Parasitic helminth genomics

M. BLAXTER[1], M. ASLETT[1,2], D. GUILIANO[1], J. DAUB[1], *and* THE FILARIAL GENOME PROJECT[3]

[1] *Institute of Cell, Animal and Population Biology, University of Edinburgh, King's Buildings, Edinburgh EH9 3JT, UK*
[2] *The European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK*
[3] *The Filarial Genome Project includes the laboratories of Steven A. Williams, Clark Science Center, Smith College, Northampton, MA 01063 USA; Kunthala Jayaraman, Center for Biotechnology, Anna University, Madras 600025 India; Reda Ramzy, Research & Training Center on Vectors of Diseases, Ain Shams University, Abbassia, Cairo 11566 Egypt; Alan Scott, Dept. Molecular Microbiology and Immunology, Johns Hopkins School of Hygiene, 615 North Wolfe St., Baltimore MD 21205 USA; Tania Supali, Department of Parasitology, Faculty of Medicine, University of Indonesia, Salewba 6, Jakarta, 10430 Indonesia; and Barton Slatko, New England Biolabs, 32 Tozer Road, Beverly, MA 01915 USA*

SUMMARY

The initiation of genome projects on helminths of medical importance promises to yield new drug targets and vaccine candidates in unprecedented numbers. In order to exploit this emerging data it is essential that the user community is aware of the scope and quality of data available, and that the genome projects provide analyses of the raw data to highlight potential genes of interest. Core bioinformatics support for the parasite genome projects has promoted these approaches. In the *Brugia* genome project, a combination of expressed sequence tag sequencing from multiple cDNA libraries representing the complete filarial nematode lifecycle, and comparative analysis of the sequence dataset, particularly using the complete genome sequence of the model nematode *C. elegans*, has proved very effective in gene discovery.

Key words: genomics, parasitic genomes, *Caenorhabditis elegans*, *Brugia malayi*, *Schistosoma*, expressed sequence tag, genome mapping, bioinformatics.

Abbreviations: BAC bacterial artificial chromosome, YAC yeast artificial chromosome, EST expressed sequence tag, WWW world wide web.

INTRODUCTION

Parasitic helminths are complex, advanced organisms which have evolved to exploit the food-rich niches of their hosts' internal milieux. Despite early concepts of parasites as essentially degenerate organisms, they appear to have (mostly) retained the intricate biochemistry of their free living ancestors, and have developed new pathways to cope with host nutritional limitations and host immune attack, amongst other pressures. Until very recently, molecular biological analysis of parasitic helminths was limited to the cloning of potential vaccine candidate antigen genes, and the illumination of some specific facets of parasite biochemistry, often pertaining to drug metabolism. Two revolutions have changed helminth molecular biology: the advent of mass sequencing and the zeitgeist which accompanies the concept of genome projects. Mass sequencing has allowed extensive gene discovery programmes to be initiated and executed at low cost, resulting in the flooding of the public databases with tens of thousands of parasitic helminth sequences. Techniques (and robotic technologies) are available to clone, analyse and integrate data from large numbers of fragments of genomic DNA or cDNA. The inception of genome projects on model organisms such as the nematode *Caenorhabditis elegans*, and the launching of the human genome project, has changed the way it is possible to think about an organism. As all organisms' DNA is essentially similar, the choice of donor species becomes irrelevant: the only consideration is 'Is there a need for a genome initiative, and is there a research or commercial community ready to exploit its outcomes?'.

The parasitic helminths are not a natural group (Winnepenninckx *et al.* 1995). The parasitic flatworms (phylum Platyhelminthes) are not at all closely related to the parasitic Nematoda. Even within each of these phyla, different parasitic groups can be but distantly related. With this in mind, it is obvious that generalizations may be a little wide, and indeed may not be possible. Here we focus on the human filarial nematode parasite, *Brugia malayi*, because it is the organism we work most closely on, and because it is the parasite for which most data exist (Blaxter, 1995; Blaxter *et al.* 1997). However, from a genomics point of view, 2 features do link these phyla: they have large genome sizes compared to the parasitic protozoa, and they have metazoan body plans including highly differentiated tissues and complex development.

The genome sizes of most nematodes are of the order of 100 million base pairs (Mb) but range from 0·5 to 5 times this value (Sulston & Brenner, 1974; Sim *et al.* 1987; Rothstein, Stoller & Rajan, 1988; Hammond & Bianco, 1992; Grisi *et al.* 1995). Schistosome genomes are estimated to be 270 Mb.

https://doi.org/10.1017/S0031182099004060 Published online by Cambridge University Press

In comparison, the genome of *E. coli* is 4 Mb, *Theileria* spp. are 10 Mb, *Leishmania* is 35 Mb, and the human genome is 3000 Mb. The number of genes predicted for these parasites is similarly large. While *E. coli* has 3000 genes, and yeast 6000, the metazoan parasites are expected to have between 15000 and 20000 protein-coding genes. Humans are predicted to have about 100000 genes (Adams *et al.* 1995). Within the 15–20000 genes encoded by these parasites will be sets for basic metabolic activities, sets involved in building and maintaining the particular body plan of the organism, and sets involved in host interaction. The goal of the parasitic helminth genome initiatives is to identify the parasite-specific and host-interactive gene sets in as rapid and efficient manner as possible.

For most parasite genomes, the route chosen has been to sequence randomly selected cDNA clones to generate Expressed Sequence Tags (ESTs) (Reddy *et al.* 1993; Chakrabarti *et al.* 1994; El-Sayed *et al.* 1995; Franco *et al.* 1995; Wan, Blackwell & Ajioka, 1995; Blaxter *et al.* 1996; Levick *et al.* 1996; Brandao *et al.* 1997). This approach allows the sampling of the genes expressed by an organism (at a particular stage, or in a particular tissue). As a large proportion of genomic DNA is non-coding (either intergenic or intronic regions), EST sequencing is more efficient in terms of identifying genes (Adams *et al.* 1991). It has the drawback that each gene is represented in a cDNA library at approximately the abundance of its mRNA. This means the cDNAs from highly expressed genes (such as those encoding housekeeping enzymes, or cytoskeletal proteins) will be selected and sequenced repeatedly, while rare transcripts (such as those derived from genes controlling differentiation and development) will be selected rarely, if at all. For small genomes, such as those of *Plasmodium*, *Trypanosoma*, or *Leishmania*, where there are few or no introns and the genes are densely arrayed on the chromosomes, it is almost as efficient to sequence random genomic DNA fragments for gene discovery as it is to sequence cDNAs (El Sayed & Donelson, 1997). In a well constructed genomic DNA library each gene is represented in equimolar quantities, and has an equal chance of being sequenced. As an EST project progresses, the probability of identifying new genes drops as the sequence set grows.

Currently, there are over 27000 parasitic helminth ESTs in the public databases. These no doubt contain some nuggets of valuable ore: parasite enzymes with radically different active site environments, ripe for drug design, or highly expressed, novel, secreted proteins which might be part of a subunit vaccine. The exponential growth of this dataset poses significant problems for its thorough exploitation. When only a few sequences are available, it is possible to keep track of and analyse them all ‘manually’. With the volume and complexity of current nematode and schistosome datasets, it is clear that new tools have to be used. Sequence similarity search tools, such as BLAST (Altschul *et al.* 1990), are the commonest routes to identifying potential genes of interest in parasite datasets. These searches can be performed *de novo* using a sequence of interest from another species (such as a representative of a class of enzyme being sought in a parasite) or can be used at one remove through the intermediary of an annotated genome database. In a genome database, the sequences will be annotated (often using BLAST sequence similarity data) and can be searched using keywords.

USING THE EST DATASETS TO IDENTIFY POTENTIAL DRUG TARGETS AND VACCINE CANDIDATES

The growing EST datasets define a large number of genes. For the *Brugia* ESTs we estimate that the overall redundancy of the sequencing is 2 to 2·5, suggesting that we have identified approximately 7000 *Brugia* genes, or nearly one-half of the total expected gene complement. Other EST projects on parasites have generated sequence sets with similar redundancy. For *C. elegans*, the exhaustive 5′ and 3′ read EST project initiated by the Kohara laboratory has resulted in the generation of over 75000 ESTs. These, when compared to the genome sequence, appear to define about half of the genes of this model nematode.

There are both operational and informatics rationales for performing ongoing analyses of EST datasets. There is a diminishing return, in terms of new genes identified, as a single library is sequenced extensively. It is important therefore to assess the redundancy of sequences derived from each library in an ongoing fashion in order to maintain the gene discovery rate. The quality of any particular cDNA library can be measured by its primary titre (how many independent recombinant clones are there), its mean insert size (and the size range; this is related to the proportion of full-length transcripts in the library) and its redundancy (the mean representation of each gene). Different libraries can vary significantly in all these parameters, and in order to best exploit the limited resources available for parasite genomics, stringent quality checks are required. For the *Brugia* initiative, we have performed periodic redundancy estimates on the ESTs from each library. When the internal redundancy of a library dataset exceeds 3 (that is, only one new gene is discovered for each three ESTs generated), we reassess the utility of continued sequencing from the library. At this stage it may be cost effective to screen out the most abundant cDNAs (which may comprise up to 2% each of the ESTs, and over 40% of the total dataset) by hybridization, or to generate new libraries subtracted with cDNAs from other stages.

https://doi.org/10.1017/S0031182099004060 Published online by Cambridge University Press

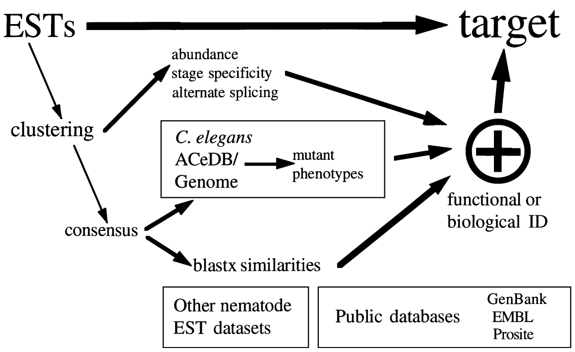

Fig. 1. Finding target genes in helminth genome sequence datasets.

Inter-library comparisons also serve to identify those libraries which have been derived from lifecycle stages (or tissues) where there is greatest diversity of gene expression. Further sequencing can focus on these libraries.

The size of the EST datasets (for example, 16000 sequences for *Brugia*) means that significant pre-analysis must be carried out if the genes defined by the ESTs are to be exploited properly. All the parasite genome projects are developing techniques for grouping ESTs which derive from the same gene into clusters, constructing consensus sequences from these clusters, and using the consensuses to infer biological function of the encoded proteins. In order to define a potential target gene (for drug development or vaccine testing) the clustered ESTs are subjected to a bioinformatic process summarized in Fig. 1. Utilizing all sources of information available to the research community, the sequences are compared to each other, to the public databases and to databases of motifs and patterns. The insights arising from these studies are integrated with biological information on the organism to identify first-pass candidates for further testing.

For the *Brugia* project, we have initiated an extensive analysis and annotation process which we hope will best serve the community wishing to exploit the genome information (Fig. 2) (Blaxter *et al*. 1996; Blaxter *et al*. 1997). The process of picking and sequencing clones has been streamlined, and in general is based on a microtitre plate format. High throughput sequencing is complemented by automated analysis of sequence read quality. The *Brugia* ESTs are deposited directly in the public databases. Further analysis is then performed on these public ESTs. First, they are compared to each other, and grouped into clusters on the basis of sequence identity. Because the ESTs are single-pass sequences, which may contain misreads or ambiguities, this clustering process has to be carefully monitored to ensure that low quality read segments and chimaeric sequences (from clones resulting from misligation of two cDNAs) do not result in the conflation of 2 genes into 1. The clusters are presumed to define genes, and consensus sequences are derived from each. The consensus sequences will tend to yield improved read lengths and read qualities for each gene. All this information (clone identity, stage, EST sequence, cluster information, consensus) are fed into the cognate genome project database (for *Brugia*, FilDB, based on ACeDB). The consensus sequence is then used to search the public databases. For *Brugia*, the primary comparator is the complete genome sequence of *C. elegans* and the sequences available from other nematodes. These databases are searched using public and local resources, using the BLAST family of algorithms. The output from BLAST is parsed into FilDB using tools developed for the *C. elegans* project, and the clusters annotated automatically. Within FilDB, genes can be examined for their expression patterns (which stage-specific libraries have ESTs been found in), levels of abundance (the number of ESTs), and putative function (BLAST similarity data).

About 40% of the genes we have identified in *Brugia* are novel, in that they have no detectable, informative similarity to other sequences. Many others identify only putative genes of unknown function in the *C. elegans* dataset. For these genes, which may include the new drug targets and unique vaccine components necessary for future studies, we perform additional analyses, looking for the presence of peptide motifs, and using more sensitive search strategies to try to define important features.

The database also integrates other information on genes and gene products in the form of bibliographic references and direct functional data arising from more conventional research programmes. The task of annotating the genome data is a huge one, and curation of the genome dataset is an important issue for future funding. A coordinated nomenclature system has been proposed for filaria (Blaxter *et al*. 1997), and other nematodes (Bird & Riddle, 1994), and similar naming schedules are in existence for other parasites.

For *Brugia*, we have the luxury of not only the *C. elegans* sequence, but also growing EST datasets from other filaria (*Onchocerca volvulus*, *Wuchereria bancrofti* and *Loa loa*) and other nematodes (*Strongyloides stercoralis* (Moore *et al*. 1996), *Pristionchus pacificus*, *Meloidogyne javanica*, *Globodera rostochiensis*, *Pristionchus pacificus*). Comparison between these datasets is helping to define specific and general targets for further study. Acquisition of EST datasets from other nematodes, particularly from groups not currently represented, will enhance and extend this approach (see below).

For filarial nematodes, as for other species, there is a wealth of information in the non-genomics literature which identifies classes of molecule which may have promise for pharmacological and immunological development. In particular, secreted or excreted products contain important enzymatic activities, and/or are protective in vaccination trials, in

https://doi.org/10.1017/S0031182099004060 Published online by Cambridge University Press

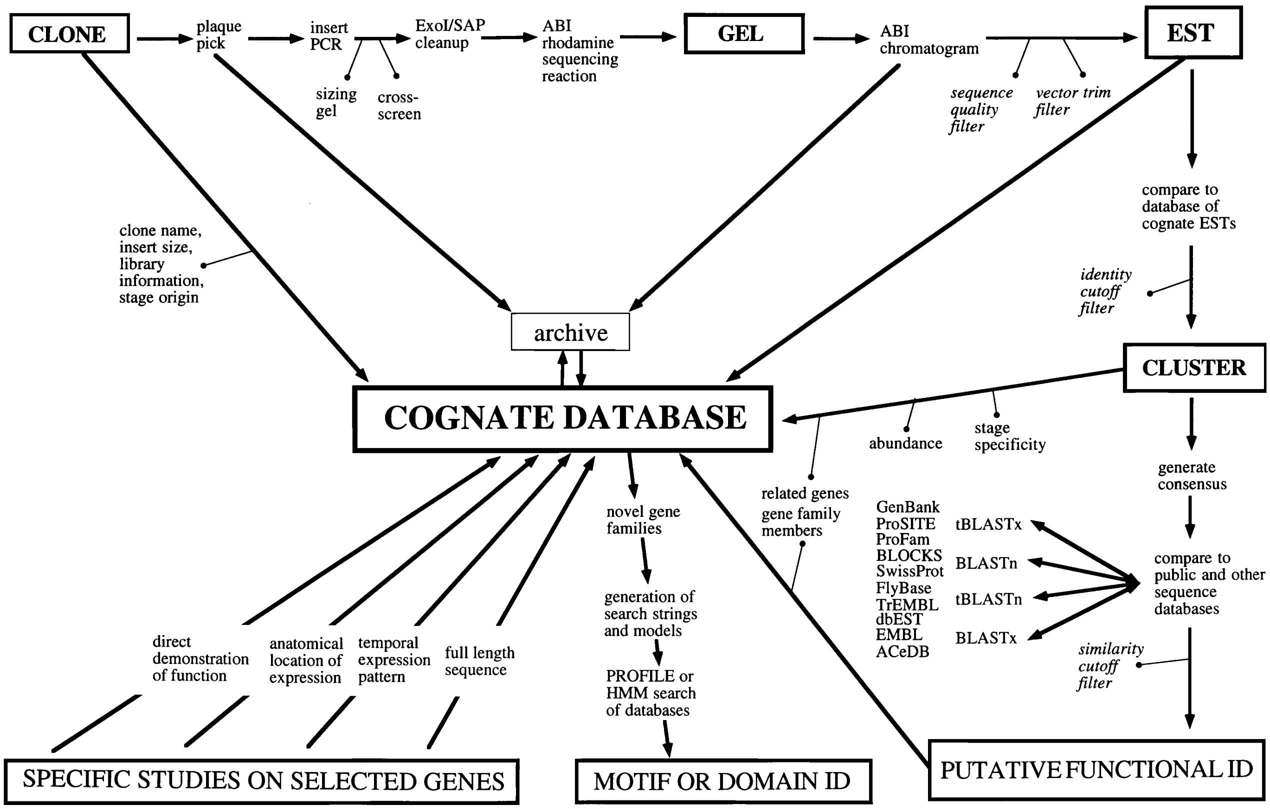

Fig. 2. Gene discovery in *Brugia malayi*: the annotation of genome sequence data.

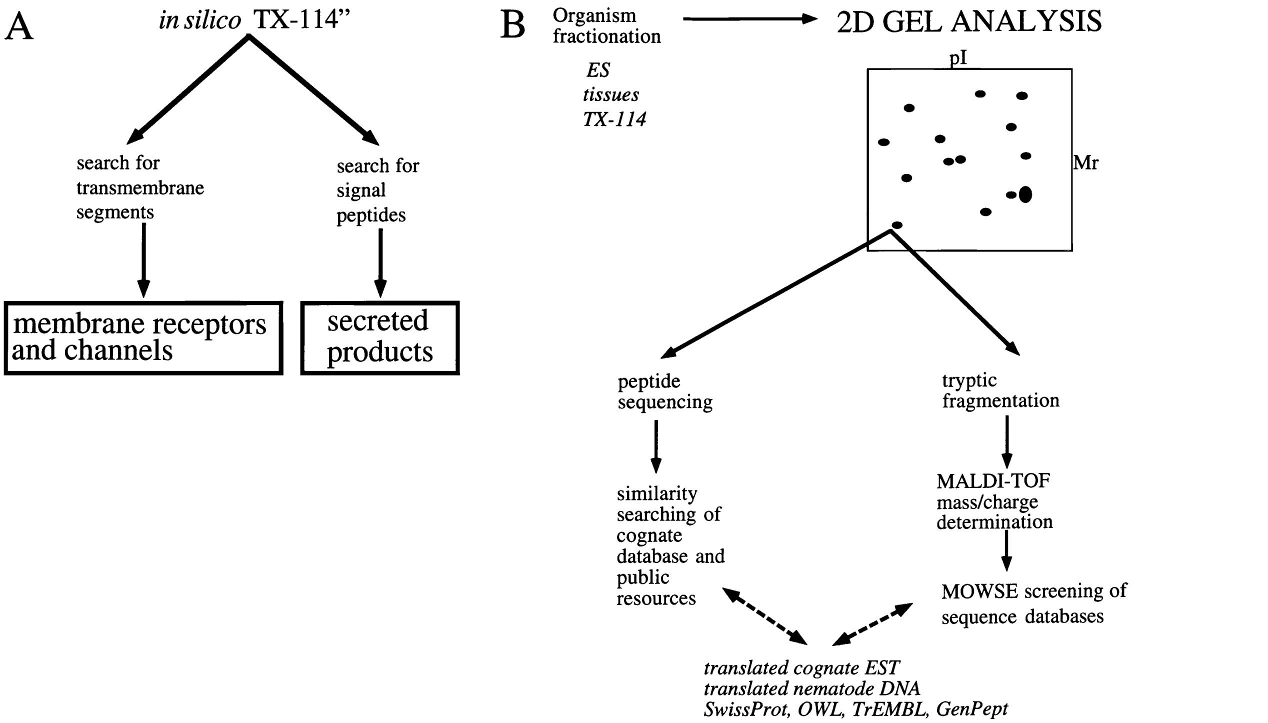
A
in silico TX-114"
search for transmembrane segments
search for signal peptides
membrane receptors and channels
secreted products
B
Organism fractionation
ES
tissues
TX-114
2D GEL ANALYSIS
pI
Mr
peptide sequencing
tryptic fragmentation
similarity searching of cognate database and public resources
MALDI-TOF mass/charge determination
MOWSE screening of sequence databases
translated cognate EST
translated nematode DNA
SwissProt, OWL, TrEMBL, GenPept

Fig. 3. (A) The *in silico* selection of membrane and secreted proteins. (B) Proteomics approaches to identifying parasite gene products.

many systems (Selkirk *et al.* 1994). It is possible, using the EST datasets, to identify putative secreted products without any prior knowledge of their function. The detergent Triton X114 has been used for many years to separate membrane proteins from soluble components *in vitro* (Etges, Bouvier & Bordier, 1986). A similar process can be carried out *in silico*, using the computer to search the EST datasets for proteins with predicted signal peptides (and thus destined for secretion) or with putative transmembrane domains (Fig. 3A) (von Heijne, 1985, 1986). Linking analysis of the proteins of a

https://doi.org/10.1017/S0031182099004060 Published online by Cambridge University Press

Table 1. Access to parasite genome project data and resources

Desired information or reagent	Access address/route
Parasite-genome www site	http://www.ebi.ac.uk/parasites/parasite-genome.html
Access to parasite genome www sites	http://www.ebi.ac.uk/parasites/paratable.html
Parasite-genome computing resources	http://www.ebi.ac.uk/parasites/genomecompute.html
The *Caenorhabditis elegans* genome project	http://www.sanger.ac.uk/projects/C_elegans/ ● the *C. elegans* ACeDB database is available online at http://www.sanger.ac.uk/Projects/C_elegans/webace_front_end.shtml
Sequence similarity search of parasite DNA sequences	● through the Parasite Genome world wide web blast server at http://www.ebi.ac.uk/parasites/parasite_blast_server.html ● through email to blast@ncbi.nlm.nih.gov ● through the NCBI www server at http://www.ncbi.nlm.nih.gov/BLAST/
Text search of EST sequence records	● through the NCBI dbest www server at http://www2.ncbi.nlm.nih.gov/dbST/dbest_query.html ● through the NCBI ENTREZ www server at http://www3.ncbi.nlm.nih.gov/Entrez/index.html
Retrieval of EST sequences	● through the NCBI retrieve email server at retrieve@ncbi.nlm.nih.gov ● through the NCBI ENTREZ www server at http://www3.ncbi.nlm.nih.gov/Entrez/index.html

parasite (the proteome) with the genome data is possible through N-terminal sequencing, or mass spectrographic mass/charge ratio determination of protease digests, of single protein spots from parasite products separated on high resolution two dimensional gels (Fig. 3B) (Pappin, Hojrup & Bleasby, 1993). N-terminal sequence data, or the predicted amino acid composition derived from mass/charge data, can be used to search parasite-specific or general sequence databases.

THE WHO PARASITE GENOMES RESOURCE CENTRE

We have been involved in trying to develop and implement such tools for the Parasite Genome projects sponsored by the World Health Organisation. We aim to assist the project database curators in their tasks by developing and installing analysis tools, and to promote the Parasite Genome projects to the wider community, by providing internet access to the data (Blaxter & Aslett, 1997).

The WHO Parasite Genome world wide web (WWW) site, based at the European Bioinformatics Institute, offers links to the individual Parasite Genome project WWW sites (some of which are based on the EBI server) and access to the Parasite Genome BLAST server. The WWW site includes information on genome computing resources available on the WWW (Table 1).

The Parasite Genome BLAST server is a public-access resource which allows the searching of a number of parasite databases with a user-supplied sequence. Currently the server will search against 16 different DNA databases culled from the public GenBank/EMBL database (*Brugia malayi* DNA, *Onchocerca volvulus*, all filarial nematodes, all nematodes other than *C. elegans*, African trypanosomes, *Trypanosoma cruzi*, *Leishmania major*, *Leishmania* spp., all kinetoplastid protozoa, *Schistosoma mansoni*, all other *Schistosoma* spp., *Toxoplasma gondii*, *Cryptosporidium*, *Plasmodium falciparum*, all other *Plasmodium* spp. and all apicomplexan DNA). The databases include all cDNAs (including ESTs) and genomic DNAs from each organism or group of organisms, and results are returned, in standard BLAST output format, by E-mail to the user. This server can be used to identify distant parasite homologues of genes of interest, when searching the full public databases would yield a bewildering array of hits to sequences from other organisms, with the parasite gene languishing at the end of a long list. All the databases used are updated regularly, and are available for download by anonymous file transfer from the parasite genome site for use in local search routines.

Each of the WHO-sponsored parasite genome projects has constructed a genome database, using the *C. elegans* database engine software, ACeDB (Thierry-Mieg & Durbin, 1992; Durbin & Thierry-Mieg, 1994). ACeDB is very powerful and is being used extensively for many genome initiatives in addition to that of *C. elegans*. ACeDB-WWW interfaces have been developed, and implementation of these for the parasite genome databases is planned. These databases allow the integration of genetic map, physical map, sequence, bibliographical and biological information in a single environment. The parasite-genome support centre also performs batch BLAST searches for the parasite genome databases, and provides a service to update these databases. Individual projects also use the Parasite Genome site to distribute additional datasets, such as the filariasis bibliography, Bibliofil.

Genome computing is a universally applicable science: it needs only a dedicated researcher and a reasonably fast computer with an internet link. To promote parasite genome bioinformatics, the WHO Parasite Genome support has visited endemic country laboratories to assist with computing and informatics issues, and is sponsoring international workshops in parasite genome bioinformatics.

THE *CAENORHABDITIS ELEGANS* GENOME

C. elegans is a small freeliving bacteriovorous nematode. Its only real-world significances are that it can be a pest in mushroom farms, and it may play a role in soil ecology. However, its significance to parasitic nematology is immense (Politz & Philip, 1992; Burglin, Lobos & Blaxter, 1998). As a nematode it carries out all the basic functions required by the nematode body plan. It has a nematode metabolism, and is sensitive to many nematicides. Its development, anatomy and neurobiology are understood at a single cell level, and over 2000 loci have been defined by mutational genetics (Riddle *et al*. 1997). A toolkit of methods (from *in situ* hybridization to transgenesis to laser ablation of individual cells) has been developed (Epstein & Shakes, 1996). In addition, the *C. elegans* genome has been sequenced in its entirety (Coulson *et al*. 1988; Sulston *et al*. 1992; Wilson *et al*. 1994; Hodgkin, Plasterk & Waterston, 1995; Waterston, Sulston & Coulson, 1997; The *C. elegans* Sequencing Consortium 1998).

This important milestone in genome analysis (it is the first sequenced animal genome) was achieved by two teams, based in St. Louis, USA and Cambridge, UK. A physical map of the genome was first built using cosmid (insert size 35 kb) and yeast artificial chromosome (YAC; insert size > 150 kb to 1 Mb) clones. This map contains over 99 % of the genome ordered with respect to the chromosomes, and served as a substrate for the sequencing effort (Coulson *et al*. 1988). The map contains 17000 fingerprinted cosmids and 3000 YACs, which have been linked to the cosmid contigs by hybridization. The few 'gaps' in the map appear to arise from there being regions of DNA which do not clone efficiently in either yeast or bacterial vectors, and have been closed as the sequencing project has progressed. The genome is 100 million base pairs, and is arranged as 5 autosomes and a sex chromosome (sex determination is through an XX–XO mechanism).

The sequence was derived from a minimally overlapping set of cosmids spanning the genome, augmented by YAC clones where there were gaps in cosmid coverage (Sulston *et al*. 1992; Wilson *et al*. 1994; Waterston *et al*. 1997). Remaining gaps have been filled by combinations of long-range PCR, direct sequencing and cloning in other vector systems (including fosmids). The sequence error rate is estimated to be 1 in 10000 bases. The sequence is extensively annotated, with predicted genes, repetitive DNA and other features being added to the records before full submission to the databases.

There are predicted to be 18000 protein coding genes in *C. elegans* and about 1000 RNA genes (tRNAs and the like) (Hodgkin & Herman, 1998). Gene prediction is based on algorithms trained to recognize features of the *C. elegans* genome (such as splice sites and codon bias), and uses EST data extensively. Prediction is not yet perfect, and is being continually refined as specific data accumulates. Protein coding genes are almost always interrupted by introns, but these are generally quite small (down to 37 bases) and thus gene density remains high (about one gene per 6 kb). The predicted genes include many which are easily recognized as homologues of known genes in other organisms (such as housekeeping enzymes) but there is a large class (40 % of all genes) for which no obvious homologues can be found. The genome is littered with the remains of dead and dying transposons and there are several other repeat families. One striking feature of gene organisation in *C. elegans* is that many genes (20 % of the total) appear to be arranged as operons, where a single promoter drives transcription of two or more genes (Speith *et al*. 1993; Blumenthal & Steward, 1997). The downstream genes in operons are trans-spliced to a family of variant spliced leader exons. The significance of the operonic arrangement of genes in *C. elegans* is not yet clear, as there is often little functional or sequence similarity between operon partners.

The complete sequence has been deposited in GenBank/EMBL, and is freely available. The fully annotated sequence is available within ACeDB (see Table 1).

WALKING BETWEEN GENOMES

With limited resources, it is not going to be possible to determine the complete genome sequence of all disease organisms, particularly if they have large genomes. However, with 1 or 2 exemplars to hand it may be possible to utilize partial genomic information to walk between genomes and focus in on genes of interest without building another complete sequence map (Fig. 4). The genes which are targets for new drugs and those responsible for genetic resistance to old ones can be identified in 1 model species, and the information gathered from these studies used to search other genomes for homologous sequences. A successful vaccine candidate in 1 species can be sought in a second.

In order to transfer between species, it is useful to have a way-map of the expected distances between them. The development of molecular phylogenies

https://doi.org/10.1017/S0031182099004060 Published online by Cambridge University Press

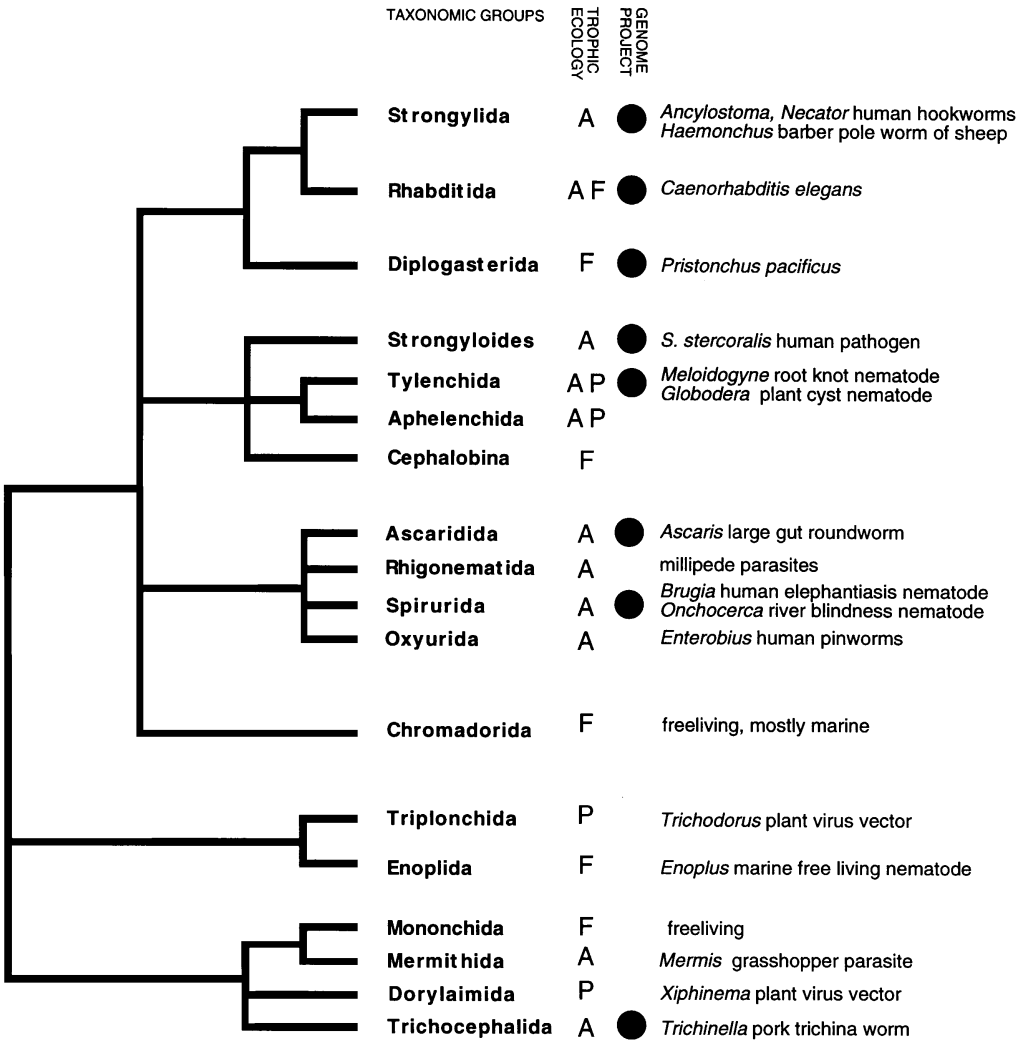
TAXONOMIC GROUPS
TROPHIC ECOLOGY
GENOME PROJECT
Strongylida A ● Ancylostoma, Necator human hookworms
Haemonchus barber pole worm of sheep
Rhabditida A F ● Caenorhabditis elegans
Diplogasterida F ● Pristonchus pacificus
Strongyloides A ● S. stercoralis human pathogen
Tylenchida A P ● Meloidogyne root knot nematode
Globodera plant cyst nematode
Aphelenchida A P
Cephalobina F
Ascaridida A ● Ascaris large gut roundworm
Rhigonematida A millipede parasites
Spirurida A ● Brugia human elephantiasis nematode
Onchocerca river blindness nematode
Oxyurida A Enterobius human pinworms
Chromadorida F freeliving, mostly marine
Triplonchida P Trichodorus plant virus vector
Enoplida F Enoplus marine free living nematode
Mononchida F freeliving
Mermithida A Mermis grasshopper parasite
Dorylaimida P Xiphinema plant virus vector
Trichocephalida A ● Trichinella pork trichina worm

Fig. 4. The interrelationships of the major groups of nematodes. This phylogenetic tree is a cartoon based on analysis of small subunit ribosomal RNA sequences from a large number of nematode taxa. Taxa for which a genome or EST project is underway are marked ●. The trophic ecology of each taxon is indicated with a letter: F free living, A animal parasite and P plant parasite. *C. elegans* is closely related to the strongylid nematodes. The ascarids, spirurids and oxyurids are all closely related. *Strongyloides* is most closely related to the cephalobid free living nematodes, and to the plant parasitic tylenchids. *Trichinella* and *Trichuris* are only distantly related to *C. elegans* and are members of a group which includes insect parasites (the mermithids) and plant parasites (the dorylaims) as well as free living nematodes.

for schistosomes (Rollinson *et al.* 1997) and other platyhelminths, and for nematodes (Blaxter *et al.* 1998), makes the selection of stepping stones in traversing phylogenetic diversity more easy. Within the nematodes, molecular analyses highlight the multiple independent origins of plant and animal parasitism. In order to fully exploit the *C. elegans* and other genome initiatives we would suggest that it would be sensible, and very cost effective, to acquire significant EST datasets from across the phylum, picking one or two species from each major clade for analysis (Fig. 4). As a pilot project, we have generated 150–220 ESTs from a set of animal parasitic nematodes representing the Strongylida, the Ascaridida and the Triocephalida (Fig. 5). These EST datasets, while small, amply demonstrate the utility of the approach. The ESTs from *Ascaris suum* are derived from a muscle/body wall library and contain a high proportion of highly expressed hypodermal and muscle genes with *C. elegans* homologues, consistent with the great deal known about these tissues. The proportion of novel genes identified was relatively small (16%). In contrast, 149 ESTs from adult *Trichuris muris*, a gut parasite only distantly related to *C. elegans*, identified 139 genes, 51% of which had no informative similarity to any sequence in the databases. For *Necator americanus*, the ESTs (from an adult library) define 166 different genes, including many abundant products predicted to be secreted. ESTs from larval *Toxocara canis* have also been used to define putative components of the secretory material (K. Tetteh, A. Loukas and R. Maizels, personal communication).

https://doi.org/10.1017/S0031182099004060 Published online by Cambridge University Press

A

Library	*Necator*	*Trichuris*	*Ascaris*
Total ESTs	205	149	141
clusters containing >1 EST	21	8	22
clusters containing only one EST	145	131	65
Individual genes identified	166	139	87
Redundancy of the dataset	1.25	1.072	1.62

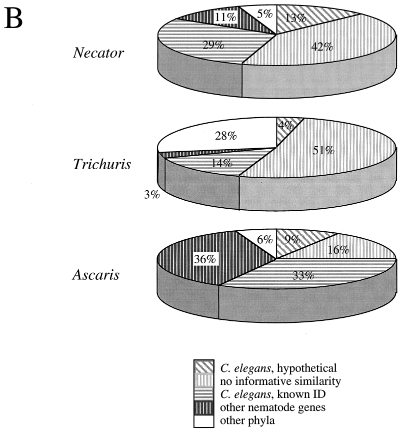

Fig. 5. Small EST datasets from selected nematode parasites. (A) The number of sequences, the redundancy and the number of genes identified in three EST projects from *Necator americanus*, *Trichuris muris* and *Ascaris suum*. (B) Pie charts showing the pattern of identification of genes by BLAST search in the three EST datasets.

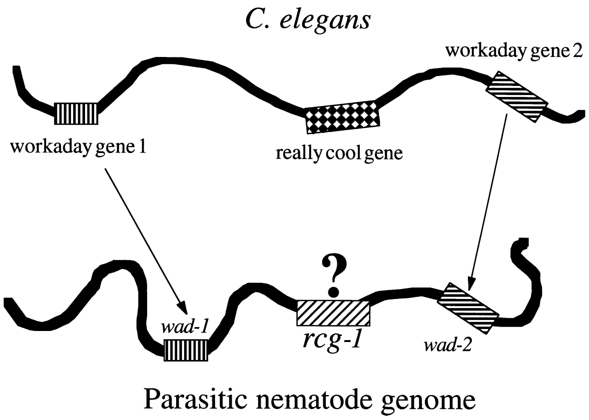

Fig. 6. Walking between genomes. The demonstration of a close syntenic relationship between a conserved workaday gene and a gene of interest in one species of nematode (*C. elegans* for example) may permit the cloning of the gene of interest in a parasite by virtue of the conservation of gene arrangement.

These EST datasets can be used to promote research on the parasites, by identifying candidate genes for further study. They are also useful for the *C. elegans* project, as they can be used to confirm genes which have been predicted from the genome sequence. For example, 13% of the *N. americanus* ESTs are clear homologues of *C. elegans* genes with no known function. In the EST datasets we also have the first examples outside *C. elegans* of genes defined to have important roles in the biology of the model nematode. For example, the *Brugia* dataset includes homologues of several different UNC genes (UNC for uncoordinated: these genes affect neuromuscular function) and of genes involved in the sex determination pathway (HER-1, TRA-1; I. Kamal and D. Guiliano, unpublished observations). These genes can be used to further refine models of gene function in *C. elegans* and may also yield insights into the biology of the parasites.

CLONING BY SYNTENY

Comparison between mouse and human genomes reveals that linkages between genes and gene order are often conserved. Some conservation of synteny is also evident when mammalian and avian genomes are compared. While we have no time axis for the nematode radiation, it might be expected that some conservation of synteny could be found between different nematode groups. The significance of this is that it might offer a route to cloning and analysis of genes too diverged to be identified by low stringency hybridization, degenerate PCR or EST database searching (Fig 6.). For example, the sex determination gene TRA-2 is poorly conserved between *C. elegans* and the congeneric *C. briggsae*, but was cloned from *C. briggsae* by isolating genomic clones which carried homologues of conserved genes found next to TRA-2 in *C. elegans* (Kuwabara & Shah, 1994). Cloning by synteny may be a fruitful approach to isolating parasite genes of interest now that the *C. elegans* genome is completely sequenced, and all synteny relationships known. Comparison of *C. elegans* and *C. briggsae* sequences also serve to identify highly conserved promoter regions upstream of genes (Heschl & Baillie, 1990; Gilleard, Barry & Johnstone, 1997), and this approach may be extensible to parasite genes.

We would predict, from the molecular phylogeny, that the strongylid nematodes would be most likely to have retained synteny relationships with *C. elegans*, and that the other animal parasites would be more or less rearranged. Of particular interest is the conservation of operonic organization. While the conservation of 1 operon (involving 2 ribosomal proteins, RPP-1 and RPL-27a) has been demonstrated between genera (Evans *et al.* 1997), no operons have yet been found in nematodes distantly related to *C. elegans*, and their biological significance remains unclear.

MAPPING THE NUCLEAR GENOME OF *BRUGIA MALAYI*

The *Brugia malayi* genome is 100 million base pairs (Sim *et al.* 1987), has an AT content of 71% (Rothstein *et al.* 1988), and is organized as six chromosomes (5 autosomes and a XY sex determination pair) (Sakaguchi *et al.* 1983). The chromosomes cannot be separated with current pulsed field gel technology, and are probably each > 12 million base pairs in size (Sim *et al.* 1987). No genes have yet

https://doi.org/10.1017/S0031182099004060 Published online by Cambridge University Press

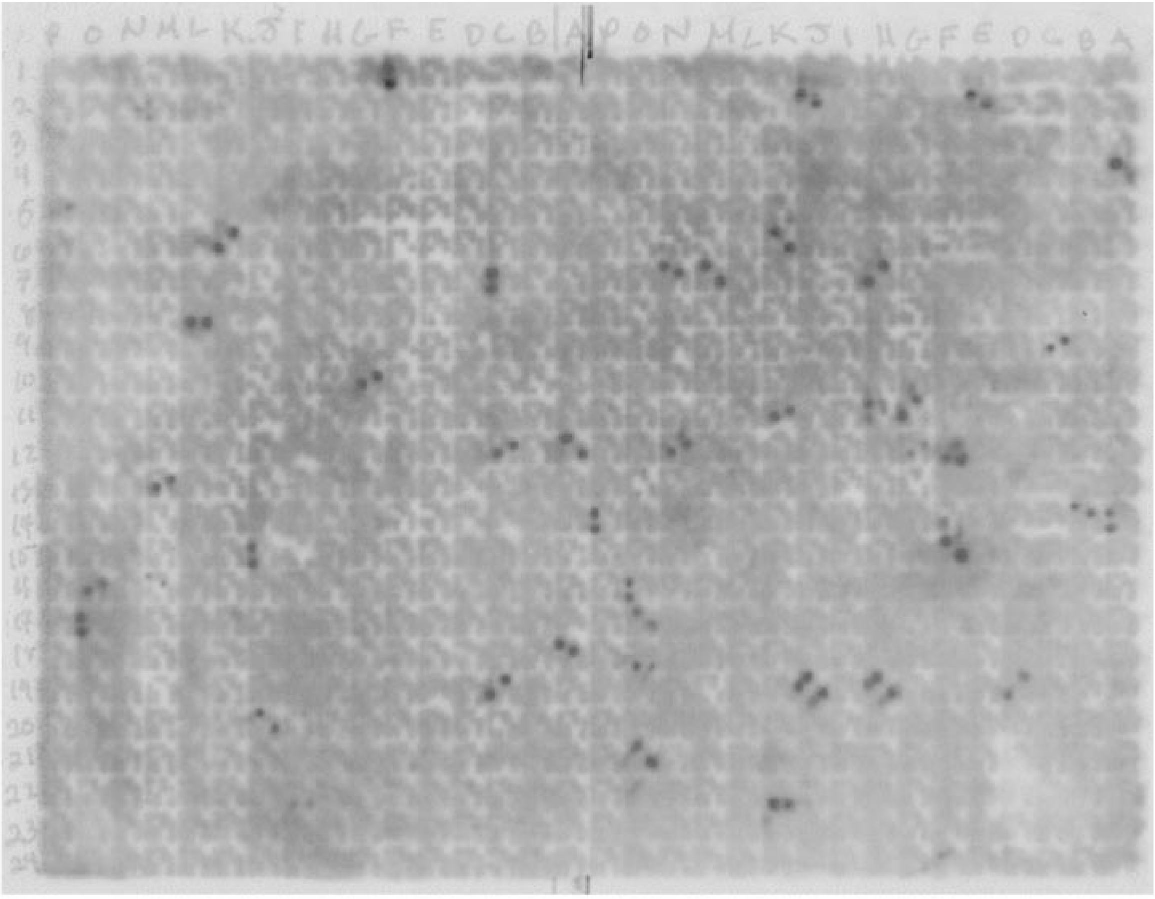
P O N M L K J I H G F E D C B A
A P O N M L K J I H G F E D C B A
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24

Fig. 7. Mapping genes to the *Brugia* BAC library. A probe derived from the *Brugia malayi* small subunit ribosomal RNA gene was labelled and hybridized to the *Brugia* BAC library filter. The hybridization was visualized using chemiluminescence. The ribosomal RNA repeat comprizes approximately 1% of the genome, and 46 clones out of 4600 gridded indicates that approximately 1% of the library contains ribosomal RNA genes. Each clone is spotted twice on the filter.

been mapped to *Brugia* chromosomes by FISH or other *in situ* mapping techniques, although these techniques have been successfully applied to *Schistosoma mansoni* (Tanaka *et al*. 1995). The repetitive DNA content, at approximately 15%, is similar to that of *C. elegans*, but the *Brugia* genomes differ in that nearly 10% of the genome is made up of a single, tandemly repeated sequence, the *Hha*I repeat. This repeat has been used as a diagnostic PCR target because of its high copy number (30000 copies per genome) (Piessens, McReynolds & Williams, 1987; Williams *et al*. 1987). *Brugia* also carries 2 other genomes: the mitochondrial genome and the genome of an endosymbiotic bacterium.

In order to build a physical map of *Brugia* a set of large insert DNA libraries is being constructed (Blaxter, 1995; Blaxter *et al*. 1997). A bacterial artificial chromosome (BAC) library, with inserts of 60–100 kb, was constructed from microfilarial DNA. The library has 4600 clones and represents an approximately three-fold coverage of the nuclear genome. It has been picked and gridded as high-density arrays, and filtermats printed with these arrays are being used as substrates in a 3-pronged approach to mapping the genome. Additional libraries are being constructed to complement this one, including additional BAC libraries, and a large-insert yeast artificial chromosome library.

Each genome project laboratory is hybridizing selected EST clusters and genes of interest to the BAC library filters (Fig. 7). The ESTs are being chosen on the basis of abundance, stage specificity and interest. Probes are generated by PCR from the clones and labelled with a non-radioactive tag. Positive hybridizations are detected using an avidin-enzyme conjugate and luminescent substrates. The mapping of ESTs to the filters is consistent with the expected size of the library (approximately 40% of hybridizations are negative) and over 100 genes have been mapped in this way by the Filarial Genome Project participating laboratories.

A second approach being followed is a random sampling without replacement strategy (Palazzolo *et al*. 1991; Hoheisel *et al*. 1993; Mizukami *et al*. 1993) utilizing end probes generated by PCR from the junctions between *Brugia* inserts and the BAC vector. These BAC ends are labelled and hybridized to the filters. Positive clones are recorded, and

https://doi.org/10.1017/S0031182099004060 Published online by Cambridge University Press

further non-hybridising clones selected for the next round. The BAC ends are also sequenced to provide information for later development of PCR-based markers. The 2 sets of hybridization data (EST and end probes) are being integrated in the generation of a sequence-tagged site map. This process is ongoing, but the first generation map (with > 90% of the 46 000 BAC clones tagged) is expected to be complete by summer 1999.

A third approach is directed chromosome walking from selected genes of interest. This labour-intensive approach is being followed for certain genomic regions where there is special interest in conservation of gene order or proximity, such as the homeobox gene cluster (A. Aboobaker and M. Blaxter, unpublished).

The mitochondrial genome is expected to comprise about 14 kb, like those of *Onchocerca volvulus* (Keddie *et al*. 1998), *C*. *elegans* and *A*. *suum* (Okimoto *et al*. 1992). About 65% of the mitochondrial genome is represented in the *B*. *malayi* EST dataset, and this information is being used to clone and sequence the complete mitochondrial genome by direct PCR (M. Blaxter, unpublished).

The presence of an endosymbiont within *B*. *malayi*, and other filaria, has been noted for many years (McLaren, 1972; McLaren *et al*. 1975), but has become significant as the genome initiative has progressed. The endosymbiont, being a eubacterium, has a metabolism distinct from its nematode host, and is thus a promising drug target (Henkle-Dührsen *et al*. 1998). It remains to be demonstrated unequivocally that the symbiosis is mutualistic, but tetracycline treatment does reduce filarial infectivity and severely blocks fecundity in infected rodent models (Bosshardt *et al*. 1993; Hoerauf *et al*. 1998). The endosymbiont is closely related to the *Wolbachia* endosymbionts of insects and other arthropods, and appears to be maintained by transovarial transmission (Sironi *et al*. 1995; Bandi *et al*. 1998). There is no evidence for recent horizontal spread of the endosymbiont through the filaria: rather the endosymbiont and nematode host phylogenies are mainly congruent, suggesting an ancient and stable vertical transmission. This is in stark contrast to the situation in insects, where *Wolbachia* has spread recently as a horizontally transmitted epidemic (O'Neill, 1995; Werren, Zhang & Guo, 1995; Werren, 1997). A small number of the *Brugia* ESTs appear to derive from endosymbiont genes (including groEL, 16S and 23S ribosomal RNAs). Of greater concern is the possibility that the endosymbiont genome, through its lower AT content (~ 50%) will be preferentially cloned in the bacterial systems used for maintaining the *Brugia* genomic mapping libraries. The size of the endosymbiont genome is unknown, but is likely to be of the order of 1–2 million bases. There are multiple endosymbionts per cell, particularly in the hypodermis and female gonad, and thus the proportion of endosymbiont DNA to nuclear DNA may be high. Preliminary screening of the *Brugia* BAC library suggests that it may comprise between 2 and 12% endosymbiont DNA. A map of the *Wolbachia* genome is being constructed by a walking strategy.

SYNTENY CONSERVATION BETWEEN *BRUGIA* AND *C. ELEGANS*

We have begun to use the *Brugia* genomics resources to investigate issues of operonic organization and synteny conservation. A survey of ribosomal protein genes in *C*. *elegans* revealed that a large proportion of these (> 50%) are members of operons. This proportion is much greater than that found for the whole genome (~ 20%). In addition, 80% of the ribosomal genes which are in operons are the first gene in the operon. One of the first results of the *Brugia* EST programme was the identification of ESTs coding for most of the ribosomal proteins (Blaxter *et al*. 1996). Examination of the ESTs also identified several corresponding to the *C*. *elegans* operon partners for these ribosomal protein genes. We reasoned that these ESTs could be used as probes to try to identify ribosomal protein gene-containing operons in *Brugia*. However, we have as yet been unable to identify any conserved operons involving these ribosomal protein genes in *Brugia*. Analysis of the genomic organization of a number of other genes of interest, whose homologues are in operons in *C*. *elegans*, has similarly failed to yield an operon (D. Guiliano and M. Blaxter, unpublished observations). As we believe that the *Brugia* genome will be as gene-dense as that of *C*. *elegans*, and the operonic organization is argued to arise in part from a need to crowd genes into the chromosomes, we are continuing this search.

Brugia adults secrete a small protein, MIF-1, with significant similarity to mammalian macrophage migration inhibition factor, a cytokine with a regulatory role in recruitment of cells in the immune response. This gene was identified by the EST programme, and has subsequently been studied in some detail, as it may play an important role in the modulation of the host immune response by *Brugia* (Pastrana *et al*. 1998). *C*. *elegans* has two MIF homologues. These are closely related to each other and to a second *Brugia* MIF (MIF-2). A MIF-related gene in mice has dopachrome tautomerase function and the two *C*. *elegans* MIF genes and *Brugia* MIF-2 are more closely related to this enzyme, while MIF-1 is more closely related to the mammalian cytokine MIF. A BAC carrying the genomic copy of the MIF-1 gene was isolated and sequenced. Comparison of this 65 kb sequence with the *C*. *elegans* genome revealed that of the 7 identifiable genes, 6 had *C*. *elegans* homologues which were located close to each other on chromosome I. Two of the *Brugia* genes have the same head-

https://doi.org/10.1017/S0031182099004060 Published online by Cambridge University Press

to-head organization as their *C. elegans* counterparts. There are no obvious operons. The conservation of synteny is not complete, as there are large genomic regions not present in the *Brugia* BAC sequence, and the BAC contains one gene (for which there is a *Brugia* EST) which has no *C. elegans* counterpart. This surprizing conservation of synteny suggests that the cloning by synteny approach may indeed be applicable, and that large scale sequence analysis of the *Brugia* genome may reveal patterns and processes in genome evolution not observable in the closer comparison of *C. elegans* and *C. briggsae*.

ACKNOWLEDGEMENTS

We would like to thank our colleagues in the Filarial Genome Project for sharing data prior to publication, and for making the process of sequencing and mapping a true community effort. Steven Jones and Laurent Baron of the Sanger Centre were instrumental in the progress of the Edinburgh EST effort. We would like to thank Boris Dobrokhotov of the WHO TDR and Mike Ashburner and Graham Cameron of the EBI for their support. David Pritchard, Richard Grencis and Tim Geary provided nematode cDNA libraries. Mark Blaxter is funded by the Darwin Trust, Jennifer Daub and David Guiliano are funded by the MRC and Martin Aslett is funded by the World Health Organisation/World Bank/UNDP TDR. The Filarial Genome Project is also supported by the WHO TDR and New England Biolabs.

REFERENCES

ADAMS, M. D. et al. (1995). Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377**, 3–174.

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

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

BANDI, C., ANDERSON, T. J. C., GENCHI, C. & BLAXTER, M. L. (1998). Phylogeny of *Wolbachia*-like bacteria in filarial nematodes **265**, 2407–2413.

BIRD, D. M. & RIDDLE, D. L. (1994). A genetic nomenclature for parasitic nematodes. *Journal of Nematology* **26**, 138–143.

BLAXTER, M. L. (1995). The Filarial Genome Project. *Parasitology Today* **11**, 811–812.

BLAXTER, M. L. & ASLETT, M. A. (1997). Internet resources for the parasite genome projects. *Trends in Genetics* **13**, 40–41.

BLAXTER, M. L., DAUB, J., WATERFALL, M., GUILIANO, D., WILLIAMS, S., JAYARAMAN, K., RAMZY, R., SLATKO, B. & SCOTT, A. (1997). The Filarial Genome Project. *COST 819*. Brussels, The Commission of the European Community.

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

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

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

BLUMENTHAL, T. & STEWARD, K. (1997). RNA processing and gene structure. *C. elegans II*. Cold Spring Harbor, NY, Cold Spring Harbor Laboratory Press.

BOSSHARDT, S. C., MCCALL, J. W., COLEMAN, S. U., JONES, K. L., PETIT, T. A. & KLEI, T. R. (1993). Prophylactic activity of tetracycline against *Brugia pahangi* infection in jirds (*Meriones unguiculatus*). *Journal of Parasitology* **79**, 775–777.

BRANDAO, A., URMENYI, T., RONDINELLI, E., GONZALEZ, A., DE MIRANDA, A. B. & DEGRAVE, W. (1997). Identification of transcribed sequences in the *Trypanosoma cruzi* genome project. *Memorias do Instituto Oswaldo Cruz* **92**, 863–866.

BURGLIN, T., LOBOS, E. & BLAXTER, M. L. (1998). *Caenorhabditis elegans* as a model for parasitic nematodes. *International Journal for Parasitology* **28**, 395–411.

CHAKRABARTI, D., REDDY, G. R., DAME, J. B., ALMIRA, E. C., LAIPIS, P. J., FERL, R. J., YANG, T. P., ROWE, T. C. & SCHUSTER, S. M. (1994). Analysis of expressed sequence tags from *Plasmodium falciparum*. *Molecular and Biochemical Parasitology* **66**, 97–104.

COULSON, A., WATERSTON, R., KIFF, J., SULSTON, J. & KOHARA, Y. (1988). Genome linking with yeast artificial chromosomes. *Nature* **335**, 184–186.

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

EL-SAYED, N. M. A., ALARCON, C. M., BECK, J. C., SHEFFIELD, V. C. & DONELSON, J. E. (1995). cDNA expressed sequence tags of *Trypanosoma brucei rhodesiense* provide new insights into the biology of the parasite. *Molecular and Biochemical Parasitology* **73**, 75–90.

EL SAYED, N. M. & DONELSON, J. E. (1997). A survey of the *Trypanosoma brucei rhodesiense* genome using shotgun sequencing. *Molecular and Biochemical Parasitology* **84**, 167–178.

EPSTEIN, H. F. & SHAKES, D. C. (1996). Caenorhabditis elegans: *Modern Biological Analysis of an Organism*. San Diego, CA, Academic Press.

ETGES, R., BOUVIER, J. & BORDIER, C. (1986). The major surface protein of *Leishmania* promastigotes is a protease. *Journal of Biological Chemistry* **261**, 9098–9101.

EVANS, D., ZORIO, D., MCMORRIS, M., WINTER, C. E., LEA, K. & BLUMENTHAL, T. (1997). Operons and SL2 trans-splicing exist in nematodes outside the genus *Caenorhabditis*. *Proceedings of the National Academy of Sciences, USA* **94**, 9751–9756.

https://doi.org/10.1017/S0031182099004060 Published online by Cambridge University Press

FRANCO, G. R., ADAMS, M. D., SOARES, M. B., SIMPSON, A. J., VENTER, J. C. & PENA, S. D. (1995). Identification of new *Schistosoma mansoni* genes by the EST strategy using a directional cDNA library. *Gene* **152**, 141–147.

GILLEARD, J. S., BARRY, J. D. & JOHNSTONE, I. L. (1997). *cis* Regulatory requirements for hypodermal cell-specific expression of the *Caenorhabditis elegans* cuticle collagen gene *dpy-7*. *Molecular and Cellular Biology* **17**, 2301–2311.

GRISI, E., BURROWS, P. R., PERRY, R. N. & HOMINICK, W. M. (1995). The genome size and chromosome complement of the potato cyst nematode. *Globodera pallida*. *Fundamental and Applied Nematology* **18**, 67–70.

HAMMOND, M. P. & BIANCO, A. E. (1992). Genes and genomes of parasitic nematodes. *Parasitology Today* **8**, 299–305.

HENKLE-DÜHRSEN, K., ECKELT, V. O., WILDENBURG, G., BLAXTER, M. & WALTER, R. D. (1998). Molecular characterisation of a catalase from intracelluaIr bacteris in *Onchocerca volvulus*. *Molecular and Biochemical Parasitology* **96**, 69–81.

HESCHL, M. F. P. & BAILLIE, D. L. (1990). Functional elements and domains inferred from sequence comparisons of a heat shock gene in two nematodes. *Journal of Molecular Evolution* **31**, 3–9.

HODGKIN, J. & HERMAN, R. K. (1998). Changing styles in *C. elegans* genetics. *Trends in Genetics* **14**, 352–357.

HODGKIN, J., PLASTERK, R. H. A. & WATERSTON, R. H. (1995). The nematode *Caenorhabditis elegans* and its genome. *Science* **270**, 410–414.

HOERAUF, A., NISSEN-PÄHLE, K., SCHMETZ, C., HENKLE-DÜHRSEN, K., BLAXTER, M. L., BÜTTNER, D., AL-QUAOUD, K. M., LUCIUS, R. & FLEISCHER, B. (1999). Endosymbiotic bacteria in the filarial nematode *Litomosoides sigmodontis* are targets for tetracycline therapy which results in filarial immunity. *Journal of Clinical Investigation* **103**, 11–18.

HOHEISEL, J. D., MAIER, E., MOTT, R., McCARTHY, L., GRIGORIEV, A. V., SCHALKWYK, L. C., NIZETIC, D., FRANCIS, F. & LEHRACH, H. (1993). High resolution cosmid and P1 maps spanning the 14 Mb genome of the fission yeast *S. pombe*. *Cell* 109–120.

KEDDIE, E. M., HIGAZI, T. & UNNASCH, T. R. (1998). The mitochondrial genome of *Onchocerca volvulus*: Sequence, structure and phylogenetic analysis. *Molecular and Biochemical Parasitology* **95**, 111–127.

KUWABARA, P. & SHAH, S. (1994). Cloning by synteny: identifying *C. briggsae* homologues of *C. elegans* genes. *Nucleic Acids Research* **22**, 159–164.

LEVICK, M. P., BLACKWELL, J. M., CONNOR, V., COULSON, R. M. R., MILES, A., SMITH, H. E., WAN, K.-L. & AJIOKA, J. W. (1996). An expressed sequence tag analysis of a full length, spliced-leader cDNA library from *Leishmania major* promastigotes. *Molecular and Biochemical Parasitology* **76**, 345–348.

McLAREN, D. J. (1972). Ultrastructural studies on microfilariae (Nematoda: Filarioidea). *Parasitology* **65**, 317–332.

McLAREN, D. J., WORMS, M. J., LAURENCE, B. R. & SIMPSON, M. G. (1975). Microorganisms in filarial larvae (Nematoda). *Transactions of the Royal Society of Tropical Medicine and Hygiene* **69**, 509–514.

MIZUKAMI, T., CHANG, W. I., GARKAVTSEV, I., KAPLAN, N., LOMBARDI, D., MATSUMOTO, T., NIWA, O., KOUNOSU, A., YANAGIDA, M., MARR, T. G. & BEACH, D. (1993). A 13 kb resolution cosmid map of the 14 Mb fission yeast genome by nonrandom sequence-tagged site mapping. *Cell* **73**, 121–132.

MOORE, T. A., RAMACHANDRAN, S., GAM, A. A., NEVA, F. A., LU, W., SAUNDERS, L., WILLIAMS, S. A. & NUTMAN, T. B. (1996). Identification of novel sequences and codon usage in *Strongyloides stercoralis*. *Molecular and Biochemical Parasitology* **79**, 243–248.

O'NEILL, S. L. (1995). *Wolbachia pipientis*: symbiont or parasite? *Parasitology Today* **11**, 168–169.

OKIMOTO, R., MacFARLANE, J. L., CLARY, D. O. & WOLSTENHOLME, D. R. (1992). The mitochondrial genomes of two nematodes, *Caenorhabditis elegans* and *Ascaris suum*. *Genetics* **130**, 471–498.

PALAZZOLO, M. J., SAWYER, S. A., MARTIN, C. H., SMOLLER, D. A. & HARTL, D. L. (1991). Optimised strategies for sequence-tagged site selection in genome mapping. *Proceedings of the National Academy of Sciences, USA* **88**, 8034–8038.

PAPPIN, D. J. C., HOJRUP, P. & BLEASBY, A. J. (1993). Rapid identification of proteins by peptide-mass fingerprinting. *Current Biology* **3**, 327–332.

PASTRANA, D. V., RAGHAVAN, N., FITZGERALD, P., EISINGER, S. W., METZ, C., BUCALA, R., SCHLEIMER, R. P., BICKELF, C. & SCOTT, A. L. (1998). Filarial nematode parasites secrete a homologue of the human cytokine macrophage migration inhibition factor (MIF). *Infection and Immunity* **66**, 5955–5963.

PIESSENS, W. F., McREYNOLDS, L. A. & WILLIAMS, S. A. (1987). Highly repeated DNA sequences as species-specific probes for *Brugia*. *Parasitology Today* **3**, 378–379.

POLITZ, S. M. & PHILIP, M. (1992). *Caenorhabditis elegans* as a model for parasitic nematodes: A focus on the cuticle. *Parasitology Today* **8**, 6–12.

REDDY, G. R., CHAKRABARTI, D., SCHUSTER, S. M., FERL, R. J., ALMIRA, E. C. & DAME, J. B. (1993). Gene sequence tags from *Plasmodium falciparum* genomic DNA fragments produced by the ''genease'' activity of mung bean nuclease. *Proceedings of the National Academy of Sciences, USA* **90**, 9867–9871.

RIDDLE, D., BLUMENTHAL, T., MEYER, B. & PRIESS, J. (Eds) (1997). *C. elegans II*. Cold Spring Harbor, NY, Cold Spring Harbor Laboratory Press.

ROLLINSON, D., KAUKAS, A., JOHNSTON, D. A., SIMPSON, A. J. & TANAKA, M. (1997). Some molecular insights into schistosome evolution. *International Journal for Parasitology* **27**, 11–28.

ROTHSTEIN, N., STOLLER, T. J. & RAJAN, T. V. (1988). DNA base composition of filarial nematodes. *Parasitology* **97**, 75–79.

SAKAGUCHI, Y., TADA, I., ASH, L. R. & AOKI, Y. (1983). Karyotypes of *Brugia pahangi* and *Brugia malayi* (Nematoda: Filaroidea). *Journal of Parasitology* **69**, 1090–1093.

SELKIRK, M. E., TANG, L., OU, X., COOKSON, E. & CHACON, M. R. (1994). Filarial anti-oxidant enzymes: Mediators of parasite persistence and potential targets for vaccination. *Parasite* **1**, 19–20.

SIM, B. K. L., SHAH, J., WIRTH, D. F. & PIESSENS, W. F. (1987). Characterisation of the filarial genome. *Filariasis*. Chichester, Wiley.

https://doi.org/10.1017/S0031182099004060 Published online by Cambridge University Press

SIRONI, M., BANDI, C., SACCHI, L., DI SACCO, B., DAMIANI, G. & GENCHI, C. (1995). Molecular evidence for a close relative of the arthropod endosymbiont *Wolbachia* in a filarial worm. *Molecular and Biochemical Parasitology* **74**, 223–227.

SPEITH, J., BROOKE, G., KUERSTEN, S., LEA, K. & BLUMENTHAL, T. (1993). Operons in *C. elegans*: Polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. *Cell* **73**, 521–532.

SULSTON, J., DU, Z., THOMAS, K., WILSON, R., HILLIER, L., STADEN, R., HALLORAN, N., GREEN, P., THIERRY-MIEG, J., QIU, L., DEAR, S., COULSON, A., CRAXTON, M., DURBIN, R., BERKS, M., METZSTEIN, M., HAWKINS, T., AINSCOUGH, R. & WATERSTON, R. (1992). The *C. elegans* genome sequencing project: A beginning. *Nature* **356**, 37–41.

SULSTON, J. E. & BRENNER, S. (1974). The DNA of *Caenorhabditis elegans*. *Genetics* **77**, 95–104.

TANAKA, M., HIRAI, H., LoVERDE, P. T., NAGAFUCHI, S., FRANCO, G. R., SIMPSON, A. J. & PENA, S. D. (1995). Yeast artificial chromosome (YAC)-based genome mapping of *Schistosoma mansoni*. *Molecular and Biochemical Parasitology* **69**, 41–51.

THE C. ELEGANS SEQUENCING CONSORTIUM (1998). Genome Sequence of the nematode *C. elegans*. A platform for investigating biology. *Science* **282**, 2012–2017.

THIERRY-MIEG, J. & DURBIN, R. (1992). ACeDB, a *C. elegans* database. *Cahiers IMABIO* **5**, 15–24.

VON HEIJNE, G. (1985). Signal sequences. The limits of variation. *Journal of Molecular Biology* **184**, 99–105.

VON HEIJNE, G. (1986). A new method for predicting signal sequence cleavage sites. *Nucleic Acids Research* **14**, 4683–4690.

WAN, K.-L., BLACKWELL, J. M. & AJIOKA, J. W. (1995). *Toxoplasma gondii* expressed sequence tags: insight into tachyzoite gene expression. *Molecular and Biochemical Parasitology* **75**, 179–186.

WATERSTON, R., SULSTON, J. E. & COULSON, A. R. (1997). The Genome. *C. elegans II*. Cold Spring Harbor, NY, Cold Spring Harbor Laboratory Press. 23–46.

WERREN, J. H. (1997). Biology of *Wolbachia*. *Annual Reviews of Entomology* **42**, 587–609.

WERREN, J. H., ZHANG, W. & GUO, L. R. (1995). Evolution and phylogeny of *Wolbachia*: reproductive parasites of arthropods. *Proceedings of the Royal Society of London Series B* (*Biological Sciences*) **261**, 55–71.

WILLIAMS, S. A., DeSIMONE, S. M., POOLE, C. B. & McREYNOLDS, L. A. (1987). Development of DNA probes to identify and speciate filarial parasites. *Molecular Paradigms for Eliminating Helminthic Parasites*. New York, Alan R. Liss Inc. 205–214.

WILSON, R., AINSCOUGH, R., ANDERSON, K., BAYNES, C., BERKS, M., BONFIELD, J., BURTON, J., CONNELL, M., COPSEY, T., COOPER, J., COULSON, A., CRAXTON, M., DEAR, S., DU, Z., DURBIN, R., FAVELLO, A., FRASER, A., FULTON, L., GARDNER, A., GREEN, P., HAWKINS, T., HILLIER, L., JIER, M., JOHNSTON, L., JONES, M., KERSHAW, J., KIRSTEN, J., LAISSTER, N., LATRIELLE, P., LIGHTNING, J., LLOYD, C., MORTIMORE, B., O'CALLAGHAN, M., PARSONS, J., PERCY, C., RIFKEN, L., ROOPRA, A., SAUNDERS, D., SHOWNKEEN, R., SIMS, M., SMALDON, N., SMITH, A., SMITH, M., SONNHAMMER, E., STADEN, R., SULSTON, J., THIERRY-MIEG, J., THOMAS, K., VAUDIN, M., VAUGHAN, K., WATERSTON, R., WATSON, A., WEINSTOCK, L., WILKINSON-SPROAT, J. & WOHLDMAN, P. (1994). 2·2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans*. *Nature* **368**, 32–38.

WINNEPENNINCKX, B., BACKELJAU, T., MACKEY, L. Y., BROOKS, J. M., DeWACHTER, R., KUMAR, S. & GAREY, J. R. (1995). 18S rRNA data indicate that Aschelminthes are polyphyletic in origin and consist of at least three distinct clades. *Molecular Biology and Evolution* **12**, 1132–1137.

https://doi.org/10.1017/S0031182099004060 Published online by Cambridge University Press