International Journal for Parasitology 41 (2011) 881–894

Contents lists available at ScienceDirect

International Journal for Parasitology

journal homepage: www.elsevier.com/locate/ijpara

NEMBASE4: The nematode transcriptome resource

Benjamin Elsworth [a], James Wasmuth [b], Mark Blaxter [a,*]

[a] Institute of Evolutionary Biology, The University of Edinburgh, EH9 3JT, UK
[b] Department of Comparative Biology and Experimental Medicine, Faculty of Veterinary Medicine, 3330, Hospital Drive, University of Calgary, Calgary, Alberta, Canada T2N 4N1

ARTICLE INFO

Article history:
Received 21 December 2010
Received in revised form 11 March 2011
Accepted 14 March 2011
Available online 21 April 2011

Keywords:
Nematode
Transcriptome
Genome
Expressed sequence tag
Database
PartiGene

ABSTRACT

Nematode parasites are of major importance in human health and agriculture, and free-living species deliver essential ecosystem services. The genomics revolution has resulted in the production of many datasets of expressed sequence tags (ESTs) from a phylogenetically wide range of nematode species, but these are not easily compared. NEMBASE4 presents a single portal into extensively functionally annotated, EST-derived transcriptomes from over 60 species of nematodes, including plant and animal parasites and free-living taxa. Using the PartiGene suite of tools, we have assembled the publicly available ESTs for each species into a high-quality set of putative transcripts. These transcripts have been translated to produce a protein sequence resource and each is annotated with functional information derived from comparison with well-studied nematode species such as *Caenorhabditis elegans* and other non-nematode resources. By cross-comparing the sequences within NEMBASE4, we have also generated a protein family assignment for each translation. The data are presented in an openly accessible, interactive database. To demonstrate the utility of NEMBASE4, we have used the database to examine the uniqueness of the transcriptomes of major clades of parasitic nematodes, identifying lineage-restricted genes that may underpin particular parasitic phenotypes, possible viral pathogens of nematodes, and nematode-unique protein families that may be developed as drug targets.

1. Introduction

Nematode genomics has thrived in the decade following the sequencing of the complete genome of the free-living rhabditid *Caenorhabditis elegans* in 1998 (The *C. elegans* Genome Sequencing Consortium, 1998). The genome sequences of two additional free-living species (*Caenorhabditis briggsae* (Stein et al., 2003) and *Pristionchus pacificus* (Dieterich et al., 2008)) and four parasitic species (*Brugia malayi* (Ghedin et al., 2007), *Meloidogyne incognita* (Abad et al., 2008), *Meloidogyne hapla* (Opperman et al., 2008) and *Trichinella spiralis* (Mitreva et al., 2011) have since been published, and many additional nematode genomes are 'in progress' (see http://www.nematodegenomes.org/). These genome sequences have assisted in defining the genetic toolkit that underpins nematode biology and, in the case of *C. elegans*, also fostered forward and reverse genetic investigations of basic biological processes such as ageing and embryogenesis. The complete sequencing of animal-parasitic (*B. malayi* and *T. spiralis*) and plant-parasitic (*M. incognita* and *M. hapla*) nematode genomes was undertaken in order to identify the particular genetic adaptations these species have made to the parasitic mode of life, and thus better inform efforts to control or eradicate the diseases they cause. However, identification of the key genes and genetic processes that permit a parasitic mode of life is difficult with so few genomes available for comparison (Blaxter, 2003).

To overcome this constraint of limited diversity of whole genome sequences and due to the experimental complexity and cost of the generation of a whole genome sequence for a target species, many research programmes have instead used the expressed sequence tag (EST) approach (Wasmuth et al., 2008). ESTs are single-pass reads derived from cDNA representing the expressed genes of an organism (or tissue or cell type). Surprisingly, despite having haploid genomes of up to 10 gigabases (Gb) or more, the transcriptome represented in mature mRNAs of most Metazoa is derived from only 20 to 50 megabases (Mb) of the genome. The EST approach samples only this subset of the genome and avoids complex bioinformatic issues of gene prediction (identification of coding exons amongst the 75–99% of non-coding DNA). From a non-normalised cDNA resource the frequency at which a particular gene transcript is sampled also reports on its steady-state mRNA concentration and thus on the level of expression of the gene. Therefore, for a small investment a research programme can generate a sequence dataset that represents many of the expressed genes of the target organism and a first estimate of the pattern of expression of these genes.

* Corresponding author. Address: Institute of Evolutionary Biology, Ashworth Laboratories, King's Buildings, The University of Edinburgh, Edinburgh EH9 3JT, UK. Fax: +44 131 650 7489.
E-mail address: mark.blaxter@ed.ac.uk (M. Blaxter).

0020-7519 © 2011 Australian Society for Parasitology Inc. Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.ijpara.2011.03.009

The EST approach has its limitations. Because cDNA fragments are selected at random for sequencing, an EST dataset cannot be guaranteed to identify all of the 15,000–25,000 transcription units present in a genome. Indeed, for transcripts expressed at low levels (e.g. one transcript per cell), the number of randomly selected ESTs required to guarantee their identification is very large. Also, because some genes are expressed only in very restricted circumstances such as in early development or in response to particular environmental challenges, an EST approach would have to sample many cDNA preparations from different life stages, tissues and environments to identify these conditionally expressed genes. Analysis of ESTs can also be problematic. Because ESTs are single-pass DNA sequencing reads, they may contain errors. Assembly of the individual ESTs into putative transcripts or 'unigenes' requires careful attention to the kinds of errors possible, and downstream functional annotation of these assembled ESTs must also consider residual errors (Parkinson et al., 2004a; Parkinson and Blaxter, 2004; Wasmuth and Blaxter, 2004).

More than one million ESTs for over 60 species have been generated (Supplementary Table S1). Most of these EST datasets have been generated and analysed individually, using a range of tools and analytical parameters. These individual analyses are often tours de force of extraction of maximal biological information and insight from limited resources (Blaxter et al., 1996; Daub et al., 2000; McCarter et al., 2003; Harcus et al., 2004; Mitreva et al., 2004a,b, 2005), and have played significant roles in promoting modern molecular genetic research on nematode parasites in particular. However, the large datasets now available for nematodes are a rich substrate for data mining across the diversity of the phylum. By comparing across species, we can identify genes putatively unique to a species or species group and associate these with features of the species' biology or pathogenesis. Comparative analyses of assembled EST datasets across species also permit identification of biochemical or regulatory pathways uniformly present or absent in groups of species and thus estimation of the physiology of the nematodes and their likely sensitivity or resistance to particular drugs. The partial nature of EST collections means conclusions concerning the absence of genes or pathways must remain conditional (absence of evidence is not the same as evidence of absence), but cross species correlation of patterns of presence/absence can lend support to hypotheses of loss.

Previously we (Parkinson et al., 2004b,c; Wasmuth et al., 2008) and others (Martin et al., 2009) have compared a limited number of species' EST datasets and thereby identified novel families of parasite-specific genes and biochemical pathways with the potential for drug disruption. These data have been made openly accessible to researchers through web portals into the NEMBASE3 (Wasmuth et al., 2009) and Nematode.net (Martin et al., 2009) databases. The technologies of DNA sequencing are now undergoing a further revolution with the introduction of ultra-high throughput instruments that generate data at a very small fraction of the cost of traditional Sanger capillary EST sequencing. This revolution has been rapidly exploited by nematode genome researchers, and the coming years will see a flood of ultra-deep transcriptome sequencing and whole genome sequencing from nematodes. Here we present NEMBASE4, an analysis of the current Sanger sequencing-derived EST data. We have updated the core NEMBASE3 with all current publicly available EST datasets and a set of previously unpublished datasets. NEMBASE4 includes nearly 700,000 ESTs and 240,000 putative transcripts. Proteins derived from fully sequenced nematode genomes are also included for comparative purposes. A streamlined interface and updated functional analyses facilitate data mining and identification of new research targets.

2. Materials and methods

2.1. Programs and databases

We used updated versions of the PartiGene suite of programs to assemble and annotate these data. PartiGene (Parkinson et al., 2004a) version 3.0.6 (M. Blaxter and R. Schmid, unpublished data) was used for sequence clustering and databasing. prot4EST (Wasmuth and Blaxter, 2004) version 3 (Wasmuth, unpublished data) was used for derivation of peptide translations from consensus sequences. annot8r (Schmid and Blaxter, 2008) version 3.1 was used for Gene Ontology (GO), Enzyme Commission (EC) and Kyoto Encyclopaedia of Genes and Genomes (KEGG) annotation. The updated versions of these scripts are available from http://www.nematodes.org/. Sequence similarity comparisons were performed using BLAST version 2.18 (Altschul et al., 1997) and the NCBI non-redundant protein database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/; July 2009) and the EBI UniProt database (http://www.ebi.ac.uk/uniprot/; UniRef100; July 2009). *C. elegans* (version WS172), *C. briggsae* (version WS172) and *B. malayi* (version 1) protein datasets and annotations were downloaded from WormBase (Harris et al., 2010). Identification of protein domains was achieved using InterProScan (Zdobnov and Apweiler, 2001) and the InterPro database. Protein tribes were inferred using TRIBE-MCL (Enright et al., 2002). The web interface was built in Hypertext Mark-up Language (html) and PHP:Hypertext Preprocessor (PHP) language using the Postgres database management tool, Apache server and custom PHP and Common Gateway Interference (CGI) scripts (see http://www.nematodes.org/NEMBASE4/).

2.2. Nematode EST sequence data

Core data were taken from the NEMBASE3 database (Wasmuth et al., 2008). New nematode EST sequence data were downloaded from EMBL/GenBank/DDBJ in January 2009 (Supplementary Table S1) using custom Perl scripts. For each nematode species in the public nucleotide sequence databases, the number of EST sequences was ascertained and all species with more than 15 sequences were selected for analysis (Supplementary Table S1). Each species' ESTs were filtered for length (sequences <100 bases were discarded) and for quality (eliminating sequences with biologically unfeasible sequence patterns that more likely resulted from Sanger sequencing technology errors, such as long runs of alternating polynucleotides). For species already present in NEMBASE3, only ESTs submitted since the last update of that database were added.

2.3. EST clustering

ESTs were clustered using CLOBB (Parkinson et al., 2002) within the PartiGene package (Parkinson et al., 2004a). CLOBB yields unique identifiers for each cluster and as these identifiers are maintained between updates, for species in NEMBASE3 the existing set of cluster identifiers was retained and added to. In the PartiGene schema, each derived consensus sequence has a two letter species identifier, followed by C for nucleotide consensus (replaced by P for the derived peptide sequence), and a unique five-digit number. As each cluster can result in more than one consensus (in the case of alternative splicing, for example), the resultant consensuses are indicated by an underscore and a number following the five-digit identifier.

2.4. Derivation of protein translations

We translated the cluster consensus sequences using the error-correcting routines built into prot4EST (Wasmuth and Blaxter,

2004). Briefly, prot4EST uses a tiered application of BLAST similarity matches, identification of coding phase by use of optimised models of codon use and identification of longest open reading frames (ORFs) to both correct frameshifts and other simple substitution errors and derive a best-estimate translation of error-prone EST consensuses. We call this peptide resource NEMPEP4.

2.5. Transcriptome annotation

Consensus and protein sequences derived from each set of clustered ESTs were annotated using BLAST searches of nematode and other databases, and decoration with protein domain, GO, EC classifiers and KEGG pathway functional information using InterProScan (Zdobnov and Apweiler, 2001) and annot8r (Schmid and Blaxter, 2008). For each species, the cDNA libraries from which the ESTs were derived were also identified and these data added to the PartiGene database.

2.6. Protein tribes

We added the complete proteomes of *C. elegans*, *C. briggsae*, *Caenorhabditis remanei*, *Caenorhabditis brenneri*, *Caenorhabditis japonica*, *P. pacificus* and *B. malayi* to NEMPEP4, deleting all consensus-derived peptides that were equivalent to entries in the whole proteome data. Some EST-consensus derived peptides were retained, particularly for *B. malayi*, because the genome annotation for species is not always complete. Protein tribes were inferred using all-against-all BLAST comparisons between all of the derived protein translations and genome-derived proteomes parsed by TRIBE-MCL. We used nine inflation values within TRIBE-MCL to identify proteins that formed highly similar (inflation value 5) to less similar (inflation value 1.1) groups. Protein tribe membership was mapped across a phylogenetic tree of the species analysed and tribes limited to particular clades were identified. Peptide sequences were also compared using BLAST (Johnson et al., 2008), with a version of the NCBI protein database that excluded all nematode proteins, to identify any that had significant similarity to non-nematode proteins. These similarities were used to identify tribes that also had members outside the Nematoda.

2.7. User interface to NEMBASE4

NEMBASE4 data were stored in a relational (Structured Query Language; SQL) database using the Postgres database management system. The web interface to NEMBASE4 was written in PHP and CGI. These scripts facilitate and automate the formulation of queries against the underlying Postgres database and format results for browsing across the internet. For pathway analysis we made use of the tools of the KEGG database, linked through EC number annotations of NEMPEP4 protein entities and KEGG Application Programming Interface (API) scripts.

2.8. Example analyses

2.8.1. Sequence alignment

Sequences downloaded from NEMBASE4 were aligned by eye. Similar sequences identified by BLAST searches using the NCBI interface were aligned using their COBALT tool (Papadopoulos and Agarwala, 2007) or ClustalX (Thompson et al., 1997). The alignments used are available as Supplementary data S1, S2 and S3. WebLogos showing conservation of residues across the aligned sequences were developed using the University of Berkley, USA, WebLogo service (http://weblogo.berkeley.edu/) (Crooks et al., 2004).

2.8.2. Phylogenetic analyses

Phylogenies were estimated from the alignments using MrBayes version 3.1.2 (Ronquist and Huelsenbeck, 2003) (with parameters "prset aamodelpr = mixed; mcmc printfreq = 1000 samplefreq = 100 nchains = 4 savebrlens = yes;"). For the nanos analysis, 10,000,000 generations were analysed, for the RDRP analyses, 5,000,000 generations were used, and for the HemH/FC analyses, 1,000,000 generations were used. Each analysis was checked using Tracer 1.4 (http://tree.bio.ed.ac.uk/software/tracer/), and the last ~60% of generations after stationarity used for estimation of the consensus tree and posterior probability support for nodes. Trees were visualised using FigTree 1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/).

3. Results

3.1. Summary

The number and diversity of nematode EST datasets has continued to rise since our last compendium published in 2008 (Wasmuth et al., 2008). We have assembled 679,480 raw ESTs from 62 species (Fig. 1) into 233,295 clusters (Supplementary Table S1). Individual species have from 17 ESTs and 17 clusters (*Globodera mexicana*) to 78,935 ESTs and 25,911 clusters (*Ancylostoma caninum*), and an average of 10,959 ESTs and 3,763 clusters (Fig. 2). Due to the partial nature of ESTs and the likely heterozygosity present in the populations of nematodes sampled for sequencing, some clusters that we were unable to assemble may have been derived from the same transcription unit, inflating the estimated number of distinct gene objects. The magnitude of these effects is unknown, but it has been estimated to be in the region of 10% over-estimation (Wylie et al., 2004). Nevertheless, these data represent a major portion of the expected 15,000–22,000 protein coding genes expected (The *C. elegans* Genome Sequencing Consortium, 1998; Blaxter et al., 2004; Ghedin et al., 2007; Abad et al., 2008) from the best-sampled of these nematode species.

We have extensively annotated these EST clusters. They were first translated to protein sequence using prot4EST, a tool that uses available evidence to identify the most biologically likely ORF and, where possible, correct sequencing error. Using prot4EST, 99.4% of the 237,181 clusters yielded a translation (Fig. 2). The EST clusters were then annotated using annot8r, yielding a total of 378,557 GO annotations, 35,753 EC annotations and 97,148 KEGG pathway annotations. Overall 38.2% of the clusters (38.4% of those with protein translations) had GO, EC or KEGG annotations. In total, 107,209 clusters (45.2%) were decorated with 318,376 protein domain annotations.

We also performed extensive BLAST searches against custom databases to add 944,803 similarity annotations to the data. Notably 22,239 clusters had BLAST similarity matches but were not annotated with domain, GO, EC or KEGG annotations. A collection of protein tribes was built using TRIBE-MCL to cluster translated protein sequences from the EST clusters and the genome-sequencing derived proteomes of *C. elegans*, *C. briggsae*, and *B. malayi* (a total of 377,839 proteins) based on BLAST similarity data. We extracted tribes using a range of inflation values, generating tribes with high between-sequence similarity (inflation value 5) to lower between-sequence similarity (inflation value 1.1). There is no one best inflation value that captures all protein families, thus reporting the results across this biologically relevant span facilitates identification of even distant relationships and rapidly evolving families, as well as highly conserved, slowly evolving ones. At an inflation value of 2.5, we defined 130,892 tribes, 17.8% of which had more than one sequence and 15.5% of which contained sequences from more than one species (Figs. 1 and 3). Compared

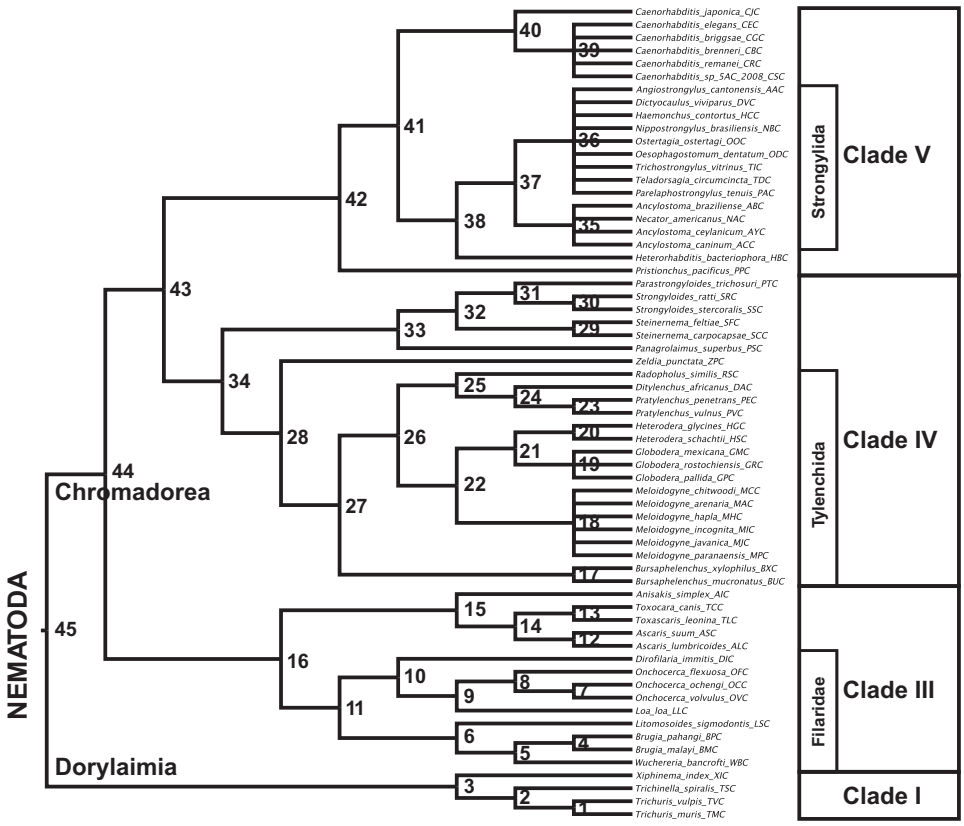

Fig. 1. Phylogenetic relationships of the nematode species analysed. The phylogenetic relationships of the taxa studied were estimated from analyses using nuclear ssrRNA sequences (Blaxter et al., 1998; Meldal et al., 2007); Clades are indicated as defined by Blaxter et al. (1998) and groups analysed further in the text are indicated. Nuclear small subunit ribosomal data do not robustly resolve the interrelationships of the Strongylida and *Meloidogyne*, and these groups are left unresolved. The three-letter identifier allocated to each species is shown after the species name. Internal nodes are numbered for ease of reference as they do not necessarily correspond to named taxa. Analyses of tribes from nodes 18 to 27 are analysed in Supplementary Table S1, and from node 37 in Table 1.

with our previous analyses of 37 nematode species in NEMBASE3, we identified 7,249 additional multi-species tribes and converted 14,344 singletons or single-species tribes into multi-species tribes. Thus including the 25 additional species and additional data for existing species has significantly improved our estimates of shared and private protein sequence diversity in the Nematoda. As predicted from NEMBASE3 (Wasmuth et al., 2008), the dimensions of nematode protein sequence diversity space are very large and still only poorly estimated even with over 250,000 protein sequences.

3.1.1. The NEMBASE4 interface: mining the nematode transcriptomes

NEMBASE4 is stored in a Postgres relational database and is thus amenable to complex querying through use of the SQL language. For custom data mining, we make an SQL file of the database openly available through the project website (http://www.nematodes.org/downloads/databases/NEMBASE4/). We also make NEMBASE4 accessible through a web interface at http://www.nematodes.org/NEMBASE4. Using PHP scripting, we assist the user in constructing and executing selected queries over the database (Fig. 4).

The Welcome Page summarises the data available and presents the user with a set of options for searching and browsing. The user can review all of the data in the database by species and then retrieve all of the clusters for a single species. Alternatively the user can retrieve single clusters through their unique name or the accession number of one of their constituent sequences. The user can search the annotation data that decorates each cluster by searching the definition lines of BLAST-match hits or the GO, EC, KEGG and InterPro domain annotations. Each of these searches can be limited to custom selections of species using a systematic tree-based, interactive menu. Each EST sequence is derived from a particular library and we offer the ability to search by EST library sets, grouped by the lifecycle stage or tissue from which they derive, and by relative abundance of ESTs per library. This abundance search facilitates identification of clusters under- or over-expressed in particular stages or tissues. These searches each result in a list of clusters that match the selected query criteria.

To search the tribe information, we offer a phylogenetic tree-based interface that permits selection of tribes restricted to different subclades of the Nematoda, and at different inflation values in TRIBE-MCL. Individual tribes can be viewed on a Tribe page, revealing a list of the member sequences.

From lists of clusters, the user can navigate to a Cluster page that summarises key data about each cluster, including the number of constituent sequences, the libraries from which they derive, the consensus sequence(s), a map of the alignment of each sequence to the consensus and a selection of annotation data. The Cluster page is linked to the Protein page, which reports on the protein sequence derived and annotations that have been made on that sequence. Additionally the Protein page lists the tribes of which the protein sequence is a member and shows links to the most closely related sequences in other nematodes in the NEMBASE4 dataset.

The Pathways page permits the selection of pathways from the KEGG-defined set and target species based on a minimum BLAST E-value match and then requests from the KEGG servers a pathway diagram decorated with indications of which pathway components are represented in the chosen data.

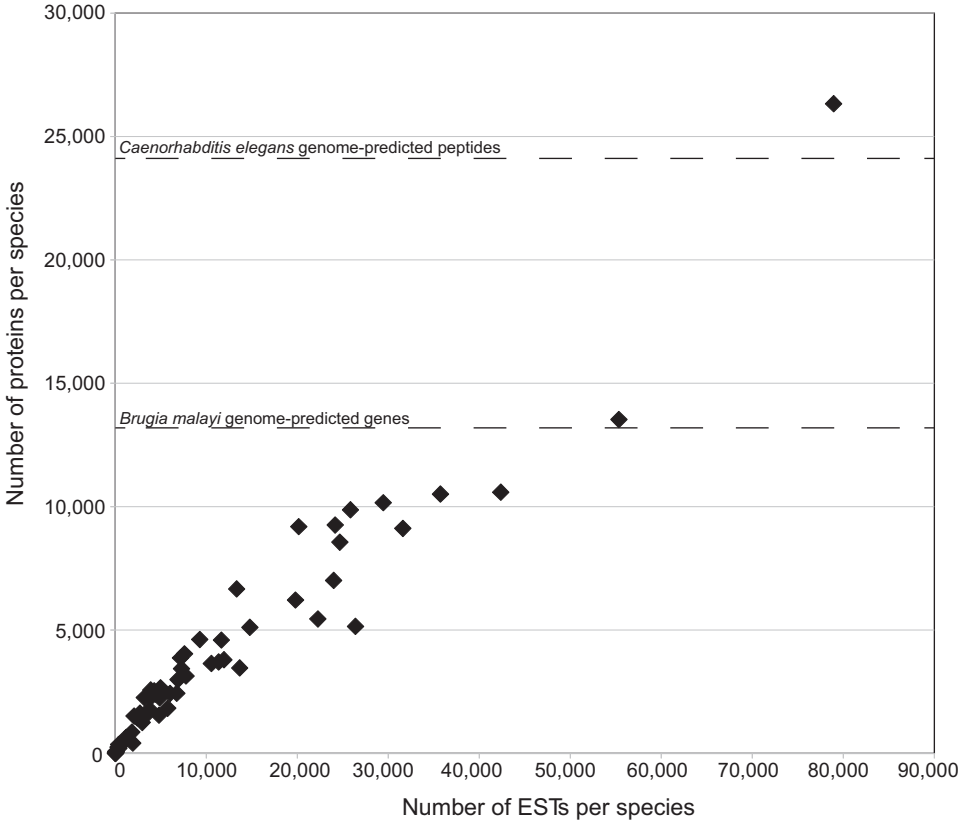

Fig. 2. Expressed sequence tags (ESTs) and clusters for 62 species of Nematoda. Increasing numbers of ESTs (*x*-axis) per species increases the number of cluster-derived protein translations inferred (*y*-axis). Most species have fewer clusters than the numbers of genes found in completely sequenced nematode genomes (the numbers of peptides derived from the *Caenorhabditis elegans* and *Brugia malayi* genomes are indicated), except *Ancylostoma caninum* (78,935 ESTs and 26,330 peptides) where intraspecific variation and fragmentation of transcripts across clusters may have elevated the numbers of clusters inferred.

Users can also search the database for expression patterns of clusters using the Lifecycle Stage and Gender search pages. As each EST is derived from a specific library, if the library has a lifecycle stage and sex attributed to it, the gene represented by the cluster is expressed in that stage. The interface allows users to place numerical cutoffs on the numbers of ESTs per stage to facilitate identification of stage-biased rather than simply stage-specifically expressed genes.

We also allow users to search the nucleotide consensus and protein sequence data using the BLAST algorithms and a query sequence of their choice. The NEMBASE4 database and sequence data files derived from it are available for download.

3.2. NEMBASE4 in action

To illustrate how the NEMBASE4 database might be employed in hypothesis generation and testing, we here present use-cases where the web interface and other online and freely available tools have been used to investigate questions of interest to nematode parasitologists. In each case we defined a hypothesis or open question that might be part of a wider research programme and one researcher spent one day per question exploring NEMBASE4 to address and answer these questions.

3.2.1. Cataloguing genes with signatures of horizontal gene transfer (HGT) in Tylenchina

In analysis of the host–parasite interface between sedentary phytoparasitic tylenchines and their hosts, a series of plant cell-wall degrading and modifying enzymes that are secreted by the nematodes have been identified. Intriguingly, some of these enzymes have the hallmarks of horizontal gene transfer into the nematode genome, as their closest homologues are not in other nematodes or even in other Metazoa but in plants, and plant fungal and bacterial symbionts and pathogens. It is thus hypothesised that the plant parasitic nematodes have acquired these genes from other organisms in their local environments due to the adaptive advantage they offer. These candidate horizontally-transferred genes are fully integrated into the nematode genome and have acquired spliceosomal introns (Blaxter, 2007). Previous surveys have identified sets of candidate horizontally transferred genes in tylenchine nematodes and suggest that these acquisitions occurred in a remote ancestor of extant species (Scholl et al., 2003; Ledger et al., 2006; Mitreva et al., 2009). The enzymes encoded by these genes are good targets for nematicides as they are distinct from those of the plant hosts and of humans and other animals in the food chain.

We therefore posed the question: Which tylenchine genes other than cellulases have signatures of horizontal gene transfer?

Putative horizontal transfer events can be highlighted in NEMBASE4 by identifying protein tribes from the group of interest (e.g. plant parasitic tylenchine nematodes) that have no counterparts in other nematodes (i.e. the tribes are restricted to tylenchines), but do have significant BLAST similarity matches to non-nematode taxa. There are 55 such tribes in NEMBASE4 within the Tylenchina (Supplementary Table S2). The non-nematode species matched include Metazoa, Protozoa, Fungi, Bacteria, Viridlantae and viruses. The matches to Metazoa were on average poorer (mean negative exponent of *E*-value 10.2, S.D. 5.9) than matches to Viridiplantae (mean 21.0, S.D. 21.8), Bacteria (mean 41.1, S.D. 25.2), Fungi (mean 71.5, S.D. 85.6) and Protozoa (mean 19). The single match to

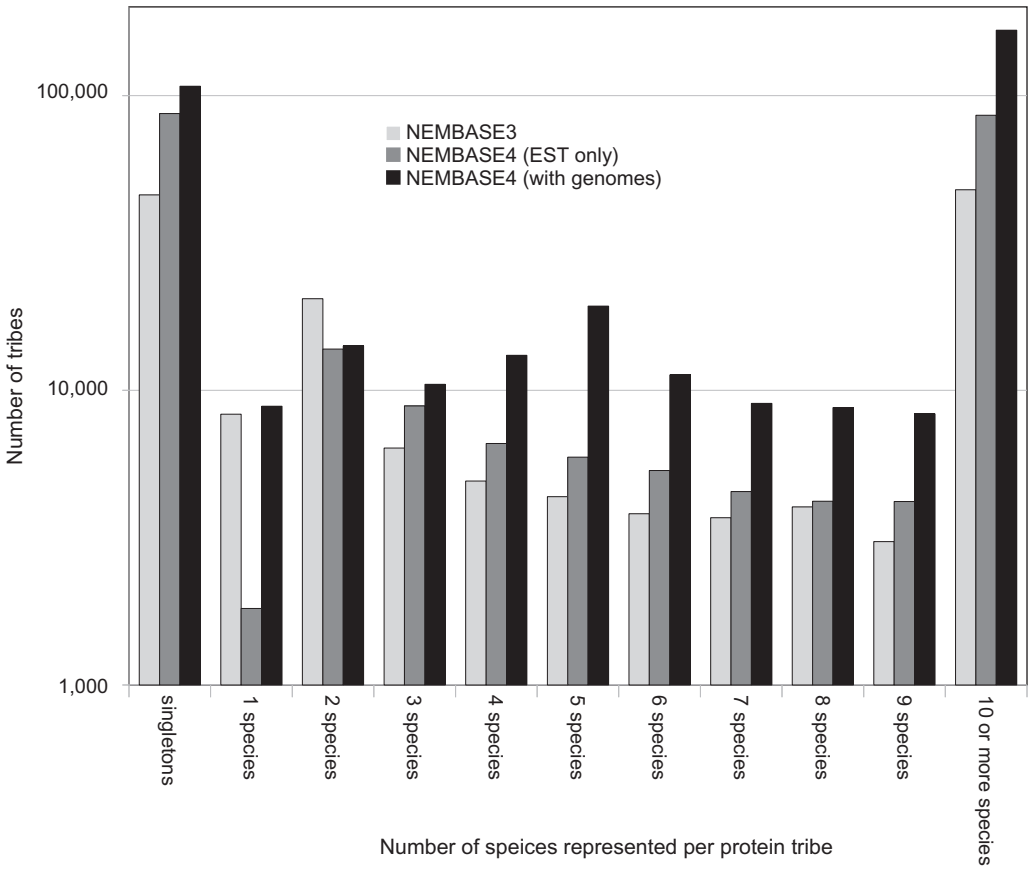

Fig. 3. Protein tribes. The tribes defined in NEMBASE4 using TRIBE-MCL (Enright et al., 2002) and an inflation value of 2.5 include over 120,000 that have proteins from 10 or more of the 62 species analysed. The inclusion of additional species to NEMBASE4 compared with NEMBASE3 resulted in many more protein tribes that include members from more than one species. The inclusion of the genome-sequence derived proteomes (of *Caenorhabditis* spp. and others) results in a peak of protein tribes with five species contributing members. EST: expressed sequence tag.

viruses had an *E*-value exponent of −35. The metazoan matches are mostly to unnamed protein products defined by genome projects and are suggestive of weak, but significant, similarities to deeply conserved protein domains. The highly-significant matches to plant, fungal and bacterial matches include similarities to proteins with roles in cellulose and other cell wall degradation, and chorismate biochemistry, as expected from published surveys (Scholl et al., 2003). Additional enzymes putatively involved in thiamine synthesis and lipid metabolism, and several tribes that have high-scoring matches to plant and bacterial proteins of no known function, are of obvious interest: are these mediators of additional nematode–plant interactions? This list of tribes will complement efforts to understand and model the acquisition of genes by lateral transfer in plant parasitic nematodes (Mitreva et al., 2009).

This search for HGT candidates also illustrated the wealth of discoveries still to be made in these EST datasets. One tribe (inf11-10477) had closest matches to virally-derived proteins. Could its members derive from nematode viruses? Viruses have been conspicuous by their absence from the roster of nematode pathogens, with a single instance recently reported (Felix et al., 2011).

The two members of tribe inf11-10477 (*Heterodera schachtii* HSC00105 and *Globodera pallida* GPC02272) encode proteins highly similar to viral RNA-directed RNA polymerases (RDRP) from Picornavirales, single-stranded, positive strand RNA viruses with no DNA stage. Picornavirales include pathogens of wasps and other arthropods, vertebrates and plants. The bee virus (Cox-Foster et al., 2007) to which the nematode sequences are most similar is a member of Dicistroviridae. We identified the 50 most-similar RDRP proteins in GenBank, selected individual representatives of each major virus species (from the Dicistroviridae (arthropods), Iflaviridae (arthropods), and Secoviridae (plants)), and aligned them (Supplementary Data S1). Phylogenetic analysis of this alignment shows that the nematode RDRP sequences form a clade distinct from other Picornaviridae (Fig. 5). The sequences do not obviously derive from a host plant virus, as viruses from related hosts (beans and peas for *H. scachtii*'s host, soybean, and tomato for *G. pallida*'s host, potato) are quite distinct. These plant parasitic nematodes do not have arthropod vectors and there is no closely related arthropod-derived sequence. The cDNA libraries from which the sequences were derived were constructed on different continents by different teams, so laboratory contamination seems unlikely. We therefore conclude that these sequences are the first evidence of a virus naturally infecting tylenchine nematodes. This has exciting prospects for development of control measures for these devastating parasites.

3.2.2. Are there conserved genes underpinning parasitism in Strongylida?

Ancestors of tylenchine nematodes acquired genes from their environment that are likely to promote their survival as plant parasites. The analysis of taxon-restricted tribes can also reveal genes that underpin the unique biology of other clades of nematode. The Strongylida are a monophyletic clade of vertebrate-parasitic species within Clade V (Fig. 1). They have radiated rapidly to parasitise most land and many marine vertebrates. The genetic tricks that underpin this successful radiation might include interference with or evasion of host immune recognition or effector systems, or

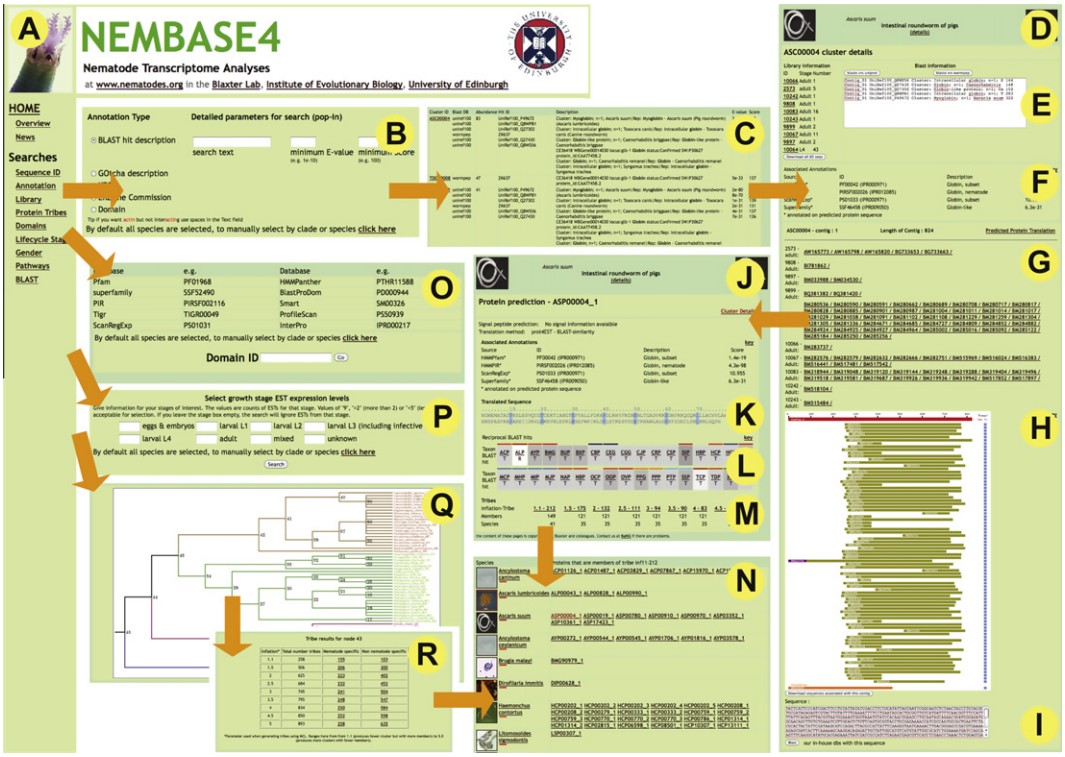

Fig. 4. The web interface to NEMBASE4 at http://www.nematodes.org/NEMBASE4. The cartoon illustrates the structure of the web interface to NEMBASE4, showing the different paths available to the user for querying and browsing the underlying data. (A) The Home page and other subsequent pages all carry a left-side strap of links to analysis pages. (B) Selecting "Annotation" brings up a page that facilitates text- and other metric-based searches of NEMBASE4. The radio buttons allow a choice amongst searching the BLAST search match definition lines, Gene Ontology (GO) descriptions, Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway names, Enzyme Commission (EC) codes and domain names. All of these searches can be limited by species or a group of species using a drop-down interactive tree (as can many other searches of NEMBASE4). (C) The results of searching the BLAST annotation with the term 'globin'. Matching cluster identifiers are shown together with description lines. (D) Selecting one cluster identifier brings up the Cluster page, where the user can browse. (E) The BLAST matches for the cluster sequences in UniRef100 and to *Caenorhabditis elegans*, and to the representation of different libraries in the Expressed sequence tags (ESTs) making up this cluster. (F) Any functional annotations available for the cluster are displayed. (G) The individual ESTs making up the cluster are hyperlinked to the EMBL database. (H) An illustration of the alignment of the ESTs with reference to the consensus (this can be viewed in base-level detail in a new window); and (I) the consensus sequence for the cluster. (J) From the cluster page (D) the user can jump to the protein translation page that includes, in addition to functional annotations; (K) the predicted protein sequence; (L) the best BLAST matches between this sequence and others in NEMBASE4, highlighted according whether the match is a reciprocal best BLAST match (and thus more likely to be a true orthologue) or just a top hit; and (M) the protein Tribes of which that the protein is a member. Tribes are shown for each of the nine inflation (or stringency) levels analysed. (N) The user can select a single tribe and be shown all of its members, sorted by species. Each member is hyperlinked back to its protein description page. The user can search NEMBASE in many other ways including: (O) by protein domain identifier (where the user is interested in a particular known domain from one of the 10 databases supported); (P) by the stage specificity of expression (using the stage-specificity metadata attributed to each EST library; users can also search by gender of expression, or for expression in a specific library); (Q) by phylogenetic restriction of the protein tribes (see Fig. 1). (R) By picking a node on the model tree relating all of the taxa studied, the user is shown a table of the numbers of tribes at each of the nine inflation values that are restricted to that node and only present in Nematoda (i.e. those which have no matches in UniRef100 that are not nematodes) or are also found outside Nematoda (i.e. those which have significant matches in UniRef100 that are from taxa other than Nematoda). Users can also search the KEGG pathway annotations directly, or search the sequences in the database using a local BLAST server.

tissue invasion systems that permit traversal of epithelia and basement membranes. Was horizontal gene transfer also a part of Strongylid evolution?

We thus pose the question: What protein tribes are unique to strongyles, are they potential targets for intervention, and do they show evidence of horizontal transfer into this parasitic clade?

There were 180 tribes at inflation 1.1 restricted to Strongylida (node 37 in Fig. 1 and Table 1), and 370 tribes at inflation 5. As would be expected, the number of tribes that were Strongylida-restricted increases as the stringency of clustering is increased, but the proportion of these that have matches to non-nematode sequences increases to nearly 20% of tribes defined with an inflation value of 5. This counter-intuitive pattern is the result of these tribes containing members that have significant matches to sequences from patent applications, where the species of origin is noted as "unknown". In fact, many of these tribes match patents describing hookworm vaccine candidate antigens. From this we conclude that all of these tribes are reasonable candidates for exploration as vaccine components or drug targets.

To answer the question concerning horizontal transfer, the four tribes at inflation 1.1 that had matches in non-nematodes were examined in more detail (Table 2). One matches the *lacZ* alpha peptide from cloning vectors and likely derives from poor trimming of sequences, and another has a single marginally significant match to a peptide predicted from marine metagenomic data with no functionally-informative matches or annotations.

Tribe inf11-1399 contains 30 sequences from six species. One has a marginally significant match to a late embryogenesis abundant (LEA)-like protein from the arthropod *Polypedilum vanderplanki*. LEA proteins are proteins of unordered structure that are found in many anhydrobiotic organisms, including plant seeds (where they were first identified) and nematodes, such as *Aphelenchus avenae*. LEA-like proteins have been found in the *Caenorhabditis* genome sequences and may be part of the ability of many nematodes to resist desiccation and freezing. NEMBASE4 does contain clusters with LEA-like sequence, but one characteristic of LEA protein sequences is their low complexity and in this case the match may be uninformative.

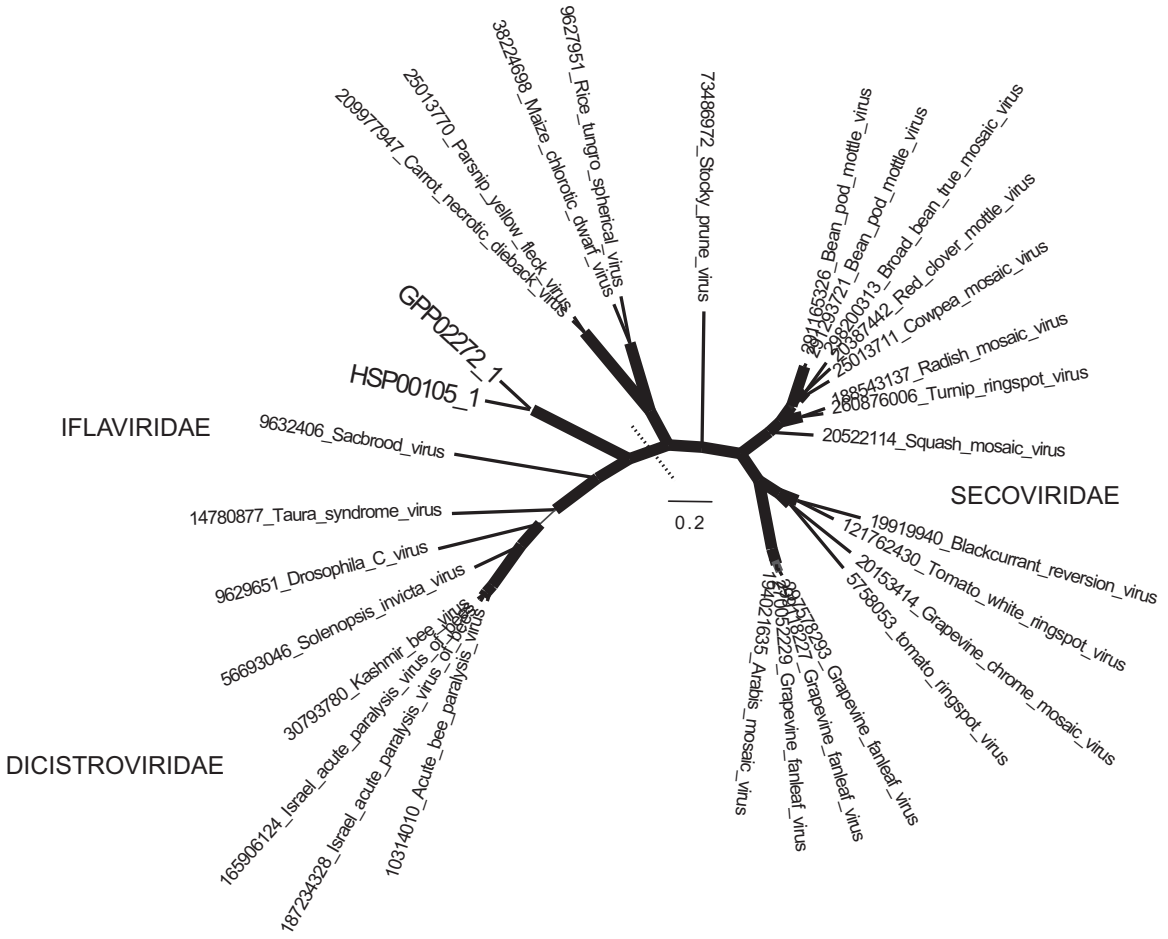

Fig. 5. Phylogenetic analysis of Picornavirus-like RNA-directed RNA polymerase genes identified in tylenchine nematodes. The unrooted phylogeny, developed in MrBayes, shows distinction between the arthropod-infecting Dicistrioviridae (taura syndrome, *Drosophila* C and the hymenopteran viruses) and Iflaviridae (represented by sacbrood virus), and the plant-pathogenic Secoviridae (all viruses to the right of the dotted line). The nematode sequences from *Globodera pallida* and *Heterodera scahchtii* form a clade distinct from these two groups. Branches with Bayesian posterior probabilities (pp) of 0.98 < pp < 1.00 are highlighted in bold.

Table 1
Tribes unique to the Strongylida within Nematoda at all inflation values.

Inflation[a]	Total number of tribes	Nematode-specific tribes	Non-nematode-specific tribes
1.1	180	176	4
1.5	226	212	14
2	263	237	26
2.5	299	261	38
3	319	270	49
3.5	335	280	55
4	350	288	62
4.5	360	294	66
5	370	299	71

[a] Inflation values used in TRIBE-MCL (Enright et al., 2002) analyses.

Tribe inf11-806 contains peptides derived from only two clusters, *Dictyocaulus viviparus* DVC03184 and *A. caninum* ACC12960, that have significant similarity to a NANOS-like protein from the crustacean *Parhyale hawaiiensis*. NANOS is a key player in the anterior–posterior patterning of animal zygotes, and has deeply conserved functions in the determination of the germline. NANOS functions through the binding of RNA, and NANOS proteins contain a distinct, conserved zinc-finger domain of ~55 amino acids (defined in Pfam05741, http://pfam.janelia.org/family/zf-nanos). The vast majority of NANOS proteins have a single zf-nanos domain, with only *P. hawaiiensis* NANOS having two. NANOS has been identified in species across the Metazoa including Cnidaria, Lophotrochozoa, Ecdysozoa and Deuterostomia but not previously from nematodes. The complete genomes of *C. elegans* and *C. briggsae* contain loci tagged as *nanos*-like, but the zinc finger domains they

Table 2
Non-nematode-specific tribes at inflation 1.1 unique to Strongylida within Nematoda.

Tribe	Number of members	Members that hit non nematode sequences	Minimum *E*-value	Top non-nematode hit description (species of origin)	GI[a] of top hit
inf11-1399	30	1	6e-07	PvLEA1 protein (*Polypedilum vanderplanki*; arthropod)	90,959,527
inf11-4078	9	1	2e-07	Hypothetical protein (marine metagenome; unknown)	134,777,650
inf11-6393	4	3	4e-12	*lacZ* alpha peptide	
inf11-8606	2	2	6e-12	*nanos* (*Parhyale hawaiensis*; arthropod)	161,898,489

[a] GI: NCBI unique identifier.

contain are poorly described by the zf-nanos model, and include deletions that are likely to be functionally significant. ACC12960 has two zf-nanos domains. DVC03184 has a single zf-nanos domain, but is a truncated sequence, matching only the 5′ (N-terminal) zf-nanos domain and flanking sequence of ACC12960. Alignment and phylogenetic analysis of representative NANOS and NANOS-like proteins (see Supplementary data S2) using Bayesian inference revealed a clade of vetebrate *nanos-1* zf-nanos domains, a clade of non-vertebrate domains and a third more diverse clade that included vertebrate NANOS-*2* and NANOS-*3* domains, as well as *Caenorhabditis*, *B. malayi* and *P. hawaiiensis* domains (Supplementary Fig. S1). The ACC12960 and DVC03184 zf-nanos domains were most closely related to each other (with the two *A. caninum* domains apparently a recent duplication) but these three sequences were not robustly placed in either the NANOS-*1* or NANOS-*2/3* clade. The highest BLAST-based match, to *P. hawaiiensis*, was likely because this was the only other sequence in the database that has two tandem zf-nanos domains, rather than close relationship between the *P. hawaiiensis* and strongylid genes.

The presence of these NANOS-like proteins in *D. viviparus* and *A. caninum* is intriguing. The lack of high similarity to vertebrate NANOS does not support a model of recent horizontal transfer from a vertebrate host, implying that they have long been resident in nematode genomes. The absence of zf-nanos domain genes in the fully sequenced genomes of *P. pacificus* and the *Meloidogyne* spp. must be due to multiple, independent losses, which might be thought unlikely a priori. However, if the nematode (and *P. hawaiiensis*) NANOS-like genes have been derived from NANOS-*2/3*-like ancestors, and subjected to rapid divergent evolution, this might explain both the extreme branch lengths in this part of the phylogeny, and the failure to recover a gene tree that matches the expected species tree. We searched the emerging genome sequence data from two strongylid nematodes for sequences matching these zf-nanos domains and identified highly similar sequences (45 of 50 residues identical) in *Heligmosomoides polygyrus* (from the Blaxter laboratory; http://www.nematodegenomes.org) and *Nippostrongylus brasiliensis* (from the Wellcome Trust Sanger Institute; http://www.sanger.ac.uk/resources/downloads/helminths/nippostrongylus-brasiliensis.html). This gene is thus part of strongylid nematode genomes. It will be informative to investigate the roles of these NANOS-like proteins in strongylid biology (where an involvement in germline development would suggest ancient retention). Interrupting the binding of these strongylid-restricted NANOS homologues might be a viable route to chemical abbreviation of infections.

We investigated only four of the 180 Strongylida-unique tribes in this limited exploration of NEMBASE4 and identified a potential developmental genetic novelty. There is a rich seam of additional tribes to be investigated for this group and indeed across the phylum. However, we found no evidence for horizontal gene transfer playing a significant role in strongylid nematode evolution and can provisionally reject horizontal gene transfer as a source of novel phenotypes in this group.

3.2.3. The evolution of heme auxotrophy in filarial nematodes

Many filarial nematodes contain an obligate endosymbiotic bacterium, *Wolbachia* (Fenn and Blaxter, 2004). This relationship has been utilised in recent efforts to identify drug targets in biochemical pathways absent from the nematode and present in *Wolbachia*, such as heme biosynthesis (Rao et al., 2005; Slatko et al., 2010). The genome of the *Wolbachia* from *B. malayi* has been sequenced (Foster et al., 2005), and heme biosynthesis identified as a possible essential part of the symbiotic relationship. Heme is an essential component of many proteins. The heme biosynthesis pathway (Fig. 6A), part of the KEGG 'porphyrin and chlorophyll metabolism' pathway, is believed to be completely missing from nematodes (and many other animals), except for the enzyme HemH or ferrochelatase (HemH/FC), which catalyses the terminal step of the pathway, the conversion of protoporphyrin IX to heme by inserting the Fe atom. The reliance on *Wolbachia* for heme is a promising drug target (Slatko et al., 2010) and might have resulted in changes in the nematodes' heme processing abilities. We thus formulated the following research question: Do *Wolbachia*-containing filarial nematodes differ from other species in their heme biosynthesis pathway?

Of the nine filarial nematodes present in NEMBASE4, seven are known to contain Wolbachia (*B. malayi*, *Brugia pahangi*, *Dirofilaria immitis*, *Litomosoides sigmodontis*, *Onchocerca ochengi*, *Onchocerca volvulus* and *Wuchereria bancrofti*) and two are not (*Loa loa* and *Onchocerca flexuosa*). We used the KEGG pathways search facility from the web interface to NEMBASE4 and identified clusters corresponding to four enzymes in the pathway (Table 3). Three enzymes are each present in only one, non-filarial species. Closer inspection of the two *A. caninum* proteins (the HemB-like ACP10593_1 and HemF-like ACP18701_1) revealed that both are closely related to enzymes from fungi. We were not able to identify similar genes in the emerging *H. polygyrus*, *Haemonchus contortus* and *N. brasiliensis* genomes, suggesting they may derive from a contamination event. The HemE-like MPP01279_1 from *Meloidogyne* paranaensis is most similar to HemE-like proteins from other metazoans, suggesting it may be a nematode gene, but we were unable to detect homologues in the *M. hapla* or *M. incognita* genomes.

We identified HemH/FC homologues in the filarial nematodes *L. loa* and *O. volvulus*, and in the Clade IV nematode *Strongyloides ratti* (Fig. 6A). The tribe associated with these clusters (tribe inf11-5740) also includes two genomic *B. malayi* HemH/FC-like proteins. The sequences from this tribe are most closely related to HemH/FC proteins from alphaproteobacteria and not metazoans, in agreement with previous findings (Slatko et al., 2010). We also searched the NEMBASE4 for 'ferrochelatase' annotations and identified proteins belonging to another tribe (inf11-3019) containing 15 proteins from 11 species. These sequences were related to other metazoan HemH/FC and we were able to identify homologues in many nematodes (Table 4; Supplementary data S3). Filarial nematodes thus contain up to three HemH/FC, two in the nuclear genome and one in the *Wolbachia* genome. We screened the emerging draft genome sequences from filarial nematodes (*L. sigmodontis* and *D. immitis* from the Blaxter laboratory (http://www.nematodegenomes.org/), and *W. bancrofti*, *L. loa* and *O. volvulus* from the Filarial Worms Sequencing Project, Broad Institute of Harvard and MIT, USA (http://www.broadinstitute.org/)) and confirmed the presence of two nuclear HemH/FC in *L. loa*, *D. immitis* and *L. sigmodontis*. The absence of the alphaproteobacterial HemH/FC in Clade V nematodes, despite the deep sampling of their transcriptomes by ESTs and the availability of several *Caenorhabditis* sp. genomes, suggests that this enzyme is not present in Clade V. While one member of the alphaproteobacterial-type enzymes was identified in Clade IV (*S. ratti*), neither extensive EST collections nor the complete genomes of *M. hapla* and *M. incognita* encode similar proteins. Thus, contrary to Slatko et al. (2010), HemH/FC is not absent from non-filarial nematodes, but an alphaproteobacterial-like isoform has limited distribution in the phylum, including filaria and *Strongyloides*.

We aligned nematode, *Wolbachia* and representative other HemH/FC proteins (Supplementary data S3) and analysed those using Bayesian phylogenetics (Fig. 6B), revealing that the ancestry of HemH/FC across the nematodes is more complex than previously thought. The *Wolbachia*-derived sequences are grouped with HemH from *Anaplasma* and *Ehrlichia*, as would be expected. The 'metazoan-like' HemH/FC sequences from nematodes (tribe inf11-3019) form an isolated clade that is weakly associated with other metazoan HemH/FC enzymes. One extraordinarily divergent

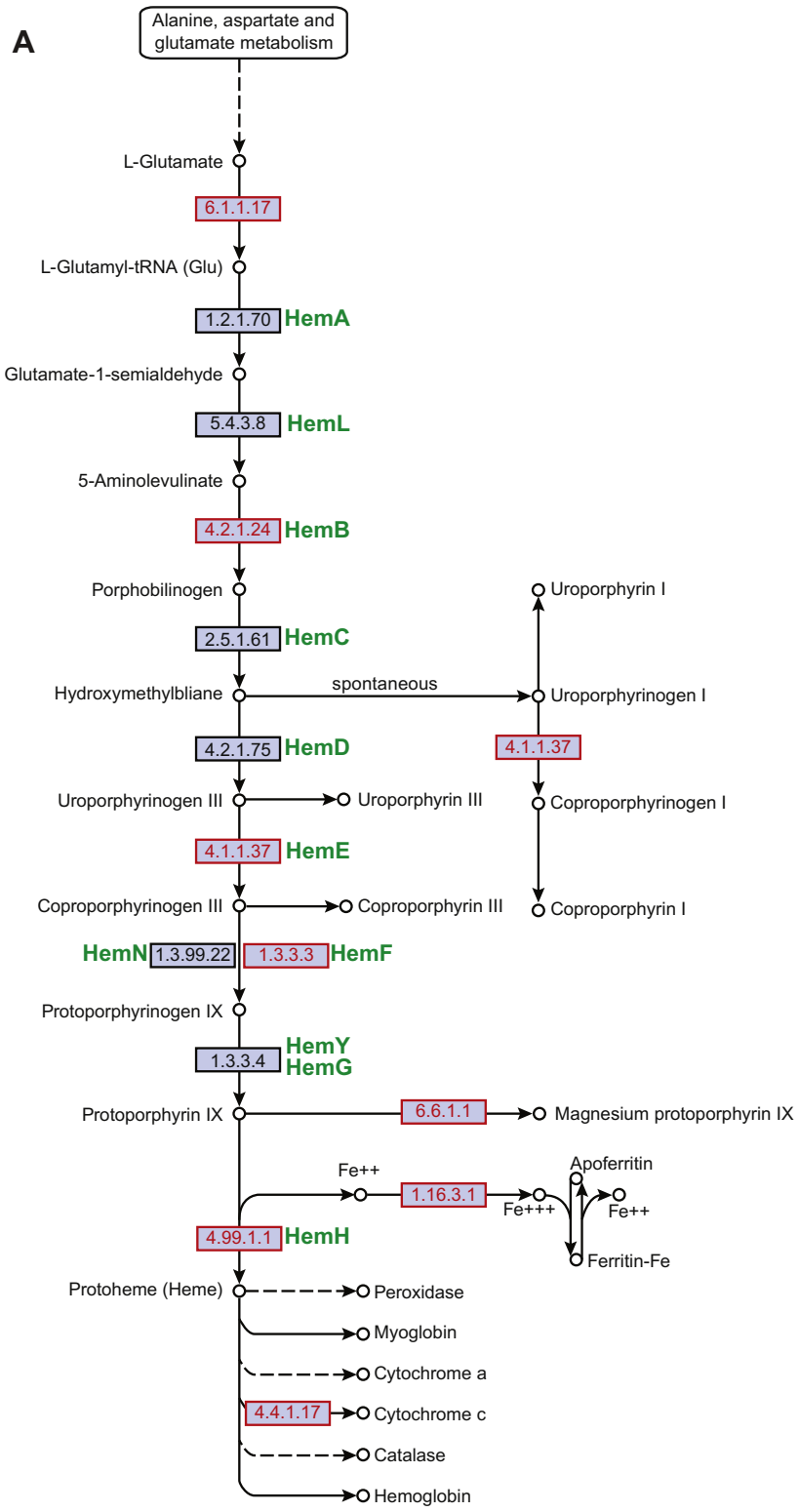

Fig. 6. The heme biosynthesis pathway. (A) The heme synthesis pathway redrawn from the Porphyrin and Chlorophyll Metabolism Kyoto Encyclopaedia of Genes and Genomes (KEGG) map (map00860). Each of the eight steps in heme synthesis are labelled in green (HemA to HemH). The first seven of these steps were believed to be missing in most nematodes. Enzymes that have at least one match from NEMBASE4 are highlighted in red. (B) Phylogenetic analysis of the enzyme performing the terminal step of heme biosynthesis, HemH/ferrochelatase. This Bayesian phylogeny shows the three subtypes of HemH/FC within nematodes: the *Wolbachia* type in black, the alphaproteobacterial type in blue (containing tribe inf11-5740), and the metazoan type in red and green (nematodes in green, containing tribe inf11-3019). NEMBASE4 proteins are listed with their ID and the tribe to which they belong for inflation 1.1. The branches are coloured corresponding to the Bayesian posterior probability (pp), with pp of $0.98 < pp < 1.00$ coloured red; pp values are noted next to each node. The tree is unrooted. The scale indicates inferred changes per site.

sequence from *C. briggsae* may be a misprediction and this may have masked support for the nematode-other metazoan link. Finally the filarial and *S. ratti* HemH/FC sequences nest robustly within a clade of alphaproteobacterial (*Rhizobium*, *Roseibium*) sequences.

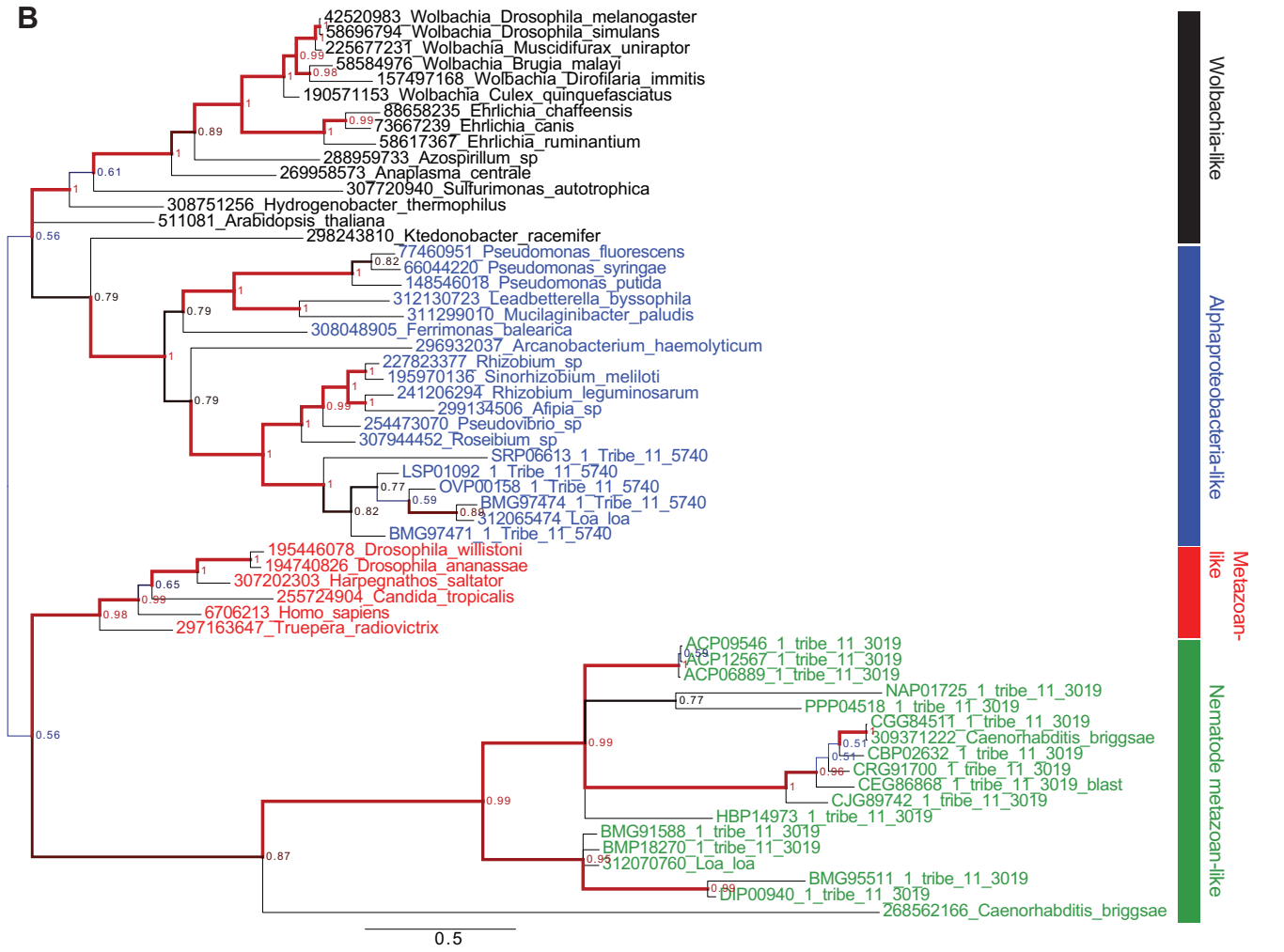

Fig. 6 (*continued*)

Table 3
Heme biosynthesis pathway enzymes: HemA–HemH pathway enzymes.

Enzyme	EC number	Common name	Clade III[a]	Clade IV[a]	Clade V[a]
HemA	1.2.1.70	Glutamyl-tRNA reductase			
HemB	4.2.1.24	Porphobilinogen synthase			ACP10593_1
HemC	2.5.1.61	Hydroxymethylbilane synthase			
HemD	4.2.1.75	Uroporphyrinogen-III synthase			
HemE	4.1.1.37	Uroporphyrinogen decarboxylase		MPP01279_1	
HemF	1.3.3.3/1.3.99.22	Coproporphyrinogen oxidase			ACP18701_1
HemG	1.3.3.4	Protoporphyrinogen oxidase			
HemH	4.99.1.1	Ferrochelatase	LSP01092_1 OVP00158_1	SRP06613_1	

EC: Enzyme Commission.
[a] Major nematode clades (see Fig. 1).

Even with the limited genomic evidence it would appear that there has either been a lateral gene transfer event in the last common ancestor of Clade IV and III nematodes (with subsequent loss from many taxa), or that there have been two independent acquisitions of this alphaproteobacterial sequence in Strongyloidoidea and Onchocercinae.

It has been demonstrated that *C. elegans* cannot use protoporphyrin IX for growth (Rao et al., 2005), implying that their metazoan-type HemH/FC is non-functional. This may also hold true for the homologues found in nematodes. The *Wolbachia*-type HemH/FC has been shown to be functional (Wu et al., 2009). The implications of these observations for drug target work focused on the heme biosynthesis pathway are significant: if there are multiple, very divergent copies of HemC/FC within filarial nematodes, designing drugs that target this step of the heme biosynthesis pathway must assess all three potential targets.

4. Discussion

Nematode EST programmes have been successful in identifying many genes of interest in target species, be they vaccine candidates, drug targets or potential host–parasite interaction mediators. Here

Table 4
Heme biosynthesis pathway enzymes: HemH/ferrochelatase proteins in NEMBASE4 and draft genomes.

Species	Metazoa-like	Alphaproteobacteria-like	*Wolbachia*-like
Clade III[a]			
Brugia malayi	BMP18270_1 BMG91588_1 BMG95511_1	BMG97471_1 BMG97474_1	58584976[b]
Litomosoides sigmodontis	Yes[c]	LSP01092_1	Yes[c]
Onchocerca volvulus	[No][d]	OVP00158_1	[No][d]
Loa loa	312070760[b]	312065474[b]	No[e]
Wuchereria bancrofti	[No][d]	ADHD01000089[b]	ADHD01000089[b]
Dirofilaria immitis	DIP00940_1	Yes[c]	Yes[c]
Clade IV[a]			
Strongyloides ratti		SRP06613_1	
Clade V[a]			
Ancylostoma caninum	ACP06889_1	ACP09546_1 ACP12567_1	
Caenorhabditis brenneri	CBP02632_1		
Caenorhabditis briggsae	CGG84511_1		
Caenorhabditis elegans	CEG86868_1		
Caenorhabditis japonica	CJG89742_1		
Caenorhabditis remanei	CRG91700_1		
Necator americanus	NAP01725_1		
Pristionchus pacificus	PPP04518_1		
Heterorhabditis bacteriophora	HBP14973_1		

[a] Major nematode clades (see Fig. 1).
[b] GenBank sequence identifiers from NCBI.
[c] Unpublished genome data from the Blaxter laboratory, Institute of Evolutionary Biology, The University of Edinburgh, UK (available at http://www.nematodegenomes.org/).
[d] Identified on the Filarial Worms Sequencing Project, Broad Institute of Harvard and MIT, USA (http://www.broadinstitute.org/) website; these sequences are partial, thus the absence of a match may be due to the draft sequence.
[e] *Loa loa* does not contain a *Wolbachia* endosymbiont.

we have shown that comprehensive analyses of the totality of these data can yield additional information not evident in analyses of single species. The reasons for this are many fold, and include the partial nature of EST data (they cannot represent all of the expressed genes of an organism), the power of cross-species comparisons (for sifting evolutionarily-conserved and thus interesting patterns from the background of neutral variation) and the utility of having collated data in one analysis environment.

NEMBASE4 offers significant data completeness and programming improvements over NEMBASE3. Nematode.net (Martin et al., 2009) offers an alternative assembly of nematode EST data. Within Nematode.net one can also search for clusters by annotation and view pathway information. However, Nematode.net currently collates data only for the core 37 taxa from NEMBASE3. The additional species representation and phylogenetically-aware searching of the protein tribes defined at different inflation values extends the usefulness of the resource. More recently, Cantacessi and colleagues (2010a) have introduced a transcriptome assembly workflow that includes Roche 454 read handling capabilities and used it for nematode data analysis, but this workflow processes single species and does not include a public-facing database portal.

Our three example analyses show the power of the NEMBASE4 integrated comparative resource for hypothesis generation and testing. We have added significantly to the roster of potentially laterally-transferred genes in the plant parasitic Tylenchina, and in addition identified a new virus family that is the first in these nematodes. A search for genes with similar signatures of lateral transfer in the Strongylida did not reveal any candidates, suggesting that the tylenchine pattern of incorporation of environmentally-acquired genes into the parasitic genome is not an universal feature of nematode parasites. Since non-tylenchine plant–parasitic nematodes have also been shown to incorporate laterally-acquired genes into their genomes; lateral transfer is not a specialism of the tylenchines. A fuller understanding of the dynamics of gene acquisition will assist in development of models evaluating the importance of this mechanism in evolution and in the development of nematicidal interventions aimed at disrupting these novel functions. Biochemical pathway analysis is facilitated by NEMBASE4, illustrated by the demonstration of patchy representation of heme biosynthesis genes across the phylum. Interestingly, we demonstrate that many filarial nematodes have three heme ferrochelatase enzymes (two from the nuclear genome and one from their *Wolbachia* endosymbiont), emphasising the importance of this druggable target for the nematodes. Obviously these findings require further experimental verification and exploration, but the power of NEMBASE4 for highlighting biological novelties is clear.

The era of Sanger dideoxy EST sequencing is probably approaching its end. The next generation platforms can now sequence cDNA (and genomes) for a tiny fraction of the cost and effort that clone-based EST projects entail. In particular the Roche 454 Titanium chemistry has replaced Sanger dideoxy sequencing for transcriptome projects, as it offers reasonable read lengths (360–400 bases average for cDNA) and massive production (1 million reads per 12 h run) for the same reagent cost as ~5000 Sanger dideoxy ESTs. De novo transcriptome assemblies from Roche 454 data (Kumar and Blaxter, 2010) are being published (Cantacessi et al., 2010b,c; Wang et al., 2010b). Illumina GAIIx and HiSeq 2000 instruments deliver shorter reads (up to 150 bases) in vast numbers and transcript sequencing protocols (called RNASeq) are widely used for transcript quantification in organisms with sequenced genomes (Wang et al., 2009). The promise of Illumina technology for de novo transcriptome sequencing has yet to be realised due to the difficulties of assembling these short reads but first attempts show promise (Mizrachi et al., 2010; Wang et al., 2010a).

The challenge for resources such as NEMBASE4 is to scale our analysis technologies to deliver integrated analyses of partial genome data for many more species and to integrate next generation data with the existing Sanger dideoxy ESTs. The challenge for parasitologists is to exploit these data for directed research programmes and NEMBASE4 will assist in this goal.

Acknowledgements

We thank all of the researchers who have deposited EST information in dbEST, and the authors of the open source or freely available programs that we have used. We also thank other members of the Blaxter laboratory and GenePool boinformatics team for helpful ideas. This project would have been much more difficult without the Edinburgh Compute and Data Facility. BE is supported by a Biotechnology and Biological Sciences Research Council, UK, PhD studentship. JW is supported by a Restracomp Fellowship awarded by the Hospital for Sick Children, Toronto, Canada. EST data reported here for the first time were generated by the GenePool Genomics Facility, UK (http://genepool.bio.ed.ac.uk/) with the support of the Wellcome Trust, the UK Medical Research Council and the Natural Environment Research Council, UK.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ijpara.2011.03.009.

References

Abad, P., Gouzy, J., Aury, J.M., Castagnone-Sereno, P., Danchin, E.G., Deleury, E., Perfus-Barbeoch, L., Anthouard, V., Artiguenave, F., Blok, V.C., Caillaud, M.C., Coutinho, P.M., Dasilva, C., De Luca, F., Deau, F., Esquibet, M., Flutre, T., Goldstone, J.V., Hamamouch, N., Hewezi, T., Jaillon, O., Jubin, C., Leonetti, P., Magliano, M., Maier, T.R., Markov, G.V., McVeigh, P., Pesole, G., Poulain, J., Robinson-Rechavi, M., Sallet, E., Segurens, B., Steinbach, D., Tytgat, T., Ugarte, E., van Ghelder, C., Veronico, P., Baum, T.J., Blaxter, M., Bleve-Zacheo, T., Davis, E.L., Ewbank, J.J., Favery, B., Grenier, E., Henrissat, B., Jones, J.T., Laudet, V., Maule, A.G., Quesneville, H., Rosso, M.N., Schiex, T., Smant, G., Weissenbach, J., Wincker, P., 2008. Genome sequence of the metazoan plant–parasitic nematode *Meloidogyne incognita*. Nat. Biotechnol. 26, 882–884.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Blaxter, M., 2003. Nematoda: genes, genomes and the evolution of parasitism. Adv. Parasitol. 54, 102–195.

Blaxter, M., Whitton, C., Thompson, M., Daub, J., Guiliano, D., Stirton, M., Jieru, Y., Aboobaker, A., Parkinson, J., 2004. Comparative nematode genomics. In: Cook, R., Hunt, D.J. (Eds.), Nematology Monographs and Perspectives, vol. 2. E.J. Brill, Leiden, pp. 557–571.

Blaxter, M., 2007. Symbiont genes in host genomes: fragments with a future? Cell Host Microbe 2, 211–213.

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. Mol. Biochem. Parasitol. 77, 77–96.

Blaxter, M.L., De Ley, P., Garey, J.R., Liu, L.X., Scheldeman, P., Vierstraete, A., Vanfleteren, J.R., Mackey, L.Y., Dorris, M., Frisse, L.M., Vida, J.T., Thomas, W.K., 1998. A molecular evolutionary framework for the phylum Nematoda. Nature 392, 71–75.

Cantacessi, C., Jex, A.R., Hall, R.S., Young, N.D., Campbell, B.E., Joachim, A., Nolan, M.J., Abubucker, S., Sternberg, P.W., Ranganathan, S., Mitreva, M., Gasser, R.B., 2010a. A practical, bioinformatic workflow system for large data sets generated by next-generation sequencing. Nucleic Acids Res. 38, e171.

Cantacessi, C., Mitreva, M., Campbell, B.E., Hall, R.S., Young, N.D., Jex, A.R., Ranganathan, S., Gasser, R.B., 2010b. First transcriptomic analysis of the economically important parasitic nematode, *Trichostrongylus colubriformis*, using a next-generation sequencing approach. Infect. Genet. Evol. 10, 1199–1207.

Cantacessi, C., Mitreva, M., Jex, A.R., Young, N.D., Campbell, B.E., Hall, R.S., Doyle, M.A., Ralph, S.A., Rabelo, E.M., Ranganathan, S., Sternberg, P.W., Loukas, A., Gasser, R.B., 2010c. Massively parallel sequencing and analysis of the *Necator americanus* transcriptome. PLoS Negl. Trop. Dis. 4, e684.

Cox-Foster, D.L., Conlan, S., Holmes, E.C., Palacios, G., Evans, J.D., Moran, N.A., Quan, P.L., Briese, T., Hornig, M., Geiser, D.M., Martinson, V., vanEngelsdorp, D., Kalkstein, A.L., Drysdale, A., Hui, J., Zhai, J., Cui, L., Hutchison, S.K., Simons, J.F., Egholm, M., Pettis, J.S., Lipkin, W.I., 2007. A metagenomic survey of microbes in honey bee colony collapse disorder. Science 318, 283–287.

Crooks, G.E., Hon, G., Chandonia, J.M., Brenner, S.E., 2004. WebLogo: a sequence logo generator. Genome Res. 14, 1188–1190.

Daub, J., Loukas, A., Pritchard, D.I., Blaxter, M., 2000. A survey of genes expressed in adults of the human hookworm, *Necator americanus*. Parasitology 120, 171–184.

Dieterich, C., Clifton, S.W., Schuster, L.N., Chinwalla, A., Delehaunty, K., Dinkelacker, I., Fulton, L., Fulton, R., Godfrey, J., Minx, P., Mitreva, M., Roeseler, W., Tian, H., Witte, H., Yang, S.P., Wilson, R.K., Sommer, R.J., 2008. The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism. Nat. Genet. 40, 1193–1198.

Enright, A.J., Van Dongen, S., Ouzounis, C.A., 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30, 1575–1584.

Felix, M.A., Ashe, A., Piffaretti, J., Wu, G., Nuez, I., Belicard, T., Jiang, Y., Zhao, G., Franz, C.J., Goldstein, L.D., Sanroman, M., Miska, E.A., Wang, D., 2011. Natural and experimental infection of *Caenorhabditis* nematodes by novel viruses related to nodaviruses. PLoS Biol. 9, e1000586.

Fenn, K., Blaxter, M., 2004. Are filarial nematode *Wolbachia* obligate mutualist symbionts? Trends Ecol. Evol. 19, 163–166.

Foster, J., Ganatra, M., Kamal, I., Ware, J., Makarova, K., Ivanova, N., Bhattacharyya, A., Kapatral, V., Kumar, S., Posfai, J., Vincze, T., Ingram, J., Moran, L., Lapidus, A., Omelchenko, M., Kyrpides, N., Ghedin, E., Wang, S., Goltsman, E., Joukov, V., Ostrovskaya, O., Tsukerman, K., Mazur, M., Comb, D., Koonin, E., Slatko, B., 2005. The *Wolbachia* genome of *Brugia malayi*: endosymbiont evolution within a human pathogenic nematode. PLoS Biol. 3, e121.

Ghedin, E., Wang, S., Spiro, D., Caler, E., Zhao, Q., Crabtree, J., Allen, J.E., Delcher, A.L., Guiliano, D.B., Miranda-Saavedra, D., Angiuoli, S.V., Creasy, T., Amedeo, P., Haas, B., El-Sayed, N.M., Wortman, J.R., Feldblyum, T., Tallon, L., Schatz, M., Shumway, M., Koo, H., Salzberg, S.L., Schobel, S., Pertea, M., Pop, M., White, O., Barton, G.J., Carlow, C.K., Crawford, M.J., Daub, J., Dimmic, M.W., Estes, C.F., Foster, J.M., Ganatra, M., Gregory, W.F., Johnson, N.M., Jin, J., Komuniecki, R., Korf, I., Kumar, S., Laney, S., Li, B.W., Li, W., Lindblom, T.H., Lustigman, S., Ma, D., Maina, C.V., Martin, D.M., McCarter, J.P., McReynolds, L., Mitreva, M., Nutman, T.B., Parkinson, J., Peregrin-Alvarez, J.M., Poole, C., Ren, Q., Saunders, L., Sluder, A.E., Smith, K., Stanke, M., Unnasch, T.R., Ware, J., Wei, A.D., Weil, G., Williams, D.J., Zhang, Y., Williams, S.A., Fraser-Liggett, C., Slatko, B., Blaxter, M.L., Scott, A.L., 2007. Draft genome of the filarial nematode parasite *Brugia malayi*. Science 317, 1756–1760.

Harcus, Y.M., Parkinson, J., Fernandez, C., Daub, J., Selkirk, M.E., Blaxter, M.L., Maizels, R.M., 2004. Signal sequence analysis of expressed sequence tags from the nematode *Nippostrongylus brasiliensis* and the evolution of secreted proteins in parasites. Genome Biol. 5, R39.

Harris, T.W., Antoshechkin, I., Bieri, T., Blasiar, D., Chan, J., Chen, W.J., De La Cruz, N., Davis, P., Duesbury, M., Fang, R., Fernandes, J., Han, M., Kishore, R., Lee, R., Muller, H.M., Nakamura, C., Ozersky, P., Petcherski, A., Rangarajan, A., Rogers, A., Schindelman, G., Schwarz, E.M., Tuli, M.A., Van Auken, K., Wang, D., Wang, X., Williams, G., Yook, K., Durbin, R., Stein, L.D., Spieth, J., Sternberg, P.W., 2010. WormBase: a comprehensive resource for nematode research. Nucleic Acids Res. 38, D463–D467.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., Madden, T.L., 2008. NCBI BLAST: a better web interface. Nucleic Acids Res. 36, W5–W9.

Kumar, S., Blaxter, M.L., 2010. Comparing de novo assemblers for 454 transcriptome data. BMC Genomics 11, 571.

Ledger, T.N., Jaubert, S., Bosselut, N., Abad, P., Rosso, M.N., 2006. Characterization of a new beta-1,4-endoglucanase gene from the root-knot nematode *Meloidogyne incognita* and evolutionary scheme for phytonematode family 5 glycosyl hydrolases. Gene 382, 121–128.

Martin, J., Abubucker, S., Wylie, T., Yin, Y., Wang, Z., Mitreva, M., 2009. Nematode.net update 2008: improvements enabling more efficient data mining and comparative nematode genomics. Nucleic Acids Res. 37, D571–D578.

McCarter, J.P., Mitreva, M.D., Martin, J., Dante, M., Wylie, T., Rao, U., Pape, D., Bowers, Y., Theising, B., Murphy, C.V., Kloek, A.P., Chiapelli, B.J., Clifton, S.W., Bird, D.M., Waterston, R.H., 2003. Analysis and functional classification of transcripts from the nematode *Meloidogyne incognita*. Genome Biol. 4, R26.

Meldal, B.H., Debenham, N.J., De Ley, P., De Ley, I.T., Vanfleteren, J.R., Vierstraete, A.R., Bert, W., Borgonie, G., Moens, T., Tyler, P.A., Austen, M.C., Blaxter, M.L., Rogers, A.D., Lambshead, P.J., 2007. An improved molecular phylogeny of the Nematoda with special emphasis on marine taxa. Mol. Phylogenet. Evol. 42, 622–636.

Mitreva, M., Jasmer, D.P., Appleton, J., Martin, J., Dante, M., Wylie, T., Clifton, S.W., Waterston, R.H., McCarter, J.P., 2004a. Gene discovery in the adenophorean nematode *Trichinella spiralis*: an analysis of transcription from three life cycle stages. Mol. Biochem. Parasitol. 137, 277–291.

Mitreva, M., McCarter, J.P., Martin, J., Dante, M., Wylie, T., Chiapelli, B., Pape, D., Clifton, S.W., Nutman, T.B., Waterston, R.H., 2004b. Comparative genomics of gene expression in the parasitic and free-living nematodes *Strongyloides stercoralis* and *Caenorhabditis elegans*. Genome Res. 14, 209–220.

Mitreva, M., McCarter, J.P., Arasu, P., Hawdon, J., Martin, J., Dante, M., Wylie, T., Xu, J., Stajich, J.E., Kapulkin, W., Clifton, S.W., Waterston, R.H., Wilson, R.K., 2005. Investigating hookworm genomes by comparative analysis of two *Ancylostoma* species. BMC Genomics 6, 58.

Mitreva, M., Smant, G., Helder, J., 2009. Role of horizontal gene transfer in the evolution of plant parasitism among nematodes. Methods Mol. Biol. 532, 517–535.

Mitreva, M., Jasmer, D.P., Zarlenga, D.S., Wang, Z., Abubucker, S., Martin, J., Taylor, C.M., Yin, Y., Fulton, L., Minx, P., Yang, S.P., Warren, W.C., Fulton, R.S., Bhonagiri, V., Zhang, X., Hallsworth-Pepin, K., Clifton, S.W., McCarter, J.P., Appleton, J., Mardis, E.R., Wilson, R.K., 2011. The draft genome of the parasitic nematode *Trichinella spiralis*. Nat. Genet. 43, 228–235.

Mizrachi, E., Hefer, C.A., Ranik, M., Joubert, F., Myburg, A.A., 2010. De novo assembled expressed gene catalog of a fast-growing Eucalyptus tree produced by Illumina mRNA-Seq. BMC Genomics 11, 681.

Opperman, C.H., Bird, D.M., Williamson, V.M., Rokhsar, D.S., Burke, M., Cohn, J., Cromer, J., Diener, S., Gajan, J., Graham, S., Houfek, T.D., Liu, Q., Mitros, T., Schaff, J., Schaffer, R., Scholl, E., Sosinski, B.R., Thomas, V.P., Windham, E., 2008. Sequence and genetic map of *Meloidogyne hapla*: a compact nematode genome for plant parasitism. Proc. Natl. Acad. Sci. USA 105, 14802–14807.

Papadopoulos, J.S., Agarwala, R., 2007. COBALT: constraint-based alignment tool for multiple protein sequences. Bioinformatics 23, 1073–1079.
Parkinson, J., Guiliano, D.B., Blaxter, M., 2002. Making sense of EST sequences by CLOBBing them. BMC Bioinformatics 3, 31.
Parkinson, J., Anthony, A., Wasmuth, J., Schmid, R., Hedley, A., Blaxter, M., 2004a. PartiGene – constructing partial genomes. Bioinformatics 20, 1398–1404.
Parkinson, J., Blaxter, M., 2004. Expressed sequence tags: analysis and annotation. Methods Mol. Biol. 270, 92–102.
Parkinson, J., Mitreva, M., Whitton, C., Thomson, M., Daub, J., Martin, J., Schmid, R., Hall, N., Barrell, B., Waterston, R.H., McCarter, J.P., Blaxter, M.L., 2004b. A transcriptomic analysis of the phylum Nematoda. Nat. Genet. 36, 1259–1267.
Parkinson, J., Whitton, C., Schmid, R., Thomson, M., Blaxter, M., 2004c. NEMBASE: a resource for parasitic nematode ESTs. Nucleic Acids Res. 32, D427–D430.
Rao, A.U., Carta, L.K., Lesuisse, E., Hamza, I., 2005. Lack of heme synthesis in a free-living eukaryote. Proc. Natl. Acad. Sci. USA 102, 4270–4275.
Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.
Schmid, R., Blaxter, M.L., 2008. Annot8r: rapid assignment of GO, EC and KEGG annotations. BMC Bioinformatics 9, 180.
Scholl, E.H., Thorne, J.L., McCarter, J.P., Bird, D.M., 2003. Horizontally transferred genes in plant-parasitic nematodes: a high-throughput genomic approach. Genome Biol. 4, R39.
Slatko, B.E., Taylor, M.J., Foster, J.M., 2010. The *Wolbachia* endosymbiont as an anti-filarial nematode target. Symbiosis 51, 55–65.
Stein, L.D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M.R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., Coulson, A., D'Eustachio, P., Fitch, D.H., Fulton, L.A., Fulton, R.E., Griffiths-Jones, S., Harris, T.W., Hillier, L.W., Kamath, R., Kuwabara, P.E., Mardis, E.R., Marra, M.A., Miner, T.L., Minx, P., Mullikin, J.C., Plumb, R.W., Rogers, J., Schein, J.E., Sohrmann, M., Spieth, J., Stajich, J.E., Wei, C., Willey, D., Wilson, R.K., Durbin, R., Waterston, R.H., 2003. The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. PLoS Biol. 1, E45.
The *C. elegans* Genome Sequencing Consortium, 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282, 2012–2018.
Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25, 4876–4882.
Wang, X.W., Luan, J.B., Li, J.M., Bao, Y.Y., Zhang, C.X., Liu, S.S., 2010a. De novo characterization of a whitefly transcriptome and analysis of its gene expression during development. BMC Genomics 11, 400.
Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63.
Wang, Z., Abubucker, S., Martin, J., Wilson, R.K., Hawdon, J., Mitreva, M., 2010b. Characterizing *Ancylostoma caninum* transcriptome and exploring nematode parasitic adaptation. BMC Genomics 11, 307.
Wasmuth, J., Schmid, R., Hedley, A., Blaxter, M., 2008. On the extent and origins of genic novelty in the phylum nematoda. PLoS Negl. Trop. Dis. 2, e258.
Wasmuth, J.D., Blaxter, M.L., 2004. Prot4EST: Translating Expressed Sequence Tags from neglected genomes. BMC Bioinformatics 5, 187.
Wu, B., Novelli, J., Foster, J., Vaisvila, R., Conway, L., Ingram, J., Ganatra, M., Rao, A.U., Hamza, I., Slatko, B., 2009. The heme biosynthetic pathway of the obligate *Wolbachia* endosymbiont of *Brugia malayi* as a potential anti-filarial drug target. PLoS Negl. Trop. Dis. 3, e475.
Wylie, T., Martin, J.C., Dante, M., Mitreva, M.D., Clifton, S.W., Chinwalla, A., Waterston, R.H., Wilson, R.K., McCarter, J.P., 2004. Nematode.net: a tool for navigating sequences from parasitic and free-living nematodes. Nucleic Acids Res. 32, D423–D426.
Zdobnov, E.M., Apweiler, R., 2001. InterProScan – an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17, 847–848.