Nucleic Acids Research, 2006, Vol. 34, Database issue ***D475–D478***
doi:10.1093/nar/gkj061

WormBase: better software, richer content

Erich M. Schwarz[1,*], Igor Antoshechkin[1], Carol Bastiani[1], Tamberlyn Bieri[3], Darin Blasiar[3], Payan Canaran[4], Juancarlos Chan[1], Nansheng Chen[4], Wen J. Chen[1], Paul Davis[5], Tristan J. Fiedler[4], Lisa Girard[1], Todd W. Harris[4], Eimear E. Kenny[1], Ranjana Kishore[1], Dan Lawson[5], Raymond Lee[1], Hans-Michael Müller[1], Cecilia Nakamura[1], Phil Ozersky[3], Andrei Petcherski[1], Anthony Rogers[5], Will Spooner[4], Mary Ann Tuli[5], Kimberly Van Auken[1], Daniel Wang[1], Richard Durbin[5], John Spieth[3], Lincoln D. Stein[4] and Paul W. Sternberg[1,2]

[1]Division of Biology, 156-29 and [2]Howard Hughes Medical Institute, California Institute of Technology, Pasadena, CA, 91125, USA, [3]Genome Sequencing Center, Washington University School of Medicine, St Louis, MO 63108, USA, [4]Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY, 11724, USA and [5]Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SA, UK

Received September 13, 2005; Revised and Accepted October 6, 2005

ABSTRACT

WormBase (http://wormbase.org), the public database for genomics and biology of *Caenorhabditis elegans*, has been restructured for stronger performance and expanded for richer biological content. Performance was improved by accelerating the loading of central data pages such as the omnibus Gene page, by rationalizing internal data structures and software for greater portability, and by making the Genome Browser highly customizable in how it views and exports genomic subsequences. Arbitrarily complex, user-specified queries are now possible through Textpresso (for all available literature) and through WormMart (for most genomic data). Biological content was enriched by reconciling all available cDNA and expressed sequence tag data with gene predictions, clarifying single nucleotide polymorphism and RNAi sites, and summarizing known functions for most genes studied in this organism.

DESCRIPTION

WormBase is the central public database for *Caenorhabditis elegans* biology. It began as a web interface for its predecessor, the genomic database ACeDB (1,2). During the last half decade, it has been expanded to cover classical genetics and cell biology (3), functional genomics (4) and the *Caenorhabditis briggsae* genomic sequence (5,6). New releases of WormBase are built every three weeks by amalgamating physical and genome sequence data from the *C.elegans* Sequencing Consortium (Sanger Institute and Washington University, St Louis), genetic map data curated by the *Caenorhabditis* Genetics Center (University of Minnesota and Oxford University) and diverse biological data curated by the WormBase Consortium. Every 10th release is maintained as a permanently available, stable data source ('freeze') for reproducible bioinformatics. In the last year, the WormBase Consortium has worked to make WormBase more reliably useful and stable, while continuing to add new biology and preparing to handle an expected five nematode genomic sequences in 2006.

USABILITY

Much of WormBase is organized around two key data hubs, the Gene page and the Genome Browser. Both of these can summarize large amounts of data in a single view. However, as the contents of WormBase grew, the Gene page became increasingly slow to load in users' web browsers. We revised our software so that Gene pages are pre-built and stored ready for use; as a result, most Gene pages now loaded in <10 s. We also redesigned WormBase so that its software and data releases are packaged for easy uploading and updating. This allowed us to construct and maintain several mirror sites at the Institute of Molecular Biology and Biotechnology (Crete, Greece), the California Institute of Technology, and the Center for Computational Biology and Bioinformatics (Daejeon, South Korea). It also allows us to run WormBase on laptop computers for network-independent use and efficient software development. These improvements were made possible by clarifying internal data structures that are invisible to the user but critical for effective database management. For example, classical loci and coding sequences were consolidated into a single Gene data object that can stably

*To whom correspondence should be addressed. Tel: +1 626 394 7078; Fax: +1 626 568 8012; Email: emsch@its.caltech.edu

represent genes regardless of fluctuations in their classical or molecular names.

We revised the Genome Browser display (7) so that different subsets of genomic data ('tracks'), as well as different sections of the Browser's display framework, can be alternatively shown or hidden at the user's option. This allows a user to construct economical and individualized views of any section of the genome, ranging in size from a few nucleotides to 1 Mb in length. These views can be bookmarked as stable URLs, or exported as publication-quality scalable-vector graphics images. Protein motifs (8–11) now have their own data track, showing the domain organization of proteins in the context of intron/exon structure, interspecies conservation, single nucleotide polymorphisms (SNPs), PCR reagents, RNAi results and other genomic features. Moreover, users can import and display their own data tracks seamlessly beside the core WormBase ones, either by uploading their own annotations from a local text file or by invoking a remote URL; using remote URLs enables collaborative genomic analyses by multiple users sharing a common data repository.

SEARCH ENGINES

We developed Textpresso, a tool for searching the full content of *C.elegans* articles for meaningful word relationships, and incorporated it into WormBase (12). We recently expanded the Textpresso ontology with four new categories: 'reporter gene', 'restriction enzyme', 'second messenger' and 'vector'. We also added new terms to the 'drugs and small molecules' and 'organism' categories. The literature searchable by Textpresso within WormBase contains 6259 full-text articles, including 5571 from the core *C.elegans* literature; this body of literature is automatically updated and expanded every week. Textpresso also contains 18 642 abstracts, including 8450 from international and regional *C.elegans* meetings. While Textpresso was first designed for use by WormBase, it has proven useful to several other model organism databases (e.g. http://www.yeastgenome.org/textpresso and http://www.ciliate.org/textpresso) and is being extended to non-genomic disciplines (such as neuroscience; http://www.textpresso.org/neuro). Textpresso has been made available to the Generic Model Organism Database software project (http://www.gmod.org) as open source code.

WormMart is a data warehousing system (13) that allows users to construct complex queries on WormBase and obtain results in HTML or tab-delimited text format. WormMart supersedes the 'Batch Sequences' and 'Batch Genes' reports, and facilitates arbitrarily complex queries such as 'Find all genes in *C.elegans* that have an orthologue in *C.briggsae*, are located in chromosome III, have reduced fertility in an RNAi screen, and have annotated untranslated regions (UTRs)'. In addition to gene-centric queries, WormMart supports querying over-expression patterns, RNAi phenotypes, mutant phenotypes, variations (alleles) and literature citations. WormMart is based on BioMart (http://www.ebi.ac.uk/biomart), the core software driving the EnsMart query engine at EnsEMBL (13).

GENOMIC BIOLOGY

Even for *C.elegans*, with a relatively compact and well-determined genomic sequence, it is a continuing challenge to detect the existence and correct structures of ~21 600 genes (14). In the past year, 2405 gene structures have been revised or newly identified. Approximately 5000 cDNAs have been connected to protein-coding sequences, resulting in 948 more protein-coding sequences becoming completely confirmed by cDNA data (a 17.2% increase). Essentially all *C.elegans* expressed sequence tag (EST) and cDNA sequences in public databases have been incorporated into WormBase gene structures. The number of introns identifiable from cDNA sequences but absent from existing gene structures was lowered considerably (from 746 to 121). Many other data besides cDNA sequences were also used to identify correct gene structures: detailed studies from individual research papers, personal communications to WormBase staff, Twinscan predictions (15), SL1/2 (16) and TEC-RED sequences (17), multiple alignments of protein families (18), and *C.briggsae* homologies (5). 5′- and 3′-UTRs in WormBase are now automatically generated as part of full-length coding transcripts, taking into account additional data such as *trans*-spliced 5′ leader sequences (16) and polyadenylation sites (19); 1800 new instances of *trans*-splicing have been identified.

SNPs (20,21) have been systematically overhauled. As originally published, *C.elegans* SNPs have often been inconsistent or incomplete: clone positions have changed over the years as sequence changes have been made, and published flanking sequences were often too short to uniquely map them to either clones or chromosomes. We thus went back through the original data and generated new flanking sequences that are unique in the genome. Similarly, we remapped all RNAi experiments, while adding two new large-scale datasets from an ORFeome library-based RNAi screen (22) and a full-genome RNAi profiling of early embryogenesis (23). This brought the total number of large-scale RNAi data points in WormBase from 27 112 to 58 778, and the number of distinct RNAi phenotypes from 78 to 119. We also continued adding microarray data to WormBase. The WS145 database release contained 2 984 398 microarray data points from 19 papers, describing 234 independent experiments, compared with 1 690 379 data points from 15 papers and 113 experiments from a year earlier.

Functional genomics is a growing part of WormBase, with the incorporation of protein–protein interaction and isolated promoter data: these currently include 5534 yeast two-hybrid interactions covering 15% of the *C.elegans* proteome (24) and 6538 promoter sequences cloned in the MultiSite Gateway system (25).

CELLULAR AND ORGANISMAL BIOLOGY

Molecular data become more useful when accompanied by human-readable, concise descriptions of gene function (26). In WormBase, 3064/7864 (39%) of genes that have been named (i.e., that are not simply anonymous, little-studied gene predictions) now have such descriptions. For genes with at least one reference, 58% have concise descriptions (2421/4133); for genes with five or more references, 74% have concise descriptions (925/1248); for those with >10 references, 76% have concise descriptions (635/839); with >100 references, 86% have concise descriptions (85/99); and with >200 references, 88% have concise descriptions (29/33). Thus, for those genes that are information-rich, we have

~75% coverage with our concise descriptions. In addition, we are also annotating gene functions with structured, computationally tractable gene ontology (GO) terms (27). 5806 gene–GO term linkages have so far been identified, from data in 718 references. Meanwhile, the entire genome has been scanned with automatic mappings to GO terms from RNAi phenotypes and from Interpro domains (28), yielding a total of 23 688 annotations.

There is far more important biology of *C.elegans* than WormBase can expect to describe in a reasonable time by traditional approaches. We thus developed a new semi-automated annotation strategy and tested it by mass-extracting genetic interactions from the primary literature. Extraction began with a Textpresso advanced query (12) for sentences containing ⩾2 'gene', ⩾1 'association' and ⩾1 'regulation' categories. A curator then read the individual sentences and identified individual gene–gene interactions. In this way, ~26 000 sentences were retrieved by Textpresso from ~4400 papers. From these, ~10 000 interactions or possible interactions were identified, including: 5439 genetic interactions (54%); 1820 non-genetic interactions (18%); and 2739 possible interactions (27%). These represented ~2000 unique, previously unannotated gene pairs.

WORMBOOK

Two encyclopedic volumes describing *C.elegans* biology were published in 1988 and 1997 (29,30); while still invaluable, both predate the last decade of research and the rise of functional genomics with web-based bioinformatics. WormBook is a new, online collection of original reviews on topics related to all aspects of *C.elegans* biology, as well as a repository for experimental protocols used by *C.elegans* researchers. WormBook is freely available as HTML or PDF documents (www.wormbook.org). WormBook provides a text companion to WormBase with contributions by >100 expert biologists reviewing and synthesizing the facts presented in WormBase. When complete, WormBook will have hypertext links for genes, alleles, proteins and literature citations to WormBase and PubMed. Conversely, researchers using these linked primary databases will have reciprocal access to WormBook, facilitating the exchange of ideas and promoting further research. Over 20 completed WormBook chapters and >60 preprints of WormBook chapters had been released by September 2005.

FUTURE DIRECTIONS

We anticipate that WormBase will be called upon to manage genomic sequences from multiple *Caenorhabditis* species (31,32). Work on this began in 2005 with a gene set prediction for *Caenorhabditis remanei*, in which several different gene prediction sets were generated, tested against *C.elegans* genomic and *C.remanei* EST sequences, combined and hierarchically selected for the best possible automatic prediction. This yielded a total of 26 253 predicted *C.remanei* genes. We also intend to expand the classical biological content of WormBase by systematically annotating mutant alleles with an extensive phenotype ontology (33) adopted for nematodes to allow better searches of gene function. We plan to make cells, cell groups and biological processes more significant entry points into the content of WormBase.

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. 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.
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. 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.
6. 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.
7. 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.
8. 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.
9. Wootton,J.C. (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem*., **18**, 269–285.
10. Lupas,A. (1996) Prediction and analysis of coiled-coil structures. *Methods Enzymol*., **266**, 513–525.
11. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol*., **305**, 567–580.
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. Kasprzyk,A., Keefe,D., Smedley,D., London,D., Spooner,W., Melsopp,C., Hammond,M., Rocca-Serra,P., Cox,T. and Birney,E. (2004) EnsMart: a generic system for fast and flexible access to biological data. *Genome Res*., **14**, 160–169.
14. Zhang,M.Q. (2002) Computational prediction of eukaryotic protein-coding genes. *Nature Rev. Genet*., **3**, 698–709.
15. Wei,C., Lamesch,P., Arumugam,M., Rosenberg,J., Hu,P., Vidal,M. and Brent,M.R. (2005) Closing in on the *C.elegans* ORFeome by cloning TWINSCAN predictions. *Genome Res*., **15**, 577–582.
16. Blumenthal,T. and Gleason,K.S. (2003) *Caenorhabditis elegans* operons: form and function. *Nature Rev. Genet*., **4**, 112–120.
17. 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.
18. Batzoglou,S. (2005) The many faces of sequence alignment. *Brief. Bioinformatics*, **6**, 6–22.
19. Hajarnavis,A., Korf,I. and Durbin,R. (2004) A probabilistic model of 3′ end formation in *Caenorhabditis elegans*. *Nucleic Acids Res*., **32**, 3392–3399.

20. Wicks,S.R., Yeh,R.T., Gish,W.R., Waterston,R.H. and Plasterk,R.H. (2001) Rapid gene mapping in *Caenorhabditis elegans* using a high density polymorphism map. *Nature Genet*., **28**, 160–164.
21. Swan,K.A., Curtis,D.E., McKusick,K.B., Voinov,A.V., Mapa,F.A. and Cancilla,M.R. (2002) High-throughput gene mapping in *Caenorhabditis elegans*. *Genome Res*., **12**, 1100–1105.
22. Rual,J.F., Ceron,J., Koreth,J., Hao,T., Nicot,A.S., Hirozane-Kishikawa,T., Vandenhaute,J., Orkin,S.H., Hill,D.E., van den Heuvel,S. *et al*. (2004) Toward improving *Caenorhabditis elegans* phenome mapping with an ORFeome-based RNAi library. *Genome Res*., **14**, 2162–2168.
23. Sonnichsen,B., Koski,L.B., Walsh,A., Marschall,P., Neumann,B., Brehm,M., Alleaume,A.M., Artelt,J., Bettencourt,P., Cassin,E. *et al*. (2005) Full-genome RNAi profiling of early embryogenesis in *Caenorhabditis elegans*. *Nature*, **434**, 462–469.
24. Li,S., Armstrong,C.M., Bertin,N., Ge,H., Milstein,S., Boxem,M., Vidalain,P.O., Han,J.D., Chesneau,A., Hao,T. *et al*. (2004) A map of the interactome network of the metazoan *C.elegans*. *Science*, **303**, 540–543.
25. Dupuy,D., Li,Q.R., Deplancke,B., Boxem,M., Hao,T., Lamesch,P., Sequerra,R., Bosak,S., Doucette-Stamm,L., Hope,I.A. *et al*. (2004) A first version of the *Caenorhabditis elegans* promoterome. *Genome Res*., **14**, 2169–2175.
26. Stein,L. (2001) Genome annotation: from sequence to biology. *Nature Rev. Genet*., **2**, 493–503.
27. Gene Ontology Consortium (2001), Creating the gene ontology resource: design and implementation. *Genome Res*., **11**, 1425–1433.
28. Camon,E., Magrane,M., Barrell,D., Binns,D., Fleischmann,W., Kersey,P., Mulder,N., Oinn,T., Maslen,J., Cox,A. *et al*. (2003) The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res*., **13**, 662–672.
29. Wood,W.B. (1988) *The Nematode Caenorhabditis elegans*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
30. 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.
31. Kiontke,K., Gavin,N.P., Raynes,Y., Roehrig,C., Piano,F. and Fitch,D.H. (2004) *Caenorhabditis* phylogeny predicts convergence of hermaphroditism and extensive intron loss. *Proc. Natl Acad. Sci. USA*, **101**, 9003–9008.
32. Cho,S., Jin,S.W., Cohen,A. and Ellis,R.E. (2004) A phylogeny of *Caenorhabditis* reveals frequent loss of introns during nematode evolution. *Genome Res*., **14**, 1207–1220.
33. Gkoutos,G.V., Green,E.C., Mallon,A.M., Hancock,J.M. and Davidson,D. (2005) Using ontologies to describe mouse phenotypes. *Genome Biol*., **6**, R8.