A survey of genes expressed in adults of the human hookworm, Necator americanus

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Parasitology

Publisher Rights Statement:
RoMEO green

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
A survey of genes expressed in adults of the human hookworm, *Necator americanus*

J. DAUB¹, A. LOUKAS¹†, D. I. PRITCHARD² and M. BLAXTER¹*  

¹ Institute of Cell, Animal and Population Biology, University of Edinburgh, Edinburgh EH9 3JT, UK  
² Division of Life Science, University of Nottingham, Nottingham NG7 2RD, UK

(Received 14 May 1999; revised 7 August 1999; accepted 25 August 1999)

**Summary**

Hookworms are gut-dwelling, blood-feeding nematodes that infect hundreds of millions of people, particularly in the tropics. As part of a program aiming to define novel drug targets and vaccine candidates for human parasitic nematodes, genes expressed in adults of the human hookworm *Necator americanus* were surveyed by the expressed sequence tag approach. In total 161 new hookworm genes were identified. For the majority of these, a function could be assigned by homology. The dataset includes proteases, protease inhibitors, a lipid binding protein, C-type lectins, an anti-bacterial factor, globins and other genes of interest from a drug or vaccine development viewpoint. Three different classes of small, secreted proteins were identified that may be involved in the host–parasite interaction, including potential potassium channel blocking peptides. One third of the genes were novel. These included highly expressed, secreted (glyco)proteins which may be part of the excretory–secretory products of these important pathogens. Of particular interest are a family of 9 genes with similarity to the immunomodulatory protein, neutrophil inhibitory factor, that may play a role in establishing an immunocompromised niche for this successful parasite.

Key words: expressed sequence tags, hookworm, *Necator americanus*, *Ancylostoma duodenale*, *Caenorhabditis elegans*, ASP.

**Introduction**

Human hookworms are intestinal, blood-feeding strongyloid nematodes. It is estimated that there are over 1200 million cases annually, and the blood loss, anaemia and growth stunting that results from hookworm infection is calculated to be responsible for the loss of over 22 million disability adjusted life years (DALYs) in developing and underdeveloped countries (Bundy, 1997; Chan, 1997). The burden of hookworm infection appears to be increasing. The 2 nematode species responsible (*Necator americanus* and *Ancylostoma duodenale*) are closely related, and are susceptible to anthelminthic treatment. However, rapid reinfection from the environment, and the threat of the development of drug resistance in heavily treated communities, makes the development of new drugs and a subunit vaccine a priority in eradication strategies. Hookworm infections, like those of many other helminths, are highly allergenic, with T-helper 2 type responses predominating (Maizels et al. 1993). The mechanisms underlying this bias, and the roles of parasite allergens in initiating or maintaining it, are largely unknown.

Despite the importance of hookworms, few genes have been described from either human or model animal-infective species (Harrop et al. 1995a, 1996b; Hawdon et al. 1995b, 1996; Bin et al. 1999). The search for novel targets requires a source of genetic information defining potential targets and reagents for testing these targets. In particular, enzymes and effectors involved in establishing and maintaining the localized niche in which the hookworm feeds (Stannsens et al. 1996), and in nutrient digestion may be of interest as drug targets. Similarly, proteins secreted by the nematodes and thus accessible to the host immune system may identify candidate antigens for vaccine development (Hotez et al. 1987, 1996).

One route to rapid gene discovery is through the analysis of expressed sequence tags (ESTs), sequences generated from randomly selected cDNAs that can be used to survey and define the genes expressed by an organism (or stage or tissue) (Adams et al. 1991; McCombie et al. 1992; Waterston et al. 1992; Adams et al. 1995; Blaxter et al. 1996, 1999). The genome of hookworms would be expected to have about 20000 different protein-coding genes, like the closely related *Caenorhabditis elegans* (The *C. elegans* Genome Sequencing Consortium, 1998). Complete genome sequencing of a hookworm, while
feasible, is currently prohibitive in terms of cost. EST analysis in contrast is relatively cheap, and rapidly identifies the highly expressed genes. EST analysis of the human filarial nematode Brugia malayi has identified about one third of the genes of these parasites from only 16000 ESTs (Blaxter et al. 1996, 1999). The efficiency of this process has prompted us to perform EST analyses on additional parasitic nematode species, including Ascaris suum, Trichuris muris and Trichinella spiralis (M. Blaxter and J. Daub unpublished observations). Here we present an EST dataset from adult N. americanus that defines 161 new genes and that includes several candidates for further study as drug target or vaccine component molecules.†

MATERIALS AND METHODS

Expressed sequence tag generation from the Necator americanus adult cDNA library

A N. americanus mixed adult cDNA library was constructed in Lambda Zap Express following the manufacturer’s instructions (Stratagene, La Jolla, CA) from parasites adapted to and maintained in hamsters (Pritchard et al. 1999). The cDNA inserts are EcoRI/XhoI fragments and the library has 84% recombinant phage.

The cDNA library was used to infect XL1-Blue cells (Stratagene, La Jolla, CA) and randomly chosen recombinant clones picked. The cDNA inserts were amplified by PCR in 20 µl reactions using universal vector primers M13forward and M13reverse and Taq polymerase (Promega Corporation, Madison, WI). Inserts > 150 bp were selected for sequencing and 15 µl of PCR products were cleaned by treatment with shrimp alkaline phosphatase (1/µl) and exonuclease I (1:5 U; 30 min at 37 °C; L. Baron, personal communication). Then 5 µl of each insert were sequenced using the 5’ universal vector primer M13 reverse and ABI rhodamine dye terminators (Perkin–Elmer Corporation, Norwalk, CT). Sequencing reactions were analysed using an ABI 377 automated sequencer. The clones are archived and are freely available to the research community.

Bioinformatics

Base calling on sequences were checked, and vector and poor 3’ sequence removed, manually. Edited sequences were compared to public databases (GenBank non-redundant nucleotide and protein databases and dbEST) using the BLAST family of algorithms (Altschul et al. 1990). Sequences were clustered using AssemblyLign (Oxford Molecular, Oxford, UK). Where possible, a putative functional identity was assigned to the sequences. ESTs with no significant similarity to any sequences in the databases (defined as maximal BLASTX scores of < 80, with a probability < 1 x e^-8) were designated as novel. One methodological issue arises through the fact that ESTs are by definition single pass sequences and thus (i) may contain errors and (ii) may not be full length. In performing analysis of encoded peptide sequence we were sensitive to the quality of the sequence read (in general sequence prediction was excellent up to 550 bases and fell off thereafter) and excluded from further analysis regions where the sequence was poor by comparison to other fully sequenced genes. In the case of clusters with more than 1 EST, the overlap between the sequences offers additional confirmation of quality.

Sequences (typically peptide sequences translated from the ESTs) were aligned to homologues from other species using ClustalW (Thompson & Higgins, 1994) as implemented in MacVector (Oxford Molecular, Oxford, UK). Alignments were edited by hand and verified against secondary and tertiary structure models (where available). Alignments were analysed for phylogenetic content using maximum parsimony and neighbour joining algorithms as implemented in PAUP* 4b2 (Swofford, 1993; Swofford et al. 1996). The alignments generated are available from the NecatorWeb worldwide web site at http://www.ed.ac.uk/~mbx/NecatorWeb/Necator.html.

RESULTS AND DISCUSSION

Overall features of the N. americanus EST dataset

Of 259 clones selected, 211 were successfully sequenced. The insert sizes of the clones ranged from 150 to ~3000 bp, and the average sequence read was 450 bp. In total 43% of the inserts were sequenced in full. Cluster analysis of the ESTs suggests that they are derived from 161 different genes, giving an overall redundancy of 1:31 ESTs per cluster. Twenty-three clusters of > 1 EST and 138 clusters containing only 1 EST were found. Each cluster has been given a unique NAC (Necator americanus cluster) identifying number (e.g. NAC00042 describes a cluster encoding an antibacterial factor homologue) and the GenBank/dbEST submissions have been annotated with these cluster numbers (Blaxter et al. 1997) (Table 1). The database records can thus be retrieved using this NAC identifier, with the advantage that all records pertaining to each cluster will be returned. Of these putative genes, none had been sequenced previously from N. americanus, though homologues of 19 had been identified in other hookworms, or other

† Sequences described in this paper have been deposited in GenBank with the Accession numbers AI856935–AI857145.
Homologues of C Small, secreted proteins

Genes of interest (for cuticle collagens see Table 2)

polymerase chain reaction analysis of levels of gene

been confirmed by more extensive sequencing

we have noted that early patterns of abundance derived from small datasets have, in general, been confirmed by more extensive sequencing (Blaxter et al. 1996, 1999). Reverse transcriptase–polymerase chain reaction analysis of levels of gene

expression through the filarial life-cycle have also confirmed the patterns derived from EST cluster analysis (Gregory, Blaxter & Maizels, 1997).

Significant or informative database matches were found for 112 (70%) of the clusters. Comparison with the genome of C. elegans yielded matches for 106 (66%) of the clusters. Twenty-one clusters had significant similarity to genes (not including ESTs) from nematodes other than C. elegans, of which 19 were to strongylid genes and 2 to ascaridid genes. There are many B. malayi EST clusters with similarity to the N. americanus ESTs (data not shown).

strongylid nematodes. The clustering process permits the confirmation of sequence of overlapping reads and also defines genes expressed at high levels. The small size of the dataset makes unequivocal definition of highly expressed genes problematic, as there is a significant stochastic element in the selection of clones for sequencing. However, in analysis of ESTs from the filarial nematode B. malayi we have noted that early patterns of abundance derived from small datasets have, in general, been confirmed by more extensive sequencing (Blaxter et al. 1996, 1999). Reverse transcriptase–polymerase chain reaction analysis of levels of gene

Table 1. Genes of interest in the Nippostrongylus americanus adult EST dataset

<table>
<thead>
<tr>
<th>Cluster number</th>
<th>Representative accession number*</th>
<th>Gene name</th>
<th>Putative identification</th>
<th>Insert length (bp)</th>
<th>Sequence or consensus length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Activation-associated proteins</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAC00019</td>
<td>A1856949</td>
<td>Na-asp-2</td>
<td>Activation associated secreted protein</td>
<td>1400</td>
<td>764</td>
</tr>
<tr>
<td>NAC00055</td>
<td>A1856975</td>
<td>Na-asp-3</td>
<td></td>
<td>950</td>
<td>757</td>
</tr>
<tr>
<td>NAC00136</td>
<td>A1857041</td>
<td>Na-asp-4</td>
<td></td>
<td>850</td>
<td>671</td>
</tr>
<tr>
<td>NAC00008</td>
<td>A1856940</td>
<td>Na-asp-5</td>
<td></td>
<td>1000</td>
<td>883</td>
</tr>
<tr>
<td>NAC00214</td>
<td>A1857125</td>
<td>Na-asp-6</td>
<td></td>
<td>850</td>
<td>710</td>
</tr>
<tr>
<td>NAC00093</td>
<td>A1857004</td>
<td>Na-asp-7</td>
<td></td>
<td>750</td>
<td>564</td>
</tr>
<tr>
<td>NAC00129</td>
<td>A1857034</td>
<td>Na-asp-8</td>
<td></td>
<td>950</td>
<td>547</td>
</tr>
<tr>
<td>NAC00002</td>
<td>A1856936</td>
<td>Na-asp-9</td>
<td></td>
<td>1000</td>
<td>471</td>
</tr>
<tr>
<td>NAC00004</td>
<td>A1856937</td>
<td>Na-asp-10</td>
<td></td>
<td>2000</td>
<td>538</td>
</tr>
<tr>
<td>Small, secreted proteins</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAC00042</td>
<td>A1856966</td>
<td>Na-abf-1</td>
<td>Anti-bacterial peptide</td>
<td>341</td>
<td>341</td>
</tr>
<tr>
<td>NAC00064</td>
<td>A1856981</td>
<td>Na-sxc-1</td>
<td>SXC domain; kaliseptine-like</td>
<td>234</td>
<td>234</td>
</tr>
<tr>
<td>NAC00118</td>
<td>A1857025</td>
<td>Na-sxc-2</td>
<td>SXC domain; kaliseptine-like</td>
<td>216</td>
<td>216</td>
</tr>
<tr>
<td>NAC00075</td>
<td>A1856989</td>
<td>Na-sxc-3</td>
<td>SXC domain; kaliseptine-like</td>
<td>180</td>
<td>180</td>
</tr>
<tr>
<td>NAC00200</td>
<td>A1856950</td>
<td>Na-tyr-1</td>
<td>Trypsin inhibitor</td>
<td>554</td>
<td>554</td>
</tr>
<tr>
<td>Genes of interest (for cuticle collagens see Table 2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAC00128</td>
<td>A1857033</td>
<td>Na-bhp-20</td>
<td>Lipid binding protein</td>
<td>672</td>
<td>672</td>
</tr>
<tr>
<td>NAC00041</td>
<td>A1856965</td>
<td>Na-glb-2</td>
<td>Globin</td>
<td>750</td>
<td>561</td>
</tr>
<tr>
<td>NAC00088</td>
<td>A1857001</td>
<td>Na-glb-3</td>
<td>Globin</td>
<td>700</td>
<td>506</td>
</tr>
<tr>
<td>NAC00134</td>
<td>A1857039</td>
<td>Na-glb-4</td>
<td>Globin</td>
<td>750</td>
<td>691</td>
</tr>
<tr>
<td>NAC00022</td>
<td>A1856952</td>
<td>Na-hsp-1</td>
<td>20 kDa heat shock protein</td>
<td>520</td>
<td>520</td>
</tr>
<tr>
<td>NAC00165</td>
<td>A1857064</td>
<td>Na-hsp-2</td>
<td>20 kDa heat shock protein</td>
<td>650</td>
<td>490</td>
</tr>
<tr>
<td>NACA0014</td>
<td>A1856945</td>
<td>Na-hsp-3</td>
<td>20 kDa heat shock protein</td>
<td>606</td>
<td>606</td>
</tr>
<tr>
<td>NAC00034</td>
<td>A1856959</td>
<td>Na-cpb-1</td>
<td>Cathepsin B</td>
<td>2000</td>
<td>453</td>
</tr>
<tr>
<td>NAC00017</td>
<td>A1856948</td>
<td>Na-cpb-2</td>
<td>Cathepsin B</td>
<td>1000</td>
<td>360</td>
</tr>
<tr>
<td>NAC00230</td>
<td>A1857115</td>
<td>Na-apr-1</td>
<td>Aspartyl protease</td>
<td>1000</td>
<td>509</td>
</tr>
<tr>
<td>NAC00063</td>
<td>A1856980</td>
<td>Na-ctl-1</td>
<td>C-type lectin</td>
<td>577</td>
<td>577</td>
</tr>
<tr>
<td>NACA0019</td>
<td>A1857143</td>
<td>Na-ctl-2</td>
<td>C-type lectin</td>
<td>N.D.</td>
<td>530</td>
</tr>
<tr>
<td>Ribosomal proteins</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAC00188</td>
<td>A1857083</td>
<td>Na-rpl-10</td>
<td>Ribosomal protein L10</td>
<td>750</td>
<td>525</td>
</tr>
<tr>
<td>NAC00210</td>
<td>A1857098</td>
<td>Na-rpl-11</td>
<td>Ribosomal protein L11</td>
<td>750</td>
<td>558</td>
</tr>
<tr>
<td>NAC00186</td>
<td>A1857081</td>
<td>Na-rpl-27a</td>
<td>Ribosomal protein L27a</td>
<td>482</td>
<td>482</td>
</tr>
<tr>
<td>NAC00148</td>
<td>A1857053</td>
<td>Na-rpl-32</td>
<td>Ribosomal protein L32</td>
<td>357</td>
<td>357</td>
</tr>
<tr>
<td>NAC00031</td>
<td>A1856957</td>
<td>Na-rps-8</td>
<td>Ribosomal protein S8</td>
<td>750</td>
<td>635</td>
</tr>
<tr>
<td>NACA0021</td>
<td>A1856951</td>
<td>Na-rps-15</td>
<td>Ribosomal protein S15</td>
<td>N.D.</td>
<td>479</td>
</tr>
<tr>
<td>NAC00091</td>
<td>A1857003</td>
<td>Na-rps-18</td>
<td>Ribosomal protein S18</td>
<td>700</td>
<td>435</td>
</tr>
<tr>
<td>NACA0003</td>
<td>A1857132</td>
<td>Na-rps-29</td>
<td>Ribosomal protein S29</td>
<td>N.D.</td>
<td>258</td>
</tr>
<tr>
<td>Homologues of C. elegans proteins</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAC00151</td>
<td>A1857056</td>
<td>Na-des-1</td>
<td>Homologue of Ce-des-1</td>
<td>1200</td>
<td>482</td>
</tr>
<tr>
<td>NAC00135</td>
<td>A1857040</td>
<td>Na-sem-5</td>
<td>Homologue of Ce-sem-5</td>
<td>1200</td>
<td>670</td>
</tr>
<tr>
<td>NAC00126</td>
<td>A1857031</td>
<td>Na-unc-37</td>
<td>Homologue of Ce-unc-37</td>
<td>578</td>
<td>578</td>
</tr>
</tbody>
</table>

* For each cluster the sole, or lowest-numbered, EST sequence accession number is given. To identify all the ESTs clustered, and to examine a list of all similarities detected, please see the NecatorWeb on the worldwide web at: http://www.ed.ac.uk/~mbx/NecatorWeb/Necator.html
Each cluster (whether it contains 1 or several ESTs) has been named after the lowest-numbered clone, following the general guidelines promoted by the Filarial Genome Project (Blaxter et al. 1997). The clustered EST dataset, with analysis and comparisons, including multiple alignments of genes discussed here, is available on the NecatorWeb worldwide web site (Daub & Blaxter, 1999).

(1) **Activation-associated secreted protein (ASP) homologues.** Twenty-six of the ESTs (12.5%) encode 9 distinct homologues of *Ancylostoma caninum* ASP, a secreted product released on activation of dog hookworm third stage infective larvae (Hawdon, Jones & Hotez, 1995a). A homologue was recently described from *N. americanus* infective larvae (Bin et al. 1999) and others have been described from the strongyloid *Haemonchus contortus* (Schallig et al. 1997). We have named these genes activation-associated secreted proteins to retain the acronym ASP. *A. caninum* ASP is internally repetitive, with two 210 amino acid degenerate repeats sharing 28% identity. In particular, all the Cys residues in the two domains are conserved, along with several Gly and other residues. An alignment of the *A. caninum* ASP domains was used as a template against which to align the *N. americanus* ASPs, neutrophil inhibitory factor (*Ac-NIF*) from *A. caninum* (Moyle et al. 1994), 2 excretory–secretory products from *Haemonchus contortus* (the 24 kDa *Hc-ASP-2* and the 40 kDa *Hc-ASP-1*) (Schallig et al. 1997) and families of related genes from *C. elegans* and filarial nematodes (Fig. 1).

The ASPs were previously shown to have similarity to a family of vespid allergens (V5 family), *Heloderma horridum* lizard venom (helothermine), plant pathogenesis-related proteins, mammalian cysteine-rich salivary proteins (CRISPs), and mammalian testis glycoproteins (TPX-1) of mostly unknown biological function (Bin et al. 1999). ASP genes are of 2 kinds, the canonical *Ac-ASP*-1-like 2-domain type, and the *Ac-NIF*-like single domain type (Moyle et al. 1994). Based on the insert length of the cDNAs, the *N. americanus* adult ASPs are all single domain proteins. Seven of the 9 have identifiable secretory leader peptides: the remaining 2 are 5’ partial cDNAs. They are very divergent in sequence, but retain the conserved Cys and Gly residues noted within *A. caninum* ASP, and conform to the BLOCKS database definition of the V5/helothermine/CRISP/TPX-1 protein family (Henikoff & Henikoff, 1992; Henikoff et al. 1998; Bin et al. 1999). *Ac-NIF* has 7 potential *N*-linked glycosylation sites, but the other ASPs have either 1 (*Hc-ASP-2* and *Na-ASP-4*, -5 and -6) or none.

Phylogenetic analysis of the aligned sequences suggests that *Na-ASP*-4, -5, -6, -7 and -9 are much more closely related to each other, than to the other strongyloid sequences (Fig. 1). The filarial ASPs and a group of *C. elegans* single-domain ASPs which are found in close genomic proximity to each other on cosmids F49E11 and C39E9, form distinct subfamilies within the diversity of nematode ASPs. *C. elegans* also has a 2-domain ASP homologue (F11C7.3), but only domain b is marginally more similar to the 2-domain strongyloid ASPs. Within *Na-ASP-3* there are 2 classes of sequence which differ consistently in 4 out of 550 bases of overlap, resulting in 3 amino acid changes. It is not known whether these differences are allelic or define 2 very closely related genes. In peptide sequencing from purified *Ac-NIF*, several variant peptides were reported, and the existence of several NIF-like genes inferred (Moyle et al. 1994). However, the aligned sequences suggest that most of the variant residues reported derive mainly from technical errors in sequencing, as they correspond to absolutely conserved Cys or Gly residues, or highly conserved aromatic residues. There are 2 remaining variant peptides which may derive from additional *A. caninum* NIF-like/ASP genes.

*Ac-NIF* has potent effects on human and canine neutrophils (Muchowski et al. 1994; Rieu et al. 1994, 1996; Barnard et al. 1995; Zhang & Plow, 1996). NIF interferes with neutrophil recruitment to sites of inflammation by blocking recognition of CD11b/CD18 leukocyte integrins, and is thus likely to play a part in the hookworms’ strategy of host immune avoidance. As recombinant NIF (glycosylated in the yeast *Pichia pastoris*) has similarly potent effects (Moyle et al. 1994), this activity is likely to reside in the peptide structure. The additional ASP homologues identified here may similarly be involved in mediation of host immune responses by interference with integrin function. The separation by sequence similarity of larval, 2-domain ASPs from adult, single domain ASPs may indicate different function, and point to the different needs, in terms of host manipulation, of invading larvae versus resident adults.

(2) **Small secreted effector molecules.** The ESTs identify 3 classes of small secreted peptide which *N. americanus* adults may use to create an immuno- and bio-chemical holdfast, and also resist the effects of both host digestive enzymes and gut flora.

(i) Anti-bacterial factor (ABF). A cysteine-rich peptide factor in the pseudocoelomic fluid of *Ascaris suum* (*As-ABF*) has potent anti-bacterial activity (Kato & Komatso, 1996). NAC00042 encodes a homologue of this gene. Using the *A. suum* and *N. americanus* sequences, 4 ABF genes can be defined in the *C. elegans* genome, in 2 pairs on cosmids C50F2 (Ce-abf-1 and -2, chromosome I) and T22H6.5 (Ce-abf-3 and -4, chromosome X) (Fig. 2) (The *C. elegans* Genome Sequencing Consortium, 1998). NAC00042 is most closely related to *As-ABF* and *Ce-ABF-2* (69–73% pairwise identity over the
mature peptide region) while Ce-ABF-3 and -4 are less closely related. Three of the C. elegans genes (abf-1, -3, and -4) have conserved introns in phase 0 between amino acids 52 and 53 in the alignment of Fig. 2. The ABF thus appear to be a conserved nematode anti-bacterial immunity system. The significance of the observed substitutions in the ABF sequences for the potency or range of anti-bacterial activity is unknown.

(ii) Six-cysteine domain (SXC) proteins. Three of the clusters encode peptides with a 6-cysteine domain (SXC) first identified in surface coat proteins of the dog ascariid Toxocara canis (Gems et al. 1995; Gems & Maizels, 1996; Blaxter, 1998) (Fig. 3). The SXC domain is found in many additional nematode genes including additional Toxocara surface components (unpublished observations, Loukas), EST’s from Brugia malayi, and over 70 genes from C. elegans (Blaxter, 1998). In general SXC proteins are extracellular, in that they have putative secretory leader peptides. Many SXC domains are at the C-terminus of proteins, where they tend to be found as pairs (or quartets). The N-terminal segments of these proteins can be identified as having putative function (in C. elegans these include tyrosinases, myeloperoxidases, and zinc metalloproteases, while in T. canis a lipid-binding protein (Gems et al. 1995) has C-terminal SXC domains). Other SXC proteins appear to be mucins, as the constituent SXC domains are separated by oligo-serine or -threonine repetitive regions. In C. elegans there are also several SXC domain proteins where all of the mature peptide is predicted to be SXC domains with few or no amino acids separating them. The N. americanus EST’s encode different single-SXC domain proteins (Fig. 3.) These are unusual in that they comprise only a secretory leader peptide and the SXC domain. There are 2 C. elegans SXC genes with similar structure. The small size of the putative mature protein suggests that these SXC could act as signal molecules, like other small 6-cysteine domains. For example, epidermal growth factor (EGF) was first identified as a small peptide ligand, but the EGF molecule is utilized in many different proteins as a structural module (Greenwald, 1985).

The only peptides with sequence conforming to the general SXC consensus identified outside the nematodes come from sea anemones. One is attached to the C-terminus of a zinc metalloprotease (Pan et al. 1998). The others are single SXC domains which are part of sea anemone venom, where they act as potent potassium channel blockers (Schweitz et al. 1995). These K-channel blockers are similar in structure to other anemone venom components.
Fig. 2. Anti-bacterial factor homologues and NAC00042. The predicted peptide sequence of cluster NAC00042 was aligned to ABF homologues from Caenorhabditis elegans and Ascaris suum (Kato & Komatso, 1996) using ClustalW. The C50F2 genes were not predicted by the C. elegans genome project (The C. elegans Genome Sequencing Consortium, 1998) and have been designated Ce-abf-1 (bases 10785–9816; an intron is predicted from bases 10647–9936), and Ce-abf-2 (bases 9548–9342). On cosmid T22H6, gene T22H6.5 (bases 28376–28819; an intron is predicted from bases 28526–28579) has been named Ce-abf-3, and another previously unidentified gene, Ce-abf-4, is found in close proximity (bases 29869–30328; an intron is predicted from bases 29824–29874). Aligned residues with >80% identity are boxed and shaded, while residues with >80% similarity are shaded. A consensus derived from the aligned sequences is given below the alignment. –, Gaps inserted to improve the alignment. The position of the phase 0 introns in Ce-ABF-1, -3, and -4 are indicated by r.

Fig. 3. Six cysteine domain protein homologues. The Nippostrongylus americanus single-SXC domain peptides are aligned with 2 homologues from Caenorhabditis elegans (from the chromosome III cosmid T20G5), and kaliseptines from 3 cnidarians (Anemone sulcata, Heteractis magnifica and Stichodactyla helianthus). Aligned residues with >80% identity are boxed and shaded, while residues with >80% similarity are shaded. A consensus derived from the aligned sequences is given below the alignment. –, Gaps inserted to improve the alignment.

such as BgK from Bunodosma granulifera (Cotton et al. 1997). The tertiary structure of BgK has been determined by NMR, and reveals that the cysteines are disulphide-linked in the order 1 → 6, 2 → 5 and 3 → 4 (Cotton et al. 1997; Dauplais et al. 1997); whether this is also true of nematode SXC domains is unknown, but is not structurally impossible. There is functional conservation of a functional Tyr-Lys diad motif between BgK and other K-channel toxins (Dauplais et al. 1997), but this is not universally present in nematode SXC, or the N. americanus examples identified here. As N. americanus adults might be expected to interfere with the local and systemic immune system, and local peristaltic activity, it is possible that these 2 SXC proteins act as secreted antagonists of the K channels on gut muscle and immune cells.

(iii) A small, secreted protease inhibitor. NAC00020 encodes a protease inhibitor of the bovine pancreatic trypsin (BPTI)/Kunitz inhibitor
New genes from Necator americanus adults

Fig. 4. Trypsin inhibitor homologue NAC00020. The trypsin inhibitor homologue NAC00020 is shown aligned to bovine pancreatic trypsin inhibitor, kalicludines from the cnidarian Anemonia sulcata and dendrotoxin I from Dendroaspis polylepis polylepis. Aligned residues with > 80% identity are boxed and shaded, while residues with > 80% similarity are shaded. A consensus derived from the aligned sequences is given below the alignment. –, Gaps inserted to improve the alignment.

Fig. 5. Lipid binding protein (LBP-20) homologues and NAC00020. Lipid binding protein homologues were identified in the Caenorhabditis elegans genome sequence (6 genes in 3 clusters of 3, 2 and 1 gene), in EST sequences from C. briggsae (clone pk03d09) and Pristionchus pacificus (clone rs05f10 encodes a second LBP-20 homologue but the sequence is not of good quality and it has thus been left out of this analysis). Unpublished LBP-20 sequences from the filarial parasites Loa loa and Acanthocheilonema viteae were supplied by Judith Allen and Jan Bradley (Av-LBP-20) and David Guiliano and Amy Klion (Ll-LBP-20). The homologues were aligned and subjected to phylogenetic analysis using maximum parsimony. The tree figured is a phylogram of the consensus bootstrap tree (100 replicates) with branch lengths given above the branches, and percentage bootstrap support below. The filarial LBP-20 form a well supported group, and the pattern of relatedness of the strongylid and rhabditid LBP-20 suggests a recent amplification of these genes in this lineage.

class (Fig. 4). The open reading frame in the EST has a putative signal peptide (residues 16–30 in Fig. 4), and thus the gene appears to encode a single, secreted inhibitor domain. BPTI/Kunitz domains are common features of larger proteins, where they may play purely structural roles. Dendrotoxin (snake venom toxin; DTX) is a voltage-sensitive potassium channel blocker which, despite having significant similarity to BPTI, has no trypsin inhibitor activity. The sequence motifs responsible for this difference have been mapped to a Lys-Ala pair at residues 15–16 in mature BPTI (50–51 in the alignment of Fig. 4), and an Ile at residue 19 (54 in Fig. 4). In DTX these are replaced by Tyr–Glu and Pro respectively. The N. americanus inhibitor differs from both these patterns in that it has an Arg–Gly pair, followed by an Arg. In the venom secreted by A. sulcata there are at least 3 related DTX-class potassium channel blockers (kalicludines 1–3) which, unusually, also have anti-trypsin activity (Schweitz et al. 1995). Comparison of these toxins with NAC00020 and BPTI shows that the N. americanus peptide has some features in common with both BPTI and DTX families, and thus may have pharmacological effects similar to those of the kalicludines. It is striking that 2 of the small secreted peptides of N. americanus adults appear to have activities similar to those found in sea anemones, perhaps pointing to convergence on a physiology requiring disabling of the local nervous system and inhibition of muscular activity. Peptides corresponding to these potential secreted mediators have been synthesized and are being tested in immunological and electrophysiological assays (D. Pritchard, unpublished).

(3) Functionally identified genes. (i) Lipid binding protein (LBP) homologue. Cluster NAC000128 (2 ESTs) encodes a homologue of a family of nematode-specific retinol-binding proteins, first identified as immunogenetic surface proteins in Onchocerca volvulus (Tree et al. 1995), but also found in B. malayi, C. elegans (6 different genes), C. briggsae, Globodera rostochiensis (a plant parasite) and Pristionchus
Toxocara canis

rooted using the globins of resampling analysis of the shortest MP trees found, figured is a cladogram derived from a bootstrap methods, which yielded congruent results. The tree maximum parsimony (MP) and neighbour joining alignment was analysed for phylogenetic content using Hunt, Blaxter, Raes, Vanfleteren, Moens and Burr. The Mermis nigrescens and et al the 2 domains of myoglobin; GLBC, cuticle globin; GLBA and GLBB, indicating the isoform of globin (GLBM, body wall modified SwissProt code, with the first 4 letters was retained. The sequences are designated by a proportion of resampling analyses in which that group 1998). Numbers below the branches indicate the outgroup for the other taxa analysed (Blaxter et al 1998). Numbers below the branches indicate the proportion of resampling analyses in which that group was retained. The sequences are designated by a modified SwissProt code, with the first 4 letters indicating the isoform of globin (GLBM, body wall myoglobin; GLBC, cuticle globin; GLBA and GLBB, the 2 domains of Ascaris suum and Pseudoterranova decipiens pseudocoelomic globin; GLBP, pseudocoelomic globin and GLBE, eye globin), followed by a 5 letter species tag (NECAM, Necator americanus; SYNTN, S. trachea; OSTOS, O. ostertagi; TRICO, Trichostrongylus colubriformis; NIPBR, Nippostrongylus brasiliensis; CAEBR, Caenorhabditis briggsae; CAERE, C. remanei; CAEEL, C. elegans; ASCSU, A. suum; TOXCA, T. canis; PSEDE, P. decipiens; and MERNI, M. nigrescens). The N. americanus globins have been additionally identified with their cluster number and bold type.

pacificus (a free-living diplogasterid nematode) (Fig. 5). These antigens bind retinol and other lipids (Kennedy et al. 1997). They are predicted to have a simple alpha helical structure and to bind lipids in an internal hydrophobic pocket. A reporter gene construct in C. elegans fused to the promoter of one of the LBP homologues displayed somatic muscle expression (Hope, 1991), whereas expression has been mapped by immunohistochemistry to the hypodermis of O. volvulus (Tree et al. 1995). They are postulated to play a role in lipid uptake and transport through the cuticle in filaria, and may interact with the host immune system by sequestering immunomodulatory lipids. Ov-LBP-20 is a promising onchocerciasis immunodiagnostic candidate (Bradley et al. 1991, 1998).

(ii) Globins (GLB). Strongylid nematodes are known to express globins at relatively high levels (Blaxter, 1993; Blaxter, Ingram & Tweedie, 1994a; Graef et al. 1996). Two isoforms have been described: a myoglobin-like intracellular globin and an extracellular cuticle globin. The EST dataset includes 3 globin genes. NAC00088 encodes a putative myoglobin (GLBM) isoform. NAC00041 and NAC00134 encode 2 different cuticle globin (GLBC) isoforms. These new globins were compared to those of other strongylids and rhabditids, and the analysis suggests that the duplication of the cuticle globin gene is a recent event within the hookworms (Fig. 6). Like other strongylid globins, these sequences encode proteins with a high-affinity oxygen binding signature consisting of a tyrosine residue at helix B residue 10 and a Glu or Leu at helix E residue 7 (Davenport, 1949; Smith & Lee, 1963; Lee & Smith, 1965; De Baere & Perutz, 1993; Kloek et al. 1993b; Yang et al. 1995). This predicted high affinity is consistent with a continued requirement for oxygen in the near-anaerobic conditions of the small intestine. The globins may capture oxygen from ingested host blood, or abstract it from the mucosa (Blaxter, 1993).

(iii) Small heat shock proteins (HSP). Three clusters define 3 different small heat shock proteins of the HSP-16 or HSP-20 family (Stringham-Durovic et al. 1992; Tweedie et al. 1993). Analysis of available nematode sequences (both genomic and EST) resulted in the definition of 20 different related HSP genes from 8 species, including 8 from C. elegans. The N. americanus genes are most closely related to HSP-20 from Nippostrongylus brasiliensis (Tweedie et al. 1993) and appear to represent an amplification of this gene family in the genome of strongylid nematodes. The C. elegans genome contains a small family of 5 related genes (HSP-16, 16-2, 16-41, 16-48 (Stringham-Durovic et al. 1992) and F08H9.4) which appear to be the result of an independent amplification event. Similarly, in filarial nematodes, a family of 4 HSP genes can be identified in the Brugia malayi EST dataset (Blaxter et al. 1999), with related HSPs in other filaria.

(iv) Collagens (COL). Sixteen ESTs encode 8 different collagens (Table 2). Seven of these encode nematode cuticle collagens which can be assigned to collagen gene families on the basis of conserved cysteine residues in the non-Gly-X-Y regions of the open reading frames (Johnstone, 1994; Kramer, 1994a, 1997). One of these genes (Na-COL-6; NAC00052, a probable COL-8 family member) is unusual in that it encodes a peptide with the full complement of N-terminal and C-terminal conserved non-Gly-X-Y regions (including a signal peptide and a procollagen protease cleavage site) but
Necator americanus cuticle collagens and their Caenorhabditis elegans collagen family allocations

<table>
<thead>
<tr>
<th>Gene or family name</th>
<th>Representative EST accession number†</th>
<th>Cluster number</th>
<th>N-terminal Cys-rich motif</th>
<th>C-terminal Cys-rich motif</th>
</tr>
</thead>
<tbody>
<tr>
<td>COL-1</td>
<td>AI856954</td>
<td>Na-col-1</td>
<td>NAC00026</td>
<td>PGEKGIAPKYCALDDGGIFFEDGTRR*</td>
</tr>
<tr>
<td>Na-col-6</td>
<td>AI857008</td>
<td>Na-col-2</td>
<td>NAC00097</td>
<td>SGERGCIPKYCALDDGGIFFEDGTMRRR*</td>
</tr>
<tr>
<td>Na-col-3</td>
<td>AI857079</td>
<td>Na-col-3</td>
<td>NAC00184</td>
<td>NGEGKDGGHCPPRTPPGY*</td>
</tr>
<tr>
<td>Na-col-4</td>
<td>AI857005</td>
<td>Na-col-4</td>
<td>NAC00094</td>
<td>DGAKGSDHCPVARTPPG*</td>
</tr>
<tr>
<td>Na-col-5</td>
<td>AI856944</td>
<td>Na-col-5</td>
<td>NAC00013</td>
<td>PGT'GGSCHCPPRTPAGP*</td>
</tr>
<tr>
<td>Na-col-6</td>
<td>AI856972</td>
<td>Na-col-6</td>
<td>NAC00052</td>
<td>GADAAYCPGPRSSYKA*</td>
</tr>
<tr>
<td>Na-col-7</td>
<td>AI856984</td>
<td>NAC00067</td>
<td>N.D.</td>
<td>GADAAYCPGPRSSYKA*</td>
</tr>
</tbody>
</table>

† For each cluster the sole, or lowest-numbered EST sequence accession number is given. To identify all the ESTs clustered, please see: http://www.ed.ac.uk/~mbx/NecatorWeb/Necator.html

* Termination codon.

N.D., Not determined (the cDNA clone does not include the N-terminus).

Cluster Na-col-1 is the sole Na-col cluster, conforming to the general pattern expected for COL-1 family members. Na-col-1 has only 4 Gly-X-Y repeats between.

Cluster Na-col-2 and Na-col-6, while conforming to the general pattern expected for COL-1 family members, have significant C-terminal extensions compared to the C. elegans genes. Similar extensions are found in other C. elegans COL-8 like collagens (for example C15A11.5, M18.1, T15B7.3 and T15B7.4).

All the previously described strongylid collagens were from the COL-1 family (Shamansky et al. 1989). The eighth cluster (NAC00034) encodes the C-terminal, non-Gly-X-Y, globular domain of an alpha basement membrane collagen (Kramer, 1994b).

Cluster Na-col-2 encodes a component of the proteasome (NAC00227). Na-col-2 and Na-col-6 are most similar to genes from C. elegans identified by mutational genetics. NAC00135 encodes what is probably the direct N. americanus homologue of sem-5. Sem-5 is a gene involved in determination of the vulval muscles, and encodes a SH2-SH3 domain protein which mediates intracellular signalling processes (Clark et al. 1992). NAC00126 encodes a direct N. americanus homologue of amp-1, a gene involved in the development of the amphidial and excretory glands (Clark et al. 1992). Na-col-6 and Na-col-7 are homologues of Na-col-1 and Na-col-2, respectively.

(v) Cathepsin B proteases (CPB). Two clusters, NAC00017 and NAC00082 encode cathepsin B-like proteases, most similar to families of cathepsin B-like enzymes identified from A. caninum and H. contortus (Fig. 7). NAC00017 covers 150 amino acids of the mature protease domain, while the sequence for NAC00082 extends from the signal peptide, through the divergent pro-region to the beginning of the protease domain. In the 35 amino acid overlap between these 2 clusters it is clear that 2 different but related proteases have been identified. NAC00082 is most similar to the A. caninum proteases (Harrop et al. 1995b), while NAC00017 is more similar to H. contortus (Ray & McKerrow, 1992) and C. elegans (Ray & McKerrow, 1992; Larminie & Johnstone, 1996), representatives of this enzyme class (Fig. 7). The presence of multiple cathepsin B-like enzymes identified in H. contortus (Ray & McKerrow, 1992) and C. elegans (Ray & McKerrow, 1992; Larminie & Johnstone, 1996), the 35 amino acid overlap between the divergent pro-region of the mature proteases from Na-col-2 and Na-col-6, extends to the signal peptide, and the sequence divergence between these 2 clusters is clear that 2 different but related proteases have been identified. NAC00017 and NAC00082 are most similar to families of cathepsin B-like enzymes, most similar to families of cathepsin B-like enzymes identified from A. caninum (Ray & McKerrow, 1992) and C. elegans (Ray & McKerrow, 1992). The presence of multiple cathepsin B-like enzymes identified in H. contortus (Ray & McKerrow, 1992) and C. elegans (Ray & McKerrow, 1992; Larminie & Johnstone, 1996), the 35 amino acid overlap between the divergent pro-region of the mature proteases from Na-col-2 and Na-col-6, extends to the signal peptide, and the sequence divergence between these 2 clusters is clear that 2 different but related proteases have been identified.

(vi) Other genes. Also identified in the ESTs are a component of the proteasome (NAC00227), a serine-threonine protein kinase (NAC00086), as well as many housekeeping genes (such as ribosomal proteins and intermediary metabolism enzymes) and mitochondrial genes (such as ribosomal proteins). NAC00135 encodes what is probably the direct N. americanus homologue of sem-5. Sem-5 is a gene involved in determination of the vulval muscles, and encodes a SH2-SH3 domain protein which mediates intracellular signalling processes (Clark et al. 1992). NAC00126 encodes a component of the proteasome (NAC00227). Na-col-2 and Na-col-6 are most similar to genes from C. elegans identified by mutational genetics. NAC00135 encodes what is probably the direct N. americanus homologue of sem-5. Sem-5 is a gene involved in determination of the vulval muscles, and encodes a SH2-SH3 domain protein which mediates intracellular signalling processes (Clark et al. 1992). NAC00126 encodes a direct N. americanus homologue of amp-1, a gene involved in the development of the amphidial and excretory glands (Clark et al. 1992).
Fig. 7. Cathepsin B proteases NAC00017 and NAC00082. The predicted proteins encoded by NAC00017 (Na-cpb-2) and NAC00082 (Na-cpb-1) are shown, aligned closely related proteases from Ancylostoma caninum (CPB1, accession AAC46377; CPB2, AAD17297 and CPB3, AAC46878; Harrop et al. 1995b) and Caenorhabditis elegans (gut cysteine protease GCP1 (Ray & McKerrow, 1992), accession AAB33058). Residues conserved in “75% of the sequences are shaded, and residues 100% conserved are given as a consensus below the aligned sequences.

Direct homologue of unc-37, a gene identified as a transcriptional regulator of the Groucho family that interacts with the homeodomain gene unc-4 in specifying neural fates (Pflugrad et al. 1997). These N. americanus homologues will aid in identification of evolutionarily conserved domains of these important proteins.

Twenty-four clusters have significant similarity to ‘hypothetical genes’ predicted by the C. elegans genome sequencing project (The C. elegans Genome Sequencing Consortium, 1998). These hypothetical genes are predicted on the basis of coding potential, base composition bias, and splicing predictions. In many cases they have not been confirmed (in C. elegans) by any additional corroborating evidence, such as cognate ESTs (Durbin & Thierry-Mieg, 1994). The N. americanus ESTs thus provide a first confirmation that the predictions for these genes are correct, and can serve to point to possible conserved functional residues. In addition, abundant expression of a N. americanus homologue might indicate similar importance for the C. elegans gene.

(5) Abundant novel transcripts. Four clusters with more than 1 EST did not have homologues in the public sequence databases. Of these, 3 (NAC00056 [3 ESTs], NAC00098 [4 ESTs] and NAC00133 [2 ESTs]) have predicted secretory leader peptides (Nielsen et al. 1997) at their N-termini and encode small polypeptides (16–25 kDa). NAC00056 has 1, and NAC00098 3, N-linked glycosylation sites. We would suggest that these may represent secreted (glyco)proteins, possibly part of the excretory–secretory antigens of adult N. americanus. The absence of C. elegans homologues might also indicate that these genes are specific adaptations to mammalian parasitism.

Conclusions

Adult N. americanus successfully colonize the human gut, despite the presence of competing gut flora and the host immune system. The 161 genes defined here offer clues to the molecular bases for this success. In analysing the ESTs, 2 sorts of information can
inform the choice of candidate genes for future work. Knowledge of the biology of the nematode–host interaction, in particular feeding, immune interactions and competition with gut flora, can suggest the sorts of molecules that might be involved. Genes identified as belonging to known classes of enzymes or effectors can be identified rapidly by comparison to databases. Secondly, the EST dataset itself can inform choice, as genes expressed at high levels by the parasite (because their protein products are required in relatively high quantities) will be over-represented in the ESTs. While these genes may be of unknown function, their abundance alone recommends them for further study. One aspect of the methodology used in this study is worthy of note. Many EST projects (for example the Kohara lab C. elegans EST program (Kohara, 1996)) have selected against smaller inserts (< 500 bp). In this study, many novel and interesting genes were defined by full length transcripts < 500 bp, and these will have been missed in other work. Indeed, in the C. elegans EST dataset many of the small ribosomal proteins, and other short genes are under-represented.

This project has identified many genes which are promising by these criteria. There are proteases (potential digestive enzymes), a lipid binding protein (perhaps involved in nutrient uptake, and/or immune evasion), globins (which may act to ensure aerobicity), heat shock proteins (stress response genes), a protease inhibitor (that may counter host trypsin), potential potassium channel blockers (disabling the local immune and nervous systems), ASP-like proteins and C-type lectins (possibly interacting with immune effector cells) and an anti-bacterial peptide (possibly preventing infection by, or reducing competition from, gut flora). These genes deserve further study because of their functional identification.

As would be expected from their close phylogenetic relationship (Blaxter et al. 1998), in many cases the most similar genes in the databases are from C. elegans. The sequencing of the genome of this small free-living rhabditid has identified around 19 000 protein coding genes (The C. elegans Genome Sequencing Consortium, 1998). The prediction of these genes relies on sequence features (start and stop codons, splice sites) and C. elegans EST sequences, as well as similarity to other genes (Durbin & Thierry-Mieg, 1994). For many of the C. elegans predicted genes there are no cognate ESTs or informative similarities, and thus the N. americanus dataset offers a new route to confirming the reality of several C. elegans genes.

A large proportion (30%) of genes identified in this study have no informative database match. While this proportion is likely to decrease as the other nematode genome projects progress, and our ability to detect distant similarities with informatics tools improves, these genes offer a set of potentially hookworm-specific targets for immunotherapy and drug development. Within this set of novel genes are a few (5) which are expressed at high levels; three of these have predicted signal peptides. These may be components of the secretory products of the nematodes, and may be involved in novel aspects of immune evasion, anti-coagulation or other processes.

This study was funded by the Medical Research Council, UK and the Darwin Trust, Edinburgh, UK. David Giuliano provided valuable informatics assistance.

REFERENCES


BLAXTER, M. L., DE LEY, P., GAREY, J., LIU, L. X.,
SCHLEDMAN, P., VIERSTRAETE, A., VANFLETEREN, J.,
MACKEY, L. Y., DORRIS, M., FRISSE, L. M., VIDA, J. T. &


BLAXTER, M. L., RAGHAVAN, N., GHOSH, I., GUILIANO, D.,


BRADLEY, J. E., ATOGHO, B. M., ELSON, L., STEWART, G. R. &


NectactorWeb/Nectactor.html

DAUPHAIN, M., LECOQ, A., SONG, J., COTTON, J., JAMIN, N.,


FRENKEL, M. J., DOPHEIDE, T. A. A., WAGLAND, B. M. &


GEMS, D. G., FERGUSON, C. J., ROBERTSON, B. D., NIEVES, R.,

GRAAF, D. C. D., BERGHEN, P., MOENS, L., MAREZ, T. M. D.,


