82–86 Nucleic Acids Research, 2001, Vol. 29, No. 1

WormBase: network access to the genome and biology of *Caenorhabditis elegans*

Lincoln Stein*, Paul Sternberg[1], Richard Durbin[2], Jean Thierry-Mieg[3] and John Spieth[4]

Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA, [1]Howard Hughes Medical Institute and California Institute of Technology, Pasadena, CA, USA, [2]The Sanger Centre, Hinxton, UK, [3]National Center for Biotechnology Information, Bethesda, MD, USA and [4]Genome Sequencing Center, Washington University, St Louis, MO, USA

Received September 18, 2000; Accepted October 4, 2000

ABSTRACT

WormBase (http://www.wormbase.org) is a web-based resource for the *Caenorhabditis elegans* genome and its biology. It builds upon the existing ACeDB database of the *C.elegans* genome by providing data curation services, a significantly expanded range of subject areas and a user-friendly front end.

DESCRIPTION

Caenorhabditis elegans (informally known as 'the worm') is a small, soil-dwelling nematode that is widely used as a model system for studies of metazoan biology (1). *Caenorhabditis elegans*' popularity results from the confluence of several factors: its developmental program is understood at the single-cell level (2,3), its complete genome is known (4), and it is highly amenable to genetic manipulation, including RNA inhibition (RNAi) intervention (5).

WormBase is a collaborative effort to capture, curate and distribute information about *C.elegans* biology. It is an outgrowth of ACeDB (http://www.acedb.org), the database used in the course of the *C.elegans* sequencing project to coordinate the sequencing effort and to integrate the worm sequence with the genetic and physical maps. Unlike the original ACeDB project, however, WormBase is heavily committed to the curation and interpretation of the *C.elegans* literature, and has moved from a genome-centric perspective to one that more evenly balances the worm genome with other aspects of its biology.

WormBase, like its predecessor, uses a high-level, object-oriented framework to organize and present *C.elegans* information. The researcher navigates through a series of biologically meaningful object classes such as Locus, Sequence, Cell and Paper. Each of these object classes is associated with at least one WormBase web page that is specialized for its display, and many classes have multiple alternative representations designed to meet different research needs. For example, if a researcher is reviewing a page of information on a particular genetically-mapped locus, he can easily switch between a text display of the mutant phenotype, strains and alleles, a graphical form that shows the position of the locus on the genetic map and a page that shows the locus in the context of the physical map.

WormBase uses HTML linking to represent the relationships between objects. For example, a segment of genomic sequence object is linked with the several predicted gene objects contained within it, and each predicted gene is linked to its conceptual protein translation.

The major components of the resource are described below.

The *C.elegans* genome

WormBase contains the 'essentially complete' genome of *C.elegans*, which now stands at 99.3 Mb of finished DNA interrupted by approximately 25 small gaps. In addition, the resource contains a reference set of predicted and confirmed genes curated by the WormBase staff, conceptual translations, alternative splicing patterns inferred from EST overlaps, and the underlying raw data used to make these determinations. WormBase also maintains a regularly updated list of DNA and protein similarity matches between the *C.elegans* genome and sequences from other species.

The biologist can gain access to the genome in a number of ways: he can search the genome by specifying the name of a well-known marker, such as the name of a clone, a predicted gene or a genetic locus. The researcher may also enter the genome via a BLAST search. WormBase has several alternative displays of genomic information, including a purely graphical display, a mixture of graphics and HTML (Fig. 1), and a purely tabular representation. The displays are designed to minimize the amount of extraneous information displayed to the user. For example, the display for a predicted gene does not, by default, show the gene's DNA or conceptual translation. However, these data are easily accessed with a single click on a 'pop down' icon.

Cells and their lineages

WormBase contains the complete lineages for the male and hermaphrodite organisms and information that describes each cell and its primary biological function. Biologists can search for cells by their standard nomenclature, or by way of a pedigree browser that displays the complete lineage in the manner of an expandable outline in a word processor (Fig. 2). The lineage display provides flexible search functionality; for example, researchers can use the lineage browser to identify all cells

*To whom correspondence should be addressed. Tel: +1 516 367 8380; Fax: +1 516 367 8389; Email: lstein@cshl.org

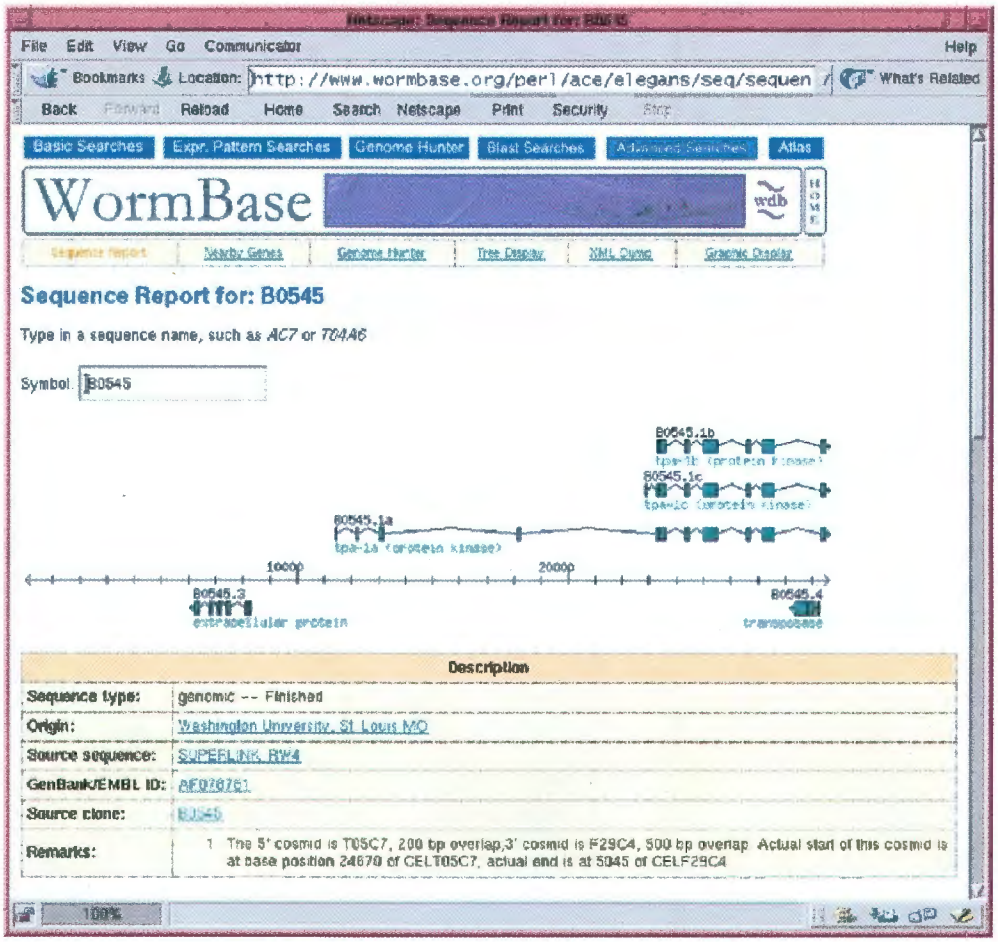

Figure 1. The sequence browser displays alternative splicing patterns predicted by EST overlaps.

involved in the genesis of the amphid organ. From there they can jump to a page that gives information on the function and spatial location of each cell, and move onward from there to a summary of the genes that are known to be specifically expressed in the amphid organ.

The default display for each cell summarizes its topological position in the worm's anatomy as well as its temporal position in the lineage and provides information on cell function and interactions. For approximately a third of the cells in the adult, WormBase displays schematics that show their position in an idealized worm.

Genetic maps

A series of pages give researchers access to up to date genetic maps, and list known mutants, alleles and phenotypes. Links to the *C.elegans* Genetics Center allow users to order strains and to submit new genetic mapping information. WormBase provides two versions of the graphical genetic map. The first version is a static HTML image map that will work correctly on all web browsers. The second is an interactive Java applet that provides scrolling and zooming functionality, but requires a plug-in to work properly on some browsers.

Another tool available in WormBase allows the user to move between the genetic map and the genome sequence, providing the means to associate genetically-mapped loci with predicted genes. If the researcher wishes, he can drill down to the raw genetic mapping data, and see the genotypes of individuals produced by mapping crosses.

Expression patterns

WormBase contains the results from reporter gene fusion experiments performed by a number of groups, chiefly those of Ian Hope (6). These experiments relate cells, organs and stages of the life cycle with the expression of tagged genes. The user interface (Fig. 3) allows researchers to search for genes that are expressed in arbitrary combinations of organs, cells and life stages, and displays micrographic data as well as text summaries of the findings.

Bibliographic references

WormBase contains an extensive bibliography of papers published in *C.elegans* biology going back to the mid-1970s, as well as unpublished abstracts from the biannual Worm Meetings and brief reports contributed to the Worm Breeder's

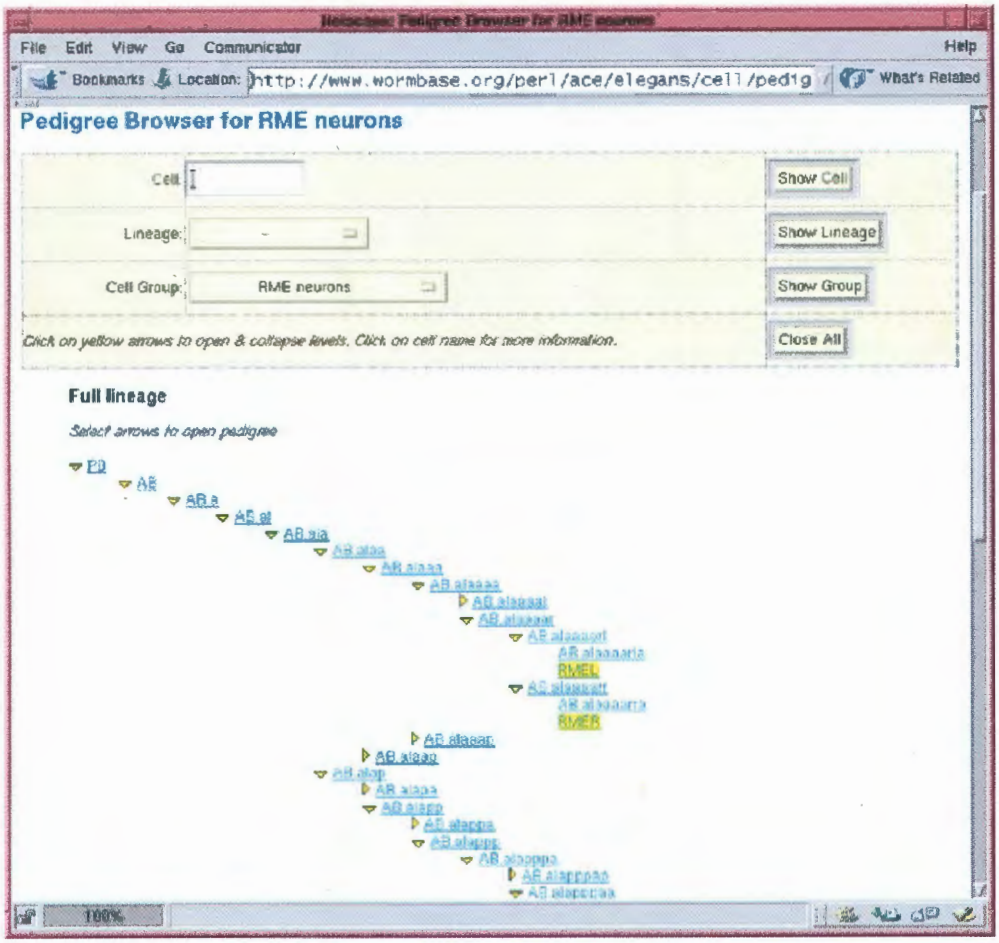

Figure 2. The lineage browser allows users to selectively display portions of the *C.elegans* cell lineage.

Gazette newsletter. Each paper is cross-referenced with the loci, sequences and cells that it refers to, as well as to the authors, their affiliations and contact information.

External links

WormBase is extensively linked to other sources of information, including GenBank/EMBL, SWISS-PROT, and the Worm Proteome Database (WPD). In the future, as we incorporate more specialized data sets such as RNA inhibition studies and microarray experiments, we will be adding links to sites that provide experimental data and further information.

DATA ACCESS

WormBase is designed to meet the needs of (i) casual users from outside the *C.elegans* community whose interaction with the resource will be fleeting, (ii) *C.elegans* biologists who will use the resource more intensively and in a more sophisticated fashion, and (iii) bioinformatics specialists who have specialized requirements.

Interactive access to the database is available at its primary web site, http://www.wormbase.org, as well as at its mirror site at the Sanger Centre, http://wormbase.sanger.ac.uk. The database can be browsed by casual users using a simple keyword search interface that appears on the home page. With this interface biologists can find database objects based on their names, accession numbers or any text that appears within the object itself.

Caenorhabditis elegans biologists and other researchers with more directed questions to ask will more typically use one of the specialized search pages, such as the lineage browser, the expression pattern search page or BLAST search. There is also a simple interface that allows for the direct retrieval of an object usings its class name and accession number.

For more advanced users we provide direct access to the database in a number of ways. Each object in the WormBase database has a representation in XML (eXtensible Markup Language), a standard data syntax used for transmitting information across the World Wide Web. By linking to the WormBase web site using a published interface, researchers can obtain XML representations of predicted genes, genetic maps, cell lineages, expression patterns, or indeed any of the objects in WormBase. These XML representations are easily parsed, and in fact form the core of the Distributed Annotation

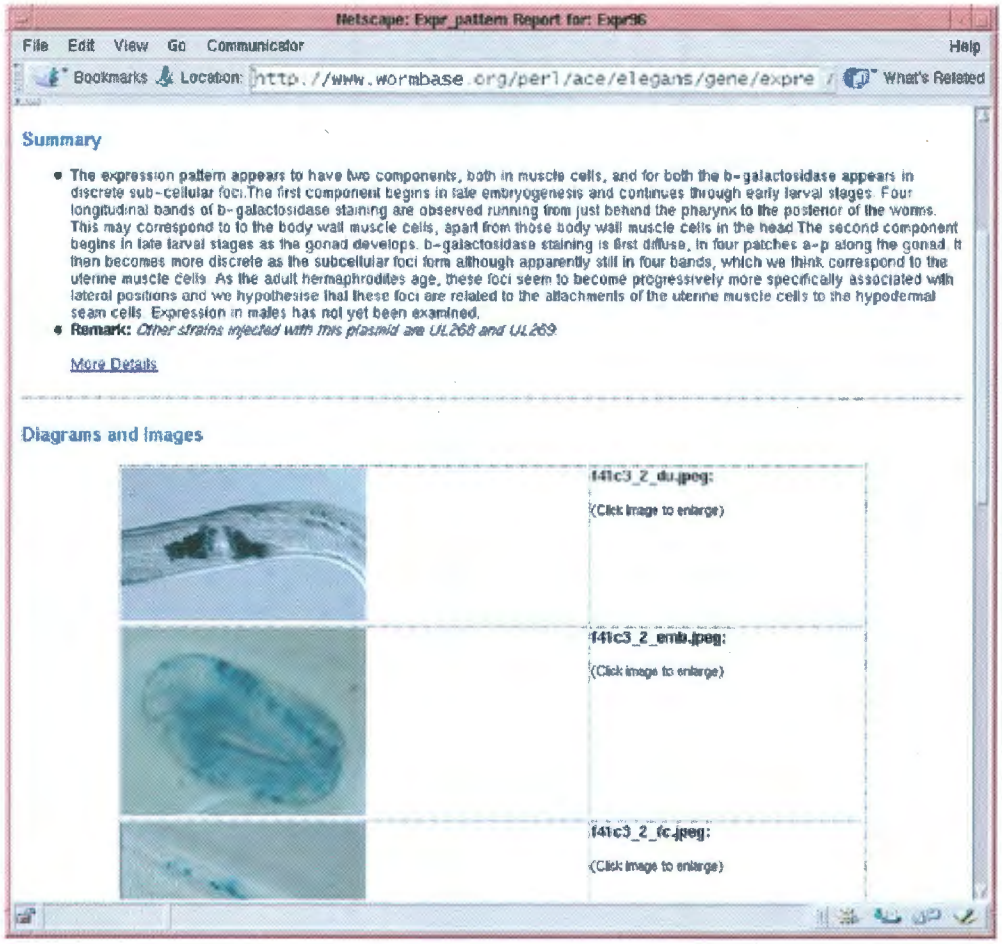

Figure 3. The expression pattern search page allows users to search for and retrieve the results of reporter gene studies.

System, a system developed by members of the WormBase group for exchanging genome annotations with other research groups (http://das.wustl.edu).

Another method of access is via a search page that accepts AQL (Acedb Query Language) queries (http://www.sanger.ac.uk/Software/Acedb/whelp/AQL/). This SQL-like language provides knowledgeable researchers with the ability to pose *ad hoc* queries against the WormBase database and integrate the results in various ways. This querying facility can also be used remotely via programmers' APIs (application programming interfaces) written in Perl and Java (7).

Finally, the WormBase database is available for bulk download, either in the form of various flat files, or as ACeDB format files and supporting software. This latter feature allows researchers to set up local mirrors of WormBase and to browse the data using the stand-alone ACeDB graphical user interfaces that run on Microsoft Windows and UNIX systems.

WormBase data is available to the public without license restrictions. The software is available under the GNU Public Licensing (GPL).

FUTURE DIRECTIONS

WormBase is very much a work in progress. Over the next year, we plan to incorporate the following features into the resource.

Improved curation of transcription splice patterns

A large number of alternative splicing patterns have been identified by the examination of *C.elegans* ESTs (8,9). We will be incorporating these data into WormBase along with an improved viewer that allows researchers to view the underlying data that supports the predicted intron/exon junctions and splice patterns.

RNA inhibition experiments

We are currently curating the results of several large-scale RNAi studies. These data will be searchable by phenotype, the genomic region affected and life stage affected. The raw data from the experiments, which consist of movie clips and static images, will be available either directly or indirectly via linking.

Gene ontology descriptions

We plan to attach Gene Ontology (GO) (10) terms to each of the predicted genes in the WormBase, allowing the database to be searched using a GO browser and linked to other databases that are indexed in this way.

Worm atlas

We have begun to assemble a photographic atlas of *C.elegans* using high magnification Nomarski images. Each image is entered into the database, and indexed in such a way that it can be used to reconstruct a view of any desired anatomic region in the organism. The researcher can navigate within the image, or zoom in on a feature of interest. Our plan is to index the location of each major cell in the atlas and to use this as a navigation tool. Researchers will be able to click on a cell of interest to learn about the role of the cell, what genes are known to be expressed within it, and so forth. Conversely, the atlas will be used to indicate the positions of cells that satisfy a query, for example cells that express a particular gene or genes.

ACKNOWLEDGEMENTS

This work was supported in part by NIH grant P41HG02223, as well as funding from the UK Medical Research Council and the National Library of Medicine.

REFERENCES

1. Riddle,D.L., Blumenthal,T., Meyer,B.J. and Preiss,J.R. (1997) *C. elegans II.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
2. Kimble,J. and Hirsh,D. (1979) Post-embryonic cell lineages of the hermaphrodite and male gonads in *Caenorhabditis elegans. Dev. Biol.*, **70**, 396–417.
3. Sulston,J.E., Schierenberg,E., White,J.G. and Thomson,J.N. (1983) The embryonic cell lineage of the nematode *Caenorhabditis elegans. Dev. Biol.*, **100**, 64–119.
4. The *C.elegans* Sequencing Consortium (1998) Genome sequence of the Nematode *C. elegans*: platform for investigating biology. *Science*, **282**, 2012–2018.
5. Fire,A., Xu,S.Q., Montgomery,M.K., Kostas,S.A., Driver,S.E. and Mello,C.C. (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans. Nature*, **391**, 806–811.
6. Hope,I.A., Arnold,J.M., McCarroll,D., Jun,G., Krupa,A.P. and Herbert,R. (1998) Promoter trapping identifies real genes in *C. elegans. Mol. Gen. Genet.*, **260**, 300–308.
7. 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.
8. Kent,W.J. and Zahler,A.M. (2000) The intronerator: exploring introns and alternative splicing in *Caenorhabditis elegans. Nucleic Acids Res.*, **28**, 91–93.
9. Kohara,Y. (1996). Large scale analysis of *C. elegans* cDNA. *Tanpakushitsu Kakusan Koso*, **41**, 715–720.
10. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T., Harris,M.A., Hill,D.P., Issel-Tarver,L., Kasarskis,A., Lewis,S., Matese,J.C., Richardson,J.E., Ringwald,M., Rubin,G.M. and Sherlock,G. (2000) Gene Ontology: Tool for the Unification of Biology. *Nature Genet.*, **25**, 25–29.

Nucleic Acids Research, 2001, Vol. 29, No. 4

CORRIGENDUM

WormBase: network access to the genome and biology of *Caenorhabditis elegans*

Nucleic Acids Res. (2001) **29**, 82–86.

The authors wish to note that, due to an error during manuscript preparation, the list of authors in the above article was incorrectly stated. The correct and full list of authors is:

L. Stein, M. Mangone, E. Schwarz, R. Durbin, J. Thierry-Mieg, J. Spieth and P. Sternberg