million people worldwide. Even though there is serious impact of this disease, very little is known about this parasite and its relationship with its hosts at the molecular level. With the advent of NGS technologies, these neglected species can be rapidly studied at molecular level.

Here, we applied our bioinformatics approach (Chapter 3) with minor modifications to analyse the first transcriptome of the third larval stage of *S. stercoralis*. We concentrated our attention on the systematic annotation of all the proteins because of the novelty of data, instead of only ES proteins as discussed in Chapters 3 and 4.

The results of the analysis are presented in Publication 5, with Additional Files available on CD.

OPEN ACCESS Freely available online

PLoS NEGLECTED TROPICAL DISEASES

The Transcriptome Analysis of *Strongyloides stercoralis* L3i Larvae Reveals Targets for Intervention in a Neglected Disease

Antonio Marcilla[1]*, Gagan Garg[2], Dolores Bernal[3], Shoba Ranganathan[2,4]*, Javier Forment[5], Javier Ortiz[6], Carla Muñoz-Antolí[1], M. Victoria Dominguez[7], Laia Pedrola[8], Juan Martinez-Blanch[8], Javier Sotillo[1], Maria Trelis[1], Rafael Toledo[1], J. Guillermo Esteban[1]

1 Área de Parasitología, Departamento de Biologia Celular y Parasitología, Universitat de València, Burjassot, Valencia, Spain, **2** Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, New South Wales, Australia, **3** Departamento de Bioquímica y Biología Molecular, Universitat de València, Burjassot, Valencia, Spain, **4** Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore, **5** Servicio de Bioinformática, Instituto de Biologia Molecular y Celular de Plantas, Universitat Politécnica de València, Ingeniero Fausto Elio, Valencia, Spain, **6** Unidad de Bioinformática, SCSIE, Universitat de València, Burjassot, Valencia, Spain, **7** Servicio de Microbiología y Parasitología, Hospital General de Castellón, Castellón, Spain, **8** LifeSequencing S.L., Parc Científic Universitat de València, Paterna, Valencia, Spain

Abstract

Background: Strongyloidiasis is one of the most neglected diseases distributed worldwide with endemic areas in developed countries, where chronic infections are life threatening. Despite its impact, very little is known about the molecular biology of the parasite involved and its interplay with its hosts. Next generation sequencing technologies now provide unique opportunities to rapidly address these questions.

Principal Findings: Here we present the first transcriptome of the third larval stage of *S. stercoralis* using 454 sequencing coupled with semi-automated bioinformatic analyses. 253,266 raw sequence reads were assembled into 11,250 contiguous sequences, most of which were novel. 8037 putative proteins were characterized based on homology, gene ontology and/or biochemical pathways. Comparison of the transcriptome of *S. strongyloides* with those of other nematodes, including *S. ratti*, revealed similarities in transcription of molecules inferred to have key roles in parasite-host interactions. Enzymatic proteins, like kinases and proteases, were abundant. 1213 putative excretory/secretory proteins were compiled using a new pipeline which included non-classical secretory proteins. Potential drug targets were also identified.

Conclusions: Overall, the present dataset should provide a solid foundation for future fundamental genomic, proteomic and metabolomic explorations of *S. stercoralis*, as well as a basis for applied outcomes, such as the development of novel methods of intervention against this neglected parasite.

Citation: Marcilla A, Garg G, Bernal D, Ranganathan S, Forment J, et al. (2012) The Transcriptome Analysis of *Strongyloides stercoralis* L3i Larvae Reveals Targets for Intervention in a Neglected Disease. PLoS Negl Trop Dis 6(2): e1513. doi:10.1371/journal.pntd.0001513

Editor: John Pius Dalton, McGill University, Canada

Received August 23, 2011; **Accepted** December 20, 2011; **Published** February 28, 2012

Funding: This work was supported by project PS09/02355 from the Fondo de Investigación Sanitaria (FIS), Spanish Ministry of Science and Innovation (Madrid, Spain) and FEDER and project PROMETEO/2009/081 from Conselleria d'Educació, Generalitat Valenciana (Valencia, Spain). GG would like to acknowledge Macquarie University for an Australian Post-graduate Award scholarship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: antonio.marcilla@uv.es (AM); shoba.ranganathan@mq.edu.au (SR)

Introduction

Strongyloidiasis caused by *Strongyloides stercoralis* is a soil-transmitted helminthiasis distributed worldwide, affecting more than 100 million people, with endemic areas in Southeast Asia, Latin America, sub-Saharan Africa, and parts of the southeastern United States [1,2]. Recently, it was classified as one of the most neglected tropical diseases (NTD) [3]. Chronic infections in endemic areas may be maintained asymptomatically for decades through the autoinfective cycle with the filariform larvae L3 [1][4,5]. The diagnosis of these chronic infections requires more sensitive diagnostic methods, particularly in low-level infections and immunocompromised patients [1].

Epidemiological studies in developed countries have identified endemic areas where misdiagnosis, inadequate treatment and the facilitation of hyperinfection syndrome by immunosupression (i.e. by the administration of steroids) are too frequent and can cause a high mortality rate ranging from 15 to 87% [5,6]. Among these areas, an endemic area with chronic patients have been described at the Valencian Mediterranean coastal region of Spain related to environmental conditions [7].

The diagnosis of strongyloidiasis is suspected when clinical signs and symptoms, or eosinophilia is observed [8], but current definitive diagnosis of strongyloidiasis is usually made on the basis of detection of larvae in agar plate coproculture and serological diagnosis by ELISA [9,10]. Those methods have the drawbacks of

Author Summary

Strongyloides stercoralis (Nematoda) is an important parasite of humans, causing Strongyloidiasis, considered as one of the most neglected diseases, affecting more than 100 million people worldwide. Chronic infections in endemic areas can be maintained for decades through the autoinfective cycle with the L3 filariform larvae. In these areas, misdiagnosis, inadequate treatment and the facilitation of hyperinfection syndrome by immunosupression are frequent and contribute to a high mortality rate. Among the affected areas, chronic patients have been described in the Valencian Mediterranean coastal region of Spain. Despite its serious impact, very little is known about this parasite and its relationship with its hosts at the molecular level, and more effective diagnostic tests and treatments are needed. Next generation sequencing technologies now provide unique opportunities to rapidly advance in these areas. In this study, we present the first transcriptome of *S. stercoralis* L3i using 454 sequencing followed by semi-automated bioinformatic analyses. Our study identifies 8037 putative proteins based on homology, gene ontology, and/or biochemical pathways, including putative excretory/secretory proteins as well as potential drug targets. The present dataset provides a useful resource and adds greatly to our understanding of a human parasite affecting both developed and developing countries.

being time consuming and requiring expertise in the first case, and of low specificity due to remaining antibodies from previous infection or cross-reactive antibodies [11]. A recent paper has described a promising coproantigen ELISA based on a polyclonal rabbit antiserum raised against excretory/secretory (ES) antigens from the closely relative *Strongyloides ratti* [12], but the identification of *S. stercoralis* specific ES proteins that could be new potential targets for diagnosis is still required.

Control of strongyloidiasis has relied mostly on the treatment of infected individuals with only three anthelmintic drugs: thiabendazole (no longer available), albendazole, and more recently ivermectin [3,13]. A recent study by Suputtamongkol *et al.* (2011) has confirmed that both a single and double dose of oral ivermectin are more effective than a 7-day course of high dose albendazole for patients with chronic infection due to *S. stercoralis* [14]. The risk of developing genetic resistance against the current drugs administered (if used excessively and at suboptimal dosages) exists and is based on the experience with drug resistance in parasitic nematodes of livestock [15]. Thus, the current focus is on the discovery of novel drugs against human parasites like *S. stercoralis*. Such a discovery effort could be strengthened with an integrated genomic and bioinformatics approach, using functional genomic and phenomic information available for the free-living nematode *Caenorhabditis elegans* (see WormBase; www.wormbase.org). This nematode, which is the best characterized metazoan organism [16,17], is considered to be related to nematodes of the order Strongylida (to which *Strongyloides* belong) [18]. Recent studies have reported that nearly 60% of genes in strongyloides have orthologues/homologues in *C. elegans*, with a wide range of biological pathways being conserved between parasitic nematodes and *C. elegans* [19]. The comparison of molecular data sets between nematodes should therefore allow the identification of specific biological pathways as potential new targets for nematocidal drugs [20].

As pointed out recently by Cantacessi *et al.* (2011) [20], advances in genomic sequencing like Next Generation Sequencing (NGS) and annotation as well as the integrated use of '-omic' technologies are now shedding light on our understanding of the systems biology of nematodes on an unprecedented scale, and is likely to provide unique opportunities for the development of entirely new strategies for the treatment and control of neglected parasitic diseases. New bioinformatic tools based on robust assembly protocol for NGS data, along with compilation of a dataset of experimentally determined ES proteins of parasitic helminths, and annotation software like KAAS [21], allow efficient and up-to-date homology-based predictions [22].

To date, there are few molecular and genomic studies on *Strongyloides* species, and only the transcriptome from *S. ratti* adults has become recently available (http://worm1.liv.ac.uk/file_summary.html) [23]. In fact, 39166 ESTs are currently available in the NCBI database of November 2011 (27366 from *S. ratti* and 11392 from *S. stercoralis*). Yoshida *et al.* (2011) have obtained 162 unique singletons and contigs from *S. venezuelensis* [24], and a recent study by Ramanathan *et al.* (2011) has described DNA microarray for *S. stercoralis* and used them to compare infective third-stage larvae (L3i) with non-infective first stage larvae (L1), with 935 differentially expressed genes identified [25].

In the present study, we have explored and functional annotated the transcriptome of L3i of *S. stercoralis* by 454 sequencing coupled to semi-automated bioinformatic analyses and predicted potential therapeutic targets for strongyloidiasis.

Materials and Methods

Accession numbers

The nucleotide sequence data obtained for this study are available in the GenBank database under accession number ERP000798.

The assembled data from this study can be requested from the corresponding author.

Parasite material and ethical issues

Fecal samples were obtained at the Hospital La Ribera, Alzira, Valencia (Spain) from an infected individual in compliance with Spanish ethical regulations [7], and approved by the Ethics Committee in human research from the Universitat de Valencia. Oral consent from the patient was obtained (she was happy to participate in the study but felt uncomfortable with signing a form), and documented as a tick on the case record form following the Hospital Reviewing Board protocols. Samples were cultured on Agar Petri dishes and L3i larvae were harvested and concentrated by centrifugation for 5 min at 1000 g, washed three times in 1 ml of phosphate buffered saline (PBS) pH 7.2 containing protease inhibitors (10 mM EDTA, and 1 mM PMSF) and samples were processed for RNA isolation.

RNA isolation, cDNA synthesis and 454 sequencing

Total RNA from around 500 larvae was prepared using Vantage^{TM} Total RNA purification kit (Marligen Biosciences, Ijamsville, MD, USA) following the manufacturers' instructions and treated with Ambion DNA-free^{TM} DNase (Ambion/Applied Biosystems, Austin, TX). The integrity of the RNA was verified by gel electrophoresis and the yield determined using the nanoDrop ND-1000 UV-VIS spectrophotometer v.3.2.1 (NanoDrop Technologies, Wilmington, DE).

The cDNA library was constructed from 0.5 µg total RNA using MINT cDNA Synthesis Kit (Cat#SK001, Evrogen). First strand cDNA synthesis starts from 3′-primer comprising oligo(dT) to enrich mRNA as template. Double strand cDNA synthesis was performed using 17 cycles of PCR amplification. Total cDNA was

digested with restriction enzyme *GsuI* in order to remove Poly (A) tails. cDNA obtained was used to perform a library with the required sequencing adaptors and was then sequenced using the Genome Sequencer (GS) FLX instrument (Roche Diagnostics) [26].

Bioinformatic analyses of sequence data

The overall bioinformatics analysis strategy followed was as described originally by Nagaraj *et al.* [27,28], implemented in the analysis pipelines ESTExplorer [29] and EST2Secretome [28]. This workflow approach has been successfully used for the analysis of transcriptomic data from *Dictyocaulus viviparus* [30], *Fasciola hepatica* [31], *Clonorchis sinensis* [32] and *Opisthorchis viverrini* [32]. However, to better identify non-classically secreted proteins from helminth parasites [33,34], we have recently implemented a novel analysis strategy for short reads applied on *Strongyloides ratti* [22] (see Figure S1).

FASTA and associated quality files were extracted from the SFF file after removing the sequence adapters. These reads were preprocessed and their contigs were assembled using MIRA v.3.2 (http://chevreux.org/projects_mira.html) [35] with the following parameters:

-job=denovo,est,accurate,454 -fasta

-OUT:rrol = 1:rld = 1:orc = 1:org = 0:ora = 0:ors = 0:otf = 0:otc = 0 -GE:not = 1 -CO:asir = 1

-LR:mxti = 1 -AS:sd = 0:uess = 0:urd = 0:ard = 1 -SK:mmhr = 2:mnr = yes 454_SETTINGS

-DP:ure = 0 -CO:mrpg = 10 -AS:bdq = 40

-CL:pvlc = 0:cpat = 0:mbc = 1:mbcgs = 30:mbcmfg = 30:mbcmeg = 30:qc = 0 -ED:ace = 1

-AL:egp = no

-ALIGN:bip = 20:bmax = 120:mo = 10.

Contigs generated from MIRA were aligned and reassembled into second order contigs using the Contig Assembly Program v.3 (CAP3) [36], employing a minimum sequence overlap length cut-off of 40 bases and an identity threshold of 90%. Following the assembly of *S. stercoralis* reads into second order contigs by CAP3 and contigs by MIRA, this contig dataset was matched using BLASTX with the NCBI non-redundant sequence database; http://www.ncbi.nlm.nih.gov, BLASTN with Nematode.net *S. stercoralis* ESTs (www.nematode.net/) and BLASTN with dbEST *Strongyloides* ESTs (www.ncbi.nlm.nih.gov/dbEST/), using permissive (E-value: $<1E^{-05}$), moderate ($<1E^{-15}$) and/or stringent ($<1E^{-30}$) search strategies.

S. stercoralis contigs were conceptually translated into putative proteins using the program ESTScan [37]. Putative protein sequences were subjected to secretome analysis using TMHMM (a membrane topology prediction program) [38] to predict trans-membrane domains, SignalP 3.0 (signal peptide prediction program) [39], SecretomeP (a prediction programme used to identify non-classical secretory proteins in mammals [40], but used in the case of parasitic helminths as well [41]), and TargetP (mitochondrial protein prediction program) [42]. Briefly, excretory/secretory (ES) proteins were selected based on the presence of a signal peptide at the N-terminus using SignalP 3.0 (employing both the neural network and hidden Markov models) or predicted as secretory using SecretomeP, predicted as non-mitochondrial by TargetP and absence of transmembrane domains. In addition to computational prediction of ES proteins were identified and collated based on sequence homology (BLASTP, E-value$<1E^{-15}$) to known ES proteins found in parasitic helminths secretome studies.

Putative proteins were classified functionally using InterProScan [43], employing the default search parameters. Based on their homology to conserved domains and protein families, predicted proteins were classified into Gene Ontology (GO) categories (http://www.geneontology.org/) based on molecular function, cellular component and biological process using interpro terms. Putative proteins were also subjected to pathway analysis, utilizing KEGG-Automatic Annotation Server (KAAS) [21], which maps the putative proteins to biochemical pathways in which they are involved and categories of Brite objects like enzymes, transcription factors and translation factors.

Putative proteins were subjected to BLAST2GO software to identify homologues from the most abundant ES transcripts [44]. BLASTP (Wormpep v 224) was used to identify *C. elegans* known proteins homologues present in *S. stercoralis* proteins using moderate search strategy (E-value: $<1E^{-15}$). These proteins were also searched for sequence homology (BLASTP, E-value$<1E^{-05}$) in human (host) proteins. All the proteins which were found homologous to *C. elegans* proteins and non-homologous to human proteins were mapped to *C. elegans* RNAi phenotypes and known drug targets present in the DrugBank database (http://drugbank.ca/), a unique bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information [45].

Results

The transcriptome of *S. stercoralis* L3i larvae

Initially a total of 253266 short reads (82490223 bases) were generated with 325±132.4 bases (average length ± standard deviation), with a GC content of 31.84%. These short reads were pre-processed, which resulted in 237341 (93.7%) quality short reads (EBI Sequence Read Archive [SRA] accession ID ERP000798). High quality reads were assembled into 12333 contigs using MIRA as described in the pipeline (Figure S1). Using CAP3, we were able to achieve 507 second order contigs, leaving 10845 MIRA contigs not assembled further by CAP3. We considered 11250 (99.1%) contigs with a minimum length of 90 bases, discarding sequences yielding peptides <30 amino acids, for further secretory protein prediction and analysis. These contigs were conceptually translated into 8037 proteins by ESTScan (Table 1; sequences available from http://biolinux.uv.es/marcilla/).

Putative proteins were annotated based on protein families and domains using Interproscan and mapped to biochemical pathways using KAAS [21]. Of the 8037 putative proteins, we were able to annotate 4494 (55.91%) proteins with protein domains and

Table 1. Expressed sequence tag (EST) data for the *S. stercoralis* L3i.

No. of EST Clusters by MIRA	12,333
Average length (±standard deviation)	538.7 (±260.7)
Re clustered unigenes after CAP3 assembly	11,352
Containing an Open Reading Frame	8,037
Returning InterProScan results	4,494
Gene Ontology	3,534
Biological process	1,905
Cellular component	1,068
Molecular function	3,083
Prediction of biological pathways (KAAS)	1,559

doi:10.1371/journal.pntd.0001513.t001

families (Table 1). The most represented Interpro terms are shown in Table S1.

A total of 3534 proteins were annotated with GO terms (3083 {Molecular Function}, 1068 {Cellular Component} and 1905 {Biological Process}) based on Interpro term annotations (Tables 1 and S1). We established pathway associations for 1559 (19.39%) putative proteins (Table 1).

All the contigs generated by using MIRA+CAP3 were checked for homologous proteins against the non-redundant nucleotide database (NR-NCBI), existing *Strongyloides* expressed sequence tags (ESTs) present in dbEST, *S. stercoralis* ESTs available from dbEST and nematode.net, *S. ratti* cDNA sequencing data from the University of Liverpool (available at http://worm1.liv.ac.uk/file_summary.html), and also for homologous proteins in *C. elegans* and human data (Figure S1). Similarity searches were done using using BlastX and BlastP algorithms at different E values (Table 2). A total of 3412 (42.45%) *S. stercoralis* putative proteins had homologues in the free-living nematode, *Caenorhabditis elegans* using stringent match conditions (E value: $<1E^{-15}$). The recent availability of *S. ratti* transcriptome data prompted us to compare these with our data and 3855 similar putative proteins (47.96%) were found. As *S. stercoralis* infects humans, we checked the similarity of *S. stercoralis* proteins with known human proteins using BlastP at different E values. Our results showed that 3759 putative proteins were similar to human ones using a permissive search strategy (E-value: $<1E^{-05}$), discarding them as potential targets for treatment (Table 2).

Predicted proteins were also categorized according to their inferred molecular function, cellular localization and association with biological pathways. Mapping to KEGG BRITE objects [46] is shown in Table 3. Enzymes were by far the most abundant category, with 720 putative proteins, followed by chromosome, spliceosome and ribosome components (with 90, 89 and 73 putative proteins, respectively). 73 putative protein kinases and 72 peptidases were also identified by BRITE (Table 3). These 72 peptidases corresponded to 60 different enzymes from 9 groups, including calpains, cathepsins, different proteasome components and aminopeptidases, and other "nematode common" proteases such as astacin, legumain, and insulysin (Table S2).

All the putative proteins were grouped according to KEGG pathways [46] into five categories, with metabolic proteins being the most abundant, followed by genetic information processing, environmental processing and cellular processes (Table 4). In the first group, the most abundant putative proteins were related to carbohydrate metabolism (201 proteins, 2.5%), amino acid metabolism (174; 2.16%) and lipid metabolism (104; 1.29%). Also 23 putative proteins were related to drug metabolism (Table 4). In the second group, the most abundant proteins were related to translation (195; 2.42%), meanwhile 144 putative proteins (1.79%) related to signal transduction were the most abundant in the group of cellular processes (Table 4).

Prediction of ES proteins

We next analyzed ES proteins, which are key molecules to understand host-parasite interactions [47]. Molecules from the secretome contribute to important processes like parasite feeding, tissue penetration or larval migration, and could participate in blocking and/or evading host immune responses [48]. ES prediction was carried out in Phase III of the pipeline (Fig. S1). Firstly, 247 (3.07%) proteins were predicted as classical secreted proteins using SignalP [39]. The remaining 7785 (96.86%) proteins, which were predicted as non-secretory by SignalP were processed by SecretomeP [40] for prediction of non-classical secretory proteins, with 252 (3.14%) proteins identified here. The classical and non-classical secretory proteins (499, 6.21%) from these two programs were analyzed by TargetP [42] for mitochondrial proteins. Only 7 proteins were predicted as mitochondrial proteins using TargetP at 95% specificity. These seven proteins were removed from the set of 499 secreted proteins, with 492 secretory proteins passed to TMHMM [38] for the prediction of transmembrane proteins. 161 (2%) proteins, were predicted as transmembrane proteins having one or more transmembrane helices, and removed from the secretory protein dataset. A total of 331 (4.12%) proteins were finally predicted as ES proteins from the computational prediction pipeline. Proteins that were considered non-secretory by SecretomeP and SignalP were matched to our in-house dataset of 1080 non redundant experimentally determined parasitic helminth proteins [22] using the BLASTP [49] similarity search. We found an additional 882 (10.97%) putative proteins similar to known ES proteins by this homology search approach (E value: $<1E^{-15}$) (Table S3). From those proteins, 50 have been recently described in the ES from infective larvae of the related species *S. ratti* [50] (data not shown).

Among the most abundant transcripts encoding ES proteins appeared a major antigen; cytoskeletal proteins like myosin heavy chain, troponin, tropomyosin, actin; galectins; enzymes like trehalase, PEPCK, GAPDH, enolase, as well as phosphatases and kinases; proteases like Metalloproteinase, Calpain-1and Cathepsin L; stress proteins like HSPs; calcium binding proteins; detoxifying enzymes along with elongation factors, histones, ubiquitins and signaling molecules (Table S4. Thus, for annotation and analyses in Phase III, we compiled a total of 1213 ES proteins, which is 15.09% of our putative proteins.

S. stercoralis proteins as drug targets

We found 4234 (52.68%) *S. stercoralis* putative proteins which had no homologues present in humans (Table 2) and therefore are

Table 2. Sequence homology inferred between *S. stercoralis* current dataset and other datasets.

Dataset	Hits (1e-05)	Hits (1e-15)	Hits (1e-30)
Strongyloides EST from dbEST (BLASTN)	4267	3935	3646
S. stercoralis EST from dbEST (BLASTN)	2903	2682	2519
S. stercoralis EST from nematode.net (BLASTN)	2225	2020	1904
S. ratti putative proteins (BLASTP)	3855	3052	2081
C. elegans proteins (BLASTP)	4475	3,412	2163
Human proteins (BLASTP)	3759	2583	1489
NR database (BLASTX)	7633	5948	4023

doi:10.1371/journal.pntd.0001513.t002

Table 3. Functions of putative proteins inferred from the transcriptome of the *S. stercoralis* L3i.

BRITE object name	Number of putative proteins mapped
Enzymes	720
Chromosome	90
Spliceosome	89
Ribosome biogenesis	73
Protein kinases	73
Peptidases	72
Ubiquitin system	72
Ribosome	72
Chaperones and folding catalysts	61
Cytoskeleton proteins	53
Proteasome	41
Ion Channels	38
Transcription factors	36
Translation factors	34
DNA repair and recombination proteins	32
GTP-binding proteins	25
DNA replication proteins	20
Glycosyltransferases	15
Lipid biosynthesis proteins	14
Cellular antigens	11
Transporters	11
Secretion system proteins	10
Glycan binding proteins	9
SNAREs	7
Nuclear receptors	6
Prenyltransferases	6
G protein coupled receptors	4
CAM ligands	4
Proteoglycans	4
Enzyme linked receptors	2
Cytokine receptors	2
Cell adhesion molecules (CAMs)	2
Bacterial toxins	1

doi:10.1371/journal.pntd.0001513.t003

preferred targets for parasite intervention strategies. These human dissimilar proteins of *S. stercoralis* were checked for known drug targets, which have lethal RNAi phenotypes present in *C. elegans*, not present in human and similar to known drug targets, data available from DrugBank 3.0 database [44], a unique bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information. The database (available at http://drugbank.ca/) contains 6707 drug entries (as of November 2011).

We found 14 contigs and singletons corresponding to four different proteins. These could represent potential therapeutic targets for strongyloidiasis as shown in Table 5. Sequence comparison demonstrate that these proteins are homologous to 2,3-bisphosphoglycerate independent phosphoglycerate mutase from *Ascaris suum* (with 1 contig and 1 singleton), hypothetical protein CBG01975 from *Caenorhabditis briggsae* similar to glutamate synthase [NADPH] from *Ascaris suum* (1 contig and 3 singletons), isocitrate lyase from *S. stercoralis* (5 singletons), and alcohol dehydrogenase I from the fungus *Candida albicans* WO-1 (2 singletons) (Table 5).

With a comparative analysis searching protein domain mapping or sequence similarity with other drug targets, we found seven additional potential targets for treatment, including well known drug targets as tubulin β, γ-amino butyric acid A (GABA) receptor, glutamate-gated chloride channel or GST (Table S5). Only one of those proteins, homologous to *Ancylostoma caninum* metalloprotease precursor, which is also predicted to be secretory, was not found similar either to *C. elegans* or human proteins (Table S5).

Discussion

Strongyloides stercoralis can replicate within the host (autoinfection) allowing the infection to remain undiagnosed and untreated for years, resulting in perpetuating parasite dispersal, increasing the risk of infection and eventually the appearance of resistances [3]. Uncontrolled multiplication of the parasite (hyperinfection) can be life-threatening in immunocompromised individuals. We also face serious endemic recurring infections in the future if this infection is not controlled in transition economies like China, India, Southeast Asia and Latin America [3] where the use of immunosuppressive therapy is becoming common. As pointed out by Olsen *et al.* (2009) [3], there is an urgent need to employ modern molecular methods to improve and simplify diagnosis, differentiate species and strains to facilitate epidemiological studies of *S. stercoralis*.

The present study provides the first detailed analysis of the transcriptome of the human pathogenic *S. stercoralis* L3i larvae and has identified specific molecules predicted to play key biological functions in this parasite. A total of 12,333 contigs were inferred from the present EST dataset, thus increasing the number of predicted proteins currently available (for this stage/species) in public databases by approximately 141-fold [we obtained 8037 conceptually translated proteins, and there are currently 57 proteins in Genbank as of November 2011]. This quantity of contigs is similar to the numbers obtained with other nematodes like *Trichostrongylus colubriformis* [19], *N. americanus* and *Ancylostoma caninum* [51], *Haemonchus contortus* [52], *Dictyocaulus viviparous* [20], and *Teladorsagia circumcinta* [53]. The subset (55.91%) of *S. stercoralis* sequences with orthologues/homologues in public databases was slightly higher to that reported in similar transcriptomic studies of other animal-parasitic helminths such as *Necator americanus* [51,54]. It is noteworthy to mention that 44.09% of the putative proteins of *S. stercoralis* L3i transcriptome remain unannotated, warranting further genomic and functional characterization studies.

With the exception of three metabolic proteins (citrate synthase, arginine kinase and ATP:guanido phosphotransferase) all proteins identified in a previous proteomic study with *S. stercoralis* L3i [55] were included in the transcriptome described here. In addition, 41 antigenic proteins including SiR and tropomyosin were present in the transcriptome (as searched in Table S1), confirming its value as a tool for searching targets for immunodiagnosis.

It is well characterized that upon infection, infective larvae (L3i) must penetrate skin as quickly as possible and then migrate within the host. In this context, proteases play an essential role. Among the proteins identified in our study, 60 different putative proteases were annotated in nine groups. These include nine metalloproteinases and three aspartic proteases, some of them assumed to play a major role in skin penetration in *Strongyloides stercoralis* [56,57], and in other *Strongyloides* species like *S. venezuelensis* [58] or

Table 4. KEGG pathways of putative proteins inferred from the transcriptome of *S. stercoralis* L3i.

Parent KEGG pathway	No of putative proteins (%)	Top KEGG pathway in the category
Metabolism		
Carbohydrate metabolism	201 (2.50)	Glycolysis/Gluconeogenesis
Energy metabolism	79 (0.98)	Oxidative phosphorylation
Lipid metabolism	104 (1.29)	Fatty acid metabolism
Nucleotide metabolism	100 (1.24)	Purine metabolism
Amino acid metabolism	174 (2.16)	Valine, leucine and isoleucine degradation
Metabolism of other amino acids	48 (0.60)	Glutathione metabolism
Glycan biosynthesis and metabolism	30 (0.37)	N-Glycan biosynthesis
Metabolism of Cofactors and Vitamins	46 (0.57)	One carbon pool by folate
Metabolism of Terpenoids and Polyketides	5	Terpenoid backbone biosynthesis
Biosynthesis of other secondary metabolites	1	Caffeine metabolism
Xenobiotic biodegradation and metabolism	23 (0.29)	Drug metabolism – other enzymes
Genetic Information processing		
Transcription	83 (1.03)	Spliceosome
Translation	195 (2.42)	Ribosome
Folding, sorting and degradation	178 (2.21)	Protein processing in endoplasmic reticulum
Replication and repair	34 (0.42)	Nucleotide excision repair
Environmental information processing		
Membrane transport	3	ABC transporters
Signal transduction	144 (1.79)	MAPK signaling pathway
Signaling, molecules and interaction	7	Neuroactive ligand-receptor interaction
Cellular processes		
Transport and catabolism	132 (1.64)	Endocytosis
Organism systems		
Immune system	7	Natural killer cell mediated cytotoxicity
Endocrine system	11 (0.13)	Progesterone-mediated oocyte maturation
Development	3	Dorso-ventral axis formation
Environmental adaptation	6	Circadian rhythm – mammal

doi:10.1371/journal.pntd.0001513.t004

S. ratti [22,23]. In *S. venezuelensis*, Yoshida *et al.* (2011) [24] have recently identified an astacin-like metalloproteinase as being specific of L3i in a transcriptomic study. Another abundant group was the cysteine proteases, including cathepsin B, legumain and calpain, proteins characterized as immunomodulators of host response and promising vaccine and drug targets [59–61]. Similar results have been reported for *Ascaris suum*, where 456 peptidases have been identified in its draft genome [62].

Kinases are also an important group of proteins considered to be good druggable targets from the medical and chemical viewpoints, since they play essential functions in the parasite, in mediating signal transduction [63–65]. In *S. stercoralis* L3i transcriptome analysis 73 putative kinases including 11 putative tyrosine kinases were identified (Table 3 and Table S1).

In our study, we have compiled 1213 putative ES proteins among the 8037 (15.09%) *S. stercoralis* annotated proteins using a new semi-automated computational approach, recently developed and applied to predict the secretome of *S. ratti* adults [22]. In a mixture of *S. ratti* parasitic females, free-living males and free-living females, Garg and Ranganathan (2011) compiled 2572 putative ES proteins, being 12.3% of the total putative proteins, which is less than that found in *S. stercoralis* L3i larvae [22]. This could be due to higher secretion processes in larvae in comparison to adults, required by penetration and migration in the host. Supporting this notion, Soblik et al. (in press) have recently described the presence of 586 ES proteins in all the stages of *S. ratti* by proteomic analysis, 196 of which are also found in L3i [50]. When comparing larval *S. ratti* ES proteins with our predicted *S. stercoralis* L3i ES proteins, we find that 50 out of the 196 proteins identified from *S. ratti* were also detected in *S. stercoralis* L3i, supporting the value of the prediction.

In *S. stercoralis* L3i, the most abundant transcripts encoding ES proteins include cytoskeletal proteins (i.e. myosin heavy chain, actin, tropomyosin, tubulin or paramyosin), metabolic enzymes (i.e. Trehalase, PEPCK, PGK, PGM, GAPDH, enolase), proteases, stress-response proteins, detoxifying enzymes, proteaseome components, most of them identified previously in *S. stercoralis* by proteomic studies [55]. These ES proteins play a major role in infection since they are present at the host-parasite interface and regulate host immune system [66]. ESPs also are among the target choice of new therapeutic solutions for helminth infections [67], as confirmed in the case of ivermectin (the currently the drug of choice for treating strongyloidosis) which has been shown to act reducing the secretion of ESPs from the ES apparatus in *Brugia malayi* microfilariae [68].

Recent studies using microarrays have identified highly expressed molecules in *S. stercoralis* L3i in comparison to L1

Table 5. Predicted therapeutic targets from *S. stercoralis* L3i.

No	Cluster ID	Description from NR Database	WBGene ID	*C. elegans* RNAi lethal phenotype	Interpro hits	*C. elegans* Gene Ontology	Drugbank targets hits	ESP
1	Contig 19 Singleton 1215	2,3-Bisphosphoglycerate independent phosphoglycerate mutase (*Ascaris suum*)	WBGene 00019001	Slow growth; embryonic lethal; larval arrest; organism morphology abnormal; egg laying abnormal; locomotion abnormal; life span abnormal	IPR006124 IPR017850	BIOLOGICAL PROCESSES: glucose catabolic process; positive regulation of growth rate; embryonic development ending in birth or egg hatching; nematode larval development; body morphogenesis; locomotion; determination of adult life span; oviposition. MOLECULAR FUNCTION: phosphoglycerate mutase CELLULAR COMPONENT: cytoplasm	2,3-Bisphosphoglycerate independent phosphoglycerate mutase	yes
2	Contig 462 Singletons 600, 2555, 3841	Hypothetical protein CBG01975 (*Caenorhabditis briggsae*) Glutamate synthase [NADPH] (*Ascaris suum*)	WBGene 00012326	Slow growth; embryonic lethal; larval arrest	IPR000583 IPR002932 IPR017932 IPR002489	BIOLOGICAL PROCESSES: electron transport; glutamate biosynthetic process; nitrogen compound metabolic process; embryonic development ending in birth or egg hatching; nematode larval development; positive regulation of growth rate. MOLECULAR FUNCTION: oxidoreductase activity, acting on the CH-NH2 group of donors, NAD or NADP as acceptor FAD binding; glutamate synthase activity, NADH or NADPH as acceptor; iron-sulfur cluster binding	Ferredoxin-dependent glutamate synthase 2	No
3	Singletons 77, 154, 179, 2466, 10457	Isocitrate lyase (*Strongyloides stercoralis*)	WBGene 00001564	Life span abnormal	IPR000918 IPR006254 IPR015813	BIOLOGICAL PROCESSES: glyoxylate cycle; carboxylic acid metabolic process; embryonic development; determination of adult life span. MOLECULAR FUNCTION: isocitrate lyase activity; malate synthase activity	Isocitrate lyase	No
4	Singletons 2046, 10787	Alcohol dehydrogenase I (*Candida albicans* WO-1)	WBGene 00010790	Life span abnormal	IPR002085 IPR011032 IPR013149	BIOLOGICAL PROCESSES: metabolic process determination of adult life span MOLECULAR FUNCTION: zinc ion binding oxidoreductase activity	Alcohol dehydrogenase	No

Putative proteins homologous to *C. elegans* proteins with lethal RNAi phenotype and with no homologue in the human host.
doi:10.1371/journal.pntd.0001513.t005

larvae, including cytochrome bc1, Hsp-90 and FAR-1, which potentially constitute new targets for intervention [25], all of which were present in our transcriptome data, but did not appear as druggable targets following our pipeline, possibly as these are not present in DrugBank, where only lethal RNAi phenotypes are included. Other important targets if interfered with, would still lead to expulsion of live worms form a host, like motility genes. In agreement with this, Garg and Ranganathan (2011) [22] have recently identified 19 contigs as putative drug targets in the *S. ratti* adult transcriptome, including myosin heavy chain, which is also one of the most abundant transcript of ES proteins in *S. stercoralis* (Table S4). This protein along with others like a metalloproteinase precursor, major sperm protein or triosephosphate isomerase (also identified in the *S. stercoralis* transcriptome in our study) did not appear as druggable molecules in our study, due to the presence of these proteins in host cells as well. In this context, efficient drugs as antihelmintics like benzimidazoles (they inhibit tubulin ß resulting in impaired microtubule formation during cell division) have much more affinity for tubulin in helminth cells than the tubulin found in the cells of mammals [69]. We found 11 potential targets for treatment against L3i larvae. As already mentioned, these are the first evolutive phase of *S. stercoralis* in the host, and constitute a good target for treatment. From these target molecules, four, with no homologues in the host, suggesting parasite specificity, are: 2,3-bisphosphoglycerate independent phosphoglycerate mutase, glutamate synthase, isocitrate lyase and alcohol dehydrogenase I. Only the first one was predicted as present in ES. Further studies are required to confirm whether these molecules are good drug targets for strongyloidiasis. Next-generation sequencing technologies are improving genomic and transcriptomic studies, and complemented by proteomic investigations, should allow the characterization of differential gene expression and essential pathways in all the developmental stages of *S. stercoralis*. The transcriptomic dataset described here constitutes the basis for future investigations enlightening the search for control measures for one of the most neglected diseases.

Supporting Information

Figure S1 Bioinformatics workflow used for transcriptomic data analysis. Bioinformatics workflow comprising Phase I (pre-processing and assembly), II (Nucleotide level annotation), III (prediction of excretory/secretory (ES) proteins) and IV (Protein-level annotation).
(DOC)

Table S1 *Strongyloides stercoralis* L3i gene ontology mapping using Interproscan.
(XLS)

Table S2 Peptidases found in the *S. stercoralis* L3i transcriptome.
(XLS)

Table S3 BLAST results of proteins found homologous to experimentally verified secretory proteins of parasitic helminths at 1e-^{15}.
(XLS)

Table S4 Top 50 abundant transcripts across *Strongyloides stercoralis* L3i excretory/secretory proteins using Blast2Go [44].
(DOCX)

Table S5 *S. stercoralis* L3i putative proteins found similar to known therapeutic targets in parasitic nematodes either by protein domains mapping or sequence similarity.
(DOCX)

Acknowledgments

The Bioinformatics Core Service of the IBMCP (UPV-CSIC) is acknowledged for its help in bioinformatics analyses.

Author Contributions

Conceived and designed the experiments: AM DB RT JGE. Performed the experiments: JS CMA MT DB AM. Analyzed the data: GG JF JO AM DB SR RT JGE. Contributed reagents/materials/analysis tools: MVD LP JMB JF JO GG SR. Wrote the paper: AM GG DB SR.

References

1. Siddiqui AA, Berk SL (2001) Diagnosis of *Strongyloides stercoralis* infection. Clin Infect Dis 33(7): 1040–1047.
2. Bethony J, Brooker S, Albonico M, Geiger SM, Loukas A, et al. (2006) Soil-transmitted helminth infections: Ascariasis, trichuriasis, and hookworm. Lancet 367(9521): 1521–1532.
3. Olsen A, van Lieshout L, Marti H, Polderman T, Polman K, et al. (2009) Strongyloidiasis–the most neglected of the neglected tropical diseases? Trans R Soc Trop Med Hyg 103(10): 967–972.
4. Keiser PB, Nutman TB (2004) *Strongyloides stercoralis* in the immunocompromised population. Clin Microbiol Rev 17(1): 208–217.
5. Vadlamudi RS, Chi DS, Krishnaswamy G (2006) Intestinal strongyloidiasis and hyperinfection syndrome. Clin Mol Allergy 4: 8.
6. Marcos LA, Terashima A, Canales M, Gotuzzo E (2011) Update on strongyloidiasis in the immunocompromised host. Curr Infect Dis Rep 13(1): 35–46.
7. Oltra-Alcaraz C, Igual-Adell R, Sanchez P, Blasco M, Sanchez O, et al. (2004) Characteristics and geographical profile of strongyloidiasis in healthcare area 11 of the valencian community (spain). J Infect 49(2): 152–158.
8. Seybolt L, Christiansen D, Barnett E (2006) Diagnostic evaluation of newly arrived asymptomatic refugees with eosinophilia. Clin Infect Dis 42(3): 363–367.
9. Dreyer G, FernandesSilva E, Alves S, Rocha A, Albuquerque R, et al. (1996) Patterns of detection of *Strongyloides stercoralis* in stool specimens: Implications for diagnosis and clinical trials. J Clin Microbiol 34(10): 2569–2571.
10. Sato Y, Otsuru M, Takara M, Shiroma Y (1986) Intradermal reactions in strongyloidiasis. Int J Parasitol 16(1): 87–91.
11. Hirata T, Uchima N, Kishimoto K, Zaha O, Kinjo N, et al. (2006) Impairment of host immune response against *Strongyloides stercoralis* by human T cell lymphotropic virus type 1 infection. Am J Trop Med Hyg 74(2): 246–249.
12. Sykes AM, McCarthy JS (2011) A coproantigen diagnostic test for *Strongyloides* infection. PLoS Negl Trop Dis 5(2): e955.
13. Igual-Adell R, Oltra-Alcaraz C, Soler-Company E, Sanchez-Sanchez P, Matogo-Oyana J, et al. (2004) Efficacy and safety of ivermectin and thiabendazole in the treatment of strongyloidiasis. Expert Opin Pharmacother 5(12): 2615–2619.
14. Suputtamongkol Y, Premasathian N, Bhumimuang K, Waywa D, Nilganuwong S, et al. (2011) Efficacy and safety of single and double doses of ivermectin versus 7-day high dose albendazole for chronic strongyloidiasis. PLoS Negl Trop Dis 5(5): e1044.
15. Wolstenholme A, Fairweather I, Prichard R, von Samson-Himmelstjerna G, Sangster N (2004) Drug resistance in veterinary helminths. Trends Parasitol 20(10): 469–476.
16. Riddle DL, Blumenthal T, Meyer BJ, Priess JR, eds (1997) *C. elegans* II. Cold Spring Harbor Laboratory Press, New York, USA.
17. Sugimoto A (2004) High-throughput RNAi in *Caenorhabditis elegans*: Genome-wide screens and functional genomics. Differentiation 72(2–3): 81–91.
18. Blaxter ML, De Ley P, Garey JR, Liu LX, Scheldeman P, et al. (1998) A molecular evolutionary framework for the phylum nematoda. Nature 392(6671): 71–75.
19. Cantacessi C, Mitreva M, Campbell BE, Hall RS, Young ND, et al. (2010) First transcriptomic analysis of the economically important parasitic nematode, *Trichostrongylus colubriformis*, using a next-generation sequencing approach. Infect Genet Evol 10(8): 1199–1207.
20. Cantacessi C, Gasser RB, Strube C, Schnieder T, Jex AR, et al. (2011) Deep insights into *Dictyocaulus viviparus* transcriptomes provides unique prospects for new drug targets and disease intervention. Biotechnol Adv 29(3): 261–271.
21. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: An automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 35(Web Server issue): W182–5.
22. Garg G, Ranganathan S (2011) *In silico* secretome analysis approach for next generation sequencing transcriptomic data. BMC Genomics 12(S3): S14.
23. Mello LV, O'Meara H, Rigden DJ, Paterson S (2009) Identification of novel aspartic proteases from *Strongyloides ratti* and characterisation of their evolutionary relationships, stage-specific expression and molecular structure. BMC Genomics 10: 611.
24. Yoshida A, Nagayasu E, Nishimaki A, Sawaguchi A, Yanagawa S, et al. (2011) Transcripts analysis of infective larvae of an intestinal nematode, *Strongyloides venezuelensis*. Parasitol Int 60(1): 75–83.
25. Ramanathan R, Varma S, Ribeiro JM, Myers TG, Nolan TJ, et al. (2011) Microarray-based analysis of differential gene expression between infective and noninfective larvae of *Strongyloides stercoralis*. PLoS Negl Trop Dis 5(5): e1039.
26. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437(7057): 376–380.
27. Nagaraj SH, Gasser RB, Ranganathan S (2007) A hitchhiker's guide to expressed sequence tag (EST) analysis. Brief Bioinform 8(1): 6–21.
28. Nagaraj SH, Gasser RB, 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(9): e301.
29. Nagaraj SH, Deshpande N, Gasser RB, Ranganathan S (2007) ESTExplorer: An expressed sequence tag (EST) assembly and annotation platform. Nucleic Acids Res 35(Web Server issue): W143–7.
30. Ranganathan S, Nagaraj SH, Hu M, Strube C, Schnieder T, et al. (2007) A transcriptomic analysis of the adult stage of the bovine lungworm, *Dictyocaulus viviparus*. BMC Genomics 8: 311.
31. Young ND, Hall RS, Jex AR, Cantacessi C, Gasser RB (2010) Elucidating the transcriptome of *Fasciola hepatica* - a key to fundamental and biotechnological discoveries for a neglected parasite. Biotechnol Adv 28(2): 222–231.
32. Young ND, Campbell BE, Hall RS, Jex AR, Cantacessi C, et al. (2010) Unlocking the transcriptomes of two carcinogenic parasites, *Clonorchis sinensis* and *Opisthorchis viverrini*. PLoS Negl Trop Dis 4(6): e719.
33. Ranganathan S, Menon R, Gasser RB (2009) 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 27(4): 439–448.
34. Robinson MW, Menon R, Donnelly SM, Dalton JP, Ranganathan S (2009) 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 8(8): 1891–1907.
35. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WE, et al. (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res 14(6): 1147–1159.
36. Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9(9): 868–877.
37. Iseli C, Jongeneel CV, Bucher P (1999) ESTScan: A program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc Int Conf Intell Syst Mol Biol. pp 138–148.
38. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden markov model: Application to complete genomes. J Mol Biol 305(3): 567–580.
39. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340(4): 783–795.
40. Bendtsen JD, Jensen LJ, Blom N, Von Heijne G, Brunak S (2004) Feature-based prediction of non-classical and leaderless protein secretion. Protein Eng Des Sel 17(4): 349–356.
41. Bennuru S, Semnani R, Meng Z, Ribeiro JM, Veenstra TD, et al. (2009) *Brugia malayi* excreted/secreted proteins at the host/parasite interface: Stage- and gender-specific proteomic profiling. PLoS Negl Trop Dis 3(4): e410.
42. Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300(4): 1005–1016.
43. Zdobnov EM, Apweiler R (2001) InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17(9): 847–848.
44. Conesa A, Götz S, Garcia-Gomez JM, Terol J, Talon M, et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18): 3674–3676.
45. Knox C, Law V, Jewison T, Liu P, Ly S, et al. (2011) DrugBank 3.0: A comprehensive resource for 'omics' research on drugs. Nucleic Acids Res 39(Database issue): D1035–41.
46. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38(Database issue): D355–60.
47. Bernal D, Carpena I, Espert AM, De la Rubia JE, Esteban JG, et al. (2006) Identification of proteins in excretory/secretory extracts of *Echinostoma friedi* (trematoda) from chronic and acute infections. Proteomics 6(9): 2835–2843.

48. Maizels RM, Yazdanbakhsh M (2003) Immune regulation by helminth parasites: Cellular and molecular mechanisms. Nat Rev Immunol 3(9): 733–744.
49. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3): 403–410.
50. Soblik H, Younis AE, Mitreva M, Renard BY, Kirchner M, et al. (2011) Life cycle stage-resolved proteomic analysis of the excretome/secretome from *Strongyloides ratti*: Identification of stage-specific protease. Mol Cell Proteomics (in press).
51. Cantacessi C, Mitreva M, Jex AR, Young ND, Campbell BE, et al. (2010) Massively parallel sequencing and analysis of the *Necator americanus* transcriptome. PLoS Negl Trop Dis 4(5): e684.
52. Cantacessi C, Campbell BE, Young ND, Jex AR, Hall RS, et al. (2010) Differences in transcription between free-living and CO2-activated third-stage larvae of *Haemonchus contortus*. BMC Genomics 11: 266.
53. Dicker AJ, Nath M, Yaga R, Nisbet AJ, Lainson FA, et al. (2011) *Teladorsagia circumcincta*: The transcriptomic response of a multi-drug-resistant isolate to ivermectin exposure in vitro. Exp Parasitol 127(2): 351–356.
54. Rabelo EM, Hall RS, Loukas A, Cooper L, Hu M, et al. (2009) Improved insights into the transcriptomes of the human hookworm *Necator americanus*–fundamental and biotechnological implications. Biotechnol Adv 27(2): 122–132.
55. Marcilla A, Sotillo J, Perez-Garcia A, Igual-Adell R, Valero ML, et al. (2010) Proteomic analysis of *Strongyloides stercoralis* L3 larvae. Parasitology 137(10): 1577–1583.
56. Brindley PJ, Gam AA, McKerrow JH, Neva FA (1995) Ss40: The zinc endopeptidase secreted by infective larvae of *Strongyloides stercoralis*. Exp Parasitol 80(1): 1–7.
57. Gomez Gallego S, Loukas A, Slade RW, Neva FA, Varatharajalu R, et al. (2005) Identification of an astacin-like metallo-proteinase transcript from the infective larvae of *Strongyloides stercoralis*. Parasitol Int 54(2): 123–133.
58. Maruyama H, Nishimaki A, Takuma Y, Kurimoto M, Suzuki T, et al. (2006) Successive changes in tissue migration capacity of developing larvae of an intestinal nematode, *Strongyloides venezuelensis*. Parasitology 132(Pt 3): 411–418.
59. Smooker PM, Jayaraj R, Pike RN, Spithill TW (2010) Cathepsin B proteases of flukes: The key to facilitating parasite control? Trends Parasitol 26(10): 506–514.
60. Dalton JP, Brindley PJ, Donnelly S, Robinson MW (2009) The enigmatic asparaginyl endopeptidase of helminth parasites. Trends Parasitol 25(2): 59–61.
61. Ohta N, Kumagai T, Maruyama H, Yoshida A, He Y, et al. (2004) Research on calpain of *Schistosoma japonicum* as a vaccine candidate. Parasitol Int 53(2): 175–181.
62. Jex AR, Liu S, Li B, Young ND, Hall RS, et al. (2011) *Ascaris suum* draft genome. Nature (in press). doi: 10.1038/nature10553.
63. Campbell BE, Boag PR, Hofmann A, Cantacessi C, Wang CK, et al. (2011) Atypical (RIO) protein kinases from *Haemonchus contortus*–promise as new targets for nematocidal drugs. Biotechnol Adv 29(3): 338–350.
64. Campbell BE, Boag PR, Hofmann A, Cantacessi C, Wang CK, et al. (2011) Atypical (RIO) protein kinases from *Haemonchus contortus*–promise as new targets for nematocidal drugs. Biotechnol Adv 29(3): 338–350.
65. Campbell BE, Hofmann A, McCluskey A, Gasser RB (2011) Serine/threonine phosphatases in socioeconomically important parasitic nematodes–prospects as novel drug targets? Biotechnol Adv 29(1): 28–39.
66. Hewitson JP, Grainger JR, Maizels RM (2009) Helminth immunoregulation: The role of parasite secreted proteins in modulating host immunity. Mol Biochem Parasitol 167(1): 1–11.
67. Bungiro R, Cappello M (2011) Twenty-first century progress toward the global control of human hookworm infection. Curr Infect Dis Rep 13(3): 210–217.
68. Moreno Y, Nabhan JF, Solomon J, Mackenzie CD, Geary TG (2010) Ivermectin disrupts the function of the excretory-secretory apparatus in microfilariae of *Brugia malayi*. Proc Natl Acad Sci U S A 107(46): 20120–20125.
69. Lacey E (1990) Mode of action of benzimidazoles. Parasitol Today 6(4): 112–115.

6.2 Conclusions

This collaborative work is the direct application of our bioinformatic approach discussed in Chapter 3 for secretome analysis using transcriptomic data with minor modifications. Overall, the present dataset should provide a solid foundation for future fundamental genomic, proteomic, and metabolomics explorations of *S. stercoralis* and the development of novel therapeutic solutions for strongyloidiasis.

Initially a total of 253266 short reads were generated from output of cDNA sequencing using 454 Roche sequencing technology. The reads were assembled into 11352 contigs using MIRA and CAP3. These contigs were conceptually translated into 8037 proteins. 1213 (15.09%) proteins were predicted as ES proteins. We found 14 contigs and singletons corresponding to 4 different proteins which have lethal RNAi phenotypes present in *C. elegans*, not present in human and similar to known drug targets.

The successful implementation of our computational approach to parasitic nematode, *S. stercoralis* has provided a platform for extending our analysis to the transcriptome of another novel parasitic trematode (*Echinostoma caproni*) (Chapter 7).

Chapter 7: Transcriptome characterization of the model organism, *Echinostoma caproni*

7.1 Summary

Human echinostomiasis is an intestinal parasitic disease caused by one of at least sixteen trematode flukes from the genus *Echinostoma*. *Echinostoma caproni* is an important food-borne human trematode parasite, affecting more than 40 million people worldwide. As very little is known knowledge about this parasite and its relationship with its hosts at the molecular level, a transcriptome analysis of the adult stage of *E. caproni* would provide clues regarding host-parasite interactions and will serve as a model organism for combatting human echinostomiasis.

As in Chapter 6, we applied our bioinformatics approach (Chapters 3 and 4) with minor modifications to analyse the first transcriptome of the adult stage of *E. caproni*. The recently available NGS assembler, iAssembler replaced the sequential running of MIRA and CAP3 (Chapter 6) and the results are presented in Publication 6 (Additional Files available on CD).

Transcriptome characterization of the model organism *Echinostoma caproni*

GAGAN GARG[1], DOLORES BERNAL[2], MARIA TRELIS[3], JAVIER FORMENT[4], JAVIER ORTIZ[5], LAIA PEDROLA[6], JUAN MARTINEZ-BLANCH[6], J. GUILLERMO ESTEBAN[3], SHOBA RANGANATHAN[1,7], RAFAEL TOLEDO[3] AND ANTONIO MARCILLA*[3]

[1]Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney SW 2109, Australia.

[2]Departamento de Bioquímica y Biología Molecular, Universitat de València, C/ Dr. Moliner, 50, 46100 Burjassot (Valencia), Spain.

[3]Área de Parasitología, Departamento de Biologia Celular y Parasitología, Universitat de València, Av. V.A. Estellés, s/n, 46100 Burjassot (Valencia), Spain.

[4]Servicio de Bioinformática, Instituto de Biologia Molecular y Celular de Plantas, Universitat Politécnica de València, Ingeniero Fausto Elio, s/n, 46022 Valencia, Spain.

[5]Unidad de Bioinformática, SCSIE, Universitat de València, C/ Dr. Moliner, 50, 46100 Burjassot (Valencia), Spain

[6]LifeSequencing S.L., Parc Científic Universitat de València, Bldg. 2 Biotech, C/ Cat. A. Escardino, 9, 46980 Paterna (Valencia), Spain

[7]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117597

*Corresponding author: Dr. Antonio Marcilla

Área de Parasitología, Departamento de Biologia Celular y Parasitología, Universitat de València, Av. V.A. Estellés, s/n, 46100 Burjassot (Valencia), Spain.

Tel: (34) 963544491

Fax: (34) 963544769

Email: antonio.marcilla@uv.es

Abstract

Background

Echinostomes are cosmopolitan parasites that infect a large number of different warm-blooded hosts, both in nature and the laboratory. They also constitute an important group of food-borne trematodes of public health importance chiefly in the Southeast Asia and the Far East. In addition, echinostomes are an ideal model for the study of several aspects of the biology of intestinal helminths, since they present a number of advantages such as their large worm size or the facility for the maintenance of their life cycles in the laboratory among other features. Recently, several studies have pointed out their great value in regard to the study of intestinal helminth-vertebrate host relationships. The knowledge of their genome, transcriptome and proteome can have a deep impact in developing control strategies for other intestinal helminths.

Results

We present the first transcriptome of the adult stage of *E.caproni* using 454 sequencing coupled to a semi-automated bioinformatic analyses. 557236 raw sequence reads were assembled into 28577 contiguous sequences using iAssembler. 23296 putative proteins were characterized based on homology, gene ontology and/or biochemical pathways. Comparisons of the transcriptome of *E. caproni* with those of other trematodes revealed similarities in transcription for molecules inferred to have key roles in parasite-host interactions. Enzymatic proteins like kinases and peptidases were abundant. 3415 putative excretory/secretory proteins were compiled including non-classical secretory proteins. Potential drug targets were also identified.

Conclusions

Overall, the present dataset should provide a solid foundation for future fundamental genomic, proteomic, and metabolomics explorations of *E. caproni*, as well as a basis for

applied outcomes such as the development of novel methods of intervention against this model organism and related parasites.

Background

Nearly 40 million people are infected with food-borne trematodes, including the intestinal flatworms (Trematoda: Echinostomatidae) [1-3]. Echinostomes are cosmopolitan parasites that infect a large number of different warm-blooded hosts, both in nature and the laboratory. The broad host specificity of echinostomes toward the definitive host is the result of phylogenetic, physiological, and ecological accommodations between the parasite and the host in an evolutionary dynamic process [4]. A total of 20 species belonging to nine genera of Echinostomatidae are known to cause human infections around the world [5, 6]. They constitute an important group of food-borne trematodes of public health importance chiefly in the Southeast Asia and the Far East [6].

Echinostomes are an ideal model for the study of several aspects of the biology of intestinal helminths since they present a number of advantages such as their large worm size or the facility for the maintenance of their life cycles in the laboratory among other features. For these reasons, they have been used for decades as experimental models. Recently, a number of findings using echinostomes as models have shown that echinostomes may be of great importance for future developments in parasitology, particularly in regard to the study of intestinal helminths–vertebrate host relationships.

Echinostoma caproni has a wide range of definitive hosts, although its compatibility differs considerably between rodent species on the basis of worm survival and development [7]. Because of these characteristics, the *E. caproni*-rodent systems are highly suitable for elucidating aspects of the host specific components that determine the course of infections with intestinal helminths [7].

Although the extensive use of these host-parasite models, little is known about the proteome of echinostomes adult worms [8-10]. Several antigenic components of the ESP of *E. caproni* were identified by Sotillo *et al*. (2008) [11], and very recently we have identified 29 proteins of the excretory/secretory proteome of *E. caproni* adult worms since few proteins of this trematode are present in the database [12].

As pointed out recently by Cantacessi *et al*., (2012) [13], the advent and integration of high-throughput "-omic" technologies (e.g., genomics, transcriptomics, proteomics and metabolomics) are becoming fundamental to explore the systems biology of helminths, providing unique opportunities for the development of entirely new strategies for the treatment and control of neglected parasitic diseases. New bioinformatic tools based on robust assembly protocol for next generation sequencing (NGS) data, along with compilation of a dataset of experimentally determined excretory/secretory (ES) proteins of parasitic helminths, and annotation software like KAAS [14], allow efficient and up-to-date homology-based predictions [15].

To date, there are few molecular and genomic studies on *Echinostoma* species [9], with the genome of *E. caproni* under sequencing (http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?study=ERP000904). In fact, only 358 ESTs are currently available in the NCBI database of July 2012 (from *E. paraensei*).

In the present study, we have explored and functional annotated the transcriptome of the adult stage of *E. caproni* by 454 sequencing coupled to semi-automated bioinformatic analyses and predicted potential therapeutic targets for trematodiasis.

Results

Pre-processing and assembly of *E.caproni* transcriptome

Initially a total of 557236 short reads (190721704 bases) were generated from sff file with 342±156 bases (average length ± standard deviation) with a average GC content of 42.7%. These short reads were pre-processed for removal of vector, adapter and other contaminant

sequences. This resulted in 399576 (71.7%) high quality short reads. High quality reads were assembled into 28577 unigenes (contigs + singletons) using iAssembler [16]. Summary of assembly is shown in Table 1. After the assembly all the unigenes were conceptually translated into 23296 proteins using ESTScan [17] (sequences available from http://biolinux.uv.es/marcilla/). Unigenes were also searched for sequence similarity against non redundant database (NR, NCBI) and other trematoda ESTs using BLAST [18] at permissive (E-value: $<1E^{-05}$), moderate ($<1E^{-15}$) and/or stringent ($<1E^{-30}$) search strategies (Table 2). As shown in Table 2, roughly half of the unigenes had correspondence when searching in BLASTX, obtaining the highest homology score when compared to related trematodes like *Fasciola gigantica* and *Fasciola hepatica* (40.8% and 40.2%, respectively), being the majority (60%) of the *E. caproni* transcripts species-specific.

Annotation of *E. caproni* proteins

Putative proteins were annotated based on protein families and domains using Interproscan [19] and mapped to biochemical pathways using KAAS [14]. Out of 23296 putative proteins, we were able to annotate 8217 (35.27%) proteins with protein domains and families. The most represented Interpro terms are shown in Table 3. We established pathway associations to 2266 (9.72%) putative proteins with 1350 unique KEGG orthologs. Maximum number of *E. caproni* proteins belonged to metabolism and genetic information processing, being these pathways important for parasite survival (Table 4). *E. caproni* proteins were found associated with important biological processes, being enzymes, splicesome and ribosome (Table 5). A total of 6246 (26.81%) proteins were annotated with GO terms (5607 {Molecular Function}, 2180 {Cellular Component}, and 3475 {Biological Process}) based on Interpro terms annotations (Supplementary file 1).

All the proteins were also searched for sequence similarity against human, *C. elegans*, *B. malayi*, *S. mansoni*, *S. japonicum* and other trematode species proteins, using BLASTP at

permissive (E-value: <$1E^{-05}$), moderate (<$1E^{-15}$) and/or stringent (<$1E^{-30}$) search strategies (Table 6).

Prediction of excretory-secretory (ES) proteins

ES protein predictions were carried out using Phase III of the workflow (Figure 1). Firstly, 1102 (4.73%) proteins were predicted as classical secreted proteins using SignalP [20]. The remaining 22194 (95.27%) proteins, which were predicted as non-secretory by SignalP, were processed by SecretomeP [21] for prediction of non-classical secretory proteins. A total of 1253 (5.38%) proteins were predicted as non-classical secretory proteins using SecretomeP. The classical and non-classical secretory proteins (2355; 10.1%) from these two programs were analyzed by TargetP [22] for mitochondrial proteins. Only 39 proteins were predicted as mitochondrial proteins using TargetP at 95% specificity. These thirty nine proteins were removed from the set of 2355 secreted proteins, and 2316 secretory proteins were passed to TMHMM [23] for the prediction of transmembrane proteins. 558 (2.39%) proteins, predicted as transmembrane proteins having one or more transmembrane helices, were removed from the secretory protein dataset. A total of 1758 (7.55%) proteins were finally predicted as ES proteins from the computational prediction pipeline.

Proteins that were considered non-secretory by SecretomeP were matched to our in-house dataset of 1485 non redundant experimentally verified parasitic helminth proteins, using BLASTP similarity search. We found additional 1657 (7.11%) proteins similar to known ES proteins by this homology search approach at E value (1e-15). Thus, for annotation and analyses in Phase III, we compiled a total of 3415 ES proteins, which constitutes 14.66% of our putative proteins.

***E. caproni* proteins as drug targets**

We found 16244 (69.73 %) *E. caproni* putative proteins which had no homologues present in humans (Table 6) at permissive (E-value: <$1E^{-05}$), and therefore are preferred targets for parasite intervention strategies. These human dissimilar proteins of *E. caproni* were

checked for known drug targets, available from DrugBank 3.0 [24]. We found 8 unigenes that are not present in human and similar to known drug targets. These could represent potential therapeutic targets for human trematodiasis as shown in Table 7.

E. caproni proteins interactions

We checked our putative protein dataset for interacting partners based on data obtained from IntAct [25], BioGRID [26] and DIP [27] using BLASTP (Supplementary file 2). 5754 (24.7%) of the proteins were found to have interacting proteins, having sequence identity ranging from 100% to as low as 19%.

Discussion

The present study provides the first detailed analyses of the transcriptome of the model trematode *E. caproni* and identifies some groups of molecules predicted to play key biological functions in this parasite. Although the genome sequencing of this trematode is currently underway, no assembly is available yet (see above). To better identify peptides in proteomic searches we have constructed the transcriptome of *E. caproni* adults using the Roche/454 technology [28]. The comparison among immature stages and adult transcriptomes could help in the knowledge of the molecular basis of pathological changes in hosts. Comprehensive transcriptomic data are also improving the identification and classification of somatic, excretory/secretory and tegumental proteins from *E. caproni*, as it occurs with other related trematodes [29, 30].

Comparing the *E. caproni* transcriptome with the coding sequences of other trematodes like *F. hepatica* and *F. gigantica* confirmed the high diversity of trematode transcriptomes. As described for other helminths, high expression dynamics of the transcriptome may well suggest an impact of the adaptation to parasitism [31]. A total of 28577 unigenes were inferred from the present EST dataset, thus increasing the number of predicted proteins currently available (for this stage/species) in public databases by more than 1200-fold [Genebank as in August 2012]. The number and GC content of unigenes is similar to the

one obtained in transcriptome analysis of other related species like *F. hepatica* [29], and *Fasciola gigantica* [30].

It is noteworthy to mention that close to 50% of the putative proteins of *E. caproni* adult transcriptome remain unannotated, warranting further genomic and functional characterization studies.

In our study, we compiled 3415 putative ES proteins among the 23296 (14.66%) *E. caproni* annotated proteins using a new semi-automated computational approach. With this pipeline the secretome of *Strongyloides ratti* adults [15] and *S. stercoralis* L3 larvae [32] have been recently obtained, increasing the identification of non-classical secretory proteins. ES proteins (known as secretome) are crucial for host-parasite interactions, predicted as good diagnostic tools, since they constitute the first parasitic material in contact with the host [33]. In previous studies we could identify 29 proteins in the secretome of *E. caproni* by a proteomic approach [Sotillo 2010], and now thanks to the transcriptome, ten times more proteins can be identified (data not shown). This fact reflects the importance of having good datasets to improve identifications in organism with no genome sequence available.

We found eight potential targets for treatment against *E. caproni* adults, none of them predicted as secreted (Table 7). As already mentioned, adults are the evolutive phase of *E. caproni* in the host intestine, and constitute a good target for treatment, since they have no homologues in the host, suggesting parasite specificity. Further studies are required to confirm whether these molecules are good drug targets for trematodiasis. Next-generation sequencing technologies are improving genomic and transcriptomic studies, and complemented by proteomic investigations, should allow the characterization of differential gene expression and essential pathways in all the developmental stages of the model organism *E. caproni*. The transcriptomic dataset described here constitutes the basis

for future investigations enlightening the search for control measures for intestinal helminthiasis.

Conclusions

Here we present the first transcriptome of the adult stage of the model organism *E. caproni* using 454 sequencing coupled to a semi-automated bioinformatic analysis. This dataset will provide a solid foundation for future explorations of *E. caproni*, as well as a basis for the development of novel therapeutic solutions against trematodiasis.

Methods

Accession numbers

The nucleotide sequence data obtained for this study are available in the GenBank database under accession number ERP001572.

The assembled data from this study can be requested from the corresponding author.

Parasite material and experimental infections

The species used in this study and the first and second intermediate snail hosts have been previously described [34]. Encysted metacercariae of *E. caproni* were removed from the kidneys and pericardial cavities of experimentally infected *Biomphalaria glabrata* snails and used to infect ICR mice (*Mus musculus*). Male mice, weighing 32-40g, were infected through a stomach tube with 75 metacercariae each of *E. caproni*. The animals were maintained under standard conditions with food and water *ad libitum*. Adult worms of *E. caproni* were collected from the intestines of mice, thoroughly washed with phosphate-buffered saline (PBS; pH 7.4) containing protease inhibitors (10mM EDTA, and 1mM PMSF) and samples were processed for RNA isolation.

RNA isolation, cDNA synthesis and 454 sequencing

Total RNA from 3 adults was prepared using Vantage^{TM} Total RNA purification kit (Marligen Biosciences, Ijamsville, MD, USA) following the manufacturers' instructions

and treated with Ambion DNA-free™ DNase (Ambion/Applied Biosystems, Austin, TX). The integrity of the RNA was verified by gel electrophoresis and the yield determined using the nanoDrop ND-1000 UV-VIS spectrophotometer v.3.2.1 (NanoDrop Technologies, Wilmington, DE).

The cDNA library was constructed from 0.5 µg total RNA using MINT cDNA Synthesis Kit (Cat#SK001, Evrogen). First strand cDNA synthesis starts from 3'-primer comprising oligo(dT) to enrich mRNA as template. Double strand cDNA synthesis was performed using 17 cycles of PCR amplification. Total cDNA was digested with restriction enzyme *GsuI* in order to remove Poly (A) tails. cDNA obtained was used to perform a library with the required sequencing adaptors and was then sequenced using the Genome Sequencer (GS) FLX instrument (Roche Diagnostics) [Roche 454].

Bioinformatic analysis of sequence data

The overall bioinformatics analysis strategy followed in this study is similar to the applied to transcriptomic analysis of *S. ratti* [15] and *S. stercoralis* [32]. FASTA and associated quality files were preprocessed for the removal of sequence adapters. Cleaned reads were assembled using iAssembler. Unigenes (Contigs + Singletons) generated were matched using BLASTX with the NCBI non-redundant sequence database; http://www.ncbi.nlm.nih.gov, BLASTN with dbEST [35] trematode ESTs (www.ncbi.nlm.nih.gov/dbEST/), BLASTN with *Schistosoma mansoni* ESTs, BLASTN with *Brugia malayi* transcripts, BLASTN with other trematode transciptomes (*Clonorchis sinensis* [36], *Fasciola hepatica* [37], *Fasciola gigantica* [29], *Opisthorchis viverrini* [36]), BLASTN with newly sequenced draft trematode genomes (*Clonorchis sinensis* [37], *Schistosoma haematobium* [38]) and BLASTN with newly sequenced *Ascaris suum* draft genome [39] using permissive (E-value: $<1E^{-05}$), moderate ($<1E^{-15}$) and/or stringent ($<1E^{-30}$) search strategies.

E. caproni unigenes were conceptually translated into putative proteins using the program ESTScan. Putative protein sequences were subjected to secretome analysis using TMHMM (a membrane topology prediction program) to predict transmembrane domains, SignalP 3.0 (signal peptide prediction program), SecretomeP (non-classical secretory proteins prediction program) and TargetP (mitochondrial protein prediction program). Briefly, excretory/secretory (ES) proteins were selected based on the presence of a signal peptide at the N-terminus using SignalP 3.0 (employing both the neural network and hidden Markov models) or predicted as secretory using SecretomeP, predicted as non-mitochondrial by TargetP and absence of transmembrane domains. In addition to computational prediction of ES proteins were identified and collated based on sequence homology (BLASTP, E-value<$1E^{-15}$) to known ES proteins found in parasitic helminth secretome studies.

Putative proteins were classified functionally using InterProScan, employing the default search parameters. Based on their homology to conserved domains and protein families, predicted proteins were classified into Gene Ontology (GO) categories (http://www.geneontology.org/) based on molecular function, cellular component and biological process using interpro terms. Putative proteins were also subjected to pathway analysis, utilizing KAAS, which maps the putative proteins to biochemical pathways in which they are involved and categories of BRITE objects like enzymes and translation factors. Putative proteins were subjected to BLASTP (Wormpep) to identify *C. elegans* known proteins homologues present in *E. caproni* proteins. These proteins were also searched for sequence homology in human (host) proteins. All the proteins which were found non-homologous to human proteins were searched for known drug targets present in DrugBank.

List of abbreviations

BRITE: Biomolecular Relations in Information Transmission and Expression

KEGG: Kyoto Encyclopedia of Genes and Genomes

KAAS: KEGG automatic annotation server

Competing Interests

The authors declare that they have no competing interests.

Authors' contributions

Conceived and designed the experiments: AM GG SR DB RT JGE. Performed the experiments: MT DB AM. Analyzed the data: GG JF JO AM DB SR RT JGE. Contributed reagents/materials/analysis tools: LP JMB JF JO GG SR. Wrote the paper: GG SR DB AM.

Acknowledgements

GG would like to acknowledge Macquarie University for an APA scholarship and PGRF. The Bioinformatics Core Service of the IBMCP (UPV-CSIC) is acknowledged for its help in bioinformatics analyses.

This work was supported by projects SAF2010-16236 and PS09/02355 from Spanish Ministry of Science and Innovation (Madrid, Spain) and FEDER and project PROMETEO/2009/081 from Conselleria d'Educació, Generalitat Valenciana (Valencia, Spain).

References

1. Fürst T, Keiser J, Utzinger J: **Global burden of human food-borne trematodiasis: a systematic review and meta-analysis.** *Lancet Infect Dis* 2012 **12**: 210-221.
2. Toledo R, Esteban JG, Fried B: **Recent advances in the biology of echinostomes**. *Adv Parasitol*, 2009, **69**:147-204.
3. Toledo R, Esteban JG, Fried B. **Current status of food-borne trematode infections**. *Eur J Clin Microbiol Infect Dis* 2012, **31**:1705-1718.
4. Huffman JE, Fried B. **Echinostoma and echinostomiasis**. *Adv Parasitol* 1990, **29**:215–269.
5. Chai, JY. **Intestinal flukes**. In: Murrell KD, Fried B (eds). Food-borne parasitic zoonoses: fish and plant-borne parasites. World class parasites, vol 11. Springer, New York, 2009, pp 53–115.
6. Chai, JY. **Echinostomes in humans**. In: Toledo R, Fried B (eds). The Biology of

Echinostomes: From the molecule to the community. Springer, New York, 2009

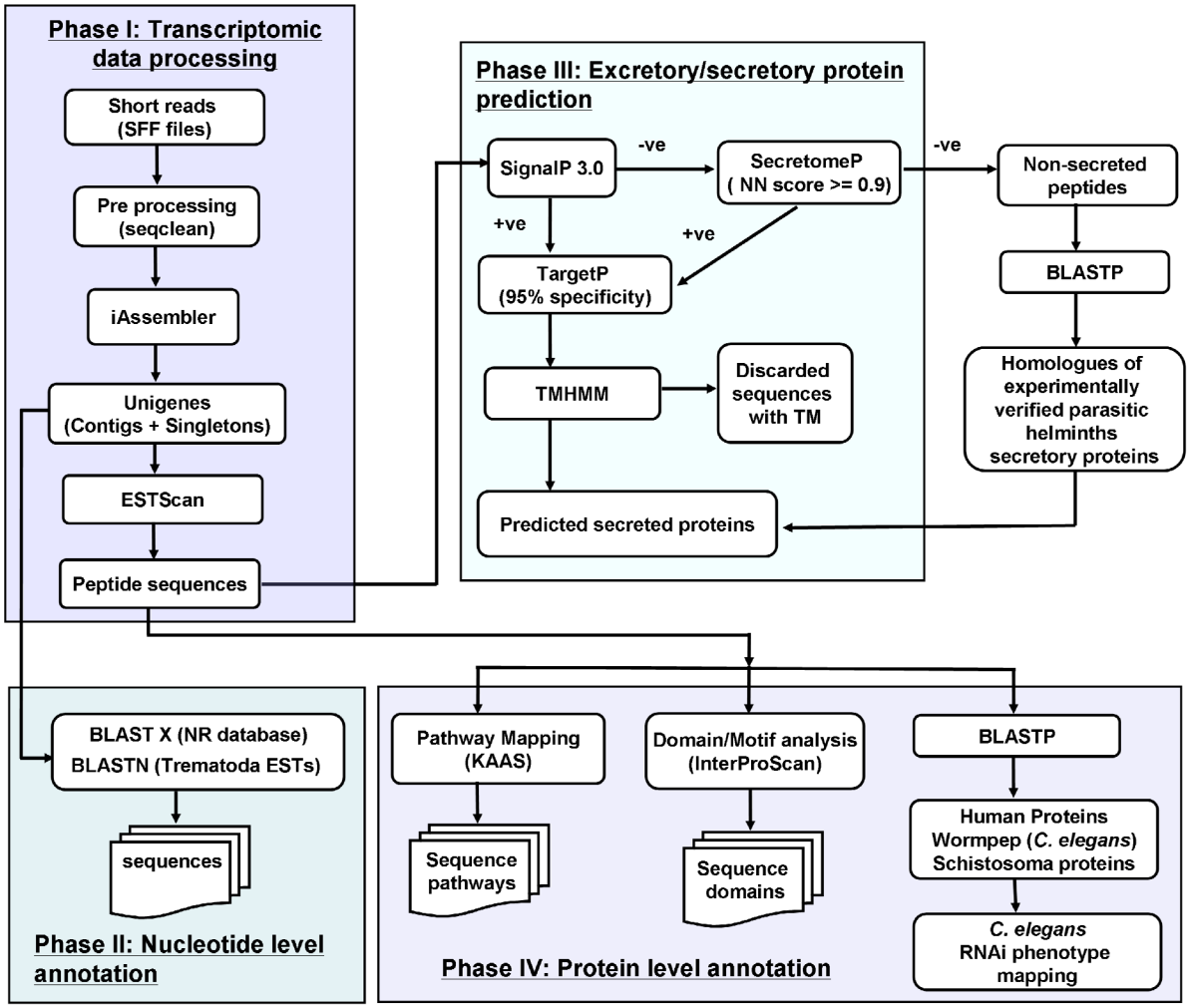
7. Toledo R, Fried B: **Echinostomes as experimental models for interactions between adult parasites and vertebrate hosts**. *Trends Parasitol* 2005, **21**:251-254.
8. Bernal D, Carpena I, Espert AM, De la Rubia JE, Esteban JG, Toledo R, *et al.* **Identification of proteins in excretory/secretory extracts of *Echinostoma friedi* (Trematoda) from chronic and acute infections**. *Proteomics* 2006, **6**: 2835-2843.
9. Marcilla, A. (2009). **Echinostomes: genomics and proteomics**. In: Fried B, Toledo R (eds). The Biology of Echinostomes: from the molecule to the community. Springer, New York, 2009, pp 207-228.
10. Toledo R, Bernal MD, Marcilla A: **Proteomics of foodborne trematodes**. *J Proteomics* 2011, **74**(9):1485-1503.
11. Sotillo J, Valero L, Sanchez Del Pino MM, Fried B, Esteban JG, Marcilla A, Toledo R: **Identification of antigenic proteins from *Echinostoma caproni* (Trematoda) recognized by mouse immunoglobulins M, A and G using an immunoproteomic approach.** *Parasite Immunol* 2008, 30:271-279.
12. Sotillo J, Trudgett A, Halferty L, Marcilla A, Esteban JG, Toledo R: ***Echinostoma caproni*: differential tegumental responses to growth in compatible and less compatible hosts.** *Exp Parasitol* 2010, 125(3):304-309.
13. Cantacessi C, Campbell BE, Jex AR, Young ND, Hall RS, Ranganathan S, Gasser RB: **Bioinformatics meets parasitology**. *Parasite Immunol* 2012, **34**:265-275.
14. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M: **KAAS: an automatic genome annotation and pathway reconstruction server.** *Nucleic Acids Res* 2007, **35:**W182-185.
15. Garg G, Ranganathan S: **In silico secretome analysis approach for next generation sequencing transcriptomic data.** *BMC Genomics* 2011, **12 Suppl 3:**S14.
16. 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.
17. 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.
18. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-404.
19. Zdobnov EM, Apweiler R: **InterProScan--an integration platform for the signature-recognition methods in InterPro.** *Bioinformatics* 2001, **17:**847-848

20. Bendtsen JD, Nielsen H, von Heijne G, Brunak S: **Improved prediction of signal peptides: SignalP 3.0.** *J Mol Biol* 2004, **340:**783-795.
21. 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:**349-356
22. Emanuelsson O, Nielsen H, Brunak S, von Heijne G: **Predicting subcellular localization of proteins based on their N-terminal amino acid sequence.** *J Mol Biol* 2000, **300:**1005-1016.
23. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes**. *J Mol Biol* 2001, **305**:567-580.
24. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, et al: **DrugBank 3.0: a comprehensive resource for 'omics' research on drugs.** *Nucleic Acids Res* 2011, **39:**D1035-1041.
25. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, et al: **IntAct--open source resource for molecular interaction data.** *Nucleic Acids Res* 2007, **35:**D561-565.
26. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, Nixon J, Van Auken K, Wang X, Shi X, et al: **The BioGRID Interaction Database: 2011 update.** *Nucleic Acids Res* 2011, **39:**D698-704.
27. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D: **DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions.** *Nucleic Acids Res* 2002, **30:**303-305.
28. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al: **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* 2005, **437**:376-380.
29. 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.** *Biotechnol Adv* 2010, **28:**222-231.
30. Young ND, Jex AR, Cantacessi C, Hall RS, Campbell BE, Spithill TW, Tangkawattana S, Tangkawattana P, Laha T, Gasser RB: **A portrait of the transcriptome of the neglected trematode, Fasciola gigantica--biological and biotechnological implications.** *PLoS Negl Trop Dis* 2011, **5:**e1004
31. 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.

32. Marcilla A, Garg G, Bernal D, Ranganathan S, Forment J, Ortiz J, Munoz-Antoli C, Dominguez MV, Pedrola L, Martinez-Blanch J, *et al.*: **The Transcriptome Analysis of *Strongyloides stercoralis* L3i Larvae Reveals Targets for Intervention in a Neglected Disease.** *PLoS Negl Trop Dis* 2012, **6:**e1513.
33. Hewitson JP, Grainger JR, Maizels RM: **Helminth immunoregulation: The role of parasite secreted proteins in modulating host immunity**. *Mol Biochem Parasitol* 2009, **167**: 1–11.
34. Toledo R, Espert A, Munoz-Antoli C, Marcilla A, Fried B, Esteban JG. **Kinetics of antibodies and antigens in serum of mice experimentally infected with *Echinostoma caproni* (Trematoda: Echinostomatidae)**. *J Parasitol* 2005, **91**: 978-80.
35. Boguski MS, Lowe TM, Tolstoshev CM: **dbEST--database for "expressed sequence tags".** *Nat Genet* 1993, **4:**332-333.
36. 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 Negl Trop Dis* 2010, **4:**e719.
37. Wang, X., Chen, W., Huang, Y., Sun, J., Men, J., Liu, H., Luo, F., Guo, L., Lv, X., Deng, C., Zhou, C., Fan, Y., Li, X., Huang, L., Hu, Y., Liang, C., Hu, X., Xu, J. & Yu, X. **The draft genome of the carcinogenic human liver fluke *Clonorchis sinensis*.** *Genome Biol* 2012, **12**: R107.
38. Young ND, Jex AR, Li B, Liu S, Yang L, Xiong Z, Li Y, Cantacessi C, Hall RS, Xu X, et al: **Whole-genome sequence of *Schistosoma haematobium*.** *Nat Genet* 2012, **44:**221-225
39. Jex AR, Liu S, Li B, Young ND, Hall RS, Li Y, Yang L, Zeng N, Xu X, Xiong Z, Chen F, Wu X, Zhang G, Fang X, Kang Y, Anderson GA, Harris TW, Campbell BE, Vlaminck J, Wang T, Cantacessi C, Schwarz EM, Ranganathan S, Geldhof P, Nejsum P, Sternberg PW, Yang H, Wang J, Gasser RB. ***Ascaris suum* draft genome**. *Nature* 2012, **479**:529-533.

Figures

Figure 1 - Bioinformatics workflow used for transcriptomic data analysis.
Bioinformatics workflow comprising, Phase I (pre-processing and assembly), II (Nucleotide level annotation), III (prediction of excretory/secretory (ES) proteins) and IV (Protein-level annotation).



Tables

Table 1. Summary of *E. caproni* dataset assembly.

Statistics	***E. caproni* dataset**
Number of raw reads	557236
Number of cleaned reads	399576
Number of unigenes (contigs + Singletons)	28577
Maximum length of unigene	3410
Minimum length of unigene	101
Average length of unigene	527.9
Total number of bases present in unigenes	15086394
Average GC content of unigenes	45.3 %
Number of unigenes greater than 1000 bases	2323
N50 value for assembly	557
N90 value for assembly	337

Table 2. Sequence homology inferred between *E. caproni* current dataset and other datasets.

Dataset	**Hits (1e-05)**	**Hits (1e-15)**	**Hits (1e-30)**
Trematode ESTs (dbEST)	8716	4922	3193
Schistosoma mansoni ESTs	3468	1638	927
Brugia malayi transcripts	1409	596	291
NR database (NCBI) (BLASTX)	14172	10189	6164
Clonorchis sinensis	3963	1665	780
Fasciola hepatica	11489	6516	3324
Fasciola gigantica	11682	5799	2812
Opisthorchis viverrini	3864	1460	581
Clonorchis sinensis [draft genome]	3986	1689	896
Ascaris suum [draft genome]	833	208	63
Schistosoma haematobium [draft genome]	1875	1075	796

Table 3. Top 10 most represented protein domains found in putative proteins using Interproscan.

InterPro description	InterPro code	Number of putative proteins (%)
Trematode Eggshell synthesis	IPR012615	226 (0.97)
EF-Hand 2	IPR018249	196 (0.84)
Ferritin/ribonucleotide reductase-like	IPR009078	192 (0.82)
EF-Hand 1, calcium binding site	IPR018247	151 (0.65)
Tubulin	IPR000217	145 (0.62)
Thioredoxin-like fold	IPR012336	134 (0.57)
RNA recognition motif domain	IPR000504	127 (0.54)
Peptidase C1A, papain	IPR013128	125 (0.53)
Heat shock protein Hsp70	IPR001023	118 (0.50)
Protein kinase-like domain	IPR011009	116 (0.49)

Table 4. KEGG pathways of putative proteins inferred from the transcriptome of *E. caproni*.

Parent KEGG pathway	**No. of putative proteins (%)**	**Top KEGG pathway in the category**
Metabolism		
Carbohydrate metabolism	335	Glycolysis/Gluconeogenesis
Energy metabolism	120	Oxidative phosphorylation
Lipid metabolism	78	Glycerophospholipid metabolism
Nucleotide metabolism	142	Purine metabolism
Amino acid metabolism	123	Arginine and proline metabolism
Metabolism of other amino acids	36	Glutathione metabolism
Glycan biosynthesis and metabolism	66	N-Glycan biosynthesis
Metabolism of Cofactors and Vitamins	57	Porphyrin and chlorophyll metabolism, Nicotinate and nicotinamide metabolism
Metabolism of Terpenoids and Polyketides	4	Terpenoid backbone biosynthesis
Xenobiotics biodegradation and metabolism	21	Drug metabolism – other enzymes
Genetic Information processing		
Transcription	159	Spliceosome
Translation	459	Ribosome

Folding, sorting and degradation	269	Protein processing in endoplasmic reticulum
Replication and repair	119	DNA replication
Environmental information processing		
Membrane transport	6	ABC transporters
Signal transduction	84	Wnt signalling pathway
Signaling molecules and interaction	8	ECM-receptor interaction
Cellular processes		
Transport and catabolism	176	Lysosome
Organismal systems		
Immune system	8	Natural killer cell mediated cytotoxicity
Endocrine system	23	Progesterone-mediated oocyte maturation
Development	2	Dorso-ventral axis formation

Table 5. Functions of putative proteins inferred from the transcriptome of *E. caproni*.

BRITE object	**Number of putative proteins represented (%)**
Enzymes	899
Spliceosome	206
Ribosome	193
Ribosome biogenesis	135
Ubiquitin system	132
Chromosome	139
Chaperones and folding catalysts	105
Cytoskeleton proteins	92
DNA repair and recombination proteins	91
Peptidases	84
Translation factors	80
DNA replication proteins	59
Proteasome	59
Glycosyltransferases	39
Protein kinases	35
GTP-binding proteins	32
Transcription factors	17
Secretion system proteins	17
Lipid biosynthesis proteins	14
Transporters	11
Glycan Binding Proteins	11

SNAREs	10
Cellular antigens	10
Prenyltransferases	9
CAM ligands	8
Ion Channels	4
Proteoglycans	3
Nuclear receptors	1
G Protein-Coupled Receptors	1

Table 6. Sequence homology inferred between *E. caproni* putative proteins and other organisms proteins.

Dataset	**Hits (1e-05)**	**Hits (1e-15)**	**Hits (1e-30)**
C. elegans proteins	6084	3936	2273
S. mansoni proteins	9616	7220	4708
Brugia malayi proteins	5967	3840	2141
Human proteins	7052	4820	2797
Clonorchis sinensis	10329	7640	4918
Fasciola hepatica	12777	10223	7096
Fasciola gigantica	11902	9687	6686
Opisthorchis viverrini	10544	7872	5051
Schistosoma haematobium [ORFs]	2880	1296	400

Table 7. Predicted therapeutic targets from *E. caproni* putative proteins with no homologue in the host, human and homologous to known drug targets in DrugBank.

S.No.	Cluster ID	Description from NR database	Interpro hits	DrugBank targets hits	Secreted
1	UN13060	dienelactone hydrolase family protein [*Sphingomonas* sp. S17]	IPR002925	Carboxymethylenebutenolidase	No
2	UN16933	RNA polymerase sigma factor RpoD [*Sphingomonas* sp. S17]	IPR000943 IPR007624 IPR007630 IPR013324	RNA polymerase pricipal sigma factor	No
3	UN18702	acetylglutamate kinase [*Sphingomonas* sp. S17]	IPR001048	Acetylglutamate kinase	No
4	UN24923	glu/Leu/Phe/Val dehydrogenase, dimerization domain protein [*Sphingomonas* sp. S17]	IPR006096 IPR016211	L-phenylalanine dehydrogenase	No
5	UN25736	pyruvate, phosphate dikinase [*Sphingomonas* sp. S17]	IPR002192	Pyruvate, phosphate dikinase	No
6	UN26651	glutamine synthetase, type I	IPR008147	Glutamine	No

		[*Sphingomonas* sp. S17]		synthetase	
7	UN27335	glucose-fructose oxidoreductase [*Sphingobium japonicum* UT26S]	IPR000683	Glucose-- fructose oxidoreductase	No
8	UN27705	NADH:flavin oxidoreductase / NADH oxidase family protein [*Sphingomonas* sp. S17]	IPR001155	Pentaerythritol tetranitrate reductase	No

Additional files (available on CD)

Additional file 1 (*.xls)
Title: Gene Ontology distribution of *E. caproni* proteins

Description: Interpro domains and Gene Ontology distribution across *E. caproni* proteins.

Additional file 2 (*.xls)
Title: Interaction partners of *E. caproni* proteins

Description: Sequence similarity BLAST results of *E. caproni* proteins with proteins from interaction databases.

7.2 Conclusions

Again, as in Chapter 6, we analysed the transcriptome of a novel organism and thus carried out the analysis and annotation of all proteins, although the therapeutic target discovery was based on the set of ES protens.

A total of 557236 raw reads generated using 454 sequencing, were assembled into 28577 unigenes using iAssembler. These unigenes were conceptually translated into 23296 proteins. All the proteins were functionally annotated in terms of pathways, gene ontology, followed by secretory peptides and therapeutic target prediction. 3415 (14.66%) proteins were predicted as ES proteins. We found eight unigenes that are not present in human and are similar to known drug targets. These could represent potential therapeutic targets for human echinostomiasis.

The present study provides the first detailed analysis of the transcriptome of *E. caproni* and identified some groups of molecules predicted to play key biological functions in this parasite. The analysis 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.

The successful implementation of our computational approach to a parasitic nematode (*S. stercoralis*) and a parasitic trematode (*E. caproni*) illustrate the dynamic naure of bioinformatics analysis, where new approaches such as iAssembler are constantly developed and analysis piplelines have to be easily updated to incorporate current strategies.

The transcriptome analysis decubed so far has focused on helminth parasites, but the bioinformatics protocol (Chapter 4) is generic and can be adapted easily to to extend our analysis to organisms other than helminths, as described in Chapter 8.

Chapter 8: High-throughput functional annotation and data mining of fungal genomes to identify therapeutic targets

8.1 Summary

With the advent of NGS approaches, there is a huge explosion in nucleotide and protein sequence data, especially in the area of genomics and transcriptomics. A large number of reference genomes have now been sequenced. Despite this increase in sequence data, there is a huge gap in the annotation of these newly sequenced genomes and several proteins remain unannotated, as hypothetical proteins.

Annotation and extraction of secretory proteins from the proteome using labour intensive wet-lab techniques is prohibitive. In this chapter, we developed a computational protocol for the annotation of the hypothetical proteome and prediction and analysis of secreted proteins as therapeutic targets, from genome sequencing projects.

Our current computational approach (Chapters 4 and 5), augmented by relevant computational tools such as SPAAN [228] was applied to analyse and annotate proteome data from two pathogenic fungi, *Cryptococcus gattii* and *Cryptococcus neoformans var. grubii*.

High-Throughput Functional Annotation and Data Mining of Fungal Genomes to Identify Therapeutic Targets

55

Gagan Garg and Shoba Ranganathan

Abstract

With the advent of next-generation sequencing approaches and mass spectrometry techniques, there is a huge explosion in nucleotide and protein sequence data. Despite this increase in sequence data, several proteins remain unannotated, such as hypothetical proteins. Annotation and extraction of secretory proteins from the proteome using labor-intensive wet-lab techniques is prohibitive. Computational tools can be used to provide putative functionality, prior to experimental validation. This chapter introduces a bioinformatics workflow system using the best currently available free computational tools for the annotation of hypothetical proteins and prediction and analysis of secreted proteins as therapeutic targets, applied to pathogenic fungi, *Cryptococcus gattii,* and *Cryptococcus neoformans var. grubii*.

Keywords

Annotation • Drug targets • Interproscan • Protein domains • BRITE • FASTA • KEGG • KAAS • SPAAN

Introduction

The proteome is the entire set of proteins expressed by an organism. It is a valuable tool for studying molecular function, development and progression of different life stages, and much more. With many fungal genomes sequenced or under sequencing led to the identification of whole genome and protein sequences. Usually genes are predicted by using gene prediction tools followed by prediction of coding regions. These putative proteins are annotated based on the similarity of other organisms' proteins in the same taxonomic class. Many proteins still remain unannotated using these practices. This chapter

G. Garg
Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, NSW 2109, Australia
e-mail: gagan.garg@mq.edu.au

S. Ranganathan (*)
Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, NSW 2109, Australia

Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 8 Medical Drive, Singapore 117597, Singapore
e-mail: shoba.ranganathan@mq.edu.au

V.K. Gupta et al. (eds.), *Laboratory Protocols in Fungal Biology: Current Methods in Fungal Biology*, Fungal Biology, DOI 10.1007/978-1-4614-2356-0_55,

outlines an approach to functionally annotate putative proteins in terms of pathways, gene ontology, and protein domains from recently sequenced fungal genomes, followed by secretory peptides and therapeutic target prediction.

Materials

Data

1. Hypothetical protein or nucleotide fungal sequences to be annotated are obtained from NCBI[1] or from other fungal genomes sources in FASTA format. We downloaded a total of 6,210 proteins for *Cryptococcus gattii* (Serotype B) [1] and 6,967 proteins for *Cryptococcus neoformans var. grubii* (Serotype A) [2].
2. Mouse and human proteins[2] are obtained from NCBI proteins database.
3. Known drug targets dataset[3] are obtained from DrugBank [3, 4].

Software

Protein Level Annotation

1. NCBI's ORF finder[4] or EMBOSS [5] getorf tool (for local set up) for prediction of open reading frames from nucleotide sequences.
2. Fungal genomes Blast[5] to search for proteins homologous to known fungal proteins.
3. Blast2go[6] [6, 7] for gene ontology annotation of proteins.
4. KAAS[7] [8] for pathway mapping of proteins.
5. iPath2[8] [9] for graphical representation of pathway associations of proteins in the dataset.
6. Interproscan[9] [10] for protein domains mapping.
7. BLASTP[10] [11] for sequence similarity search of the query sequence against different datasets.

Secretory Proteins Prediction

1. SignalP[11] [12] for prediction of classically secreted proteins (CSP).
2. SecretomeP[12] [13] for prediction of nonclassical secreted proteins (NCSP).
3. TargetP[13] [14] for prediction of mitochondrial proteins.
4. TMHMM[14] [15] for prediction of transmembrane proteins.

Therapeutic Target Prediction

1. SPAAN [16] to predict the probability of a protein to act as adhesins. This tool is available under free academic license from the developers of the tool.

Methods

The protocol described here is a general approach to functionally annotate hypothetical proteins based on similarity searches. However, many parts of the workflow (shown in Fig. 55.1) are independent of others, so some parts can be deleted according to the requirements. We applied the protocol to download (refer to the Sect. Data) *C. gattii* and *C. grubii* proteins.

Translate Nucleotide Sequences into Protein Sequences

1. Submit nucleotide sequences (if used) in FASTA format at NCBI ORF finder web server for conceptual translation into putative

[1] http://www.ncbi.nlm.nih.gov.
[2] http://www.ncbi.nlm.nih.gov/protein.
[3] http://www.drugbank.ca/downloads.
[4] http://www.ncbi.nlm.nih.gov/projects/gorf/.
[5] http://www.ncbi.nlm.nih.gov/sutils/genom_table.cgi?organism=fungi.
[6] http://www.blast2go.org/.
[7] www.genome.jp/kaas.
[8] http://pathways.embl.de/.
[9] http://www.ebi.ac.uk/Tools/pfa/iprscan/.
[10] ftp://ftp.ncbi.nih.gov/blast/executables/blast+/LATEST/.
[11] http://www.cbs.dtu.dk/services/SignalP/.
[12] http://www.cbs.dtu.dk/services/SecretomeP/.
[13] http://www.cbs.dtu.dk/services/TargetP/.
[14] http://www.cbs.dtu.dk/services/TMHMM/.

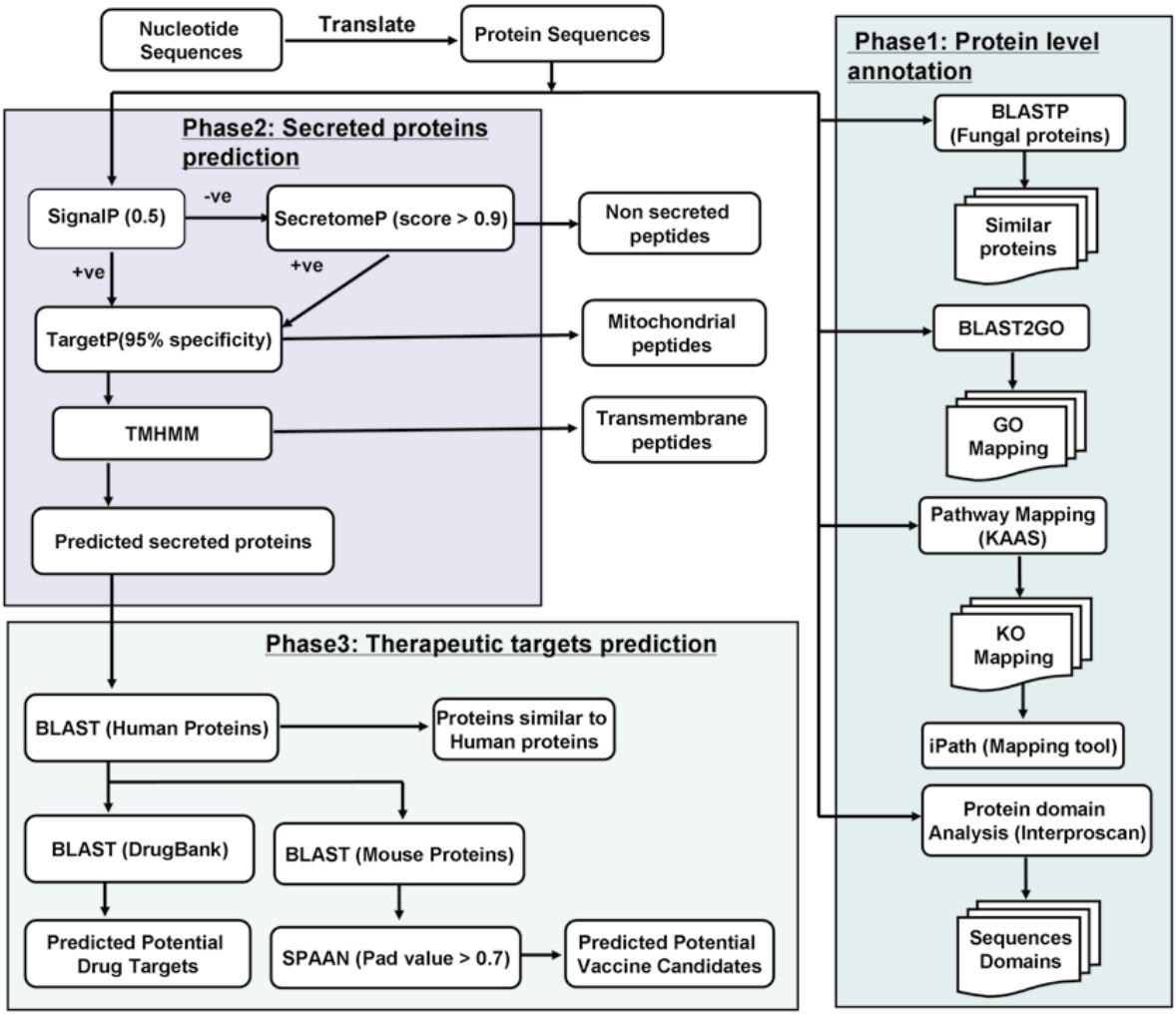


Fig. 55.1 The workflow protocol described here is a general approach to functionally annotate hypothetical proteins based on similarity searches. Many parts of the workflow are independent of others, so some parts can be deleted according to the requirements

proteins from predicted open reading frames. Use standard code (translation table 1) or the alternative yeast nuclear code (translation table 12), depending on the fungal organism. Consider the minimum length of 100 bases for nucleotide sequences translation to get reliable predictions in the following steps. For local setup, install EMBOSS getorf package on your local machine.

After all the nucleotide sequences get translated into proteins, run them through each phase of the work flow shown in Fig. 55.1.

Protein Level Annotation

For annotation of proteins, BlastP against known fungal proteins, Blast2go, KAAS, and Interproscan are used as shown in Phase 1 of Fig. 55.1.

1. Paste protein sequence on fungal genomes blast page. Choose query and database as protein and blast program as BlastP. We found 1,278 (20.6 %) *C. gattii* proteins dissimilar to *Saccharomyces cerevisiae* proteins. 508 (7.3 %) *C. grubii* proteins were found dissimilar to Aspergillus known proteins.

2. Load protein sequences into Blast2go software for gene ontology annotation.[15] We were able to annotate 135 (2.17 %) hypothetical proteins.
3. Submit protein sequences to KAAS web server. Uncheck the nucleotide box. Choose the same organism for proteins or the closest organism from organism list box using bi-directional best hit method. KAAS will map our KEGG orthology (KO) terms along with association of proteins with pathways and BRITE objects. All the files can be downloaded on local machine. We were able to map 2,288 (36.8 %) proteins with KO terms for *C. gattii* and 2,399 (34.4 %) for *C. grubii.*
4. Submit unique KO ids from KAAS mapping to iPath for graphical representation of pathway associations in proteins sample on global pathway maps. This tool provides extensive map customization capabilities. After plotting, the map can be downloaded on local machine.
5. Submit protein sequences in FASTA format to Interproscan for protein domains mapping. This program is computationally very expensive to run locally, especially for large datasets. We were able to annotate 6,001 (96.6 %) proteins for *C. gattii* and 6,614 (94.9 %) proteins for *C. grubii* with protein domains.

Secretory Proteins Prediction

For prediction of ES proteins, a combination of four tools, SecretomeP, SignalP, TargetP, and TMHMM is used as shown in Phase 2 of Fig. 55.1. All the methods described here have been implemented previously in fungal studies [17, 18].

1. Submit protein sequence in FASTA format at SignalP 3.0 server. Use eukaryotes as an organism group for fungal proteins. SignalP is based on neural networks (NN) and Hidden Markov models (HMM). We recommend the use of both methods for reliable prediction results with standard output. Truncation field needs to be set according to the protein sequence. In the output, choose sequence having D score in SignalP-NN result and signal peptide probability in SignalP-HMM greater than 0.5 as classically secreted. These thresholds have been very well tested in other studies [19]. We were able to find 419 (6.7 %) proteins as CSP for *C. gattii* and 462 (6.6 %) for *C. grubii* using SignalP.
2. Submit protein sequences found nonsecretory in the previous step to SecretomeP 2.0 server. This server predicts nonclassical secretory proteins based on large number of amino acid features along with results of other feature prediction servers such as SignalP to obtain information on various post-translational and localization aspects of the protein. SecretomeP run SignalP as well but SignalP is run separately in the previous step because of the stringent cut offs of D score and signal peptide probability used in the protocol for reliable results. This is not the case for SignalP running inside SecretomeP. SecretomeP is based on neural network and we recommend selecting protein as nonclassical secreted where NN score is greater than equal to 0.9. We were able to find 94 (1.5 %) proteins as NCSP for *C. gattii* and 106 (1.5 %) for *C. grubii* using SecretomeP.
3. Combine protein sequences predicted as classically and nonclassically secreted and submit to TargetP 1.1 server. Select Non-plant as organism group and specificity greater than 0.95 in cut offs section. Consider a protein sequence as mitochondrial if Loc column is M in the output. Delete these proteins from the set of secretory proteins predicted in previous steps. A total of 19 (0.3 %) for *C. gattii* and 20 (0.3 %) for *C. grubii* were predicted as mitochondrial proteins.
4. Submit final proteins dataset after the deletion of mitochondrial proteins to TMHMM 2.0 server. Consider a protein sequence as non-transmembrane protein if number of predicted TMHs (transmembrane helices) is 0 or 1 in the final output. Such proteins are finally considered as secretory proteins. We consider proteins having one transmembrane helix in our final

[15]A detailed documentation of how to run Blast2go is available at http://www.blast2go.org/start_blast2go.

set of secretory proteins as these can be surface proteins having therapeutic value as potential vaccine candidates. A total of 384 (6.2 %) for *C. gattii* and 440 (6.3 %) for *C. grubii* were finally predicted as ES proteins.

In addition to the computational approach (shown in Fig. 55.1) for the prediction of ES proteins, sequence similarity search can be performed against known fungal ES proteins for the prediction of ES proteins, using BlastP.

Therapeutic Target Prediction

Lot of fungal species are pathogenic to humans. ES proteins of pathogens play a key role during pathogenic infections [20]. ES proteins predicted in Phase 2 can be checked computationally for therapeutic value.

This phase of the protocol is a tricky one. All the steps of this phase need to be performed on a local machine and command line operation is necessary. No specific web server like other parts of protocol is available. Different components are combined together for therapeutic targets prediction as shown in Phase 3 in Fig. 55.1.

1. To be a potential therapeutic target, a protein should not be present in human. To find out the secretory proteins similar to human proteins, search secretory proteins for sequence similarity against human proteins using BlastP. Sample command line for this operation is as follows:
   blastall -i query -d human.fa –m 8 –e 1e-08 -o blast.out

   Here query is the input file of protein sequences in fasta format, human.fa is the database file used for blast search, blast.out is the blast output file. Use –m 8 option to provide blast result in tabular format, which is easy to parse. E-value threshold is 1e-08. The command line shown here can be altered according to the datasets.[16]

   Proteins found dissimilar to human proteins are searched for therapeutic value in terms of drug targets and potential vaccine candidates. We found 245 (3.9 %) secreted proteins for *C. grubii* and 261 (3.8 %) for *C. gattii*, similar to human proteins.
2. For drug target prediction, human dissimilar ES proteins are searched for sequence similarity against known drug targets from DrugBank using BlastP. Use the same command line mentioned above for this operation by changing the input, database, and output files. We found six potential drug targets for *C. gattii* and four for *C. grubii* in respective secreted proteins, mappable to known drug targets.
3. To test human dissimilar ES proteins for a potential vaccine candidate, a protein should not be present in mouse along with humans because most of the vaccine candidates are tested on mouse before they are tested on humans, so proteins that are found dissimilar to humans are tested for similarity against mouse proteins. Use the same command line mentioned above for this operation changing the input, database, and output files. We found further 86 (1.4 %) for *C. gattii* and 108 (1.5 %) for *C. grubii* secreted proteins similar to human dissimilar ES proteins.

   Adhesins are cell surface proteins that are present during host pathogen invasion. These proteins play an important part in pathogenicity. Due to their important role in pathogenic infection, these proteins are good vaccine candidates.
4. To predict the probability of a protein to act as adhesins run SPAAN program according to program guidelines for proteins found dissimilar to human and mouse proteins. Consider a protein to be predicted as adhesion if Pad-value in the SPAAN output is greater than 0.7. This tool has been applied previously to fungal proteins for prediction of adhesins and adhesin-like molecules [21]. We predict 33 (0.5 %) for *C. gattii* and 35 (0.5 %) for *C. grubii* as potential vaccine candidates.

Detailed result files from our fungal protein annotation are available from http://estexplorer.biolinfo.org/fungal_annotation/

[16]Detailed description of all command line parameters of blast is available at http://www.ncbi.nlm.nih.gov/books/NBK1763/.

Notes

1. All the tools except ORF finder and iPath used in the protocol are available free to install locally under academic license. Although some of the tools available as web servers, it is recommended to install these tools locally on a Linux machine by following tool installation guidelines for big sequence datasets.
2. All the databases used for blast search locally needs to converted to blastable format before use by formatdb (provided in blast executables).
3. All the therapeutic targets predicted using this protocol are preliminary predictions which need to be further validated by additional computation analysis such as structural modeling and by experimental assays.

Acknowledgments We would like to thank Mr. Ben Herbert, for introducing us to the pathogenic Cryptococcal fungal genomes. GG acknowledges the award of Australian Postgraduate Award scholarship from Macquarie University.

References

1. *Cryptococcus gattii* Sequencing Project, Broad Institute of Harvard and MIT. http://www.broadinstitute.org/. Accessed 20 Jul 2011.
2. *Cryptococcus neoformans var. grubii* H99 Sequencing Project, Broad Institute of Harvard and MIT. http://www.broadinstitute.org/. Accessed 22 Jul 2011.
3. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D et al (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 36:D901–D906
4. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P et al (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res 34:D668–D672
5. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276–277
6. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ et al (2008) High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res 36:3420–3435
7. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–3676
8. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 35:W182–W185
9. Letunic I, Yamada T, Kanehisa M, Bork P (2008) iPath: interactive exploration of biochemical pathways and networks. Trends Biochem Sci 33:101–103
10. Zdobnov EM, Apweiler R (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17:847–848
11. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410
12. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340:783–795
13. Bendtsen JD, Jensen LJ, Blom N, Von Heijne G, Brunak S (2004) Feature-based prediction of non-classical and leaderless protein secretion. Protein Eng Des Sel 17:349–356
14. Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300:1005–1016
15. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580
16. Sachdeva G, Kumar K, Jain P, Ramachandran S (2005) SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks. Bioinformatics 21:483–491
17. Jain P, Podila G, Davis M (2008) Comparative analysis of non-classically secreted proteins in *Botrytis cinerea* and symbiotic fungus *Laccaria bicolor*. BMC Bioinformatics 9:O3
18. Choi J, Park J, Kim D, Jung K, Kang S, Lee YH (2010) Fungal secretome database: integrated platform for annotation of fungal secretomes. BMC Genomics 11:105
19. Nagaraj SH, Gasser RB, 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
20. Ranganathan S, Garg G (2009) Secretome: clues into pathogen infection and clinical applications. Genome Med 1:113
21. Upadhyay SK, Mahajan L, Ramjee S, Singh Y, Basir SF, Madan T (2009) Identification and characterization of a laminin-binding protein of *Aspergillus fumigatus*: extracellular thaumatin domain protein (AfCalAp). J Med Microbiol 58:714–722

8.2 Conclusions

Using the protein part of our computational protocol developed for transcriptome analysis, we demonstrate how different computational tools can be used together for the annotation of hypothetical proteins and prediction and analysis of secreted proteins as therapeutic targets. All the programs used in our approach are open source tools that are freely available for academic purposes. The therapeutic targets identified by using the protocol will provide a foundation for understanding the role of ES proteins in pathogenic fungi as well as for the development of novel therapeutic solutions to eradicate the clinical infections.

The approach is a generalized method which can be applied to any organism, although its main application is for organisms whose genomes are not yet sequenced, with limited functional knowledge.

Chapter 9: Conclusions and future directions

9.1 Summary

ES proteins are an important class of proteins in many organisms, spanning bacteria to human beings and the choice of new therapeutic solutions for different clinical infections, especially in the case of parasitic infections. These proteins are the subject of intense research as they are present at the host-parasite interface and act as immunoregulators to control host immune recognition for parasite survival inside the host organism. In case of parasites especially parasitic helminths, transcriptomics has been used extensively to understand the molecular basis of parasitism and for developing novel therapeutic strategies against parasitic infections using ES protein prediction. ES proteins are accredited as one of the richest sources for discovery of novel therapeutic solutions for parasitic infections in parasite biology using a range of experimental and computational methods and algorithms for the analysis.

A review of the different methods and computational tools used in secretome analysis was necessary for the development of new analysis pipeline with new and updated bioinformatic tools. Based on the review, a preliminary secretome analysis of *E. multilocularis* and *E. granulosus* ESTs was carried out by integrating the bioinformatic tool, SecretomeP, for the identification of non-classically and classically secreted proteins in the existing pipeline (EST2Secretome). This preliminary analysis resulted in the prediction of five therapeutic targets against echinococcosis and provided us a basis for further development of computational approach for secretome analysis.

With the introduction of NGS, short read sequences are generated extensively for different organisms. EST2Secretome [79], a computational prediction and annotation pipeline for ES proteins was developed by our group in 2008 to address this important area of bioinformatics research. However, this webserver was designed to handle only ESTs generated from Sanger sequencing and has the following limitations: (i) the assembly of short reads, (ii) the prediction of non-classical secretory proteins and (iii) pathway mapping using KOBAS, which contains pathways that are not regularly updated. To address these deficiencies, we developed a comprehensive secretome analysis bioinformatic workflow, capable of handling large amounts of EST as well as NGS transcriptomic data. The workflow was validated with 454 *S. ratti* transcriptome data

available from University of Liverpool. Functional annotation includes protein domain mapping, pathway mapping, gene ontology and therapeutic targets prediction.

Critical evaluation of the efficiency and usefulness of our bioinformatic workflow in analysing transcriptomic data was necessary for the validation of its annotation capabilities. Therefore, we selected the largest datasets available for helminths (ESTs for 78 species), from dbEST. The EST data were cleaned, assembled and translated into proteins. From these putative proteins, ES proteins were predicted, which were further verified by homology matching to our in-house dataset of experimentally determined parasitic helminth ES proteins. All the ES proteins were functionally annotated in terms of similarity to other known proteins, protein families, domains and biochemical pathways, followed by therapeutic targets prediction. All the EST clusters (unigenes) and excretory/secretory proteins datasets generated as a part of the analysis are freely available to scientific community. A BLAST server has also been implemented to map sequence similarity of submitted proteins against predicted helminth ES proteins. All the services are freely available to the scientific community at http://estexplorer.biolinfo.org/hsd

We have worked collaboratively with an experimental parasitology group and applied our computational approach for the analysis of transcriptome data from parasitic helminths, causing human diseases, strongyloidiasis and echinostomiasis In the first study, 253,266 raw sequence reads were generated from third larval stage (L3i) of *S. stercoralis* and a detailed transcriptomic analysis was carried out. This was the first reported transcriptome dataset of the infective third larval stage of *S. stercoralis* using 454 sequencing coupled to our bioinformatic analyses. Besides ES protein prediction, we carried out functional annotation of all putative proteins translated from 11250 contiguous sequences, of which most were novel. Comparisons of the transcriptome of *S. stercoralis* with those of other nematodes revealed similarities in the transcription of molecules inferred to have key roles in parasite-host interactions. Overall, this analysis elucidated both relatively conserved and potentially novel genes that should assist researchers to understand biology of *S. stercoralis*, and could lead to develop novel intervention strategies.

In the second study, the first transcriptome generated for the adult stage of *E. caproni* was analysed. More than 0.5 million raw reads generated by 454 sequencing, were assembled into 28577 contiguous sequences. Our bioinformatic workflow was employed to generate 23296 putative proteins. 3415 putative excretory/secretory proteins were compiled

including non-classical secretory proteins. Potential therapeutic targets were also identified. Overall, the large-scale analysis paved the way for future research focussing on disease and molecular biology of *E. caproni,* along with studies to identify novel disease control strategies.

The identification and analysis of ES proteins was carried out as a part of transcriptome analyses in parasites. With the boom in sequencing technologies, sequencing data has been generated on a large scale, especially in the area of genomics and transcriptomics. Despite this increase in sequence data, several proteins remain unannotated, as hypothetical or unknown proteins. To fill this gap, we have developed a bioinformatics workflow system using the best currently available free computational tools for the annotation of hypothetical proteins and the prediction and analysis of secreted proteins as therapeutic targets. This protocol was applied to pathogenic fungi, *C. gattii* and *C. neoformans var. grubii*. The approach can be applied to functionally annotate putative proteins in terms of pathways, gene ontology and protein domains from other recently sequenced fungal genomes, followed by secretory peptides and therapeutic target prediction. This application of our computational approach to organisms other than helminths illustrates the generic nature of the bioinformatic strategies developed in this thesis.

Overall, by developing computational workflows to analyse large scale transcriptomic or proteomic data, we have enabled experimental biologists to select specific genes and/or proteins to carry out detailed and directed functional assays for elucidating the molecular complexities of host-parasite interactions.

9.2 Significance and contributions

The study and analysis of secretome data was primarily reviewed, based on the currently available methods and pipelines and written up in the form of a review article. Based on this review and preliminary data analysis, we have developed an *in silico* approach for secretome analysis approach using transcriptomic data. The approach is implemented to analyse the large scale EST data of helminths available from dbEST and a database (Helminth Secretome Database, HSD) has been developed, with the data freely available to the scientific community. HSD provides a comprehensive workflow system for EST and NGS ES protein data management and analysis. The *in silico* secretome analysis approach lead to the identification of several important proteins and therapeutic targets during 454 transcriptome analysis of parasitic nematode *S. stercoralis* and parasitic trematode *E.*

caproni. Finally, we extend our workflow to *C. gattii* and *C. neoformans var. grubii* (pathogenic fungi) for the applicability the approach to protein or nucleotide sequence data from newly sequenced parasite genomes.

Overall the large-scale analysis approach developed will lead to rapid annotation, better understanding of parasite biology and development of novel drugs or vaccines for parasite intervention and control.

9.3 Future directions

Based on the outcomes, our computational approach (Chapter 4) for secretome analysis can be logically extended to focus on the analysis of pathogens other than helminths (as shown in Chapter 8) for the annotation and therapeutic targets prediction from protein or nucleotide sequences. The workflow can be extended for microarray analysis; new tools could be incorporated for SNP discovery like autoSNP [229] and SNPServer [230] and to construct a relational database like HSD (Chapter 5) to enable the efficient mining of the identified polymorphisms. Other areas to explore would be the differential gene expression analysis across different cDNA libraries of an organism from different cell types or environmental conditions.

With NGS technologies, new reference genomes are generated with an unprecedented rate, *de novo* transcriptome assembly part could be coupled with alignment of raw sequence reads to their cognate or related genomic sequences for more reliable assembly. This goal can be achieved by the inclusion of short read aligner tools like Bowtie [231] and MAQ [232]. This will lead to alternative splicing, gene discovery projects and comparative genomics studies in neglected organisms like parasitic helminths. All the transcriptomic data sets analysed as a part of this thesis were from single stage of infection. With the low cost sequencing by NGS platforms, it is possible to capture transcriptomics data for multiple infection stages at the same time. This will lead to quantitative analysis of trancriptomics data. Using transcriptomic data from different life stages will allow us to infer quantitative expression values to carry out the analysis further.

On the other hand, novel therapeutic targets for parasite control and intervention have been reported in the analyses described in this thesis. Importantly, numerous putative ES proteins containing helminth specific domains and their absence from the respective host are critical requirements for novel therapeutic targets. Large-scale experimental validation

is necessary to further select a few molecules from this list for complete characterization. Results from experimental assays will further enable the determination of threshold parameters for computational transcriptome data analysis and the validation of the computational protocols.

References

1. Skach WR: **The expanding role of the ER translocon in membrane protein folding**. *J Cell Biol* 2007, **179**(7):1333-1335.
2. Maizels RM, Yazdanbakhsh M: **Immune regulation by helminth parasites: cellular and molecular mechanisms**. *Nat Rev Immunol* 2003, **3**(9):733-744.
3. Frith MC, Pheasant M, Mattick JS: **The amazing complexity of the human transcriptome**. *Eur J Hum Genet* 2005, **13**(8):894-897.
4. 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***. *Mol Biochem Parasitol* 2005, **142**(1):32-46.
5. 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.
6. 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.
7. McCombie WR, Adams MD, Kelley JM, FitzGerald MG, Utterback TR, Khan M, Dubnick M, Kerlavage AR, Venter JC, Fields C: ***Caenorhabditis elegans* expressed sequence tags identify gene families and potential disease gene homologues**. *Nat Genet* 1992, **1**(2):124-131.
8. Bouck A, Vision T: **The molecular ecologist's guide to expressed sequence tags**. *Mol Ecol* 2007, **16**(5):907-924.
9. Dong Q, Kroiss L, Oakley FD, Wang BB, Brendel V: **Comparative EST analyses in plant systems**. *Methods Enzymol* 2005, **395**:400-418.
10. Jongeneel CV: **Searching the expressed sequence tag (EST) databases: panning for genes**. *Brief Bioinform* 2000, **1**(1):76-92.
11. Bourdon V, Naef F, Rao PH, Reuter V, Mok SC, Bosl GJ, Koul S, Murty VV, Kucherlapati RS, Chaganti RS: **Genomic and expression analysis of the 12p11-p12**

amplicon using EST arrays identifies two novel amplified and overexpressed genes. *Cancer Res* 2002, **62**(21):6218-6223.

12. Verdun RE, Di Paolo N, Urmenyi TP, Rondinelli E, Frasch AC, Sanchez DO: **Gene discovery through expressed sequence Tag sequencing in *Trypanosoma cruzi***. *Infect Immun* 1998, **66**(11):5393-5398.
13. Santos TM, Johnston DA, Azevedo V, Ridgers IL, Martinez MF, Marotta GB, Santos RL, Fonseca SJ, Ortega JM, Rabelo EM *et al*: **Analysis of the gene expression profile of *Schistosoma mansoni* cercariae using the expressed sequence tag approach**. *Mol Biochem Parasitol* 1999, **103**(1):79-97.
14. 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.
15. Boguski MS, Lowe TM, Tolstoshev CM: **dbEST--database for "expressed sequence tags"**. *Nat Genet* 1993, **4**(4):332-333.
16. Crick F: **Central dogma of molecular biology**. *Nature* 1970, **227**(5258):561-563.
17. Bentley DR: **Whole-genome re-sequencing**. *Curr Opin Genet Dev* 2006, **16**(6):545-552.
18. 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**(7057):376-380.
19. Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, LeProust EM, Peck BJ, Emig CJ, Dahl F *et al*: **Multiplex amplification of large sets of human exons**. *Nat Methods* 2007, **4**(11):931-936.
20. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM: **Accurate multiplex polony sequencing of an evolved bacterial genome**. *Science* 2005, **309**(5741):1728-1732.
21. **The Polonator** (http://www.polonator.org/index.htm)
22. **Pacific Biosciences** (http://www.pacificbiosciences.com/)
23. **The HeliScope Single Molecule Sequencer** (http://www.helicosbio.com)
24. **Ion torrent semiconductor technology** (http://www.iontorrent.com/technology/)
25. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM: **Accurate multiplex polony sequencing of an evolved bacterial genome**. *Science* 2005, **309**(5741):1728-1732.

26. Poinar HN, Schwarz C, Qi J, Shapiro B, Macphee RD, Buigues B, Tikhonov A, Huson DH, Tomsho LP, Auch A *et al*: **Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA**. *Science* 2006, **311**(5759):392-394.
27. Goldberg SM, Johnson J, Busam D, Feldblyum T, Ferriera S, Friedman R, Halpern A, Khouri H, Kravitz SA, Lauro FM *et al*: **A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes**. *Proc Natl Acad Sci U S A* 2006, **103**(30):11240-11245.
28. Mardis ER: **Next-generation DNA sequencing methods**. *Annu Rev Genomics Hum Genet* 2008, **9**:387-402.
29. Nyren P, Lundin A: **Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis**. *Anal Biochem* 1985, **151**(2):504-509.
30. 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 Negl Trop Dis* 2010, **4:**e719.
31. 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.** *Biotechnol Adv* 2010, **28:**222-231.
32. Martin J, Abubucker S, Heizer E, Taylor CM, Mitreva M: **Nematode.net update 2011: addition of data sets and tools featuring next-generation sequencing data**. *Nucleic Acids Res* 2012, **40**(Database issue):D720-728.
33. Adessi C, Matton G, Ayala G, Turcatti G, Mermod JJ, Mayer P, Kawashima E: **Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms**. *Nucleic Acids Res* 2000, **28**(20):E87.
34. Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G: **BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies**. *Nucleic Acids Res* 2006, **34**(3):e22.
35. 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**(7218):53-59.
36. Young ND, Jex AR, Cantacessi C, Hall RS, Campbell BE, Spithill TW, Tangkawattana S, Tangkawattana P, Laha T, Gasser RB: **A portrait of the transcriptome of the neglected trematode, *Fasciola gigantica*--biological and biotechnological implications.** *PLoS Negl Trop Dis* 2011, **5:**e1004.

37. Jex AR, Liu S, Li B, Young ND, Hall RS, Li Y, Yang L, Zeng N, Xu X, Xiong Z *et al*: ***Ascaris suum*** **draft genome**. *Nature* 2011, **479**(7374):529-533.
38. Young ND, Jex AR, Li B, Liu S, Yang L, Xiong Z, Li Y, Cantacessi C, Hall RS, Xu X, et al: **Whole-genome sequence of *Schistosoma haematobium*.** *Nat Genet*, **44:**221-225
39. **SOLiD Sequencing applications** Available at: http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/publications-literature.printable.html
40. Marguerat S, Wilhelm BT, Bahler J: **Next-generation sequencing: applications beyond genomes**. *Biochem Soc Trans* 2008, **36**(Pt 5):1091-1096.
41. Wold B, Myers RM: **Sequence census methods for functional genomics**. *Nat Methods* 2008, **5**(1):19-21.
42. 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.
43. Bonin-Debs AL, Boche I, Gille H, Brinkmann U: **Development of secreted proteins as biotherapeutic agents**. *Expert Opin Biol Ther* 2004, **4**(4):551-558.
44. Huxley-Jones J, Foord SM, Barnes MR: **Drug discovery in the extracellular matrix**. *Drug Discov Today* 2008, **13**(15-16):685-694.
45. Bellafiore S, Shen Z, Rosso MN, Abad P, Shih P, Briggs SP: **Direct identification of the *Meloidogyne incognita* secretome reveals proteins with host cell reprogramming potential**. *PLoS Pathog* 2008, **4**(10):e1000192.
46. 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.
47. 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.
48. Loei H, Tan HT, Lim TK, Lim KH, So JB, Yeoh KG, Chung MC: **Mining the gastric cancer secretome: identification of GRN as a potential diagnostic marker for early gastric cancer**. *J Proteome Res* 2012, **11**(3):1759-1772.
49. Gronborg M, Kristiansen TZ, Iwahori A, Chang R, Reddy R, Sato N, Molina H, Jensen ON, Hruban RH, Goggins MG *et al*: **Biomarker discovery from pancreatic**

cancer secretome using a differential proteomic approach. *Mol Cell Proteomics* 2006, **5**(1):157-171.

50. Tjalsma H, Bolhuis A, Jongbloed JD, Bron S, van Dijl JM: **Signal peptide-dependent protein transport in *Bacillus subtilis*: a genome-based survey of the secretome**. *Microbiol Mol Biol Rev* 2000, **64**(3):515-547.
51. Dombkowski AA, Cukovic D, Novak RF: **Secretome analysis of microarray data reveals extracellular events associated with proliferative potential in a cell line model of breast disease**. *Cancer Lett* 2006, **241**(1):49-58.
52. Welsh JB, Sapinoso LM, Kern SG, Brown DA, Liu T, Bauskin AR, Ward RL, Hawkins NJ, Quinn DI, Russell PJ *et al*: **Large-scale delineation of secreted protein biomarkers overexpressed in cancer tissue and serum**. *Proc Natl Acad Sci U S A* 2003, **100**(6):3410-3415.
53. 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**(5845):1756-1760.
54. Berriman M, Haas BJ, LoVerde PT, Wilson RA, Dillon GP, Cerqueira GC, Mashiyama ST, Al-Lazikani B, Andrade LF, Ashton PD *et al*: **The genome of the blood fluke *Schistosoma mansoni***. *Nature* 2009, **460**(7253):352-358.
55. **The *Schistosoma japonicum* genome reveals features of host-parasite interplay**. *Nature* 2009, **460**(7253):345-351.
56. Feng L, Reeves PR, Lan R, Ren Y, Gao C, Zhou Z, Cheng J, Wang W, Wang J, Qian W *et al*: **A recalibrated molecular clock and independent origins for the cholera pandemic clones**. *PLoS One* 2008, **3**(12):e4053.
57. Nickel W: **The mystery of nonclassical protein secretion. A current view on cargo proteins and potential export routes**. *Eur J Biochem* 2003, **270**(10):2109-2119.
58. Gronborg M, Kristiansen TZ, Iwahori A, Chang R, Reddy R, Sato N, Molina H, Jensen ON, Hruban RH, Goggins MG *et al*: **Biomarker discovery from pancreatic cancer secretome using a differential proteomic approach**. *Mol Cell Proteomics* 2006, **5**(1):157-171.
59. Khwaja FW, Svoboda P, Reed M, Pohl J, Pyrzynska B, Van Meir EG: **Proteomic identification of the wt-p53-regulated tumor cell secretome**. *Oncogene* 2006, **25**(58):7650-7661.

60. Bumann D, Aksu S, Wendland M, Janek K, Zimny-Arndt U, Sabarth N, Meyer TF, Jungblut PR: **Proteome analysis of secreted proteins of the gastric pathogen *Helicobacter pylori***. *Infect Immun* 2002, **70**(7):3396-3403.
61. Marouga R, David S, Hawkins E: **The development of the DIGE system: 2D fluorescence difference gel analysis technology**. *Anal Bioanal Chem* 2005, **382**(3):669-678.
62. Lilley KS, Friedman DB: **All about DIGE: quantification technology for differential-display 2D-gel proteomics**. *Expert Rev Proteomics* 2004, **1**(4):401-409.
63. Volmer MW, Stuhler K, Zapatka M, Schoneck A, Klein-Scory S, Schmiegel W, Meyer HE, Schwarte-Waldhoff I: **Differential proteome analysis of conditioned media to detect Smad4 regulated secreted biomarkers in colon cancer**. *Proteomics* 2005, **5**(10):2587-2601.
64. Liu H, Lin D, Yates JR, 3rd: **Multidimensional separations for protein/peptide analysis in the post-genomic era**. *Biotechniques* 2002, **32**(4):898, 900, 902 passim.
65. Kislinger T, Gramolini AO, MacLennan DH, Emili A: **Multidimensional protein identification technology (MudPIT): technical overview of a profiling method optimized for the comprehensive proteomic investigation of normal and diseased heart tissue**. *J Am Soc Mass Spectrom* 2005, **16**(8):1207-1220
66. Washburn MP: **Utilisation of proteomics datasets generated *via* multidimensional protein identification technology (MudPIT)**. *Brief Funct Genomic Proteomic* 2004, **3**(3):280-286.
67. Mbeunkui F, Fodstad O, Pannell LK: **Secretory protein enrichment and analysis: an optimized approach applied on cancer cell lines using 2D LC-MS/MS**. *J Proteome Res* 2006, **5**(4):899-906.
68. Sardana G, Marshall J, Diamandis EP: **Discovery of candidate tumor markers for prostate cancer *via* proteomic analysis of cell culture-conditioned medium**. *Clin Chem* 2007, **53**(3):429-437.
69. Washburn MP, Ulaszek RR, Yates JR, 3rd: **Reproducibility of quantitative proteomic analyses of complex biological mixtures by multidimensional protein identification technology**. *Anal Chem* 2003, **75**(19):5054-5061.
70. Ivakhno S, Kornelyuk A: **Quantitative proteomics and its applications for systems biology**. *Biochemistry (Mosc)* 2006, **71**(10):1060-1072.
71. Fenselau C: **A review of quantitative methods for proteomic studies**. *J Chromatogr B Analyt Technol Biomed Life Sci* 2007, **855**(1):14-20.

72. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R: **Quantitative analysis of complex protein mixtures using isotope-coded affinity tags**. *Nat Biotechnol* 1999, **17**(10):994-999.
73. Cuomo A, Moretti S, Minucci S, Bonaldi T: **SILAC-based proteomic analysis to dissect the "histone modification signature" of human breast cancer cells**. *Amino Acids* 2011, **41**(2):387-399.
74. Zieske LR: **A perspective on the use of iTRAQ reagent technology for protein complex and profiling studies**. *J Exp Bot* 2006, **57**(7):1501-1508.
75. Silverman JM, Chan SK, Robinson DP, Dwyer DM, Nandan D, Foster LJ, Reiner NE: **Proteomic analysis of the secretome of *Leishmania donovani***. *Genome Biol* 2008, **9**(2):R35.
76. Higa LM, Caruso MB, Canellas F, Soares MR, Oliveira-Carvalho AL, Chapeaurouge DA, Almeida PM, Perales J, Zingali RB, Da Poian AT: **Secretome of HepG2 cells infected with dengue virus: implications for pathogenesis**. *Biochim Biophys Acta* 2008, **1784**(11):1607-1616.
77. Michalski A, Cox J, Mann M: **More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS**. *J Proteome Res* 2011, **10**(4):1785-1793.
78. 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:**1891-1907.
79. 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.
80. Bendtsen JD, Nielsen H, von Heijne G, Brunak S: **Improved prediction of signal peptides: SignalP 3.0.** *J Mol Biol* 2004, **340:**783-795.
81. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes**. J Mol Biol 2001, 305:567-580.
82. Brenner S: **The genetics of *Caenorhabditis elegans***. *Genetics* 1974, **77**(1):71-94.
83. Nagaraj SH, Gasser RB, Ranganathan S: **A hitchhiker's guide to expressed sequence tag (EST) analysis**. *Brief Bioinform* 2007, **8**(1):6-21.
84. **Univec** Available at : http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html

85. **EMvec vector database** Available at: ftp://ftp.ebi.ac.uk/pub/databases/emvec/
86. **Seqclean** Available at: http://www.tigr.org/
87. 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.
88. Chou HH, Holmes MH: **DNA sequence quality trimming and vector removal**. *Bioinformatics* 2001, **17**(12):1093-1104.
89. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-4.
90. 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 Res* 1997, **25**(17):3389-3402.
91. 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.
92. 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.
93. Huang CQ, Gasser RB, Cantacessi C, Nisbet AJ, Zhong W, Sternberg PW, Loukas A, Mulvenna J, Lin RQ, Chen N *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**(6):e246.
94. Huang X, Madan A: **CAP3: A DNA sequence assembly program.** *Genome Res* 1999, **9:**868-877.
95. 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:**1147-1159
96. **Newbler** Available at: http://454.com/products/analysis-software/index.asp
97. 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 Res* 2005, **33**(Database issue):D71-74.
98. Schuler GD: **Pieces of the puzzle: expressed sequence tags and the catalog of human genes**. *J Mol Med (Berl)* 1997, **75**(10):694-698.

99. Cariaso M, Folta P, Wagner M, Kuczmarski T, Lennon G: **IMAGEne I: clustering and ranking of I.M.A.G.E. cDNA clones corresponding to known genes**. *Bioinformatics* 1999, **15**(12):965-973.

100. Eckman BA, Aaronson JS, Borkowski JA, Bailey WJ, Elliston KO, Williamson AR, Blevins RA: **The Merck Gene Index browser: an extensible data integration system for gene finding, gene characterization and EST data mining**. *Bioinformatics* 1998, **14**(1):2-13.

101. Yee DP, Conklin D: **Automated clustering and assembly of large EST collections**. *Proc Int Conf Intell Syst Mol Biol* 1998, **6**:203-211.

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

103. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities**. *Genome Res* 1998, **8**(3):186-194.

104. Sutton GG, White O, Adams MD, Kerlavage AR: **TIGR Assembler: A new tool for assembling large shotgun sequencing projects.** *Genome Science and Technology* 1995 Jan 01; **1**(1): 9-19.

105. Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J: **An optimized protocol for analysis of EST sequences**. *Nucleic Acids Res* 2000, **28**(18):3657-3665.

106. Pop M: **DNA sequence assembly algorithms,** in: McGraw-Hill (Ed.), McGraw-Hill 2006 Yearbook of Science and Technology, *McGraw-Hill*, New York, 2005.

107. Sutton G, Dew I: **Shotgun Fragment Assembly**, in: I. Rigoutsos, G. Stephanopoulos (Eds.), Systems Biology: *Genomics*, Oxford University Press, New York, 2007, pp. 79–117.

108. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mobarry CM, Reinert KH, Remington KA *et al*: **A whole-genome assembly of Drosophila**. *Science* 2000, **287**(5461):2196-2204.

109. 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 Res* 2010, **38:**e171

110. 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.

111. Cantacessi C, Gasser RB, Strube C, Schnieder T, Jex AR, Hall RS, Campbell BE, Young ND, Ranganathan S, Sternberg PW, Mitreva M: **Deep insights into *Dictyocaulus viviparus* transcriptomes provides unique prospects for new drug targets and disease intervention.** *Biotechnol Adv* 2011, **29:**261-271
112. Kumar S, Blaxter ML: **Comparing *de novo* assemblers for 454 transcriptome data.** *BMC Genomics* 2010, **11:**571.
113. 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.
114. Pevzner PA, Tang H, Waterman MS: **An Eulerian path approach to DNA fragment assembly**. *Proc Natl Acad Sci U S A* 2001, **98**(17):9748-9753.
115. Li R, Li Y, Kristiansen K, Wang J: **SOAP: short oligonucleotide alignment program**. *Bioinformatics* 2008, **24**(5):713-714.
116. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I: **ABySS: a parallel assembler for short read sequence data**. *Genome Res* 2009, **19**(6):1117-1123.
117. Pevzner PA, Tang H, Waterman MS: **An Eulerian path approach to DNA fragment assembly**. *Proc Natl Acad Sci U S A* 2001, **98**(17):9748-9753.
118. Zerbino DR, Birney E: **Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs**. *Genome Res* 2008, **18**(5):821-829.
119. Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB: **ALLPATHS: *de novo* assembly of whole-genome shotgun microreads**. *Genome Res* 2008, **18**(5):810-820.
120. Strahm Y, Powell D, Lefevre C: **EST-PAC a web package for EST annotation and protein sequence prediction**. *Source Code Biol Med* 2006, **1**:2.
121. 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.
122. Tang Z, Choi JH, Hemmerich C, Sarangi A, Colbourne JK, Dong Q: **ESTPiper--a web-based analysis pipeline for expressed sequence tags**. *BMC Genomics* 2009, **10**:174.
123. 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.
124. Fukunishi Y, Hayashizaki Y: **Amino acid translation program for full-length cDNA sequences with frameshift errors**. *Physiol Genomics* 2001, **5**(2):81-87.

125. Min XJ, Butler G, Storms R, Tsang A: **OrfPredictor: predicting protein-coding regions in EST-derived sequences**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W677-680.
126. 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:**349-356.
127. Emanuelsson O, Nielsen H, Brunak S, von Heijne G: **Predicting subcellular localization of proteins based on their N-terminal amino acid sequence.** *J Mol Biol* 2000, **300:**1005-1016.
128. Nakai K, Horton P: **PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization**. *Trends Biochem Sci* 1999, **24**(1):34-36.
129. Jagla B, Schuchhardt J: **Adaptive encoding neural networks for the recognition of human signal peptide cleavage sites**. *Bioinformatics* 2000, **16**(3):245-250.
130. Kall L, Krogh A, Sonnhammer EL: **A combined transmembrane topology and signal peptide prediction method**. *J Mol Biol* 2004, **338**(5):1027-1036.
131. Chen Y, Yu P, Luo J, Jiang Y: **Secreted protein prediction system combining CJ-SPHMM, TMHMM, and PSORT**. *Mamm Genome* 2003, **14**(12):859-865.
132. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M *et al*: **UniProt: the Universal Protein knowledgebase**. *Nucleic Acids Res* 2004, **32**(Database issue):D115-119.
133. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank**. *Nucleic Acids Res* 2005, **33**(Database issue):D34-38.
134. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic Acids Res* 2007, **35**(Database issue):D61-65.
135. Kent WJ: **BLAT--the BLAST-like alignment tool**. *Genome Res* 2002, **12**(4):656-664.
136. Wu TD, Watanabe CK: **GMAP: a genomic mapping and alignment program for mRNA and EST sequences**. *Bioinformatics* 2005, **21**(9):1859-1875.
137. Lee BT, Tan TW, Ranganathan S: **MGAlignIt: A web service for the alignment of mRNA/EST and genomic sequences**. *Nucleic Acids Res* 2003, **31**(13):3533-3536.
138. Harris TW, Antoshechkin I, Bieri T, Blasiar D, Chan J, Chen WJ, De La Cruz N, Davis P, Duesbury M, Fang R *et al*: **WormBase: a comprehensive resource for nematode research**. *Nucleic Acids Res* 2010, **38**(Database issue):D463-467.

139. Chen N, Harris TW, Antoshechkin I, Bastiani C, Bieri T, Blasiar D, Bradnam K, Canaran P, Chan J, Chen CK *et al*: **WormBase: a comprehensive data resource for Caenorhabditis biology and genomics**. *Nucleic Acids Res* 2005, **33**(Database issue):D383-389.
140. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000**. *Nucleic Acids Res* 2000, **28**(1):45-48.
141. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI**. *Nucleic Acids Res* 2011, **39**(Database issue):D52-57.
142. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic Acids Res* 2000, **28**(1):235-242
143. 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.
144. Gilbert D: **Biomolecular interaction network database**. *Brief Bioinform* 2005, **6**(2):194-198.
145. 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 Res* 2005, **33**(Database issue):D418-424.
146. Bader GD, Hogue CW: **BIND--a data specification for storing and describing biomolecular interactions, molecular complexes and pathways**. *Bioinformatics* 2000, **16**(5):465-477.
147. 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 Res* 2003, **13**(10):2363-2371.
148. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update**. *Nucleic Acids Res* 2004, **32**(Database issue):D449-451.
149. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, Nixon J, Van Auken K, Wang X, Shi X *et al*: **The BioGRID Interaction Database: 2011 update**. *Nucleic Acids Res* 2011, **39**(Database issue):D698-704.
150. 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 Res* 2010, **38**(Database issue):D532-539.

151. 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**. *Nat Genet* 2000, **25**(1):25-29.
152. 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.
153. Groth D, Lehrach H, Hennig S: **GOblet: a platform for Gene Ontology annotation of anonymous sequence data**. *Nucleic Acids Res* 2004, **32**(Web Server issue):W313-317.
154. Zehetner G: **OntoBlast function: From sequence similarities directly to potential functional annotations by ontology terms**. *Nucleic Acids Res* 2003, **31**(13):3799-3803.
155. Beisvag V, Junge FK, Bergum H, Jolsum L, Lydersen S, Gunther CC, Ramampiaro H, Langaas M, Sandvik AK, Laegreid A: **GeneTools--application for functional annotation and statistical hypothesis testing**. *BMC Bioinformatics* 2006, **7**:470.
156. **Gene Ontology Tools** Available at http://www.geneontology.org/GO.tools.shtml
157. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P *et al*: **InterPro: an integrated documentation resource for protein families, domains and functional sites**. *Brief Bioinform* 2002, **3**(3):225-235.
158. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P *et al*: **The InterPro Database, 2003 brings increased coverage and new features**. *Nucleic Acids Res* 2003, **31**(1):315-318.
159. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L *et al*: **InterPro, progress and status in 2005**. *Nucleic Acids Res* 2005, **33**(Database issue):D201-205.
160. Sigrist CJ, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N: **PROSITE, a protein domain database for functional characterization and annotation**. *Nucleic Acids Res* 2010, **38**(Database issue):D161-166.
161. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL: **The Pfam protein families database**. *Nucleic Acids Res* 2002, **30**(1):276-280.
162. Zdobnov EM, Apweiler R: **InterProScan--an integration platform for the signature-recognition methods in InterPro.** *Bioinformatics* 2001, **17:**847-848

163. Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P *et al*: **PRINTS and its automatic supplement, prePRINTS**. *Nucleic Acids Res* 2003, **31**(1):400-402.
164. Bru C, Courcelle E, Carrere S, Beausse Y, Dalmar S, Kahn D: **The ProDom database of protein domain families: more emphasis on 3D**. *Nucleic Acids Res* 2005, **33**(Database issue):D212-215.
165. Servant F, Bru C, Carrere S, Courcelle E, Gouzy J, Peyruc D, Kahn D: **ProDom: automated clustering of homologous domains**. *Brief Bioinform* 2002, **3**(3):246-251.
166. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A: **PANTHER: a library of protein families and subfamilies indexed by function**. *Genome Res* 2003, **13**(9):2129-2141.
167. Letunic I, Doerks T, Bork P: **SMART 7: recent updates to the protein domain annotation resource**. *Nucleic Acids Res* 2012, **40**(Database issue):D302-305.
168. Schultz J, Milpetz F, Bork P, Ponting CP: **SMART, a simple modular architecture research tool: identification of signaling domains**. *Proc Natl Acad Sci U S A* 1998, **95**(11):5857-5864.
169. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, Chothia C, Gough J: **SUPERFAMILY--sophisticated comparative genomics, data mining, visualization and phylogeny**. *Nucleic Acids Res* 2009, **37**(Database issue):D380-386.
170. Haft DH, Selengut JD, White O: **The TIGRFAMs database of protein families**. *Nucleic Acids Res* 2003, **31**(1):371-373.
171. Lees J, Yeats C, Redfern O, Clegg A, Orengo C: **Gene3D: merging structure and function for a Thousand genomes**. *Nucleic Acids Res* 2010, **38**(Database issue):D296-300.
172. Lima T, Auchincloss AH, Coudert E, Keller G, Michoud K, Rivoire C, Bulliard V, de Castro E, Lachaize C, Baratin D *et al*: **HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot**. *Nucleic Acids Res* 2009, **37**(Database issue):D471-478.
173. Nikolskaya AN, Arighi CN, Huang H, Barker WC, Wu CH: **PIRSF family classification system for protein functional and evolutionary analysis**. *Evol Bioinform Online* 2006, **2**:197-209.

174. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 2000, **28:**27-30.
175. Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F, Kaipa P, Karthikeyan AS, Kothari A, Krummenacker M *et al*: **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases**. *Nucleic Acids Res* 2010, **38**(Database issue):D473-479.
176. Schomburg I, Chang A, Schomburg D: **BRENDA, enzyme data and metabolic information**. *Nucleic Acids Res* 2002, **30**(1):47-49.
177. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B *et al*: **Reactome knowledgebase of human biological pathways and processes**. *Nucleic Acids Res* 2009, **37**(Database issue):D619-622.
178. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: **KEGG for representation and analysis of molecular networks involving diseases and drugs.** *Nucleic Acids Res* 2010, **38:**D355-360.
179. 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:**D354-357.
180. 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:**W720-724.
181. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M: **KAAS: an automatic genome annotation and pathway reconstruction server.** *Nucleic Acids Res* 2007, **35:**W182-185.
182. Goesmann A, Haubrock M, Meyer F, Kalinowski J, Giegerich R: **PathFinder: reconstruction and dynamic visualization of metabolic pathways**. *Bioinformatics* 2002, **18**(1):124-129.
183. Sakharkar KR, Sakharkar MK, Chow VT: **A novel genomics approach for the identification of drug targets in pathogens, with special reference to *Pseudomonas aeruginosa***. *In Silico Biol* 2004, **4**(3):355-360.
184. Gao Z, Li H, Zhang H, Liu X, Kang L, Luo X, Zhu W, Chen K, Wang X, Jiang H: **PDTD: a web-accessible protein database for drug target identification**. *BMC Bioinformatics* 2008, **9**:104.

185. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, et al: **DrugBank 3.0: a comprehensive resource for 'omics' research on drugs.** *Nucleic Acids Res* 2011, **39:**D1035-1041.
186. Aguero F, Al-Lazikani B, Aslett M, Berriman M, Buckner FS, Campbell RK, Carmona S, Carruthers IM, Chan AW, Chen F *et al*: **Genomic-scale prioritization of drug targets: the TDR Targets database**. *Nat Rev Drug Discov* 2008, **7**(11):900-907.
187. Chen X, Ji ZL, Chen YZ: **TTD: Therapeutic Target Database**. *Nucleic Acids Res* 2002, **30**(1):412-415.
188. Mitreva M, Jasmer DP, Zarlenga DS, Wang Z, Abubucker S, Martin J, Taylor CM, Yin Y, Fulton L, Minx P *et al*: **The draft genome of the parasitic nematode *Trichinella spiralis***. *Nat Genet* 2011, **43**(3):228-235
189. Wang X, Chen W, Huang Y, Sun J, Men J, Liu H, Luo F, Guo L, Lv X, Deng C *et al*: **The draft genome of the carcinogenic human liver fluke *Clonorchis sinensis***. *Genome Biol* 2011, **12**(10):R107.
190. **Broad Insititute (**http://www.broadinstitute.org/**)**
191. **DOE Joint Genome Institute** (http://www.jgi.doe.gov/)
192. **The Institute for Genomic Research** (http://www.jcvi.org/)
193. **The Genome Institute at Washington University** (http://genome.wustl.edu/)
194. 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.
195. Moreno Y, Geary TG: **Stage- and gender-specific proteomic analysis of *Brugia malayi* excretory-secretory products**. *PLoS Negl Trop Dis* 2008, **2**(10):e326.
196. 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 Biol* 2007, **8**(4):R65.
197. Blaxter M: **Nematodes: the worm and its relatives**. *PLoS Biol* 2011, **9**(4):e1001050.
198. Blaxter ML: **Nematoda: genes, genomes and the evolution of parasitism**. *Adv Parasitol* 2003, **54**:101-195.
199. Coghlan A: **Nematode genome evolution**. *WormBook* 2005:1-15.
200. Leroy S, Duperray C, Morand S: **Flow cytometry for parasite nematode genome size measurement**. *Mol Biochem Parasitol* 2003, **128**(1):91-93.

201. Witherspoon DJ, Robertson HM: **Neutral evolution of ten types of mariner transposons in the genomes of *Caenorhabditis elegans* and *Caenorhabditis briggsae***. *J Mol Evol* 2003, **56**(6):751-769.
202. 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.
203. Viney ME: **A genetic analysis of reproduction in *Strongyloides ratti***. *Parasitology* 1994, **109 (Pt 4)**:511-515.
204. Viney ME, Matthews BE, Walliker D: **Mating in the nematode parasite *Strongyloides ratti*: proof of genetic exchange**. *Proc Biol Sci* 1993, **254**(1341):213-219.
205. Viney ME, Lok JB: **Strongyloides spp**. *WormBook* 2007:1-15.
206. Siddiqui AA, Berk SL: **Diagnosis of *Strongyloides stercoralis* infection**. *Clin Infect Dis* 2001, **33**(7):1040-1047.
207. Bethony J, Brooker S, Albonico M, Geiger SM, Loukas A, Diemert D, Hotez PJ: **Soil-transmitted helminth infections: ascariasis, trichuriasis, and hookworm**. *Lancet* 2006, **367**(9521):1521-1532.
208. Olsen A, van Lieshout L, Marti H, Polderman T, Polman K, Steinmann P, Stothard R, Thybo S, Verweij JJ, Magnussen P: **Strongyloidiasis--the most neglected of the neglected tropical diseases?** *Trans R Soc Trop Med Hyg* 2009, **103**(10):967-972.
209. Keiser PB, Nutman TB: ***Strongyloides stercoralis* in the Immunocompromised Population**. *Clin Microbiol Rev* 2004, **17**(1):208-217.
210. Vadlamudi RS, Chi DS, Krishnaswamy G: **Intestinal strongyloidiasis and hyperinfection syndrome**. *Clin Mol Allergy* 2006, **4**:8.
211. Toledo R, Fried B: **Echinostomes as experimental models for interactions between adult parasites and vertebrate hosts**. *Trends Parasitol* 2005, **21**(6):251-254.
212. **Divison of Parasitic diseases, Centers for Disease Control and Prevention, USA** (http://www.dpd.cdc.gov/dpdx/HTML/Echinostomiasis.htm)
213. **Sanger Institute** (http://www.sanger.ac.uk/)
214. **Divison of Parasitic diseases, Centers for Disease Control and Prevention, USA** (http://www.dpd.cdc.gov/dpdx/html/frames/a-f/echinococcosis/body_Echinococcosis_page1.htm)
215. ***Cryptococcus gattii* Sequencing Project**, Broad Institute of Harvard and MIT Available at http://www.broadinstitute.org/

216. ***Cryptococcus neoformans var. grubii* H99 Sequencing Project**, Broad Institute of Harvard and MIT Available at http://www.broadinstitute.org/
217. Lum G, Min XJ: **FunSecKB: the Fungal Secretome KnowledgeBase**. *Database (Oxford)* 2011, **2011**:bar001.
218. Choi J, Park J, Kim D, Jung K, Kang S, Lee YH: **Fungal secretome database: integrated platform for annotation of fungal secretomes**. *BMC Genomics* 2010, **11**:105.
219. Kwon-Chung KJ, Bennett JE: **High prevalence of *Cryptococcus neoformans* var. *gattii* in tropical and subtropical regions**. *Zentralbl Bakteriol Mikrobiol Hyg A* 1984, **257**(2):213-218.
220. Sorrell TC: ***Cryptococcus neoformans* variety *gattii***. *Med Mycol* 2001, **39**(2):155-168.
221. Byrnes EJ, 3rd, Bildfell RJ, Dearing PL, Valentine BA, Heitman J: ***Cryptococcus gattii* with bimorphic colony types in a dog in western Oregon: additional evidence for expansion of the Vancouver Island outbreak**. *J Vet Diagn Invest* 2009, **21**(1):133-136.
222. Byrnes EJ, 3rd, Bildfell RJ, Frank SA, Mitchell TG, Marr KA, Heitman J: **Molecular evidence that the range of the Vancouver Island outbreak of *Cryptococcus gattii* infection has expanded into the Pacific Northwest in the United States**. *J Infect Dis* 2009, **199**(7):1081-1086.
223. Datta K, Bartlett KH, Baer R, Byrnes E, Galanis E, Heitman J, Hoang L, Leslie MJ, MacDougall L, Magill SS *et al*: **Spread of *Cryptococcus gattii* into Pacific Northwest region of the United States**. *Emerg Infect Dis* 2009, **15**(8):1185-1191.
224. Bartlett KH, Kidd SE, Kronstad JW: **The emergence of *Cryptococcus gattii* in British Columbia and the Pacific Northwest**. *Curr Infect Dis Rep* 2008, **10**(1):58-65.
225. MacDougall L, Kidd SE, Galanis E, Mak S, Leslie MJ, Cieslak PR, Kronstad JW, Morshed MG, Bartlett KH: **Spread of *Cryptococcus gattii* in British Columbia, Canada, and detection in the Pacific Northwest, USA**. *Emerg Infect Dis* 2007, **13**(1):42-50.
226. Fraser JA, Giles SS, Wenink EC, Geunes-Boyer SG, Wright JR, Diezmann S, Allen A, Stajich JE, Dietrich FS, Perfect JR *et al*: **Same-sex mating and the origin of the Vancouver Island *Cryptococcus gattii* outbreak**. *Nature* 2005, **437**(7063):1360-1364.

227. oftus BJ, Fung E, Roncaglia P, Rowley D, Amedeo P, Bruno D, Vamathevan J, Miranda M, Anderson IJ, Fraser JA *et al*: **The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans***. *Science* 2005, **307**(5713):1321-1324.
228. Sachdeva G, Kumar K, Jain P, Ramachandran S: **SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks**. *Bioinformatics* 2005, **21**(4):483-491.
229. 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 Res* 2005, **33**(Web Server issue):W493-495.
230. 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**(3):421-422.
231. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biol* 2009, **10**(3):R25.
232. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores**. *Genome Res* 2008, **18**(11):1851-1858.