



400 000 nematode ESTs on the Net

John Parkinson¹, Makedonka Mitreva², Neil Hall³, Mark Blaxter¹ and James P. McCarter²

¹Institute of Cell, Animal and Population Biology, Ashworth Laboratories, King's Buildings, West Mains Rd, University of Edinburgh, Edinburgh, UK EH9 3JT

²Genome Sequencing Center, Washington University School of Medicine, Campus Box 8501, 4444 Forest Park Blvd, St Louis, MO 63108, USA

³Pathogens Sequencing Unit, Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK CB10 1SA

The parasitic nematode expressed sequence tag (EST) project, a collaboration between University of Edinburgh and the Wellcome Trust Sanger Institute in the UK and the Genome Sequencing Center, St Louis, MO, USA, is currently generating sequence information from >30 different species of nematode. Over 400 000 nematode ESTs are now available and at least another 130 000 are planned. Here, an update is provided on the status of the project and describes the database tools being developed to disseminate these data.

With the availability of high throughput sequencing, genome projects have been initiated for a range of different parasites (a comprehensive list of completed and ongoing genome projects is available at <http://wit.integratedgenomics.com/GOLD/> [1]). For some of the smaller genomes, sequencing is relatively inexpensive and offers the potential of providing an understanding of a parasite's biology and its interactions with host(s), and in the identification of novel vaccine candidates and drug targets [2,3]. For organisms with larger genomes, such as nematodes, the current costs of sequencing can be prohibitive. This is further exacerbated when considering that nematodes represent a highly diverse group of organisms; therefore, a single 'representative nematode' genome sequence is inadequate for understanding their biology. The Institute for Genomic Research (Rockville, MD, USA) has been funded to generate an expected 98% coverage of the genome of the human pathogen *Brugia malayi* (Box 1a) which will complement the genome information available for the free-living nematodes *Caenorhabditis elegans* (Box 1b) and *Caenorhabditis briggsae* (Box 1c). However, to allow sampling from a greater diversity of nematodes, expressed sequence tags [ESTs, single pass-reads derived from randomly selected complementary DNA (cDNA) clones] can be used to provide a cost-effective route to gene discovery [4,5]. Previously, we have outlined the plans to provide EST sequence data for several different parasitic nematodes and to make these data available on the web [6–8]. Here, we provide an update on the status of ongoing sequencing projects and provide a synopsis of the online tools available to access these data.

At present, two sister projects are involved in high throughput nematode EST generation. The first involves a collaboration between the nematode genomics group at the University of Edinburgh, UK (ED), and the Pathogens Sequencing Unit of the Wellcome Trust Sanger Institute, Cambridge, UK (PSU), and the second project is based at the Genome Sequencing Center in St Louis, MO, USA (GSC). These projects have already generated 173 000 ESTs and at least another 130 000 are planned. Including the ~190 000 sequences derived from *C. elegans* [9], there are >400 000 nematode ESTs in the public dbEST database (Table 1). The species selected for sequencing come from four out of the five major nematode clades (Fig. 1). The availability of sequence data from free-living nematodes and from both animal- and plant-parasitic nematodes will not only yield insights into the design of

Box 1. Websites of interest

- (a) **The Institute for Genomic Research *Brugia malayi* Genome Project**
<http://www.tigr.org/tdb/e2k1/bma1/>
- (b) **The *Caenorhabditis elegans* Genome Project at the Wellcome Trust Sanger Institute, Cambridge, UK**
http://www.sanger.ac.uk/projects/c_elegans/
- (c) **The *Caenorhabditis briggsae* Genome Project at the Genome Sequencing Center, Washington University Medical School, St Louis, MO, USA**
<http://genome.wustl.edu/projects/cbriggsae/>
- (d) **NEMBASE**
<http://www.nematodes.org>
- (e) **NemaGene**
<http://www.nematode.net>
- (f) **The Filarial Genome Network**
<http://nema.cap.ed.ac.uk/fgn/filgen1.html>
- (g) **National Center for Biotechnology Information**
<http://www.ncbi.nlm.nih.gov/>
- (h) **Sequence Retrieval System at the European Bioinformatics Institute**
<http://srs.ebi.ac.uk/>
- (i) **PHRAP**
<http://www.phrap.org/>
- (j) **Wormpep92**
The current release (as of 5 December 2002) contains 21 118 entries of which 2818 are predicted to have splice variants.
http://www.sanger.ac.uk/Projects/C_elegans/wormpep/
- (k) **Sequences hosted by NemaGene**
http://www.nematode.net/FTP/cluster_ftp/index.php

Corresponding author: John Parkinson (john.parkinson@ed.ac.uk).

Table 1. Nematode expressed sequence tag projects^a

Clade ^a	Species	Description	Sequencing centre	Submitted ESTs ^b	Clusters/ sequences (redundancy) ^c	Planned ESTs	Available libraries	Database ^d	
I	<i>Trichinella spiralis</i>	Human muscle parasite	GSC	10 372	3458/10 153 (2.94)	23 000	3	NemaGene	
	<i>Trichuris muris</i>	Mouse threadworm	ED,PSU	2125	1322 (1.61)	20 000	1	NEMBASE	
	<i>Trichuris trichuria</i>	Human threadworm	ED,PSU	–	Pending	5000	0	(NEMBASE)	
III	<i>Ascaris lumbricoides</i>	Human gut parasite	ED,PSU	1822	892 (2.04)	12 000	1	NEMBASE	
	<i>Ascaris suum</i>	Swine gut parasite	GSC,ED,PSU	31 933	6628/29 624 (4.47)	40 000	22	NEMBASE	
	<i>Brugia malayi</i>	Human lymphatic parasite	FGP,GSC	23 887	8392/18 741 (2.23)	23 887	20	NEMBASE	
	<i>Dirofilaria immitis</i>	Canine heart worm	GSC	3949	Pending	5000	2	(NemaGene)	
	<i>Litomosoides sigmodontis</i>	Rodent model filarial parasite	ED,JA	198	Pending	2000	1	(NEMBASE)	
	<i>Onchocerca volvulus</i>	Human filarial parasite	FGP	14 922	3504/7911 (2.26)	14 922	9	NEMBASE	
	<i>Toxocara canis</i>	Canine gut parasite	GSC,ED,RM	4206	1466 (2.87)	5000	3	NEMBASE	
Iva	<i>Parastrongyloides trichosuri</i>	Possum gut parasite	GSC	7963	Pending	10 000	7	(NemaGene)	
	<i>Strongyloides ratti</i>	Rodent gut parasite	GSC	9024	Pending	21 500	5	(NemaGene)	
	<i>Strongyloides stercoralis</i>	Human gut parasite	GSC	11 392	3311/10 908 (3.29)	11 392	2	NemaGene	
Ivb	<i>Globodera rostochiensis</i>	Potato cyst nematode	GSC,JJ,GS ^e	5934	Pending	TBD	1	(NemaGene)	
	<i>Globodera pallida</i>	Potato cyst nematode	JJ,GS ^e	1832	Pending	TBD	1	(NEMBASE)	
	<i>Heterodera glycines</i>	Soy bean cyst	GSC ^e	4327	1790/4307 (2.41)	11 500	4	NemaGene	
	<i>Meloidogyne arenaria</i>	Root knot nematode	GSC	3334	Pending	95 000 ^f	1	(NemaGene)	
	<i>Meloidogyne hapla</i>	Root knot nematode	GSC	11 049	Pending	–	3	(NemaGene)	
	<i>Meloidogyne incognita</i>	Root knot nematode	GSC,GS	12 752	1603/5661 (3.53)	–	2	NemaGene	
	<i>Meloidogyne javanica</i>	Root knot nematode	GSC	5600	Pending	–	4	(NemaGene)	
	<i>Pratylenchus penetrans</i>	Plant lesion nematode	GSC	2048	Pending	2048	1	(NemaGene)	
	<i>Zeldia punctata</i>	Free-living	GSC	395	195/378 (1.94)	1000	1	NemaGene	
V	<i>Ancylostoma caninum</i>	Dog hookworm	GSC	9362	2656/5553 (2.09)	10 000	4	NemaGene	
	<i>Ancylostoma ceylanicum</i>	Human hookworm	GSC	5371	Pending	6000	2	(NemaGene)	
	<i>Ancylostoma duodenale</i>	Human hookworm	GSC	–	Pending	4000	1	(NemaGene)	
	<i>Caenorhabditis briggsae</i>	Free-living	GSC	2424	NA	2424	–	– ^g	
	<i>Caenorhabditis elegans</i>	Free-living	NIG, GSC, TIGR, Sanger	189 632	NA	–	–	– ^h	
	<i>Haemonchus contortus</i>	Sheep gut parasite	GSC,ED,PSU	6312	1970/5181 (2.63)	20 000	7	NEMBASE	
	<i>Necator americanus</i>	Human hookworm	ED,PSU	4766	2298 (2.07)	20 000	3	NEMBASE	
	<i>Nippostrongylus brasiliensis</i>	Rodent gut parasite	ED,RM	1234	750 (1.64)	2000	3	NEMBASE	
	<i>Ostertagia ostertagi</i>	Cattle gut parasite	GSC	7009	Pending	10 000	5	(NemaGene)	
	<i>Pristionchus pacificus</i>	Free living	GSC	8818	2603/4979 (1.91)	15 000	4	NemaGene	
	<i>Teladorsagia circumcincta</i>	Sheep gut parasite	ED,PSU	315	Pending	20 000	1	(NEMBASE)	
			Total (excluding <i>C. elegans</i>)	212 627	Total planned ESTs (excluding <i>C. elegans</i> and <i>Globodera</i> spp.)	412 673			

^a The phylum Nematoda has previously been defined into five major clades [10,11]. Abbreviations: ED, Mark Blaxter Nematode Genomics Laboratory, University of Edinburgh, UK in conjunction with the Wellcome Trust Sanger Institute, Cambridge, UK; EST, expressed sequence tag; FGP, Filarial Genome Project (Box 1f); GS, Geert Smant, Department of Plant Science, Wageningen University, Wageningen, The Netherlands; GSC, Genome Sequencing Center, Washington University School of Medicine, St Louis, MO, USA; JA, Judith Allen, Institute of Cell, Animal and Population Biology, University of Edinburgh, UK; JJ, John Jones, The Scottish Crop Research Institute, Dundee, UK; NA, not available; NIG, Yuji Kohara, National Institute of Genetics, Mishima, Japan; PSU, Pathogen Sequencing Unit, The Wellcome Trust Sanger Institute, Cambridge, UK; RM, Rick Maizels, Institute of Cell, Animal and Population Biology, University of Edinburgh, UK; TBD, to be determined; TIGR, The Institute for Genomic Research, Gaithersburg, MA, USA.

^b ESTs deposited as at 27 November 2002.

^c As a result of the time lag between submitting sequences and generating partial genomes, for certain species, not all deposited sequences have been clustered. For entries where there are two numbers (e.g. 1880/3979), the second number refers to the number of sequences used for generating the clusters. The number in parentheses refers to sequence redundancy (number of sequences/number of clusters).

^d Partial genome databases for the majority of species are split between the two sites: NEMBASE (Box 1d) and NemaGene (Box 1e). For species where a partial genome has yet to be created, the database in which it is due to appear is listed in brackets.

^e See Ref. [12].

^f This figure represents an estimate of the number of planned ESTs for all *Meloidogyne* spp.

^g <http://genome.wustl.edu/gsc/projects/briggsae.shtml>.

^h <http://www.wormbase.org> http://www.ddbj.nig.ac.jp/c-elegans/html/ce_index.html.

novel drug targets and vaccine candidates [13], but will also provide an exciting source of data to examine the evolutionary history of this important taxonomic group.

The ESTs are deposited in the public databases as they are generated (Box 1g,h). To maximize transcript discovery in each species, several different cDNA libraries are prepared and used for EST sampling. The numbers and nature of these libraries varies between species and according to availability of material. For example, there are ESTs from 22 different sex- and tissue-specific libraries for *Ascaris suum*, whereas *Necator americanus* ESTs have been generated from only three different stage-specific libraries.

Individual ESTs tend to be short (up to 700 bp) and prone to sequencing errors. However, because it is possible

to obtain more than one sequence for the same gene, ESTs from the same species can be grouped on the basis of sequence similarity into clusters that are putatively derived from one gene. This leads to a reduction in the number of errors and an increase in the effective length of the derived transcript sequence. This clustering process permits large EST datasets to be analyzed in a whole-transcriptome context; we call the cluster datasets 'partial genomes'. Once created, partial genomes greatly facilitate data mining and allow genomic-style comparisons to be performed, although one must always be aware of the incomplete nature of the partial genome.

Databases describing the partial genomes of nematodes for which significant sequence information has been generated are available at ED (Box 1d) and at GSC

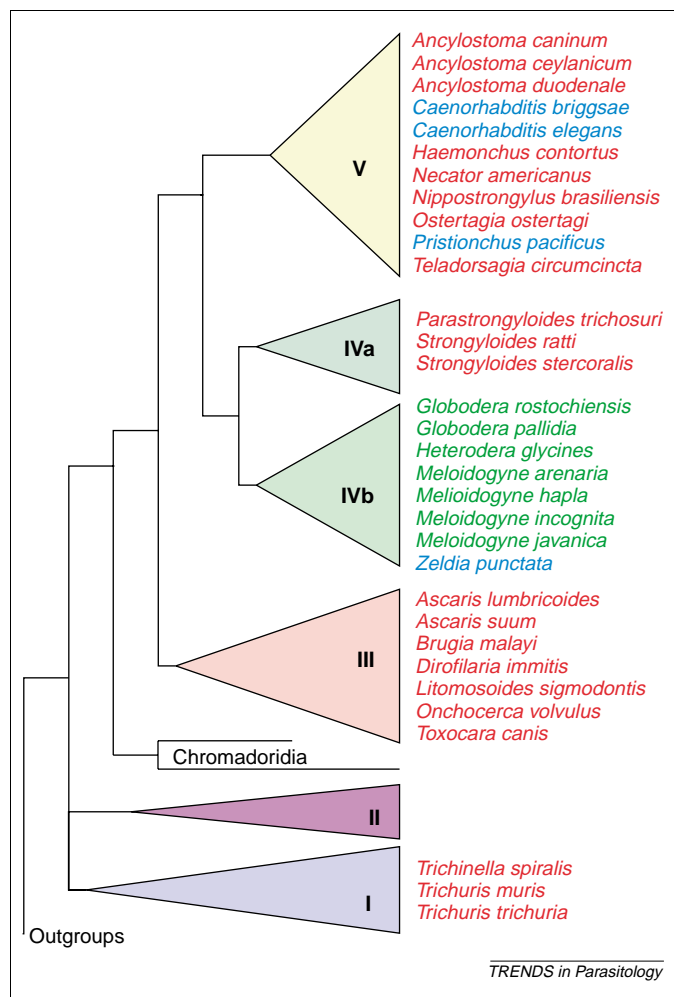


Fig. 1. Expressed sequence tag datasets from diverse species from the phylum Nematoda. Nematoda can be divided into five major clades [10,11] and four have been sampled using the expressed sequence tag (EST) strategy. Nematode species for which significant sequence datasets exist are listed on the right. Key: blue font, free-living nematode; green font, plant-parasitic nematode; red font, animal-parasitic nematode. Note that no sequencing is being performed on any clade II organism within this collaboration because of a lack of appropriate material.

(Box 1e). The species presented by both databases do not currently overlap, although in the future NEMBASE is expected to host all nematode EST datasets. The processes used to generate the partial genomes differ between the two sites. For NEMBASE, sequences are uploaded from dbEST and grouped on the basis of sequence similarity into clusters using CLOBB, a program that is capable of maintaining cluster integrity between consecutive builds [14]. After clustering, sequences are assembled into contigs (representing putative gene sequences) using the fragment assembly program PHRAP (Box 1i). In the NemaGene process, raw sequence traces are processed by the base-calling program PHRED [15,16] to create sequence files that are then assembled into draft clusters using PHRAP. Individual clusters containing more than four ESTs are manually edited using a combination of in-house scripts and the contig editing tool Consed [17]. In both processes, the final derived set of contigs represents the partial genome and is treated to further annotation

(e.g. BLAST similarity to other sequences and peptide predictions).

Generation of partial genomes is a time-consuming process and hence performed only when sufficient new sequence data warrants an update. To date, the partial genomes for 16 species have been assembled (Table 1). However, as we near the end of sequence generation, efforts will focus on sequence analysis and we expect that the ESTs from the 31 non-*Caenorhabditis* spp. will be clustered and made available online as partial genomes (at NEMBASE and/or NemaGene) by 2005. The level of redundancy (ESTs per gene) varies according to species (from 1.6 for *Trichuris muris* to 4.5 for *A. suum*; Table 1). Redundancy is dependent on many factors, including number of ESTs, number of libraries sampled and the complexity of the organism's expressed genome. For *A. suum*, the high redundancy value of almost 4.5 is almost certainly due to fact that the majority of libraries used for sequencing were derived from adult material. From these figures, it is estimated that 10 000 sequences will yield a partial genome of ~3000–4000 genes, whereas 20 000 sequences should yield a partial genome of 5000–8000 genes. Current estimates suggest that the number of genes in *C. elegans* is ~20 000 (Box 1j). These sequencing projects will therefore provide 10–40% of the genes of each species.

Once clustered and assembled into partial genomes, several methods have been developed that enable data mining of this information. At NemaGene, clusters can be retrieved either on the basis of simple annotation derived from BLAST searches (via a text-based query e.g. kinase) or on the basis of sequence similarity using a locally installed BLAST server. NemaGene clusters are also available by file transfer protocol (FTP) (Box 1k). NEMBASE also allows cluster retrieval via sequence similarity and annotation but, in addition, offers the ability to retrieve clusters on the basis of library expression (e.g. all the clusters that are expressed in library A but not in library B) and also on the basis of similarity profiles using an in-house, Java-based tool, SimiTri [18]. In addition to these search methods, clusters can also be viewed by their relative expression.

The large numbers of sequences involved necessitates the use of an automated approach for generating and annotating the partial genomes. This can occasionally lead to the incorrect clustering of sequences (e.g. sequencing errors can lead to splitting of clusters of sequences that were originally derived from the same gene) or misleading annotation (e.g. even genes that share as much as 60% sequence identity can be found to have different functions). To circumvent these problems, NEMBASE attempts to provide access to as much information as possible (e.g. links to BLAST output, links to the raw trace files and links to the alignment files used to create the contigs), so that the user can make up their own mind regarding the division and annotation of the partial genomes. Hence, these partial genomes provide a valuable entry point into exploring the biology of these nematodes and as a whole will provide useful insights into the biology of parasitism.

Initial analysis of datasets on NEMBASE reveals that for four out of the nine nematode species available, three

out of the top five most abundantly expressed genes do not have an ascribed function. Initial comparisons of the datasets with the *C. elegans* proteome also reveal that 40–60% of the derived genes for each species do not share any sequence similarity with *C. elegans* (data not shown). In one extreme case, out of 5181 ESTs derived from the sheep parasite *Haemonchus contortus*, over 10% are from a small family of genes that only shares similarity to *C. elegans* hypothetical genes (F54D5.3 and F54D5.4). This gene is now the subject of an intensive programme of research.

In the short term, database development is concentrating on improved methods of annotation including prediction of peptides from the consensus sequences. Future development will attempt to incorporate features of nematode biology such as identification of nematode-specific gene families and analysis of metabolic pathways.

Acknowledgements

We thank Claire Whitton (ED), Marian Thomson (ED), Jen Daub (ED), Claire Murphy (GSC) and Brandi Chiapelli (GSC) for their invaluable contributions in making libraries, preparing the clones and undertaking preliminary sequencing, and Mike Dante (GSC) for his role in building and maintaining the NemaGene clusters. The authors acknowledge the support of Bart Barrell (PSU) who helped initiate the project and oversee the sequencing performed at PSU. Work in ED is supported by the Medical Research Council and the Wellcome Trust; the PSU is supported by the Wellcome Trust. J.M. was supported by a Helen Hay Whitney/Merck Postdoctoral Fellowship. EST sequencing at GSC was supported by NIH NIAID research grant AI 46593 to Robert Waterston and by a National Science Foundation Plant Genome award 0077503 to David Bird and Sandra Clifton. We thank our research colleagues who have supplied nematode materials and libraries (Judith Appleton, Prema Arasu, Thomas Baum, David Bird, Bernadette Connolly, Richard Davis, Eric Davis, Lou Gasbarree, Tim Geary, Godelieve Gheysen, Warwick Grant, Richard Grecis, John Hawdon, Doug Jasmer, Vadim Kapulkin, Andrew Kloek, David Knox, Rick Maizels, Thomas Nutman, David Pritchard, Alan Scott, Geert Smant, Ralf Sommer, Mark Viney, Gary Weil, Valerie Williamson and Dante Zarlenga).

References

- 1 Bernal, A. *et al.* (2001) Genomes online database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res.* 29, 126–127
- 2 Gardner, M.J. *et al.* (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419, 498–511
- 3 Katinka, M.D. *et al.* (2001) Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* 414, 450–453
- 4 Blaxter, M.L. *et al.* (1999) Parasitic helminth genomics. *Parasitology* 118, S39–S51
- 5 Johnston, D.A. *et al.* (1999) Genomics and the biology of parasites. *Bioessays* 21, 131–147
- 6 Parkinson, J. *et al.* (2001) 200000 nematode expressed sequence tags on the Net. *Trends Parasitol.* 17, 394–396
- 7 McCarter, J. *et al.* (2000) Rapid gene discovery in plant parasitic nematodes via expressed sequence tags. *Nematology* 2, 719–731
- 8 McCarter, J.P. *et al.* (2002) Nematode gene sequences, update for June 2002. *J. Nematology* 34, 71–74
- 9 Kohara, Y. (1996) [Large scale analysis of *C. elegans* cDNA] *Tanpakushitsu Kakusan Koso* 41, 715–720
- 10 Blaxter, M.L. *et al.* (1998) A molecular evolutionary framework for the phylum Nematoda. *Nature* 392, 71–75
- 11 Dorris, M. *et al.* (1999) Molecular analysis of nematode diversity and the evolution of parasitism. *Parasitol. Today* 15, 188–193
- 12 Popeijus, H. *et al.* (2000) Analysis of genes expressed in second stage juveniles of the potato cyst nematodes *Globodera rostochiensis* and *G. pallida* using the expressed sequence tag approach. *Nematology* 2, 567–574
- 13 Lizotte-Waniewski, M. *et al.* (2000) Identification of potential vaccine and drug target candidates by expressed sequence tag analysis and immunoscreening of *Onchocerca volvulus* larval cDNA libraries. *Infect. Immun.* 68, 3491–3501
- 14 Parkinson, J. *et al.* (2002) Making sense of EST sequences by CLOBBing them. *BMC Bioinformatics* 3, 31
- 15 Ewing, B. *et al.* (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* 8, 175–185
- 16 Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 8, 186–194
- 17 Gordon, D. *et al.* (1998) Consed: A graphical tool for sequence finishing. *Genome Res.* 8, 195–202
- 18 Parkinson, J. and Blaxter, M. (2003) SimiTri – visualising similarity relationships for groups of sequences. *Bioinformatics* 19, 390–395