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

J. DAUB[1], A. LOUKAS[1]†, D. I. PRITCHARD[2] *and* M. BLAXTER[1]*

[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 strongylid 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, and result in significantly skewed immune responses, 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*. 1995*a*, 1996*b*; Hawdon *et al*. 1995*b*, 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

* Corresponding author: Ashworth Laboratories, ICAPB, King's Buildings, University of Edinburgh, Edinburgh EH9 3JT, UK. Tel: +44 131 650 6760. Fax: +44 131 650 5450. E-mail: mark.blaxter@ed.ac.uk

† Present address: Molecular Parasitology Unit, Queensland Institute for Medical Research, Royal Brisbane Hospital, QLD 4029, Australia.

https://doi.org/10.1017/S0031182099005375 Published online by Cambridge University Press

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 *Eco*RI/*Xho*I 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/U) 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 $< 1 \times e^{-8}$) 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.

https://doi.org/10.1017/S0031182099005375 Published online by Cambridge University Press

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

Cluster number	Representative accession number*	Gene name	Putative identification	Insert length (bp)	Sequence or consensus length (bp)
Activation-associated proteins					
NAC00019	AI856949	*Na-asp-2*	Activation associated secreted protein	1400	764
NAC00055	AI856975	*Na-asp-3*		950	757
NAC00136	AI857041	*Na-asp-4*		850	671
NAC00008	AI856940	*Na-asp-5*		1000	883
NAC00214	AI857125	*Na-asp-6*		850	710
NAC00093	AI857004	*Na-asp-7*		750	564
NAC00129	AI857034	*Na-asp-8*		950	547
NAC00002	AI856936	*Na-asp-9*		1000	471
NAC00004	AI856937	*Na-asp-10*		2000	538
Small, secreted proteins					
NAC00042	AI856966	*Na-abf-1*	Anti-bacterial peptide	341	341
NAC00064	AI856981	*Na-sxc-1*	SXC domain; kaliseptine-like	234	234
NAC00118	AI857025	*Na-sxc-2*	SXC domain; kaliseptine-like	216	216
NAC00075	AI856989	*Na-sxc-3*	SXC domain; kaliseptine-like	180	180
NAC00020	AI856950	*Na-tyi-1*	Trypsin inhibitor	554	554
Genes of interest (for cuticle collagens see Table 2)					
NAC00128	AI857033	*Na-lbp-20*	Lipid binding protein	672	672
NAC00041	AI856965	*Na-glb-1*	Globin	750	561
NAC00088	AI857001	*Na-glb-2*	Globin	700	506
NAC00134	AI857039	*Na-glb-3*	Globin	750	691
NAC00022	AI856952	*Na-hsp-1*	20 kDa heat shock protein	520	520
NAC00165	AI857064	*Na-hsp-2*	20 kDa heat shock protein	650	490
NACA0014	AI856945	*Na-hsp-3*	20 kDa heat shock protein	606	606
NAC00034	AI856959	*Na-col-8*	Basement membrane collagen	900	578
NAC00082	AI856996	*Na-cpb-1*	Cathepsin B	2000	453
NAC00017	AI856948	*Na-cpb-2*	Cathepsin B	1000	360
NAC00230	AI857115	*Na-apr-1*	Aspartyl protease	1000	509
NAC00063	AI856980	*Na-ctl-1*	C-type lectin	577	577
NACA0019	AI857143	*Na-ctl-2*	C-type lectin	N.D.	530
Ribosomal proteins					
NAC00188	AI857083	*Na-rpl-10*	Ribosomal protein L10	750	525
NAC00210	AI857098	*Na-rpl-11*	Ribosomal protein L11	750	558
NAC00186	AI857081	*Na-rpl-27a*	Ribosomal protein L27a	482	482
NAC00148	AI857053	*Na-rpl-32*	Ribosomal protein L32	357	357
NAC00031	AI856957	*Na-rps-8*	Ribosomal protein S8	750	635
NACA0021	AI856951	*Na-rps-15*	Ribosomal protein S15	N.D.	470
NAC00091	AI857003	*Na-rps-18*	Ribosomal protein S18	700	435
NACA0003	AI857132	*Na-rps-29*	Ribosomal protein S29	N.D.	258
Homologues of *C. elegans* proteins					
NAC00151	AI857056	*Na-des-1*	Homologue of *Ce-des-1*	1200	482
NAC00135	AI857040	*Na-sem-5*	Homologue of *Ce-sem-5*	1200	670
NAC00126	AI857031	*Na-unc-37*	Homologue of *Ce-unc-37*	578	578

* 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

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 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).

https://doi.org/10.1017/S0031182099005375 Published online by Cambridge University Press

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, 1995*a*). A homologue was recently described from *N. americanus* infective larvae (Bin *et al*. 1999) and others have been described from the strongylid *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 strongylid 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 strongylid 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

https://doi.org/10.1017/S0031182099005375 Published online by Cambridge University Press

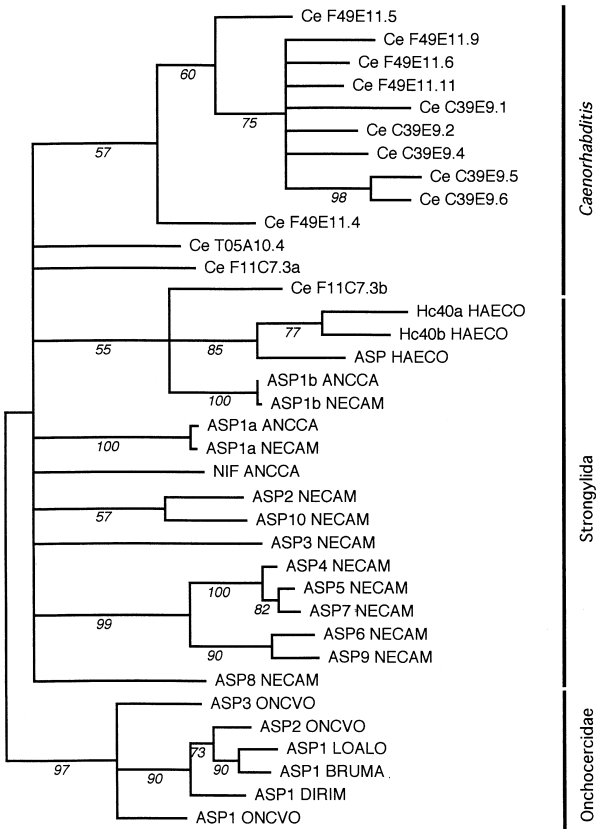

Fig. 1. Activation-associated protein homologues. Nine different clusters were identified that showed similarity to the *Ancylostoma* activation-associated secreted protein family. The predicted peptides from these clusters were aligned to ASP homologues from *Caenorhabditis elegans* (Ce), strongylid and filarial nematodes. The alignment was subjected to maximum parsimony and neighbour joining analyses, and a bootstrap consensus tree derived from the MP analysis is figured. The italic figures below the branches indicate percent bootstrap support. Branching orders with $< 50\,\%$ support are collapsed to form polytomies. Branch lengths are drawn relative to the number of inferred changes. The NJ trees found were congruent. The sequences used were Ce_F11C7.3b, Ce_F11C7.3a, Ce_F49E11.4, Ce_F49E11·5, Ce_F49E11.6, Ce_F49E11.9, Ce_F49E11.11, Ce_C39E9.1, Ce_C39E9.2, Ce_C39E9.4, Ce_C39E9.5, Ce_C39E9.6, Ce_T05A10.4: ASP homologues from *C. elegans* indicated by their cosmid.gene designation; Hc40a_HAECO, Hc40b_HAECO: the 2 domains of *Haemonchus contortus* Hc40; ASP_HAECO: the single-domain ASP homologue from *H. contortus*; NIF_ANCCA: neutrophil inhibitory factor from *A. caninum*; ASP1a_ANCCA, ASP1b_ANCCA: the 2 domains of ASP from *A. caninum*; ASP1a_NECAM, ASP1b_NECAM: the 2 domains of *N. americanus* ASP; ASP1_ONCVO, ASP2_ONCVO, ASP3_ONCVO: 3 ASP homologues from *Onchocerca volvulus* EST project, ASP1_LOALO: an ASP from *Loa loa* (D. Guiliano & A. Klion, unpublished); ASP1_DIRIM: ASP homologue from *Dirofilaria immitis*, ASP1_BRUMA: ASP homologue from the *Brugia malayi* genome project; ASP2_NECAM, ASP3_NECAM, ASP4_NECAM, ASP5_NECAM, ASP6_NECAM, ASP7_NECAM, ASP8_NECAM, ASP9_NECAM, ASP10_NECAM: ASP homologues identified in this study (see Table 1).

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 ascaridid *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), ESTs 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* ESTs 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 domain 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,

https://doi.org/10.1017/S0031182099005375 Published online by Cambridge University Press

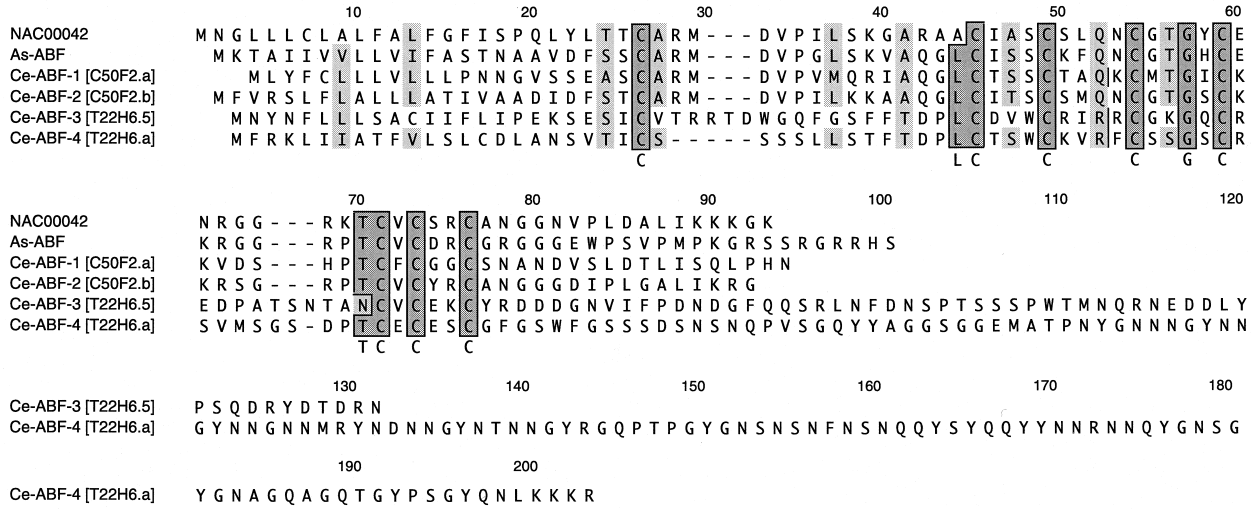

Fig. 2. Anti-bacterial factor homologues and NAC00042. The predicted peptide sequence of cluster NAC00042 was aligned to ABF homologues from *Caenorhabditis elegans* and *Ascaris suum* (Kato & Komatso, 1996) using ClustalW. The C50F2 genes were not predicted by the *C. elegans* genome project (The *C. 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 |.

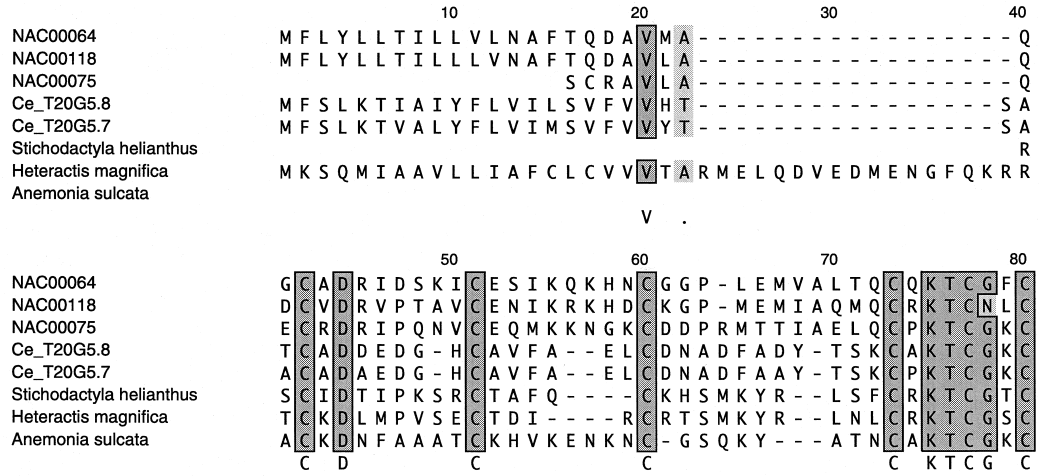

Fig. 3. Six cysteine domain protein homologues. The *Nippostrongylus americanus* single-SXC domain peptides are aligned with 2 homologues from *Caenorhabditis elegans* (from the chromosome III cosmid T20G5), and kaliseptines from 3 cnidarians (*Anemone sulcata*, *Heteractis magnifica* and *Stichodactyla 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 such as scorpion charybdotoxin (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

https://doi.org/10.1017/S0031182099005375 Published online by Cambridge University Press

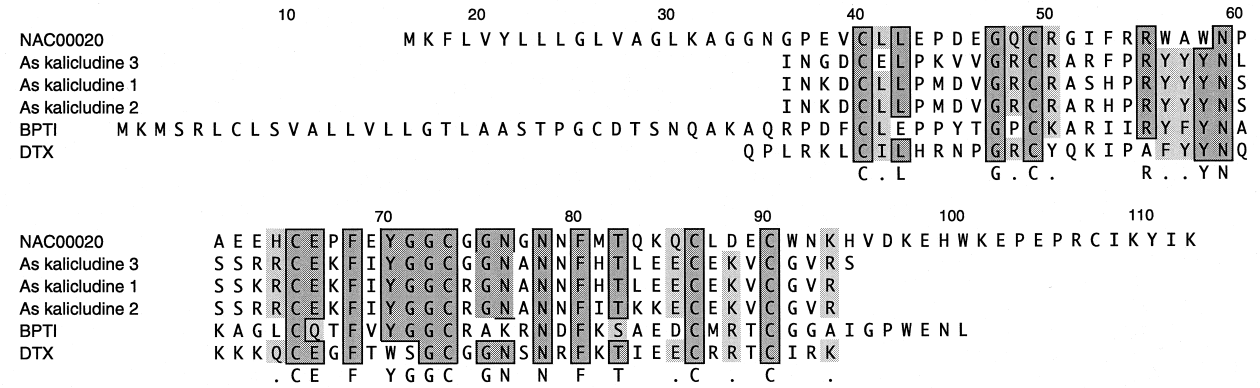

Fig. 4. Trypsin inhibitor homologue NAC00020. The trypsin inhibitor homologue NAC00020 is shown aligned to bovine pancreatic trypsin inhibitor, kalicludines from the cnidarian *Anemonia sulcata* and dendrotoxin I from *Dendroaspis polylepis polylepis*. Aligned residues with $> 80\,\%$ identity are boxed 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.

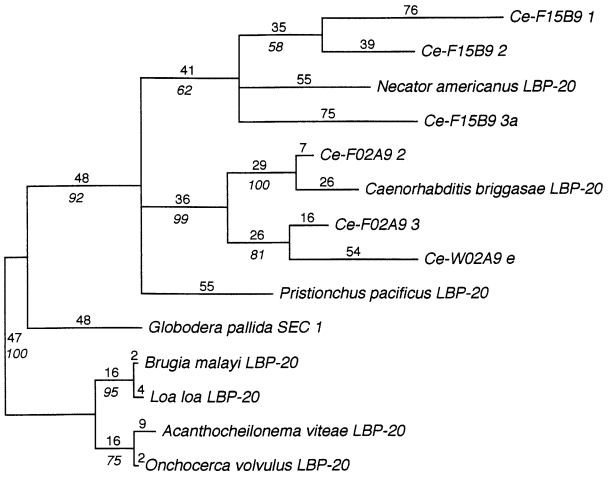

Fig. 5. Lipid binding protein (LBP-20) homologues and NAC00128. Lipid binding protein homologues were identified in the *Caenorhabditis elegans* genome sequence (6 genes in 3 clusters of 3, 2 and 1 gene), in EST sequences from *C. briggsae* (clone pk03d09) and *Pristionchus pacificus* (clone rs04h05; 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 immunogenic 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*

https://doi.org/10.1017/S0031182099005375 Published online by Cambridge University Press

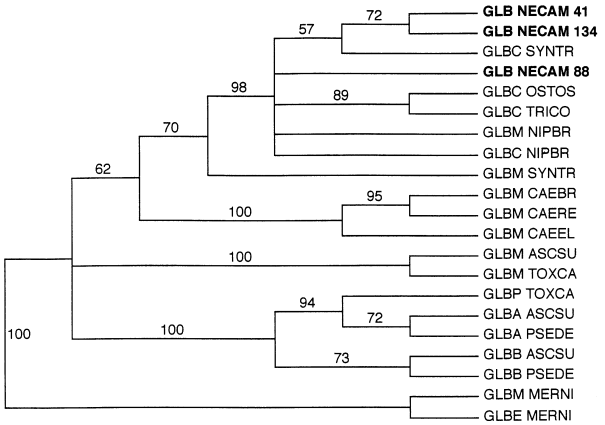

Fig. 6. Globin homologues NAC00088, NAC00041 and NAC00134. The predicted protein products of the 3 globin-like EST clusters were aligned to other nematode globin sequences (Frenkel *et al*. 1992; Sherman *et al*. 1992; Blaxter, 1993; Kloek *et al*. 1993*a*, 1996; Blaxter *et al*. 1994*a*, *b*; Graaf *et al*. 1996). Globins from *Toxocara canis*, *Ostertagia ostertagi*, *Syngamus trachea*, and *Mermis nigrescens* are from unpublished data of Hunt, Blaxter, Raes, Vanfleteren, Moens and Burr. The alignment was analysed for phylogenetic content using maximum parsimony (MP) and neighbour joining methods, which yielded congruent results. The tree figured is a cladogram derived from a bootstrap resampling analysis of the shortest MP trees found, rooted using the globins of *Mermis nigrescens*, which is an 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*; SYNTR, *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, 1994*a*; Graaf *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*. 1993*b*; 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-16-1, -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, 1994*a*, 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

https://doi.org/10.1017/S0031182099005375 Published online by Cambridge University Press

Table 2. *Necator americanus* cuticle collagens and their *Caenorhabditis elegans* collagen family allocations

Gene or family name	Representative EST accession number†	Cluster number	N-terminal Cys-rich motif	C-terminal Cys-rich motif
COL-1			NPAPNLQCEGCCLP	PGEKGICPKYCALDGGIFFEDGTRR*
Na-col-1	AI856954	NAC00026	RQAGMCDDCCAP	SGERGICPKYCALDGGIFFEDGTMRRR*
COL-6			VNAEPAAVCCTCNQ	NGEKGDCGHCPPPRTPPGY*
Na-col-2	AI857008	NAC00097	RQYPELCCSCGI	DGAKGSCDHCPVARTPPGY*
Na-col-3	AI857079	NAC00184	N.D.	PGTGGSCDHCPPPRTAPGY*
COL-8			EQCNCGPKSEGCPA	GADAAYCPCPGRSYKA*
Na-col-4	AI857005	NAC00094	N.D.	GKDGAYCPCPRTTGYRSRQKKASEKLSAA*
Na-col-5	AI856944	NAC00013	N.D.	GGDGAYCPCPPRSTVLALKKTVAVDSFSATDS
				NAVKKRVARRMNPHLNKAAGLS*
Na-col-6	AI856972	NAC00052	PHCKCGAFPTACPA	GGDGAYCPCPPRTGRYSRQGPRGISRHRKRLV
				PVPKKRVARRPNGPSTARNRIIQHKTAYRKQ*
Na-col-7	AI856984	NAC00067	N.D.	GADAEYCPCPPRRKRRL*

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

has only 4 Gly-X-Y repeats between. *Na*-COL-5 and -6, while conforming to the general pattern expected for COL-8 family members, have significant C-terminal extensions compared to the canonical *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, 1994*b*).

(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.* 1995*b*), while NAC00017 is more similar to *H. contortus* (Pratt *et al.* 1992) and *C. elegans* (Ray & McKerrow, 1992; Larminie & Johnstone, 1996), representatives of this enzyme class (Fig. 7). The presence of multiple cathepsins B in *N. americanus* is not unexpected, as there has been an amplification of this cathepsin class in all strongylids examined. In other species these enzymes play a role in haemoglobin degradation and digestion, and are located, in *A. caninum*, in the amphidial and excretory glands. Cluster NAC00230 encodes an aspartyl protease.

(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 mitochondrially encoded genes (Table 1). There are 2 C-type lectin homologues (NAC00063 and NACA0019). NACA0019 has greater similarity to mammalian P-selectin than to any of the ~120 *C. elegans* C-type lectins (data not shown), and may be an immunomodulatory protein that has evolved convergently (in structure and function) with the host.

(4) *Clusters with similarity to genes from* C. elegans *genome sequence*. Two clusters 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 hermaphrodite vulval muscles, and encodes a SH2–SH3 domain protein which mediates intracellular signalling processes (Clark *et al.* 1992). NAC00126 encodes the

https://doi.org/10.1017/S0031182099005375 Published online by Cambridge University Press

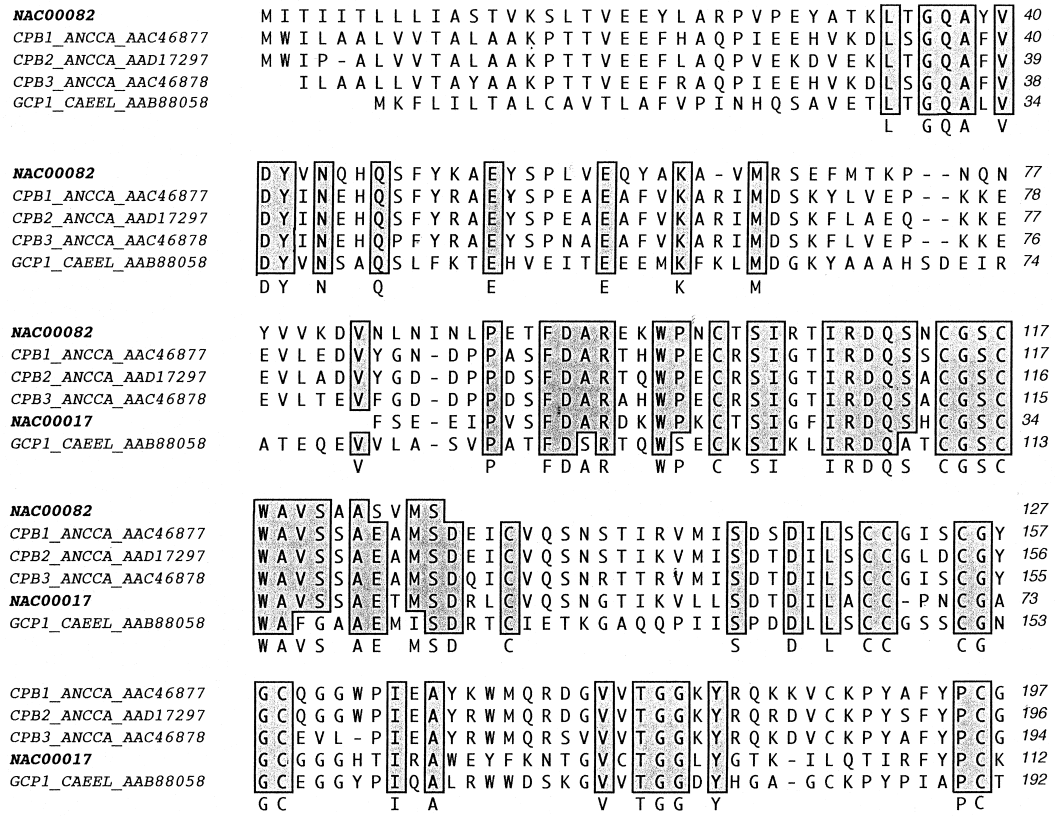

Fig. 7. Cathepsin B proteases NAC00017 and NAC00082. The predicted proteins encoded by NAC00017 (*Na-cpb-2*) and NAC00082 (*Na-cpb-1*) are shown, aligned with closely related proteases from *Ancylostoma caninum* (CPB1, accession AAC46377; CPB2, AAD17297 and CPB3, AAC46878; Harrop *et al*. 1995*b*) and *Caenorhabditis elegans* (the 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. For example, cluster NAC00054 (2 ESTs) encodes a 169 amino acid protein which is 59 % identical (and 70 % similar) to the gene F22B5.4. Both these predicted proteins appear to be type II membrane proteins, lacking signal peptides but sharing a central 20 amino acid, hydrophobic, potential membrane-spanning region.

(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

https://doi.org/10.1017/S0031182099005375 Published online by Cambridge University Press

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 Guiliano provided valuable informatics assistance.

REFERENCES

ADAMS, M. D., KERLAVAGE, A. R., FLEISCHMANN, R. D., FULDNER, R. A., BULT, C. J., LEE, N. H., KIRKNESS, E. F., WEINSTOCK, K. G., GOCAYNE, J. D., WHITE, O., SUTTON, G., BLAKE, J. A., BRANDON, R. C., CHIU, M. W., CLAYTON, R. A., CLINE, R. T., COTTON, M. D., EARLE-HUGHES, J., FINE, L. D., FITZGERALD, L. M., FITZHUGH, W. M., FRITCHMAN, J. L., GEOGHAGEN, N. S. M., GLODEK, A., GNEHM, C. L., HANNA, M. C., HEDBLOM, E., HINKLE, P. S. JR, KELLEY, J. M., KLIMEK, K. M., KELLEY, J. C., LIU, L.-I., MARMAROS, S. M., MERRICK, J. M., MORENO-PALANQUES, R. F., MCDONALD, L. A., NGUYEN, D. T., PELLEGRINO, S. M., PHILLIPS, C. A., RYDER, S. E., SCOTT, J. L., SAUDEK, D. M., SHIRLEY, R., SMALL, K. V., SPRIGGS, T. A., UTTERBACK, T. R., WEIDMAN, J. F., LI, Y., BARTHLOW, R., BEDNARIK, D. P., CAO, L., CEPEDA, M. A., COLEMAN, T. A., COLLINS, E.-J., DIMKE, D., FENG, P., FERRIE, A., FISCHER, C., HASTINGS, G. A., HE, W.-W., HU, J.-S., HUDDLESTON, K. A., GREENE, J. M., GRUBER, J., HUDSON, P., KIM, A., KOZAK, D. L., KUNSCH, C., JI, H., LI, H., MEISSNER, P. S., OLSEN, H., RAYMOND, L., WEI, Y.-F., WING, J., XU, C., YU, G.-L., RUBEN, S. M., DILLON, P. J., FANNON, M. R., ROSEN, C. A., HASELTINE, W. A., FIELDS, C., FRASER, C. M. & VENTER, J. C. (1995). Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature, London* **377** (Suppl.), 3–174.

ADAMS, M. D., KELLEY, J. M., GOCAYNE, J. D., DUBNICK, M., POLYMEROPOULOS, M. H., XIAO, H., MERRIL, C. R., WU, A., OLDE, B., MORENO, R. F., KERLAVAGE, A. R., MCCOMBIE, W. R. & VENTER, J. C. (1991). Complementary DNA sequencing: expressed sequence tags and the human genome project. *Science* **252**, 1651–1656.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410.

BARNARD, J. W., BIRO, M. G., LO, S. K., OHNO, S., CAROZZA, M. A., MOYLE, M., SOULE, H. R. & MALIK, A. B. (1995). Neutrophil inhibitory factor prevents neutrophil-dependent lung injury. *Journal of Immunology* **155**, 4876–4881.

BIN, Z., HAWDON, J., QIANG, S., HAINAN, R., HUIQING, Q., WEI, H., SHU-HUA, X., TIEHUA, L., XING, G., ZHENG, F. & HOTEZ, P. (1999). *Ancylostoma* secreted protein 1 (ASP-1) homologues in human hookworms. *Molecular and Biochemical Parasitology* **98**, 143–149.

BLAXTER, M. L. (1993). Nemoglobins: divergent nematode globins. *Parasitology Today* **9**, 353–360.

BLAXTER, M. L. (1998). *Caenorhabditis elegans* is a nematode. *Science* **282**, 2041–2046.

BLAXTER, M. L., ASLETT, M., DAUB, J., GUILIANO, D. & THE FILARIAL GENOME PROJECT. (1999). Parasitic helminth genomics. *Parasitology* (in the Press)

BLAXTER, M. L., DE LEY, P., GAREY, J., LIU, L. X., SCHELDEMAN, P., VIERSTRAETE, A., VANFLETEREN, J., MACKEY, L. Y., DORRIS, M., FRISSE, L. M., VIDA, J. T. & THOMAS, W. K. (1998). A molecular evolutionary framework for the phylum Nematoda. *Nature, London* **392**, 71–75.

BLAXTER, M. L., GUILIANO, D. B., SCOTT, A. L. & WILLIAMS, S. A. (1997). A unified nomenclature for filarial genes. *Parasitology Today* **13**, 416–417.

BLAXTER, M. L., INGRAM, L. & TWEEDIE, S. (1994*a*). Sequence, expression and evolution of the globins of the nematode *Nippostrongylus brasiliensis*. *Molecular and Biochemical Parasitology* **68**, 1–14.

BLAXTER, M. L., RAGHAVAN, N., GHOSH, I., GUILIANO, D., LU, W., WILLIAMS, S. A., SLATKO, B. & SCOTT, A. L. (1996). Genes expressed in *Brugia malayi* infective third stage larvae. *Molecular and Biochemical Parasitology* **77**, 77–96.

BLAXTER, M. L., VANFLETEREN, J., XIA, J. & MOENS, L. (1994*b*). Structural characterisation of an *Ascaris* myoglobin. *Journal of Biological Chemistry* **269**, 30181–30186.

BRADLEY, Y. E., HELM, R., LAHAISE, M. & MAIZELS, R. M. (1991). cDNA clones of *Onchocerca volvulus* low molecular weight antigens provide immunologically specific diagnostic probes. *Molecular and Biochemical Parasitology* **46**, 219–228.

BRADLEY, J. E., ATOGHO, B. M., ELSON, L., STEWART, G. R. & BOUSSINESQ, M. (1998). A cocktail of recombinant *Onchocerca volvulus* antigens for serologic diagnosis with the potential to predict the endemicity of onchocerciasis infection. *American Journal of Tropical Medicine and Hygiene* **59**, 877–882.

BUNDY, D. A. P. (1997). This wormy world – then and now. *Parasitology Today* **13**, 407–408.

CHAN, M.-S. (1997). The global burden of intestinal nematode infection – fifty years on. *Parasitology Today* **13**, 438–443.

CLARK, S. G., STERN, M. J. & HORVITZ, H. R. (1992). *C. elegans* cell-signalling gene *sem-5* encodes a protein with SH2 and SH3 domains. *Nature* **356**, 340–344.

COTTON, J., CREST, M., BOUET, F., ALESSANDRI, N., GOLA, M., FOREST, E., KARLSSON, E., CASTANEDA, O., HARVEY, A. L., VITA, C. & MENEZ, A. (1997). A potassium-channel toxin from the sea anemone *Bunodosoma granulifera*, an inhibitor for Kv1 channels. Revision of the amino acid sequence, disulfide bridge assignment chemical synthesis and biological activity. *European Journal of Biochemistry* **244**, 192–202.

DAUB, J. & BLAXTER, M. (1999). NecatorWeb: *Necator americanus* ESTs. http://www.ed.ac.uk/~mbx/NecatorWeb/Necator.html

DAUPLAIS, M., LECOQ, A., SONG, J., COTTON, J., JAMIN, N., GILQUIN, B., ROUMESTAD, C., VITA, C., DE MEDEIROS, C. L. C., ROWAN, E. G., HARVEY, A. L. & MÉNEZ, A. (1997). On the convergent evolution of animal toxins. Conservation of a diad of functional residues in potassium channel-blocking toxins with unrelated structures. *Journal of Biological Chemistry* **272**, 4302–4309.

DAVENPORT, H. E. (1949). The haemoglobins of *Nippostrongylus muris* (Yokagawa) and *Strongylus* spp. *Proceedings of the Royal Society of London, B* **136**, 271–280.

DE BAERE, I. & PERUTZ, M. (1993). Formation of two hydrogen bonds from the globin to the heme-linked oxygen molecule in *Ascaris* hemoglobin. *Proceedings of the National Academy of Sciences, USA* **91**, 1594–1597.

DURBIN, R. & THIERRY-MIEG, J. (1994). The ACeDB genome database. *Computational Methods in Genome Research*. (ed. Suhai, S.) pp. 34–40. Plenum, New York.

FRENKEL, M. J., DOPHEIDE, T. A. A., WAGLAND, B. M. & WARD, C. W. (1992). The isolation, cloning and characterisation of a globin-like host-protective antigen from the excretory-secretory products of *Trichostrongylus colubriformis*. *Molecular and Biochemical Parasitology* **50**, 27–36.

GEMS, D. & MAIZELS, R. M. (1996). An abundantly expressed mucin-like protein *Toxocara canis* infective larvae: the precursor of the TES-120 surface coat glycoproteins. *Proceedings of the National Academy of Sciences, USA* **93**, 1665–1670.

GEMS, D. G., FERGUSON, C. J., ROBERTSON, B. D., NIEVES, R., PAGE, A. P., BLAXTER, M. L. & MAIZELS, R. M. (1995). An abundant *trans*-spliced mRNA from *Toxocara canis* infective larvae encodes a 26-kDa protein with homology to phosphatidylethanolamine binding proteins. *Journal of Biological Chemistry* **270**, 18517–18522.

GRAAF, D. C. D., BERGHEN, P., MOENS, L., MAREZ, T. M. D., RAES, S., BLAXTER, M. L. & VERCRUYSSE, J. (1996). Isolation, characterisation and immunolocalisation of a globin-like antigen from *Ostertagia ostertagi* adults. *Parasitology* **113**, 63–70.

GREENWALD, I. (1985). *lin-12*, a nematode homoeotic gene, is homologous to a set of mammalian proteins that includes epidermal growth factor. *Cell* **43**, 583–590.

GREGORY, W. F., BLAXTER, M. L. & MAIZELS, R. M. (1997). Differentially expressed abundant, *trans*-spliced cDNAs from larval *Brugia malayi*. *Molecular and Biochemical Parasitology* **86**, 85–96.

HARROP, S. A., PROCIV, P. & BRINDLEY, P. J. (1995*a*). Amplification and characterization of cysteine proteinase genes from nematodes. *Tropical Medicine and Parasitology* **46**, 119–122.

HARROP, S. A., PROCIV, P. & BRINDLEY, P. J. (1996). Acasp a gene encoding a cathepsin D-like aspartic protease from the hookworm *Ancylostoma caninum*. *Biochemical Biophysical Research Communications* **227**, 294–302.

HARROP, S. A., SAWANGJAROEN, N., PROCIV, P. & BRINDLEY, P. J. (1995*b*). Characterisation and localisation of cathepsin B proteinases expressed by adult *Ancylostoma caninum* hookworms. *Molecular and Biochemical Parasitology* **71**, 163–171.

HAWDON, J. M., JONES, B. F., HOFFMAN, D. R. & HOTEZ, P. J. (1996). Cloning and characterization of Ancylostoma-secreted protein. A novel protein associated with the transition to parasitism by infective hookworm larvae. *Journal of Biological Chemistry* **271**, 6672–6678.

HAWDON, J. M., JONES, B. F. & HOTEZ, P. J. (1995*a*). Cloning and characterization of a cDNA encoding the

catalytic subunit of a cAMP-dependent protein kinase from *Ancylostoma caninum* third stage infective larvae. *Molecular and Biochemical Parasitology* **69**, 127–130.

HAWDON, J. M., JONES, B. F. & HOTEZ, P. (1995*b*). Cloning and characterisation of a cDNA encoding the catalytic subunit of a cAMP-dependent protein kinase from *Ancylostoma caninum* third stage larvae. *Molecular and Biochemical Parasitology* **69**, 127–130.

HENIKOFF, S. & HENIKOFF, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences, USA* **89**, 10915–10919.

HENIKOFF, S., PIETROKOVSKI, S. & HENIKOFF, J. G. (1998). Superior performance in protein homology detection with the blocks database servers. *Nucleic Acids Research* **26**, 309–312.

HOPE, I. A. (1991). 'Promoter trapping' in *Caenorhabditis elegans*. *Development* **113**, 399–408.

HOTEZ, P. J., HAWDON, J. M. & CAPPELLO, M. (1996). Molecular approaches to vaccinating against hookworm disease. *Pediatric Research* **40**, 515–521.

HOTEZ, P. J., LE TRANG, N. & CERAMI, A. (1987). Hookworm antigens: the potential for vaccination. *Parasitology Today* **3**, 247–249.

JOHNSTONE, I. L. (1994). The cuticle of the nematode *Caenorhabditis elegans*: a complex collagen structure. *BioEssays* **16**, 1–8.

KATO, Y. & KOMATSO, S. (1996). ASABF, a novel, cysteine-rich antibacterial peptide isolated from the nematode *Ascaris suum*. Purification, primary structure and molecular cloning of cDNA. *Journal of Biological Chemistry* **271**, 30493–30498.

KENNEDY, M. W., GARSIDE, L. H, GOODRICK, L. E., MCDERMOTT, L., BRASS, A., PRICE, N. C., KELLY, S. M., COOPER, A. & BRADLEY, J. E. (1997). The Ov20 protein of the parasitic nematode *Onchocerca volvulus*. A structurally novel class of small helix-rich retinol-binding proteins. *Journal of Biological Chemistry* **272**, 29442–29448.

KLOEK, A. P., MCCARTER, J. P., SETTERQUIST, R. A., SCHEDL, T. & GOLDBERG, D. E. (1996). *Caenorhabditis* globin genes: rapid intronic divergence contrasts with conservation of silent exonic sites. *Journal of Molecular Evolution* **43**, 101–108.

KLOEK, A. P., SHERMAN, D. R. & GOLDBERG, D. E. (1993*a*). Novel gene structure and evolutionary context of *Caenorhabditis elegans* globin. *Gene* **129**, 215–221.

KLOEK, A. P., YANG, J., MATTHEWS, F. S., FRIEDEN, C. & GOLDBERG, D. E. (1993*b*). The tyrosine B10 hydroxyl is crucial for oxygen avidity of *Ascaris* hemoglobin. *Journal of Biological Chemistry* **268**, 17669–17671.

KOHARA, Y. (1996). Large scale analysis of *C. elegans* cDNA. *Tanpakushitsu Kakusan Koso* **41**, 715–720.

KRAMER, J. (1994*a*). Structures and functions of collagens in *Caenorhabditis elegans*. *FASEB Journal* **8**, 329–336.

KRAMER, J. M. (1994*b*). Genetic analysis of the extracellular matrix in *C. elegans*. *Annual Review of Genetics* **28**, 95–116.

KRAMER J. M. (1997). Extracellular matrix. *C. elegans II*. Cold Spring Harbor Press, Cold Spring Harbor.

LARMINIE, C. G. C. & JOHNSTONE, I. L. (1996). Isolation and characterisation of four developmentally regulated cathepsin-B like cysteine protease genes from the nematode *Caenorhabditis elegans*. *DNA and Cell Biology* **15**, 75–82.

LEE, D. L. & SMITH, M. H. (1965). Hemoglobins of parasitic animals. *Experimental Parasitology* **16**, 392–424.

MAIZELS, R. M., BUNDY, D. A. P., SELKIRK, M. K., SMITH, D. F. & ANDERSON, R. M. (1993). Immunological modulation and evasion by helminth parasites in human populations. *Nature, London* **365**, 797–805.

MCCOMBIE, W. R., ADAMS, M. D., KELLEY, J. M., FITZGERALD, M. G., UTTERBACK, T. R., KHAN, M., DUBNICK, M., KERLAVAGE, A. R., VENTER, J. C. & FIELDS, C. (1992). *Caenorhabditis elegans* expressed sequence tags identify gene families and potential disease gene homologues. *Nature, Genetics* **1**, 124–131.

MOYLE, M., FOSTER, D. L., MCGRATH, D. E., BROWN, S. M., LAROCHE, Y., DE MEUTTER, J., STANSSENS, P., BOGOWITZ, C. A., FRIED, V. A., ELY, J. A., SOULE, H. R. & VLASUK, J. P. (1994). A hookworm glycoprotein that inhibits neutrophil function is a ligand of the integrin CD11b/CD18. *Journal of Biological Chemistry* **269**, 10008–10015.

MUCHOWSKI, P. J., ZHANG, L., CHANG, E. R., SOULE, H. R., PLOW, E. F. & MOYLE, M. (1994). Functional interaction between the integrin agonist neutrophil inhibitory factor and the I domain of CD11b/CD18. *Journal of Biological Chemistry* **269**, 26419–26423.

NIELSEN, H., ENGELBRECHT, J., BRUNAK, S. & VON HEIJNE, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering* **10**, 1–6.

PAN, T., GROGER, H., SCHMID, V. & SPRING, J. (1998). A toxin homology domain in an astacin-like metalloproteinase of the jellyfish *Podocoryne carnea* with a dual role in digestion and development. *Development, Genes and Evolution* **208**, 259–266.

PFLUGRAD, A., MEIR, J. Y.-J., BARNES, T. & MILLER, D. M. (1997). The groucho-like transcription factor UNC-37 functions with the neural specificity gene *unc-4* to govern motor neuron identity in *C. elegans*. *Development* **124**, 1699–1709.

PRATT, D., ARMES, L. G., HAGEMAN, R., REYNOLDS, V., BOISVENUE, R. J. & COX, G. N. (1992). Cloning and sequence analysis of four distinct cysteine proteases expressed by *Haemonchus contortus* adult worms. *Molecular and Biochemical Parasitology* **51**, 209–218.

PRITCHARD, D. I., BROWN, A., KASPER, G., MCELROY, P., LOUKAS, A., HEWITT, C., BERRY, C., FULLKRUG, R., BECK, E. (1999). A hookworm allergen that strongly resembles calreticulin. *Parasite Immunology* **20** (in the Press).

RAY, C. & MCKERROW, J. H. (1992). Gut-specific and developmental expression of a *Caenorhabditis elegans* cysteine protease gene. *Molecular and Biochemical Parasitology* **51**, 239–250.

RIEU, P., SUGIMORI, T., GRIFFITH, D. L. & ARNAOUT, M. A. (1996). Solvent-accessible regions on the metal ion-dependent adhesion site face of integrin CR3 mediate its binding to the neutrophil inhibitory factor. *Journal of Biological Chemistry* **271**, 15858–15861.

RIEU, P., UEDA, T., HARUTA, I., SHARMA, C. P. & ARNAOUT, M. A. (1994). The A-domain of beta 2 integrin CR3

(CD11b/CD18) is a receptor for the hookworm-derived neutrophil adhesion inhibitor NIF. *Journal of Cell Biology* **127**, 2081–2091.

SCHALLIG, D. H. F., VAN LEEUWEN, M. A. W., VERSTREPEN, B. E. & CORNELISSEN, A. W. C. A. (1997). Molecular characterisation and expression of two putative protective excretory secretory proteins of *Haemonchus contortus*. *Molecular and Biochemical Parasitology* **88**, 203–213.

SCHWEITZ, H., BRUHN, T., GUILLEMARE, E., MOINIER, D., LANCELIN, J.-M., BÉRESS, L. & LAZDUNSKI, M. (1995). Kalicludines and Kaliseptine. Two different classes of sea anemone toxins for voltage-sensitive K^+ channels. *Journal of Biological Chemistry* **270**, 25121–25126.

SHAMANSKY, L. M., PRATT, D., BOISVENUE, R. J. & COX, G. N. (1989). Cuticle collagen genes of *Haemonchus contortus* and *Caenorhabditis elegans* are highly conserved. *Molecular and Biochemical Parasitology* **37**, 73–86.

SHERMAN, D. R., KLOEK, A. P., KRISHMAN, B. R., GUINN, B. & GOLDBERG, D. E. (1992). *Ascaris* hemoglobin gene: plant-like structure reflects the ancestral globin gene. *Proceedings of the National Academy of Sciences, USA* **89**, 11696–11700.

SMITH, M. H. & LEE, D. L. (1963). Metabolism of haemoglobin and haematin compounds in *Ascaris lumbricoides*. *Proceedings of the Royal Society of London, B* **157**, 234–257.

STANNSENS, P., BERGUM, P. W., GANSEMANS, Y., JESPERS, L. Y. L., HUAG, S., MAKI, S., MESSENS, J., LAUWEREYS, M., CAPPELLO, M., HOTEZ, P. J., LASTERS, I. & VLASUK, G. P. (1996). Anticoagulant repertoire of the hookworm *Ancylostoma caninum*. *Proceedings of the National Academy of Sciences, USA* **93**, 2149–2154.

STRINGHAM-DUROVIC, E. G., DIXON, D. K., JONES, D. & CANDIDO, E. P. M. (1992). Temporal and spatial expression patterns of the small heat shock (*hsp-16*) proteins in transgenic *Caenorhabditis elegans*. *Molecular Biology of the Cell* **3**, 21–33.

SWOFFORD, D. L. (1993). *PAUP: Phylogenetic Analysis Using Parsimony, Version 3.1*. Champaign, IL, Illinois Natural History Society.

SWOFFORD, D. L., OLSEN, G. J., WADDELL, P. J. & HILLIS, D. M. (1996). Phylogenetic inference. In *Molecular Systematics* (ed. Hillis, D. M., Moritz, C. & Mable, B. C.), pp. 407–514. Sinauer Associates, Sunderland, MA.

THE *C. ELEGANS* GENOME SEQUENCING CONSORTIUM. (1998). Genome sequence of *Caenorhabditis elegans*: a platform for investigating biology. *Science* **282**, 2012–2018.

THOMPSON, J. D. & HIGGINS, D. G. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673–4680.

TREE, T. I. M., GILLESPIE, A. J., SHEPLEY, K. J., BLAXTER, M. L., TUAN, R. S. & BRADLEY, J. E. (1995). Characterisation of an immunodominant glycoprotein antigen of *Onchocerca volvulus* with homologues in other filarial nematodes and *Caenorhabditis elegans*. *Molecular and Biochemical Parasitology* **69**, 185–195.

TWEEDIE, S., GRIGG, M. E., INGRAM, L. & SELKIRK, M. E. (1993). The expression of a small heat shock protein homologue is developmentally regulated in *Nippostrongylus brasiliensis*. *Molecular and Biochemical Parasitology* **61**, 149–154.

WATERSTON, R., MARTIN, C., CRAXTON, M., HUYNH, C., COULSON, A., HILLIER, L., DURBIN, R., GREEN, P., SHOWNKEEN, R., HALLORAN, N., METZSTEIN, M., HAWKINS, T., WILSON, R., BERKS, M., DU, Z., THOMAS, K., THIERRY-MIEG, J. & SULSTON, J. (1992). A survey of expressed genes in *Caenorhabditis elegans*. *Nature, Genetics* **1**, 114–123.

YANG, J., KLOEK, A. P., GOLDBERG, D. E. & MATHEWS, F. S. (1995). The structure of *Ascaris* hemoglobin domain I at 2·2 Å resolution: molecular features of oxygen avidity. *Proceedings of the National Academy of Sciences, USA* **92**, 4224–4228.

ZHANG, L. & PLOW, E. F. (1996). Overlapping, but not identical, sites are involved in the recognition of C3bi, neutrophil inhibitory factor, and adhesive ligands by the alpha M beta 2 integrin. *Journal of Biological Chemistry* **271**, 18211–18216.