D506–D510 *Nucleic Acids Research, 2007, Vol. 35, Database issue* *Published online 11 November 2006*
doi:10.1093/nar/gkl818

WormBase: new content and better access

Tamberlyn Bieri*, Darin Blasiar, Philip Ozersky, Igor Antoshechkin[1], Carol Bastiani[1], Payan Canaran[3], Juancarlos Chan[1], Nansheng Chen[3], Wen J. Chen[1], Paul Davis[4], Tristan J. Fiedler[3], Lisa Girard[1], Michael Han[4], Todd W. Harris[3], Ranjana Kishore[1], Raymond Lee[1], Sheldon McKay[3], Hans-Michael Müller[1], Cecilia Nakamura[1], Andrei Petcherski[1], Arun Rangarajan[1], Anthony Rogers[4], Gary Schindelman[1], Erich M. Schwarz[1], Will Spooner[3], Mary Ann Tuli[4], Kimberly Van Auken[1], Daniel Wang[1], Xiaodong Wang[1], Gary Williams[4], Richard Durbin[4], Lincoln D. Stein[3], Paul W. Sternberg[1,2] and John Spieth

Genome Sequencing Center, Washington University School of Medicine, St Louis, MO 63108, USA, [1]Division of Biology, 156-29 and [2]Howard Hughes Medical Institute, California Institute of Technology, Pasadena, CA 91125, USA, [3]Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA and [4]Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK

Received September 14, 2006; Revised October 3, 2006; Accepted October 4, 2006

ABSTRACT

WormBase (http://wormbase.org), a model organism database for *Caenorhabditis elegans* and other related nematodes, continues to evolve and expand. Over the past year WormBase has added new data on *C.elegans*, including data on classical genetics, cell biology and functional genomics; expanded the annotation of closely related nematodes with a new genome browser for *Caenorhabditis remanei*; and deployed new hardware for stronger performance. Several existing datasets including phenotype descriptions and RNAi experiments have seen a large increase in new content. New datasets such as the *C.remanei* draft assembly and annotations, the Vancouver Fosmid library and TEC-RED 5′ end sites are now available as well. Access to and searching WormBase has become more dependable and flexible via multiple mirror sites and indexing through Google.

DESCRIPTION

Much of our understanding of biological processes and human biology comes from a vast amount of data generated in the study of a few model organisms. The National Human Genome Research Institute (NHGRI; http://www.genome.gov/) in the United States and the British Medical Research Council have funded model organism databases (MODs) to store and integrate information for individual model organisms. WormBase is one of these MODs for the small, soil nematode, *Caenorhabditis elegans*. WormBase originated in 2001 and initially was simply a web-based interface for its predecessor, ACeDB (1,2) (http://www.acedb.org/), which contained primarily the genetic and physical maps, and the genome sequence of *C.elegans*. Over the years, WormBase has evolved and expanded through developing and utilizing better software, and providing richer content (3) by bringing in data from classical genetics, cell biology and functional genomics (4,5), and the genomic sequences of other closely related nematodes (6,7).

WormBase is maintained by the International WormBase Consortium, a group of ∼30 scientists located at four sites (http://www.wormbase.org/wiki/index.php/WormBase_Consortium). The data are organized into a single database that is available for downloading (http://www.wormbase.org/wiki/index.php/Downloads) and web browsing. Access to, and use of, the database and resources are freely available, subject to an Acceptable Use Policy (http://www.wormbase.org/wiki/index.php/Acceptable_use_policy) and copyrights (http://www.wormbase.org/wiki/index.php/WormBaseWiki:Copyrights). There is also a User's Guide (http://www.its.caltech.edu/~wormbase/userguide/), an evolving Wiki site (http://www.wormbase.org/wiki) and an email-based Help Desk (wormbase-help@wormbase.org)

WormBase is not a static database. It is constantly growing with new data continually being added and made available to users through new builds and releases of the database every three weeks. Every 10th release is maintained as a permanently available, stable data source for reference. This review is an overview of major new content, software and performance improvements, and how the user community helps to focus efforts of WormBase.

*To whom correspondence should be addressed. Tel: +1 314 286 1957; Fax: +1 314 286 1810; Email: tbieri@watson.wustl.edu

NEW CONTENT

WormBase's content has doubled in the last three years as measured by the size of its underlying database (Figure 1). This growth is expected to increase as research using *C.elegans* expands and more nematode genome sequencing projects are completed.

Additions to existing data types

Phenotypes. Over the past year one of the largest percentage increases in an existing data type has been for phenotype objects. There have been major changes in phenotype curation, including a new and expanded Phenotype Ontology, the evaluation and consolidation of preexisting phenotype classes, and improvements in phenotype curation methods. All of these allow more accurate and detailed phenotype information to be stored in WormBase, and make querying for phenotype-related information more efficient. Consequently, the number of phenotype objects in WormBase has increased dramatically over the past year from 119 to 1282.

Gene Ontology. Gene products annotated with Gene Ontology (GO) terms are another existing data type to see a large percentage increase. WormBase is a member of the Gene Ontology Consortium (8), and is developing GO content and annotating *C.elegans*' genes in GO terms. Currently there are over 50 000 GO annotations for over 10 000 genes in WormBase. Most of these are automated mappings, although there are curator-performed, manual annotations for ~700 genes with over 800 unique papers cited as references. The automatic GO annotations are of two types. The first is the automated assignment of GO terms to genes based on the RNAi-knockdown phenotype. For example, an RNAi phenotype of Egl (egg-laying defective) is automatically assigned the GO biological process term 'oviposition'. These are based on manual mapping of phenotypes to GO terms from large-scale RNAi screens. The second type of automated GO annotation uses the InterPro2GO mappings (http://www.geneontology.org/external2go/interpro2go) provided by the European Bioinformatics Institute (http://www.ebi.ac.uk/). WormBase has recently begun determining InterPro domains with each build to ensure that they are completely up to date. These are GO term assignments based on protein family, domain, repeat, etc. In WormBase the GO annotations can now be found in a Gene Ontology section on the Gene page, which is the most common entry point for users (Supplementary Figure 1a). A summary of all the GO annotations (Supplementary Figure 1b) is reached through the GO summary link in the 'Gene Ontology' section. Details for individual annotations (Supplementary Figure 1c) can also be obtained through links in the section.

RNAi. RNAi data have also increased significantly over the past year, specifically from papers that describe individual RNAi experiments. Curation of these experiments is slow and labor intensive compared with papers containing large-scale datasets. Nonetheless, there has been a 150% increase in the individual RNAi experiments in WormBase. There are currently 2754 such experiments. Including curation of large-scale datasets, there are 63 740 total, curated RNAi experiments in WormBase, compared with 59 882 a year ago. There is still more to do, however. Approximately 45% of 871 *C.elegans* papers that contain RNAi data have been curated, compared with just over 20% a year ago. RNAi results are summarized in the Function section of the Gene page (Supplementary Figure 2a). A link to view all RNAi experiments that target the gene takes the user to a table (Supplementary Figure 2b) containing the details and results of each RNAi experiment.

Microarray data. The June 2006 database release (WS160) has microarray data from 33 papers describing 417 experiments, which represents almost a doubling in the number of curated, microarray papers and a 78% increase in the number of curated experiments in WormBase over the past year.

To support microarray research in the worm community, the probe sets for two new microarray platforms available from Agilent (http://www.chem.agilent.com/scripts/pds.asp?lPage=29452) and Washington University's Genome

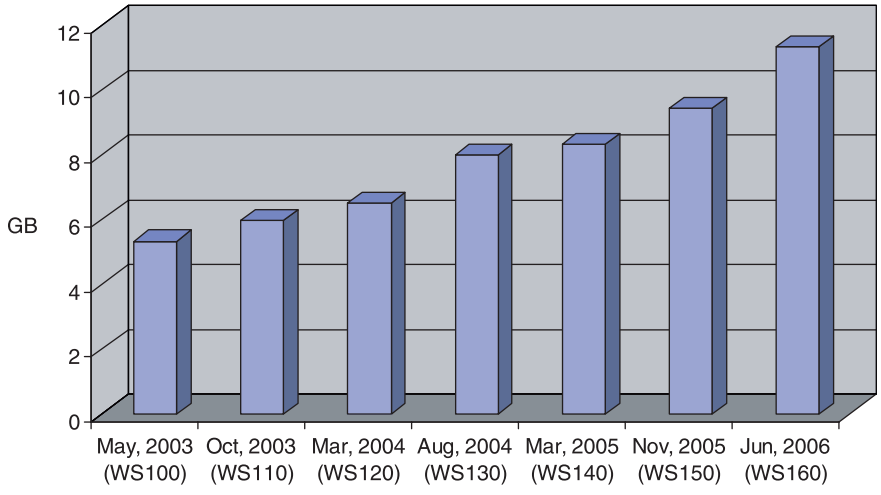

Figure 1. The increase in content of WormBase as measured by the size of the underlying database. Plotted are the sizes in gigabytes of the archived releases of the WormBase database.

Sequencing Center (http://www.genome.wustl.edu/genome/celegans/microarray/ma_gen_info.cgi) are now in WormBase. These are in addition to the Affymetrix chip data previously available in WormBase. All of these data can be downloaded using the WormBase data-mining tool, WormMart (http://www.wormbase.org/biomart/martview). The probes are mapped to the genome sequence and displayed as a track on the Genome Browser (9) (Supplementary Figure 3a). They are also mapped to corresponding gene models (Supplementary Figure 3b) with detailed reports (Supplementary Figure 3c) linked to the Gene page.

Intellectual lineage. The Intellectual Lineage was conceived as a way to indicate the heritage of the *C.elegans* research community. The types of connections a person can have are as follows: Supervised, Supervised by and Collaborated with. There are 1461 people that have connections to other people with 3729 person-to-person connections. This is up from 1030 people and 2626 person-to-person connections from a year ago. The lineage for an individual can be found on their Person page (Supplementary Figure 4), while the global view of the entire worm community lineage is available at WormBase (http://www.wormbase.org/presentations/2005/2005-intellectual_lineage.html).

New datasets

Caenorhabditis remanei annotations. A preliminary, draft assembly and automated annotations of the *C.remanei* genome are now available on a Genome Browser (http://wormbase.org/db/seq/gbrowse/remanei/) and can be download from the WormBase ftp site (ftp://ftp.wormbase.org/pub/wormbase). Annotations include gene predictions from several *ab initio* and alignment-based algorithms. Sequence improvement is ongoing at the Washington University Genome Sequencing Center with a final draft assembly and new annotations anticipated early in 2007. A summary of the best BLASTP match to *C.remanei* is now displayed on *C.elegans* and *Caenorhabditis briggsae* Gene pages (Supplementary Figure 5).

Vancouver fosmids. The fosmid library constructed and mapped to the genome by Don Moerman and colleagues at the *C.elegans* Reverse Genetics Core Facility located at the University of British Columbia is now available on the Genome Browser in the YAC, Fosmids and Cosmids track (Supplementary Figure 6) and searchable in WormBase. This library was made to supplement the aging cosmid and YAC libraries used in the original mapping and sequencing of the genome, some of which are now over 20 years old, and are more difficult to work with than fosmids.

TEC-RED. This relatively new technique involves *trans*-spliced exon coupled 5′ RNA end determination, which identifies 5′ ends of expressed genes in nematodes [details of the procedure can be found in reference (10)]. The TEC-RED data of Hwang *et al*. (10) is now in WormBase and can be viewed in the Genome Browser (Supplementary Figure 7). These data have been used extensively to update numerous gene structures.

IMPROVED ACCESS AND USABILITY

Fast and robust access

Community use of WormBase continues to grow. There are now over 2 million page hits per month, a 40% increase over the past year, with the Gene page being the most frequently accessed page. This increase prompted the deployment of a multiply-redundant, load-balancing server system, which now distributes requests across three back-end servers, each capable of serving the complete WormBase site. Two additional servers handle specialized tasks such as BLAST requests and query pages. There is also robust caching that delivers the most frequently requested pages from a high-speed, in-memory and on-disk cache. These improvements have resulted in dramatic increases in reliability and performance. Inserting low-cost machines into the existing infrastructure can easily accommodate increased demand in the future.

Access to the data at WormBase is becoming more diverse and flexible. In addition to the main WormBase site located at the Cold Spring Harbor Laboratory (http://www.wormbase.org/), there are now two new mirror sites: one at the Welcome Trust Sanger Institute in England (http://wormbase.sanger.ac.uk/) and one at the University of Marseille in France (http://crfb-3.univ-mrs.fr/) that provide flexibility and backup if the main server experiences problems. This increases the total number of mirror sites to four with the existing sites at the California Institute of Technology (http://caltech.wormbase.org/) and the Institute of Molecular Biology and Biochemistry in Greece (http://imbb.wormbase.org/). There are links to all the mirror sites and the main and development site on the WormBase homepage (Supplementary Figure 8).

Another portal to WormBase is through WormBook (http://www.wormbook.org/). WormBook content is extensively linked to WormBase with Genes, Proteins and Cells linked to the relevant pages in WormBase. Links from WormBase back to WormBook have recently been implemented to provide users easy access to background and in-depth information.

Improved searching

More recently, access to WormBase can be gained through search engines such as Google, which are now able to index the site. Previously their access had been restricted due to load issues on the old, single server infrastructure. Researchers worldwide can now search the vast content of highly curated data without having to actually visit the WormBase site. Google searches can be limited to WormBase results by using the format 'site: www.wormbase.org [search terms]'. Google's page caching also provides an additional layer of redundancy during unexpected server interruptions at the WormBase site.

Searching for strains at the *Caenorhabditis* Genetic Center (CGC) has been improved and is now hosted at WormBase. In addition to searching by strain name, users may now search for strains carrying a specific gene or allele, or other information found in the strain class (i.e. species, mutagen, data received and remarks).

Improved data presentation

Data layout and access has become more intuitive with new or updated web pages. Information from the Protein Data Bank's TargetDB (11) (http://targetdb.pdb.org/), which provides status and tracking information on the production and solution of structures, is now displayed on Gene pages. Also on the Gene page, a GO summary (Supplementary Figure 1) has been added as well as three types of evolutionary data: InParanoid clusters of orthologous genes (12) (http://inparanoid.cgb.ki.se/), Treefam's (13) (http://www.treefam.org/) curated ortholog and paralog assignments (Figure 2), and *C.elegans*/*C.briggsae* mutual best BLASTP matches (Supplementary Figure 5). The Gene pages also include the phenotype of RNAi experiments, and physical and genetic interactions, which are now displayed in a convenient table (Supplementary Figure 2a).

WORMBASE IS COMMUNITY DRIVEN

The focus and direction of WormBase is driven by the user community. Two ways this occurs is through user surveys and the WormBase advisory board. User surveys are commissioned periodically to gauge user concerns and direct future efforts for data curation. The most recent survey was conducted in the fall of 2005 (http://www.wormbase.org/db/misc/2005_survey). The results were discussed at the WormBase advisory board in November 2005 and summarized in the January 2006 WormBase Newsletter (http://www.wormbase.org/announcements/newsletters/pdf/2006-01.pdf). One of the concerns of many of the respondents was the speed of WormBase. The Advisory Board felt improving response time should be a major priority, and WormBase responded with the new hardware structure, additional servers and page caching described above.

The WormBase Advisory board meets once a year with the entire WormBase staff to review progress and prioritize projects for the upcoming year. The board consists of several members of the *C.elegans* research community, as well as people with expertise in bioinformatics and databases.

WormBase also reaches out to the user community by having staff members at all the international, regional and topical *C.elegans* meetings, as well as other meetings, often presenting talks and posters, but mainly to receive input from users regarding problems as well as new features or datasets they would like to see WormBase host. Some requests are for information for which WormBase simply does not have access, such as large-scale studies before publication.

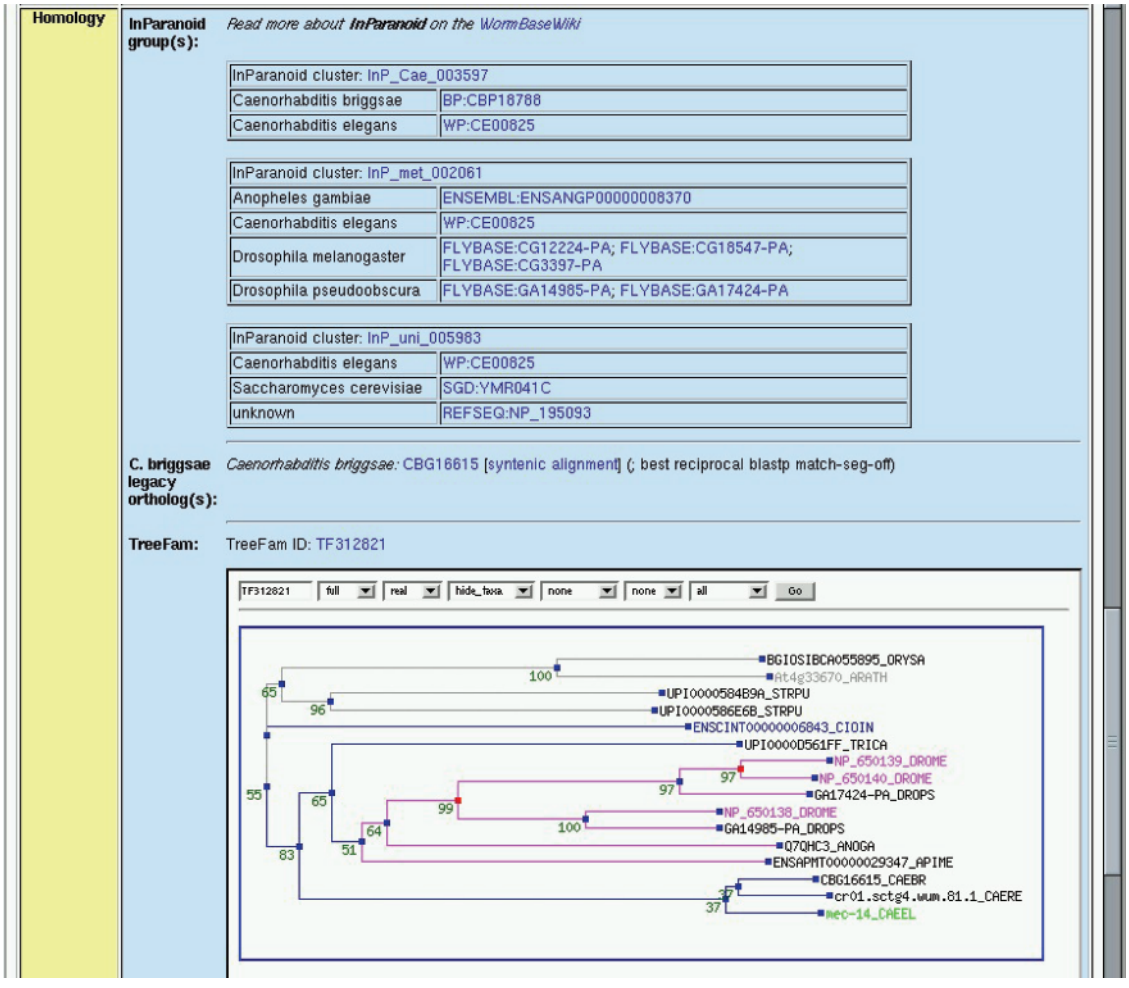

Figure 2. Screen shot of the Homology section of the *mec-14* Gene page showing InParanoid groups, *C.briggsae* orthologs and a TreeFam tree.

WormBase strives to have data available at the time of publication and encourages authors to contact WormBase about upcoming publications.

The WormBase Wiki site (http://www.wormbase.org/wiki) provides a mechanism for users to easily add content to WormBase. Individuals can post experimental protocols, make posting about meetings and job openings, or add useful information about their favorite gene(s). A long-term goal is for the Wiki site to assume the role of the print version of the Worm Breeder's Gazette.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

P.W.S. is an investigator with the Howard Hughes Medical Institute. WormBase is supported by grant P41-HG02223 from the US National Human Genome Research Institute and by the British Medical Research Council. Funding to pay the Open Access publication charges for this article was provided by grant P41-HG02223 from the US National Human Genome Research Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Eeckman,F.H. and Durbin,R. (1995) ACeDB and macace. *Methods Cell Biol*., **48**, 583–605.
2. Stein,L.D. and Thierry-Mieg,J. (1998) Scriptable access to the *Caenorhabditis elegans* genome sequence and other ACEDB databases. *Genome Res*., **8**, 1308–1315.
3. Schwarz,E.M., Antoshechkin,I., Bastiani,C., Bieri,T., Blasiar,D., Canaran,P., Chan,J., Chen,N., Chen,W.J., Davis,P. *et al*. (2006) WormBase: better software, richer content. *Nucleic Acids Res*., **34**, D475–D478.
4. Chen,N., Lawson,D., Bradnam,K., Harris,T.W. and Stein,L.D. (2004) WormBase as an integrated platform for the *C.elegans* ORFeome. *Genome Res*., **14**, 2155–2161.
5. Chen,N., Harris,T.W., Antoshechkin,I., Bastiani,C., Bieri,T., Blasiar,D., Bradnam,K., Canaran,P., Chan,J., Chen,C.K. *et al*. (2005) WormBase: a comprehensive data resource for *Caenorhabditis* biology and genomics. *Nucleic Acids Res*., **33**, D383–D389.
6. 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.
7. Harris,T.W., Chen,N., Cunningham,F., Tello-Ruiz,M., Antoshechkin,I., Bastiani,C., Bieri,T., Blasiar,D., Bradnam,K., Chan,J. *et al*. (2004) WormBase: a multi-species resource for nematode biology and genomics. *Nucleic Acids Res*., **32**, D411–D417.
8. Gene Ontology Consortium. (2001) Creating the gene ontology resource: design and implementation. *Genome Res*., **11**, 1425–1433.
9. 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.
10. Hwang,B.J., Müller,H.M. and Sternberg,P.W. (2004) Genome annotation by high-throughput 5′ RNA end determination. *Proc. Natl Acad. Sci. USA*, **101**, 1650–1655.
11. Chen,L., Oughtred,R., Berman,H.M. and Westbrook,J. (2004) TargetDB: a target registration database for structural genomics projects. *Bioinformatics*, **20**, 2860–2862.
12. Remm,M., Storm,C.E.V. and Sonnhammer,E.L.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol*., **314**, 1041–1052.
13. Li,H., Coghlan,A., Ruan,J., Coin,L.J., Hériché,J.-K., Osmotherly,L., Li,R., Lui,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.