Published online 12 November 2009 Nucleic Acids Research, 2010, Vol. 38, Database issue D463–D467
doi:10.1093/nar/gkp952

WormBase: a comprehensive resource for nematode research

Todd W. Harris[1,*], Igor Antoshechkin[2], Tamberlyn Bieri[3], Darin Blasiar[3], Juancarlos Chan[2], Wen J. Chen[2], Norie De La Cruz[1], Paul Davis[4], Margaret Duesbury[4], Ruihua Fang[2], Jolene Fernandes[2], Michael Han[4], Ranjana Kishore[2], Raymond Lee[2], Hans-Michael Müller[2], Cecilia Nakamura[2], Philip Ozersky[3], Andrei Petcherski[2], Arun Rangarajan[2], Anthony Rogers[4], Gary Schindelman[2], Erich M. Schwarz[2], Mary Ann Tuli[2], Kimberly Van Auken[2], Daniel Wang[2], Xiaodong Wang[2], Gary Williams[4], Karen Yook[2], Richard Durbin[4], Lincoln D. Stein[1], John Spieth[3] and Paul W. Sternberg[2,5]

[1]Ontario Institute For Cancer Research, Toronto, ON, Canada M5G0A3, [2]California Institute of Technology, Division of Biology 156-29, Pasadena, CA 91125, [3]Genome Sequencing Center, Washington University School of Medicine, St Louis, MO 63108, USA, [4]Sanger Institute, Wellcome Trust Genome Campus Hinxton, Cambridgeshire CB10 1SA, UK and [5]Howard Hughes Medical Institute, California Institute of Technology Pasadena, CA 91125, USA

Received September 15, 2009; Revised October 11, 2009; Accepted October 12, 2009

ABSTRACT

WormBase (http://www.wormbase.org) is a central data repository for nematode biology. Initially created as a service to the *Caenorhabditis elegans* research field, WormBase has evolved into a powerful research tool in its own right. In the past 2 years, we expanded WormBase to include the complete genomic sequence, gene predictions and orthology assignments from a range of related nematodes. This comparative data enrich the *C. elegans* data with improved gene predictions and a better understanding of gene function. In turn, they bring the wealth of experimental knowledge of *C. elegans* to other systems of medical and agricultural importance. Here, we describe new species and data types now available at WormBase. In addition, we detail enhancements to our curatorial pipeline and website infrastructure to accommodate new genomes and an extensive user base.

DESCRIPTION

Caenorhabditis elegans is a free-living soil nematode, well-established as a genetic model system due to its small size (1 mm in length), rapid generation time (3.5 days), compact genome (100 MB, ~20 000 protein-coding genes), invariant cell lineage and described neural connectivity (1; WormBook: http://www.wormbook.org/). Initially used for studies of development and neurobiology, *C. elegans* is now widely used to address a broad range of biological questions from perspectives ranging from subcellular to systems and single gene to whole genome. WormBase aims to capture the full spectrum of this information, placing it in an intellectually rich and interactive context to facilitate new insights and hypotheses.

AVAILABLE SPECIES

WormBase now includes information from a total of nine species: five from the *Caenorhabditis* genus [*C. elegans*, *C. briggsae*, *C. remanei*, *C. brenneri* and *C. japonica*; (2,3)]; the filarial agent *Brugia malayi* (4); the plant parasites *Meloidogyne incognita* (5) and *Meloidogyne hapla* (6); and the diplogastrid *Pristionchus pacificus* (7).

We provide a consistent and familiar user interface for each species at WormBase. First, we provide a Genome Browser (e.g. http://www.wormbase.org/db/seq/gbrowse/c_elegans) using the open source GBrowse software (8). Second, genome, gene and protein sets (when available) are accessible for sequence similarity searches using BLAST and BLAT (http://www.wormbase.org/db/searches/blast_blat). Finally, Gene and Protein Summaries give additional details for all known genes. When available, additional annotations—such as ortholog and paralog assignments, phenotypic

*To whom correspondence should be addressed. Tel: +1 406 222 2450; Email: info@toddharris.net

descriptions and controlled ontology terms—are presented on these pages.

Comparative studies are facilitated by a wide range of precalculated data. We provide users with ortholog assignments derived from the Ensembl Compara pipeline (9), homology groups from InParanoid (10) and inline tree displays from TreeFam (11). With these data, users can effectively step from one species to another, and quickly compare gene model structures and protein primary sequences and tertiary structures when known. Further, multispecies alignments are shown on a separate Synteny Browser (http://www.wormbase.org/cgi-bin/gbrowse_syn; for details see http://gmod.org/wiki/GBrowse_syn).

ENHANCEMENTS TO THE CURATION PROCESS

WormBase relies on a thorough manual curation pipeline to extract data from the corpus of *C. elegans* literature. This process ensures high quality and consistency of data and provides an atomized evidence trail for all data entered into the database, but it is time consuming and labor intensive.

We continue to refine our curation strategy for greater efficiency, breadth and depth of coverage. Notably, we have automated nearly 90% of 'first-pass' curation, which aims to flag and extract data types contained in each reference. We implemented these improvements using a two-tiered approach that combines programmatic natural language processing with web-based data submission forms targeted to authors. This process now identifies 27 distinct data types (for details, see http://www.wormbase.org/wiki/index.php/Curated_data_types). Five of these data types (alleles, small-scale RNAi experiments, transgenes, gene interactions and antibodies) are identified automatically through use of the text mining system Textpresso (12). In addition, we employ Textpresso for fact extraction, notably for curation of Gene Ontology (GO) cellular component terms (13).

The identification and flagging of two data types (small and large-scale RNAi experiments, and phenotype analysis) have been automated by Support Vector Machines (SVMs; 14). We are testing SVMs on all curated data types and will adopt this method to classify and index papers for those data types that can be efficiently and reproducibly identified. For those data types that are not amenable to SVMs, we are exploring other statistical methods (such as hidden Markov models and conditional random fields) or rule-based methods. For example, we now use rule-based methods to identify papers that discuss *C. elegans* homologs of genes associated with human disease.

A second refinement has been to integrate the research community directly into curation. Using our database of public biographical information, we automatically e-mail authors of new *C. elegans* papers, and ask them to use a concise and time-efficient web form to identify which data types are relevant to their papers. So far, 413 out of 864 authors have responded, giving us a very large pool of expert first-pass curators and greatly speeding curation. Community input will further support our switch to an automated first-pass pipeline because it will help us assess the success and failure rates of automation.

DATA CURATION UPDATE

The most prominent curated data at WormBase are discussed below and new data types added in the last 2 years are summarized. See http://www.wormbase.org/wiki/index.php/Curated_data_types for a complete list of curated data types.

Concise descriptions

WormBase now provides concise summaries for over 5597 genes (~25% of the protein-coding genes in *C. elegans*). These prose descriptions provide a quick overview of gene function and are particularly useful for researchers not familiar with *C. elegans* as an experimental system.

Genes and sequence changes

WormBase relies on a curation anomaly pipeline to indicate regions that require attention. This has resulted in 2094 *C. elegans*, 150 *C. briggsae* and 661 *C. brenneri* gene structure changes in the last year. These changes include 127 new *C. elegans* genes and 443 new *C. elegans* isoforms. Most of these changes have been based either on RNA-Seq transcript data (from 454 or Illumina sequencing) or on comparative gene analysis. There have been 11 changes to the *C. elegans* canonical genomic sequence made after re-inspection of the original sequence traces. Ten of these changes were 1 bp in size; the largest was an insertion of 1469 bp.

Alleles

WormBase now contains over 10 000 alleles with molecular information, greatly facilitating structure–function studies. Sequencing efforts in polymorphic strains have now identified over 170 000 single nucleotide polymorphisms.

Gene Ontology

WormBase is an active participant of the GO consortium (http://www.geneontology.org). We continue to expand the breadth and depth of gene ontology annotation at WormBase via the association of biological process, cellular component and molecular function controlled vocabulary terms to gene products. Associations are generated by several methods, including manual annotation by curators reading the primary literature, sequence similarity (e.g. mapping of InterPro motifs to GO terms provided by the InterPro2GO project) and mapping of terms from other ontologies to the GO (e.g. mapping of phenotype terms to GO terms). Currently 1800 unique GO terms have been associated to over 14 000 genes; the GO terms for ~2000 of these genes have been manually curated.

Phenotype Ontology

In addition to annotating species other than *C. elegans*, WormBase also contains phenotypic and ethological data from strains of *C. elegans* that are not Bristol N2.

These strains must be considered as alternative 'wild type' rather than mutants. For instance, the wild strains RC301 and CB4932 exhibit clumping behavior that is missing in N2, but which is common in freshly isolated *C. elegans* (15). This has driven us to revise our phenotype ontology so that it is no longer N2-centric, allowing us to annotate normal and mutant phenotypes of both new species and non-N2 elegans strains. We have accordingly replaced the word 'abnormal' in term names and definitions and its replacement with the less-biased term 'variant'. We have also been actively expanding the phenotype ontology to reflect the richness of described phenotypes. Since the WS190 release we have added 168 new terms, bringing the total number of terms to 1845. Eighty-eight percent of these terms are now defined in detail using uniform language, compared with 42% coverage in WS190.

This expansion of the available number of terms has enabled a more complete description of phenotypes for alleles, transgenes and strains. We increased the number of alleles with associated phenotype ontology terms almost 40% since the WS190 release. Currently, 6482 alleles carry phenotype associations; 17 904 allele–phenotype associations in total have been generated representing a major advance in the controlled description of phenotypes in *C. elegans*. We have added 122 new strain–phenotype associations to 36 strains, and 232 transgene–phenotype associations to 101 transgenes.

The improved precision of the phenotype ontology has prompted both GO and phenotype curators to generate accurate mappings of Phenotype Ontology to GO terms using entity–quality (EQ) relationships. This pipeline is envisioned to generate GO assignments to genes based on their allele–phenotype assignments.

Anatomy Ontology and function

WormBase has expanded the controlled anatomy ontology terms and relations among them. This vocabulary includes descriptions of the entire cell lineage of *C. elegans*. We have used this ontology to describe the function of cells and tissues. That is, we describe the phenotypes caused by physical or genetic perturbation of anatomic structures. For example, a cell may be physically ablated or the genes expressed in it may be changed by tissue-specific gene expression. To date, we have annotated 350 such experiments. Complimentary to anatomy ontology, we are now working to represent neural anatomic relationships using the network browser N-Browse (16).

Transcription factor-binding sites

We have expanded WormBase to deal with position-specific matrices defining regulatory elements in nematode genomes, with 58 currently annotated (http://www.wormbase.org/db/seq/position_matrix/?list=all). Our data model for matrices can accommodate either frequency counts [in position frequency matrices (PFMs)] or logarithmically weighted tables [in position weight matrices (PWMs); 17]. We use the transcription factor-binding site (TFBS) module of Lenhard and Wasserman (18) to map frequency to weight data, given a background nucleotide frequency of 34% CG. We are also increasing the level of manual curation of experimentally determined sequence features, including TFBSs.

USER INTERFACE AND SITE INFRASTRUCTURE

In the past 2 years, we have added a number of significant new data displays. For example, Gene Pages now indicate genes implicated in human disease with their corresponding OMIM entry. A new orthology display lists precalculated ortholog assignments (http://www.wormbase.org/db/orthology/gene). Finally, we unified our display of the three ontologies available at WormBase (Gene, Phenotype and Anatomy) with a single search and browsing interface (http://www.wormbase.org/db/ontology/search).

We routinely evaluate and revise our service infrastructure to match both size and performance needs of the database as well as a growing user base. One approach to meeting increased demand has been the establishment of mirror sites across the globe. We recently upgraded our principal mirror at Caltech (http://caltech.wormbase.org). Behind the scenes, we have expanded our server pool as necessary to ensure a suitable browsing experience, employing a robust reverse proxy load-balancing system to maintain service even in the event of server failure.

ARCHIVAL RELEASES

With every fifth release of the database (currently an interval of ~6 months) we create and permanently archive a frozen release of the database and website. In the past 2 years, we created releases such as WS180, WS190, WS200 and WS205. Each of these is available at a unique URL (e.g., http://ws205.wormbase.org). We encourage users to use and cite these referential releases of the database.

INTEGRATION WITH THIRD-PARTY RESOURCES

Although the data housed at WormBase is extensive, it is not comprehensive. When appropriate, we provide cross-links to external resources for additional information. For example, we draw data from Reactome (19) to provide users with detailed pathway and process data from directly within WormBase.

One notable new partner is the journal *Genetics* (http://www.genetics.org). Using the text indexing tool Textpresso (14), *Genetics* now automatically markups new papers, generating links back to WormBase. This provides online readers of *Genetics* with expansive information on genes, proteins, sequences and phenotypes directly from a refereed journal article.

WormBase is also tightly linked to the open-access online publication WormBook (http://www.wormbook.org; 20). WormBase objects referenced in WormBook chapters are cross-linked to their corresponding entries at WormBase. From WormBase, users can navigate

from objects such as genes to more in-depth discussion at WormBook on a wide range of topics.

COMMUNITY OUTREACH

Finding effective means of communication with end users is a challenge for all who build and maintain biological databases. Users are often reluctant or too busy to contact database maintainers with bug reports, feature requests or general questions on how to use the site. To address these concerns, we have established several new channels of communication with our end users.

A Twitter feed gives us a quick mechanism for letting users know about site status (http://www.twitter.com/wormbase). A separate feed allows interested parties to follow updates to our core software (http://twitter.com/wormbase_commit).

The general aggregator FriendFeed (http://friendfeed.org/wormbase) provides both an easy method for users to stay up-to-date with happenings at the site and as a real-time Help Desk. Users can post questions to the WormBase FriendFeed and receive nearly instantaneous response from WormBase staff. Instead of being sequestered on a mailing list, this feedback is publicly available and searchable for the benefit of all users.

Finally, the Worm Community Forum (http://www.wormbase.org/forums) gives users a place for general discussion related to nematode biology, including experimental protocols, reagents, microscopy and even job postings and meeting details. The forum now has over 800 members and receives several posts per day.

FUTURE DIRECTIONS

We are actively developing the next generation of the WormBase website. Design considerations take into account scalability and usability employing modern web interface design and navigation patterns. Development of a comprehensive Perl API to all databases and services that comprise WormBase forms the core of this rearchitecture. A subset of this API will be accessible as a RESTful (21) web service. This new interface will give users a high degree of flexibility in how they navigate and interrogate the data available at WormBase, as well as give other database providers an easy mechanism to embed WormBase data in their own web sites. Beta-testing will begin in late Fall 2009.

In the coming year, we will add a number of significant new data sets and visualizations. First, WormBase will incorporate all data generated under the modENCODE (http://modencode.org/) project, making it available on the *C. elegans* Genome Browser. Second, we will add additional genomes to the resource, including the ruminant pathogen *Haemonchus contortus*. Third, we will begin curating metabolic process and pathway data using the BioCyc format (22). Interested users will be able to perform metabolic analyses using up-to-date nematode genomes and annotations. Finally, in an effort to make the substantial amount of microarray data already available in WormBase easier to query, we will deploy a SPELL (Serial Patterns of Expression Levels Locator) database and interface (23).

FUNDING

US National Institutes of Health (Grant no. P41 HG02223); US National Human Genome Research Institute (Grant no. P41-HG02223 to WormBase); British Medical Research Council (to WormBase); P.W.S. is an investigator with the Howard Hughes Medical Institute. Funding for open access charge: US National Human Genome Research Institute (Grant P41-HG02223).

Conflict of interest statement. None declared.

REFERENCES

1. Riddle,D.L., Blumenthal,T., Meyer,B.J. and Priess,J.R. (1997) *C. elegans II*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
2. *C. elegans* genome sequencing consortium. (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, **282**, 2012–2018.
3. 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.
4. 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.
5. Abad,P., Gouzy,J., Aury,J.M., Castagnone-Sereno,P., Danchin,E.G.J., Deleury,E., Perfus-Barbeoch,L., Anthouard,V., Artiguenave,F., Blok,V.C. *et al.* (2008) Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*. *Nat. Biotechnol.*, **26**, 909–915.
6. Opperman,C.H., Bird,D.M., Williamson,V.M., Rokhsar,D.S., Burke,M., Cohn,J., Cromer,J., Diener,S., Gajan,J., Graham,S. *et al.* (2008) Sequence and genetic map of *Meloidogyne hapla*: a compact nematode genome for plant parasitism. *Proc. Natl Acad. Sci. USA*, **105**, 14802–14807.
7. Dieterich,C., Clifton,S.W., Schuster,L.N., Chinwalla,A., Delehaunty,K., Dinkelacker,I., Fulton,L., Fulton,R., Godfrey,J., Minx,P. *et al.* (2008) The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism. *Nat. Genet.*, **40**, 1193–1198.
8. 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.
9. Hubbard,T.J.P., Aken,B.L., Ayling,S., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Clarke,L. *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.
10. Berglund,A.C., Sjölund,E., Ostlund,G. and Sonnhammer,E.L. (2008) InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res.*, **36**, D263–D266.
11. Li,H., Coghlan,A., Ruan,J., Coin,L.J., Hériché,J., Osmotherly,L., Li,R., Liu,T., Zhang,Z., Bolund,L. *et al.* (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.*, **34**, D572–D580.
12. Müller,H.M., Kenny,E.E. and Sternberg,P.W. (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.*, **2**, E309.
13. Van Auken,K., Jaffery,J., Chan,J., Müller,H.M. and Sternberg,P.W. (2009) Semi-automated curation of protein subcellular localization: a text mining-based approach to gene ontology (GO) cellular component curation. *BMC Bioinform.*, **10**, 228.

14. Chen,D., Müller,H.M. and Sternberg,P.W. (2006) Automatic document classification of biological literature. *BMC Bioinform.*, **7**, 370.
15. de Bono,M. and Bargmann,C.I. (1998) Natural variation in a neuropeptide Y receptor homolog modifies social behavior and food response in *C. elegans*. *Cell*, **94**, 679–689.
16. Kao,H.L. and Gunsalus,K.C. (2008) Browsing multidimensional molecular networks with the generic network browser (N-Browse). *Curr. Protoc. Bioinform.*, Chapter 9, Unit 9.11.
17. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.
18. Lenhard,B. and Wasserman,W.W. (2002) TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics*, **18**, 1135–1136.
19. Matthews,L., Gopinath,G., Gillespie,M., Caudy,M., Croft,D., de Bono,B., Garapati,P., Hemish,J., Hermjakob,H., Jassal,B. *et al.* (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, **37**, D619–D622.
20. Girard,L., Fiedler,T.J., Harris,T.W., Carvalho,F., Antoshechkin,I., Han,M., Sternberg,P.W., Stein,L.D. and Chalfie,M. (2007) WormBook: the online review of *Caenorhabditis elegans* biology. *Nucleic Acids Res.*, **35**, D472–D475.
21. Richardson,L. and Ruby,S. (2007) *RESTful Web Services*. O'Reilly Media, Inc., Sebastapol, CA.
22. Caspi,R., Foerster,H., Fulcher,C.A., Kaipa,P., Krummenacker,M., Latendresse,M., Paley,S., Rhee,S.Y., Shearer,A.G., Tissier,C. *et al.* (2008) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **36**, D623–D631.
23. Hibbs,M.A., Hess,D.C., Myers,C.L., Huttenhower,C., Li,K. and Troyanskaya,O. (2007) Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics*, **23**, 2692–2699.