Duplication and divergence: The evolution of nematode globins

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Final published version

Published In:
Journal of nematology

Publisher Rights Statement:
Freely available via Pub Med.

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.
Duplication and Divergence: The Evolution of Nematode Globins

P. W. HUNT,1,2* J. McNALLY,2 W. BARRIS,3 M. L. BLAXTER1

Abstract: In common with many other groups, nematodes express globins with unknown functions. Nematode globin-like genes can be divided into class 1 globins, similar to vertebrate myoglobins, and a wide range of additional classes. Here we show that class 1 nematode globins possess a huge amount of diversity in gene sequence and structure. There is evidence for multiple events of gene duplication, intron insertion and loss between species, and for allelic variation effecting both synonymous and non-synonymous sites within species. We have also examined gene expression patterns in class I globins from a variety of species. The results show variation in the degree of gene expression, but the tissue specificity and temporal specificity of expression may be more conserved in the phyllum. Because the structure-function relationships for the binding and transport of oxygen by globins are well understood, the consequences of genetic variation causing amino acid changes are explored. The gene family shows great promise for discovering unique insights into both structure-function relationships of globins and their physiological roles.

Key words: Genomics, Globin, Evolution, Genetic variation, Gene expression.

Globins are found ubiquitously in Archaea, Eubacteria and Eukaryota (Vinogradov et al. 2006). Globin genes have been identified in most fully sequenced archaeal, eubacterial and eukaryote genomes. However there are a few examples where globin genes have not been found, including some pathogenic bacteria. In the Metazoa, many members of the phylum Nematoda possess characterised globin genes, demonstrably similar to the well-known myoglobins of vertebrates (Blaxter 1993). Some of these genes have a signal peptide-encoding sequence, and the protein products have been demonstrated to be extracellularly located, while others are intracellular. These genes are grouped as class I nematode globins (Hoogewijs et al. 2007). Genome sequencing of Caenorhabditis elegans and close relatives has revealed the existence of a second class of globin-like genes that have just-detectable similarity to myoglobins, and specific roles in neural tissues (Hoogewijs et al. 2007). The roles of class I globins appear to lie mainly in oxygen binding, though whether this reflects a role as a transport protein or as some other system (such as enzymatic catalysis or sequestration) remains unclear (Vinogradov et al. 2008). One route to understanding and unravelling the physiology and evolution of globins is to survey their presence and variability across a wide range of taxa exhibiting different lifestyles, and to use this associative data to focus hypotheses of function.

Nematode globins have been the subject of previous analyses for three main reasons. The first is their abundance in adult nematode tissues (and in some secretions): many strongyle species are bright red (Blaxter et al. 1994a), and the mermithid Mermis ni- grescens has a red, globin-containing ‘eye spot’ (Burr et al. 2000). The second is the discovery that some nematode class I globins have very different oxygen binding properties to their vertebrate homologues, showing, for example, thousand-fold increased affinities for oxygen. These properties made nematode globins a target for studies of general processes of haemoglobin structure-function relationships (Yang et al. 2005). Thirdly, globins were also identified as targets of host immunity in parasitic nematodes, a finding that may relate to both their abundance and presence in secretions (Frenkel et al. 1992; Vercauteren et al. 2003). The genomic structure of nematode class I globins also elicited interest. Vertebrate globins have a conserved three-exon structure, where the positions of the introns fit well with subdomains of the globin fold defined from structural studies. These intron positions have been used as strong evidence for an introns-early, genes-in-pieces model of protein evolution. Nematode globin genes were surprising because they had central introns that were clearly non-homologous to the two introns of vertebrate globins (Sherman et al. 1992; Moens et al. 1992). The identification of two distinct central intron positions (in C. elegans and Ascaris suum globins) lent weight to an introns-late, insertional origin.

Most nematode class I globins isolated previously have been identified by directed searches. There is now a wealth of genomic and transcriptomic data for nematodes, and so we have used the resource of nematode expressed sequence tag (EST) projects and genome projects to identify many additional class I globin genes in a wide variety of parasitic and non-parasitic nematodes. In addition, we survey globin allelic diversity within one species (the strongylo Helmonchus contortus) to probe patterns of within-species diversification in class I globins. We also use EST and quantitative PCR analyses to examine tissue- and stage-specific expression patterns of class I globins in many species across the phyllum.

Materials and Methods

Nomenclature: We name each globin gene identified with the three letter gene name *glb* preceded by

---

Received for publication December 18, 2008.
1Institute of Evolutionary Biology, University of Edinburgh, King’s Buildings, Edinburgh EH9 3JT, United Kingdom.
2CSIRO, Locked Mail Bag 1, Armidale, NSW 2350, Australia.
3CSIRO, 306 Carmody Road, St Lucia QLD 4067, Australia.

The authors gratefully acknowledge the gift of biological materials from K. Teteuk and E. Riga, and the sharing of sequence information from L. Liu, and preliminary data by D. Hoogewijs, S. McClure and D. Emery, technical assistance from J. Kenny and funding support from the Leverhulme Trust and CSIRO. The genomic data for *H. contortus* were generated by Matt Berriemann and colleagues of the Wellcome Trust Sanger Institute, Cambridge UK, Supplementary material for this article is available at the website: http://www.nematodes.org/downloads_area/supplementary_information/Hunt_2009/peter.hunt@csiro.au

This paper was edited by Paula Agudelo.
a three letter abbreviation of the species name (e.g. *Tca*-*glbm* for *Toxocara canis* myoglobin), following precedence (Blaxter 1997). We also refer to globin cDNAs with this italicised notation. For proteins, we follow the SwissProt format: Roman uppercase followed by a five letter species abbreviation (e.g. GLBM_TOXCA). Where genes and proteins can be classified into one of the known subgroups of class I nematode globins they are further designated: *glbm/GLBM* for body wall or intracellular globin, *gbc/GLBC* for cuticular, secreted globin, *glbp/GLBP* for pseudocoelomic globin (and, as some ascarid nematode GLBP proteins consist of two distinct globin-like domains, these are further distinguished as GLBA and GLBB), and *gbf/GLBE* for eye globin. For globins of uncertain subgroup affinity, we follow the name with the cluster or contig number (e.g. *Tca-glbm*; see below for definition of clustering and contigs). When discussing gene structure and protein residues, we have used the alpha-numeric coding system for referring to helices and interhelix residues. We have followed the alignment of GLBA_ASCSU (*Ascaris D1*) and GLBM_PHYCA (*Physeter catodon* [sperm whale] myoglobin) based on crystal structures presented in Yang *et al.* 1995 for our alignment of amino acid sequence (Supplementary Table 2), and residues are named after the corresponding GLBM_PHYCA residues. There are a number of places in the alignment where there is no corresponding GLBM_PHYCA residue. Where these amino acid residues are mentioned, we name them after the preceding GLBM_PHYCA residue followed by a letter as in the ef interhelix in Figure 2 and supplementary Table 2.

**Directed isolation of new class I nematode globin genes**

*Syngamus trachea* globin genes: *S. trachea* adult pairs (a gift of Dr. E. Riga, Aberystwyth [now at Washington State University]) were obtained from fledging crows and frozen at -80°C. Genomic DNA, RNA and protein were prepared simultaneously from four pairs using TRISOLV reagent (Biotecx Laboratories, Inc.). Total RNA was reverse transcribed into first strand cDNA and PCR performed with degenerate primers SG1 and SG3 designed from an alignment of previously isolated strongylid globins (Blaxter et al. 1994a; Vanfleteren et al. 1994) and, oligo d(T). After sequencing the amplicons obtained, additional reverse primers were designed and used with the nematode spliced leader, SL1 (5'-GGTGTATTTATTCCAGGTTTGAG-3') to obtain the 5' part of the cDNA. Full length cDNA was subsequently isolated with specific primers. Analysis of the sequences revealed two distinct genes had been cloned. The two genes cloned (*Str-glbm* and *Str-glbh*) were identified as a myoglobin and a cuticle globin respectively. Genomic DNA corresponding to these two loci was isolated by PCR and cloned before being sequenced on both strands.

*Toxocara canis* globin genes: Whole adult *T. canis* (a gift of K. Tetteh, Edinburgh) were ground in liquid nitrogen and RNA extracted using the Ultraspec RNA isolation system (Biotecx Laboratories, Inc.). Oligo-dT primed cDNA was generated using the GeneAmp™ kit (Perkin Elmer). Body wall globin cDNAs (*Tca-glbm*) were isolated by amplification of first strand cDNA using an upstream primer *TcG-S* (5'-AATGGCCGAC GCCATGCTTG-3') based on partial cDNA sequence (kindly provided by Dr L. Liu, Harvard) and a GC-anchored oligo d(T) primer DGDT (5'-GGCATGAACATGTT YGARMAYTAY CC-3' and APG-dR 5'-GGARTKTYT CRAACATRTGYTTRTA-3'); derived from an alignment of the four previously published ascarid pseudocoelomic globin domains (Dixon *et al.* 1991; Sherman *et al.* 1992), in combination with DGDT and SL1 primers respectively. PCR products (approximately 200 bp for APG-dR - SL1 and 500 bp for APG-dF - DGDT) were cloned and sequenced. To confirm the size and sequence of the predicted cDNA a second RT-PCR experiment was conducted using a new 5' primer TcpG-S (5'-TGCGATCTTTTGCGTTGTTTG-3') and DGDT. This PCR product was also cloned and sequenced. Both *T. canis* globin transcripts are trans-spliced to SL1. Genomic DNA from individual *T. canis* nematodes was amplified with primers derived from cDNA, and fragments cloned and sequenced. Introns were predicted with reference to the relevant cDNA sequence. Intron position in the proteins was derived through alignment of the encoded proteins with the *A. suum* pseudocoelomic globin domain A (GLBA_ASCSU), for which the three dimensional structure has been resolved (Yang *et al.* 1995).

**Identification of class I nematode globin genes from EST data**

Model-like genes were identified in EST datasets clustered into putative genes using PartiGene in the NEMBASE3 database (Parkinson *et al.* 2004; Wasmuth *et al.* 2008) or, for *H. contortus* data, StackPACK in WormSIS - CSIRO. For some analyses, multiple clusters assembled algorithmically were collapsed together to form new consolidated clusters by eye. This was especially necessary for some *Teladorsagia circumcincta* clusters. A description of the EST clusters is given in Table 1a, and Supplementary Table 1 lists the GenBank accessions for ESTs comprising the clusters. From each cluster of ESTs a consensus sequence or contig was predicted. Table 1b lists other globin sequences used in the phylogenetic and protein structure analyses described below. Contigs from WormSIS and those which have been adjusted are indicated in Table 1a: these sequences can be provided by the authors on request. NEMBASE3 data are freely available at http://www.nematodes.org/nembase3.

**Phylogenetic analysis:** We assembled an alignment of inferred amino acid sequences from the EST clusters...
<table>
<thead>
<tr>
<th>Species</th>
<th>Cluster</th>
<th>Cluster origin</th>
<th>Number of ESTs</th>
<th>SNPs identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ancylostoma caninum</td>
<td>cn01487</td>
<td>NEMBASE</td>
<td>2</td>
<td>excluded</td>
</tr>
<tr>
<td>Ancylostoma caninum</td>
<td>cn03829</td>
<td>NEMBASE</td>
<td>3</td>
<td>excluded</td>
</tr>
<tr>
<td>Ascaris lumbricoides</td>
<td>cn00043</td>
<td>NEMBASE</td>
<td>41</td>
<td>1</td>
</tr>
<tr>
<td>Ascaris suum</td>
<td>cn00004</td>
<td>NEMBASE</td>
<td>82</td>
<td>13</td>
</tr>
<tr>
<td>Ascaris suum</td>
<td>cn00019</td>
<td>NEMBASE</td>
<td>5</td>
<td>excluded</td>
</tr>
<tr>
<td>Ascaris suum</td>
<td>cn00780</td>
<td>NEMBASE</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Ascaris suum</td>
<td>cn17423</td>
<td>NEMBASE</td>
<td>2</td>
<td>excluded</td>
</tr>
<tr>
<td>Ancylostoma ceylanicum</td>
<td>orphan</td>
<td>NEMBASE</td>
<td>1</td>
<td>excluded</td>
</tr>
<tr>
<td>Ancylostoma ceylanicum</td>
<td>cn00272</td>
<td>NEMBASE</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>Ancylostoma ceylanicum</td>
<td>cn00544</td>
<td>NEMBASE</td>
<td>4</td>
<td>excluded</td>
</tr>
<tr>
<td>Ancylostoma ceylanicum</td>
<td>cn01816</td>
<td>NEMBASE</td>
<td>2</td>
<td>excluded</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>Clones of ZK637.13</td>
<td>WormBase</td>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>Haemonchus contortus</td>
<td>orphans</td>
<td>NEMBASE</td>
<td>2</td>
<td>excluded</td>
</tr>
<tr>
<td>Haemonchus contortus</td>
<td>cn00868</td>
<td>WormSIS</td>
<td>24</td>
<td>43</td>
</tr>
<tr>
<td>Haemonchus contortus</td>
<td>cn01319</td>
<td>WormSIS</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>Haemonchus contortus</td>
<td>cn01329</td>
<td>WormSIS</td>
<td>29</td>
<td>66</td>
</tr>
<tr>
<td>Haemonchus contortus</td>
<td>cn01356</td>
<td>WormSIS</td>
<td>2</td>
<td>excluded</td>
</tr>
<tr>
<td>Haemonchus contortus</td>
<td>cn01377</td>
<td>WormSIS</td>
<td>23</td>
<td>53</td>
</tr>
<tr>
<td>Haemonchus contortus</td>
<td>cn01583</td>
<td>WormSIS</td>
<td>4</td>
<td>excluded</td>
</tr>
<tr>
<td>Haemonchus contortus</td>
<td>cn01747</td>
<td>WormSIS</td>
<td>15</td>
<td>7</td>
</tr>
<tr>
<td>Haemonchus contortus</td>
<td>cn08501</td>
<td>NEMBASE</td>
<td>3</td>
<td>excluded</td>
</tr>
<tr>
<td>Heterodera glycines</td>
<td>orphan</td>
<td>NEMBASE</td>
<td>1</td>
<td>excluded</td>
</tr>
<tr>
<td>Heterodera glycines</td>
<td>cn06223</td>
<td>NEMBASE</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Litomosoides sigmodontis</td>
<td>orphan</td>
<td>NEMBASE</td>
<td>1</td>
<td>excluded</td>
</tr>
<tr>
<td>Meloidogyne chaiuodonti</td>
<td>cn00803</td>
<td>NEMBASE</td>
<td>2</td>
<td>excluded</td>
</tr>
<tr>
<td>Meloidogyne hapla</td>
<td>cn02564</td>
<td>NEMBASE</td>
<td>3</td>
<td>excluded</td>
</tr>
<tr>
<td>Meloidogyne incognita</td>
<td>cn00586</td>
<td>NEMBASE</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Meloidogyne javanica</td>
<td>orphan</td>
<td>NEMBASE</td>
<td>1</td>
<td>excluded</td>
</tr>
<tr>
<td>Necator americanus</td>
<td>orphan</td>
<td>NEMBASE</td>
<td>1</td>
<td>excluded</td>
</tr>
<tr>
<td>Necator americanus</td>
<td>cn00041</td>
<td>NEMBASE</td>
<td>3</td>
<td>excluded</td>
</tr>
<tr>
<td>Necator americanus</td>
<td>cn00088</td>
<td>NEMBASE</td>
<td>2</td>
<td>excluded</td>
</tr>
<tr>
<td>Nippostrongylus brasiiliensis</td>
<td>orphans</td>
<td>NEMBASE</td>
<td>1</td>
<td>excluded</td>
</tr>
<tr>
<td>Nippostrongylus brasiiliensis</td>
<td>cn00124</td>
<td>NEMBASE</td>
<td>2</td>
<td>excluded</td>
</tr>
<tr>
<td>Nippostrongylus brasiiliensis</td>
<td>cn00197</td>
<td>NEMBASE</td>
<td>19</td>
<td>17</td>
</tr>
<tr>
<td>Nippostrongylus brasiiliensis</td>
<td>cn00328</td>
<td>NEMBASE</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>Ostertagia ostertagi</td>
<td>orphan</td>
<td>NEMBASE</td>
<td>1</td>
<td>excluded</td>
</tr>
<tr>
<td>Ostertagia ostertagi</td>
<td>cn00190</td>
<td>NEMBASE</td>
<td>40</td>
<td>38</td>
</tr>
<tr>
<td>Ostertagia ostertagi</td>
<td>cn00214</td>
<td>NEMBASE</td>
<td>2</td>
<td>excluded</td>
</tr>
<tr>
<td>Ostertagia ostertagi</td>
<td>cn03902</td>
<td>NEMBASE</td>
<td>2</td>
<td>excluded</td>
</tr>
<tr>
<td>Ostertagia ostertagi</td>
<td>cn03548</td>
<td>NEMBASE</td>
<td>2</td>
<td>excluded</td>
</tr>
<tr>
<td>Onchocerca volvulus</td>
<td>cn00634</td>
<td>NEMBASE</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>Parascaris falcipennis</td>
<td>orphan</td>
<td>NEMBASE</td>
<td>1</td>
<td>excluded</td>
</tr>
<tr>
<td>Strongyloides stercoralis</td>
<td>cn00221</td>
<td>NEMBASE</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Strongyloides stercoralis</td>
<td>cn02295</td>
<td>NEMBASE</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>Toxocara canis</td>
<td>cn0537c</td>
<td>NEMBASE</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>Toxocara canis</td>
<td>cn0563c</td>
<td>NEMBASE</td>
<td>4</td>
<td>excluded</td>
</tr>
<tr>
<td>Teladorsagia circumcincta</td>
<td>orphan</td>
<td>NEMBASE</td>
<td>1</td>
<td>excluded</td>
</tr>
<tr>
<td>Teladorsagia circumcincta</td>
<td>cn00008</td>
<td>NEMBASE</td>
<td>34</td>
<td>57</td>
</tr>
<tr>
<td>Teladorsagia circumcincta</td>
<td>cn00009</td>
<td>NEMBASE</td>
<td>23</td>
<td>50</td>
</tr>
<tr>
<td>Teladorsagia circumcincta</td>
<td>cn00032</td>
<td>NEMBASE</td>
<td>4</td>
<td>excluded</td>
</tr>
<tr>
<td>Teladorsagia circumcincta</td>
<td>cn00804</td>
<td>NEMBASE</td>
<td>3</td>
<td>excluded</td>
</tr>
<tr>
<td>Teladorsagia circumcincta</td>
<td>cn00173</td>
<td>NEMBASE</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Teladorsagia circumcincta</td>
<td>cn01113</td>
<td>NEMBASE</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>Trichuris muris</td>
<td>cn00180</td>
<td>NEMBASE</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Trichuris muris</td>
<td>cn01615</td>
<td>NEMBASE</td>
<td>2</td>
<td>excluded</td>
</tr>
<tr>
<td>Trichinella spiralis</td>
<td>cn03467</td>
<td>NEMBASE</td>
<td>3</td>
<td>excluded</td>
</tr>
<tr>
<td>Trichinella spiralis</td>
<td>cn00408</td>
<td>NEMBASE</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>Trichurus vulpis</td>
<td>cn00077</td>
<td>NEMBASE</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Trichurus vulpis</td>
<td>cn00688</td>
<td>NEMBASE</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Xiphinema index</td>
<td>cn00721</td>
<td>NEMBASE</td>
<td>3</td>
<td>excluded</td>
</tr>
<tr>
<td>Zeldia punctata</td>
<td>orphan</td>
<td>NEMBASE</td>
<td>1</td>
<td>excluded</td>
</tr>
</tbody>
</table>

* Some ESTs were not clustered with any other sequence. These are called “orphans” in this column of the table.

* Where there were fewer than 5 ESTs in a cluster, these were not used to infer SNPs and are identified as “excluded” in this column of the table (see materials and methods).

* These clusters encode cDNA which is nearly identical to genes described in Table 1b.

* For GenBank accessions, see Supplementary Table 1.

* These clusters differed from those presented in NEMBASE, see materials and methods.
(B) – Other nematode globins.

<table>
<thead>
<tr>
<th>Species</th>
<th>Gene Name*</th>
<th>GenBank Accession(s)</th>
<th>Corresponding cluster</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ascaris suum</td>
<td>As-glbm</td>
<td>U17337</td>
<td>cn00004</td>
<td>(Blaxter et al. 1994b)</td>
</tr>
<tr>
<td>Ascaris suum</td>
<td>As-glbp (GLBA, GLBB)†</td>
<td>L03531</td>
<td>cn00019</td>
<td>(Sherman et al. 1992)</td>
</tr>
<tr>
<td>Brugia malayi</td>
<td>Bma-glgb</td>
<td>NW_001893041</td>
<td>N/A</td>
<td>(Hoogewijs et al. 2008)</td>
</tr>
<tr>
<td>Caenorhabditis briggsae</td>
<td>Cbr-glgb</td>
<td>U48298, U48299, U48291</td>
<td>N/A</td>
<td>(Kloeck et al. 1996)</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>Ceg-glβ</td>
<td>NM_066573</td>
<td>N/A</td>
<td>(Kloeck et al. 1996)</td>
</tr>
<tr>
<td>Caenorhabditis remanei</td>
<td>Creg-glβ</td>
<td>U48295, U48294, U48293, U48292</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Mermis nigenscens</td>
<td>Mnr-glbm</td>
<td>AF138291, AF138296</td>
<td>N/A</td>
<td>(Burr et al. 2000)</td>
</tr>
<tr>
<td>Mermis nigenscens</td>
<td>Mnr-glbe</td>
<td>AF138295, AF138294, AF138293, AF138292,</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>Nippostrongylus brasilensis</td>
<td>Nbr-glbm</td>
<td>L20895, L25872, L25873</td>
<td>cn00095**</td>
<td>(Blaxter et al. 1994a)</td>
</tr>
<tr>
<td>Nippostrongylus brasiliensis</td>
<td>Nbr-glbc</td>
<td>L20896, L25874</td>
<td>cn00197**</td>
<td>(Blaxter et al. 1994a)</td>
</tr>
<tr>
<td>Pseudodermatocoeus decipiens</td>
<td>Pde-glbp (GLBA, GLBB)†</td>
<td>M6298</td>
<td>N/A</td>
<td>(Dixon et al. 1992)</td>
</tr>
<tr>
<td>Syngamus trachea</td>
<td>Str-glbm</td>
<td>AF370722</td>
<td>N/A</td>
<td>This report</td>
</tr>
<tr>
<td>Syngamus trachea</td>
<td>Str-glbe</td>
<td>AF370721</td>
<td>N/A</td>
<td>This report</td>
</tr>
<tr>
<td>Toxocara canis</td>
<td>Tca-glbp</td>
<td>AF370726</td>
<td>cn00537</td>
<td>This report</td>
</tr>
<tr>
<td>Toxocara canis</td>
<td>Tca-glbm</td>
<td>AF370724, AF370723</td>
<td>cn00563</td>
<td>This report</td>
</tr>
<tr>
<td>Trichostrongylus colubriformis</td>
<td>Tcc-glbe</td>
<td>M63263</td>
<td>N/A</td>
<td>(Frenkel et al. 1992)</td>
</tr>
</tbody>
</table>

* - Name used in this report, ** - Not an exact match, † - Di-domain globins domains are separated for phylogenetic and structural analyses (Figures 1 and 2).

(see next section) and nematode globins from GenBank. The full alignment is available in Supplementary Table 2. Predicted amino acid sequences from the consensus of each cluster were aligned by eye. For phylogenetic reconstruction, globins represented by single ESTs (orphans in Table 1a) were not used, as the quality of the predicted amino acid sequence could not be affirmed. We did not utilise the genome-project predicted Pristionchus pacificus globin sequences (see http://www.pristionchus.org) as our predictions disagree with those derived computationally (See Supplementary Figures 4-7. Preliminary trees were generated using a neighbour joining (using default settings in PAUP*, Unix version (Swofford 2000)), and these showed that the sequences from genera Trichinella, Trichuris, Xiphinema and Mermis formed a separate group. This observation concurs with traditional and molecular phylogenetic analyses, which place these taxa in the Dorylaimia, distinct from all other taxa studied, which are in the Chromadorea (Meldal et al. 2007). This was also observed in analyses using maximum parsimony and so this group was defined as the outgroup for further analyses. These subsequent analyses were undertaken using Bayesian, parsimony and neighbour joining methods. Bayesian analyses were carried out on the portion of the alignment corresponding to the globin domain of the A. suum myoglobin (i.e. excluding the secretory leader peptides and polar zipper extensions) in MrBayes 3.1.2 (Ronquist and Huelsenbeck 2003). A flat prior was assumed on the underpinning amino acid substitution model, and two chains of MCMC analysis were run for 2.5 million generations. Inspection of the chains, sampled every 100 generations, in Tracer (version 1.4.1; A. Rambaut; http://tree.bio.ed.ac.uk/software/tracer/) showed that stationarity had been achieved after ~50,000 generations, and the split frequency between the chains was less than 0.001. A consensus tree was derived from the last 2.4 million generations, and Bayesian posterior probability support estimated. The consensus tree presented in figure 1 was examined and annotated in FigTree (version 1.2.2; Andrew Rambaut; http://tree.bio.ed.ac.uk/software/figtree/). Maximum parsimony analyses were carried out on the reduced alignment (as for MrBayes) in PAUP* version 4.10b (Swofford) by full heuristic search, with reliability of the resulting tree estimated by performing 10,000 bootstrap resamplings. Neighbour joining analyses were carried out in PAUP* version 4.10b (Swofford) using the BioNJ method, with reliability of the resulting tree estimated by performing 10,000 bootstrap resamplings.

Identification and verification of single nucleotide polymorphisms (SNPs) in globin genes Prediction of SNPs from EST clusters: An analysis program (findsnp1.pl) was written using perl. The program interrogates alignments, ignoring sequence differences occurring in the beginning and ending of each sequence in the alignment in order to minimize the effects of low quality sequence. SNPs are predicted when one or more sequences have non-consensus base pairs at a position in the alignment. For the results reported in this paper, we report predicted SNPs only from those alignments where there were five or more ESTs, and also from only those positions in alignments where there were five or more sequences aligned. Because of the real risk of a single observation being due to sequencing errors, rather than true allelic polymorphism, the number of observed alternative bases had to be found in 2 or more sequences for the SNP to be considered. For all of our alignments (the largest cluster contains 73 sequences), this “two-or-more” rule approximates rejecting each SNP if the lower limit of the 99% confidence interval for the minor allele frequency drops below zero. We also excluded SNP polymorphisms from the data where
there was more than one minor allele and none of these were observed more than once. As a consequence, we have reported some SNPs to be bi-allelic when there has been only a single observation of third or fourth alternative alleles.

Confirmation of SNPs in two Haemonchus contortus globin genes: Analyses of H. contortus globin clusters showed that Hco-glb_cn01320 and Hco-glb_cn01319 were very similar, and that both had a large number of predicted SNPs. We attempted to amplify fragments of both these genes from three individual H. contortus DNA from two males and the head region of one female H. contortus adults were extracted using Chelex resin (BioRad) as described previously (Hunt et al. 2008). The primers cn1320-F (5’-AAAGCTTCACTGCCGATGAC-3’) and cn1320-R2 (5’-AGCATGAGGGAGTCCAAGAG-3’) were used to amplify a segment of Hco-glb_cn01320 and the primers cn1319-F (5’-CCATCCAGAAAATCGCAAAT-3’) and cn1319-E (5’-CCATTATACATGTGGAAGACCGAGGCTTGCGAGATG-3’) were used for Hco-glb_cn01319.

As part of the work it was determined that cn1320-F is not specific for Hco-glb_cn01320, and can be used to amplify a segment of Hco-glb_cn01319 using cn1319-E as a reverse primer. PCR products amplified using proof reading Pfu DNA polymerase (Promega) were subsequently cloned into the pGEM-Teasy vector (Promega). Several clones (Figure 4A) from each amplification were sequenced and aligned by eye with both the EST.
cluster consensus and individual reads from the H. contortus genome sequencing project (http://www.sanger.ac.uk/Projects/H_contortus/). The alignment was analysed using PAUP* as described above. By aligning the whole genome shotgun (WGS) sequences obtained with the EST cluster alignment, three categories of SNPs were identified; those which were evident in both alignments, those in the EST alignment only and those in the WGS only.

Gene expression analyses: Multiplex RT-PCR: T. canis tissues for RT-PCR experiments were obtained by dissecting live nematodes, and rapidly freezing the tissues in liquid nitrogen. Frozen tissue was either ground in liquid nitrogen (whole females, males and head sections) or homogenised in Ultraspec RNA reagent using microhomogenisers. RNA and resultant cDNA was then isolated as above. Multiplex RT-PCR for detecting globin transcripts in T. canis tissue was performed by co-amplifying either Tca-glbm or Tca-glbp with T. canis ribosomal protein L3 (Tca-rpl-3) (GenBank accession U17358) transcript as an internal control. Primers used were TcpG-S and TcpG-E (5’-GGTTGCTATGACTGCTTTCATGTGG-3’) for Tca-glbp or TcG-S and TcG-E (5’-GAAATGGTCTAATGGGGT-3’) for Tca-glbm with TcR-1F (5’-CGTTTATCGCATTCAAGGCTGG-3’) and TcR-2R (5’-GCAATCCTCGCTAAGATGTTCAGC-3’) for Tca-rpl-3. Products were analysed on agarose gels.

Expression levels of EST clusters: For genes identified in the EST datasets, the tissue and stage of origin of the libraries from which the ESTs were derived was ascertained, and these data used to estimate tissue- and stage-specificity of expression of each gene. Expression levels derived in this way are expressed as proportions of the total numbers of ESTs generated from the relevant libraries.

**RESULTS**

Toxocara canis globin genes: The dog ascaridid T. canis lives as an adult in the dog gut, but infective larvae can cause human disease if eggs are ingested. The presence of globins in this nematode was not known, but the related A. suum has two well characterised isoforms. Two globin genes, Tca-glbm and Tca-glbp, were identified. The predicted protein GLBM_TOXCA has no secretory leader peptide and a high amino acid identity (84%) to A. suum myoglobin (GLBM_ASCSU). The central region (B6 to F9) of the predicted protein sequences differ by ten residues, five of which are highly conserved substitutions (Figure 2). This classifies GLBM_TOXCA as an intracellular, tissue or myoglobin type molecule. In contrast GLBP_TOXCA has a secretory leader sequence and is highly similar to the peripheral fluid hemoglobins of A. suum (Sherman et al. 1992) (GLBP ASCSU; GLBA domain 75% identity; GLBB domain 67% identity) and *Pseudoterranova decipiens* (Dixon et al. 1991) (GLBP PSEDE; GLBA domain 70% identity; GLBB domain 60% identity). GLBP_TOXCA however, has only a single globin domain and lacks the predicted polar zipper sequence (de Baere et al. 1992) possessed by GLBP_ASCSU and GLBP_PSEDE. GLBP_TOXCA is classified as a peri-teric (pseudocelomic) fluid type hemoglobin.

**Syngamus trachea globin genes:** The strongylid parasitic nematode *S. trachea* lives in the airways of avian hosts, a highly oxygenated environment, so globins from this species were cloned to enable comparison with other strongylids such as *H. contortus* which parasitise the micro-aerobic, lumenal surface of the gastrointestinal tract. PCR was performed on oligo-dT primed (primer DGDT) first strand cDNA using primers SG1 and DGDT, and SG3 and DGDT. Two products of differing size were amplified and cloned. Sequencing revealed that these corresponded to distinct products both with nucleotide sequence similarity with nematode globins. Oligonucleotides designed to the unique regions downstream from the predicted termination codon in each gene were then used in PCR with SL1 as the upstream primer and SL1-DGDT cDNA as target. Products of the expected size (~500 base pairs) were cloned and sequenced and the complete cDNA sequences of the two transcripts assembled.

Reverse transcription and SL1-DGDT amplification of *S. trachea* poly(A)+ RNA resulted in a smear of products from ~100 bp to >3 kb, with two more abundant bands discernable. As the globins are abundant protein products of the nematodes, we reasoned that the abundant transcripts might encode them. The larger of the abundant bands (~700 bp) was excised and blunt end cloned. Several clones were sequenced and proved to encode distinct proteins including a homologue of the *Nippostrongylus brasiliensis* Hsp20 gene (Tweedie et al. 1993) and a gene with a cysteine-rich repeat motif with homologues in *C. elegans* and *Brugia malayi* (Fuhrman et al. 1995; Blaxter 1996), as well as globin transcripts. Two globin cDNAs were identified. *Str-glbm* encodes a protein of 158 amino acids and *Str-glbp* encodes a protein of 170 amino acids, which has an N-terminal hydrophobic extension predicted to be a secretory signal peptide.

**Nematode globins defined by clustering ESTs:** Using an E-value cutoff of 1xl0^-6, we distinguished 85 clusters from NEMBASE3 (http://www.nematodes.org/nembase3/) (Wasmuth et al. 2008) that had significant matches to HMMPfam IPR000971 (Table 1). These genes are from 25 species, with the number of clusters for each taxon varying between 12 (*T. circumcincta*) and 1 (12 species). A cluster of *C. elegans* ESTs corresponding to the globin gene Ce-glb-1 (ZK037.13) was also included for comparison. The total number of ESTs available for each species varies, and for those with low numbers it is possible that further globin sequences remain undefined.

Protein predictions derived from ascaridid (A. suum, *A. lumbricoides* and T. canis) clusters were aligned with
GLBM_TOXCA and GLBP_TOXCA. Tentatively, this EST clustering shows that *A. suum* has one additional extracellular globin quite distinct from the published sequence (*As-glb _cn00780*).

*H. contortus* globins were originally divided into 22 clusters, but re-analysis using an alternative clustering algorithm placed these into 8, based on more conservative criteria. For *H. contortus*, the globin EST cluster

![Partial alignment of nematode globin sequences showing homology to GLBP_ASCSU domain A with identical residues shaded grey.](image)

**Fig. 2.** Partial alignment of nematode globin sequences showing homology to GLBP_ASCSU domain A with identical residues shaded grey. The three dimensional structure of this gene "Ascaris D1" is illustrated directly above the alignment, and the three dimensional structure and sequence of *Phyre catodon* myoglobin GLBM_PHYCA is shown above D1. Helix numbering below the Helix designation letters follows the *P. catodon* structure. The numbering at the top of the figure indicates the portion of the alignment from which the figure is derived. The entire alignment is available as Supplementary Table 2. Arrows indicate the key ligand coordinating residue B10 and E7. The heme coordinating residue F8 and the ligand pocket gating residue cd1.

GLBM_PHYCA and GLBP_PHYCA. Tentatively, this EST clustering shows that *A. suum* has one additional extracellular globin quite distinct from the published sequence (*As-glb _cn00780*).
consensus sequences were analysed using SignalP (Emanuelsson et al. 2007) to check for secretory leader peptides. The results indicate that Hco-glb.cn00868, cn01320, cn01319 and cn01747 are likely to be extracellular globins and Hco-glb.cn01356, cn08501, cn01583 and cn01377 are most likely to be intracellular.

After adjusting alignments 61 globin EST clusters were defined (Table 1A and Supplementary Table 1). From literature and database searches we also found thirteen sequences from previously published work (Table 1B). Combining the EST clusters, the four new sequences generated above and the sequences from the literature there are 72 distinguishable nematode globin sequences now identified when we exclude six EST clusters that are near identical to published sequences. Of the 72 putative genes, 62 had sufficient sequence to construct alignments and explore relationships between genes (Supplementary Table 2).

**Evolution of nematode globin genes:** Using nematode globin genes from published reports, our data for *T. canis* and *S. trachea*, and globins defined by EST clustering (Table 1 - excluding orphan sequences), we inferred a gene genealogy and compared this to published species trees constructed using small subunit ribosomal RNA (SSU rRNA) sequences (Meldal et al. 2007). As expected, due to a high degree of amino acid dissimilarity the consensus globin trees are not well resolved and many internal nodes are polytomous (Figure 1 and Supplementary figures 1 and 2). However, some clades are well supported in both the Bayesian analysis and by bootstrap analysis of the maximum parsimony (ML) and minimum evolution (ME) trees, and in some cases these groupings conflict with the family level structure of the published SSU rRNA species relationship tree (Meldal et al. 2007). In the Bayesian analysis, the most significant discrepancy between the SSU rRNA RNA tree and the globin tree is the separation of the majority of strongylid sequences into the clade marked “S” in Figure 1. This clade is separate from the major chromadorean clade and *Meloidogyne* spp. (Tylenchida) and *Caenorhabditis* spp. are shown as sister groups to the exclusion of all Strongylida. This conflicts with the SSU rRNA tree where Strongylida and Caenorhabditidae are sister groups to the exclusion of Tylenchida; this may be a long-branch attraction artefact, and the group containing both Caenorhabditidae and Tylenchida has low statistical support. In the ML and ME trees shown in Supplementary figures 1 and 2, the Caenorhabditidae/Tylenchida clade has less than 50% bootstrap support and so is not shown.

Within the Strongylida, gene duplication has been rampant. There are multiple genes in all genera sampled through EST projects (*Ancylostoma, Haemonchus, Nippostrongylus, Ostertagia, Teladorsagia*). In some cases, there is an indication of homology between genes from separate taxa (e.g. clades M1 and M2 in Figure 1, which contain sequences from three trichostrongylid taxa with high bootstrap support). Some strongylid-derived protein sequences are significantly divergent from all other sequences in the group and form a separate clade (for example, GLB_ANCCA cn01487 and GLBM_SYNTR). These genes could define a separate family of globins perhaps lost in some strongylids such as *H. contortus*.

**Correspondence of EST clusters with genes:** The EST clusters we have generated may not correspond exactly to separate genes because of the complication introduced by sequencing errors and allelic variation. These sources of variation may contribute to a division of EST sequences, derived from the transcripts of a single gene, into more than one cluster. Our phylogenetic analysis provides evidence that some clusters are likely to be derived from separate genes when clusters from other taxa are consistently grouped as sister sequences to the exclusion of other sequences from the same taxon. The groups M1 and M2 described above are examples. In other instances, the correspondence of clusters to genes is more ambiguous (e.g. clade C1 in Figure 1 which contains three clusters from *H. contortus*). We undertook a number of investigations to try and establish whether the seven *H. contortus* predicted genes actually represent separate genes or are representative of extremely divergent alleles at fewer than seven loci.

First, alignment of the predicted position of SNPs (see below) within each cluster was compared to the alignment of the consensus nucleotide sequences of each cluster. If allelic variation had been the major determinant of division of sequences into multiple clusters, it would be expected that SNP positions would rarely correspond to divergent nucleotides between clusters. This was not the case; some divergent nucleotides corresponded to intra-cluster SNP predictions whilst others did not.

Second, alignment of cluster consensus sequences with raw sequence data from the incomplete *H. contortus* genome project called Sanger reads in figure 4A and below (http://www.sanger.ac.uk/Projects/H_contortus/) was undertaken to see if there was a correspondence of genomic loci and EST cluster sequences. For the four intracellular globin EST clusters, there was a significant match for only *Hco-glb.cn08501* to genomic sequences (Sanger reads haem-1155c08.q1k, and haem-1033p03.q1k). Because the other three intracellular globin EST clusters display many differences to *Hco-glb.cn08501* and are not matched to any Sanger read by Blastn, this implies that these four clusters must define at least two “real” genes. For the four extracellular globins, there was evidence of four divergent genes based on genomic sequence. No genomic sequence was a close match to *Hco-glb.cn00868*, whilst there were matches to *Hco-glb.cn01320* (haem-70g17.q1k, haem-57j02.q1k, haem-1056p10.q1k), *Hco-glb.cn01319* (haem-439c10.q1k, haem-804l05.q1k) and *Hco-glb.cn01747* (haem-1075m01.q1k, haem-1099k13.p1k).
Divergent intron and 3’UTR sequences allowed alignment of Hco-glb cn01320, Hco-glb cn01319 and Hco-glb cn01747 to separate genomic contigs, and those matching Hco-glb cn01319 and Hco-glb cn01320 were also used in phylogenetic analysis where they grouped separately (Figure 4A). This was an interesting observation as Hco-glb cn01320 and Hco-glb cn01319 share a high level of exonic nucleotide identity, but can be separated based on genomic sequence; implying that these "genes" have arisen from a recent duplication. We investigated this further by sequencing clones from both a cDNA library and from genomic DNA amplified using primers we designed to be specific for either Hco-glb cn01320 or Hco-glb cn01319. The sequence of the clones obtained was quite similar using either set of primers or template DNA, but amongst a high degree of apparent allelic variation (described below), the sequences obtained using the Hco-glb cn01319 primers are distinct from those obtained using Hco-glb cn01320 primers. One way of illustrating this is via phylogenetic analysis, and this is shown in Figure 4A.

**Nematode globin gene structure:** Genomic sequences are available for class I globins from A. suum, B. malayi, C. elegans, Caenorhabditis briggsae, Caenorhabditis remaneii, Mermis nigrescens, N. brasiliensis, P. pacificus, P. decipiens, S. trachea, T. canis and H. contortus, and intron positions assigned (Table 2). The occurrence of conserved intron positions, relative to the predicted protein three-dimensional structure, in globin genes is well documented, with 5’ and 3’ introns at B12.2 and G7.0 occurring in many taxa across the plant and animal kingdoms, including in many nematode globins (Table 2). In the Ppa-glb genes (see Supplementary Figures 4-7) the predicted B helix is shorter than in the other globins, and in our amino acid alignment (Supplementary Table 2) residue B12 is represented by a gap. Because of this, the Ppa-glb introns I could be considered to be at B11.2. Central introns are found in nematodes and plants, and the placement of these is not conserved between these groups (Sherman et al. 1992; Moens et al. 1992). In nematodes, the central intron may be at E8.1 (ascarid nematodes; Sherman et al. 1992) and we observed this intron position in both T. canis genes described here. The B. malayi globin (see Supplementary Figure 3) also probably has an intron at E8.1, but the intron donor sequence at this position is not optimal, and the alternative position of E9.0 cannot be ruled out. In other nematode clades the second intron may be located at E3.2, as in Caenorhabditis spp. and Strongylida (Blaxter et al. 1994a; Kloek et al. 1996) and we observed this intron in the S. trachea and H. contortus globin genes described here and the P. pacificus genes predicted from genome sequence also have introns at E3.2. Intron loss is common in nematode globins, whereas it is relatively rarely observed in globins from chordates and angiosperms where divergent gene families from distantly related species often retain the same gene structure (Fuchs et al. 2005; Hunt et al.

### Table 2. Intron positions in Class I globin genes from nematodes.

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>B12.2</th>
<th>E3.2</th>
<th>E8.1</th>
<th>ef4b.0</th>
<th>ef4d.0</th>
<th>G7.0</th>
<th>H14.2</th>
<th>Inter-domain introns **</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>As-glb</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>Blaxter et al., 2008</td>
</tr>
<tr>
<td>As-glb GLBA</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>Sherman et al., 1992</td>
</tr>
<tr>
<td>Av-glb GLBB</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>Hoogewijs et al., 2008</td>
</tr>
<tr>
<td>Bma-glb ***</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>Kloeck et al., 1996</td>
</tr>
<tr>
<td>Cbr-glb</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>Kloeck et al., 1996</td>
</tr>
<tr>
<td>Ce-glb-1</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>Kloeck et al., 1996</td>
</tr>
<tr>
<td>Cre-glb</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>Kloeck et al., 1996</td>
</tr>
<tr>
<td>Mni-glbm</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>Burr et al., 2000</td>
</tr>
<tr>
<td>Mni-glbh</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>Burr et al., 2000</td>
</tr>
<tr>
<td>Nhr-glb</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>Blaxter et al., 1994a</td>
</tr>
<tr>
<td>Ppa-glb</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>Blaxter et al., 1994a</td>
</tr>
<tr>
<td>Ppa-glb GLBA</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>Dixon et al., 1992</td>
</tr>
<tr>
<td>Ppa-glb GLBB</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>This report</td>
</tr>
<tr>
<td>Str-glbm</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>This report</td>
</tr>
<tr>
<td>Str-glbh</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>This report</td>
</tr>
<tr>
<td>Tca-glbh</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>This report</td>
</tr>
<tr>
<td>Hco-glb cn01320</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>This report</td>
</tr>
<tr>
<td>Hco-glb cn01319</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>This report</td>
</tr>
<tr>
<td>Hco-glb cn01747</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>This report</td>
</tr>
<tr>
<td>Hco-glb cn08501</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>This report</td>
</tr>
<tr>
<td>Hco-glb</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>This report</td>
</tr>
</tbody>
</table>

* "yes" is reported where the intron is present in the gene sequence; "no" is reported where it is not. Two additional intron positions are known for P. pacificus globin which is not shown in the table (see text). ** - introns separating the secretory leader peptide-encoding domain from the globin domain, and/or introns separating adjacent globin domains (GLBP_ASCSU and GLBP_PSEDE). *** - see text regarding Bma-glb intron positions; E9.0 and E8.1 cannot be distinguished given the available information. **** - All four Ppa-glb genes have the same predicted gene structure.
C. elegans has no 3' and 5' introns, retaining the central intron alone (Kloek et al. 1996), whereas in A. suum myoglobin there are no introns (Blaxter et al. 1994b). Both S. trachea globins and four predicted H. contortus globins (Hco-glb cn01320, Hco-glb cn01319, Hco-glb cn01747, Hco-glb cn08501) have an additional intron at position ef4b.0, giving them four introns in total (see methods for description of the modified nomenclature). The four predicted P. pacificus globins have five introns, with additional introns at positions ef4d.0 and H14.2. The ef4d.0 intron in Ppa-glb genes does not appear to be homologous to the ef4b.0 intron from Strongylids. These differing positions are not likely to be artefacts of ambiguous amino acid alignment as theaf interhelix region is quite conserved between these groups (Figure 2). Combined with the observation of introns between secretory leader encoding sequences and the main globin domain in a variety of species, and inter-domain introns in the di-domain globins from A. suum and P. decipiens, there have clearly been multiple examples of intron insertion in nematode globin genes.

Our assigned intron positions for Bma-glbm and Ppa-glb differ from those published previously (Hoogewijs, 2008) because we have revised the gene structure predictions of these genes in the B. malayi and P. pacificus genomes and because of differences in our alignments of P. catadon myoglobin with the nematode globins. Irrespective of the exact intron positions, the P. pacificus genes have additional introns in unique positions for class I nematode globins, further indicating that the late acquisition of introns has occurred in this gene family.

Allelic variation in nematode globins (polymorphism in T. canis globins): Twelve clones of Tca-glbp were generated by RT-PCR from two separate RNA pools, from a total of five animals (or 10 haploid chromosome sets). Gonadal tissue was excluded by dissection. Consistent variability was observed between sequences, and we propose that five alleles were identified of a possible 16 which differ bi-allelically at four nucleotide positions. The variant nucleotides are a silent A to G transition at the 3rd position H2, and two adjacent non-synonymous, C to T transitions at the 1st and 2nd positions of a codon in the C-terminal region of the predicted protein which cannot be aligned with the A. suum D1 structure. These non-synonymous SNPs could cause this codon to encode serine (TGT), leucine (CTT), proline (CCT) or phenylalanine (TTT), depending on the sequence, however only TCT and CTT sequences have been observed in the clones we sequenced.

The gene structures of both Tca-glbm and Tca-glbp have been elucidated. Sequencing of five clones of a region containing intron 4 from Tca-glbp identified two alleles with different length introns (Figure 4B). The polymorphism is complex and may have arisen following either two deletion events in one allele (represented by 2

![Figure 3](image_url)

**FIG. 3.** The number of predicted SNPs from 24 globin EST clusters, illustrating how the relative abundance of genetic variation differs between genes and between species. Transition and transversion SNPs are shown in separate columns for comparison.
FIG. 4. Allelic variation in nematode globins. (A) Clones derived from a genomic DNA fragment of *Hco-glb<sub>cn01319</sub>* or *Hco-glb<sub>cn01320</sub>* were sequenced from either 3 individuals (288, 1025 and 1007 – cn01320) or 2 individuals (1025 and 1007 – cn01319). These were aligned with sequence reads from the *H. contortus* genome sequencing project, and a phylogenetic analysis undertaken. The relationships between clones are illustrated by cladograms with the result of minimum evolutionary distance to the left “Distance” and maximum parsimony to the right “Parsimony”. The results suggest that cn01320 and cn01319 are separate genes as these group separately in the tree. (B) Comparison of two alleles of *Tca-glbp* that differ in the length of intron 4. A number of SNPs can also be seen.
clones) or by two insertion events in the other (3 clones).

In comparison to the EST sequence data, one predicted SNP (TC00537) was in a position distinct from the five SNPs observed by multiple sequencing of *Tca-glbm* cDNAs (data not shown).

**Allelic variation in nematode globins (polymorphism in a H. contortus globin):** Our in silico prediction of SNPs based on EST alignment revealed 43 synonymous and 27 non-synonymous SNPs for *Hco-glb*_cn01320. Three indels causing premature stop codons presumably leading to truncated translation were also found. The non-synonymous SNPs caused either amino acid changes (24) or premature stop codons (3). When aligned with GLBP_ASCSU, for which the 3 dimensional structure has been partially resolved (Yang et al. 1995), the amino acid changes are predicted in a number of helices and inter-helix regions as described in Table 3. When we sequenced a portion of the gene (exons 3, 4 and 5 and a small portion of exon 2) from 3 individuals, we again observed 21 of the 33 synonymous SNPs and 5 of the 7 non-synonymous SNPs that had been predicted from EST analysis. An additional 11 synonymous and 1 non-synonymous SNPs were observed that had not been predicted from EST sequence alignment, and a further 66 SNPs and 2 insertion-deletion mutations were observed over three intron regions. The residues affected by the non-synonymous SNPs were in the ef loop, helix F10, H11 and two in parts of the alignment C-terminal of the H-helix aligned with the GLBP_ASCSU crystal structure (Table 3).

**Gene expression patterns of T. canis globins:** Both *Tca-glbm* and *Tca-glbp* were initially cloned by RT-PCR using RNA derived from whole adult female nematodes. Further analysis of expression using tissues dissected from live adult females and from male RNA indicated that both globins were expressed in both sexes (Figure 5A, B). The tissue distribution of expression differed between isoforms; *Tca-glbm* is expressed in the female reproductive tract and both *Tca-glbp* and *Tca-glbm* are expressed in the body wall of adult females, in males and in the head region of females. No expression of either gene was evident from intestinal tissue.

**Gene expression patterns for nematode globins derived from EST data:** The cDNA libraries from which the publicly

---

**Table 3.** Details of non-synonymous SNPs and coding region indels predicted in *Hco-glb*_cn01320.

<table>
<thead>
<tr>
<th>Gene Region</th>
<th>SNP or indel</th>
<th>AA change</th>
<th>3D position</th>
<th>Present in Re-sequencing</th>
<th>minor allele frequency</th>
<th>number of ESTs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exon0</td>
<td>A/G</td>
<td>Truncated translation</td>
<td>SLP</td>
<td>NA</td>
<td>0.33</td>
<td>6</td>
</tr>
<tr>
<td>Exon0</td>
<td>G/T</td>
<td>Truncated translation</td>
<td>SLP</td>
<td>NA</td>
<td>0.33</td>
<td>6</td>
</tr>
<tr>
<td>Exon0</td>
<td>G/T</td>
<td>L-H</td>
<td>SLP</td>
<td>NA</td>
<td>0.17</td>
<td>12</td>
</tr>
<tr>
<td>Exon0</td>
<td>A/T</td>
<td>L-H</td>
<td>SLP</td>
<td>NA</td>
<td>0.12</td>
<td>17</td>
</tr>
<tr>
<td>Exon0</td>
<td>A/C</td>
<td>I-L</td>
<td>SLP</td>
<td>NA</td>
<td>0.39</td>
<td>18</td>
</tr>
<tr>
<td>Exon0</td>
<td>C/T</td>
<td>L-F</td>
<td>SLP</td>
<td>NA</td>
<td>0.20</td>
<td>20</td>
</tr>
<tr>
<td>Exon0</td>
<td>G/T</td>
<td>V-F</td>
<td>SLP</td>
<td>NA</td>
<td>0.13</td>
<td>24</td>
</tr>
<tr>
<td>Exon0</td>
<td>G/T</td>
<td>F-C</td>
<td>SLP</td>
<td>NA</td>
<td>0.08</td>
<td>25</td>
</tr>
<tr>
<td>Exon0</td>
<td>G/T</td>
<td>V-G</td>
<td>SLP</td>
<td>NA</td>
<td>0.08</td>
<td>25</td>
</tr>
<tr>
<td>Exon0</td>
<td>-/C</td>
<td>Truncated translation</td>
<td>SLP</td>
<td>NA</td>
<td>0.18</td>
<td>22</td>
</tr>
<tr>
<td>Exon0</td>
<td>A/T</td>
<td>F-Y</td>
<td>SLP</td>
<td>NA</td>
<td>0.16</td>
<td>25</td>
</tr>
<tr>
<td>Exon0</td>
<td>G/C</td>
<td>F-L</td>
<td>SLP</td>
<td>NA</td>
<td>0.14</td>
<td>24</td>
</tr>
<tr>
<td>Exon0</td>
<td>C/T</td>
<td>T-I</td>
<td>SLP</td>
<td>NA</td>
<td>0.16</td>
<td>25</td>
</tr>
<tr>
<td>Exon1</td>
<td>-/T</td>
<td>Truncated translation</td>
<td>A1</td>
<td>NA</td>
<td>0.15</td>
<td>26</td>
</tr>
<tr>
<td>Exon1</td>
<td>G/T</td>
<td>D-Y</td>
<td>A2</td>
<td>NA</td>
<td>0.07</td>
<td>27</td>
</tr>
<tr>
<td>Exon1</td>
<td>-/A</td>
<td>Truncated translation</td>
<td>A12</td>
<td>NA</td>
<td>0.11</td>
<td>28</td>
</tr>
<tr>
<td>Exon2</td>
<td>C/T</td>
<td>R-C</td>
<td>C5</td>
<td>NA</td>
<td>0.11</td>
<td>27</td>
</tr>
<tr>
<td>Exon2</td>
<td>G/T</td>
<td>R-L</td>
<td>C5</td>
<td>NA</td>
<td>0.17</td>
<td>29</td>
</tr>
<tr>
<td>Exon2</td>
<td>A/T</td>
<td>R-K</td>
<td>C6</td>
<td>NA</td>
<td>0.14</td>
<td>29</td>
</tr>
<tr>
<td>Exon2</td>
<td>A/T</td>
<td>R-N</td>
<td>C6</td>
<td>NA</td>
<td>0.11</td>
<td>28</td>
</tr>
<tr>
<td>Exon2</td>
<td>C/G</td>
<td>S-R</td>
<td>cd loop</td>
<td>NA</td>
<td>0.07</td>
<td>29</td>
</tr>
<tr>
<td>Exon2</td>
<td>A/G</td>
<td>A-T</td>
<td>D2</td>
<td>NA</td>
<td>0.07</td>
<td>29</td>
</tr>
<tr>
<td>Exon3</td>
<td>A/T</td>
<td>F-Y</td>
<td>cd loop</td>
<td>Yes</td>
<td>0.35</td>
<td>26</td>
</tr>
<tr>
<td>Exon4</td>
<td>G/T</td>
<td>S-A</td>
<td>F10</td>
<td>Yes</td>
<td>0.36</td>
<td>25</td>
</tr>
<tr>
<td>Exon5</td>
<td>C/T</td>
<td>P-S</td>
<td>gh loop</td>
<td>No</td>
<td>0.08</td>
<td>24</td>
</tr>
<tr>
<td>Exon5</td>
<td>A/G</td>
<td>A-T</td>
<td>H6</td>
<td>No</td>
<td>0.13</td>
<td>24</td>
</tr>
<tr>
<td>Exon5</td>
<td>A/G</td>
<td>I-V</td>
<td>H11</td>
<td>Yes</td>
<td>0.13</td>
<td>24</td>
</tr>
<tr>
<td>Exon5</td>
<td>A/G</td>
<td>T-A</td>
<td>post H</td>
<td>Yes</td>
<td>0.26</td>
<td>25</td>
</tr>
<tr>
<td>Exon5</td>
<td>C/G</td>
<td>T-S</td>
<td>post H</td>
<td>Yes</td>
<td>0.09</td>
<td>23</td>
</tr>
</tbody>
</table>

1 - “Exon0” represents the part of this predicted gene encoding the secretory leader peptide (SLP). In other extracellular globins there is an intron separating this from the remainder of the gene, however we have no evidence from genome sequencing for this intron in *Hco-glb*_cn1320.

2 - SLP residues are those predicted to be part of the secretory leader peptide; “post H” refers to residues C-terminal of the H-helix aligned with the GLBP_ASCSU crystal structure, and the other designations refer to positions aligned with the GLBP_ASCSU crystal structure.

3 - NA for this table is for “not attempted”. Re-sequencing was conducted using a partial gene fragment.

4 - The minor allele frequency was calculated by dividing the number of EST sequences containing the rarer allele by the number of EST sequences in the alignment at the position indicated. Sequences containing a residue or gap not represented by the two most common alleles were excluded.
available globin ESTs were derived from various life cycle stages, tissues and other treatments. Supplementary Table 3 shows expression of the nematode globins with details of the cDNA libraries from which they were derived. Some estimates of globin expression in these nematodes and nematode tissues can be obtained by dividing the number of globin clones by the number of sequenced clones from the library. In this way, we predict that strongyloid nematodes and Ascaris spp. express globins to the highest levels in the dataset (Table 4). Plant parasitic nematodes and free-living nematodes have lower levels of expression.

The A. suum dataset includes libraries constructed using cDNA from dissected tissues; and so a comparison with our qPCR analysis of T. canis globins can be made. As with our T. canis results, the intracellular globin is most highly expressed in the hypodermis/body wall muscle and head tissues. The Ascaris library results also show that the L_4 larval stage expresses As-glb_n00004 to a high level. Considering the high abundance of...
protein, it is surprising that the di-domain extracellular globin from *A. suum* (*As-glbp*) is not highly expressed in the female gonad, but rather has a very similar pattern of expression to *As-glbl* (*As-glb* 15 male head SL1 TOPO v1 Murphy Chiapelli McCarter) suggesting that the functions of *As-glbp* and the di-domain extracellular globin from *A. suum* (*As-glbp* and *As-glb*) are divergent. Another *A. suum* cluster, *As-glbl* 00780, appears to encode a single domain globin from *A. suum* that has higher sequence similarity to *Tca-glbl* than *As-glbl* 00019; it is interesting that cDNAs from this cluster are predominantly from libraries obtained from gut and ovarian tissue, which is a pattern dissimilar to either *Tca-glblm* or *Tca-glblp*. *A. suum* also had a predicted second intracellular globin (*As-glbl_17423*), which is expressed at a very low level, the sequenced clones are both from a female gut-derived library.

The majority of cDNA libraries with a high abundance of globin cDNA clones were from strongylid nematodes. There were 8 clusters of ESTs from four strongylid nematode species (*H. contortus, T. circumcincta, O. ostertagi* and *N. brasiliensis*) that had globin clone abundances above 1%. All these libraries were derived from *L* 4 or adult stage nematodes. When compared to *L* 4 or egg derived libraries, it appears that globin expression in strongylids is enhanced during parasitic life stages (Figure 5C). Considering *H. contortus* alone, no globin ESTs were detected in *L* 3 or egg libraries, whilst abundances in an *SL1 trans-spliced L 4 library ranged from 3.7 per 1000 for *Hco-glbl Gn01356* to 44.0 for *Hco-glbl Gn00868*. In two adult libraries abundances ranged from 0.43 per 1000 in one library to 8.4 in a day 11 adult library for *Hco-glbl Gn01320*. Interestingly, *Hco-glbl Gn00868* (extracellular), *Hco-glbl Gn01356* (intracellular) and *Hco-glbl Gn01377* (intracellular) are all highly expressed in *L* 4 and appear to be *L* 4 specific; no other *H. contortus* clusters contained EST sequences from the *L* 4 stage library.

### DISCUSSION

Globins are almost ubiquitous genes found in the genomes of organisms from all four eukaryotic kingdoms, Eubacteria and Archaea (Vinogradov et al. 2006). Nematodes have been known to express globins for some time (Adduco 1889; Blaxter 1993; Frenkel et al. 1992), and they have been demonstrated to have divergent gene structures and functions (Blaxter et al. 1994b; Burr et al. 2000; Kloek et al. 1996; Moens et al. 1992; Sherman et al. 1992; Vinogradov and Moens, 2008). In the work described here we have shown that nematode globin genes display a great deal of sequence and gene structure diversity. There is evidence for both recent and ancient globin gene duplications, especially in the Strongylid and Ascarid groups. The presence of introns in a fourth, “non-standard” position in Strongylid globin genes from *S. rachaea* and *H. contortus*, and of other additional introns in the *P. pacificus* globin genes show that introns have been inserted subsequent to the putative assembly of a proto-globin gene from gene fragments in the earliest eukaryotes (Gilbert et al. 1986). These introns are not associated with additions of new structural domains in contrast to the introns separating secretory leader peptide domains or globin domains of di-domain globins. Clearly some globin introns can be “late” in addition to others being “early”. In *C. elegans* a larger family of globin like genes have been identified (Vinogradov et al. 2006; Hoogewijs et al. 2007). The globins discussed above are most similar to the *C. elegans* globin ZK637.15 defined as Class I globins in Hoogewijs, et al., 2008. The gene structures exhibited by the other Class II globins and globin-like genes are also divergent from the conserved pattern observed in globins from Chordates and other groups (Hoogewijs, et al., 2008).

The function of globins in invertebrate animals remains enigmatic (Vinogradov and Moens, 2008).
Although there have been some illuminating investigations (Burr et al. 2000; Kimura et al. 1999; Minning et al. 1999), these suggest many divergent functions rather than a single, global purpose for these proteins.

Within the nematode taxa considered here, we have shown that the pattern of expression of globins follows some consistent patterns. Expression in parasitic species seems to be increased in adult and pre-adult stages, though these do not necessarily correspond to parasitic stages, as in M. nigrescens (Burr et al. 2000). This pattern is not conserved in C. elegans however where Ce-glb-1 is expressed at roughly equivalent levels in the L3, adult and alternate L3 dauer stage, and is also expressed in eggs, L1 and L2 (Hoogewijs, et al., 2007). It is also of interest that Ce-glb-1 gene expression does not respond to hypoxia as other globins do in plants and other invertebrate animals (e.g. Kimura, et al., 1999; Hunt, et al., 2001), but is instead responsive to perturbations in insulin signalling (Hoogewijs, et al., 2007). The parasitic nematode gene expression patterns may indicate divergence from Ce-glb-1 if the main stimulus for increased expression is the movement from normoxic to hypoxic environments, however there are a whole range of factors encountered by parasitic larvae on their entrance to the host, so changes in insulin signalling could well be occurring and influencing gene expression at this stage. This is an area of investigation worth pursuing in future work.

Expression of globins in the hypodermis also seems to be a common observation for Class I nematode globins. In contrast, the Class II globins described are expressed predominantly in neurons (Hoogewijs, et al., 2008), including one from C. elegans (glb-5) which has been shown to be important for the avoidance of high oxygen concentrations (Persson, et al., 2009; McGrath, et al., 2009).

Other aspects of the investigation we have undertaken support the hypothesis of a diversity of functions in nematode globins. First we note that many nematode globins have putative secretory leader peptides and therefore have roles outside the cell rather than internally whilst other globins are clearly located intracellularly. Second, there is significant sequence divergence in the genes we have identified and many globins have divergent residues at the E7 ligand coordinating position. Glutamine and leucine are the most common E7 residues observed in our alignment (Figure 2), and these residues have both divergent side-chain properties and have been shown to impart divergent ligand binding kinetics to globins when substituted for one another (Hargrove et al. 1996). Third, we have observed a high degree of genetic diversity between globins from different species and between multiple globin genes within species. This variation is highly suggestive of actively evolving proteins being selected for divergent function. It seems very likely that the further investigation of these proteins will reveal a diversity of both biochemical and cellular functions which may be useful for biotechnological applications or as targets for the control of parasites through immunological or pharmaceutical means.

The impressive degree of gene diversity we have observed is almost matched by the level of within-gene allelic diversity. The abundance of SNP in the H. contortus cn01320 gene (157 SNP per 1000 bp in exons and UTRs, and far higher in intron sequence) far exceeds that observed in a large segment of the C. elegans genome (1.1 SNP per 1000 bp (Wicks et al. 2001)). The abundance of putative SNP in most globin genes we analysed using in silico methods was also high, except for the C. elegans globin gene (ZK637.13) because the ESTs we used were all obtained from cDNA originating from the laboratory strain N2. Estimates of nucleotide diversity in coding regions of nematode genes have been attempted for a variety of nematodes, and these range from 5.8/1000 for wild C. elegans (Cutter et al. 2006) to 68/1000 for both the obligate out-crossing C. remanei (Cutter et al. 2006) and 20/1000 for the mean in silico predicted SNP from 1,548 EST clusters in our H. contortus database WormSIS (data not shown). Therefore the available evidence suggests allelic diversity in some globin-encoding genes is higher than that exhibited by other nematode genes.

This observed allelic diversity has two important consequences. Firstly for evolutionary studies, these genes are not appropriate for studies of nematode taxon relationships. The rate of evolution is far too high for meaningful comparisons. We have also tried to use Hco-glb.cn01320 for population-level studies with little success, mostly because of the high level of diversity within isolates (data not presented). Secondly, the use of globins as vaccine or drug targets could be problematic. Although use of these genes as a vaccine target has met with limited success allelic divergence could have affected the consistency of results (Frenkel et al. 1992; Claerebout, et al., 2005; S. McClure and D. Emery, personal communication). A second consideration for this application would be the observation of multiple genes in economically important species, perhaps necessitating the use of multivalent approaches.

In conclusion, we have demonstrated that nematode globins show a huge amount of diversity in sequence and gene structure, though the timing and tissue specificity of expression may be more highly conserved. There is evidence for multiple gene duplications, multiple intron insertions and losses and for allelic variation at both synonymous and non-synonymous sites. The gene family shows great promise for discovering both unique insights into both globin structure-function relationships and for cellular roles in an important animal phylum.
**LITERATURE CITED**


of larval and adult *Ostertagia ostertagi* by immunoscreening of cDNA libraries. Molecular and Biochemical Parasitology 126:201–208.


