nature genetics

The draft genome of the parasitic nematode *Trichinella spiralis*

Makedonka Mitreva[1,2], Douglas P Jasmer[3], Dante S Zarlenga[4], Zhengyuan Wang[1], Sahar Abubucker[1], John Martin[1], Christina M Taylor[1], Yong Yin[1,7], Lucinda Fulton[1,2], Pat Minx[1], Shiaw-Pyng Yang[1,7], Wesley C Warren[1,2], Robert S Fulton[1,2], Veena Bhonagiri[1], Xu Zhang[1], Kym Hallsworth-Pepin[1], Sandra W Clifton[1,2], James P McCarter[2,5], Judith Appleton[6], Elaine R Mardis[1,2] & Richard K Wilson[1,2]

Genome evolution studies for the phylum Nematoda have been limited by focusing on comparisons involving *Caenorhabditis elegans*. We report a draft genome sequence of *Trichinella spiralis*, a food-borne zoonotic parasite, which is the most common cause of human trichinellosis. This parasitic nematode is an extant member of a clade that diverged early in the evolution of the phylum, enabling identification of archetypical genes and molecular signatures exclusive to nematodes. We sequenced the 64-Mb nuclear genome, which is estimated to contain 15,808 protein-coding genes, at ~35-fold coverage using whole-genome shotgun and hierarchal map–assisted sequencing. Comparative genome analyses support intrachromosomal rearrangements across the phylum, disproportionate numbers of protein family deaths over births in parasitic compared to a non-parasitic nematode and a preponderance of gene-loss and -gain events in nematodes relative to *Drosophila melanogaster*. This genome sequence and the identified pan-phylum characteristics will contribute to genome evolution studies of Nematoda as well as strategies to combat global parasites of humans, food animals and crops.

Currently, no complete genome sequence information exists from lineages spanning the phylum Nematoda (**Supplementary Fig. 1**). Yet, such information is essential in understanding the evolution of the Nematoda analogous to the way that a basal chordate informed our understanding of vertebrate evolution[1]. To this end, we generated the genome sequence of *T. spiralis*, a food-borne, zoonotic parasite, to reveal molecular characters and evolutionary trends between this organism, evolutionarily distant parasitic and non-parasitic nematodes and a member of the next closest sequenced relatives, the arthropods. In so doing, we identified commonalities that link nematodes to other Metazoa members, as well as distinctions that define the Nematoda and differentiate *T. spiralis* from the other species investigated. The *Trichinella* assembly is 64 million bp in length and encodes at least 15,808 proteins, which makes this genome substantially smaller than that of the prototypical nematode, *C. elegans*.

Trichinellosis is a worldwide zoonotic disease. The nematode *T. spiralis*, the most common cause of human trichinellosis, is a member of a clade that diverged early in the evolution of the Nematoda. It differs substantially in biological and molecular characters from other crown groups[2–4]. The lineage giving rise to the genus *Trichinella* last shared a common ancestor approximately 275 million years ago (Lower Permian Period), whereas the diversification of extant *Trichinella* species occurred as recently as 16–20 million years ago (Miocene Epoch)[5].

The life cycle of *Trichinella* spp. (**Supplementary Fig. 2**) begins when muscle tissue containing first-stage larvae (ML) is ingested by the new host. The ML rapidly develop into adults in the intestine, where they mate and produce newborn larvae (NBL). The NBL migrate from the intestines through the lymphatic system, eventually to the blood, and then they invade striated skeletal muscle cells to complete the cycle and become infectious to the next host. Intense inflammation is a primary cause of disease and involves myositis, myocarditis and encephalitis, the intensity of which depends on the number of parasites ingested. Currently, the genus consists of eight distinct species and/or genotypes that are further categorized as encapsulated or non-encapsulated, predicated upon the formation of a collagen envelope around the infected muscle cell. This capsule is believed to be a host-derived structure induced only by species that infect placental mammals and is unique to this genus. In addition to the formation of a collagen capsule, and contrary to most other parasitic nematodes, *T. spiralis* shows little host specificity among mammals, completes its entire life cycle in a single host, does not have a free-living stage and lives as an intracellular parasite within a single striated muscle cell. As such, this genus presents biological characteristics that markedly differ from what is common among most other nematodes.

Herein we compare the molecular characteristics of nematodes and other metazoans using the entire *T. spiralis* genome. This comparative

[1]The Genome Center, Washington University School of Medicine, St. Louis, Missouri, USA. [2]Department of Genetics, Washington University School of Medicine, St. Louis, Missouri, USA. [3]Department of Veterinary Microbiology and Pathology, Washington State University, Pullman, Washington, USA. [4]US Department of Agriculture, Animal Parasitic Disease Laboratory, Beltsville, Maryland, USA. [5]Divergence, Inc., St. Louis, Missouri, USA. [6]James A. Baker Institute for Animal Health, College of Veterinary Medicine, Cornell University, Ithaca, New York, USA. [7]Present address: Monsanto Company, St. Louis, Missouri, USA. Correspondence should be addressed to R.K.W. (rwilson@genome.wustl.edu) or M.M. (mmitreva@genome.wustl.edu).

Received 16 April 2010; accepted 21 January 2011; published online 20 February 2011; doi:10.1038/ng.769

approach allowed us to identify conserved protein and gene sequences with apparent archetypical standing for the phylum Nematoda. We found that intrachromosomal rearrangements were common throughout the phylum; however, this was in contrast to other characters such as protein family deaths and births, which showed a clear demarcation between a parasitic and non-parasitic nematode. In addition, unlike in *D. melanogaster*, the levels of gene loss and gain in each nematode species indicate that these events may have played a substantially larger role in the evolution of this phylum. The identification of these and other conserved characteristics, predicated in part upon this work, will advance more targeted research on pathogens from a phylum harboring thousands of pathogens that infect humans, animals and plants. The advances may one day provide holistic strategies to treat and control diseases caused by pathogens from across the Nematoda.

RESULTS

Sequencing, assembly and gene organization

Data were generated from whole-genome shotgun sequencing and hierarchal map–assisted sequencing[6]. The assembly totaled 64 Mb (Online Methods, **Supplementary Note** and **Supplementary Table 1**), which is in line with recent genome size estimates made by flow cytometry (1 C = 71 Mb)[6,7]. The data provided a coverage level of 35-fold, with 15% of the supercontigs encompassing 90% of the genome. The *T. spiralis* fingerprint clone map enabled construction of 9 ultracontigs comprised of 69 supercontigs representing 49 Mb, or 76% of the genome.

The repeat content of the *T. spiralis* genome is estimated at 18%. The repeats have a low GC content (27%) relative to the genome overall (34%) and to protein coding regions (43%). The 15,808 protein-coding sequences occupy 26.6% of the genome at an average density of 272 genes per Mb. Although 15% of *C. elegans* genes are organized in operons[8], spatial relationships of genes in *T. spiralis* do not readily support the existence of operons (**Supplementary Note**). This observation validated prior studies indicating similar findings[4]. As such, the existence of operons in this nematode remains an open question. Further, *T. spiralis* lacks both the canonical SL1 trans-spliced leader found in most nematodes and the SL2 trans-spliced leader that is spliced onto transcripts from downstream genes in *C. elegans* operons. To date, at least 15 distinct spliced leaders encoded by 19 SL RNA genes have been identified in *T. spiralis*[4]; however, these putative splice leaders show sequence variability at nearly all base positions, and we found them to be present in only 1% of the complementary DNAs (cDNAs) examined. It is likely, therefore, that the canonical SL1 and SL2 spliced leader sequences were not part of the genetic repertoire in nematodes that diverged early in the evolution of the Nematoda. This hypothesis is supported in part by our inability to identify canonical SL1 and SL2 sequences among *Trichuris muris* expressed sequence tags (EST) as well (data not shown). After comparison to an extensive collection of proteins from other species, 45% (7,251) of the predicted protein coding genes were *T. spiralis* specific, of which 12% had EST confirmation (**Supplementary Fig. 3**). The amino acid composition of the predicted proteins in *T. spiralis* is similar to that observed in other nematodes[9], organisms (**Supplementary Table 2**) and taxa[10]. In agreement with previous studies[11], nematodes show a correlation between amino acid usage and the degree of codon degeneracy ($R = 0.74$).

Genome evolution

The availability of the genome from a member of Dorylaimia expanded our abilities to evaluate genome evolution among highly divergent crown clades and to potentially identify factors underlying lineage diversification. We evaluated changes associated with nematode evolution in relation to (i) genome organization; (ii) births and deaths of gene families; (iii) gene duplications and deletions that have occurred within gene families; and (iv) linear organization of orthologous genes.

We evaluated organizational characteristics by comparing the genomes of *T. spiralis* and *C. elegans*. The number of predicted genes in *T. spiralis* is notably lower than the 20,140 genes identified in *C. elegans*, even though the two genomes show similar repeat content and gene density. A comparison of approximately ~3,400 predicted orthologous genes (based on reciprocal best BLAST hits) showed that *T. spiralis* has a significantly shorter average intron size (191 bp compared to 391 bp; $P = 6.5 \times 10^{-69}$) amidst an average exon size that is relatively similar for the two species (179 bp for *T. spiralis* and 226 bp for *C. elegans*; $P = 7.0 \times 10^{-3}$). Focusing only on predicted orthologous genes with 20 or more exons, the mean total length for all exons was significantly higher in *C. elegans* ($P = 0.001$). Comparisons of the domains in the Pfam database that are contained in orthologous pairs showed that *C. elegans* had significantly more domains compared to the orthologous *T. spiralis* genes (876 genes compared to 755 genes; $P < 0.01$). These differences coincide with the smaller size of the *T. spiralis* genome; however, we cannot rule out the possibility for higher numbers of gene fragments in *T. spiralis* resulting from less refined genome annotation.

Delineating gene family emergence and extinction within phylogenetically related organisms can identify molecular determinants that underlie species (and pathogen) adaptation and lineage or species evolution. Such an approach has been used in analyzing nematode EST[12–14]. Here we measured potential emergence and extinction events of protein families across Nematoda. The analysis included species from four major lineages that collectively span the phylum (*C. elegans, Meloidogyne incognita*[15], *Brugia malayi*[16] and *T. spiralis*). These species represent nematodes that are non-parasitic, parasitic in plants and parasitic in animals, respectively, thus representing diverse trophic ecologies. Arthropod (*D. melanogaster*[17]) and yeast (*Saccharomyces cerevisae*[18]) species were used as outgroups. Markov clustering[19] of the complete protein catalog (87,406 proteins) comprising all six species generated 12,163 protein families (**Supplementary Table 3**). Inter-specific protein families overlaid onto species phylogeny identified 702 protein families at the node between Nematoda and the outgroups (**Fig. 1a** and **Supplementary Table 4**). Of these nematode families, 274 families were common among all four members of the Nematoda. We screened the genes in the 274-family core nematode group (1,990 genes) against all available nematode ESTs and cDNAs and found 73% shared homology to nematode transcriptome data from 27 nematode genera and only 5% shared sequence homology to arthropods using the same cutoff value. These numbers do not preclude gains that may have occurred before the appearance of the Nematoda or gains relative to *Drosophila* that may still be present in other arthropods. In contrast, we identified 88 protein family deaths as common among the four nematodes relative to *D. melanogaster*. Protein family deaths outnumbered births for all three parasitic species, whereas in the non-parasitic species *C. elegans*, births outnumbered deaths four to one. The methods used here will allow future assessment of this tendency with the availability of additional genomes from other parasitic and non-parasitic nematodes. We observed emergence of new protein families in all nematode lineages, albeit less so in *B. malayi*. Accordingly, it is now possible to explore the relevance of protein families identified in the evolution of lineages within the Nematoda and across phyla.

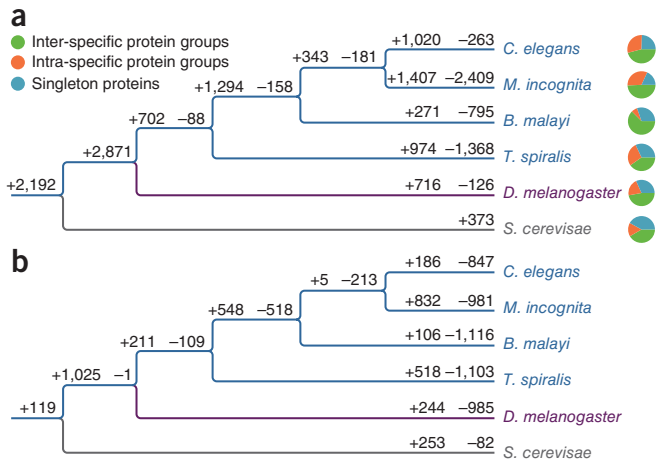

Figure 1 Protein and gene family changes associated with the origin and evolution of Nematoda. (**a**) Protein family changes. At the branch of each lineage, the '+' number indicates family birth events, and the '–' number indicates family death events represented by all members indicated for that lineage. For example, there are 702 protein family births ancestral to the phylum Nematoda and 88 protein family deaths in common among the four nematodes in comparison to arthropods (represented by *D. melanogaster*). We reconstructed these events from 12,206 interspecific orthologous families (63,273 proteins). (**b**) Gene duplications and losses over the evolution of the common protein families. We reconstructed the gene duplication and loss events using 858 orthologous multi-member protein families (containing 8,260 proteins) conserved among all six species. At the branch of each lineage, the '+' number indicates the number of gene duplication events, and the '–' number indicates the number of gene loss events for that lineage.

Similarly, quantitative changes in protein family members (duplications and deletions) can reflect evolutionary determinants of lineage and species diversity. We evaluated 858 families (8,260 genes) common to the four nematode species and two outgroup species defined above (**Fig. 1b**); 674 families had no obvious duplications or deletions, 70 had only deletions, 105 had only duplications and 9 had both. Nematode species had higher numbers of events compared to *D. melanogaster* (**Fig. 1b**). Among the nematodes, *M. incognita* had the highest number of both duplications and deletions, likely due to 30% of the genome being duplicated, resulting in more species-specific events[15]. An example for *T. spiralis* involves the secreted DNase II–like protein family, a member of which has been evaluated as a vaccine candidate[20] and which has been implicated in host-parasite interactions. The genome shows more extensive expansion of this family (an estimated 125 genes) than previously realized (**Supplementary Note** and **Supplementary Fig. 4**).

To provide additional examples, we compared protein families in *C. elegans* with sequence homologs in *T. spiralis*. Ten families were relatively expanded and five families were contracted in *T. spiralis* ($P < 0.001$) (**Supplementary Table 5**). These families can be grouped into (i) those present before the separation of nematodes and arthropods (nine families) and (ii) those putatively born coincident with this separation (six families) and possibly the origin of nematodes. The six protein families in this later group included four that are relatively expanded in *T. spiralis*: a retrotransposon (2:201 Ce (*C. elegans*): Ts (*T. spiralis*)); a translation initiation factor 2C, putatively related to lipid metabolism (2:140 Ce:Ts); a zinc finger C2H2 type protein (1:14, Ce:Ts); and a hypothetical protein (1:44, Ce:Ts) associated with defective egg laying in *C. elegans*. Two protein families are relatively contracted in *T. spiralis*: a major sperm protein (33:1, Ce:Ts) and a protein of unknown function, DUF1647 (18:1, Ce:Ts).

Comparisons of orthologous protein families outlined in sections two and three facilitated assessment of a nematode genome (*T. spiralis*) from a basally positioned clade (clade 2) with those from highly divergent clades (clades 8, 9 and 12)[21] and an outgroup member (*D. melanogaster*). Results consistently demonstrated similar and extensive levels of disparity in orthologous family sizes between *T. spiralis* and either *C.elegans* or *D. melanogaster*, whereas members of clades 8, 9, and 12 showed higher levels of shared attributes with *C. elegans* only (**Fig. 2**).

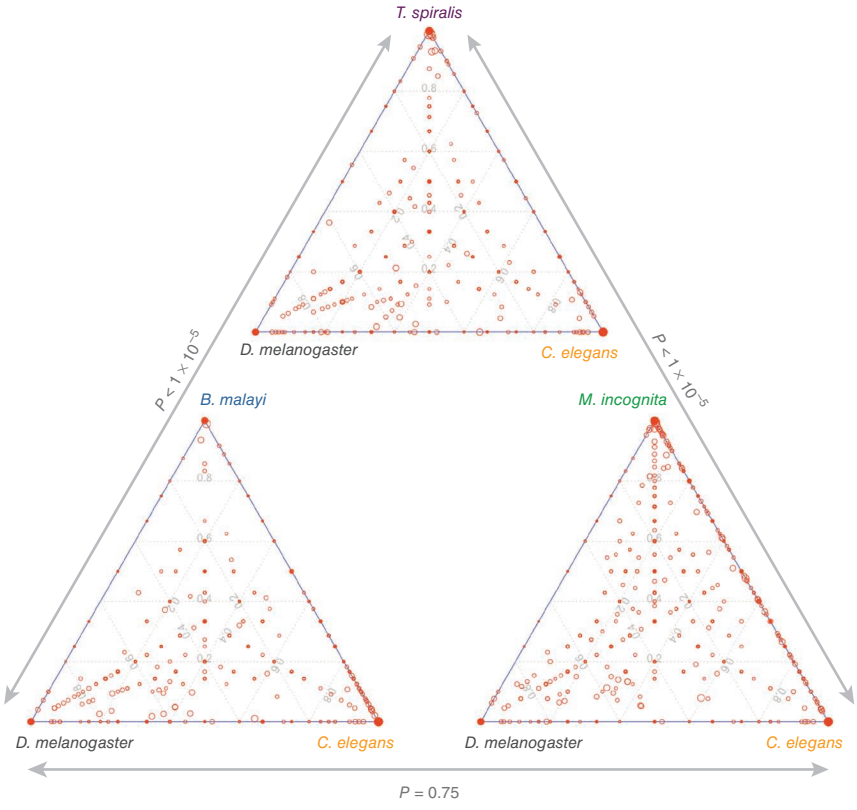

Figure 2 Comparison of orthologous protein families among nematodes that span the phylum. Orthologous families comprised of each of the three parasites and *D. melanogaster* and *C. elegans* are plotted separately. The size of the dot represents the size of the orthologous family; the position represents the composition of the family based on the three represented species. With the assumption that evolutionarily close species have similar orthologous family size (fewer duplications and deletions), these plots illustrate that *T. spiralis* is equally distinct from both *C. elegans* and *D. melanogaster*, whereas the two other parasites share greater commonality with *C. elegans*. *P* values (derived using a χ^2 test in pairwise plot comparison) indicate a greater number of families present in *C. elegans* compared to *D. melanogaster* and show that statistically significantly ($P < 1 \times 10^{-5}$) fewer families are biased to *C. elegans* when *T. spiralis* is present in the orthologous family.

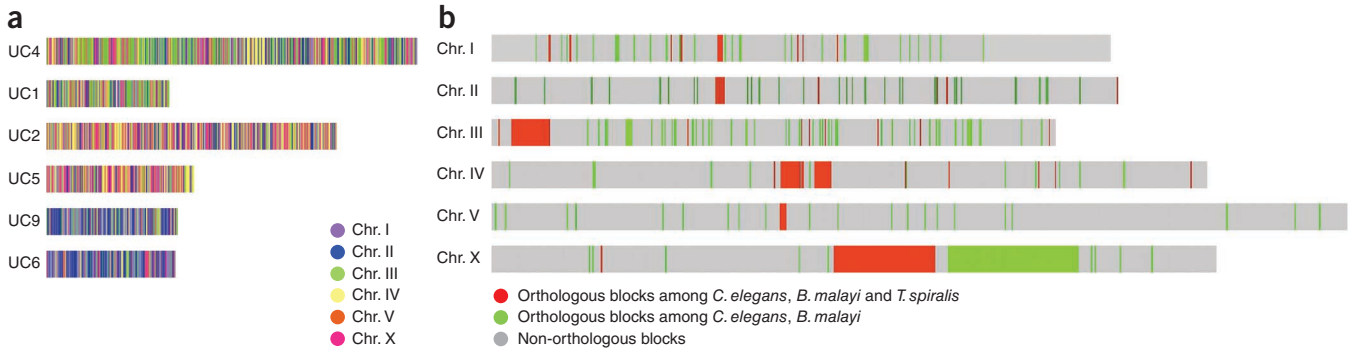

Figure 3 Genes from *T. spiralis* show macrosyntenic relationships with predicted orthologs from other nematodes. (**a**) *T. spiralis* genes on the six largest ultracontigs with orthologs in *C. elegans*, colored to indicate the *C. elegans* chromosome on which the ortholog is located. The correlation was strong ($R = 0.95$, $R = 0.76$ and $R = 0.99$) and was even stronger when the X chromosome was excluded ($R = 0.97$, $R = 0.97$ and $R = 0.99$). For example, $R = 0.95$ indicates that genes from both *T. spiralis* ultracontigs 1 and 4 are strongly associated with one predominant *C. elegans* chromosome, chromosome 3, and this organization is not a result of random gene distribution. (**b**) Orthologous segments shared among nematode species shown on the *C. elegans* chromosomes. Red segments are considered to be ancestral orthologous segments among nematodes. The size of segments corresponds to the *C. elegans* orthologous segment that may be different than the orthologous segment in the other two species (**Supplementary Table 7**).

Information in the next section provides independent measures, based on genome organization, to support this data which previously was indicated by rRNA sequence comparisons[21].

Next we evaluated the nematode genomes across the phylum regarding extent and limits to evolutionary changes and functional associations that may depend on gene arrangements. Comparisons between *C. elegans* and *B. malayi* (~350 million years of separation) indicated that intra-rather than inter-chromosomal rearrangements preferentially characterize genome evolution evident between these species[16]. We used the *T. spiralis* genes organized on the six longest ultracontigs to extend this analysis. For *B. malayi*, *T. spiralis* genes showed macrosyntenic relationships with predicted orthologs from *C. elegans* ($P < 0.0001$), albeit to a lesser extent (**Fig. 3a**). Because the X chromosome in *T. spiralis* is diploid only in females of these species (female $2n = 12$ (XX) and male $2n = 11$ (XO)), we also calculated the correlation coefficient when the X chromosome was excluded. This resulted in improved support for macrosynteny. This non-random distribution of orthologous genes is consistent with that observed in several nematode species[22–24].

Assuming a constant tendency toward randomness, genome reassortment is expected to occur at a rate commensurate with evolutionary distance. Using syntenic blocks of *C. elegans* for standardization, we measured the dynamics of nematode chromosome reassortment among multiple nematode pairs[25]. We observed the highest syntenic conservation score between *C. elegans* and *C. briggsae* (0.752), less so between *C. elegans* and *B. malayi* (0.508) and the least between *C. elegans* and *T. spiralis* (0.28) (**Supplementary Table 6**). Because sequences for non–*C. elegans* genomes have varying levels of fragmentation, it was not possible to use entirely complementary gene sets in the pairwise comparisons (we did not consider orthologous genes on different scaffolds). Nevertheless, the relative syntenic conservation values were consistent with the perceived evolutionary distance of the species investigated. The approximate 72% of the *T. spiralis* genome organization that lacked demonstrable congruence with the *C. elegans* genome provided a tentative estimate on the limits of evolutionary diversity of this kind across the Nematoda.

Despite an anticipated tendency toward randomization, the existence of syntenic blocks suggests functional constraints to genome evolution. We investigated this possibility with a high-level orthology map created with coding exons as anchors[26] from *C. elegans*, *B. malayi* and *T. spiralis*. We identified 196 orthologous segments (**Supplementary Table 7**); 155 of these segments were shared among *C. elegans* and *B. malayi*, 5 were shared among *B. malayi* and *T. spiralis* and 36 were shared among all three species, putatively defined as ancestral orthologous segments. No segments were shared exclusively between *C. elegans* and *T. spiralis* (**Fig. 3b**). These results are again consistent with the perceived evolutionary distance among these organisms based on all pairwise comparisons. The genes within the 36 ancestral segments accounted for ~50% of the genes in all segments for *C. elegans* and *B. malayi* but accounted for 97% of the genes in *T. spiralis*. Over half of the ancestral segments are located on *C. elegans* chromosomes 3 and 4. These ancestral segments tended to localize more centrally in the chromosomes ($P = 0.001$)[27]. This tendency was also suggested by the two-species orthologous segments, although it was less evident there (different at $P = 0.1$). The overall patterns highlighted likely reflect basic properties that influence the evolution of genome organization in nematodes.

The nematode species from the lineages evaluated span recent and early radiation events within the phylum Nematoda. Hence, the quantitative and qualitative measures of genomic diversity will help to define both the extent and limits of genome organizational diversity across the Nematoda and help clarify molecular determinants of nematode lineages and species. Nevertheless, the results based on Markov clustering of predicted orthologous protein families will exclude other forms of diversity such as nucleotide substitutions, insertions and deletions. As such, the documented differences reflect but a small component of the total genomic diversity within the Nematoda.

Molecular determinants archetypical of the phylum Nematoda

We evaluated the molecular determinants for traits that characterize the archetypical nematode[12,14]. To identify proteins and protein sequences that are broadly conserved among the four nematodes that span the phylum, we further compared worm-derived proteins to those of arthropod and yeast outgroups. The 12,163 orthologous protein families were partitioned into (i) orthologous protein sequences that are broadly conserved among all of the four nematode species and any of the two outgroups (2,517 families, 14,801 nematode proteins); (ii) those conserved exclusively among the four nematodes (274 families, 1,990 nematode proteins); and (iii) those that are conserved between any nematode and any outgroup (4,980 families, 30,729 proteins) (**Supplementary Table 3**). We evaluated 328 protein families represented by a single copy gene in all six species by querying the *C. elegans* database for RNAi phenotypes. The exclusion

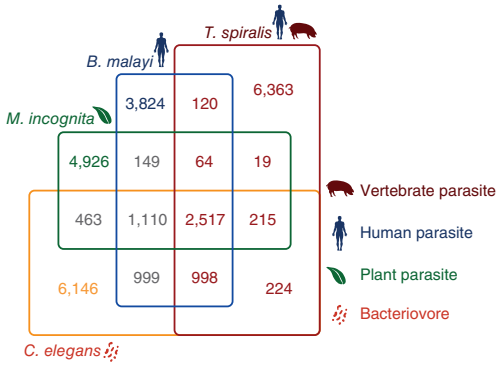

Figure 4 Distribution of orthologous families among the four nematode representatives spanning the phylum Nematoda. The lineages represented in the Nematoda are Rhabditida (*C. elegans*), Tylenchina (*M. incognita*), Spirurina (*B. malayi*) and Dorylaimia (*T. spiralis*). The trophic ecology of each of the four nematode species used in this study for the pan-phylum analysis is indicated next to the species name. The 2,517 orthologous groups are conserved in all four nematodes. Sixty-four orthologous groups are conserved among the parasitic species but not in the free-living *C. elegans*. The enrichment of functional categories related to certain orthologous groups compared to the complete functional repertoire for the four nematode species is presented in **Supplementary Tables 8** and **9**.

of multi-member protein families from this evaluation precluded cases where compensation by other family members might obscure RNAi phenotypes. Of the 328 *C. elegans* genes, 232 (71%) had associated RNAi phenotypes (significant enrichment at $P < 0.00001$) consistent with a gene set essential to core cellular and biochemical functions of eukaryotes (**Supplementary Table 8**).

Of the 2,517 nematode protein families (**Fig. 4**), 274 were detected in all four nematodes only (**Supplementary Note**), and we refer to these collectively as Nematode Orthologous Groups (NOGs) (**Supplementary Table 9** and **Supplementary Fig. 5**). These NOGs were significantly enriched ($P < 0.00001$) for genes with RNAi phenotypes in *C. elegans* and likely represent a gene set essential to core cellular and biochemical functions of nematodes.

The 274 NOGs encoded 189 multi-copy gene families and 85 single-copy gene families (scNOGs). Sixty-eight of the scNOGs had RNAi information and 21 had observable RNAi phenotypes (**Table 1** and **Supplementary Table 9**). There was no enrichment of RNAi phenotypes in the *C. elegans* genes in scNOGs compared to all *C. elegans* genes ($P < 0.05$). Nevertheless, among the 21 genes with phenotypes, 8 had known tissue localization and only 1 was neuronal. Of the remaining 64 genes, 17 had known expression patterns, of which 10 were neuronal. Therefore, the biological importance of the scNOGs may be underestimated by RNAi information because nervous tissue is relatively insensitive to RNAi (for example, see ref. 28).

Nematode-specific amino acid sequences in scNOG proteins may have practical importance for functional investigations. As such, we evaluated the scNOGs sequences for molecular features by forced alignment with non-nematode homologs (human, chicken, frog and zebrafish) associated with the same Pfam entries. We categorized the scNOGs into two groups: (i) those involving nematode-specific insertions and deletions (InDels) (for example, see ref. 29) relative to non-nematode homologs (15 proteins) (**Supplementary Fig. 6a**) and (ii) those involving unique patterns of conservation independent of InDels (70 proteins) (**Supplementary Figs. 6b** and **7**) (for example, see ref. 14). Sequence variation exclusive of conserved motifs was generally higher among the nematode proteins than among the vertebrate proteins, even though evolutionarily, each comparison spanned similar predicted lengths of time, consistent with a previous report[30] (**Supplementary Fig. 8**). Therefore, pan-Nematoda–specific conservation has persisted despite the high evolutionary rate in adjacent sequences of these NOGs.

The nematode-specific amino acid sequences in NOGs may have fundamental importance across the Nematoda. For instance, the predicted subunit of an electron transfer complex (**Supplementary Fig. 6a**) has well-defined insertions, and a severe RNAi phenotype is associated with the *C. elegans* member of this NOG. As such, comparative information from the vertebrate homolog may guide experiments to dissect the functional roles of the NOG insertions. Furthermore, a sequence containing amino acid insertions in one protein interaction partner may be compensated by deletions in the other protein interaction partner. We indeed identified that the interaction partner of the complex to

Table 1 Pan-phylum single-copy genes with the *C. elegans* ortholog having severe RNAi phenotype

						Structural annotation	
T. spiralis gene	Ortholog in *B. malayi*	*C. elegans*	*M. incognita*	Descriptor[a]	*C. elegans* RNAi[b]	TM[c]	SP[d]
Tsp_14949	14972.m07791	F39H12.2	Minc14650	Hypothetical, WD40 repeat-like	Emb	–	–
Tsp_03879	14330.m00196	F28F8.6	Minc16561	Machedo-Joseph protein	Emb	–	–
Tsp_09591	14058.m00575	M05B5.2	Minc04214	Hypothetical protein	Lon Unc thin Gro	Y	Y
Tsp_02563	14972.m07706	F53B6.1	Minc01712a	Tetraspanin family protein	Lva Dpy Bmd Bli	Y	–
Tsp_07476	14379.m00149	W01A8.4	Minc03402	NADH dehydrogenase (ubiquinone) 1 beta subcomplex 4	Lva Emb Bmd	Y	–
Tsp_05829	14961.m05209	ZK899.2	Minc06660	Hypothetical protein	Lva Emb Lvl Gro	Y	–
Tsp_10274	13068.m00024	F44G4.2	Minc14463	NADH dehydrogenase (ubiquinone) 1 beta subcomplex 2	Lva Emb RBS	–	–
Tsp_05872	14972.m06963	ZK682.5	Minc05446a	Leucine Rich Repeat family protein	Lva Gro	Y	Y
Tsp_05373	13756.m00013	C45B2.7	Minc06522	Patched related family protein 4	Prl Unc Lva Dpy Emb Lvl	Y	–
Tsp_09505	14968.m01485	W08F4.6	Minc18112	Hypothetical protein	Prl Unc Lva Lvl Bmd Ela	–	–
Tsp_10877	14992.m10900	T19B10.2	Minc10356	Hypothetical protein	Prl Unc Tsla Rup Gro	–	Y
Tsp_11032	13644.m00292	C09H10.7	Minc15358	Hypothetical protein	Pvl Da, Emb Stp	–	–
Tsp_10369	13847.m00044	F10E7.6	Minc16059	Hypothetical protein	Sck Clr Ela Gro	Y	–
Tsp_01966	14972.m07319	W04G3.2	Minc11161	Lipocalin protein	Unc Lva Lvl Bmd	–	Y
Tsp_10030	14961.m05181	Y8G1A.2	Minc07816	Innexin membrane protein	Unc Rup Stp Gro	Y	–

[a]Descriptor, annotation based on KEGG Orthology and Interpro. [b]RNAi phenotype description (http://www.wormbase.org/). [c]TM, transmembrane. [d]SP, signal peptide for secretion.

which that protein belongs (long chain Acyl-CoA dehydrogenase, with which interaction has been confirmed experimentally[31]) has deletions in the non-nematode protein (**Supplementary Note**, **Supplementary Figs. 9** and **10**).

This series of analyses identified genes and proteins that may have fundamental importance in all nematode species. Two categories of nematode-specific sequences are responsible for delineation as scNOGs. Therefore, scNOGs, and most likely other NOGs, contain pan-phylum nematode-specific sequences incorporated either into universally conserved protein structures or into protein structures that are unique to the Nematoda. Evidence reflecting biological importance highlights the potential for NOGs to serve as targets for control of parasitic nematodes that infect humans, animals and plants while potentially limiting risk to the host.

Core- and phylogenetically-restricted functional categories

A question of central importance is whether or not parasitic nematodes (and potentially other parasites) have evolved independently or have preferentially retained common solutions to challenges of parasitism despite their exploitation of widely divergent trophic ecologies (for example, see ref. 32). Much interest in this context has focused on (i) secretory proteins, (ii) molecular functions and (iii) biochemical pathways that are conserved or taxonomically restricted.

Although not all secretory proteins from parasitic nematodes are involved in interactions with the host, constituents of this protein category are prime candidates for examining the host-pathogen interface. Here, we sought proteins that are broadly conserved among nematodes or among parasitic nematodes. We sorted these proteins into orthologous protein groups shared among species representing diverse parasite lineages and then subgrouped them into those with secretory peptides (**Supplementary Fig. 11**). We interrogated predicted secretory protein orthologs with previously identified secreted proteins using an orthogonal approach based on excretory-secretory products in *T. spiralis* and *B. malayi* identified by tandem mass spectrographic analysis[33,34]. We identified only two proteins as secretory and common to each parasite member (including vertebrate and plant parasites) but which were absent from the non-parasitic *C. elegans*: (i) a serine peptidase member of the prolyl oligopeptidase family that can be critical for invasion of the mammalian host cells by protozoan parasites[35] and (ii) a cyanate hydratase that in other organisms hydrolyzes and detoxifies environmental cyanate[36]. Our results suggest that the number of conserved secretory proteins broadly involved in nematode interactions with hosts may be relatively few. Nevertheless, this number is likely to increase when reducing our analysis to subgroupings of parasitic nematodes, as we found when we interrogated proteomes for any two of the three parasitic species here.

Among the *T. spiralis* genes analyzed, 35% (5,456 out of 15,808) could be assigned one or more Gene Ontology (GO) terms. We assigned putative molecular functions to 90% of this 35%, biological processes to 68% and cellular components to 45%. The remaining two-thirds of the genes in *T. spiralis* represent uncharacterized and possibly new functions in the parasite. A set of 25 molecular functions were significantly enriched or depleted (at $P < 0.01$) when we compared intra- or inter-specific orthologous groups to the complete repertoire of GO terms for *T. spiralis* (**Supplementary Table 10** and **Supplementary Fig. 12**). Among the orthologous families confined only to *T. spiralis* and *C. elegans,* rhodopsin-like receptor activity was enriched, a possible consequence of the number of genes involved in G-protein–coupled receptor protein signaling pathways. In orthologous groups with members only from *T. spiralis* and *B. malayi*, the enriched category involved steroid-binding proteins.

Among a total of 71 molecular GO categories identified, 42 were enriched and 29 were depleted in the 2,517 nematode orthologous families (including *C. elegans*) by comparison to the complete proteomes of the four nematode species (**Supplementary Table 11**). When considering the 64 orthologous groups conserved among the three parasitic nematodes, nine GO categories were statistically enriched or depleted; ATP binding was the only depleted category, whereas DNA- and RNA-binding, aspartic-type endopeptidase and prolyl oligopeptidase activities were among those enriched (**Supplementary Table 12**). Therefore, commonalities in molecular functions may exist even among parasites from widely diverse ecological niches. Further light will be shed on genetic associations among parasitic and non-parasitic nematodes as more robust comparisons among species from each category begin to surface.

Guided by the possibility that parasitic nematodes undergo reductive genome evolution because of reliance on the metabolic capacity and homeostatic buffering of their host, we compared *T. spiralis* genes encoding enzymes to similar genes from the other parasites and the non-parasitic *C. elegans*[37,38] (**Supplementary Fig. 13**) and the NemaCyc viewer (**Supplementary Fig. 14**). We found that the parasitic species had fewer KOs (Kyoto Encyclopedia of Genes and Genomes (KEGG) database orthology) associated with their genes (~522–548) compared to *C. elegans* (704) (**Table 2** and **Supplementary Table 13**). The number of genes correlated with the number of associated KOs. Therefore, we examined the KOs in relation to nematode lineages used in this study. Among the 785 KOs associated with the nematode

Table 2 Genes and KEGG Orthologies (KOs) represented in metabolic patways in four nematodes

				C. elegans		*M. incognita*		*B. malayi*		*T. spiralis*	
Pathway	KOs in KEGG reference pathway	Represented KOs in nematodes	Conserved KO in nematodes	Genes	KOs	Genes	KOs	Genes	KOs	Genes	KOs
1. Metabolism	2,258	785	337	2,480	704	1,822	525	1,132	548	1,069	515
1.1 Carbohydrate metabolism	550	192	92	626	167	499	130	294	133	252	145
1.2 Energy metabolism	408	131	71	235	123	210	97	144	107	123	87
1.3 Lipid metabolism	325	144	52	710	122	380	98	218	101	199	87
1.4 Nucleotide metabolism	174	78	35	306	74	294	52	182	51	182	53
1.5 Amino acid metabolism	484	188	75	607	174	430	129	250	114	266	124
1.6 Metabolism of other amino acids	126	55	26	222	50	119	39	73	41	76	39
1.7 Glycan biosynthesis and metabolism	160	83	30	163	74	153	54	95	63	89	55
1.8 Biosynthesis of polyketides and nonribosomal peptides	4	2	1	5	2	6	1	4	2	1	1
1.9 Metabolism of cofactors and vitamins	301	91	31	392	80	298	55	174	57	185	56
1.10 Biosynthesis of secondary metabolites	55	25	13	234	20	115	18	59	19	47	18
1.11 Xenobiotics biodegradation and metabolism	178	61	27	548	55	249	40	125	37	119	38

species evaluated herein, 337 were shared among all four species (core nematode KOs, CNKs). The pathway that had most of the KOs as CNKs was the energy metabolism pathway (53% of all KOs were conserved across all four species), and the pathway with the least KOs was the metabolism of cofactors and vitamins pathway (34% of the KOs were in all four species). Among the energy metabolism pathways, there were 96 KOs related to oxidative phosphorylation, 52 of which were conserved among all four nematodes. This result supports previous observations in which parasite enzymes involved in oxidative phosphorylation exhibited sequence divergence from similar host proteins. These differences were largely associated with nematode-specific insertions[14,29]. Despite the high level of conservation, the number of CNKs among all four nematodes was very low (34%), suggesting that different adaptations distinguish nematodes with distinct modes of existence.

DISCUSSION

Here we present the genome sequence of *T. spiralis*, a member of Dorylaimia and a lineage that diverged early in the evolution of the phylum Nematoda. The draft sequence of *T. spiralis* covered over 90% of the estimated genome and expected genes. Coupled with genomes from nematode lineages depicting more recent episodes of divergence, the *T. spiralis* data provide new perspectives on genomic evolution that more broadly spans the Nematoda.

The *T. spiralis* genome sequence and the accompanying genome-mining analysis address four key issues. First, details of genomic diversity that were deduced among species have outlined molecular determinants, where the magnitude of change likely reflects molecular elements that have figured decisively in both the lineage and species evolution of Nematoda (for example, see refs. 39–41). It has been argued that such drastic differences can be related to functional diversification, speciation and species adaptation. Given the modest number of nematode species with available genomes, we fully expect that as additional nematode genome sequences become available, a much greater resolution of differences will occur. Nonetheless, the results presented here helped resolve many specific genomic characteristics that can be further investigated in this context. Second, host characteristics may select for common parasite characteristics of otherwise widely disparate nematode species. The similarities in the steroid-binding protein family common to two parasites of humans and mammals, *T. spiralis* and *B. malayi*, were distinct from a large family of related nuclear hormone receptors in *C. elegans*, many of which are homologous to steroid-binding receptors in other organisms[42]. This distinction provides support for convergent enrichment of common steroid-binding receptors in these two parasites of humans and other mammals, possibly dictated by characteristics of the host environment, as previously suggested[43]. Third, the new databases guided discovery of genes and proteins that appear to have fundamental importance to all nematode species (archetypical characteristics). Accordingly, the NOGs were significantly enriched for genes with RNAi phenotypes in *C. elegans*. Success in circumscribing archetypical nematode characteristics from pan-phylum databases will serve to refocus research on characteristics that have the broadest application for controlling pathogens of humans, animals and plants. Fourth, these results provide a valuable resource to investigate the biology of the intracellular pathogen *T. spiralis*. One example involves a DNase II gene family of *T. spiralis*, which includes secreted proteins previously implicated in host-parasite interactions and immune control[20]. The curious expansion and diversification of this family in comparison to other nematodes can now be related to unique characteristics of *T. spiralis* and possibly the lineages it represents. A second example centers on why species within this genus have separated into those that generate protective capsules from those which do not, a characteristic which is not host related. There are innumerable anticipated applications of the genome data toward elucidating the biology, methods for immune control and treatments of this parasite. The comparative value of this genome sequence will extend these applications well beyond this species and phylum.

URLs. RepeatMasker, http://repeatmasker.org/; RNAmmer, http://www.cbs.dtu.dk/cgi-bin/nph-sw_request?rnammer; Rfam database, http://selab.janelia.org/software.html; BER, http://ber.sourceforge.net/; PHYLIP, http://evolution.genetics.washington.edu/phylip.html.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession codes. The *T. spiralis* Whole Genome Shotgun project (project id 12603) has been deposited at DNA Data Bank of Japan, EMBL and GenBank under the accession ABIR00000000. The version described in this paper is the second version, (contigs, ABIR02000001–ABIR02009267; scoffolds, GL622784–GL629646; proteins, EFV46182–EFV62561).

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

ACKNOWLEDGMENTS

We thank A. Cutter and members of the Genome Center for discussion and helpful comments on the manuscript. This work was supported by a National Human Genome Research Institute grant to R.K.W. (HG003079) and a National Institute of Allergy and Infectious Diseases grant to M.M. (81803) and J.A. (14490).

AUTHORS CONTRIBUTIONS

M.M., D.P.J., J.P.M., D.S.Z., E.R.M. and R.K.W. initiated the project; J.A. and D.S.Z. provided all the worms for the shotgun and D.P.J. for the cDNA sequencing; L.F. and R.S.F. directed sequencing and sequence improvement; S.-P.Y., P.M. and W.C.W. assembled the genome and evaluated the assembly; V.B., X.Z. and K.H.-P. directed annotation; M.M., Z.W., S.A., J.M., Y.Y. and C.M.T. contributed to most of the specific analysis presented in this manuscript; M.M., D.P.J., D.S.Z. and S.W.C. directed the project and assembled the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Putnam, N.H. *et al.* The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**, 1064–1071 (2008).
2. Lavrov, D.V. & Brown, W.M. Trichinella spiralis mtDNA: a nematode mitochondrial genome that encodes a putative ATP8 and normally structured tRNAs and has a gene arrangement relatable to those of coelomate metazoans. *Genetics* **157**, 621–637 (2001).
3. 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).
4. Pettitt, J., Müller, B., Stansfield, I. & Connolly, B. Spliced leader trans-splicing in the nematode *Trichinella spiralis* uses highly polymorphic, noncanonical spliced leaders. *RNA* **14**, 760–770 (2008).
5. Zarlenga, D.S., Rosenthal, B.M., La Rosa, G., Pozio, E. & Hoberg, E.P. Post-Miocene expansion, colonization, and host switching drove speciation among extant nematodes of the archaic genus *Trichinella*. *Proc. Natl. Acad. Sci. USA* **103**, 7354–7359 (2006).

6. Mitreva, M. & Jasmer, D.P. Advances on sequencing the genome of the Clade I nematode *Trichinella spiralis*. *Parasitology* **135**, 869–880 (2008).
7. Zarlenga, D.S., Rosenthal, B., Hoberg, E. & Mitreva, M. Integrating genomics and phylogenetics in understanding the history of *Trichinella* species. *Vet. Parasitol.* **159**, 210–213 (2009).
8. Blumenthal, T. & Gleason, K.S. *Caenorhabditis elegans* operons: form and function. *Nat. Rev. Genet.* **4**, 112–120 (2003).
9. Cutter, A.D., Wasmuth, J.D. & Blaxter, M.L. The evolution of biased codon and amino acid usage in nematode genomes. *Mol. Biol. Evol.* **23**, 2303–2315 (2006).
10. King, J.L. & Jukes, T.H. Non-Darwinian evolution. *Science* **164**, 788–798 (1969).
11. Mitreva, M. *et al.* Codon usage patterns in Nematoda: analysis based on over 25 million codons in thirty-two species. *Genome Biol.* **7**, R75 (2006).
12. Wasmuth, J., Schmid, R., Hedley, A. & Blaxter, M. On the extent and origins of genic novelty in the phylum Nematoda. *PLoS Negl. Trop. Dis.* **2**, e258 (2008).
13. Parkinson, J. *et al.* A transcriptomic analysis of the phylum Nematoda. *Nat. Genet.* **36**, 1259–1267 (2004).
14. Yin, Y. *et al.* Molecular determinants archetypical to the phylum Nematoda. *BMC Genomics* **10**, 114 (2009).
15. Abad, P. *et al.* Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*. *Nat. Biotechnol.* **26**, 909–915 (2008).
16. Ghedin, E. *et al.* Draft genome of the filarial nematode parasite *Brugia malayi*. *Science* **317**, 1756–1760 (2007).
17. Adams, M.D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195 (2000).
18. Goffeau, A. *et al.* Life with 6000 Genes. *Science* **274**, 546–567 (1996).
19. Li, L., Stoeckert, C.J. Jr. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
20. Vassilatis, D.M. *et al.* Analysis of a 43-kDa glycoprotein from the intracellular parasitic nematode *Trichinella spiralis*. *J. Biol. Chem.* **267**, 18459–18465 (1992).
21. Holterman, M. *et al.* Phylum-wide analysis of SSU rDNA reveals deep phylogenetic relationships among nematodes and accelerated evolution toward crown clades. *Mol. Biol. Evol.* **23**, 1792–1800 (2006).
22. Stein, L.D. *et al.* The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol.* **1**, E45 (2003).
23. Lee, K.-Z., Eizinger, A., Nandakumar, R., Schuster, S.C. & Sommer, R.J. Limited microsynteny between the genomes of *Pristionchus pacificus* and *Caenorhabditis elegans*. *Nucleic Acids Res.* **31**, 2553–2560 (2003).
24. Guiliano, D.B. *et al.* Conservation of long-range synteny and microsynteny between the genomes of two distantly related nematodes. *Genome Biol.* **3**, RESEARCH0057 (2002).
25. Vergara, I.A. & Chen, N. Using OrthoCluster for the detection of synteny blocks among multiple genomes. *Curr. Protoc. Bioinformatics* **Chapter 6**, Unit 6.10 6.10.1–18 (2009).
26. Dewey, C.N. Aligning multiple whole genomes with Mercator and MAVID. *Methods Mol. Biol.* **395**, 221–236 (2007).
27. *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
28. Fraser, A.G. *et al.* Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference. *Nature* **408**, 325–330 (2000).
29. Wang, Z. *et al.* Systematic analysis of insertions and deletions specific to nematode proteins and their proposed functional and evolutionary relevance. *BMC Evol. Biol.* **9**, 23 (2009).
30. Mushegian, A.R., Garey, J.R., Martin, J. & Liu, L.X. Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes. *Genet. Res.* **8**, 590–598 (1998).
31. Schilling, B. *et al.* Proteomic analysis of succinate dehydrogenase and ubiquinol-cytochrome c reductase (Complex II and III) isolated by immunoprecipitation from bovine and mouse heart mitochondria. *Biochim. Biophys. Acta* **1762**, 213–222 (2006).
32. Bird, D.M. & Opperman, C.H. The secret(ion) life of worms. *Genome Biol.* **10**, 205 (2009).
33. Robinson, M.W. & Connolly, B. Proteomic analysis of the excretory-secretory proteins of the *Trichinella spiralis* L1 larva, a nematode parasite of skeletal muscle. *Proteomics* **5**, 4525–4532 (2005).
34. Moreno, Y. & Geary, T.G. Stage- and gender-specific proteomic analysis of *Brugia malayi* excretory-secretory products. *PLoS Negl. Trop. Dis.* **2**, e326 (2008).
35. Santana, J.M., Grellier, P., Schrevel, J. & Teixeira, A.R.L. A *Trypanosoma cruzi*–secreted 80 kDa proteinase with specificity for human collagen types I and IV. *Biochem. J.* **325**, 129–137 (1997).
36. Sung, Y.C. & Fuchs, J.A. Characterization of the Cyn operon in *Escherichia coli* K12. *J. Biol. Chem.* **263**, 14769–14775 (1988).
37. Martin, J. *et al.* Nematode.net update 2008: improvements enabling more efficient data mining and comparative nematode genomics. *Nucleic Acids Res.* **37**, D571–D578 (2009).
38. Wylie, T. *et al.* NemaPath: online exploration of KEGG-based metabolic pathways for nematodes. *BMC Genomics* **9**, 525 (2008).
39. Panhuis, T.M., Clark, N.L. & Swanson, W.J. Rapid evolution of reproductive proteins in abalone and *Drosophila*. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **361**, 261–268 (2006).
40. Kocher, T.D. Adaptive evolution and explosive speciation: the cichlid fish model. *Nat. Rev. Genet.* **5**, 288–298 (2004).
41. Givnish, T.J. *et al.* Molecular evolution, adaptive radiation, and geographic diversification in the amphiatlantic family Rapateaceae: evidence from ndhF sequences and morphology. *Evolution* **54**, 1915–1937 (2000).
42. Sluder, A.E., Mathews, S.W., Hough, D., Yin, V.P. & Maina, C.V. The nuclear receptor superfamily has undergone extensive proliferation and diversification in nematodes. *Genome Res.* **9**, 103–120 (1999).
43. Sluder, A.E. & Maina, C.V. Nuclear receptors in nematodes: themes and variations. *Trends Genet.* **17**, 206–213 (2001).

ONLINE METHODS

Sequencing, assembly and annotation. Rats were infected orally with muscle tissue containing first stage larvae (ML) of *T. spiralis* strain ISS 195. Infections were allowed to precede a minimum of 30 days, then the muscle tissue was digested and the parasite was collected. Genomic DNA was extracted from muscle larvae of *T. spiralis* using standard protocols. Whole genome shotgun, BAC and EST libraries were generated[3,6]. The assembly was performed using the PCAP package[44]. The physical map for *T. spiralis* was constructed using 26,784 clones (**Supplementary Note**).

The repeats were masked using RECON[45] and RepeatMasker (see URLs). Then the ribosomal RNA genes were identified using RNAmmer (see URLs). Transfer RNA genes were identified with tRNAscan-SE[46]. Noncoding RNAs were identified by sequence homology searches of the Rfam database (see URLs). Protein-coding genes were predicted using a combination of *ab initio* programs[47] and FgenesH (Softberry, Corp) and the evidence-based program EAnnot[48]. A consensus gene set from the above prediction algorithms was generated using a logical, hierarchical approach. Gene product naming was determined by BER (see URLs). The signal peptide for secretion and the trans-membrane-domain–containing proteins were identified using Phobius[49].

Protein families and genome evolution. OrthoMCL[19] was used to predict orthologous groups of proteins. Phylogenetic trees were built for protein families with one member from each of the six species using PHYLIP (version 3.69; see URLs) after aligning the family members with MUSCLE (version 3.7; ref. 50). The consensus tree of the trees was used as the phylogeny of the species. Death and birth of each protein family overlaid over species phylogeny was constructed using PHYLIP-DOLLOP by treating each protein family as a character. Gene duplication and deletion events of the families having members from each of the six species were reconstructed using URec[51], and a neighbor joining tree of each family was generated using PHYLIP-NEIGHBOR.

The dynamics of nematode chromosome reassortment among multiple nematode pairs was measured using OrthoCluster[25] and using syntenic blocks of *C. elegans* for standardization. For the identification of the ancestral orthologous regions, we used exons that are orthologous among species as map 'anchors'[52] (**Supplementary Note**).

Nematode-specific molecular features. A profile was built for each of the 85 scNOGs using HMMBUILD[53]. The profiles were calibrated using hmmcalibrate and each profile was used to search the Pfam database (release 23.0). Hits better than 0.1 were considered. The selected non-nematode species were of evolutionary distances similar to *C. elegans* and *T. spiralis*: human, chicken, zebrafish and frog. After identification of the non-nematode families that were associated with the same Pfam as the scNOGs, the multi-fasta files were aligned using MUSCLE. These alignments were used to build a distance matrix using PHYLIP-PROTDIST. RNAi source data were from Wormmart from Wormbase release 180. The core nematode groups were screened against nematode (~1.1 million ESTs and/or Roche/454 cDNAs) and arthropod (5.3 million ESTs) transcript data and sequence homology at 35 bits, and 55% identity cut-off was accepted as significant.

Structural annotation and comparison of interaction partners. The three-dimensional structure was modeled using the Rosetta3.0 software suite[54–56]. A total of 40,000 decoys were generated using the full-atom scoring method[57] for each sequence. Several of the decoys with a small radius of gyration and low all-atom energy (that is, the bottom of the energy well) were compared using TM-align[58] and MAMMOTH[59]. The position of the insertions was mapped onto the models generated. The secondary structure predictions calculated for the Rosetta *ab initio* program were added to the sequence alignment generated by MUSCLE[50]. The functional importance of the insertions in the electron transfer complex was further dissected by comparing interacting proteins. Two protein-protein interaction databases, IntAct[60] and MINT[61], were used to see if this protein or its orthologs were involved in a protein-protein interaction.

Functional associations and taxonomic restrictions. Default parameters for InterProScan (v16.1) were used to search against the InterPro database[62], and Gene Ontology (GO[63]) annotations were obtained with no additional curation (IEA associations only). These annotations have been displayed graphically by AmiGO and can be accessed at Nematode.net[37]. The significant enrichment of GO terms was computed based on the hypergeometric distribution using FUNC[64] (including false discovery rate, FDR). A probability refinement was done to remove the GO terms identified as significant due to their children terms. We used the FDR computed by FUNC to reduce false discovery. Therefore, unless specified otherwise, the GO term enrichment was selected based on both P value < 0.05 (after refinement) and FDR < 0.1.

The gene products were associated with a specific biochemical pathway using the KEGG pathway mappings[65]. WU-BLAST matches of the genes against KEGG database version 46.0 was used for pathway mapping with an E-value filter of 1×10^{-10}. Graphical presentation of the pathway associations was done using NemaPath[38]. The *C. elegans* NemaCyc viewer is based on mapping a BLASTP alignment of the KEGG's genes database against the predicted *T. spiralis* genes. Scores stronger than 1×10^{-10} were considered.

44. Huang, X., Wang, J., Aluru, S., Yang, S.-P. & Hillier, L. PCAP: a whole-genome assembly program. *Genome Res.* **13**, 2164–2170 (2003).
45. Bao, Z. & Eddy, S.R. Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Res.* **12**, 1269–1276 (2002).
46. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
47. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
48. Ding, L. *et al.* EAnnot: a genome annotation tool using experimental evidence. *Genome Res.* **14**, 2503–2509 (2004).
49. Käll, L., Krogh, A. & Sonnhammer, E.L.L. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **338**, 1027–1036 (2004).
50. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
51. Gorecki, P. & Tiuryn, J. URec: a system for unrooted reconciliation. *Bioinformatics* **23**, 511–512 (2007).
52. Dewey, C.N. Aligning multiple whole genomes with Mercator and MAVID. *Methods Mol. Biol.* **395**, 221–236 (2007).
53. Eddy, S.R. A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **23**, 205–211 (2009).
54. Andre, I., Bradley, P., Wang, C. & Baker, D. Prediction of the structure of symmetrical protein assemblies. *Proc. Natl. Acad. Sci. USA* **104**, 17656–17661 (2007).
55. Rohl, C.A., Strauss, C.E., Misura, K.M. & Baker, D. Protein structure prediction using Rosetta. *Methods Enzymol.* **383**, 66–93 (2004).
56. Qian, B. *et al.* High-resolution structure prediction and the crystallographic phase problem. *Nature* **450**, 259–264 (2007).
57. Misura, K.M., Chivian, D., Rohl, C.A., Kim, D.E. & Baker, D. Physically realistic homology models built with ROSETTA can be more accurate than their templates. *Proc. Natl. Acad. Sci. USA* **103**, 5361–5366 (2006).
58. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
59. Ortiz, A.R., Strauss, C.E. & Olmea, O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.* **11**, 2606–2621 (2002).
60. Kerrien, S. *et al.* IntAct–open source resource for molecular interaction data. *Nucleic Acids Res.* **35**, D561–D565 (2007).
61. Chatr-aryamontri, A. *et al.* MINT: the Molecular INTeraction database. *Nucleic Acids Res.* **35**, D572–D574 (2007).
62. Mulder, N.J. *et al.* InterPro, progress and status in 2005. *Nucleic Acids Res.* **33**, D201–D205 (2005).
63. The Gene Ontology Consortium. The gene ontology project in 2008. *Nucleic Acids Res.* **36**, D440–D444 (2008).
64. Prüfer, K. *et al.* FUNC: a package for detecting significant associations between gene sets and ontological annotations. *BMC Bioinformatics* **8**, 41 (2007).
65. Kanehisa, M. *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36**, D480–D484 (2008).

doi:10.1038/ng.769