nature genetics

A transcriptomic analysis of the phylum Nematoda

John Parkinson[1,2], Makedonka Mitreva[3], Claire Whitton[2], Marian Thomson[2], Jennifer Daub[2], John Martin[3], Ralf Schmid[2], Neil Hall[4,6], Bart Barrell[4], Robert H Waterston[3,6], James P McCarter[3,5] & Mark L Blaxter[2]

The phylum Nematoda occupies a huge range of ecological niches, from free-living microbivores to human parasites. We analyzed the genomic biology of the phylum using 265,494 expressed-sequence tag sequences, corresponding to 93,645 putative genes, from 30 species, including 28 parasites. From 35% to 70% of each species' genes had significant similarity to proteins from the model nematode *Caenorhabditis elegans*. More than half of the putative genes were unique to the phylum, and 23% were unique to the species from which they were derived. We have not yet come close to exhausting the genomic diversity of the phylum. We identified more than 2,600 different known protein domains, some of which had differential abundances between major taxonomic groups of nematodes. We also defined 4,228 nematode-specific protein families from nematode-restricted genes: this class of genes probably underpins species- and higher-level taxonomic disparity. Nematode-specific families are particularly interesting as drug and vaccine targets.

Nematodes, or roundworms, are a highly diverse group of organisms[1]. What nematodes lack in obvious morphological disparity, they make up for in abundance, accounting for 80% of all individual animals on earth[2], and diversity, with estimates ranging from 100,000 to 1 million extant species[3]. They exploit a wide variety of niches and include free-living terrestrial and marine microbivores, meiofaunal predators, herbivores, and plant and animal parasites. On the basis of small subunit ribosomal RNA (SSU rRNA) phylogenetics[1,4], nematodes can be divided into three major clades: Dorylaimia (clade I)[1,4], Enoplia (clade II) and Chromadorea (which includes Rhabditida, also known as Secernentea). Rhabditida can be further divided into Spirurina (clade III), Tylenchina (clade IV) and Rhabditina (clade V; **Fig. 1**). Parasitism of both animals and plants seems to have arisen multiple times during nematode evolution, and all major clades include parasites.

Most nematode diseases are intractable problems. Infections of humans by nematodes result in substantial human mortality and morbidity, especially in tropical regions of Africa, Asia and the Americas: 2.9 billion people are infected. Morbidity from nematodes is substantial and rivals diabetes and lung cancer in worldwide disability adjusted life year measurements[5]. Although mortality is low in proportion to the huge number of infections, deaths may still total 100,000 annually. The most important parasites include hookworms, *Ascaris* and whipworms (>1 billion infections each) and the filarial nematodes that cause elephantiasis and African river blindness (120 million infections). Parasitic nematodes also cause substantial losses in livestock and companion animals and are responsible for $80 billion in annual crop damage worldwide[6].

Much of what we know about the molecular and developmental biology of nematodes stems from the study of the free-living soil rhabditine nematode *Caenorhabditis elegans* (**Fig. 1**). *C. elegans* is a versatile and tractable model organism, contributing substantially to understanding of important medical fields including cancer, ageing, neurobiology and parasitic diseases[7–9]. *C. elegans* was the first multicellular organism whose genome was completely assembled[7]. Despite the wealth of information available for *C. elegans*, and its sister species *Caenorhabditis briggsae*[10], comparatively little is known about other members of this important phylum.

Two projects were initiated to generate new sequence data for nematode parasites spanning the phylogenetic disparity of the phylum[11]. We used expressed-sequence tags (ESTs; sequences derived from randomly selected cDNA clones) as they are a cost-effective route to gene discovery[12]. We generated 265,494 sequences from 30 different species of nematode, the largest collection of ESTs representing the full diversity of a single phylum. In addition to identifying traits that may be species- or phylum-specific, this collection offers an unparalleled opportunity to explore and elucidate evolutionary and functional relationships. Here we present an overview of the sequence data arising from the parasitic nematode EST project and place them in the context of *C. elegans* genomic biology.

RESULTS

265,494 ESTs from nematodes other than *Caenorhabditis*

Nematode EST projects have generated more than 250,000 ESTs from 30 target species (**Fig. 1** and **Table 1** online; refs. 11,13–20 and

[1]Hospital for Sick Children, 555 University Avenue; Departments of Biochemistry and Medical Genetics and Microbiology, University of Toronto, Toronto, Ontario M5G 1X8, Canada. [2]School of Biological Sciences, University of Edinburgh, Edinburgh, EH9 3JT, UK. [3]Genome Sequencing Center, Washington University School of Medicine, St Louis, Missouri 63108, USA. [4]Pathogen Sequencing Unit, Wellcome Trust Sanger Institute, Cambridge, CB10 1SA, UK. [5]Divergence, St Louis, Missouri 63141, USA. [6]Present addresses: The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA (N.H.); Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA (R.H.W.). Correspondence should be addressed to J.P. (jparkin@sickkids.ca).

Published online 14 November 2004; doi:10.1038/ng1472

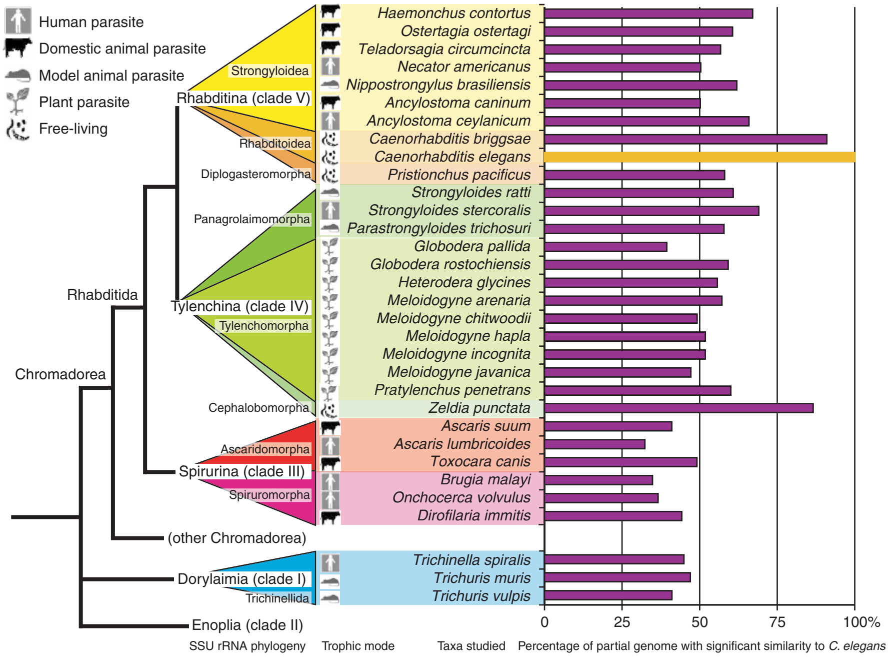

Figure 1 EST data sets from across the phylum Nematoda. (**a**) Species are grouped into major taxonomic groups based on SSU rRNA phylogeny[1,4]. This differs from 'traditional' phylogenies but is consistent with current morphological and developmental evidence. The trophic biology of each targeted species is indicated by a small icon. (**b**) The proportion of each species' partial genome that has significant similarity (a match with a raw BLASTX score ≥ 50) to the complete proteome of *C. elegans*. Owing to the difference in criteria used to define significant similarity, these numbers differ slightly from those previously reported[17,18].

J.P.M. and M.L.B., unpublished data). For each species, we grouped ESTs into clusters and predicted consensus sequences for each cluster (putative gene). These sequences together form the 'partial genome' of each species. **Figure 2a** shows the level of redundancy (ESTs per gene) associated with each partial genome. We observed diminishing returns, in terms of new gene discovery, as we sequenced more ESTs from one species. Redundancy was greatest for *Ascaris suum*, the most heavily sampled species. The number of genes per species ranged from 208 for the smallest EST set (*Zeldia punctata*, 388 ESTs) to more than 9,500 (*Brugia malayi*, 25,067 ESTs). We defined 93,645 putative genes. This is probably a slight overestimate, as the clustering process may split some allelic variation into distinct genes (most parasitic nematode populations used were outbred), different splice forms may not have been clustered together, and nonoverlapping ESTs derived from the same mRNA will not have clustered. This inflation is probably minor (~5%), based on comparisons to the complete *C. elegans* proteome[17,18] and previous analyses of subsets of these data[18,19]. If, as seems likely, most nematodes have ~20,000 protein-coding genes[7,10,21], we have tags for 1–50% of the expected gene complement for each species, with a mean of ~16%. The total number of putative genes triples the number of nematode genes defined[22,23].

Genomic disparity across the phylum Nematoda

The different genes found in the genome(s) of an organism or group of organisms can be thought of as occupying a 'genespace'. More

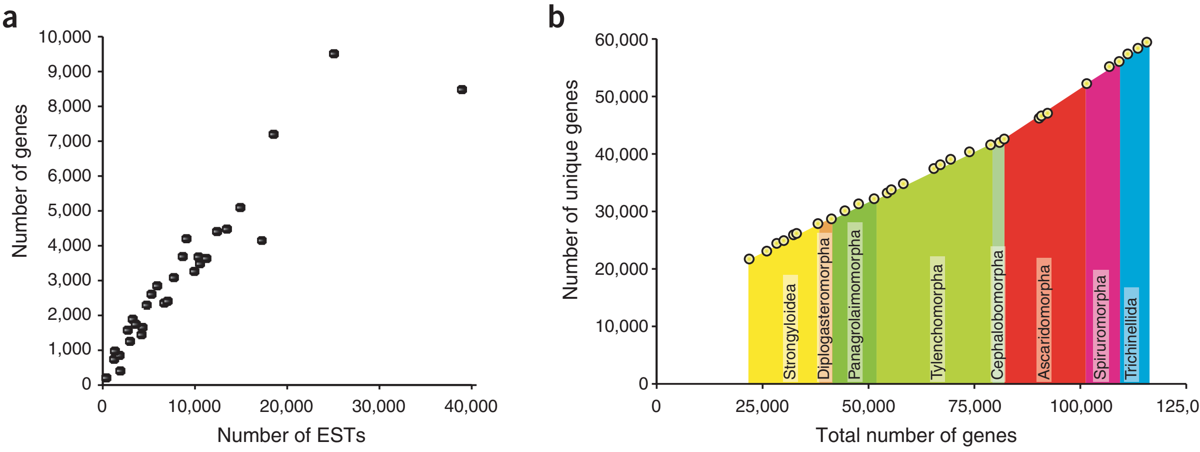

Figure 2 Gene discovery in nematode EST data sets. (**a**) Gene discovery rates in nematode EST data sets. This graph shows the relationship between the number of ESTs sequenced and the number of genes discovered. Each point represents an individual organism's data set. See **Table 1** for details. (**b**) Exploration of genespace in the phylum Nematoda. The cumulative number of different genes (those that have no significant similarity to any other gene) in the EST and proteome data sets from the phylum Nematoda. Each point represents the addition of one nematode species. The first point represents the ~22,000 *C. elegans* proteins. As each partial genome data set was added, increasing the total number of genes (*x* axis), the number of different genes (*y* axis) increased. There was no apparent fall-off in the rate of discovery of new genes, suggesting that nematode genespace may be very large. The colors indicate the systematic origin of each species group (see **Fig. 1**).

Table 1 Summary information of sequence data derived from 30 different species of nematodes

Clade	Genus and species	ESTs	Gene (clusters)	Libraries	Unique to species (%)	Unique to clade (%)	Unique to Nematoda (%)
Dorylaimia (clade I)	*Trichuris muris*	2,713	1,577	1	489 (31)	561 (40.6)	775 (49.1)
	Trichinella spiralis	10,384	3,680	7	1,369 (37.2)	1,428 (38.8)	1,819 (49.4)
	Trichuris vulpis	2,958	1,257	1	461 (36.7)	606 (48.2)	690 (54.9)
Spiruria (clade III)							
Ascaridomorpha	*Ascaris lumbricoides*	1,822	853	1	206 (24.2)	464 (54.6)	611 (71.6)
	Ascaris suum	38,944	8,482	24	3,365 (39.7)	3,797 (44.8)	5,235 (61.7)
	Toxocara canis	4,206	1,447	5	338 (23.4)	402 (27.8)	768 (53.1)
Spiruromorpha	*Brugia malayi*	25,067	9,511	24	4,465 (46.9)	5,010 (52.6)	6,572 (69.1)
	Dirofiliaria immitis	3,585	1,747	2	656 (37.6)	839 (48.1)	1,070 (61.2)
	Onchocerca volvulus	14,922	5,097	9	1,914 (37.6)	2,211 (43.4)	3,242 (63.6)
Tylenchina (clade IV)							
Panagrolaimomorpha	*Parastrongyloides trichosuri*	7,712	3,086	6	746 (24.2)	860 (27.9)	1,564 (50.7)
	Strongyloides ratti	9,932	3,264	10	748 (22.9)	920 (28.2)	1,548 (47.4)
	Strongyloides stercoralis	11,236	3,635	2	556 (15.3)	699 (19.2)	1,401 (38.5)
Tylenchomorpha	*Globodera pallida*	1,317	977	3	425 (43.5)	509 (52.1)	630 (64.5)
	Globodera rostochiensis	5,905	2,851	2	694 (24.3)	979 (34.3)	1,387 (48.6)
	Heterodera glycines	18,524	7,198	10	2,195 (30.5)	2,614 (36.3)	3,594 (49.9)
	Meloidogyne arenaria	3,251	1,892	1	308 (16.3)	968 (35.3)	946 (50)
	Meloidogyne chitwoodi	7,036	2,409	2	700 (29.1)	1,037 (43.1)	1,346 (55.9)
	Meloidogyne hapla	13,462	4,479	4	1,141 (25.5)	1,759 (39.3)	2,388 (53.3)
	Meloidogyne incognita	12,394	4,408	4	1,049 (23.8)	1,814 (41.2)	2,405 (54.6)
	Meloidogyne javanica	5,282	2,609	5	641 (24.6)	1,169 (44.8)	1,531 (58.7)
	Pratylenchus penetrans	1,908	408	1	88 (21.6)	114 (28)	196 (48)
Cephalobomorpha	*Zeldia punctata*	388	208	1	15 (7.2)	15 (7.2)	43 (20.7)
Rhabditina (clade V)							
Strongyloidea	*Ancylostoma caninum*	9,079	4,203	3	1,453 (34.6)	1,910 (45.5)	2,648 (63)
	Ancylostoma ceylanicum	10,544	3,485	9	730 (20.9)	1,037 (29.7)	1,696 (48.7)
	Haemonchus contortus	17,268	4,146	12	797 (19.2)	1,039 (25)	2,012 (48.5)
	Necator americanus	4,766	2,294	3	689 (30)	920 (40.1)	1,476 (64.3)
	Nippostrongylus brasiliensis	1,234	742	3	145 (19.5)	174 (23.4)	382 (51.5)
	Ostertagia ostertagi	6,670	2,355	10	500 (21.2)	681 (28.9)	1,369 (58.1)
	Teladorsagia circumcincta	4,313	1,655	4	362 (21.9)	548 (33.1)	1,008 (60.9)
Diplogasteromorpha	*Pristiochus pacificus*	8,672	3,690	3	1,108 (30)	1,128 (30.5)	1,915 (51.9)
Total		265,494	93,645	172	28,353		52,267

Species are grouped into major taxonomic groups[1,4] (**Fig. 1**). The number of putative genes (size of partial genome) associated with each species is given by the number of clusters (derived from the ESTs). 'Unique to species' indicates genes with no significant (BLAST score <50) sequence similarity to a gene outside that species. 'Unique to clade' indicates genes that share significant sequence similarity to at least one other member of the same clade but do not have any similarity to any gene from out with the clade in question. 'Unique to Nematoda' indicates genes that do not share significant sequence similarity with any non-nematode protein.

complex genomes occupy a larger genespace, in general, and larger groups of organisms (*e.g.*, phyla) have a genespace that is the union of the constituent species' genespaces. Analysis of bacterial genespace from complete genomes showed that sequencing additional eubacterial genomes has yielded diminishing returns in terms of novelty[24]. If nematode genespace is similarly limited, the fact that the genomes of *C. elegans* and *C. briggsae* have been completely sequenced means that sampling additional genomes will result in a low rate of gene discovery as additional genomes are sampled. We carried out an exhaustive series of cross-species BLAST analyses to estimate the extent of nematode genes. We found that 30–70% of the genes from each species had no non-nematode homolog (**Table 1**). The partial nature of the consensus sequences derived from the ESTs may preclude finding matches: short sequences will not reach our score cutoff, and some consensuses may cover only 3′ untranslated regions. Of the 64,685 cluster consensuses longer than 400 bp, 29,118 (45%) had no significant match to non-nematode sequences; for consensuses less than 400 bp in length, 78% seemed to be new. Thus, even excluding short sequences, nearly half the predicted genes seem to be new. The rate of discovery of genetic novelty has not yet started to decline with the analysis of new genomes (**Fig. 2b**), implying that nematode genespace may be much larger than bacterial genespace.

Of the 93,645 putative genes identified in this study, 14,630 (~15%) had significant sequence similarity to putative genes in all of the five major nematode taxonomic groups (*i.e.*, homologs of these genes were identified in all major clades; **Table 1**). Most of these (13,368; 91%) also had homologs outside Nematoda and therefore are probably involved in core metabolic or structural pathways. We found that 1,262 genes are nematode-specific but widely represented within the phylum. These genes may have roles unique to the nematode body plan and life history and are good targets for pan-nematode control drugs.

These findings raise an important issue: if organisms from the same major taxonomic group share only ~60% of their genes, then individual taxa may have widely divergent biology. Within the genus *Caenorhabditis*, *C. elegans* and *C. briggsae* share ~90% of their genes at the level of discrimination used here[10]. Our survey suggests that this level of genetic novelty may be universal across the Nematoda. Lineage-specific genes could have completely new functions, could have similar functions to genes in other organisms but use a completely different mechanism (analogous genes), or could have

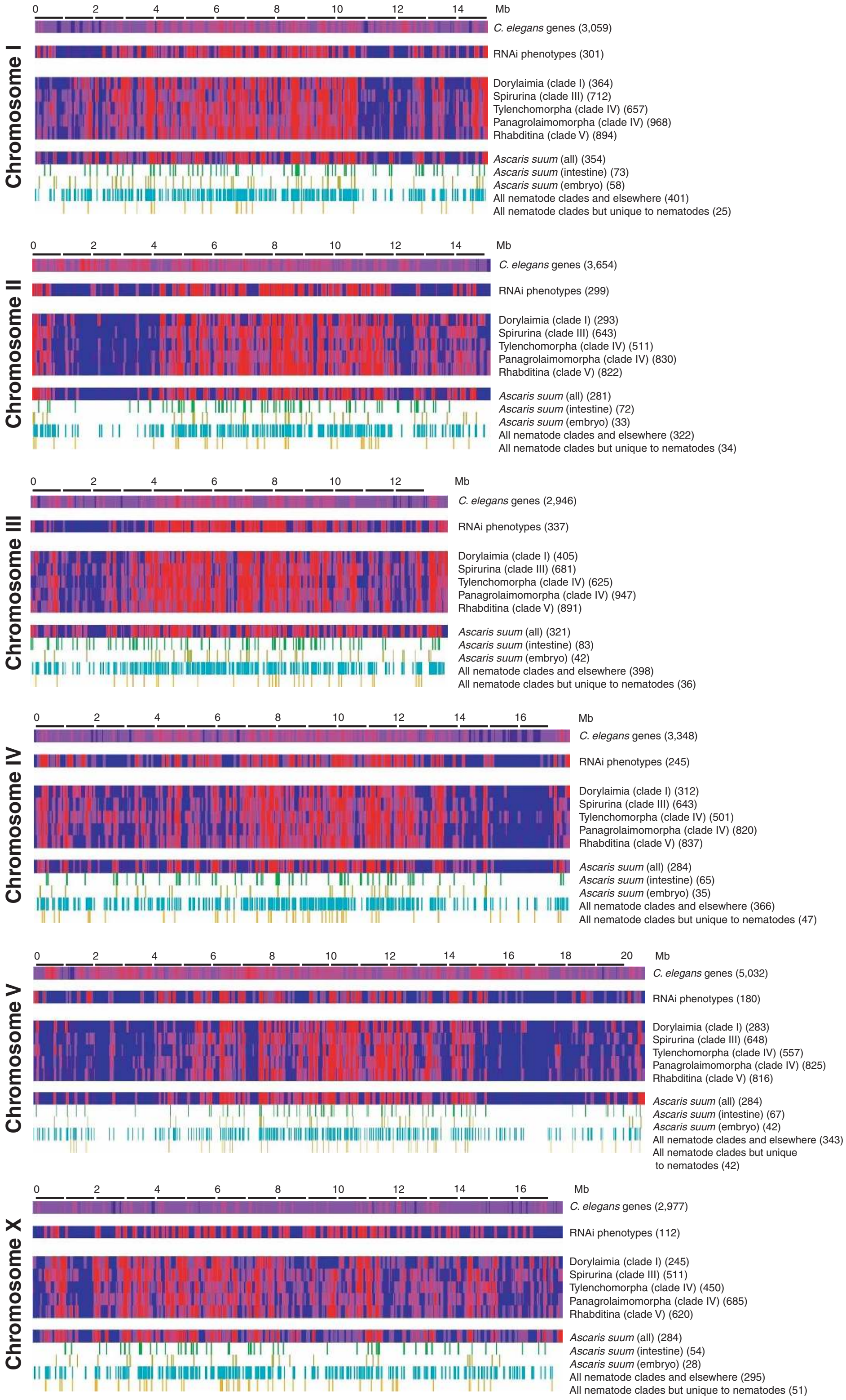

Figure 3 Chromosomal location of *C. elegans* homologs of other nematode genes. Each panel represents a different *C. elegans* chromosome (autosomes I–V and the X sex chromosome). Track 1: the average gene density per 100-kb division (brightest red, >40 genes per 100 kb; blue, <10 genes per 100 kb). Track 2: the relative abundance of genes with visible RNAi phenotypes (red, highest abundance; blue, no genes represented). Tracks 3–7: the abundance of *C. elegans* genes with homologs in the pooled partial genome data sets of the five major clades of the Nematoda. Tracks 8–10: the abundance of *C. elegans* genes with homologs in the complete partial genome, in tissue (intestine) and in stage-specific (embryo) subsets from *A. suum*. Track 11: the positions of individual *C. elegans* genes that have homologs in all five major clades. Track 12: the positions of individual *C. elegans* genes that have homologs in all five major nematode clades but are not significantly similar to any non-nematode gene (a subset of the genes plotted in track 11). The number of genes contributing to each track plot is given in brackets.

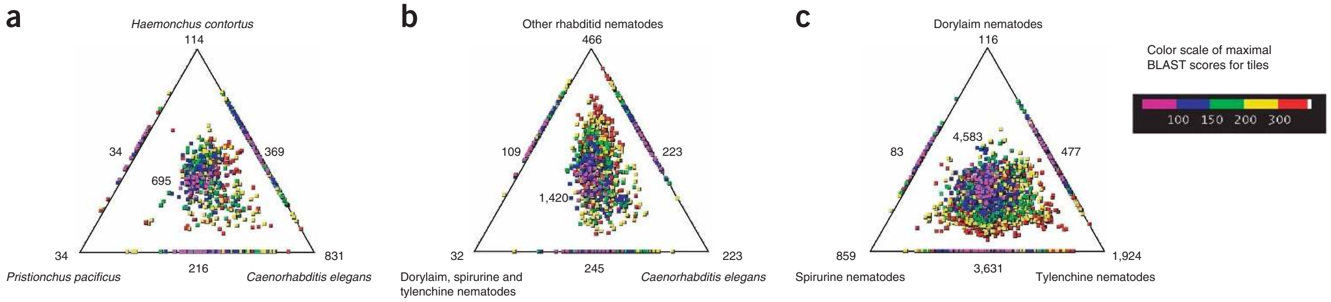

Figure 4 Comparing partial genomes across the Nematoda. SimiTri plots provide a two-dimensional representation of the degree of similarity of an entire data set of sequences with those of three different organisms[32]. The plots allow estimation of relationships of whole sequence data sets and highlight genes with patterns of conservation that differ from the main trend in a data set (see **Supplementary Methods** for a more detailed description). (**a**) *A. caninum* compared with *H. contortus*, *P. pacificus* and *C. elegans*. This plot shows that although *A. caninum* has more matches to *C. elegans* (because its complete genome is available), overall, the sampled transcriptome from *A. caninum* is closer to that of *H. contortus*. (**b**) *A. caninum* compared with other rhabditine nematodes (excluding *C. elegans*), nematodes from other clades and *C. elegans*. (**c**) All partial genomes from rhabditid nematodes (excluding *C. elegans* species) compared to dorylaim, spirurine and tylenchine nematode partial genomes. The numbers at each vertex indicate the number of genes matching only that database. The numbers on the edges indicate the number of genes matching the two linked databases. The number in each triangle indicates the number of genes with matches to all three databases.

similar features, such as tertiary structure, that enable them to use similar molecular mechanisms.

Genomic conservation across the phylum Nematoda

Since the last common ancestor of Nematoda, ∼750–650 million years ago[25,26], nematodes have evolved to occupy many niches. The success of nematodes today reflects the expression of successful complexes of genetic traits, some of which are derived from the common ancestor. A contrasting tendency towards evolution of new genetic function underpinning particular fitness advantage in particular habitats will have resulted in divergence in gene complement. We examined the patterns of gene gain, retention and loss across the phylum by comparing partial genomes both within and between major clades.

The complete genome of *C. elegans* yields a predicted proteome of more than 22,000 polypeptides, some of which derive from alternative splicing and more than 75% of which have some experimental verification[27,28]. We carried out extensive comparisons of the 93,645 new nematode genes with this data set; comparisons with the genome of *C. briggsae*[10] yielded similar results (data not shown). For each *C. elegans* chromosome, we examined the patterns of gene density and the density of genes with known RNA interference (RNAi) phenotypes (**Fig. 3**). As described previously, each autosome has a greater density of genes in the autosomal centers[7,29], and RNAi phenotypes cluster to the centers[28,30]. The autosomal arms are enriched in *C. elegans*–specific genes, in tandemly duplicated gene families and in repetitive and transposable elements[7]. The sex chromosome (X) has a more even distribution of genes and RNAi phenotypes along its length.

We compared the nematode partial genomes to the *C. elegans* proteome in this genomic context (**Fig. 3**). Although we cannot make definitive statements concerning the absence of homologs, matches, when summed over major clades, should identify overall themes. Chromosomal centers were enriched, relative to arms, in the density of similarity matches to all the major taxonomic groups. The pattern of match density faithfully reflected the distribution of genes with RNAi phenotypes[28,30] rather than overall gene density. Thus, taking chromosome (chr) II as an example, each of the five major clades had a high density of matches from ∼4 Mb to 12 Mb (the chromosomal center), with additional peaks of RNAi phenotype genes at 0–0.5 Mb, 3.5 Mb, 13 Mb and 14–15 Mb mirrored by matches in the nematode partial genomes (**Fig. 3**).

There were additional regions of high density of matches that did not coincide with RNAi phenotype genes. On chr II, we observed high match densities in all major clades at 1.0–1.2 Mb and 13.6 Mb, where RNAi phenotype genes were rare. We noted similar regions on chr I (1.2–2.0 Mb), chr III (0.8–1.0 Mb), chr IV (1.8 Mb) and chr V (∼1 Mb and ∼20 Mb; **Fig. 3**). On the X chromosome there was a cluster of conserved genes at ∼17–18.5 Mb that did not have a high frequency of RNAi phenotypes. These conserved genes may not yield RNAi phenotypes in *C. elegans* because of gene family redundancy, or because they are involved in nematode-specific phenotypes that are not assessed by currently applied assays under select laboratory conditions. Conversely, there were a few chromosomal regions where a high density of RNAi phenotype genes was not matched by a high density of matches to the partial genomes: one example is on chr II at 2.8 Mb (**Fig. 3**).

Comparison between autosomes showed that chr V had ∼60% fewer matches per *C. elegans* gene than the other autosomes. This adds to the known peculiarities of chr V (ref. 31), which also has more *C. elegans*–specific genes, has fewer RNAi phenotypes per gene, and is significantly less likely to match genome survey sequences from the filarial nematode *B. malayi*[21].

Mapping genomic divergence within the phylum Nematoda

C. elegans is a rhabditine nematode[1], closely related to the Diplogasteromorpha (represented here by *Pristionchus pacificus*) and the vertebrate-parasitic Strongyloidea (represented here by seven species; **Fig. 1** and **Table 1**)[4]. The proportion of the partial genome of each sampled species that contained *C. elegans* homologs roughly followed the species' relationship to *C. elegans*, with rhabditine species having the highest mean proportion of homologs (**Figs. 1** and **4**). Comparative analyses of the full partial genome data sets yield a set of similarity relationships congruent with the SSU rRNA phylogeny[4]. We used SimiTri analyses, which plot the relative similarity of entire gene data sets against three data sets of interest[32], to examine these relationships further. Comparison of the partial genomes of the hookworm *Ancylostoma caninum* and *Haemonchus contortus* (another strongyloid parasite), *P. pacificus* and *C. elegans* (**Fig. 4a**) showed that although

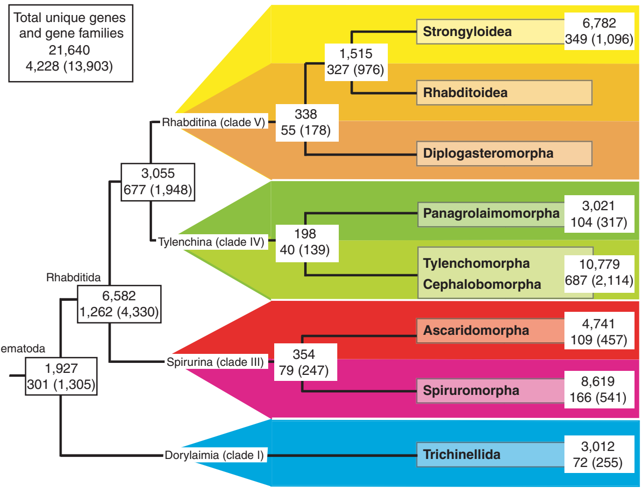

Figure 5 Evolutionary origins of unique genes and gene families in the phylum Nematoda. The inferred positions of origin for the nematode-specific genes and gene families were mapped across the robust SSU rRNA phylogeny[1,4]. For each node, the upper number shows the number of genes unique to each clade, and the lower number shows the predicted number of unique gene families (with the number of individual predicted genes included in these families in brackets). In the absence of complete genome sequences from most of the Nematoda, this mapping places the origin of each gene or gene family at the highest possible node: adding complete genome sequences will tend to move the node of origin of some genes lower in the tree.

genes with matches to *C. elegans* comprised the largest subset, most genes were more similar to *H. contortus* than to *P. pacificus*. A number of genes had high-scoring matches in *P. pacificus* and *H. contortus* but were lacking from the complete *C. elegans* proteome, probably indicative of gene loss in the *C. elegans* lineage. Extending these analyses to span the major nematode clades showed the expected global trend for *A. caninum* genes to be most similar to homologs from other rhabditids (including other species from Strongyloidea; **Fig. 4b**) and least similar to homologs from the dorylaims *Trichinella spiralis* and *Trichuris muris* (**Fig. 4c**)[19].

Tylenchomorpha, Cephalobomorpha and Panagrolaimomorpha species also had a high proportion of matches to rhabditine sequences, but as the predicted phylogenetic distance from *C. elegans* increased, there was a corresponding decrease in the number of genes sharing significant similarity (**Fig. 1**). The cephalobomorph *Z. punctata* seems to be an exception, but this anomaly is probably due to sampling from a single spliced leader-PCR–based library that biases towards short, conserved transcripts[16]. Only ~45% of the genes from the dorylaims *T. muris* and *T. spiralis* had significant similarity to genes from *C. elegans*, but a similar percentage of their genes shared similarity with *Drosophila melanogaster* (data not shown). Thus *T. spiralis* and *T. muris* may be good choices for deeper genomic analysis of the relationships of Nematoda to other metazoan phyla. Overall, these results suggest that *C. elegans* will be an effective genomic model for other rhabditid nematodes, but that accumulated differences will make extrapolation to distantly related nematodes, such as dorylaims, more challenging. There are many nematode genes, found in species across the phylum, that are absent from *C. elegans*. Gene loss has therefore been an important part of *C. elegans* genome evolution, as was suggested by the finding that *C. elegans*' depleted HOX gene complement is a result of lineage-specific losses[33].

It has previously been suggested that there is a high-level ordering of genes within *C. elegans* chromosomes, with, for example, muscle-expressed genes being located in close proximity to each other more often than would be expected by chance[34], and RNAi phenotyping suggesting that large chromosomal domains of genes have similar biological roles[28]. We investigated whether nematode-specific or stage-specific genes were clustered at a megabase level but did not find evidence for linkage of these classes of genes. For example, we mapped homologs of *A. suum* genes that had stage- or tissue-specific expression patterns to the *C. elegans* genome (**Fig. 3**). The tissue-specific (intestine) or stage-specific (embryo, L3, L4) genes showed the same general pattern of distribution as all *A. suum* genes.

Nematode-specific genes and gene families

Putative nematode-specific targets for drugs with diminished risk of toxicity to hosts or other nontarget organisms may be found in the class of nematode-specific genes. We found that 30–50% of each of our chosen species' partial genomes was unique. Of the 52,267 genes for which no homolog was identified outside the phylum (**Table 1**), 21,640 had significant similarity with a sequence from another nematode species. Mapping these nematode-specific genes onto the phylogeny showed an incremental evolution of novelty (**Fig. 5**). Most unique genes were associated with shallow-level taxonomic groups, but a considerable proportion had a deeper origin. Some deep splits in Nematoda were associated with few unique genes (*e.g.*, Panagrolaimomorpha plus Tylenchomorpha/Cephalobomorpha has only 198), perhaps reflecting relatively rapid divergence of these daughter clades shortly after the origin of the ancestral tylenchine.

We clustered the predicted polypeptides associated with these nematode-specific genes using Tribe-MCL[35] into putative gene families and mapped the latest possible origin of these families onto the phylogeny (**Fig. 5**). In general, the number of new gene families at each node of the tree reflected the number of genes associated with the smallest daughter clade. The two largest groups of unique gene families occur at the base of Rhabditida and at the node connecting Spirurina with Tylenchina plus Rhabditina. This possibly reflects both the relatively large number of taxa and the number of ESTs generated for these three clades (*e.g.*, the Tylenchina data set included several closely related *Meloidogyne* species[17]). Most unique gene–origin events seemed to occur relatively early in the nematode lineage. For example, more than 6,500 genes had homologs in each of the three taxonomic groups in the Rhabditida, and these included 4,330 genes in 1,262 nematode-specific families (**Fig. 5**).

We examined some nematode-specific gene families in more detail. For many we identified *C. elegans* members, permitting exploration of possible function through published RNAi data[28]. For example, a family with ten members from all major clades in our data set had a single *C. elegans* homolog identified by RNAi to be essential for postembryonic larval development (**Supplementary Fig. 1** online). Another was limited to Rhabditida and had a *C. elegans* member with an RNAi phenotype of inhibition of postembryonic growth (**Supplementary Fig. 1** online). In both cases, the degree of sequence conservation suggests that the RNAi function may be ascribed to the other genes: the *C. elegans* phenotype recommends these for further investigation as targets for nematicides.

Domain and functional analysis of nematode proteins

We used InterPro[36] to identify known protein domains in the partial genomes. Because the *C. elegans* proteome has been extensively investigated for protein domains[7,37], many domains of unknown function

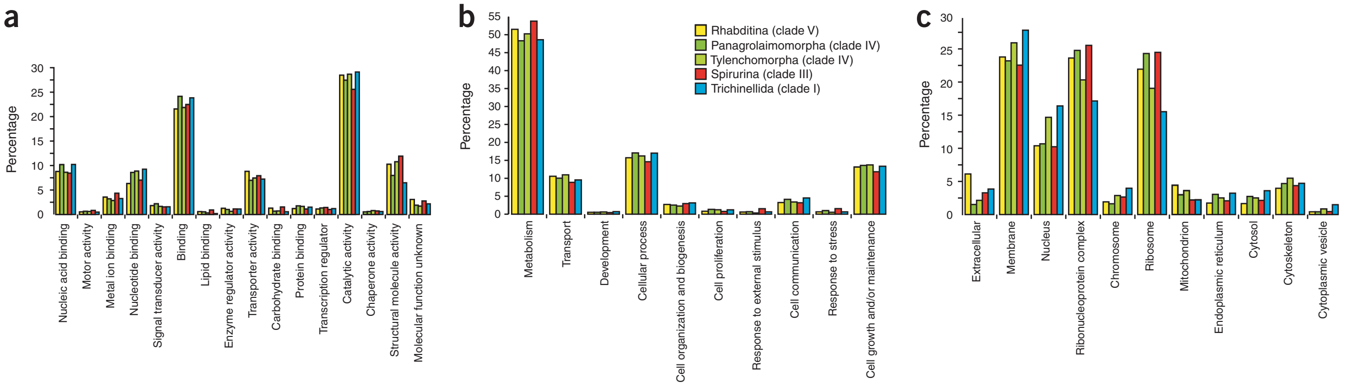

Figure 6 Functional annotation of genes using Gene Ontology terms. Each sequence was compared with the InterPro database of domains and these matches were used to assign high-level Gene Ontology terms. The data is summarized by major clade (see **Fig. 1**). The *x* axes show the percentage of all the Gene Ontology terms for each assignment: (**a**) 'molecular function' assignments; (**b**) 'biological process' assignments; and (**c**) 'cellular component' assignments.

have been defined that are exclusive to *C. elegans* and *C. briggsae*. Many of the matches we discovered were to these nematode-restricted domains. For each species, 30–50% of the polypeptides were predicted to contain at least one previously identified domain (**Supplementary Table 1** online). Fewer polypeptides from both spirurine and dorylaim species than from tylenchine and strongyloid species contained a domain, reflecting the *C. elegans* bias in InterPro. The number of unique domains associated with each species increased with size of its partial genome (**Supplementary Table 1** online).

Comparison of the most abundant domains associated with each clade showed that, with the exception of the protein kinase domain, the abundant domains did not correlate well with those previously identified in the complete proteomes of *C. elegans* and *C. briggsae* (**Supplementary Table 2** online)[7,10]. These differences may have arisen from the unavoidable bias in the types of genes sampled by ESTs. We minimized this bias by grouping the partial genomes and their domain contents by major clade (**Supplementary Table 2** online). Cuticle collagens (IPR008160 and IPR002048) are abundant in *C. elegans* and *C. briggsae* (~170 in each) but were poorly represented in the dorylaim partial genomes, possibly reflecting the derivation of these data sets from nonmoulting stages. Collagens have a temporally restricted expression pattern in *C. elegans*, with most expression in larval stages[38]. The strongyloid partial genomes were enriched for peptidases (IPR000169, IPR001254, IPR001353 and IPR00668), and dorylaim sequences were enriched for potential proteinase inhibitors (IPR008197 and IPR008198) and for chymotrypsin (IPR001254). This may reflect the parasitic niche of the sampled species (the host intestine), where peptidases and inhibitors may be required for feeding and survival in such a hostile environment. Also in Strongyloidea, the ShK metridin-like toxin domain (IPR003582) was highly represented, perhaps reflecting involvement in parasitic interactions[15]. EGF-like domains (IPR006209) are one of the more common domains in *C. elegans* and *C. briggsae*[7,10]. Although EGF-like domains were found in other Rhabditina, dorylaim and panagrolaimomorph organisms had the highest relative abundance. EGF-like domains are associated with membrane-bound or secreted proteins involved in signaling and recognition and may therefore be involved in manipulation of host responses.

Analysis of InterPro domain representation in nematode-specific genes identified a set of domains that may be important in parasitism. In the Strongyloidea, the thirteenth-most abundant InterPro domain is the allergen V5/Tpx-1 related domain IPR001283, found in many secreted proteins, most notably in the *Ancylostoma* secreted proteins (asp) that have immunomodulatory activity[39]. Additional asp-like proteins are present in other clades[15]. Although IPR001283 is found in organisms other than nematodes, genes containing this domain have undergone lineage-specific amplification and divergence in Strongyloidea. An abundant domain found in nematode-specific gene families of particular prominence is the 'transthyretin-like' IPR001534: this family had 394 members from all species (only lacking in *Nippostrongylus brasiliensis*), of which 377 had no significant BLAST similarity outside the Nematoda. It is prominent in *C. elegans* and *C. briggsae*. Mammalian transthyretins transport thyroid hormones, and many of the nematode genes have secretory leader polypeptides, suggestive of a role in hormonal signaling in nematodes also.

To compare the biological functions of the genes associated with each nematode, we used InterPro matches to assign Gene Ontology terms[40]. The high-level Gene Ontology profiles for each clade were very similar (**Fig. 6**), but we noted some differences between clades. Dorylaim and panagrolaimomorph nematodes had a lower incidence of structural proteins (**Fig. 6a**). Rhabditine nematodes had an elevated number of predicted extracellular proteins, and pirurine, tylenchine and rhabditine nematodes had an increased proportion of ribonuclear proteins (**Fig. 6c**)[20]. The two groups with the highest proportion of nuclear-localized predicted polypeptides were Tylenchina and Dorylaimia. In both, parasite-secreted proteins are known to localize to the host nucleus (**Fig. 6c**).

Metabolic pathway analyses

There is a general perception that parasites have lost function (undergone reductive evolution) as they came to rely on the metabolic capacity and homeostatic buffering of their hosts. But many parasitic nematodes spend part of their life cycle outside any host, or have multiple phylogenetically and metabolically different hosts, and therefore may experience evolutionary pressure to maintain or even expand metabolic and regulatory functions[41]. We compared the partial genome of each species with the KEGG database[42] to determine the extent of metabolic pathway representation (**Supplementary Table 3** online). For most pathways, the number of enzymes associated with each major clade correlated with the number of sequences generated (**Table 1** and **Supplementary Table 3** online). The general congruence between the major clades suggested that many pathways are conserved within the nematodes despite their diversity. Some differences were noted, however. Spiruria and Dorylaimia had 17 enzymes (34 clusters) from fatty acid biosynthesis pathway 1 (using acyl carrier protein-bound precursors) but lacked pathway-2 enzymes (using coenzyme A-bound precursors) completely, whereas Tylenchina had only pathway-2 enzymes (44 clusters mapping to three enzyme types), and Rhabditina had both. No valine or methionine biosynthesis enzymes were identified in the animal-parasitic Spiruria, suggesting that these may be essential amino acids. N-glycan degradation enzymes were notably abundant in Tylenchina, but less evident elsewhere. As the complete genome of *C. elegans* encodes many N-glycan degradation enzymes, this suggests that this pathway is particularly highly expressed in these plant parasites. Enzymes involved in inositol metabolism were also prominent in Tylenchina (and in the complete *C. elegans* proteome) but absent in other sampled species. These predictions of taxonomically restricted biochemical pathways may serve to direct drug target definition.

DISCUSSION

We used ~250,000 ESTs to predict more than 90,000 genes for a suite of important human, animal and plant parasitic, and two free-living, nematodes. Comparison of each species' partial genome with the complete genomes of *C. briggsae* and *C. elegans* and with genome data from other phyla identified a spectrum of genes and gene families, some of which were deeply conserved, others were pan-nematode but nematode-unique, and others were taxonomically restricted. This data set aids the annotation of the *C. elegans* genome in confirming gene predictions and identification by alignment of homologs of conserved and thus functionally important residues. Highly conserved genes discovered in species across the phylum may have important function in *C. elegans*, but some such genes currently have no known RNAi phenotypes, perhaps showing the limitation of on-plate assays. We identified tens of thousands of potential targets for drug and vaccine development. Many of these are nematode-specific but conserved across the phylum, offering the prospect of new pan-nematode treatments. We have deposited our sequence data in public databases as it is generated and offered our analyses openly over the Internet since the inception of the project, and so many of the genes we identified have already been selected by other researchers in parasitology and *C. elegans* biology for further study.

We look forward to further expansions of the nematode genome data sets. We are still sequencing additional ESTs from target species, and other projects, including enoplid and chromadorid taxa, are also underway or planned. The genus *Caenorhabditis* will soon have five nearly complete genome sequences, and the *B. malayi* genome is nearing completion at The Institute for Genomic Research[43]. Genome sequencing is planned for *H. contortus*, *Meloidogyne hapla*, *P. pacificus* and *T. spiralis*. Our survey indicates that model species cannot show the genetic and genomic diversity of even their own phylum, and that continuing, phylogenetically informed genome sequencing is essential for advances in genomics, evolution and infectious disease biology.

METHODS

Large-scale EST generation across the phylum Nematoda. We selected a portfolio of nematode species based on criteria of phylogenetic spread, availability of material and health, economic or scientific importance, after consultation with the research and funding communities. The number of ESTs sequenced for each species varied because of factors such as availability of material, quality and source of libraries, and perceived importance of the organism. For each species, we aimed to accumulate an EST data set spanning multiple life cycle stage–specific and, where possible, tissue-specific cDNA libraries. Previous experience with the filarial parasite *B. malayi* showed that sampling from throughout the nematode life cycle was essential to maximize rates of gene discovery[44]. To this end, we constructed 172 different cDNA libraries, using various cDNA synthesis technologies and vectors (**Supplementary Methods** online). Members of the research community were generous in providing both biological materials and libraries.

Sequencing and data processing. Sequencing and EST processing at the Genome Sequencing Center was carried out as described[17,18,45]. Sequencing and processing at the Wellcome Trust Sanger Institute was carried out as described[46]. Before submitting them to dbEST, we processed the sequences to assess quality, trim vector, remove contaminants and cloning artifacts, and identify BLAST similarities using Genome Sequencing Center pipelines[45] and the trace2dbEST pipeline[47]. All ESTs have been submitted to dbEST[12].

Sequence clustering, annotation and database creation. For each species, we downloaded sequences from dbEST in May 2003 and parsed them through PartiGene, a software pipeline designed to analyze and organize EST data sets[47]. We first checked sequences for contaminating vector sequence and trimmed poly(dA) tails. We then clustered sequences into groups (putative genes) on the basis of sequence similarity using CLOBB[48]. We assembled clusters to yield consensus sequences using PHRAP (P. Green, unpublished data). We then subjected each consensus sequence to a series of BLAST analyses[49] against a suite of protein and nucleotide databases derived from public databases (see **Supplementary Methods** online for details) and the thirty sets of consensus sequences (partial genomes) for each nematode species analyzed here. We defined significant matches as those having a raw BLAST score ≥ 50 (this corresponds to an expect value of 10^{-5} to 10^{-6}, depending on the size and composition of the databases). Although this cutoff may miss some homologous matches, it is sufficiently inclusive to identify most domain matches. Results were processed and stored in a local installation of a PostgreSQL database[22,23].

Predicted proteome analysis: Gene Ontology, domains and metabolic pathways. For each consensus sequence, we obtained polypeptide predictions using the prot4EST package in the PartiGene pipeline[47]. Predicted polypeptide sequences were compared to InterPro (data version 7.0) to identify functional domains using InterProScan[36]. An InterPro annotation was assigned to each domain and translated into Gene Ontology[40] codes. These results were parsed into a local installation of AmiGO[40] from which broader functional categories were derived. Protein families were identified using TRIBE-MCL[35] using default parameters.

To map predicted polypeptides to metabolic pathways, we compared them using BLASTP to the KEGG database (version 29)[42]. We retained each match meeting a cut-off of an expect statistic $\leq 1 \times 10^{-10}$. When one cluster matched several closely related enzymes, we considered the top match and all the matches within a range of 30% of the top score.

URLs. See http://www.nematode.net/ for more information on Genome Sequencing Center trace files and clone ordering, and see http://www.nematodes.org/ for more information on Wellcome Trust Sanger Institute trace files and clone ordering. PHRAP is available at http://www.phrap.org.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank all our nematology and parasitology colleagues for supplying materials (see **Supplementary Methods** online) and for their enthusiasm for this project and A. Anthony, J. Wasmuth and A. Hedley for trace2dbest, PartiGene and prot4EST software. The UK arm of the project was funded by the Wellcome Trust, and the US arm by the National Institutes of Health (National Institute of Allergy and Infectious Diseases). J.P.M. was supported by a Helen Hay Whitney/Merck fellowship. Sequencing at The Wellcome Trust Sanger Institute was carried out by C. Churcher, T. Chillingworth, P. Cummings, Z. Hance, K. Jagels, S. Moule and S. Whitehead. Most of the computational analyses were done using facilities at the Center for Computational Biology, Hospital for Sick Children, Toronto.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Genetics* website for details).

Received 19 August; accepted 15 October 2004
Published online at http://www.nature.com/naturegenetics/

1. De Ley, P. & Blaxter, M.L. Systematic position and phylogeny. in *The Biology of Nematodes* (ed. Lee, D.) 1–30 (Taylor & Francis, London, 2002).
2. Platt, H.M. Foreword. in *The Phylogenetic Systematics of Free-living Nematodes* (Lorenzen, S.) (The Ray Society, London, 1994).
3. Lambshead, P.J.D. Recent developments in marine benthic biodiversity research. *Oceanis* **19**, 5–24 (1993).
4. Blaxter, M.L. *et al.* A molecular evolutionary framework for the phylum Nematoda. *Nature* **392**, 71–75 (1998).
5. Chan, M.-S. The global burden of intestinal nematode infections - fifty years on. *Parasitol. Today* **13**, 438–443 (1997).
6. Barker, K.R., Hussey, R.S. & Krusberg, L.R. *Plant and Soil Nematodes: Societal Impact and Focus on the Future* (Committee on National Needs and Priorities in Nematology, Society of Nematologists, Marceline, Missouri, USA, 1994).
7. The *C. elegans* Genome Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
8. Wood, W.B. (ed.) *The Nematode Caenorhabditis elegans* 667 (Cold Spring Harbor Laboratory Press, New York, 1988).
9. Riddle, D. Blumenthal, T., Meyer, B. & Priess, J. (eds.) *C. elegans II* 1222 (Cold Spring Harbor Laboratory Press, New York, 1997).
10. Stein, L.D. *et al.* The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. *PLoS Biol.* **1**, E45 (2003).
11. Parkinson, J., Mitreva, M., Hall, N., Blaxter, M. & McCarter, J.P. 400000 nematode ESTs on the Net. *Trends Parasitol.* **19**, 283–286 (2003).
12. Boguski, M.S., Lowe, T.M. & Tolstoshev, C.M. dbEST - database for "expressed sequence tags". *Nat. Genet.* **4**, 332–333 (1993).
13. Lizotte-Waniewski, M. *et al.* Identification of potential vaccine and drug target candidates by expressed sequence tag analysis and immunoscreening of *Onchocerca volvulus* larval cDNA libraries. *Infect. Immun.* **68**, 3491–3501 (2000).
14. Tetteh, K.K., Loukas, A., Tripp, C. & Maizels, R.M. Identification of abundantly expressed novel and conserved genes from the infective larval stage of Toxocara canis by an expressed sequence tag strategy. *Infect. Immun.* **67**, 4771–4779 (1999).
15. Daub, J., Loukas, A., Pritchard, D.I. & Blaxter, M. A survey of genes expressed in adults of the human hookworm, *Necator americanus*. *Parasitology* **120**, 171–184 (2000).
16. Blaxter, M.L. *et al.* Genes expressed in *Brugia malayi* infective third stage larvae. *Mol. Biochem. Parasitol.* **77**, 77–96 (1996).
17. McCarter, J.P. *et al.* Analysis and functional classification of transcripts from the nematode *Meloidogyne incognita*. *Genome Biol.* **4**, R26 (2003).
18. Mitreva, M. *et al.* Comparative genomics of gene expression in the parasitic and free-living nematodes *Strongyloides stercoralis* and *Caenorhabditis elegans*. *Genome Res.* **14**, 209–220 (2004).
19. Mitreva, M. *et al.* Gene discovery in the adenophorean nematode Trichinella spiralis: an analysis of transcription from three life cycle stages. *Mol. Biochem. Parasitol.* **137**, 277–291 (2004).
20. Harcus, Y.M. *et al.* Signal sequence analysis of expressed sequence tags from the nematode Nippostrongylus brasiliensis and the evolution of secreted proteins in parasites. *Genome Biol.* **5**, R39 (2004).
21. Whitton, C. *et al.* A genome sequence survey of the filarial nematode *Brugia malayi*: repeats, gene discovery, and comparative genomics. *Mol. Biochem. Parasitol.* **137**, 215–227 (2004).
22. Parkinson, J., Whitton, C., Schmid, R., Thomson, M. & Blaxter, M. NEMBASE: a resource for parasitic nematode ESTs. *Nucleic Acids Res.* **32**, D427–D430 (2004).
23. Wylie, T. *et al.* Nematode.net: a tool for navigating sequences from parasitic and free-living nematodes. *Nucleic Acids Res.* **32**, D423–D426 (2004).
24. Tatusov, R.L. *et al.* The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29**, 22–28 (2001).
25. Blaxter, M.L. *Caenorhabditis elegans* is a nematode. *Science* **282**, 2041–2046 (1998).
26. Wang, D.Y., Kumar, S. & Hedges, S.B. Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proc. R. Soc. Lond. B Biol. Sci.* **266**, 163–171 (1999).
27. Reboul, J. *et al.* C. elegans ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat. Genet.* **34**, 35–41 (2003).
28. Kamath, R.S. *et al.* Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421**, 231–237 (2003).
29. Barnes, T.M., Kohara, Y., Coulson, A. & Hekimi, S. Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans*. *Genetics* **141**, 159–179 (1995).
30. Simmer, F. *et al.* Genome-wide RNAi of *C. elegans* using the hypersensitive *rrf-3* strain reveals novel gene functions. *PLoS Biol.* **1**, E12 (2003).
31. Gregory, W.F. & Parkinson, J. *Caenorhabditis elegans* – applications to nematode genomics. *Comp. Funct. Genomics* **4**, 194–202 (2003).
32. Parkinson, J. & Blaxter, M.L. SimiTri - visualising similarity relationships for large groups of sequences. *Bioinformatics* **19**, 390–395 (2002).
33. Aboobaker, A.A. & Blaxter, M.L. Hox gene loss during dynamic evolution of the nematode cluster. *Curr. Biol.* **13**, 37–40 (2003).
34. Roy, P.J., Stuart, J.M., Lund, J. & Kim, S.K. Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans*. *Nature* **418**, 975–979 (2002).
35. Enright, A.J., Van Dongen, S. & Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
36. Zdobnov, E.M. & Apweiler, R. InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).
37. Hutter, H. *et al.* Conservation and novelty in the evolution of cell adhesion and extracellular matrix genes. *Science* **287**, 989–994 (2000).
38. Johnstone, I.L. & Barry, J.D. Temporal reiteration of a precise gene expression pattern during nematode development. *EMBO J.* **15**, 3633–3639 (1996).
39. Hawdon, J.M., Jones, B.F., Hoffman, D.R. & Hotez, P.J. Cloning and characterization of *Ancylostoma*-secreted protein. A novel protein associated with the transition to parasitism by infective hookworm larvae. *J. Biol. Chem.* **271**, 6672–6678 (1996).
40. The Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Res* **11**, 1425–1433 (2001).
41. Blaxter, M. Nematoda: Genes, genomes and the evolution of parasitism. *Adv. Parasitol.* **54**, 102–195 (2003).
42. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**, D277–D280 (2004).
43. Ghedin, E., Wang, S., Foster, J.M. & Slatko, B.E. First sequenced genome of a parasitic nematode. *Trends Parasitol.* **20**, 151–153 (2004).
44. Williams, S.A. *et al.* The filarial genome project: analysis of the nuclear, mitochondrial and endosymbiont genomes of *Brugia malayi*. *Int. J. Parasitol.* **30**, 411–419 (2000).
45. Hillier, L.D. *et al.* Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* **6**, 807–828 (1996).
46. Whitton, C., Daub, J., Thompson, M. & Blaxter, M. Expressed sequence tags: medium throughput protocols. in *Parasite Genomics* (ed. Melville, S.E.) (Humana, New York, in the press).
47. Parkinson, J. *et al.* PartiGene - constructing partial genomes. *Bioinformatics* **20**, 1398–1404 (2004).
48. Parkinson, J., Guiliano, D. & Blaxter, M. Making sense of EST sequences by CLOBBing them. *BMC Bioinf.* **3**, 31 (2002).
49. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).