Published online 21 October 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D571–D578***
doi:10.1093/nar/gkn744

Nematode.net update 2008: improvements enabling more efficient data mining and comparative nematode genomics

John Martin, Sahar Abubucker, Todd Wylie, Yong Yin, Zhengyuan Wang and Makedonka Mitreva*

The Genome Center, Department of Genetics, Washington University School of Medicine, St Louis, MO 63108, USA

Received August 8, 2008; Revised September 19, 2008; Accepted October 3, 2008

ABSTRACT

Nematode.net (http://nematode.net) is a publicly available resource dedicated to the study of parasitic nematodes. In 2000, the Genome Center at Washington University (GC) joined a consortium including the Nematode Genomics group in Edinburgh, and the Pathogen Sequencing Unit of the Sanger Institute to generate expressed sequence tags (ESTs) as an inexpensive and efficient solution for gene discovery in parasitic nematodes. As of 2008 the GC, sampling key parasites of humans, animals and plants, has generated over 500 000 ESTs and 1.2 million genome survey sequences from more than 30 non-*Caenorhabditis elegans* nematodes. Nematode.net was implemented to offer user-friendly access to data produced by this project. In addition to sequence data, the site hosts: assembled NemaGene clusters in GBrowse views characterizing composition and protein homology, functional Gene Ontology annotations presented via the AmiGO browser, KEGG-based graphical display of NemaGene clusters mapped to metabolic pathways, codon usage tables, NemFam protein families which represent conserved nematode-restricted coding sequences not found in public protein databases, a web-based WU-BLAST search tool that allows complex querying and other assorted resources. The primary aim of Nematode.net is the dissemination of this diverse collection of information to the broader scientific community in a way that is useful, consistent, centralized and enduring.

INTRODUCTION

Parasitic nematode infection represents a significant socio-economic concern on multiple levels. Human parasitic nematodes have a major, long-term impact (directly and indirectly) on human health and cause substantial suffering, particularly in children. The four most prevalent soil-transmitted species (*Ascaris lumbricoides, Trichuris trichura* and the hookworms, *Necator americanus* and *Ancylostoma ceylanicum*) alone infect nearly 3 billion people (1). Furthermore, there are species that in combination with their endosymbiont *Wolbachia* are causal agents of river blindness (*Onchocerca volvulus*), or elephantiasis (*Brugia malayi, Wucheria bancrofti*). Morbidity from nematodes is substantial, and rivals diabetes and lung cancer in worldwide disability-adjusted life year (DALY) measurements (2,3). While mortality is low in proportion to the number of infections, deaths may still total 100 000 annually. In addition to the direct threat to human health, parasitic nematodes also represent a sizeable problem for livestock and food crops. Its estimated that plant infections are responsible for $100 billion in annual crop damage, and that worldwide sales of livestock parasiticides (predominantly anthelmintics) are over $1 billion per year (4).

Nematode.net was established to provide access to, and information-mining tools for, genomic-scale data from parasitic nematodes (5). Nematode.net has proven to be a valuable community resource for human/mammalian parasitologists and plant nematologists since its inception in early 2000, hosting research on over 500 000 nematode expressed sequence tags (ESTs) and 1.2 million genome survey sequences (GSSs) spanning 32 species (27 parasitic and five free-living). Over the 12 months ending 12 December 2007, more than 44 000 unique users accessed

*To whom correspondence should be addressed. Tel: +1 314 286 1118; Fax: +1 314 286 1810; Email: mmitreva@watson.wustl.edu

Nematode.net with over 4900 returning on multiple days. This is comparable to other, similar database websites such as PlasmoDB which garnered ~27 000 unique visitors over the same period of time (http://plasmodb.org). Nematode.net also mediates public requests for cDNA clones generated by the Parasitic Nematode Genomics Project. Since 2003 the project has shipped 377 individual clones and 58 full plates (384 clones each) to 37 investigators at 31 different institutions around the globe. This paper will focus on improvements and additions made to Nematode.net over the last few years as well as those planned for the future.

RECENT IMPROVEMENTS

Since last reported in 2004 Nematode.net has undergone a number of changes that include the addition of several new analytical tools and various improvements in site design. Additionally, the influx of new data over the last 4 years has provided significantly broader coverage of the phylum nematoda than was previously available (Table 1).

NemaGene

The NemaGene cluster collection was previously displayed only through static table dumps directly from a MySQL database which displayed only member contig consensus sequences, names of the EST members of those contigs and a list of NR homologies. Now that information is further supplemented with a new GBrowse (6) 'clusterhub' view displaying library, gender and developmental stage information for member reads/contigs, as well as displaying graphically how member reads and contigs align under the footprint of the entire cluster. Additionally, the GBrowse glyphs representing each element are linked out to external resources (i.e. ESTs link to GenBank entries, member contigs link back to their top-level NemaGene cluster pages, etc). Furthermore, NemaGene cluster entries are now integrated with *Caenorhabditis elegans* gene pages [WormBase; (7)] in cases where a relationship is believed to exist. Bidirectional links between Nematode.net and WormBase allow users to easily navigate between related entries on both sites. Nematode.net links to *C. elegans* gene pages are displayed at the bottom of each NemaGene cluster page, and conversely WormBase gene pages are linked to NemaGene pages via the 'Nematode ESTs (non-*Caenorhabditis*)' track in each gene's GBrowse view.

NemaBlast

Our WU-BLAST service (8), NemaBlast, has been migrated to a powerful compute cluster providing better support for multiple, concurrent users. Two previous blast functions have been merged into a single NemaBlast application, which clearly directs users to the appropriate function. Users are able to blast their own sequence against virtual combinations of sequences grouped by library, or EST contigs and genes grouped by species and phylogenetic clade [phylogeny based on (9)].

Table 1. Growth of data hosted by Nematode.net over the last 4 years

	2004	2008
EST sequencing totals	215 127	509 161
GSS sequencing totals	0	1 208 119
Gene Ontology classifications	7 species	29 species
Codon usage tables	2 species	30 species
NemaGene clusters	12 158	118 770
GSS derived genes	0	32 952

NemaPath

A significant addition to Nematode.net is the introduction of our NemaPath metabolic pathway viewer. Our NemaPath server is a web-based, pathway-level visualization tool for illustrating mappings of EST contigs to Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic pathways (10). Currently 33 species of nematodes (30 parasites and three free-living) have been mapped in NemaPath (Figure 1). The NemaPath pipeline consists of two parts: (i) KEGGscan—a backend tool to align and evaluate nematode transcriptomic sequence (EST contigs) against the annotated KEGG Genes database; and (ii) NemaPath—a web viewing application that paints mapped sequence onto the KEGG metabolic pathway image maps based on user-defined confidence levels of primary sequence similarity. In summary, KEGGscan runs WU-BLAST to align queries (contigs of assembled ESTs or full-length genes) against a version of KEGG Genes database [modified to include only entries having Enzyme Commision ids (11)]. These alignments along with their e-value and bit score are deposited into the NemaPath database. Upon loading the database builds an index of hits to each EC based on a user-defined e-value threshold. Once the index is built, NemaPath highlights the corresponding EC nodes in the KEGG pathway maps in various shades of green (dark shades indicate more hits to that EC node at the given e-value cut-off). This approach is similar to the KAAS genome annotation server (12), but differs in that we provide a broader association of input data to every possible enzymatic function whereas KAAS confidently assigns input data to a KO group, which is subsequently linked to specific gene products in the KEGG pathway diagrams via manual curation.

NemaPath also supports comparative views showcasing the differences between the mappings of any two species represented in the NemaPath database (Figure 1B). At any time the user can select a second species to layer onto a previously drawn map, and once chosen, the pathway will be redrawn displaying nodes for both species in distinctive colors. This useful feature allows visitors to distinguish enzymatic reactions common between sampled species and reactions that are not shared (caveat: an under-representation of data for one or both species greatly impacts the confidence in any conclusions drawn from this comparative view). In addition, Nemantode.net provides a phylogeny-based view (Figure 1C) showing, for a chosen pathway, which nematode taxa have been putatively identified to have gene products encoding

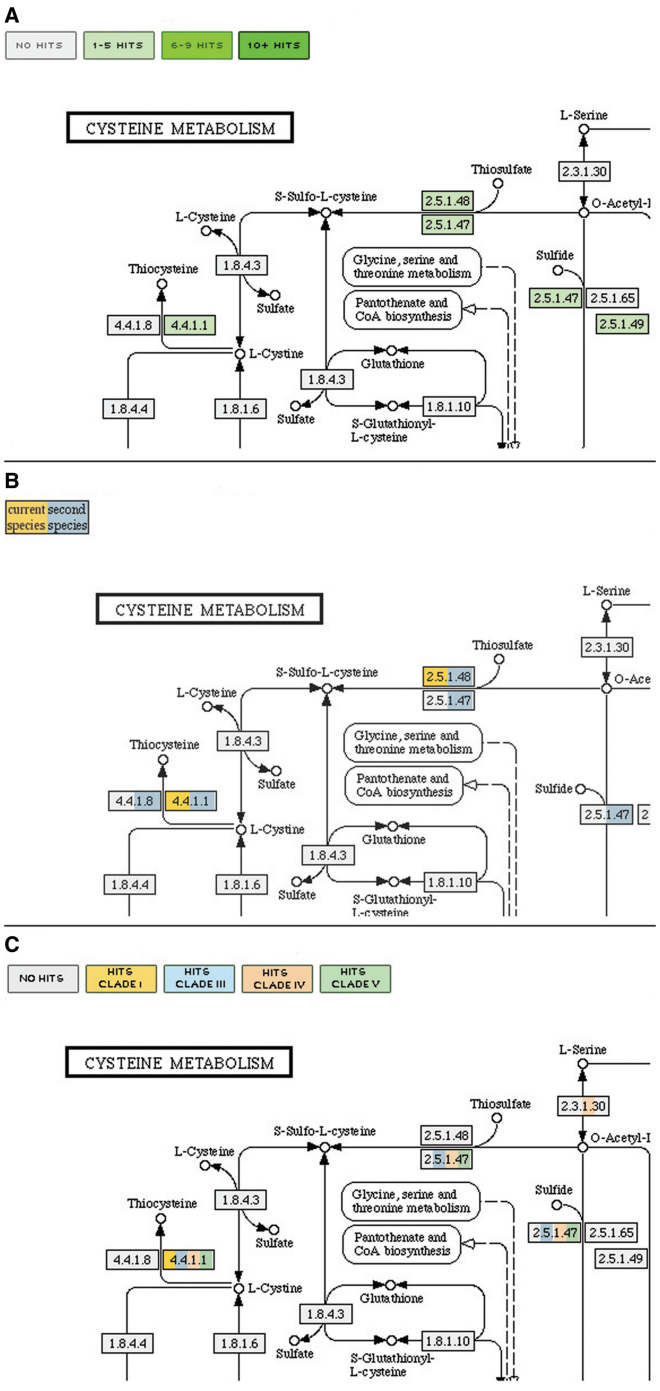

Figure 1. NemaPath pathway view. (**A**) In this case we are looking at a section of the Cysteine metabolism pathway. Green shaded boxes show gene products that have been putatively identified in the current organism by primary sequence homology to a member of KEGG's genes database with that same EC assignment. (**B**) Comparative view highlighting populated gene products for two user-selected species, (yellow and blue) on the same section of this pathway map. (**C**) Phylogeny-based comparison highlighting clades that have gene products putatively encoding the specific ECs. The phylogeny is based on reference 9.

specific enzymes. This clade-based comparison uses the nematode phylogeny based on 18S rRNA (9). Furthermore, the server also provides another comparative view showing EC node population by nematodes' food source and mode of parasitism, showing which EC nodes have been discovered in plant parasites, animal parasites and free-living species. Finally, when viewing *C. elegans* mappings, nodes have been outlined for genes with RNA interference (RNAi) phenotype information mapped to them. This feature will scale as RNAi information becomes available for other organisms.

Finally, we have supplemented access to NemaPath pathways by providing a 'treeview' entry portal for NemaPath processed organisms. The treeview allows a user to drill down to specific pathways for a given organism and be shown all populated EC ids. The user can click on the 'NemaPath viewer' link and be taken directly to the selected pathway with their chosen EC number highlighted in red. This allows quick access to a populated EC number in a pathway of interest.

PHYLUM-RESTRICTED MOLECULAR FEATURES

Several parasitic nematode draft genomes are complete (13) or underway (14). The ongoing draft genomes are expected to become available within the next 3–5 years, therefore ESTs and GSSs will continue to be the main data resource available for most parasitic nematode species for the next few years. However, EST and GSS data are not represented in protein domain databases as a result of quality-control measures in place that restrict sequences they incorporate only to include those found in Swiss-Prot (15). As Swiss-Prot lacks putative open reading frames from ESTs and GSSs, the vast majority of parasitic nematode sequences are missing. Therefore, some of our effort has been focused on providing tools and disseminating information on this pool of data that is underrepresented in public protein databases, as well as identifying molecular features that are of interest in the development of new control strategies (as described below).

NemFam

Proteins in nematode transcriptomes have experienced drastic changes (16,17) related to functional diversification, speciation and species adaptation (18–22). Among them are nematode-specific proteins that bear crucial importance for understanding nematode evolution and parasitism (23–25). Proteins that are specific to the pathogen or have sufficiently diverged in the host as to be functionally absent or altered can be good targets for drugs with low toxicity to the host or for environmentally safe pesticides (26). Hence, another addition to Nematode.net is the inclusion of the NemFam collection of nematode-related, conserved regions of proteins that are not well represented in Pfam (27). NemFam was constructed using over 214 000 polypeptides from 32 nematode species (27 of which are parasites of vertebrates or plants). Each polypeptide represents a full-length protein or a putative translation (28) of a single EST contig generated by assembling identical or nearly identical overlapping sequences using a per-species 'clustering' process (29,30) on all available ESTs for each organism. Subsequently, the full set of 214 000 polypeptides were grouped into families using a

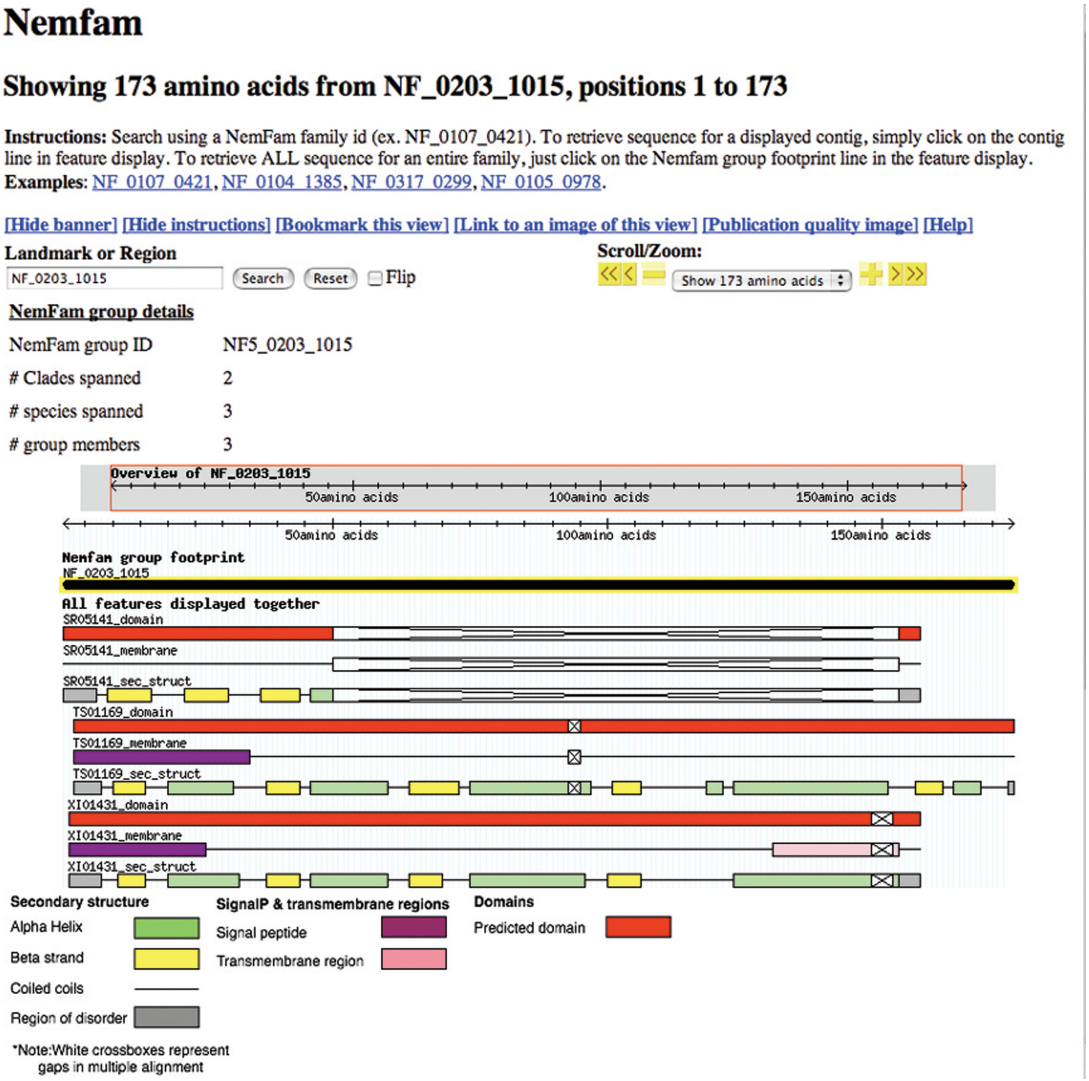

Figure 2. NemFam GBrowse screenshot. In this example, details for the family NF_203_1015 are displayed. The chosen track is displaying all available features compressed together (feature specific tracks are also available). Also, some basic information about the family is printed near the top of the page.

clustering method that relies on the Markov Cluster Algorithm (31). Similar pan-phylum clustering of nematode-originated proteins was recently published (32), and presents an additional valuable resource available to the community through an updated version of the NemBase database (33). Our families were multiply aligned, and families having at least 10% of their full alignment length shared by three or more distinct nematode species became candidate members of NemFam. These candidates were then screened against Pfam (release 20.0; 8296 families), and those groups with no members hitting Pfam were assembled into the final NemFam family set.

The final sets of NemFam families were subjected to various structural analysis tools to further characterize them. Each member of each family was run through PHOBIUS (34) to find transmembrane regions and signal peptides, SSEP-domain (35) to identify putative domain structure, PSIPRED (36) to predict secondary structure, and DISOPRED (37) to locate regions of disorder. All the features of each member were then localized back to the families' footprint and displayed via a GBrowse server (Figure 2).

Each NemFam family page displays the number of clades and species spanned by its members, and graphically shows the alignment of each member to the footprint of the whole family. Individual tracks exist for domain predictions, signal peptide and transmembrane predictions and secondary structure predictions. Selecting the family footprint displays the amino acid sequence for all members of that family, and clicking on any of the family members takes the user to the sequence for that individual member (which is itself linked to the NemaGene 'clusterhub' entry for that member). NemFam is rebuilt as new sequence data become available, followed by functional and structural annotation for new family members.

GENE EXPRESSION IN PARASITIC NEMATODES

EST-based expression data

For most parasitic nematode genes little is known about the distribution of expression by tissue or organ. The relatively large size of some nematode species allows for the dissection of specific organs. The intestine is a focus because it is a major surface for interaction with the host and is also a validated target site for parasite control. We have produced the largest collection of nematode intestinal genes currently available from the clade V blood-feeding parasite *Haemonchus contortus* and the clade III nematode *Ascaris suum*, which is thought to feed on the semi-digested contents of its host intestine. Using tools implemented within Nematode.net, comparative studies were made among intestinal genes from the two parasites and the free-living bacterivore *C. elegans*. From this study, a number of proteins were identified that may represent core nematode intestinal functions among the lineages investigated (38). Intestinal genes from each studied species, and a set of core intestinal protein families, were mapped to metabolic pathways via the NemaPath pipeline, and functional classifications were displayed via AmiGO. This section presents a research model of Nematode.net as host for the display of specialized analysis results and ancillary information.

Microarray-based expression data

Based on the large amount of sequence data that has become available for the first time in many nematodes, we and other labs have started to investigate gene expression profiles in various species using microarray technology. For example, the 20 109 ESTs that we generated from the soybean cyst nematode (*Heterodera glycines*) together with the 1900 previously published sequences formed 6860 contigs. These contigs are represented by 7530 probesets on the Affymetrix Soybean Genome Array GeneChip (Affymetrix, Santa Clara, CA). Using this GeneChip, we determined that the gene expression profiles associated with *C. elegans* dauer and *H. glycines* J2i are not well conserved. The data showed striking differences in the underlying biochemistry and physiology of developmentally arrested *C. elegans* dauers and *H. glycines* J2i (39). The transcriptomic data that we have generated from the filarial pathogen *Brugia malayi* was used in the design of the Filarial Expression Array v1 and v2, studies that have led to better understanding of parasitism and provided novel insights for the development of next generation control strategies. Several other expression experiments are also underway. Resources from microarray experiments such as those described above are hosted in the Microarray section of Nematode.net.

FROM PARTIAL TO COMPLETE PARASITIC NEMATODE GENOMES

A number of draft genomes have recently become available for parasitic nematodes. *Brugia malayi* (13), *Ancylostoma caninum* (40) and *A. suum* (Mitreva *et al.*, unpublished data) draft assemblies are now available, with *Trichinella spiralis* coming in the months ahead, and *Trichuris suis* and others at various early stages of production (http://www.genome.gov/10002154; http://www.sanger.ac.uk/Projects/Helminths/). As a pilot for our plans to host this type of information, we have implemented a Gbrowse-based view of *T. suis* genes mined from an early assembly of that organism.

Using the MAKER suite of annotation software (41) we built an initial set of gene calls. Then, regions of the assembly upon which MAKER made no predictions were subject to blastx against a collection of nematode proteins. Regions with strong alignments were sent to GeneWise (42) for further analysis. This final set of MAKER plus GeneWise gene predictions were used to annotate the *T. suis* early assembly, and these annotations are now available for public viewing through a GBrowse instance available within the NemaBrowse section of Nematode.net (Figure 3). Similar GBrowse instances will be made available for all upcoming annotated parasitic nematode genomes.

PLANS FOR THE FUTURE

Our ongoing goal is to provide the nematode research community with useful, consistent and lasting integrated databases and the tools required to explore them. With over 44 000 unique users in the past year, Nematode.net has proven itself useful to the scientific community. Moving forward, we have a number of planned improvements being implemented.

The primary objectives of these improvements are: (i) build the infrastructure necessary to support the influx of new data and analyses spawned by new technologies (e.g. pyrosequencing); (ii) improve integration within the site allowing users to make meaningful connections between related data; and (iii) improve the site interface to allow easy and effective navigation. The planned flow of data from origination, through various analyses and to its final display is illustrated in Figure 4.

We plan three conceptual points of entry into our data: (a) an organism-centric portal allows the user to explore information we have collected for a single species. From descriptive summaries of the species itself, to genome statistics, to complex metabolic reconstructions; (b) an analysis-centric portal provides the user with comparative views of the results of a single analysis across two or more species. It will provide for two species comparisons, pan-phylum comparisons and phylogenetically restricted comparisons. We will make every effort to create visualizations that clearly compare and contrast results across whatever species (or collections of species) are being examined; and (c) a data mining portal will allow users to access data starting with only a gene name or other piece of information. Advanced search features will allow the user to specify combinations of attributes, and return linked lists of other views that are appropriate to the user's request. For example, the user may enter a gene name and be presented with a list of all metabolic pathways in which that gene plays a role. Or the user may pick a metabolic pathway, and a species name, and be

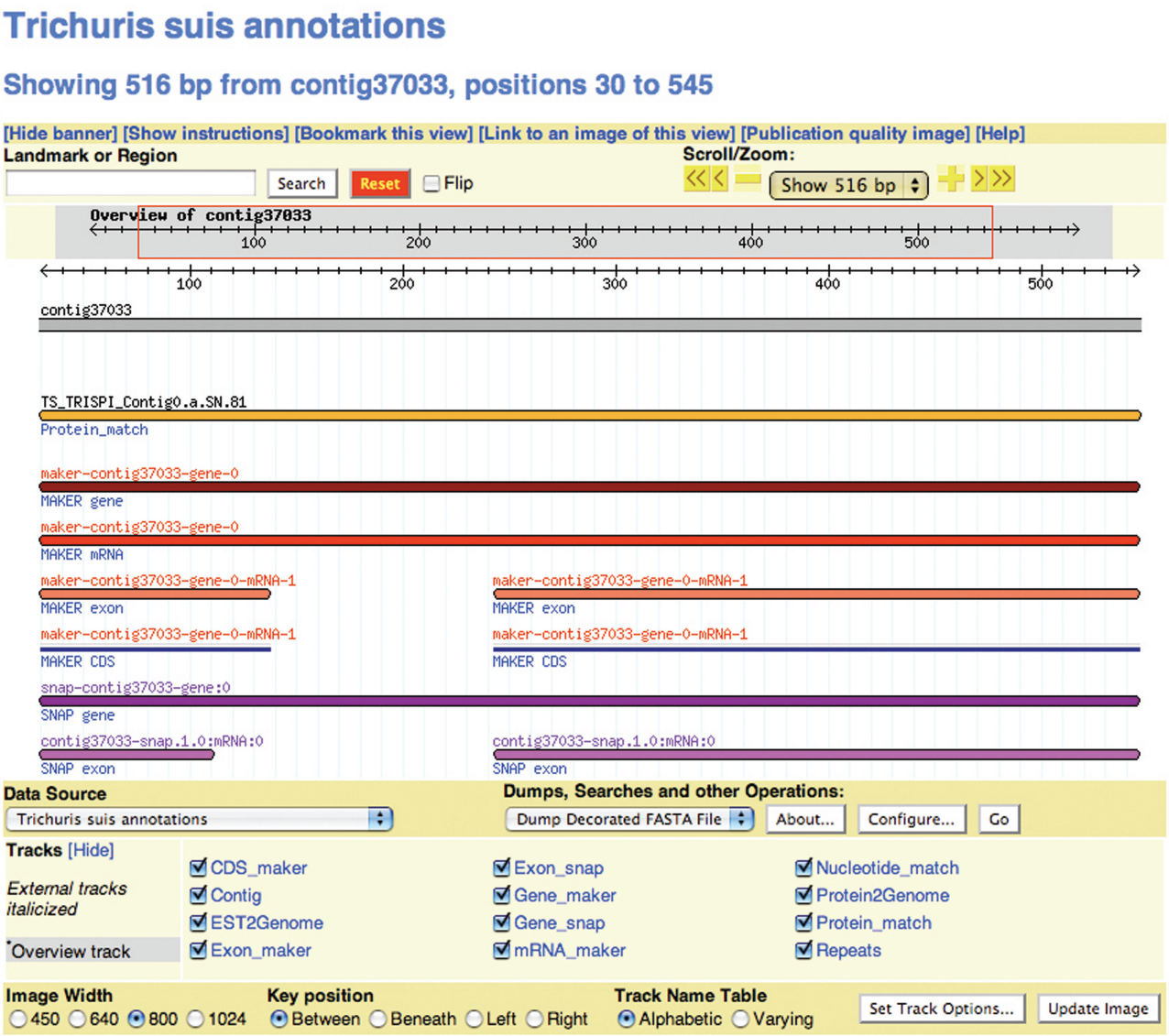

Figure 3. This screenshot displays a typical annotation of a parasitic nematode genome via GBrowse. In this case we are looking at the annotation of a *Trichuris suis* contig made by the MAKER software application.

presented with a list of all proteins, genes, clusters, contigs and reads from that species that have been mapped to that pathway.

Most resources will be made available via FTP. We will host raw data, analysis summaries and results and various software tools for community use on our FTP site. The results of analytical visualizations served elsewhere on the site will be made available as tab-delimited text files whenever possible. Heavily trafficked analyses will be precomputed and placed on the FTP site at regular intervals.

More static analyses will be recomputed on a quarterly basis as supporting resources themselves are updated. (i.e. once per quarter we will update to new versions of KEGG, GenBank's NR, GO term definitions and will recompute data as necessary to stay current).

The success of Nematode.net is due, in large part, to extensive interaction with the parasitic nematode research community including the annual User's Survey and an e-mail-based Help Desk. To increase the interactions both on the data use level and data input level, we will create an extensive online User's Guide, an evolving Wiki site, where protocols, analytical pipelines and other information can be freely shared, and a mailing list to provide news and notices to interested users.

ACKNOWLEDGEMENTS

Sequence generation has been aided by numerous collaborators in the nematology community, cDNA and WGS library creation by Irina Ronko and Michael Becker, and the dedicated members of the sequence production group at the GC. We would like to thank, James P. McCarter who initiated and supervised the first version of

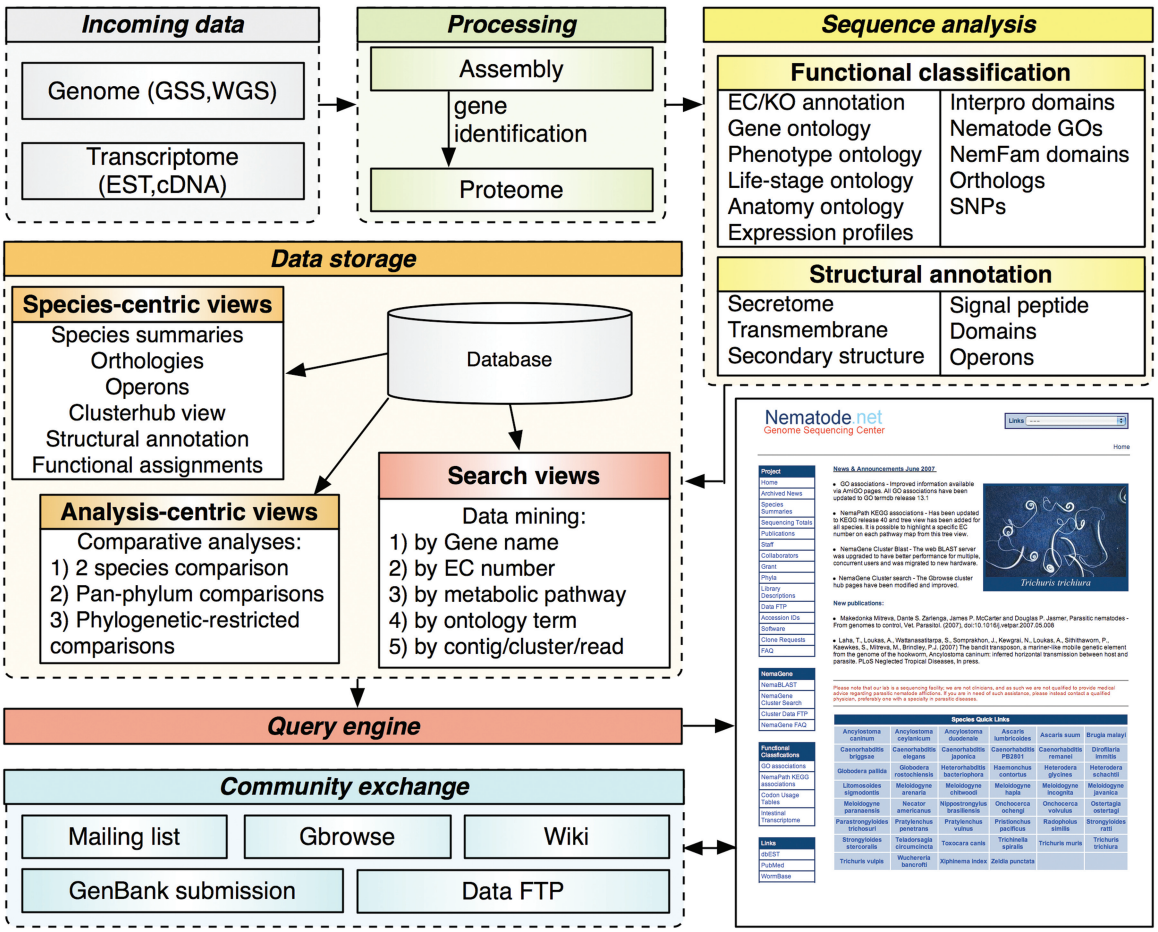

Figure 4. This flowchart depicts the movement of data from origination, through analyses, to its final display in the planned version of Nematode.net. Data generation begins with either genomic (GSS, WGS) or transcriptomic (EST, cDNA) sequence data that are assembled, and then run through a robust process of gene identification. Once the proteome is determined, various structural annotations and functional classifications are computed, and an exhaustive enzymatic pathway reconstruction will be performed with the initial focus on metabolic pathways, but eventually to include other pathways. This information is then spooled into a relational database whose schema supports multiple analysis and species centric views. As much as possible, views that are expected to be in high demand will be precomputed to keep user requests for information responsive and fast. On top of this layer, a robust query engine will support data mining on a large number of attributes, including gene name, EC number, metabolic pathway, ontology term and/or by cluster/contig/read name. Advanced queries also allow data mining based on valid combinations of the above-mentioned attributes.

Nematode.net and offered advice throughout and John Spieth, who heads the Wormbase efforts at the GC.

FUNDING

The US National Institute for Allergy and Infectious Disease (grant number AI46593 to M.M). Funding for open access charges: US National Institute for Allergy and Infectious Disease (grant U01-AI46593).

Conflict of interest statement. None declared.

REFERENCES

1. WHO (2005) Deworming for health and development. Report of the third global meeting of the partners for parasite control. World Health Organization.
2. Bethony,J., Brooker,S., Albonico,M., Geiger,S.M., Loukas,A., Diemert,D. and Hotez,P.J. (2006) Soil-transmitted helminth infections: ascariasis, trichuriasis, and hookworm. *Lancet*, **367**, 1521–1532.
3. Chan,M.S. (1997) The global burden of intestinal nematode infections – fifty years on. *Parasitol. Today*, **13**, 438–443.
4. Jasmer,D.P., Groverse,A. and Smant,G. (2003) Parasitic nematode interactions with mammals and plants. *Annu. Rev. Phytopathol.*, **41**, 245–270.
5. Wylie,T., Martin,J., Dante,M., Mitreva,M., Clifton,S.W., Chinwalla,A., Waterston,R.H., Wilson,R.K. and McCarter,J.P. (2004) Nematode.net: a tool for navigating sequences from parasitic and free-living nematodes. *Nucleic Acids Res.*, **32**, D423–D426.
6. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
7. Rogers,A., Antoshechkin,I., Bieri,T., Blasiar,D., Bastiani,C., Canaran,P., Chan,J., Chen,W.J., Davis,P., Fernandes,J. *et al.* (2008) WormBase 2007. *Nucleic Acids Res.*, **36**, D612–D617.
8. Gish,W. (1996–2002) http://blast.wustl.edu.
9. Blaxter,M.L., De Ley,P., Garey,J.R., Liu,L.X., Scheldeman,P., Vierstraete,A., Vanfleteren,J.R., Mackey,L.Y., Dorris,M., Frisse,L.M. *et al.* (1998) A molecular evolutionary framework for the phylum Nematoda. *Nature*, **392**, 71–75.
10. Kanehisa,M. and Goto,S. (2006) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

11. IUBMB (1992) *Enzyme Nomenclature: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*. Academic Press, San Diego.
12. Moriya,Y., Itoh,M., Okuda,S., Yoshizawa,A.C. and Kanehisa,M. (2007) KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.*, **35**, W182–W185.
13. Ghedin,E., Wang,S., Spiro,D., Caler,E., Zhao,Q., Crabtree,J., Allen,J.E., Delcher,A.L., Guiliano,D.B., Miranda-Saavedra,D. *et al.* (2007) Draft genome of the filarial nematode parasite. *Brugia malayi*. *Science*, **317**, 1756–1760.
14. Mitreva,M., Zarlenga,D.S., McCarter,J.P. and Jasmer,D.P. (2007) Parasitic nematodes–from genomes to control. *Vet. Parasitol.*, **148**, 31–42.
15. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
16. Stein,L.D., Bao,Z., Blasiar,D., Blumenthal,T., Brent,M.R., Chen,N., Chinwalla,A., Clarke,L., Clee,C., Coghlan,A. *et al.* (2003) The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol.*, **1**, E45.
17. Parkinson,J., Mitreva,M., Whitton,C., Thomson,M., Daub,J., Martin,J., Hall,N., Barrell,B., Waterston,R.H., McCarter,J.P. *et al.* (2004) A transcriptomic analysis of the phylum Nematoda. *Nat. Genet.*, **36**, 1259–1267.
18. Panhuis,T.M., Clark,N.L. and Swanson,W.J. (2006) Rapid evolution of reproductive proteins in abalone and Drosophila. *Philos. Trans. R Soc. Lond. B Biol. Sci.*, **361**, 261–268.
19. Peng,J. and Huang,C.H. (2006) Rh proteins vs Amt proteins: an organismal and phylogenetic perspective on CO2 and NH3 gas channels. *Transfus. Clin. Biol.*, **13**, 85–94.
20. Givnish,T.J., Evans,T.M., Zjhra,M.L., Patterson,T.B., Berry,P.E. and Sytsma,K.J. (2000) Molecular evolution, adaptive radiation, and geographic diversification in the amphiatlantic family Rapateaceae: evidence from ndhF sequences and morphology. *Int. J. Org. Evol.*, **54**, 1915–1937.
21. Kocher,T.D. (2004) Adaptive evolution and explosive speciation: the cichlid fish model. *Nat. Rev. Genet.*, **5**, 288–298.
22. Jang,C.S., Jung,J.H., Yim,W.C., Lee,B.M., Seo,Y.W. and Kim,W. (2007) Divergence of genes encoding non-specific lipid transfer proteins in the poaceae family. *Mol. Cells*, **24**, 215–223.
23. Lilley,C.J., Urwin,P.E. and Atkinson,H.J. (1999) Characterization of plant nematode genes: identifying targets for a transgenic defence. *Parasitology*, **118**, S63–S72.
24. Davis,E.L., Hussey,R.S. and Baum,T.J. (2004) Getting to the roots of parasitism by nematodes. *Trends Parasitol.*, **20**, 134–141.
25. Curtis,R.H.C. (2007) Plant parasitic nematode proteins and the host parasite interaction. *Brief. Funct. Genomic Proteomic.*, **6**, 50–58.
26. McCarter,J.P. (2004) Genomic filtering: an approach to discovering novel antiparasitics. *Trends Parasitol.*, **20**, 462–468.
27. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.-R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L.L. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.
28. Wasmuth,J.D. and Blaxter,M.L. (2004) prot4EST: translating expressed sequence tags from neglected genomes. *BMC Bioinformatics*, **5**, 187.
29. McCarter,J., Dautova Mitreva,M., Martin,J., Dante,M., Wylie,T., Rao,U., Pape,D., Bowers,Y., Theising,B., Murphy,C.V. *et al.* (2003) Analysis and Functional Classification of Transcripts from the Nematode *Meloidogyne incognita*. *Genome Biol.*, **4**, R26.
30. Mitreva,M., McCarter,J.P., Martin,J., Dante,M., Wylie,T., Chiapelli,B., Pape,D., Clifton,S.W., Nutman,T.B. and Waterston,R.H. (2004) Comparative genomics of gene expression in the parasitic and free-living nematodes *Strongyloides stercoralis* and *Caenorhabditis elegans*. *Genome Res.*, **14**, 209–220.
31. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
32. Wasmuth,J., Schmid,R., Hedley,A. and Blaxter,M. (2008) On the extent and origins of genic novelty in the phylum nematoda. *PloS Negl. Trop. Dis.*, **2**, e258.
33. Parkinson,J., Whitton,C., Schmid,R., Thomson,M. and Blaxter,M. (2004) NEMBASE: a resource for parasitic nematode ESTs. *Nucleic Acids Res.*, **32**, D427–D430.
34. Kall,L., Krogh,A. and Sonnhammer,E.L.L. (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, **338**, 1027–1036.
35. Gewehr,J.E. and Zimmer,R. (2006) SSEP-Domain: protein domain prediction by alignment of secondary structure elements and profiles. *Bioinformatics*, **22**, 181–187.
36. McGuffin,L.J., Bryson,K. and Jones,D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.
37. Ward,J.J., McGuffin,L.J., Bryson,K., Buxton,B.F. and Jones,D.T. (2004) The DISOPRED server for the prediction of protein disorder. *Bioinformatics*, **20**, 2138–2139.
38. Yin,Y., Martin,J., Abubucker,S., Scott,A.L., McCarter,J.P., Wilson,R.K., Jasmer,D.P. and Mitreva,M. (2008) Intestinal transcriptomes of nematodes: comparison of the parasites *Ascaris suum* and *Haemonchus contortus* with the free-living *Caenorhabditis elegans*. *PLoS Negl. Trop. Dis.*, **2**, e269.
39. Elling,A., Mitreva,M., Recknor,J., Gai,X., Martin,J., Maier,T., McDermott,J., Hewezi,T., McK Bird,D., Davis,E. *et al.* (2007) Divergent evolution of arrested development in the dauer stage of *Caenorhabditis elegans* and the infective stage of *Heterodera glycines*. *Genome Biol.*, **8**, R211.
40. Abubucker,S., Martin,J., Yin,Y., Fulton,L., Yang,S.-P., Hallsworth-Pepin,K., Johnston,J.S., Hawdon,J., McCarter,J.P., Wilson,R.K. *et al.* (2008) The canine hookworm genome: Analysis and classification of *Ancylostoma caninum* survey sequences. *Mol. Biochem. Parasitol.*, **157**, 187–192.
41. Cantarel,B.L., Korf,I., Robb,S.M.C., Parra,G., Ross,E., Moore,B., Holt,C., Sanchez Alvarado,A. and Yandell,M. (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.*, **18**, 188–196.
42. Birney,E., Clamp,M. and Durbin,R. (2004) GeneWise and Genomewise. *Genome Res.*, **14**, 988–995.