Open Access

Analysis and functional classification of transcripts from the nematode *Meloidogyne incognita*

James P McCarter*[†], Makedonka Dautova Mitreva*, John Martin*, Mike Dante*, Todd Wylie*, Uma Rao[‡], Deana Pape*, Yvette Bowers*, Brenda Theising*, Claire V Murphy*, Andrew P Kloek[†], Brandi J Chiapelli[†], Sandra W Clifton*, David McK Bird[‡] and Robert H Waterston*[§]

Addresses: *Genome Sequencing Center, Department of Genetics, Box 8501, Washington University School of Medicine, St. Louis, MO 63108, USA. [†]Divergence Inc., 893 North Warson Road, St. Louis, MO 63141, USA. [‡]The Center for the Biology of Nematode Parasitism, North Carolina State University, Raleigh, NC 27695, USA. [§]Department of Genome Sciences, University of Washington, 1705 NE Pacific St, Seattle, WA 98195, USA.

Correspondence: James P McCarter. E-mail: jmccarte@watson.wustl.edu

Published: 31 March 2003

Genome ***Biology*** 2003, **4**:R26

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/4/R26

Received: 10 December 2002
Revised: 17 February 2003
Accepted: 28 February 2003

Abstract

Background: Plant parasitic nematodes are major pathogens of most crops. Molecular characterization of these species as well as the development of new techniques for control can benefit from genomic approaches. As an entrée to characterizing plant parasitic nematode genomes, we analyzed 5,700 expressed sequence tags (ESTs) from second-stage larvae (L2) of the root-knot nematode *Meloidogyne incognita*.

Results: From these, 1,625 EST clusters were formed and classified by function using the Gene Ontology (GO) hierarchy and the Kyoto KEGG database. L2 larvae, which represent the infective stage of the life cycle before plant invasion, express a diverse array of ligand-binding proteins and abundant cytoskeletal proteins. L2 are structurally similar to *Caenorhabditis elegans* dauer larva and the presence of transcripts encoding glyoxylate pathway enzymes in the *M. incognita* clusters suggests that root-knot nematode larvae metabolize lipid stores while in search of a host. Homology to other species was observed in 79% of translated cluster sequences, with the *C. elegans* genome providing more information than any other source. In addition to identifying putative nematode-specific and *Tylenchida*-specific genes, sequencing revealed previously uncharacterized horizontal gene transfer candidates in *Meloidogyne* with high identity to rhizobacterial genes including homologs of nodL acetyltransferase and novel cellulases.

Conclusions: With sequencing from plant parasitic nematodes accelerating, the approaches to transcript characterization described here can be applied to more extensive datasets and also provide a foundation for more complex genome analyses.

Background

Root-knot nematode species, including *Meloidogyne incognita*, are the most important of the plant parasitic nematodes, infecting almost all cultivated plants, and are responsible for billions of dollars in crop losses annually [1,2]. They are obligatory sedentary endoparasites with a 1- to 2-month life cycle. Embryos develop in a proteinaceous matrix extruded by the adult female, and hatch as second-stage larvae (L2) that move through the soil and invade the plant root. Within the root, the worm establishes a feeding site and undergoes three additional molts to become an adult. *M. incognita* is a mitotic parthenogenetic species. Males develop but appear to play no role in reproduction [3]. Females swell to a pear shape and are incapable of moving once committing to a root feeding site.

The *Meloidogyne* L2 larvae, the infective stage where the worm is away from the host plant (also referred to as second-stage juvenile in the literature), is more accessible than the rest of the life cycle, and is an interesting stage biologically with the worm completing multiple steps required for survival. On hatching from the eggshell, L2 worms are able to locate and migrate towards a potential host plant, penetrate the root behind its tip in the zone of elongation, and migrate intercellularly through the vascular cylinder by separating cells at the middle lamella [4]. The migration is enabled by a combination of stylet protrusion (mechanical force) and secretion of cell-wall-degrading enzymes from specialized glands [5-8]. Upon completion of migration, secretions from the nematode's glands, and potentially other cues, induce root cells to alter their development and gene expression, undergoing abnormal growth and repeated endomitotic rounds of replication to form a feeding site made up of giant cells [9,10]. The L2 feeds from the giant cells for 10-12 days, then ceases feeding and molts three times over the next two days to form the adult. L2 undergo significant change following establishment of the feeding site, including swelling of the body and a switch in gland activity from subventral to dorsal dominance [11].

Until recent years, molecular characterization of *Meloidogyne* genes has been limited [12,13], particularly because the species' obligate parasitic life cycle makes studies difficult. Both basic understanding of root-knot nematode biology and applied research toward new means of nematode control are now beginning to benefit from the rapid identification of transcribed genes in the species. The generation of expressed sequence tags (ESTs) by single-pass random sequencing of cDNA libraries is a powerful tool for rapid gene transcript identification in metazoans [14-17] including parasitic nematodes of humans and animals [18-23]. High-throughput projects on two dozen nematode species have now brought the total number of publicly available roundworm ESTs to nearly 400,000, with half the sequences coming from parasites [24-27]. As a part of these efforts, EST sequencing from plant parasitic nematodes is in progress [28] and pilot EST datasets from the root-knot nematode *M. incognita* and the cyst nematodes *Globodera rostochiensis* and *G. pallida* second-stage larvae have recently been analyzed [29,30].

Important to the characterization and understanding of these sequences is the creation and implementation of bioinformatics approaches (such as clustering, functional classification, similarity analysis) that can be applied uniformly across the ever-increasing multiple nematode datasets. We present here an analysis of 5,713 ESTs from *M. incognita* L2 including creation of NemaGene clusters to reduce sequence redundancy, identification of abundant transcripts, and functional classification of gene products based on assignments to InterPro domains, the Gene Ontology hierarchy, and KEGG biochemical pathways. Building on the availability of the complete genome sequence, gene homologs of the free-living nematode *Caenorhabditis elegans* [31] were identified for *M. incognita* clusters and correlated with known RNA interference (RNAi) phenotypes. Genes specific to plant parasitic nematodes (*Tylenchida* species) as well as prokaryotic-like horizontal gene transfer candidates were also examined.

Results and discussion

As part of a larger effort to examine expressed gene sequences from parasitic nematodes, we have generated and submitted to GenBank's EST database 5,713 ESTs from a *M. incognita* L2 library. Sequences, which include both 5′ and 3′ reads, averaged 481 nucleotides, resulting in 2.82 million submitted nucleotides. Here we present a first analysis applying semi-automated bioinformatics tools to genome data from a plant parasitic nematode, thereby laying the groundwork for more extensive analyses.

NemaGene cluster analysis

To reduce data redundancy, improve base accuracy and transcript length, and determine gene representation within the library, ESTs from the *M. incognita* L2 library were grouped by sequence identity into contigs and clusters by a method using Phrap and BLAST. 'Contig' member ESTs appear to derive from identical transcripts while 'cluster' members may derive from the same gene yet represent different transcript splice isoforms (that is, ESTs form contigs, contigs form clusters). Beginning with 5,713 traces, automated screens and manual inspection of misassembled contigs resulted in the elimination of 52 ESTs as potential chimeric sequences. The remaining 5,661 ESTs formed 1,798 contigs and 1,625 clusters. Clusters varied in size from a single EST (723 cases) to 77 ESTs (1 case) (Figure 1). By eliminating data redundancy during contig building, the total number of nucleotides used for further analysis was reduced from 2.82 million to 1.99 million. To a first approximation, this project generated sequence from as many as 1,625 genes, for a new gene discovery rate of 29%, with only 13% of ESTs being singletons. This number may, however, overestimate gene discovery as a single gene could be represented by multiple non-overlapping clusters. While library redundancy reduces the number of new genes discovered, 65% of clusters still have 10 or fewer EST members. Such redundancy is desirable to increase base accuracy and transcript length within contigs. Additionally, 122 clusters have

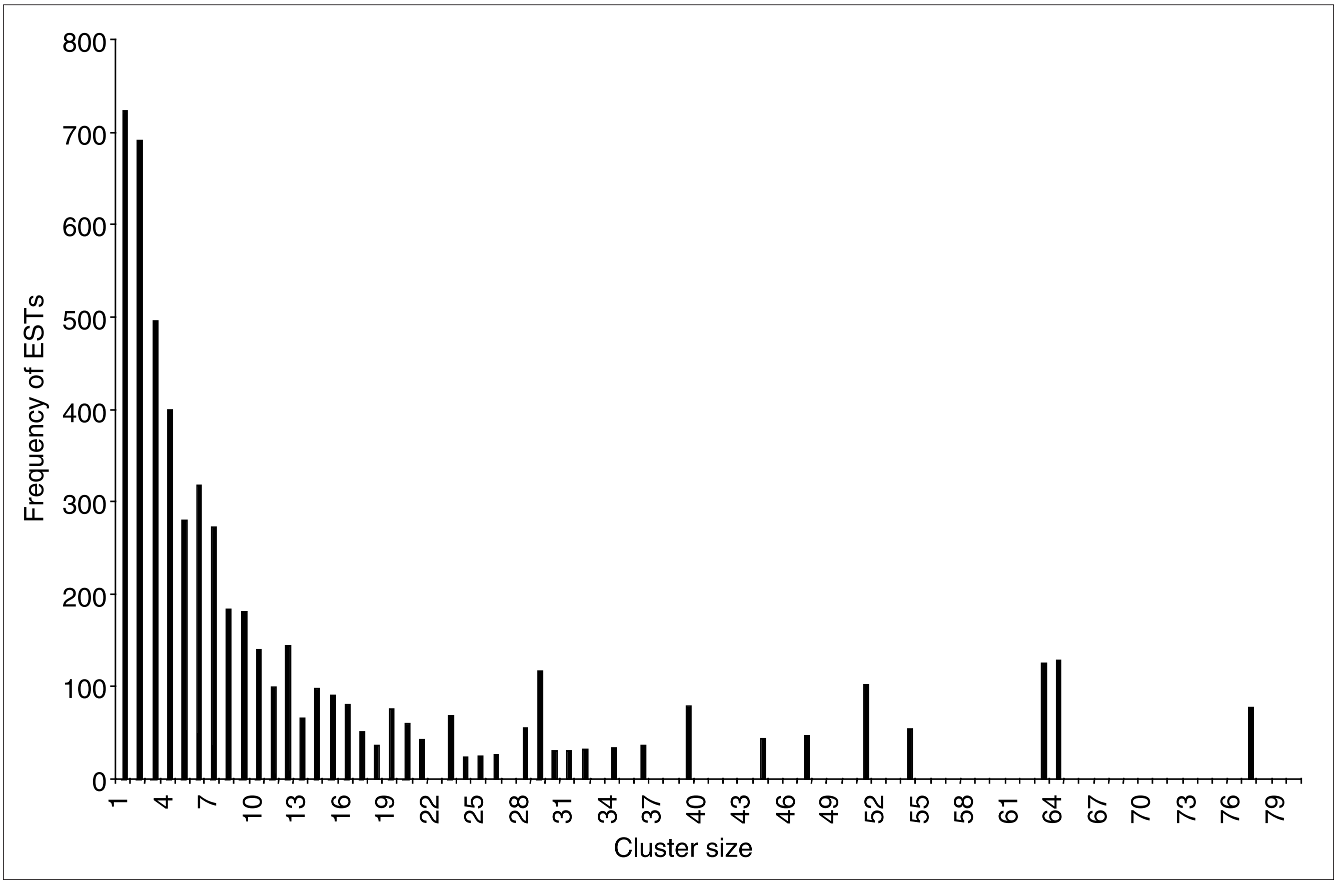

Figure 1
Histogram showing the distribution of ESTs by cluster size. For example, there were seven clusters of size 14 containing a sum of 98 ESTs. Distribution of contig sizes is not shown.

multiple contig members, revealing potential splice isoforms. Contig building was successful in significantly increasing the length of assembled transcript sequences from 481 ± 108 nucleotides for submitted ESTs alone to 611 ± 174 nucleotides for multi-member contigs. The longest sequence also increased from 780 to 2,353 nucleotides. Sampling of another 5,661 ESTs from the same source is estimated to result in the discovery of only 329 new clusters, a new gene discovery rate of only 6% (ESTFreq, W. Gish, personal communication). Further sampling will therefore await library normalization. This same clustering methodology is being applied to ESTs from other nematode species [32].

Transcript abundance and highly represented genes

The 25 most abundant EST clusters accounted for 18% of all ESTs generated. A high level of representation in a cDNA library generally correlates with high transcript abundance in the original biological sample [33], although artifacts of library construction can result in selection for or against representation of some transcripts. Transcripts abundantly represented in the library include genes encoding cytoskeleton proteins (such as myosin, actin, UNC-87, troponin T) and proteins that carry out core eukaryotic energetic and metabolic processes (for example ADP/ATP translocase, lactate dehydrogenase) (Table 1). Sixty-four ESTs had significant homology to the putative fatty-acid-binding protein Sec-2, confirming the abundant expression of this gene reported in L2 cDNA libraries from *M. incognita* [29] and the cyst nematodes *G. rostochiensis* and *G. pallida* cDNA [30]. Sec-2 is secreted by plant-parasitic nematodes at relatively high levels [34]. Several abundantly expressed genes are also horizontal gene transfer candidates (see below).

Functional classification based on Gene Ontology assignments

To categorize transcripts by putative function, we have utilized the Gene Ontology (GO) classification scheme [35,36]. GO provides a dynamic controlled vocabulary and hierarchy that unifies descriptions of biological, cellular and molecular functions across genomes. InterProScan was used to match *Meloidogyne* clusters to characterized protein domains (5,875 entries) in the InterPro database [37]. Existing mappings of InterPro domains allowed placement of *Meloidogyne* clusters into the GO hierarchy, viewed locally with the

Table 1

The most abundantly represented transcripts in the *M. incognita* cDNA library

			Non-redundant GenBank			
	Cluster	ESTs	Best identity descriptor	Accession SW/TR*	E-value	*C. elegans* gene Wormpep
1	MI00951.cl	77	*C. elegans* UNC-87, thin filament associated	P37806	5e-87	F08B6.4
2	MI00033.cl	64	*C. elegans* MLC-1, myosin light chain	P19625	3e-74	C36E6.3
3	MI00502.cl	64	*G. pallida* SEC-2, sec-2 protein	Q94569	3e-67	F02A9.3†
4	MI00049.cl	63	*C. elegans* HSP-12, heat shock protein 20	P34328	2e-36	C14B9.1
5	MI01047.cl	63	Novel	-	-	-
6	MI00984.cl	54	*M. javanica* CAP-1, calponin homolog	P91763	2e-126	F28H1.2†
7	MI01045.cl	51	*Rhizobium* NODL, nodulation protein L	P28266	3e-56	-
8	MI00702.cl	51	*C. elegans* NHL repeat	P91268	4e-104	F21F3.1
9	MI00046.cl	47	*C. elegans* MIP/Aquaporin-3 water channel	Q21473	1e-54	M02F4.8
10	MI00487.cl	44	*C. elegans* ACT-2, actin 2	P10986	2e-240	M03F4.2
11	MI00784.cl	39	*C. elegans* MUP-2 troponin-T	Q20694	7e-107	F53A9.10
12	MI01043.cl	39	*C. elegans* cytidylyl transferase	Q9BL56	3e-06	Y65B4A.8
13	MI00775.cl	36	*C. elegans* NLP-21	Q9U2B9	5e-17	Y47D3B.2
14	MI01042.cl	34	*C. elegans* ADP/ATP Translocase	P91410	1e-54	T01B11.4
15	MI00483.cl	32	*M. incognita* ENG-1, Beta-1,4-endoglucanase	Q9UA57	1e-305	-
16	MI01040.cl	31	Novel	-	-	-
17	MI00027.cl	30	*C. elegans* MLC-3, myosin light chain family	P53014	2e-71	F09F7.2
18	MI01113.cl	29	Human APG-5, apoptosis specific protein	O60875	1e-16	F08.H9.4†
19	MI00774.cl	29	*Dictyostelium* ACRA, adenylate cyclase	Q9U9S7	2e-20	C24A8.3†
20	MI00721.cl	29	*C. elegans* LDH-1, l-lactate dehydrogenase	Q27888	7e-124	F13D12.2
21	MI00040.cl	29	*C. elegans* GST-7, glutathione *S*-transferase	P91254	7e-42	F11G11.2
22	MI01038.cl	28	Mouse TNRC11, Opa repeat	Q62006	5e-19	H20J18.1
23	MI00629.cl	28	*C. elegans* C4-type steroid receptor zinc finger	O16890	2e-23	F13A2.8
24	MI01036.cl	26	Novel	-	-	-
25	MI01034.cl	25	*C. elegans* arginine kinase phosphotransferase	Q10454	1e-91	F46H5.3

*SW/TR is SWISS-PROT and TrEMBL Proteinknowledgebase [105]. †*C. elegans* homolog present but with a lower probability match than the best GenBank descriptor.

AmiGO browser. Of 1,625 clusters, 1,280 (79%) have homologies beyond *M. incognita*, 693 (43%) align to InterPro domains, and 475 (29%) map to the GO hierarchy. These 475 clusters represent generally conserved genes containing domains with characterized biochemical and physiological function in other species. The actual mappings are more complicated than one-to-one: the 693 clusters with InterPro alignments match to 379 InterPro domains, and the 475 clusters with GO assignments have 764 mappings to 127 GO categories.

Gene Ontology representation of *M. incognita* clusters is shown for each organizing principle of GO: biological process (Table 2a, Figure 2a), cellular component (Table 2b, Figure 2b), and molecular function (Table 2c, Figure 2c). Table 2 and Figure 2 provide a breakdown of representation by major GO categories. A complete listing of GO mappings is available as additional data with the online version of this article. While hatched L2 before plant invasion are a long-lived non-feeding dispersal stage [4], GO categories reveal numerous transcripts encoding metabolic enzymes, including those involved in biosynthetic pathways. Distributions of clusters by GO categories can be compared to findings from other species using the TIGR gene index [38,39] which includes information for three nematodes - the free-living *C. elegans* and the human filarial parasites *Brugia malayi* and *Onchocerca volvulus*. Table 3 compares observed GO representation among nematode species. The most striking initial differences in *M. incognita* GO representation from the other three species were for molecular function, where 52% of *Meloidogyne* clusters had ligand-binding/carrier mappings versus 24-28% for the other species, and cellular component, where 15% of *M. incognita* clusters had extracellular mappings versus 0-2% for the other species. *Meloidogyne* extracellular mappings (15 clusters) were all within the category of SCP/Tpx-1/Ag5/PR-1/Sc7 extracellular

Table 2

Gene Ontology mappings

(a) Biological process

Categories and subcategories	Representation			% Representation of total		
Metabolism	133			75%		
Protein metabolism and modifications		57			32%	
Protein modification			25			14%
Protein biosynthesis			15			8%
Protein degradation			14			8%
Protein folding			3			2%
Glycoprotein metabolism			1			1%
Catabolism		24			13%	
Protein degradation			14			8%
Glycolysis			8			4%
Phosphate metabolism		23			13%	
Kinase			19			11%
Phosphatase			4			2%
Biosynthesis		17			10%	
Protein biosynthesis			15			8%
Electron transport		21			12%	
Nucleic acid metabolism		16			9%	
Transcription			13			7%
RNA metabolism			2			1%
DNA metabolism			1			1%
Carbohydrate metabolism		11			6%	
Glycolysis			8			4%
Amino acid and derivative metabolism		4			2%	
One-carbon compound metabolism		3			2%	
Oxygen and radical metabolism		3			2%	
Nitrogen metabolism		1			1%	
Secondary metabolism		1			1%	
Transport	24			13%		
Ion transport (including channels)		8			4%	
Protein transport and trafficing		4			2%	
Amino acid transport		2			1%	
Cell communication	21			12%		
Signal transduction		20			11%	
Intracellular signaling cascade			14			8%
Cell surface receptor linked signal transduction			4			2%
Response to external stimulus		1			1%	

(b) Cellular component

Categories and subcategories	Representation				% Representation of total			
Cell	79				81%			
Intracellular		62				64%		
Cytoplasm			42				43%	
Ribosome				29				30%
Cytoskeleton				5				5%
Mitochondria				5				5%
Proteasome				2				2%
Translation factor				1				1%
Nucleus			15				15%	

Table 2 *(continued)*

(b) Cellular component

Categories and subcategories	Representation	% Representation of total
Unspecified	3	3%
Plasma membrane	1	1%
Membrane	22	23%
Unspecified	16	16%
Mitochondrial membrane	4	4%
Integral membrane	2	2%
Extracellular	15	15%
Unlocalized	3	3%

(c) Molecular function

Categories and subcategories	Representation	% Representation of total
Ligand binding / carrier	135	52%
Nucleic acid binding	44	17%
Nucleotide binding	40	15%
Calcium binding	22	8%
Protein binding	12	5%
Carbohydrate binding	7	3%
Electron transport	3	1%
Lipid binding	3	1%
Heavy metal binding	1	<1%
Oxygen binding	1	<1%
Oxygen transport	1	<1%
Enzyme	101	39%
Hydrolase	37	14%
Transferase	26	10%
Oxidoreductase	22	8%
Kinase	15	6%
Phosphatase	8	3%
Helicase	4	2%
Lyase	4	2%
Aldolase	2	1%
Ligase	2	1%
Isomerase	1	<1%
Monooxygenase	1	<1%
Transporter	14	5%
Channel/pore	5	2%
Carrier	4	2%
Intracellular transporter	3	1%
Ion transporter	3	1%
Oxygen transporter	1	<1%
Signal transducer	9	3%
Receptor	5	2%
Receptor signaling protein	3	1%
Structural molecule	5	2%
Enzyme regulator	4	2%
Cell adhesion	1	<1%
Motor	1	<1%
Transcriptional regulator	1	<1%

(a) 178 clusters generated 336 multiple mappings. Percentage representation is based on 178. (b) 97 clusters generated 107 multiple mappings. Percentage representation is based on 97. (c) 261 clusters generated 321 multiple mappings. Percentage representation is based on 261.

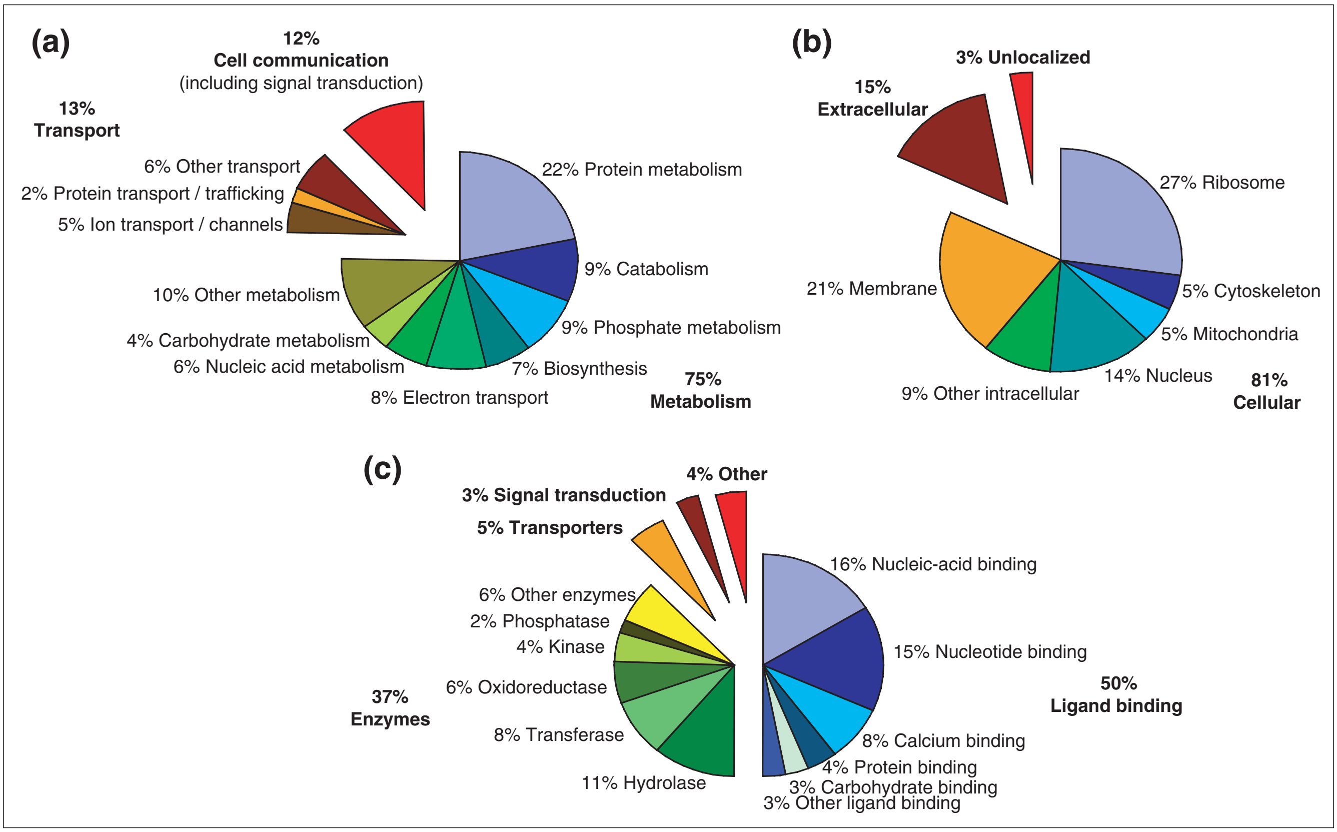

Figure 2

Percentage representation of gene ontology (GO) mappings for *M. incognita* clusters. **(a)** Biological process; **(b)** cellular component; **(c)** molecular function. More detailed information is provided in Table 2 (see also Additional data files). Note that individual GO categories can have multiple mappings. For instance, GO:0015662: P-type ATPase (cluster-MI00952, Interpro domain IPR004014) is a nucleic-acid-binding protein, a hydrolase enzyme and a transporter.

proteins (InterPro domain IPR001283) and showed homology to the genes *vap-1* from *H. glycines* and *Mi-msp-1* from *M. incognita* [40,41], both venom allergen antigen 5 family members with homologs in numerous nematodes including hookworms and *C. elegans* [42]. Categories that particularly contributed to the abundance of ligand-binding/carrier mappings for *Meloidogyne* included EF-hand calcium binding (22 clusters), RNA recognition motif (18 clusters), and a variety of ATP-binding domains (20 clusters). Differences in the distribution of GO mappings may be attributable to the more extensive stage representation available for the other species. Comparisons of relative expression levels for genes among different *M. incognita* stages will begin to be possible as EST collections from other life-cycle stages are generated and analyzed.

Functional classification based on KEGG analysis

As an alternative method of categorizing clusters by biochemical function, clusters were assigned to metabolic pathways using the Kyoto Encyclopedia of Genes and Genomes database (KEGG [43]) using enzyme commission (EC) numbers as the basis for assignment. EC numbers were assigned to 258 clusters (16% of total), of which 176 (11%) had mappings to KEGG biochemical pathways (361 total and 212 unique mappings). Out of 82 possible metabolic pathways 56 were represented (Table 4). For a complete listing of KEGG mappings see Additional data files. Pathways well represented by the *M. incognita* clusters include: glycolysis/gluconeogenesis (10 enzymes represented), citrate cycle (7), fatty-acid metabolism and biosynthesis (11), pyrimidine metabolism (7), lysine degradation (8), arginine and proline metabolism (8) and tryptophan metabolism (8). Lysine, arginine and tryptophan are essential amino acids in *C. elegans* whereas proline is not [44]. Pathways not represented in *Meloidogyne* include alkaloid biosynthesis II and riboflavin (vitamin B2) metabolism. *C. briggsae* is incapable of synthesizing riboflavin [45] but *C. elegans* does appear to have a homolog of a riboflavin kinase (R10H10.6) and *M. incognita* may have at least one enzyme involved in riboflavin processing (see below).

Nematodes are believed to be unique among animals in utilizing the glyoxylate cycle to generate carbohydrates from the beta-oxidation of fatty acids (reviewed in [46]). The

comment reviews reports deposited research refereed research interactions information

Table 3

Comparison of gene ontology mappings among nematode species

		% Representation			
Gene Ontology	Categories and subcategories	*M. incognita*	*C. elegans*	*B. malayi*	*O. volvulus*
Biological process	Cell growth and maintenance	88	68	91	93
	Cell communication	12	16	3	4
Cellular component	Cell	81	96	99	98
	Extracellular	15	2	-	-
	Unlocalized	3	0.6	-	1
Molecular function	Ligand binding / carrier	52	28	24	28
	Enzyme	39	35	33	31
	Transporter	5	13	6	13
	Signal transducer	3	7	2	3
	Structural molecule	2	5	17	15
	Enzyme regulator	2	1	2	-
	Cell adhesion	0.4	0.3	-	-
	Motor	0.4	1	2	3
	Transcriptional regulator	0.4	4	1	1

GO mappings for *C. elegans*, *B. malayi* and *O. volvulus* were obtained from [39].

glyoxylate pathway, generally found in plants and microorganisms, is similar to the citrate cycle, but relies on two critical enzymes, malate synthase and isocitrate lyase, to bypass two decarboxylation steps. Nematodes appear to use this pathway for energy production from stored lipids during starvation or non-feeding stages [47,48] such as *Meloidogyne* pre-infective L2. Eight *M. incognita* L2 clusters map to five glyoxylate pathway enzymes. These include homologs of malate synthase (MI00879.cl, EC 4.1.3.2, BLASTX probability of 2e-31), several enzymes not shared with the citrate cycle (for example, formate tetrahydrofolate ligase, EC 6.3.4.3, 5e-38), as well as two shared with the citrate cycle (for example, malate dehydrogenase, EC 1.1.1.37, 4e-29). Isocitrate lyase (EC 4.1.3.1) was not observed in this EST collection, but the first putative *Tylenchida* homologs of this gene have subsequently appeared from our further EST sequencing (*M. hapla* BM883225, and *M. javanica* BI324412). In *C. elegans*, two genes each encode unusual bifunctional enzymes containing both isocitrate lyase and malate synthase domains [49]. Since the isocitrate lyase domain lies within the amino-terminal half of the *C. elegans* bifunctional enzyme and none of the *Meloidogyne* EST reads stretches across both domains, further sequencing of the 3′ end of cDNA clones from the *M. hapla* or *M. javanica* isocitrate lyase ESTs will be necessary to determine whether the *Meloidogyne* genus contains a bifunctional glyoxylate enzyme homolog similar to that of *C. elegans*. The presence of glyoxylate pathway enzymes in *Meloidogyne* L2 provides experimental support for the model describing this larval stage as the functional equivalent of the *C. elegans* dauer larva [41]. These ESTs and their corresponding cDNA clones will be useful reagents for the further study of the glyoxylate pathway in different stages of the *Meloidogyne* life cycle. For instance, energy metabolism would be expected to change markedly upon plant invasion and intracellular migration toward the feeding site, and might include a decrease in expression of transcripts specific to the glyoxylate pathway.

Distribution of BLAST database matches and homologs in *C. elegans*

Figure 3 is a Venn diagram combining the results of BLAST searches versus three databases for the 79% (1,280/1,625) of *M. incognita* clusters which had matches to sequences from other species. Strikingly, in the majority of cases where homologies were found (740/1,280), matches were found in all three of the databases surveyed - *C. elegans* proteins, other nematode sequences, and non-nematode sequences. Gene products in this category are generally widely conserved across metazoans and many are involved in core biological processes. This category should continue to expand as additional complete genomes become available [50,51].

The 20% of contigs (353) that had no homology may contain novel or diverged amino-acid coding sequences that are specific to *Meloidogyne* species or even to *M. incognita* only. Alternatively, clusters which containing mostly 3′ or 5′ untranslated regions (UTRs) would lack BLASTX homology because they are non-coding or contain too short a coding sequence to result in significant homology. To examine this latter possibility contig consensus sequences with and without BLASTX homology were examined to determine their longest open reading frame (ORF). The distribution of ORF sizes indicates that clusters without homology contain two populations; one population of novel protein-coding sequences with a similar distribution of ORF sizes to that found in sequences with homology, and a second population

Table 4

KEGG biochemical pathway mappings for *M. incognita* clusters

KEGG categories represented	Clusters	Enzymes
1.1 Glycolysis/gluconeogenesis	13	10
1.2 Citrate cycle (TCA cycle)	11	7
1.3 Pentose phosphate cycle	8	6
1.4 Pentose and glucuronate interconversions	3	3
1.5 Fructose and mannose metabolism	8	6
1.6 Galactose metabolism	6	5
1.7 Ascorbate and aldarate metabolism	6	3
1.8 Pyruvate metabolism	18	9
1.9 Glyoxylate and dicarboxylate metabolism	8	5
1.10 Propanoate metabolism	11	6
1.11 Butanoate metabolism	11	6
2.1 Oxidative phosphorylation	12	3
3.1 Fatty acid biosynthesis (path 1)	1	1
3.2 Fatty acid biosynthesis (path 2)	5	3
3.3 Fatty acid metabolism	20	7
3.4 Synthesis and degradation of ketone bodies	2	1
3.5 Sterol biosynthesis	1	1
3.6 Bile acid biosynthesis	6	3
3.8 Androgen and estrogen metabolism	3	3
4.1 Purine metabolism	6	5
4.2 Pyrimidine metabolism	9	7
4.3 Nucleotide sugars metabolism	5	4
5.1 Glutamate metabolism	4	4
5.2 Alanine and aspartate metabolism	3	2
5.3 Glycine, serine and threonine metabolism	6	5
5.4 Methionine metabolism	3	2
5.5 Cysteine metabolism	3	2
5.6 Valine, leucine and isoleucine degradation	9	5
5.7 Valine, leucine and isoleucine biosynthesis	1	1
5.8 Lysine biosynthesis	1	1
5.9 Lysine degradation	13	8
5.10 Arginine and proline metabolism	14	8
5.11 Histidine metabolism	6	3
5.12 Tyrosine metabolism	8	5
5.13 Phenylalanine metabolism	8	6
5.14 Tryptophan metabolism	22	8
5.15 Phenylalanine/tyrosine/tryptophan biosynthesis	2	2
5.16 Urea cycle and metabolism of amino groups	1	1
6.1 beta-Alanine metabolism	8	3
6.3 Aminophosphonate metabolism	1	1
6.4 Selenoamino acid metabolism	5	3
6.6 D-Glutamine and D-glutamate metabolism	1	1
6.7 D-Arginine and D-ornithine metabolism	4	3
6.9 Glutathione metabolism	8	4
7.1 Starch and sucrose metabolism	9	5
7.2 Glycoprotein biosynthesis	2	1
7.4 Aminosugars metabolism	3	3
8.1 Glycerolipid metabolism	9	4
8.2 Inositol phosphate metabolism	1	1
8.5 Sphingoglycolipid metabolism	3	3
8.8 Prostaglandin and leukotriene metabolism	2	1
9.3 Vitamin B6 metabolism	1	1
9.4 Nicotinate and nicotinamide metabolism	13	2
9.5 Pantothenate and CoA biosynthesis	3	2
9.8 One carbon pool by folate	3	3
9.11 Ubiquinone biosynthesis	8	4
10.20 Tetrachloroethene degradation	0	0
10.21 Styrene degradation	0	0
12.3 Aminoacyl-tRNA biosynthesis	0	0

KEGG categories not represented	Clusters	Enzymes
2.5 Methane metabolism	0	0
2.6 Nitrogen metabolism	0	0
2.7 Sulfur metabolism	0	0
6.2 Taurine and hypotaurine metabolism	0	0
6.5 Cyanoamino acid metabolism	0	0
7.3 Glycoprotein degradation	0	0
7.7 Glycosaminoglycan degradation	0	0
8.3 Sphingophospholipid biosynthesis	0	0
8.4 Phospholipid degradation	0	0
9.2 Riboflavin metabolism	0	0
9.7 Folate biosynthesis	0	0
9.10 Porphyrin and chlorophyll metabolism	0	0
10.2 Flavonoids, stilbene and lignin biosynthesis	0	0
10.3 Alkaloid biosynthesis I	0	0
10.4 Alkaloid biosynthesis II	0	0
10.6 Streptomycin biosynthesis	0	0
10.7 Erythromycin biosynthesis	0	0
10.14 Gamma-hexachlorocyclohexane degradation	0	0
10.18 1,2-Dichloroethane degradation	0	0

Categories eliminated	
2.2 Photosynthesis	Plants
2.3 Carbon fixation	Plants
2.4 Reductive carboxylate cycle (CO_2 fixation)	Plants
7.6 Peptidoglycan biosynthesis	Bacterial cell wall

comment
reviews
reports
deposited research
refereed research
interactions
information

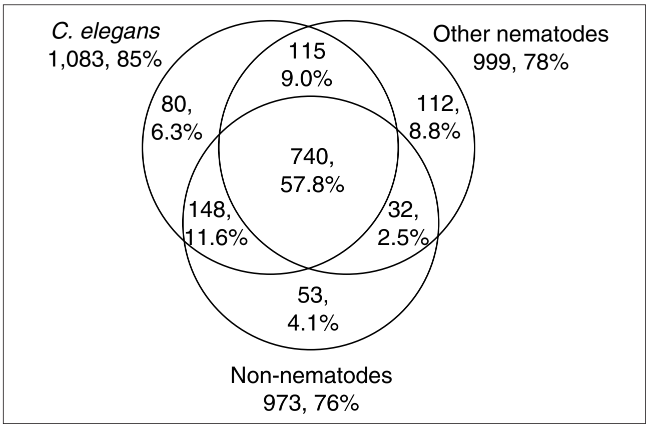

Figure 3
Venn diagram showing distribution of *M. incognita* BLAST matches by database. Databases used were: for *C. elegans*, Wormpep v.54 and mitochondrial protein sequences; for other nematodes, all GenBank nucleotide data for nematodes except *C. elegans* and *M. incognita*; and for non-nematodes, SWIR v.21 with all nematode sequences removed.

of UTR sequences containing random or generally short ORFs (Figure 4). The combined distribution is bimodal (relatively high left shoulder) with a mean ORF size of 140 amino acids versus a mean ORF size of 172 amino acids for sequences with homology. A further characterization of novel *M. incognita* genes could begin by examining those with longer ORFs as these are most likely to be real coding regions.

In contrast to these findings for *M. incognita* where most clusters had homology, BLAST searches with EST clusters from the filarial nematode *B. malayi* showed far fewer database matches with the same e-value cut-off of 10^{-5} [52] - 57% versus 79%. Part of this difference is due to the use of more extensive databases in the *M. incognita* search. For instance, the *Meloidogyne* search included all dbEST sequences in the 'other nematode' set, resulting in matches for 61% of all clusters, whereas the *Brugia* search used only protein sequences in GenBank and saw matches in only around 12% of cases. However, even matches in *C. elegans* were fewer for *B. malayi* (50% versus 67%), where nearly identical databases were used. *Brugia*, *Meloidogyne* and *Caenorhabditis* represent three separate major nematode clades (III, IV and V, respectively) [53]. Possible explanations for the discrepancy in matches are that the *Brugia* clusters contain a large fraction of non-coding sequences (that is, 5′ and 3′ UTR, unspliced introns) or have undergone more rapid molecular evolution and diversification. Alternatively, since the *Brugia* ESTs derive from 12 different libraries they may represent rarer transcripts than are contained in the *M. incognita* collection. A correlation between stage of expression and molecular conservation has been observed in *C. elegans* [54].

As expected, the *C. elegans* genome [31] was the best source of information for interpreting *M. incognita* sequences with 85% of all clusters with matches showing homology to a *C. elegans* gene product (Figure 3). Table 5 presents the 15 gene products with the highest level of conservation (e-240 to e-115) between *M. incognita* and *C. elegans*; these include gene products involved in cell structure (for example, actin, myosin), protein biosynthesis (for example, ribosomal proteins) and glycolysis (for example, lactate dehydrogenase, enolase). Representation of these clusters in the *M. incognita* L2 EST collection varied from common (77 ESTs) to rare (1 EST). None of these most conserved gene products was nematode specific. Out of all clusters 281 (17%) had homology only to nematodes, either *C. elegans* (80), other nematodes (53), or both (148). The most conserved of these nematode-specific proteins had a probability value of e-77. Included among the most conserved nematode-specific proteins were previously characterized nematode-specific domains including the transthyretin-like domain IPR001534 [55] (MI00092.cl), as well as uncharacterized *C. elegans* hypothetical proteins (for example, MI01590.cl = TrEMBL Q19251; MI00719.cl = TrEMBL P90889).

Thirteen *M. incognita* clusters lacked homology to any *C. elegans* protein in Wormpep (v.54) yet had significant homology to regions of the *C. elegans* genome by TBLASTX. Such matches might reveal unpredicted protein-coding regions within the genome. Most of the clusters, including MI00112.cl, MI0000518.cl, MI01572.cl (matching to *C. elegans* LG V:10343341..10344858), MI01502.cl (LG X:16624802..16624921), MI00768.cl (LG III:2421909..2421700) matched regions of the genome where genes were predicted in later versions of Wormpep (WP 88, WP 73 and WP 65, respectively) indicating the usefulness of ESTs from other nematodes in predicting *C. elegans* coding regions. In fact, ESTs from our parasitic nematode sequencing project are being continually mapped to the *C. elegans* genome [56] and used by Wormpep curators for this purpose. We are further investigating other regions of homology such as MI00899.cl (LG II:7443833..7443537) to determine whether modifications to current *C. elegans* gene-structure predictions are necessary.

Nematodes process many mRNAs by *trans*-splicing to SL1 and other splice leader sequences [57,58] and in *C. elegans* use of different splice leaders is tied to genome organization in operons [59]. SL1 is the predominant nematode splice leader and is highly conserved across many species. Use of SL1 by transcripts is estimated at 70% in *C. elegans* [60], more than 80% in *Ascaris lumbricoides* [61], and approximately 60% in *G. rostochiensis* (Ling Qin, personal communication). SL1 has previously been observed in *M. incognita* [12], although genes with non *trans*-spliced 5′ ends have also been cloned [5,6]. Only 33 of our *M. incognita* contigs have an SL1 sequence at their 5′ end. This limited detection of SL1 is not surprising as both the poor processivity of reverse transcriptase and the positioning of the vector sequence primer near the beginning of the insert result in

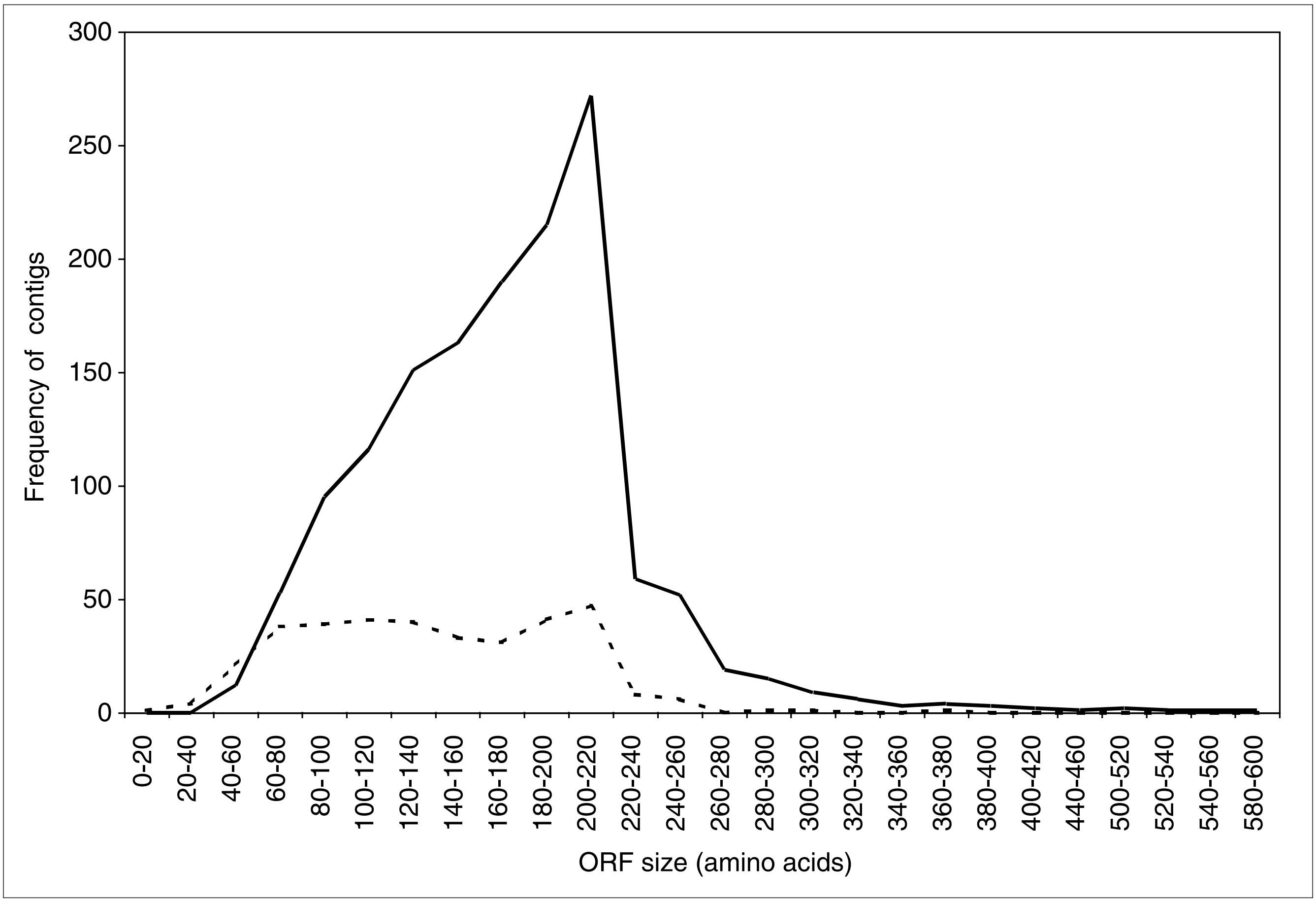

Figure 4
Distribution of contigs by size of longest ORF. Solid line, contigs with any database homology by BLASTX (1,445). Dotted line, contigs without database homology (353).

low representation of the initial 5′ nucleotides of a transcript among EST collections. As an alternative method of determining which *M. incognita* genes may have an SL1 splice leader, contigs were compared by BLASTN to our recently sequenced ESTs from a *M. arenaria* egg library produced by PCR with an SL1 primer sequence. Of the *M. incognita* contigs 188 had high-level nucleotide identity (better than 1e-30) to this collection of SL1-containing *Meloidogyne* genes. With ESTs now available in our collection from four *Meloidogyne* and numerous other SL1-PCR cDNA libraries [32], it should be possible to address whether or not SL1-splicing of individual genes is conserved across nematode species.

Comparison to *C. elegans* genes with known RNAi phenotypes

The technique of RNAi, whereby the introduction of a sequence-specific double-stranded RNA leads to degradation of matching mRNAs [62], has allowed the systematic surveying of thousands of *C. elegans* genes for phenotypes following transient gene knockout [63-65]. Such information is potentially transferable to understanding which genes have crucial roles in parasitic nematodes where high-throughput RNAi is not yet possible. A list of 7,212 *C. elegans* RNAi experiments surveying 4,786 genes was compared to the list of all *M. incognita* clusters with significant homology to *C. elegans* proteins. Using the criterion that the *C. elegans* gene was the best match available for one of the *Meloidogyne* clusters and RNAi experimental information was available, 539 genes were revealed. A specific phenotype by RNAi was apparent for 221 (41%) of these genes, whereas 318 (59%) remained wild type (see Additional data files for the complete list of *C. elegans* RNAi phenotypes for genes with *M. incognita* homologs). By comparison, RNAi surveys of all predicted genes on a *C. elegans* chromosome have found a smaller percentage of genes with phenotypes: 14% for chromosome I [63] and 13% for chromosome III [64]. Surveys of expressed genes reveal an intermediate level of 27% with phenotypes [65]. Further, selecting for *C. elegans* genes with expressed *Meloidogyne* homologs led to enrichment for genes with severe phenotypes by RNAi such as embryonic lethality or sterility as compared to the overall dataset (Figure 5) (For a complete

Table 5

Most conserved nematode genes between *M. incognita* and *C. elegans*

M. incognita cluster/contig*	ESTs per cluster	Wormpep accession	*C. elegans* gene	Assignment	E-value
MI00487.cl / MI01030	44	CE13150	T04C12.5	ACT-2, actin 2	1e-240
MI00951.cl / MI01122	77	CE20658	F08B6.4	UNC-87, calponin	1e-193
MI00892.cl / MI00892	7	CE02619	F10C1.2	Intermediate filament protein	1e-180
MI00666.cl / MI00666	4	CE07537	T25F10.6	Calponin like protein	2e-155
MI00750.cl / MI00805	5	CE12204	K12F2.1	MYO-3, myosin heavy chain	1e-148
MI00701.cl / MI00820	4	CE03403	F52H3.7	LEC-2, galactoside-binding lectin	8e-143
MI00590.cl / MI00661	3	CE18478	B0250.1	Ribosomal protein L2	4e-134
MI00081.cl / MI00081	2	CE09349	F11C3.3	UNC-54, myosin heavy chain	3e-127
MI00721.cl / MI01033	4	CE02181	F13D12.2	LDH-1, l-lactate dehydrogenase	4e-125
MI01008.cl / MI01008	16	CE25005	F54H12.1	Aconitate hydratase	5e-122
MI00918.cl / MI00918	8	CE15900	F25H5.4	EFT-2, elongation factor Tu family	2e-119
MI01789.cl / MI01789	1	CE25977	T01A4.1	Guanylyl cyclase	8e-119
MI01065.cl / MI01065	4	CE00664	F56F3.5	Ribosomal protein S3a	8e-117
MI00900.cl / MI00900	7	CE16333	T03E6.7	cathepsin-like protein	4e-115
MI00792.cl / MI00792	5	CE03684	T21B10.2	Enolase	7e-115
MI00809.cl / MI00809	6	CE03368	F49C12.8	RPN-7, proteasome regulatory particle	9e-115

*Contig shown is the consensus sequence within the cluster which generated the most significant E-value score.

tally of all observed phenotypes see Additional data files). A correlation between sequence conservation and severe phenotype by RNAi had previously been shown by comparison of *C. elegans* to genomes from the distant phyla *Saccharomyces*, *Drosophila* and human [63,64]. Here we show a similar trend following detection of homology to expressed genes in other nematode species. Applying RNAi techniques directly to parasitic nematodes is challenging owing to the organisms' generally longer and more complex life cycles, including the requirement for passage through a host organism. Progress has been made recently in assaying RNAi effects in both plant [66] and animal [67] parasitic nematodes. Further success may allow for a more high-throughput examination of phenotypes resulting from transient gene knockout in parasites.

Tylenchida-specific genes and horizontal gene transfer candidates

Fifty-three *M. incognita* clusters showed homology to sequences from other nematode species yet lacked either *C. elegans* or non-nematode homologs. Twenty of these clusters showed conservation only to gene products from other *Tylenchida* species. MI00244.cl, for example, had homology to 47 ESTs in our collection from other *Tylenchida* species including root-knot nematodes *M. javanica*, *M. hapla* and *M. arenaria*, cyst nematodes *H. glycines* and *G. rostochiensis*, and the lesion nematode *Pratylenchus penetrans* with E-values from 7e-78 to 3e-05. The best homology to any *C. elegans* protein was an extremely weak match (E-value = 0.017) to hypothetical protein M01H9.3b. Genes in this collection may be rapidly evolving so that homologs are only detected in closely related species. Alternatively, genes may be special adaptations to plant parasitism. No annotation is available for any of these genes, but alignments with sequences from related species can define domains for further characterization.

In 1998, it was discovered that plant parasitic nematodes possess genes encoding beta-1,4-endoglucanase enzymes (cellulases) and that by far the strongest non-*Tylenchida* homologs for these enzymes were prokaryotic cellulases from *Pseudomonas*, *Clostridium* and other microbes. Following isolation from *G. rostochiensis* and *H. glycines* [5], cellulases have been identified in *M. incognita* [6], *G. tabacum* [68], *H. schachtii* [69], and *P. penetrans* [70]. Additional prokaryotic-like sequences identified in plant parasitic nematodes include other cell-wall-degrading enzymes such as xylanase [7], pectate lyase [8,71] and polygalacturonase [72], and evidence is accumulating that these sequences have been acquired by horizontal gene transfer [11]. The known *Meloidogyne* cellulase (MI00483.cl), potentially novel cellulases (MI00537.cl, MI01196.cl, MI01381.cl, MI01842.cl), and pectate lyase (MI00592.cl, MI00520.cl) were represented in the *M. incognita* EST clusters.

MI01045.cl, the seventh largest *Meloidogyne* EST cluster, is a new horizontal gene transfer candidate with homology to nodL acetyltransferase from *Rhizobium leguminosarum* (1e-53). Nod factor is responsible for the induction of nodules in nitrogen-fixing plants and nodL has an essential role in Nod factor biosynthesis [73]. Experimental demonstration of a *trans*-spliced leader on the *Meloidogyne nodL* mRNA and

comment
reviews
reports
deposited research
refereed research
interactions
information

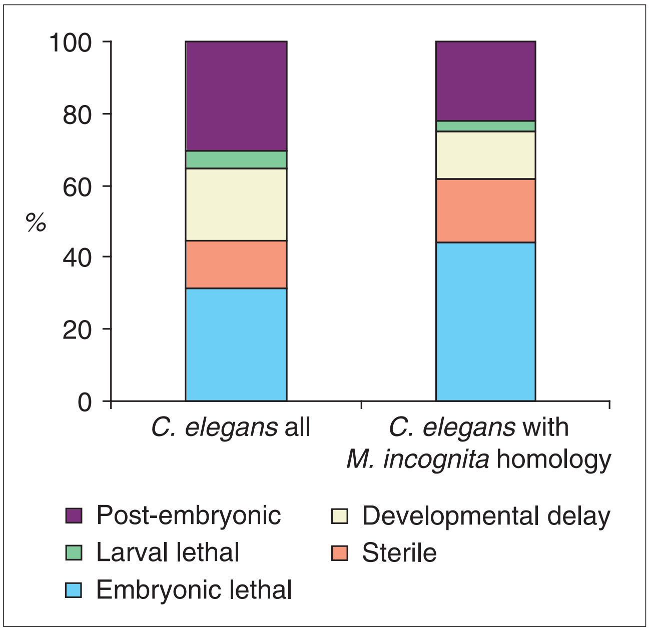

Figure 5
A comparison of phenotype distribution between all RNAi-surveyed *C. elegans* genes with phenotypes (4,786) versus only those *C. elegans* genes with homology to *M. incognita* (221).

the presence of introns in the gene confirm that it is not a bacterial contaminant and more extensive characterization is underway (E.H. Scholl, J.L. Thorne, J.P.M. and D.M.B., unpublished work). It is possible that root-knot nematodes have adapted a portion of Nod factor biology to the induction of feeding sites, rather than nodules, in plants.

To identify further horizontal gene transfer candidates from the *M. incognita* EST clusters, the subset of clusters with homology to sequences in other *Tylenchida* and in non-nematodes but not in non-*Tylenchida* nematodes were examined. In addition to those sequences already characterized, four additional clusters of interest were identified. MI00109.cl shows homology to a group of hypothetical proteins from alpha-proteobacteria: *Sinorhizobium meliloti* NP_386252 (3e-44); *Novosphingobium aromaticivorans* ZP_00095448 (3e-38); *Mesorhizobium loti* NP_107072 (5e-37). The finding of multiple *Tylenchida* genes with close homologs in rhizobacteria suggests the possibility of horizontal transfer of cassettes of genes or multiple transfer events between nitrogen-fixing soil bacteria and plant parasitic nematodes. MI01406.cl and MI00267.cl show homology to two hypothetical proteins from the Actinomycetales - *Amycolatopsis mediterranei* CAC42207 (5e-29) and *Streptomyces lavendulae* AAD32751 (2e-24). Providing some clue to function, the clusters as well as the hypothetical proteins are more distant homologs (1e-05 to 1e-08) of a putative riboflavin aldehyde-forming enzyme from *Agaricus bisporus*, CAB85691 (D.C. Eastwood, GenBank direct submission, 2000), an annotation based on homology (5e-05) to the characterized enzyme from *Schizophyllum commune* [74]. A weak but common motif between all of the proteins is discernible.

Conclusions

As recently as February 2000 only 22 ESTs from plant parasitic nematodes had been deposited in dbEST. As of October 2002, that number has risen to 46,876, including 42,210 from Washington University and collaborators. Included are 32,735 sequences from *Meloidogyne* species (*M. incognita* 12,752, *M. hapla* 11,049, *M. javanica* 5,600, *M. arenaria* 3,334), as well as ESTs from cyst nematode species (*G. rostochiensis* 5,934, *H. glycines* 4,327, *G. pallida* 1,832), and the lesion nematode (*P. penetrans* 2,048). The majority of these sequences have been isolated from L2 and egg libraries, but sequencing from more diverse stages is now underway.

The only previous analysis of root knot nematodes ESTs [29] used 914 ESTs from *M. incognita* L2 without clustering and with non-automated assignment of genes to categories. The two datasets share some overlap, with 35% (316/914) of the previously analyzed ESTs finding matches in 16% (261/1,625) of the NemaGene clusters analyzed in this paper, many with strong homology (< 1e-40). This overlap was less than expected given the redundancy of the cDNA library analyzed here, at nearly 6,000 ESTs, and suggest that: first, libraries made by different methods are likely to result in different representation from an mRNA pool (either different genes or other portions of the same genes as a result of different 5′ processivity); and second, that *M. incognita* L2 are likely to have a substantial number of unsampled messages awaiting generation of new libraries or library normalization. The semi-automated clustering, sequence homology searching and scripted assignment of sequences to functional categories presented here is a scalable approach to analysis that can be applied to larger datasets.

In addition to applying the approaches presented here to larger and more diverse datasets, further topics in *Meloidogyne* genome analysis have yet to be explored. The availability of ESTs representing different developmental stages of *Meloidogyne* will allow an examination of changes in gene representation between stages, and in turn an understanding of the relative importance of various metabolic processes at different stages of development. EST sequences and their corresponding clones can be further used to study relative expression level between stages and conditions using microarrays [75] and amplified fragment length polymorphism (AFLP) approaches [76]. Contig sequences within clusters can also be compared directly for evidence of alternative splicing, another feature which might correlate with developmental stage. Other topics where bioinformatics analysis of available ESTs can improve current knowledge of *Meloidogyne* molecular biology include the identification of secreted and transmembrane proteins through secretion signal sequence detection [77], the creation of a more accurate codon usage/bias and amino-acid usage tables [78], the identification of conserved genes and pathways used in dauer/infective stages across nematode species [79], the definition and study of nematode-specific domains [55], and improved phylogenies based on sampling from multiple genes [53].

While ESTs do not provide information on genome organization in *Meloidogyne* (no genome sequence or physical map is yet available), they can shed light on the organization of the *C. elegans* genome. For instance, *C. elegans* autosomes are organized into central regions dense with predicted genes, highly expressed genes and known mutants, whereas the chromosome arms contain more repetitive sequences and have a higher meiotic recombination rate [31,80]. By using the expanding collection of ESTs from nematodes at various evolutionary distances from *C. elegans*, the hypothesis that genes on the autosome arms are more rapidly evolving can be tested more systematically. Mapping of ESTs from other nematode species can also detect genes contained in the *C. elegans* genome yet not previously recognized, and therefore missing from Wormpep, as well as recognized genes where not all exons have been correctly predicted.

In conclusion, the 5,713 ESTs analyzed here in 1,625 clusters probably represent 6-10% of the genes in the *M. incognita* genome. This initial study, which will be expanded as further sequences are generated, demonstrates that EST generation is an effective method for the discovery of the new genes in plant parasitic nematodes. Further, functional categorization and comparison to known sequences allows the identification of important biological processes at specific developmental stages as well as unusual sequences, such as horizontal gene transfer candidates.

Materials and methods

Source material and library production

To obtain *M. incognita* L2 larvae, a population of nematodes maintained on Rutgers tomato were harvested, eggs were isolated and hatched by standard protocols [81]. Briefly, galled roots were removed from sandy soil, rinsed, and shaken in 15% bleach for 3 min to break roots and free egg masses. Contents were filtered with a large excess of water through a No. 200 sieve to remove root and soil fragments, and a No. 500 sieve to retain nematode eggs. Decanted eggs in small volume were applied above a 40% sucrose solution in a 50 ml conical tube and spun at 2,000 rpm for 10 min. Eggs banded at the sucrose/water interface and were removed by pipette. Following rinsing, sucrose banding was repeated. Harvested eggs were hatched over 4 days on top of a moist filter paper barrier (3 Crown Shopmaster heavy-duty wipes). Hatched larvae migrated through the paper and were collected in a water-filled petri dish below. By microscopic examination, collected worms were predominantly live moving L2, but rare dead L2 and eggs could be found. Total RNA was isolated from collected L2 by the Trizol method (Life Technologies, Gaithersburg, MD) with a yield of 380 μg from around 1 ml of packed L2 worms. Poly(A)$^+$ RNA was isolated from total RNA using the Promega Isolation System II (Promega, Madison, WI) following the manufacturer's instructions with a yield of 4.04 μg. The cDNA library (named Bird_Rao_*Meloidogyne_incognita*_J2) was constructed using the Zap Express cDNA Synthesis Kit and Gigapack III Gold Cloning Kit, 200403 (Stratagene, Cedar Creek, TX). Inserts were directionally cloned between an *Eco*RI site (5′) and a *Xho*I site (3′); however, sequencing indicates that ~22% of clones are in reverse orientation. The non-directionality of the library does not interfere with either clustering or homology detection as both orientations are examined. The titer of the non-amplified phage library was 70,000 recombinants. In preparation for high-throughput sequencing the pBK-CMV phagemid was excised in bulk from the Zap Express phage using the ExAssist Interference-Resistant Helper Phage protocol 211203 (Stratagene). Resulting plasmids were replicated in the helper phage-resistant host cell XLOLR with kanamycin selection. It is expected that the majority of messages in this whole-animal library derive from the tissues that make up most of the mass of the L2 animal including hypodermis/cuticle, intestine, muscle, esophageal and rectal gland, and esophagus/pharnyx [82].

Sequence production and dbEST submission

Clone processing and sequencing was performed as in Hiller *et al.* [83] with some modifications. Single bacterial colonies from the plasmid library were picked from agar trays into 384-well plates containing media, kanamycin, and 7% glycerol using a Q-bot robotic colony picker (Genetix, Christchurch, UK). Plates were incubated overnight at 37°C and stored at -80°C. To prepare template plasmid DNA from each sample, bacterial inoculates were transferred from 384-well storage to 96-well growth blocks containing 1 ml medium per sample and grown overnight. All subsequent sample and reagent transfers were done using a stationary 96-channel Hydra (Robbins Scientific, Sunnyvale, CA). DNA isolation was performed using a fast and inexpensive microwave-based protocol [84]. Sequencing reactions using the T3 (5′) primer employed BigDye terminator chemistry (Applied Biosystems, Foster City, CA) and the cycle sequencing reactions were performed with 96 x 4-block thermocyclers (MJ Research, Waltham, MA). Samples were loaded on ABI377 (96-lane slab gel) sequencers (Applied Biosystems).

Following gel image analysis and DNA sequence extraction, sequence data were processed in an automated pipeline to: assess EST quality; trim flanking vector sequences; mask repetitive elements; remove contaminated ESTs; identify similarities by BLAST; identify cloning artifacts; and determine which portion of the EST to submit [83]. The resulting sequences were annotated with similarity information and sequence quality information and submitted to dbEST. Clones are named for their 96-well plate identity and position during processing (for instance ra40e04.y1). Names are mapped to stored clone location in 384-well plate format. Clones can be ordered at [85]. From 7,818 attempted reads, 5,854 sequences (75%) passed quality and contamination filters and were submitted to dbEST [86]. Most submissions (5,713) were made between March and June of 2000. An additional 141 ESTs originally failed as bacterial contaminants (by an overly inclusive filter) have since been submitted (September 2001), but are not included in this analysis. EST sequences are available from GenBank, EMBL and DDJB under the accession numbers AW440989-AW441125, AW570643-AW571393, AW588598-AW588988, AW589050-AW589115, AW735503-AW735730, AW782981-AW783662, AW827629-AW830045, AW870657-AW871697, and BI-773381-BI773521. Submissions total approximately 2.8 million nucleotides.

A failure rate of 25% is typical for high-throughput sequencing and resulted from poor overall trace quality (~21% of all reads), missing insert (~0.3%), small insert size (~0.06%), and *E. coli* contamination (~0.1%). To further exclude bacterial contamination we have closely examined cases where strong amino-acid homology to prokaryotic genes is observed (see Horizontal gene transfer candidates). Many of these genes have already been confirmed as of *M. incognita* origin by cloning from genomic DNA, *in situ* localization and the finding of homologs in other *Tylenchida* nematodes. In all of these cases, the high level of identity observed at the amino-acid level does not extend to nucleotide level, and GC content and codon usage is typical of other *M. incognita* transcripts (E.H. Scholl, J.L. Thorne, J.P.M. and D.M.B., unpublished work).

To estimate the number of 5′ versus 3′ reads, we examined the 4,198 ESTs with detectable homology on either sense or antisense strands at time of submission (BLASTX search versus the SWIR non-redundant protein database, Sanger Centre). Most ESTs (78%) showed translated amino-acid homology consistent with sequencing from the 5′ end of the transcript, while 22% showed homology consistent with 3′ end sequencing. The mean submitted read length was 481 nucleotides with a standard deviation of 108. Longest and shorted submitted reads were 49 and 780, respectively. Since our submission filter includes a quality cut-off at the distal end of the read (Phred Score < 12 [87,88]), additional sequence can sometimes be obtained by direct examination of the sequencing trace available at [89].

Clustering for NemaGene *Meloidogyne incognita* v 2.0

Clustering was performed by first building 'contigs' of ESTs with identical or nearly identical overlapping sequence and second, by bringing together related contigs to form 'clusters'. Contig member ESTs should all derive from identical transcripts whereas cluster members might derive from the same gene yet represent different transcript splice isoforms or transcripts from multigene families with extremely high sequence identity. The raw traces for submitted ESTs were base-called using Phred [87] and assembled to form contigs using Phrap (P. Green, personal communication). Although Phrap is a program intended for genome assembly, it has been applied previously to ESTs with modifications [90]. To determine initial assembly quality, the largest contigs were inspected using the assembly viewer Consed [91]. Misassemblies bringing unrelated ESTs together into giant contigs usually resulted from the alignment of long poly(A) tails. To eliminate these assemblies of otherwise dissimilar ESTs, Phrap parameters (forcelevel 1, minmatch 20 and minscore 100) were adjusted and Phrap was rerun.

Once acceptable assembly parameters were obtained, Phrap was run to generate a first-draft assembly. Contigs with only one member EST (singletons) were removed from consideration until the trimming and cluster building stage. All contigs with more than three member ESTs was screened for misassemblies using Consed tools and newly written scripts. Misassemblies were recognized by: regions of high quality unaligned sequence; multiple runs of poly(A) and/or poly(T) (at least 15 nucleotides with no more than a one non-A/T base); internal poly(A) and/or poly(T) runs (> 50 nucleotides from either end of a contig and ≥ 15 or more nucleotides long with no more than one non-A/T base; internal stretches of low consensus quality (> 30 nucleotides from either end of a contig and ≥ 50 nucleotides where 90% of the nucleotides had a consensus quality below Phred 20). Contigs flagged for possible misassembly were manually edited in Consed and potentially chimeric ESTs and other suspect ESTs were identified and removed from the pool of traces. Chimerism can result from multiple-insert cloning or mistracking of sequence gel lanes. The project was reassembled with Phrap and screened again as above. All contigs with more than three members were examined again in Consed to eliminate additional misassemblies not resolved by the initial screens. In total, around 450 contigs were examined manually and around 200 were edited. For each contig, a consensus sequence of all EST members was generated. Contigs (now including singleton EST contigs) were then trimmed to high quality and any internal consensus position with a calculated quality value below 12 was changed to an N (unknown base).

Following the creation of contigs by Phrap, the contig consensus sequences were compared using WU-BLASTN (G = 2 E = 1 v = 100 F = F) [92,93] and grouped on the basis of similarity to form clusters of related contigs. Contigs with overlaps of 100 bases or more with nucleotide-nucleotide identities of 93% or more were clustered together. For further analysis, new assemblies based on clusters were not formed; rather, each cluster retained all the consensus sequences of its contig members. NemaGene *Meloidogyne incognita* v 2.0 represents our second complete attempt at generating clusters for this species and is used as the basis for all subsequent analysis in this manuscript. Scripts have been written to allow the addition of new data while retaining the original contig and cluster naming scheme. Additional NemaGene versions of *M. incognita* will be built as additional ESTs become available for the species. A comparison of the NemaGene clustering approach to other EST clustering methods will be considered in a separate manuscript. NemaGene *Meloidogyne incognita* v 2.0 is available for searching at [94] and FTP at [95].

Sequence analysis

Following clustering, comparative analyses were performed using WU-BLASTX and WU-TBLASTX [92,93] with 1,798 contig consensus sequences (themselves grouped into 1,625 cluster groups) as queries versus multiple databases including SWIR v.21 (5/19/2000) non-redundant protein database and Wormpep v.54 *C. elegans* protein database (Wellcome Trust Sanger Institute, unpublished work), *C. elegans* mitochondrial protein sequences, and six internally constructed databases using intersections of data from the GenBank nucleotide database and dbEST [96]. These include: nemnoele (all nucleotide data from the phylum *Nematoda* with *C. elegans* removed); nemnoelenomi (nemnoele with *M. incognita* removed); nemnoelenomel (nemnoele with all *Meloidogyne* species removed); nemnoelenotyl (nemnoele with all *Tylenchida* species removed); yestylnomel (all *Tylenchida* species except *Meloidogyne*); mj (only *M. javanica* sequences). An additional database, nrnonem, is an amino-acid database of all non-nematode proteins derived from SWIR v.21. WU-BLASTX (translated nucleotide query versus protein database) parameters were S = 100 M = PAM120 V =0 W = 4 T = 17. WU-TBLASTX (translated nucleotide query versus translated nucleotide database, each in all six reading frames) parameters were Q = 10 R = 2 gapw = 10. Homologies were reported for e-value scores of 1e-5 and better. By creating intersections of various database search results, contigs/clusters could be organized by their distribution of homologies (for example, clusters which have *M. javanica* matches but not *C. elegans* matches). Data analysis was performed in a Unix environment using Perl and Bourne shell scripts. The program ESTFreq (W. Gish, personal communication) was used to estimate novel sequences expected from a second sampling and the program Translate (S. Eddy, personal communication) was used to translate nucleotide consensus sequences for ORF analysis.

Functional assignments

To assign putative functions to clusters, the integrated protein domain recognition program InterProScan [97,98] was run locally to search translated contig consensus sequences versus all InterPro protein domains (as of 2 April 2002) [99]. The Prosite, Prints, Pfam, ProDom, and Smart search components of InterProScan were used with default parameters. The GO categorization scheme (go_200205-assocdb.sql) of classification by biological process, cellular component, and molecular function was used to classify clusters based on the existing mappings of InterPro domains to the GO hierarchy [36]. Mappings were stored in a local MySQL database and displayed using the AmiGO browser (16 May 2002) [100] (*M. incognita* mappings at [101]).

As an alternative means of assigning function to clusters, clusters were also assigned to metabolic pathways using KEGG [102,103]. Assignments were made by requiring that the highest-scoring BLAST match in SWIR v.21 have an assigned enzyme commission (EC) number [104]. EC number mappings to KEGG pathways were then used to putatively assign clusters into biochemical pathways. Nonspecific pathway mappings (for example, kinases, EC 2.7.1.-) were eliminated, as were misleading pathway assignments (for example, plant carbon fixation, KEGG 2.3, where the assigned protein had only a peripheral 'feed-in' role in the pathway). Assignments were not made to KEGG regulatory pathways as proteins in these pathways lack EC numbers.

C. elegans homologs with RNAi phenotype

To identify cases where *M. incognita* and *C. elegans* share homologous genes which have been surveyed in *C. elegans* for knockout phenotype using RNAi, a list of all 7,212 available *C. elegans* RNAi experiments (5 May 2002) from WormBase [56] was compared to the list of all *M. incognita* clusters with significant homology matches to the *C. elegans* Wormpep v.54 protein database. Redundant RNAi experiments were removed to consolidate the WormBase list to 6,107 and experiments performed on the same gene with different phenotypic outcomes were consolidated later. For any given *M. incognita* cluster, only the best *C. elegans* matches, ranked by BLAST score, were considered.

Nematode origin of the cDNA sequences

To insure that sequences generated originate from *M. incognita* and are not contaminants, multiple steps purifying material and cross-checking sequence origin have been incorporated into the project: the starting material is purified and freed of plant material; poly(A) selection during library production is highly selective for eukaryotic transcripts, though it is possible for AT-rich prokaryotic transcripts to be cloned; analyzed sequences have been filtered for prokaryotic homology resulting in the removal of eight *E. coli* contaminants (0.14%), a typical background for cDNA cloning; 96% of the clusters with detectable homology have nematode homologs (1,227/1,280), 17% have only nematode homology, and in the vast majority of cases, higher conservation is seen to a nematode sequence than any non-nematode sequence; additional confirmation of nematode origin comes from the presence of an SL1 *trans*-spliced leader sequence on some genes; all sequences with strong amino-acid homology to prokaryotic genes were closely examined and in no cases were the high levels of identity maintained at the nucleotide level (as would be the case with a contaminating sequence). While it can be stated with confidence that the vast majority of the sequence analyzed originates from *M. incognita*, we cannot rule out the possibility that the collection may include a very small number of contaminating sequences.

Additional data files

The following files are available as additional data with the online version of this article: a complete listing of gene ontology mappings for *M. incognita* clusters organized into (a) biological process, (b) cellular component, (c) molecular function (Additional data file 1); complete KEGG biochemical pathway mappings for *M. incognita* clusters (Additional data file 2); a complete list of *C. elegans* RNAi phenotypes for genes with *M. incognita* homologs (Additional data file 3); classification by RNAi phenotype of *C. elegans* genes with *M. incognita* homologs (Additional data file 4).

Acknowledgements

Meloidogyne EST sequencing at Washington University was supported by NIH-NIAID research grant AI 46593 to R.W. and NSF Plant Genome award 0077503 to D.B. and S.C. J.M. was supported by a Helen Hay Whitney/Merck Postdoctoral Fellowship. We thank John Spieth, Jeremy Williams and Ian Korf for helpful databases and scripts, Jennifer Schaff for technical support and The Institute for Genomic Research for use of TIGR gene index information. J.M., A.K. and B.C. are employees and equity holders of Divergence Inc. This research was not funded by Divergence Inc.

References

1. Agrios GN: **Plant diseases caused by nematodes.** In *Plant Pathology.* Edited by Agrios GN. New York: Academic Press; 1997: 565-597.
2. Sasser JN, Freckman DW: **A world perspective on nematology: the role of the society.** In *Vistas on Nematology.* Edited by Veech JA, Dickson DW. Hyattsville: Society of Nematology; 1987: 7-14.
3. Taylor AL, Sasser JN: *Biology, Identification and Control of Root-knot Nematodes (*Meloidogyne *species).* Raleigh, NC: United States Agency for International Development; 1978.
4. Wyss U, Grundler FMW, Munch A: **The parasitic behaviour of second-stage juveniles of *Meloidogyne incognita* in roots of *Arabidopsis thaliana*.** *Nematologica* 1992, **38:**98-111.
5. Smant G, Stokkermans JP, Yan Y, de Boer JM, Baum TJ, Wang X, Hussey RS, Gommers FJ, Henrissat B, Davis EL, *et al.*: **Endogenous cellulases in animals: isolation of beta-1, 4-endoglucanase genes from two species of plant-parasitic cyst nematodes.** *Proc Natl Acad Sci USA* 1998, **95:**4906-4911.
6. Rosso MN, Favery B, Piotte C, Arthaud L, De Boer JM, Hussey RS, Bakker J, Baum TJ, Abad P: **Isolation of a cDNA encoding a beta-1,4-endoglucanase in the root-knot nematode *Meloidogyne incognita* and expression analysis during plant parasitism.** *Mol Plant Microbe Interact* 1999, **12:**585-591.
7. Dautova M: **Population and molecular genetics of root-knot nematodes.** *PhD thesis.* Wageningen: Wageningen University and Research Centre; 2001.
8. Popeijus H: **The identification of cell wall degrading enzymes in *Globodera rostochiensis*.** *PhD thesis.* Wageningen: Wageningen University and Research Centre; 2002.
9. Bird DM, Koltai H: **Plant parasitic nematodes: habitats, hormones, and horizontally-acquired genes.** *J Plant Growth Reg* 2000, **19:**183-194.
10. Bird DM: **Manipulation of host gene expression by root-knot nematodes.** *J Parasitol* 1996, **82:**881-888.
11. Davis EL, Hussey RS, Baum TJ, Bakker J, Schots A, Rosso MN, Abad P: **Nematode parasitism genes.** *Annu Rev Phytopathol* 2000, **38:**365-396.
12. Ray C, Abbott AG, Hussey RS: **Trans-splicing of a *Meloidogyne incognita* mRNA encoding a putative esophageal gland protein.** *Mol Biochem Parasitol* 1994, **68:**93-101.
13. Van der Eycken W, de Almeida Engler J, Van Montagu M, Gheysen G: **Identification and analysis of a cuticular collagen-encoding gene from the plant-parasitic nematode *Meloidogyne incognita*.** *Gene* 1994, **151:**237-242.
14. Milner RJ, Sutcliffe JG: **Gene expression in rat brain.** *Nucleic Acids Res* 1983, **11:**5497-5520.
15. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, *et al.*: **Complementary DNA sequencing: expressed sequence tags and human genome project.** *Science* 1991, **252:**1651-1656.
16. Waterston R, Martin C, Craxton M, Huynh C, Coulson A, Hillier L, Durbin R, Green P, Shownkeen R, Halloran N, *et al.*: **A survey of expressed genes in *Caenorhabditis elegans*.** *Nat Genet* 1992, **1:**114-123.
17. McCombie WR, Adams MD, Kelley JM, FitzGerald MG, Utterback TR, Khan M, Dubnick M, Kerlavage AR, Venter JC, Fields C: ***Caenorhabditis elegans* expressed sequence tags identify gene families and potential disease gene homologues.** *Nat Genet* 1992, **1:**124-131.
18. Blaxter ML, Raghavan N, Ghosh I, Guiliano D, Lu W, Williams SA, Slatko B, Scott AL: **Genes expressed in *Brugia malayi* infective third stage larvae.** *Mol Biochem Parasitol* 1996, **77:**77-93.
19. Moore TA, Ramachandran S, Gam AA, Neva FA, Lu W, Saunders L, Williams SA, Nutman TB: **Identification of novel sequences and codon usage in *Strongyloides stercoralis*.** *Mol Biochem Parasitol* 1996, **79:**243-248.
20. Tetteh KK, Loukas A, Tripp C, Maizels RM: **Identification of abundantly expressed novel and conserved genes from the infective larval stage of *Toxocara canis* by an expressed sequence tag strategy.** *Infect Immun* 1999, **67:**4771-4779.
21. Daub J, Loukas A, Pritchard DI, Blaxter M: **A survey of genes expressed in adults of the human hookworm, *Necator americanus*.** *Parasitology* 2000, **120:**171-184.
22. Hoekstra R, Visser A OM, Tibben J, Lenstra JA, Roos MH: **EST sequencing of the parasitic nematode *Haemonchus contortus* suggests a shift in gene expression during transition to the parasitic stages.** *Mol Biochem Parasitol* 2000, **110:**53-68.
23. Lizotte-Waniewski M, Tawe W, Guiliano DB, Lu W, Liu J, Williams SA, Lustigman S: **Identification of potential vaccine and drug target candidates by expressed sequence tag analysis and immunoscreening of *Onchocerca volvulus* larval cDNA libraries.** *Infect Immun* 2000, **68:**3491-3501.
24. Blaxter M, Aslett M, Guiliano D, Daub J: **Parasitic helminth genomics. Filarial Genome Project.** *Parasitology* 1999, **118:**S39-S51.
25. McCarter JP, Bird DM, Clifton SW, Waterston RH: **Nematode gene sequences, December 2000 Update.** *J Nematol* 2000, **32:**331-333.
26. Parkinson J, Whitton C, Guiliano D, Daub J, Blaxter M: **200000 nematode expressed sequence tags on the Net.** *Trends Parasitol* 2001, **17:**394-396.
27. McCarter JP, Clifton S, Bird DM, Waterston RH: **Nematode gene sequences, Update for June 2002.** *J Nematol* 2002, **34:**71-74.
28. McCarter J, Abad P, Jones JT, Bird D: **Rapid gene discovery in plant parasitic nematodes via Expressed Sequence Tags.** *Nematology* 2000, **2:**719-731.
29. Dautova M, Rosso MN, Abad P, Gommers FJ, Bakker J, Smant G: **Single pass cDNA sequencing - a powerful tool to analyse gene expression in preparasitic juveniles of the southern root-knot nematode *Meloidogyne incognita*.** *Nematology* 2001, **3:**129-139.
30. Popeijus H, Blok VC, Cardle L, Bakker E, Phillips MS, Helder J, Smant G, Jones JT: **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* 2000, **2:**567-574.
31. The *C. elegans* Sequencing Consortium: **Genome sequence of the nematode *C. elegans*: a platform for investigating biology.** *Science* 1998, **282:**2012-2018.
32. **Nematode.net Genome Sequencing Center** [http://www.nematode.net]
33. Audic S, Claverie JM: **The significance of digital gene expression profiles.** *Genome Res* 1997, **7:**986-995.
34. Prior A, Kennedy MW, Blok VC, Robertson WM, Jones JT: **Functional characterisation of a secreted protein from the potato cyst nematode *Globodera pallida*.** *Nematologica* 1998, **44:**514.
35. **Gene Ontology Consortium** [http://www.geneontology.org]

36. The Gene Ontology Consortium: **Gene ontology: tool for the unification of biology.** *Nat Genet* 2000, **25:**25-29.
37. **InterPro** [http://www.ebi.ac.uk/interpro]
38. **The Institute for Genomic Research: TIGR gene indices** [http://www.tigr.org/tdb/tgi]
39. Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R, White J: **The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species.** *Nucleic Acids Res* 2001, **29:**159-164.
40. Ding X, Shields J, Allen R, Hussey RS: **Molecular cloning and characterization of a venom allergen AG5-like cDNA from *Meloidogyne incognita*.** *Int J Parasitol* 2000, **30:**77-81.
41. Bird DM, Opperman CH: ***Caenorhabditis elegans*: a genetic guide to parasitic nematode biology.** *J Nematol* 1998, **30:**299-308.
42. Blaxter M: **Genes and genomes of *Necator americanus* and related hookworms.** *Int J Parasitol* 2000, **30:**347-355.
43. **Kyoto Encyclopedia of Genes and Genomes** [http://www.genome.ad.jp/kegg/kegg2.html]
44. Vanfleteren JR: **Nematodes as nutritional models.** In *Nematodes as Biological Models*. Edited by Zuckerman BM. vol. II. New York: Academic Press; 1980: 47-77.
45. Nicholas WL, Hansen E, Dougherty EC: **The B-vitamins required by *Caenorhabditis briggsae* (*Rhabditidae*).** *Nematologica* 1962, **8:**129-135.
46. Barrett J, Wright DJ: **Intermediary metabolism.** In *The Physiology and Biochemistry of Free-living and Plant-Parasitic Nematodes.* Edited by Perry RN, Wright DJ. Oxford: CABI Publishing; 1998: 331-353.
47. Reversat G: **Consumption of food reserves by starved second-stage juveniles of *Meloidogyne javanica* under conditions including osmobiosis.** *Nematologica* 1981, **27:**207-214.
48. Wadsworth WG, Riddle DL: **Developmental regulation of energy metabolism in *Caenorhabditis elegans*.** *Dev Biol* 1989, **132:**167-173.
49. Liu F, Thatcher JD, Barral JM, Epstein HF: **Bifunctional glyoxylate cycle protein of *Caenorhabditis elegans*: a developmentally regulated protein of intestine and muscle.** *Dev Biol* 1995, **169:**399-414.
50. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, *et al.*: **Life with 6000 genes.** *Science* 1996, **274:**563-567.
51. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, *et al.*: **The genome sequence of *Drosophila melanogaster*.** *Science* 2000, **287:**2185-2195.
52. Blaxter ML, Daub J, Guiliano D, Parkinson J, Whitton C, Project. FG: **The *Brugia malayi* genome project: expressed sequence tags and gene discovery.** *Trans R Soc Trop Med Hyg* 2002, **96:**7-17.
53. Blaxter ML, De Ley P, Garey JR, Liu LX, Scheldeman P, Vierstraete A, Vanfleteren JR, Mackey LY, Dorris M, Frisse LM, *et al.*: **A molecular evolutionary framework for the phylum *Nematoda*.** *Nature* 1998, **392:**71-75.
54. Castillo-Davis CI, Hartl DL: **Genome evolution and developmental constraint in *Caenorhabditis elegans*.** *Mol Biol Evol* 2002, **19:**728-735.
55. Sonnhammer EL, Durbin R: **Analysis of protein domain families in *Caenorhabditis elegans*.** *Genomics* 1997, **46:**200-216.
56. **WormBase** [http://www.wormbase.org]
57. Krause M, Hirsh D: **A trans-spliced leader sequence on actin mRNA in *C. elegans*.** *Cell* 1987, **49:**753-761.
58. Takacs AM, Denker JA, Perrine KG, Maroney PA, Nilsen TW: **A 22-nucleotide spliced leader sequence in the human parasitic nematode *Brugia malayi* is identical to the trans-spliced leader exon in *Caenorhabditis elegans*.** *Proc Natl Acad Sci USA* 1988, **85:**7932-7936.
59. Blumenthal T, Evans D, Link CD, Guffanti A, Lawson D, Thierry-Mieg J, Thierry-Mieg D, Chiu WL, Duke K, Kiraly M, *et al.*: **A global analysis of *Caenorhabditis elegans* operons.** *Nature* 2002, **417:**851-854.
60. Blumenthal T, Steward K: **RNA processing and gene structure.** In *C. elegans.* Edited by Riddle DL, Blumenstiel B, Meyer BJ, Priess JR. Plainview: Cold Spring Harbor Laboratory Press; 1997: 117-145.
61. Nilsen TW: **Trans-splicing of nematode premessenger RNA.** *Annu Rev Microbiol* 1993, **47:**413-440.
62. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC: **Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*.** *Nature* 1998, **391:**806-811.
63. Fraser AG, Kamath RS, Zipperlen P, Martinez-Campos M, Sohrmann M, Ahringer J: **Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference.** *Nature* 2000, **408:**325-330.
64. Gonczy P, Echeverri C, Oegema K, Coulson A, Jones SJ, Copley RR, Duperon J, Oegema J, Brehm M, Cassin E, *et al.*: **Functional genomic analysis of cell division in *C. elegans* using RNAi of genes on chromosome III.** *Nature* 2000, **408:**331-336.
65. Maeda I, Kohara Y, Yamamoto M, Sugimoto A: **Large-scale analysis of gene function in *Caenorhabditis elegans* by high-throughput RNAi.** *Curr Biol* 2001, **11:**171-176.
66. Urwin PE, Lilley CJ, Atkinson HJ: **Ingestion of double-stranded RNA by preparasitic juvenile cyst nematodes leads to RNA interference.** *Mol Plant Microbe Interact* 2002, **15:**747-752.
67. Hussein AS, Kichenin K, Selkier PM: **Suppression of selected acetylcholinesterase expression in *Nippostrongylus brasiliensis* by RNA interference.** *Mol Biochem Parasitol* 2002, **122:**91-94.
68. Goellner M, Smant G, de Boer JM, Baum TJ, Davis EL: **Isolation of beta-1,4-endoglucanase genes from *Globodera tabacum* and their expression during parasitism.** *J Nematol* 2000, **32:**154-165.
69. de Meutter J, Tytgat T, van der Schueren E, Smant G, Schots A, Coomans A, van Montagu M, Gheysen G: **Cloning of two endoglucanase genes from *Heterodera schachtii*.** *Med Fac Landbouww Univ Gent* 1998, **63/3a:**619-623.
70. Uehara T, Kushida A, Momota Y: **PCR-based cloning of two beta-1,4-endoglucanases from the root-lesion nematode *Pratylenchus penetrans*.** *Nematology* 2001, **3:**335-341.
71. Doyle EA, Lambert KN: **Cloning and characterization of an esophageal-gland-specific pectate lyase from the root-knot nematode *Meloidogyne javanica*.** *Mol Plant Microbe Interact* 2002, **15:**549-556.
72. Jaubert S, Laffaire JB, Abad P, Rosso MN: **A polygalacturonase of animal origin isolated from the root-knot nematode *Meloidogyne incognita*.** *FEBS Lett* 2002, **522:**109-112.
73. Surin BP, Downie JA: **Characterization of the *Rhizobium leguminosarum* genes nodLMN involved in efficient host-specific nodulation.** *Mol Microbiol* 1988, **2:**173-183.
74. Chen H, McCormick DB: **Riboflavin 5′-hydroxymethyl oxidation. Molecular cloning, expression, and glycoprotein nature of the 5′-aldehyde-forming enzyme from *Schizophyllum commune*.** *J Biol Chem* 1997, **272:**20077-20081.
75. Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, Eizinger A, Wylie BN, Davidson GS: **A gene expression map for *Caenorhabditis elegans*.** *Science* 2001, **293:**2087-2092.
76. Qin L, Overmars H, Helder J, Popeijus H, van der Voort JR, Groenink W, van Koert P, Schots A, Bakker J, Smant G: **An efficient cDNA-AFLP-based strategy for the identification of putative pathogenicity factors from the potato cyst nematode *Globodera rostochiensis*.** *Mol Plant Microbe Interact* 2000, **13:**830-836.
77. Wang X, Allen R, Ding X, Goellner M, Maier T, de Boer JM, Baum TJ, Hussey RS, Davis EL: **Signal peptide-selection of cDNA cloned directly from the esophageal gland cells of the soybean cyst nematode *Heterodera glycines*.** *Mol Plant Microbe Interact* 2001, **14:**536-544.
78. Stenico M, Lloyd AT, Sharp PM: **Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases.** *Nucleic Acids Res* 1994, **22:**2437-2446.
79. Jones SJ, Riddle DL, Pouzyrev AT, Velculescu VE, Hillier L, Eddy SR, Stricklin SL, Baillie DL, Waterston R, Marra MA: **Changes in gene expression associated with developmental arrest and longevity in *Caenorhabditis elegans*.** *Genome Res* 2001, **11:**1346-1352.
80. Barnes TM, Kohara Y, Coulson A, Hekimi S: **Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans*.** *Genetics* 1995, **141:**159-179.
81. Hussey RS, Barker KR: **A comparison of methods of collecting inocula of *Meloidogyne spp.*, including a new technique.** *Plant Dis Reporter* 1973, **57:**1025-1028.
82. Eisenback JD: **Detailed morphology and anatomy of second-stage juveniles, males and females of the genus *Meloidogyne* (root-knot nematodes).** In *An Advanced Treatise on* Meloidogyne*: Biology and Control.* Edited by Sasser JN ,Carter CC. Raleigh: North Carolina State University Graphics; 1985: 47-77.
83. Hillier LD, Lennon G, Becker M, Bonaldo MF, Chiapelli B, Chissoe S, Dietrich N, DuBuque T, Favello A, Gish W, *et al.*: **Generation and analysis of 280,000 human expressed sequence tags.** *Genome Res* 1996, **6:**807-828.
84. Marra MA, Kucaba TA, Hillier LW, Waterston RH: **High-throughput plasmid DNA purification for 3 cents per sample.** *Nucleic Acids Res* 1999, **27:**e37.

85. **Nematode.net Genome Sequencing Center: clone requests** [http://www.nematode.net/Clone.Requests]
86. **NCBI expressed sequence tags database** [http://www.ncbi.nlm.nih.gov/dbEST]
87. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment.** *Genome Res* 1998, **8:**175-185.
88. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities.** *Genome Res* 1998, **8:**186-194.
89. **Nematode.net Genome Sequencing Center: EST data** [http://genome.wustl.edu/traceviewer/traceview.php]
90. Ewing B, Green P: **Analysis of expressed sequence tags indicates 35,000 human genes.** *Nat Genet* 2000, **25:**232-234.
91. Gordon D, Abajian C, Green P: **Consed: a graphical tool for sequence finishing.** *Genome Res* 1998, **8:**195-202.
92. **WU-BLAST** [http://blast.wustl.edu]
93. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
94. **GSC Nematode EST data** [http://www.nematode.net/NemaGene/index.php]
95. **GSC Nematode EST data** [http://www.nematode.net/FTP/index.php]
96. Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA, Wagner L, *et al.*: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2001, **29:**11-16.
97. **InterProScan** [ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan]
98. Zdobnov EM, Apweiler R: **InterProScan - an integration platform for the signature-recognition methods in InterPro.** *Bioinformatics* 2001, **17:**847-848.
99. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD, *et al.*: **The InterPro database, an integrated documentation resource for protein families, domains and functional sites.** *Nucleic Acids Res* 2001, **29:**37-40.
100. **AmiGO** [http://www.godatabase.org/cgi-bin/go.cgi]
101. **GSC GO mappings** [http://genome.wustl.edu/go_meloidogyne_incognita/cgi-bin/go.cgi]
102. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res* 2000, **28:**27-30.
103. Bono H, Ogata H, Goto S, Kanehisa M: **Reconstruction of amino acid biosynthesis pathways from the complete genome sequence.** *Genome Res* 1998, **8:**203-210.
104. IUBMB: *Enzyme Nomenclature: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*. San Diego: Academic Press; 1992.
105. **Swiss-Prot and TrEMBL** [http://us.expasy.org/sprot]