2012

WormBase 2012: More genomes, more data, new website

Karen Yook
California Institute of Technology

Tamberlyn Bieri
Washington University School of Medicine in St. Louis

Bill Nash
Washington University School of Medicine in St. Louis

Philip Ozersky
Washington University School of Medicine in St. Louis

John Spieth
Washington University School of Medicine in St. Louis

Published online 8 November 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D735–D741***
doi:10.1093/nar/gkr954

WormBase 2012: more genomes, more data, new website

Karen Yook[1,*], Todd W. Harris[2], Tamberlyn Bieri[3], Abigail Cabunoc[2], Juancarlos Chan[1], Wen J. Chen[1], Paul Davis[4], Norie de la Cruz[2], Adrian Duong[2], Ruihua Fang[1], Uma Ganesan[1], Christian Grove[1], Kevin Howe[4], Snehalata Kadam[1], Ranjana Kishore[1], Raymond Lee[1], Yuling Li[1], Hans-Michael Muller[1], Cecilia Nakamura[1], Bill Nash[3], Philip Ozersky[3], Michael Paulini[4], Daniela Raciti[1], Arun Rangarajan[1], Gary Schindelman[1], Xiaoqi Shi[2], Erich M. Schwarz[1], Mary Ann Tuli[5], Kimberly Van Auken[1], Daniel Wang[1], Xiaodong Wang[1], Gary Williams[4], Jonathan Hodgkin[6], Matthew Berriman[5], Richard Durbin[5], Paul Kersey[4], John Spieth[3], Lincoln Stein[2] and Paul W. Sternberg[1,7,*]

[1]California Institute of Technology, Division of Biology 156-29, Pasadena, CA 91125, USA, [2]Ontario Institute for Cancer Research, 101 College Street, Suite 800, Toronto, ON, Canada M5G0A, [3]The Genome Institute, Washington University, School of Medicine, St Louis, MO 63108, USA, [4]EMBL-European Bioinformatics Institute Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, [5]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, [6]Genetics Unit, Department of Biochemistry, University of Oxford, South Parks Road, Oxford, OX1 3QU, UK and [7]Howard Hughes Medical Institute, California Institute of Technology, Pasadena, CA 91125, USA

Received September 14, 2011; Accepted October 12, 2011

ABSTRACT

Since its release in 2000, WormBase (http://www.wormbase.org) has grown from a small resource focusing on a single species and serving a dedicated research community, to one now spanning 15 species essential to the broader biomedical and agricultural research fields. To enhance the rate of curation, we have automated the identification of key data in the scientific literature and use similar methodology for data extraction. To ease access to the data, we are collaborating with journals to link entities in research publications to their report pages at WormBase. To facilitate discovery, we have added new views of the data, integrated large-scale datasets and expanded descriptions of models for human disease. Finally, we have introduced a dramatic overhaul of the WormBase website for public beta testing. Designed to balance complexity and usability, the new site is species-agnostic, highly customizable, and interactive. Casual users and developers alike will be able to leverage the public RESTful application programming interface (API) to generate custom data mining solutions and extensions to the site. We report on the growth of our database and on our work in keeping pace with the growing demand for data, efforts to anticipate the requirements of users and new collaborations with the larger science community.

INTRODUCTION

Caenorhabditis elegans is a millimeter long, free-living, soil nematode used as a model organism for biology research for nearly four decades [(1,2); http://www.wormbook.org]. WormBase curates, stores and displays genomic and genetic data about nematodes with primary emphasis on *C. elegans* and related *Caenorhabditis* nematodes (3). WormBase started as a web-based interface for ACeDB, which was built to contain genetic and physical maps of *C. elegans* as well as the genome sequence itself (4–7). Now in its 11th year, WormBase has expanded to house numerous nematode genomes, experimental observations, reagents and literature. Over the past 2 years, we have enhanced our database by adding new graphics, more data types, more data overall and new curation tools to increase the efficiency of capturing and annotating all

*To whom correspondence should be addressed. Tel: +626 395 8325; Fax: +626 449 0756; Email: kyook@wormbase.org
Correspondence may also be addressed to Paul W. Sternberg. Tel: +626 395 2181; Fax: +626 568 8012; Email: pws@caltech.edu

these new data. We have continued to expand our outreach to other model organism databases (MODs) sharing insight and setting up tools for curation pipelines. We have also expanded, both in number and depth, our collaborations with other biological resources, leading to better synchronization of biological data across multiple online resources. Finally, as we have transitioned from a single species to a multi-species resource, we built a new website released as a public beta version in September 2011.

GENOMES

Reference genome

The *C. elegans* reference genome has been updated using data from the modENCODE (8) RNASeq data submission and verified by a private submission of high-throughput-sequencing data from Julie Ahringer and Matt Berriman (personal communication). This update has resulted in a net increase of the genome by 66 bp, with the correction of 151 loci and 100 gene models.

New genomes

Initially housing just the *C. elegans* genome, WormBase now has genomic sequences for seven *Caenorhabditis* species with the recent integration of *C. angaria* (9), and *C.* sp. 11. We also work with the research communities of a number of other nematodes of agricultural and medical interest, acting as a portal for the storage and display of their data. We currently provide data files and genome browsers for: *Brugia malayi* (10), *Pristionchus pacificus* (11), *Haemonchus contortus* (M. Berriman *et al.*, unpublished data), *Strongyloides ratti* (M. Berriman *et al.*, unpublished data), *Meloidogyne incognita* (12), *M. hapla* (13), *Ascaris suum* (Jex *et al.*, Draft Ascaris suum genome. Nature, in press) and *Trichinella spiralis* (14). Genomes expected in the near future include *Steinernema carpocapsae* (A. Dillman, manuscript in preparation) and *Heterorhabditis bacteriophora* (X. Bai *et al.*, manuscript in preparation). Groups wishing to submit new genomes to WormBase should consult http://wiki.wormbase.org/index.php/Genome_Standards.

Data availability and releases

With the ensuing flood of new genomic data, we have created a data warehouse with highly standardized file names, paths and contents for all species at WormBase as well as other nematode species of interest to the community. We anticipate this data warehouse will become a valuable clearinghouse in its own right, for both the *C. elegans* and broader nematode research community.

Due to the significant increase in data, new releases of WormBase now occur on a bi-monthly schedule and are available for download in various formats from the project FTP site (ftp://ftp.wormbase.org/pub/wormbase). A permanent archive of the database and website is created every fifth release and is available at a unique URL (http://ws225.wormbase.org). We encourage users to use and cite these referential releases.

New graphics

WormBase now incorporates images from a 3D virtual reconstruction of the anatomy of *C. elegans* (http://caltech.wormbase.org/virtualworm). The 3D model represents an adult hermaphroditic worm at cellular resolution and was manually constructed using the open-source 3D graphics software, Blender (version 2.49; http://www.blender.org). The model consists of 684 3D objects, representing 680 cells and 953 somatic nuclei, and is an initial draft version of a virtual *C. elegans*, depicting the morphology and spatial positioning of every cell, to the best of collective knowledge. Individual cell and tissue models have been created via interpolation/extrapolation of descriptions from WormAtlas (http://www.WormAtlas.org) and the '*C. elegans* Atlas' book (15), as well as from available micrographs (DIC or fluorescence) or other descriptors of anatomical structure. The Blender file allows the user to browse the virtual worm and learn more about the anatomy of *C. elegans*, for example by allowing users to select parts of the worm to display the names of individual cells and tissues. This Blender file also provides a variety of visualization options such as applying transparency, color, or hiding cells to make viewing easier. Video tutorials are available at http://caltech.wormbase.org/virtualworm/Instructional_Videos.html. Images from this file have been incorporated into both gene and expression pattern pages on the new WormBase website.

EXPANSION OF DATA

Active import and curation of new types of *C. elegans* data continues to be one of the primary activities in the maintenance and development of WormBase. The past 2 years have seen the incorporation of modENCODE (8) data along with other large-scale data sets; the development of a Worm Phenotype Ontology [WPO; (16)]; adaptation of Serial Patterns of Expression Levels Locator [SPELL; (17)] to house microarray data; and the incorporation of new data classes such as molecules, images and human disease connections. We discuss these data types below.

modENCODE data

modENCODE data was added to the primary *C. elegans* Genome Browser in June 2010; curators are using modENCODE data for sequence curation and have devised strategies to integrate these data into WormBase. modENCODE data sets include UTRome features, pseudogene curation targets, Highly Occupied Target (HOT) regions, polyA sites, ncRNA genes and aggregate coding gene models. These data sets have been subjected to rigorous internal quality control and fully integrated into the database.

Gene model curation

WormBase continues to maintain a manual gene curation program whereby gene structures are corrected in line with all currently available data for a given locus. This is managed and streamlined via the use of the Sequence Curation Tool (CT) an in-house developed software

suite [see below; (18)]. The integration of large data sets such as modENCODE has provided valuable extra evidence for gene model curation. RNASeq data from modENCODE has been used to discover anomalies that highlight potential cases where adjacent genes could be merged. Resolving these anomalies alone has so far resulted in the improvement of over 100 gene models.

Representation of miRNAs has been rationalized and extended so that there is now a clear distinction between mature miRNA products and primary transcripts. Integration of additional large datasets included polyA sites generated by a project not associated with modENCODE (19). Combining these with the modENCODE data has resulted in the assignment of polyA sites to >80% of coding genes. genBlastG (20) gene models for *C. briggsae*, *C. brenneri* and *C. remanei* have also been incorporated into the database. These gene models were computed by projection of *C. elegans* gene models, and have been helpful for the curation of these genomes.

Whole genome sequencing data

One of the key challenges faced by WormBase is the rapid growth of *C. elegans* strain variation data generated by Whole Genome Sequencing (WGS) projects. The strains from which these data sets are derived vary, ranging from wild isolates to laboratory-manipulated mutants. We continue to investigate and develop mechanisms for the efficient storage, processing and visualization of these data sets. The acknowledged canonical resource for the management and archiving of variation data is dbSNP (21). We strongly encourage projects to submit their data to dbSNP, and continue to act as a submission broker in cases where a laboratory lacks the technical resources to conform to the dbSNP submission protocols. While dbSNP acts as the primary repository for the data, WormBase adds curated and computationally derived value, for example putative gene consequence, and provides full cross-referencing back to the dbSNP primary records. To date, WGS data from six projects (one ongoing) have been integrated into WormBase and submitted to dbSNP [Andersen *et al.*, manuscript in preparation; Moerman and Waterston, manuscript in preparation; (22–25)] This amounts to a total of about 400 000 variations.

Worm phenotype ontology

We have continued to develop the WPO and have added 115 new phenotype terms this past year, bringing the total number of terms to 1985. New terms are added in parallel to the curation process, allowing us to remain up-to-date with the field. The WPO was published as a resource for the scientific community (16). Currently, the Biological General Repository for Interaction Datasets [BioGrid; http://thebiogrid.org; (26)] database is utilizing the WPO for the annotation of phenotypes associated with genetic interactions in *C. elegans*.

Microarray data

All *C. elegans* related microarray datasets from Gene Expression Omnibus [GEO; (27)] and ArrayExpress (28) have been imported into WormBase. Probe-centric microarray data are mapped to the latest version of the *C. elegans* genome for each WormBase release to generate gene-centric data, which are stored in a MySQL-based SPELL database [http://spell.caltech.edu:3000/; (17)]. These displays also include expression levels from RNAseq datasets.

Images

We are now extracting published images from expression pattern analyses and will expand this curation to include images of other data types. To make the process more efficient, effort has been devoted to automating image acquisition. To display published images, permission for each individual image has to be obtained from the publisher. To date, permission has been obtained from 27 major publishers and WormBase is negotiating with several others. We are also working on automating the process of requesting permission. Before this project began, 7228 images were directly submitted by a small number of laboratories engaged in large-scale projects. These images will be added to over 2000 images now extracted from the literature. Each image is manually curated and associated with a gene, anatomical structure and cellular component.

Molecules

Molecule curation captures small molecules and drugs that modify or cause phenotypes in a mutant background or RNAi-based experiments, and/or cause changes in gene-regulation activity. This data class has been populated with molecules from ChEBI (http://www.ebi.ac.uk/chebi/), the National Library of Medicine (http://www.nlm.nih.gov/mesh/MBrowser.html), the Comparative Toxicogenomic Database (CTD; http://ctd.mdibl.org/) and Small Molecule Metabolite (http://www.SMMID.org), which act as sources of IDs, names and synonyms for assigning molecule annotations to WB data. Over 600 molecule connections to gene and RNAi and variation phenotype objects have been created since the beginning of this data type curation.

Human disease gene orthologs

WormBase provides curated, concise descriptions of genes based on the reading of published literature. These are free-text and include information about gene orthology, function and expression. Since *C. elegans* is an important animal model that is increasingly used for the study of human disease, we write these gene descriptions with emphasis on the orthologies to human disease genes, and how their study in *C. elegans* has informed the disease field. This information will be highlighted with a special 'Human disease relevance' tag, for the benefit of both the *C. elegans* and non-*C. elegans* researcher. We plan to facilitate queries to serve as a portal through which one can access relevant information from the nematode field,

for example, a query using either a human gene name or disease name will lead the user to the relevant *C. elegans* gene.

INCREASING THE EFFICIENCY OF ANNOTATION AND CURATION

The need for efficient curation necessitates the development of customized curation tools. We have developed tools to improve the rate and accuracy of curation. In addition, we are actively developing automated and non-automated methods for identifying papers that contain relevant data for curation.

Improving sequence curation

To facilitate more accurate gene structure curation we recently developed the Sequence Curation Tool [CT; (18)]. The CT consists of three components: (i) a Perl based program that reads GFF files and identifies inconsistencies, or anomalies, between existing gene models and evidence such as the protein and transcript alignments with the genome, and other types of genomic features (e.g. repeat sequences); (ii) a MySQL database of these anomalies and information on which anomalies have been investigated previously; and (iii) a Perl/TK graphical user interface (GUI) for reading and displaying potential gene structure problems from the MySQL database and allowing the curator to select and edit regions of the genome that contain a high incidence of anomalies. There currently are 28 anomaly types that are identified by the CT including EST alignments not matching an exon, a frame-shifted protein alignment, weak splice sites and RNASeq alignment spanning a novel intron.

Cross-linking to orthology data provided by other groups continues to be improved and extended, and encompasses InParanoid7 (29), OMA (30), TreeFam (31), EnsEMBL-Compara (32), Panther (33) and eggNOG (34). The OMIM resource (35) has also been used to annotate worm genes orthologous to human genes associated with disease (see above).

Improving literature curation

To facilitate data extraction and curation from the literature we developed the Ontology Annotator (OA). The OA was inspired by and is similar to Phenote (http://phenote.org/), which was developed by Berkeley Bioinformatics Open-Source Projects (BBOP; http://berkeleybop.org/). The OA provides curation interfaces for a number of data types: phenotype, gene regulation, gene interactions, images, Gene Ontology (GO; http://www.geneontology.org/) and transgenes, among others. This tool offers the capabilities of Phenote, for example, the ability to annotate data using ontologies. In addition, it is web-based, providing easy access for curators, and allows entered data to be stored in a local database. These features allow curators to query and edit data whenever required, and to access data from other projects, that use the OA, as soon as they are entered into the local database.

Improving the identification of papers for curation

Identifying papers containing specific data types is a major effort for any literature curation database. Over the past few years we have investigated and incorporated various methods of automated data type identification, ranging from computational methods such as relatively simple string searching algorithms, to statistical machine learning methods such as hidden Markov models (HMM) (H-M. Muller, personal communication) or Support Vector Machines [SVMs; (36)], to author participation via a web form.

Automated methods are currently used to identify over 25 data types (http://www.wormbase.org/wiki/index.php/Curated_data_types). Nine of these data types, including alleles, RNAi experiments, transgenes and images, are identified automatically using either pattern matching or matches to category lexica through use of the text mining system, Textpresso [http://www.texptresso.org; (37)]. In addition to identifying the data type, Textpresso is employed for extracting information for gene interactions, GO cellular component annotation (38), transgenes, physical interactions and images.

A second automated method using an SVM algorithm is employed to flag papers containing data types such as antibody, molecular lesions, corrections to gene structures, gene regulation, gene expression patterns, gene product interactions, gene–gene interactions, RNAi and allele-based phenotypes, and phenotypes due to the overexpression of a gene. While SVM has proved very useful for identifying some data types, such as GO cellular component, other data types, such as gene expression, are not as successfully flagged by this algorithm and will need more work to be detected by automated identification (Fang *et al.*, manuscript in preparation).

Author participation

For the past 3 years, we have reached out to authors to ask for help in flagging their papers for the presence of specific data types. Authors are contacted via an e-mail that contains a link to a data declaration form that asks them to indicate the types of information their paper contains and to provide details. When the form is submitted, curators at WormBase receive an e-mail alert depending on the data type declared by the author. We have had a 40% ($n = 2355$) feedback rate through this pipeline over the last 2 years. This flagging pipeline has served as a useful safety net for capturing papers that have been missed through other flagging mechanisms.

OUTREACH

Extending automated pipelines to other model organism databases

Motivated by our success in employing an SVM-based flagging pipeline for certain WormBase data types, we extended this effort to FlyBase (http://flybase.org/) and *Saccharomyces* Genome Database (SGD; http://www.yeastgenome.org/) to achieve the same automated flagging goals (Fang *et al.*, manuscript in preparation). We set up

an SVM flagging pipeline for a number of relevant data types curated by FlyBase curators, with promising results. During the course of setting up these pipelines we found that training papers from different species for similar data types can be used together to significantly improve the performance of SVM for identifying papers for a single organism. Specifically, we found that the addition of WormBase RNAi training papers to the RNAi training set of FlyBase increased the recall of known positive papers while the precision in identifying new positive papers remained constant for the SVM analysis.

Extending curation tools to other model organism databases

At the request of The Arabidopsis Information Resource (TAIR; http://www.arabidopsis.org/), we modified and implemented our semi-automated, Textpresso-based GO Cellular Component Curation (CCC) pipeline (34) for *Arabidopsis* by creating a curation pipeline and interface for TAIR curators. Among changes we implemented for TAIR, the most important were: (i) additions to the cellular component category to include plant-specific terms, and (ii) the addition of filtering steps to avoid examining text mining results from previously curated papers. An extension of our semi-automated GO CCC pipeline is also being modified and implemented for dictyBase, which includes helping to establish semi-automated paper acquisition for dictyBase.

COLLABORATION

Ensembl genomes

WormBase has recently formalized its partnership with the Ensembl Genomes project (http://www.ensembl.org/) at the European Bioinformatics Institute (this issue). Ensembl Genomes aims to work with communities interested in non-vertebrate species to develop genome-oriented resources. WormBase will explore opportunities for exploiting technologies developed in Ensembl Genomes in the context of other genome projects, at the same time contributing to their development.

BioGRID

In August of 2010, we began a collaboration with the BioGRID Interaction Database (26) to exchange physical and genetic interaction data for *C. elegans*. Previous physical interaction curation at WormBase consisted of data from several large-scale yeast one- and two-hybrid assays and annotation performed in the context of GO Molecular Function curation. As a result of this collaboration, we hope to begin adding all protein–protein interactions to WormBase. These data will be displayed on the respective gene pages in WormBase along with a link to the corresponding interaction page at BioGRID.

Genetics Society of America

We are collaborating with the Genetics Society of America (GSA; http://www.genetics-gsa.org/) to identify nematode-specific biological entities, e.g. gene names, alleles, anatomy terms, etc., within published GENETICS papers, and to convert these entities into embedded direct links to WormBase (39). Entities from over 10 data classes are marked up and linked back to WormBase. This project pioneered the development of a markup pipeline to link GSA articles to MODs; SGD and FlyBase are now using this method for their respective GSA papers. As part of the markup pipeline, we ensure that the links are unambiguous by employing critical, curator-based quality control (QC), a step that is lacking in many automated text markup tools. We have made significant progress in making the QC step time-efficient by using automated scripts, employing online tools that scan for erroneous and uninformative links, and soliciting authors' help in identifying entities that are not yet part of our database.

WEBSITE OVERHAUL

New website

To accommodate the increasing demands on the resource and the diversifying needs of the user community, the WormBase website application has been entirely re-designed. A beta version of the new website (http://beta.wormbase.org/) was released in September of 2011. While WormBase is not a wiki-based database, community participation is encouraged; the new site employs a number of novel features to capture community input. For example, in-line and ubiquitous submission forms atomized to pages allow users to easily report issues pertaining to annotations and see when curators act upon those issues. Public or private comments can be left on any entity in the database as a light-weight, low participation-barrier community annotation system. We plan to use this system to more easily collect and incorporate community-submitted annotations, a task particularly important for species that lack extensive curation. Finally, social media features aim to discover additional patterns in the data; anonymous aggregate browsing history is being used to develop an Amazon-style suggestion system to present possibly related entities when users are browsing the site. A powerful and extensive API using the RESTful design pattern makes every piece of data in WormBase addressable at unique URIs; data miners and developers will be able to leverage this interface for querying the resource or easily embedding WormBase data in third party websites.

FUTURE DIRECTIONS

Having successfully transitioned from a single-species resource to one that begins to represent the diversity of the nematode phylogeny, we are now providing a database service to a much broader audience. To accommodate our current and new audiences, one future enhancement to the

site will be the creation of new web pages that aim to display comprehensive views of the biology of nematodes. These pages will complement our current gene-centric view of the data by using complex queries and data calls to synthesize pages that pull together information from the database related to a defined biological process. In addition to these enhanced views of the data, we will be expanding the 3D *C. elegans* anatomical model. The model will be more fully incorporated into WormBase, enabling WormBase users to visually navigate the adult *C. elegans* anatomy from the web browser as well as access and extract key pieces of information relevant to the anatomy object in question. We also plan to construct and integrate models for the adult male as well as the four larval stages. With the ongoing enhancements to the database and the constant growth in data, we will be continuing to refine and extend our new web architecture in anticipation of the demands for access to these data.

FUNDING

This work is supported by the US National Institutes of Health (Grant no. P41 HG02223); US National Human Genome Research Institute (Grant no. P41-HG02223) to WormBase; and British Medical Research Council (Grant no. G070119) 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 no. P41-HG02223).

Conflict of interest statement. None declared.

REFERENCES

1. Brenner,S. (1974) The genetics of *Caenorhabditis elegans*. *Genetics*, **77**, 71–94.
2. 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.
3. Harris,T.W., Antoshechkin,I., Bieri,T., Blasiar,D., Chan,J., Chen,W.J., La Cruz,N., De, Davis,P., Duesbury,M., Fang,R. *et al*. (2010) WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res*., **38**, D463–D467.
4. Eeckman,F.H. and Durbin,R. (1995) Caenorhabditis elegans: Modern Biological Analysis of an Organism. *Methods in Cell Biol*., **48**, 583–605.
5. 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.
6. *C. elegans* Genome Sequencing Consortium. (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, **282**, 2012–2018.
7. 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.
8. Gerstein,M.B., Lu,Z.J., Nostrand,E.L., Van, Cheng,C., Arshinoff,B.I., Liu,T., Yip,K.Y., Robilotto,R., Rechtsteiner,A., Ikegami,K. *et al*. (2010) Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science*, **330**, 1775–1787.
9. Mortazavi,A., Schwarz,E.M., Williams,B., Schaeffer,L., Antoshechkin,I., Wold,B.J. and Sternberg,P.W. (2010) Scaffolding a Caenorhabditis nematode genome with RNA-seq. *Genome Res*., **20**, 1740–1747.
10. 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.
11. 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.
12. 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. Biotech*., **26**, 909–915.
13. 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.
14. Mitreva,M., Jasmer,D.P., Zarlenga,D.S., Wang,Z., Abubucker,S., Martin,J., Taylor,C.M., Yin,Y., Fulton,L., Minx,P. *et al*. (2011) The draft genome of the parasitic nematode *Trichinella spiralis*. *Nat. Genet*., **43**, 228–235.
15. Hall,D. and Altun,Z. (2008) *C. elegans atlas*. Cold Spring Harbor Laboratory Press, USA.
16. Schindelman,G., Fernandes,J.S., Bastiani,C.A., Yook,K. and Sternberg,P.W. (2011) Worm Phenotype Ontology: Integrating phenotype data within and beyond the *C. elegans* community. *BMC Bioinform*., **12**, 32.
17. Hibbs,M.A., Hess,D.C., Myers,C.L., Huttenhower,C., Li,K. and Troyanskaya,O.G. (2007) Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics*, **23**, 2692–2699.
18. Williams,G.W., Davis,P.A., Rogers,A.S., Bieri,T., Ozersky,P. and Spieth,J. (2011) Methods and strategies for gene structure curation in WormBase. *Database*, **2011**, baq039.
19. Jan,C.H., Friedman,R.C., Ruby,J.G. and Bartel,D.P. (2011) Formation, regulation and evolution of *Caenorhabditis elegans* 3'UTRs. *Nature*, **469**, 97–101.
20. She,R., Shih-Chieh Chu,J., Uyar,B., Wang,J., Wang,K. and Chen,N. (2011) genBlastG: extending BLAST to be a high performance gene finder. *Bioinformatics*, **27**, 2141–2143.
21. Sherry,S.T. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*., **29.**, 308–311.
22. Sarin,S., Bertrand,V., Bigelow,H., Boyanov,A., Doitsidou,M., Poole,R.J., Narula,S. and Hobert,O. (2010) Analysis of multiple ethyl methanesulfonate-mutagenized *Caenorhabditis elegans* strains by whole-genome sequencing. *Genetics*, **185**, 417–430.
23. Flibotte,S., Edgley,M.L., Chaudhry,I., Taylor,J., Neil,S.E., Rogula,A., Zapf,R., Hirst,M., Butterfield,Y., Jones,S.J. *et al*. (2010) Whole-genome profiling of mutagenesis in *Caenorhabditis elegans*. *Genetics*, **185**, 431–441.
24. Zuryn,S., Gras,S., Le, Jamet,K. and Jarriault,S. (2010) A strategy for direct mapping and identification of mutations by whole-genome sequencing. *Genetics*, **186**, 427–430.
25. Grishkevich,V., Hashimshony,T. and Yanai,I. (2011) Core promoter T-blocks correlate with gene expression levels in *C. elegans*. *Genome Res*., **21**, 707–717.
26. Stark,C., Breitkreutz,B.-J., Chatr-Aryamontri,A., Boucher,L., Oughtred,R., Livstone,M.S., Nixon,J., Auken,K., Van, Wang,X., Shi,X. *et al*. (2011) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res*., **39**, D698–D704.
27. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M. *et al*. (2011) NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic Acids Res*., **39**, D1005–D1010.
28. Parkinson,H., Sarkans,U., Kolesnikov,N., Abeygunawardena,N., Burdett,T., Dylag,M., Emam,I., Farne,A., Hastings,E., Holloway,E. *et al*. (2011) ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res*., **39**, D1002–D1004.
29. Ostlund,G., Schmitt,T., Forslund,K., Köstler,T., Messina,D.N., Roopra,S., Frings,O. and Sonnhammer,E.L.L. (2010) InParanoid

7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.*, **38**, D196–D203.
30. Altenhoff,A.M., Schneider,A., Gonnet,G.H. and Dessimoz,C. (2011) OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res.*, **39**, D289–D294.
31. Li,H., Coghlan,A., Ruan,J., Coin,L.J., Heriche,J.K., 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.
32. 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.
33. Mi,H., Guo,N., Kejariwal,A. and Thomas,P.D. (2007) PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic Acids Res.*, **35**, D247–D252.
34. Muller,J., Szklarczyk,D., Julien,P., Letunic,I., Roth,A., Kuhn,M., Powell,S., Mering,C., von, Doerks,T., Jensen,L.J. *et al.* (2010) eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res.*, **38**, D190–D195.
35. Amberger,J., Bocchini,C.A., Scott,A.F. and Hamosh,A. (2009) McKusick's online Mendelian inheritance in man (OMIM). *Nucleic Acids Res.*, **37**, D793–D796.
36. Chen,D., Müller,H.-M. and Sternberg,P.W. (2006) Automatic document classification of biological literature. *BMC Bioinform.*, **7**, 370.
37. 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.
38. 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.
39. Rangarajan,A., Schedl,T., Yook,K., Chan,J., Haenel,S., Otis,L., Faelten,S., De Pellegrin-Connelly,T., Isaacson,R., Skrzypek,M.S. *et al.* (2011) Toward an interactive article: integrating journals and biological databases. *BMC Bioinform.*, **12**, 175.