GENOMIC-BIOINFORMATIC INVESTIGATIONS OF KEY GASTROINTESTINAL PARASITES OF SOCIO-ECONOMIC IMPORTANCE AND THEIR IMPLICATIONS

Cinzia Cantacessi

DVM (Bari, Italy)

Submitted in fulfilment of the requirements of the degree of Doctor of Philosophy

April 2011

Department of Veterinary Science,

The University of Melbourne

Produced on archival quality paper
SUMMARY

The objective of the present thesis was to characterise the transcriptomes (= messenger RNAs transcribed in an organism at any one time) of selected gastrointestinal parasitic nematodes of major socioeconomic importance (orders Ascaridida and Strongylida) using sequencing and bioinformatic methods. The thesis consists of a literature review, seven chapters describing original research findings (each of which includes a detailed abstract) and a general discussion. A detailed review of the literature (Chapter 1) revealed that very little is known about the transcriptomes of many of socioeconomically important parasitic nematodes, which represents a major limitation for the development of novel strategies for treatment and control of the diseases that they cause. In addition, in spite of the availability of a vast array of bioinformatic tools for world-wide web (www)-based analyses of transcriptomic datasets, there has been no critical assessment of the performance of such tools for the analysis of sequence datasets produced by high-throughput sequencing platforms. Thus, the main aims of the present thesis were: (i) to use conventional and advanced sequencing approaches (i.e., Sanger sequencing and next-generation sequencing [NGS], respectively) and publicly available, www-based bioinformatic tools for investigating the transcriptomes of selected gastrointestinal nematodes of animal health importance, (ii) to establish a practical and user-friendly bioinformatic workflow system for the analysis of NGS datasets, and (iii) to construct a bioinformatic-phylogenetic framework for the classification of key groups of molecules encoded in the transcriptomes of free-living and parasitic nematodes as well as other eukaryotes.

Using microarray-based analysis of expressed sequence tags (ESTs) and subsequent bioinformatic analyses using a web-based bioinformatic pipeline (i.e., ESTExplorer), molecules enriched in adult males and females of the pig roundworm (Ascaris suum) (Chapter 2) and conserved between the canine hookworm (Ankylostoma caninum) and the barber’s pole worm of small ruminants (Haemonchus contortus) (Chapter 5) were investigated. The performance of ESTExplorer for the analysis of large transcriptomic datasets produced for the adult stages of the strongylid nematodes Trichostrongylus colubriformis (Chapter 3) and Nectator americanus (Chapter 6), as well as from free-living and parasitic third-stage larvae (L3s) of H. contortus (Chapter 4) was critically assessed. The results from these analyses showed that, due to the sizes of the data files (i.e., gigabytes to terabytes) generated by NGS platforms and the lengths of the sequences produced (i.e., usually ≤400 bases), www-based interfaces are not suited to the analyses of large-scale transcriptomic datasets. Therefore, a practical and user-friendly bioinformatic workflow (accessible via the www) for the analysis of NGS datasets was established using custom-written Perl, Python and Unix shell computer scripts, and tested using transcriptomic sequence datasets produced from four different developmental stages of the porcine nodule worm Oesophagostomum dentatum (Chapter 7). The results from the analyses of this and other transcriptomic datasets investigated in the present thesis (Chapters 2-7) revealed that, amongst genes encoding protein kinases, phosphatases and various types of proteases), transcripts encoding Ankylostoma-secreted proteins (= ASPs, which belong to the SCP/Tpx-1/Ag5/PR-1/Sc7 [SCP/TAPS] family of proteins) were consistently represented (Chapter 8). Members of this protein family have been characterised from a range of eukaryotic organisms (including parasitic nematodes and flatworms, arthropods, plants, snakes and humans) and are currently under investigation as potential vaccines against some gastrointestinal parasitic nematodes, such as hookworms. As the precise biological roles of SCP/TAPS proteins is still unclear, Chapter 8 investigated the relationships of SCP/TAPS from a range of eukaryotes using a phylogenetic-bioinformatic approach, in order to provide a framework for future structural and functional studies of this group of molecules (Chapter 8).
In conclusion, the present thesis has contributed significantly by providing entirely new insights into the molecular biology of key parasites of socioeconomic importance and has established a new, user-friendly bioinformatic approach for the analyses of NGS datasets (Chapter 9). The results provide a sound basis for future fundamental explorations of, for example, developmental and reproductive processes in parasitic nematodes and for studies of their proteomes, metabolomes and genomes. The results also provide a foundation for applied outcomes, such as the development of novel treatment and intervention strategies against parasitic nematodes and the diseases that they cause.
DECLARATION

The work described in the thesis was performed in the Department of Veterinary Science of the University of Melbourne between October 2007 and October 2010. The scientific work was performed solely by the author with the exception of the assistance which has been specifically acknowledged. The thesis is less than 100,000 words in length, exclusive of tables, figures, references and appendices. No part of this thesis has been submitted for any other degree or diploma.

..........................................................

Cinzia Cantacessi

April 2011
ACKNOWLEDGEMENTS

I would like to take this opportunity to express my deepest gratitude to my supervisor, Professor Robin B. Gasser for his continued support and encouragement throughout the course of my PhD program. I consider myself extremely fortunate to have been given the opportunity to undertake this study in Robin’s ‘laboratory’. His infectious enthusiasm and depth of knowledge are constantly ‘transmitted’ to both student and collaborator alike, and were extremely motivational over the past three years. I would also like to thank him for epitomizing great examples of both teacher and researcher, something I would, one day, like to become.

I would like to thank my co-supervisor, Dr Bronwyn E. Campbell, for her patience, constant help and support, both professionally and personally. Thank you for the discussions, the suggestions, the supervision and, perhaps most importantly, the great friendship!

Many thanks to Professor Domenico Otranto (University of Bari, Italy), for introducing me to the ‘world of parasites’, for his valuable advice and constant encouragement over many years.

The members of the Gasser Laboratory have contributed immensely to make my experience at The University of Melbourne an enjoyable one. I would particularly like to thank Drs Aaron R. Jex and Neil D. Young for their helpful discussions and valuable advice on the bioinformatic analyses of many datasets investigated during the course of my PhD. Special thanks to Mr Ross S. Hall, for having taught me the obscure language of command lines and computer scripts with endless patience and enthusiasm. The completion of this project would not have been possible without him. Thank you also for all the good times in the office and for sharing all your morning teases with me!

I would also like to thank all of the people in the Parasitology section at the University of Melbourne who have been great friends or have contributed towards the completion of my candidature. Special thanks to Min Hu, Nathan Bott, Abdul Jabbar, Florian Roeber, Zia-ud-Din Sandhu, Jasmin Hufschmid, Christine Andersen and Professor Ian Beveridge for their support and suggestions.

My deepest and sincerest thanks to my friends, all of whom have transformed my time in Melbourne into the most enjoyable of experiences. Thanks to Anna, Yvette, Constantine, Yvonne and Eduard Gasser and Betty, Helen and Polly Haralambous for having made a place on the other side of the world feel like home. Thanks to Aradhana Pangasa for being such a lovely and understanding friend and to Sonja Gauci and Maria Giovanna de Tommaso for the good times we spent together.

During my PhD candidature I was fortunate to receive help and support from many well-regarded scientists from around the world. Sincerer thanks to Dr F.C. Zou and Professor Xingquan Zhu (Younmnan Agricultural University, Younmnan, China) for supplying the sequence data for gender-enriched molecules in *Ascaris suum* (Chapter 2). Thanks to Dr Makedonka Mitreva (Genome Sequencing Center, St Louis, USA) for providing the sequence data for *Trichostrongylus colubriformis* (Chapter 3), *Necator americanus* (Chapter 6) and *Oesophagostomum dentatum* (Chapter 7). Special thanks to the late Paul Presidente and Jodi Zawadzki (Department of Primary Industries, Victoria) for help with the experiment described in Chapter 3. I am also grateful to Dr Bennett Datu for producing and providing the *Ancylostoma caninum* material used in Chapter 5, Professor Elida Rabelo (University of Minas Geiras, Minas Geiras, Brazil) for
supplying the *Necator americanus* material (Chapter 6) and to Dr Anja Joachim (University of Vienna, Vienna, Austria) for producing and providing the *Oesophagostomum dentatum* material used in Chapter 7. Sincerest thanks to Drs Weiwei Zhong and Boanerges Aleman-Meza (Rice University, Houston, USA) for their help and advice with the prediction of probabilistic genetic interaction networks and to Dr Garry Anderson (University of Melbourne) for advice on statistical analyses. Helpful comments from Professor Shoba Ranganathan (Macquarie University, Sydney, New South Wales), Drs Alex Loukas and Jason Mulvenna (James Cook University, Cairns, Queensland), Professor Paul Sternberg (California Institute of Technology, Pasadena, USA), Drs Maria Doyle and Stuart Ralph (Bio21 Molecular Science and Biotechnology Institute, Melbourne, Victoria), Peter Geldhof and Aline Visser (Gent University, Gent, Belgium), Alasdair Nisbet and Jacqueline Matthews (Moredun Research Institute, Scotland, UK) are gratefully acknowledged.

Personal funding was provided by the Melbourne Research Scholarship, the International Postgraduate Research Scholarship, and the Sir Clunies-Ross (2008) and Dr Sue Newton (2009) awards (University of Melbourne), for which I am extremely grateful. Research funding was also provided through grants awarded to Professors Robin Gasser (Australian Research Council, Australian-American Fulbright Commission, Australian Academy of Science), Xingquan Zhu (Natural Science Foundation of China) and Paul Sternberg (Howard Hughes Medical Institute) and Dr Makedonka Mitreva (National Human Genome Research Institute and National Institute of Health).

My most heartfelt acknowledgment must go to my partner Matt. I met you at the beginning of my PhD studies and your support, encouragement, and companionship have turned this journey into a pleasure. For all that, and for being everything I am not, you have my everlasting love.

My final and deepest thanks go to my family for their endless support and understanding. For always being there when I needed them most, and never once complaining about how infrequently I visit, they deserve far more credit than I can ever give them.

Grazie per tutto il supporto, l’incoraggiamento, le lunghe chiacchierate al telefono e la pazienza che avete sempre avuto. Grazie per aver sempre capito e compreso. Grazie per avere sempre creduto in me soprattutto quando io non ho piu’ creduto in me stessa. Vi voglio bene!
PREFACE AND DISSEMINATION OF RESEARCH FINDINGS

Scientific papers published or submitted by the author in collaboration with supervisors and other colleagues are listed in the following:

*Peer-reviewed articles published in international scientific journals:*


*Conference proceedings and seminars given:*


**Cantacessi C**, Loukas A, Campbell BE, Ong EK, Zhong W, Sternberg PW, Gasser RB (2008) Transcriptional conservation between *Ancylostoma caninum* and


# TABLE OF CONTENTS

The present thesis contains 9 chapters with a reference list, tables and figures at the end of each chapter. Chapters 2-8 describe original research findings and each of them contains specific sections and subsections.

## Chapter 1 – Literature review

1.1 Introduction .........................................................................................................................1
1.2 Taxonomy and systematics (orders Ascaridida and Strongylida) .................................2
1.3 Epidemiology, life cycle and pathogenesis ........................................................................3
   1.3.1 Nematodes within clade III (Ascaris sp.) .................................................................3
   1.3.2 Nematodes within clade V .......................................................................................4
      1.3.2.1 Nematodes within the superfamily Ancylostomatoidea ......................................4
      1.3.2.2 Nematodes within the superfamily Trichostrongylidida ......................................5
      1.3.2.3 Nematodes within the superfamily Strongyloidea .............................................6
1.4 Aspects of immunology ....................................................................................................6
1.5 Diagnosis ..........................................................................................................................7
1.6 Treatment strategies and vaccine research ......................................................................8
1.7 Conventional techniques for investigating the transcriptome of parasitic nematodes .........................................................................................................................12
1.8 Next-generation sequencing technologies ....................................................................15
1.9 Current bioinformatic tools for the analysis of EST data ..............................................17
   1.9.1 Sequence assembly .................................................................................................18
   1.9.2 Conceptual translation and annotation ................................................................. 18
   1.9.3 Similarity searches, and current online tools and databases ............................. 19
   1.9.4 The free-living nematode, Caenorhabditis elegans, as a model for parasitic nematodes ...................................................................................................... 20
1.10 Conclusions from the literature review and aims of the thesis ..............................23
1.11 References ......................................................................................................................25

## Chapter 2 – Using a worldwide-web (www)-based workflow platform for the bioinformatic analysis of abundant gender-enriched transcripts of adult Ascaris suum (Nematoda)

Abstract ................................................................................................................................. 57
2.1 Introduction ...................................................................................................................... 58
2.2 Materials and methods ................................................................................................. 59
2.3 Results and discussion ................................................................................................. 60
2.4 References ...................................................................................................................... 67
Chapter 3 – Use of a next-generation sequencing approach to elucidate the transcriptome of the economically important parasitic nematode, *Trichostrongylus colubriformis*

Abstract ............................................................................................................................................. 88
3.1 Introduction ...................................................................................................................................... 89
3.2 Materials and methods .................................................................................................................. 90
  3.2.1 Parasite material ...................................................................................................................... 90
  3.2.2 Preparation of 3’-cDNA from *Trichostrongylus colubriformis* for 454 sequencing ................................................................................................................................. 91
  3.2.3 Bioinformatic analysis of sequence data ................................................................................ 91
3.3 Results ............................................................................................................................................ 92
3.4 Discussion ....................................................................................................................................... 94
3.5 References ...................................................................................................................................... 99

Chapter 4 – The application of next-generation sequencing to explore differential transcription between free-living and pre-parasitic third-stage larvae of *Haemonchus contortus*

Abstract .............................................................................................................................................. 115
4.1 Introduction ..................................................................................................................................... 116
4.2 Materials and methods ................................................................................................................ 117
  4.2.1 Parasite material ..................................................................................................................... 117
  4.2.2 Library construction and next-generation sequencing .......................................................... 118
  4.2.3 Bioinformatic analyses ........................................................................................................... 118
  4.2.4 Identification of transcripts unique to either L3 or xL3 by in silico subtraction and verification of specificity by PCR-coupled sequencing ........................................................................... 119
  4.2.5 Probabilistic functional gene networking .............................................................................. 120
4.3 Results ........................................................................................................................................... 120
4.4 Discussion ..................................................................................................................................... 122
4.5 References ..................................................................................................................................... 129

Chapter 5 – Oligonucleotide microarray and bioinformatic analyses to explore transcriptional conservation between *Ancylostoma caninum* and *Haemonchus contortus*

Abstract .............................................................................................................................................. 143
5.1 Introduction ..................................................................................................................................... 144
5.2 Materials and methods ................................................................................................................ 145
5.3 Results .......................................................................................................................................... 148
5.4 Discussion ..................................................................................................................................... 150
5.5 References .................................................................................................................................... 155
Chapter 6 – Massively parallel sequencing and analysis of the transcriptome of the human hookworm Necator americanus

Abstract .................................................................................................................................................. 167
6.1 Introduction ........................................................................................................................................... 168
6.2 Materials and methods .......................................................................................................................... 169
  6.2.1 Parasite material ............................................................................................................................... 169
  6.2.2 RNA isolation, cDNA synthesis and 454 sequencing ....................................................................... 169
  6.2.3 Bioinformatic analyses ...................................................................................................................... 170
  6.2.4 Prediction of essentiality and drug targets ......................................................................................... 171
6.3 Results .................................................................................................................................................. 171
6.4 Discussion ........................................................................................................................................... 173
6.5 References .......................................................................................................................................... 179

Chapter 7 – A practical, bioinformatic workflow system for the analysis of large datasets generated by next-generation sequencing

Abstract ...................................................................................................................................................... 194
7.1 Introduction .......................................................................................................................................... 195
7.2 Materials and methods .......................................................................................................................... 196
  7.2.1 Sequence datasets ............................................................................................................................. 196
  7.2.2 Bioinformatic components for the construction of the workflow system ....................................... 197
7.3 Results .................................................................................................................................................. 198
  7.3.1 Assembly and detailed annotation and analyses of the Oesophagostomum dentatum datasets ............ 198
7.4 Discussion ............................................................................................................................................ 201
  7.4.1 Technical considerations ................................................................................................................... 201
  7.4.2 Biological interpretations from the annotated dataset ....................................................................... 202
  7.4.3 Concluding remarks ....................................................................................................................... 205
7.5 References .......................................................................................................................................... 206

Chapter 8 – A framework for the classification of the “SCP/TAPS” proteins and its implications

Abstract ...................................................................................................................................................... 223
8.1 Introduction .......................................................................................................................................... 224
8.2 SCP/TAPS proteins in plants: defence-related proteins ......................................................................... 226
8.3 SCP/TAPS proteins of animals: from insect allergens to cancer cells ................................................. 228
8.4 SCP/TAPS from helminthes: ‘activation-associated’ secreted proteins ................................................. 230
  8.4.1 Nematodes within clade III (including the orders Spirurida and Ascaridida) ..................................... 230
  8.4.2 Nematodes within clade IV (including the order Tylenchida) ......................................................... 231
  8.4.3 Nematodes within clade V (including the orders Strongylida and...
Rhabditida) ........................................................................................................... 232
8.4.3.1 Superfamily Ancylostomoidea (order Strongylida). 232
8.4.3.2 Superfamily Trichostrongyloidea (order Strongylida) 233
8.4.3.3 The free-living nematode, Caenorhabditis elegans (order Rhabditida) .......................................................... 234
8.4.4 Trematodes (i.e., subclass Digenea) ..................................................... 235

8.5 Phylogenetic analysis of eukaryote SCP/TAPS: a proposed classification system ........................................................................................................... 235
8.6 Concluding remarks ...................................................................................... 237
8.7 References ..................................................................................................... 240

Chapter 9 – General discussion ........................................................................... 257
References ........................................................................................................... 267
List of appendices ................................................................................................. 277
CHAPTER 1

Literature review

1.1 Introduction

Infectious diseases caused by bacteria, viruses, fungi and parasites are major causes of death and disability in animals, and have significant social and economic impacts on millions of people [1,2]. Thanks to access to an array of effective drugs and vaccines, infectious diseases account for only one of ten deaths in the richest countries of the world. However, in developing countries, six of ten people still die due to complications associated with these diseases [1,2]. Parasitic nematodes of the gastrointestinal tracts of humans and livestock are of major socioeconomic significance worldwide [3-7]. Amongst these nematodes, the soil-transmitted helminths (STHs) _Ancylostoma duodenale_, _Necator americanus_, _Ascaris_ spp. and _Trichuris trichiura_ are estimated to infect almost one-sixth of the global population [8,9], whereas gastrointestinal parasites of livestock (including _Haemonchus contortus_, _Ostertagia ostertagi_ and _Trichostrongylus_ spp.) cause economic losses estimated at billions of dollars per annum, due to poor productivity, failure to thrive, repeated anthelmintic treatment and deaths [4,10,11]. In addition to their socioeconomic impact, genetic resistance against the main classes of anthelmintics in a range of nematodes of livestock [12] has stimulated research towards developing alternative intervention and control strategies against these parasites. Despite the wealth of information on aspects of parasite taxonomy and systematics, biology, epidemiology, immunology and anthelmintics [13,14], there is a paucity of knowledge of the molecular mechanisms that govern essential biological processes in parasitic nematodes. Gaining an improved understanding of the molecular biology of these organisms offers a possible avenue to assist the discovery and development of novel methods of treatment and control.

Advances in genomic and bioinformatic technologies are now providing the opportunity to explore the molecular basis of fundamental aspects of the biology of parasitic nematodes, such as developmental and reproductive processes. In particular, studies of the transcriptomes of parasites (= the study of expressed sequence tags [ESTs]; i.e., short sequences of complementary DNAs [cDNAs] synthesized from the messenger RNA [mRNA]; [15] have become instrumental in various areas of molecular parasitology, including gene discovery and characterization, and for gaining insights into aspects of gene expression, regulation and function [16,17]. The purpose of this Chapter was to review relevant information of aspects of the molecular biology of
gastrointestinal nematodes of socio-economic significance (i.e., selected species of the orders Ascaridida and Strongylida, including the ‘roundworm’ *Ascaris suum*, the ‘hookworms’ *Ancylostoma caninum* and *Necator americanus*, the ‘barber’s pole worm’, *Haemonchus contortus*, the ‘black scour worm’, *Trichostrongylus colubriformis* and the ‘nodule worm’ *Oesophagostomum dentatum*), to describe a range of molecular approaches for transcriptomic studies of parasites, and to indicate the prospects and implications of investigating fundamental and applied aspects of their biology at the molecular level. Based on the conclusions from the literature review, the aims of the thesis were formulated.

### 1.2 Taxonomy and systematics (orders Ascaridida and Strongylida)

The phylum Nematoda (nematodes or roundworms) is one of the most diverse phyla in the kingdom Animalia and includes > 28,000 species, of which > 16,000 are parasites of animals or plants [13,18]. According to the classification proposed by Chitwood [19], the phylum Nematoda consists of two classes, the Adenophorea and the Secernentea [19,20]. Within the latter class, the order Rhabditida includes both free-living and entomopathogenic species of nematodes, whereas the order Tylenchida includes species of parasitic nematodes of invertebrates, particularly insects, mites and leeches, and plants [13,20]. Conversely, species within the orders Ascaridida, Oxyurida, Spirurida and Strongylida are parasitic in humans and vertebrate animals [13]. Within the Ascaridida, the superfamily Ascaridoidea consists of medium to large-sized nematodes with three lips [21,22], which parasitize a range of vertebrate hosts, such as equids (e.g., *Parascaris equorum*), pigs (*As. suum*), humans (*Ascaris* sp.) and rodents (*Baylisascaris laevis*) [13,20]. According to a molecular phylogenetic analysis of sequences of the small subunit of the ribosomal DNA (rDNA), nematodes of the order Ascaridida are proposed to belong to ‘clade III’ of the phylum Nematoda [23] (Fig. 1.1).

Within the order Strongylida, the superfamily Ancylostomoidea includes species of blood-feeding nematodes (called the ‘hookworms’) which inhabit the small intestine of their vertebrate hosts and are characterized by large, globular buccal capsules, which enable them to attach to the intestinal wall [13,20]. The superfamily Strongyloidea includes, amongst others, gastrointestinal parasites of equines (family Strongylidae), the ‘nodular worms’ of pigs (Chabertiidae) and the ‘gapeworm’ of birds (Syngamidae) [13,20]. Members of this superfamily are characterized by complex buccal capsules, often with a series of leaf-like structures on the border of the labial region (= *corona radiata*) [24]. Conversely, the buccal capsule of species of parasitic nematodes of the superfamily Trichostrongyloidea is absent or greatly reduced, and lips and *corona radiata* are vestigial or absent [25-27]. Members of the superfamily
Trichostrongyloidea are parasitic in a broad range of mammalian species, particularly wild and domestic ruminants [13,20]. According to the classification proposed by Blaxter et al. [23], parasitic nematodes of the order Strongylida, together with free-living nematodes of the sub-order Rhabditina (e.g., Caenorhabditis elegans) and order Diplogasterida (e.g., Pristionchus pacificus), belong to ‘clade V’ of the phylum Nematoda [23] (Fig. 1.1).

1.3 Epidemiology, life cycle and pathogenesis

1.3.1 Nematodes within clade III (Ascaris sp.)

Ascaris lumbricoides (infecting humans) and As. suum (infecting pigs) are large (usually 10-15 cm) roundworms found in the intestine of their vertebrate hosts. Although these two parasites are often considered as separate species, there has been continued controversy as to whether they represent distinct species [28-30]. It has been estimated that ~1.4 billion people in 150 countries are infected with Ascaris sp., with high prevalences (~25-90%) recorded in China, Southeast Asia, Africa and Latin America [3,28,31-34]. The life cycle of Ascaris sp. is direct, with thick-shelled eggs (50-75 X 35-60 µm in size; [20,35,36]) being passed in the faeces of the infected host; under suitable environmental conditions (i.e., 28-33 °C; [37]), the first stage larva (L1) develops within the egg to then moult (via the second-stage larva, L2) to the infective, third-stage larva (L3) within 2-6 weeks [38]. Under favourable conditions, larvated (= embryo-onated) eggs of Ascaris sp. can remain viable in the environment for up to five years [39,40]. These eggs are ingested by the host and hatch in the intestine. Larvae invade the intestinal wall and migrate to the liver and lungs (‘hepato-pulmonary migration’) and up the trachea; they are then swallowed and reach the intestine, where they develop to adult males and females within 21-29 days [39]. Following the ingestion of eggs by the host, adult worms become mature in 50-55 days, and eggs appear in the faeces of the host after 60-62 days [20,39]. Earthworms, such as Lumbricus terrestris, may serve as paratenic hosts for Ascaris sp. [41].

Although usually asymptomatic, heavy infections by Ascaris sp. may cause severe pulmonary disease, or partial or complete obstruction of biliary and intestinal tracts [28]. The migration of Ascaris larvae through the liver of the infected host results in the development of foci of inflammation/hepatocytic necrosis and, subsequently, interstitial hepatitis, known as ‘milk spots’ or ‘white spots’, representing areas of localized fibrotic thickening of the interlobular septa of the liver [42]. The presence of numerous adult worms in the small intestinal tract of the host may lead to malnutrition, colic and/or non-specific gastrointestinal symptoms [43]; however, the
often profound lesions in the liver of pigs caused by the migration of the larvae usually result in a high level of condemnation of livers following slaughter [44].

1.3.2. Nematodes within clade V

1.3.2.1 Nematodes within the superfamily Ancylostomoidea

The hookworms *Ancylostoma duodenale* and *N. americanus* of humans are estimated to infect ~740 million people in rural areas of the tropics and subtropics [3], with the highest prevalence (~17%) recorded in areas of China and sub-Saharan Africa [3,45], and causing an estimated disease burden of 22 million disability-adjusted life years (DALYs) [46]. Whilst *N. americanus* is the most widely distributed hookworm of humans globally [47], a related species, *An. caninum*, is a cosmopolitan hookworm of the intestine of dogs and other canids [13,20]. The life cycle of these parasitic nematodes is direct, with female hookworms excreting thin-shelled eggs (50-80 X 36-42 μm in size) which are passed in the faeces of the host [20,48]. Under suitable environmental conditions (i.e., 23-33 ºC), the ‘rhabditiform’ L1s, characterized by a pointed, tapered tail, an elongated buccal cavity and an oesophagus with a valved bulb, hatch from the eggs [20,48]. The L1s feed on bacteria and, within 2 days, moult to L2s and subsequently to the L3s within 4-5 days. This latter stage retains the cuticle of the L2 (i.e., sheath) and is referred as to ‘filariform’ larva [48]. The infection occurs when the L3s penetrate the skin of the vertebrate host after cuticular shedding [49]; subsequently, larvae enter the subcutaneous tissues and migrate via the circulatory system to the heart and lungs, where they moult to fourth-stage larvae (L4s). From the lungs, the larvae migrate (via the trachea and pharynx) to the small intestine, where they develop to adult males and females within 2-7 weeks depending on the species [13,20,50,51]. The adult stages attach by their buccal capsule to the intestinal mucosa, rupture capillaries and feed on blood [52,53]. Although skin penetration is considered the main route, ingestion of L3s may also lead to infection [54]. L3s of *Ancylostoma* spp. can undergo developmental arrest (= hypobiosis) in the somatic tissues of the vertebrate host and, following activation during pregnancy, undergo transmammary transmission to the offspring [55-57].

The pathogenesis of hookworm disease is mainly a consequence of the blood loss, which is caused by tissue damage and direct ingestion of red cells by the adult worm [52]. These focal lesions are characterized by local hemorrhage, tissue cytolysis and neutrophilic immune response [52]. The clinical manifestations of the disease mainly relate to iron-deficiency anaemia, which can cause physical and mental retardation and sometimes death in children as well as maternal mortality, impaired lactation, prematurity and low birth rates [5,58,59].
1.3.2.2 Nematodes within the superfamily Trichostrongyloidea

The trichostrongyloid nematodes of small ruminants *H. contortus* and *Trichostrongylus* spp. (including *T. colubriformis*) are responsible for substantial production losses in the livestock industries worldwide [60]. *H. contortus* is the most important parasitic nematode of small ruminants in tropical and subtropical areas or summer rainfall areas, whereas *Trichostrongylus* spp. are dominant parasites in winter rainfall areas due to their ability to develop and survive at lower temperatures than *H. contortus* [61]. The life cycles of *H. contortus* and *T. colubriformis* are similar and direct, with morulated eggs (66-79 X 43-46 and 79-101 X 39-47 μm in size, respectively; [20,62,63]) being laid by females in the abomasum (*H. contortus*) or small intestine (*T. colubriformis*) of the mammalian host [62,63]. Under suitable environmental conditions (i.e., 100% humidity, 26 °C and 18 to 21 °C for *H. contortus* and *T. colubriformis*, respectively; [62,64]), L1s hatch from eggs to then develop (via the L2s) to infective L3s. The cuticle of the L2 is retained as a sheath around the L3 and protects it from desiccation [13,62,64]. Small ruminants acquire the infection by ingesting L3s from contaminated pastures. Once ingested, the L3s pass through the stomachs and undergo an ‘exsheathment’ process to then establish, via the parasitic L4s, as adult males and females in the abomasum (*H. contortus*) or small intestine (*T. colubriformis*) within ~3 weeks [13,20,62,64]. The exsheathment process is triggered by stimuli from the host and may include (depending on the species of nematode) dissolved, gaseous CO₂ and undissociated carbonic acid (*H. contortus*) or hydrochloric acid and pepsin (*T. colubriformis*) in the abomasum; the L3s respond to these stimuli by producing an exsheathment fluid which determines the detachment of the sheath from the bodies of the larvae [7,65-67].

The infection by *H. contortus* is typically accompanied by clinical signs linked to the haematophagous activity of this species of parasites. The main symptoms of acute haemonchosis are anaemia, variable degrees of oedema, lethargy, decreased liveweight gain, impaired wool and milk production as well as decreased reproductive performance, often leading to death in severely affected animals [68,69]. The pathogenesis of *T. colubriformis* infection is triggered by the presence of adult parasites in mucus-covered tunnels in the epithelial surface of the small intestine [70] and usually consists of extensive villous atrophy combined with hyperplasia of the submucosal glands, mucosal thickening and erosion as well as infiltration of lymphocytes and neutrophils into the affected areas [70-74]. Clinical signs of trichostrongylosis include malabsorption, weight loss, progressive emaciation and diarrhoea (= scouring or ‘black scour’).
1.3.2.3 Nematodes within the superfamily Strongyloidea

The superfamily Strongyloidea includes, amongst others, important parasitic nematodes of equines (e.g., ‘large strongyles’ = sub-family Strongylinae and ‘small strongyles’ or cyathostomins = sub-family Cyathostominae), ruminants (e.g., gastrointestinal nematodes of the sub-family Chabertiinae) and pigs (e.g., the ‘nodular worms’ = sub-family Oesophagostominae) [13,20]. Within the latter sub-family, *Oe. dentatum* is a common parasite of the large intestine of domestic and wild pigs, responsible for severe production losses to the livestock industry. The life cycle of nematodes of the genus *Oesophagostomum* is direct, with adult female nematodes releasing thin-shelled eggs (70-75 X 40-45 μm in size; [13,20,75,76]), which are passed in the faeces of the host into the environment, where they develop rapidly to L1s and L2s, characterized by long, attenuated tails, rhabditiform oesophagus and a small number of intestinal cells [20,75,76]. Under optimal environmental conditions (15-20 °C; [77]), the L2s develop to infective L3s within 3-5 days [13]. These larvae retain the cuticle of the L2s and are characterized by a strongyliform oesophagus and a conical tail [20,75,76]. Pigs acquire the infection by ingesting infective L3s, which exsheath in the small intestine and rapidly invade the mucosa and become encysted in transparent capsules (= ‘nodules’); within the mucosal layer, the L3s develop to L4s [78], to then emerge into the intestinal lumen; L4s migrate from the small intestine posteriorly to the caecum and colon where they develop, *via* a further moult, to adult males and females within 7-10 days [13]. Eggs appear in the faeces of the host ~20 days following the ingestion of L3s [79].

The pathogenesis of oesophagostomiasis relates to the formation of nodules in the mucosa of the intestine, which is often accompanied by mild inflammatory reactions due to the accumulation of eosinophils and neutrophils in the sites of lesions [80]. Infections are often characterized by a reduction in appetite, growth rate and feed conversion efficiency during the period of nodule formation [81]. In intense infections (e.g., following the ingestion of > 200,000 larvae), necrotic enteritis associated with diarrhoea has been observed [80].

1.4 Aspects of immunology

Several studies have attempted to define the cells and molecules implicated in the hosts’ immune responses against gastrointestinal infections by parasitic nematodes [4,44,82-96]. The primary immunological reactions stimulated by parasitic nematodes of the orders Ascaridida and Strongylida are largely dependent on the processes and mechanisms of invasion of and establishment in the host [4]. For instance, the hepato-
pulmonary migration of *As. suum* larvae is responsible for the development of a specific (IgG and IgA) serum antibody response in infected pigs [83], a marked blood eosinophilia [44] and a transient blastogenesis of peripheral blood lymphocytes [44]. In addition, the migration of *As. suum* larvae through the liver of the mammalian host causes cellular infiltrations of neutrophils, eosinophils and mononuclear cells in damaged tissues [44,87]. Similarly, migrating L3s of hookworms are known to stimulate a marked peripheral blood eosinophilia in the mammalian host, both systemic and in the lungs [97,98]. Conversely, gastrointestinal parasitic nematodes that do not undergo extensive tissue migration usually stimulate a mucosal immune response at the site of infection [99]. For instance, the primary invasion of the abomasum of small ruminants by larvae of *H. contortus* and *Trichostrongylus* spp. leads to a localised IgE-mediated immune response [99]. However, it has been observed that the infection of pigs with larvae of *Oe. dentatum* is associated with a systemic production of IgG antibodies [100-103], followed by the formation of eosinophilic cysts (= nodules), containing the larvae, within the intestinal mucosa [80].

Despite differences in immune responses stimulated by larval stages of gastrointestinal parasitic nematodes, adult stages of species of the orders Ascaridida and Strongylida appear to stimulate similar immunological reactions in their mammalian host(s), which include (i) increased production of mucus by the gastrointestinal epithelium of the host, (ii) eosinophilia and increased presence of mast cells and leucocytes in the site of infection and (iii) production of specific antibodies [4,85]. A consensus has emerged that the immunological reactions against primary infections by gastrointestinal parasitic nematodes are regulated by a T helper (Th) 2-type immune response which, in turn, determines the secretion of several types of cytokines, including IL-4, IL-5, IL-9 and IL-13 [4,88,93,95,103-105]. In contrast, immunological responses in individuals with chronic infections by gastrointestinal parasitic nematodes have been shown to be regulated predominantly by a Th1-type immune response (characterized by production of IL-2, IL-18 and interferon-γ) [90,91]. In particular, individuals infected chronically by hookworms, show a significant alteration of the immune response to helminth infections, characterised by a dysfunction of the antigen-presenting ability of dendritic cells, which results in a ‘hypo-responsiveness’ of the antigen-induced proliferation of T-lymphocytes [106].

1.5 Diagnosis

The parasitological diagnosis of infections by gastrointestinal parasitic nematodes is usually made based on the detection and identification of the parasite eggs in the faeces of the host using, for instance, the formalinethanol acetate
sedimentation, the Kato-Katz and/or the McMaster techniques. Such techniques can allow an estimation of the burden of the infection [53]. With the exception of the large and thick-shelled eggs of *Nematodirus* spp. (~152-260 X 67-120 µm; [13,20,107]), eggs of most strongylids are very similar morphologically, often making specific diagnosis of strongylid infections impossible [53,108,109]. To overcome this limitation, strongylid eggs in the faeces can be cultured to allow L3s to develop, so that L3s can be identified to the genus level [53,108,109]. Various serological and immunological methods, such as the complement fixation test, fluorescent-antibody test and enzyme-linked immunosorbent assay, have been evaluated for the diagnosis of infections with strongylids, such as hookworms [53,110-113]. However, these tests do not permit the accurate diagnosis of infections with multiple species of worms and/or reliably distinguish between current and past infections [53]. In addition, cross-reactivity can be a common limitation of serological and immunological assays [53].

Polymerase chain reaction (PCR)-based diagnostic techniques have also been established for the diagnosis of infection by parasitic nematodes [53,114-117]. Different molecular markers have been employed that allow specific and sensitive diagnoses; these markers include, amongst others, regions of the rDNA and mitochondrial DNA (mtDNA) [118,119]. Importantly, numerous studies (e.g., [117,120-138]) have demonstrated the utility of the first and second internal transcribed spacers (ITS-1 and ITS-2) of the rDNA as genetic markers for the specific identification of a range of gastrointestinal nematodes, including ascaridoïds (i.e., *Ascaris* spp.; [139]), hookworms (i.e., *N. americanus*, *A. duodenale*, *A. caninum*, *A. ceylanicum* and *A. braziliense*; [53]) and other strongylids (i.e., *Trichostrongylus* spp. and *Tenidens deminutus*; [53]).

### 1.6 Treatment strategies and vaccine research

Presently, treatment of infections by gastrointestinal ascaridoïd and strongylid nematodes relies largely on the administration of anthelmintic drugs [140]. Such drugs, which are grouped into classes on the basis of their chemical structure and mode of action, include imidazothiazoles/tetrahydropyrimidines (e.g., levamisole and pyrantel), benzimidazoles (e.g., albendazole and mebendazole) and macrocyclic lactones (e.g., ivermectin and moxidectin) [140].

The imidazothiazoles/tetrahydropyrimidines act by binding to the nicotinic acetylcholine receptors, resulting in an over-stimulation, blockage of the neuromuscular junctions and rigid paralysis of the worms. The parasites are unable to move in the intestinal tract and are eliminated by the peristaltic action of the intestines. Benzimidazole compounds are active against a range of species of parasitic nematodes.
[141]. They act by blocking the formation of the microtubular matrix by binding to the cytoskeletal protein tubulin, which is essential for various biological processes in the cell, including chromosome movement and cell division [12,140,142]. The macrocyclic lactones act by opening glutamate-gated chloride channels, increasing chloride ion conductance, leading to defects in neurotransmission and flaccid paralysis [142].

The relatively low cost, ease of administration and efficacy of anthelmintic drugs against a wide spectrum of human and animal gastrointestinal parasitic nematodes has led to their extensive use and, consequently, to the appearance of resistance [12]. Indeed, resistance against all of the major families of broad-spectrum drugs active against gastrointestinal parasites of livestock has been reported, particularly in Africa, Australia, New Zealand, Asia and South America [12,142-145]. In relation to benzimidazoles, three mutations in the gene encoding the beta-tubulin isotype 1 in *H. contortus* were proposed to be involved in the mechanism of resistance against them [146]. Although it was hypothesized that the less frequent use of anthelmintics in humans (compared with their extensive use in livestock) should negatively affect the selection of resistant populations of parasitic nematodes [147-150], some studies [151-156] have reported a reduction in efficacy of mebendazole and pyrantel against *N. americanus* and *A. duodenale* in areas of Mali, North Western Australia and Zanzibar, which had been proposed to be attributed to emerging anthelmintic resistance. Given the lack of knowledge of the molecular mechanisms linked to resistance in parasitic nematodes [12], much attention is now directed towards the identification of new drug targets and the development of novel and effective nematocides [157-158] as well as new effective strategies to prevent its emergence [159-161].

Much research has been directed toward developing vaccines against a range of parasitic nematodes [10,11,59,162-167]. Early efforts in the development of vaccines against nematodes resulted in the success of a radiation-attenuated larval vaccine (i.e., Dictol) against *Dictyocaulus viviparus*, the bovine lungworm [168]. Radiation-attenuated larvae were also used as the basis of a vaccine against *H. contortus* infection in sheep [169-171]. In experimental trials, this vaccine was demonstrated to be highly effective, with up to 98% protection being achieved in lambs (7-8 months of age) and adult sheep against challenge infection with L3s of *H. contortus* [169-171]. However, in field trials, this vaccine did not achieve protection against naturally acquired infection [172]. These results, together with the discontinuation (for commercial rather than scientific reasons) of the manufacture of an effective radiation-attenuated larval vaccine against *A. caninum* in dogs [173] indicated the need to search for effective immunogenic molecules as anti-nematode vaccines.
A number of protein components of the epithelial cell surface membrane of the digestive tract of different species of gastrointestinal parasitic nematodes have been evaluated as vaccine candidates both in experimental murine models and in livestock [11,165,166,174]. For instance, a 110 kDa integral membrane aminopeptidase of *H. contortus*, which is heavily glycosylated and localized in the brush border of the epithelial cells of the gut of the adult worm, was shown to be effective in reducing the intensity of *H. contortus* infection in different breeds and ages of sheep [175-178]. However, vaccine efficacy is limited to native proteins administered multiple times, usually in Freund's adjuvant [179]. Another peptidase complex (P1), separated from H11 by ion-exchange chromatography, was identified [180] and shown to represent a ubiquitous component of the microvillar membrane of the intestinal cells of *H. contortus* [180]. Although vaccination with this protein complex was shown to result in a significant reduction (69%) in the number of *H. contortus* eggs in the faeces from vaccinated sheep following *H. contortus* challenge infection, P1 led only to a ~22-38% reduction in the intensity of infection [175]. Conversely, vaccination with the glucose-binding glycoprotein complex (H-gal GP complex), separated by lectin affinity chromatography from other integral membrane proteins from the gut of *H. contortus*, was demonstrated to result in ~53-72% protection and a > 90% reduction in the number of eggs in the faeces from vaccinated sheep [181]. However, the vaccination of lambs (9 months of age) with recombinant H-gal GP failed to stimulate the development of a protective immunity against challenge infection with *H. contortus* L3s [182].

Other vaccine candidates have represented molecules in the excretory/secretory products (ES) from worms [59,165,183]. For instance, proteases in ES from parasitic nematodes represent a major focus of vaccine development because of their inferred roles in the digestion of nutrients acquired from the host and in the penetration and migration through host tissues [183]. Of such proteases, metalloproteases and aspartic and cysteine proteases have been the focus of attention for blood-feeding nematodes, such as *H. contortus* and *An. caninum* and *N. americanus* [59,165,184-190]. For instance, the vaccination of dogs with recombinant forms of an aspartyl- and cysteine-protease from *An. caninum* (designated Ac-APR-1 and Ac-CP-2, respectively) was shown to result in a partial protection against this hookworm, characterised by an absence of clinical signs and a reduced fecundity of the adult worms in the dogs [187,191]. In addition, vaccination of hamsters with the *N. americanus* homologue of Ac-CP-2 (i.e., Na-CP-2) was shown to induce a partial protection, associated with a ~30-46% reduction of the intensity of infection following challenge infection with L3s [192]. Similarly, vaccination with a cysteine protease-
enriched fraction (thiol sepharose binding fraction-TSBP), prepared from membrane extracts from the microvillar surface of the intestinal cells of adult *H. contortus* [193], was shown to reduce the intensity of infection by 47% as well as the number of eggs in faeces by 77% in sheep following a single challenge infection [194].

Proteases from larval stages have also been the focus of vaccine research, because of their proposed role(s) in the invasion of the host [183]. In hookworms, for example, one of the best characterised proteases in larval ES is an astacin-like zinc metalloprotease from *An. caninum*, called Ac-MTP-1 [195,196], which has been demonstrated to degrade fibronectin, laminin and collagen [196]. Based on the results of a vaccine trial conducted in hamsters, this protein has been proposed as a potential candidate for the development of a multi-epitope vaccine [192]. Also, two cysteine-rich secretory proteins, known as ‘*Ancylostoma*-secreted proteins’ (ASPs), which belong to the pathogenesis related protein (PRP) superfamily [197-199] and are a major component of ES of hookworm L3s, represent vaccine candidates [200]. Although the function of these molecules has not yet been assessed in detail [200], they have been suggested to play active roles in tissue invasion and modulation of the host immune responses [197,201]. A study showed that the vaccination of hamsters with a recombinant ASP (i.e., Ac-ASP-2) from the infective L3 of *An. ceylanicum* and expressed in *Pichia pastoris* was effective in reducing significantly the intensity of hookworm infection in hamsters challenged orally with infective larvae of *An. caninum* [202]. Ac-ASP-2 exhibits a high degree of amino acid similarity with a low-molecular weight antigen from adult *H. contortus*, called ‘Hc24’. In sheep, vaccination with Hc24 was shown to result in a significant reduction in the number of eggs in faeces and intensity of infection (i.e., ~70%) following challenge infection with infective L3s of *H. contortus* [203]. More generally, ASP homologues have been proposed to play key regulatory roles of fundamental biological processes in a range of organisms, including plants [204], arthropods [205] and trematodes [206]. However, the phylogenetic and evolutionary relationships amongst members of this group of molecules are presently unclear. Given the fundamental roles that ASP homologues are proposed to play in a range of eukaryotes, a clear classification, based on molecular analyses of sequence data derived from a range of eukaryote species, should provide a framework to explore aspects of the structure, function and interactions of these molecules, which will ultimately provide a foundation for the development of strategies to disrupt key biological pathways linked to ASPs.

Taken together, the results of studies focusing on the identification of suitable immunogens and the development of effective vaccines against gastrointestinal parasitic nematodes show that significant progress has been made over the years.
However, there is still a paucity of information on parasite-host interactions at the molecular level. A better understanding of these interactions should assist future research aimed at identifying novel vaccine and drug targets. Advances in genomic technologies are now providing the unique opportunities to explore the molecular biology of parasitic nematodes, parasite-host interactions and disease on a global scale. Indeed, the advent and integration of high-throughput ‘-omics’ technologies (e.g., genomics, transcriptomics, proteomics, metabolomics and lipidomics) are revolutionizing the way biology is done, allowing the systems biology of organisms to be elucidated. In particular, sequencing provides a powerful approach for detailed explorations of transcription (i.e., ‘transcriptomics’) and associated molecular processes in an organism, its tissues and its cells (see Table 1.1).

1.7 Conventional techniques for investigating the transcriptomes of parasitic nematodes

The genome of any living organism includes coding regions that are transcribed mRNAs, which are subsequently translated into proteins. Techniques, such as Northern blot [207], quantitative real-time transcription PCR (qRT-PCR; [208]) and differential display (DD; [209]) have been used to define patterns of transcription of single genes or small numbers of molecules in parasitic nematodes including As. suum [210], T. colubriformis [211], H. contortus [212-215], Oe. dentatum [216], the ‘brown stomach worm’ of cattle Os. ostertagi [217] and the filarioid nematode Brugia malayi [218]. Recently, techniques that allow ‘global’ analyses of gene transcription have become increasingly popular. For instance, serial analysis of gene expression (SAGE) was one of the first techniques allowing large-scale identification of transcripts and comparison of levels of transcription [219]. The SAGE technique is based on the generation of a short specific tag (14 bp) from each mRNA present in the sample; these tags are used for the construction of a SAGE library [219]. The sequencing of these tags allows a relatively high-throughput determination of their frequencies in the library, which are correlated with relative amounts of the corresponding mRNAs [219]. Despite its demonstrated utility in investigations of yeast [220] and humans [221,222], the application of the SAGE method for studies of gene transcription in parasitic nematodes of socio-economic importance has remained limited (see [223]). A single study [224] used SAGE to sequence and analyse ~3,000 transcripts from adult H. contortus, of which ~60% had homologues in public databases.

In the last years, the analysis of ESTs datasets has proven to be the most widely used approach for investigations of the transcriptomes of parasitic nematodes. In vitro, mRNAs can be reverse transcribed, resulting in stable complementary DNAs (cDNAs);
ESTs are usually single pass DNA sequence reads derived from cloned cDNAs [15,225]. Traditional sequencing [226,227] has been used, by which a DNA polymerase, a primer and four types of deoxynucleotide triphosphates (dNTPs) are used to synthesize the complementary strand to the template sequence [226-228]. The advent of EST sequencing marked a revolution in the field of parasitology and has been used in a range of studies aimed at investigating fundamental molecular processes in parasitic nematodes as well as drug and vaccine target discovery [17,229-236]. For nematodes of animals, applications range from the analyses of stage- and gender-enriched molecules (e.g., *B. malayi* [229] and *Oe. dentatum* [216], respectively) to global analyses of gene transcription (e.g., in *N. americanus*; [47,230,237]).

Also cDNA microarray technology [238] represented a significant advance for large-scale studies of the transcriptomes of parasitic nematodes [239]. In cDNA microarrays, several thousands of oligonucleotides, usually cDNAs, EST clones or fragments of PCR products (which correspond to previously characterised genes/transcripts) are ‘spotted’ (= ‘arrayed’) on to glass slides or chips in precise positions. The mRNAs from different stages or tissues are labelled with different fluorescent or radioactive markers and hybridized to the spots on the array. The relative abundance of hybridization for each mRNA population is then determined by comparing the relative signal intensity of each fluorescent marker [238]. Supported by the increasing amount of sequence data available in public databases for parasitic nematodes, microarray technology has allowed comparisons of levels of transcription of large numbers of mRNAs in, for instance, different tissues, developmental stages and sexes of parasitic nematodes to be performed, ultimately providing researchers with the opportunity to identify molecules considered to play essential roles in fundamental biological pathways of survival, development and reproduction of these parasites [239]. Clearly, the availability of microarray technology has resulted in an expanded knowledge of the transcriptomes of of socio-economically important parasitic nematodes, including *As. suum* [240,241], *H. contortus* [242], *T. vitrinus* [243], *Oe. dentatum* [244], *Teladorsagia circumcincta* [245], *An. caninum* [201,246] and *B. malayi* [247-251]. In addition, the combined use of suppressive-subtractive hybridization (SSH) and microarray analysis has been useful in enabling rapid comparisons of transcriptional profiles between/among life cycle stages, genders and/or species of parasitic nematodes [201,241,244,245,252,253]. A number of studies have employed these tools to elucidate molecular aspects of development, reproduction and host-parasite interactions in gastrointestinal parasitic nematodes. For instance, in *As. suum*, a cDNA microarray analysis of transcripts derived from L4s isolated from the jejunum (i.e., L4-I) and ileum (L4-I) of experimentally infected pigs revealed molecules encoding
actins, myosins and dehydrogenases to be highly expressed in L4-] [240]. As L4s displaced to the ileum are eliminated from the host as a consequence of a Th2-dependent immunological response [254-256], a role of these molecules in biological pathways linked to the survival of *As. suum* in the mammalian host was hypothesized [240]. In another study using SSH followed by a cDNA microarray analysis, Huang et al. [241] identified enolases and protein kinases and phosphatases to be abundantly transcribed in *As. suum* L3s, in which the probes spotted on to the array were derived from mRNAs from different developmental stages of this nematode [241]. Some information on tissue-specific transcription in *As. suum* is also available from a recent study which explored a relatively large number of ESTs from the intestine of the adult stage of this nematode and compared the data with an EST dataset from the intestine of *H. contortus* [257]. The results showed that ~20% of the peptides predicted from each of these datasets were conserved [257], thus suggesting that differences exist among the transcriptomes of parasitic nematodes belonging to different phylogenetic groups. This conclusion was not particularly surprising. Clearly, deep explorations of transcriptomes from different developmental stages, sexes and tissues of these and other parasitic nematodes are required to provide better insights into various aspects of their molecular, cellular and developmental biologies as well as parasite-host interactions and parasitism.

Despite the substantial economic impact of gastrointestinal parasitic nematodes of humans and livestock [1-11,258], there is only a small number of transcriptomic studies, and global analyses are lacking. For instance, SSH-based microarray analyses of both sexes of *T. vitrinus* [243], *Oe. dentatum* [244], *H. contortus* [242] provided initial insights into gender-enriched transcription. These studies showed that transcripts encoding protein kinases and phosphatases, major sperm proteins (MSPs) and enzymes involved in carbohydrate metabolism were consistently male-enriched, whereas vitellogenins, heat-shock proteins and chaperonins were enriched in females [242-244]. Given the economic impact of these nematodes in the livestock industry [4,7,10,11,258] and the problems associated with genetic resistance to various classes of nematocides [12,259,260], an enhanced knowledge of the molecules transcribed in other developmental stages of these parasites could provide a basis for molecular mechanisms linked to drug resistance as well as for of the prediction or identification of novel drug targets.

Knowledge of the complement of molecules transcribed in the larval stages of gastrointestinal parasitic nematode should also aid the elucidation of pathways associated with infectivity and interactions with the vertebrate host. The molecular mechanisms that regulate the transition from the free-living to the parasitic stage of
nematodes may allow the development of novel strategies to disrupt this transition. Previous studies have analysed differences in transcription between the ensheathed, free-living L3 and the exsheathed L3 of *H. contortus* [7,215,231,261] and the related strongylid, *An. caninum* [201,246]. The results of a cDNA microarray analysis followed by qRT-PCR of differentially transcribed molecules showed that, amongst others, the majority of transcripts encoding ASPs were up-regulated in free-living L3s compared with parasitic, serum-stimulated larvae of *An. caninum* [246]. However, using SSH-based microarray analysis, a recent study [201] showed a substantial 'up-regulation' in the numbers and levels of transcripts encoding ASPs in serum-activated L3s [201].

Considering that *An. caninum* and *H. contortus* are both blood-feeding nematodes, deeper investigations of the molecular mechanisms associated with the transition from the free-living to the parasitic stage of *H. contortus* should provide insights into conserved pathways of development for these strongylid nematodes. In addition, with the exception of *An. caninum* [246,262,263], investigations of the transcriptomes of other hookworms are presently limited [47,230,237]. To date, molecular studies of this group of blood-feeding nematodes have mainly involved *An. caninum*, because of its use as a “model” for human hookworms [201,246,262-264]. Clearly, detailed knowledge and understanding of the molecules transcribed in all stages of different species of hookworms, including *N. americanus* and *An. duodenale* of humans, should facilitate the identification of conserved pathways linked to development, survival, reproduction, parasite interactions and disease and could assist in the discovery of new intervention strategies.

Recent advances in sequencing technologies [265-268] now provide the unique opportunity to perform *de novo* analyses of the whole transcriptomes of different species, sexes and/or developmental stages of nematodes of socio-economic importance. Indeed, although 'next-generation' (also called “massively parallel”) sequencing methods were introduced only recently (Roche 454 GS FLX; [265]), their capacity to generate millions or hundreds of millions of sequences in parallel has placed them at the forefront of the genomic and transcriptomic research [269-271]. They are thus powerful tools for investigating the transcriptomes of parasitic nematodes on a scale like never before.

### 1.8 Next-generation sequencing technologies

Next-generation sequencing (NGS) has substantially reduced the costs of generating large sequence datasets. Currently available NGS sequencing platforms include the 454/Roche ([265]; [www.454.com]), the Illumina/Solexa ([266];
www.illumina.org and the SOLiD (= Supported Oligonucleotide Ligation and Detection) ([267]; www.appliedbiosystems.com) (Table 1.2).

The 454/Roche platform ([265]; www.454.com) uses a sequencing-by-synthesis approach (Fig. 1.2). The main principles of this technology are an in vitro sample preparation and a ‘miniaturization’ of sequencing chemistries which enable the generation of sequence data in a ‘massively parallel’ manner [272]. For transcriptomic studies, cDNA is randomly fragmented (by ‘nebulization’) into 500-1000 base pair (bp) fragments. For the preparation of the library, an adaptor is ligated to each end of these fragments, which are then mixed with a population of agarose beads whose surfaces anchor oligonucleotides complementary to the 454-specific adapter sequence, such that each bead is associated with a single fragment. Each of these complexes is transferred into individual oil-water micelles containing amplification reagents and is then subjected to an emulsion PCR (emPCR) step, during which ~10 million copies of each cDNA are produced and bound to individual beads. In the sequencing phase, the beads anchoring the cDNAs are deposited on a pico-titre plate, together with other enzymes required for the pyrophosphate sequencing reaction (i.e., ATP sulfurylase and luciferase). The sequencing is carried out by flowing sequencing reagents (nucleotide and buffers) over the plate [272]. To date, the 454 sequencing technology is still the only platform able to generate ‘long’ reads (100-600 bp) and is thus largely used for de novo genomic or transcriptomic studies.

After the introduction of the 454 technology, the first Illumina (formerly Solexa) sequencer became available ([266]; www.illumina.com). This technology, now able to generate billions of bases per run, has features that differ significantly from the 454 approach [266]. After fragmentation of cDNA sample into a shotgun library, Illumina-specific adaptors are ligated in vitro to each cDNA template; one terminus of the template is covalently attached to the surface of a glass slide (or flow cell). Attached to the flow cell are primers complementary to the other end of the template, which bend the cDNAs to form bridge-like structures. During the amplification step (bridge-PCR), clonal clusters, each consisting of ~1000 amplicons, are generated and immobilized to a single physical location on the slide. Subsequently, the cDNAs are linearised, and the sequencing reagents are directly added to the flow cell, with four fluorescently labelled nucleotides. After each incorporation of a fluorescent base, the flow cell is interrogated with a laser in several locations, which results in several image acquisitions at the end of a single synthesis cycle [272]. This technology is considered ideal for re-sequencing projects, targeted sequencing, single nucleotide polymorphism (SNP) analyses and gene transcription studies.
The SOLiD platform ([267]; www.appliedbiosystems.com) was introduced recently. Although this platform relies on chemistries similar to the 454 (www.454.com), the sequencing process employs the enzyme DNA ligase, instead of a polymerase [272]. Briefly, after an emPCR step, the adaptor sequences of the cDNA templates bind to complementary primers that are covalently anchored to a glass slide. Subsequently, a set of four fluorescently labelled di-probes (octamers of random sequence, except known dinucleotides at the 3’-terminus) is added to the sequencing reaction. In case an octamer is complementary to the template, it will be ligated, and the two specific nucleotides can be called; subsequently, an image is acquired and the fluorescent dye is removed, so that other octamers can be ligated. After multiple ligations (e.g., 7 ligations for a 35 bp read), the newly synthetized cDNA is removed and the primer is inactivated. This process is repeated multiple times from different starting points of the cDNA templates, so that each position is sequenced at least twice. This technique, known as ‘two-base-calling’, allows the correction of sequencing errors, thus providing accurate base calling [272]. Because of the short read-length, the range of applications of the SOLiD system is considered similar to that of the Illumina technology and includes (targeted) re-sequencing projects, SNP detection and gene transcription studies.

In the last years, numerous studies have demonstrated the utility of NGS technologies for investigating, for instance, aspects of the systematics, population genetics and molecular biology of pathogens, including viruses [273], bacteria [274], arthropods [275], protozoa [276-278], nematodes [263,279] and trematodes [280,281]. In particular, the 454 platform was used recently for the de novo sequencing of the transcriptomes of important trematodes of humans and animals (i.e., Clonorchis sinensis, Opisthorchis viverrini and Fasciola hepatica; [280,281]). More than 50,000 unique and novel sequences were characterised for each of these parasites, demonstrating the capacity of this technology to generate huge datasets. The development of bioinformatic tools has become crucial for the detailed analyses of such datasets.

1.9 Current bioinformatic tools and pipelines for the analysis of EST data

The increasing number of ESTs in public databases has been accompanied by an expansion of bioinformatic tools for the analysis of sequence data, both at the cDNA and protein levels. This expansion has resulted in the development of a number of web-based programs and/or integrated pipelines [16,282-287]. The principles, methods and protocols for the analysis of EST data, together with currently available bioinformatic tools and pipelines, have been reviewed [285,289]. In brief, following the acquisition of
sequence data, ESTs are firstly screened for sequence repeats, contaminants and/or adaptor sequences [285,289]. Vector sequences (residual from the cloning step) are eliminated prior to the assembly using information available in web-based non-redundant vector databases, such as UniVec (http://www.ncbi.nlm.nih.gov/VectScreen/UniVec.html) and EMVEC (http://www.ebi.ac.uk/Tools/blastall/vectors.html). Following the pre-processing of ESTs, sequences are 'clustered' (= assembled) into contiguous sequences (of maximum length) based on sequence similarity.

1.9.1 Sequence assembly

The main goal of sequence assembly is to determine, with confidence, the sequence of a target transcript/gene. This process involves the alignment and merging of DNA fragments to form long contiguous sequences (i.e., contigs) [285,288]. Long- (e.g., generated by Sanger sequencing, 454 platform) and short-reads (e.g., Illumina and SOLiD platforms) are assembled using the algorithms 'overlap-layout-consensus' [290] and 'de Bruijn graph' [291,292], respectively.

For the former algorithm [290], all pair-wise overlaps among reads are computed and stored in a graph; all graphs are then used to compute a layout of reads and a consensus sequence of contigs [293,294]. Some of the assemblers designed to support long-read assembly include PHRAP [295], the Contig Assembly Program v.3 (CAP3; [282]), the TIGR assembler [296], the Parallel Contig Assembly Program (PCAP; [297]) and the Mimicking Intelligent Read Assembly program (MIRA; [298]).

For the 'de Bruijn graph' [291,292], reads are fragmented into short segments, denominated 'k-mers', where 'k' represents the number of nucleotides in each segment; overlaps between or among k-mers are captured and stored in graphs, which are subsequently used to generate the consensus sequences [293,294]. Examples of programs specifically designed for the assembly of short-reads include the Short Sequence Assembly by k-mer search and 3' read extension (SSAKE; [299]), Velvet [292], the exact de novo assembler (EDENA; [300]), Euler-SR [301], the assembly by short sequencing (ABySS; [302]) and the short oligonucleotide analysis package (SOAP; [303]).

1.9.2 Conceptual translation and annotation

Following assembly, the contigs and single reads (or singletons) are compared with known sequence data available in public databases, in order to assign a predicted identity to each query sequence if significant matches are found [285]. In addition, assembled nucleotide sequences are usually conceptually translated into predicted
proteins using algorithms that identify protein-coding regions (ORFs) from individual contigs. Examples of such algorithms are OrfPredictor [304], ESTScan [283] and DECODER [305]. Once peptides are predicted, protein analyses, including amino acid sequence comparisons with data available in public databases, and known protein domains are then inferred [285]. For instance, the software InterProScan [286] provides an integrated tool for the characterization of a protein family, or an individual protein sequence, domain and/or functional site by comparing sequences with information available in the databases PROSITE [306], PRINTS [307], Pfam [308], ProDom [309], SMART [310] and/or Gene Ontology (GO; [311]). In addition, other programs are available for the prediction of transmembrane domains (e.g., TMHMM; [312]) and/or signal peptide motifs (e.g., SignalP; [313]).

1.9.3 Similarity searches, and current online tools and databases

Different types of the Basic Local Alignment Software Tool (BLAST; [314]) are used for comparing the nucleotide sequence data with DNA or cDNA (BLASTn), or amino acid (BLASTx) sequences or conceptually translated peptides with protein sequences (BLASTp), available in public databases [285]. Public databases represent comprehensive collections of nucleotide and amino acid sequences. Due to the rapid progress in the discovery and characterization of novel genes and proteins, online public databases have become one of the primary resources for sequence data analysis and annotation. For example, the International Nucleotide Sequence Database Collaboration includes three ‘sister’ databases, namely GenBank [315], the Enterprise Management Technology Transfer nucleotide database (EMBL; [316]) and the DNA Databank of Japan (DDBJ; [317]). In these databases, all publicly available nucleotide sequences are stored and curated; in addition, each sequence is stored as a separate record and linked to information, such as primary source, references and predicted and/or experimentally verified biological features. For ESTs, raw sequence data are often stored in subdivisions of these nucleotide databases, such as the UniGene [318] and the Sequence Read Archive [319] databases. Various databases, which exclusively store known amino acid sequence data, are also available. For instance, the Protein Data Bank (PDB; [320]), maintained by the Research Collaboratory for Structural Bioinformatics, represents the primary source for protein structures, whereas the SWISS-PROT database [320] is a protein sequence database for a number of prokaryotes and eukaryotes. The TrEMBL [322] division of SWISS-PROT contains a non-redundant set of translations for all coding sequences in the EMBL nucleotide sequence database that do not correspond to existing SWISS-PROT entries. In addition to these comprehensive general databases, there is a number of specialized collections
of gene and protein information on specific organisms. Examples include the databases for *Saccharomyces cerevisiae* (yeast) ([http://www.yeastgenome.org/](http://www.yeastgenome.org/); [323]), *Drosophila melanogaster* (v vinegar fly) ([http://flybase.org/](http://flybase.org/); [324]), *Mus musculus* (mouse) ([http://www.informatics.jax.org/](http://www.informatics.jax.org/); [325]) and *C. elegans* (free-living nematode) (WormBase at [http://www.wormbase.org](http://www.wormbase.org); [326]). WormBase is a comprehensive repository of information on *C. elegans* and related nematodes, such as *C. briggsae* [326,327]. Here, essentially all information and data on classical genetics, cellular biology, structural and functional genomics of these free-living nematodes are stored and continually curated [326-329].

Recently, a web-based bioinformatic pipeline (called ESTExplorer) was established for the automated analysis and annotation of EST datasets (both at the nucleotide and amino acid levels) ([16]; see Fig. 1.3), and shown to substantially accelerate and facilitate the analyses of sequences (generated using conventional Sanger sequencing) compared with traditional database searches [330]. However, thus far, no studies have assessed the performances of this pipeline for the analyses of large transcriptomic (e.g., RNA-seq) datasets generated using NGS technologies.

1.9.4 The free-living nematode, Caenorhabditis elegans, as a model for parasitic nematodes

Presently, the analysis and annotation of sequence data derived from parasitic nematodes relies heavily on information available for *C. elegans* (in WormBase). The latter nematode is simple in its anatomy (959 somatic cells in the hermaphrodite and 1031 in the male), has a short life cycle (~3 days) and is easy to culture in vitro [331]. The genome of *C. elegans* is ~97 Mb in size [332]; it contains five autosomal and a single sex chromosome. The karyotype is n=12 (10A:2X) for the the hermaphrodite and 2n=11 (10A:X) for the male. Since the first publication on the genetics of *C. elegans* [331], the amount of biological and molecular data has expanded massively [333], so that the establishment of an up-to-date database to rapidly access information of this free-living nematode had become necessary. Currently, WormBase ([www.wormbase.org](http://www.wormbase.org)) contains detailed and curated information on ~19,000 *C. elegans* genes and associated data on, for instance, transcription/expression profiles in different developmental stages, tissues and cells, mutants and their phenotypes, genetic and physical maps, SNPs, information on gene-gene and protein-protein interactions as well as all peer-reviewed literature pertaining to *C. elegans*.

The introduction of the technique of RNA interference (RNAi; [334]) has represented a revolution in the study of gene function in metazoan organisms and has led to detailed information of the functions of ~96% genes in *C. elegans* [335-339]. The
principle of RNAi relies on the introduction of double-stranded RNA (dsRNA) into the cells of a living organism, which induces the degradation of the homologous (target) mRNA [334]. The dsRNA can be introduced directly into C. elegans by injection [334], by soaking worms in solution [340] or by feeding worms Escherichia coli expressing a dsRNA fragment of a target gene [341]; it can also be introduced using a transgene expressing dsRNA [342,343]. This gene silencing approach opened up avenues for large-scale studies of molecular function in C. elegans [335-338,342,344,345] as well as for comparative studies (e.g., comparison with parasitic nematodes or humans; [346-349]).

In addition to the RNAi technique, the transgenesis of C. elegans by has been widely used for assessing and proving gene function [350,351]. This technique can involve the microinjection of expression constructs, which usually comprise plasmid or cosmid DNA, often incorporating green fluorescent protein (GFP; [352]) into the syncytium (mitotically active) region of the adult hermaphrodite gonad (= ‘gonadal microinjection’); alternatively, the DNA constructs can be transferred directly into target cells via high density microparticles of gold or tungsten (= ‘biolistics’ or ‘particle bombardment’) [353]. Introduced DNA does not usually integrate into the chromosome, rather it forms a multi-copy extrachromosomal array, which can be inherited. GFP allows the study of a number of (temporal and spatial) biological processes, including gene expression, protein localization and dynamics, protein-protein interactions, cell division, chromosome replication and organization, intracellular transport pathways, organelle inheritance and biogenesis [354]. In C. elegans, GFP was first used as a marker for gene expression [352], where the GFP coding sequence was placed under the control of a promoter for the me-7 tubulin gene; in C. elegans expressing this regulated GFP, the pattern of fluorescence in vivo was similar to that characterised previously using an antibody probe [352]. Since this first study, GFP has been used in gene expression studies of C. elegans and led to a Nobel Prize (to Dr Martin Chalfie, 2008; http://nobelprize.org/nobel_prizes/chemistry/laureates/2008/#). Today, GFP is applied widely to analyses of gene expression and localisation in a broad range of biological systems [355].

In addition to studies of gene expression and localization, patterns of gene transcription during key developmental and reproductive processes have also been investigated in C. elegans using microarray technology [356-358]. In the first study [356], three groups of molecules were demonstrated to have high expression levels in the germline tissues of C. elegans, i.e. the ‘germline intrinsic’ molecules (expressed in the germline of hermaphrodites producing either sperm or oocytes, and proposed to
play key roles in biological processes linked to meiosis, stem cell recombination and germline development) and molecules highly expressed exclusively in oocytes-producing and sperm-producing hermaphrodites, respectively [356]. The latter group included a large number of molecules, such as protein kinases and phosphatases, associated with spermatogenesis, in accordance with other studies investigating gender-enriched transcriptional patterns in parasitic nematodes [242-244]. Previously, genetic studies had indicated that ~50-70% of genes in parasitic nematodes have orthologues in C. elegans [23,235], which led to the grouping of this free-living nematodes into ‘clade V’ of the phylum Nematoda, together with parasitic nematodes of the order Strongylida [23]. These results, together with similarities in various characteristics (such as body plan and moulting) between C. elegans and some parasitic nematodes [7,359] indicate that this free-living nematodes provides a useful system for comparative investigations of many conserved biochemical and molecular pathways linked to development in related nematodes.
1.10 Conclusions from the literature review and aims of the thesis

In the last decades, significant progress has been made in the study of various aspects of epidemiology, pathogenesis, immunobiology, diagnosis, treatment and the control of diseases caused by parasitic nematodes of socio-economic importance. However, before integrated approaches for control can be developed, much more knowledge and understanding of molecular processes that underpin parasite development, host-parasite interactions and disease are needed. The study of the complement of molecules transcribed in different developmental stages, tissues and cells of these parasites, and of similarities and differences in the transcriptomes among different species of nematodes, could provide a sound basis toward this discovery effort. Various studies have elucidated patterns of transcription of small numbers of genes or proteins in some socioeconomically important parasitic nematodes, including As. suum, H. contortus and other trichostrongylids, An. caninum, N. americanus and Oe. dentatum. However, detailed knowledge is still lacking. The availability of NGS technologies now provides unprecedented opportunities to explore the molecular biology of parasitic nematodes on a scale never before possible. This massive technology capacity now requires critical evaluations and the development of enhanced and practical bioinformatic pipelines for the assembly and analyses of NGS datasets. This thesis tackles some of these aspects. The main objectives were to significantly advance our knowledge and understanding of the molecular biology of key gastrointestinal nematodes (orders Ascaridida and Strongylida) through transcriptomic studies and to assess, establish and integrate bioinformatic tools for the rapid and efficient analysis of NGS datasets.

The specific research aims were to:

- Use a conventional, publicly available, worldwide-web (www)-based bioinformatic workflow platform to analyse abundant, gender-enriched transcripts in the adult stage of the porcine roundworm, As. suum (order Ascaridida) (Chapter 2);
- Employ advanced sequencing technology to gain first insights into the transcriptome of T. colubriformis, an important small intestinal nematode of sheep (order Strongylida) (Chapter 3);
- Extend the use of this sequencing technology (Chapter 3) to H. contortus, a related nematode of the abomasum of the sheep (Chapter 4);
- Use microarray and bioinformatic tools to explore transcriptional conservation between H. contortus (Chapter 4) and the canine hookworm, An. caninum (family Ancylostomatidae) (Chapter 5);
• Gain deep insights into the transcriptome of *N. americanus*, a human hookworm, through the use of massively parallel sequencing and bioinformatic analyses (Chapter 6);

• Establish a practical and user-friendly bioinformatic workflow system for the analyses of large-scale datasets generated by massively parallel sequencing (Chapter 7);

• Construct a bioinformatic-phylogenetic framework for the classification of pathogenesis related proteins (PRPs), including ASPs, encoded in the transcriptomes of parasitic worms and other eukaryotes (Chapter 8), and

• Discuss the relevance of the results in relation to the current state of knowledge in relevant fields and propose areas for future investigation (Chapter 9).
1.11 References


63. Monnig HO. The life histories of Trichostrongylus instabilis and T. rugatus of sheep in South Africa. 11-12th Annual Report of the Director of Veterinary Education and Research, Union of South Africa 1926;231-51.


148. Roos MH. The role of drugs in the control of parasitic nematode infections: must we do without? Parasitology 1997;114:S137-44.


177. Newton SE, Morrish LE, Martin PJ, Montague PE, Rolph TP. Protection against multiply drug-resistant and geographically distant strains of *Haemonchus*


Sarkar G, Sommer SS. Access to a messenger RNA sequence or its protein product is not limited by tissue or species specificity. Science 1989;244:331-4.


<table>
<thead>
<tr>
<th>Year</th>
<th>Milestone</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>1965</td>
<td>Sequencing of the first RNA molecule</td>
<td>[360]</td>
</tr>
<tr>
<td>1977</td>
<td>Development of the Northern blot technique and the Sanger sequencing method</td>
<td>[207,226,227]</td>
</tr>
<tr>
<td>1989</td>
<td>First report of reverse transcription PCR experiments for transcriptome analysis</td>
<td>[361]</td>
</tr>
<tr>
<td>1991</td>
<td>First high-throughput EST sequencing study</td>
<td>[15]</td>
</tr>
<tr>
<td>1996</td>
<td>Reports of the use of the microrray and Serial Analysis of Gene Expression (SAGE) methods for transcriptome studies</td>
<td>[238,362]</td>
</tr>
<tr>
<td>2005</td>
<td>First next-generation sequencing platform introduced to the market (454/Roche)</td>
<td>[265]</td>
</tr>
<tr>
<td>2006</td>
<td>First sequencing of a transcriptome using a NGS technology (454/Roche)</td>
<td>[363]</td>
</tr>
</tbody>
</table>
Table 1.2 Technical features of next-generation sequencing platforms (i.e. 454/Roche, Illumina/Solexa and SOLiD).*

<table>
<thead>
<tr>
<th></th>
<th>454/Roche</th>
<th>Illumina/Solexa</th>
<th>SOLiD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Platform</td>
<td>Genome Sequencer FLX</td>
<td>Genome Analyzer IIx</td>
<td>SOLiD 3 Plus System</td>
</tr>
<tr>
<td>Sequencing method</td>
<td>Emulsion PCR of bead-bound oligos</td>
<td>Isothermal bridge amplification on flowcell</td>
<td>Emulsion PCR of bead-bound oligos</td>
</tr>
<tr>
<td>Sequencing chemistry</td>
<td>Pyrosequencing using polymerase</td>
<td>Ligation ('dual-base encoding' octamers)</td>
<td>Reversible terminator using polymerase</td>
</tr>
<tr>
<td>Base pairs sequenced (per run)</td>
<td>0.5 Gbp</td>
<td>~30-50 Gbp</td>
<td>~60 Gbp</td>
</tr>
<tr>
<td>Read length</td>
<td>400-500 bp (titanium chemistry)</td>
<td>50 bp</td>
<td>100 bp</td>
</tr>
<tr>
<td></td>
<td>or ~250 bp (standard chemistry)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Run time</td>
<td>~12 hours</td>
<td>~2-9 days</td>
<td>~3 days</td>
</tr>
<tr>
<td>Advantages</td>
<td>Long read length</td>
<td>No emulsion PCR required</td>
<td>High accuracy due to dual base calls</td>
</tr>
<tr>
<td>Disadvantages</td>
<td>Unreliable determination of homopolymer regions and large repeats</td>
<td>Long run time, short read length</td>
<td>Short read length</td>
</tr>
<tr>
<td>Machine cost</td>
<td>~$500,000</td>
<td>~$600,000</td>
<td>~$600,000</td>
</tr>
</tbody>
</table>

*Data tabulated on October 2010.
Fig 1.1 Evolutionary relationships within the phylum Nematoda based on the analysis of sequences representing the small subunit (SSU) of the ribosomal DNA from 53 nematode taxa, using the neighbour joining and maximum parsimony algorithms (modified from [23]).
Fig. 1.2 Schematic representation of the cDNA sequencing reaction performed by the 454 platform. Specific cDNA samples are fractioned into small cDNA fragments; specific adapter sequences are added to both ends of each fragment. Each fragment is then captured by cDNA capture beads; during the emulsion PCR step, several million copies of each fragment are produced. The beads with amplified fragments are then loaded on a PicoTiterPlate™ for sequencing, which is performed by flowing sequencing reagents (nucleotides and buffers) over the plate (from www.454.com).
Fig. 1.3 Schematic representation of the ESTExplorer workflow. Phase I is dedicated to EST sequence pre-processing and assembly, whereas Phases II and III provide annotation at the nucleotide and the amino acid level, respectively (modified from [330]).
CHAPTER 2

Using a worldwide-web (www)-based workflow platform for the bioinformatic analysis of abundant gender-enriched transcripts of adult Ascaris suum (Nematoda)¹

Abstract

Expressed sequence tag (EST) data representing transcripts with a high-level of differential hybridization in suppressive-subtractive hybridization (SSH)-based microarray analysis between adult female and male Ascaris suum were subjected to detailed bioinformatic analysis. A total of 361 ESTs clustered into 209 sequences, of which 52 and 157 represented transcripts that were enriched in female and male A. suum, respectively. Thirty (57.7%) of the ‘female’ subset of 52 sequences had orthologues/homologues in other parasitic nematodes and/or Caenorhabditis elegans, 13 (25%) exclusively in other parasitic nematodes and nine (17.3%) had no significant similarity to any other organism for which sequence data are currently available; the C. elegans orthologues encoded molecules involved in reproduction as well as embryonic and gamete development, such as vitellogenins and chitin-binding proteins. Of the ‘male’ subset of 157 sequences, 73 (46.5%) had homologues in other parasitic nematodes and/or C. elegans, 57 (37.5%) in other parasitic nematodes only, and 22 (14.5%) had no significant similarity to any other organism; the C. elegans orthologues encoded predominantly major sperm proteins (MSPs), kinases and phosphatases, actins, myosins and an Ancylostoma-secreted protein-like molecule. The findings of the present chapter should support further genomic investigations of As. suum.

¹ New nucleotide sequence data used in this chapter have been deposited in the GenBank database under accession numbers G0254595-G0254953.
2.1 Introduction

As described in Chapter 1, diseases caused by parasitic nematodes (= roundworms) are of major socioeconomic importance worldwide. For example, *Ascaris* sp. (or *As. lumbricoides*) infects hundreds of millions of people globally and causes the disease ascariasis, which particularly affects children [1]. Similarly, ascariasis of pigs due to *As. suum* infection can cause major production and financial losses linked to reduced feed conversion efficiency as well as losses to the meat industry associated with the condemnation of ‘milk-spot’ livers [2]. In addition to the socioeconomic impact of such parasites in animals, genetic resistance in a range of nematode groups against all main classes of anthelmintics (drugs) has become a major problem globally [3]. Therefore, there is a need to discover new intervention methods against parasitic nematodes. Gaining an improved understanding of fundamental, molecular aspects of parasite development and reproduction provides a possible avenue to assist this discovery effort.

Compared with the free-living nematode *Caenorhabditis elegans* (see WormBase; www.wormbase.org), there is a paucity of information on the molecular biology of development and reproduction in parasitic nematodes [4-7]. The entire genome sequence of *C. elegans* is known [8], the functions of more than 96% of the *C. elegans* genes have been assessed [9-13] by double-stranded RNA interference (RNAi, or gene silencing; [14]), and microarray analyses have explored developmental and gender-enriched gene expression [15,16]. Current evidence indicates that ~50-70% of genes in parasitic nematodes have orthologues in *C. elegans* [17,18], and there is similarity in other features (such as moultmg and basic body plan) between *C. elegans* and parasitic nematodes, indicating that some molecular pathways are conserved [6,19]. Understanding the pathways linked to basic nematode biology and development has important implications for finding new ways of disrupting these pathways and thus facilitates the identification of new drug targets.

Advances in genomic and bioinformatic technologies are also providing opportunities to explore developmental and reproductive processes in parasitic nematodes at the molecular level [5,7,20-29]. Considering the major socioeconomic impact of ascariasis in humans and pigs [30,31], *As. suum* provides a useful experimental model system, because: (i) the adult stage of this nematode is large, allowing investigations of individual organ systems and tissues; (ii) various developmental stages of the parasite can be produced in experimentally infected pigs; (iii) larvae can be maintained in *in vitro* culture for relatively long periods of time (weeks), and eggs can be stored at 4 °C for years [32]; (iv) RNAi achieves “cross-species” gene silencing for selected genes [33,34]; and, (v) the genome sequence is imminent (http://www.sanger.ac.uk/Projects/Helminths/).
Recent reports describe differential transcription between stages and the sexes of *As. suum* using suppressive-subtractive hybridization (SSH)-based microarray analyses [35-37]. In the latter study, Huang et al. [37] undertook a detailed bioinformatic analysis of ~500 expressed sequence tags (ESTs) demonstrated to be enriched in the infective, third-stage (L3) larva, mapped them to biochemical pathways and predicted genetic interactions based on comparisons with *C. elegans* and/or other organisms, as a foundation for functional studies. Although the former studies [35,36] profiled gender-enriched transcripts in *As. suum*, no detailed characterisation of these transcripts was undertaken, thus limiting the interpretation of the results. In the present chapter, this work is extended by undertaking a comprehensive bioinformatic exploration and functional annotation of a key subset of abundant female- and male-enriched transcripts of the adult stage of *As. suum*, as a foundation for future genomic studies focused on elucidating developmental and/or reproductive processes in this nematode.

2.2 Materials and methods

From an SSH-coupled microarray analysis [36], conducted essentially as described by Nisbet and Gasser [38], ESTs representing transcripts \( n = 361 \) with substantial differential hybridization (≥8-16-fold) between adult male and female *As. suum* were subjected to bioinformatic analyses. Differential transcription established by microarray analysis was verified based on reverse transcription-coupled polymerase chain reaction (RT-PCR) analysis (according to [37]) of a subset of 17 transcripts. In the first instance, the semi-automated platform ESTExplorer available at http://estexplorer.biolinfo.org [28,29] was used for the processing of EST data. In brief, all ESTs were pre-processed (SeqClean, RepeatMasker), aligned/clustered using the Contig Assembly Program v.3, CAP3 [39], employing a minimum sequence overlap length “cut-off” of 30 bases and an identity threshold of 95% for the removal of flanking vector and adapter sequences, and assembled. Following the pre-processing of the ESTs, contigs and singletons for each female and male -enriched dataset were subjected to BLASTx (NCBI: www.ncbi.nlm.nih.gov) and BLASTn (EMBL-EBI Parasite Genome Blast Server: www.ebi.ac.uk) to identify putative homologues in *C. elegans*, other nematodes and organisms other than nematodes (e-value cut-off: ≤ 1e-05). WormBase (www.wormbase.org) was interrogated extensively for relevant information on *C. elegans* orthologues/homologues, including RNAi phenotypic, transcriptomic, proteomic and interactomic data. Comparison (at the amino acid sequence level) of *As. suum* ESTs with molecules available for *C. elegans*, other parasitic nematodes and non
nematodes in current databases was performed using SimiTri [40], which provides a two-dimensional display of similarity relationships.

Following the conceptual translations of ESTs into peptides using ESTScan [29], gene ontology (GO) annotations were performed using BLAST2GO [41]. Peptides were mapped via InterProScan (domain/motifs) and to respective pathways in *C. elegans* using KOBAS (KEGG Orthology-Based Annotation System; [42]. The open reading frames (ORFs) inferred from ESTs with orthologues in *C. elegans* were also subjected to “secretome analysis” using the program SignalP v.2.0 (available at www.cbs.dtu.dk/services/SignalP/), employing both the neural network and hidden Markov models to predict signal peptides and/or anchors [43-45]. Also, transmembrane domains were predicted using the program TMHMM (www.cbs.dtu.dk/services/TMHMM/; [46-48]), and subcellular localization inferred employing the program WoLF PSORT (http://wolfsort.org/; [49].

The method developed by Zhong and Sternberg [50] was used to predict the interactions among *C. elegans* orthologues of the gender-enriched molecules from *As. suum*. Genomic data (regarding interactions, phenotypes, expression and gene ontology) from selected *C. elegans* gene orthologues/homologues, also incorporating data from *Drosophila melanogaster* (vinaegar fly), *Saccharomyces cerevisiae* (yeast), *Mus musculus* (mouse) and *Homo sapiens* (human), were integrated using a naïve Bayesian model to predict genetic interactions among *C. elegans* genes using the recommended, stringent cut-off value of 4.6 [50,51]. The predicted networks resulting from the analyses were saved in a graphic display file (gdf) format and examined using the graph exploration system available at http://graphexploration.cond.org/.

### 2.3 Results and discussion

From a total of 2,163 clones from the gender-enriched cDNA libraries [35,36], a subset of 361 ESTs was selected based on a substantial gender-specific hybridization signal in a microarray analysis. Of this subset, 122 ESTs had an eight to 168-fold higher signal intensity in females than males, and 239 ESTs had a 16 to 266-fold increased signal in males than females [36]. Henceforth, ESTs that related to female-enriched and male-enriched transcripts in *As. suum* are designated as ‘female’ and ‘male’ ESTs, respectively. RT-PCR analysis of eight and nine molecules selected from the ‘female’ and ‘male’ EST sets, respectively, independently confirmed the gender-specificity of microarray hybridization (results not shown). Following bioinformatic pre-processing, the ‘female’ ESTs were 202-876 bp (mean: 477 bases ± 284 [= standard deviation]), whereas the ‘male’ ESTs were 189-705 bases in length (mean: 510 bp ± 175). Following
clustering, the mean lengths of contigs were 643 (± 197) bases and 636 (± 252) bases for respective EST sets.

Upon clustering, the 122 ‘female’ and 239 ‘male’ ESTs were represented by 52 (= 21 contigs + 31 singletons) and 157 (43 contigs + 114 singletons) distinct sequences. Of these clustered sequences, 51 (98%) and 152 (97%) could be conceptually translated into peptide sequences using ESTScan. InterProScan analysis of the 203 peptides allowed 96 (47%) of them to be mapped to either a protein domain or motif. All of the clustered sequences were compared against the non-redundant (nr) genomic database for parasitic nematodes available at the EMBL-EBI Parasite Genome Blast Server (www.ebi.ac.uk) using BLASTn and against the C. elegans proteomic database available at www.wormbase.org using BLASTx. Of 52 ‘female’ sequences, 30 (57.7%) had orthologues/homologues in other parasitic nematodes and/or C. elegans (Table 2.1; GenBank accession numbers GO254595-GO254715), 13 (25%) exclusively in other parasitic nematodes, including ascaridoids (i.e., As. lumbricoides and Toxocara canis) and nine (17.3%) had no significant similarity to molecules in any other organism for currently available sequence data. Of the 157 ‘male’ sequences, 73 (46.5%) had orthologues/homologues in other parasitic nematodes and/or C. elegans (Table 2.2; accession numbers GO254716-GO254953), 57 (37.5%) in other parasitic nematodes only, including As. lumbricoides and T. canis, and 22 (14.5%) had no significant similarity to molecules in any other organism (e-value cut-off: < 1e-05) for which sequence data are currently available. All 103 sequences with orthologues/homologues in C. elegans were annotated in detail using ESTExplorer and then categorized (at the protein level), following interrogation against three databases containing protein sequences from various organisms. Specifically, data were compared with protein sequences available for (i) C. elegans (from WORMPEP v.167; cf. www.wormbase.org), (ii) parasitic nematodes (available protein sequences and peptides from conceptually translated ESTs), and (iii) organisms other than nematodes (from NCBI nr protein database; cf. www.ncbi.nlm.nih.gov). A three-way comparison of the gender-enriched molecules of As. suum has been figuratively displayed using SimiTri (Fig. 2.1).

As gene ontology (GO) provides a hierarchy that unifies the descriptions of biological, cellular and molecular functions [52], this approach was used to predict the classification and gene function of C. elegans molecules representing the subsets of 52 ‘female’ and 157 ‘male’ sequences from As. suum. The approach developed by Beissbarth and Speed [53] was used to identify statistically significant GO terms within both gender-enriched EST datasets. A summary of the GO categories of these molecules is given in Appendix 2.1. Twenty-two (42.3%) of the former subset could be assigned to GO categories ‘biological process’ (n = 20 terms), ‘cellular component’ (n = 11) and
molecular function’ (n = 26), with the most common subcategories being ‘lipid transport’ (GO:0006869; 40%) and ‘chitin metabolic process’ (GO:0006030; 40%) within ‘biological process’, ‘extracellular region’ (GO:0005576; 73%) within ‘cellular component’ and ‘lipid transporter activity’ (GO:0005319; 31%) and ‘chitin binding’ (GO:0008061; 31%) within ‘molecular function’ (p < 0.05). Fifty-two (33.1%) of the latter subset could be assigned functionally to ‘biological process’ (n = 25), ‘cellular component’ (n = 9) and/or ‘molecular function’ (n = 74). The most common subcategories were ‘protein amino acid dephosphorylation’ (GO:0006470; 28%) and ‘metabolic process’ (GO:0008152; 28%) within ‘biological process’, ‘myosin complex’ (GO:0016459; 22%) and ‘6-phosphofructo-kinase complex’ (GO:0005945; 22%) within ‘cellular component’, and ‘protein binding’ (GO:0005515; 17.6%) and ‘structural molecule activity’ (GO:0005198; 13.5%) within ‘molecular function’ (p < 0.05) (Appendix 2.1).

KOBAS analysis of the C. elegans homologues inferred from the 30 ‘female’ molecules were predicted to be involved predominantly in focal adhesion (21%), and other replication, recombination and repair proteins (21%). Eleven (37%) of these 30 molecules were associated with (non-wild-type) RNAi phenotypes, mainly embryonic lethality (Emb) (three types; 55%) (Table 2.1). KOBAS pathway mapping of C. elegans homologues inferred from the 73 ‘male’ molecules predicted their involvement to be mainly in insulin-signalling pathway (12%) and long-term potentiation (11%), although 24 (33%) of them could not be mapped to a specific metabolic pathway (see Table 2.2). Functionally, 37 (51%) of these 73 molecules were associated with (non-wild-type) RNAi phenotypes in C. elegans, the majority displaying sterility (Ste) (35%), embryonic lethality (Emb) (four types; 33%) and larval arrest (Lva) (27%) (Table 2.2).

The genetic interaction predictions for C. elegans orthologues linked to nine ‘female’ and 28 ‘male’ sequences from As. suum revealed, with high statistical support, interactions among 52 and 680 genes, respectively (Appendix 2.2). The GO classification of the 52 C. elegans genes predicted to interact with those representing the nine ‘female’ ESTs were: ‘embryonic development ending in birth or egg hatching’, ‘reproduction’, ‘positive regulation of growth rate’ and/or ‘nematode larval development’ as the most common terms for ‘biological function’, ‘cytoplasm’ and ‘nucleus’ for ‘cellular component’ and ‘protein binding’ and ‘nucleotide binding’ for ‘molecular function’ (p < 0.05). The GO classifications of the 680 C. elegans genes predicted to interact with those representing the 28 ‘male’ ESTs were: ‘embryonic development ending in birth or egg hatching’, ‘nematode larval development’, ‘reproduction’ and ‘positive regulation of growth rate’ for ‘biological process’, ‘nucleus’
and 'integral to membrane' for 'cellular component' and 'ATP binding' and 'protein binding' for 'molecular function' (p < 0.05) (see Appendix 2.2 and www.wormbase.org).

Most 'female' sequences from As. suum encoded C. elegans vitellogenins (n = 6) and chitin-binding proteins (n = 4) (Table 2.1). Vitellogenins are large proteins which are expressed abundantly in the maternal intestinal epithelium [54] and accumulated in the developing oocyte, where they play a key role in lipid transport and metabolism and are essential for the survival and growth of the embryo [55]. Vitellogenins are commonly found among the most abundant transcripts in female-enriched EST datasets in parasitic nematodes, including strongylids [29,38,51,56] and filarioids [57-59]. In the present study, C. elegans homologues of the As. suum vitellogenins vit-2, vit-5 and vit-6 were predicted to interact, with high statistical support, with eight other genes involved in 'embryonic development ending in birth or egg hatching', 'reproduction', 'positive regulation of growth rate' and 'nematode larval development' (see Appendix 2.2).

Four C. elegans homologues (encoded by the C. elegans genes cbd-1, C39D10.7, cpg-1 and R02F2.4; see Table 2.1) containing chitin-binding peritrophin A domains (InterPro: IPR002557; Pfam: PF01607.13) were also amongst the molecules abundantly transcribed in the female of As. suum (see Table 2.1). The peritrophin-A domain is commonly found in chitin binding proteins, particularly the peritrophic matrix (or peritrophic membrane) of the gut of insects and other animals, in which they play important roles in the facilitation of the digestive processes and the protection of the gut from invasion by micro-organisms and parasites [60]. In C. elegans, a number of protein sequences containing one or more peritrophin-A domains have been identified (see www.wormbase.org). Among these, the chondroitin proteoglycans (e.g., CPG-1/CE)-1 and CPG-2) have been shown to be essential for the synthesis of the eggshell and early embryonic development [61]. The gut of mid-to-late embryo of C. elegans has been demonstrated to contain lysosome-related organelles with birefringent content, which has been hypothesized to play essential roles in the storage of fat in 'gut-granules' [62]. A C. elegans multdrug resistance factor (i.e., MRP-4) has been localized to the germline (i.e., oocytes and early embryos), in which it might play a role in the processes leading to the formation of birefringent gut granule-associated material in the developing larva [62]. A homologue of the C. elegans mrp-4 was amongst the molecules differentially transcribed in female As. suum (Table 2.1), and another homologue of mrp-4 was also found to be enriched in the adult female Brugia malayi [59]. Although the specific function of mrp-4 in the germline of C. elegans is presently unknown, its high level of transcription in females of parasitic nematodes suggests an
essential role in the physiology of the reproductive process, thus deserving further investigation.

The majority of *As. suum* male-enriched molecules were orthologues/homologues of *C. elegans* major sperm proteins (MSPs; n = 7) and protein kinases and phosphatases (n = 11) (Table 2.2), which were also highly represented in other species of strongyloid nematodes, such as *Trichostrongylus vitrinos*, *Oesophagostomum dentatum* and *Haemonchus contortus* [38,51,56] and the filarioid *B. malayi* [58,59,63]. The MSPs have been relatively well studied in *As. suum* [64-66], and they have also been recognized as bipartite signalling molecules, activating pathways linked to oocyte production and maturation in *C. elegans* [67,68]. However, MSPs are predominantly involved in nematode sperm motility through their polymerization in dense filaments which are required for the movement of the plasma membrane of the amoeboid sperm cells [64-66]. The correct assembly of MSPs in the motility apparatus of the amoeboid sperm of *As. suum* has been shown to be dependent upon the activity of a group of serine/threonine kinases (i.e., MSP polymerization-activating kinases, MPAKs), which interact with the MSPs to induce membrane associated polymerization [69].

Protein kinases and phosphatases were also amongst the most highly represented molecules in the male-enriched ESTs from *As. suum*, in accordance with the findings of previous analyses of gender-enriched molecules in other parasitic nematodes, such as *Oe. dentatum* [56], *T. vitrinos* [29,38] and *B. malayi* [59]. In *C. elegans*, such kinases and phosphatases are abundant in sperm-producing germline tissue, where they are likely to be involved in regulating sperm maturation by post-translational modification [15] and in signalling cascades and/or protein modification within the oocyte following fertilization [70]. In particular, a group of molecules homologous to *C. elegans* serine/threonine specific protein phosphatases PP1 (encoded by *gsp-3, gsp-4, pph-1* and F23B12.1; see Table 2.2) was male-enriched in *As. suum* as well as in other parasitic nematodes [51,56,70,71]. Indeed, male-enriched serine/threonine protein phosphatases have been characterized in *T. vitrinos* (i.e., *Tv-stp-1*; [71]) and *Oe. dentatum* (i.e., *Od-mpp1*; [70]). Functional analysis of *C. elegans* homologues of *Od-mpp1* in *C. elegans* showed a significant reduction (30-40%) in the numbers of F2 progeny produced [70], confirming the finding from a previous study in which gene silencing by dsRNAi of these genes resulted in the impairment of the function of the sperm of the hermaphrodite worm [72]. Given their essentiality, such serine/threonine protein phosphatases might represent targets for novel anthelmintics [73,74].

64
A significant group of molecules (n = 13) in the male-enriched ESTs from *As.
*suum was represented by enzymes predicted (by KOBAS analysis) to be involved in
amino acid and carbohydrate metabolism (see Table 2.2). This involvement could be
associated with the high energy supply, required for the production and motility of the
sperm [56]. Genes encoding molecules essential for cellular movement (i.e., actins and
myosins) were also represented in the male-enriched ESTs from *As. suum* (n = 4; see
Table 2.2). This class is also enriched in the transcriptome of male *B. malayi* [59],
whereas differential transcription was not recorded for other nematode species
[38,51,56]. In the trematodes *Schistosoma mansoni* and *S. japonicum*, actins and
myosins are amongst the most up-regulated molecules in adult male worms [75-77]. As
actin polymers provide mechanical stability for the cytoskeleton and serve as tracks for
motor proteins, such as myosins, it has been proposed that the bias in the expression of
these molecules in male schistosomes might be associated with (i) males being
significantly larger than females (with a greater volume of tegument and muscle), (ii)
males being more physically active than females, and (iii) the preferential transcription
of actins and myosins in the male having an important developmental role in both
sexes during copulation [75]. However, as the content of tegumental and muscular
tissues does not seem to be biased according to gender in *As. suum* and *B. malayi*, actins
and myosins might play a specific, but yet unknown role in the male reproductive
process in these nematode species. This aspect warrants investigation using a
functional genomic approach, such as RNAi or cross-species RNAi [33,34].

A *C. elegans* homologue encoding a defence-related protein containing a sperm
coating protein (SCP) - extracellular domain (i.e., vap-I; cf. Table 2.2) was among the
molecules differentially transcribed in male *As. suum*. The SCP proteins belong to a
large protein family, called SCP/Tpx-1/Ag5/PR-1/Sc7 (SCP/TAPS; Pfam: PF00188). SCP/TAPS family members have been identified in various eukaryotes, including
plants, arthropods, snakes, mammals and free-living and parasitic helminths [78]. The
*As. suum* orthologue/homologue identified herein shows high sequence homology to
molecules identified in other parasitic nematodes, including hookworms, other
strongylids and filarioïds, and is phylogenetically linked to the parasite clade of the
SCP/TAPS (not shown). In ascaridoid nematodes, SCP/TAPS molecules have been
identified previously in an EST dataset representing the arrested larval stage of *T. canis*,
the common roundworm of canids [79]. However, in parasitic nematodes, these
molecules have been studied mainly in the hookworms *Ancylostoma caninum* and
*Necator americanus* and are commonly referred as to *Ancylostoma*-secreted proteins
(i.e., ASPs; [78]). Due to their abundance in the excretory/secretory (ES) products in
serum-activated third stage larvae (aL3s) of *An. caninum* and to the high transcriptional
levels of mRNAs encoding ASPs in aL3s compared with non-activated L3s, these molecules have been hypothesized to play a major role in the transition from the free-living to the parasitic stage of hookworms [80]. Other ASP orthologues/homologues have been characterised in the adult stage of hookworms, and they might play a role in the initiation, establishment and/or maintenance of the host-parasite relationship [78,81-83]. A gender-biased transcription of ASP homologues/orthologues has been reported previously for the strongylids *T. vitrinus* [38], *Oe. dentatum* [56] and *Ostertagia ostertagi* [84]. However, in the latter species, the protein homologues *Oo-ASP1* and *Oo-ASP2* were localised to the reproductive tract of both male and female worms [84]. Although the role of ASPs in the physiology of reproduction is currently unknown, the high transcriptional level of mRNAs encoding ASPs in male parasitic nematodes is an intriguing aspect, which warrants further investigation. It might also be possible to assess the function of ASPs in *As. suum*, as gene silencing has been reported recently for this nematode [33]. In addition, opportunities to investigate the function of ASPs will be enhanced when the complete genomes and transcriptomes of a wide range of parasitic nematodes are available, which would also underpin meaningful proteomic analyses of differentially expressed molecules.
2.4 References


32. Douvres FW, Urban JF Jr. Factors contributing to the in vitro development of *Ascaris suum* from second-stage larvae to mature adults. J Parasitol 1983;69:549-


57. Michański ML, Monsey JD, Cistola DP, Weil GJ. An embryo-associated fatty acid-binding protein in the filarial nematode Brugia malayi. Mol Biochem Parasitol 70


82. Zhan B, Liu Y, Badamchian M, Williamson A, Feng J, et al Molecular characterisation of the *Ancylostoma*-secreted protein family from the adult stage


<table>
<thead>
<tr>
<th>EST code</th>
<th>Size (bases)</th>
<th>In silico peptide analysis</th>
<th>Description of <em>C. elegans</em> homologue (gene code, gene name)</th>
<th>RNAi phenotypes&lt;sup&gt;b&lt;/sup&gt;</th>
<th>InterProScan analysis</th>
<th>KOBAS analysis</th>
<th>Other ascaridoid nematodes [non-ascaridoid nematodes]&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Contigs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contig1</td>
<td>717</td>
<td>A/0/Ex</td>
<td>Unnamed protein with 12 chitin-binding peritrophin-A domains (H02115.2, <em>cbd-1</em>)</td>
<td>Emb&lt;sup&gt;1,2&lt;/sup&gt;, Ste, pattern of transgene expression abnormal, receptor mediated endocytosis defective</td>
<td>Invertebrate chitin-binding proteins</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contig4, Contig17</td>
<td>1250, 290</td>
<td>Q/0/Ex</td>
<td>Vitellogenin precursor (K07H8.6, <em>vit-6</em>)</td>
<td>Nicotine hypersensitive</td>
<td>Lipovitellin-phosvitin complex, superhelical domain, beta-sheet shell regions</td>
<td>Other replication, recombination and repair proteins</td>
<td>[Cbr, Hco, Nbr, Oos, Ppa] [Cbr]</td>
</tr>
<tr>
<td>Contig5, Contig18</td>
<td>386, 543</td>
<td>Q/0/Ex Q/0/PI_Ex</td>
<td>Vitellogenin (C42D8.2, <em>vit-2</em>)</td>
<td>Age</td>
<td>Other replication, recombination and repair proteins</td>
<td>Tca [Hco]</td>
<td></td>
</tr>
<tr>
<td>Contig8, Contig12, Contig19</td>
<td>456, 872, 628</td>
<td>Q/0/Ex Q/0/Ex</td>
<td>Unnamed protein with 11 chitin-binding peritrophin-A domains</td>
<td>Invertebrate chitin-binding proteins</td>
<td>Aminosugars metabolism</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contig9, Contig14</td>
<td>719, 837</td>
<td>S/2/Ex Q/0/Ex</td>
<td>Unnamed protein (R06C7.4, <em>cpg-3</em>)</td>
<td>Age, Emb&lt;sup&gt;1&lt;/sup&gt;, Gro</td>
<td>Lipovitellin-phosvitin complex, superhelical domain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contig21</td>
<td>478</td>
<td>Q/0/Ex</td>
<td>Vitellogenin (K09F5.2, <em>vit-1</em>)</td>
<td>Age, Emb&lt;sup&gt;1&lt;/sup&gt;, Gro</td>
<td>Lipovitellin-phosvitin complex, superhelical domain</td>
<td>Tca</td>
<td></td>
</tr>
<tr>
<td><strong>Singletons</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Folate biosynthesis [Ppa] Tca [Ppa] [Hco, Oos, Ppa, Tvi]</td>
</tr>
<tr>
<td>FE274092_f1027; FE274113_f1080; FE274287_f578; FE274360_f1735</td>
<td>433, 479, 384, 551</td>
<td>Q/0/Ex Q/0/Ex</td>
<td>Vitellogenin precursor (K07H8.6, <em>vit-6</em>)</td>
<td>Nicotine hypersensitive</td>
<td>Lipovitellin-phosvitin complex; beta-sheet shell regions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accession</td>
<td>Chr</td>
<td>Thy/Loc</td>
<td>Description</td>
<td>Function</td>
<td>Reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>-----</td>
<td>---------</td>
<td>-------------</td>
<td>----------</td>
<td>-----------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274100_f1042</td>
<td>710</td>
<td>Q/0/Ex</td>
<td>Putative NAD+ dependent epimerases (C01F1.3a)</td>
<td>Early larval arrest, Egl, Let, Lva, Unc, transgene expression increased, transgene subcellular localization abnormal</td>
<td>[Bma, Cbr, Mha, Min, Wba]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274109_f1070</td>
<td>417</td>
<td>Q/0/Ex</td>
<td>Aspartyl-tRNA synthetase (B0464.1, drs-1)</td>
<td>Emb1,2, Lva, Sck, Ste, pattern of transgene expression abnormal, receptor mediated endocytosis defective</td>
<td>[Cbr, Hgf]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274126_f113</td>
<td>541</td>
<td>S/0/Ex</td>
<td>Uncharacterized protein with conserved cysteine (Y5F2A.1, ttr-16)</td>
<td>Transthyretin-like</td>
<td>Valine, leucine and isoleucine biosynthesis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274184_f271</td>
<td>712</td>
<td>Q/0/Ex</td>
<td>Vitellogenin (C42D8.2, vit-2)</td>
<td>Age</td>
<td>von Willebrand factor, type D</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274236_f405</td>
<td>704</td>
<td>Q/0/Ex</td>
<td>Unnamed protein (F18A1.7)</td>
<td>Emb1</td>
<td>Tca</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274247_f435</td>
<td>876</td>
<td>Q/0/PI</td>
<td>Alpha-crystallin family of the HSP proteins (F52E1.7, hsp-17)</td>
<td>HSP20-like chaperones</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274276_f548</td>
<td>470</td>
<td>Q/0/PI</td>
<td>Unnamed protein with 3 chitin-binding peritrophin-A domains (C07G2.1, cpg-1)</td>
<td>Invertebrate chitin-binding proteins</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274281_f562</td>
<td>781</td>
<td>S/0/Pl</td>
<td>Leucine rich repeat (K07A12.2)</td>
<td>Emb1,2</td>
<td>Proteoglycans</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274310_f628</td>
<td>554</td>
<td>Q/0/Ex</td>
<td>Uncharacterized conserved protein (C50B8.3)</td>
<td>Reduced brood size</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274312_f632</td>
<td>524</td>
<td>S/1/Ex</td>
<td>Unnamed protein (F52E1.5)</td>
<td></td>
<td>Signal transduction mechanisms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274317_f641</td>
<td>372</td>
<td>Q/0/Pl</td>
<td>Transcription factor zerknullt and related HOX domain proteins (R06F6.6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274322_f650</td>
<td>666</td>
<td>Q/0/Ex_Cy_Nu</td>
<td>Putative secreted TIL-domain protease inhibitor (C25E10.9, swm-1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274343_f691</td>
<td>707</td>
<td>A/0/Ex</td>
<td>Unnamed protein with 6 chitin-binding peritrophin-A domains (R02F2.4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274357_f723</td>
<td>700</td>
<td>Q/2/Ex</td>
<td>Multidrug resistance associated protein/mitoxantrone resistance protein, ABC superfamily (F21G4.2, mrp-4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274383_f811</td>
<td>488</td>
<td>S/0/Ex</td>
<td>Oxoglutarate/malate carrier protein (K11G12.5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Abbreviations used in proteomic analysis: Non-secretory protein (Q), secretory protein (S)/ predicted number of transmembrane domains/ predominant cellular location: extracellular (Ex); nuclear (Nu), plasma membrane (Pl), cytoplasm (Cy).

*Abbreviations of RNAi phenotypes (alphabetical): lifespan abnormal (Age), egg laying defective (Egl), embryonic lethal (Emb1), embryo osmotic integrity abnormal (Emb2), pleiotropic defects severe (Emb3), slow growth (Gro), larval lethal (Let), larval arrest (Lar), sick (Sick), sterile (Ste), uncoordinated (Unc).

*Abbreviations of nematode species (alphabetical): Ascaris lumbricoides (Alu), Brugia malayi (Bma), Brugia pahangi (Bpa), Caenorhabditis briggsae (Cbr), Haemonchus contortus (Hco), Heterodera glycines (Hgl), Meloidogyne hapla (Mha), Meloidogyne incognita (Min), Nippostrongylus brasiliensis (Nbr), Ostertagia ostertagi (Oos), Parastrongyloides trichosuri (Ptr), Pristionchus pacificus (Ppa), Strongyloides ratti (Sra), Toxocara canis (Tca), Trichostrongylus vitrinus (Tvi), Wuchereria bancrofti (Wba).
Table 2.2 Bioinformatic analysis of expressed sequence tags (ESTs) representing male-enriched molecules in adult *Ascaris suum* with homologues in *Caenorhabditis elegans* and other nematodes.

<table>
<thead>
<tr>
<th>EST code</th>
<th>Size (bases)</th>
<th>In silico peptide analysisa</th>
<th>Description of <em>C. elegans</em> homologue (gene code, gene name)</th>
<th>RNAi phenotypesb</th>
<th>InterProScan analysis</th>
<th>KOBAS analysis</th>
<th>Other ascaridoid nematodes [non-ascaridoid nematodes]c</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contig1</td>
<td>616</td>
<td>Q/O/Ex</td>
<td>Phosphoenolpyruvate carboxykinase (R11A5.4c)</td>
<td></td>
<td>PEP carboxykinase-like</td>
<td>Alanine and aspartate metabolism; pyruvate metabolism; Citrate cycle (TCA cycle); Insulin signaling pathway; Adipocytokine signaling pathway; PPAR signaling pathway</td>
<td>[Aca, Ace, Cbr, Hco, Mch, Mha, Min, Oos, Ovo, Ppa, Sst, Tci, Tvi, Wba]</td>
</tr>
<tr>
<td>Contig2</td>
<td>728</td>
<td>Q/O/Ex</td>
<td>Protein tyrosine phosphatase (C02B10.6)</td>
<td>(Phosphotyrosine protein) phosphatases II</td>
<td></td>
<td>Signal transduction mechanisms</td>
<td>[Cbr]</td>
</tr>
<tr>
<td>Contig3</td>
<td>806</td>
<td>Q/O/Ex</td>
<td>Unnamed protein (Y69E1A.2)</td>
<td>Emb¹</td>
<td></td>
<td>General function prediction only</td>
<td>[Bma, Cbr, Hco, Ppa, Tci]</td>
</tr>
<tr>
<td>Contig5</td>
<td>747</td>
<td>Q/O/Pl</td>
<td>Ribosomal Protein, Large subunit family member (D1007.12, rpl-24.1)</td>
<td></td>
<td></td>
<td></td>
<td>Alu [Lsi]</td>
</tr>
<tr>
<td>Contig6,</td>
<td>958, 859</td>
<td>Q/O/Pl/Q/O/Ex</td>
<td>Major Sperm Protein family member (T13F2.11, msp-78)</td>
<td>Emb¹, fat content reduced, fat content increased</td>
<td></td>
<td></td>
<td>Alu</td>
</tr>
<tr>
<td>Contig40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contig7</td>
<td>444</td>
<td>Q/O/Ex</td>
<td>Alanine aminotransferase (C32F10.8a)</td>
<td></td>
<td></td>
<td>Alanine and aspartate metabolism; Carbon fixation; Glutamate metabolism</td>
<td>Alu, Tca [Cbr, Ptr, Sra]</td>
</tr>
<tr>
<td>Contig8</td>
<td>515</td>
<td>Q/O/Ex</td>
<td>Uncharacterized conserved protein (T05G5.7, rmd-1)</td>
<td>Emb¹</td>
<td>Protein prenyltransferase PapD-like</td>
<td>Tca [Ace, Gro]</td>
<td></td>
</tr>
<tr>
<td>Contig9</td>
<td>630</td>
<td>Q/4/Ex</td>
<td>Uncharacterized protein, contains MSP domain (F21H7.5)</td>
<td></td>
<td></td>
<td>Alu [Hco]</td>
<td></td>
</tr>
<tr>
<td>Contig13</td>
<td>893</td>
<td>Q/O/Ex</td>
<td>Pyrophosphate dependent phosphofructo-1-kinase (Y71H10A.1a)</td>
<td>Emb¹</td>
<td>Phosphofructokinase</td>
<td>Glycolysis / Gluconeogenesis; Insulin signaling pathway; Galactose metabolism; Pentose phosphate pathway;</td>
<td>[Ace, Bma, Cbr, Gpa, Hco, Mha, Ptr, Tsp]</td>
</tr>
<tr>
<td>Contig</td>
<td>Genomic Location</td>
<td>Gene Name</td>
<td>Protein Description</td>
<td>GO Terms</td>
<td>Genomic Feature</td>
<td>Location</td>
<td>GenBank Accession</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------</td>
<td>-----------</td>
<td>---------------------</td>
<td>-----------</td>
<td>-----------------</td>
<td>----------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Contig14</td>
<td>447</td>
<td>Q/1/Ex</td>
<td>Sperm-specific protein (ZC168.6)</td>
<td>Bmd, Dpy, Unc, unclassified</td>
<td>Thiamin diphosphate-binding fold (THDP-binding)</td>
<td>PapD-like</td>
<td></td>
</tr>
<tr>
<td>Contig16</td>
<td>664</td>
<td>Q/1/Ex</td>
<td>Collagen (F30B5.1, dpy-16)</td>
<td>Gro, Ste, Unc, oocyte morphology abnormal, reduced brood size</td>
<td></td>
<td></td>
<td>Biosynthesis of siderophore group nonribosomal peptides</td>
</tr>
<tr>
<td>Contig18</td>
<td>498</td>
<td>Q/5/Ex</td>
<td>Branched chain alpha-keto acid dehydrogenase E1 (F27D4.5, tag-173)</td>
<td></td>
<td>Thiamin diphosphate-binding fold (THDP-binding)</td>
<td></td>
<td>Pyruvate metabolism; Glycolysis / Gluconeogenesis; Valine, leucine and isoleucine biosynthesis; Butanoate metabolism</td>
</tr>
<tr>
<td>Contig20</td>
<td>740</td>
<td>Q/4/Ex</td>
<td>Unnamed protein (W03D8.9)</td>
<td></td>
<td>Thiamin diphosphate-binding fold (THDP-binding)</td>
<td></td>
<td>Glycine, serine and threonine metabolism</td>
</tr>
<tr>
<td>Contig21</td>
<td>761</td>
<td>Q/0/Ex</td>
<td>Phosphoenolpyruvate carboxykinase (W05G11.6d)</td>
<td></td>
<td>PEP carboxykinase-like</td>
<td></td>
<td>Pyruvate metabolism; Citrate cycle (TCA cycle); Insulin signaling pathway; Adipocytokine signaling pathway; PPAR signaling pathway</td>
</tr>
<tr>
<td>Contig29</td>
<td>500</td>
<td>Q/0/Ex</td>
<td>Heat Shock Protein family member (Y22D7AL.5, hsp-60)</td>
<td>Emb³, Ste</td>
<td></td>
<td></td>
<td>Single hybrid motif</td>
</tr>
<tr>
<td>Contig30</td>
<td>580</td>
<td>Q/1/Ex</td>
<td>Dihydropyrimidine acetyltransferase (F23B12.5)</td>
<td>Emb¹, Gro, Stp, pattern of transgene expression abnormal, receptor mediated endocytosis defective</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contig33</td>
<td>491</td>
<td>Q/1/Ex</td>
<td>Unnamed protein (Y69E1A.2)</td>
<td>Emb¹, Emb¹,², Cyk, Lva, Lvl, Pvl, Sck, Ste, transgene subcellular localization abnormal, distal tip cell</td>
<td></td>
<td></td>
<td>N-Glycan degradation</td>
</tr>
<tr>
<td>Contig34</td>
<td>491</td>
<td>Q/0/Ex</td>
<td>Actin (M03F4.2, act-4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contig</td>
<td>Seq No.</td>
<td>Type</td>
<td>Description</td>
<td>Annotation</td>
<td>Family</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>---------</td>
<td>------</td>
<td>-------------</td>
<td>------------</td>
<td>--------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contig35</td>
<td>477</td>
<td>Q/1/Ex</td>
<td>Uncharacterized protein, contains MSP domain (F53B6.4)</td>
<td>PapD-like migration abnormal</td>
<td>Alu [Hco, Lsi, Nbr, Ppa]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contig37</td>
<td>656</td>
<td>Q/1/Pl</td>
<td>GTP specific succinyl CoA synthetase (C50F7.4)</td>
<td>Succinyl-CoA synthetase domains; Glutathione synthetase ATP-binding domain-like</td>
<td>[Cbr, Hgl, Tmu]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contig38</td>
<td>1240</td>
<td>Q/0/Cy_Nu</td>
<td>Pyruvate dehydrogenase (T05H10.6b)</td>
<td>Emb&lt;sup&gt;1,2&lt;/sup&gt;, Clr, Sck; Gro, pattern of transgene expression abnormal receptor mediated endocytosis defective, reduced brood size, oogenesis abnormal</td>
<td>Thiamin diphosphate-binding fold (THDP-binding) Alanine and aspartate metabolism; pyruvate metabolism; Glycolysis / Gluconeogenesis; Valine, leucine and isoleucine biosynthesis; Butanoate metabolism</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contig39, Contig43</td>
<td>474, 582</td>
<td>Q/0/Pl, Q/0/Ex</td>
<td>Asparagine synthase (M02D8.4c)</td>
<td>Adenine nucleotide alpha hydrolases-like</td>
<td>Alanine and aspartate metabolism; Nitrogen metabolism</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274450_m1040, FE274778_m828</td>
<td>566, 502</td>
<td>Q/0/Ex, Q/0/Ex</td>
<td>C-type lectin (W04E12.6, clec-49)</td>
<td>C-type lectin-like Glycan Binding Proteins; Proteoglycans; Cell adhesion molecules (CAMs)</td>
<td>Alu</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274469_m1109, FE274478_m113</td>
<td>358, 464</td>
<td>A/1/Ex, Q/0/Ex</td>
<td>Unnamed protein (F22B5.4), Fructose biphosphatase aldolase (T05D4.1)</td>
<td>Unclassified Aldolase Glycolysis / Gluconeogenesis; Pentose phosphate pathway; Fructose and mannose metabolism; Carbon fixation</td>
<td>[Ptr, Sra, Sst] Tca [Bma, Hco, Lsi, Ptr, Tct]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274483_m114</td>
<td>608</td>
<td>Q/0/Ex</td>
<td>Serine threonine specific protein phosphatase PP1 (W09C3.6, gsp-3)</td>
<td>Ste Metallo-dependent phosphatases Long-term potentiation; Insulin signaling pathway; Regulation of actin</td>
<td>[Ace, Bma, Cbr, Gpa, Gro, Hco, Hgl, Min, Ppa, Sra, Sst, Tmu, Tvi, Xin]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274487_m1148</td>
<td>454</td>
<td>Q/O/Ex</td>
<td>Asparagine synthase (M02D8.4c)</td>
<td>Adenine nucleotide alpha hydrolases-like RNA-binding domain, RBD</td>
<td>cytoskeleton; Focal adhesion</td>
<td>[Aca, Cbr, Min, Ppa]</td>
<td></td>
</tr>
<tr>
<td>FE274492_m1158</td>
<td>671</td>
<td>Q/O/Ex</td>
<td>Polyadenylate binding protein (Y106G6H.2, pab-1)</td>
<td>Bmd, Fgc, Gro, Pvl, Rup, Ste, Stp, transposon silencing abnormal, reproductive system morphology abnormal, reduced brood size, pattern of transgene expression abnormal, receptor mediated endocytosis defective</td>
<td>Proteins, nucleotide metabolism</td>
<td>[Ace, Cbr, Hco, Hgl, Mar, Min, Mpa, Pvu]</td>
<td></td>
</tr>
<tr>
<td>FE274497_m1166</td>
<td>617</td>
<td>Q/O/Ex</td>
<td>Unnamed protein (B0207.11)</td>
<td>Galactose oxidase, central domain</td>
<td>Ubiquitin enzymes</td>
<td>[Alu]</td>
<td></td>
</tr>
<tr>
<td>FE274496_m1168</td>
<td>484</td>
<td>Q/O/Ex</td>
<td>KELch-repeate containing protein (T16H12.6, kel-10)</td>
<td>Sck, fat content increased, small, sterile progeny</td>
<td></td>
<td>[Ace, Tci]</td>
<td></td>
</tr>
<tr>
<td>FE274500_m117, FE274641_m403</td>
<td>480, 328</td>
<td>S/O/Ex</td>
<td>Uncharacterized protein, contains MSP domain (K07F5.9, ssp-10)</td>
<td>Sck, fat content increased, small, sterile progeny</td>
<td></td>
<td>[Bma, Lsi, Ptr, Tci]</td>
<td></td>
</tr>
<tr>
<td>FE274508_m119</td>
<td>507</td>
<td>Q/O/Pl</td>
<td>Translation initiation factor 4F (F57B9.6, inf-1)</td>
<td>P-loop containing nucleoside triphosphatase hydrolases</td>
<td></td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>Accession</td>
<td>Start</td>
<td>Type</td>
<td>Description</td>
<td>Function</td>
<td>Genetic Interactions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>-------</td>
<td>------</td>
<td>-------------</td>
<td>----------</td>
<td>---------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274515_m1211</td>
<td>572</td>
<td>Q/0/Ex</td>
<td>Transcription factor of the Forkhead HNF3 family (R13H8.1, daf-16)</td>
<td>Subcellular localization abnormal</td>
<td>[Aca, Lsi, Ovo]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274539_m1297</td>
<td>689</td>
<td>Q/0/Mi</td>
<td>Serpin (F20D6.4, srp-7)</td>
<td>Emb&lt;sup&gt;1&lt;/sup&gt;</td>
<td>Serpins</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274540_m13</td>
<td>193</td>
<td>Q/0/Ex</td>
<td>Pyrophosphate dependent phosphofructo-1-kinase (Y71H10A.1a)</td>
<td>Emb&lt;sup&gt;1&lt;/sup&gt;</td>
<td>Phosphofructokinase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274547_m1324</td>
<td>488</td>
<td>Q/0/Ex</td>
<td>Unnamed protein (Y69E1A.2)</td>
<td>Emb&lt;sup&gt;1&lt;/sup&gt;</td>
<td>Metallo-dependent phosphatases</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274550_m1329</td>
<td>651</td>
<td>Q/0/Ex</td>
<td>Serine threonine specific protein phosphatase PP1 (T03F1.5, gsp-4)</td>
<td>Emb&lt;sup&gt;1&lt;/sup&gt;</td>
<td>Long-term potentiation; Insulin signaling pathway; Regulation of actin cytoskeleton; Focal adhesion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274556_m158</td>
<td>320</td>
<td>Q/0/Ex</td>
<td>Ribosomal Protein, Small subunit family member (F37C12.11, rps-21)</td>
<td>Emb&lt;sup&gt;1,4&lt;/sup&gt; Gro, Lva, Sck, Ste, transgene subcellular localization abnormal</td>
<td>[Bma, Cbr, Hco, Ovo, Sra, Tci, Tvl]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274564_m182</td>
<td>679</td>
<td>Q/0/Ex</td>
<td>Secreted surface protein (T05B4.12)</td>
<td>Emb&lt;sup&gt;1,3&lt;/sup&gt;, Cyk, Lva, Lvl, Pvl, Sck, Ste, transgene</td>
<td>Fatty acid biosynthesis; N-Glycan degradation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274566_m187</td>
<td>383</td>
<td>Q/1/Ex</td>
<td>Actin (M03F4.2, act-4)</td>
<td>Emb&lt;sup&gt;1&lt;/sup&gt;</td>
<td>[Bma, Ovo]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>ID</td>
<td>Q/C</td>
<td>Description</td>
<td>Localization/Function</td>
<td>Similarities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>------</td>
<td>------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------</td>
<td>--------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274574_m200</td>
<td>341</td>
<td>Q/0/Ex</td>
<td>Uncharacterized protein, contains MSP domain (ZC168.6)</td>
<td>PapD-like</td>
<td>Alu</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274582_m211</td>
<td>569</td>
<td>Q/0/Pl</td>
<td>Uncharacterized protein, contains PDZ domain (C25G4.6)</td>
<td>Lva, Pvl, Ste, Stp Emb1, Let, Lvl, Ste, transgene</td>
<td>Tca</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274588_m229</td>
<td>608</td>
<td>S/0/Ex</td>
<td>AAA+ ATPase (H15N14.2, nsf-1)</td>
<td>Subcellular localization abnormal, distal tip cell migration abnormal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274590_m235</td>
<td>626, 700</td>
<td>Q/0/Ex</td>
<td>Serine threonine specific protein phosphatase PP1 (C05A2.1, pph-1)</td>
<td>Metallo-dependent phosphatases</td>
<td>[Bma, Hgl, Min, Tvu]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274590_m237</td>
<td>520</td>
<td>Q/0/Ex</td>
<td>Ribonuclease H (ZK1290.6, rnh-1.1)</td>
<td>Ribonuclease H-like</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274594_m252</td>
<td>277</td>
<td>S/0/Ex</td>
<td>Defense-related protein containing SCP domain (F11C7.3, vap-1)</td>
<td>PR-1 like</td>
<td>[Bma, Lsi, Mch, Ovo, Wba]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274599_m268</td>
<td>673</td>
<td>Q/0/Ex</td>
<td>Uncharacterized conserved protein (Y67D2.3)</td>
<td>Snake toxin-like</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274605_m282</td>
<td>595</td>
<td>Q/0/Ex</td>
<td>Casein kinase (Y71F9AL.2)</td>
<td>Reduced brood size, spontaneous mutation rate increased</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FE274610_m301</td>
<td>637</td>
<td>Q/0/Ex</td>
<td>Branched chain alpha dehydrogenase (Y39E4A.3a)</td>
<td>Thiamin diphosphate-binding fold</td>
<td>[Ace, Cbr, Sra, Tmu]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

82
| FE274630_m373 | 539 | Q/0/Ex | MinK related peptide (C29F5.4, mps-1) | Emb*1, unclassified Emb*4, Lva, Ste, transgene subcellular localization abnormal (THDP-binding) |
| FE274631_m374 | 530 | Q/0/Pl | Small ribosomal subunit S2 protein (C49H3.11, rps-2) | dsRNA-binding domain-like Ribosome [Bma, Ovo, Ptr, Tmu, Wba, Xin] |
| FE274634_m388 | 573 | Q/0/Pl Ex | Immunoglobulin superfamily protein (Y48A6A.1, zig-5) | Immunoglobulin Protein kinases [Alu [Ace, Cbr, Bma, Bpa, Gpa, Gro, Hco, Hgl, Mar, Mch, Mha, Min, Oos, Ppa, Sra, Sst, Xin] |
| FE274647_m419 | 534 | Q/0/Ex | Molecular chaperone (C47E8.5, daf-21) | Esp, Ste, unclassified Ribosomal protein S5 domain 2-like | Progesterone-mediated oocyte maturation; Antigen processing and presentation; Prostate cancer; Chaperones and folding catalysts |
| FE274654_m445 | 666 | Q/0/Ex | Casein kinase (C39H7.1) | Protein kinase-like (PK-like) RING/U-box Protein kinases [Bma, Cbr, Hgl, Lsi, Sra, Tmu, Tvi, Tvu] |
| FE274658_m453 | 455 | Q/0/Ex | Unnamed protein (F58E6.1) | Ste, reduced brood size, oocyte morphology abnormal Ubiquitin enzymes [Bma] |
| FE274661_m463 | 702 | Q/0/Pl | Myosin class IV heavy chain (F47G6.4, spe-15) | P-loop containing nucleoside triphosphate hydrolases |
| FE274668_m478 | 639 | Q/0/Ex | Unnamed protein (C27D8.2) | Metallo-dependent phosphatases Long-term potentiation; Insulin signaling pathway; Regulation of actin cytoskeleton; Focal adhesion Glycolysis / Gluconeogenesis; Pentose phosphate pathway; Fructose and mannose metabolism; Carbon fixation [Lsi] |
| FE274675_m490 | 635 | Q/0/Pl Ex | Serine threonine specific protein phosphatase PP1 (F23B12.1) | |
| FE274683_m506 | 277 | Q/0/Ex | Fructose biphosphatase aldolase (T05D4.1) | Aldolase [Ace, Cbr, Gro, Hgl, Hsc, Ppa, Ptr, Tci, Tsp, Tvi, Xin] |
| FE274687_m518 | 579 | Q/0/Ex | Unnamed protein (K07H8.5) | Egl Purine metabolism [Ovo] |
| FE274705_m572 | 620 | Q/0/Pl | Unnamed protein (C18G1.9) | Emb*1, Bmd, Lva, Slu, Stp, Unc, pattern of transgene expression Tight junction; Cytoskeleton proteins [Tca [Bma, Cbr, Gro, Hco, Mar, Mch, Min, Ovo, Ppa, Sra, Sst]] |
| FE274706_m577 | 692 | Q/0/Ex | Myosin class II heavy chain (K12F2.1, myo-3) | |

83
<table>
<thead>
<tr>
<th>mmPLOCID</th>
<th>acc</th>
<th>Q/O</th>
<th>Gene</th>
<th>Function/Pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>FE274723_m629</td>
<td>589</td>
<td>Q/O</td>
<td>Dihydrolipoamide acetyltransferase (F23B12.5)</td>
<td>Emb(^a), Gro, Stp, pattern of transgene expression abnormal, receptor mediated endocytosis defective CoA-dependent acyltransferases Glycolysis / Gluconeogenesis; Alanine and aspartate metabolism; Pyruvate metabolism</td>
</tr>
<tr>
<td>FE274740_m675</td>
<td>689</td>
<td>Q/O</td>
<td>Cyclic AMP-dependent protein kinase (F47F2.1)</td>
<td>Protein kinase-like (PK-like) Long-term potentiation; Insulin signaling pathway; Protein kinases; Taste transduction; Progesterone-mediated oocyte maturation; Olfactory transduction; Apoptosis; Hedgehog signaling pathway; MAPK signaling pathway; Melanogenesis; Calcium signaling pathway; GnRH signaling pathway; Wnt signaling pathway; Gap junction Protein kinases [Ace, Chr, Lsi, Min, Ovo, Ptr, Sst, Tvu, Xin]</td>
</tr>
<tr>
<td>FE274742_m681</td>
<td>455</td>
<td>Q/O</td>
<td>Protein tyrosine kinase (F53G12.6, spe-8)</td>
<td>Egl, anoxia hypersensitiv e Protein kinase-like (PK-like)</td>
</tr>
<tr>
<td>FE274787_m847</td>
<td>305</td>
<td>Q/O</td>
<td>Major sperm protein (T13F2.11, msp-78)</td>
<td>Emb(^a), fat content reduced, fat content increased Tight junction [Lsi, Mpa]</td>
</tr>
<tr>
<td>FE274789_m853</td>
<td>357</td>
<td>Q/O</td>
<td>Mitochondrial phosphate carrier protein (F01G4.6)</td>
<td>Emb(^a), Gro, Lva, Pvl, Ste, reduced brood size Other ion-coupled transporters; PPAR signaling pathway [Aca, Ace, Gro, Hco, Nam, Ppa, Ptf, Sra]</td>
</tr>
<tr>
<td>FE274814_m951</td>
<td>500</td>
<td>Q/0/Ex</td>
<td>Unnamed protein (F49E11.7)</td>
<td>Emb&lt;sup&gt;1&lt;/sup&gt;, Lva, reduced brood size</td>
</tr>
</tbody>
</table>

*Abbreviations used in proteomic analysis: Non-secretory protein (Q), secretory protein (S)/ predicted number of transmembrane domains/ predominant cellular location: extracellular (Ex); nuclear (Nu), plasma membrane (Pl). |

*Abbreviations of RNAi phenotypes (alphabetical): lifespan abnormal (Age), body morphology defect (Bmd), clear (Clr), cytokinesis abnormal (Cyk), dumpy (Dpy), egg laying defective (Egl), embryonic lethal (Emb<sup>1</sup>), general pace of development abnormal early emb (Emb<sup>2</sup>), asymmetric cell division abnormal early emb (Emb<sup>3</sup>), pleiotropic defects severe early emb (Emb<sup>4</sup>), pathogen susceptibility increased (Esp), fewer germ cells (Fgc), slow growth (Gro), larval lethal (Let), larval arrest (Lva), early larva lethal (Lvl), protruding vulva (Pvl), exploded through vulva (Rup), sick (Sck), sluggish (Slu), sterile (Ste), sterile progeny (Stp), uncoordinated (Unc). |

Fig 2.1 Relationships of expressed sequence tags (ESTs) encoding female (red dots) and male (blue dots)-enriched molecules in *Ascaris suum* with homologous proteins from *Caenorhabditis elegans*, parasitic nematodes and non-nematodes, displayed in a SimiTri plot [40].
Fig. 2.2 Top twenty genetic interactions (black squares) predicted for a subset of *Caenorhabditis elegans* orthologues of expressed sequence tags (ESTs) representing female (n = 5; red squares) and male (n = 8; blue squares)-enriched molecules in *Ascaris suum.*
CHAPTER 3

Use of a next-generation sequencing approach to elucidate the transcriptome of the economically important parasitic nematode, *Trichostrongylus colubriformis*¹

Abstract

*Trichostrongylus colubriformis* (Strongylida), a small intestinal nematode of small ruminants, is a major cause of production and economic losses in many countries. The aims of the present chapter were to define the transcriptome of the adult stage of *T. colubriformis*, using 454 sequencing technology and bioinformatic analyses, and to predict the main pathways that key groups of molecules are linked to in this nematode. A total of 21,259 contigs was assembled from the sequence data produced from a normalised cDNA library; 7,876 of these contigs had known orthologues in the free-living nematode *Caenorhabditis elegans*, and encoded, amongst others, proteins with ‘transthryretin-like’ (8.8%), ‘RNA recognition’ (8.4%) and ‘metrindin-like ShK toxin’ (7.6%) motifs. Bioinformatic analyses inferred that relatively high proportions of the *C. elegans* homologues are involved in biological pathways linked to ‘peptidases’ (4%), ‘ribosome’ (3.6%) and ‘oxidative phosphorylation’ (3%). Highly represented were peptides predicted to be associated with the nervous system, digestion of host proteins or inhibition of host proteases. Probabilistic functional gene networking of the complement of *C. elegans* orthologues (n = 2,126) assigned significance to particular subsets of molecules, such as protein kinases and serine/threonine phosphatases. The present chapter represents the first, comprehensive insight into the transcriptome of adult *T. colubriformis*, which provides a foundation for fundamental studies of the molecular biology and biochemistry of this parasitic nematode as well as prospects for identifying targets for novel nematocides. Future investigations should focus on comparing the transcriptomes of different developmental stages, both genders and various tissues of this parasitic nematode for the prediction of essential genes/gene products that are specific to nematodes.

¹ New nucleotide sequence data used in this chapter have been deposited in the GenBank Sequence Read Archive database under accession number SRP002574.
3.1 Introduction

Parasitic nematodes of livestock animals are of major socio-economic importance worldwide due to the diseases and associated production losses that they cause (Chapter 1). The nematode *Trichostrongylus colubriformis* (Strongylida, Trichostrongylidae), is amongst the most important parasites of small ruminants, and can be a major cause of economic losses [1]. Its life cycle is direct, with morulated eggs being passed in the faeces of the host. Under suitable environmental conditions (i.e., 18 to 21°C, 100% humidity; [2]), the first-stage larvae (L1s) hatch from eggs to then develop (*via* the second stage, L2) to infective, third-stage larvae (L3s). The cuticle of the L2 is retained as a sheath around the L3 and protects it from desiccation [2]. Infective L3s are ingested with herbage by the host, pass through the forestomachs and undergo an exsheathment process. This process is triggered by the pepsin/HCl in the abomasum, stimulating receptors in the L3 to produce exsheathment fluids [2]. The exsheathed L3 penetrate the mucosa of the small intestine and moult to the fourth-stage larva (L4), which return to the intestinal lumen and develop to adult males and females within ~3 weeks following the ingestion of L3s [2]. Adult *T. colubriformis* live in mucus-covered tunnels in the mucosal surface of the small intestine, where they feed on chyme components [3]. Heavy infections are associated with severe enteritis, characterized by extensive villus atrophy, mucosal thickening and erosion and infiltration of lymphocytes and neutrophils into affected mucosal areas [3]. Clinical signs of trichostrongylosis include malabsorption, weight loss and diarrhoea (= scouring or “black scour”).

Traditionally, the control of *T. colubriformis* infection and trichostrongylosis has relied heavily on the administration of anthelmintics. The excessive and suppressive use of such drugs [4,5] has led to major problems with anthelmintic resistance [6,7]. Attempts to develop effective vaccines to circumvent resistance problems have largely been unsuccessful to date [7,8]. Therefore, there is a continuous need to identify molecular targets for the development of new and efficacious nematocides. A detailed understanding of the complement of molecules transcribed in the adult stage of this parasitic nematode could provide a basis for the identification or prevalidation of essential genes and gene products for the subsequent design of such nematocides.

Investigations of the transcriptome of parasitic nematodes using different approaches [9] is gradually leading to a better understanding of the biochemical and molecular processes involved in parasite development, reproduction and
interactions with their host/s [9-14]. In particular, next-generation sequencing technologies, such as 454-Roche (www.454.com; [15]), ABI-SOLiD (www.appliedbiosystems.com; [16]), Illumina-Solexa (www.illumina.com; [17]) and Helicos (www.helicosbio.com; [18]) are changing the way we discover and define parasite transcriptomes and genomes [19,20]. These advances in sequencing techniques are reflected in the development of enhanced computational methods for the pre-processing, assembly and annotation of sequence data [21-23]. Furthermore, the availability of the entire genome sequences of other helminths, such as the free-living nematode Caenorhabditis elegans, for which detailed information of, for example, molecular and biochemical aspects of development, metabolism and reproduction is available (see www.wormbase.org) for comparative purposes, are allowing the elucidation of fundamental aspects of the biology of parasitic nematodes of public and veterinary health importance [9,11,24].

Despite the substantial economic impact of trichostrongylosis in livestock (e.g., [25]), no genomic and transcriptomic information for T. colubriformis is available in public databases. Gaining an improved understanding of fundamental molecular pathways linked to parasite survival in the environment, development and reproduction in the vertebrate host and host-parasite interactions will assist in finding new ways of disrupting these pathways and thus facilitate the identification of new drug targets. The present chapter (i) describes the production of the first, large-scale transcriptomic dataset for adult T. colubriformis using a next-generation sequencing-based approach, (ii) subjects these data to detailed bioinformatic exploration, and (iii) predicts key pathways and groups of molecules involved in fundamental metabolic pathways of the biology in this nematode.

3.2 Materials and methods

3.2.1 Parasite material

Merino lambs (8–12 weeks of age), maintained under helminth-free conditions, were inoculated intra-ruminally with 10,000 infective third-stage larvae (L3) of T. colubriformis (McMaster strain; Animal Ethics Approval Number 707528, The University of Melbourne). The patency of the infection (~21–25 days) was established based on the detection of strongylid eggs in the faeces using the McMaster flotation method [26]. For the collection of adult worms, infected lambs were euthanised with an overdose of pentobarbitone sodium (Lethobarb, Virbac Pty. Ltd.), administered intravenously 30 days after
inoculation. Adult worms were immediately collected from the first 4 m of the small intestine, washed extensively in phosphate-buffered saline (PBS; pH 7.4), and snap frozen in liquid nitrogen for subsequent storage at -70 °C.

3.2.2 Preparation of 3′-cDNA from Trichostrongylus colubriformis for 454 sequencing

Total RNA from adult female and male worms was prepared using TRIzol Reagent (GibcoBRL, Life Technologies or Invitrogen, Carlsbad, CA) following the manufacturers’ instructions and treated with Ambion Turbo DNase (Ambion/Applied Biosystems, Austin, TX). The integrity of the RNA was verified using the Bioanalyzer 2100 (Agilent Technologies, Cedar Creek, Texas), 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 using the SMART™ kit (Clontech/Takara Bio, CA). An optimized PCR cycling protocol (over 20 cycles) was used to amplify full-length cDNAs employing primers complementary to the SMART IIA-Probe and custom oligo(dT) and the Advantage-HF 2 polymerase mix (Clontech/Takara). The cDNA was then normalised by denaturation-reassociation, treated with duplex-specific nuclease (Trimmer kit, Evrogen, CA) and amplified over 14 cycles. Subsequently, the 5′- and 3′- adaptors were removed by digestion with the exonuclease Mme1 and streptavidin-coated paramagnetic beads [27]. The normalised cDNA (500-700 bases) was then amplified using 9 cycles of Long and Accurate (LA)-PCR [28] and then sequenced in a Genome Sequencer™ (GS) FLX instrument (Roche Diagnostics), employing a standard protocol [15].

3.2.3 Bioinformatic analysis of sequence data

Sequences from the normalised cDNA library for T. colubriformis were aligned and clustered using the Contig Assembly Program v.3, CAP3 [29], employing a minimum sequence overlap length cut-off of 30 bases and an identity threshold of 95%, and assembled. Following the pre-processing of the expressed sequence tags (ESTs), T. colubriformis contigs in the dataset were subjected to BLASTx (NCBI; www.ncbi.nlm.nih.gov) and BLASTn (EMBL-EBI Parasite Genome Blast Server; www.ebi.ac.uk) to identify putative homologues in C. elegans, other nematodes and organisms other than nematodes (e-value cut-off: ≤ 1e-05). WormBase WS200 (www.wormbase.org) was interrogated extensively for relevant information on C. elegans orthologues/homologues, including RNAi phenotypic, transcriptomic, proteomic and interactomic data. T. colubriformis contigs were conceptually translated into peptides using the program ESTScan.
Peptides were classified using InterProScan (domain/motifs) and gene ontology (GO; [30]), and mapped to respective pathways in C. elegans using the KEGG-Orthology Based Annotation System (KOBAS; [31]). The open reading frames (ORFs) inferred from ESTs with orthologues in C. elegans were also subjected to “secretome analysis” using the program SignalP v.2.0 (available at www.cbs.dtu.dk/services/SignalP/), employing both the neural network and hidden Markov models to predict signal peptides and/or anchors [32-34]. Also, transmembrane domains were predicted using the program TMHMM (www.cbs.dtu.dk/services/TMHMM/; [35-37]). Sequence comparisons of peptides predicted from the ESTs of T. colubriformis with those available for C. elegans (WORMPEP v. 202), other parasitic nematodes and organisms other than nematodes in current databases (i.e., www.wormbase.org and www.ncbi.nlm.nih.gov), was performed using SimiTri [38], which provides a two-dimensional display of similarity relationships.

The method developed by Zhong and Sternberg [39] was used to predict the interaction networks among C. elegans orthologues of T. colubriformis contigs. Genomic data (regarding interactions, phenotypes, expression and GO) from C. elegans gene orthologues/homologues, also incorporating data from Drosophila melanogaster (vinegar fly), Saccharomyces cerevisiae (yeast), Mus musculus (mouse) and Homo sapiens (human), were integrated using a naïve Bayesian model to predict genetic interactions among C. elegans genes using the recommended, stringent cut-off value of 4.6 [12,39]. The predicted networks resulting from the analyses were saved in a graphic display file (gdf) format and examined using the graph exploration system available at http://graphexploration.cond.org/.

3.3 Results

A total of 2,674,406 expressed sequence tags (ESTs; average of 328.4 bp ± 276.9 bases in length) were generated. After filtering of the sequences <100 nucleotides in length, the CAP3 assembly yielded 21,259 contigs (average length: 495 bases ± 224.9; Appendix 3.1; sequences available from http://www.nematode.net/ or http://research.vet.unimelb.edu.au/gasserlab/index.html). A total of 2,692 contigs (13%) matched known nucleotide sequences available in current databases, and 7,876 (37%) had known C. elegans homologues. The results of the conceptual translation of nucleotide into amino acid sequences, signal peptide and transmembrane domains predictions and InterProScan, GO and KOBAS
(pathway mapping) analyses are listed in Table 3.1. The findings from homology searches of the proteins predicted for *T. colubriformis* with those available for *C. elegans*, other parasitic nematodes and organisms other than nematodes are displayed in Fig. 3.1. In total, 15,475 proteins were inferred from all 21,259 contigs, of which 6,492 and 5,393 matched known *C. elegans* and other parasitic nematode homologues, respectively (Fig. 3.1); 15,291 predicted peptides of *T. colubriformis* mapped to known proteins with 2,417 different domains (Table 3.1; Appendix 3.2). 'Transthyretin-like' (IPR001534; 8.8%), 'RNA recognition motif, RNP-1' (IPR000504; 8.4%) and 'metridin-like ShK toxin' (IPR003582; 7.6%) were the domains most commonly detected (Table 3.2). The GO annotation revealed that 5,860 predicted peptides of *T. colubriformis* could be assigned to 320 'biological process', 3,892 to 124 'cellular component' and 10,941 to 446 'molecular function' terms (Table 3.3; Appendix 3.3). The most represented GO terms were 'translation' (GO:0006412; 11.5%) and 'metabolic process' (GO:0008152; 8.4%) for 'biological process', 'intracellular' (GO:0005622; 25.7%) and 'ribosome' (GO:0005840; 14.6%) for 'cellular component' and 'ATP binding' (GO:0005524; 6.5%) and 'structural constituent of ribosome' (GO:0003735; 5.6%) for 'molecular function' (Table 3.3; Appendix 3.3). Pathway mapping using KOBAS predicted 5,150 peptides of *T. colubriformis* to be involved in 235 distinct pathways, of which most were represented by 'peptidases' (n = 211; 4%), 'ribosome' (n = 184; 3.6%), and 'oxidative phosphorylation' (n = 153; 3%) (Appendix 3.4). The 211 predicted peptides of *T. colubriformis* which were assigned to biological pathways linked to 'peptidases' (Appendix 3.4) had significant homology (at the amino acid level; e-value cut-off: ≤1e-05) to a total of 38 unique *C. elegans* peptidases (listed in Table 3.4).

Probabilistic functional gene networking predicted 2,126 orthologues in *C. elegans* to interact directly with 2,847 other genes (range: 1-271; Appendix 3.5). In particular, a subset of 42 of these orthologues were each predicted to interact directly with ≥100 other genes (Table 3.5). The majority of these orthologues had embryonic (n = 29; 69%), larval (n = 18; 43%) and/or adult (n = 8; 19%) lethal RNAi phenotypes in *C. elegans*. Genes encoding serine/threonine protein kinases had the highest representation (n = 6; 14.3%), followed by GTPases (n = 5; 11.9%), serine/threonine protein phosphatases (n = 3; 7.1%), hedgehog proteins (n = 3; 7.1%), transcription and translation factors (n = 2; 6.5%) and other proteins (n = 1; ~2.4% for each) (Table 3.5).
3.4 Discussion

The present chapter has provided the first, detailed analysis of the transcriptome of adult *T. colubriformis* and identified some groups of molecules predicted to play pivotal roles in essential biological processes in this parasite. The percentage (∼40%) of *T. colubriformis* sequences with orthologues/homologues in public databases was similar to that reported in similar transcriptomic studies of other parasitic helminths [40-42], and is likely to reflect the paucity of sequence data available for this group. In addition, of the ∼22,000 contigs assembled here, ∼25% did not have predicted ORFs. The likely explanation for this result is technical and would appear to relate to a 3’-bias in sequences derived from the normalised cDNA library for *T. colubriformis*. Future investigations should compare the data from non-normalised libraries with those from the present chapter. Among the *T. colubriformis* proteins encoded in the transcriptome, those with ‘transthyretin-like’, ‘RNA recognition motif, RNP-1’ and ‘metridin-like ShK toxin’ motifs were the most commonly identified. The ‘transthyretin-like’ proteins (TTLs; [43]) represent one of the largest protein families encoded by genes specific to nematodes [44]. The function of the TTLs differs from that of the ‘transthyretin’ and the ‘transthyretin-related’ proteins, which are known to be carriers of lipophilic substances or hormones [43]. Members of the TTL family have been detected in nematodes, including *Xiphinema index*, *Heteroder a glycines*, *Meloidogyne incognita* and *Radophilis similis* of plants [44-47], and *Brugia malayi* of humans [48], and *Ostertagia ostertagi* (related to *T. colubriformis*) of ruminants [49,50]. In nematodes, a role of TTL proteins in the nervous system has been hypothesized, supported also by the observation that *ttl* genes are expressed in the ventral nerve of *R. similis* (i.e., *Rsttl-2*), in the tail and head neurons as well as the hypodermis of *C. elegans* (gene code R13A5.6; [www.wormbase.org]), and based on the sequence similarity between TTL proteins and the *C. elegans* ‘neuropeptide-like proteins’ (NLPs) [43].

Interestingly, 12 proteins predicted in the transcriptome of *T. colubriformis* showed high sequence similarity to NLPs of *C. elegans*, including NLP-12 (gene code M01D7.5) ([www.wormbase.org](http://www.wormbase.org)). Based on the observation that *nlp-12* mRNA has been localized to a single neurone in the posterior end of *T. colubriformis* [51], it was speculated that *nlp-12* is expressed in a cell of the pre-anal ganglion, which includes both motor- and inter-neurones [51]. Transcription of *nlp-12* was also detected in L3s, and adult males and females of *T. colubriformis*, suggesting that the expression of NLP-12 might not be developmentally regulated and supporting the hypothesis for a key role of this
protein in the nervous system of nematodes [51]. Neuropeptides have been the focus of a number of studies of parasitic nematodes (e.g., [52,53]), particularly because the neuromuscular system represents the target of several anthelmintic compounds, such as piperazine, levamisole and macrocyclic lactones [54]. In the potato cyst nematode, *Globodera pallida*, the silencing by RNAi of five characterized genes encoding FMRFamide-like peptides (FLPs, a family of neuropeptides similar to the NLPs; cf. [52]) resulted in impaired locomotory behaviour of the infective juvenile stage, which led to the hypothesis that RNAi-mediated *flp* gene silencing might represent a novel approach for the control of plant parasites [55]. Some success with RNAi-based silencing of selected genes in *T. colubriformis* L3s [56] and some *nlp* genes (e.g., *nlp*-10, *nlp*-11, *nlp*-12) expressed in neural tissues of *C. elegans*, which resulted in non-wildtype phenotypes such as defects in embryonic development and aldicarb resistance (www.wormbase.org), suggests that there is scope for investigating the functions of *nlp* genes in the neuromuscular system in this trichostrongyloid.

The ‘metridin-like ShK toxin’ domain was the second most represented protein motif amongst the peptides inferred from the transcriptome of *T. colubriformis* (see Table 3.2). This domain, which is named after ‘metridin’, a toxin from the brown sea anemone, *Metridium senile*, and ‘ShK’, a structurally defined polypeptide from the sea anemone, *Stoichactis helianthus* [57], is found in one or more copies as a C-terminal domain in the metallo-peptidases of *C. elegans* (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR003582#PUB00023590; [58]). Indeed, the vast majority of peptides of *T. colubriformis* were predicted to be involved in protein catabolism (see Table 3.4). Previously, serine- and metallo-proteases have been identified in the excretory/secretory products (ES) from L4s and adults of *T. vitrinus* [59,60], and shown to be active at various pHs [59]. It was suggested that these proteases might facilitate the survival of the parasite in the host by mediating, for example, tissue penetration, feeding and/or immune-evasion by (i) digesting antibodies [61]; (ii) cleaving cell-surface receptors for cytokines [62] and/or (iii) causing the direct lysis of immune cells [63]. Proteases expressed on the epithelial surface of the gut of nematodes have been the focus of a number of studies, aimed at exploring their potential as vaccine candidates, particularly in blood-feeding nematodes [64-69]. In one of these studies [66], antibodies raised against the gut proteases of hookworms were demonstrated to bind to the nematode intestine during the blood-meal and shown to neutralize the proteolytic activity of these enzymes *in vitro*. Although attempts have been made to neutralize gut proteases of strongyloid nematodes that are not obligate
blood-feeders, the results have been not conclusive [70,71]; this outcome suggests that these nematodes do not ingest sufficient (IgG) antibodies for the vaccine to be effective [64]. It has been proposed that secreted proteases might be attractive targets for the development of vaccines against strongylid nematodes that are not obligate blood-feeders [66]. Interestingly, the predicted peptidases (n = 211; see Table 3.4 and Appendix 3.4) identified in the transcriptome of *T. colubriformis* did not have a signal peptide indicative of excretion/secretion, with the exception of the *C. elegans* homologues TRY-1, SEL-12 and IMP-2 (see Table 3.4). A possible explanation for the absence of such a signal is that these proteases could be excreted/secreted using a “non-classical” pathway that does not involve signal peptide cleavage [72]. Alternatively, these proteases could be released bound to secreted proteinase inhibitors [60]. Indeed, the ‘proteinase inhibitor I2, Kunitz metazoa’ was amongst the ten most represented protein motifs identified for *T. colubriformis* (Table 3.2). Some protease inhibitors have been studied in *T. colubriformis*. For instance, an aspartyl protease inhibitor (*Tco-api-I*) has been identified as a major allergen by immunochemical analysis of somatic antigens of *T. colubriformis* L3 using IgE purified from the serum of sheep grazed on worm-contaminated pastures [73], and is proposed to play a key role in the establishment of parasites in the vertebrate host by inhibiting pepsin activity during the transit of L3s through the gastric environment. In adult worms, proteinase inhibitors may also be implicated in the evasion of the host immune response through the cleavage of leucocyte elastases, mast cell proteases and cathepsins released from stimulated polymorphonuclear neutrophils [cf. 65]. Given that a significant mucosal (IgA and IgE) immune response can be induced in sheep vaccinated with *T. colubriformis* native or recombinant antigens delivered across the epithelium of the jejunal or rectal lymphoid tissue [74], further investigations of the potential of secreted protease inhibitors as vaccine targets are warranted.

Other secreted proteins, such as those containing a ‘sperm-coating protein (SCP)-like extracellular domain’ (InterPro: IPR014044), also called SCP/Tpx-1/Ag5/PR-1/Sc7 (SCP/TAPS; Pfam accession number no. PF00188), were amongst the top 20 most represented molecules in the transcriptome of *T. colubriformis* (see Table 3.2 and Appendix 3.2). This is the first record of SCP/TAPS proteins in this parasitic nematode. SCP/TAPS belong to a large group of proteins that include the *Ancylostoma*-secreted proteins (ASPs; [75]). In parasitic nematodes, ASPs were first characterized for hookworms [76,77] and subsequently from related strongylid nematodes [78]. As ASPs are abundant in
the excretory/secretory (ES) products of the infective L3, they are thought to play an important role in the transition from the free-living to the parasitic stage of a nematode during its invasion of the host [76,77,79,80]. In adult hookworms, ASPs have been proposed to play an immunomodulatory function during the invasion of the host, the migration through tissues, attachment to the intestinal wall and blood-feeding [81]. In trichostrongyloid nematodes that do not feed on blood, homologues of the hookworm ASPs have been identified in the Ostertagia ostertagi (brown stomach worm). In this species, two molecules (designated Oo-ASP-1 and Oo-ASP-2), were characterized as major antigens in a protective ES-thiol fraction of partially purified ES products. These molecules were shown to be highly expressed in L4s and adult males of Os. ostertagi, a finding that was consistent with their transcriptional profile [82]. Similarly, in T. vitrinus and Oesophagostomum dentatum (the nodule worm of pigs), ESTs with high sequence similarity to different types of asps were demonstrated to be male-enriched [83,84]. The function of the ASPs in trichostrongyloid parasites is currently unknown [75]. A clear understanding of the expression patterns of ASPs in larval stages and both genders of these nematodes would provide clues as to the roles of these molecules in the life cycle and interactions with the host.

A probabilistic functional gene network of protein-encoding C. elegans orthologues inferred from the transcriptome of T. colubriformis was constructed to link genes that are known to function together in C. elegans. Amongst the orthologues predicted to each interact with >100 other genes, serine/threonine protein kinase and phosphatase genes were the most abundantly represented group overall (n = 9; 21.4%) (Table 3.5). In nematodes these molecules are known or inferred to be involved in sperm production by the adult males, as suggested from previous studies of C. elegans [85], T. vitrinus [83], Haemonchus contortus (the ‘barber pole’ worm of small ruminants) [12,86] and Oe. dentatum [84,87]. In the latter nematode, a catalytic subunit of a serine/threonine protein phosphatase (PP1) was characterized (Od-mpp1), and gene silencing by RNA interference (RNAi) of the corresponding C. elegans homologue resulted in a significant reduction (30-40%) in the numbers of F2 progeny produced [87]. In an independent study [88], gene silencing of these genes by double-stranded RNAi in C. elegans hermaphrodites resulted in impaired sperm function. Indeed, protein kinases and phosphatases are known to be highly expressed in the sperm-producing germline tissue of C. elegans [85], thus suggesting key roles in regulating sperm maturation, following the expulsion of the organelles involved in protein synthesis from maturing
spermatids [85]. In *T. vitrinus*, a serine/threonine protein phosphatase
(designated *Tv-stp-1*), was shown to be transcribed specifically in adult males,
but not in adult females or any of the larval stages [89], whereas the transcript
_corresponding to the *H. contortus* homologue of *Tv-stp-1* (i.e., *Hc-stp-1*) was
detected in L4s and adult males of this species [89]. In addition, sequence
conservation among *Od-mpp1*, *Tv-stp-1*, *Hc-stp-1* and genes encoding
serine/threonine protein phosphatases in *C. elegans* suggest a similar biological
function [86,89]. Similarly, the *T. colubriformis* gene orthologues of
serine/threonine phosphatases identified in the present study are likely to play
key roles in the reproductive processes of this species, but still need detailed
investigation.

The transcriptomic dataset described here constitutes a basis for future
investigations of essential pathways of development and reproduction in *T.
colubriformis*, a statement supported by functional gene networking inferring
that 95% of *C. elegans* orthologues that interacted with > 100 other genes were
linked to lethality, growth defects or sterility based on gene silencing (see Table
3.5). Clearly, next-generation sequencing technologies, particularly 454-Roche
(www.454.com), Illumina-Solexa (www.illumina.com) and Helicos
(www.helicosbio.com), might assist future genomic and transcriptomic studies,
aimed, for instance, at exploring the sequence variability of mRNA transcripts
encoding surface antigens in different populations of infective L3s of *T.
colubriformis* [8], thus assisting in the elucidation of an aspect of immune evasion
in this nematode. Complemented by proteomic exploration, the characterisation
of the transcriptomes of all developmental stages and both sexes of *T.
colubriformis* from non-normalised cDNA libraries will allow a global study of
differential gene expression. In addition, the development of methods for RNAi in
the L3 of *T. colubriformis* [56] offers opportunities for investigating the
function(s) of molecules (e.g., serine/threonine phosphatases) predicted to play
crucial roles in parasitic nematodes. Future investigations should also focus on
inferring the functions of orthologous molecules from trichostrongyloid
nematodes using genetic complementation in *C. elegans* [90]. Improved
knowledge of fundamental molecular pathways in nematodes should provide a
sound basis for the discovery and prevalidation of targets for drug design.
3.5 References


40. Chapter 4. The application of next-generation sequencing technology to explore differential transcription between free-living and pre-parasitic third-stage larvae of Haemonchus contortus.


Table 3.1 Number of expressed sequence tags (ESTs) determined from the cDNA library of the adult stage of *Trichostrongylus colubriformis* prior to and after assembly, and the results of bioinformatic analyses.

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of EST clusters</td>
<td>21,259</td>
</tr>
<tr>
<td>Average length (± range)</td>
<td>495 bp ± 224.9</td>
</tr>
<tr>
<td>Containing an Open Reading Frame (no. of full-length sequences)</td>
<td>15,475 (12,472)</td>
</tr>
<tr>
<td>Signal peptides</td>
<td>274</td>
</tr>
<tr>
<td>Containing transmembrane domains</td>
<td>588</td>
</tr>
<tr>
<td>Returning InterProScan results</td>
<td>15,475 (2,417 domains)</td>
</tr>
<tr>
<td><strong>Gene Ontology</strong></td>
<td></td>
</tr>
<tr>
<td>Biological process</td>
<td>5,860 (320 terms)</td>
</tr>
<tr>
<td>Cellular component</td>
<td>3,892 (124 terms)</td>
</tr>
<tr>
<td>Molecular function</td>
<td>10,941 (446 terms)</td>
</tr>
<tr>
<td>Prediction of biological pathways (KOBAS)</td>
<td>235</td>
</tr>
<tr>
<td><strong>Homology searches</strong></td>
<td></td>
</tr>
<tr>
<td>With known orthologues</td>
<td>2,692 (13%)</td>
</tr>
<tr>
<td>With known homologues</td>
<td>7,876 (37%)</td>
</tr>
</tbody>
</table>
Table 3.2 The thirty most represented protein domains inferred from peptides, conceptually translated from individual contigs of expressed sequence tags (ESTs) for the adult stage of *Trichostrongylus colubriformis.*

<table>
<thead>
<tr>
<th>InterPro code</th>
<th>InterPro description</th>
<th>No. of ESTs (%)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPR001534</td>
<td>Transthryretin-like</td>
<td>212 (8.8)</td>
</tr>
<tr>
<td>IPR000504</td>
<td>RNA recognition motif, RNP-1</td>
<td>202 (8.3)</td>
</tr>
<tr>
<td>IPR003582</td>
<td>Metridin-like ShK toxin</td>
<td>187 (7.6)</td>
</tr>
<tr>
<td>IPR001680</td>
<td>WD40 repeat</td>
<td>166 (6.9)</td>
</tr>
<tr>
<td>IPR016040</td>
<td>NAD(P)-binding domain</td>
<td>141 (5.8)</td>
</tr>
<tr>
<td>IPR001304</td>
<td>C-type lectin</td>
<td>136 (5.6)</td>
</tr>
<tr>
<td>IPR002068</td>
<td>Heat shock protein Hsp20</td>
<td>132 (5.5)</td>
</tr>
<tr>
<td>IPR002223</td>
<td>Proteinase inhibitor I2, Kunitz metazoa</td>
<td>131 (5.4)</td>
</tr>
<tr>
<td>IPR002110</td>
<td>Ankyrin</td>
<td>123 (5.1)</td>
</tr>
<tr>
<td>IPR019781</td>
<td>WD40 repeat, subgroup</td>
<td>115 (4.7)</td>
</tr>
<tr>
<td>IPR007087</td>
<td>Zinc finger, C2H2-type</td>
<td>110 (4.6)</td>
</tr>
<tr>
<td>IPR019782</td>
<td>WD40 repeat 2</td>
<td>104 (4.3)</td>
</tr>
<tr>
<td>IPR018249</td>
<td>EF-HAND 2</td>
<td>95 (4)</td>
</tr>
<tr>
<td>IPR013032</td>
<td>EGF-like region, conserved site</td>
<td>92 (3.8)</td>
</tr>
<tr>
<td>IPR012677</td>
<td>Nucleotide-binding, alpha-beta plait</td>
<td>84 (3.5)</td>
</tr>
<tr>
<td>IPR011009</td>
<td>Protein kinase-like</td>
<td>84 (3.5)</td>
</tr>
<tr>
<td>IPR014044</td>
<td>SCP-like extracellular</td>
<td>79 (3.3)</td>
</tr>
<tr>
<td>IPR002130</td>
<td>Peptidyl-prolyl cis-trans isomerase, cyclophilin-type</td>
<td>77 (3.2)</td>
</tr>
<tr>
<td>IPR002423</td>
<td>Chaperonin Cpn60/TCP-1</td>
<td>75 (3.1)</td>
</tr>
<tr>
<td>IPR001395</td>
<td>Aldo/keto reductase</td>
<td>69 (2.9)</td>
</tr>
<tr>
<td>IPR011046</td>
<td>WD40 repeat-like</td>
<td>68 (2.8)</td>
</tr>
<tr>
<td>IPR008978</td>
<td>HSP20-like chaperone</td>
<td>66 (2.7)</td>
</tr>
<tr>
<td>IPR001650</td>
<td>DNA/RNA helicase, C-terminal</td>
<td>65 (2.7)</td>
</tr>
<tr>
<td>IPR016187</td>
<td>C-type lectin fold</td>
<td>63 (2.6)</td>
</tr>
<tr>
<td>IPR001128</td>
<td>Cytochrome P450</td>
<td>60 (2.5)</td>
</tr>
<tr>
<td>IPR009072</td>
<td>Histone-fold</td>
<td>60 (2.5)</td>
</tr>
<tr>
<td>IPR0008355</td>
<td>Major sperm protein</td>
<td>60 (2.5)</td>
</tr>
<tr>
<td>IPR000242</td>
<td>Protein-tyrosine phosphatase, receptor/non-receptor type</td>
<td>59 (2.4)</td>
</tr>
<tr>
<td>IPR017986</td>
<td>WD40 repeat, region</td>
<td>59 (2.4)</td>
</tr>
<tr>
<td>IPR000557</td>
<td>Calponin repeat</td>
<td>58 (2.4)</td>
</tr>
</tbody>
</table>

* Percentage is calculated on the total number of InterPro domains that could be mapped in the present dataset.
Table 3.3 Functions predicted for proteins encoded in the transcriptome of the adult stage of *Trichosrongylus colubriformis* based on GO. The parental (= level 2) GO categories were assigned according to InterPro domains with homology to functionally annotated molecules.

<table>
<thead>
<tr>
<th>GO category</th>
<th>GO code</th>
<th>GO description</th>
<th>No. of ESTs (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Biological process</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0008152</td>
<td>Metabolic process</td>
<td>2,034 (13.3)</td>
<td></td>
</tr>
<tr>
<td>GO:0009987</td>
<td>Cellular process</td>
<td>1,846 (12.1)</td>
<td></td>
</tr>
<tr>
<td>GO:0051179</td>
<td>Localization</td>
<td>431 (2.9)</td>
<td></td>
</tr>
<tr>
<td>GO:0065007</td>
<td>Biological regulation</td>
<td>282 (1.8)</td>
<td></td>
</tr>
<tr>
<td>GO:0050789</td>
<td>Regulation of biological process</td>
<td>270 (1.8)</td>
<td></td>
</tr>
<tr>
<td>GO:0016043</td>
<td>Cellular component organization</td>
<td>150 (1)</td>
<td></td>
</tr>
<tr>
<td>GO:0044085</td>
<td>Cellular component biogenesis</td>
<td>109 (0.7)</td>
<td></td>
</tr>
<tr>
<td>GO:0010926</td>
<td>Anatomical structure formation</td>
<td>87 (0.6)</td>
<td></td>
</tr>
<tr>
<td>GO:0050896</td>
<td>Response to stimulus</td>
<td>45 (0.3)</td>
<td></td>
</tr>
<tr>
<td>GO:0022610</td>
<td>Biological adhesion</td>
<td>13 (0.1)</td>
<td></td>
</tr>
<tr>
<td>GO:0032501</td>
<td>Multicellular organismal process</td>
<td>7 (0.05)</td>
<td></td>
</tr>
<tr>
<td>GO:0032502</td>
<td>Developmental process</td>
<td>5 (0.03)</td>
<td></td>
</tr>
<tr>
<td>GO:0002376</td>
<td>Immune system process</td>
<td>1 (0.01)</td>
<td></td>
</tr>
<tr>
<td>GO:0016265</td>
<td>Death</td>
<td>1 (0.01)</td>
<td></td>
</tr>
<tr>
<td><strong>Cellular component</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0005623</td>
<td>Cell</td>
<td>1,713 (11.2)</td>
<td></td>
</tr>
<tr>
<td>GO:0044464</td>
<td>Cell part</td>
<td>1,713 (11.2)</td>
<td></td>
</tr>
<tr>
<td>GO:0043226</td>
<td>Organelle</td>
<td>761 (5)</td>
<td></td>
</tr>
<tr>
<td>GO:0032991</td>
<td>Macromolecular complex</td>
<td>589 (3.9)</td>
<td></td>
</tr>
<tr>
<td>GO:0005576</td>
<td>Extracellular region</td>
<td>58 (0.4)</td>
<td></td>
</tr>
<tr>
<td>GO:0031975</td>
<td>Envelope</td>
<td>48 (0.3)</td>
<td></td>
</tr>
<tr>
<td>GO:0031974</td>
<td>Membrane-enclosed lumen</td>
<td>40 (0.3)</td>
<td></td>
</tr>
<tr>
<td>GO:0044456</td>
<td>Synapse part</td>
<td>10 (0.01)</td>
<td></td>
</tr>
<tr>
<td>GO:0045202</td>
<td>Synapse</td>
<td>10 (0.01)</td>
<td></td>
</tr>
<tr>
<td>GO:0044421</td>
<td>Extracellular region part</td>
<td>9 (0.01)</td>
<td></td>
</tr>
<tr>
<td><strong>Molecular function</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0005488</td>
<td>Binding</td>
<td>2,358 (15.4)</td>
<td></td>
</tr>
<tr>
<td>GO:0003824</td>
<td>Catalytic activity</td>
<td>2,032 (13.3)</td>
<td></td>
</tr>
<tr>
<td>GO:0005198</td>
<td>Structural molecule activity</td>
<td>284 (1.9)</td>
<td></td>
</tr>
<tr>
<td>GO:0005215</td>
<td>Transporter activity</td>
<td>202 (1.3)</td>
<td></td>
</tr>
<tr>
<td>GO:0009055</td>
<td>Electron carrier activity</td>
<td>75 (0.5)</td>
<td></td>
</tr>
<tr>
<td>GO:0030234</td>
<td>Enzyme regulator activity</td>
<td>67 (0.4)</td>
<td></td>
</tr>
<tr>
<td>GO:0030528</td>
<td>Transcription regulator activity</td>
<td>67 (0.4)</td>
<td></td>
</tr>
<tr>
<td>GO:0045182</td>
<td>Translator regulator activity</td>
<td>35 (0.2)</td>
<td></td>
</tr>
<tr>
<td>GO:0060089</td>
<td>Molecular transducer activity</td>
<td>33 (0.2)</td>
<td></td>
</tr>
<tr>
<td>GO:0016209</td>
<td>Antioxidant activity</td>
<td>15 (0.1)</td>
<td></td>
</tr>
<tr>
<td>GO:0010860</td>
<td>Proteasome regulator activity</td>
<td>2 (0.01)</td>
<td></td>
</tr>
<tr>
<td>GO:0016530</td>
<td>Metallochaperone activity</td>
<td>1 (0.01)</td>
<td></td>
</tr>
</tbody>
</table>
Table 3.4 *Caenorhabditis elegans* peptidase homologues of *Trichostrongylus colubriformis* predicted peptides.

<table>
<thead>
<tr>
<th><em>C. elegans</em> homologue (gene code)</th>
<th>Description</th>
<th>No. of <em>T. colubriformis</em> homologues</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAS-31 (F58B4.1)</td>
<td>Meprin A metalloprotease</td>
<td>43</td>
</tr>
<tr>
<td>NEP-1 (ZK20.6)</td>
<td>Thermolysin-like zinc metallopeptidases</td>
<td>28</td>
</tr>
<tr>
<td>PAS-5 (F25H2.9)</td>
<td>Subunit of the core 20S proteasome subcomplex</td>
<td>20</td>
</tr>
<tr>
<td>CPR-6 (C25B8.3)</td>
<td>Cysteine proteinase Cathepsin L</td>
<td>18</td>
</tr>
<tr>
<td>CLP-1 (C06G4.2)</td>
<td>Ca2+-dependent cysteine protease (calpain)</td>
<td>17</td>
</tr>
<tr>
<td>PAS-3 (Y110A7A.14)</td>
<td>Subunit of the core 20S proteasome subcomplex</td>
<td>15</td>
</tr>
<tr>
<td>CPZ-1 (F32B5.8)</td>
<td>Cysteine proteinase Cathepsin L</td>
<td>15</td>
</tr>
<tr>
<td>NAS-5 (T23H4.3)</td>
<td>Meprin A metalloprotease</td>
<td>6</td>
</tr>
<tr>
<td>TRY-1 (ZK546.15)</td>
<td>Trypsin</td>
<td>5*</td>
</tr>
<tr>
<td>F44E7.4</td>
<td>Zn2+-dependent endopeptidases</td>
<td>5</td>
</tr>
<tr>
<td>PBS-3 (Y38A8.2)</td>
<td>Subunit of the core 20S proteasome subcomplex</td>
<td>4</td>
</tr>
<tr>
<td>PBS-2 (C47B2.4)</td>
<td>Subunit of the core 20S proteasome subcomplex</td>
<td>3</td>
</tr>
<tr>
<td>PBS-7 (F39H11.5)</td>
<td>Subunit of the core 20S proteasome subcomplex</td>
<td>3</td>
</tr>
<tr>
<td>UBH-1 (F46E10.8)</td>
<td>Ubiquitin C-terminal hydrolase</td>
<td>3</td>
</tr>
<tr>
<td>IMP-1 (C36B1.12)</td>
<td>Uncharacterized conserved protein, contains PA domain</td>
<td>2</td>
</tr>
<tr>
<td>PES-9 (R11H6.1)</td>
<td>Metalloexopeptidases</td>
<td>1</td>
</tr>
<tr>
<td>PAM-1 (F49E8.3)</td>
<td>Puromycin-sensitive aminopeptidase</td>
<td>1</td>
</tr>
<tr>
<td>PAS-7 (ZK945.2)</td>
<td>Subunit of the core 20S proteasome subcomplex</td>
<td>1</td>
</tr>
<tr>
<td>PBS-4 (T20F5.2)</td>
<td>Subunit of the core 20S proteasome subcomplex</td>
<td>1</td>
</tr>
<tr>
<td>APP-1 (W03G9.4)</td>
<td>Xaa-Pro aminopeptidase</td>
<td>1</td>
</tr>
<tr>
<td>SPG-7 (Y47G6A.10)</td>
<td>Metalloprotease</td>
<td>1</td>
</tr>
<tr>
<td>CLP-2 (T04A8.16)</td>
<td>Ca2+-dependent cysteine protease</td>
<td>1</td>
</tr>
<tr>
<td>BLI-4 (K04F10.4)</td>
<td>Subtilisin-like proprotein convertase</td>
<td>1</td>
</tr>
<tr>
<td>CPR-1 (C52E4.1)</td>
<td>Cysteine proteinase Cathepsin L</td>
<td>1</td>
</tr>
<tr>
<td>TRY-4 (F31D4.6)</td>
<td>Trypsin</td>
<td>1</td>
</tr>
<tr>
<td>ULP-3 (Y48A5A.2)</td>
<td>Sentrin-specific cysteine protease</td>
<td>1</td>
</tr>
<tr>
<td>ADM-4 (ZK154.7)</td>
<td>Tumor necrosis factor-alpha-converting enzyme</td>
<td>1</td>
</tr>
<tr>
<td>MATH-33 (H19N07.2)</td>
<td>Ubiquitin carboxyl-terminal hydrolase</td>
<td>1</td>
</tr>
<tr>
<td>DPF-1 (T23F1.7)</td>
<td>Dipeptidyl aminopeptidase</td>
<td>1</td>
</tr>
<tr>
<td>DPF-3 (K02F2.1)</td>
<td>Dipeptidyl aminopeptidase</td>
<td>1</td>
</tr>
<tr>
<td>MAP-2 (Y116A8A.9)</td>
<td>Metallopeptidase</td>
<td>1</td>
</tr>
<tr>
<td>CLPP-1 (ZK970.2)</td>
<td>ATP-dependent Clp protease</td>
<td>1</td>
</tr>
<tr>
<td>CLP-1 (C06G4.2)</td>
<td>Ca2+-dependent cysteine protease</td>
<td>1</td>
</tr>
<tr>
<td>DPF-5 (R11E3.8)</td>
<td>Dipeptidyl aminopeptidase</td>
<td>1</td>
</tr>
<tr>
<td>CPL-1 (T03E6.7)</td>
<td>Cysteine proteinase Cathepsin L</td>
<td>1</td>
</tr>
<tr>
<td>NAS-9 (C37H5.9)</td>
<td>Meprin A metalloprotease</td>
<td>1</td>
</tr>
<tr>
<td>SEL-12 (F35H12.3)</td>
<td>Presenilin</td>
<td>1*</td>
</tr>
<tr>
<td>IMP-2 (T05E11.5)</td>
<td>Uncharacterized conserved protein, contains PA domain</td>
<td>1*</td>
</tr>
</tbody>
</table>

*Protein/s with predicted signal peptides
Table 3.5 Description and gene ontology (GO) classifications (according to the categories ‘biological process’, ‘cellular component’ and ‘molecular function’) of the *Caenorhabditis elegans* orthologues of *Trichosstrongylus colubriformis* molecules predicted to interact with ≥100 other genes.

<table>
<thead>
<tr>
<th>Gene name (code)</th>
<th>Description</th>
<th>No. predicted interacting genes</th>
<th>Biological process</th>
<th>Gene Ontology</th>
<th>Molecular function</th>
<th>RNAi phenotypes*</th>
</tr>
</thead>
<tbody>
<tr>
<td>glp-1 (F02A9.6)</td>
<td>Member of the LIN-12/Notch family of receptors</td>
<td>271</td>
<td>Cell differentiation</td>
<td>Integral to membrane</td>
<td>Calcium-ion binding</td>
<td>Emb, Ste, Lvl</td>
</tr>
<tr>
<td>lin-12 (R107.8)</td>
<td>Member of the LIN-12/Notch family of receptors</td>
<td>260</td>
<td></td>
<td></td>
<td></td>
<td>Slo, Ste, Unc</td>
</tr>
<tr>
<td>cdc-42 (R07G3.1)</td>
<td>RHO GTPase</td>
<td>255</td>
<td>Small GTPase mediated signal transduction</td>
<td>Intracellular</td>
<td>GTP-binding</td>
<td>Emb, Lvl, Ste</td>
</tr>
<tr>
<td>let-60 (ZK792.6)</td>
<td>Member of the GTP-binding RAS protooncogene family</td>
<td>236</td>
<td>Small GTPase mediated signal transduction</td>
<td>Intracellular</td>
<td>GTP-binding</td>
<td>Let, Emb, Lvl, Lva, Ste</td>
</tr>
<tr>
<td>let-23 (ZK1067.1)</td>
<td>Transmembrane tyrosine kinase</td>
<td>203</td>
<td>Protein amino acid phosphorylation</td>
<td>Membrane</td>
<td>ATP-binding</td>
<td>Unclassified</td>
</tr>
<tr>
<td>ced-10 (C09G12.8)</td>
<td>GTPase</td>
<td>194</td>
<td>Small GTPase mediated signal transduction</td>
<td>Intracellular</td>
<td>GTP-binding</td>
<td>Let, Emb, Lvl, Lva</td>
</tr>
<tr>
<td>crb-1 (F11C7.4)</td>
<td>Homolog of Drosophila CRUMBS</td>
<td>190</td>
<td></td>
<td></td>
<td>Calcium-ion binding</td>
<td>Slo</td>
</tr>
<tr>
<td>ima-3 (F32E10.4)</td>
<td>Importin alpha nuclear transport factor</td>
<td>189</td>
<td>Intracellular protein transport</td>
<td>Nucleus</td>
<td>Protein transporting activity</td>
<td>Emb, Lva, Ste</td>
</tr>
<tr>
<td>gei-4 (W07B3.2)</td>
<td>Glutamine/asparagine-rich protein</td>
<td>159</td>
<td></td>
<td></td>
<td></td>
<td>Let, Emb, Lvl, Lva</td>
</tr>
<tr>
<td>dbl-1 (T25F10.2)</td>
<td>Member of the transforming growth factor beta (TGFbeta) superfamily</td>
<td>148</td>
<td></td>
<td></td>
<td>Growth factor activity</td>
<td>Slo</td>
</tr>
<tr>
<td>hmp-2 (K05C4.6)</td>
<td>Beta-catenin</td>
<td>143</td>
<td></td>
<td></td>
<td>Binding</td>
<td>Emb, Slo</td>
</tr>
<tr>
<td>act-4 (M03F4.2)</td>
<td>Actin</td>
<td>134</td>
<td></td>
<td></td>
<td></td>
<td>Emb, Lvl, Lva, Ste</td>
</tr>
<tr>
<td>gsp-2 (F56C9.1)</td>
<td>Serine/threonine protein phosphatase</td>
<td>134</td>
<td></td>
<td></td>
<td>Hydrolase activity</td>
<td>Let, Emb, Sck</td>
</tr>
<tr>
<td>mig-2 (C35C5.4)</td>
<td>Member of the Rho family of GTP-binding</td>
<td>132</td>
<td>Small GTPase mediated signal transduction</td>
<td>Intracellular</td>
<td>GTP-binding</td>
<td>Unclassified, embryonic defects</td>
</tr>
<tr>
<td>Gene</td>
<td>Gene ID</td>
<td>Protein Name/Function</td>
<td>Molecular Function</td>
<td>Cellular Localization</td>
<td>Molecular Activity</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>----------</td>
<td>---------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------</td>
<td>-----------------------</td>
<td>-----------------------------</td>
<td></td>
</tr>
<tr>
<td>ras-1</td>
<td>C44C11.1</td>
<td>Ras-related GTPase</td>
<td>Small GTPase mediated signal transduction</td>
<td>Intracellular</td>
<td>GTP-binding</td>
<td></td>
</tr>
<tr>
<td>wrt-6</td>
<td>ZK377.1</td>
<td>Hedgehog-like protein</td>
<td>Cell communication</td>
<td></td>
<td>Peptidase activity</td>
<td></td>
</tr>
<tr>
<td>kin-19</td>
<td>C03C10.1</td>
<td>Casein kinase (serine/threonine/tyrosine protein kinase)</td>
<td>Protein amino acid phosphorylation</td>
<td></td>
<td>Unc, small ATP binding</td>
<td></td>
</tr>
<tr>
<td>cdk-1</td>
<td>T05G5.3</td>
<td>Protein kinase</td>
<td>Protein amino acid phosphorylation</td>
<td></td>
<td>ATP binding</td>
<td></td>
</tr>
<tr>
<td>fib-1</td>
<td>T01C3.7</td>
<td>Fibrillarin</td>
<td>RNA processing</td>
<td></td>
<td>Nucleus</td>
<td></td>
</tr>
<tr>
<td>cye-1</td>
<td>C37A2.4</td>
<td>G1/S-specific cyclin E</td>
<td>Regulation of transcription</td>
<td></td>
<td>ATP binding</td>
<td></td>
</tr>
<tr>
<td>psa-4</td>
<td>F01G4.1</td>
<td>DNA/RNA helicase</td>
<td>Protein transport</td>
<td></td>
<td>GTP binding</td>
<td></td>
</tr>
<tr>
<td>rho-1</td>
<td>Y51H4A.3</td>
<td>Ras-related small GTPase</td>
<td>Protein transport</td>
<td></td>
<td>Peptidase activity</td>
<td></td>
</tr>
<tr>
<td>grd-11</td>
<td>K02E2.2</td>
<td>Hedgehog-like protein</td>
<td>Proteolysis</td>
<td></td>
<td>Small, Unc Hydrolase activity</td>
<td></td>
</tr>
<tr>
<td>grd-2</td>
<td>F46B3.5</td>
<td>Hedgehog-like protein</td>
<td>Proteolysis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>let-92</td>
<td>F38H4.9</td>
<td>Serine/threonine protein phosphatase</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kin-3</td>
<td>B0205.7</td>
<td>Casein kinase II</td>
<td>Protein amino acid phosphorylation</td>
<td></td>
<td>ATP binding</td>
<td></td>
</tr>
<tr>
<td>goa-1</td>
<td>C26G6.2</td>
<td>G-protein alpha subunit</td>
<td>G-protein coupled receptor protein signaling pathway</td>
<td></td>
<td>ATP binding</td>
<td></td>
</tr>
<tr>
<td>act-2</td>
<td>T04C12.5</td>
<td>Actin</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>psa-1</td>
<td>F33H2.5</td>
<td>Chromatin remodeling factor</td>
<td>Nucleotide and nucleic acid metabolic process</td>
<td></td>
<td>Protein binding</td>
<td></td>
</tr>
<tr>
<td>hmg-1.2</td>
<td>F47D12.4</td>
<td>HMG box-containing protein</td>
<td></td>
<td></td>
<td>DNA binding</td>
<td></td>
</tr>
<tr>
<td>par-5</td>
<td>M117.2</td>
<td>Multifunctional chaperone</td>
<td></td>
<td></td>
<td>Protein domain specific binding</td>
<td></td>
</tr>
<tr>
<td>rac-2</td>
<td>K03D3.10</td>
<td>Ras-related small GTPase</td>
<td>Small GTPase mediated signal transduction</td>
<td></td>
<td>GTP-binding Unclassified</td>
<td></td>
</tr>
<tr>
<td>nmy-2</td>
<td>F20G4.3</td>
<td>Myosin class II heavy chain</td>
<td>Myosin complex</td>
<td></td>
<td>ATP binding</td>
<td></td>
</tr>
<tr>
<td>pal-1</td>
<td>C38D4.6</td>
<td>Transcription factor</td>
<td></td>
<td></td>
<td>Transcription factor activity</td>
<td></td>
</tr>
<tr>
<td>ife-3</td>
<td>B0348.6</td>
<td>Translation initiation factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene</td>
<td>Description</td>
<td>Position</td>
<td>Function</td>
<td>Specificity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>------------------------------------</td>
<td>----------</td>
<td>-----------------------------------------------</td>
<td>------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ftt-2 (F52D10.3)</td>
<td>Multifunctional chaperone</td>
<td>103</td>
<td>Protein domain specific binding</td>
<td>Let, Emb, Lvl, Lva, Ste, Slo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>air-2 (B0207.4)</td>
<td>Serine/threonine protein kinase</td>
<td>102</td>
<td>Protein amino acid phosphorylation</td>
<td>ATP binding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>his-47 (B0035.7)</td>
<td>Histone 2A</td>
<td>101</td>
<td>Nucleosome assembly</td>
<td>DNA binding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>plc-1 (C14B9.4)</td>
<td>Serine/threonine protein kinase</td>
<td>101</td>
<td>Protein amino acid phosphorylation</td>
<td>ATP binding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>his-46 (B0035.9)</td>
<td>Histone H4</td>
<td>100</td>
<td>Nucleosome assembly</td>
<td>DNA binding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>air-1 (K07C11.2)</td>
<td>Serine/threonine protein kinase</td>
<td>100</td>
<td>Protein amino acid phosphorylation</td>
<td>ATP binding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gsp-1 (F29F11.6)</td>
<td>Serine/threonine protein phosphatase</td>
<td>100</td>
<td>Hydrolase activity</td>
<td>Emb, Lvl, Ste, Sck</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Abbreviations of RNAi phenotypes (alphabetical): adult lethal (Let), embryonic lethal (Emb), larval arrest (Lva), larval lethal (Lvl), sick (Sck), slow growth (Slo), sterile (Ste), uncoordinated (Unc)
Fig. 3.1 Relationships of individual contigs of expressed sequence tags (EST) from the adult stage of *Trichostrongylus colubriformis* with protein homologues from *Caenorhabditis elegans*, other parasitic nematodes and organisms other than nematodes, displayed in a SimiTri plot [38].
CHAPTER 4

The application of next-generation sequencing to explore differential transcription between free-living and pre-parasitic third-stage larvae of *Haemonchus contortus*¹

Abstract

The disease caused by *Haemonchus contortus* (= haemonchosis), a blood-feeding nematode of small ruminants, is of major economic importance worldwide. The infective third-stage larva (L3) of this gastric nematode is enclosed in a cuticle (sheath) and, once ingested with herbage by the host, undergoes an exsheathment process that marks the transition from the free-living (L3) to the parasitic (xL3) stage. This chapter explored changes in gene transcription associated with this transition and predicted, based on comparative analysis, functional roles for key transcripts in the metabolic pathways linked to larval development. Totals of 101,305 (L3) and 105,553 (xL3) expressed sequence tags (ESTs) were determined using 454 sequencing technology, and then assembled and annotated; the most abundant transcripts encoded transthryretin-like, calcium-binding EF-hand, NAD(P)-binding and nucleotide-binding proteins as well as homologues of *Ancylostoma*-secreted proteins (ASP). Using an *in silico*-subtractive analysis, 560 and 685 sequences were shown to be uniquely represented in the L3 and xL3 stages, respectively; the transcripts encoded ribosomal proteins, collagens and elongation factors (in L3), and mainly peptidases and other enzymes of amino acid catabolism (in xL3). *Caenorhabditis elegans* orthologues of transcripts that were uniquely transcribed in each L3 and xL3 were predicted to interact with a total of 535 other genes, all of which were involved in embryonic development. The present chapter indicated that some key transcriptional alterations taking place during the transition from the L3 to the xL3 stage of *H. contortus* involve genes predicted to be linked to the development of neuronal tissue (L3 and xL3), formation of the cuticle (L3) and digestion of host haemoglobin (xL3). Future efforts using next-generation sequencing and bioinformatic technologies should provide the efficiency and depth of coverage required for the determination of the complete transcriptomes of different developmental stages and/or tissues of *H. contortus* as well as the genome of this important parasitic nematode. Such advances should lead to a significantly improved understanding of the molecular biology of *H. contortus* and, from an applied perspective, to novel methods of intervention.

¹ This chapter is dedicated to the late Paul JA Presidente.
4.1 Introduction

Parasitic nematodes of livestock animals are of major socio-economic importance worldwide due to the diseases and production losses that they cause (Chapter 1). *Haemonchus contortus* (order Strongylida) is one of the most important nematodes of the abomasum (stomach) of small ruminants. The disease (= haemonchosis) caused by this parasite represents ~15% of all gastrointestinal diseases of small ruminants worldwide and causes major financial losses (www.fao.org). *Haemonchus contortus* is a blood-feeding worm that causes anaemia and associated complications, leading to death in severely affected animals [1]. In the abomasum of the ruminant host, the oviparous adult females release eggs *via* the faeces into the environment, in which first-stage larvae (L1s) develop and then hatch (within ~1 day, depending on conditions). The L1s develop into second-stage larvae (L2s), which then moult to become infective third-stage larvae (L3s). The cuticle from the L2 is retained as a sheath around the L3 and protects this stage from environmental pressures. Infective L3s are ingested with herbage by the host, pass through the forestomachs and undergo an exsheathment process to then establish, *via* the parasitic fourth-stage larvae (L4s), to dioecious haematophagous adults within ~3 weeks [2].

The exsheathment process marks the transition from the free-living (L3) to the parasitic (xL3) larval stage, at which *H. contortus* grows, develops and commences feeding on the blood of the host [1]. Early studies examined aspects of this process in nematodes in response to a host stimulus. Evidence indicated that at least two separate pathways control these responses [3-5]. These pathways are stimulated by gaseous CO$_2$ and mediated by carbonic anhydrase, leading to the secretion of the neurotransmitter noradrenaline [3]. Noradrenaline then mediates the activation of genes required for further development and the release of exsheathment fluid to induce moulting of the L2 cuticle through a separate pathway. In the absence of CO$_2$ stimulation, both pathways appear to be inhibited in L3s by an analogue of the "insect juvenile hormone", which is implicated in the regulation of development in insects [3-6]. It has been suggested that the CO$_2$ cue is sensed by chemosensory neurons of the amphids [7,8], which are sensory organs located in the head of L3s and are exposed to the external environment *via* a pore [9]. The exsheathment process can be induced *in vitro* [10] and is recognised to suitably represent the process *in vivo* in the host animal [10,11]. In spite of knowledge of the exsheathment process, the role of carbonic anhydrase and the chemoreception are poorly understood in *H. contortus*, and the exact nature of the regulation and components of the "downstream" molecular pathways are not yet known [1].
Despite these substantial knowledge gaps for *H. contortus*, Rogers and Petronijevic [4] did hypothesize early that there is likely to be a specific pattern of gene expression during the development of nematodes, whereby genes which control continuous “house-keeping” processes and maintain survival are expressed on a constitutive basis, whereas genes which have specific functions in one or more stages of development are regulated for expression exclusively at specific developmental time points. Although studies of *H. contortus* have shown that the pattern of transcription differs between free-living and parasitic stages [12-14], there has been no detailed study of the molecular alterations occurring during the early phase of transition to parasitism in this parasite, although there is some information for the canine hookworm, *Ancylostoma caninum* [15-18]. Insights into the transcriptome of *H. contortus* during this critical phase of establishment in its host would enhance knowledge of developmental processes at the molecular level and might identify new intervention targets.

Advances in genomic, proteomic and bioinformatic technologies [19-22] now provide opportunities for exploring the molecular basis of developmental switches in *H. contortus* and other nematodes. In particular, the combined use of next-generation sequencing, such as 454 technology, SOLiD and Illumina/Solexa [19,23-25] and improved bioinformatic algorithms for the analysis and annotation of expressed sequence tag (EST) datasets [26] are suited to elucidate molecular changes at the transcriptomic level. The present chapter (i) provides the first detailed insights into the transcriptome of *H. contortus* during its transition from the L3 to the exsheathed, parasitic xL3 stage using a next generation sequencing-based approach, (ii) predicts, employing a number of bioinformatic approaches, the functional roles of these molecules in larval development and the metabolic pathways linked to this transition, and (iii) discusses the implications of the findings in relation to the fundamental, developmental biology of nematodes as well as applied aspects of developing new methods of nematode control.

4.2 Materials and methods

4.2.1 Parasite material

L3s of *H. contortus* ("Haecon5 strain") were cultured [27] from the faeces from an experimental sheep with a nonspecific infection of this strain using a protocol described previously [28]. Animal ethics approval (AEC no. 0707528) was given by The University of Melbourne, and the care and maintenance of sheep followed this institution's guidelines. The L3s were sieved (mesh size: 20 μm), washed extensively in sterile water and then stored at 10 °C for 18 days. L3s (n = 500,000) were exsheathed
[10], consistently achieving an exsheathment rate of 99%. Exsheathment using CO₂ in vitro [10] is recognised to represent the process in vivo within the host animal [10,11]. L3s and xL3s were each snap frozen in liquid nitrogen and then stored at -70 °C until use.

4.2.2 Library construction and next-generation sequencing

For each L3 and xL3 of *H. contortus*, a normalised cDNA library was constructed (Eurofins MWG Operon, Ebersberg, Germany). In brief, total RNA was extracted from either L3 or xL3 using the mirVana isolation kit (Ambion). The integrity of each RNA sample was verified using a Bioanalyzer 2100 (Agilent Technologies), and the yield determined spectrophotometrically (ND-1000 UV-VIS v.3.2.1, NanoDrop Technologies). RNA was then treated with DNase I (DNAfree, Ambion) and polyA+ RNA purified from ~120 ng of total RNA. For each library constructed, first-strand cDNA was synthesized using an oligo(dT)-adapter primer, and the second-strand was produced using a random (N)ₖ-adapter primer. The resultant double-stranded cDNA was amplified (21 cycles) by Long and Accurate PCR (LA-PCR) [29] and an aliquot (~200 ng) examined electrophoretically. Normalisation was conducted using one cycle of denaturation and reassociation of the cDNA, followed by purification of the reassociated double-stranded cDNA on a hydroxylapatite column. The normalised cDNA (500-700 bases) was then amplified using 9 cycles of LA-PCR and sequenced in a Genome Sequencer™ (GS) FLX Instrument (Roche Diagnostics) employing a standard protocol [19].

4.2.3 Bioinformatic analyses

Expressed sequence tags (ESTs) determined from the L3 and xL3 libraries were subjected to separate bioinformatic analyses. For each stage, all ESTs were pre-processed (SeqClean [30]; RepeatMasker [31]), aligned and then clustered using the Contig Assembly Program v.3 (CAP3) [32], employing a minimum sequence overlap of 30 bases and an identity threshold of 95% (for the removal of the flanking adapter sequences), and assembled. The small number of sequences (n = 55; 0.24% of 23,245 contigs) with a perfect match to those available for *Ovis aries* [GenBank: GI:3288836-GI:280977729; e-value cut-off: < 1e -15] were eliminated from each dataset. Both unassembled and assembled EST datasets were compared, at the nucleotide level using BLASTn (e-value cut-off: < 1e -05), with genomic sequence data publicly available for *H. contortus* (http://www.sanger.ac.uk; 21st August 2008). Contigs and singletons in each EST dataset, and the corresponding genomic sequence(s), were subjected to BLASTx (NCBI: www.ncbi.nlm.nih.gov) and BLASTn (EMBL-EBI Parasite Genome BLAST Server: www.ebi.ac.uk) searches to identify putative orthologues/homologues in
Caenorhabditis elegans, other nematodes, and organisms other than nematodes (e-
value cut-off: < 1e-05). WormBase (www.wormbase.org) was interrogated for relevant
information on C. elegans orthologues/homologues, including RNA interference (RNAi)
phenotypes as well as transcriptomic, proteomic and interactomic data.
Orthologues/homologues predicted from individual ESTs were always consistent with
those inferred from genomic sequences (not shown). ESTs with no match to any
sequences in the NCBI and/or EMBL-EBI databases were subjected to BLASTn analysis
against the genomic data for H. contortus to increase the likelihood of identifying
orthologues/homologues in currently available databases.

Following the conceptual translation of contigs and singletons into peptides
using ESTScan [33], functional annotation was performed by gene ontology (GO) using
BLAST2GO [34]. Domains/motifs within translated peptides were identified via
InterProScan [35] and linked to pathways in C. elegans using the KEGG Orthology-
Based Annotation System (KOBAS; [36]). The open reading frames (ORFs) inferred
from sequences with orthologues in C. elegans were also subjected to 'secretome
analysis' using the program SignalP v.2.0 (available at
www.cbs.dtu.dk/services/SignalP/), employing both the neural network and hidden
Markov models to predict signal peptides and/or anchors [37-39]. In addition,
transmembrane domains were inferred using the program TMHMM
(www.cbs.dtu.dk/services/TMHMM/; [40-42]).

4.2.4 Identification of transcripts unique to either L3 or xL3 by in silico subtraction, and
verification of specificity by PCR-coupled sequencing

In order to identify transcripts unique to each stage, the L3 and xL3 datasets
were subtracted from one another (in both directions) using in silico subtraction (at
both the nucleotide and the amino acid levels) employing the BLASTn and tBLASTx
algorithms, set at a stringent cut-off (e-value cut-off: < 1e-15). Subsets of molecules
(sequences) that were present in either L3 (n = 10) or xL3 (n = 10) or in both stages (n
= 10) were selected for subsequent experimental verification of specificity by reverse-
transcription (rt)-PCR. Forward and reverse primers were designed to selected contig
sequences and corresponding genomic sequences (available at
http://www.sanger.ac.uk) (Appendix 4.1), ensuring that at least one intronic region in
the genomic sequence was included (per sequence) to enable the detection by rtPCR of
any possible residual genomic DNA in the RNA used. The specificity of each primer
designed was evaluated in silico by BLASTn analysis against all presently available
databases, including that containing H. contortus genomic data. The rtPCR was carried
out as described previously [43] with the following modifications: (i) the cycling
conditions were: 95 °C, 2 min (initial denaturation) followed by 35 cycles of 95 °C for 1 min (denaturation), 55 °C for 1 min (annealing) and 72 °C for 1 min (extension), and a final extension of 72 °C for 5 min; (ii) in addition to part of the β-tubulin (250 bp) gene being used as a positive control, a portion of the elongation factor-1α (216 bp) gene was included for the amplification (cf. Appendix 4.1). Negative control reactions using template RNA from L3 and xL3 were also included. All amplicons produced were sequenced [44] to demonstrate unequivocally that they represented the correct sequence. The Fisher’s exact test was used to confirm that the number of molecules selected from each dataset (obtained following in silico subtraction and containing sequences unique to either L3 or xL3) for the verification of differential transcription by rtPCR was representative statistically (p < 0.001).

4.2.5 Probabilistic functional gene networking

The method developed by Zhong and Sternberg [45] was used to predict the interaction networks among *C. elegans* orthologues of molecules transcribed in either L3 or xL3. Data regarding interactions, phenotypes, expression and GO for selected *C. elegans* gene orthologues/homologues, also incorporating data from *Drosophila melanogaster* (vinegar fly), *Saccharomyces cerevisiae* (yeast), *Mus musculus* (mouse) and *Homo sapiens* (human), were integrated using a naïve Bayesian model to predict genetic interactions among *C. elegans* genes using the recommended, stringent cut-off value of 4.6 [44,45]. The predicted networks resulting from the analyses were saved in a graphic display file (gdf) format, examined using the graph exploration system available at http://graphexploration.cond.org/ and drawn using Adobe Illustrator CS2 (Adobe Systems Inc.). The genes predicted to interact with selected transcripts unique to L3 and xL3 were also classified according to the ‘Biological process Gene Ontology annotations of their interacting partners’ using the PROSTIDIN website (http://crfb.univ-mrs.fr/webdistin/; [46]).

4.3 Results

Totals of 101,305 (L3) and 105,553 (xL3) ESTs (average of 210–216 bases in length) were generated using 454 technology. The vast majority (95%) of these ESTs matched publicly available genomic DNA sequences of *H. contortus*. A summary of the characteristics of the raw and assembled data is given in Table 4.1. For L3, 9,046 proteins were inferred from a total of 20,066 assembled nucleotide sequences (Table 4.2), of which 3,066 matched known proteins with 1,465 different domains (Fig. 4.1a; Appendix 4.2). For xL3, 9,001 proteins were inferred from a total of 20,116 assembled nucleotide sequences (Table 4.2), of which 2,885 mapped to known proteins with 1,394
distinct domains (Fig. 4.1a; Appendix 4.2). For both L3 and xL3, the ‘transthyretin-like’ (8%; IPR001534), ‘calcium-binding EF-hand’ (6.7%; IPR002048), ‘nicotinamide adenine dinucleotide phosphate (NAD(P))-binding domain’ [NAD(P)-binding] (5%; IPR016040), and ‘nucleotide-binding, alpha beta plait’ (5%; IPR012677) domains were most commonly detected (see Fig. 4.1a; Appendix 4.2).

The subsequent annotation of the proteins inferred for the L3 stage revealed 1,229 different Gene Ontology (GO) terms, of which 334 were linked to ‘biological process’, 116 ‘cellular component’ and 779 ‘molecular function’ (see Fig. 4.1b; Appendix 4.3). For xL3, 1,213 GO terms were identified, of which 330 represented ‘biological process’, 93 ‘cellular component’ and 790 ‘molecular function’. For both L3 and xL3, the most common GO terms were ‘translation’ (9.2%; GO:0006412) and ‘metabolic process’ (7.1%; GO:0008152) for ‘biological process’, ‘intracellular’ (18.7%; GO:0005622) and ‘ribosome’ (15.1%; GO:0005840) for ‘cellular component’ and ‘structural constituent of ribosome’ (9%; GO:0003735) and ‘catalytic activity’ (6.3%; GO:0003824) for ‘molecular function’ (see Fig. 4.1b; Appendix 4.3). Pathway mapping using KOBAS predicted a total of 253 different pathways, of which the most represented were ‘ribosome’, ‘oxidative phosphorylation’, ‘chaperones and folding catalysts’ (see Appendix 4.4).

Subtractive bioinformatic analysis identified 560 and 685 nucleotide sequences as being uniquely transcribed in the L3 and xL3 stages, respectively [available via http://research.vet.unimelb.edu.au/gasserlab/index.html]. The uniqueness of selected (n = 30) transcripts was verified by independent rtPCR-coupled sequencing of amplicons. Of the sequences unique to L3, 336 had orthologues in C. elegans (Appendix 4.5), 15.7% of which encoded ribosomal proteins (n = 32), collagens and elongation factors (n = 21). Of the 555 functional domains predicted for all proteins, ‘glutathione S-transferase/chloride channel’ (4.7%; IPR017933), ‘homeodomain-like’ (3.7%; IPR009057) and ‘MORN motif’ (3%; IPR003409; Fig. 4.1a and Appendix 4.2) were abundant. Among the 204 GO terms identified in the L3 ‘unique’ dataset, the most common were ‘isoprenoid biosynthetic process’ (4.6%; GO:0008299) ‘DNA recombination’ (3%; GO:0006310) and ‘DNA replication’ for ‘biological process’; ‘alpha DNA polymerase, primase complex’ (1.1%; GO:0005658) for ‘cellular component’; and, ‘cation binding’ (6.8%; GO:0443169), ‘metallo carboxypeptidase activity’ (5%; GO:0004181) and ‘4-hydroxyphenylpyruvate dioxygenase activity’ (1.8%; GO:0003868) for ‘molecular function’ (Fig. 4.1b and Appendix 4.3). KOBAS mapping revealed links to 21 different pathways, of which ‘biosynthesis and biodegradation of secondary metabolites’ and ‘cell motility and secretion’ were the most commonly represented (Appendix 4.4).
Of the 685 sequences unique to xL3, 421 had known orthologues in *C. elegans*, the largest number (i.e., 5%) representing peptidases and other enzymes of the amino acid catabolism (n = 21) (see Appendix 4.5). Among the 484 functional domains identified in the InterProScan analysis, ‘globin’ or ‘globin-like’ domains (19.7%; IPR000971, IPR012292 and IPR009050) had the greatest representation (Fig. 4.1a; Appendix 4.2). The GO analysis identified a total of 188 terms, the most common being ‘oxygen transport’ (16.9%; GO:0015671), ‘regulation of phosphorylation’ (9.1%; GO:0042325) and ‘nitrogen compound metabolic process’ (4.7%; GO:0006807) for ‘metabolic process’; ‘transcription factor TFIIA complex’ (10.3%; GO:0005672) for ‘cellular component’; and, ‘oxygen binding’ (11%; GO:0019825), ‘phosphoric ester hydrolase activity’ (6.3%; GO:0042578) and ‘dopamine beta-monooxygenase activity’ (2.6%; GO:0004500) for ‘molecular function’ (Fig. 4.1b; Appendix 4.3). Among the seven KOBAS pathways predicted for the xL3-specific data, ‘cytokine-cytokine receptor interaction’ and ‘D-arginine and D-ornithine metabolism’ were most commonly represented (Appendix 4.4).

Twenty-one *C. elegans* orthologues (i.e., genes *elc-1*, *rps-22*, *sod-2*, *rps-14*, *rpl-24.1*, *rpl-32*, *pas-4*, *ced-10*, *T10B11.2*, *krs-1*, *rpl-31*, *pas-5*, *cul-3*, *F21D5.7*, *rps-27*, *sno-1*, *rps-11*, Y37E3.8a, *rpl-5*, *col-97* and F57A8.2b) unique to L3 (cf. Appendix 4.6) were predicted to interact directly with a total of 535 other genes (range: 1-227, see Fig. 4.2a and Appendix 4.7). In contrast, seven *C. elegans* orthologues (i.e., *ncbp-2*, F57B10.3a, *vab-9*, *ncs-1*, *cpr-6*, Y52B11A.2a and *nhr-80*) unique to xL3 (Appendix 4.8) were inferred to interact with 45 other genes (range 1-27; see Fig. 4.2b and Appendix 4.7). The functional classification of the *C. elegans* orthologues of molecules which were unique to L3 revealed one predominant group linked to ‘embryonic development’ (within ‘biological process’); this cluster contained molecules associated with ‘regulation of development’ and other pathways, including ‘positive regulation of growth’, ‘RNA metabolism’ and ‘biosynthesis’ (Fig. 4.3a; Appendix 4.7). Orthologues unique to xL3 also revealed one main cluster associated with ‘embryonic development’ (within ‘biological process’) (Fig. 4.3b; Appendix 4.7).

### 4.4 Discussion

In *H. contortus*, the transition to parasitism was proposed to be accompanied by differences in the levels of transcription of particular genes that may facilitate the invasion of the host and the evasion of the immune response [12]. Next-generation sequencing technology was used here to investigate the transcriptome of the *H. contortus* L3 before and after the exsheathment process *in vitro*. Interestingly, ~75% of the sequences determined did not have orthologues/homologues in any other
organism for which sequence data are available. However, annotation was enhanced through the mining of genomic data available for *H. contortus* (http://www.sanger.ac.uk; August 2008), ultimately increasing the percentage of annotated sequences to ~50% for the L3 and xL3 datasets (see Results section). The likely explanation for this result is technical and would appear to relate to a 3'-bias in sequence reads [47], thus affecting the prediction of ORFs and, thus, the identification of orthologues/homologues. In the absence of the complete genome sequence for *H. contortus*, genomic mining provided the sole option for maximizing the mapping and subsequent annotation of sequences.

The integrated approach of using both transcriptomic and available genomic data allowed the annotation of a total of 11,302 sequences, of which 5,117 matched unique *C. elegans* orthologues. Of the 3,010 unique InterPro domains identified, the most represented in both L3 and xL3 were the 'transthyretin-like' (IPR001534), 'calcium-binding EF-hand' (IPR002048), 'NAD(P)-binding' (IPR016040) and 'sperm-coating protein (SCP)-like extracellular' (IPR014044) motifs (Appendix 4.2). Most 'transthyretin' proteins identified to date are conserved (across large evolutionary distances) [48] and have been recognized as enzymes of the purine catabolism that catalyse the conversion of 5-hydroxyisourea (HIU) to OHCU [49,50]. Of the 4,000 'nematode-specific' protein families encoded by 'nematode-restricted' genes, the 'transthyretin-like' proteins (TTLs) [51] represent one of the largest groups [52]. Indeed, members of the TTL family have been detected in plant parasitic nematodes, such as *Xiphinema index*, *Heterodera glycines*, *Meloidogyne incognita* and *Radophillus similis* [51,53-55], the filarial nematode of humans, *Brugia malayi* [56], and *Ostertagia ostertagi* of ruminants (which is related to *H. contortus*) [57,58]. In *Os. ostertagi*, at least 18 *ttl* genes have been identified by database mining; most of these genes are constitutively transcribed from the free-living L3 stage through to adult males and females [58]. In parasitic nematodes, TTLs might play a role as carriers of lipophilic substances or hormones [52]. More recently, Jacob and coworkers [51] localized the expression of *ttl* genes to the ventral nerve of *R. similis* (i.e., Rs-ttl-2) and to the nervous system and hypodermis of *C. elegans* (gene code R13A5.6; www.wormbase.org). These findings, together with the sequence similarity between TTLs and some neuropeptides, suggest that these proteins play a yet uncharacterized role in the nematode nervous system [51]. In *H. contortus*, a TTL has been identified previously using a proteomic approach and shown to be the most immunogenic protein in excretory/secretory products (E/S) of the adult stage [59]. TTLs are also highly abundant in E/S of the canine hookworm, *An. caninum* [60]. These latter two studies propose a central role for
these proteins in the parasite-host interplay [59,60], which warrants detailed investigations.

Some host-parasite interactions have been reported to rely on the essential role played by calcium-mediated signalling pathways in protein secretion, motility, cell invasion and/or differentiation [61-63]. These functions are controlled by various, specialized subcellular structures (such as the Golgi apparatus, and some channels and transporters) for the uptake and release of calcium, which acts as a secondary messenger for the activation of calcium-dependent proteins [64], particularly those containing ‘EF-hand’ domains. In the present study, the large number of sequences (equating to 6.7% overall) with such domains in the L3 and xL3 stages of *H. contortus* might reflect a substantial need for calcium ions in the cascade of developmental events occurring during larval growth, particularly those linked to the maturation of the nervous system, as proposed to function in the *C. elegans* dauer stage [65]. Similarly, the redox reactions catalysed by oxidoreductases with a NAD(P)-binding domain are essential for all developmental processes [66]. The NAD+, a coenzyme found in all living cells, is involved in redox reactions and carries electrons from one reaction to another. In contrast, the main function of NADP+ is as a reducing agent in anabolism, with this coenzyme being involved in key pathways, such as fatty acid synthesis and glucose metabolism [66].

Another relatively abundant group of sequences in both L3 and xL3 encoded key SCP/Tpx-1/Ag5/PR-1/Sc7 proteins (designated SCP/TAPS; Pfam accession number no. PF00188), characterized by one or more SCP-like extracellular domains (InterPro: IPR014044; \( p < 1\times10^{-05} \) [67]. In parasitic nematodes, such proteins are also called *Ancylostoma*-secreted proteins or activation-associated proteins (ASPs), mainly because they were originally discovered in hookworms [18,68]. Although the numbers of sequences encoding SCP/TAPS (187 and 207, assembled into three and seven EST clusters, respectively) were similar in L3 and xL3 (cf. Table 4.3), there were some qualitative differences in transcription between these two stages. Homologues of the molecules designated Hc24 and Hc40, previously identified in excretory/secretory (ES) products of adult *H. contortus* [69,70], were encoded in both of these stages. However, other previously undescribed SCP/TAPS, such as an inferred protein homologue (encoded by *C. elegans* gene F09B9.5) with no known homologues in any other nematode, were encoded only by sequences in xL3 (cf. Table 4.3). The finding of entirely novel SCP/TAPS supports a previous hypothesis, formulated based on observations from transcriptomic and proteomic analyses, that a broader range of molecules of this group is expressed in *H. contortus* [59]. This finding also suggests a diversified, active and specific involvement of SCP/TAPS in the exsheathment process,
establishment of *H. contortus* in its host and/or the parasite-host interplay. Current
evidence indicates that the transcriptional profile for *H. contortus* in the transition from
L3 to xL3 is quite distinct from that of *An. caninum*, in that there was a massive
expansion in SCP/TAPS in the serum-activated L3 (i.e., xL3) in the hookworm [18].
Although there were methodological differences between studies, particularly in
relation to the use of normalised cDNA libraries herein, an apparent difference in the
diversity of transcripts encoding SCP/TAPS might reflect distinct roles for various
types of these molecules in *H. contortus* compared with *An. caninum* during the
invasion process in the host. Specifically, L3s of *H. contortus* are ingested by the
ruminant host, exsheath in the reticulo-rumen and establish in the abomasum, in which
the xL3s start feeding on blood to subsequently develop into dioecious adults [1]. In
contrast, infective L3s of *An. caninum* penetrate skin of the canid host, exsheath and
then migrate as xL3s via the circulatory system and the lung to finally reside as adults
in the small intestine; the dioecious hookworm adults attach via their buccal capsule to
the intestinal mucosa, rupture capillaries and feed on blood [71]. Some xL3s also
undergo somatic migration to then encyst in tissues, such as muscle and the mammary
gland, and become arrested as L3s [72-74]. Therefore, the added biological complexity
in the hookworm life cycle, particularly in the initial phases of invasion and migration
as an xL3, compared with the relatively simple gastric infection process for *H.
contortus*, supports the current proposal of a very distinct arsenal of SCP/TAPS (both
qualitatively and quantitatively) between the two nematodes during their transition to
parasitism. Nonetheless, only a direct comparison of transcription between L3 and xL3
for the two parasites using the same or a similar approach as employed herein will
allow this hypothesis to be tested conclusively.

The *in silico* subtraction approach employed in the present study identified 560
and 421 sequences that were specific to *H. contortus* L3 and xL3, respectively. The
specificity of a subset of these transcripts was unequivocally verified by rtPCR-coupled
sequencing (see Results section). Most of the subset of ‘L3-specific’ molecules with
known orthologues in *C. elegans* encoded ribosomal proteins (*n* = 32), collagens (*n* =
10) and elongation factors (*n* = 11) (cf. Appendix 4.6). In nematodes, the synthesis of
collagens has been observed to increase significantly prior to a moult [75], whereas
ribosomal proteins have been reported to participate in various cellular processes
besides protein biosynthesis. For instance, they can act as components of the
translation apparatus and also regulate cell proliferation and apoptosis [76]. Indeed,
genetic interaction networking for *C. elegans* orthologues of *H. contortus* L3-specific
transcripts predicted a number of genes coding for ribosomal proteins (i.e., *rps-11, rps-
14, rps-22, rps-27, rpl-5 and rpl-24.1*) to interact directly with a GTPase (i.e., *ced-10,*
which is required for phagocytosis during programmed cell death and for migration of the distal tip cells of the somatic gonad [77] (see Fig. 4.2a; Appendix 4.7). In addition, the protein CED-10 has been demonstrated to be essential in the cascade of events which leads to the development of the nervous system in the growing and developing larva, by guiding the migrating cells and axonal growth cones to their final position in the body [78]. A direct interaction was also predicted between ced-10 and another C. elegans orthologue of an L3-specific transcript (i.e., pas-4) encoding a proteasome alpha-type seven subunit of the core 20S proteasome subcomplex (see Fig. 4.2a; Appendix 4.7; [www.wormbase.org]) involved in the ‘ubiquitin proteasome system’ (UPS) [79]. Loss-of-function in this system is known to be the primary cause or secondary consequence of pathological conditions of the nervous system (e.g., [80]). Moreover, gene silencing of pas-4 in C. elegans has been shown to result in a wide variety of defects, including embryonic and larval lethality, sterility, abnormal locomotion, slow growth, and abnormal transgene expression and subcellular localization [79] ([www.wormbase.org]). Here, both ced-10 and pas-4 were predicted to interact with a C. elegans orthologue of SUMO (i.e., smo-1; Fig. 4.2a; Appendix 4.7). SUMO is a small ubiquitin-like moiety that, when attached to protein substrates, regulates subcellular localization and activity [81,82]. The loss of function of smo-1 results in developmental defects of the nervous and reproductive systems as well as embryonic or larval lethality [81,82]. Based on the genetic interaction networks predicted for orthologues of H. contortus L3-specific molecules, it could be proposed that axon guidance and synapse formation, which are primarily controlled by UPS, through the regulation of protein turnover at the growth cone and the synapse [80], are crucial in the biology of the L3 of this parasitic nematode. This statement is supported by the fact that the exsheathment process is triggered by gaseous CO₂, detected by chemosensory neurons of amphids, located in the anterior end of the L3 [7,8], ultimately leading to the secretion of the neurotransmitter noradrenaline [1,3].

Of the H. contortus xL3-specific transcripts, those predicted to be involved in mechanisms regulating chemotaxis during larval development were also highly represented. For example, the C. elegans orthologue ncs-1 encodes a neuronal calcium sensor protein, whose expression in C. elegans is primarily associated with two sensory organs, the head amphids and tail phasmds [83]. This orthologue was predicted to interact with 11 distinct genes involved in axon guidance, neuron projection, embryonic and/or larval development (see Fig. 4.2b; Appendix 4.7). However, the largest number of genetic interactions was predicted for a C. elegans orthologue of a H. contortus xL3-specific transcript encoding a subunit of the nuclear cap-binding complex (i.e., ncbp-2; Fig. 4.2b; Appendix 4.7). This complex includes RNA-binding proteins that
bind to the 5'-cap within the nucleus; when RNA is exported to the cytoplasm, the nuclear cap-binding protein complex is replaced by cytoplasmic cap-binding complex [84]. Although ncbp-2 has been demonstrated to play key roles in mRNA decay, embryonic/larval development, reproduction and vulval morphogenesis [85], its precise role in the biology of the developing larva is still unclear.

The largest number of C. elegans orthologues of H. contortus xL3-specific transcripts encoded peptidases and other enzymes of the amino acid catabolism (cf. Appendices 4.5 and 4.8). A similar spectrum of proteases and other molecules linked to catalytic activity have been shown also to be highly represented in the transcriptome of the serum-activated xL3 stage of An. caninum by comparison to its L3 [18]. This finding, for two haematophagous bursate nematodes with differing biologies, is likely to reflect the key roles that these molecules play in host tissue invasion, degradation and/or digestion [1,18]. In particular, a cysteine protease (i.e., cpr-6) was amongst the C. elegans orthologues of xL3-specific molecules in H. contortus (see Appendix 4.8). This finding is supported by previous evidence, showing that cysteine proteases play a crucial role in the catabolism of globin peptides by the cleavage of haemoglobin in blood-feeding nematodes (e.g., An. caninum and Necator americanus) [86-89]. Given this role, these proteases have been considered as promising candidates for developing recombinant vaccines against H. contortus as well as hookworms [90,91]. Typically, these proteases include a 'globin domain' that protects the haeme iron from rapid oxidation and regulates oxygen homoeostasis [92] in the gut of the host, which is characterized by a reduced oxygen tension [93]. Indeed, this domain (represented by InterPro codes IPR000971, IPR012292 and IPR009050; cf. Fig. 4.1a; Appendix 4.2) was identified as the most abundant motif amongst the xL3-specific molecules. In H. contortus, transcripts encoding cathepsin B cysteine protease-like (CBL) proteins have been reported to be most abundant (~16%) in the intestine of the adult female [94]. In addition, a comparison of H. contortus isolates from Europe and North America has revealed a minor to moderate nucleotide diversity in cbl genes, which has been proposed to reflect antigenic variation among CBLs [94]. Clearly, C. elegans provides a useful surrogate system [95] to express H. contortus cysteine-proteases and to assess their function(s).

In conclusion, the present chapter has provided new and exciting insights into the molecular biology of the L3 stage of H. contortus and has elucidated transcriptional alterations taking place during the transition from the free-living to the parasitic stage of this nematode. Although approximately half of the sequences generated by 454 sequencing remains uncharacterised, as a consequence of the absence of complete genomic information for this parasite, future sequencing efforts using, for example,
Illumina technology, should provide the efficiency and depth of coverage required to define the complete transcriptomes of all developmental stages and various tissues as well as the genome of *H. contortus*. The determination of the genome sequence of *H. contortus* has major potential to accelerate large-scale studies of genes and gene products involved in nematode development and reproduction, parasite-host interactions and the disease caused by the parasite. Importantly, comprehensive genomic and transcriptomic data will also underpin future proteomic and metabolomic studies of *H. contortus*. Such an integrated approach should lead to important conceptual advances in our understanding of various aspects of nematode biology and should have major implications, in the medium to long term, for the development of novel strategies for parasite intervention, resulting in biotechnological outcomes (such as drugs and diagnostic tests). For instance, future work could focus on defining a spectrum of key molecules involved in pathways linked to the development of the nervous system in different stages of *H. contortus* and assessing their potential as drug targets. Moreover, determining the structure and function of SCP/TAPS homologues/orthologues could establish their role(s) in the invasion of and establishment of this parasite in the host animal, providing the prospect of interfering with the host-parasite relationship. Although the present chapter focused on *H. contortus*, the research findings and the integrated technological approach employed should find broad applicability to other parasitic nematode of major global importance.
4.5 References


Table 4.1 Numbers of expressed sequence tags (ESTs) determined from cDNA libraries representing the ensheathed (= L3) or exsheathed (= xL3) third larval stage of *Haemonchus contortus*, and numbers of sequences (including contigs and singletons) and clusters and their lengths (in brackets) before and after assembly.

<table>
<thead>
<tr>
<th>Numbers of</th>
<th>L3</th>
<th>xL3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequences before assembly of ESTs (average length)</td>
<td>101,305</td>
<td>105,553</td>
</tr>
<tr>
<td>Sequences after assembly</td>
<td>11,824</td>
<td>11,671</td>
</tr>
<tr>
<td>Contigs (average length ± standard deviation)</td>
<td>6,299 (352.7 ± 91.9)</td>
<td>5,574 (361.6 ± 99.8)</td>
</tr>
<tr>
<td>Singletons (average length ± standard deviation)</td>
<td>5,525 (274.1 ± 45.7)</td>
<td>6,097 (278.7 ± 60.2)</td>
</tr>
<tr>
<td>Clusters matching genomic sequences</td>
<td>10,823</td>
<td>11,021</td>
</tr>
<tr>
<td>Total number of unique clusters (ESTs + GSS)</td>
<td>20,066</td>
<td>20,116</td>
</tr>
</tbody>
</table>
Table 4.2 Terms (i.e. 'biological process', 'cellular component' and/or 'molecular function') representing proteins inferred to be encoded by either the ensheathed (= L3) or exsheathed (= xL3) third larval stage of *Haemonchus contortus*.

<table>
<thead>
<tr>
<th>Term</th>
<th>Numbers of ESTs / genomic sequences in L3</th>
<th>Numbers of ESTs / genomic sequences in xL3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequences with ORFs</td>
<td>6847 / 2199</td>
<td>6696 / 2305</td>
</tr>
<tr>
<td>InterPro</td>
<td>1876 / 1190</td>
<td>1614 / 1271</td>
</tr>
<tr>
<td>GO</td>
<td>1382 / 505</td>
<td>1171 / 540</td>
</tr>
<tr>
<td>KOBAS (pathway mapping)</td>
<td>2144 / 688</td>
<td>1788 / 1021</td>
</tr>
<tr>
<td><em>C. elegans</em> homologues/orthologues</td>
<td>2975 / 2630</td>
<td>2922 / 2782</td>
</tr>
<tr>
<td>Homologues in organisms other than nematodes</td>
<td>2069 / 928</td>
<td>1689 / 1023</td>
</tr>
<tr>
<td>No known homologues in available databases</td>
<td>3274 / 4968</td>
<td>3349 / 4709</td>
</tr>
</tbody>
</table>
Table 4.3 Expressed sequence tags (ESTs) inferred to encode SCP/TAPS proteins in the ensheathed (= L3) or exsheathed (= xL3) third larval stage of *Haemonchus contortus*.

<table>
<thead>
<tr>
<th>EST code</th>
<th>Length (bases)</th>
<th>Homologue</th>
<th>e-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contig875</td>
<td>530</td>
<td><em>Ancylostoma caninum</em> Ancylostoma-secreted protein 1 (ASP-1)</td>
<td>1e - 09</td>
</tr>
<tr>
<td>FJISXER02CFZ04</td>
<td>250</td>
<td><em>Haemonchus contortus</em> Hc24</td>
<td>1e - 22</td>
</tr>
<tr>
<td>FJISXER02CG110</td>
<td>263</td>
<td><em>Ancylostoma caninum</em> Ancylostoma-secreted protein 2 (ASP-2)</td>
<td>6e - 17</td>
</tr>
<tr>
<td>xL3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contig990</td>
<td>337</td>
<td><em>Ancylostoma caninum</em> Ancylostoma-secreted protein 1 (ASP-1)</td>
<td>2e - 41</td>
</tr>
<tr>
<td>Contig1141</td>
<td>271</td>
<td><em>Meloidogyne arenaria</em> venom allergen-like protein-1 (MaVAP-1)</td>
<td>6e - 09</td>
</tr>
<tr>
<td>Contig2101</td>
<td>431</td>
<td><em>Caenorhabditis elegans</em> hypothetical protein F09B9.5</td>
<td>2e - 28</td>
</tr>
<tr>
<td>Contig2955</td>
<td>256</td>
<td><em>Ostertagia ostertagi</em> activation associated secreted protein ASP-4 (Oo-ASP-4)</td>
<td>9e - 16</td>
</tr>
<tr>
<td>Contig3360</td>
<td>315</td>
<td><em>Haemonchus contortus</em> Hc40</td>
<td>3e - 24</td>
</tr>
<tr>
<td>Contig4207</td>
<td>351</td>
<td><em>Haemonchus contortus</em> Hc40</td>
<td>4e - 53</td>
</tr>
<tr>
<td>Contig4530</td>
<td>355</td>
<td><em>Ancylostoma ceylanicum</em> Ancylostoma-secreted protein 1 (ASP-1)</td>
<td>4e - 13</td>
</tr>
<tr>
<td>FJISXER05F8XC0</td>
<td>246</td>
<td><em>Ancylostoma caninum</em> Ancylostoma-secreted protein 5 (ASP-5)</td>
<td>1e - 05</td>
</tr>
<tr>
<td>FJISXER06G7M9K</td>
<td>242</td>
<td><em>Meloidogyne arenaria</em> venom allergen-like protein-1 (MaVAP-1)</td>
<td>4e - 05</td>
</tr>
<tr>
<td>FJISXER06GY7FZ</td>
<td>266</td>
<td><em>Haemonchus contortus</em> Hc40</td>
<td>2e - 17</td>
</tr>
</tbody>
</table>
Fig. 4.1 Venn diagram displaying the number of protein domains (InterPro) (a) and Gene Ontology (GO) terms (b) common and uniquely represented in expressed sequence tags (ESTs; numbers in black) and genome survey sequences (GSS; in dark grey) data for the ensheathed (= L3) and exsheathed (= xl3) third larval stage of Haemonchus contortus. Descriptions of the 25 most abundant domains (a) and GO terms (b) are given in the boxes.
Fig. 4.2 Genetic interaction networks predicted for *Caenorhabditis elegans* orthologues of expressed sequence tags (ESTs) unique to either the ensheathed (= L3) (a) or exsheathed (= xL3) (b) larval stages of *Haemonchus contortus*. Representing L3: ced-10 was linked to axon guidance; krs-1 to lysine biosynthesis; pas-4 and pas-5 to proteasome system; cul-3, elc-1 and smo-1 to ubiquitin; F21D5.7 to protein export; rpl-5, rps-14, rpl-24.1, rpl-32 and Y37E3.8a to ribosome. Representing xL3: F57B10.3a was linked to glycolysis; ncs-1 to olfactory signal transduction; cpr-6 to antigen processing and presentation.
Fig. 4.3 Categorization of genes predicted to interact with *Caenorhabditis elegans* orthologues of expressed sequence tags (ESTs) unique to each ensheathed (= L3) (a) and exsheathed (= xL3) larval stage of *Haemonchus contortus*, clustered according to ‘biological process’ (Gene Ontology). The Gene Ontology hierarchies for individual clusters are given when known.
CHAPTER 5

Oligonucleotide microarray and bioinformatic analyses to explore transcriptional conservation between *Ancylostoma caninum* and *Haemonchus contortus*

Abstract

This chapter describes the use of an oligonucleotide microarray platform for the parasitic nematode *Haemonchus contortus* to identify transcripts that are 'conserved' between serum-activated and non-activated L3s of *Ancylostoma caninum* (aL3 and L3, respectively) and *H. contortus* by cross-species hybridization (CSH) at high stringency as well as extensive bioinformatic analyses of the cross-hybridizing expressed sequence tags (ESTs). The microarray analysis revealed significant differential hybridization between aL3 and L3 for 32 molecules from *An. caninum*, of which 29 were shown to have homologues in the free-living nematode *Caenorhabditis elegans* and/or *An. caninum* and the other three molecules had no homologues in current gene databases. 'Non-wildtype' RNAi phenotypes were recorded for 13 of the *C. elegans* homologues. A subset of 16 *C. elegans* homologues/orthologues (i.e., genes abce-1, act-2, C08H9.2, C55F2.1, calu-1, col-181, cpr-6, elo-2, asp-1, K07E3.4, rpn-2, sel-9, T28C12.4, hsb-1, Y57G11C.15 and ZK593.1) were predicted to interact genetically with a total of 156 (range: 1-88) other genes. Gene ontology (GO) analysis of the interacting genes revealed that the most common subcategories were signal transduction (7%), intracellular protein transport and glycolysis (6.2%) within 'biological process'; nuclear (25.7%) and intracellular (19.8%) within 'cellular component'; and ATP-binding (14.4%) and protein-binding (8.4%) within 'molecular function'. The potential roles of key molecules in the two blood-feeding parasitic nematodes are discussed in relation to the known roles of their homologues/orthologues in *C. elegans*. The CSH approach used may provide a tool for the screening of genes conserved across a range of different taxa of parasites for which DNA microarray platforms are not available.
5.1 Introduction

Parasitic nematodes of humans and animals are of major socio-economic importance world-wide (Chapter 1). For example, hundreds of millions of people are infected with geohelminths (soil-transmitted worms), such Ancylostoma duodenale and Necator americanus (blood-feeding hookworms), Trichuris trichiura (‘whipworm’) and Ascaris spp. (‘roundworm’) [1], causing serious adverse effects on human health, particularly in children. Similarly, parasitic nematodes of livestock, such as sheep and cattle, also cause substantial economic losses due to subclinical and clinical diseases, with billions of dollars lost annually on the control of gastro-intestinal nematodes. Traditionally, the control of nematodes has relied predominantly on the treatment with anthelmintic drugs, including benzimidazoles, closantel, pyrantel pamoate, levamisole and/or macrocyclic lactones [1,2]. However, based on the experience with nematodes of livestock [3], there is a risk that nematodes of humans could develop genetic resistance to these drugs, if used excessively [4-6]. For blood-feeding nematodes, there has been a strong focus on developing effective vaccines [7-12]. Recent studies have shown that the infective stage of hookworms, the third-stage larvae (L3), release immunogenic proteins that can be used in vaccines in recombinant form [8]. These immunogens include Ancylostoma-secreted proteins (ASPs) which belong to the pathogenesis-related protein (PRP) superfamily [13,14]. Presently, two ASPs (namely Na-ASP-1 and Na-ASP-2) are under investigation as vaccine candidates against necatoriasis in humans [15,16]. Vaccines against blood-feeding nematodes of livestock have also been pursued. Antigens or immunogens have been purified from worm extracts or in vitro-secretions, and some have achieved protective immunity in vaccination trials in sheep or cattle [17]. For instance, the integral membrane glycoprotein complex H11 from the gut of Haemonchus contortus (the barber’s pole worm) has been demonstrated to be highly immunogenic and protective against homologous challenge infection in sheep [18]. In spite of extensive efforts in these applied areas and some key successes, to date, there has been a relatively limited focus on exploring fundamental molecular aspects of the development and reproduction, host invasion and the host-parasite relationship for socio-economically important parasitic nematodes [19]. Detailed insights into such molecular processes, particularly those common among related parasitic nematodes and associated with the early phase in the transition from free-living, third-stage larvae (L3s) to the parasitic stages and the invasion of the host, provide opportunities for finding new ways of killing them or disrupting their development or reproduction in the host [19-21]. For instance, particular subsets of molecules involved in parasite-specific biological pathways could be common and conserved between or among relatively disparate nematodes of the same group or order.
Hookworms (superfamily Ancylostomatidae) belong to a large order of nematodes, called the Strongylida [22]. This order also comprises the superfamilies Strongyoidea, Metastrongyoidea and Trichostrongyoidea, which, according to a molecular study [23], belong to “clade V”. Like hookworms, Haemonchus species are also blood-feeding nematodes [22]. For instance, H. contortus belongs to the subfamily Trichostrongylidae and infects small ruminants but, unlike hookworms, infects its host exclusively via the oral route. Despite biological, ecological and evolutionary differences between these nematodes, both of them (at the adult stage) feed on blood in the gastrointestinal tract of the host, causing pathogenic effects, and their pattern of development through four larval stages to the dioecious adults is similar.

Recent advances in genomic and bioinformatic technologies now provide an opportunity to experimentally investigate “conserved elements” between these nematode species without knowledge of their entire genome sequences. For example, cross-species hybridization (CSH) by microarray analysis provides the opportunity to study transcriptional conservation among species [24,25]. This analytical approach is particularly useful for species for which no or limited genome information and expressed sequence tag (EST) data are available [25]. Recently, oligonucleotide microarrays (Agilent) were designed and applied to An. caninum to investigate in detail transcription during its developmental transition to parasitism (in vitro) [21] or to H. contortus to explore gender-enriched transcription [26]. The availability of these arrays, together with the extensive genomic and biological information on the free-living nematode Caenorhabditis elegans (which, like members of the order Strongylida, is also considered to belong to clade V; www.wormbase.org; [27]) for comparative analyses, provides opportunities for investigating particular transcripts conserved among An. caninum, H. contortus and/or also other nematodes. The present chapter describes the use of an oligonucleotide microarray platform established for H. contortus [26], to identify transcripts that are ‘conserved’ between both An. caninum and H. contortus by hybridization at high stringency for subsequent characterisation using various bioinformatic analyses.

5.2 Materials and methods

Third-stage, infective larvae (L3s) of An. caninum were cultured from the faeces from infected dogs, purified and stored as described previously [21]. The dogs were from Brisbane and environs (Queensland) where An. caninum is endemic in the canine host [28]. The specific identity of the L3s in each batch was confirmed using a PCR-coupled sequencing approach [29]; the sequences of the first and second internal
transcribed spacers (ITS-1 and ITS-2) determined were the same as those reported previously (GenBank accession numbers Y19181 and AJ001591, respectively; [30]).

For molecular analyses, two distinct batches of L3s from different dogs and geographical locations were produced. For each isolate, half of the L3s were activated (designated aL3s) in 25 mM S-methylglutathione in RPMI-C medium (Gibco) containing 15% (v/v) dog serum for 3 hrs, according the method originally described by Hawdon et al. [31] and the other half remained untreated (incubated in the same medium without serum). In aL3s, pharyngeal pumping was verified by feeding ~100 larvae with fluorescein-isothiocyanate (FITC)-labeled bovine serum albumin (BSA) (10 mg/ml) for 3 hrs and detection by epifluorescence microscopy [32]. Each batch of 40,000 L3s or aL3s was washed twice in phosphate-buffered saline (PBS, pH 7.4; 23 °C) and subsequently frozen at -80 °C.

Larvae were suspended in 100 µl Trizol reagent (Invitrogen) and homogenized in a 1.5 ml tube using an RNase-free, disposable, in-tube pestle and subjected to three rapid (1 min) freeze/thaw cycles. Trizol was added to a final volume of 500 µl, before snap freezing in liquid nitrogen and storage for ≤ 1 month at -80 °C. Total RNA was isolated from each batch of L3s and aL3s as described previously [21], its quality was verified using a Bioanalyzer 2100 (Agilent Technologies) and the yield determined using a spectrophotometer (ND-1000 UV-VIS v.3.2.1, NanoDrop Technologies). Each RNA sample (~10 µg) was treated with 2 U of DNase I (Promega), incubated at 37 °C for 30 min prior to heat denaturation of the enzyme (75 °C for 5 min) and then frozen at -70 °C. Using a similar approach to that described by Hu et al. [33], no genomic DNA had been detected in any of the RNA samples used [21].

Complementary RNA (cRNA) probes were produced from 200 ng of total RNA from L3s and aL3s of _An. caninum_ from each of the two isolates. Each sample was reverse-transcribed and simultaneously labelled with cyanine 3 or cyanine 5 (Low RNA Input Linear Amplification Kit PLUS Two-Color kits, Agilent). The quality of cRNAs (purified using RNeasy mini-spin columns, Qiagen) was verified spectrophotometrically and the amount determined employing the Bioanalyzer (RNA 6000 LabChip Kit, Agilent).

Microarray analysis was conducted using the oligonucleotide array constructed recently and representing multiple developmental stages, tissues and both sexes of _H. contortus_ [26] using recommended protocols (Agilent Technologies). This array represents 1,885 ESTs (three oligonucleotides spotted for each) of _H. contortus_. For each of two biological replicates, 200 ng each of cyanine 3-labelled _An. caninum_ L3 cRNA and cyanine 5-labelled aL3 cRNA (or vice versa) were mixed and used for the hybridization. In brief, slides were hybridized (in quadruplicate, using 250 µl of hybridization mix) at 65 °C for 17 hrs, disassembled into GE wash buffer 1 (Agilent) and washed again for 1
min with the same buffer, in prewarmed GE wash buffer 2 (Agilent) (37 °C for 1 min) and then washed for 1 min in stabilization and drying solution to protect the cyanine dyes from degradation by atmospheric ozone (Agilent). Slides were removed from the final solution and dried. A “dye-flip” [34] was performed to control for any bias in hybridization signal between the cyanine-labelled cRNAs (produced for two distinct mRNA populations). The slides were scanned using a microarray scanner B (Agilent). The data were normalised and subjected to statistical analysis as described previously [26]. The data were examined for differential hybridization (≥ 2-fold) between L3 and aL3 cDNAs from An. caninum, using a probability value (p-value) of ≤ 0.001. Evaluation of covariance revealed low (<10%) intra- and inter-array variations for this microarray analysis procedure [26]. The CSH results were verified independently using a reverse transcription-PCR-coupled sequencing method [35]. In brief, a subset (25%) of cross-hybridizing molecules from cDNA derived from mRNA from H. contortus (adult) or An. caninum (L3 or aL3) were amplified (individually) using primer pairs designed to corresponding oligonucleotides on the array. Then, the specificity and identity of individual amplicons were confirmed by direct sequencing using the same primers (separately) as employed for the primary amplification. The sequences were compared with those available in current gene databases.

Molecules displaying differences in hybridization signal between L3s and aL3s on the H. contortus microarray were identified and subjected to BLASTn (The Wellcome Sanger Institute: www.nematode.net) to identify putative homologues in both H. contortus and An. caninum and to BLASTx (NCBI: www.ncbi.nlm.nih.gov) and BLASTn (EMBL-EBI Parasite Genome Blast Server: www.ebi.ac.uk) analyses to identify putative homologues in C. elegans, other nematodes and other organisms (e-value cut-off ≤ 1e-05). WormBase (www.wormbase.org) was interrogated extensively for relevant information on C. elegans homologues/orthologues, including RNA interference (RNAi; [36]) phenotypes, transcriptomic, proteomic and interactomic data. ESTs with homologues in C. elegans and other nematodes were also analysed using the KEGG Orthology-Based Annotation System (KOBAS) (www.kobas.cbi.pku.edu.cn), which predicted the biochemical pathways in which molecules are involved. The open reading frames (ORFs) inferred from selected ESTs with orthologues in C. elegans were also subjected to “secretome analysis” using the program SignalP v.2.0 (www.cbs.dtu.dk/services/SignalP/), employing both the neural network and hidden Markov models to predict signal peptides and/or anchors [37-39]. Also, transmembrane domains were predicted using the program TMHMM (www.cbs.dtu.dk/services/TMHMM/; [40-42], and subcellular localization predicted using the program Wolf PSORT (http://wolfpsort.org/; [43]).
The approach described by Zhong and Sternberg [44] was used to predict the genetic interaction network for the genes representing the cRNAs which hybridized to spots (all three oligonucleotides) on the *H. contortus* microarray. Genomic data (regarding interactions, phenotype, and gene ontology) from *C. elegans* gene homologues/orthologues, also incorporating data from the vinegar fly (*Drosophila melanogaster*), yeast (*Saccharomyces cerevisiae*), mouse (*Mus musculus*) and human, were integrated using a naïve Bayesian network model, and the predicted networks resulting from the analyses were saved in a graphic display file (gdf) format and examined using the graph exploration system available at http://graphexploration.cond.org/. Images were labelled and saved in the joint photographic experts group (jpeg) format. Molecules were also functionally classified based on biological, cellular and molecular functions using the gene ontology (GO) approach [45].

5.3 Results

The oligonucleotide microarray for *H. contortus* was used to identify and study conserved transcripts in the L3 of *An. caninum* upon serum stimulation *in vitro* (reflecting the transition from the free-living, infective L3 to the parasitic stage [aL3] which invades the host). The analysis revealed differential hybridization between aL3 and L3 for 32 molecules (each represented by a triplet of overlapping and/or non-overlapping oligonucleotides) (1.7% of the total number on the array). Of these, 20 (62.5%) had a significantly greater signal intensity for *An. caninum* aL3 cRNA and 12 (37.5%) for L3 cRNA. Reverse transcription-PCR-coupled sequencing of a subset of 8 of the 32 (25%) cDNAs from each *H. contortus* and *An. caninum* was conducted to verify their identities. Individual sequence tags from the eight amplicons (91 to 204 bases) derived from *H. contortus* matched the EST sequences (i.e., EST codes BF060177, BF422911, CB332051, CB016255, CB331997, CA957659, CB332588 and CB015015, respectively) from which individual primers pairs had been designed (see Table 5.1). Seven of the eight sequences (85 to 200 bases) derived from amplicons from *An. caninum* had significant matches (e-values: 1.0e-12 to 9.6e-05) to the corresponding *H. contortus* ESTs, with amino acid identities of 87 to 93% (not shown); one sequence tag did not match any EST in current databases. While *in silico* predictions had revealed significant matches to sequences from *An. caninum* (see Table 5.1), the sequence tags determined had higher identities to *H. contortus* homologues/orthologues available in current databases.

Of the 20 molecules with an increased signal intensity (2.0- to 14.0-fold) upon hybridization with *An. caninum* aL3 cRNA, comparative analyses revealed 19 and 15 to
have homologues or orthologues in *C. elegans* and *An. caninum*, respectively (see Table 5.1). For one molecule (designated BM173829), no *C. elegans* or *An. caninum* homologue/orthologue was identified in any of the current public gene databases. For 11 (~58%) of the *C. elegans* homologues (Table 5.1), ‘non-wildtype’ RNAi phenotypes, including lifespan abnormal (*Age*), body morphology defect (*Bmd*), cell death abnormal (*Ced*), clear (*Clr*), cytokinesis abnormal (*Cyk*), dumpy (*Dpy*), two types of embryonic defect (*Emb*), slow growth (*Gro*), larval lethal (*Let*), larval arrest (*Lva*), early larval lethal (*Lvl*), moult defect (*Mlt*), protruding vulva (*Pvl*), sick (*Sck*), sluggish (*Slu*), sterile (*Ste*), sterile progeny (*Stp*) and/or uncoordinated (*Unc*) (see [www.wormbase.org](http://www.wormbase.org)) were recorded. KOBAS analysis predicted these molecules to be involved in or linked to adherens junctions (*C. elegans* homologue *act-2*), urea metabolism (*Y46G5A.19*), mitochondrial-associated proteins (*ZK953.1*, *F59B8.2* and *K07E3.4*), protein folding and export (*fkb-3* and *Y57G11C.15*) and other enzymes (*cpr-6*). For 10 of these homologues (52.6%), a role in a specific metabolic pathway could not be predicted (see Table 5.1). A GO analysis of the proteins inferred from genes *C08H9.2*, *calu-1*, *elo-2*, *Y55F3AM.1*, *rpn-2* and *sel-9* revealed the biological process terms of ‘growth, embryonic development ending in birth or egg hatching’, ‘larval development’, ‘positive regulation of body size and reproduction’ (cf. [www.wormbase.org](http://www.wormbase.org)).

Of the 12 ESTs with statistically higher signal intensities (2.0- to 4.0-fold) using *An. caninum* L3 compared with aL3 cRNA, 10 had *C. elegans* homologues (Table 5.1). For two of these homologues (namely *abce-1* and *C55F2.1*; accession codes *CA033958* and *CA033560*, respectively), ‘non-wildtype’ RNAi phenotypes, including *Dpy*, *Emb*, *Gro*, *Lva*, *Ste*, *Stp* and *Unc*, were recorded in *C. elegans* (see Table 5.1 and [www.wormbase.org](http://www.wormbase.org)). Based on KOBAS analysis, none of these molecules was predicted to be involved in a specific metabolic pathway (Table 5.1). Four molecules were inferred to be involved in growth and larval development (*abce-1*), fatty acid biosynthesis (*dhs-17*), phosphate transport (*col-181*) and reproduction (*C55F2.1*) (cf., [www.wormbase.org](http://www.wormbase.org)), while the other six could not be assigned GO terms.

Sixteen *C. elegans* orthologues (i.e., genes *abce-1*, *act-2*, *C08H9.2*, *C55F2.1*, *calu-1*, *col-181*, *cpr-6*, *elo-2*, *asp-1*, *K07E3.4*, *rpn-2*, *sel-9*, *T28C12A*, *hsb-1*, *Y57G11C.15* and *ZK953.1*) were predicted to interact directly with a total of 156 (range: 1-88, see Appendix 5.1) other genes; in particular, a direct genetic interaction was predicted among *T28C12A* and *hsb-1* and *calu-1* and *col-181* (see Fig. 5.1). To predict the classification and gene function of the 156 interactors, a detailed GO analysis was conducted. The summary of the GO categories for ‘biological process’, ‘cellular component’ and ‘molecular function’ of these molecules is displayed in Fig. 5.2. The 156 molecules could be assigned to a total number of 129 different terms for ‘biological
process’, 101 for ‘cellular component’ and 215 for ‘molecular function’. The most common subcategories were: ‘signal transduction’ (7%), ‘intracellular protein transport’ and ‘glycolysis’ (6.2%) within ‘biological process’; ‘nucleus’ (25.7%) and ‘intracellular’ (19.8%) within ‘cellular component’; ‘ATP binding’ (14.4%) and ‘protein binding’ (8.4%) within ‘molecular function’ (cf. Appendix 5.2).

5.4 Discussion

The present chapter focused on investigating transcripts conserved between H. contortus (different developmental stages) and An. caninum (free-living and serum activated larvae). A total number of 32 (1.7%) molecules of An. caninum were found to cross-hybridize with H. contortus oligonucleotides on the microarray, with a significantly higher signal intensity for 20 (62.5%) in aL3 and 12 (38.5%) in L3. Twenty-nine of these molecules (i.e., 19 aL3 and 10 for L3; cf. Table 5.1) were shown to have C. elegans orthologues, whereas three had no known C. elegans or An. caninum orthologues in current gene databases (not shown). ‘Non-wildtype’ RNAi phenotypes were recorded for 13 of the C. elegans homologues (i.e., 11 and 2 for An. caninum aL3 and L3 cRNA, respectively; cf. Table 5.1), including two types of embryonic defect (Emb), larval lethal (Let), larval arrest (Lar) and early larval lethal (Lvl) (see www.wormbase.org and Table 5.1). GO annotation for ‘biological process’ of the C. elegans orthologues of the conserved transcripts between H. contortus and An. caninum revealed terms ‘growth’, ‘embryonic development ending in birth or egg hatching’, ‘larval development’, ‘positive regulation of body size’ and ‘reproduction’ (see, among others, CO8H9.2, calu-1, elo-2, Y55F3AM.1, rpn-2, sel-9 and C55F2.1 at www.wormbase.org), suggesting that some of the molecules involved in the moult and process are conserved between free-living and the blood-feeding parasitic nematodes considered herein [20,21,46].

Sixteen C. elegans orthologues (namely abce-1, act-2, CO8H9.2, C55F2.1, calu-1, col-181, cpr-6, elo-2, asp-1, K07E3.4, rpn-2, sel-9, T28C12.4, hsb-1, Y57G11C.15 and ZK593.1), encoding RNAse inhibitors, actins, virgins, synthases, collagen, proteases, elongases, transport proteins and kinases, were predicted to interact with or among a total of 156 other genes (see Appendix 5.1). In a recent microarray-based study [47], conducted to compare the transcription profile of Ascaris suum (which belongs to the order Ascaridida) infective L3s versus all other developmental stages, nine L3-enriched As. suum genes were found to interact with a total number of 296 C. elegans orthologues. Forty-two (27%) of the present interactors (n = 156) were common to the interaction network (n = 296) for the As. suum molecules ([48]; Appendix 5.1). These common interactors were all involved in embryonic development, egg hatching, larval
development and/or growth, indicating that some conserved pathways are common to
distantly-related nematode species (in this case, of the orders Strongylida and
Ascaridida).

Of the *An. caninum* genes which cross-hybridized with *H. contortus* probes on the
microarray, the RNAse inhibitor *abce-1* [ATP-binding cassette (ABC) superfamily] was
found to be highly transcribed in *An. caninum* L3s and predicted to interact with 20 other
genes all involved in fundamental processes, including translational initiation,
embryonic and larval development, growth, gamete generation and reproduction (see
Appendices 5.1 and 5.2 and www.wormbase.org). The *abce* homologue identified in the
present study had 85% similarity to the *H. contortus* CA033958 EST, whereas no *An.
caninum* homologue was found in public databases (see Table 5.1). ABC class E (ABCE)
proteins constitute one of the largest protein families in both prokaryotes and
eukaryotes, being conserved (65-95%) across a range of different phyla of eukaryotes
(including free-living nematodes, flies, yeasts and mammals) [48-50]. The *abce* genes of
eukaryotes have been traditionally assigned to RNase L inhibitors, playing crucial roles
in regulating RNA turnover and stability [51,52]. A recent analysis of *C. elegans*,
conducted in order to study the functions or roles of ABCE in the context of a developing
multicellular organism, showed that this group of molecules is involved in a range of
events, including translation and transcriptional control, regulation of growth and vulval
development [50]. The present results, together with RNAi data for *C. elegans* (i.e., Emb;
www.wormbase.org), suggest that *abce* genes are conserved also across different taxa of
parasitic nematodes, and their products play essential functions in well-conserved
pathways, other than RNase L inhibitors and RNA regulators which remain to be
characterised.

The carboxylesterase gene T28C12.4 had 65% similarity (at the nucleotide level)
to molecule CB015015 of *H. contortus* and was abundantly transcribed in *An. caninum*
L3s, although no homologues for the latter species were available in current databases
(cf. Table 5.1). Carboxylesterases are believed to play an important regulatory role in key
processes, such as lipid metabolism, vitellogenesis and reproduction, in both
prokaryotes and eukaryotes [53]. In *C. elegans*, they have been found mainly to be
enriched in the intestine for the metabolism of carbohydrates [53]. While no
information is available for this group of enzymes in other nematodes, their genetic
similarity (65%) suggests functional conservation between *C. elegans* and the two blood-
feeding nematodes considered herein. The genetic interaction network for the
carboxylesterase gene T28C12.4 revealed a direct link to the heat-shock factor binding
protein gene (*hsb-1*; Fig. 5.1), which was shown to be significantly up-regulated in the
aL3 stage of *An. caninum* (compared with L3). In *Homo sapiens*, *hsb-1* is thought to play a
role as a negative regulator of the heat shock response, suppressing the activation of the stress response during the ageing process [54]. Currently, no information is available on the specific function of this gene in *C. elegans* and its role in ‘regulating’ or ‘influencing’ the activity of T28C12.4 carboxylesterase remains to be investigated in detail.

Of the 32 molecules inferred to be conserved between *An. caninum* and *H. contortus*, the homologue/orthologue of the collagen family member *col-181* was shown to be highly transcribed in *An. caninum* L3s. Due to significant commonality in the biochemical and basic structural patterns in the surface components between *C. elegans* and some parasitic nematodes [55-57] and their immunogenic properties, collagens have been investigated as potential vaccine candidates [58]. Previously, collagen genes have been demonstrated to be structurally conserved between *C. elegans* and the parasitic nematodes *H. contortus*, *Teladorsagia circumcincta*, *As. suum* and *Brugia pahangi*, stimulating the hypothesis that they are conserved throughout the phylum Nematoda [58]. The evidence of transcriptional conservation for collagen homologues among *C. elegans*, *H. contortus* and *An. caninum* (70%; see Table 5.1) provides some additional support for this proposal.

Also, actins are known to regulate embryonic morphogenesis in *C. elegans* by providing forces that drive cellular rearrangements and formation of membrane protrusion through regulated polymerization or actomyosin contraction [59]. Based on this knowledge, the finding of the actin gene act-2 amongst the cross-hybridizing gene homologues highly transcribed in *An. caninum* aL3s supports its fundamental role in the chain of events which characterize the developing larva. At least five actin genes which localize to specific cell types have been characterized thus far in *C. elegans* (see www.wormbase.org). However, all isoforms show a low level of amino acid sequence variation, suggesting relative conservation for this class of molecules (reviewed by [60]). Bioinformatic analysis of act-2 confirmed that this gene is highly conserved (i.e., 97%) across free-living and the blood-feeding nematodes considered herein and has homologues in a range of different orders of nematode parasites, including ascaridoids and spiruroids (see Table 5.1). This finding, along with the information available on RNAi of this gene in *C. elegans* (i.e., Unc; [61]), provides a basis for further investigations into the role of this molecule in regulating larval development in parasitic nematodes.

The long chain fatty acid elongase elo-2 belongs to a group of conserved molecules inferred to be highly upregulated in aL3s of *An. caninum* (compared with L3) (see Table 5.1). In *C. elegans*, this molecule has been reported to cooperate in the production of 20-carbon polynsaturated fatty acid, being involved in the regulation of normal growth and development [62]. The interaction network inferred for the *H. contortus* and *An. caninum elo-2* orthologues in *C. elegans* included eight other genes
(tag-331, elo-3, elo-1, elo-4, elo-6, elo-7, unc-64, alr-1), all of which are required for rapid
growth, large body size, fertility, and for the “quantitative regulation” of the ultradian
defaecation rhythm (see Fig. 5.1, Appendix 5.1 and www.wormbase.org).

Also the *C. elegans* homologues aspartic protease *asp-I* and cysteine protease *cpr-
6* were up-regulated in aL3s of *An. caninum* (compared with L3). Proteases have been
shown to be substantially upregulated in L3s of *An. caninum* following serum-activation
[21] and enriched in the infective L3 of the ascaridoid nematode *Anisakis simplex* [63].
Aspartic and cysteine proteases are relatively conserved proteolytic enzymes that have
been characterized for a variety of species (e.g., [64,65]). Such proteases function in the
intra-cellular and extra-cellular degradation of proteins, which includes the digestion of
nutrients, processing of peptide hormones, antigens, and immunoglobulins. In parasitic
helminths, aspartic and cysteine proteases play key roles during the migration and
feeding in the host animal and, together with serine and metalloproteinase, are mainly
associated with host invasion of and/or migration through tissues [66]. *H. contortus* and
*An. caninum* share similar proteolytic pathways for the digestion of haemoglobin [67],
supporting the positive cross-hybridization results for molecules involved in these
pathways. Due to their immunogenic properties, such proteases are currently under
investigation as vaccine candidates for both of these blood-feeding nematodes [11,12].

One of the transcripts conserved between *H. contortus* and *An. caninum* and ‘up-
regulated’ in *An. caninum* L3s had a *C. elegans* orthologue (T05A10.4) encoding a sperm
coating protein (SCP) (see Table 5.1). The SCP domain in this protein typifies members of
the PRP-superfamily, which have been proposed to play a fundamental role(s) in the
transition of hookworms from the free-living to the parasitic stage in the host [20,21]. A
recent microarray study [21] demonstrated a substantial expansion of PRPs during this
transition. Therefore, it is possible that the orthologue (T05A10.4) in *An. caninum* might
be involved in the developmental switch to parasitism, but this hypothesis requires
testing.

Biotechnological advances provide invaluable tools for exploring molecular
mechanisms regulating parasite biology and development. Despite the progress in
genome sequencing of selected parasitic nematodes [68,69], knowledge of the molecular
biology of such nematodes is limited and presently relies heavily on information
available for *C. elegans* (see www.wormbase.org; [27]). In this context, the CSH approach
employed in the present study can assist genomic research in cases where the
infrastructure (such as cDNA libraries and genomic sequence data) are not available. For
instance, the *H. contortus* microarray previously designed [26] and employed herein
might constitute a useful tool for investigating transcriptomic profiles (in different
stages, both sexes and in various tissue types) linked to parasite-specific genes with no
homologues/orthologues in the free-living nematode *C. elegans*, such as some of those identified in the present study. In addition, the investigation of molecules common across a range of parasitic nematodes employing next-generation sequencing technologies should assist in elucidating fundamental molecular aspects of parasite biology and development as well as parasite-host interactions, with the potential of leading to new intervention strategies against parasitic nematodes.
5.5 References


155


Table 5.1 *Haemonchus contortus* probes (listed according to KOBAS classification) linked to differential hybridization signal in the microarray between cDNAs from activated (AL3s) and non-activated third stage larvae (L3s) of *Ancylostoma caninum*. *Ancylostoma caninum* and *Caenorhabditis elegans* homologues were predicted in silico from other nematodes and other organisms based on comparative analyses (BLASTx or BLASTn) with sequences in current databases. *C. elegans* homologues with known RNAi phenotypes have been also included.

<table>
<thead>
<tr>
<th>EST code representing a probe</th>
<th>Size (bp)</th>
<th>A. caninum homologue predicted</th>
<th>Differential hybridization signal</th>
<th>In silico peptide analysis of H. contortus ESTs</th>
<th>Description of C. elegans homologue (Gene code, gene name/ E-value/ percentage of similarity)</th>
<th>RNAI phenotype in C. elegans&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Homologues in other strongyrid nematodes [non-strongyrid nematodes]&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Homologues in mammals&lt;sup&gt;d&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contig249&lt;sup&gt;a&lt;/sup&gt;</td>
<td>640</td>
<td>C2217486</td>
<td>2.42</td>
<td>S/0/Ex</td>
<td>Actin (act-2/ 1e-116/ 97%)</td>
<td>Ste, Pvl, Lva, Sck, Lvd, thin, Cyk, transgene subcellular localization abnormal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arginine and proline metabolism; beta-alanine metabolism; urea cycle and metabolism of amino groups</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Spermidine synthetase (Y4656A.19/ 5e-87/ 81%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BF060177&lt;sup&gt;e&lt;/sup&gt;</td>
<td>647</td>
<td>AW626946</td>
<td>2.48</td>
<td>S/0/Pl</td>
<td></td>
<td>[Asu, Bma, Cbr Hgl, Mch, Min, Mha]</td>
<td></td>
<td>Hsa, Mmu, Mmus, Rno</td>
</tr>
<tr>
<td>Carbon fixation; glycolysis/ gluconeogenesis; insulin signaling pathway; purine metabolism; pyruvate metabolism; type II diabetes mellitus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pyruvate kinase (Z95393.1/ 2e-5B/ 64%)</td>
<td></td>
<td>Ace, Sra, Tvi [Bma, Cbr, Hgl, Mha, Sja, Xiu]</td>
<td>Bta, Gga, Hsa, Mfa, Mmul, Mmus</td>
</tr>
<tr>
<td>Citrate cycle (TCA cycle); glutathione metabolism; reductive carboxylic cycle (CO2 fixation)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NADP-dependent isocitrate dehydrogenase (F59B8.2/ 4e-107/ 84%)</td>
<td>Age</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CB332051&lt;sup&gt;e&lt;/sup&gt;</td>
<td>665</td>
<td>C2276976, C2275802 (contig)</td>
<td>2.74</td>
<td>S/0/Pl</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glyoxylate and dicarboxylic metabolism; one carbon pool by folate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Formyltetrahydrofolate synthetase (K07E3.4/ 3e-53/ 88%)</td>
<td></td>
<td>[Cbr, Hgl, Gro, Min, Mar]</td>
<td>Bta, Hsa, Mmus, Rno</td>
</tr>
<tr>
<td>Protein folding and associated processing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>FKBP-type peptidyl-prolyl cis-trans isomerase (Jkb-3/ 2e-75/ 71%)</td>
<td>Age</td>
<td>Oos, Ptr, Sra, Sst, Tci [Cbr, Gro, Hgl, Hsc, Mha, Min, Ovo Wba]</td>
<td>Bta, Hsa, Mmus, Rno</td>
</tr>
<tr>
<td>Protein export</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Transport protein Sec61, Sck, Ste</td>
<td>Sck, Ste</td>
<td>Ace, Oos, Sra, Sst, Tci</td>
<td>Bta, Cfa, Eca, Gga,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>Homologues/orthologues with increased transcription in *A. caninum* AL3s

<sup>b</sup>Adherens junctions; focal adhesion; leukocyte transendothelial migration; regulation of actin cytoskeleton; tight junction

<sup>c</sup>Homo sapiens, Mus musculus, Rattus norvegicus

<sup>d</sup>Homologues in mammals: Homo sapiens, Mus musculus, Rattus norvegicus
<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Accession</th>
<th>Value</th>
<th>Category</th>
<th>Description</th>
<th>Reference Genomes</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>CB174876</td>
<td>CW704361</td>
<td></td>
<td>contig</td>
<td>alpha subunit (Y57G11C.15/ E6- 61/ 98%)</td>
<td>[Chr, Gro, Hgl, Mar, Ppa]</td>
<td>Hsa, Mmup</td>
</tr>
<tr>
<td>BQ666459,</td>
<td>BM130427,</td>
<td>2.12</td>
<td>S/0/Mi</td>
<td>Cysteine protease (cpn-6/ 8e- 89/ 76%)</td>
<td>Nam, Nbr, Oos, Tci, Ptr</td>
<td>Hsa, Ppy</td>
</tr>
<tr>
<td>BM1744392</td>
<td></td>
<td></td>
<td>contig</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BQ666459</td>
<td>311</td>
<td>2.68</td>
<td>S/0/Ex</td>
<td>Vigilin (C08H9.2/ 2e- 25/ 63%)</td>
<td>Gro, Bmd, Dpy, Shu, Unc, Pvl</td>
<td>Hsa, Mmus, Rno</td>
</tr>
<tr>
<td>BM250320</td>
<td>499</td>
<td>3.80</td>
<td>S/0/Mi</td>
<td>Cytochrome oxidase c subunit one (3e- 63/ 85%)</td>
<td>Ace, Tci [Bma, Cbr, Ovo, Ppa]</td>
<td></td>
</tr>
<tr>
<td>BM077436</td>
<td>437</td>
<td>3.66</td>
<td>S/0/Ex</td>
<td>Heat shock factor binding protein (hsu-1/ 6e- 13/ 91%)</td>
<td>Ace, Tci [Asu, Bma, Cbr, Ppa, Xin]</td>
<td>Eca, Gga, Hsa, Mmus, Ssc, Rno</td>
</tr>
<tr>
<td>BQ666710,</td>
<td>698</td>
<td>2.14</td>
<td>S/0/Ex</td>
<td>Reticulocalbin, calumenin, DNA supercoiling factor, and related Ca2+-binding</td>
<td>Lvl, Lva, Mlt, Unc, Clr</td>
<td>Oos, Sra [Asu, Bma, Cbr, Mch, Mha, Tca]</td>
</tr>
<tr>
<td>BM077653</td>
<td></td>
<td>2.79</td>
<td>S/0/Ex</td>
<td>Long chain fatty acid elongase (elo-2/ 1e- 38/ 63%)</td>
<td>Gro, fat content reduced, lipid composition abnormal, posterior body wall shortened interval, reproductive system physiology abnormal, Sck, reduced brood size</td>
<td>Tci, Sra [Asu, Bma, Cbr, Mar, Min]</td>
</tr>
<tr>
<td>BM077853</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BQ666459</td>
<td>338</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BM077653</td>
<td>524</td>
<td>2.42</td>
<td>S/0/Ex</td>
<td>Glutamine synthetase (ghs-3/ 6e- 85/ 82%)</td>
<td>Sst, Sra [Chr, Asu, Xin, Pbr, Gro, Tca]</td>
<td>Bta, Cja, Hsa, Mmus, Ssc, Rno</td>
</tr>
<tr>
<td>CA957659</td>
<td>494</td>
<td>2.94</td>
<td>S/0/Ex</td>
<td>Putative mitochondrial ribosomal protein mrps25</td>
<td>Emb1</td>
<td>Nbr [Bma, Cbr, Hgl, Gpa, Gro, Pbr, Tma, Tsp]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accession</td>
<td>Gene ID</td>
<td>Freq.</td>
<td>Coverage</td>
<td>Source</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>---------</td>
<td>-------</td>
<td>----------</td>
<td>--------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>BF060222</td>
<td>460</td>
<td>2.97</td>
<td>S/0/Ex</td>
<td>(Y55F3AM.1/9e-61/74%) 26S proteasome regulatory complex, subunit RPNI/PSMD1 (rpn-2/3e-40/57%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CB332588+</td>
<td>661</td>
<td>2.20</td>
<td>S/0/Ex</td>
<td>Putative cargo transport protein EMP24 (p24 protein family) (sel-9/4e-87/86%) aspartyl protease (asp-1/1e-23/61%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AW670717</td>
<td>280</td>
<td>2.01</td>
<td>S/0/Ex</td>
<td>F-box domain (F42A10.7/2e-40/55%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AW670783</td>
<td>451</td>
<td>2.53</td>
<td>S/0/Pl</td>
<td>F-box domain (F42A10.7/2e-40/55%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Gene homologues/orthologues with increased transcription in A. caninum L3s**

<table>
<thead>
<tr>
<th>Accession</th>
<th>Gene ID</th>
<th>Freq.</th>
<th>Coverage</th>
<th>Source</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA033958</td>
<td>407</td>
<td>2.96</td>
<td>S/0/Ex</td>
<td>RNAse L inhibitor, ABC superfamily (abce-1/7e-07/85%)</td>
<td></td>
</tr>
<tr>
<td>CB015678</td>
<td>631</td>
<td>2.72</td>
<td>S/0/Ex</td>
<td>Collagen family member (col-181/3e-12/70%)</td>
<td></td>
</tr>
<tr>
<td>Contig713</td>
<td>515</td>
<td>2.68</td>
<td>S/0/Pl</td>
<td>Unclassified protein T05A10.4 (2e-13/73%)</td>
<td></td>
</tr>
<tr>
<td>CA034292</td>
<td>538</td>
<td>2.22</td>
<td>Q/0/Ex</td>
<td>Predicted membrane protein, contains DoH and Cytochrome b-561/ferric reductase transmembrane domains (C05D12.1/7e-68/62%)</td>
<td></td>
</tr>
<tr>
<td>CB015015+</td>
<td>538</td>
<td>2.19</td>
<td>S/0/Ex</td>
<td>Carboxylesterase and related proteins (T28C12.4/1e-65/65%)</td>
<td></td>
</tr>
<tr>
<td>CB021109</td>
<td>634</td>
<td>2.13</td>
<td>S/0/Ex</td>
<td>Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases (dhs-17/3e-58/57%)</td>
<td></td>
</tr>
<tr>
<td>BF060184</td>
<td>500</td>
<td>2.13</td>
<td>S/0/Ex</td>
<td>Ion channel protein (K11H3.7/3e-61/81%)</td>
<td></td>
</tr>
<tr>
<td>CB012436</td>
<td>680</td>
<td>2.08</td>
<td>S/0/Ex</td>
<td>Dehydrogenases with</td>
<td></td>
</tr>
</tbody>
</table>

**References**

- Saj [Asu, Bma, Chr, Hgl, Hsh, Mha, Wba]
- Eca, Hsa, Mmus, Rno
- Hsa, Gga, Mmus, Rno
- Bta, Eco, Gga, Hsa, Mmus, Oar
- Bta, Eco, Hsa, Mma, Mmus
| CW712370 |  
|---------|---------|---------|---------|---------|
| CA033560 | 310     | 2.01    | S/0/Ex  | different specificities (related to short-chain alcohol dehydrogenases (dhs-17/2e-79/70%) AICAR transformylase/IMP cyclohydrolase/methylglyoxal synthase (C55F2.1/3e-14/39%) | Ste | Oos | Mmus |
| CA994809 | 561     | BM077786 | 2.32    | Uncharacterized protein with conserved cysteine (trr-16, Y5F2A.1/1e-44/63%) | Ace, Nam, Oos, Ptr, Tci [Bma, Xin] |

*a Abbreviations used in proteomic analysis: Non-secretory protein (Q), secretory protein (S)/ predicted number of transmembrane domains/ predominant cellular location: extracellular (Ex); nuclear (Nu), plasma membrane (Pl).

*b Abbreviations of RNAi phenotypes (alphabetical): lifespan abnormal (Age), body morphology defect (Bmd), cell death abnormal (Ced), clear (Clr), cytokinesis abnormal (Cyk), dumpy (Dpy), embryonic lethal (Emb*), transgene expression increased, passage through meiosis abnormal early emb (Emb*), slow growth (Gro), larval lethal (Let), larval arrest (Lva), early larva lethal (LxL), mouthing defect (Mlt), protruding vulva (Pvl), sick (Sck), sluggish (Sli), sterile (Ste), sterile progeny (Stp), uncoordinated (Unc).

*c Abbreviations of nematode species (alphabetical): Ancylostoma ceylanicum (Ace), Ascaris lumbricoides (Alu), Brugia malayi (Bma), Brugia pahangi (Bpa), Caenorhabditis briggsae (Cbr), Globodera pallida (Gpa), Globodera rostochiensis (Gro), Heterodera glycines (Hgl), Heterodera schachtii (Hsc), Litomosoides sigmodontis (Lsi), Meloidogyne arenaria (Mar), Meloidogyne chitwoodi (Mch), Meloidogyne hapla (Mha), Meloidogyne javanica (Mja), Meloidogyne incognita (Min), Meloidogyne paranaensis (Mpa), Necator americanus (Nam), Nippostrongylus brasiliensis (Nbr), Ostertagia ostertagi (Oos), Onchocerca volvulus (Ovo), Parastrongyloides trichosuri (Ptr), Pristionchus pacificus (Ppa), Pratylenchus penetrans (Ppe), Pratylenchus vulnus (Pvu), Schistosoma japonicum (Sja), Schistosoma mansoni (Sma), Strongyloides ratti (Srr), Strongyloides stercoralis (Sst), Toxocara canis (Tca), Teladorsagia circumcincta (Tci), Trichuris muris (Tmu), Trichinella spiralis (Tsp), Trichostongyulus virens (Tv), Wuchereria bancrofti (Wbr), Xiphinema index (Xin).

*d Abbreviations of mammalian species (alphabetical): Bos taurus (Bta), Canis familiaris (Cfa), Equus caballus (Eca), Gallus gallus (Gga), Homo sapiens (Hsa), Macaca fascicularis (Mja), Macaca mulatta (Mmul), Mus musculus (Mmus), Ovis aries (Oar), Pongo pygmeus (Ppy), Rattus norvegicus (Rno), Sus scrofa (Ssc).

*e Molecules subjected to reverse transcription-PCR-coupled sequencing.
Fig. 5.1 Genetic interactions predicted for a subset of sixteen *C. elegans* genes (*abce-1, act-2, C08H9.2, C55F2.1, calu-1, col-181, cpr-6, elo-2, asp-1, K07E3.4, rpn-2, sel-9, T28C12.4, hsb-1, Y57G11C.15 and ZK593.1*, in red) with homologues/orthologues in *Ancylostoma caninum* and *Haemonchus contortus* identified based on cross-species hybridization (CSH) in microarray. Interactors are indicated in black.
Fig. 5.2 Summary of predicted functions and locations for gene products inferred from 156 genes interacting with a subset of sixteen *C. elegans* homologues (*abe-1*, *act-2*, *C08H9.2*, *C55F2.1*, *calu-1*, *col-181*, *cpr-6*, *elo-2*, *asp-1*, *K07E3.4*, *rpn-2*, *sel-9*, *T28C12.4*, *hsb-1*, *y57G11C.15* and *ZK593.1*; Appendix 5.1), classified according to the gene ontology (GO) categories ‘cellular component’ ‘biological process’ and ‘molecular function’ as well as subcategories within.
CHAPTER 6
Massively parallel sequencing and analysis of the transcriptome of the human hookworm *Necator americanus*

Abstract

The blood-feeding hookworm *Necator americanus* infects hundreds of millions of people worldwide. In order to elucidate fundamental molecular biological aspects of this hookworm, the transcriptome of the adult stage of *Necator americanus* was explored using next-generation sequencing and bioinformatic analyses. A total of 19,997 contigs were assembled from the sequence data; 6,771 of these contigs had known orthologues in the free-living nematode *Caenorhabditis elegans*, and most of them encoded proteins with WD40 repeats (10.6%), proteinase inhibitors (7.8%) or calcium-binding EF-hand proteins (6.7%). Bioinformatic analyses inferred that the *C. elegans* homologues are involved mainly in biological pathways linked to ribosome biogenesis (70%), oxidative phosphorylation (63%) and/or proteases (60%); most of these molecules were predicted to be involved in more than one biological pathway. Comparative analyses of the transcriptomes of *N. americanus* and the canine hookworm *Ancylostoma caninum* revealed qualitative and quantitative differences. For instance, proteinase inhibitors were inferred to be highly represented in the former species, whereas SCP/Tpx-1/Ag5(PR-1/Sc7 proteins (= SCP/TAPS or Ancylostoma-secreted proteins) were predominant in the latter. In *N. americanus*, essential molecules were predicted using a combination of orthology mapping and functional data available for *C. elegans*. Further analyses allowed the prioritization of 18 predicted drug targets which did not have homologues in the human host. These candidate targets were inferred to be linked to mitochondrial (e.g., processing proteins) or amino acid metabolism (e.g., asparagine t-RNA synthetase). This chapter provides detailed insights into the transcriptome of the adult stage of *N. americanus* and examines similarities and differences between this species and *A. caninum*. Future efforts should focus on comparative transcriptomic and proteomic investigations of the other predominant human hookworm, *Ancylostoma duodenale*, for both fundamental and applied purposes, including the prevalidation of anti-hookworm drug targets.

---

1 The nucleotide sequence data produced for this chapter are available in the GenBank Sequence Read Archive database under accession SRA012052. The contigs assembled from these data are available at www.nematode.net.
6.1 Introduction

As described in Chapter 1, soil-transmitted helminths (= geohelminths) are responsible for neglected tropical diseases (NTDs) mostly in developing countries [1]. In particular, the blood-feeding hookworms *Necator americanus* and *Ancylostoma duodenale* (Nematoda) infect ~740 million people in rural areas of the tropics and subtropics [2], causing an estimated disease burden of 22 million disability-adjusted life years (DALYs) [3]. Geographically, *N. americanus* is the most widely distributed hookworm of humans globally [4]. The life cycle is direct, with thin-shelled eggs passed in the faeces from the infected host. Under suitable environmental conditions (e.g., 26 °C and 100% humidity; [5]), the eggs hatch and develop through two free-living larval stages to the infective, third-stage (L3; filariform) larvae. The latter larvae penetrate human skin and migrate via the circulatory system and lung to finally reside as adults usually in the duodenum. The adult stages attach by their buccal capsule to the intestinal mucosa, rupture capillaries and feed on blood. The pathogenesis of hookworm disease is mainly a consequence of the blood loss, which occurs during attachment and feeding. The disease (= necatoriasis) is commonly characterized by iron-deficiency anaemia, which can cause physical and mental retardation and sometimes deaths in children, adverse maternal-foetal outcomes [6,7] and, in chronically infected individuals, can result in a significant alteration of their immune response to helminths [8].

Traditionally, the control of hookworm disease has relied mostly on the treatment of infected individuals with anthelmintics, such as albendazole, mebendazole, pyrantel pamoate and/or levamisole. With mass treatment strategies now in place in a number of countries [9,10], there is an increased potential for hookworms to develop genetic resistance against the compounds administered, if they are used excessively and at suboptimal dosages. Thus, given the experience with drug resistance in parasitic nematodes of livestock [11], it is prudent to maintain a continual focus on the discovery of novel drugs against hookworms of humans. Such a discovery effort could be underpinned by an integrated genomic-bioinformatic 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 characterised metazoan organism [12,13], is considered to be relatively closely related to nematodes of the order Strongylida (to which hookworms belong) [14]. Current evidence indicates that ~60% of genes in strongylids (or hookworms) have orthologues/homologues in *C. elegans* [15,16], and that a range of biological pathways...
is conserved between strongyloid nematodes/hookworms and this free-living nematode [17,20]. Therefore, conducting comparative explorations of molecular data sets between these nematodes should identify nematode-specific biological pathways, which, if essential for the development and survival, could provide new targets for nematocidal drugs.

Next generation sequencing technologies, such as ABI-SOLiD, Illumina/Solexa (www.illumina.com; [21]), Helicos (www.helicosbio.com; [22]) and 454/Roche (www.454.com; [23]), together with the recent progress in bioinformatics, are providing unique opportunities for the high-throughput transcriptomic and genomic explorations of nematodes in far more detail than previously possible [24] and at a substantially lower cost than using conventional (Sanger) sequencing. To date, genomic and molecular studies of hookworms have mainly involved the canine hookworm, Ancylostoma caninum [19,25–27], because of its use as a model for human hookworms [27,28]. In contrast, genomic datasets for N. americanus are scant, representing a major constraint to progress in molecular research of this nematode [4]. The present chapter (i) describes a detailed exploration and functional annotation of the transcriptome of the adult stage of N. americanus by 454 sequencing coupled to semi-automated bioinformatic analyses, (ii) compares the transcriptome of N. americanus to currently available transcriptomic data for An. caninum, and (iii) infers the essentiality of key genes and gene products in order to predict putative drug targets.

6.2 Materials and methods

6.2.1 Parasite material

The “Shanghai strain” of N. americanus (kindly provided by Drs Bin Zhan and Peter Hotez) was produced in golden hamsters (Mesocricetus auratus; infected for 94 days) at the Universidade Federal de Minas Gerais, Brazil. The infection experiment was conducted according the animal ethics guidelines of the Universidade Federal de Minas Gerais.

6.2.2 RNA isolation, cDNA synthesis and 454 sequencing

Total RNA from 30 adult worms was prepared using TRIzol Reagent (GibcoBRL, Life Technologies, USA) following the manufacturer’s instructions and then treated with Ambion Turbo DNase (Ambion/Applied Biosystems, Austin, TX). The integrity of the RNA was verified using the Bioanalyzer 2100 (Agilent Technologies, USA), 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 using
the SMART™ kit (Clontech/Takara Bio, CA) from ~100 ng of total RNA. An optimized PCR cycling protocol (over 20 cycles) was used to amplify full-length cDNAs, employing primers complementary to the SMART IIA-Probe and custom oligo(dT), and the Advantage-HF 2 polymerase mix (Clontech/Takara). The cDNA was normalised by denaturation-reassociation, treated with duplex-specific nuclease (Trimmer kit, Evrogen, CA) and amplified over 11 cycles. Subsequently, the 5’- and 3’- adaptors were removed by digestion with the exonuclease Mme1 and streptavidin-coated paramagnetic beads [29]. The normalised cDNA (500–700 bases) was then amplified using 9 cycles of Long and Accurate (LA)-PCR [30] and then sequenced in a Genome Sequencer™ (GS) Titanium FLX instrument (Roche Diagnostics) employing a standard protocol [23].

6.2.3 Bioinformatic analyses

Expressed sequence tags (ESTs) generated from the normalised cDNA library for *N. americanus* were assembled and annotated using a standard bioinformatic pipeline [31]. Briefly, sequences were aligned and assembled using the Contig Assembly Program v.3 (CAP3; [32], employing a minimum sequence overlap length of 50 nucleotides and an identity threshold of 95%. ESTs (n = 2,200; www.ncbi.nlm.nih.gov) from adult *N. americanus* available from previous studies [4,16,33,34] were included for comparative analysis. Following the pre-processing of the ESTs, contigs and singletons from the present dataset were subjected to analysis by BLASTx (NCBI, www.ncbi.nlm.nih.gov) and BLASTn (EMBL-EBI Parasite Genome Blast Server, www.ebi.ac.uk) to identify putative homologues in *C. elegans*, other nematodes, and organisms other than nematodes (e-value cut-off: ≤1e-05). WormBase Release WS200 (www.wormbase.org) was interrogated extensively for relevant information on *C. elegans* orthologues/homologues, including transcriptomic, proteomic, RNAi phenotypic and interactomic data. Gene ontology (GO) annotations were conducted using BLAST2GO [35]. Peptides were mapped by InterProScan [36] and linked to respective pathways in *C. elegans* using the KEGG Orthology-Based Annotation System (KOBAS; [37]). The protein sequences inferred from open reading frames (ORFs) of the ESTs with orthologues in *C. elegans* were also subjected to “secretome analysis” using the program SignalP v.2.0 (available at www.cbs.dtu.dk/services/SignalP/), employing both the neural network and hidden Markov models to predict signal peptides and/or anchors [38–40]. Also, transmembrane domains were inferred using the program TMHMM (www.cbs.dtu.dk/services/TMHMM/; [41–43]). Protein sequences inferred from contigs for *N. americanus* were compared with those predicted for *C. elegans* and from a similar-sized, publicly available EST dataset for adult *An. caninum* produced by
454 sequencing (GenBank accession numbers EW741128-EW744730; EX534506-EX567272); protein similarities were displayed using SimiTri [44].

6.2.4 Prediction of essentiality and drug targets

All protein sequences predicted from contigs for N. americanus were compared with protein sequences available in the OrthoMCL 2.0 database (www.OrthoMCL.org) by BLASTp (e-value cut-off: <1e-05). A subset of C. elegans protein homologues was then selected based on: (i) an association with a lethal RNAi phenotype; (ii) the presence/absence of gene paralogues (based on OrthoMCL orthology grouping); and (iii) GO annotation to terms linked to enzyme or G protein-coupled receptor (GPCR) activity (i.e., GO:0003824 or GO:0004930, or a sub-term thereof). The following information was obtained: (i) network connectivity score (cf. http://www.functionalnet.org/wormnet/Wormnet_v1_index.html; see [45]); (ii) presence of mammalian orthologues (based on OrthoMCL orthology grouping (iii) essentiality information (i.e., association with non-wildtype RNAi phenotypes) in other model organisms (including Saccharomyces cerevisiae, Mus musculus and Drosophila melanogaster) based on OrthoMCL groups. Each predicted drug target was selected based on (i) the presence of orthologues linked to non-wildtype RNAi or mutant phenotypes in S. cerevisiae, M. musculus and D. melanogaster, (ii) the absence of orthologues/homologues from the human host and (iii) its network connectivity score [45].

To predict the potential of selected C. elegans orthologues of N. americanus contigs as drug targets (= “druggability”), the InterPro domains inferred from the predicted proteins were compared with those linked to known small molecular drugs which follow the ‘Lipinsky rule of 5’ regarding bioavailability [46,47]. Similarly, GO terms inferred from the predicted proteins were mapped to Enzyme Commission (EC) numbers, and a list of enzyme-targeting drugs was compiled based on data available in the BRENDA database (www.brenda-enzymes.info; [48,49]). The C. elegans orthologues included in the list were ranked according to the ‘severity’ of the non-wild-type RNAi phenotypes (i.e., adult lethal, embryonic and/or larval lethal, sterile and other defects) in C. elegans (cf. www.wormbase.org) defined in previous studies [50,51].

6.3 Results

A total of 116,839 ESTs (287±235 bases in length) were generated by 454 sequencing. After removing the ESTs of <100 bases, 63,523 ESTs were assembled into 19,997 contigs (369 bases ± 215.31). Of these, 6,771 (33.9%) had known C. elegans orthologues, and 2,287 (11.4%) matched known nucleotide sequences from various
nematodes, including *Brugia malayi*, *Haemonchus contortus*, *Pristionchus pacificus*, *N. americanus*, *An. caninum*, *An. duodenale* and *Nippostrongylus brasiliensis* (73.2%), other invertebrates (21.3%) and some vertebrates (5.5%) available in current databases. All of the previously published ESTs for *N. americanus* ([www.ncbi.nlm.nih.gov; [4,16,33,34]]) represented a subset (12.4%) of the present dataset (not shown). The number of ORFs in the *N. americanus* EST data, predicted peptides and their signal, transmembrane and/or InterPro domains as well as the results of GO and KOBAS (pathway mapping) searches are given in Table 6.1. A total of 12,799 proteins were predicted from the 19,997 contigs, of which 7,214 mapped to known proteins defined by 2,381 different domains (Table 6.1 and Appendix 6.1), the most abundant being ‘WD40’ (IPR0011680; 10.6%), ‘proteinase inhibitors’ (IPR000215; 7.8%) and ‘EF-hand’ molecules (IPR018248; 6.7%) (Table 6.2). The subsequent annotation of the inferred proteins revealed 887 different GO terms, of which 314 were ‘biological process’, 117 ‘cellular component’ and 456 ‘molecular function’ (Table 6.3 and Appendix 6.2). The predominant terms were ‘translation’ (GO:0006412; 20.3%) and ‘metabolic process’ (GO:0008152; 14.9%) for ‘biological process’; ‘intracellular’ (GO:0005622; 25.1%) and ‘ribosome’ (GO:0005840; 17%) for ‘cellular component’, and, ‘ATP binding’ (GO:0005524; 18.9%) and ‘structural constituent of ribosome’ (GO:0003735; 17.9%) for ‘molecular function’ (Table 6.3 and Appendix 6.2). Proteins inferred from the *N. americanus* contigs were predicted to be involved in 235 different biological pathways, of which the vast majority represented ‘ribosome biogenesis’ (n = 163, 70%), ‘oxidative phosphorylation’ (n = 148, 63%) and ‘proteases’ (n = 140, 60%) (see Appendix 6.3).

For comparative analyses, publicly available EST data for the adult stage of *An. caninum* was included. For this dataset, the same bioinformatic analyses described in the Methods section were conducted. From 15,755 contigs of *An. caninum*, a total of 12,622 proteins were inferred, of which 4,534 matched those encoded by *N. americanus* ORFs (Fig. 6.1); 8,650 of these predicted proteins could be mapped to known molecules with 2,546 different motifs (Table 6.1 and Appendix 6.1). The protein motifs ‘SCP-like extracellular’ (IPR014044; 9.5%), ‘ankyrin’ (IPR002110; 7%) and ‘allergen V5/Tpx-1 related’ (IPR0011283; 6%) were most commonly recorded in the *An. caninum* dataset (Table 6.2). Differences in the numbers of IPR domains identified in the *N. americanus* and *An. caninum* predicted peptides were calculated using a Chi-square test (p < 0.05) and are indicated in Table 6.2. GO annotation of the *An. caninum* predicted peptides revealed 323 different terms for ‘biological process’, 119 for ‘cellular component’ and 500 for ‘molecular function’ (Table 6.3 and Appendix 6.2). The terms ‘metabolic process’ (GO:0008152; 7.4%) and ‘proteolysis’ (GO:0006508; 6.6%) had the highest representation for ‘biological process’, as did ‘intracellular’
(GO:0005622; 7.5%) and ‘membrane’ (GO:0016020; 6.6%) for ‘cellular component’; and, ‘ATP binding’ (GO:0005524; 13.4%) and ‘catalytic activity’ (GO:0003824; 9%) for ‘molecular function’ (Table 6.3 and Appendix 6.2). Using the protein data, a total of 235 different biological pathways were predicted, of which ‘proteases’ (n = 219, 93%), ‘other enzymes’ (n = 164, 70%) and ‘protein kinases’ (n = 151, 54%) were the most predominant (see Appendix 6.3).

From the *N. americanus* dataset, 5,498 proteins matched known proteins encoded by orthologues available in the OrthoMCL 2.0 database (www.OrthoMCL.org); 372 of these proteins had homologues in *C. elegans*, and 278 (277 enzymes and one G-PCR) of them were linked to adult lethal, embryonic and/or larval lethal and sterile RNAi phenotypes (Appendix 6.4). A subset of 18 molecules in *N. americanus* with homologues in *C. elegans* but not in humans were defined, also considering RNAi phenotype/s [i.e., adult lethal (n = 2), larval and/or embryonic lethal (n = 16), sterile (n = 4) and other defects (n = 12); cf. Table 6.4], as drug target candidates. These proteins could be mapped to 54 ‘druggable’ InterPro domains, and 212 EC numbers were linked to ‘druggable’ enzymes; a total of 3,320 effective drugs were predicted (Appendix 6.4).

### 6.4 Discussion

Next-generation sequencing and integrated bioinformatic analyses have provided detailed and biologically relevant insights into the transcriptome of the adult stage of *N. americanus*. A total of 12,799 ORFs 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 27-fold [4]. Amongst the InterPro domains identified, ‘WD40’, ‘proteinase inhibitors’ and ‘EF-hand’ motifs were the most abundant, followed by ‘proteases’ and ‘protein kinases’. WD40 repeats (also known as WD or beta-transducin repeats) are short (~40 amino acid) motifs found in the proteomes of all eukaryotes and implicated in a variety of functions, ranging from signal transduction and transcription regulation to cell-cycle control and apoptosis [52,53]. WD40 motifs act as sites for protein-protein interactions; proteins containing WD40 repeats are known to serve as platforms for the assembly of protein complexes or mediators of a transient interplay with other proteins, such as the ubiquitin ligases, involved in the onset of the anaphase during cell mitosis [54]. Similarly, proteins containing ‘EF-hand’ domains are involved in a number of protein-protein interactions regulated by various specialized systems (e.g., Golgi system, voltage-dependent calcium channels and calcium transporters) for the uptake and release of calcium, which acts as a secondary messenger for their activation [55]. In *C. elegans*, both EF-hand and WD40 proteins are known to be required for the maturation of the nervous system and the
formation of ciliated sensory neurons, in particular of the chemoreceptors located in the amphids [8,56]. The amphids of parasitic nematodes are, besides having the chemoreceptive activity, also known to play a role as secretory organs, primarily to provide an appropriate substrate for the transmission of neuronal potentials [57]. However, in *N. americanus*, a group of specialized amphidial neuronal cells (= amphidial glands; [57]) expresses a group of aspartic proteases (i.e., cathepsin D-like *Na*-APR-1 and *Na*-APR-2) which are proposed to degrade host haemoglobin and serum proteins in the buccal capsule of adult worms [58]. In the dog hookworm, *An. caninum*, the amphidial glands have also been shown to produce a proteinase inhibitor (called ‘ancylostomatin’) that acts as an anticoagulant to promote the flow of host blood and tissue fluids into the buccal capsule and the intestine of the parasite [59]. Although proteinase inhibitors, such as the ‘kunitz-type’ molecules, were significantly more abundant in the transcriptome of adult *N. americanus* [4,33] than *Ancylostoma* spp., they have been better characterized in the latter parasites [60–63] for which both single and multiple kunitz-domain proteins have been described [61]. For instance, a cDNA coding a single kunitz-domain proteinase inhibitor (named AceKI-1) was isolated from *An. ceylanicum*. The corresponding recombinant protein has been shown to act as a tight-binding inhibitor of the serine proteases chymotrypsin, pancreatic elastase, neutrophil elastase and trypsin [60] and confers partial protection against hookworm-associated growth delay in hamsters [62]. Recently, a kunitz-type cDNA was shown to be enriched in the adult male of *An. braziliense* [63]. Although their precise biological function remains to be determined, kunitz-type proteinase inhibitors of hookworms appear to play pivotal roles in preventing homeostasis and inhibiting host proteases (e.g., pancreatic and intestinal enzymes; [60,64]).

Proteases were also highly represented in the transcriptome of *N. americanus* (6.1%) as well as that of *An. caninum* (4.6%) (see Table 6.2). These proteases included cysteine, aspartic and metallo-proteases, which are known to function in multi-enzyme cascades to digest haemoglobin and other serum proteins [65,66]. In *N. americanus*, cysteine proteases with high sequence homology to the protein cathepsin B were localized to the gut of adult worms and the corresponding mRNAs shown to be upregulated in the adult stage compared with the infective L3 stage, thus strongly suggesting that these enzymes are involved in blood-feeding [67]. In *An. caninum*, a cysteine protease (*Ac-CP-1*) with 86% amino acid sequence identity to those characterized in *N. americanus*, was shown to be expressed in the cephalic and excretory glands [68] and was detected in the excretory/secretory products (ES) [69] of adult worms; thus, it has been proposed that *Ac-CP-1* functions as an extracorporeal digestive enzyme at the site of attachment [67]. Another cysteine protease (*Ac-CP-2*)
was localized to the brush border membrane of the intestine and demonstrated to be involved in the digestion of haemoglobin [65]. The *N. americanus* homologue of *Ac-CP-2* (i.e., *Na-CP-2*) digests haemoglobin [66] and, expressed as a recombinant protein in *Escherichia coli* and injected subcutaneously into experimental hamsters, has been shown to induce a significant reduction in adult worm burden following challenge infection with L3s of *N. americanus* [28], suggesting that the immunogenic response directed against this protein severely impairs the digestion of host proteins by the adult worms. However, recently, a cathepsin-like cysteine protease has been isolated and characterised in the human filarial nematode *Brugia malayi* and shown by double-stranded RNAi to play an essential role in the early development and maturation of embryos of this nematode [70]. Therefore, it is possible that the abundant transcripts encoding proteases in both adult *N. americanus* and *An. caninum* also reflect a key role of these enzymes in embryogenesis. Proteases have also been isolated from larval stages of both *An. caninum* and *N. americanus* [71,72]. For instance, a metalloprotease in ES of the activated third-stage larvae (L3) of *An. caninum* has been characterised and demonstrated to be released specifically in response to stimuli that induce feeding [73]. The corresponding cDNA, isolated from an L3 expression library, encoded a zinc-metalloprotease (*Ac-MTP-1*) of the astacin family, that has been proposed to (i) regulate developmental changes associated with the transition from the free-living to the parasitic L3 and the subsequent moult to the fourth-stage larva (L4) [72]; (ii) activate host TGF-β during the infection, which, in turn, could stimulate parasite development directly, determine tissue predilection sites [74] and/or inhibit neutrophil infiltration at the site of penetration [75]; and, (iii) facilitate skin penetration or tissue migration by the invading L3 [72,76] and/or degrade the cuticular proteins of the sheath surrounding the infective, free-living L3 [77]. In *N. americanus*, serine proteases have been isolated from ES of the L3 stage and proposed to play a central role in the evasion of the host immune response [71]. Interestingly, a significant number (n = 135, 30%) of *N. americanus* proteases and protease inhibitors were not predicted to possess signal peptides indicative of secretion (cf. Table 6.1 and 6.2). The likely explanation for this result is technical and would appear to relate to a 3′-bias in sequence reads [78], thus affecting the prediction of ORFs as well as the identification of signal peptide sequences at the 5′-ends.

Other groups of molecules, such as Ancylostoma-secreted proteins (ASPs), have been proposed to have an immunomodulatory function during the invasion of the host, the migration through tissues, attachment to the intestinal wall and blood-feeding [79]. In the present study, ASPs were amongst the ten most abundant groups of molecules in the *N. americanus* dataset, and are most abundant in *An. caninum* (cf. Table 6.2). ASPs
belong to a large group of proteins, the ‘sperm-coating protein (SCP)-like extracellular proteins’, also called SCP/Tpx-1/Ag5/PR-1/Sc7 (SCP/TAPS; Pfam accession number no. PF00188), characterised by the presence of a single or double ‘SCP-like extracellular domain’ (InterPro: IPR014044). In An. caninum, double and a single SCP-domain ASPs, called Ac-ASP-1 and Ac-ASP-2, respectively, were identified as major components of ES from serum-activated, infective L3s and proposed to be secreted in response to one or more host-specific signals during the infection process [80,81], as also hypothesized in a transcriptomic analysis of serum-activated L3s [19]. In N. americanus, homologues of Ac-ASP-1 and Ac-ASP-2 (i.e., Na-ASP-1 and Na-ASP-2, respectively) have been identified in the L3 stage [82–84]. Results from crystallography [85], combined with the observation that Na-ASP-2 induces neutrophil and monocyte migration [86], suggest that this molecule has a role as an antagonistic ligand of complement receptor 3 (CR3) and alters the immune cascade by preventing the binding of chemotaxin [85]. Because of its immunogenic properties, Na-ASP-2 is under investigation as a vaccine candidate against necatoriasis [7,28,81,87]. In adult An. caninum, at least four other ASPs have been identified to date and named Ac-ASP-3, Ac-ASP-4, Ac-ASP-5 and Ac-ASP-6 [72]. Another SCP/TAPS molecule, designated neutrophil inhibitor factor (NIF), has been isolated and shown to play an immunomodulatory role by blocking the adhesion of activated neutrophils to vascular endothelial cells and the subsequent release of H₂O₂ from activated neutrophils [88] and by interfering with the function of integrin receptors located on the cell surface, which results in the inhibition of platelet aggregation and adhesion [89]. Subsequently, NIF was shown to be transcribed abundantly in the intestines of both An. caninum and N. americanus [34]. The present chapter revealed that, although highly represented in the transcriptome of adult N. americanus, ASPs were much more abundant in An. caninum (cf. Results section). One of the possible explanations for this finding is that, although the An. caninum dataset was generated from adult worms recovered from their natural host (i.e., dog), the specimens of N. americanus were recovered from a Chinese strain of the golden hamster (M. auratus), which is not a natural host for this parasite [90,91]. Indeed, adults of N. americanus recovered from hamsters with patent infections are smaller and less fecund than from the human host [91]. These phenetic differences in this parasite might be associated with variation in transcriptional profiles. However, the difference in prevalence of particular transcripts, such as those of asps, between An. caninum and N. americanus might reflect their distinct roles in the modulation of the host immune response between the two hookworms, an hypothesis that requires testing.
A benefit of investigating the transcriptome of parasitic nematodes using predictive algorithms is that potential drug targets can be inferred and/or prioritized. The present chapter identified a subset of 278 ‘druggable’ proteins, of which 18 did not match any human homologues (cf. Results section). Of these 18 molecules, mitochondrial-associated proteins were significantly represented (i.e., encoded by the *C. elegans* orthologues W01a8.4, *ucr-1*, F26E4.6 and Y71H2aM.4; cf. Table 6.4). Mitochondria are essential organelles with central roles in diverse cellular processes, such as apoptosis, energy production via oxidative phosphorylation, ion homeostasis, and the synthesis of haeme, lipid, amino acids, and iron-sulfur ions [92]. In *C. elegans*, defects in the mitochondrial respiratory chain lead to or are associated with a wide variety of abnormalities, including embryonic, larval and adult lethality, sterility and embryonic defects [92]. Despite their essential roles in numerous fundamental biological processes, knowledge of mitochondrial genes and proteins in parasitic nematodes has been utilized mainly to study their systematics, population genetics and ecology [93–95]. However, that some mitochondrial-associated proteins are predicted to be essential in *N. americanus* and significantly different from human homologues provides a context for the discovery of new drug targets in mitochondrial pathways and chemical compounds that disrupt these pathways [95,96]. Amongst the other *N. americanus* orthologues of essential *C. elegans* genes, *nrs-2* encodes an asparaginyl-tRNA synthetase (AsnRS), which is a class II aminoacyl-tRNA synthetase that catalyzes the attachment of asparagine to its cognate tRNA and is required for protein biosynthesis [97]; loss of *nrs-2* function via RNAi has been shown to result in a number of phenotypes, including adult and larval lethality and/or larval arrest [97]. In parasitic nematodes, information on amino acid biosynthesis is limited [98]. Although a number of parasitic helminths, including the nematode *Heligmosomoides polygyrus* [sic. *H. bakeri*] and the trematode *Fasciola hepatica*, have been reported to excrete asparagine during *in vitro* incubation [99,100], the role of asparagine synthetases in essential biological processes is currently unknown. However, in a study investigating the molecular mechanisms of induced cell differentiation in human pro-myelocytic leukemia, asparagine synthetase transcription was reported to be significantly reduced in maturing monocytes/macrophages [101]: therefore, an active role of asparagine synthetases in the development and growth of cancer cells has been suggested, which led to the hypothesis that the induction of a down-regulation of asparagine synthetases might be a new strategy for the treatment of blast cell leukaemia [101]. This finding raises questions about the role(s) of asparagine synthetases in cell differentiation and maturation in parasitic nematodes and the potential of inhibitors of these enzymes as anti-hookworm drugs.
The present chapter has provided new insights into the transcriptome of *N. americanus*, elucidated similarities and differences between the transcriptomes of *N. americanus* and the related canine hookworm, *An. caninum*, and predicted a panel of novel drug targets and nematocides. All except one of the essential ‘druggable’ proteins (n = 18) inferred for *N. americanus* were present in the *An. caninum* (and *C. elegans*) but not in the mammalian hosts, suggesting relative sequence conservation for these targets among nematodes. The prediction of such targets is particularly important, considering the risk of emerging drug resistance in parasitic nematodes [102,103]. Clearly, transcriptomic and genomic studies, such as that carried out here can facilitate and expedite the prevalidation of targets for nematocidal drugs, although the lack of genomic and transcriptomic data for many nematodes, including the human hookworm *An. duodenale*, impairs the comparative exploration of essential biological pathways in parasitic nematodes of major public health significance [6]. Furthermore, the present analysis has inferred qualitative and quantitative differences in the transcriptome between *N. americanus* and *An. caninum*, raising questions as to the suitability of the latter species as a model for the former. Although these differences require experimental validation, there is a need to define the transcriptome of *An. duodenale* as a foundation for comparative investigations with a perspective on the identification of new and hookworm-specific drug targets.
6.5 References


179


20. Chapter 5. Oligonucleotide microarray and bioinformatic analyses to explore transcriptional conservation between *Ancylostoma caninum* and *Haemonchus contortus*.


Table 6.1 Summary of the expressed sequence tag (EST) data for the adult stage of *Necator americanus* determined following 454 sequencing and detailed bioinformatics annotation and analyses.

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of EST clusters</td>
<td>19,997</td>
</tr>
<tr>
<td>Average length (± standard deviation)</td>
<td>369 bp ± 215.31</td>
</tr>
<tr>
<td>Containing an Open Reading Frame</td>
<td>12,799</td>
</tr>
<tr>
<td>Signal peptides</td>
<td>274</td>
</tr>
<tr>
<td>Returning InterProScan results</td>
<td>7,214 (2,381 domains)</td>
</tr>
<tr>
<td>Gene Ontology</td>
<td>2,950 (887 terms)</td>
</tr>
<tr>
<td>Biological process</td>
<td>4,830 (314 terms)</td>
</tr>
<tr>
<td>Cellular component</td>
<td>3,087 (117 terms)</td>
</tr>
<tr>
<td>Molecular function</td>
<td>8,671 (456 terms)</td>
</tr>
<tr>
<td>Prediction of biological pathways (KOBAS)</td>
<td>235</td>
</tr>
</tbody>
</table>
Table 6.2 The thirty most abundant protein domains inferred using the InterProScan software from peptides inferred for *Necator americanus* and *Ancylostoma caninum*. The arrows infer statistically significant ($p < 0.05$; chi-square) higher (▲) or lower (▼) number of genes encoding proteins (with particular InterPro domains) common to *N. americanus* and *An. caninum*.

<table>
<thead>
<tr>
<th>InterPro domain</th>
<th>No. of <em>Na</em> EST clusters (%)</th>
<th>No. of <em>Ac</em> EST clusters (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WD40</td>
<td>315 (10.6)▼</td>
<td>553 (14.5)▲</td>
</tr>
<tr>
<td>EF-HAND</td>
<td>196 (6.7)</td>
<td>187 (2.6)</td>
</tr>
<tr>
<td>Proteinase inhibitors</td>
<td>230 (7.8)▲</td>
<td>126 (3.3)▼</td>
</tr>
<tr>
<td>Proteases</td>
<td>179 (6.1)</td>
<td>177 (4.6)</td>
</tr>
<tr>
<td>Protein kinases</td>
<td>131 (4.4)▼</td>
<td>388 (10.1)▲</td>
</tr>
<tr>
<td>NAD(P)-binding domain</td>
<td>114 (3.9)▼</td>
<td>160 (4.2)▲</td>
</tr>
<tr>
<td>Transthyretin-like</td>
<td>97 (3.3)▲</td>
<td>19 (0.5)▼</td>
</tr>
<tr>
<td>Galectin, carbohydrate recognition domain</td>
<td>95 (3.2)▲</td>
<td>66 (1.7)▼</td>
</tr>
<tr>
<td>SCP-like extracellular</td>
<td>94 (3.2)▼</td>
<td>362 (9.5)▲</td>
</tr>
<tr>
<td>Peptidyl-prolyl cis-trans isomerase</td>
<td>91 (3.1)▲</td>
<td>25 (0.6)▼</td>
</tr>
<tr>
<td>RNA recognition motif, RNP-1</td>
<td>83 (2.8)▼</td>
<td>198 (6.2)▲</td>
</tr>
<tr>
<td>Mitochondrial substrate/solute carrier</td>
<td>88 (3)</td>
<td>88 (2.3)</td>
</tr>
<tr>
<td>Thioredoxin fold</td>
<td>81 (2.7)</td>
<td>70 (1.8)</td>
</tr>
<tr>
<td>Allergen V5/Tpx-1 related</td>
<td>64 (2.2)▼</td>
<td>232 (6.6)▲</td>
</tr>
<tr>
<td>Zinc finger, C2H2-type</td>
<td>64 (2.2)▼</td>
<td>185 (4.8)▲</td>
</tr>
<tr>
<td>Aldo/keto reductase</td>
<td>60 (2)▲</td>
<td>9 (0.2)▼</td>
</tr>
<tr>
<td>Scr homology-3 domain</td>
<td>57 (2)</td>
<td>98 (2.6)</td>
</tr>
<tr>
<td>Actin/actin like</td>
<td>56 (2)</td>
<td>49 (1.3)</td>
</tr>
<tr>
<td>Short-chain dehydrogenase/reductase SDR</td>
<td>51 (1.7)</td>
<td>51 (6.2)</td>
</tr>
<tr>
<td>Metridin-like ShK toxin</td>
<td>47 (1.6)▲</td>
<td>6 (0.1)▼</td>
</tr>
<tr>
<td>Histone-fold</td>
<td>44 (1.5)▲</td>
<td>19 (0.5)▼</td>
</tr>
<tr>
<td>Nucleotide binding, alpha beta plait</td>
<td>43 (1.4)▼</td>
<td>80 (2.1)▲</td>
</tr>
<tr>
<td>Heat shock protein Hsp20</td>
<td>41 (1.4)▲</td>
<td>14 (0.4)▼</td>
</tr>
<tr>
<td>Chaperonin Cpn60/TCP-1</td>
<td>39 (1.3)</td>
<td>50 (1.3)</td>
</tr>
<tr>
<td>Cytochrome P450</td>
<td>39 (1.3)▲</td>
<td>4 (0.1)▼</td>
</tr>
<tr>
<td>Ankyrin</td>
<td>38 (1.2)▼</td>
<td>271 (7)▲</td>
</tr>
<tr>
<td>Annexin repeat</td>
<td>37 (1.2)</td>
<td>53 (1.4)</td>
</tr>
<tr>
<td>Ubiquitin-conjugating enzyme, E2</td>
<td>37 (1.2)▲</td>
<td>15 (0.4)▼</td>
</tr>
<tr>
<td>Tetratricopeptide repeat</td>
<td>36 (1.2)</td>
<td>20 (0.5)</td>
</tr>
<tr>
<td>Protein-tyrosine phosphatase, receptor/non-receptor type</td>
<td>35 (1.2)▼</td>
<td>76 (2)▲</td>
</tr>
</tbody>
</table>
Table 6.3 The twenty most abundant Gene Ontology (GO) terms (according to the categories ‘biological process’, ‘cellular component’ and ‘molecular function’) for peptides inferred for *Necator americanus* and *Ancylostoma caninum*.

<table>
<thead>
<tr>
<th>GO term</th>
<th>GO code</th>
<th>No. of <em>Na</em> EST clusters (%)</th>
<th>No. of <em>Ac</em> EST clusters (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Biological process</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Translation</td>
<td>GO:0006412</td>
<td>599 (20.3)</td>
<td>146 (3.8)</td>
</tr>
<tr>
<td>Metabolic process</td>
<td>GO:0008152</td>
<td>438 (14.9)</td>
<td>284 (7.4)</td>
</tr>
<tr>
<td>Proteolysis</td>
<td>GO:0006508</td>
<td>329 (11.2)</td>
<td>254 (6.6)</td>
</tr>
<tr>
<td>Oxidation reduction</td>
<td>GO:0055114</td>
<td>197 (6.7)</td>
<td>85 (2.2)</td>
</tr>
<tr>
<td>Protein amino acid phosphorylation</td>
<td>GO:0006468</td>
<td>147 (5)</td>
<td>159 (4.2)</td>
</tr>
<tr>
<td>Regulation of transcription, DNA-dependent</td>
<td>GO:0006355</td>
<td>137 (4.6)</td>
<td>53 (1.4)</td>
</tr>
<tr>
<td>Transport</td>
<td>GO:0006810</td>
<td>134 (4.5)</td>
<td>114 (3)</td>
</tr>
<tr>
<td>ATP synthesis coupled proton transport</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein folding</td>
<td>GO:0006457</td>
<td>104 (3.5)</td>
<td>48 (1.3)</td>
</tr>
<tr>
<td>Carbohydrate metabolic process</td>
<td>GO:0005975</td>
<td>101 (3.4)</td>
<td>100 (2.6)</td>
</tr>
<tr>
<td>Small GTPase mediated signal transduction</td>
<td>GO:0007264</td>
<td>62 (2.1)</td>
<td>38 (1)</td>
</tr>
<tr>
<td>Ubiquitin-dependent protein catabolic process</td>
<td>GO:0006511</td>
<td>62 (2.1)</td>
<td>30 (0.8)</td>
</tr>
<tr>
<td>Intracellular protein transport</td>
<td>GO:0006886</td>
<td>59 (2)</td>
<td>52 (1.4)</td>
</tr>
<tr>
<td>Vesicle-mediated transport</td>
<td>GO:0016192</td>
<td>54 (1.8)</td>
<td>39 (1)</td>
</tr>
<tr>
<td>Nucleosome assembly</td>
<td>GO:0006334</td>
<td>53 (1.8)</td>
<td>21 (0.5)</td>
</tr>
<tr>
<td>Protein transport</td>
<td>GO:0015031</td>
<td>50 (1.7)</td>
<td>40 (1)</td>
</tr>
<tr>
<td>Response to oxidative stress</td>
<td>GO:0006979</td>
<td>48 (1.6)</td>
<td>9 (0.2)</td>
</tr>
<tr>
<td>Protein amino acid dephosphorylation</td>
<td>GO:0006470</td>
<td>47 (1.6)</td>
<td>45 (1.2)</td>
</tr>
<tr>
<td>Protein polymerization</td>
<td>GO:0051258</td>
<td>46 (1.6)</td>
<td>15 (0.4)</td>
</tr>
<tr>
<td><strong>Cellular component</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intracellular</td>
<td>GO:0005622</td>
<td>798 (25.1)</td>
<td>297 (7.5)</td>
</tr>
<tr>
<td>Ribosome</td>
<td>GO:0005840</td>
<td>499 (17)</td>
<td>88 (2.3)</td>
</tr>
<tr>
<td>Membrane</td>
<td>GO:0016020</td>
<td>296 (9.7)</td>
<td>251 (6.6)</td>
</tr>
<tr>
<td>Nucleus</td>
<td>GO:0005634</td>
<td>280 (9.5)</td>
<td>174 (4.6)</td>
</tr>
<tr>
<td>Integral to membrane</td>
<td>GO:0016021</td>
<td>185 (6.3)</td>
<td>143 (3.7)</td>
</tr>
<tr>
<td>Cytoplasm</td>
<td>GO:0005737</td>
<td>141 (4.8)</td>
<td>122 (3.2)</td>
</tr>
<tr>
<td>Extracellular region</td>
<td>GO:0005576</td>
<td>86 (2.9)</td>
<td>156 (4)</td>
</tr>
<tr>
<td>Nucleosome</td>
<td>GO:0000786</td>
<td>51 (1.7)</td>
<td>18 (0.5)</td>
</tr>
<tr>
<td>Protein complex</td>
<td>GO:0043234</td>
<td>46 (1.6)</td>
<td>15 (0.4)</td>
</tr>
<tr>
<td>Endoplasmic reticulum</td>
<td>GO:0005783</td>
<td>38 (1.3)</td>
<td>35 (0.9)</td>
</tr>
<tr>
<td>Mitochondrion</td>
<td>GO:0005739</td>
<td>36 (1.2)</td>
<td>5 (0.1)</td>
</tr>
<tr>
<td>Cytoskeleton</td>
<td>GO:0005856</td>
<td>33 (1.1)</td>
<td>14 (0.4)</td>
</tr>
<tr>
<td>Microtubule</td>
<td>GO:0005874</td>
<td>31 (1)</td>
<td>15 (0.4)</td>
</tr>
<tr>
<td>Proton-transporting two-sector ATPase complex, catalytic domain</td>
<td>GO:0033178</td>
<td>29 (1)</td>
<td>19 (0.5)</td>
</tr>
<tr>
<td>Proton-transporting two-sector ATPase complex, proton-transporting domain</td>
<td>GO:0033177</td>
<td>27 (0.9)</td>
<td>8 (0.2)</td>
</tr>
<tr>
<td>GO:0005743</td>
<td>Mitochondrial inner membrane</td>
<td>22 (0.7)</td>
<td>7 (0.2)</td>
</tr>
<tr>
<td>GO:0045261</td>
<td>Proton-transporting ATP synthase complex, catalytic core F(1)</td>
<td>18 (0.6)</td>
<td>9 (0.2)</td>
</tr>
<tr>
<td>GO:0030131</td>
<td>Clathrin adaptor complex</td>
<td>15 (0.5)</td>
<td>9 (0.2)</td>
</tr>
<tr>
<td>GO:0005839</td>
<td>Proteasome core complex</td>
<td>15 (0.5)</td>
<td>15 (0.4)</td>
</tr>
<tr>
<td>GO:0005853</td>
<td>Eukaryotic translation elongation factor 1 complex</td>
<td>15 (0.5)</td>
<td>8 (0.2)</td>
</tr>
</tbody>
</table>

**Molecular function**

| GO:0005524 | ATP binding | 558 (18.9) | 514 (13.4) |
| GO:0003735 | Structural constituent of ribosome | 527 (17.9) | 91 (2.4) |
| GO:0003824 | Catalytic activity | 429 (14.5) | 346 (9) |
| GO:0016491 | Oxidoreductase activity | 317 (10.8) | 185 (4.8) |
| GO:0005515 | Protein binding | 311 (10.5) | 212 (5.6) |
| GO:0005488 | Binding | 287 (9.3) | 237 (6.2) |
| GO:0008270 | Zinc ion binding | 287 (9.3) | 214 (5.6) |
| GO:0003677 | DNA binding | 242 (8.2) | 121 (3.2) |
| GO:0004252 | Serine-type endopeptidase inhibitor activity | 230 (7.8) | 25 (0.7) |
| GO:0003676 | Nucleic acid binding | 204 (6.9) | 192 (5) |
| GO:0005525 | GTP binding | 202 (6.8) | 95 (2.5) |
| GO:0005509 | Calcium ion binding | 169 (5.8) | 79 (2) |
| GO:0009055 | Electron carrier activity | 140 (4.7) | 59 (1.5) |
| GO:0020037 | Heme binding | 134 (4.5) | 17 (0.4) |
| GO:0003723 | RNA binding | 124 (4.2) | 90 (2.3) |
| GO:0008270 | Iron ion binding | 123 (4.2) | 18 (0.5) |
| GO:0000166 | Nucleotide binding | 113 (3.8) | 138 (3.6) |
| GO:0004190 | Aspartic-type endopeptidase activity | 101 (3.4) | 27 (0.7) |
| GO:0005529 | Sugar binding | 99 (3.4) | 18 (0.5) |
| GO:0003700 | Transcription factor activity | 98 (3.3) | 47 (1.2) |
Table 6.4 Description of *Caenorhabditis elegans* orthologues of *Necator americanus* contigs for which inferred peptides were associated with ‘druggable’ InterPro domains and/or Enzyme Commission (EC) numbers, and examples of candidate nematocidal compounds linked to these domains predicted using the BRENDA database. These genes are not present in *H. sapiens*. The presence of known *Ancylostoma caninum* orthologues is also indicated (✓).

<table>
<thead>
<tr>
<th><em>C. elegans</em> gene ID</th>
<th>Gene name</th>
<th>RNAi phenotypes</th>
<th>Protein description</th>
<th><em>An. caninum</em> orthologue</th>
<th>Druggable IPR domain (description)</th>
<th>Examples of compounds in the BRENDA database</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBGene00003816</td>
<td>nrs-2</td>
<td>Lethal, larval lethal, larval arrest, sterile</td>
<td>Asparagine synthetase</td>
<td>✓</td>
<td></td>
<td>1-methyl-4-(1-methylethyl)-7-oxabicyclo[2.2.1]heptane</td>
</tr>
<tr>
<td>WBGene00012166</td>
<td>W01A8.4</td>
<td>Lethal, embryonic lethal, larval arrest, sterile</td>
<td>Mitochondrial NADH dehydrogenase (ubiquinone) complex (complex I) subunit V0 subunit of V-ATPase</td>
<td>✓</td>
<td></td>
<td>1-Geranyl-2-methylbenzimidazole</td>
</tr>
<tr>
<td>WBGene00021952</td>
<td>vha-19</td>
<td>Larval lethal, maternal sterile, sick</td>
<td>Elongation factor Tu</td>
<td>✓</td>
<td></td>
<td>Efrageptin</td>
</tr>
<tr>
<td>WBGen00007001</td>
<td>tufm-2</td>
<td>Embryonic lethal, slow growth, sick, sterile progeny</td>
<td>AAA ATPase</td>
<td>✓</td>
<td></td>
<td>Leu-Gly-Asn repeat-enriched protein</td>
</tr>
<tr>
<td>WBGene00010562</td>
<td>cdc-48.3</td>
<td>Embryonic lethal, larval lethal, larval arrest, slow growth, sick</td>
<td>Mitochondrial processing protease enhancing protein</td>
<td>✓</td>
<td>IPR001431 (Insulinase family)</td>
<td>1-beta-D-ribofuranosyl-1,2,4-triazole-3-carboxamide-5'-triphosphate</td>
</tr>
<tr>
<td>WBGene00018963</td>
<td>ucr-1</td>
<td>Embryonic lethal, larval arrest, maternal sterile</td>
<td>Acyl-CoA dehydrogenase</td>
<td>✓</td>
<td></td>
<td>1,10-phenanthroline</td>
</tr>
<tr>
<td>WBGene00001566</td>
<td>acdh-13</td>
<td>Embryonic lethal, sterile progeny, slow growth</td>
<td>Intestinal acid phosphatase</td>
<td>✓</td>
<td></td>
<td>Amyloid beta</td>
</tr>
<tr>
<td>WBGene00004020</td>
<td>pho-1</td>
<td>Embryonic lethal, larval arrest, sterile</td>
<td>Cytochrome c oxidase</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WBGene00009161</td>
<td>F26E4.6</td>
<td>Embryonic lethal, larval arrest, slow growth, maternal</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene ID</td>
<td>Gene Symbol</td>
<td>Description</td>
<td>Subunit/Activity</td>
<td>Enzyme ID</td>
<td>Activity</td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td>-------------</td>
<td>--------------------------------------------------</td>
<td>-----------------------------------------------</td>
<td>-------------------------</td>
<td>-----------------------------</td>
<td></td>
</tr>
<tr>
<td>WBGene00022169</td>
<td>Y71H2aM.4</td>
<td>sterile Embryonic lethal, larval arrest, slow growth</td>
<td>NADH:ubiquinone oxidoreductase, NDUFC2/B14.5B subunit</td>
<td></td>
<td>√</td>
<td>1-Geranyl-2-methylbenzimidazole</td>
</tr>
<tr>
<td>WBGene00000151</td>
<td>apn-1</td>
<td>Embryonic lethal AP endonuclease (family 2)</td>
<td></td>
<td></td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>WBGene00019885</td>
<td>R05D8.7</td>
<td>Embryonic lethal Reductases with broad range of substrate specificities Dehydrogenase</td>
<td></td>
<td>IPR002198 (Short-chain dehydrogenase/reductase) IPR002198 (Short-chain dehydrogenase/reductase)</td>
<td>√</td>
<td>4-Trifluoromethyl-2,3-dihydro-2,3-dihydroxybenzoate 4-Trifluoromethyl-2,3-dihydro-2,3-dihydroxybenzoate</td>
</tr>
<tr>
<td>WBGene00020089</td>
<td>R119.3</td>
<td>Embryonic lethal Aminopeptidase Uncharacterized conserved protein, contains double-stranded beta-helix domain</td>
<td></td>
<td></td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>WBGene00011803</td>
<td>T16G12.1</td>
<td>Embryonic lethal Zinc metalloprotease</td>
<td></td>
<td></td>
<td>√</td>
<td>1,10-phenanthroline</td>
</tr>
<tr>
<td>WBGene00020149</td>
<td>T01D1.4</td>
<td>Embryonic lethal Phosphoglycerate mutase</td>
<td></td>
<td></td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>WBGene00006592</td>
<td>dpy-31</td>
<td>Embryonic lethal Protein tyrosine phosphatase</td>
<td></td>
<td>IPR000387 (Tyrosine protein phosphatases)</td>
<td>√</td>
<td>1,3-difluoro-2-((E)-2-nitrovinyl)benzene</td>
</tr>
<tr>
<td>WBGene00019001</td>
<td>F57B10.3</td>
<td>Embryonic lethal, larval lethal, slow growth</td>
<td></td>
<td></td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>WBGene00016356</td>
<td>C33F10.8</td>
<td>Embryonic lethal, slow growth</td>
<td></td>
<td></td>
<td>√</td>
<td></td>
</tr>
</tbody>
</table>
Fig. 6.1 Relationships of proteins predicted for *Necator americanus* with homologues from *Ancylostoma caninum* and *Caenorhabditis elegans*, displayed in a SimiTri plot [44]. The description of proteins with most abundant InterPro domains identified in each similarity group is given in the boxes.
CHAPTER 7

A practical, bioinformatic workflow system for the analysis of large datasets generated by next-generation sequencing

Abstract
Transcriptomics (at the level of single cells, tissues and/or whole organisms) underpins many fields of biomedical science, from understanding the basic cellular function in model organisms, to the elucidation of the biological events that govern the development and progression of human diseases, and the exploration of the mechanisms of survival, drug-resistance and virulence of pathogens. Next-generation sequencing (NGS) technologies are contributing to a massive expansion of transcriptomics in all fields and are reducing the cost, time and performance barriers presented by conventional approaches. However, bioinformatic tools for the analysis of the sequence datasets produced by these technologies can be daunting to researchers with limited or no expertise in bioinformatics. The present chapter describes the construction of a practical workflow system, and critically evaluates it for the analysis and annotation of large scale sequence datasets generated by NGS. The utility of this workflow for the exploration of differences in the transcriptomes among various stages and both sexes of an economically important parasitic worm (*Oesophagostomum dentatum*) as well as the prediction and prioritization of essential molecules (including GTPases, protein kinases and phosphatases) is demonstrated. This workflow system provides a practical tool for the assembly, annotation and analysis of NGS datasets, also to researchers with a limited bioinformatic expertise. The custom-written Perl, Python and Unix shell computer scripts can be readily modified or adapted to suit many different applications. This system is now utilised routinely for the analysis of datasets from pathogens of major socio-economic importance and can, in principle, be applied to transcriptomic datasets from any organism.
7.1 Introduction

Transcriptomics is the molecular science of examining, simultaneously, the transcription of all genes at the level of the cell, tissue and/or whole organism, allowing inferences regarding cellular functions and mechanisms (Chapter 1). The ability to measure the transcription of thousands of genes simultaneously has led to major advances in all biomedical fields, from understanding the basic function in model organisms, such as the free-living nematode Caenorhabditis elegans [1-3] or the vinegar fly, Drosophila melanogaster [4-6] to studying molecular events associated with the development and progression of human diseases, including cancer [7-9] and neurodegenerative disorders [10-12], to the exploration of the mechanisms of survival, drug-resistance and virulence/pathogenicity of bacteria [13,14] and other socioeconomically important pathogens, such as parasites [15-20]. For more than a decade, transcriptomes have been determined by sequencing expressed sequence tags (ESTs) using the conventional Sanger method [21,22], whereas levels of transcription have been established quantitatively or semi-quantitatively by real-time PCR [23] and/or cDNA microarrays [24]. The use of these technologies has been accompanied by an increasing demand for analytical tools for the efficient annotation of nucleotide sequence datasets, particularly within the framework of large-scale EST projects [25]. With a substantial expansion of EST sequencing has come the development of algorithms for sequence assembly, analysis and annotation, in the form of individual programs [26-28] and integrated pipelines [29,30], some of which have been made available on the worldwide web [29,31,32]. However, the cost and time associated with large-scale sequencing using a conventional (Sanger) method and/or the design of customised analytical tools (e.g., cDNA microarray) have driven the search for alternative methods for transcriptomic studies [33].

In the last few years, there has been a massive expansion in the demand for and access to low cost, high-throughput sequencing, attributable mainly to the development of next-generation sequencing technologies (NGS), which allow massively parallelised sequencing of millions of nucleic acids [33,34]. These sequencing platforms, such as 454/Roche ([35]; http://www.454.com/) and Illumina/Solexa ([36]; http://www.illumina.com/), have transformed transcriptomics by decreasing the cost, time and performance limitations presented by previous approaches. This situation has resulted in an explosion of the number of EST sequences deposited in databases worldwide, the majority of which is still awaiting detailed functional annotation. However, the high-throughput analysis of such large datasets has necessitated significant advances in computing capacity and performance,
and in the availability of bioinformatic tools to distil biologically meaningful information from raw sequence data.

Sequences generated by NGS are significantly shorter (454/Roche: ~400 bases; Illumina/ABI-SOLiD: ~60 bases) than those determined by Sanger sequencing (0.8-1 kb), which poses a challenge for assembly. In addition, the data files generated by these technologies are often gigabytes to terabytes (1x10^9 to 1x10^12 bytes) in size, substantially increasing the demands placed on data transfer and storage, such that many web-based interfaces are not suited for large-scale analyses. The bioinformatic processing of large datasets usually requires access to powerful computers and support from bioinformaticians with significant expertise in a range of programming languages (e.g., Perl and Python). This situation has limited the accessibility of high-throughput sequencing technologies to some (smaller) research groups, and has thus restricted somewhat the “democratisation” of large-scale genomic and/or transcriptomic sequencing. Clearly, user-friendly and flexible bioinformatic pipelines are needed to assist researchers from different disciplines and backgrounds in accessing and taking full advantage of the advances heralded by NGS. Increasing the accessibility to high-throughput sequencing will have major benefits in a range of areas, including the investigation of pathogens. The exploration of the transcriptomes of pathogens has major implications in improving our understanding of their development and reproduction, survival in and interactions with the host, virulence, pathogenicity, the diseases that they cause and drug resistance [17-20,37-39], and has the potential to pave the way to novel approaches for treatment, diagnosis and control. The present chapter (i) describes the construction of a semi-automated, bioinformatic workflow system for the analysis and annotation of large-scale sequence datasets generated by NGS, (ii) demonstrates its utility by profiling differences in the transcriptome of an economically important parasite, _Oesophagostomum dentatum_ (Strongylida), throughout its development, and (iii) indicates the broader applicability of this system to different types of transcriptomic datasets.

### 7.2 Materials and methods

#### 7.2.1 Sequence datasets

For this study, original cDNA sequence datasets representing four distinct developmental stages of _Oe. dentatum_ (i.e., third-stage [L3] and fourth-stage [L4] larvae as well as adult female and male worms) were produced and stored as described previously [40]. Total RNA (10 μg) from each stage and/or sex was used to construct a normalised cDNA library; each library was sequenced using a Genome Sequencer™ (GS) Titanium FLX (Roche Diagnostics) as previously described [18]. FASTA- and associated
files, with short-read sequence quality scores of each dataset, were extracted from each SFF-file; sequence adaptors were clipped using the ‘sff_extract’ software (http://bioinf.comav.upv.es/sff_extract/index.html).

7.2.2 Bioinformatic components for the construction of the workflow system

Five components (1-5), documented in a series of peer-reviewed, international publications, were selected based on the parameters of general applicability, ease of use, versatility and efficiency. Once constructed, the workflow system was applied to the analysis of the *Oe. dentatum* datasets.

1. Assembly. The Contig Assembly Program (CAP3 v.3; [31]) was used to cluster sequences (with quality scores) into contigs and singletons from individual or combined (i.e., pooled) datasets, employing a minimum sequence overlap of 40 nucleotides and an identity threshold of 90%. This program was selected to enable the assembly of relatively long sequences and to remove redundant short-reads [41].

2. Similarity searching. BLASTn and BLASTx algorithms (42) were used to compare contigs and singletons with sequences available in public databases (i.e., NCBI [www.ncbi.nlm.nih.gov] and EMBL-EBI Parasite Genome Blast Server [www.ebi.ac.uk]; April 2010), to identify putative homologues in range of other organisms (e-value cut-off: <1e-05). For nematodes, WormBase (Release WS200; www.wormbase.org) was interrogated extensively for relevant information on *C. elegans* orthologues/homologues, including transcriptomic, proteomic, RNA interference (RNAi) phenotypes and interactomic data.

3. Prediction and annotation of peptides. The program ESTScan [32] was used to conceptually translate peptides from assembled contigs and singletons. InterProScan (available at http://www.ebi.ac.uk/InterProScan/; [27]) and gene ontology (GO; [43]) were used to classify peptides (based on their putative function/s). Biological pathways were inferred from *C. elegans* for each peptide using the KEGG Orthology-Based Annotation System software (KOBAS; [44]) and displayed using the iPPath tool (http://pathways.embl.de/data_mapping.html; [45]).

4. In silico subtraction. A BLASTn algorithm, employing a stringent cut-off (e-value cut-off: <1e-15; [17]), was used to examine differential transcription between datasets by subtraction *in silico*. Peptides corresponding to transcripts that were unique to a particular dataset were assigned parental (i.e., level 1) InterPro terms and
compared, using a BLASTp algorithm (e-value cut-off: <1e-15), with peptides inferred from the assembly of sequences from combined datasets. The subtraction approach allows qualitative (not quantitative) differences between or among samples to be established.

5. Probabilistic functional networking of protein-encoding genes, and drug target prediction. Interaction networks among C. elegans orthologues of differentially transcribed molecules were inferred using an established approach [46]. The druggability of C. elegans homologues of molecules unique to a particular O. dentatum dataset or common to all datasets was inferred using a published method [18]. Briefly, the InterPro domains of predicted proteins were compared with those linked to known, small molecular drugs, which follow the 'Lipinsky rule of 5' regarding bioavailability [47,48]. GO terms were mapped to Enzyme Commission (EC) numbers, and a list of enzyme-targeting drugs was compiled based on data available in the BRENDA database (www.brenda-enzymes.info; [49,50]). The C. elegans orthologues/homologues included in this list were ranked according to the ‘severity’ of non-wildtype RNAi phenotypes (including lethality or sterility of different developmental stages; see www.wormbase.org; release WS200).

7.3 Results

A semi-automated bioinformatic workflow system (see Fig. 7.1), incorporating five key bioinformatic components, was constructed and linked using customised Perl, Python and Unix shell computer scripts (listed in Appendix 7.1 and accessible via http://research.vet.unimelb.edu.au/gasserlab/index.html). This system was then assessed for the assembly, analysis and functional annotation of each or all of the four sequence datasets for Oe. dentatum. The specificity of the in silico subtraction step was verified using independent experimental evidence.

7.3.1 Assembly and detailed annotation and analyses of the Oesophagostumum dentatum datasets

A total of 1,826,367 sequences (244 ± 32 bases; i.e., mean length ± standard deviation) were determined for L3, L4 as well as adult female and male of Oe. dentatum. Following the clipping of adapter sequences, only sequences of >100 bases (n = 1,800,874; 98.6%) were included in further analyses. The numbers of contigs assembled for each of the four datasets are listed in Table 7.1. The assembly of the sequences of all four datasets yielded 36,233 contigs (516 ± 316 bases in length; Fig. 7.2) and 452,528 singletons (Table 7.1); sequences (n = 115) with similarity (e-value
cut-off: <1e-15) to potential host molecules were excluded. The L3 dataset had the largest number of sequence clusters with orthologues/homologues in *C. elegans* (n = 32,904; Table 7.1) and in organisms other than nematodes (n = 14,731; Table 7.1), whereas the L4 dataset included the largest number of clusters with orthologues/homologues in other parasitic nematodes (n = 38,634; Table 7.1).

Of the four assembled datasets, the L3 set included the largest number of sequence clusters with predicted open reading frames (ORFs; n = 57,818; Table 7.1), of which 27,297 (47.2%) could be annotated functionally using InterPro terms and 12,763 (22.1%) could be assigned GO terms, including 19,705 ‘biological process’, 10,926 ‘cellular component’ and 34,904 ‘molecular function’. The numbers of peptides inferred from sequence clusters in the adult female, adult male and/or L4 datasets, which could be assigned InterPro and/or GO terms, are given in Table 7.1. In total, 85,395 peptides were predicted for all sequences from all four datasets, representing 17.5% of clusters (Table 7.1); 56,940 (66.7%) of them could be mapped to known proteins defined by 31,982 different domains, the most represented being 'SCP-like extracellular' (IPR014044; 1.2% of the peptides mapping to a conserved protein motif), 'NAD(P)-binding' (IPR016040; 1.1%) and 'proteinase inhibitor I2, Kunitz metazoa' (IPR002223; 1%) (Table 7.2). GO annotation allowed 56,940 (66.7%) inferred proteins to be assigned to 19,346 ‘biological process’, 11,007 ‘cellular component’ and 35,182 ‘molecular function’ terms (Table 7.1). The predominant terms were ‘metabolic process’ (GO:0008152; 10.9%), ‘proteolysis’ (GO:0006508; 7%) and ‘translation’ (GO:0006412; 5.4%) for ‘biological process’; ‘intracellular’ (GO:0005622; 17.5%), ‘membrane’ (GO:0016020; 15.6%) and ‘nucleus’ (GO:0005634; 11.6%) for ‘cellular component’ and ‘ATP binding’ (GO:0005524; 7.5%); ‘catalytic activity’ (GO:0003824; 7%) and ‘binding’ (GO:0005488; 4.6%) for ‘molecular function’ (Table 7.3). Proteins inferred from the combined assembly were predicted to be involved in 262 different biological pathways, defined by 64 unique KEGG terms, of which ‘peptidases’ (12%), ‘other enzymes’ (8%) and ‘antigen processing and presentation’ (5.5%) were predominant (see Appendix 7.2). A display of biological pathways, defined by KEGG terms inferred from predicted peptides and mapped to the complement of known pathways in *C. elegans*, is shown in Appendix 7.3.

Using BLASTn algorithms, subsets of 3,451, 10,344, 14,380 and 7,520 nucleotide sequences were identified as being uniquely transcribed in adult female, adult male, L3 and L4, respectively (Table 7.1). The accuracy of the *in silico* subtraction process was verified using independent evidence from a previous analysis of differential transcription between adult females and males of *Oe. dentatum* using a microarray-based approach [51]. This verification showed that all 220 female-
171 male-enriched molecules characterised previously ([51]; GenBank accession numbers AM157797-AM158083) were contained exclusively within the female and male datasets, respectively, following in silico subtraction (not shown). Based on these findings, the specificity of the subtraction process, calculated using the Wilson score [52] at a confidence interval of 95%, ranged from 98 to 100%. Of the 139 parental functional domains assigned to predicted peptides unique to the adult female dataset, 'chitin-binding protein, peritrophin-A' (IPR002557; 8.6%) and 'basic-leucine zipper (bZIP) transcription factor' (IPR004827; 4.8%) were highly represented. Of the 243 protein motifs identified amongst the predicted peptides that were unique to the adult male dataset, 'PapD-like' (IPR008962; 4%) and 'major-sperm protein' (IPR000535; 3.7%) were most represented. For the L3 dataset, 220 unique protein motifs were identified, of which 'RmlC-like jelly roll fold' (IPR014710; 4.5%) and 'six-bladed beta-propeller' (IPR011042; 2.7%) had the highest representation. In contrast, of the 249 protein motifs unique to L4 dataset, 'peptidase M24, methionine aminopeptidase' (IPR0011714; 2.2%) and 'FAD-binding' (IPR016166; 1.3%) were the predominant domains (Table 7.2). The number of 'biological process', 'cellular component' and 'molecular function' terms assigned to peptides unique to each of the individually assembled datasets is given in Table 7.1. The KOBAS analysis assigned 7, 16, 18 and 23 KEGG terms to inferred peptides exclusive to the adult female, adult male, L3 and L4 datasets, respectively; of the 23 KEGG terms assigned to L4, 20 could be mapped to known pathways in C. elegans (Fig. 7.3).

Probabilistic genetic interaction networking predicted 215 C. elegans orthologues, representing sequence clusters unique to the adult female of Oe. dentatum, to interact directly with a total of 1,729 other genes (range: 1-277), including some (e.g., lin-12, mom-5, glp-1, ppk-1, tbx-2 and rnr-1; Appendices 7.4 and 7.5) that are essential to embryogenesis and reproduction (see www.wormbase.org). The 373 C. elegans orthologues of sequence clusters unique to the adult male of Oe. dentatum were predicted to interact directly with a total of 1,710 other genes (range: 1-117; Appendix 7.5). Amongst these were genes involved in sperm development (i.e., ima-3) and motility (i.e., act-2) (Appendices 7.4 and 7.5; www.wormbase.org). A total number of 387 and 323 C. elegans orthologues of L3- and L4-unique molecules, respectively, were predicted to interact with 790 (range: 1-122; cf. Appendix 7.5) and 1,058 (range: 1-59; Appendix 7.5) other genes, respectively, including some involved in embryonic and/or larval viability (i.e., scc-1, tba-4, cct-3, pfd-3 and mcm-4) and larval development (i.e., let-711) (Appendices 7.4 and 7.5; www.wormbase.org).

The 2,397 predicted peptides unique to the adult female of Oe. dentatum had significant homology (e-value cut-off: >1e-05) to 261 C. elegans
orthologues/homologues (not shown), of which 151 were associated with EC numbers linked to ‘druggable’ enzymes and/or InterPro domains (Table 7.4); of these, 92 were associated with non-wildtype RNAi phenotypes, including adult lethality (n = 3), embryonic and/or larval lethality (n = 44) and/or adult sterility (n = 65). Of the 541 C. elegans homologues of the 7,117 predicted peptides unique to the adult male of Oe. dentatum, 375 were associated with EC numbers linked to ‘druggable’ enzymes and/or InterPro domains (Table 7.4). Of these, 205 were associated with the RNAi phenotypes ‘embryonic and/or larval lethality’ and 196 to ‘sterility’ (Table 7.4). Of the 565 unique C. elegans homologues of predicted peptides unique to the L3 of Oe. dentatum, 375 were associated with EC numbers linked to ‘druggable’ enzymes and/or InterPro domains (Table 7.4); 121 of these were linked to RNAi phenotypes ‘embryonic and/or larval lethality’ and 165 to ‘sterility’ (Table 7.4). Amongst the 416 C. elegans homologues of predicted peptides unique to the L4 stage of Oe. dentatum, 283 could be associated with EC numbers linked to ‘druggable’ enzymes and/or InterPro domains (Table 7.4). Sixty-three of these homologues were associated with RNAi phenotypes ‘embryonic and/or larval lethality’ and 72 to ‘sterility’ (Table 7.4). Examples of ‘druggable’ molecules unique to each of the datasets, together with examples of effective BRENDA compounds, are given in Table 7.4 and Appendix 7.6.

7.4 Discussion

7.4.1 Technical considerations

This chapter demonstrated the utility of an integrated bioinformatic workflow system for the analysis and annotation of large sequence datasets produced by NGS. This system is considered useful for researchers with basic expertise in computer programming but without the means for developing bioinformatic pipelines or purchasing expensive soft- or hardware packages. The system constructed here was appraised according to: (i) computational time required to perform the analyses, (ii) ease of use, (iii) compatibility with different computer operating systems, (iv) ability to focus the analyses on answering relevant biological questions and (v) general applicability.

The majority of the software incorporated in the bioinformatic workflow was derived from existing application tools (e.g., CAP3 = maximum length of 50 kb) available as web-based interfaces, and originally designed for the analysis and annotation of a relatively small number of sequences. These applications were adapted here to face the challenges presented by the need to analyse large sequence datasets in a time-efficient manner. Indeed, the original sequence datasets described herein, which included a total of ~2 million sequences (244 ± 32 bases), could be analysed and
annotated using a 2 CPU Linux computer with 8 processor cores, within ~2,000 computing hours corresponding to ~240 man-hours (one computing hour = one hour of computing time on one processor core). Based on previous experience (Chapters 2-6), the same analyses, conducted using web-based interfaces, require several months to complete. However, an advantage of web-based software tools with extensive graphical interfaces is that no knowledge of computing and/or programming is required [29]. The process of developing, trouble-shooting, maintaining and updating scripts can be involved and challenging, laborious and time-consuming. On the other hand, the use of a command line (which consists of a series of standardized commands) to execute pre-existing scripts, such as the Perl, Python and Unix shell, which have been written and made available here, overcomes this limitation. Furthermore, although these scripts have been written and optimized using the Linux operational system, the output files (generated in the form of text or tab delimited files) can be readily viewed, analysed and modified in a range of different operating systems, such as Microsoft Windows and Mac OS, thus being broadly applicable.

A key goal for scientists focusing on the analyses of large NGS datasets is to distil, from large amounts of raw data, biologically meaningful information about the organism under investigation. For example, some pathogens, such as parasitic worms, have complex life cycles and thus represent a challenging group of organisms for genomic and transcriptomic studies, because different life stages can express various sets of genes which are involved in development, reproduction, host-parasite interactions and/or disease [17,37-39]. Understanding these aspects should have important implications for finding new ways of disrupting biological processes and pathways, and thus could facilitate the prediction and prioritisation of new drug and/or vaccine targets. In addition, compared with the free-living nematode C. elegans, there is a paucity of knowledge on the fundamental molecular biology of parasitic worms [17,39,53]. However, extensive information is available on the functions of C. elegans genes through the use of gene silencing and/or transgenesis (see www.wormbase.org). This knowledge, together with the results of comparative analyses of genetic datasets, revealed that parasitic nematodes usually share ~50-70% of genes with C. elegans [54,55], indicating the utility of this free-living nematode as a model to explore molecular aspects of development, survival and reproduction in some parasitic nematodes [18,38,51,56,57].

7.4.2 Biological interpretations from the annotated dataset

The bioinformatic workflow system constructed here was utilised to explore differential transcription in Oe. dentatum. Several reports indicate that this nematode
provides a unique model system for studying fundamental aspects of the molecular biology of gastrointestinal strongylid nematodes [58]. The in silico subtraction approach identified 139 and 243 protein motifs specific to the adult female and male of *Oe. dentatum*, respectively. Most of these molecules could be linked, using KOBAS analyses and genetic interaction networking, to pathways associated with reproductive processes. For instance, a large number of female-specific molecules encoded proteins containing a ‘chitin-binding protein, peritrophin A’ domain (i.e., n = 18; Table 7.2). This domain was also found to be highly represented amongst the molecules enriched in the female of the pig roundworm, *Ascaris suum* (59). These proteins are hypothesized to have crucial roles in pathways linked to developmental and reproductive processes, based on the knowledge that the corresponding *C. elegans* homologues (containing one or more peritrophin-A domains) CPG-1/CE-1 and CPG-2 are essential for the synthesis of the eggshell as well as for early embryonic development [60]. The production and maturation of oocytes has also been shown, in *C. elegans*, to be regulated by nematode-specific bipartite signalling molecules, the major-sperm proteins (MSPs) [61,62]. Numerous sequences unique to the adult male of *Oe. dentatum* represented MSPs (n = 15; cf. Table 7.2), in accordance with previous studies of male-enriched datasets of other species of strongylid nematodes, including *Trichostrongylus vitrinus* [63], *Haemonchus contortus* [38], as well as the filarioid *Brugia malayi* [64-66], and *As. suum* [59]. Based on the observation that MSPs from various nematodes, including *C. elegans*, are characterised by a significant amino acid sequence conservation (i.e., ~64%; [67]), a similar role has been proposed for these proteins in processes linked to the maturation of oocytes in the uterus of female nematodes [61,62].

In addition to molecules unique to adult female and male of *Oe. dentatum*, the predicted proteins exclusive to the larval stages of this parasite could be linked, using InterPro and/or GO classification and/or probabilistic genetic interaction networking, to biological pathways associated with larval development and/or interactions with the vertebrate host (see Table 7.2). For example, a large number of molecules unique to the L4 stage (n = 10) were inferred to represent proteases. In parasitic nematodes, proteases have been proposed to facilitate the survival of the parasite by mediating, for instance, tissue penetration, feeding and/or immune evasion [68-70]. Indeed, *Oe. dentatum* L4s are known to evoke immunological reactions that result in the encapsulation of the larvae in nodules with aggregations of neutrophils and eosinophils [58,71]. In addition, somatic extracts of and supernatants from in vitro maintenance cultures of *Oe. dentatum* L4s have been shown to induce the proliferation of porcine mononuclear cells in vitro [72]. These observations suggest an active role
for L4-specific proteases in the modulation of the host’s immune response, which (as proposed for other biological systems) could consist of: (i) the direct digestion of antibodies [68]; (ii) cleavage of cell-surface receptors for cytokines [73] and/or (iii) direct lysis of immune cells [74]. In parasitic nematodes, other molecules have been proposed to play immuno-modulatory roles during the invasion of the host, the migration through tissues as well as feeding. Amongst them, proteins containing a ‘sperm-coating protein (SCP)-like extracellular domain’ (InterPro: IPR014044), also called SCP/Tpx-1/Ag5/PR-1/Sc7 (SCP/TAPS; Pfam accession number no. PF001888), were highly represented in the transcriptome of *Oe. dentatum* (see Table 7.2). Members of the SCP/TAPS protein family have been identified in various eukaryotes, including plants, arthropods, snakes, mammals as well as free-living and parasitic helminths [75]. These molecules have been studied mainly in the hookworms *Ancylostoma caninum* and *Necator americanus*, and are commonly referred as to *Ancylostoma*-secreted proteins (i.e., ASPs; [75]). Due to their abundance in the excretory/secretory (ES) products from serum-activated L3s (= aL3s) of *An. caninum* and to the high transcriptional levels of mRNAs encoding ASPs in aL3s compared with non-activated, ensheathed L3s (L3s), these molecules have been hypothesized to play a major role in the transition from the free-living to the parasitic stage of this species [39,76]. Other ASP homologues have been characterised for the adult stage of hookworms, and suggested to play a role in the initiation, establishment and/or maintenance of the host-parasite relationship [39,77,78]. Although a male-biased transcription of ASP homologues had been reported for *Oe. dentatum* [51], results from the present study show that the transcription of SCP/TAPS molecules occurs in all developmental stages studied herein. As the sequences analysed were generated from normalised cDNA libraries, the differences in levels of transcription of genes encoding SCP/TAPS throughout the life cycle of *Oe. dentatum* could not be inferred. Future work could involve, for instance, the application of the present bioinformatic workflow tool to the analysis of ESTs generated (e.g., by Illumina sequencing) from non-normalised cDNA libraries of *Oe. dentatum*, which would allow quantitative rather than qualitative differences in transcription to be determined for genes encoding SCP/TAPS, to assist in the study of the biological function(s) of these molecules [75]. The *Oe. dentatum*-pig model could also provide a useful means of exploring the biological role/s of these molecules in the development and reproduction of this nematode as well as its interactions with the host. Several features of *Oe. dentatum*, including its short life-cycle, its ability to survive and grow in culture *in vitro* for weeks through several moults, and the possibility of rectally transplanting worms (e.g., from *in vitro* culture) into the host without the need for surgical intervention [58,79], offer
an opportunity to experimentally test hypotheses formulated based on the interpretation of results from bioinformatic analyses. Bioinformatically-guided interpretations of NGS datasets are also increasingly playing an important role in the identification of putative drug targets [80], due to the possibility of using predictive algorithms to prioritise and select sets of molecules for experimental studies both in vitro and in vivo [81-83], potentially leading to a significant reduction in the cost associated with drug discovery and development [84]. For instance, in the present study, subsets of molecules without known host (pig) homologues were identified and predicted to represent targets for intervention. Amongst them, protein kinases and phosphatases were the most abundantly represented (Table 7.4). Previously, in *Oe. dentatum*, a catalytic subunit of a serine/threonine protein phosphatase (PP1) was characterized (*Od-mpp1*); gene silencing by RNAi of the corresponding *C. elegans* homologue resulted in a significant reduction (30-40%) in the numbers of F2-progeny produced [56]. Based on these findings, it is tempting to speculate that some pathways, involving phosphatases/kinases, represent key targets for nematocidal drugs.

### 7.4.3 Concluding remarks

This chapter demonstrates, using a large test dataset derived from different stages/sexes of a parasitic worm (*Oe. dentatum*), that the bioinformatic workflow system described herein provides a practical tool for the assembly, annotation and analysis of NGS data. The custom-written Perl, Python and Unix shell computer scripts, accessible via the web, can be readily adapted to suit the requirements of researchers conducting transcriptomic studies in their particular discipline. This workflow system is now routinely used for the analysis of datasets from a range of pathogens of major socio-economic importance and has been applied more broadly to datasets representing other organisms, including mammals. Thus, this integrated system should be a user-friendly and efficient tool for biologists involved in transcriptomic studies in any field on any organism.
7.5 References


17. Chapter 4. The application of next-generation sequencing technology to explore differential transcription between free-living and pre-parasitic third-stage larvae of Haemonchus contortus.


56. Boag PR, Ren P, Newton SE, Gasser RB. Molecular characterisation of a male-specific serine/threonine phosphatase from Oesophagostomum dentatum


75. Chapter 8. A framework for the classification of “SCP/TAPS” proteins and its implications

Table 7.1 Summary of the nucleotide sequence data for the adult female, adult male, and third (L3) and fourth (L4) larval stages of *Oesophagostomum dentatum* prior to and following *in silico* subtraction as well as detailed bioinformatic annotation and analyses.

<table>
<thead>
<tr>
<th>Expression sequence tags (ESTs)</th>
<th>Female</th>
<th>Male</th>
<th>L3</th>
<th>L4</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of unassembled ESTs</td>
<td>336,131</td>
<td>490,645</td>
<td>503,566</td>
<td>496,025</td>
<td>1,826,367</td>
</tr>
<tr>
<td>Contigs (average length ± SD)</td>
<td>23,807 (483 ± 290)</td>
<td>29,043 (484 ± 289)</td>
<td>30,176 (465 ± 281)</td>
<td>26,349 (498 ± 308)</td>
<td>36,233 (516 ± 316)</td>
</tr>
<tr>
<td>Singletons</td>
<td>23,303 (233 ± 50)</td>
<td>37,248 (243 ± 45)</td>
<td>49,341 (227 ± 57)</td>
<td>36,875 (242 ± 40)</td>
<td>452,528 (244 ± 37)</td>
</tr>
<tr>
<td>Total</td>
<td>47,110</td>
<td>66,291</td>
<td>79,517</td>
<td>63,224</td>
<td>488,761</td>
</tr>
<tr>
<td>Containing an Open Reading frame (%)</td>
<td>38,504 (81.7)</td>
<td>52,787 (80)</td>
<td>57,818 (73)</td>
<td>50,533 (80)</td>
<td>85,395 (17.5)</td>
</tr>
<tr>
<td>Returning InterProScan results (%)</td>
<td>20,229 (43)</td>
<td>26,496 (40)</td>
<td>27,297 (47.2)</td>
<td>26,121 (51.7)</td>
<td>56,940 (66.7)</td>
</tr>
<tr>
<td>Gene Ontology (%)</td>
<td>9,970 (25.9)</td>
<td>12,386 (23.5)</td>
<td>12,763 (22.1)</td>
<td>12,735 (25.2)</td>
<td>25,216 (30)</td>
</tr>
<tr>
<td><em>Number of Biological process terms</em></td>
<td>17,031</td>
<td>19,510</td>
<td>19,705</td>
<td>19,645</td>
<td>19,346</td>
</tr>
<tr>
<td>Cellular component</td>
<td>8,864</td>
<td>10,091</td>
<td>10,926</td>
<td>10,649</td>
<td>11,007</td>
</tr>
<tr>
<td>Molecular function</td>
<td>30,482</td>
<td>35,934</td>
<td>34,904</td>
<td>35,241</td>
<td>35,182</td>
</tr>
<tr>
<td>With orthologues in <em>C. elegans</em></td>
<td>23,485 (50)</td>
<td>28,643 (43.2)</td>
<td>32,904 (41.4)</td>
<td>30,000 (47.4)</td>
<td></td>
</tr>
<tr>
<td>other parasitic nematodes (%)</td>
<td>17,533 (37.2)</td>
<td>21,553 (32.5)</td>
<td>23,748 (29.9)</td>
<td>38,634 (61)</td>
<td></td>
</tr>
<tr>
<td>other organisms (%)</td>
<td>12,011 (25.5)</td>
<td>13,843 (21)</td>
<td>14,731 (18.5)</td>
<td>14,332 (22.7)</td>
<td></td>
</tr>
<tr>
<td>KOBAS (number of biological pathways predicted)</td>
<td>256</td>
<td>254</td>
<td>249</td>
<td>255</td>
<td></td>
</tr>
<tr>
<td>In silico subtracted datasets</td>
<td>3,451 (671 + 2,780)</td>
<td>10,344 (2,902 + 7,442)</td>
<td>14,380 (2,752 + 11,628)</td>
<td>7,520 (1,280 + 6,240)</td>
<td></td>
</tr>
<tr>
<td>Number of ESTs (contigs + singletons)</td>
<td>2,397 (70)</td>
<td>7,117 (69)</td>
<td>7,222 (50.2)</td>
<td>4,789 (63.7)</td>
<td></td>
</tr>
<tr>
<td>Containing an Open Reading frame (%)</td>
<td>2,397 (70)</td>
<td>7,117 (69)</td>
<td>7,222 (50.2)</td>
<td>4,789 (63.7)</td>
<td></td>
</tr>
<tr>
<td>Predicted peptides</td>
<td>521 (21.7)</td>
<td>1,179 (16.6)</td>
<td>1,224 (17)</td>
<td>989 (20.7)</td>
<td></td>
</tr>
<tr>
<td>Returning InterProScan results (%)</td>
<td>376 (15.7)</td>
<td>840 (11.8)</td>
<td>760 (10.5)</td>
<td>652 (13.6)</td>
<td></td>
</tr>
<tr>
<td>Gene Ontology (%)</td>
<td>314</td>
<td>625</td>
<td>684</td>
<td>527</td>
<td></td>
</tr>
<tr>
<td><em>Number of Biological process terms</em></td>
<td>177</td>
<td>355</td>
<td>412</td>
<td>359</td>
<td></td>
</tr>
<tr>
<td>Cellular component</td>
<td>563</td>
<td>1,259</td>
<td>1,073</td>
<td>948</td>
<td></td>
</tr>
<tr>
<td>Molecular function</td>
<td>824 (23.9)</td>
<td>1,834 (17.7)</td>
<td>2,252 (15.6)</td>
<td>1,589 (21.1)</td>
<td></td>
</tr>
<tr>
<td>With homologues in <em>C. elegans</em> (%)</td>
<td>558 (16.1)</td>
<td>1,212 (11.7)</td>
<td>1,384 (9.6)</td>
<td>1,052 (14)</td>
<td></td>
</tr>
<tr>
<td>other parasitic nematodes (%)</td>
<td>159 (4.6)</td>
<td>123 (1.2)</td>
<td>176 (1.2)</td>
<td>137 (1.8)</td>
<td></td>
</tr>
<tr>
<td>other organisms (%)</td>
<td>16</td>
<td>18</td>
<td>23</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 7.2 The twenty most represented (InterPro) protein domains inferred from peptides conceptually translated from individual contigs for *Oesophagostomum dentatum* (combined assembly of data for adult female, adult male, and the third [L3] and fourth [L4] larval stages) and InterPro protein domains (level 1) assigned to predicted peptides unique to each stage or sex following *in silico* subtraction.

<table>
<thead>
<tr>
<th>InterPro description</th>
<th>InterPro code</th>
<th>Number of predicted peptides (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Combined assembly (31,982)(^a)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCP-like extracellular</td>
<td>IPR014044</td>
<td>377 (1.2)</td>
</tr>
<tr>
<td>NAD(P)-binding domain</td>
<td>IPR016040</td>
<td>365 (1.1)</td>
</tr>
<tr>
<td>Proteinase inhibitor I2, Kunitz metazoa</td>
<td>IPR002223</td>
<td>339 (1)</td>
</tr>
<tr>
<td>Zinc finger, LIM-type</td>
<td>IPR001781</td>
<td>332 (1)</td>
</tr>
<tr>
<td>WD40 repeat</td>
<td>IPR001680</td>
<td>312 (0.9)</td>
</tr>
<tr>
<td>Ankyrin</td>
<td>IPR002110</td>
<td>257 (0.8)</td>
</tr>
<tr>
<td>EF-HAND 2</td>
<td>IPR018249</td>
<td>247 (0.7)</td>
</tr>
<tr>
<td>WD40 repeat, subgroup</td>
<td>IPR019781</td>
<td>242 (0.7)</td>
</tr>
<tr>
<td>Allergen V5/Tpx-1 related</td>
<td>IPR001283</td>
<td>236 (0.7)</td>
</tr>
<tr>
<td>Protein kinase-like</td>
<td>IPR011009</td>
<td>220 (0.6)</td>
</tr>
<tr>
<td>RNA recognition motif, RNP-1</td>
<td>IPR000504</td>
<td>216 (0.6)</td>
</tr>
<tr>
<td>WD40 repeat 2</td>
<td>IPR019782</td>
<td>215 (0.6)</td>
</tr>
<tr>
<td>Protease inhibitor I4, serpin</td>
<td>IPR000215</td>
<td>207 (0.6)</td>
</tr>
<tr>
<td>Src homology-3 domain</td>
<td>IPR001452</td>
<td>201 (0.6)</td>
</tr>
<tr>
<td>Peptidase C1A, papain C-terminal</td>
<td>IPR000668</td>
<td>194 (0.6)</td>
</tr>
<tr>
<td>C-type lectin</td>
<td>IPR001304</td>
<td>183 (0.5)</td>
</tr>
<tr>
<td>Kelch repeat type 1</td>
<td>IPR006652</td>
<td>183 (0.5)</td>
</tr>
<tr>
<td>Annexin repeat</td>
<td>IPR018502</td>
<td>183 (0.5)</td>
</tr>
<tr>
<td>Protein kinase, core</td>
<td>IPR000719</td>
<td>172 (0.5)</td>
</tr>
<tr>
<td>EF-HAND 1</td>
<td>IPR018247</td>
<td>168 (0.5)</td>
</tr>
<tr>
<td><strong>Female (208)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chitin binding protein, peritrophin-A</td>
<td>IPR002557</td>
<td>18 (8.6)</td>
</tr>
<tr>
<td>Basic-leucine zipper (bZIP) transcription factor</td>
<td>IPR004827</td>
<td>10 (4.8)</td>
</tr>
<tr>
<td>DNA primase, small subunit</td>
<td>IPR002755</td>
<td>6 (2.9)</td>
</tr>
<tr>
<td>p53-like transcription factor, DNA-binding</td>
<td>IPR008967</td>
<td>5 (2.4)</td>
</tr>
<tr>
<td>DNA-binding HORMA</td>
<td>IPR003511</td>
<td>4 (2)</td>
</tr>
<tr>
<td>Acyl-CoA dehydrogenase/oxidase</td>
<td>IPR013786</td>
<td>3 (1.4)</td>
</tr>
<tr>
<td>Frizzled-like domain</td>
<td>IPR020067</td>
<td>3 (1.4)</td>
</tr>
<tr>
<td>Lipid transport protein</td>
<td>IPR001747</td>
<td>3 (1.4)</td>
</tr>
<tr>
<td>PreATP-grasp-like fold</td>
<td>IPR016185</td>
<td>3 (1.4)</td>
</tr>
<tr>
<td>UbiA prenyltransferase</td>
<td>IPR000537</td>
<td>3 (1.4)</td>
</tr>
<tr>
<td><strong>Male (402)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PapD-like</td>
<td>IPR008962</td>
<td>16 (4)</td>
</tr>
<tr>
<td>Major sperm protein</td>
<td>IPR000535</td>
<td>15 (3.7)</td>
</tr>
<tr>
<td>C-type lectin</td>
<td>IPR018378</td>
<td>6 (1.5)</td>
</tr>
<tr>
<td>Phosphoenolpyruvate carboxykinase</td>
<td>IPR008209</td>
<td>6 (1.5)</td>
</tr>
<tr>
<td>Protein of unknown function DUF236</td>
<td>IPR004296</td>
<td>6 (1.5)</td>
</tr>
<tr>
<td>Scramblase</td>
<td>IPR005552</td>
<td>6 (1.5)</td>
</tr>
<tr>
<td>ClpX, ATPase regulatory subunit</td>
<td>IPR004487</td>
<td>5 (1.3)</td>
</tr>
<tr>
<td>Galactose oxidase/kelch</td>
<td>IPR011043</td>
<td>5 (1.3)</td>
</tr>
<tr>
<td>Ribosomal protein S2</td>
<td>IPR001865</td>
<td>5 (1.3)</td>
</tr>
<tr>
<td>Amidinotransferase</td>
<td>IPR003198</td>
<td>4 (1)</td>
</tr>
<tr>
<td><strong>L3 (374)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RmlC-like jelly roll fold</td>
<td>IPR014710</td>
<td>17 (4.5)</td>
</tr>
<tr>
<td>Six-bladed beta-propeller, TolB-like</td>
<td>IPR011042</td>
<td>10 (2.7)</td>
</tr>
<tr>
<td>Protein of unknown function DUF590</td>
<td>IPR007632</td>
<td>9 (2.4)</td>
</tr>
<tr>
<td>7TM GPCR, serpentine receptor class r (Str),</td>
<td>IPR019428</td>
<td>8 (2.1)</td>
</tr>
<tr>
<td>Category</td>
<td>Domain Description</td>
<td>InterPro ID</td>
</tr>
<tr>
<td>----------</td>
<td>--------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Nematode</td>
<td>Acyltransferase ChoActase/COT/CPT</td>
<td>IPR000542</td>
</tr>
<tr>
<td></td>
<td>Putative DNA binding</td>
<td>IPR009061</td>
</tr>
<tr>
<td></td>
<td>7TM GPCR, serpentine receptor class e (Sre),</td>
<td>IPR004151</td>
</tr>
<tr>
<td>Nematode</td>
<td>Nuclear hormone receptor, ligand-binding, core</td>
<td>IPR000536</td>
</tr>
<tr>
<td></td>
<td>Coenzyme A transferase</td>
<td>IPR004165</td>
</tr>
<tr>
<td></td>
<td>Ion transport</td>
<td>IPR005821</td>
</tr>
<tr>
<td><strong>L4 (320)</strong></td>
<td>Peptidase M24, methionine aminopeptidase</td>
<td>IPR001714</td>
</tr>
<tr>
<td></td>
<td>FAD-binding, type 2</td>
<td>IPR016166</td>
</tr>
<tr>
<td></td>
<td>Oxysterol-binding protein</td>
<td>IPR000648</td>
</tr>
<tr>
<td></td>
<td>Translation protein SH3-like</td>
<td>IPR008991</td>
</tr>
<tr>
<td></td>
<td>Tubulin/FtsZ, GTPase domain</td>
<td>IPR003008</td>
</tr>
<tr>
<td></td>
<td>6-phosphogluconate dehydrogenase</td>
<td>IPR008927</td>
</tr>
<tr>
<td></td>
<td>Peptidase C13, legumain</td>
<td>IPR001096</td>
</tr>
<tr>
<td></td>
<td>Aminoacyl-tRNA synthetase</td>
<td>IPR015413</td>
</tr>
<tr>
<td></td>
<td>Adenosylcobalamin biosynthesis, ATP</td>
<td>IPR016030</td>
</tr>
<tr>
<td></td>
<td>Aspartate/other aminotransferase</td>
<td>IPR000796</td>
</tr>
</tbody>
</table>

*Total number of predicted peptides assigned to InterPro domains in each dataset.*
Table 7.3 Functions predicted for proteins encoded in the transcriptome of *Oesophagostomum dentatum* (combined assembly), based on gene ontology (GO). The parental (= level 2) GO categories were assigned according to (InterPro) domains inferred from proteins with homology to functionally annotated molecules.

<table>
<thead>
<tr>
<th>GO description (GO code)</th>
<th>Number of predicted peptides (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Biological process (19,346)</strong></td>
<td></td>
</tr>
<tr>
<td>Metabolic process (GO:0008152)</td>
<td>2,102 (10.9)</td>
</tr>
<tr>
<td>Proteolysis (GO:0006508)</td>
<td>1,361 (7)</td>
</tr>
<tr>
<td>Translation (GO:0006412)</td>
<td>1,033 (5.4)</td>
</tr>
<tr>
<td>Transport (GO:0006810)</td>
<td>816 (4.2)</td>
</tr>
<tr>
<td>Protein amino acid phosphorylation (GO:0006468)</td>
<td>763 (4)</td>
</tr>
<tr>
<td><strong>Cellular component (11,007)</strong></td>
<td></td>
</tr>
<tr>
<td>Intracellular (GO:0005622)</td>
<td>1,925 (17.5)</td>
</tr>
<tr>
<td>Membrane (GO:0016020)</td>
<td>1,717 (15.6)</td>
</tr>
<tr>
<td>Nucleus (GO:0005634)</td>
<td>1,279 (11.6)</td>
</tr>
<tr>
<td>Integral to membrane (GO:0016021)</td>
<td>1,159 (10.5)</td>
</tr>
<tr>
<td>Ribosome (GO:0005840)</td>
<td>736 (6.7)</td>
</tr>
<tr>
<td><strong>Molecular function (35,182)</strong></td>
<td></td>
</tr>
<tr>
<td>ATP binding (GO:0005524)</td>
<td>2,645 (7.5)</td>
</tr>
<tr>
<td>Catalytic activity (GO:0003824)</td>
<td>2,449 (7)</td>
</tr>
<tr>
<td>Binding (GO:0005488)</td>
<td>1,622 (4.6)</td>
</tr>
<tr>
<td>Zinc ion binding (GO:0008270)</td>
<td>1,229 (3.5)</td>
</tr>
<tr>
<td>Oxidoreductase activity (GO:0016491)</td>
<td>1,226 (3.5)</td>
</tr>
<tr>
<td>Protein binding (GO:0005515)</td>
<td>1,206 (3.4)</td>
</tr>
<tr>
<td>Nucleic acid binding (GO:0003676)</td>
<td>919 (2.6)</td>
</tr>
<tr>
<td>DNA binding (GO:0003677)</td>
<td>788 (2.2)</td>
</tr>
<tr>
<td>Structural constituent of ribosome (GO:0003735)</td>
<td>755 (2.1)</td>
</tr>
<tr>
<td>Nucleotide binding (GO:0000166)</td>
<td>717 (2)</td>
</tr>
</tbody>
</table>

*Total number of GO terms assigned to predicted peptides.*
Table 7.4 Examples of *Caenorhabditis elegans* orthologues of contigs unique to each *Oesophagostomum dentatum* adult female, adult male, and the third (L3) and fourth (L4) larval stages, following *in silico* subtraction, ranked according to the 'severity' of the RNAi phenotype/s observed, and for which inferred peptides were associated with 'druggable' (InterPro) domains and/or Enzyme Commission (EC) numbers as well as examples of candidate compounds linked to these domains, predicted using the BRENDA database. The number of the *C. elegans* orthologues predicted to interact with each of the molecules listed is also indicated.

<table>
<thead>
<tr>
<th>Contig code</th>
<th><em>C. elegans</em> gene ID</th>
<th>Gene name</th>
<th>RNAi phenotypes</th>
<th>Protein description</th>
<th>Druggable IPR domain (description)</th>
<th>Examples of BRENDA compounds</th>
<th>No. of predicted interacting genes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Female (151)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contig722</td>
<td>T23G5.1</td>
<td><em>mr-1</em></td>
<td>Embryonic lethal, embryonic defects, larval lethal, larval arrest, sterile</td>
<td>Ribonucleotide reductase</td>
<td>IPR000788 (Ribonucleotide)</td>
<td>D-phosphoserine</td>
<td>35</td>
</tr>
<tr>
<td>Contig18241</td>
<td>F44F4.2</td>
<td><em>Egg-3</em></td>
<td>Embryonic lethal, maternal sterile, sterile progeny</td>
<td>Protein tyrosine phosphatase</td>
<td>IPR000242 (Protein tyrosine)</td>
<td>4-nitrophenyl phosphate</td>
<td>-</td>
</tr>
<tr>
<td>Contig15526</td>
<td>T21E3.1</td>
<td><em>Egg-4</em></td>
<td>Embryonic lethal, maternal sterile</td>
<td>Protein tyrosine phosphatase</td>
<td>IPR000242 (Protein tyrosine)</td>
<td>4-nitrophenyl phosphate</td>
<td>-</td>
</tr>
<tr>
<td>Contig10671</td>
<td>Y110A7A.4</td>
<td>Embryonic lethal, reduced brood size</td>
<td>Thymidylate synthase</td>
<td>IPR000398 (Thymidylate)</td>
<td>5,10- methylentetrahydrofolate + deoxyuridine phosphate</td>
<td>3',5'-cAMP + diphosphate</td>
<td>26</td>
</tr>
<tr>
<td>E6SSEER01EX2TA</td>
<td>F17C8.1</td>
<td><em>acy-1</em></td>
<td>Embryonic defects, Larval arrest</td>
<td>Adenyl cyclase</td>
<td>IPR001054 (Adenyl)</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td><strong>Male (375)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contig12350</td>
<td>W03A5.1</td>
<td>Embryonic lethal, embryonic defects</td>
<td>Fibroblast/platelet-derived growth factor receptor and related receptor tyrosine kinase</td>
<td>IPR001254 (Serine proteases)</td>
<td>Cleaved azocasein</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>Contig10801</td>
<td>T04B2.2</td>
<td><em>frk-1</em></td>
<td>Embryonic lethal, embryonic defects</td>
<td>Protein tyrosine kinase</td>
<td>IPR001245 (Tyrosine protein kinase)</td>
<td>ADP + a phosphoprotein</td>
<td>-</td>
</tr>
<tr>
<td>Contig13376</td>
<td>T04B2.2</td>
<td><em>frk-1</em></td>
<td>Embryonic lethal, embryonic defects</td>
<td>Protein tyrosine kinase</td>
<td>IPR001245 (Tyrosine protein kinase)</td>
<td>ADP + a phosphoprotein</td>
<td>-</td>
</tr>
<tr>
<td>Contig</td>
<td>ID</td>
<td>Description</td>
<td>Function</td>
<td>Gene Information</td>
<td>Reaction Products</td>
<td>Direct Reads</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
<td>-----------------------------</td>
<td>--------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------</td>
<td>---------------------------------------------</td>
<td>--------------</td>
<td></td>
</tr>
<tr>
<td>Contig10782</td>
<td>ZK354.6</td>
<td>Embryonic defects</td>
<td>Casein kinase</td>
<td>IPR001245 (Tyrosine protein kinase)</td>
<td>ADP + a phosphoprotein</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Contig13084</td>
<td>C25A8.5</td>
<td>Aldicarb resistant</td>
<td>Protein tyrosine kinase</td>
<td>IPR001254 (Serine proteases)</td>
<td>Cleaved azocasein</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Contig10987</td>
<td>T04D3.4</td>
<td>gcy-35</td>
<td>Embryonic lethal, larval arrest</td>
<td>IPR001054 (Guanylate cyclase)</td>
<td>3',5'-cAMP + diphosphate</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Contig17117</td>
<td>B0240.3</td>
<td>daf-11</td>
<td>Embryonic lethal, slow growth</td>
<td>IPR001054 (Guanylate cyclase)</td>
<td>3',5'-cAMP + diphosphate</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>Contig10518</td>
<td>R01E6.1b</td>
<td>odr-1</td>
<td>Guanyle cyclase</td>
<td>IPR001054 (Guanylate cyclase)</td>
<td>3',5'-cAMP + diphosphate</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Contig10600</td>
<td>C24G6.2b</td>
<td>Slow growth</td>
<td>Fibroblast/platelet-derived growth factor receptor and related receptor tyrosine kinase</td>
<td>IPR11009 (protein kinase)</td>
<td>Cleaved azocasein</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Contig1406</td>
<td>R134.2</td>
<td>gcy-2</td>
<td>Slow growth</td>
<td>IPR001054 (Guanylate cyclase)</td>
<td>3',5'-cAMP + diphosphate</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Contig11765</td>
<td>Y46H3A.1</td>
<td>srt-42</td>
<td>Extended life span</td>
<td>IPR11009 (protein kinase)</td>
<td>ADP + a phosphoprotein</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Contig23920</td>
<td>T05G5.3</td>
<td>Embryonic lethal, maternal sterile arrest</td>
<td>Protein kinase PCTAIRE and related kinases</td>
<td>IPR000719 (protein kinase)</td>
<td>ADP + a phosphoprotein</td>
<td>139</td>
<td></td>
</tr>
<tr>
<td>Contig1501</td>
<td>K12D12.1</td>
<td>top-2</td>
<td>Embryonic lethal, maternal sterile arrest</td>
<td>IPR002205 (DNA girase)</td>
<td>Catenated DNA networks + ADP + phosphate</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>Contig2892</td>
<td>C46A5.4</td>
<td>Protruding vulva</td>
<td>DNA topoisomerase type II</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contig20741</td>
<td>C46A5.4</td>
<td>Protruding vulva</td>
<td>IPR002007 (Animal haem peroxidase)</td>
<td>2-amino-9,10a-dihydro-3H-phenoxazin-3-one</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contig25779</td>
<td>R11A5.7</td>
<td>Dumpy</td>
<td>Zinc carboxypeptidase</td>
<td>IPR000834 (Zinc carboxypeptidases)</td>
<td>4-chlorocinnamic acid + L-beta-phenyllactate</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>
Fig. 7.1 Bioinformatic analyses of the *Oesophagostomum dentatum* dataset. Stars (★) indicate analyses performed using custom-written Perl, Python and/or Unix shell scripts, accessible via http://research.vet.unimelb.edu.au/gasserlab/index.html. (1) Individual and combined expressed sequence tags (EST) datasets are assembled using CAP3 (compiled Linux 64-bit executable) to generate consensus sequences. (2) Assembled contigs with high similarity (e-value cut-off: <1e-15) to nucleotide sequences of the vertebrate host (*Sus scrofa*) are eliminated. (3) Database similarity searches (for individual or combined datasets) are carried out using BLASTn and BLASTx (compiled Linux 64-bit executable; [42]), embedded in custom-built Unix shell scripts. (4) Sequences (from the individually and combined assembled datasets) are conceptually translated into peptide sequences using ESTScan (compiled Linux 64-bit executable with a Perl wrapper). (5) Domains/motifs within translated peptides are identified via InterProScan (Perl wrapper) and linked to biological pathways in *Caenorhabditis*.
*elegans* using KOBAS (stand-alone Python application; [44]). Functional annotation of the predicted peptides is performed by gene ontology (Perl wrapper; [27]). (6) The individually assembled datasets are subtracted from one another (in both directions) using a BLASTn algorithm [42] embedded in a custom-built Unix shell script; proteins inferred from subtracted transcripts are assigned parental (i.e., level 1) InterPro terms and subtracted from one another using a BLASTp algorithm, embedded in a custom-built Unix shell script. (7) Potential drug target candidates for each of the individually assembled and/or in silico subtracted datasets are predicted and ranked according to the 'severity' of the non-wildtype RNAi phenotypes observed for the corresponding *C. elegans* orthologues/homologues (custom-built Unix shell scripts). (8) Probabilistic interaction networks among *C. elegans* orthologues of subtracted molecules are predicted (command lines).
Fig. 7.2 Length distribution of contigs after the assembly of combined sequence data for adult female, adult male, and the third (L3) and fourth (L4) larval stages of *Oesophagostomum dentatum.*
Fig. 7.3 Mapping of inferred peptides unique to either adult female, adult male, and the third (L3) and fourth (L4) larval stages of *Oesophagostomum dentatum* (following *in silico* subtraction) to known biological pathways in *Caenorhabditis elegans* using the iPath tool. KEGG codes linked to predicted peptides unique to either female, male, L3 or L4 are also given.
CHAPTER 8
A framework for the classification of “SCP/TAPS” proteins and its implications

Abstract
A wide range of proteins belonging to the SCP/TAPS “family” has been described for various eukaryotic organisms, including plants and animals (vertebrates and invertebrates, such as helminths). Although SCP/TAPS proteins have been proposed to play key roles in a number of fundamental biological processes, such as host-pathogen interactions and defence mechanisms, there is a paucity of information on their genetic relationships, structures and functions, and there is no standardised nomenclature for these proteins. A detailed analysis of the relationships of members of the SCP/TAPS family of proteins, based on key protein signatures, could provide a foundation for investigating these areas. This chapter appraises the current state of knowledge of key SCP/TAPS proteins of eukaryotes, with an emphasis on those from parasitic helminths, undertakes a comprehensive, systematic phylogenetic analysis of currently available full-length protein sequence data (considering characteristic protein signatures or motifs) to infer relationships and provides a framework (based on statistical support) for the naming of these proteins. This framework is intended to guide genomic and molecular biological explorations of key SCP/TAPS molecules associated with infectious diseases of plants and animals. In particular, fundamental investigations of these molecules in parasites and the integration of structural and functional data could lead to new and innovative approaches for the control of parasitic diseases, with important biotechnological outcomes.
8.1 Introduction

Parasites have evolved a range of strategies to invade their hosts, whilst hosts have developed immune and other defence mechanisms against pathogens (Chapter 1). Numerous host-pathogen interactions involve molecular processes, in which proteins and other molecules from a pathogen target host proteins to initiate or maintain the infection, stimulate or evade the host’s immune response and/or cause disease [1]. Various groups of proteins have been proposed to play major biological roles in the host-pathogen interplay [2]. Among these molecules are the sperm-coating protein (SCP)-like extracellular proteins, also called SCP/Tpx-1/Ag5/PR-1/Sc7 (SCP/TAPS; Pfam accession number no. PF00188). SCP/TAPS family members have been identified in various nematodes (Chapters 2-7) and other eukaryotes. SCP/TAPS proteins, which belong to the cysteine-rich secretory protein (CRISP) “superfamily” [3], include rodent sperm-coating glycoproteins (or acidic epididymal glycoproteins, proposed to be involved in sperm maturation during its passage through the epididymis) [4], mammalian testis-specific protein (Tpx-1) [5], glioma pathogenesis-related protein [6-8], venom allergen 5 from vespid wasps and the venom allergen 3 from fire ants, which mediate allergic reactions to the bites by some insects of the order Hymenoptera [9] as well as plant pathogenesis proteins (PRPs) of the PR-1 “subfamily” which are synthesized in response to infections with pathogens or other stress-inducing factors [10]. Data for a range of eukaryotes suggest that all SCP/TAPS molecules share a common primary structure; the signal peptide is followed by the SCP-extracellular domain (InterProScan: IPR014044), which is proposed to act as a Ca²⁺-chelator in various signalling processes [11]. In the SCP/TAPS proteins of yeast, the SCP-domain is variable in length, polymorphic and flanked by a threonine-rich region. The C-terminus of family members representing mammals and reptile venoms harbours the CRISP-domain, typified by ten conserved cysteine residues. In invertebrates, such as parasitic helminths (including flatworms and roundworms), this portion of the protein is smaller and contains only four to six cysteine residues [12].

SCP/TAPS homologues have been identified in a range of invertebrates, particularly roundworms (nematodes), flatworms (trematodes), arthropods, and plants. Despite the roles that SCP/TAPS proteins are proposed to play in fundamental biological processes in various eukaryotes [10,13], attempts to define criteria for the unequivocal classification of these molecules are scant and limited to SCP/TAPS proteins of the vinegar fly, Drosophila melanogaster [14], the blood fluke Schistosoma mansoni [3] and plants [15]. To date, 17 different “subfamilies” of plant PRPs have been proposed and named in the order in which homologues with defined biological and biochemical properties were discovered [10]. In parasitic nematodes, Ancylostoma-
secreted proteins or activation-associated proteins (ASPs) were first characterized from hookworms and subsequently from related strongyloid nematodes [16]. As ASPs are abundant in the excretory/secretory (ES) products of the infective third-stage larvae (L3), they are thought to play an important role in the transition from the free-living to the parasitic stage during the invasion of the host [17-20]. Because of their immunogenic properties, one ASP (namely Na-ASP-2) is under investigation as a vaccine candidate against the disease (= necatoriasis) caused by the hookworm Necator americanus in humans [21-24]. Attempts to classify ASPs have been based on the number and features of the SCP-domain of these molecules. Thus far, three types of ASPs have been described: ‘double domain ASPs’ which have two distinct but related SCP-domains, and the ‘C-type single domain ASPs’ and ‘N-type single domain ASPs’ which have the highest homology to the C- and N-terminus of the SCP-double domain ASPs, respectively [25]. C-type single domain and double-domain ASPs have been identified in a range of nematodes parasitic in animals and plants [16]. To date, N-type single domain ASPs have been characterized only for Cooperia punctata, Ostertagia ostertagi and Teladorsagia circumcincta, three gastrointestinal trichostrongyloid nematodes of ruminants [12,25,26].

Usually, new SCP/TAPS gene orthologues or protein homologues identified in parasitic nematodes (based on nucleotide and/or amino acid sequence identities or homologies) are named without consideration of the SCP/TAPS protein group as a whole. Also, names are sometimes assigned to SCP/TAPS gene orthologues/protein homologues for which only partial sequence data are provided or available, which has resulted in an unclear relationship among gene(s), transcript(s) and protein(s). The confusion and inconsistencies in classification are exacerbated, as different authors use different names for the same groups of molecules. Besides their usual designation as ASPs, SCP/TAPS molecules of parasitic nematodes have also been named ‘ASP-like’ (AL), based on their homology to ASPs, identified first in Ancylostoma [27], and ‘venom allergen-like’ (VAL; [3,28,29]), ‘venom allergen-homologues’ (VAH; [30,31]) and/or ‘venom allergen-proteins’ (VAP; [32]), based on the homology with the major allergen in the venom of the yellow jacket wasp [33]. In the free-living nematode Caenorhabditis elegans, at least 17 different ASP-related genes have been identified [34] and given various names (e.g., vap or scl = ‘SCP-like’). In C. elegans, SCP/TAPS molecules have been shown to be involved in biological aspects, such as anti-microbial activity [35], normal body formation [36] and lifespan [37].

Although SCP/TAPS proteins are considered by most authors to be of major biological importance in a wide range of eukaryotes (plants, vertebrates and invertebrates), there are inconsistencies in classification as well as the lack of
structural and functional information for most of them. The present chapter (i)
appraises the current state of knowledge of these molecules, with an emphasis on those
from parasitic helminths, (ii) provides a comprehensive, systematic phylogenetic
analysis of current protein sequence data (considering characteristic
signatures/motifs) to infer relationships and (iii) proposes a framework (based on
statistical support) for the naming of these proteins. This framework is intended to
support genomic and molecular biological explorations of key SCP/TAPS molecules
associated with infectious diseases in plants and animals. In particular, fundamental
investigations of these molecules in parasites, using novel genomic, proteomic and
bioinformatic technologies, and the integration of structural and functional data could
underpin new approaches for the control of infectious diseases, with significant
biotechnological outcomes.

8.2 SCP/TAPS proteins in plants: defence-related proteins

In response to infections by pathogens and/or exposure to other stress-
inducing factors, many plants accumulate a variety of pathogenesis-related proteins
(PRPs), which are involved in antimicrobial activity and other defence-related
mechanisms, but are not expressed in healthy plant tissues. PRPs are proposed to be
present in all monocotyledons and dicotyledonous plants and have been classified
into 17 “sub-families” [10]. The significance of inducible defence-related proteins in
pathogen-infected plants has been reviewed in detail elsewhere [10] and is thus not
covered in the present section. The main focus here is on selected members of the PR-1
subfamily, the dominant group of PRPs, induced by pathogens [38] and whose function
is the least understood. The PR-1 proteins show significant sequence similarity to
SCP/TAPS from mammals, arthropods and parasitic helminths; although their
functions are mostly unknown, their occurrence in this diverse range of organisms is
suggestive of roles extending beyond those proposed to be associated with plant
defence [39].

PR-1 proteins were first discovered in 1970 in tobacco mosaic virus (TMV)-
infected tobacco (Nicotiana tabacum; family Solanaceae) [40]. Three acidic (PR-1a, -1b
and -1c; pi ~ 4.5) and one basic (PR-1g; pi ~10.0) PR-1 proteins, with differing
biological properties, were identified and named in order of decreasing electrophoretic
mobility in non-denaturing polyacrylamide gels. Similarly, van Loon et al. [41]
proposed that any protein whose amino acid sequence resembled a characterised PR-1
from tobacco should be designated by assigning a letter to any new members within a
family, according to the order in which they had been identified. Using biochemical and
immunological tools, PR-1a, -1b and -1c were first detected in the extracellular spaces
and xylem elements of TMV-infected tobacco leaves [42]. Subsequently, Carr et al. [43] described the accumulation of these proteins in central vacuoles (= crystal idioblasts) in specialized leaf cells following the experimental TMV-infection of N. tabacum. Since tobacco PR-1 genes do not encode a known peptide sequence for vacuolar targeting, the synthesis and accumulation of acidic PR-1 proteins within crystal idioblasts in cells appear to be unique [44]. In other Solanaceae, extracellular proteins corresponding to the acidic PR-1 from tobacco are basic. For example, in tomato (Lycopersicon esculentum), basic protein homologues of acidic PR-1a, -1b and -1c were detected in the apoplastic fluid as well as in leaf homogenates infected with mycetes of Cladosporium fulvum, the causative agent of the tomato leaf mould [45]. In tomato, the conformation of PR-1b was determined by nuclear magnetic resonance; its unique molecular architecture includes four α-helices and four β-strands arranged in an antiparallel fashion between helices. The tight packing of the α-helices on both sides of the central β-sheet (α–β–α “sandwich” structure) results in a compact, bipartite molecular core, which is stabilized by hydrophobic interactions and multiple hydrogen bonds [11]. In potato (Solanum tuberosum), a direct inhibitory effect of basic PR-1 family members on fungal pathogens (i.e., Phytophthora infestans and Uromyces fabae) has been demonstrated in vitro and in vivo [46,47], but the precise mode of action as well as the cellular and molecular targets of PR-1 proteins remain unknown.

Because of difficulties in studying complex functional aspects in higher plants (such as tobacco and tomato), Arabidopsis thaliana (family Brassicaceae) has been used as a model to explore the mechanisms of the hypersensitivity response against pathogens [48] and transcriptional events linked to PR-1 in response to Botrytis cinerea infection [49]. In rice (Oryza sativa; family Poaceae), homologues of the acidic tobacco PR-1 are encoded by a multigene family whose transcription has been demonstrated to be enhanced in response to cutting, the administration of jasmonic acid, other phytohormones or protein phosphatase inhibitors together with the protein kinase inhibitor staurosporine [50]. Although the mode of action of these molecules is unclear, it has been proposed that the expression of one of the genes encoding a PR-1 protein (i.e., OsPR1a) in rice is regulated by an interaction(s) of phytohormones with light signals, which is most likely mediated via one or more protein phosphorylation/dephosphorylation events [51]. Although the roles of the PR-1 proteins are enigmatic in many cases, current sequence data indicate some conservation across invertebrates and vertebrates. Therefore, the occurrence of PR-1 protein homologues in different groups of organisms suggests that these proteins share an evolutionary origin and play crucial roles in the biology, function and survival of organisms [10].
8.3 SCP/TAPS proteins of animals: from insect allergens to cancer cells

In animals, SCP/TAPS molecules represent major components of the venom of a range of biting insects of the orders Hymenoptera and Diptera [33,52,53]. Insect venom allergy is a harmful reaction caused by stinging insects, such as bees, yellow jackets, wasps, hornets and fire ants [54]. The first exposure to insect venom does not usually result in allergic reactions, but subsequent exposures can cause sensitization and disease. Clinical signs can vary from pruritus, affecting any part of the skin, to anaphylactic shock, which can occur within minutes and result in death [55]. Individuals allergic to insect stings often show similar symptoms against a range of different insect species, as a result of sensitization to individual antigens or cross-reactivity with antigens with similar structural and biochemical features [55]. Among the major components of vespid venom with allergenic properties, Ves v 5 is a hyaluronidase which is related, based on amino acid sequence identity, to SCP/TAPS proteins in other arthropods, mammals, reptiles, fungi and plants [33]. Ves v 5 was first isolated from the venom of different species of yellow jackets, hornets and paper wasps; its crystal structure shows hydrogen bonding patterns and hydrophobic interactions which define an α-β-α core structure [33]. Homologues of Ves v 5 are predicted to have the same conformation due to a high degree of sequence similarity. Although it has been proposed that the Ves v 5 protein in the yellow jacket might act as trypsin inhibitor, via the blocking of calcium-dependent ryanodine receptors, their biochemical role in inducing venom insect allergy remains enigmatic [33].

SCP/TAPS proteins are also components of the venom from fire ants (Solenopsis spp.) and the saliva of a range of blood-feeding insects, including mosquitoes (Anopheles gambiae and Aedes aegypti), sandfly (Lutzomyia longipalpis) and the tsetse fly (Glossina morsitans morsitans) [52,53,56-58]. Given their abundance in the sialotranscriptome of adult female A. gambiae and L. longipalpis, homologues of Ves v 5 have been proposed to be associated with the physiology of the secretory process (e.g., antihaemostatic activity, “stabilizers” of the other secretory products, or assist in lubricating the insect mouth parts), thus explaining its ubiquity and relative sequence conservation [52,57]. In G. morsitans, SCP/TAPS homologues have been shown to be expressed primarily in salivary glands but could also be detected in midgut and proventriculus; therefore, the authors hypothesized a role for these proteins in the mediation of immune reactions in the gut of tsetse flies [53]. In Drosophila melanogaster, the SCP/TAPS gene family includes 26 molecules belonging to two distinct groups based on conserved positioning of introns and sequence similarities in the SCP-domain in the genes [14]. Seventy per cent of the D. melanogaster SCP/TAPS
genes have been shown to be expressed specifically in the testis (semenal vesicles and/or primary spermatocytes), thus suggesting a central role in reproduction, either by mediating interactions between germ cells and somatic cells within the male reproductive tract or between sperm and egg in the female fly [14,59].

SCP/TAPS proteins are significant components of snake venom. They belong to a group of toxins affecting smooth muscle contraction and another group affecting cyclic nucleotide-gated (CNG) ion channels [60]. Within the first group, tigrin (isolated from the venom of the Yamakagashi snake, Rhabdopsis tigrinus tigrinus), ablofin (from the Japanese mamushi snake, Agkistrodon blomhoffii), triflin (from the habu snake, Trimeresurus flavoviridis) and latismeshin (from the erabu sea snake, Laticaude semifasciata) share ~50% sequence homology to mammalian SCP/TAPS proteins, with all cysteine residues being conserved [60]. Tigrin is a helolkeeper with ion channel-blocking activity [61], whereas ablofin, latismeshin and triflin, whose crystal structure was recently determined [62], block the depolarization of the smooth muscle membrane [60]. Within the second group of the CNG ion channel blockers, pseudochetoxin (PsTx) and pseudecin have been identified in the venom of the Australian king brown snake (Pseudechis australis) and the red-bellied black snake (Pseudechis porphyriacus), respectively [63,64]. Although these latter molecules are SCP/TAPS proteins, their biological properties and activities differ substantially from other homologues. Indeed, PsTx and pseudecin are basic proteins (p1 ~10.0) that block olfactory and retinal a-subunit homotetrameric channels when applied to the exterior surface of membrane patches [60]. Studies have shown that SCP/TAPS proteins are widespread in the venoms from snakes (including those of the families Viperidae, Elapidae and Colubridae) from all five continents and Pacific and Indian Oceans [65]; although the mechanisms of action of some of these proteins have been elucidated [60], the functions of most homologues are unknown. Another SCP/TAPS protein belonging to the second group has been isolated from the venom of the cone snail Conus textile and shown to act as a substrate-specific protease in the processing of pro-conotoxins [66].

Members of the SCP/TAPS protein family have also been identified in vertebrates, such as rat (Rattus norvegicus) [67], mouse (Mus musculus) [68], rhesus monkey (Macaca mulatta) [69] and horse (Equus caballus) [70]. Due to their high level of expression in the epididymal lumen, they have been proposed to be associated with the binding of sperm to the oocyte and the fertilization process [4]. In humans, the SCP/TAPS family also includes proteins which are expressed specifically in the testes and likely to be involved in sperm maturation [5]. However, a human SCP/TAPS member (i.e., 'glioma pathogenesis-related protein'- GliPR, also called 'trypsin
inhibitor’ - P25TI and ‘related to testes-specific, vespid and pathogenesis protein 1’-RTVP-1) has also been isolated from malignant glioma cells. Its high level of expression in glioblastoma multiforme/astrocytoma has been linked to the degree of malignancy of astrocytic tumours and to the regulation of the growth, survival and malignancy of the cancer [6-8,13]. Another member (i.e., ‘glioma pathogenesis related 2’- GLIPR-2, also known as ‘Golgi associated PR-1 protein’- GAPR-1 and ‘Homo sapiens chromosome 9 open reading frame 19’- C9orf19) has been shown to be highly expressed in leukocytes, monocytes, lung, spleen and embryonic tissues [71]. The highest level of expression of GLIPR-1 is in fibrotic kidney cells [71-73]; this latter molecule has been proposed to play a role in the development of fibrosis, due to its ability to stimulate (via the recruitment of activated fibroblasts) the transformation/transition of renal epithelial to mesenchymal cells in vitro [73].

8.4 SCP/TAPS from helminths: ‘activation- associated secreted proteins’

SCP/TAPS gene orthologues and/or protein homologues are common in helminths of the phylum Nematoda and class Trematoda. Within the Nematoda, SCP/TAPS have been described in the orders Spirurida, Ascaridida, TylENCHIDA, Strongylida and Rhabditida [74], which, based on an analysis of sequence data from the small subunit (SSU) of nuclear ribosomal DNA (rDNA) [75], have been proposed to correspond to clades III, IV and V. In most cases, SCP/TAPS molecules of parasitic helminths have been referred to commonly as ASps.

8.4.1 Nematodes within clade III (including the orders Spirurida and Ascaridida)

In nematodes of clade III, three C-type single-domain ASps of Onchocerca volvulus (Spirurida; a filarial parasite of humans) have been identified and named Ov-ASP-1, Ov-ASP-2 and Ov-ASP-3 [76,77]. Each of these three proteins has been shown to display a distinct transcription pattern during the development of the parasite; specifically, the transcript Ov-asp-2 was detected in all developmental stages, Ov-asp-3 was specific to the third-stage larvae (L3s), whereas Ov-asp-1 was highly transcribed in second-stage larvae (L2s) and the L3, also being present (but at lower levels) in moulting L3s and females of adult O. volvulus [77]. Subsequently, the protein Ov-ASP-1 was studied further in L3s and shown to be localized exclusively to the granules of the glandular oesophagus [78]. This protein is secreted via degranulation during development, following the invasion of the host. Due to its immunogenicity, recombinant Ov-ASP-1 has been demonstrated to induce a "partial but significant protection" in mice against challenge infection with infective L3s of O. volvulus [78]. However, the presence of Ov-ASP-1 could not be demonstrated in the supernatants
from cultures of L3s, possibly because of a lack of a host-derived signal [78]. Recombinant Ov-ASP-1 and Ov-ASP-2 have each been shown to induce an angiogenic response following their injection into the cornea of naïve mice, suggesting that these proteins might contribute to corneal neovascularisation in onchocercal keratitis [77]. In another parasitic nematode of humans, Brugia malayi, only one SCP/TAPS molecule, designated as Bm-VAL-1, has been characterised to date [29]. The transcription of Bm-val-1 was assessed at several time points during the parasitic phase of the life cycle, revealing a uniform pattern. In contrast, the presence of this protein was shown to be restricted to the larval stages, thus raising the question as to whether it plays a role in the invasion of the host [29]. Furthermore, incubation of mosquito-derived L3s (under a variety of conditions) did not yield sufficient Bm-VAL-1 to be detected by immunochemochemical means; thus, it was proposed that the exposure to host-derived factors in vivo provides the signal required to stimulate the release of this ASP-like protein [29]. More recently, a cDNA-oligonucleotide array analysis of transcription in adult males and females of B. malayi revealed that four EST clusters representing ASPs were “up-regulated” in the male [79], suggesting a role in the reproductive physiology of this nematode. Also relating to nematodes within clade III, ASP orthologues were identified in an EST dataset representing the arrested larval stage of the ascaridoid nematode Toxocara canis, the common roundworm of canids [30]. The large insert sizes in the cDNA clones suggested to the authors the presence of two double-domain SCP molecules [30].

8.4.2 Nematodes within clade IV (including the order Tylenchida)

Within clade IV, two ASPs, Meloidogyne secretory protein (Mi-MSP-1) and a venom allergen-like protein (Mi-VAP-2), have been recorded for the root knot nematode of plants, Meloidogyne incognita [80,81]. Genes encoding both proteins showed similarly high transcription levels in pre-parasitic and early parasitic stages and low or no transcription in older juveniles and adult females. More specifically, the transcription of Mi-msp-1 was high in the parasitic, second-stage juveniles (J2) [80]. The authors hypothesized that this molecule might play a role in the establishment of the parasitic relationship between plant and nematode [80,81]. In a Southern blot analysis, a Mi-msp-1 probe hybridized to genomic DNA of two more Meloidogyne species, M. arenaria and M. javanica, but neither for M. hapla nor for C. elegans or the soy bean cyst nematode, Heterodera glycines [80]. In the latter species, two vap genes (namely Hg-vap-1 and Hg-vap-2) were isolated and the corresponding mRNAs localized by in situ hybridization specifically to the subventral oesophageal gland cells of unhatched J2s (indicating the synthesis of Hg-VAP-1 prior to hatching) as well as pre-
parasitic and parasitic J2s. The hybridization signal for both Hg-vap-1 and Hg-vap-2 was weak within the gland cells in late parasitic stages, thus reflecting the degeneration of the subventral gland cells in H. glycines during parasitism and suggesting an involvement of these molecules in parasite-plant interactions [32].

8.4.3 Nematodes within clade V (including the orders Strongylida and Rhabditida)
8.4.3.1 Superfamily Ancylostomoidea (order Strongylida)

In clade V, the ASPs have been studied relatively well in the dog hookworm, An. caninum. A double SCP-domain ASP, called Ac-ASP-1, was first identified as the major component of the ES products from serum-activated, infective L3s. The release of the molecule was proposed to be a response to a host-specific signal during the infection process [17]. Subsequently, a C-type single SCP-domain ASP, called Ac-ASP-2, was identified as a second, major component of ES from serum-activated L3s [18] and localized predominantly to granules in the glandular oesophagus [82], suggesting that it was secreted from An. caninum as it invades the host [82]. A recent transcriptomic analysis of serum-activated L3s of An. caninum demonstrated that the most abundantly represented group of mRNAs associated with serum-activation was that of the asps [20]. In adult An. caninum, a SCP/TAPS molecule, designated ‘neutrophil inhibitor factor’ (i.e., ‘NIF’) was first isolated from a cDNA library and shown to play an immunomodulatory role by blocking the adhesion of activated neutrophils to vascular endothelial cells and the release of H2O2 from activated neutrophils [83]. More recently, two other SCP/TAPS molecules with significant sequence similarity to NIF and Ac-ASP-2 were purified chromatographically and shown to inhibit the aggregation and adhesion of platelets by blocking the function of integrin receptors located on the cell surface [84]. Subsequently, both NIF and platelet inhibitors were shown to be transcribed abundantly in the intestines of both An. caninum and N. americanus [85].

Four additional ASPs were then identified in adult An. caninum and named Ac-ASP-3, Ac-ASP-4, Ac-ASP-5 and Ac-ASP-6 [34]. Ac-ASP-3 was identified as a C-type SCP-single domain in ES products, whereas Ac-ASP-4, Ac-ASP-5 and Ac-ASP-6 were double SCP-domain ASPs, originally identified by immunoscreening an An. caninum cDNA library using antibodies raised specifically against ES from the adult stage [34]. Ac-ASP-3 has been localized to the glands of the oesophagus, Ac-ASP-4 to the cuticular surface, Ac-ASP-5 to the brush border membrane of the posterior end of the gut, and Ac-ASP-6 to the cephalic and excretory glands of the adult stage of An. caninum. A recent proteomic analysis of the ES products of adult An. caninum [86] has revealed that 28% of these products comprise ASP proteins. To date, none of the ASPs present in the adult stage of An. caninum have been detected by the immunocytochemical analysis of ES products from
serum-activated L3s, although transcripts of Ac-asp-3 and Ac-asp-4 have been detected in this latter stage [34]. While the functions of these molecules in the adult worm are largely unknown, a key involvement in the initiation, establishment and/or maintenance of the host-parasite relationship has been hypothesized [20,34,86].

Ac-ASP-1 or/and Ac-ASP-2 homologues have been isolated from Ancylostoma ceylanicum (designated by the authors as Ay-ASP-1 and Ay-ASP-2, respectively), Ancylostoma duodenale (Ad-ASP-1) and N. americanus (Na-ASP-1 and Na-ASP-2, respectively) via the screening of (homologous) L3 cDNA libraries using radiolabelled Ac-asp probes [87-89]. Although the ASP-1 homologues were >85% identical at the amino acid level, the sequence of Ac-ASP-1 from An. caninum was more similar to the Na-ASP-1 from N. americanus than to that of the congeneric species An. duodenale [87]. In addition, the presence of ASPs has also been shown by nucleic acid-hybridization in the dog and cat hookworms (i.e., Ancylostoma tubaeforme, Ancylostoma braziliense and Uncinaria stenocephala) [18]. In N. americanus, results from crystallography [90], combined with the observation that Na-ASP-2 induces neutrophil and monocyte migration [91], suggest that this molecule has a role as an antagonistic ligand of complement receptor 3 (CR3) and alters the immune cascade by preventing the binding of chemotaxin [90].

8.4.3.2 Superfamily Trichostrongyloidea (order Strongylida)

In adult Haemonchus contortus (the ‘barber’s pole worm’) of small ruminants, two ASPs have been identified to date, namely Hc24 (a C-type single SCP-domain) and Hc40 (double SCP-domain) [92,93]. Hc24 was first described by Takáts et al. [94] as one of the two low molecular weight oesophageal antigens recognized by sera from sheep hyperimmunized (by infection) with L3s of H. contortus. Subsequently, the same protein was identified in the ES products from H. contortus adults [92]. In contrast to hookworms, the transcript representing Hc24 was shown to be present in L4s and adult worms but not in eggs or L3s [95]. Hc40 was detected via the immunoscreening of an adult worm cDNA library using polyclonal serum raised against membrane- or secreted proteins from the gut of adult H. contortus. However, interestingly, the mRNA encoding Hc40 was preferentially detected in tissues external to the gut [93]. Although only these two molecules have been isolated and characterized to date, an analysis of Haemonchus ES proteins revealed 9 distinct spots representing Hc24 or related molecules, which showed variation in relative immunogenicity in sheep [96]. A subsequent appraisal of the EST database available for H. contortus revealed that a range of expressed sequence tags (ESTs), encoding asps (n = 36; cf.
www.nematode.net) and differing in their predicted amino acid sequence by 5-13%, could be assigned to 22 distinct clusters [96].

In adult *Ostertagia ostertagi* (the ‘brown stomach worm’), two N-type single SCP-domain ASP homologues, designated *Oo*-ASP-1 and *Oo*-ASP-2, were characterized as major antigens in a protective ES-thiol fraction of partially purified ES products. The highest levels of transcription and expression of these molecules were detected in L4s and adults of *Os. ostertagi*. These proteins could be localized to the reproductive tract of the adult female stage [27]. In the closely related nematode *Teladorsagia circumcincta*, Nisbet et al. [97] characterized 93 ASP-encoding ESTs which were uniquely represented in L4s but not in exsheathed L3s. These ESTs from the L4 stage formed 40 contigs and encoded three different types of ASPs. In the EST dataset representing the L3 stage, only two *asp* contigs (comprising six ESTs) were identified. A proteomic analysis of ES products harvested from *T. circumcincta* L4s (five days after inoculation of sheep with L3s) also identified a number of ASPs [26]. Small numbers of ESTs (n = 2-3) encoding ASPs have also been identified in *Trichostrongylus vitrinus* (black scour worm of small ruminants) and *Oesophagostomum dentatum* (the nodule worm of pigs) [98,99], and were demonstrated to be male-enriched. Also, an analysis of PCR products from genomic DNA revealed substantial sequence variation within an individual of *Cooperia punctata* [12]. Indeed, two double SCP-domain ASPs were identified and named *Cp*-ASP-1a and *Cp*-ASP-1b, in addition to seven different single SCP-domain ASPs, of which three showed higher homology in the N-terminus and four in the C-terminus of the double domain molecules identified, respectively [12].

8.4.3.3 The free-living nematode Caenorhabditis elegans (order Rhabditida)

At least 17 genes encoding SCP/TAPS homologues have been characterized in *C. elegans* [34]. Of these, *vap-1* encodes a unique double SCP-domain protein which is predicted to be secreted by the adult stage of this nematode. However, information on its localization prior to the proposed secretion/excretion and its function is currently lacking. Other orthologues include *scl* genes encoding single SCP-domain proteins. Three *scl* genes (i.e., *scl-2, scl-20* and *scl-27*) displayed high levels of transcription in *C. elegans* in response to individual infections with multiple bacterial pathogens [35,100,101]. However, a specific function has been assessed only for *scl-2*, which is likely to act as a putative anti-microbial effector in response to infection by *Microbacterium nematophilum* [35]. Transcription for *scl-2, scl-20* and *scl-27* has been observed to be increased in response to infection by more than one pathogen [35,100,101]. Other *scl* genes have been proposed to be involved in the regulation of longevity and stress resistance (i.e., *scl-1*; [37]), ageing or stress resistance (i.e., *scl-1* to
scl-15; [102]) and normal fat storage (i.e., scl-9; [103]). In a study of *C. elegans*, aimed at investigating the possible role of *gld-1* (for ‘defective in germ line development’) in the cell proliferation in a germ line cancer, the genes *scl-19* (renamed *dct-2*, DAF-16/FOXO-controlled tumor gene) and *scl-20* (renamed *dct-4*) were inferred to suppress tumor growth promoted by *gld-1* [104]. The authors suggested a possible biological function of these molecules, based on their sequence similarity with the human GliPR proteins (see section 8.3); *scl-19* promoted apoptosis and *scl-20* was inferred to play a role in signalling and in innate immune response [104]. Another SCP/TAPS gene orthologue, designated *lon-1* (for *long* body), was demonstrated to regulate body morphogenesis [105]. Selected mutations in the *lon-1* gene resulted in worms that were 25% longer than those of the wild-type (N2) strain, whereas an over-expression resulted in a small-bodied (‘Lon’) phenotype [105,106].

8.4.4 Trematodes (i.e., subclass Digenea)

SCP/TAPS proteins have been studied comprehensively in *S. mansoni*, a dioecious trematode infecting humans [3]. Thirteen SCP/TAPS proteins (SmVAL1 to SmVAL13) have been isolated and characterized from this parasite, of which 12 display a single SCP-domain and one a double SCP-domain. However, an analysis of genomic sequence data for *S. mansoni* has predicted the existence of 15 other SmVAL proteins (http://www.genedb.org/genedb/smansonir/), and transcription has been assessed for 11 of them [3]. Transcriptional analysis of SmVAL1 to SmVAL13 in different developmental stages of *S. mansoni* revealed a significant variation in profiles, with some proteins exhibiting higher transcription in the invasive stage of the parasite. While roles in regulating protease activity, protease inhibition and/or a modulation of the immune response during the invasion of the vertebrate host have been postulated [3], they remain to be investigated in detail.

8.5 Phylogenetic analysis of eukaryote SCP/TAPS: a proposed classification system

To investigate the phylogenetic relationships of SCP/TAPS proteins, 72 full-length amino acid sequences representing 35 species of eukaryote were obtained from the GenBank (http://www.ncbi.nlm.nih.gov/) and EMBL-EBI (http://www.ebi.ac.uk/) databases (cf. Tables 8.1 and 8.2) and subjected to InterProScan analysis (http://www.ebi.ac.uk/interpro) to identify single (i.e., n = 60; cf. Table 8.1) and double SCP-domains (i.e., n = 12; cf. Table 8.2). For each domain-type, sequences were aligned separately using BioEdit [107] and adjusted manually. Amino acid sequence variation determined based on pairwise comparisons (p-distance; [108]) using MEGA v4.0 software [109] ranged from 0.034 to 0.909 (mean:
0.649) for single- and from 0.025 to 0.759 (mean: 0.606) for double-SCP domain sequences, respectively.

The data for each of the two domains were subjected to analysis by Bayesian inference (BI) using the program MrBayes v.3.1.2 [110]. The latter program incorporates a "codon-position model", which is presently recognized to be the most appropriate for the phylogenetic analysis of protein sequence datasets [111]. Furthermore, analysis by BI is time efficient to carry out, allowing all relevant parameters (i.e., topology, branch lengths and posterior probabilities) to be determined in a single run. In the present investigation, each analysis was conducted for 1,000,000 generations (ngen = 1,000,000), with every 100-th tree being saved, using the following parameters: rates = gamma,amodelpr = mixed, and the prior parameters left at the default settings. Tree and branch lengths were summarized employing the parameter ‘sumt burnin = 1000’; a consensus tree was constructed, with ‘contype = halfcompat’ nodal support being determined using consensus posterior probabilities and displayed employing the program TreeView v.1.6.6 [112]. A nodal support of ≥ 0.70 was considered significant. This procedure was applied to the set of holotype sequences (see Tables 8.1 and 8.2) to produce the phylogenetic trees shown in Figs. 8.1 and 8.2.

For single SCP-domain SCP/TAPS proteins, the most likely tree was generated using the protein PR-5 of Or. sativa as the outgroup (Fig. 8.1). The analysis identified one main clade comprising SCP/TAPS proteins from animals (nodal support: 0.96) to the exclusion of clades formed by homologues from plants (nodal supports: 0.68 and 0.91, cf. Fig. 8.1). Within the former clade, 3 sub-clades representing SCP/TAPS sequences from nematodes, snake venom and S. mansoni/archeopods were formed. Nodal support for each of these clades ranged from 0.73 to 0.88 (see Fig. 8.1). The proteins from nematodes (C-type and N-type) grouped to the exclusion of snake venom, supporting the existence of a monophyletic group of SCP/TAPS for parasitic nematodes. Conversely, a distinct group of SCP/TAPS proteins from trematodes and humans was not supported as monophyletic, based on the analysis. Indeed, two SCP/TAPS proteins from S. mansoni (i.e., SmVAL6 and SmVAL13) formed a separate cluster with GAPR1 from H. sapiens and G2 from D. melanogaster. A recent phylogenetic analysis of the SCP/TAPS protein family for S. mansoni identified two groups, with both SmVAL6 and SmVAL13 grouping to the exclusion of the remaining family members; this division could be attributed to a lack of predicted signal peptides in both of these molecules [3]. Similarly, two separate groups of genes encoded the SCP/TAPS proteins in D. melanogaster, with proteins G2 and G3 each representing a clade [14]. In the present analysis, the clade of proteins from nematodes included two
distinct sub-clades of SCP/TAPS proteins representing strongylids of animals (nodal support: 0.73) and those from filarioids and plant-parasitic nematodes (nodal support: 1.00), respectively (see Fig. 8.1).

For double SCP-domain SCP/TAPS proteins, the most likely tree was generated using sequence data available for organisms other than nematodes as the outgroup (i.e., SmVAL11; cf. Table 8.2 and Fig. 8.2). The analysis identified a monophyletic clade (nodal support: 0.94) that included all but one sequence (i.e., Ac-ASP-4). No distinct separation between SCP/TAPS from hookworms and those from other strongylid nematodes was supported.

Phylogenetic approaches, such as that used herein, are commonly utilized to construct frameworks for the classification of gene and protein families for a range of prokaryote and eukaryote organisms [113-115]. Although there is no consensus opinion as to which is the best algorithm for phylogenetic analysis/reconstruction, the present approach (considering clades with strong nodal support) provides a foundation for an improved classification of individual groups (classes, subclasses or families) of nematode SCP/TAPS proteins. Together with the positioning of newly described nematode SCP/TAPS proteins within the current tree, it would also be possible to provide a standardized nomenclature. Proteins could be named with a code, such as ‘Nem’ (for nematode), followed by an abbreviation of the species name (the first letter of the genus name followed by the first two letters of the species name - *in italics*), followed by the wording ‘SCP/TAPS’ indicating the protein family and ‘1’ or ‘2’ for single- and double SCP-domain proteins, respectively. A lower case letter might be added to indicate the order of isolation. Following these criteria, as an example, the known secretory protein *Mi-MSP-1*, the first single domain SCP/TAPS member isolated from the nematode *M. incognita* [80], would be named Nem-\*Mi-SCP/TAPS-1a. Using this approach, each SCP/TAPS protein could be positioned in the phylogenetic tree (with statistical support for its position) and then assigned a name containing all of the relevant information required.

8.6 Concluding remarks and biotechnological implications

SCP/TAPS proteins have been the focus of numerous studies aimed at understanding the mechanisms determining the interplay between hosts and pathogens. The vast number of SCP/TAPS proteins that have been identified to date, and described from a range of plant and animal species, exhibit variation in sequence, structure and/or biochemistry, which might reflect a diversification of function. However, the presence of SCP/TAPS proteins across a broad range of taxa suggests key biological roles. Unless a clear classification based on phenetic and genetic characters
for these molecules is provided, the opportunity to explore the function of those molecules in the network of the host-pathogen interaction mechanisms is somewhat compromised. The present chapter examined in detail the relationships of the SCP/TAPS proteins of parasitic nematodes (by comparing amino acid sequence data for homologues from other eukaryotic organisms), using a phylogenetic approach, based on an objective mathematical/statistical model/algorithm. The sequence data and molecular phylogenetic analysis can be transformed into quantifiable parameters, such as degrees of sequence identity, genetic distances and nodal support for groupings, which might allow the definition of criteria for the demarcation of groups of SCP/TAPS proteins. The present classification system does not take into account the localization of a molecule, participation in biochemical pathways or functional role/s assigned to a small number of SCP/TAPS proteins. Indeed, it is well accepted that functional classification should be distinct and independent from the classification based on structure, as the first should allow structurally and/or phylogenetically unrelated molecules to be included in the same functional category based on similarities in biological roles and other features [116]. Moreover, a classification based on phylogenetic relationships can be assessed in silico using software designed specifically to facilitate the grouping of molecules, unlike a functional classification which requires extensive experimentation in a defined biological system to provide detailed insights into the cellular and molecular aspects of genes and their products.

Given the fundamental biological roles that SCP/TAPS proteins have been proposed to play in a range of eukaryotic systems (e.g., defence against pathogens, allergens and induction of cell proliferation), a clear understanding of the function of such molecules in each biological system has become a priority. The major advances in proteomic, genomic and bioinformatic technologies now provide unique opportunities for a ‘systems biological’ exploration of the structure, function and molecular interactions of these proteins. Such an integrated approach will be advantageous in elucidating the fundamental molecular biology and functional roles of these proteins. Such an approach should lead to important biotechnological outcomes, including the development of drugs or vaccines designed to disrupt or interrupt key biological pathways or processes associated with SCP/TAPS proteins. For example, current evidence indicates that a range of SCP/TAPS molecules are transcribed/expressed by hookworms (as larvae) during their invasion of the host (i.e., transition from free-living to parasitic stage) and as adults in the intestine of the host (see section 8.4.3.1). The likely involvement of these molecules at the parasite-host interface might make some of them attractive as vaccine candidates. Currently, the protein Na-ASP-2 from the human hookworm *N. americanus* is being evaluated in clinical trials, aimed at assessing
its safety and tolerability as a vaccine against necatoriasis in humans [21]. Elucidating the molecular function of this molecule (as well as of other SCP/TAPS members of other parasitic nematodes), together with additional structural information on the proteins, could assist in the identification of key ‘structural antigens’; such small protein domains could then be synthesized chemically and evaluated as a “conformational epitope” vaccine [117]. The identification of structural antigenic or immunogenic epitopes, which are conserved among multiple hookworm species, has the potential to lead to the development of a “multivalent” vaccine against hookworm disease in humans.
8.7 References


85. Ranjit N, Jones MK, Stenzel DJ, Gasser RB, Loukas A. A survey of the intestinal transcriptomes of the hookworms, *Nectator americanus* and *Ancylostoma*


Table 8.1 Full-length single SCP-domain SCP/TAPS protein sequences included in the phylogenetic analysis, listed according the main eukaryote taxa (i.e. plants, arthropods, mammals, parasitic and free-living helminths). Current nomenclature, sequence accession number and available information on function (when available) are also given.

<table>
<thead>
<tr>
<th>Organism (listed in order of appearance in the text)</th>
<th>Protein name</th>
<th>Current nomenclature</th>
<th>Accession number</th>
<th>Inferred function</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Plants</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>Pathogenesis related protein 1</td>
<td>AT2G14610</td>
<td>815949 (Gene ID)</td>
<td>Defence against mycete infections</td>
<td>[49]</td>
</tr>
<tr>
<td><em>Lycopersicon esculentum</em></td>
<td>Pathogenesis related protein 1</td>
<td>SIP14</td>
<td>A22633</td>
<td>Defence against mycete infections</td>
<td>[46]</td>
</tr>
<tr>
<td><em>Nicotiana tabacum</em></td>
<td>Pathogenesis related protein 1</td>
<td>NtPR-1a</td>
<td>X05959</td>
<td>Defence against tobacco mosaic virus infections</td>
<td>[119]</td>
</tr>
<tr>
<td></td>
<td>Pathogenesis related protein 1</td>
<td>NtPR-1b</td>
<td>X03465</td>
<td>Defence against tobacco mosaic virus infections</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pathogenesis related protein 1</td>
<td>NtPR-1c</td>
<td>X05454</td>
<td>Defence against tobacco mosaic virus infections</td>
<td></td>
</tr>
<tr>
<td><em>Oryza sativa</em></td>
<td>Pathogenesis related protein 1</td>
<td>OsPR-1a</td>
<td>U82200</td>
<td>Response to cutting, application of jasmonic acid or other phytomolecules</td>
<td>[50]</td>
</tr>
<tr>
<td></td>
<td>Pathogenesis related protein 1</td>
<td>OsPR-1b</td>
<td>U89895</td>
<td>Response to cutting, application of jasmonic acid or other phytomolecules</td>
<td>[51]</td>
</tr>
<tr>
<td><em>Solanum tuberosum</em></td>
<td>Pathogenesis related protein 1</td>
<td>StPR-1b</td>
<td>AY050221</td>
<td>Defence against mycete infections</td>
<td>[118]</td>
</tr>
<tr>
<td><strong>Arthropods</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>SCP/TAPS</td>
<td>G2</td>
<td>NP608663</td>
<td>Male reproduction</td>
<td>[14]</td>
</tr>
<tr>
<td></td>
<td>SCP/TAPS</td>
<td>G3</td>
<td>NP650264</td>
<td>Male reproduction</td>
<td></td>
</tr>
<tr>
<td><em>Vespuca vulgaris</em></td>
<td>Venom allergen 5</td>
<td>Ves v 5</td>
<td>M98858</td>
<td>Venom allergen</td>
<td>[9]</td>
</tr>
<tr>
<td><em>Vespuca squamosa</em></td>
<td>Venom allergen 5 homologue</td>
<td>VsAG5</td>
<td>P35786</td>
<td>Venom allergen</td>
<td></td>
</tr>
<tr>
<td><em>Vespuca flavipalosa</em></td>
<td>Venom allergen 5 homologue</td>
<td>VfAG5</td>
<td>P35783</td>
<td>Venom allergen</td>
<td>[119]</td>
</tr>
<tr>
<td><em>Vespuca germanica</em></td>
<td>Venom allergen 5 homologue</td>
<td>VgAG5</td>
<td>P35784</td>
<td>Venom allergen</td>
<td></td>
</tr>
<tr>
<td><em>Anopheles gambiae</em></td>
<td>Venom allergen 5 homologue</td>
<td>gVaG</td>
<td>Y17702</td>
<td>Secretion of saliva</td>
<td>[120]</td>
</tr>
<tr>
<td><em>Aedes aegypti</em></td>
<td>Venom allergen 5 homologue</td>
<td>AaAG5</td>
<td>AF466605</td>
<td>Unknown</td>
<td>[121]</td>
</tr>
<tr>
<td><em>Lutzomyia longipalpis</em></td>
<td>Venom allergen 5 homologue</td>
<td>LuLoAG5</td>
<td>AF132511</td>
<td>Secretion of saliva</td>
<td>[52]</td>
</tr>
<tr>
<td><em>Glossina morsitans</em></td>
<td>Venom allergen 5 homologue</td>
<td>TAG5</td>
<td>AF259957</td>
<td>Unknown</td>
<td>[53]</td>
</tr>
<tr>
<td><em>Solenopsis invicta</em></td>
<td>Venom allergen 3</td>
<td>Vesv3</td>
<td>AF012919</td>
<td>Venom allergen</td>
<td>[9]</td>
</tr>
<tr>
<td><strong>Snakes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Agkistrodon blomhoffii</em></td>
<td>Ablomin</td>
<td></td>
<td>AF384218</td>
<td>Smooth muscle contraction blockers</td>
<td>[64]</td>
</tr>
<tr>
<td>Animal</td>
<td>Venom Allergen Homologue</td>
<td>Accession</td>
<td>Function</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------</td>
<td>---------------------------------------------------------------</td>
<td>-------------</td>
<td>---------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Laticauda semifasciata</em></td>
<td>Latisemin</td>
<td>AF384220</td>
<td>Smooth muscle contraction blockers</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Rhabdopsis tigrinus</em></td>
<td>Tigrin</td>
<td>AY093955</td>
<td>Smooth muscle contraction blockers</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Trimeresurus flaviridus</em></td>
<td>Triflin</td>
<td>AF384219</td>
<td>Smooth muscle contraction blockers</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Conus textile</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Human</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Homo sapiens</em></td>
<td>Related to testes-specific, vespid and pathogenesis protein 1</td>
<td>X91911</td>
<td>Regulation of the growth, survival and invasion of glioma cancer cells</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Trypsin inhibitor</td>
<td>BC126292</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Glioma pathogenesis-related protein</td>
<td>U16307</td>
<td>Regulation of development of fibrosis</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Parasitic helminths</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Ancylostoma caninum</em></td>
<td><em>Ancylostoma caninum</em> secreted protein-1</td>
<td>AAC35986</td>
<td>Transition from the free-living to the parasitic stage of activated third stage larvae</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Ancylostoma caninum</em> secreted protein-2</td>
<td>AY217004</td>
<td>Establishment or maintenance of host-parasite relationships</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Necator americanus</em></td>
<td><em>Necator americanus</em> activation associated secreted protein-3</td>
<td>AAP41952</td>
<td>Unknown</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Ancylostoma duodenale</em></td>
<td><em>Ancylostoma duodenale</em> activation associated secreted protein-2</td>
<td>AAP41951</td>
<td>Unknown</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Ancylostoma ceylanicum</em></td>
<td><em>Ancylostoma ceylanicum</em> activation associated secreted protein-2</td>
<td>AAP41953</td>
<td>Unknown</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Ostertagia ostertagi</em></td>
<td><em>Ostertagia</em> secreted protein 1*</td>
<td>AJ310819</td>
<td>Unknown</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Ostertagia</em> secreted protein 2*</td>
<td>AJ515523</td>
<td>Unknown</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Ostertagia</em> secreted protein 3</td>
<td>AM747038</td>
<td>Unknown</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Haemonchus contortus</em></td>
<td><em>Haemonchus</em> 24 kDa protein</td>
<td>U64793</td>
<td>Unknown</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Onchocerca volvulus</em></td>
<td><em>Onchocerca</em> secreted protein 1</td>
<td>AF020586</td>
<td>Role in corneal neovascularisation in onchocercal disease</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note**: Accession numbers are based on the latest available information from the GenBank database.
<table>
<thead>
<tr>
<th>Name</th>
<th>Species</th>
<th>Accession</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Onchocerca secreted protein 2</td>
<td><em>Ov-ASP-2</em></td>
<td>H39490</td>
<td>keratosis?</td>
</tr>
<tr>
<td>Onchocerca secreted protein 3</td>
<td><em>Ov-ASP-3</em></td>
<td>AA917267</td>
<td>Unknown</td>
</tr>
<tr>
<td><em>Brugia malayi</em></td>
<td><em>Bm-VAL-1</em></td>
<td>AF334661</td>
<td>Invasion of the vertebrate host</td>
</tr>
<tr>
<td><em>Wuchereria bancrofti</em></td>
<td><em>Wb-VAH</em></td>
<td>AAD16985</td>
<td>Unknown</td>
</tr>
<tr>
<td><em>Dirofilaria immitis</em></td>
<td><em>Di-VAL-1</em></td>
<td>AF001100</td>
<td>Unknown</td>
</tr>
<tr>
<td><em>Meloidogyne incognita</em></td>
<td><em>Mi-MSP-1</em></td>
<td>AF013289</td>
<td>Establishment of the relationships with the plant host</td>
</tr>
<tr>
<td><em>Meloidogyne incognita</em></td>
<td><em>Mi-VAP-2</em></td>
<td>EF370396</td>
<td>Unknown</td>
</tr>
<tr>
<td><em>Heterodera glycines</em></td>
<td><em>Hg-VAP-1</em></td>
<td>AF374388</td>
<td>Establishment of the relationships with the plant host</td>
</tr>
<tr>
<td><em>Heterodera glycines</em></td>
<td><em>Hg-VAP-2</em></td>
<td>AY033601</td>
<td>Establishment of the relationships with the plant host</td>
</tr>
<tr>
<td><em>Schistosoma mansoni</em></td>
<td><em>SmVAL1</em></td>
<td>AY994061</td>
<td>Protease activity, protease inhibition and/or modulation of the host immune response during the infection process</td>
</tr>
<tr>
<td><em>Schistosoma mansoni</em></td>
<td><em>SmVAL2</em></td>
<td>AY994062</td>
<td>Protease activity, protease inhibition and/or modulation of the host immune response during the infection process</td>
</tr>
<tr>
<td><em>Schistosoma mansoni</em></td>
<td><em>SmVAL3</em></td>
<td>DQ060000</td>
<td>Protease activity, protease inhibition and/or modulation of the host immune response during the infection process</td>
</tr>
<tr>
<td><em>Schistosoma mansoni</em></td>
<td><em>SmVAL4</em></td>
<td>AY994063</td>
<td>Protease activity, protease inhibition and/or modulation of the host immune response during the infection process</td>
</tr>
<tr>
<td><em>Schistosoma mansoni</em></td>
<td><em>SmVAL5</em></td>
<td>DQ269980</td>
<td>Protease activity, protease inhibition and/or modulation of the host immune response during the infection process</td>
</tr>
<tr>
<td><em>Schistosoma mansoni</em></td>
<td><em>SmVAL6</em></td>
<td>AY953433</td>
<td>Protease activity, protease inhibition and/or modulation of the host immune response during the infection process</td>
</tr>
<tr>
<td><strong>Schistosoma mansoni</strong> venom allergen</td>
<td>SmVAL7</td>
<td>DQ060001</td>
<td></td>
</tr>
<tr>
<td>---------------------------------------</td>
<td>--------</td>
<td>----------</td>
<td></td>
</tr>
<tr>
<td>inhibition and/or modulation of the host immune response during the infection process</td>
<td>Protease activity, protease inhibition and/or modulation of the host immune response during the infection process</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Schistosoma mansoni</strong> venom allergen</td>
<td>SmVAL9</td>
<td>DQ269979</td>
<td></td>
</tr>
<tr>
<td>like 9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Schistosoma mansoni</strong> venom allergen</td>
<td>SmVAL10</td>
<td>EF421456</td>
<td></td>
</tr>
<tr>
<td>like 10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Schistosoma mansoni</strong> venom allergen</td>
<td>SmVAL12</td>
<td>DQ269978</td>
<td></td>
</tr>
<tr>
<td>like 12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Schistosoma mansoni</strong> venom allergen</td>
<td>SmVAL13</td>
<td>DQ269977</td>
<td></td>
</tr>
<tr>
<td>like 13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Free-living helminths</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Caenorhabditis elegans</em></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>scp-like 10 (F49E11.5)</strong></td>
<td>SCL-10</td>
<td>Q20609</td>
<td></td>
</tr>
<tr>
<td>Ageing and stress-resistance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>scp-like 22 (T05A10.5)</strong></td>
<td>SCL-22</td>
<td>P90958</td>
<td></td>
</tr>
<tr>
<td>Resistance to pathogen infection</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[102]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[35,100,101]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 8.2 Full-length double SCP-domain SCP/TAPS protein sequences characterized for parasitic and free-living helminths and included in the phylogenetic analysis. Current nomenclature, sequence accession number and information on function (when available) are also given.

<table>
<thead>
<tr>
<th>Species (listed in order of appearance in the text)</th>
<th>Protein name</th>
<th>Current nomenclature</th>
<th>Accession number</th>
<th>Inferred function</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Ancylostoma caninum</em></td>
<td><em>Ancylostoma caninum</em> secreted protein-2</td>
<td>Ac-ASP-1</td>
<td>Q16937</td>
<td>Response to host-specific signals during the infection process</td>
<td>[17]</td>
</tr>
<tr>
<td></td>
<td><em>Ancylostoma caninum</em> secreted protein-4</td>
<td>Ac-ASP-4</td>
<td>AY217005</td>
<td>Establishment or maintenance of host-parasite relationships</td>
<td>[34]</td>
</tr>
<tr>
<td></td>
<td><em>Ancylostoma caninum</em> secreted protein-5</td>
<td>Ac-ASP-5</td>
<td>AY217006</td>
<td>Establishment or maintenance of host-parasite relationships</td>
<td>[34]</td>
</tr>
<tr>
<td></td>
<td><em>Ancylostoma caninum</em> secreted protein-6</td>
<td>Ac-ASP-6</td>
<td>AY217007</td>
<td>Establishment or maintenance of host-parasite relationships</td>
<td>[34]</td>
</tr>
<tr>
<td><em>Necator americanus</em></td>
<td><em>Necator americanus</em> Ancylostoma-secreted protein 1</td>
<td>Na-ASP-1</td>
<td>AAD13340</td>
<td>Unknown</td>
<td>[87]</td>
</tr>
<tr>
<td><em>Ancylostoma duodenale</em></td>
<td><em>Ancylostoma duodenale</em> secreted protein 1</td>
<td>Ad-ASP-1</td>
<td>AAD13339</td>
<td>Unknown</td>
<td></td>
</tr>
<tr>
<td><em>Ancylostoma ceylanicum</em></td>
<td><em>Ancylostoma ceylanicum</em> secreted protein 1</td>
<td>Ay-ASP-1</td>
<td>AAN11402</td>
<td>Unknown</td>
<td>[88]</td>
</tr>
<tr>
<td><em>Ostertagia ostertagi</em></td>
<td><em>Ostertagia ostertagi</em> Ancylostoma secreted protein 4</td>
<td>Oo-ASP-4</td>
<td>AM747039</td>
<td>Unknown</td>
<td>[16]</td>
</tr>
<tr>
<td><em>Haemonchus contortus</em></td>
<td><em>Haemonchus</em> 40 kDa protein</td>
<td>Hc40</td>
<td>AAC03562</td>
<td>Unknown</td>
<td>[92]</td>
</tr>
<tr>
<td><em>Cooperia punctata</em></td>
<td><em>Cooperia punctata</em> Ancylostoma-secreted protein 1a</td>
<td>Cp-ASP-1a</td>
<td>AF352702</td>
<td>Unknown</td>
<td>[12]</td>
</tr>
<tr>
<td></td>
<td><em>Cooperia punctata</em> Ancylostoma-secreted protein 1b</td>
<td>Cp-ASP-1b</td>
<td>AF352714</td>
<td>Unknown</td>
<td></td>
</tr>
<tr>
<td><em>Schistosoma mansoni</em></td>
<td><em>Schistosoma mansoni</em> venom allergen-like 11</td>
<td>SmVAL11</td>
<td>DQ151891</td>
<td>Protease activity, protease inhibition and/or modulation of the host immune response during the infection process</td>
<td>[3]</td>
</tr>
<tr>
<td><em>Caenorhabditis elegans</em></td>
<td>Venom allergen protein-1</td>
<td>VAP-1</td>
<td>Q19348</td>
<td>Unknown</td>
<td>[<a href="http://www.wormbase.org">www.wormbase.org</a>]</td>
</tr>
</tbody>
</table>
Fig. 8.1 Phylogenetic reconstruction of single SCP-domain SCP/TAPS protein sequence data using Bayesian inference. Accession numbers, linked to [numbers in the tree], are: NM_001056008 [1]; Q20609 [2]; AF205886 [3]; AF334661 [4]; AAD16985 [5]; AF001100 [6]; H39490 [7]; AA917267 [8]; AF013289 [9]; EF370396 [10]; AF374388 [11]; AY033601 [12]; P90958 [13]; AAC35986 [14]; AAP41951 [15]; AAP41953 [16]; AAP41952 [17]; AM747038 [18]; U64793 [19]; AY217004 [20]; AJ310819 [21]; AJ515523 [22]; X91911 [23]; U16307 [24]; AF384218 [25]; AF384219 [26]; AY093955 [27]; AF384220 [28]; AY994061 [29]; AY994062 [30]; AY994063 [31]; EF421456 [32]; DQ269978 [33]; DQ269980 [34]; DQ269979 [35]; DQ060001 [36]; DQ060000 [37]; NP650264 [38]; P35786 [39]; P35783 [40]; M98858 [41]; P35784 [42]; AF012919 [43]; Y17702 [44]; AF132511 [45]; AF466605 [46]; AF259957 [47]; BC126292 [48]; AJ491318 [49]; AY953433 [50]; DQ269977 [51]; AY039756 [52]; NP608663 [53]; AY050221 [54]; A22633 [55]; 815949 (Gene ID) [56]; U82200 [57]; U89985 [58]; X05959 [59]; X03465 [60]; X05454 [61].
Fig 8.2 Phylogenetic reconstruction of double SCP-domain SCP/TAPS protein sequence data using Bayesian inference. Accession numbers, linked to [numbers in the tree], are: DQ151891 [1]; AY217005 [2]; Q19348 [3]; Q16937 [4]; AAD13340 [5]; AAD13339 [6]; AAN11402 [7]; AM747039 [8]; AAC03562 [9]; AF352702 [10]; AF352714 [11]; AY217007 [12]; AY217006 [13].
CHAPTER 9

General discussion

Knowledge about the transcription of genes in different developmental stages of a parasite is central to a better understanding of the fundamental molecular mechanisms that govern essential biological processes and, ultimately, could assist in identifying possible avenues for the development of novel treatment and control strategies.

The present thesis (i) applied a conventional, publicly available, web-based bioinformatic workflow system for the analysis of both small- and large-scale transcriptomic datasets from selected parasitic nematodes of the orders Ascaridida and Strongylida (Chapters 2-6), (ii) developed a practical, semi-automated bioinformatic pipeline for the analysis of large-scale datasets generated by next-generation sequencing (NGS) (Chapter 7), and (iii) established a bioinformatic-phylogenetic framework for the classification of a key group of molecules (i.e., the SCP/TAPS proteins) encoded in the transcriptome of a range of parasitic worms and other eukaryotes (Chapter 8). The objectives of the present chapter were (a) to summarize the main technical and fundamental research achievements, (b) to discuss the findings and the implications from this research in relation to parasitology and a broader context, and (c) to provide a perspective on opportunities and prospects for future investigations.

Accurate bioinformatic analyses of sequence data are crucial in providing biological meaningful molecular biological information on organisms, including parasitic nematodes. Until recently, detailed bioinformatic analyses of nucleotide and protein sequences have been restricted largely to specialized laboratories with substantial computer and software capacities. The introduction of worldwide web (www)-based computer pipelines for the acquisition and automated processing of sequence data has marked a revolution in molecular parasitology, because the community could rapidly derive considerable amounts of biological information from nucleotide and amino acid sequence datasets produced by conventional (Sanger) sequencing. At the commencement of this thesis, such a web-based pipeline (called ESTExplorer; [1,2]) was employed for the detailed analysis of novel, but relatively small, EST datasets (32-361 nucleotide sequences) representing gender-enriched transcripts in Ascaris suum (Chapter 2) and molecules conserved between Haemonchus contortus and Ancylostoma
caninum (Chapter 5). This pipeline performed satisfactorily for data derived via Sanger sequencing. However, it was not able to handle the analyses of large NGS datasets (equating to several hundred Sanger-type sequencing runs or 1-100 million sequences and output file sizes of ~1x10⁹ bytes). Therefore, in Chapters 3-4 and 6-7, software originally designed for web-based analyses of relatively small datasets was adapted to face the challenges presented by the need to analyse huge NGS datasets in a time- and cost-efficient manner. The development of a user-friendly, flexible and practical bioinformatic workflow system yielded a tool that scientists, with no or very limited bioinformatic expertise, could use for the analysis of massive datasets in a straightforward and biologically useful manner (Chapter 7).

Overall, the present thesis has described the results of bioinformatic analyses of ~4,700,000 ESTs from selected representatives of the orders Ascaridida and Strongylida (Chapters 2-7). The raw sequence data generated are available for download from public databases (i.e., the GenBank and/or Sequence Read Archive [SRA] databases at NCBI [http://www.ncbi.nlm.nih.gov/]), thus representing a large resource for future genomic and fundamental molecular biological explorations of these organisms as well as the development of molecular tools, such as diagnostic tests, for applied investigations.

With the exception of the sequence data analysed in Chapters 2 and 5 (in which sequences were selected based on the results of a microarray analysis), the cDNA libraries generated and sequenced by 454 technology in this thesis were normalised. Although the normalisation process allows transcripts to be studied qualitatively, this approach does not allow differential gene expression to be investigated quantitatively (see Chapters 3, 4, 6 and 7). Exploring differential transcription among stages, sexes and tissues of parasitic nematodes and other helminths provides unique insights into molecular changes occurring, for example, during development and reproduction. Future studies involving the sequencing of non-normalised cDNA libraries by, for instance, Illumina technology (www.illumina.com; [3]) will provide an avenue to explore essential biological pathways in parasitic nematodes, such as those linked to the development of neuronal tissue, the formation of cuticle and the digestion of host haemoglobin in H. contortus (cf. Chapter 4) and in mitochondrial and amino acid metabolism in N. americanus (cf. Chapter 6). However, the incorporation of gene expression data will inevitably pose new computational challenges for the correct assembly and analysis of sequence datasets and for the accurate prediction of alternatively spliced transcripts.
Currently, due to the lack of complete genomic sequences for parasitic nematodes, newly generated transcriptomic and genomic sequence datasets need to be assembled de novo, which means that pooled reads are assembled without a bias towards known sequences [4]. Due to the amount of RNA required for NGS (~5-10 µg; [5,6]), transcriptomes from small nematodes usually originate from multiple individuals, potentially leading to an increased complexity of the sequence data acquired (linked, for instance, to single nucleotide polymorphisms [SNPs] and other types of sequence variation) and posing some challenges for the assembly. In terms of complexity, computational and time requirements, de novo assemblies are orders of magnitude slower and much more computer-memory intensive than knowledge-based (mapping) assemblies, in which reads are aligned and assembled against an existing “backbone” sequence [7]. In addition, reliable de novo assemblies are heavily dependent upon the availability of long reads (>100 bases) and of high-coverage, paired-end sequence data [7,8]. In previous studies, the complementary nature of the 454 and Illumina sequencing platforms has allowed the assembly of raw reads into large scaffolds without need for reference sequence [9-11]. Thus, clearly, the 454 sequence data assembled in the present thesis should assist future de novo assemblies of Illumina data (both transcriptomic and genomic) for the species investigated.

In the absence of a reference genome for a parasitic nematode, a correct assembly of ESTs is a crucial step for examining coding genes and, ultimately, addressing biological questions regarding gene and protein function. Knowledge of the function of genes and gene products from organisms is predicted using a process known as ‘sequence annotation’, which has been defined as “the process of gathering all available information and relating it to the sequence assembly both by experimental and computational means” [12]. In the present thesis, the annotation of sequences was based on comparisons with data available in public databases (Chapters 2-7). Typically, quality databases are regularly updated and continually curated (see [13,14]). Accurate annotation, particularly in the absence of a reference genome sequence, is dependent on the efficiency of the updates and curation. Presently, open-source programs and databases routinely employed for the bioinformatic analyses of sequence data from parasitic nematodes are available via multiple portals, thus requiring significant efforts to maintain accurate and up-to-date assembly and annotation pipelines [15]. In addition, the rate at which public databases are updated and corrected differs considerably. For instance, the Swiss-Prot database (http://au.expasy.org/sprot/) accepts corrections from its user community, whereas GenBank (http://www.ncbi.nlm.nih.gov/genbank/) only accepts corrections from the author of an entry [16], thus significantly affecting the
accuracy and speed with which new sequences are annotated. In addition, some information-management systems evolve to efficiently incorporate data from large-scale projects, but often, the annotation of single records from the literature is slow and cumbersome [17]. Given that, presently, the annotation of sequence data for parasitic nematodes relies heavily on the use of bioinformatic approaches and already annotated-curated sequence data for a wide range of organisms (see Chapters 2-7), these observations are particularly crucial and deserve further consideration. For instance, the analyses and annotation of large-scale transcriptomic sequence datasets for parasitic nematodes could be considerably facilitated through the establishment of a ‘reference’ website for molecular parasitologists. Such a website could provide regular releases of newly developed and validated bioinformatic pipelines for the analyses of sequence datasets. It could also provide links to regularly updated databases that are routinely employed for the annotation of new sequences as well as a distinct, high-quality database of curated functional annotations, supported by experimental data documented in peer-reviewed, international publications. In the future, the establishment of a ‘centralized’ consortium to facilitate the sharing and optimization of bioinformatic pipelines for sequence processing and annotation and, more broadly, to allow access to new sequence data, and experimental protocols and relevant literature would be very useful.

In the present thesis, the annotation of peptides inferred from the datasets analysed (Chapters 2-7) was performed by assigning predicted biological function/s based on comparison with existing information available for the free-living nematode Caenorhabditis elegans and for other organisms in public databases (e.g., WormBase, www.wormbase.org; InterPro, http://www.ebi.ac.uk/interpro/; Gene Ontology, http://www.geneontology.org/; OrthoMCL, http://www.orthomcl.org/; BRENDA, http://www.brenda-enzymes.org/). Using this approach, predictions for key groups of molecules were made in relation to their function and essential roles in biological processes (Chapters 2-7). Such groups included the SCP/TAPS proteins (see Chapter 8) and molecules linked to the physiology of the nervous system (Chapters 3 and 4), to the formation of the cuticle (Chapter 4), proteases and protease inhibitors (Chapters 3, 4, 6 and 7) and protein kinases and phosphatases (Chapters 2, 3 and 7). However, information gleaned now requires experimental validation. Therefore, in the future, extensive laboratory experiments need to be conducted to evaluate the functions of molecules in the parasites studied and/or in a suitable surrogate organism. RNAi has been applied to a number of parasitic nematodes of animals (orders Strongylida, Ascaridida and Spirurida), but success has been limited [18-30]. Current evidence
[25,27] indicates that a number of nematodes of animals, including *H. contortus* and *B. malayi*, lack critical components of the RNAi machinery (see [26,27,31]. Transgenesis and gene complementation studies have shown considerable promise for evaluating the function of genes from some parasitic nematodes [32-35]. Indeed, a study demonstrating successful transgenesis in the parasitic nematode *Parascrobylloides trichosuri* (Rhabditida) [36] as well as the use of *C. elegans* as a surrogate system for the analysis of the function of some genes from selected members of the Strongylida and Rhabditida [32-35] provide substantial promise and scope for the application of this methodology to genetic functional studies of selected groups of parasitic nematodes.

*C. elegans* represents a useful model for studying aspects of, for example, the neurobiology of parasitic nematodes, including *As. suum* [37-41], *An. caninum* [42], *H. contortus* [39,41,43-45] and *Trichostrongylus colubriformis* [46]. For instance, due to the similarities in amphid structures between the dauer (arrested or hypobiotic) form of *C. elegans* [47] and the L3s of various parasitic nematodes, such as *H. contortus* [44], *An. caninum* [42] and *Strongyloides stercoralis* [48], it was proposed that similar molecular mechanisms control the development of these stages [48-50]. The molecular and biochemical pathways that control dauer formation in *C. elegans* have been the subject of intense studies [41,47,51-53]. The formation of dauer in this nematode is mediated by a specific pheromone, whose secretion is determined by parallel insulin-signalling pathways in four types of sensory neurons, namely ADF, ASG, ASJ and ASJ [47,49,54]. One pathway is mediated by the ‘dauer formation’ DAF-2 (insulin-like) receptor [47,55], whereas the second is regulated by the combination of DAF-7 (transforming growth factor-8 [TGF-8]) and DAF-11 (guanylate cyclase) [41,47]; loss of function in either pathway results in dauer, indicating that these pathways function independently [41]. The *daf*-22 gene has been proposed to play a crucial role in final synthesis of the dauer-pheromone, based on the observation that *daf*-22(*m130), a dauer-defective mutant *C. elegans*, accumulates a hydrophilic precursor of precursors of this molecule with similar activity [51,56]. Although the precise molecular identity of *daf*-22 is yet known (cf. [51]), the nuclear receptor *daf*-12 is hypothesized to represent the final, common target of regulatory pathways for dauer [51]. In addition, the fork head boxO (FoxO) transcription factor, encoded by *daf*-16, is known to stimulate a dauer-specific pattern of expression that promotes the formation of dauer [57-59]. For *S. stercoralis* and *H. contortus*, homologues of *C. elegans daf*-16 have been isolated and characterised (designated *Ss-daf*-16 and *Hc-daf*-16.1 and *Hc-daf*-16.2, respectively; [34,60]. Using transgenesis, both *Ss-daf*-16 and *Hc-daf*-16.2 were shown to restore *daf*-16 function to a *C. elegans* strain carrying a null mutation at this locus [34,60]. Another study [61]
demonstrated the involvement of the insulin-like signalling pathway in a cascade of events that lead to the activation of the An. caninum L3s (following serum-stimulation), and the authors postulated the involvement of a daf-16 homologue in the molecular events linked to the resumption of feeding. The results from these studies indicate significant scope for further (functional) explorations of other molecules linked to key biological pathways in parasitic nematodes. For example, the H. contortus datasets analysed in Chapter 4 could be investigated for the presence of other members of the insulin-signalling pathway, which could then be subjected to functional studies employing transgenesis, in a similar way as described by Hu et al. [34], in the first instance. Elucidating, in detail, the function(s) of molecules involved in the transition from free-living to parasitic stages as well as among the different parasitic stages of socio-economically parasitic nematodes should have enormous implications for a fundamental understanding of their biology, ecology and adaptation, and will provide unique insights which could lead to new intervention or control strategies.

Proteases and protease inhibitors of nematodes have also been the focus of functional genetic studies, aimed at assessing their roles in a variety of important biological processes (such as moultng, cuticle formation and remodelling, embryogenesis, feeding and immune evasion) within the host animal [35,62-65]. Such studies have been stimulated by an interest in defining new drug and/or vaccine targets in parasites, such as blood-feeding nematodes [66-69]. For example, the control of hookworms in humans relies mainly on the administration of anthelmintic drugs (including benzimidazoles; [70]). Re-infections and repeated treatment with such drugs is common. Hence, there has been growing concern that emerging resistance is developing in parasitic nematodes of humans, as is already the case for nematodes of livestock [71]. Thus, developing new drugs and vaccines against nematodes of humans and other animals has been a priority [67,72-79]. Some success has been made toward developing an anti-hookworm vaccine; however, efficacy trials of the molecules as vaccine candidates (e.g., aspartyl- and cysteine-protease from An. caninum and N. americanus, designated Ac-APR-1, Ac-CP-2 and Na-CP-2, respectively) have been limited to studies in experimental animals (see [80-83]; cf. Chapter 1). Also, in early studies, the efficacy of a radiation-attenuated larval vaccine had been shown to induce up to 90% protection in dogs against challenge infection with An. caninum [84]. However, despite its efficacy, the vaccine was rapidly removed from the market due a short shelf-life, relatively high cost of production and a “failure” to induce a sterilizing immunity [84]. More recently, the efficacy of recombinant proteases against hookworm infection has been assessed [77,81,82]. However, despite some positive results [77,81,82], an
effective recombinant anti-hookworm vaccine remains to be optimized. Conversely, attempts to use recombinant gut-derived proteases to stimulate immunity against infections by strongylid nematodes that are not obligate blood-feeders have been largely unsuccessful [85,86]. In addition, it has also been proposed that the combined administration of an effective vaccine and one or more anthelmintics might result in an improved protection and a significant delay in the emergence of anthelmintic resistance [77,87], although evidence is required to support this proposal [87].

Currently, bioinformatic methods are increasingly being used to predict proteins as novel anthelmintic targets [88-91]. This approach involves ‘filtering’ [92,93] and usually includes inferring targets based on key principles and requirements (cf. [94]). First, target proteins should play a key (i.e. essential) role in fundamental biological processes of the parasite, such that the disruption of the molecule or its gene will damage and/or kill the parasite. In the absence of phenotypic data for most parasitic nematodes of animals, the essentiality of genes and/or gene products is inferred based on effective knockdown or knockout of homologues/orthologues in C. elegans [92,95]. In addition, since drugs against parasite molecules could also theoretically bind to and affect host molecules, candidate proteins should be unique to the parasite or, at least, show significant differences in sequence and structure from host homologues [94,96].

In the present thesis, the prediction of drug target candidates in parasitic nematodes was assisted by using extensive information on function and essentiality in C. elegans, D. melanogaster, Mus musculus and/or Saccharomyces cerevisiae. This information accessed via public databases, including WormBase at http://www.wormbase.org, FlyBase at http://flybase.org/, Mouse Genome Informatics at http://www.informatics.jax.org/ and Saccharomyces Genome Database at http://www.yeastgenome.org/; [97-100]; cf. Chapter 1). In addition, since most effective drugs achieve their activity by competing with endogenous small molecules for a binding site on a target protein [101], the amino acid sequences predicted from essential genes were screened for the presence of conserved ligand-binding domains [101,102]. Lists of inhibitors known based on experimental evidence, to specifically bind to such domains were compiled [102]. Interestingly, ~50% of the targets predicted for N. americanus (see Chapter 6) and O. dentatum (see Chapter 7) represented proteins belonging to the same categories, i.e. zinc metalloproteases, amino peptidases, guanosine triphosphatases (GTPases), protein tyrosine kinases (PTKs) as well as serine/threonine protein phosphatases (STPs) and kinases (STKs) (Chapters 6 and 7). Multiple cellular signalling pathways function through the activity of small GTP-binding proteins (GTPases) to regulate multiple biological processes, such as transmembrane
signal transduction, cytoskeletal reorganization, gene expression, intracellular vesicle trafficking, microtubule organization and nucleocytoplasmic transport [103]. GTPases are small (~20-28 kDa), monomeric proteins belonging to six families (i.e., Ras, Rho, Rab, Arf, Ran and Rad; [104]). These regulatory proteins act as bi-molecular switches that cycle between two conformational states (i.e., GDP-bound ["inactive" state] and GTP-bound ["active" state]) and hydrolyze GTP [105]. In humans, the aberrant regulation of GTPases is linked to a number of dysfunctions, including neurological and developmental disorders and cancer [105]. In addition, intracellular pathogenic bacteria, such as Mycobacterium tuberculosis, are known to target host GTPases to evade host immune responses to facilitate the infection process [106]. Such information has stimulated efforts to develop novel therapeutic strategies to inhibit the function of GTPases [103]. For instance, treatments with farnesyltransferase inhibitors, to block the oncogenic properties of Ras GTPases, have been shown to be effective in significantly reducing the progression of various forms of cancer, including carcinomas of the colon, pancreas and lung, neurofibrosarcoma and chronic myelogenous leukaemia, in experimental animals [107,108] and the migration and organization of the cytoskeleton of human prostate cancer cells [109]. Although the overall structure of individual small GTPases is conserved across eukaryotes, the filtering of the N. americanus (Chapter 6) and O. dentatum (Chapter 7) datasets in the present thesis allowed the identification of significant differences in sequence within specific regions of nematode and host transcripts encoding GTPases. These differences might be considered in future studies, aimed at assessing the possibility of designing and synthesizing selective and specific inhibitors against parasite GTPases. Homology modelling [110,111], X-ray crystallography and docking [112-117] studies should assist in this process.

Protein kinases (PTKs) were consistently predicted as targets for nematocides in the present thesis (Chapters 6 and 7). PTKs belong to a large family of proteins regulating development, cell division, differentiation and metabolism in many organisms; these molecules are considered the second most important group of drug targets after GPCRs (see [118,119]). The family of PTKs comprises cell surface receptors (RTK) and non-receptor or cytosolic (CTK) kinases. The unregulated activation or over-expression of PTKs is considered to play a central role in the induction of various forms of cancer [120-124]. Integrated genomic-bioinformatic-chemoinformatic approaches have been employed for the identification and screening effective PTK inhibitors as therapeutic agents for the treatment of human leukaemia [125-127]. For example, in studies aimed at identifying novel inhibitors of a human tyrosine kinase involved in the development and progression of chronic myelogenous leukemia, 15 compounds were
selected following in silico screening of a database of 200,000 known inhibitors [126]. Of these compounds, eight were shown to selectively inhibit the growth of leukemia in vitro [126]. In another study, novel and selective inhibitors of caseine kinase II (CK2) were identified via in silico screening of a database containing ~400,000 compounds, followed by in silico docking [127]. These examples indicate the advantages of using computer-aided tools for the rational prediction and design of drugs for subsequent in vitro and in vivo efficacy testing [128]. Nonetheless, it is clear that any compound shown to be efficacious must also be rigorously tested for its safety according to international guidelines ([129]; http://www.ich.org/cache/compo/276-254-1.html).

Because of the regulatory role that PTKs play in a number of signalling pathways in the cell, interference with their activity can result in the disruption of fundamental homeostatic processes in parasites [130]. In the last years, protein kinases have received particular attention as drug targets in protozoan parasites, such as species of *Plasmodium, Leishmania* and *Trypanosoma* and helminths, including *Echinococcus multilocularis* and *Schistosoma mansoni* [119,130]. In the latter two, PTK inhibitors (i.e., tyrphostins AG1024 and AG538) have been shown to significantly affect the development and survival of the adult parasite through the blockage of glucose uptake [119]. In another study, the inactivation of *S. mansoni* PTKs with herbimicin A (an Src kinase inhibitor) was demonstrated to interfere with mitosis, thus significantly affecting the expression of proteins essential for egg production, including the formation of the eggshell, in adult females [131]. Although the crystal structures of PTKs from parasitic nematodes have not yet been defined, progress has been made in the identification and design of effective inhibitors based on homology models for protein kinases from humans [130]. There is evidence that the active sites of parasite PTKs display subtle differences compared with their human counterparts [130], which is considered promising for the development of parasite-specific kinase inhibitors. However, much more study is required to establish the potential of PTK inhibitors as nematocides. This is obviously a research area worthy of pursuit.

In the future, improved bioinformatic prediction and prioritization of potential drug targets in parasitic nematodes will depend on the availability of respective complete genome sequences. Global repertoires of drug targets could be inferred. For instance, the parasite kinome (= the complete set of kinase genes in the genome) could represents a unique opportunity for the design of parasite-selective inhibitors [130]. In addition, the integration of genomic, transcriptomic and proteomic data will be crucial to identify other groups of molecules essential to parasite survival and development, which could represent drug target candidates. Clearly, next-generation sequencing will
provide the efficiency and depth-of-coverage required to rapidly define the complete genomes of eukaryotic pathogens of socio-economic importance.

In conclusion, the present thesis has elucidated the transcriptomes and molecular biology of some neglected parasitic nematodes of socio-economic importance. These transcriptomic datasets, which are now available through public databases (i.e., [http://www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov) and [http://research.vet.unimelb.edu.au/gasserlab/index.html](http://research.vet.unimelb.edu.au/gasserlab/index.html)), represent an invaluable resource and tool for the future assembly and annotation of respective genomes. Such advances in genomic sequencing and annotation as well as the integrated use of ‘omic technologies will open the door to understanding the systems biology of these nematodes on an unprecedented scale, and will provide golden opportunities for the development of entirely new strategies for the treatment and control of neglected parasitic diseases.
References


LIST OF APPENDICES

All appendices and the published articles linked to individual chapters of this thesis (in pdf format) are on a compact disc (CD) provided with this thesis.

CHAPTER 2

Appendix 2.1 Gene ontology (GO slim terms) classification according the categories 'biological process', 'cellular component' and 'molecular function' inferred for selected *Ascaris suum* male- and female-enriched molecules.

Appendix 2.2 Genes predicted to interact with each of 9 and 28 *Caenorhabditis elegans* genes orthologous to expressed sequence tags (ESTs) encoding male- and female-enriched transcripts in *Ascaris suum*, respectively, listed according decreasing cut-off scores. Gene Ontology (GO slim terms) classification, according the categories 'biological process', 'cellular component' and 'molecular function' inferred for the interacting genes, is also indicated.

CHAPTER 3

Appendix 3.1 Length distribution of contigs following the assembly of the *Trichostrongylus colubriformis* nucleotide sequence data.

Appendix 3.2 List of InterPro domains in the predicted peptide sequences encoded in the adult stage of *Trichostrongylus colubriformis*.

Appendix 3.3 List of gene ontology (GO) terms (according to the categories 'biological process', 'cellular component' and 'molecular function') linked to proteins inferred to be encoded by the adult stage of *Trichostrongylus colubriformis*.

Appendix 3.4 Predicted biological pathways linked to molecules inferred to be encoded in the adult stage of *Trichostrongylus colubriformis*.

Appendix 3.5 Graphic representation of the genetic interaction networks predicted for *Caenorhabditis elegans* orthologues (blue dots) of *Trichostrongylus colubriformis* genes (red dots).

CHAPTER 4

Appendix 4.1 The sequences of oligonucleotide primers used in reverse transcription (rt)PCR.

Appendix 4.2 InterPro domains of predicted peptide sequences encoded in either the ensheathed (= L3) or exsheathed (= xL3) third larval stage of *Haemonchus contortus*.

Appendix 4.3 Terms (i.e., 'biological process', 'cellular component', and/or 'molecular function') representing proteins inferred to be encoded by either the ensheathed (= L3) or exsheathed (= xL3) third larval stage of *Haemonchus contortus*.

Appendix 4.4 Pathways predicted for molecules inferred to be encoded in either the ensheathed (= L3) or exsheathed (= xL3) third larval stage of *Haemonchus contortus*.
Appendix 4.5 *Caenorhabditis elegans* orthologues of expressed sequence tags (ESTs) and genome survey sequences (GSS) representing either the ensheathed (= L3) or exsheathed (= xL3) third larval stage of *Haemonchus contortus*.

Appendix 4.6 Bioinformatic characterization of expressed sequence tags (ESTs) encoding molecules uniquely transcribed in the ensheathed third larval stage (= L3) of *Haemonchus contortus* with orthologues in *Caenorhabditis elegans* and other parasitic nematodes.

Appendix 4.7 A list of *Caenorhabditis elegans* orthologues of expressed sequence tags (ESTs) unique to either the ensheathed (= L3) or exsheathed (= xL3) third larval stage of *Haemonchus contortus* for which probabilistic genetic interaction networks were predicted. Interacting genes are listed according to decreased cut-off scores (see Methods).

Appendix 4.8 Bioinformatic characterization of expressed sequence tags (ESTs) encoding molecules uniquely transcribed in the ensheathed third larval stage (= xL3) of *Haemonchus contortus* with orthologues in *Caenorhabditis elegans* and other parasitic nematodes.

**CHAPTER 5**

Appendix 5.1 Genes (n = 156) predicted to interact with each of sixteen selected *Caenorhabditis elegans* genes (abce-1, act-2, C08H9.2, C55F2.1, calu-1, col-181, cpr-6, elo-2, asp-1, K07E3.4, rpm-2, sel-9, T28C12.4, hsb-1, Y57G11C.15 and ZK593.1), listed according decreasing cut-off score.

Appendix 5.2 Numbers of genes predicted to interact with each of sixteen selected *Caenorhabditis elegans* genes (abce-1, act-2, C08H9.2, C55F2.1, calu-1, col-181, cpr-6, elo-2, asp-1, K07E3.4, rpm-2, sel-9, T28C12.4, hsb-1, Y57G11C.15 and ZK593.1), classified according to gene ontology (GO) categories ‘biological process’, ‘cellular component’ and ‘molecular function’.

**CHAPTER 6**

Appendix 6.1 InterPro domains identified in the peptides predicted for *Necator americanus* and *Ancylostoma caninum*.

Appendix 6.2 Gene ontology (GO) terms (according to the categories ‘biological process’, ‘cellular component’ and ‘molecular function’) linked to peptides predicted for *Necator americanus* and *Ancylostoma caninum*.

Appendix 6.3 Biological pathways involving key peptides predicted for *Necator americanus* and *Ancylostoma caninum*.

Appendix 6.4 Description of *Caenorhabditis elegans* orthologues of *Necator americanus* contigs for which inferred peptides were associated with ‘druggable’ InterPro domains and/or Enzyme Commission (EC) numbers, and a list of candidate nematicidal compounds linked to these domains predicted using the BRENDA database. The presence (✓) or absence (X) of known orthologues in *Ancylostoma caninum*, *Haemonchus contortus*, *Homo sapiens*, *Drosophila melanogaster* and/or *Mus musculus* is also indicated.
CHAPTER 7


Appendix 7.2 Biological pathways assigned to molecules inferred to be encoded in either adult female, adult male, and the third (L3) and fourth (L4) larval stage of Oesophagostomum dentatum.

Appendix 7.3 Mapping of peptides inferred from the transcriptome of Oesophagostomum dentatum [combined assembly of data for adult female, adult male, and the third (L3) and fourth (L4) larval stages] to known biological pathways in Caenorhabditis elegans using the iPath tool.

Appendix 7.4 Examples of genetic interaction networks predicted for Caenorhabditis elegans orthologues of expressed sequence tags (ESTs) unique to either adult female, adult male, and the third (L3) and fourth (L4) larval stages of Oesophagostomum dentatum. Each of the represented C. elegans orthologues was predicted to interact with a minimum of 25 other genes.

Appendix 7.5 A list of Caenorhabditis elegans orthologues of expressed sequence tags (ESTs) unique to either adult female, adult male, and the third (L3) and fourth (L4) larval stages of Oesophagostomum dentatum, for which probabilistic genetic interaction networks were predicted. Interacting genes are listed according to decreasing cut-off score.

Appendix 7.6 Life cycle of the porcine ‘nodule-worm’ Oesophagostomum dentatum. The boxes contain examples of predicted compounds in the BREnda database against molecules unique to either adult female, adult male, and the third (L3) and fourth (L4) larval stages.
Author/s: CANTACESSI, CINZIA

Title: Genomic-bioinformatic investigations of key gastrointestinal parasites of socio-economic importance and their implications

Date: 2011


Persistent Link: http://hdl.handle.net/11343/36000

File Description: Genomic-bioinformatic investigations of key gastrointestinal parasites of socio-economic importance and their implications

Terms and Conditions: Copyright in works deposited in Minerva Access is retained by the copyright owner. The work may not be altered without permission from the copyright owner. Readers may only download, print and save electronic copies of whole works for their own personal non-commercial use. Any use that exceeds these limits requires permission from the copyright owner. Attribution is essential when quoting or paraphrasing from these works.