PROPOSAL

WHOLE GENOME SEQUENCES OF NEMATODES OF THE ORDER STRONGYLIDA

CONSORTIUM*

Robin B. GASSER, colleagues and collaborators
robinbg@unimelb.edu.au
Faculty of Veterinary Science, The University of Melbourne, Australia

Shoba RANGANATHAN
shoba@els.mq.edu.au
Biotechnology Research Institute, Macquarie University, Sydney, Australia

David BAILLIE
baillie@sfu.ca
Department of Mol. Biol. and Biochemistry, Simon Fraser University, British Columbia, Canada

Paul STERNBERG
pws@caltech.edu
California Institute of Technology, Pasadena, California, USA

Makedonka MITREVA
Elaine MARDIS
emardis@watson.wustl.edu
Richard K. WILSON
rwilson@watson.wustl.edu
Genome Sequencing Center, Washington University, St. Louis, MO, USA

*For groups supporting this proposal see page 18

Correspondence address:
Makedonka MITREVA
Research Instructor in Genetics
Washington University
St. Louis, MO 63108
Phone 314-286-1844
Fax 314-286-1810
E-mail: mmitreva@watson.wustl.edu

1. SUMMARY

The order Strongylida (clade V) represents parasites most closely related to *C. elegans*, and the group most likely to benefit from the comparative value of its genome sequence. *C. elegans* has already served as an essential guide for exploring the genomes of other nematode species. However, transcriptomic data from parasitic nematode species, generated by EST approaches, have already served as a boon for *C. elegans* biology, as they confirm gene predictions and add depth to analyses of structure-function.

Here, we propose sequencing the genomes of 20 species, members of the order Strongylida ('strongylids'). Strongylids, including human hookworms and many intestinal parasites of livestock, are relatively closely related to one another and may have resulted from a recent radiation. To advance and facilitate molecular studies of strongylid parasites, an expressed sequence tag (EST)-based gene discovery program has been carried out, as a part of a broad study of the transcriptomes of members of the phylum Nematoda. More than 60,000 ESTs have been generated for key parasites of human and domestic animals, and for model parasitic species belonging to order Strongylida (comprising 83% of all clade V non-*Caenorhabditis* ESTs), and deposited in the public databases. Although the transcriptomic data hold promise for the development of novel control strategies for these important pathogens, emerging whole genome data will be pivotal in a more comprehensive approach to identify alternative control strategies, which is increasingly pressing as levels of parasite resistance to available drugs is rising, making some treatment strategies that worked just 5 years ago, ineffective. Also, some drug treatments and environmental nematocides are being withdrawn because of the potential risk to human health and due to ecological contamination.

Progress on vaccine and drug development has already proven prolific. For example, the Human Hookworm Vaccine Initiative is beginning clinical trials of a larval hookworm antigen, ASP-2, from *Necator americanus*, as a vaccine antigen, and a hookworm anticoagulant is already in human trials. The nematode anticoagulant protein c2 (rNAPc2) is a recombinant version of a naturally occurring protein with anticoagulant properties. The mechanism of action lies in its ability to block the factor VIIa/tissue factor protease complex, which is responsible for the initiation of blood clot formation. By our initiative to sample the order Strongylida comprehensively, more drug candidates, such as parasite proteins that alter host physiology, biochemistry and/or immunology will be found.

The value of *C. elegans* as a model organism for understanding human health and disease has long been recognized. The genomes of four other *Caenorhabditis* spp. are completed or in progress. However, decoding the genomes of the most closely related parasitic nematodes represents an extraordinary and unique resource for comparative evolutionary developmental studies, offering power to identify genes and regulatory sequences by 'phylogenetic footprinting', better define proteins involved in nematode parasitism, expanding the understanding of both conserved and divergent aspects of nematode biology which will enhance the value of *Caenorhabditis* spp. as a model for understanding health and disease, host-parasite relationships, and basic biological processes.

2. INTRODUCTION

Parasitic nematodes of humans, livestock and other animals cause major (subclinical and clinical) diseases of major socio-economic importance globally. In particular, parasitic flatworms (trematodes and cestodes) and roundworms (nematodes) have a major, long-term impact (directly and indirectly) on human health and cause substantial suffering, particularly in children. The World Health Organization (WHO) estimates that 2.9 billion people are infected with nematodes. Morbidity from nematodes is substantial and surpasses diabetes and lung cancer in disability adjusted life year (DALY) measurements. Worldwide, the current financial losses caused by parasites to agriculture (domesticated animals and crops) have a major impact on farm profitability and exacerbate the global food shortage. In agriculture, most parasites are controlled mainly through the use of chemotherapeutic agents (anthelmintics). This type of control is expensive (over 3 billion US dollars are spent annually worldwide on anthelmintics) and, only partially effective. Also, the excessive and uncontrolled use of such agents has resulted in serious problems with drug resistance. Furthermore, the use of such drugs poses major risks of residue problems in meat, milk and the environment, as well as potential risks of resistance in pathogens. Given the increasingly stringent demands placed on maximum residue levels, the ongoing development of novel and improved control strategies (including non-chemical means) is crucial. Possibilities include the rational development of diagnostic tests and/or safe anti-parasitic compounds, based on a better understanding of parasite genomes, the host-parasite relationships and the molecular biology of the parasites themselves. Hence, there are major gains to be made by improving our knowledge of such pathogens, which will lead to such outcomes. Whole genome sequencing of key parasitic helminths provides a critically important foundation for a wide range of fundamental areas (including functional genomics, genetics, proteomics, systems biology, molecular biology, physiology, biochemistry, ecology, epidemiology, pathology, and many more) underpinning applied areas.

In March 2004, a meeting funded by the Wellcome Trust was convened at the Sanger Institute to discuss whole genome sequencing of parasitic helminths. The global impact of helminth diseases, particularly through the subclinical and chronic infections and pathogenic effects they cause, make them extremely important candidates for genome sequencing. The purpose of the meeting was to provide advice on the prioritization of helminth species of medical and veterinary importance, considering experimental models, where scientifically justified. The criteria for selecting a particular species included: the clinical aspect (human or veterinary) or the availability of a model system; the comparative value of a taxon; the size of the scientific community interested in or working on a species, and its potential to enable and facilitate post-genome sequencing research. Several of the candidates identified at the meeting were nominated for sequencing. At this point, it was concluded that the scientific community should proceed with individual nominations for genome sequencing, while working together to build packages (such as groups of related species) that might gain support for more ambitious plans from the major funding agencies. Based on these recommendations, we now propose the whole genome sequencing of a number of key pathogens.

Although a range of genome projects on metazoan organisms and EST sequencing (cf. http://www.nematode.net/), have provided preliminary information, very little progress has been made on parasitic helminths of socio-economic importance. By contrast, the completion of the full genome sequence of the free-living nematode *C. elegans* (http://www.wormbase.org) has provided an extremely valuable resource and a solid platform for comparative genome analyses for various nematode groups, particularly those representing clade V (1). *C. elegans* is also a powerful system for genetic and molecular investigations, since it has a rapid life-cycle and is easy to maintain *in vitro* (2, 3). The karyotype is 2n = 12 (five pairs of autosomes and one pair of sex chromosomes) for the hermaphrodite, which appears to be consistent with a range of strongylid nematodes, and the genome of *C. elegans* contains ~20000 genes. Strongylid nematodes are, based on molecular phylogenetic analysis, considered relatively closely related to *C. elegans* (cf. (2)), as supported by findings from expressed sequence tag (EST) projects carried out on 40 nematode species other than *Caenorhabditis*, including 24 mammalian parasites, 14 plant parasites

and 2 free-living bacteriovores (4). Indeed, the strongylid data sets (clade V) have the highest similarity to *C. elegans* compared with the other taxa. Both the distribution of homology matches and their relative scores support the position of the Strongylida as the most related to *C. elegans* and further related to other nematode species. For instance, clusters from Strongylida species (represented by *Ancylostoma caninum* and *A. ceylanicum;* clade V) were evaluated relative to those from *Trichinella spiralis* (clade I), *Dirofilaria immitis* (clade III), *Strongyloides stercoralis* (clade IVA) and *Meloidogyne incognita* (clade IVB). For all of the non-Strongylida species, 20.7±9.6 of *blast* matches are nematode-specific, and 6.3±4.0 of matches are to non-nematode species. In contrast, clusters representing the Strongylida are more skewed towards the nematode category, with only 3.3±0.01% of matches being non-nematode whereas 31.3±0.06 are to nematode species. The ratio of nematode-only to non-nematode-only matches in *Ancylostoma* spp. (9.6±5.6) differs from other examined nematodes (3.3±2.4) by a statistically highly significant margin ($P<0.0001$, Students *t*-test). Furthermore, when searching the *C. elegans* proteome, only 45% of *T. spiralis* clusters had raw *blastx* scores more significant than 50, compared to 60-65% for *Strongyloides* and *Ancylostoma*, respectively. Further, the 15 most conserved gene products among *C. elegans*, *T. spiralis* and *S. stercoralis* had an e-value of 1*e*-153-1*e*-103 and 1*e*-243-1*e*-187, respectively, while the top 15 *A. ceylanicum* scores ranged from 1*e*-293-1*e*-179 (5-7). The same pattern is seen for percent identity values as for e-values. Therefore, extrapolation from the biology of a well-studied nematode, such as *C. elegans* will be of the highest benefit when studying evolutionary closely related species such as the members of the order Strongylida. Also, the chance that a cloned gene from a strongylid nematode has a homologue in *C. elegans* is high, with the exception of genes associated with host-parasite interaction. As there are no reliable culturing systems available for the propagation and maintenance of the entire life cycle of strongylid nematodes *in vitro*, *C. elegans* provides a very powerful system to test the function of homologous genes.

Given these genomic resources, future research will focus on the functions of genes defined by genome and large-scale EST sequencing and on questions regarding the genetic basis of parasitism. Functional characterization will require the application of genomic, proteomic and bioinformatic technologies that have been developed in other fields, including genome mapping strategies and DNA microarray analysis. These will be greatly aided by the comparison of *C. elegans* to whole genome sequences for key parasitic helminths. In spite of major technological advances, progress on the sequencing of key parasitic helminth genomes has been too slow. We propose to change that by undertaking whole genome sequencing for a range of key parasitic helminths of major human and/or animal health importance, focusing (in the shorter term) on comparative genomic and evolutionary analyses and (in the longer term) on developing improved methods for parasite diagnosis and control. The study of sequences from a variety of nematodes is also essential in proving evolutionary context to developmental and genetic studies in the model *C. elegans*.

3. PHYLOGENETIC CONSIDERATIONS PRIOR TO WHOLE GENOME SEQUENCING

The first, most detailed multigene phylogenetic analysis of the Strongylida (Chilton et al., submitted) has recently been undertaken using extensive ribosomal RNA gene data for species from all four suborders and 7 superfamilies of the Strongylida, to test existing hypotheses proposed for the relationships of the suborders. The study demonstrated that the Strongylida is a monophyletic assemblage (Figure 1), with only the Metastrongylina (but not the other suborders), forming a distinct monophyletic clade. In contrast to all previous hypotheses, one major lineage comprises taxa which occur exclusively in the pulmonary, circulatory or nervous systems of marsupial and eutherian mammals (i.e. the Metastrongylina and the genus *Dictyocaulus*), whereas a second lineage comprises species occurring in the gastrointestinal tracts (such as those proposed herein) or perirenal tissues of vertebrates, or in the lungs of birds. The findings reveal that the predilection site of adult bursate nematodes and host type reflect the evolutionary origin of the different taxonomic groups within the Strongylida.

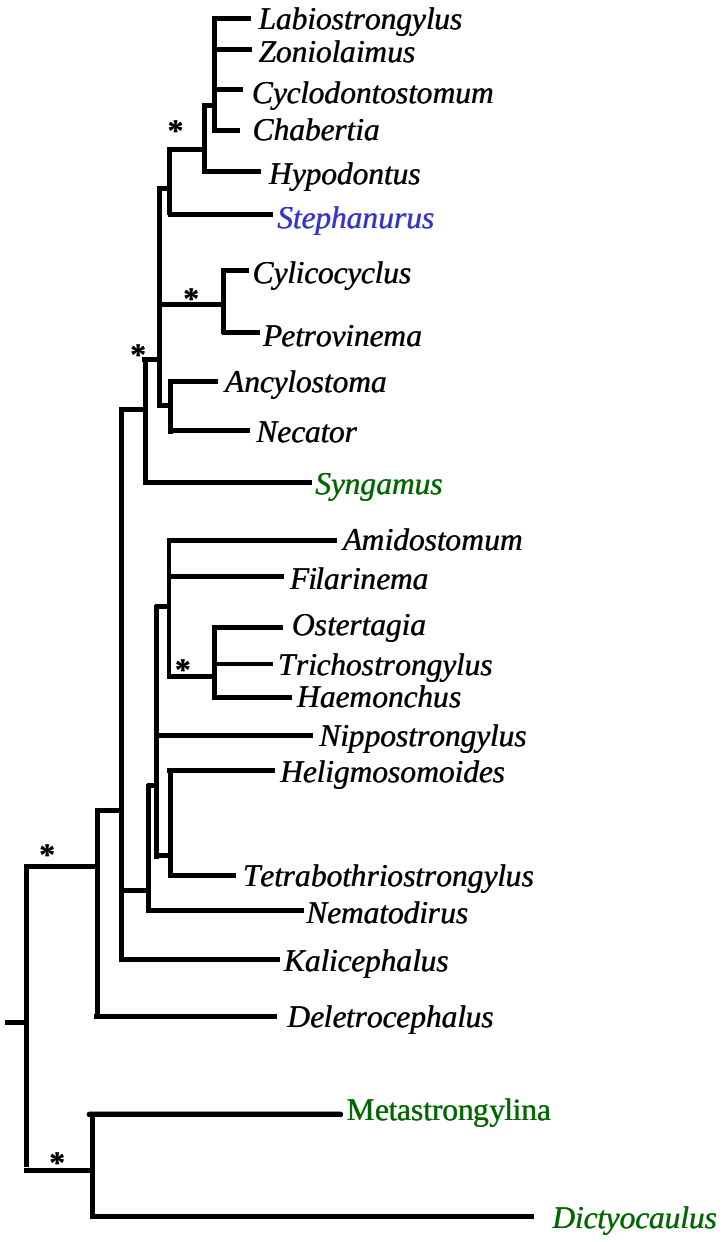

Figure 1. Molecular phylogeny of order Strongylida based on ribosomal RNA sequence data (modified from Chilton et al., submitted). Genera/groups proposed for genome sequencing are included. Parasites in gastrointestinal (black) and respiratory (green) track and perirenal tissue (blue). Bootstrap values >88% (*).

Furthermore, due to the paramount medical and veterinary importance of most members within the order Strongylida, as a part of a broader study of nematode transcriptomes, we have generated ~60,000 ESTs from 7 strongylids (Figure 2) which represent 83% of all non-*Caenorhabditis* originated clade V ESTs (8). In addition, the genomes of 8 nematode species of clade V have been sequenced or are in progress. Seven of the 8 species are non-strongylid members (Figure 2).

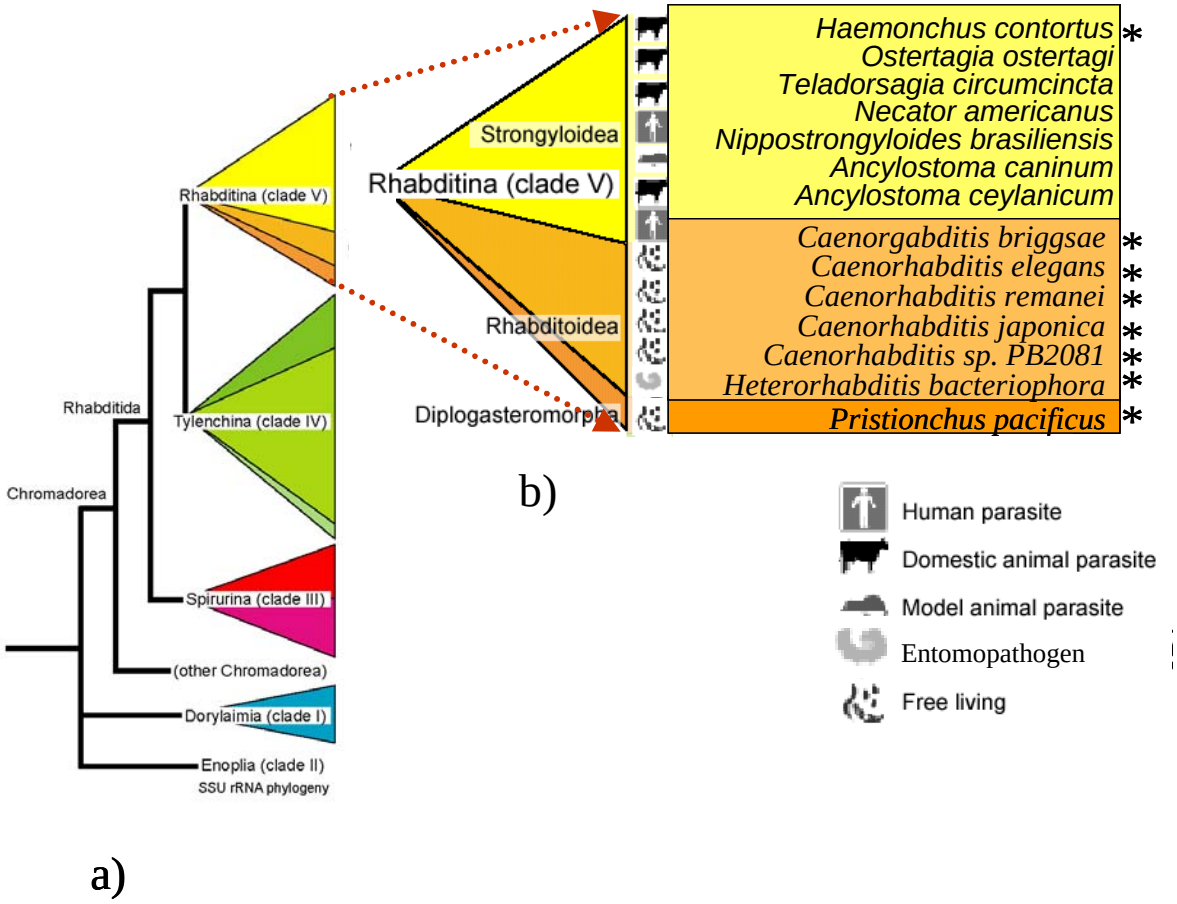

Figure 2. Phylum Nematoda. a) Species are grouped into major taxonomic groups based on phylogenic analysis of data for the small subunit of ribosomal DNA (SSU) (1). This relationship differs from 'traditional' phylogenies but is consistent with current morphological and developmental evidence. b) Species within clade V with either significant number of ESTs in the public database or genome projects (indicated by asterisks) are complete or underway.

4. SELECTED SPECIES FOR GENOME SEQUENCING

Since *C. elegans* is a member of clade V, comparative evolutionary developmental studies are particularly attractive in relation to the Strongylida. We propose sequencing of 20 species (Table 1), members of 4 out of the 7 subfamilies of the order Strongylida (Figure 1). When we categorized the species by priority, 2 tiers were identified:

- Tier 1, first priority: these parasites are clearly of major human and animal health importance;
- Tier 2, second priority: species whose genomes are worthy of sequencing, but for which lower coverage is the most cost effective yet still informative approach, since most of them have a close relative proposed in Tier 1.

Table 1. List of Strongylida species proposed for genome sequencing.

Superfamily	Tier 1	Tier 2
Ancylostomatoidea	*Ancylostoma caninum* *Necator americanus*	*Ancylostoma ceylanicum* *Ancylostoma duodenale*
Metastrongyloidea / Trichostrongyloidea	*Teladorsagia circumcincta* *Ostertagia ostertagii* *Nippostrongylus brasiliensis* *Trichostrongylus vitrinus* *Cooperia oncophora* *Nematodirus battus* *Dictyocaulus viviparus*	*Trichostrongylus colubriformis* *Trichostrongylus axei* *Hyostrongylus rubidus*
Strongyloidea	*Oesophagostomum dentatum*	*Oesophagostomum quadrispinulatum* *Oesophagostomum radiatum* *Oesophagostomum venulosum* *Oesophagostomum columbianum* *Chabertia ovina*

Hereafter, we provide information for all Tier 1 species. Parasite DNA isolates would be obtained together with collaborators (using monospecific lines where possible).

ANCYLOSTOMA CANINUM AND NECATOR AMERICANUS (ANCYLOSTOMATOIDEA)

Overview: The clade V strongylids are the parasites most closely related to *C. elegans* and the group most likely to benefit from its genome data. Strongylids include hookworms, blood feeding nematodes that infect one billion people, causing iron deficiency anemia and retarded physical and cognitive development in children (9). Zoonotic transmission to humans through contaminated soil results in cutaneous or visceral larva migrans, an extremely common condition in the United States and Europe. *Necator americanus* is the most prevalent species in human and *Ancylostoma caninum* in dogs is the most widely used model for hookworm infection (10). The genus is made up of very closely related species, which should make cross-genome comparisons relatively easy. The Bill and Melinda Gates Foundation is supporting a $16 million initiative to develop a hookworm vaccine at the Sabin Vaccine Institute and George Washington University. Genome size is estimated to be 56 Mb, based on close relatives. There are 24,750 ESTs available from several hookworms and a 1X coverage of the *A. caninum* genome (94,424 genome survey sequences) has been produced. Genomic information has already begun to accelerate

research progress on vaccines, diagnostics, drugs, and basic research, and the findings of the already broad scientific community has resulted in 3,658 hookworm-related publications worldwide.

Significance: Hookworm disease is extremely common in the tropics and sub-tropics, with an estimated 1.3 billion people infected. Hookworm is probably the most significant public health threat of all parasites except malaria, with a disease burden of 22 million disability adjusted life years lost when anemia is included, a burden comparable to measles and exceeding that of diabetes and lung cancer.

General description: Human hookworms (*Ancylostoma duodenale* and *Necator americanus*, also *Ancylostoma ceylanicum*) are intestinal blood feeders with the adult forms transmitting eggs in the feces. Larvae invade either through the skin, transiting to the lungs where they are coughed-up and swallowed, or by direct oral ingestion. Heavy infections can cause a range of symptoms, particularly in children, including chronic anemia, growth stunting, and impaired intellectual development. Hookworm induced anemia can also disrupt pregnancy, harming the health of the mother and causing intrauterine growth retardation, prematurity, and low birth weight. Hookworms (*Ancylostoma caninum*, *A. braziliense*, *A. tubaeforme*) also cause significant and common diseases of dogs and cats and can result in severe anemia and death in puppies and kittens. Zoonotic transmission to humans through contaminated soil results in cutaneous or visceral larva migrans, an extremely common condition in the United States and Europe.

Genome facts: Genome sizes are estimated to be in the 53-59 MB range, based on other clade V and Strongylida examples (11). Available ESTs: *A. ceylanicum* – 10,651 (estimated 3,840 genes), *A. caninum* – 9,331 (estimated 3,149), *N. americanus* – 4,766 (estimated 2,294 genes). G+C content, based on the EST contig consensus sequences, is 46-48%. There are also 94,429 genome survey sequences generated from *A. caninum*. Based on this coverage the repeat content is 27% and G+C content of the *A. caninum* genome is 43.2%.

Community and active labs: Several dozen labs worldwide study a variety of hookworm species including groups in the U.S., Europe, Japan, and endemic countries. Pubmed Key word Search Outcomes:

Hookworm	3,658 references
Ancylostoma	1,201
Necator	590
A. caninum	818
A. duodenale	783
A. ceylanicum	122

Curation: A description of *N. americanus* and *A. caninum* & *A. ceylanicum* ESTs has been published (7, 12). Hookworms cluster sequences, Gene Ontology mappings, Kegg enzyme mappings, and codon usage tables based on 387372, 305036 and 192756 codons for *A. ceylanicum*, *A. caninum* and *N. americanus* respectively, are available at www.nematode.net (13).

Exploitation: Needs in applied research and product development for hookworm control in humans include diagnostics, vaccines, and a wider array of anthelmintic drugs. Despite decades of research, no vaccines are available nor have any reached human trials. Promising avenues for vaccine development include secreted antigens such as ASP-1 (14) and intestinal antigens such as H11 (15), but until recently the field has been limited to antigens that could be biochemically purified. The Bill and Melinda Gates Foundation is supporting a $16 million initiative to develop a hookworm vaccine at the Sabin Vaccine Institute and George Washington University.

Benefits: The genome sequence provides a fundamental resource that ultimately will accelerate the development of novel methods to control these parasites in livestock, will define the genetic mechanisms underlying drug susceptibility and resistance with the possibility of extending the useful life of existing drugs and improved diagnostics in this area, and will provide the means for meaningful whole animal studies of the host-parasite interaction.

Sequencing Recommendation: It is important to generate as complete a genome sequence as possible from both *Necator* and *Ancylostoma* species, as they account for two most prevalent species. ***N. americanus*** is the most prevalent hookworm of humans, therefore most important from a public health standpoint. For the *Ancylostoma* species to be sequenced, we recommend ***A. caninum***. This genus is made up of very closely related species, which should make comparisons between them relatively easy. *A. caninum*

material is abundant and easily obtained (we have also already generated ~1x coverage of the genome as a part of an NIAID grant to RKW); has the most extensive molecular literature and community but lacks a small animal model (it is maintained in dogs); *A. ceylanicum* can be maintained in hamsters (making the *in vivo* work easier and cheaper) and is a human parasite, but it has been difficult to obtain sufficient amounts of DNA and has a less extensive literature and community; and *A. duodenale*, while the primary human parasite, is not easily maintained in laboratory hosts and lacks an extensive molecular literature.

TELADORSAGIA CIRCUMCINCTA (TRICHOSTRONGYLOIDEA)

Overview: Infections from gastrointestinal nematodes in livestock in countries with well-developed agricultural systems have always been an important production issue. *T. circumcincta* is one of the most prevalent sheep parasites in temperate climates. Economic losses are estimated in the millions of dollars yearly. Low production of meat, wool, and milk as well as the costs of anthelminthic treatment are the major losses in animal production worldwide, and appear to be a major constraint to efficient sheep production. Drug resistance in this parasite species is extensive and increasing. Such concerns have grown with the recent identification, in Europe and Australasia, of worms resistant to all 3 major classes of anthelmintic currently available. A vaccine against *T. circumcincta* would reduce the economic losses associated with infection, particularly in regions where multiple-drug resistant strains exist. A genome size estimate, based on flow cytometry, is 58.6 MB. There are 4,313 ESTs available, representing an estimated 1,655 genes which will be invaluable for gene finding and annotation.
Significance: *T. circumcincta*, a Clade V nematode, Order Stringylida, is economically the most important parasitic nematode of sheep in temperate regions and has developed multiple resistances to anthelmintic drugs. One major alternative to chemotherapy is vaccination, an approach already shown to be feasible in controlled experimental trials. The significance of sequencing *T. circumcincta* is not only its position as the most important ovine nematode, but also because powerful methods for studying the dynamics of the local immune response in this host-parasite system have already been developed (mainly by the Moredun Research Institute, Scotland; e.g. (16)). The power of these studies would be greatly enhanced by having the tools to monitor the molecular responses of the parasite to the changing intra-host environment by microarray and proteomic approaches. In addition, *T. circumcincta* is readily passaged and relatively inexpensive to maintain. *T. circumcincta* can also be used as a tractable model system for studying Ostertagiosis in cattle because it is comparatively cheap and these parasites are very closely related.
General description: *T. circumcincta* (aka *Ostertagia circumcincta*) - is a parasite of small ruminants (sheep and goats) infecting the abomasum. This pathogen has a major economic impact on sheep farming, causing weight loss (death) and decreased wool production. They are generally found in more temperate climatic zones. In terms of morphology, the nematodes are brownish and thread-like (up to 12mm in length). Worms develop in the gastric glands of the stomach (abomasum) and as they grow, destroy the glands. This, in turn, affects appetite, digestion and nutrient utilization.
Genome facts: Genome size is estimated to be 58.6 MB using flow cytometry (11). Available are 4,313 ESTs (estimated 1,655 genes; Parkinson et al., 2004). G+C content of protein coding exons is 48% (Mitreva et al., submitted).
Community and active labs: Several dozen labs worldwide study *T. circumcincta*, including USA, New Zealand, and Europe. Pubmed Key word Search Outcomes:

Teladorsagia	229 references
T. circumcincta	187

Curation: *T. circumcincta* cluster sequences are available at www.nematodes.org (17); and a codon usage table based on 194351 codons is available at www.nematode.net (13).
Exploitation: The availability of cheap effective chemotherapeutics has enabled the development of intensive livestock production systems that are wholly reliant upon effective chemical control. In recent years the increasing prevalence of drug resistance in small ruminants indicates lack of sustainability, which coupled with the need to reduce chemical usage in food producing animals and minimize any environmental impact has led to research examining other approaches to control. The available genome

sequence would have immediate application in genome-wide filtering to identify and prioritize molecular targets for alternative drug discovery to control parasitic (internal) or free-living (external) stages of the life-cycle. Furthermore, these datasets will be invaluable for characterizing mechanisms involved in drug resistance, such as changes in genes encoding a drug receptor or in gene expression.

Sheep develop immunity following infection with *T. circumcincta*, so vaccination is a viable concept for control. Identifying specific parasite molecules that may be involved in generating host immunity or transition to parasitism is accomplished by identifying differences between molecules found in free-living (pasture) stages from those in newly parasitic (gastric gland) stages via sensitive molecular biology techniques. Having a defined genome sequence will accelerate these studies, and opens the door for comprehensive genome-wide analysis of gene expression, e.g. by microarrays. The identified candidates are of potential use in trials for recombinant vaccine for teladorsagiosis. Furthermore, previous work identified digestive enzymes from the intestinal cells of closely related *Haemonchus contortus* as strong vaccine candidates. By comparative analysis the identified equivalent proteins from *Teladorsagia* or *Ostertagia* has been shown to be only modestly protective, probably because, unlike *Haemonchus*, adult *Teladorsagia* do not feed on blood and so in vaccinated animals do not ingest enough host antibody to seriously affect nutrient uptake. In contrast, their fourth stage larvae may be much more vulnerable. By inhabiting and thereby damaging the gastric glands they are exposed to and ingest higher concentrations of antibody. Moreover, because they grow very rapidly they should be highly sensitive to digestive interference. The defined genome sequence hence, ultimately would enable a better understanding of the host-parasite interaction, differences between life cycle stages and subsequently identifying better candidates for vaccine trials.

<u>*Benefits*</u>: Having a defined genome sequence will accelerate the development of novel methods to control these parasites in livestock, defining the genetic mechanisms underlying drug susceptibility and resistance with the possibility of extending the useful life of existing drugs and improved diagnostics in this area, as well as providing the means whole animal studies of the host-parasite interaction. Furthermore, it opens the door for comprehensive genome-wide analysis of gene expression. The sequence of *T. circumcincta* will be important in the annotation of the *Ostertagia* genome (major parasite of cattle).

OSTERTAGIA OSTERTAGI (TRICHOSTRONGYLOIDEA)

<u>*Overview*</u>: *O. ostertagi* is economically the most important parasite of cattle in the United States and most other countries. These brown stomach worms become embedded in the lining of the fourth stomach (abomasum) of cattle in late spring. Cattle under 3 years and some older animals are especially vulnerable. The damage produced in the abomasum can cause decreased acid production, loss of appetite, diarrhea, weight loss and even death. Most modern de-wormer drugs are effective in removing these embedded (inhibited) larvae. However, resistance has started to appear and recent experimental work in Belgium indicates that this can arise within as few as 6 generations of ongoing exposure to a drug. Multiple drug resistance in a very close relative *Teladorsagia circumcincta* has recently been detected in the field. The genome size estimate is ~58 MB, based on *T. circumcincta* (aka *Ostertagia circumcincta*), whose genome is 58.6 MB and on *H. contortus* at 52 MB, based on flow cytometry. There are 7,006 ESTs, representing 2,564 genes, and a broad scientific community whose research has resulted in 1,360 *Ostertagia*-related publications worldwide.

<u>*Significance*</u>: The abomasal nematode *O. ostertagi* is a Clade V nematode, Order Stringylida, and is the most economically damaging endoparasite of cattle throughout temperate regions of the world. On a global basis these worms are the major cause of parasitic gastritis (Ostertagiosis) of ruminants. Damage to the abomasum results in lost productivity throughout the life of the animal. Economic losses are estimated to be in the millions of dollars every year. The importance of this parasite is that protective immunity to ostertagiosis develops slowly, making vaccine development a more pressing priority.

<u>*General description*</u>: *O. ostertagi* are small (~10mm) reddish brown worms found in the abomasum of ruminants. The life cycle is direct - preparasitic larvae are entirely free living. No migration occurs inside the definitive host. Infected animals have acute or chronic abomastitis. Acutely affected animals can

appear normal and die suddenly. Chronically infected animals are often emaciated and have a poor haircoat.

Genome facts: The genome size estimate is ~58 Mb, based on sister nematode *Telodorsagia circumcincta* (aka *Ostertagia circumcincta*) whose genome is 58.6 MB and other trichostrongylid species such as *H. contortus* with a 52.6 MB genome size (11). There are 7,006 ESTs in the dbEST division of GenBank, representing 2.564 genes, based on clustering (17). G+C content for protein coding exons is 48% (Mitreva et al., submitted).

Community and active labs: Several dozen labs worldwide study *O. ostertagi* reflecting their global importance. Pubmed Key word Search Outcomes:

Ostertagia	1,360 references
O. ostertagi	890

Curation: *O. ostertagi* EST clusters are available on www.nematodes.org (18), and a codon usage table based on 222616 codons is available at www.nematode.net (13).

Exploitation: The majority of the *O. ostertagia*-related studies to date have involved work on individual genes. The genome information generated will have immediate and urgent application in the identification of novel target molecules for the control of these parasites by vaccination or by drug development. Fourth stage *Ostertagia* probably contains membrane proteins in addition to or different from those found in the adult stage, representing new targets for the gut antigen approach to vaccination. These larvae are now targets for vaccine development given that they grow rapidly and larval extracts enriched for gut proteins, including proteases, induce high levels of protective immunity in vaccinated animals given an experimental challenge infection. Determination of the genome sequence will open the door to applying functional genomics tools (such as proteomic analyses and microarrays) to identify targets that are expressed throughout the parasite lifecycle.

Benefits: The benefits from an *Ostertagia* genome sequence would be accelerating the development of novel methods to control these parasites in livestock, defining the genetic mechanisms underlying drug susceptibility and resistance with the possibility of extending the useful life of existing drugs and improved diagnostics in this area, as well as providing the means for meaningful whole animal studies of the host-parasite interaction. The sequence will aid in the annotation of already ongoing genome sequencing (funded by Wellcome trust) of the blood–feeding nematode *Haemonchus contortus* (member of Strongylida, clade V; (1)), and be of value for comparative analysis within the trichostrongyloid nematodes (Figure 2).

NIPPOSTRONGYLUS BRASILIENSIS (TRICHOSTRONGYLOIDEA)

Overview: Nippostrongylus brasiliensis is a gastrointestinal parasite of rats with a similar lifestyle and morphology to the human hookworms *Necator americanus* and *Ancylostoma duodenale*. The transmission is direct and thus does not require an intermediate host. Adult worms live in the intestine of the host and eggs are passed into the faeces. The intestinal phase of infection has been extensively exploited as a model system to define immune responses to enteric helminths, while the lung phase of infection is used to study pulmonary inflammation as a paradigm for asthma. In addition, the model is now being applied towards hookworm vaccine development, as the different parasite stages are found in similar host tissues, and the ability to complete the life cycle in mice makes it highly amenable to immunological analysis. For these reasons, there is an extensive knowledge database (1259 publications) for this parasite, and a real opportunity to link this knowledge to that derived from genome sequencing, with downstream analyses of gene expression as infection progresses, for example. To date, there are 1250 ESTs available, representing ~750 genes. RNAi has been developed in this species at a number of sites worldwide, and success has been reported with a number of different genes (e.g. (19)).

Significance: The intestinal phase of infection has been extensively exploited as a model system to define immune responses to enteric helminths, while the lung phase of infection is used to study pulmonary inflammation, with relevance for asthma. The potential of *N. brasiliensis* as an experimental system for functional genomics has been greatly enhanced by the demonstration of successful RNAi knockdown in

this species (19). We propose that *N. brasiliensis* represents the most appropriate model parasitic species whose genome sequence will greatly enhance the understanding of the immune response to helminth infection and research on allergic disorders such as asthma.

General description: Adult worms live in the intestine of the host and eggs are passed into the faeces. Outside the host, the eggs hatch and the L1 larvae continue to develop. After two further moults the infective L3 stage is produced. At this stage the larvae crawl up grass stems and are able to penetrate the skin of the rodent host. The larvae migrate first to the heart, then to the lungs, where they penetrate the pulmonary alveoli and develop to the L4 stage. Parasites make their way via the trachea and esophagus to the intestinal jejunum, where they develop to mature adults. Both L3 and adult *N. brasiliensis* can be maintained *in vitro* in different culture systems (20), and immune response-related work is mostly carried out in mice.

Genome facts: Genome size is estimated in the 53-59 MB range, based on other clade V and *Strongylida* examples (11). Available ESTs: *N. brasiliensis* – 1,234 (estimated 742 genes). G+C content for protein coding exons, based on the EST contig consensus sequences, is 50%. RNAi is effective in L3 and adult *N. brasiliensis*.

Community and active labs: Several dozen labs worldwide study *N. brasiliensis*. Pubmed Key word Search Outcomes:

Nippostrongylus	1,259 references
Nippostrongylus brasiliensis	1,259

Curation: A description of *N. brasiliensis* ESTs has been published, with emphasis on secreted proteins (21). *N. brasiliensis* cluster sequences are available at www.nematodes.org (18) and a codon usage table based on 75934 codons is available at www.nematode.net (13).

Exploitation: N. brasiliensis has been utilized as a model to study diverse phenomena related to understanding the interaction between the parasite and the host. However, the majority of *N. brasiliensis*-related studies to date have involved work on individual genes, and a small EST dataset representing ~742 genes has been produced. The availability of the *N. brasiliensis* genome will be crucial for identifying the full repertoire of gene products involved in infection, tissue migration, immuno-modulation, and a plethora of other facets of host-parasite interactions that will ultimately enable us to understand the specific adaptations that characterize parasitism in gastrointestinal nematodes. It will also provide a major impetus to efforts to develop a vaccine for human hookworm infection.

Benefits: The benefits would be an immediate acceleration of research efforts aimed at understanding host-parasite interactions and the development of novel methods to control nematode species that colonize the alimentary tract of their mammalian hosts. This is particularly timely for vaccine development, and for assessing the potential of different gene products expressed in both systemic and enteric phases of infection. The combination of genome sequence and RNAi knockdown in this species provide the ability to perform genome-wide RNAi screenings, and set the stage for comparative analysis with *C. elegans* phenotype information, defining the differences due to the complex host-parasite interaction of the parasitic versus free-living species.

OTHER TRICHOSTRONGYLOID NEMATODES: TRICHOSTRONGYLUS, COOPERIA AND NEMATODIRUS (TRICHOSTRONGYLOIDEA)

Overview: These are economically important gastro-intestinal parasites of sheep, goats and other ruminants worldwide, in different climatic zones. Most are relatively host specific, but some cross host species boundries (e.g., *Trichostrongylus axei*). Some species are zoonotic (*Trichostrongylus*). Control of members of this genus has become difficult worldwide due to widespread drug resistance to all major classes of anthelmintic. There is a large, active research community working on the trichostongylid nematodes within clade V. This group has arguably been the subject of a substantial research effort on vaccine and diagnostic test development, control strategies, drug efficacy trials and drug resistance. Some species are important models of parasitism as well as important pathogens in their own right. The adult parasites are small (<1 cm), with relatively high reproductive potential, meaning that large numbers of

parasite material can be produced experimentally for biochemical, immunological and molecular studies. Importantly, well-defined strains of these parasite species are available in different laboratories around the world, including our own. RNAi knockdown has been achieved for *Trichostrongylus colubriformis* (22) and there has also been recent success with *T. vitrinus* (EL-Osta, Hu and Gasser, unpublished data).

Significance: *Trichostrongylus* nematodes are of major veterinary and economic importance, causing substantial production losses (through a reduction in weight gain, meat and/or milk production) to farmers and exacerbating the global food shortage problem. Of most importance are the parasites of grazing livestock which cause significant economic problems to agriculture and detrimental effects on animal welfare. Also, a recent report, commissioned by the Department for International Development (DFID), has drafted a list of the top 80 animal diseases that have a major impact on the poor in the developing world. The trichostrongylid nematodes were ranked as a group (since they mainly occur as mixed infections) and were top of that list (i.e. considered to have a greater impact than any other disease of domestic animals). This group of parasites are also amongst the most economically important diseases of livestock in the developed world.

General description: There is a range of key species in livestock. The relative importance of particular species varies for different regions of the world and, consequently, the relative priority of each will differ. The disease syndrome observed varies depending on the species predominating. For instance, *T. vitrinus* and *T. colubriformis* cause enteritis (scouring) and major economic losses to the sheep industries globally. Subclinical infections substantially reduce productivity in livestock units. These 2 species are the economically most important and tractable species from each of the remaining genera of primary importance.

Genome facts: The genome sizes are likely to be similar to that of other *Trichostrongylus* species with estimated genomes by flow cytometry at 53-59 Mb (11). All the indications are that the genome sizes of all species within Trichostrongyloidea are in this range. Various smaller EST data sets are available. As indicated, RNA interference is effective for selected genes in selected species. Trichostrongylid nematodes of livestock provide an important resource in that they can be readily passaged.

Community and active labs: Many labs worldwide investigate gastrointestinal (trichostrongyloid) nematodes of livestock, reflecting their global importance.

Pubmed Ket word Search outcomes:

Trichostrongylus	1600 references
Cooperia	690
Nematodirus	460

Curation: ESTs are available at dbEST. The individual clones are becoming available to the community and publications describing the datasets are in preparation.

Exploitation: The genome information produced will have immediate utility the identification of novel target molecules for the control of these parasites by vaccination or by drugs. In addition, these datasets will be invaluable for characterising developmental processes and the mechanisms of drug resistance. Specific laboratories have available selected lines of some species with defined resistance to all the currently available anthelmintics. This will prove an invaluable resource for comparative genomics to seek common genes involved with this problem. Having genome sequences opens the door to comprehensive and meaningful analyses of evolutionary relationships, gene expression and applied areas, such as drug development and diagnostic tests.

Benefits: The benefits would be to accelerate the development of novel methods for the control of these parasites in livestock, defining the genetic mechanisms underlying drug susceptibility and resistance with the possibility of extending the useful life of existing drugs, improved diagnostics in this area as well as providing the means for meaningful whole animal studies of the host-parasite interactions. Importantly, detailed insights into the molecular developmental processes could lead to specific control methods, which would have a significant impact on herd or flock health through accumulated reduction in pasture contamination and hence the ingestion of infective parasite stages (larvae). Thus, detailed knowledge of the genome will also benefit fundamental studies focused on novel methods of parasite control through safe compounds or vaccines.

DICTYOCAULUS VIVIPARUS (TRICHOSTRONGYLOIDEA/METASTRONGYLOIDEA)

Overview: Lungworms of the genus *Dictyocaulus* (family Dictyocaulidae) are key parasitic nematodes causing pathological effects in different ruminant hosts and major economic losses worldwide, due to clinical disease, particularly in young animals. In cattle, *D. viviparus* is the bovine lungworm, which causes a severe and frequently fatal bronchitis (known colloquially as 'husk') and is of major importance in many countries. Severe cases of dictyocaulosis lead to emphysema and pneumonia – heavy infections can lead to complications can cause a mortality rate of 20% or more among affected animals. It is important to generate a genome sequence from *D. viviparus*, given its major economic importance and the major controversy surrounding the phylogenetic relationship of the lungworms in relation to other members within clade V. This genus is made up of closely related species, which should make comparisons between them relatively easy. Labs studying *D. viviparus* are mainly located in Europe, and cattle model is available in Sweden, the Netherlands and Germany. Lungworm-related research has resulted in nearly 1,500 publications.

Significance: *Dictyocaulus* species are of major veterinary and economic importance, causing substantial production losses to farmers. The disease caused by lungworm is parasitic bronchitis, also called husk or hoose, which is characterized by rapid shallow breathing and coughing. Severe cases lead to emphysema and pneumonia – heavy infections can lead to complications that can cause a mortality rate of 20% or more among affected animals.

General description: All members of the Dictyocaulidae have direct life-cycles. Adults of *D. viviparus* live in the airways and lay eggs containing larvae that live on the *Pilobolus* fungus, common in cattle faeces. They molt twice and the L3s are ingested by grazing cattle. Larval stages can remain inhibited in the lungs for up to 150 days. The time between infection of the cattle and the earliest time at which this parasite's eggs can be recovered in the feces is ~1 month. The genus is made from closely related species of which the cervid lungworm, considered to represent *D. eckerti*, is the parasite of prime economic importance in farmed red deer. Light infestation causes subclinical disease associated with production losses, and heavy infestation frequently is fatal. *D. filaria*, a lungworm infecting ovids is also of economic importance in small ruminants, such as sheep, particularly in farming areas of Eastern Europe, the Middle East and India. Historically, the control of dictyocaulosis in cattle in Europe has been through the use of an irradiated larval (L3 stage) vaccine (Bovilis®; Huskvac®). However, an increased use of anthelminthics has replaced vaccination on many farms. The control of dictyocaulosis in farmed ruminants now relies heavily on the use of anthelminthics, and no recombinant vaccine is currently available.

Genome facts: The genome sizes are likely to be similar to that of other trichostrongylus species with estimated genomes by flow cytometry at 53-59 Mb (11). All the indications are that the genome sizes of all species within Trichostrongyloidea are in this range. Several hundred nucleotide and protein sequences are available in the public databases.

Community and active labs: Number of labs, mainly in Europe, study *D. viviparus* and other lungworms.

Keyword	Search outcome
Dictyocaulus	830 references
Dictyocaulosis	560

Exploitation: Needs in applied research and product development for lungworm control in cattle include diagnostics, vaccines and anthelmintics. In spite of decades of research, no recombinant vaccines are available. As for hookworms, there is excellent prospect for the development of a recombinant vaccine against *D. viviparus*, because it is one of the few examples where effective protection in cattle can be achieved using a live, irradiated vaccine.

Benefits: The genome sequence would provide a fundamental resource that ultimately will accelerate the development of novel methods for the control of these parasites in livestock, will define the genetic mechanisms underlying drug susceptibility and resistance with the possibility of extending the useful life of existing drugs and improved diagnostics in this area, and will provide the means for meaningful whole

animal studies of the host-parasite interaction. Like hookworms, vaccination with live, irradiated third-stage larvae achieves protection in cattle. As the live vaccine has a limited shelf life and is expensive to produce, there is excellent prospect for the development of a recombinant vaccine. Also, there is a major need for the development of a sensitive diagnostic test for the diagnosis on an individual animal basis. The availability of the genome sequence will also enable, for the first time, the direct comparison with other members of the Strongylida, in order to address the controversy about the phylogenetic position of the lungworms in relation to the Trichostrongyloidea. The availability of the sequence will also provide a basis to study other species of lungworm.

OESOPHAGOSTOMUM DENTATUM (STRONGYLOIDEA)

Overview: Nodule worms of the genus *Oesophagostomum* are economically important intestinal parasites that are mainly host specific, but some cross host-species boundaries. They cause severe pathological lesions in the large and/or small intestines of non-primate and primate hosts. During infection, pigs display a reduction in appetite and a reduced growth rate and feed conversion efficiency in the period of nodule formation. There are selected research groups working on these nematodes. Recently, this parasite group has been the subject of a significant research effort on control strategies, drug efficacy trials and drug resistance. Some species are important models to study parasitism as well as important pathogens in their own right. The adult parasites are small (1-3 cm), with high reproductive potential, meaning that large numbers of parasite material can be produced experimentally for biochemical, immunological and molecular studies. Importantly, well-defined strains of these parasite species are available in selected laboratories (e.g., in Austria and Denmark). Recent investigations have demonstrated clearly that *Oesophagostomum dentatum* is a powerful system (*in vivo* and *in vitro*) for studying molecular biological aspects of strongylid nematodes (e.g. (23)). *O. dentatum* represent the most appropriate species for sequencing and assembly, because a unique model and extensive parasitological and molecular expertise exist. This species have been chosen also because a monospecific line and *in vitro* culture facilities/expertise are available. Studies related to *Oesophagostomum* spp. and oesophagostomiasis/osis have resulted in more than 1,000 publications.
Significance: Common parasites of ruminants, pigs, primates and rodents. *O. dentatum*, and few other *Oesophagostomum* spp., are found in domestic animals, and therefore are considered to be of pathogenic importance. We chose *O. dentatum* as a reference model since larvae can be produced *in vitro* in high numbers and exsheathment can regularly be induced. In addition, infection of pigs with *O. dentatum* is a major cause of economic losses in pig productions.
General description: There is a range of key species in livestock. The relative importance of particular species can vary for different countries and, consequently, the relative priority of each will differ. The porcine nodule worm, *O. dentatum* is of economic importance in large piggeries and provides a unique model system for studying fundamental aspects of reproductive biology in strongylid nematodes. Several characteristics, including its short life cycle and the ability to maintain worms in culture *in vitro* (24), show that *O. dentatum* is a valuable model system to investigate reproductive processes. *O. dentatum* has a direct life cycle. Fertilized eggs are passed in the feces into the environment, where they develop to infective third-stage larvae (L3) which, upon ingestion by the pig, exsheath in the stomach or small intestine, burrow into the mucosa of the large intestine and continue their development in the submucosa. This cycle is in contrast to the more intensively studied trichostrongylids of ruminants (*Haemonchus*, *Ostertagia*, *Trichostrongylus*) which live in the more anterior part of the gut and exsheath in the stomach. It is the transition from L3 to L4 that induces an immune response typified by the presence of raised nodular lesions (aggregations of neutrophils and eosinophils surrounding individual larvae in the intestinal mucosa). The L3 molt to the L4 stage within the nodules and emerge into the intestinal lumen where they develop to the fifth-stage larva (L5) and adult. The pre-patent period for *O. dentatum* is 20±1.4 days, although significantly longer periods have been described.

Importantly, Joachim and co-workers have developed a culture system which allows the development of L3 to the L4 stage and the maintenance of adult *O. dentatum in vitro*. Although currently

used for investigations of the physiological processes relating to the synthesis and mode of action of eicosanoids produced by larval stages, this system has applicability to other investigations. For example, the ability to carry out manipulations of the culture environment (i.e., testing of worm extracts or chemicals) could be valuable for the identification and functional analysis of genes activated in response to these manipulations, and may also provide a valuable tool for the characterization of molecular processes and mechanisms.

Genome facts: The genome sizes are likely to be similar to other Strongylida species with estimated genomes by flow cytometry at 53-59 Mb (11).

Community and active labs: Selected labs in Austria (Joachim), Denmark and Australia (Gasser) are experts in *O. dentatum* biology, *in vitro* cultivation and molecular biology, and have monospecific lines of this parasite.

Pubmed Key word	Search outcome
Oesophagostomum	725 references
Oesophagostomiasis/osis	360

Curation: Small numbers of ESTs are available and clustered sequences have been deposited in GenBank. The individual clones will be made available to the community and publications describing the datasets are in preparation.

Exploitation: The genome information produced will have immediate utility in the identification of novel target molecules for the control of these parasites by drugs. In addition, these datasets will be invaluable for characterizing developmental processes and the mechanisms of drug resistance. Specific laboratories have available selected lines of *O. dentatum*, and extensive expertise is available in Austrian (Joachim lab) and Australian labs (Gasser lab). This will prove an invaluable resource for comparative and functional genomics, as has already been demonstrated.

Benefits: As for other strongylids, the benefits would be to accelerate the development of novel methods for the control of these parasites in livestock, defining the genetic mechanisms underlying drug susceptibility and resistance with the possibility of extending the useful life of existing drugs, improved diagnostics in this area as well as providing the means for meaningful whole animal studies of the host-parasite interactions. Importantly, detailed insights into the molecular developmental processes could lead to specific control methods, which would have a significant impact on herd health through accumulated reduction in pasture contamination and hence the ingestion of infective parasite stages (larvae). Thus, detailed knowledge of the genome will also benefit fundamental studies focused on novel methods of parasite control through safe compounds or vaccines. The major advantage of *O. dentatum* is the availability of well-defined *in vivo* and *in vitro* systems, which are crucial for any functional genomic and molecular biological work.

5. POST– GENOMIC ACTIVITIES

The revolution in genomics, genetics and proteomics provides unique opportunities and prospects for post-genomic sequencing research. It is clear that functional genomic resources such as microarrays, RNAi, cDNA archives and libraries offer platforms to exploit the whole genome sequence data. The proposed genome sequence resources will advance our understanding of the molecular biology of pathogens using cutting-edge technologies and will lead to novel approaches for parasite diagnosis, prevention and control as well as to applications of socio-economic impact.

The socio-economic benefits flowing from this proposal will be: (i) enhanced focus on animal and human health biotechnology through the development of diagnostic assays, anti-parasite compounds and/or vaccines; (ii) improved and sustainable control of important parasites with decreased risk of drug resistance; (iii) development of a technology platform for further applications in genetics and genomics of pathogens with global significance; (iv) capturing the benefits and outcomes of fundamental research and strengthening the links between fundamental and applied research; and (v) enhancing the quality and quantity of scientifically skilled people globally.

6. WHOLE GENOME SEQUECING STRATEGY

Prior to sequencing, the genome sizes will be re-measured using flow-sorted nuclei stained with propidium iodide. *C. elegans* and *D. melanogaster* will be included as a standards.

Since the *C. elegans* genome sequence is available for comparative purposes, we propose 1) to sequence the genomes of species in Tier 1 to 6-fold coverage in plasmids, and to construct and end sequence fosmids to 0.2x coverage, and of species in Tier 2 to 2-fold coverage in plasmids; 2) species in Tier 1 will undergo two rounds of directed sequence improvement ("pre-finishing") once the initial whole genome sequence (WGS) reads are assembled; 3) automated annotation of the draft sequences will follow the final genome assembly (see section 7).

To facilitate better gene prediction and to add depth to the analysis of structure-function for species without substantial number of ESTs, up to 10,000 ESTs and 20,000 ESTs will be generated for Tier 1 and Tier 2 species, respectively (for several species there are ~10,000/species ESTs already available). To increase the diversity of identified transcripts stage- or tissue-specific cDNA libraries will be constructed.

7. AUTOMATED ANNOTATION AND DATA DEPOSITION

We will use the draft sequence to analyze the repetitive content of the genome, predict genes using a variety of different methods (*ab initio* and evidence-based), merge the different predictions into a final gene set and provide a web-based genome browser to display the annotations. Although we call this "automated" annotation, there is some manual involvement, mainly during the set up of the automated pipeline, in which parameters are tested and the output evaluated. The term "automated" is used to distinguish what we propose from a truly manual annotation effort, where each gene prediction is evaluated and adjusted manually, and which would cost orders of magnitude more.

Repeat characterization

The first step in gene prediction is masking repeated sequences. We use RECON (25) for automated, *de novo* identification and classification of repeat sequence families. Initial RECON results are checked to remove gene families and further classified before being used to mask the genome. In addition to RECON repeats, the genome is also masked for simple and low-complexity repeats using Repeatmasker. Local tandem and inverted repeats are annotated but not masked.

Gene Prediction

Protein coding genes are predicted using a combination of *ab initio* and evidence-based prediction methods. We have access to several *ab initio* predictors (Genefinder, FgenesH, SNAP, Genscan, GeneMark.hmm and HMMGene) and we test and evaluate new ones as they become available. Most of these require organism-specific tables or training. Here, we will evaluate predictions using *C. elegans* tables and then decide if species-specific tables will need to be generated or purchased. Evidence-based gene predictors available to us are FgenesH+ and Genewise, which use protein alignments, Eannot (26), which uses EST, mRNA and protein alignments, and Fgenesh2, which uses similarity between two closely-related genomes. We use a modified Ensembl pipeline to generate the alignments, requiring the building of appropriate mRNA, EST and mRNA databases. If an appropriate genome is available for predicting genes using FgenesH2, we will first determine the appropriate syntenic regions to provide to FgenesH2. We will also generate a consensus gene set, choosing the best prediction at each locus based on the underlying data and the specificity of each prediction method.

Transfer RNA genes will be predicted using tRNAscan (27). Other non-coding RNAs will be predicted based on homology to known RNA genes.

Data Display

The data will be made available to the public by deposition of traces within 24 hours of data collection and of assemblies > 1kb. All the assemblies will be added to the WU-GSC ftp site (ftp://genome.wustl.edu/pub/seqmgr). We will also make the data available to the public on a web-based

genome browser (GBrowse), to enable easier design and interpretation of experiments by the scientific community.

8. ESTIMATED COST AND TIME FRAME

Estimated total cost for this project would be $6.7M. The calculation is based on an average genome size of 55 Mb, and includes the following:

1. 6.2X coverage in a combination of plasmid (6X) and fosmid (0.2X) reads for 10 Tier 1 genomes
2. 2X coverage in plasmid reads for 10 Tier 2 genomes
3. Two rounds of pre-finishing for 10 Tier 1 genomes ($512,000 per genome)
4. Genome assembly, annotation and data deposition
5. cDNA library construction and EST generation for Tier 1 and Tier 2 species

The project should be completed in 5 years.

9. SUPPORT FOR THIS PROPOSAL

The Helminth Genome Consortium (2004 Hinxton Meeting Group)

- Rick Maizels (Edinburgh Univ), Andy Tait (Glasgow Univ), Bart Barrell (Sanger Institute), Pauline Beattie (Wellcome Trust), Matt Berriman (Sanger Institute), David Bird (North Carolina), Mark Blaxter (Edinburgh Univ.), Klaus Brehm (Wrzburg, Germany), Najib El Sayed (TIGR), Elodie Ghedin (TIGR), Neil Hall (Sanger Institute), David Johnstone (Natural History Museum, London), Rick Komuniecki (Toledo OH), Brian Kerry (Rothamsted, UK), David Knox (Moredun Institute), Phil LoVerde (Buffalo), James McCarter (Divergence, GSC), Alan Scott (Johns Hopkins), Makedonka Mitreva (GSC)

The Society of Nematology Genomics White Paper Group, www.nematologists.org

- David Bird, Mark Blaxter, Kelly Thomas, Paul Sternberg, Makedonka Mitreva and James McCarter

Wormbase and the Caenorhabditis elegans Community, www.wormbase.org

- Paul Sternberg (Caltech), Richard Durbin (Sanger Institute), John Spieth (Washington Univ.), Lincoln Stein (Cold Spring Harbor)

Edinburgh University and Sanger Institute Nematode Projects, www.nematodes.org

- Mark Blaxter, John Parkinson

The Nematode Tree of Life (NemAToL) Group, http://nematol.unh.edu/

- Jim Baldwin, Kelley Thomas, Steve Nadler, Paul De Ley, Irma De Ley, David Fitch, Byron Adams, Peter Mullin, Dan Bumbarger, Sergei Subbotin, Sven Bostrøm, Tom Powers, Manuel Mundo, Jay Burr, Robin Giblin-Davis, Krystalynne Morris, Erik Ragsdale, and Einhard Schierenberg, Walter Sudhaus, Karin Kiontke, Patricia Stock, Eric Hoberg, Alexander von Lieven, Ramon Carreno, Gaëtan Borgonie, and Mark Blaxter

10. REFERENCES

1. Blaxter, M. L., De Ley, P., Garey, J. R., Liu, L. X., Scheldeman, P., Vierstraete, A., Vanfleteren, J. R., Mackey, L. Y., Dorris, M., Frisse, L. M., Vida, J. T. & Thomas, W. K. (1998) *Nature* **392,** 71-5.
2. Gasser, R. B. & Newton, S. E. (2000) *Int J Parasitol* **30,** 509-34.

3. Newton, S. E., Boag, P. R. & Gasser, R. B. (2002) in *World Class Parasites: Volume 2. The Geohelminths: Ascaris, Trichuris and Hookworm*, ed. Kennedy, C. V. H. a. M. W. (Kluwer Academic Press, Boston), pp. 235-268.
4. Mitreva, M., Blaxter, M.L., Bird, D.M., McCarter, J.P. (2005) *Trends in Genetics* **21,** 573-81.
5. Mitreva, M., Jasmer, D. P., Appleton, J., Martin, J., Dante, M., Wylie, T., Clifton, S. W., Waterston, R. H. & McCarter, J. P. (2004) *Mol Biochem Parasitol* **137,** 277-291.
6. Mitreva, M., McCarter, J. P., Martin, J., Dante, M., Wylie, T., Chiapelli, B., Pape, D., Clifton, S. W., Nutman, T. B. & Waterston, R. H. (2004) *Genome Res* **14,** 209-220.
7. Mitreva, M., McCarter, J. P., Arasu, P., Hawdon, J., Martin, J., Dante, M., Wylie, T., Xu, J., Stajich, J. E., Kapulkin, W., Clifton, S. W., Waterston, R. H. & Wilson, R. K. (2005) *BMC Genomics* **6,** 58.
8. McCarter, J. P., Bird, D.M., Mitreva, M. (2005) *J Nematol* **in press**.
9. de Silva, N. R., Brooker, S., Hotez, P. J., Montresor, A., Engels, D. & Savioli, L. (2003) *Trends Parasitol* **12,** 547-551.
10. Hotez, P. J., Hawdon, J. M., Cappello, M., Jones, B. F., Ghosh, K., Volvovitz, F. & Xiao, S. H. (1996) *Pediatr Res* **40,** 515-21.
11. Leroy, S., Duperray, C. & Morand, S. (2003) *Mol Biochem Parasitol* **128,** 91-93.
12. Daub, J., Loukas, A., Pritchard, D. I. & Blaxter, M. (2000) *Parasitology* **120,** 171-84.
13. Wylie, T., Martin, J., Dante, M., Mitreva, M., Clifton, S. W., Chinwalla, A., Waterston, R. H., Wilson, R. K. & McCarter, J. P. (2004) *Nucleic Acids Res* **32,** D423-D426.
14. Ghosh, K. & Hotez, P. J. (1999) *J Infect Dis* **180,** 1674-1681.
15. Knox, D. P. & Smith, W. D. (2001) *Vet Parasitol* **100,** 21-32.
16. Houdijk, J. G., Kyriazakis, I., Jackson, F., Huntley, J. F. & Coop, R. L. (2003) *Int J Parasitol* **33,** 327-38.
17. Parkinson, J., Mitreva, M., Whitton, C., Thomson, M., Daub, J., Martin, J., Hall, N., Barrell, B., Waterston, R. H., McCarter, J. P. & Blaxter, M. (2004) *Nature Genetics* **36,** 1259-67.
18. Parkinson, J., Whitton, C., Schmid, R., Thomson, M. & Blaxter, M. (2004) *Nucleic Acids Res* **32,** D427-30.
19. Hussein, A. S., Kichenin, K. & Selkzer, P. M. (2002) *Mol Biochem Parasitol* **122,** 91-4.
20. Bonner, T. P. (1979) *J Parasitol* **65,** 74-8.
21. Harcus, Y. M., Parkinson, J., Fernandez, C., Daub, J., Selkirk ME, M.L., B. & Maizels, R. M. (2004) *Genome Biology* **5,** R39.
22. Issa, Z., Grant, W.N., Stasiuk, S., Shoemaker, C.B. (2005) *Int J Parasitol* **35,** 935-40.
23. Hoholm, F., Zhu, X., Ashton, F. T., Freeman, A. S., Veklich, Y., Castelletto, A., Lamont, S. & Schad, G. A. (2005) *J Parasitol* **91,** 61-8.
24. Joachim, A., Ruttkowski, B. & Daugschies, A. (2001) *Parasitol Res* **87,** 37-42.
25. Bao, Z. & Eddy, S. R. (2002) *Genome Res* **12,** 1152-5.
26. Ding, L., Sabo, A., Berkowicz, N., Meyer, R. R., Shotland, Y., Johnson, M. R., Pepin, K. H., Wilson, R. K. & Spieth, J. (2004) *Genome Res* **14,** 2503-9.
27. Lowe, T. M. & Eddy, S. R. (1997) *Nucleic Acids Res* **25,** 955-64.