The versatile worm: genetic and genomic resources for *Caenorhabditis elegans* research

Igor Antoshechkin and Paul W. Sternberg*‡*

Abstract | Since its establishment as a model organism, *Caenorhabditis elegans* has been an invaluable tool for biological research. An immense spectrum of questions can be addressed using this small nematode, making it one of the most versatile and exciting model organisms. Although the many tools and resources developed by the *C. elegans* community greatly facilitate new discoveries, they can also overwhelm newcomers to the field. This Review aims to familiarize new worm researchers with the main resources, and help them to select the tools that are best suited for their needs. We also hope that it will be helpful in identifying new research opportunities and will promote the development of additional resources.

Since its initial use for studying nervous system development and function, the nematode *Caenorhabditis elegans* has rapidly become one of the most widely used model organisms, leading to seminal discoveries in diverse fields from development, signal transduction, cell death and ageing to RNAi. *Caenorhabditis elegans* was also the first multicellular organism to have its genome sequenced, propelling it to the forefront of genomic research[1], and the recent release of the full genome sequence of a related nematode, *Caenorhabditis briggsae*, has made *C. elegans* a leading system for comparative genomics[2]. The amazing success of *C. elegans* research is to a large extent due to its biological simplicity and the advantages it offers in terms of its transparent body, invariable cell number, small genome size, rapid life cycle, mode of reproduction and ease of maintenance. However, another factor that has made an important contribution to the success of this model system is the culture of close collaboration that exists between worm researchers, producing an extensive palette of tools that have historically been freely shared.

The array of resources that are available to *C. elegans* researchers has the potential to bewilder newcomers to the field. With this in mind, this Review provides a guide to the main resources that have been specifically developed for worm research. We first review online databases and tools, followed by a discussion of the most widely used experimental and material resources (BOXES 1,2). Finally, we outline some rapidly evolving areas of *C. elegans* research that offer new opportunities and require the development of new tools and approaches, with the hope of stimulating such expansion by new and established worm researchers alike. Many other tools that are not worm-specific can also be useful to *C. elegans* researchers; however, these fall outside the scope of this article and we provide only a partial list of links and brief descriptions to encourage the reader to explore these resources (TABLE 1).

Information resources

General information. The book *The Nematode Caenorhabditis elegans*' and its successor, *C. elegans II*[3,4], are invaluable guides for both new and established *C. elegans* biologists. They provide information on a range of topics, including the history of the organism as a model system, its anatomy, genome organization and specific biological processes such as sex determination, ageing and apoptosis. Since the completion of *C. elegans II* in 1997, *C. elegans* research has expanded significantly, making it necessary to compile a new version. The online resource WormBook aims to provide comprehensive reviews on all aspects of *C. elegans* biology and up-to-date descriptions of technical procedures[5]. Unlike the two previous publications, WormBook is continually updated, with cross references to related databases and tools. It currently comprises more than 130 chapters, and this number is expected to increase significantly. WormBook is also tightly integrated with WormBase and WormAtlas (see below) through reciprocal links and a shared community forum portal.

*Division of Biology 156-29, California Institute of Technology, 1200 East California Boulevard, Pasadena, California 91125, USA.
‡Howard Hughes Medical Institute, California Institute of Technology.
Correspondence to I.A.
e-mail: igor.antoshechkin@caltech.edu

doi:10.1038/nrg2105
Published online 5 June 2007

Box 1 | *Caenorhabditis elegans* online information resources

C. elegans II	http://www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowSection&rid=ce2
WormBook	http://www.wormbook.org
WormClassroom	http://www.wormclassroom.org
WormAtlas	http://www.wormatlas.org
WormImage	http://www.wormimage.org
WormBase	http://www.wormbase.org
WormGenes	http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/index.html?worm
RNAiDB	http://www.rnai.org
PhenoBank	http://www.worm.mpi-cbg.de/phenobank2/cgi-bin/MenuPage.py
C. elegans RNAi Phenome Database	http://omicspace.riken.jp/Ce/rnai/jsp/index.jsp
InteractomeDB	http://vidal.dfci.harvard.edu/interactomedb
n-Browse	http://nematoda.bio.nyu.edu:8080/NBrowse/N-Browse.jsp
WormBase interaction browser	http://www.wormbase.org/db/seq/interaction_viewer
IntAct	http://www.ebi.ac.uk/intact/index.jsp
BioGrid	http://www.thebiogrid.org
EDGEdb	http://edgedb.umassmed.edu
NEXTDB	http://nematode.lab.nig.ac.jp/index.html
Hope Laboratory Expression Pattern Database	http://bgypc059.leeds.ac.uk/~web/databaseintro.htm
BC *C. elegans* Gene Expression Consortium	http://elegans.bcgsc.ca/perl/eprofile/index
Stanford Microarray Database	http://smd.stanford.edu/index.shtml
NCBI Gene Expression Omnibus	http://www.ncbi.nlm.nih.gov/projects/geo
EBI ArrayExpress	http://www.ebi.ac.uk/arrayexpress
C. elegans SAGE libraries	http://tock.bcgsc.bc.ca/cgi-bin/sage170
Structural Genomics of *C. elegans*	http://sgce.cbse.uab.edu/index.php
Protein Data Bank	http://www.pdb.org
NCBI PubMed	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed
Textpresso	http://www.textpresso.org
WashU GSC	http://genome.wustl.edu/genome_group.cgi?GROUP=6
WashU GSC Blast server	http://genome.wustl.edu/tools/blast
Pristionchus pacificus	http://www.pristionchus.org
Parasitic Nematode Sequencing Project	http://www.nematode.net
C. elegans WWW server	http://elegans.swmed.edu
WormBase Wiki	http://www.wormbase.org/wiki
Worm Community Forums	http://www.wormbase.org/forums
Entrez Gene	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
MapViewer	http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=6239
Ensembl	http://www.ensembl.org/Caenorhabditis_elegans/index.html
UCSC Genome Browser	http://genome.ucsc.edu/cgi-bin/hgGateway?hgsid=85966288&clade=worm&org=0&db=0
UniProt	http://www.pir.uniprot.org
InParanoid	http://inparanoid.sbc.su.se
OrthoMCL	http://orthomcl.cbil.upenn.edu
TreeFam	http://www.treefam.org
Reactome	http://www.reactome.org/cgi-bin/frontpage?DB=gk_current&FOCUS_SPECIES=Caenorhabditis+elegans&.cgifields=MULTISPECIES
MetaCyc	http://metacyc.org
KEGG	http://www.genome.jp/kegg/pathway.html
aMAZE	http://www.scmbb.ulb.ac.be/amaze
Gene Ontology	http://www.geneontology.org/index.shtml
PFam	http://www.sanger.ac.uk/Software/Pfam
RFam	http://www.sanger.ac.uk/Software/Rfam
miRBase	http://microrna.sanger.ac.uk
NCBI Tools for Data Mining	http://www.ncbi.nlm.nih.gov/Tools
Toolbox at the EBI	http://www.ebi.ac.uk/Tools
UniProt Tools	http://www.pir.uniprot.org/search/tools.shtml
ExPASy Proteomics Tools	http://www.expasy.org/tools

BC, British Columbia; CGC, *Caenorhabditis* Genetics Center; EBI, European Bioinformatics Institute; EDGE, *C. elegans* Differential Gene Expression; ExPASy, Expert Protein Analysis System; KEGG, Kyoto Encyclopaedia of Genes and Genomes; NEXTDB, Nematode Expression Pattern DataBase; NCBI, National Center for Biotechnology Information (USA); SAGE, serial analysis of gene expression; WashU, Washington University; GSC, Genome Sequencing Center.

Box 2 | ***Caenorhabditis elegans* online experimental resources**

Caenorhabditis Genetics Center (CGC)	http://biosci.umn.edu/CGC
CGC Strain Request	http://biosci.umn.edu/CGC/Strains/request.htm
CGC Nomenclature Guide	http://biosci.umn.edu/CGC/Nomenclature/nomenguid.htm
C. elegans Fosmids at Geneservice Ltd	http://www.geneservice.co.uk/products/clones/Celegans_Fos.jsp
BC GSC Fosmid Search	http://elegans.bcgsc.ca/perl/fosmid/CloneSearch
C. elegans Cosmids	http://www.its.caltech.edu/~wormbase/userguide/OtherResource/ObtainingReagents.html#elegans_genomic
C. elegans ESTs (NEXTDB)	http://nematode.lab.nig.ac.jp/index.html
C. elegans Gene Knockout Consortium	http://www.celeganskoconsortium.omrf.org
National BioResource Project of Japan	http://shigen.lab.nig.ac.jp/c.elegans/index.jsp
NemaGENETAG	http://elegans.imbb.forth.gr/nemagenetag
WorfDB	http://worfdb.dfci.harvard.edu
ORFeome at Geneservice	http://www.geneservice.co.uk/products/cdna/Celegans_ORF.jsp
ORFeome at OpenBiosystems	http://www.openbiosystems.com/GeneExpression/Non%2DMammalian/Worm/CelegansORFs/
Promoterome	http://vidal.dfci.harvard.edu/promoteromedb
Promoterome Clones at Geneservice	http://www.geneservice.co.uk/products/clones/Celegans_Prom.jsp
Promoterome Clones at OpenBiosystems	http://www.openbiosystems.com/GeneExpression/Non%2DMammalian/Worm/CelegansPromoters
Primer Pairs at Research Genetics	http://www.resgen.com/products/CelGPs.php3
Ahringer RNAi library at Geneservice	http://www.geneservice.co.uk/products/rnai/Celegans.jsp
ORFeome RNAi library at OpenBiosystems	http://www.openbiosystems.com/GeneExpression/Non%2DMammalian/Worm/CelegansORF%2DRNAi
RNAi clone mapping table	ftp://caltech.wormbase.org/pub/annots/rnai
Affymetrix Expression Profiling Microarray	http://www.affymetrix.com/products/arrays/specific/celegans.affx
Agilent Expression Profiling Microarray	http://www.chem.agilent.com/Scripts/PDS.asp?lPage=29452
WashU GSC Expression Profiling Microarray	http://www.genome.wustl.edu/genome/celegans/microarray/ma_gen_info.cgi
Affymetrix Tiling Microarray	http://www.affymetrix.com/products/arrays/specific/celegans_tiling.affx
Agilent Tiling Microarray	http://www.chem.agilent.com/scripts/pds.asp?lPage=50880
Nimblegen	http://www.nimblegen.com/products/chip/index.html
Fire Vector Kit at Addgene	http://www.addgene.org/pgvec1?f=c&cmd=showcol&colid=1
Seydoux Vectors	http://www.bs.jhmi.edu/MBG/SeydouxLab/vectors/index.html

BC, British Columbia; GSC, Genome Sequencing Center; NEXTDB, Nematode Expression Pattern Database; WashU, Washington University.

WormClassroom is another recently launched information portal that focuses on *C. elegans* as a tool for teaching. This resource includes a brief introduction to *C. elegans*, discussions of major research topics, links to experimental protocols and many other educational resources.

Anatomy. WormAtlas, which is coordinated by the Center of Anatomical Studies of *C. elegans*, is an excellent reference that is capable of answering all but the most specialized questions about *C. elegans* anatomy. It provides a comprehensive overview of anatomical features and offers several ways to explore the worm. The Handbook of Worm Anatomy, with its textbook-like layout, is perfect for newcomers to the field. It starts with a general introduction to body layout, progressing gradually into more specialized topics and individual tissues. Each chapter is accompanied by a multitude of light and electron micrographs and extensively annotated diagrams (FIG. 1). The structure of the worm can be investigated using the Slidable Worm interface, which allows an interactive examination of serial electron micrograph sections of an adult hermaphrodite. Individual neuron pages can be accessed directly through the alphabetical list of neuronal cells.

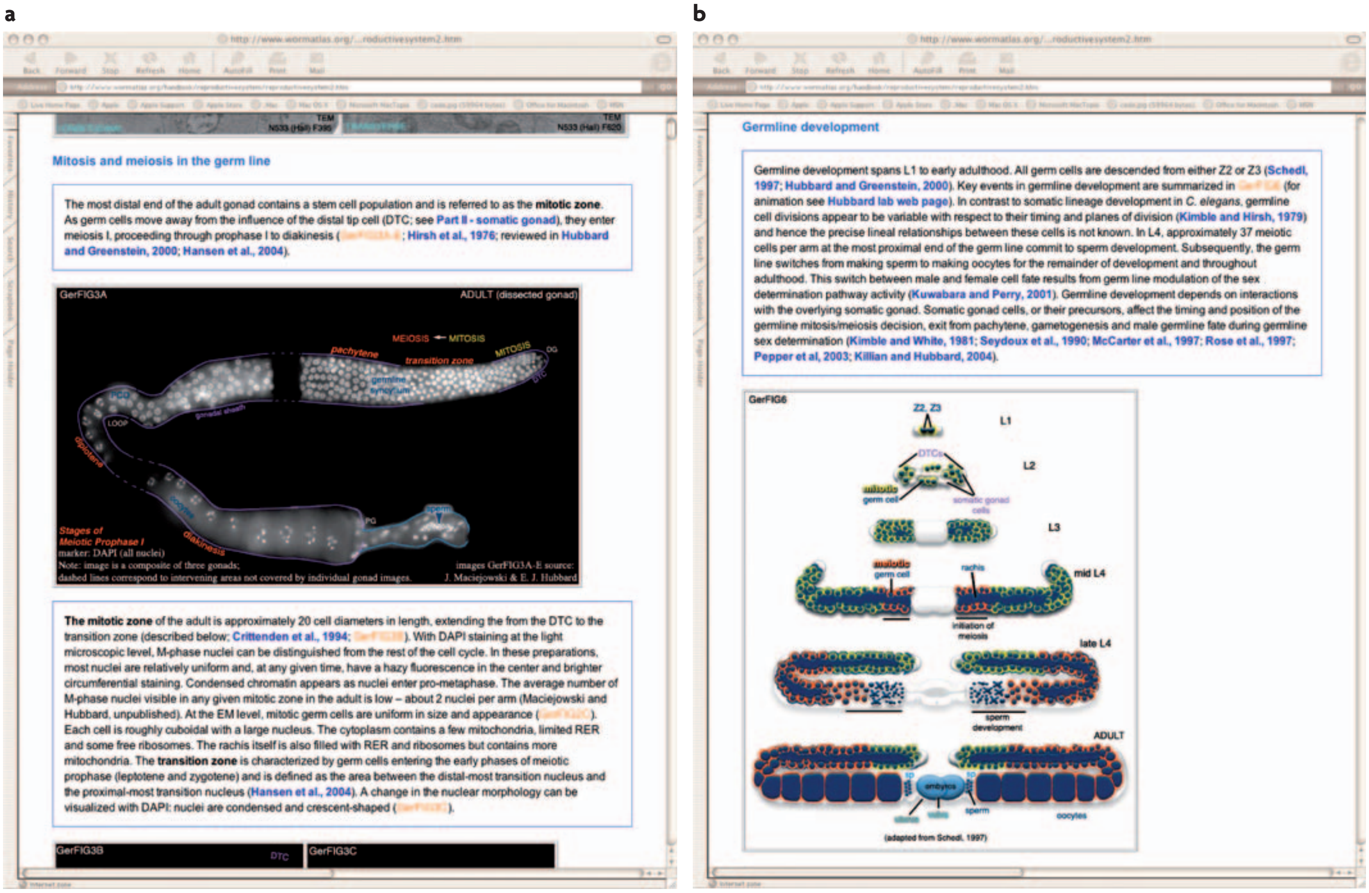
WormAtlas also offers an extensive guide to cell-identification methods that focuses mostly, but not exclusively, on neurons and comprises separate sections on head, body and tail regions, featuring images of cells viewed from different perspectives. In addition, WormAtlas contains a collection of methods for studying worm anatomy, a glossary of anatomical terms and links to cornerstone publications and external resources, such as *C. elegans* movies and software for exploring worm anatomy. WormAtlas is accompanied by a specialized database called WormImage, which stores thousands of electron micrographs contributed by the *C. elegans* community and which can be freely annotated by users.

WormBase. WormBase is the primary database for *C. elegans* and related nematodes[6], and maintains and distributes the genomic sequences for *C. elegans* and *C. briggsae* that are used by all other genomic databases. Gene-structure models are created using a combination of several automated methods along with manual curation using information that is extracted from the literature. These models are imported by such databases as Ensembl and GenBank, and the

Table 1 | **Key non-worm-specific resources used by the *Caenorhabditis elegans* community**

Name	Brief description
Integrated Databases	
NCBI Entrez Gene	An integrated repository for gene-specific information that is produced or collected at NCBI[43].
NCBI MapViewer	A map-based tool that allows one to search and display genomic information by chromosomal position.
EBI Ensembl	Generates automatic annotations of select eukaryotic genomes and provides several unique data-visualization and data-mining tools for exploring protein architecture, browsing family clusters, visualizing large-scale synteny blocks, and so on[44].
UCSC Genome Browser	Displays a segment of the genome together with multiple annotation tracks, including gene structures, alignments of mRNAs and ESTs, *Caenorhabditis briggsae* WABA and Blastz alignments, repetitive elements identified by RepeatMasker and Tandem Repeats Finder, and so on[45–48]. Gene Sorter allows the identification of genes on the basis of sequence homology, similarity of expression profiles or gene-ontology term distribution[49].
UniProt	A repository of protein-centered sequence and functional information[50]. It comprises three databases: UniParc, UniRef and UniProt Knowledge Base. UniParc and UniRef provide non-redundant sequence collections, whereas the Knowledge Base combines sequence data with various manually and automatically generated functional annotations.
Orthology	
InParanoid	The InParanoid algorithm attempts to make a distinction between gene-duplication events that occurred before and after speciation events and identifies both orthologous and paralogous genes[51]. InParanoid is accepted as one of the standard methods for orthologue identification by several model-organism databases, including WormBase.
OrthoMCL	OrthoMCL is a fully automatic graph-clustering algorithm that identifies orthologous and paralogous relationships on the basis of all-against-all BLAST similarity scores followed by Markov clustering[52]. Whereas InParanoid identifies orthologous genes between two genomes, OrthoMCL clusters orthologues from multiple species.
TreeFam	Unlike InParanoid and OrthoMCL, which infer orthologues and paralogues from BLAST scores, TreeFam orthologues and paralogues are deduced from the phylogenetic tree of a gene family. In addition, extensive manual curation is used to refine and correct automatically generated trees[53]. TreeFam-based orthologues are also incorporated into WormBase.
Pathways	
Reactome, MetaCyc, KEGG, aMAZE	Reactome manually curates human pathways that are automatically projected through orthology mapping to several model organisms, including *C. elegans*[54]. MetaCyc is now being developed for *C. elegans*[55]. KEGG and aMAZE do not currently contain *C. elegans*-specific information, but that might change in the future[56,57].
Gene Ontology	
Gene Ontology (GO)	The GO project is creating a dynamic, controlled vocabulary that can be used for automated analyses of gene function. Three independent ontologies are constructed, describing biological processes, molecular functions and cellular components pertaining to a gene[58]. Ontologies are developed by the GO Consortium, whereas gene annotations are carried out by individual model-organism databases.
Protein and RNA Families	
PFam	PFam is a database of protein families that comprises multiple sequence alignments and hidden Markov models (HMMs) that are derived from them[59,60]. It allows users to identify proteins that belong to the same family, examine functional annotation for the identified domains, visualize interdomain interactions, and so on.
RFam	RFam[61] is a collection of non-coding RNA families that are represented by multiple sequence alignments and profile stochastic context-free grammars (SCFGs), which capture both the secondary structure and the primary-sequence profile of multiple sequence alignments, and enable users to carry out searches for putative new members of non-coding RNA families.
miRBase	Formerly a part of RFam, miRBase is devoted to microRNAs (miRNAs)[62]. miRNAs are of special interest to *C. elegans* researchers because the two founding members of this family, *let-7* and *lin-4*, were discovered through mutational analysis in *C. elegans*[63–66]. miRBase provides access to all published miRNA sequences and their potential target genes through the Sequence and Target databases.
Tools	
NCBI Tools for Data Mining	The NCBI provides tools for the analysis of nucleotide and protein sequences, protein structure, gene expression, and so on. The collection includes programming utilities for interaction with the Entrez retrieval system.
Toolbox at the EBI	Tools developed at EBI cover such areas as microarray and protein functional analyses, similarity and homology searches, proteomic services, and so on.
UniProt Tools	Tools developed by individual UniProt consortium members and by several independent projects can be used for similarity searches and multiple sequence alignments, batch retrieval, proteomic and literature analyses.
ExPASy Proteomics Tools	An extensive list of tools for proteomic analysis have been developed and hosted on the ExPASy server. Many links to proteomics tools that are available from other sites are also provided.

EBI, European Bioinformatics Institute; ExPASy, Expert Protein Analysis System; KEGG, Kyoto Encyclopaedia of Genes and Genomes; NCBI, National Center for Biotechnology Information (USA); UCSC, University of California at Santa Cruz.



Mitosis and meiosis in the germ line

The most distal end of the adult gonad contains a stem cell population and is referred to as the **mitotic zone**. As germ cells move away from the influence of the distal tip cell (DTC; see Part II - somatic gonad), they enter meiosis I, proceeding through prophase I to diakinesis ([illegible]; Hirsh et al., 1976; reviewed in Hubbard and Greenstein, 2000; Hansen et al., 2004).

marker: DAPI (all nuclei)
Note: image is a composite of three gonads;
dashed lines correspond to intervening areas not covered by individual gonad images.

images GerFIG3A-E source:
J. Maciejowski & E. J. Hubbard

The mitotic zone of the adult is approximately 20 cell diameters in length, extending the from the DTC to the transition zone (described below; Crittenden et al., 1994; [illegible]). With DAPI staining at the light microscopic level, M-phase nuclei can be distinguished from the rest of the cell cycle. In these preparations, most nuclei are relatively uniform and, at any given time, have a hazy fluorescence in the center and brighter circumferential staining. Condensed chromatin appears as nuclei enter pro-metaphase. The average number of M-phase nuclei visible in any given mitotic zone in the adult is low – about 2 nuclei per arm (Maciejowski and Hubbard, unpublished). At the EM level, mitotic germ cells are uniform in size and appearance ([illegible]). Each cell is roughly cuboidal with a large nucleus. The cytoplasm contains a few mitochondria, limited RER and some free ribosomes. The rachis itself is also filled with RER and ribosomes but contains more mitochondria. The **transition zone** is characterized by germ cells entering the early phases of meiotic prophase (leptotene and zygotene) and is defined as the area between the distal-most transition nucleus and the proximal-most transition nucleus (Hansen et al., 2004). A change in the nuclear morphology can be visualized with DAPI: nuclei are condensed and crescent-shaped ([illegible]).

Germline development

Germline development spans L1 to early adulthood. All germ cells are descended from either Z2 or Z3 (Schedl, 1997; Hubbard and Greenstein, 2000). Key events in germline development are summarized in [illegible] (for animation see Hubbard lab web page). In contrast to somatic lineage development in *C. elegans*, germline cell divisions appear to be variable with respect to their timing and planes of division (Kimble and Hirsh, 1979) and hence the precise lineal relationships between these cells is not known. In L4, approximately 37 meiotic cells per arm at the most proximal end of the germ line commit to sperm development. Subsequently, the germ line switches from making sperm to making oocytes for the remainder of development and throughout adulthood. This switch between male and female cell fate results from germ line modulation of the sex determination pathway activity (Kuwabara and Perry, 2001). Germline development depends on interactions with the overlying somatic gonad. Somatic gonad cells, or their precursors, affect the timing and position of the germline mitosis/meiosis decision, exit from pachytene, gametogenesis and male germline fate during germline sex determination (Kimble and White, 1981; Seydoux et al., 1990; McCarter et al., 1997; Rose et al., 1997; Pepper et al, 2003; Killian and Hubbard, 2004).

(adapted from Schedl, 1997)

Figure 1 | **WormAtlas.** WormAtlas provides a comprehensive overview of *Caenorhabditis elegans* anatomy from general body layout to discussions of specialized tissues and individual cells. The resource features many light and electron micrographs together with artistically rendered diagrams to accompany all its pages. **a** | A micrograph of a dissected hermaphrodite gonad depicting mitosis and meiosis in the germ line. **b** | Germline development in a hermaphrodite. Both images appear in the Germ Line chapter of the Handbook of Worm Anatomy. The figure in part **a** is reproduced with permission from WormAtlas © (2005) R. Lints & D. H. Hall, from an original image by J. Maciejowski & E. J. Hubbard, New York University, USA. The figure in part **b** is reproduced with permission from WormAtlas © (2005) R. Lints & D. H. Hall.

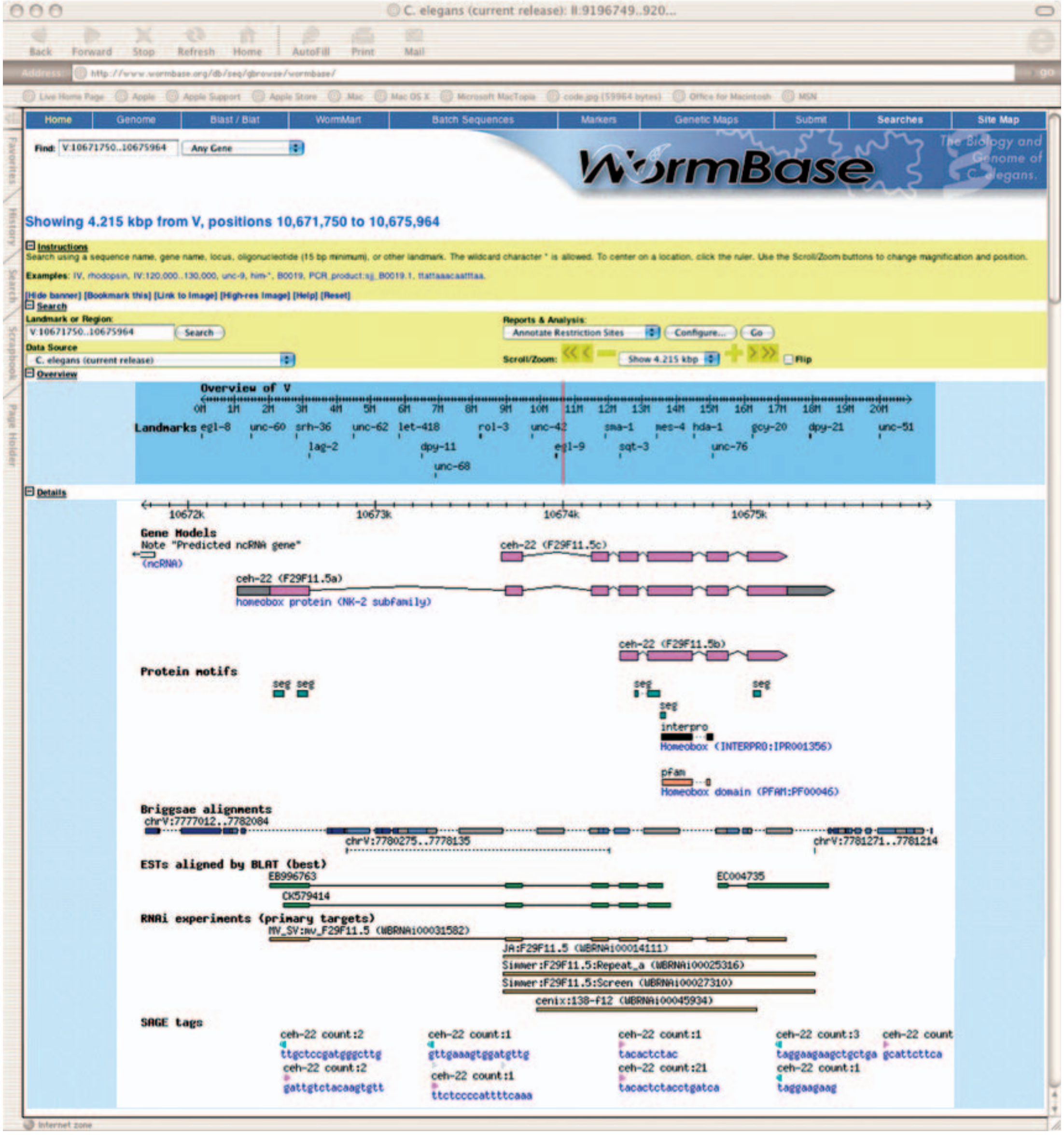
corresponding protein set, WormPep, is used by many databases including UniProt.

WormBase also contains a wealth of functional information on genes and gene products, and its user interface allows easy exploration of various types of data and their relationships. The gene page summarizes the most important facts about a gene, with related data types being grouped together in separate sections, including gene identity, genomic information, gene function, homology data, available reagents and bibliography. Many gene pages feature concise descriptions that amalgamate individual facts that have been manually extracted from cornerstone publications, and provide users with a simple summary of gene function. Links are also provided to Entrez Gene and to the US National Center for Biotechnology (NCBI) *C. elegans*-specific database, WormGenes, which offers an alternative set of gene models that have been reconstructed from mRNA and EST alignments and carries out a number of analyses, including protein-domain and homologue identification.

The WormBase Genome Browser offers an interactive and highly customizable way of exploring genomic regions, allowing users to visualize sequence features and related resources such as RNAi and microarray probes, ESTs and mRNAs, alleles and protein motifs (FIG. 2). Each data type has a dedicated web display and many have specialized search engines; for example, for expression patterns and genetic markers.

WormBase also provides several tools that allow users to make use of sophisticated data-mining strategies to discover complex relationships between various pieces of data. These include two types of highly-developed query languages, database searches that traverse data-type boundaries and a specialized data-mining tool, WormMart. WormBase also provides a dedicated data-mining server that can be programmatically accessed by users with advanced bioinformatics skills.

Finally, the availability of the *C. briggsae* genome makes WormBase a powerful platform for comparative genomics, and the addition of three other nematode



C. elegans (current release): II:9196749..920...

http://www.wormbase.org/db/seq/gbrowse/wormbase/

Find: V:10671750..10675964 Any Gene

Showing 4.215 kbp from V, positions 10,671,750 to 10,675,964

⊟ Instructions
Search using a sequence name, gene name, locus, oligonucleotide (15 bp minimum), or other landmark. The wildcard character * is allowed. To center on a location, click the ruler. Use the Scroll/Zoom buttons to change magnification and position.

Examples: IV, rhodopsin, IV:120,000..130,000, unc-9, him-*, B0019, PCR_product:sjj_B0019.1, ttattaaacaatttaa.

[Hide banner] [Bookmark this] [Link to Image] [High-res Image] [Help] [Reset]

⊟ Search
Landmark or Region: V:10671750..10675964 Search
Data Source: C. elegans (current release)
Reports & Analysis: Annotate Restriction Sites Configure... Go
Scroll/Zoom: Show 4.215 kbp □ Flip

⊟ Overview

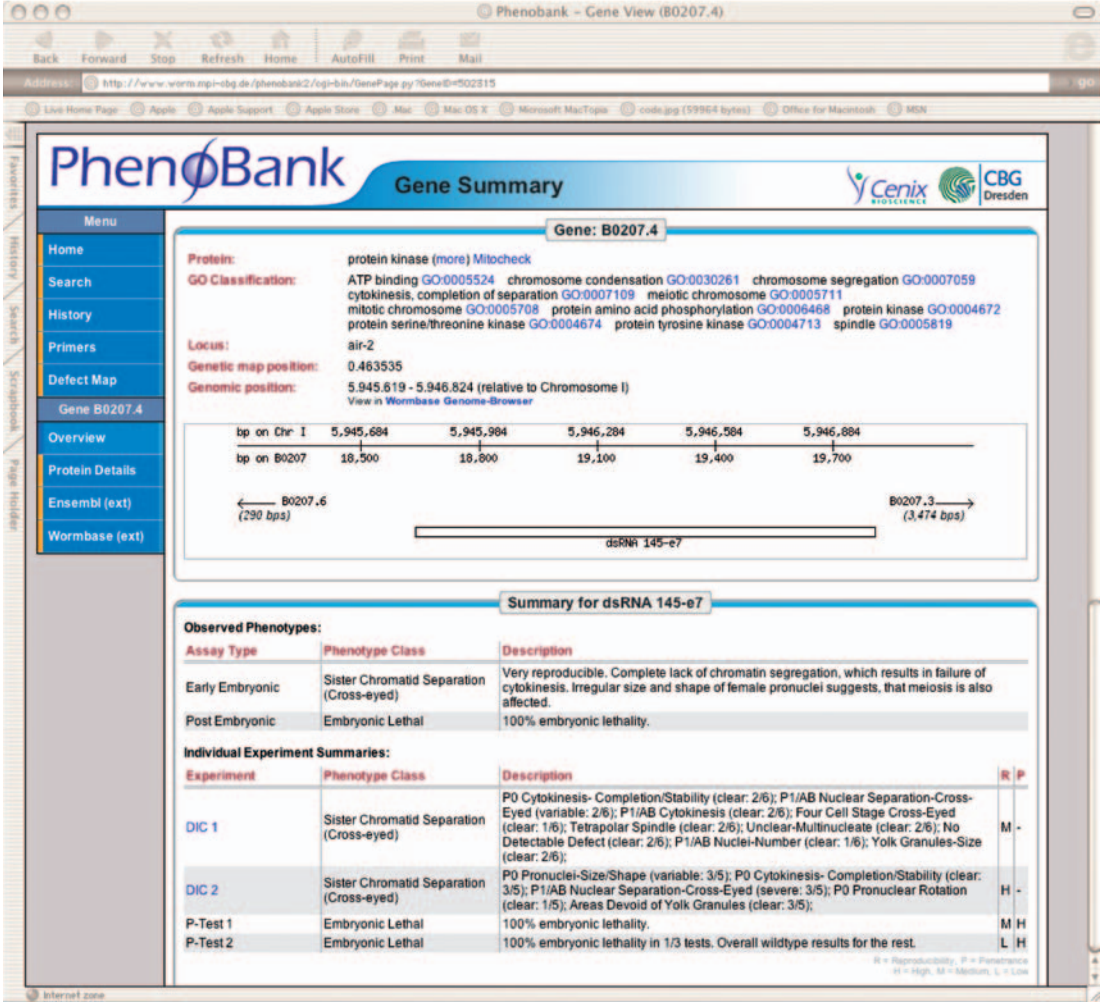
⊟ Details

Note "Predicted ncRNA gene"

Figure 2 | **WormBase Genome Browser.** Genome Browser allows visualization of a genomic region with associated sequence features and annotations. The display can be customized easily, for example, allowing the user to select tracks to be displayed, change the size of the genomic region and define strandedness. DNA sequences and annotation files in GFF format can also be retrieved. The *Caenorhabditis elegans ceh-22* genomic region on chromosome V is shown; the tracks selected are: gene models, protein motifs, alignments to the *Caenorhabditis briggsae* genome, EST alignments, probes used in RNAi experiments and SAGE (serial analysis of gene expression) tags. Figure reproduced with permission from WormBase © (1995–2005) California Institute of Technology, USA, Cold Spring Harbor Laboratory, USA, Washington University at St. Louis, USA, and The Wellcome Trust Sanger Institute, UK.

genomes in the near future will further enhance this utility (see below).

Phenotype and gene function. Several specialized databases focus on in-depth characterizations of mutant phenotypes and offer unique analysis and data-representation tools. One such database is RNAiDB, which contains phenotypic data from several large-scale RNAi studies[7] and uses a detailed embryonic phenotype-scoring system, based on 47 criteria, that makes phenotypes amenable to computational analysis. Experiments that are annotated using this approach can be clustered to produce groups of genes that are likely to be functionally related. The PhenoBlast query tool takes further advantage of this phenotype-recording method, allowing users to search for functionally related genes on the basis of their phenotypic signature similarity. Other data-access methods range from simple keyword searches for gene names to complex combinatorial queries using various phenotype attributes. RNAiDB also features more than 1,500 movies, corresponding to about 400 RNAi experiments, that enable users to examine primary data to search for additional or subtle phenotypes.



Gene: B0207.4

Protein: protein kinase (more) Mitocheck

GO Classification: ATP binding GO:0005524 chromosome condensation GO:0030261 chromosome segregation GO:0007059 cytokinesis, completion of separation GO:0007109 meiotic chromosome GO:0005711 mitotic chromosome GO:0005708 protein amino acid phosphorylation GO:0006468 protein kinase GO:0004672 protein serine/threonine kinase GO:0004674 protein tyrosine kinase GO:0004713 spindle GO:0005819

Locus: air-2

Genetic map position: 0.463535

Genomic position: 5.945.619 - 5.946.824 (relative to Chromosome I)
View in Wormbase Genome-Browser

bp on Chr I 5,945,684 5,945,984 5,946,284 5,946,584 5,946,884
bp on B0207 18,500 18,800 19,100 19,400 19,700
B0207.6 (290 bps)
B0207.3 (3,474 bps)
dsRNA 145-e7

Summary for dsRNA 145-e7

Observed Phenotypes:

Assay Type	Phenotype Class	Description
Early Embryonic	Sister Chromatid Separation (Cross-eyed)	Very reproducible. Complete lack of chromatin segregation, which results in failure of cytokinesis. Irregular size and shape of female pronuclei suggests, that meiosis is also affected.
Post Embryonic	Embryonic Lethal	100% embryonic lethality.

Individual Experiment Summaries:

Experiment	Phenotype Class	Description	R	P
DIC 1	Sister Chromatid Separation (Cross-eyed)	P0 Cytokinesis- Completion/Stability (clear: 2/6); P1/AB Nuclear Separation-Cross-Eyed (variable: 2/6); P1/AB Cytokinesis (clear: 2/6); Four Cell Stage Cross-Eyed (clear: 1/6); Tetrapolar Spindle (clear: 2/6); Unclear-Multinucleate (clear: 2/6); No Detectable Defect (clear: 2/6); P1/AB Nuclei-Number (clear: 1/6); Yolk Granules-Size (clear: 2/6);	M	-
DIC 2	Sister Chromatid Separation (Cross-eyed)	P0 Pronuclei-Size/Shape (variable: 3/5); P0 Cytokinesis- Completion/Stability (clear: 3/5); P1/AB Nuclear Separation-Cross-Eyed (severe: 3/5); P0 Pronuclear Rotation (clear: 1/5); Areas Devoid of Yolk Granules (clear: 3/5);	H	-
P-Test 1	Embryonic Lethal	100% embryonic lethality.	M	H
P-Test 2	Embryonic Lethal	100% embryonic lethality in 1/3 tests. Overall wildtype results for the rest.	L	H

R = Reproducibility, P = Penetrance
H = High, M = Medium, L = Low

Figure 3 | **PhenoBank report page.** The gene page for the *Caenorhabditis elegans* gene *air-2* (aurora/Ipl1-related kinase; B0207.4) shows information on gene function, gene ontology classification, genetic and genomic positions. A summary of RNAi experiments that target the gene is also shown, including the assay type and descriptions of observed phenotypes and their classification. Data obtained from individual experiments and the resulting aggregated summary are presented. Figure reproduced with permission from Phenobank.

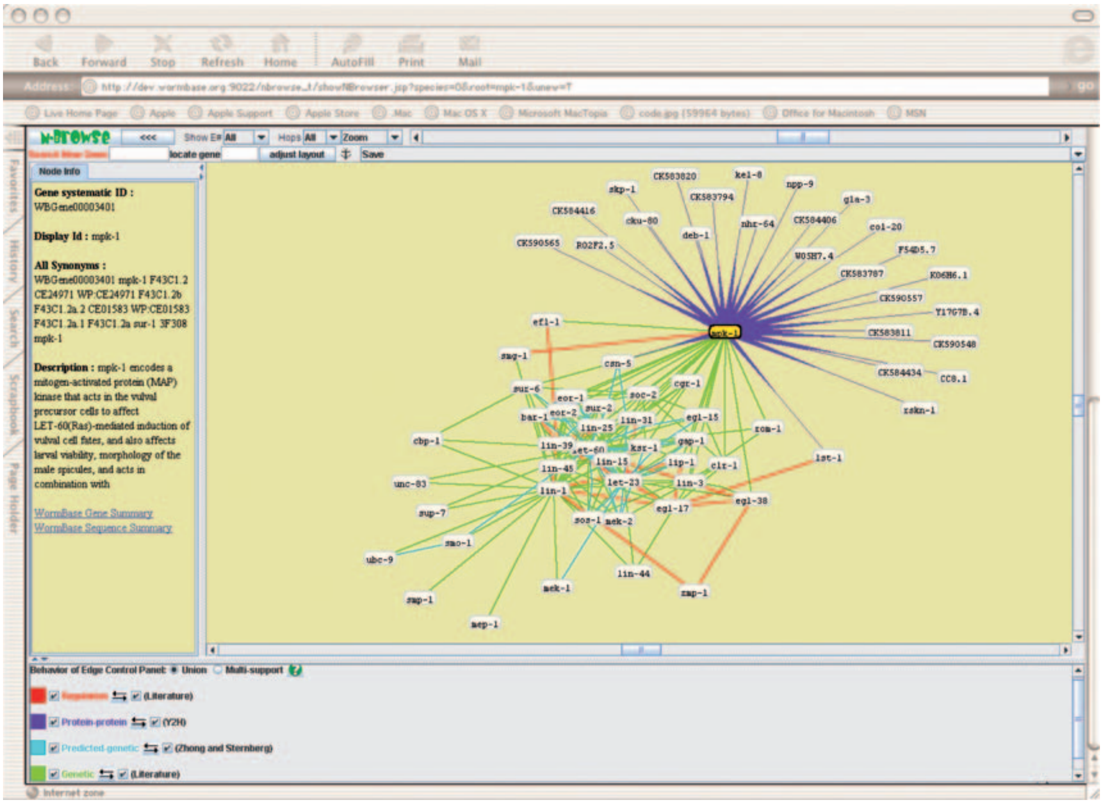
PhenoBank is devoted to the set of phenotypes that were observed in a full-genome RNAi screen for early embryonic defects[8]. The PhenoBank approach is similar to that of RNAiDB, in that it provides highly detailed descriptions of these defects. The embryonic phenotypes that are recorded during the first two cell divisions are classified using 45 defect categories that are grouped into 25 classes, enabling clustering of genes. A unified search engine allows users to retrieve data on the basis of a combination of criteria, and each experimental report page contains a gene description, phenotypic classification and a representative movie (FIG. 3).

The *C. elegans* RNAi Phenome Database uses a similar approach to catalogue terminal embryonic RNAi phenotypes, which are described by a systematic annotation method that makes use of controlled vocabulary terms. All three databases contain links to WormBase gene report pages, which makes it easy to integrate the phenotypic information provided by these resources with additional functional and genomic data.

Protein and gene interactions. Protein- and gene-interaction networks can be used to infer biological processes and the molecular functions of network components. The largest available experimental data set for *C. elegans*, Interactome, comprises more than 5,600 interactions, which were identified through a genome-wide yeast two-hybrid screen[9]. The InteractomeDB database harbours the primary data and allows users to explore relationships between multiple proteins through a graphical user interface.

Interactome data have also been imported into WormBase and can be examined using the n-Browse-based interaction viewer (FIG. 4). The WormBase interaction browser also allows users to visualize other types of data, such as information on genetic interactions and gene regulation, and relationships extracted

Controlled vocabulary
In contrast to natural language vocabularies, which have no restrictions on the terms that can be used, controlled vocabulary is a set of predefined, authorized terms that have been chosen by its designer to reduce the inherent ambiguity in human language and ensure consistency of concept descriptions.



Gene systematic ID :
WBGene00003401

Display Id : mpk-1

All Synonyms :
WBGene00003401 mpk-1 F43C1.2 CE24971 WP:CE24971 F43C1.2b F43C1.2a.2 CE01583 WP:CE01583 F43C1.2a.1 F43C1.2a sur-1 3F308 mpk-1

Description : mpk-1 encodes a mitogen-activated protein (MAP) kinase that acts in the vulval precursor cells to affect LET-60(Ras)-mediated induction of vulval cell fates, and also affects larval viability, morphology of the male spicules, and acts in combination with

WormBase Gene Summary
WormBase Sequence Summary

Behavior of Edge Control Panel: ● Union ○ Multi-support

🟥 ☑ Regulation ⇆ ☑ (Literature)
🟪 ☑ Protein-protein ⇆ ☑ (Y2H)
🟦 ☑ Predicted-genetic ⇆ ☑ (Zhong and Sternberg)
🟩 ☑ Genetic ⇆ ☑ (Literature)

Figure 4 | **n-Browse interaction viewer.** This tool allows the user to explore interaction relationships between genes and proteins using a graphical interface. Genetic, predicted and physical interactions, as well as interactions that have been extracted from the literature, are currently supported, with additional types of interaction data to be added in the future. A graph for the MAP kinase gene *mpk-1* is shown. The Node Info panel displays gene ID, name and synonyms, together with a brief description of gene function. The types of interaction that are displayed can be selected using the Control Panel. Figure reproduced with permission from n-Browse.

from literature, making it a powerful framework for exploring protein and gene interactions. Interactome data are also available through the IntAct and BioGrid databases, which collect data from species ranging from *Saccharomyces cerevisiae* to humans[10,11].

Other *C. elegans*-specific data sets are also being developed, based on a variety of experimental techniques including mass spectrometric analysis of protein complexes, co-immunoprecipitation and genetic epistasis analysis. In addition, new resources are emerging for analysing protein–DNA-interaction networks in *C. elegans*, including EDGEdb, which contains data on promoter–transcription factor interactions[12,13].

Expression patterns. To facilitate high-throughput functional analyses, several projects that are aimed at determining expression patterns on a genome-wide level are underway. The goal of the Nematode Expression Pattern Database (NEXTDB) is to construct a comprehensive expression map of *C. elegans* through EST analysis and whole-mount *in situ* hybridization. NEXTDB serves as a repository for experimental data and provides an access interface for the research community. It currently contains information on approximately 250,000 ESTs and *in situ* hybridization images for more than 11,000 cDNA clones, which can be accessed several ways, including keyword and homology searches or a chromosome map browser (FIG. 5). Data from a genome-wide RNAi screen that was carried out by the Sugimoto group[14] are also incorporated into NEXTDB, making it possible to search for expression patterns of genes that produce specific RNAi phenotypes.

Two other databases — the Hope Laboratory Expression Pattern Database and the Expression Pattern Database from the British Columbia (BC) *C. elegans* Gene Expression Consortium — use promoter fusion constructs to monitor gene expression in transgenic worms. The BC Expression Pattern Database uses defined promoter sequences fused to GFP to monitor gene expression *in vivo*, with the aim of generating reporter constructs for about 5,000 *C. elegans* genes that have human orthologues. The Hope Laboratory uses a strategy that involves fusion of random genomic DNA fragments to a reporter gene, followed by sequencing and selection of plasmids that contain appropriate fusions[15]. Recently, a new high-throughput approach that makes use of Gateway cloning and promoterome constructs (described below) has been adopted by this group[16]. More than 1,300 transgenic strains have been generated by the project to date. Expression patterns in

Gateway recombinational cloning
Gateway cloning technology is based on the site-specific recombination system of λ phage, which allows rapid transfer of DNA fragments between different vectors while maintaining the correct orientation and reading frame. It provides a highly efficient alternative to the restriction-enzyme and ligase-based cloning strategy.

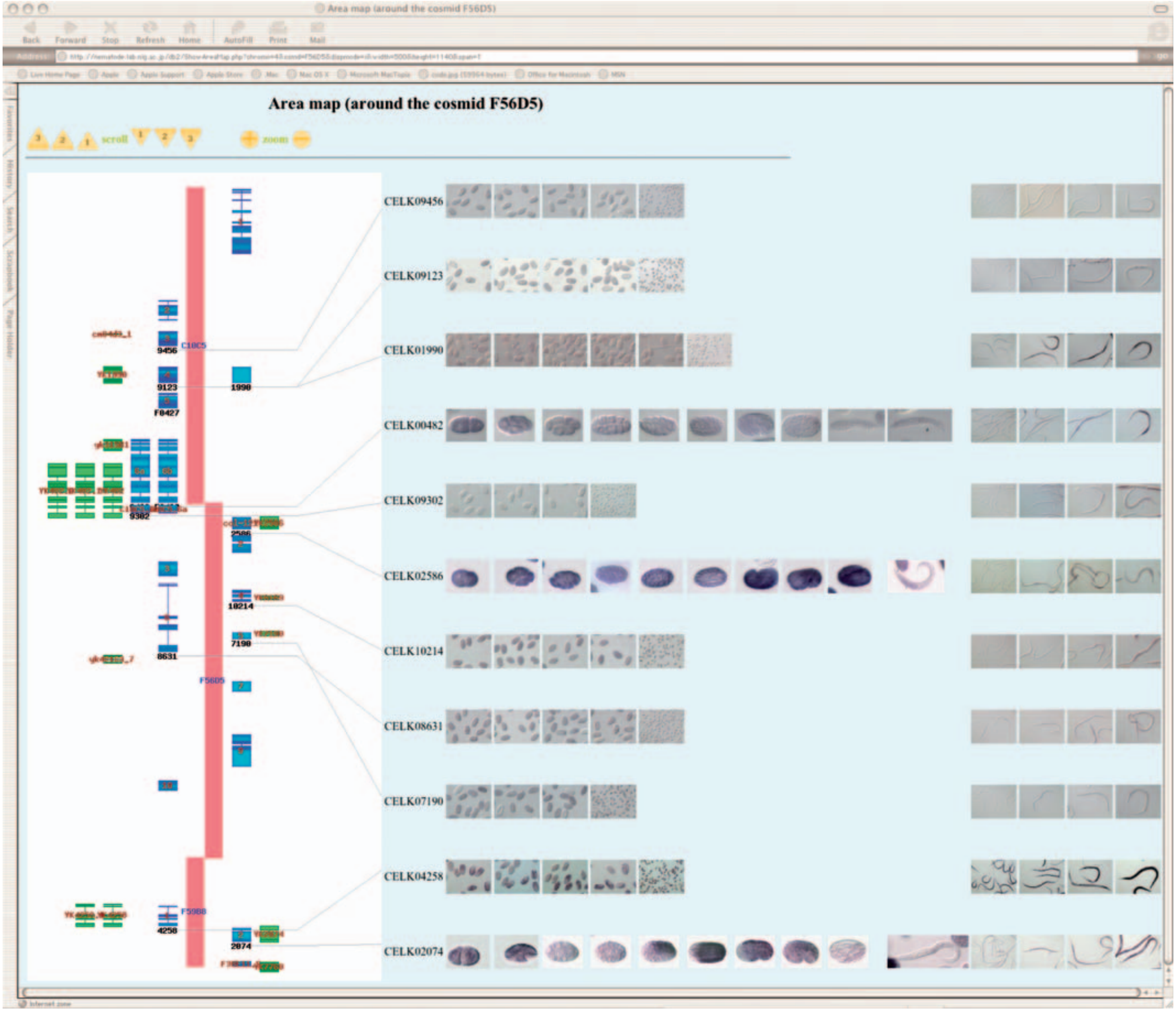


Figure 5 | **NEXTDB map view.** The area map around the cosmid F56D6 is shown, with ESTs aligned to the genome, gene clusters and results of *in situ* hybridization experiments. Gene clusters and hybridization images are linked to report pages that detail information such as protein family and gene ontology classification, homology analyses and data on clones available at NEXTDB (Nematode Expression Pattern Database). Figure reproduced with permission from NEXTDB © (2001–2005) Y. Kohara, National Institute of Genetics, Japan.

both databases can be identified using various criteria, including gene names, the tissue in which the reporters are expressed and life stage. Both projects submit their expression data to WormBase and make stable transgenic strains available to the research community through the *Caenorhabditis* Genetics Center (CGC; described below).

Expression profiling. The majority of *C. elegans* microarray data that have been produced to date were generated using PCR-product-based chips developed at Stanford University, USA. Those data are available through the Stanford Microarray Database (SMD) in different formats, from raw image files to normalized intensity values[17]. In addition to worm data, SMD stores expression information for various other species, from *S. cerevisiae* to humans, facilitating cross-species analyses. For example, studies of gene co-regulation across multiple species have been used to construct expression topomaps, which have proven to be an important tool for gene discovery and functional annotation[18,19].

Two other databases collect *C. elegans* microarray data: NCBI Gene Expression Omnibus (GEO) and the European Bioinformatics Institute (EBI) ArrayExpress. GEO archives high-throughput data generated through various methods, including microarray profiling, serial analysis of gene expression (SAGE) and mass spectrometry analyses of protein abundance[20]. This database

Expression topomap
Expression topomap is a visualization of genes that show correlated expression across a large set of microarray experiments. Co-regulated genes appear as mountains, the height of which indicates local gene density. Expression topomap can be used to infer gene functions or to identify genes that are co-regulated with known sets of genes.

provides several options for data analysis, including a data-clustering facility, a subset comparison tool and a profile neighbour search. ArrayExpress is a microarray-specific data repository that is fully compliant with MIAME (Minimum Information About a Microarray) standards[21]. It is tightly integrated with Expression Profiler, an EBI-based online platform for microarray data analysis that brings together diverse analytical tools[22]. ArrayExpress provides password-protected access to prepublication data, making it a useful collaborative platform. WormBase also collects published microarray data sets, currently including data from more than 40 publications using several microarray platforms.

SAGE is a powerful tool for gene discovery as it does not require *a priori* knowledge of gene structure. More than 30 *C. elegans* SAGE libraries generated at the British Columbia Genome Sciences Centre (BC GSC) are currently available, covering a wide variety of experimental conditions[23–25]. Tag-frequency data can be accessed through the BC GSC site or through WormBase, which carries the full complement of SAGE libraries and offers additional tools for data visualization and analysis.

Structural genomics. Protein structural information can provide invaluable insights into many aspects of biology, such as the molecular function of the gene product, its binding partners and potential interacting small molecules. Three-dimensional structure alignments can help establish relatedness of highly divergent proteins and aid in the classification of proteins into families and superfamilies. The recently initiated Structural Genomics of *C. elegans* (SGCE) project is developing an automated high-throughput system that will be used to determine crystal structures of *C. elegans* proteins[26]. More than 16,000 ORFs have been selected for structural analysis and are at various stages in the pipeline, including 19 fully resolved structures. Lists of the proteins that are undergoing each step of the procedure, from cloning, to purification, to the final structure, are available through the SGCE site, together with such information as protein solubility, yield and purity. This feature enables users to estimate when structures for particular proteins might become available. Structures are ultimately deposited to the Protein Data Bank[27] (PDB), which also offers multiple tools for further analyses. SGCE also allows users to request proteins that are expressed as part of the project, providing a valuable material resource.

Literature analysis. Literature analysis is an essential part of scientific research. NCBI PubMed enables users to search the vast majority of biological literature through its Entrez interface and is familiar to almost all biologists. *Caenorhabditis elegans* researchers also have access to a powerful search engine, Textpresso, which was originally developed specifically for the analysis of worm literature but is now being adopted by other research communities[28]. Unlike PubMed, Textpresso can access the full text of articles, and currently contains more than 8,000 papers on *C. elegans* and related nematodes, as well as abstracts from international and regional *C. elegans* meetings that are not available in PubMed. Close to 1,000 new papers are added to the database each year. The search engine utilizes ontology-based categories of biological terms, enabling the use of semantic queries that combine multiple keywords and category terms. A specialized query language is also offered by Textpresso, making it a powerful literature-mining platform.

Ontology
Like controlled vocabularies, ontologies, as used in computer science, describe objects using predefined terms. In addition, ontologies define relationships between objects, making it possible to capture the structure of a set of objects.

Related nematodes and comparative genomics. Since the release of the *C. briggsae* genomic sequence, *C. elegans* has served as one of the most powerful systems for comparative genomics[2]. To further improve the power of comparative analyses by enabling multiway comparisons and allowing the identification of rapidly evolving sequence elements, genomes of three more *Caenorhabditis* species are being sequenced. The status of this sequencing is reported on the Washington University Genome Sequencing Center (WashU GSC) web site and the newly generated sequences can be searched through the WashU GSC Blast server. All *Caenorhabditis* genomes will be ultimately incorporated into WormBase.

The relatively small size of most nematode genomes compared with those of other animals and plants makes them attractive from the point of view of whole-genome sequencing, and several other nematode species are currently in the genome-sequencing pipeline. Of particular interest is the sequencing of *Pristionchus pacificus*, as much of the research on this species focuses on comparative analyses of biological processes that are functionally conserved with *C. elegans* but significantly divergent at the molecular level, such as vulva development[29]. Another large sequencing project that is becoming increasingly important to *C. elegans* biologists is the Parasitic Nematode Sequencing Project. Its initial goal was to generate a large number of ESTs from ~30 parasitic nematodes. In appreciation of the value of comparative analyses, genomes of ten parasitic nematodes that are closely related to *C. elegans* and belong to the order Strongylida have recently been approved for sequencing, in addition to the parasite *Trichinella spiralis*[30] and the insect-killing nematode *Heterorhabditis bacteriophora.*

Community resources: Wiki and forums. Historically, the *C. elegans* WWW server has been a site that assembles various types of information on *C. elegans* and serves as an outlet for news and announcements. With the expansion of the worm research community, a more robust infrastructure to support such a portal has become necessary. New internet technologies have been developed in the last several years that have made it possible to transform such an outlet into a truly communal resource that all *C. elegans* researchers can freely contribute to. Although it is still in its infancy, the WormBase Wiki web site already carries many community-driven pages that detail experimental protocols, meeting announcements, tools for data mining and job postings. The number of such documents is expected to increase dramatically as worm researchers become more familiar with the Wiki concept (see the Wiki usage guidelines and objectives). Another exciting feature that is provided by Wiki is its support for community-based annotation. Each gene page in WormBase has a link to a corresponding Wiki page, enabling researchers to add their comments or to post data that have not yet been entered into WormBase. In addition, Wiki contains many WormBase- and WormBook-related documents including the user guide, FAQs, newsletters, links to several mailing lists, RSS feeds and commonly requested data sets.

A separate site that hosts Worm Community Forums for WormBase, WormBook and WormAtlas has also been launched recently. Unlike Wiki, which is intended for stable or slowly evolving documents, forums are designed to promote active discussions within the worm community. Neither Wiki nor the forums require special web skills or software, and all members of the worm community are encouraged to contribute actively to their development.

Experimental resources

C. elegans strains and nomenclature. The ability to carry out powerful genetic experiments is one of the main advantages of *C. elegans*. Over the years, classical mutagenesis screens have generated thousands of mutants with easily discernable phenotypes, and suppressor and enhancer screens have uncovered many more alleles that could not have been isolated in wild-type backgrounds. More sophisticated strategies are now being used to discover complex genetic interactions, such as screens for synthetic phenotypes. Improvements in targeted gene knockout technologies have also made *C. elegans* amenable to reverse genetics, and several large-scale projects aimed at generating deletion alleles for most predicted genes are underway, as described below. Many transgenic strains have also been constructed for studies of expression patterns and other biological questions. The *Caenorhabditis* Genetics Center (CGC) was established in 1978 to support this rapidly developing model system, and remains the only comprehensive resource that collects, maintains and distributes *C. elegans* strains. Frozen stocks allow permanent storage of and consequent unlimited access to ancestral material, making *C. elegans* an attractive model for evolutionary biology. CGC works in close collaboration with WormBase, which imports all information on available strains and enables users to locate these data through a specialized search interface.

Another crucial task of CGC is the coordination of *C. elegans* nomenclature. Historically, worm researchers have adhered to strict conventions that have ensured the uniformity of gene names, derived from either the mutant phenotype or the molecular nature of the gene product. This has virtually eliminated the confusions and clashes that are so common in other model organisms, and is proving to be even more valuable today as automated phenotypic data extraction and literature analyses become more common.

Synthetic phenotype
A phenotype that is produced when two or more mutations that have no discernable phenotypes on their own are combined.

Genomic and cDNA clones. The availability of genomic clones and the physical map of the *C. elegans* genome, together with the development of the germline transformation technique, have contributed greatly to the success of worm research by facilitating positional cloning of thousands of mutants isolated in genetic screens. Genomic clones are also indispensable reagents for a various applications, such as probe synthesis for RNAi analyses and generation of promoter fusions for expression studies. Original genomic clones are maintained at the Sanger Institute and are freely distributed. As an alternative to this cosmid-based collection, which displays some loss of viability among the original bacterial stocks and DNA rearrangements, a new library is being constructed by BC GSC and is distributed by Geneservice Ltd. This library utilizes a fosmid vector that is maintained at low copy number and shows low rearrangement frequency, which ensures the stability of the library. Its continuous genomic coverage eliminates the need for YAC-based clones that were previously used to bridge gaps between cosmids. Information on cosmid and fosmid clones is available through both WormBase and the BC GSC Fosmid Search interface.

Although genomic clones are perfectly suited for some experiments, spliced cDNA sequences are required for many others, including protein expression, *in vitro* transcription and synthesis of some RNAi probes. cDNA clones also serve as the ultimate tool for predicting gene structure, and the number of isolated clones correlates well with the gene-expression level in most cases. Many of these tasks are simplified by the extensive collection of cDNA clones generated by the Kohara laboratory at the National Institute of Genetics of Japan, which is accessible through NEXTDB.

Gene knockouts and transposon tagging. The ease of isolating genetic mutants is one of the indisputable advantages of *C. elegans* research. Hundreds of screens have generated thousands of alleles, which have been instrumental in unravelling many genetic pathways. However, mutants have been isolated for only about one-third of the ~20,000 *C. elegans* genes, which is insufficient for the comprehensive dissection of genetic networks.

Although gene-disruption by homologous recombination is not yet feasible in *C. elegans*, targeted gene knockouts have become possible owing to recent advances in reverse-genetics methods. Two projects, run by the *C. elegans* Gene Knockout Consortium (GKC) and the National BioResource Project of Japan (NBRP), aim to isolate deletion mutants for all *C. elegans* genes. Both use similar methods, and researchers can request the inactivation of particular genes. Mutant alleles that are generated by NBRP are stored in a local repository and are distributed directly by the project. NBRP requires strain recipients to provide feedback information such as descriptions of detected phenotypes so that this information can be made available to the research community. Researchers are also encouraged to submit the same information to WormBase. GKC deposits mutant strains to CGC, which makes them available to the research community without restrictions, and information that is generated by the GKC itself is also integrated into WormBase.

The recently launched project NemaGENETAG aims to generate a genome-wide collection of transposon-tagged *C. elegans* mutants. Such alleles can be used for the efficient generation of deletion mutants through imprecise excision of transposable elements, providing an alternative to the more laborious chemical-mutagenesis-based methods that are used by GKC and NBRP. It might also be possible to utilize such alleles for the introduction of exogenous DNA sequences into predetermined genomic locations, and methodologies that will enable such applications are being developed, including the promising MosTIC approach for genome engineering[31]. The project aims to generate ~40,000 insertion alleles, with ~4,000 alleles available already.

ORFeome and Promoterome databases. The essentially complete catalogue of *C. elegans* genes offers researchers an opportunity to move away from analyses of individual genes to studies that are focused on complex genetic interactions in the context of biological networks. Such approaches necessitate the development of new, easily customizable reagents that are suitable for large-scale functional studies. The ORFeome library, which was developed by Mark Vidal's laboratory, comprises full-length ORFs cloned into a vector that supports Gateway recombinational cloning and allows researchers to easily transfer a large number of inserts to the vectors that are appropriate for their studies[32,33]. The utility and convenience of this approach have been demonstrated by the construction of the *C. elegans* Interactome, discussed above[9]. ORFeome currently covers ~65% of *C. elegans* genes, with more then 12,500 available clones. With the continual improvement of gene predictions, the ORFeome library is being expanded with the aim of providing an almost complete coverage of *C. elegans* ORFs. The library can be searched via the WorfDB web site, and clones can be ordered through one of the two ORFeome distributors, Geneservice Ltd and OpenBiosystems. OpenBiosystems also offers an ORFeome-based library for RNAi analysis, as discussed below.

Another project that has been established by the Vidal group is Promoterome, the goal of which is to generate a library of promoter-carrying vectors that can be easily reused for various applications[34]. Promoterome aims to cover all *C. elegans* genes, facilitating functional analysis of biological networks. The project utilizes an enhanced Gateway cloning system that allows the generation of hybrid constructs that contain inserts from two entry clones. This enables, for example, the production of vectors containing promoter regions that have been fused to ORFs and GFP, facilitating high-throughput analyses of gene-expression patterns and subcellular protein localization. The current release of the Promoterome database contains clones for approximately 6,000 *C. elegans* genes, which are available through Geneservice Ltd and OpenBiosystems.

RNAi libraries. RNAi is one of the standard functional analysis methods in *C. elegans*, providing an alternative to studies of mutant alleles. To facilitate large-scale RNAi studies and promote reagent standardization, two RNAi libraries with broad genome coverage have recently been developed and made available to the *C. elegans* community. Both libraries use bacterial feeding for dsRNA delivery, which has become the method of choice for large-scale screens and is also frequently used for experiments that target individual genes[35]. The two libraries differ in the type of template that is used to produce dsRNA and the number of targeted genes, although this difference will probably be reduced in the future.

The first library, from the Ahringer group at the University of Cambridge, comprises almost 17,000 clones, covering ~87% of *C. elegans* genes. Library inserts were generated by PCR amplification of genomic DNA using primers that were originally designed for SMD microarrays, mentioned above[36,37]. The resulting PCR products represent unspliced genomic sequences that completely or partially overlap target genes. The Ahringer library is the most commonly used source of RNAi reagents, and is available as a complete set or as individual clones from Geneservice Ltd. The second library uses full-length ORF templates derived from the ORFeome library, which was described above, and covers approximately 11,000 genes. Several published studies have made use of this library, which is available from OpenBiosystems. Because the PCR primers used by both projects were designed against versions of the *C. elegans* genome that were annotated several years ago, clone names might no longer correspond to the current gene names. It is therefore important to check current information on clone-to-gene mapping, which is available from WormBase.

Microarray resources. Custom *C. elegans* microarray chips, produced by Stuart Kim's group, have been used extensively since the emergence of microarray technology. Today, they are gradually being replaced by a number of commercially available microarray platforms. Affymetrix was the first company to develop a commercial expression-profiling array with genome-wide coverage for *C. elegans* research, and their GeneChip array contains probe sets designed against about 22,000 transcripts. Another microarray that covers more then 21,000 *C. elegans* transcripts has been recently released by Agilent Technologies. Although both are manufactured using *in situ* synthesis technologies that ensure high quality of features, the Affymetrix array uses multiple (11 on average) 25-mer oligonucleotides for each feature, whereas in Agilent arrays each feature is represented by a single 60-mer oligonucleotide. A microarray chip developed by WashU GSC features pre-synthesized 60-mer oligonucleotides targeting approximately 21,500 worm transcripts that are deposited on glass slides using a significantly cheaper, although slightly less accurate, pin printing technology. NimbleGen is yet another source of *C. elegans* arrays, offering a high-density 60-mer array containing ~390,000 probes that represent the whole *C. elegans* genome (16 probes for each gene on average).

In addition to chips that are designed for gene-expression profiling, a new type of array that carries uniformly spaced oligonucleotides (50–200 bp apart) covering the whole genome is becoming widely used. These tiling arrays pave the way for new approaches such as

*Mos*TIC
(*Mos1* excision-induced transgene-instructed gene conversion). In *Mos*TIC experiments, a double-strand break (DSB) is introduced by excision of the *Mos1* transposable element, which is then repaired using a transgene containing sequences that are homologous to the DSB flanking regions as a template. Modifications that are present in the transgene are incorporated into the genome to enable efficient genome engineering.

comparative genomic hybridization to measure DNA copy number, ChIP-on-Chip analyses for mapping transcription factor binding sites, and the identification of novel transcripts. Affymetrix, Agilent and NimbleGen now produce *C. elegans* tiling arrays for such applications. NimbleGen also allows users to custom design high-density arrays to their own specifications, adding significant flexibility to both expression-profiling studies and tiling-array-based experiments.

Finally, some researchers might choose to print their own arrays, which can be advantageous — for example, when the expression levels of only a subset of genes need to be determined. PCR-amplified genomic regions are the most commonly used type of probe for this purpose, and a standard set of primers that has been used for production of SMD chips is available from Research Genetics.

Vectors for C. elegans research. Numerous *C. elegans*-specific vectors have been accumulated by the worm community. Many were developed in Andrew Fire's laboratory and are available as the Fire Lab Vector Kit. The kit also contains many plasmids that have been contributed by other researchers, making it the most comprehensive collection of *C. elegans* vectors. The current version comprises 288 plasmids that can be used for a variety of genetic experiments. A substantial part of the library is devoted to vectors for studying gene-expression patterns and protein subcellular localization, including plasmids for creating transcriptional and translational gene fusions with several versions of GFP and its colour variants. Vectors that target proteins to specific cellular compartments or that are designed for studying enhancer elements in the context of a minimal promoter are also provided. The RNAi libraries that are described above were constructed using another vector from the kit, which drives expression of dsRNA in *Esherichia coli*. This plasmid is commonly used for custom RNAi constructs, and comes with a set of controls for bacteria-mediated RNAi. The library contains many other vectors, such as hairpin constructs for testing the ability of promoters to trigger RNAi *in vivo* and plasmids for Smg-dependent control of gene expression. The kit is supplied with extensive documentation and sequence files for each of the 288 plasmids and is distributed by Addgene.

To enable germline gene expression, which is not easily attainable using vectors from the Fire laboratory collection, Geraldine Seydoux's group has made constructs that drive transgene expression in the adult germ line of *C. elegans* hermaphrodites and in early embryos. Two types of plasmid are available: the first utilizes standard restriction-enzyme-based cloning techniques to produce GFP fusions, whereas the second involves the Gateway system[32], allowing high-throughput cloning. Some of these latter vectors include an *unc-119*-rescuing fragment, which allows the construction of plasmids for biolistic transformation, which has been shown to yield low-copy integrants and is an efficient way to obtain stable transgene expression in the maternal germ line[38].

Conclusions and future directions

The range of *C. elegans* tools and resources covers a broad area of research interests, from bioinformatic analyses to wet-lab experiments. They undergo constant improvement and expansion, continually increasing their usefulness. To give some examples: the PhenoBlast search, which was originally developed for analyses of data obtained in an early embryonic RNAi screen[39], has recently been extended to several new data sets, dramatically improving its power; a new algorithm for multispecies clustering of orthologous genes, MultiParanoid, is being developed by the InParanoid group[40]; the NemaGENETAG project, which has already produced thousands of *Mos1* transposon-tagged alleles is expected to start generating mutants using the *Minos* transposable element in the near future; and the ORFeome clone collection is being continually updated to incorporate the latest gene predictions, driving the expansion of Interactome data.

Although the value of the available resources cannot be overestimated, there is room for improvement; for example, the availability of several databases devoted to phenotypic information, each using a proprietary phenotype-scoring system, illustrates the necessity for the convergence of methods of data description to facilitate comparisons across data sets. The capture of phenotypic data represents a particularly important case, as a large proportion of biological literature, especially in genetically tractable model organisms such as *C. elegans*, is devoted to phenotype description, which often serves as the primary source of information for inferring the biological functions of a gene. The use of ontology-based systems holds great promise for standardizing phenotypic descriptions, and several groups, including WormBase, are working on the development of such ontologies. In another example, the large amounts of expression-profiling data that are generated by microarray and SAGE analyses offer great potential for exploring relationships between genes, deducing gene functions and reconstructing biological pathways. Although many software packages and web-based resources already exist, their further development will be necessary to unlock the full potential of these data. The power of cell-biology studies in *C. elegans* will be greatly improved by the establishment of cell lines and the development of a dedicated antibody resource, which is currently lacking. Other areas of research that will benefit significantly from the development of new tools include comparative sequence analyses and the elucidation of regulatory sites, targeted gene inactivation and tissue-specific transgene expression using bipartite systems such as GAL4/UAS, identification of non-coding RNA genes and analyses of their functions, reconstruction of genetic networks, and data integration.

The emergence of new technologies will also require radically new approaches and tools. In the past several years, efforts have been made to develop methods involving high-throughput two-dimensional electrophoresis and chromatography–mass spectrometry for proteomic analysis. Such techniques have the potential of not only

Comparative genomic hybridization
Comparative genomic hybridization measures differences in DNA copy numbers between two genomes, usually between a reference and a sample genome. This method can be used to detect deletions, duplications and translocations, and to map associated breakpoints with unprecedented precision and speed.

ChIP-on-Chip
ChIP-on-Chip combines chromatin immunoprecipitation (ChIP) with microarray analysis, most commonly using uniform tiling arrays that cover the whole genome or chips that focus on known promoter regions. ChIP-on-Chip allows the rapid identification of binding sites of DNA-binding proteins, such as transcription factors and histones.

Smg-dependent control of gene expression
mRNAs containing premature stop codons are rapidly degraded by the nonsense-mediated decay pathway, which in *C. elegans* is represented by members of the Smg gene family. This makes it possible to control the expression of transgenes carrying aberrant 3′ UTRs by regulating the Smg-pathway activity, for example, by using a temperature-sensitive allele of *smg-1*.

Polony-based DNA sequencing
Polonies (polymerase colonies) are created either on a surface such as a slide or gel, or in emulsion attached to beads. PCR within each polony produces clusters that contain millions of template copies, which grow like bacterial colonies. The amplified template can then be sequenced by synthesis or ligation in a highly parallel fashion, enabling high-throughput genome sequencing.

measuring the concentrations of all proteins that are present in the organism, but also of identifying protein modifications, such as phosphorylation and myristoylation, that are central to the regulation of many biological processes. Many of these technologies are still in their initial stages of development, and no major resource for *C. elegans* proteomics has been developed. This situation will certainly change, and several medium-scale studies have already been published[41,42]. *Caenorhabditis elegans* metabolic pathways are also poorly studied, but recent advances in NMR and mass-spectrometry-based metabolomic data-acquisition techniques promise rapid improvements in this area, which will require the development of specialized databases and analysis tools. Another challenge that will face biologists in the future is how to leverage the extensive knowledge and resources of *C. elegans* to help understand and ultimately control the many nematode parasites. High-throughput single molecule and polony-based DNA-sequencing techniques, advances in tiling array technology and methods for studying protein–DNA interactions are just a few additional examples of rapidly developing methodologies that will create new opportunities — and challenges — for *C. elegans* biologists.

1. *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
This cornerstone paper describes the sequencing of the *C. elegans* genome, the first from a multicellular organism. The project produced the essentially complete catalogue of worm genes, enabling functional and genomic studies. Collaboration between the Sequencing Consortium and the *C. elegans* research community was essential for cross-correlation of physical and genetic maps, making possible much of the genetic research that relied heavily on positional cloning of mutants identified in genetic screens.
2. Stein, L. *et al.* The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol.* **1**, 166–192 (2003).
3. Wood, W. B. (ed.) *The Nematode Caenorhabditis elegans* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1988).
4. Riddle, D. L., Blumenthal, T., Meyer, B. J. & Priess, J. R. (eds) *C. Elegans* II (Cold Spring Harbor Laboratory Press, Plainview, 1997).
5. Girard, L. R. *et al.* WormBook: the online review of *Caenorhabditis elegans* biology. *Nucleic Acids Res.* **35**, D472–D475 (2006).
WormBook provides up-to-date reviews on all aspects of *C. elegans* biology and the experimental methods that are used to study this organism. The paper describes the goals of the project, the current content of WormBook and its future directions.
6. Bieri, T. *et al.* WormBase: new content and better access. *Nucleic Acids Res.* **35**, D506–D510 (2007).
The article is the latest instalment in a series of articles that focus on the most recent changes and additions to WormBase. As the primary database for *C. elegans* and related nematodes, WormBase accumulates comprehensive genetic, genomic and functional information on *C. elegans* genes and provides links to other databases and resources such as PubMed and Entrez Gene.
7. Gunsalus, K. C., Yueh, W. C., MacMenamin, P. & Piano, F. RNAiDB and PhenoBlast: web tools for genome-wide phenotypic mapping projects. *Nucleic Acids Res.* **32**, D406–D410 (2004).
In-depth studies of the phenotypes that are produced in genome-wide RNAi screens hold great promise for large-scale functional analyses of the genome. RNAiDB exemplifies the controlled-vocabulary-based approach to phenotype description, which enables computational analyses of observed defects, including phenotypic clustering, classification and gene searches involving phenotypic signature similarity.
8. Sonnichsen, B. *et al.* Full-genome RNAi profiling of early embryogenesis in *Caenorhabditis elegans*. *Nature* **434**, 462–469 (2005).
9. Li, S. *et al.* A map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540–543 (2004).
Gene- and protein-interaction networks and their integration with other types of large-scale data sets serve as powerful hypotheses generating tools. This article describes the first attempt to build a protein-interaction network in *C. elegans* using a high-throughput yeast two-hybrid approach.
10. Hermjakob, H. *et al.* IntAct: an open source molecular interaction database. *Nucleic Acids Res.* **32**, D452–D455 (2004).
11. Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34**, D535–D539 (2006).
12. Deplancke, B., Dupuy, D., Vidal, M. & Walhout, A. J. A Gateway-compatible yeast one-hybrid system. *Genome Res.* **14**, 2093–2101 (2004).
13. Barrasa, M. I., Vaglio, P., Cavasino, F., Jacotot, L. & Walhout, A. J. EDGEdb: a transcription factor–DNA interaction database for the analysis of *C. elegans* differential gene expression. *BMC Genomics* **8**, 21 (2007).
14. Maeda, I., Kohara, Y., Yamamoto, M. & Sugimoto, A. Large-scale analysis of gene function in *Caenorhabditis elegans* by high-throughput RNAi. *Curr. Biol.* **11**, 171–176 (2001).
15. Mounsey, A., Bauer, P. & Hope, I. A. Evidence suggesting that a fifth of annotated *Caenorhabditis elegans* genes may be pseudogenes. *Genome Res.* **12**, 770–775 (2002).
16. Hope, I. A. *et al.* Feasibility of genome-scale construction of promoter::reporter gene fusions for expression in *Caenorhabditis elegans* using a multisite Gateway recombination system. *Genome Res.* **14**, 2070–2075 (2004).
17. Ball, C. A. *et al.* The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Res.* **33**, D580–D582 (2005).
18. Kim, S. K. *et al.* A gene expression map for *Caenorhabditis elegans*. *Science* **293**, 2087–2092 (2001).
19. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255 (2003).
20. Barrett, T. *et al.* NCBI GEO: mining millions of expression profiles — database and tools. *Nucleic Acids Res.* **33**, D562–D566 (2005).
21. Parkinson, H. *et al.* ArrayExpress — a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* **33**, D553–D555 (2005).
22. Kapushesky, M. *et al.* Expression Profiler: next generation — an online platform for analysis of microarray data. *Nucleic Acids Res.* **32**, W465–W470 (2004).
23. Jones, S. J. *et al.* Changes in gene expression associated with developmental arrest and longevity in *Caenorhabditis elegans*. *Genome Res.* **11**, 1346–1352 (2001).
24. McKay, S. J. *et al.* Gene expression profiling of cells, tissues, and developmental stages of the nematode *C. elegans*. *Cold Spring Harb. Symp. Quant. Biol.* **68**, 159–169 (2003).
25. Halaschek-Wiener, J. *et al.* Analysis of long-lived *C. elegans daf-2* mutants using serial analysis of gene expression. *Genome Res.* **15**, 603–615 (2005).
26. Luan, C. H. *et al.* High-throughput expression of *C. elegans* proteins. *Genome Res.* **14**, 2102–2110 (2004).
27. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
28. Muller, H. M., Kenny, E. E. & Sternberg, P. W. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.* **2**, e309 (2004).
Textpresso is a full-text literature search platform that was developed specifically for *C. elegans* research. This paper outlines the logic behind its powerful ontology-based search engine and discusses new possibilities that are offered by the system.
29. Sommer, R. J. As good as they get: cells in nematode vulva development and evolution. *Curr. Opin. Cell Biol.* **13**, 715–720 (2001).
30. Mitreva, M. & Jasmer, D. P. Biology and genome of *Trichinella spiralis* in *WormBook* (ed. The *C. elegans* Research Community) 23 Nov 2006 (doi: 10.1895/wormbook.1.124.1).
31. Robert, V. & Bessereau, J. L. Targeted engineering of the *Caenorhabditis elegans* genome following *Mos1*-triggered chromosomal breaks. *EMBO J.* **26**, 170–183 (2007).
32. Walhout, A. J. *et al.* Gateway recombinational cloning: application to the cloning of large numbers of open reading frames or ORFeomes. *Methods Enzymol.* **328**, 575–592 (2000).
33. Lamesch, P. *et al.* *C. elegans* ORFeome version 3. 1: increasing the coverage of ORFeome resources with improved gene predictions. *Genome Res.* **14**, 2064–2069 (2004).
34. Dupuy, D. *et al.* A first version of the *Caenorhabditis elegans* Promoterome. *Genome Res.* **14**, 2169–2175 (2004).
35. Timmons, L. Delivery methods for RNA interference in *C. elegans*. *Methods Mol. Biol.* **351**, 119–125 (2006).
36. Reinke, V. *et al.* A global profile of germline gene expression in *C. elegans*. *Mol. Cell* **6**, 605–616 (2000).
Microarray analyses have become an essential part of biological research. The first *C. elegans* microarray-based expression-profiling study, which paved the way for many functional studies in *C. elegans*, is presented in this paper.
37. Jiang, M. *et al.* Genome-wide analysis of developmental and sex-regulated gene expression profiles in *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA* **98**, 218–223 (2001).
38. Praitis, V., Casey, E., Collar, D. & Austin, J. Creation of low-copy integrated transgenic lines in *Caenorhabditis elegans*. *Genetics* **157**, 1217–1226 (2001).
39. Piano, F. *et al.* Gene clustering based on RNAi phenotypes of ovary-enriched genes in *C. elegans*. *Curr. Biol.* **12**, 1959–1964 (2002).
40. Alexeyenko, A., Tamas, I., Liu, G. & Sonnhammer, E. L. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics* **22**, e9–e15 (2006).
41. Husson, S. J., Clynen, E., Baggerman, G., De Loof, A. & Schoofs, L. Discovering neuropeptides in *Caenorhabditis elegans* by two dimensional liquid chromatography and mass spectrometry. *Biochem. Biophys. Res. Commun.* **335**, 76–86 (2005).
42. Wielsch, N. *et al.* Rapid validation of protein identifications with the borderline statistical confidence via *de novo* sequencing and MS BLAST searches. *J. Proteome Res.* **5**, 2448–2456 (2006).
43. Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **33**, D54–D58 (2005).
44. Hubbard, T. *et al.* Ensembl 2005. *Nucleic Acids Res.* **33**, D447–D453 (2005).
45. Karolchik, D. *et al.* The UCSC Genome Browser Database. *Nucleic Acids Res.* **31**, 51–54 (2003).
46. Kent, W. J. & Zahler, A. M. Conservation, regulation, synteny, and introns in a large-scale *C. briggsae*–*C. elegans* genomic alignment. *Genome Res.* **10**, 1115–1125 (2000).
47. Schwartz, S. *et al.* Human–mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).

48. Jurka, J. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.* **16**, 418–420 (2000).
49. Kent, W. J. *et al.* Exploring relationships and mining data with the UCSC Gene Sorter. *Genome Res.* **15**, 737–741 (2005).
50. Bairoch, A. *et al.* The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **33**, D154–D159 (2005).
51. O'Brien, K. P., Remm, M. & Sonnhammer, E. L. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* **33**, D476–D480 (2005).
52. Chen, F., Mackey, A. J., Stoeckert, C. J. Jr & Roos, D. S. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* **34**, D363–D368 (2006).
53. Li, H. *et al.* TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.* **34**, D572–D580 (2006).
54. Joshi-Tope, G. *et al.* Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* **33**, D428–D432 (2005).
55. Krieger, C. J. *et al.* MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* **32**, D438–D442 (2004).
56. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**, D277–D280 (2004).
57. Lemer, C. *et al.* The aMAZE LightBench: a web interface to a relational database of cellular processes. *Nucleic Acids Res.* **32**, D443–D448 (2004).
58. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.* **25**, 25–29 (2000).
59. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **32**, D138–D141 (2004).
60. Finn, R. D. *et al.* Pfam: clans, web tools and services. *Nucleic Acids Res.* **34**, D247–D251 (2006).
61. Griffiths-Jones, S. *et al.* Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, D121–D124 (2005).
62. Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A. & Enright, A. J. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* **34**, D140–D144 (2006).
63. Reinhart, B. J. *et al.* The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* **403**, 901–906 (2000).
64. Lee, R. C., Feinbaum, R. L. & Ambros, V. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* **75**, 843–854 (1993).
65. Wightman, B., Ha, I. & Ruvkun, G. Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans. Cell* **75**, 855–862 (1993).
66. Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294**, 858–862 (2001).

Acknowledgements

We thank G. Schindelman for helpful discussions and assistance. I.A. is supported by a grant to WormBase from the US National Human Genome Research Institute (P41-HG02223). P.W.S. is an investigator with the Howard Hughes Medical Institute, Pasadena, USA.

Competing interests statement

The authors declare no competing financial interests.

DATABASES

The following terms in this article are linked online to:
Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
air-2 | *mpk-1*
Access to this links box is available online.