WormBase
User's Guide

Updated: June 10, 2003

Questions or Comments:
wormbase@caltech.edu

TABLE OF CONTENTS

General Description 3
Release Notes 4
Display Tools (Sub Menu) 6
Basic Search Tools 8
Simple Search 8
Browse Genome 10
Blast Search 13
Batch Downloads 16
Marker/Strain Search 17
Genetic Map 18
Advanced Search Tools 21
Advanced Search 21
ePCR 23
Gene Ontology 25
Cell Search 26
Neuron Search 29
Expression Search.....31
Expr_profile Search 32
RNAi Search 33
Interval Search 35
Genome Dumper 36
Individual Pages 37
Gene Page 37
Sequence Page 40
Gene Ontology Page 42
Microarray Page 43
Person Page 44
Other Resources provided by WormBase 46
How to Obtain Reagents 46

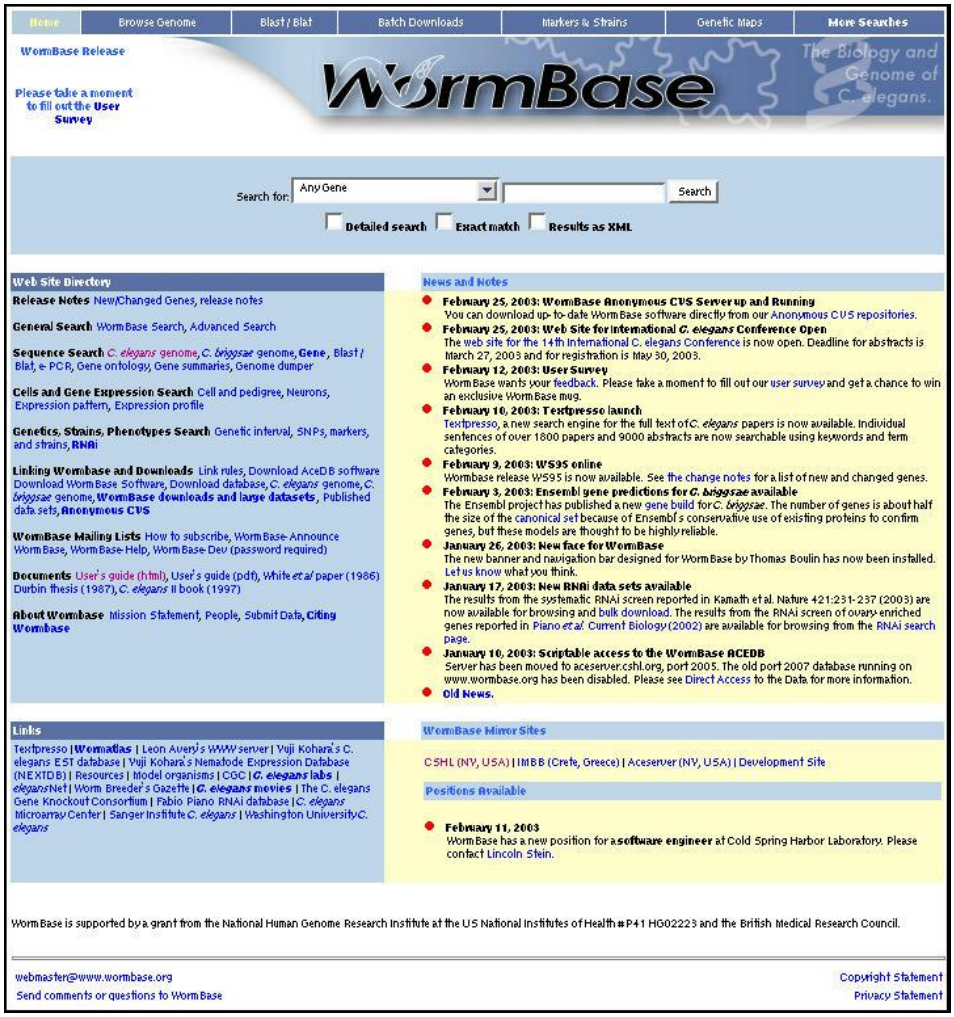
An Overview of The WormBase Homepage

GENERAL DOCUMENTATION

CAVEAT: ALWAYS UNDER CONSTRUCTION

At this writing, WormBase is very much a dynamic, ongoing project. A good consequence of this is that new features and options are added and old ones are removed at a noticeable rate. A drawback, however, is that documentation written for a given feature of the WormBase Web interface may be obsolete by the time you are reading this text.

Therefore, this documentation is based upon the main WormBase web site at Cold Spring Harbor Laboratories (http://wormbase.org). The Greek mirror (http://worm.imbb.forth.gr/), at this writing, has many of the same features as the main site but lags behind the main site in some respects; that is because the Greek mirror represents the last version of the WormBase web site that was stable enough to be ported by non-WormBase staff.



NEW RELEASE AND RELEASE NOTES

WormBase is updated every two weeks, with each release assigned a new version number (for example: WS10, ... WS51, ... WS97). Release notes can be accessed directly from the WormBase home page: see the help page for Release notes.

QUICK SEARCH:

Select a class from the pull-down menu to "Search for", and then type key words in the box.
If you leave the search box blank, everything from the selected class will appear as the search result. For example: if you select class "Cell", and type nothing in search box, all *C. elegans* cells will appear in the search result.
For anything you are not sure of, try using * as the wild card insertion; it can be placed anywhere in the keyword. Example: if you type "ad*", all data that begin with "ad" will appear in the search result.

What can be typed into the "Quick Search" box?
The following objects can be entered in the search box, which is case sensitive and space sensitive. You can search for a certain class such as "Any Gene" or "Author", or search for "Anything", which will return related objects from all classes.

- Any Gene Example: lin-3
- Author Example: Lee RYN
- Cell Example: ADFL

- Genbank Accession Number Example: AL032609

See more details on the explanation of Quick Search tool.

MENU FOR WORMBASE

The WormBase menu includes five basic options on the top of the WormBase homepage. More advanced search tools can be found under the "More Searches" option.

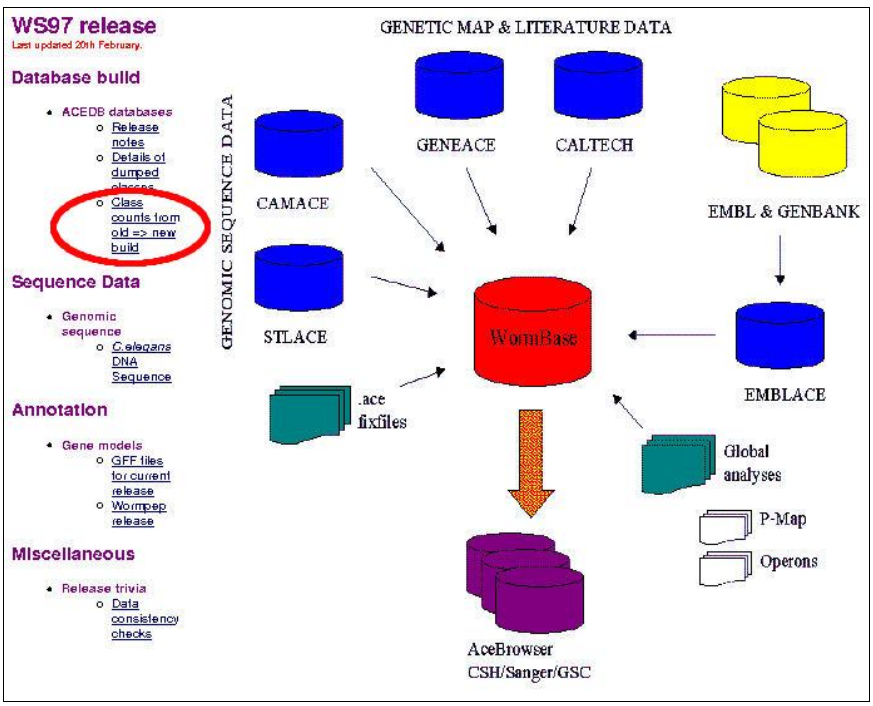
To learn how to use these search tools, please read the User's Guide help pages for Basic Search Tools and Advanced Search Tools.

Sometimes a submenu will appear when you search for certain items such as genetic map, sequence report or cell lineage. The submenu usually is for different displays of information, for example, "Tree display" versus "Graphic display", etc. For more information, please read the help page for Sub Menu.

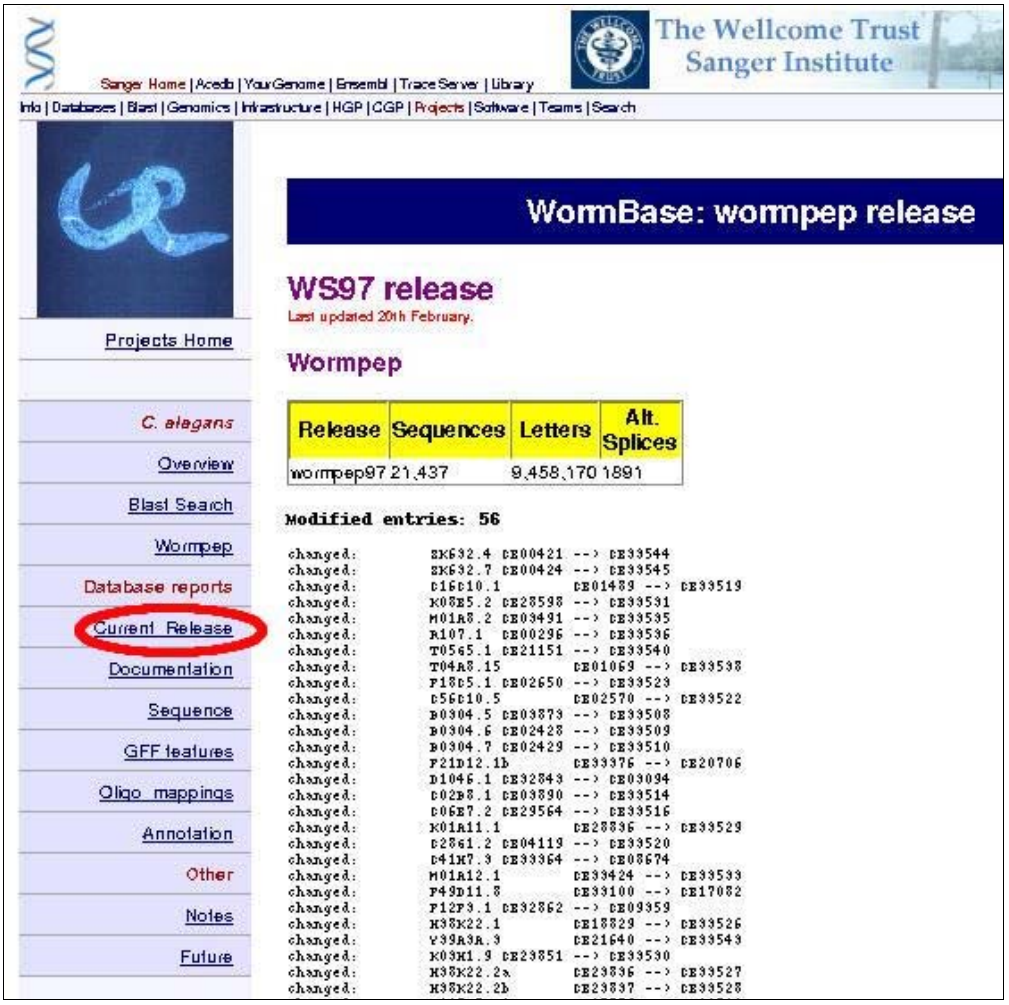
Release Notes

How can users find out what is new in each WormBase release? Actually all the information is in the notes for each new WS release, which can be directly accessed from the WormBase homepage.

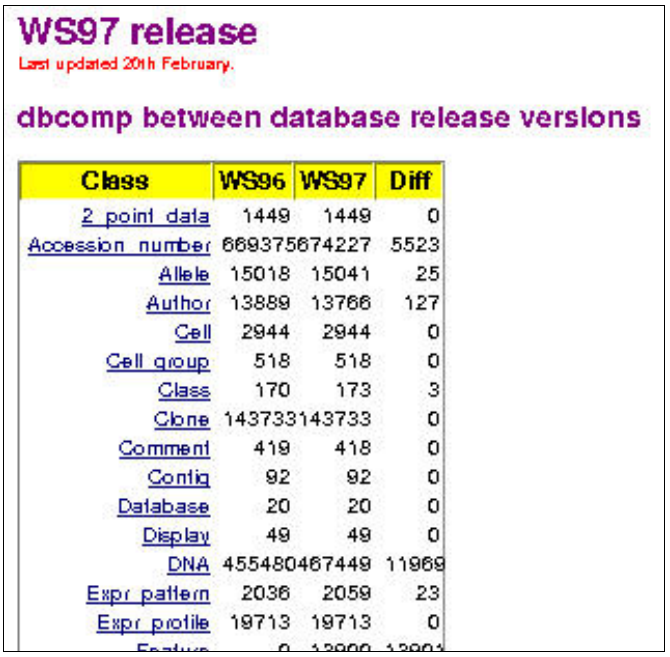
Under the Web Site Directory, go to Release Notes and click on New/Changed Genes. This will show the latest wormpep release. To access additional information about the latest WormBase release, go to the Database Reports heading to the left of the wormpep report and click on Current_Release. You will see a diagram showing how WormBase data are generated from multiple small databases and what the data pipeline looks like.



To the left of the diagram, users can access more specific information by clicking on Current_Release. ►



The "Class counts from old => new build" under "ACEDB databases" is very helpful. It is a straightforward list telling people the changes on each class of data. For example, in the release note for WS97, Class counts showed significant increases in Homol_data, Microarray_results and person_ name, meaning there are new data for these classes. ▼



WS97 release

Last updated 20th February.

dbcomp between database release versions

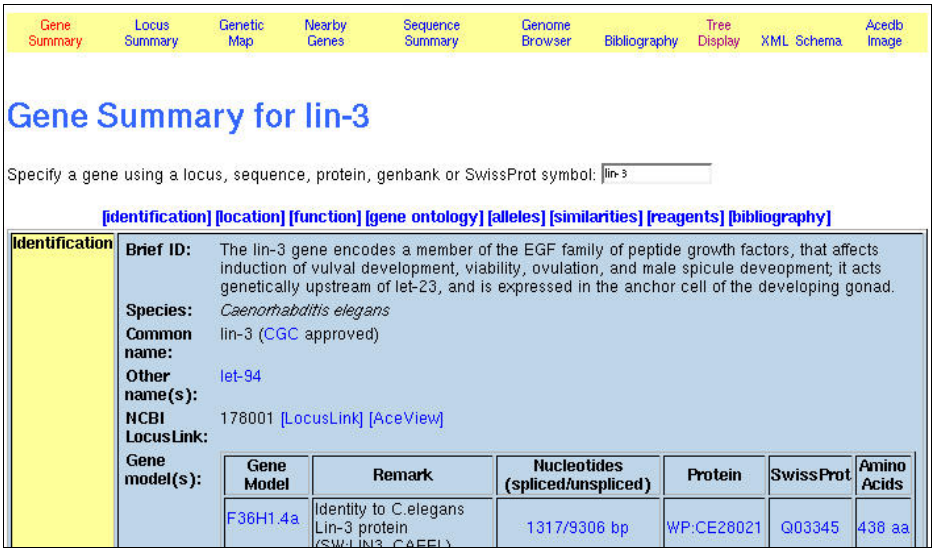
Class	WS96	WS97	Diff
2_point_data	1449	1449	0
Accession_number	669375	674227	5523
Allele	15018	15041	25
Author	13889	13766	127
Cell	2944	2944	0
Cell_group	518	518	0
Class	170	173	3
Clone	143733	143733	0
Comment	419	418	0
Contig	92	92	0
Database	20	20	0
Display	49	49	0
DNA	455480	467449	11969
Expr_pattern	2036	2059	23
Expr_profile	19713	19713	0

Various Displays of WormBase Data:

Most result pages in WormBase contain a submenu below the main menu and WormBase logo. The submenu usually is for different kinds of display. There are three major types of selections:

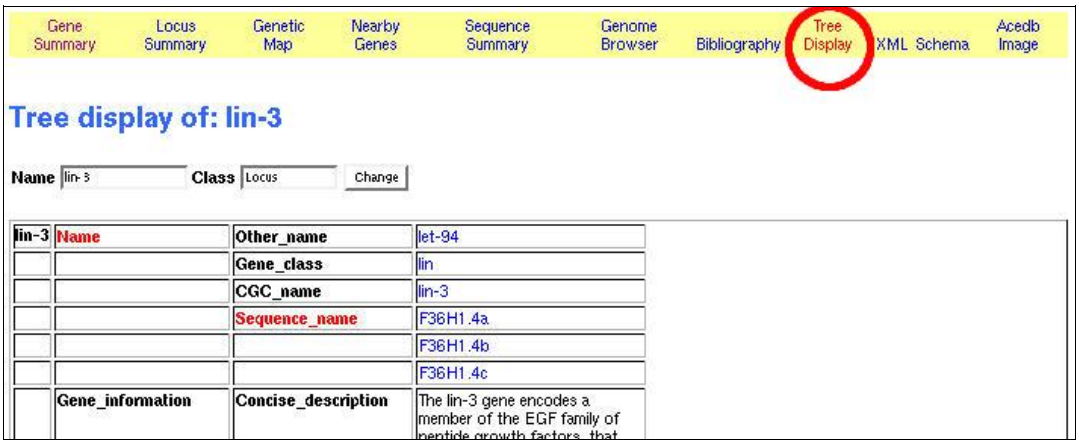
HTML Display

This kind of display is shown by default. It has a nice-looking graphic interface. The title of the page depends on the data type being displayed. For example, in a page for Expr_pattern, it is called "Expression Pattern", while in other places it may be called "Cell Report" or "Clone Report".



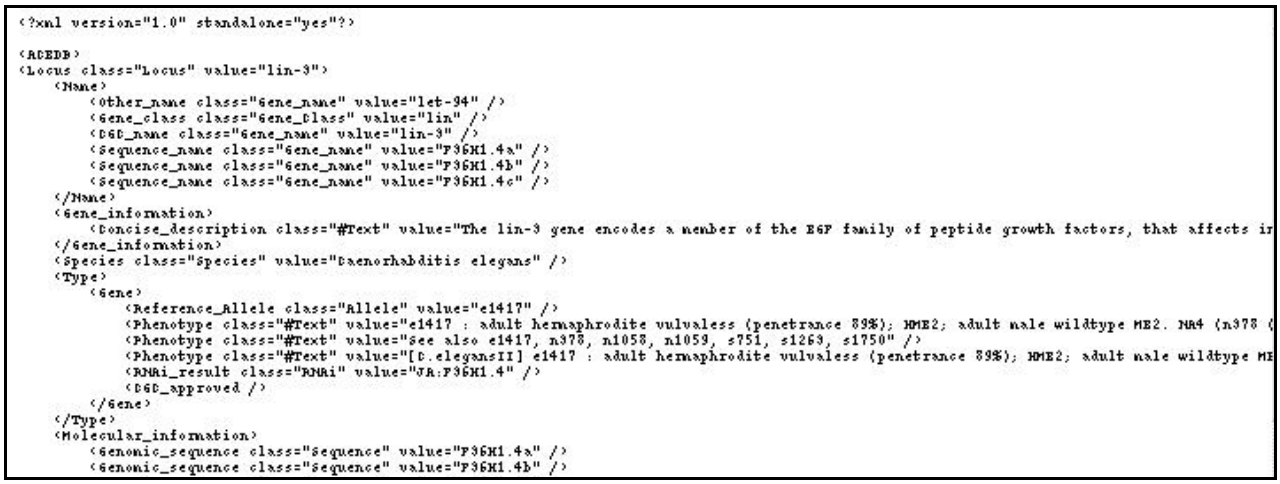
Tree Display

The Tree Display shows the data relationships according to each ACeDB data model. Shown in the diagram: red color shows tags, blue color indicates data fields that are linked elsewhere, and black color indicates text data fields.



XML

It shows the source XML (eXtensible Markup Language) code of the current page.



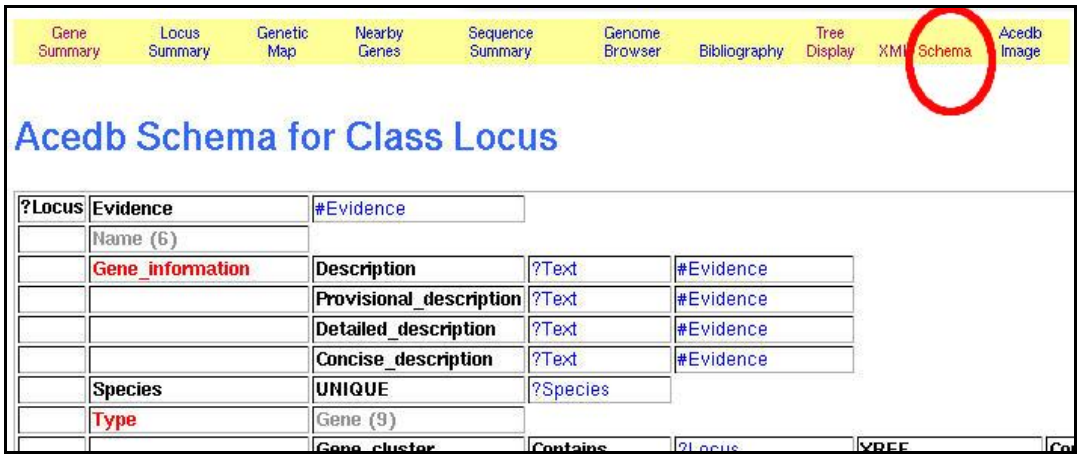
```
<?xml version="1.0" standalone="yes"?>

<ACEDB>
<Locus class="Locus" value="lin-3">
    <Name>
        <Other_name class="Gene_name" value="let-94" />
        <Gene_class class="Gene_Class" value="lin" />
        <CGC_name class="Gene_name" value="lin-3" />
        <Sequence_name class="Gene_name" value="F36H1.4a" />
        <Sequence_name class="Gene_name" value="F36H1.4b" />
        <Sequence_name class="Gene_name" value="F36H1.4c" />
    </Name>
    <Gene_information>
        <Concise_description class="#Text" value="The lin-3 gene encodes a member of the EGF family of peptide growth factors, that affects in
    </Gene_information>
    <Species class="Species" value="Caenorhabditis elegans" />
    <Type>
        <Gene>
            <Reference_Allele class="Allele" value="e1417" />
            <Phenotype class="#Text" value="e1417 : adult hermaphrodite vulvaless (penetrance 89%); MME2; adult male wildtype ME2. MA4 (n378 (
            <Phenotype class="#Text" value="See also e1417, n378, n1058, n1059, s751, s1263, s1750" />
            <Phenotype class="#Text" value="[C.elegansII] e1417 : adult hermaphrodite vulvaless (penetrance 89%); MME2; adult male wildtype ME
            <RNAi_result class="RNAi" value="JA:F36H1.4" />
            <CGC_approved />
        </Gene>
    </Type>
    <Molecular_information>
        <Genomic_sequence class="Sequence" value="F36H1.4a" />
        <Genomic_sequence class="Sequence" value="F36H1.4b" />
```

ACeDB Schema

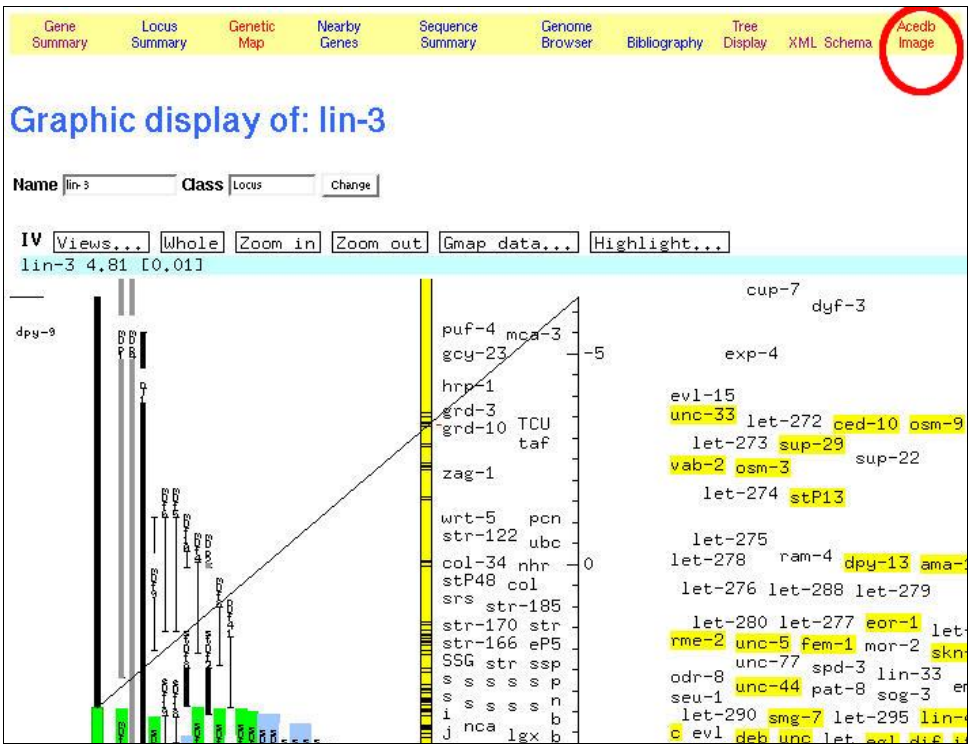
The Schema submenu shows the data model for the displaying class. For example, the "Schema" on Cell Report displays the Cell model, the "Schema" on Sequence Report displays the Sequence model...

This is a very important feature for curators. To understand more about ACeDB, the backend database of WormBase, please visit http://www.acedb.org.



ACeDB Image

This is the old display for the Acedb interface, but the links do not work.



Basic Search Tools

On the top of the WormBase homepage, the menu includes five basic search options and a link to "More Searches".

1. Simple Search
2. Browse Genome
3. Blast Search
4. Batch Downloads
5. Marker/Strain Search
6. Genetic Map

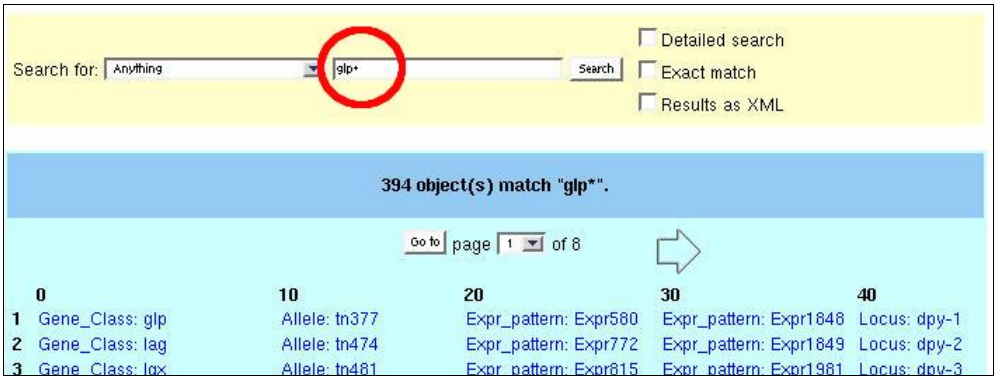
Simple Search

A simple search allows users to search WormBase for keywords contained in the database objects. The search can be broad (Anything) or can be restricted by searching for:

1) a particular object type (Author, Cell, Sequence, etc.) or
2) by using the Exact Match option.
The search can be made more comprehensive by using the Detailed Search option. This expands the search to include full text of all paper and meeting abstracts.

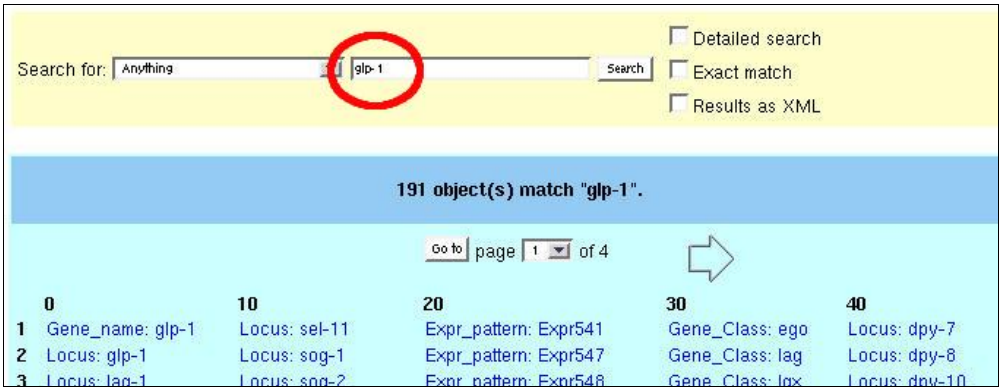
SEARCH WITH THE WILD CARD (*)

As an example, let's search using the key word glp*. Use of the wild card character * allows one to search for any objects that contain the key word glp. When searching for "Anything" in WS97, this particular search produces 394 objects.



SEARCH FOR A SPECIFIC OBJECT

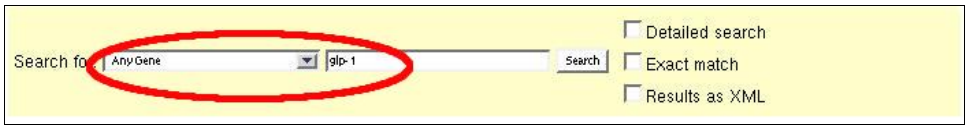
Now we can restrict the search by looking specifically for the glp-1 gene. The search result is now narrowed to 191 hits in WS97 when searching for "Anything".

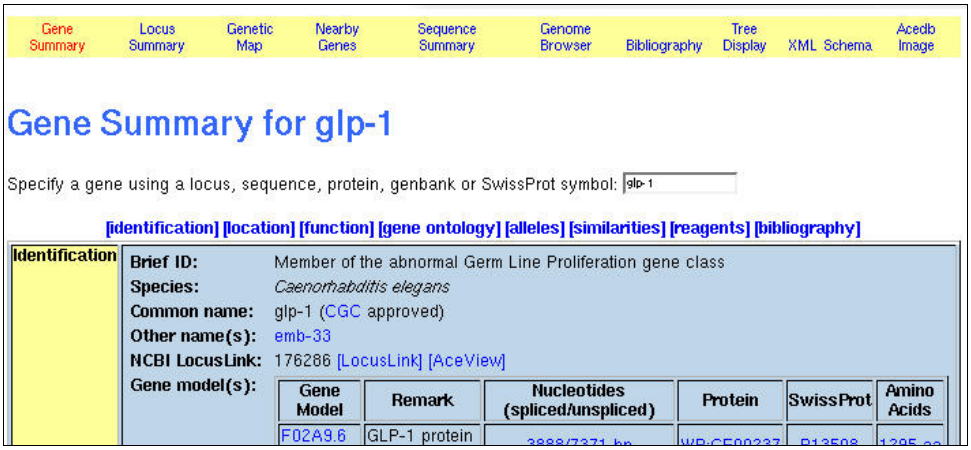


SEARCH FOR A PARTICULAR TYPE OF INFO

To look for a particular type of information, for example, GLP-1 protein sequence, we can restrict the search by searching only for Protein, *C. elegans*.

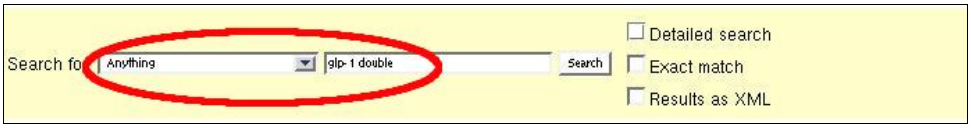
This results in a search with only one hit.

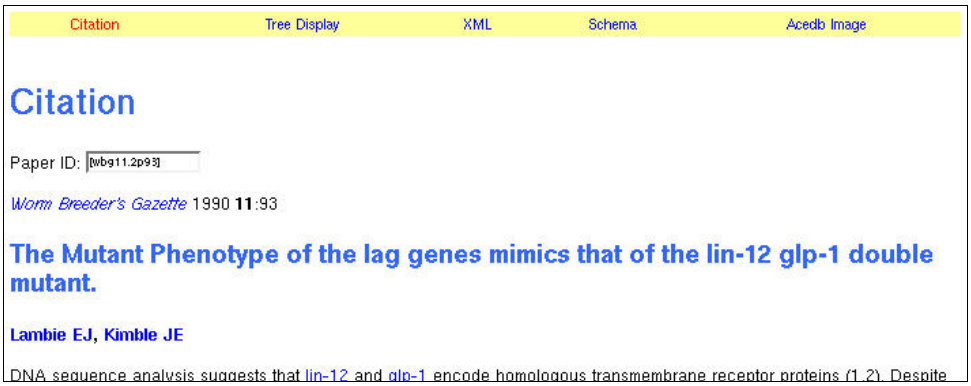




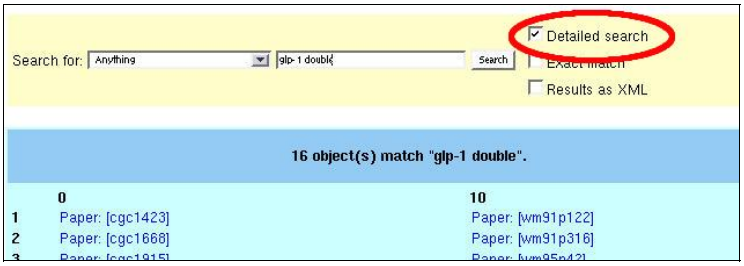
DETAILED SEARCH

Selecting the Detailed search option allows the user to expand the search to include full text of all paper and meeting abstracts as well as Worm Breeder's Gazette (wbg) articles. For example, if we search using the phrase "glp-1 double", we get only one hit in WS97. This wbg article has the phrase "glp-1 double" in its title.



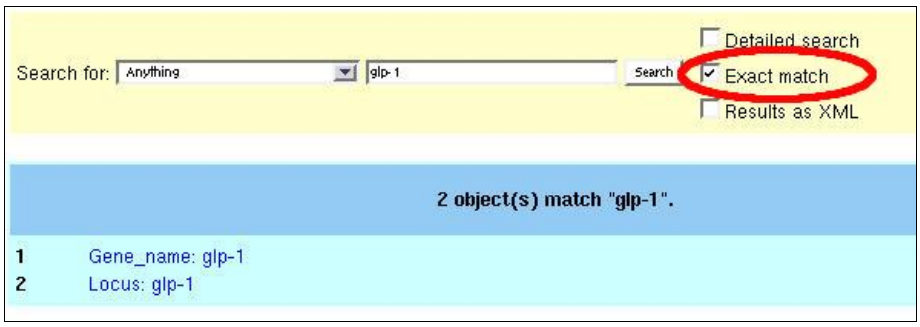


If we now expand the search by selecting the Detailed search option, we get 16 hits in WS97 where the phrase glp-1 double occurs in the body of the abstract or article.



EXACT MATCH

A search can be restricted by selecting the Exact Match option. If we search for Anything that is an exact match for glp-1, we get two hits that correspond to only two objects, Gene_name and Locus, that are uniquely identified as glp-1.



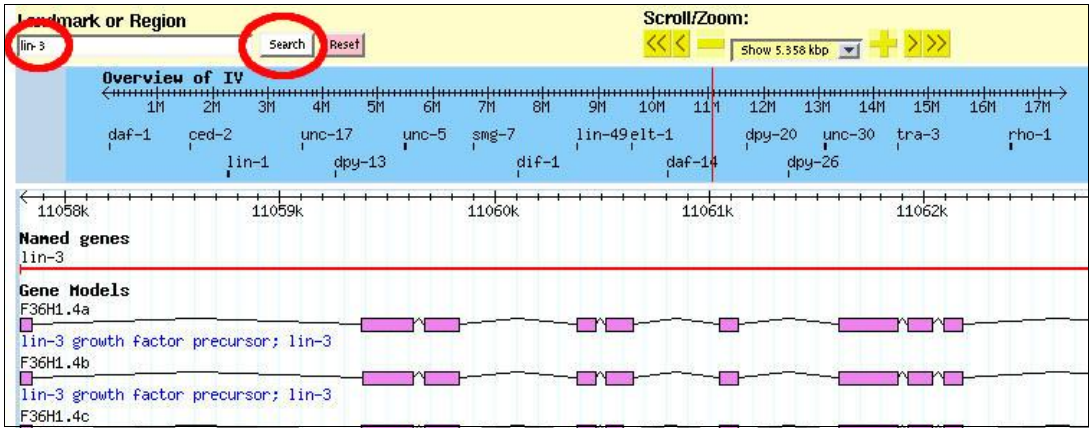
Browse Genome

OVERVIEW

You have a sequence that you want to work on -- a predicted gene, a cDNA, an STS marker -- or a sequence about which you have some descriptive traits. You want to extract it directly from WormBase and analyze its genomic context in detail. Where to go?

One place to go is the Genome Browser page, which is specifically designed for just this problem. Given some word or phrase that corresponds to a known genomic component, the Genome Browser page will give you that genomic component itself if it has been entered into WormBase, or make an effort to guess the closest relevant match, if there seems to be partial overlap between your text and an entry in WormBase.

HOW TO DO A BASIC SEARCH



Type your word or phrase of interest into the "Landmark or Region" window at the top of the search page, then either type a line return ("enter") into the window or click the "Search" button with your mouse's screen arrow.

You will either get a "no match" response, what might be an exact match, or what might be partial matches.

THEN WHAT?

For "no match" you will need to come up with some other search. One thing to consider is that you are most likely to get effective matches if you use one of the vocabularies standard in WormBase -- sequence names in the "cosmid.dot. number" format, canonical gene names, sequence families that are named in the Interpro motif system, DNA coordinates that are realistic...

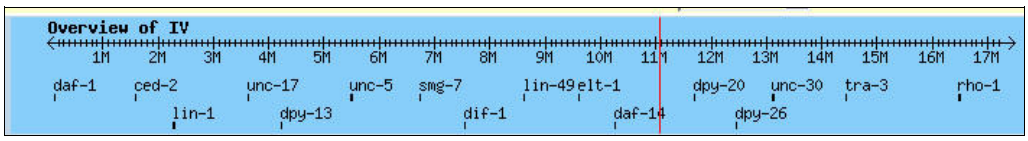
For possible matches to sequences that "may be related", you will either want to check all the results (which can be slow if there are a large number of possible sequences) or redefine your search term in some way that makes it more restrictive.

For a match you consider valid (or at least interesting), you can do several things.

MOVE ALONG THE CHROMOSOME GRAPHICALLY

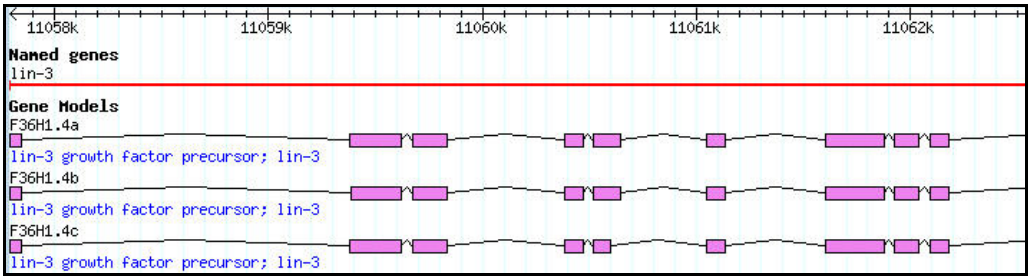
For a given sequence there will be a graphical overview provided in two layers.

The topmost layer shows the location of the sequence in the overall chromosome, with some well-spaced genetic markers interspersed along the chromosome, and with the whole chromosome length given in megabases of DNA sequence.



This view can be browsed graphically by clicking directly on the map, or by clicking buttons at the upper right corner of the map. The buttons allow outward zooming, inward zooming, small left or right steps along the map, and large left or right leaps along the map.

The bottom layer of the graphic display shows details of the DNA sequence comprising a feature or a feature's immediate vicinity. These features will grow or shrink as the map is zoomed outward or inward, and can be altered to show many different details of the genomic sequence.



EXAMINE DETAILS OF THE GENOMIC SEQUENCE

The basic view of a sequence given from a successful search hit includes the following features, by default:

1. Which genes have been assigned classical locus names ("Named Genes"; e.g., "unc-1").

2. That part of a known or predicted transcript predicted structure of predicted exon segments and intervening introns ("Gene Models"). This is the basis of all the other features. Exons are colored to distinguish the predicted protein-coding segments of exons (in magenta) and 5' or 3' untranslated regions of protein-coding transcripts (in dark grey).

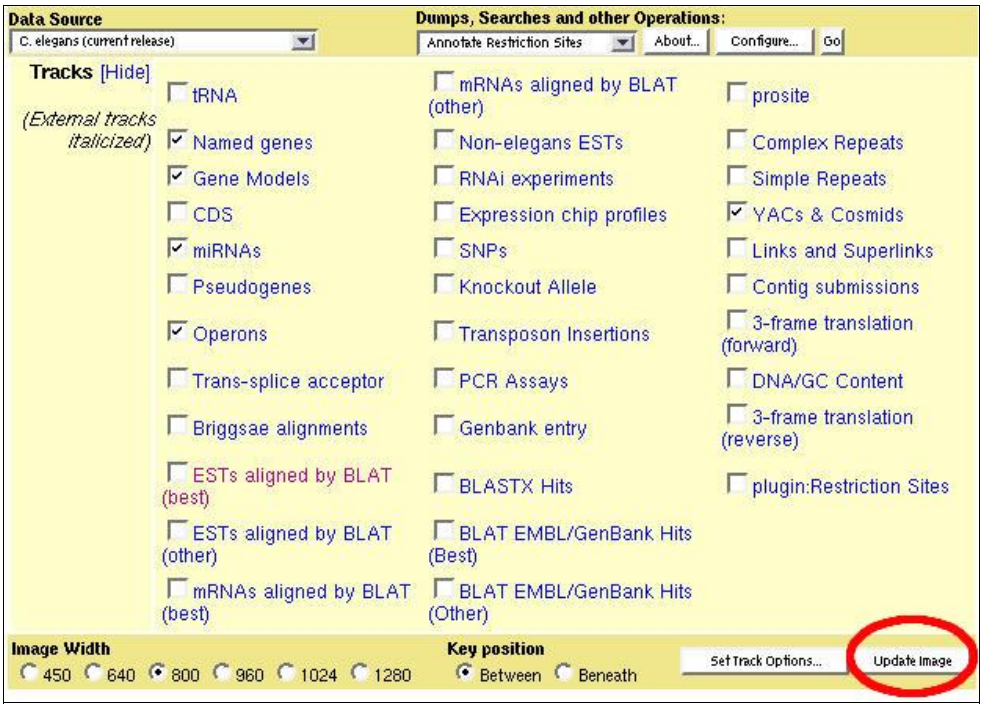
3. Non-protein coding genes ("miRNAs")

4. Sets of genes belonging to a single, multicistronic transcription unit ("Operons").

5. Which standard *C. elegans* clones, used by the *C. elegans* genome project, overlap the sequence ("YACs & Cosmids").

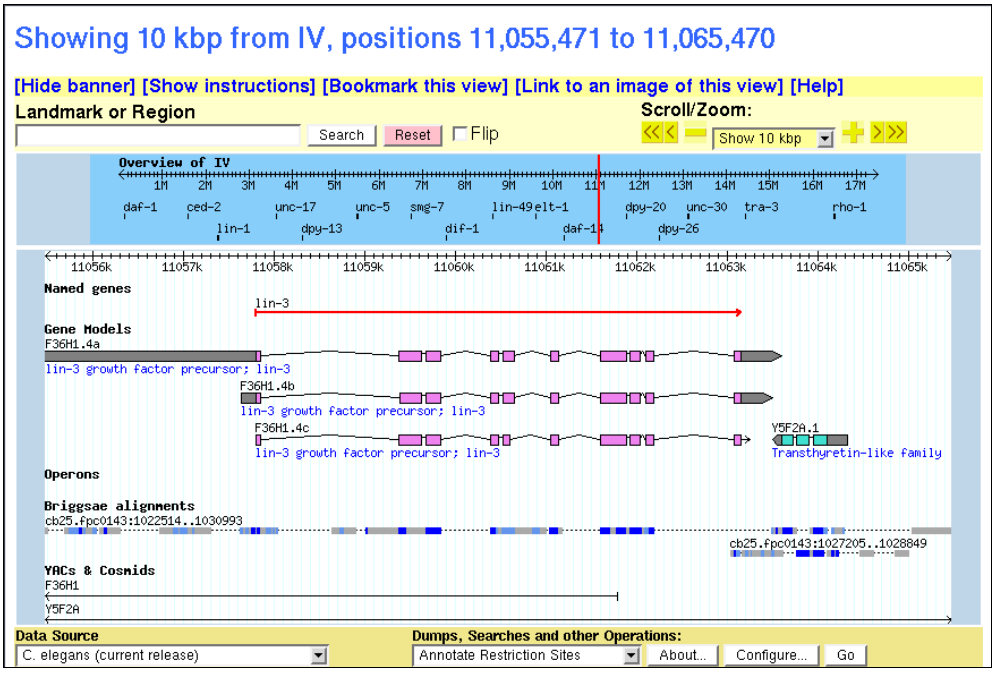
There are many other possible features that, by default, are not shown (in order to keep the initial view of a sequence simple). They can be brought to the screen by checking off the appropriate option checkbox in "Show features", and then clicking on the "Update Image" button. These features are described at:

http://www.wormbase.org/db/seq/gbrowse?help=citations

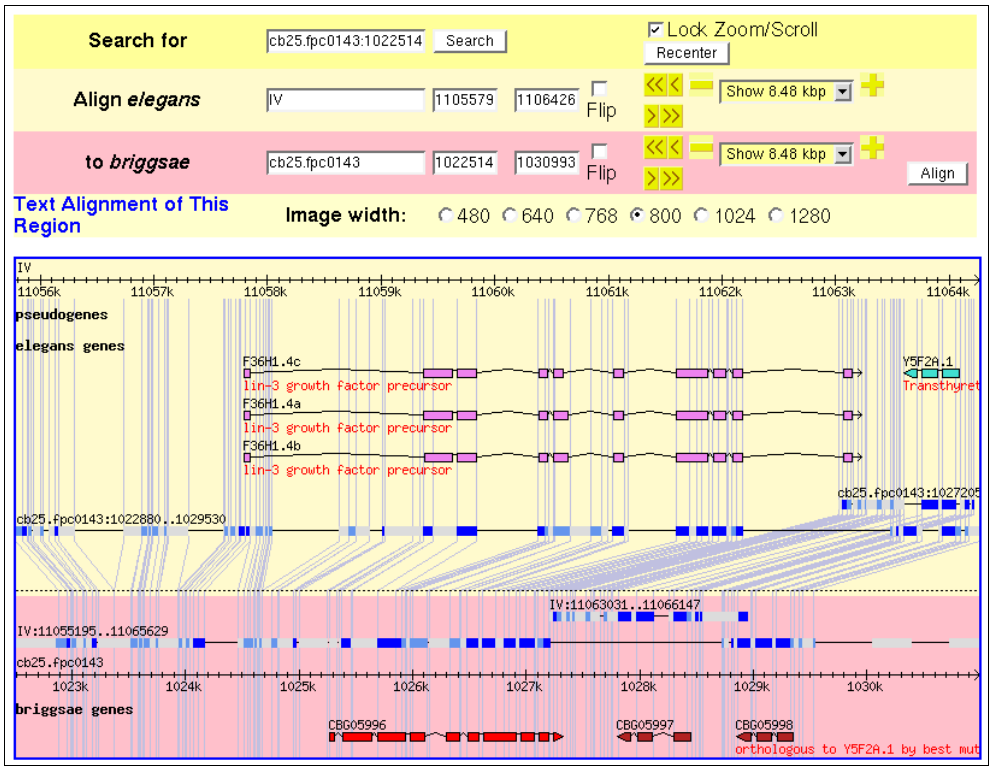


One important feature is the "DNA/GC Content" button. At most resolutions this shows the %GC content of DNA, but at the highest resolution of window (100 bp) it actually shows exact nucleotides (for both sense and antisense strand) of a given region. This is useful for designing experiments that require the exact sequence of a site (clone construction, PCR, etc.)

Another important feature is the "Briggsae alignments" button. This gives a map of the degree of similarity of genomic sequences in *C. elegans* to aligned sequences in *C. briggsae*, with dark blue indicating the strongest matches and pale blue the weakest.



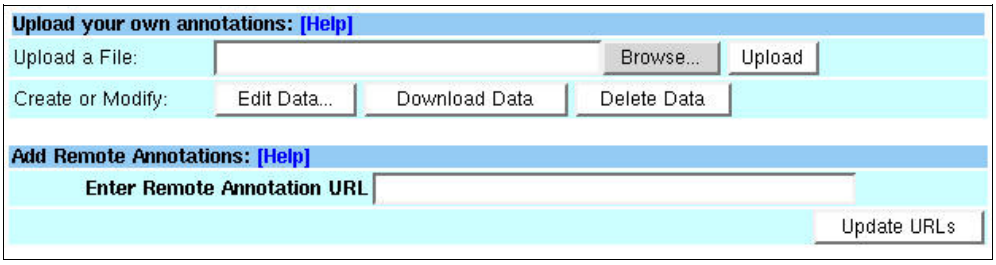
Moreover, clicking on this graphical display sends the browser to a view of *C. elegans* - *C. briggsae* homology in the Synteny Brower, in which you can actually view two aligned genomes at once.



At some point you may decide that you want to work on the sequence as a text file rather than a graphical Web page. This can be done by "dumping" a textual version of the sequence.

ADDING ANNOTATIONS TO THE GENOME▼

The "Upload your own annotations" and "Add remote annotations" features have two purposes: to allow an individual researcher to, in effect, use WormBase as an electronic notebook for recording private comments on a genomic feature of interest; and to allow expert researches to contribute specific annotations in their areas of knowledge.



For more details on how to do this, see:

http://www.wormbase.org/db/seq/gbrowse?help=annotation

http://www.wormbase.org/db/seq/gbrowse?help=annotation#remote

Blast/BLAT Search

OVERVIEW:

BLAST is a method for searching either protein or DNA databases with either a protein or a DNA query sequence. Positive matches in the database to the query sequence are shown with an alignment of the positive to the query, and with a "P value" that represents the probability (given the numerical score of the alignment) that such a query-positive match would occur purely by chance. In other words, small P values are good (very unlikely to be random, i.e., very likely to be meaningful). P values become unreliable above ~0.01, but it is worth noting that a hit which is statistically insignificant in isolation can nevertheless be a real hint of a subtle connection between two highly diverged, yet evolutionarily related sequences.

Full details of how to effectively search databases with BLAST cannot be given here. A short list, by no means complete, of references is:

Jambeck, P. and Gibas, C. (2001). Developing Bioinformatics Computer Skills. O'Reilly, Sebastopol, CA. ISBN: 1-56592-664-1.

http://www.oreilly.com/catalog/bioskills

Mount, D.W. (2001). Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY. ISBN: 0-87969-608-7.

http://www.bioinformaticsonline.org/

Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, 2cd ed. A.D. Baxevanis and B.F.F. Ouellette, eds. Wiley-Interscience, New York. ISBN: 0-471-38391-0.

http://www.wiley.com/Corporate/Website/Objects/Products/0,9049,39021,00.html

Also, information about BLAST is available from the NCBI in an overview, http://www.ncbi.nlm.nih.gov/BLAST/blast_overview.html

frequently asked questions, http://www.ncbi.nlm.nih.gov/BLAST/blast_FAQs.html

a tutorial, http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html

or a formal course. http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

BLAT, which is short for "BLAST-like alignment tool" is designed to quickly find sequences of 95% and greater similarity of length 40 bases or more.It may miss more divergent or shorter sequence alignments. It was written by James Kent. (e-mail him at kent@biology.ucsc.edu). BLAT is similar in many ways to BLAST. The program rapidly scans for relatively short matches (hits), and extends these into high-scoring pairs (HSPs). However, BLAT differs from BLAST in some significant ways. For instance, where BLAST returns each area of homology between two sequences as separate alignments, BLAT stitches them together into a larger alignment. BLAT has a special code to handle introns in RNA/DNA alignments. Therefore, whereas BLAST delivers a list of exons sorted by exon size, with alignments extending slightly be-

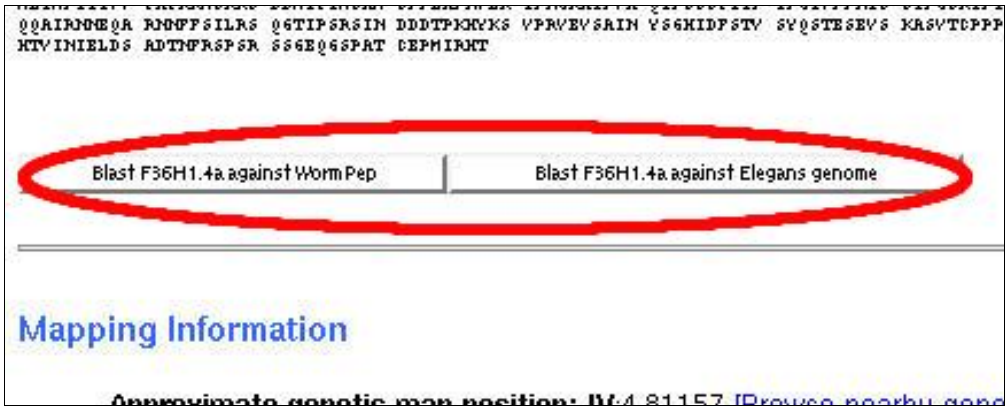
yond the edge of each exon, BLAT effectively "unsplices" mRNA onto the genome giving a single alignment that uses each base of the mRNA only once, and which correctly positions splice sites.

More information about BLAT can be found at: Kent WJ. BLAT-The BLAST-Like Alignment Tool. Genome Res 2002 Apr 12(4):656-64. http://www.genome.org/cgi/content/full/12/4/656

HOW TO USE WORMBASE BLAST SEARCH

The WormBase site provides BLAST/BLAT searches that are specifically directed to *C. elegans* DNA or protein sequences; positive hits are then linked by hypertext to their information in WormBase. The query (search) sequence can be either DNA or protein, and can either from WormBase itself or a sequence supplied by the user. Sequences are automatically provided from WormBase when a BLAST search is specifically requested for a given sequence by clicking on "BLAST against WormPep/Elegans genome".

The following is an example of "Blast" button on Sequence Report pages: ▼



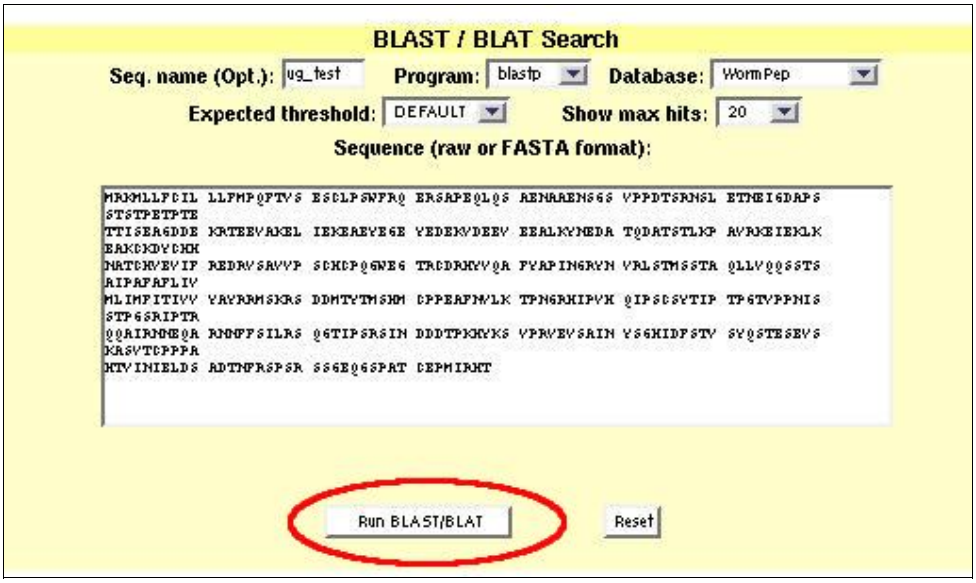
1. Sequence Input:

Back to the Blast Search page. First, sequences must either be completely plain (nothing but one-letter residues) or in the FASTA format. FASTA format sequences start with one header line in the format:

>Sequence_name
[plain residues from here on]
[end of text]

Keep in mind that only nucleotide sequences can be used as queries for BLAT searches.

The right is an example of a BLAST input. ►



2. Expected Threshold:

This is very important but easily forgotten by users: the number you entered for Expected Threshold controls the number of hits displayed in the result. As noted above, BLAST provides quantitative estimates of the probability that a given hit is purely by chance, and these E-value scores are better (from the standpoint of somebody looking for a new, nontrivial sequence similarity) the smaller they are. The WormBase BLAST interface allows the user to only see the subset of BLAST hits that have E values less than a given threshold value. Note that there can be many hits *not* shown, due to this default, that nevertheless have significant E values (e.g., E=0.002) or weak but possibly interesting E values (e.g., E=0.1). Often when one does a BLAST search and 'fails' to see any hits, it is worth checking to see if there are in fact hits whose E values were greater than 0.001.

3. Program Type:

The name "BLAST" is generically used to describe several different varieties of search programs that vary depending on

the type of query (search) sequence and the type of database being searched:

- blastP -- protein query sequence, protein database (e.g.: a given protein versus Wormpep)
- blastN -- DNA query sequence, DNA database (e.g.: a given DNA sequence versus worm genomic DNA, or versus worm ESTs)
- blastX -- DNA query sequence, protein database (e.g.: a given DNA sequence versus Wormpep)
- TblastN -- protein query sequence, DNA database (e.g.: a given protein versus worm genomic DNA, or versus worm ESTs). As mentioned before, BLAT is used to search worm genomic DNA or EST databases solely with DNA query sequences.

Where proteins and DNA are compared, six-frame translations of the DNA sequence are compared to protein.

4. Database Type:

There are three built-in databases that BLAST on WormBase is enabled to search. These complement the various versions of BLAST/BLAT search that WormBase allows a user:

- WormPep -- the most recent compilation of predicted, conceptual protein sequences, derived from the most recent set of gene models.
- elegans ESTs -- the current set of expressed sequence tag sequences available in WormBase; these are predominantly, though not entirely, from the laboratory of Yuji Kohara.
- elegans genomic -- the most recent version of the *C. elegans* genomic DNA sequence.

5. Show Maximum Hits:

In typical BLAST/BLAT searches, there is only a small number of hits that will be of interest -- for example, if one is searching for the single *C. elegans* ortholog of a given gene of interest. On the other hand, it is also possible to have many hits that are both significant and important (e.g., if one is examining a diverse protein family). On the WormBase BLAST/BLAT server, the default maximum number of hits shown is 20, but this can be set to smaller values, greater values, or to show all hits.

WHERE TO GET THE SOFTWARE:

There are two somewhat different versions of the BLAST search software. A publicly available, and reasonably compilable, open-source version is available from the NCBI.

ftp://ftp.ncbi.nlm.nih.gov/blast

The version used in WormBase, however, is a closed-source version provided from Washington University under a free academic license.

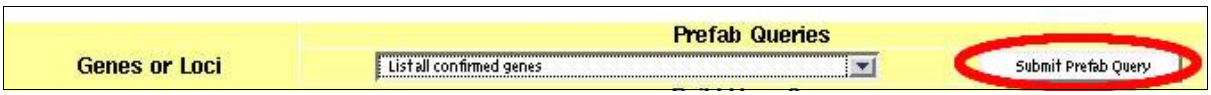
http://blast.wustl.edu

The latter version has been chosen for use in WormBase because it is thought to be superior to NCBI BLAST.

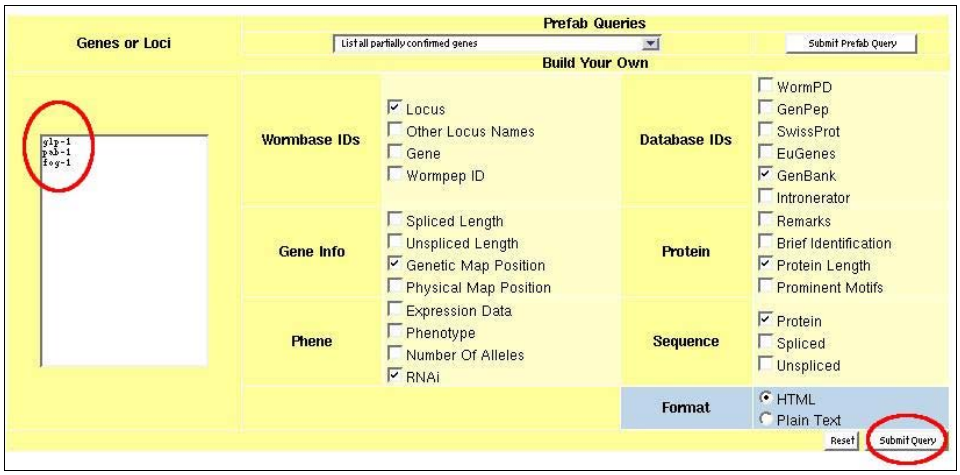
Batch Downloads

Batch information retrieval is very convenient for presentation of different types of data in the form of a table. There are two ways to do this. First, one can use a set of prefabricated queries. Alternatively, one can create one's own queries.

The list of prefabricated queries allows one to download large data sets. For example, all genes of a certain group can be retrieved. To select them go to the Prefab Query pull down menu. There you can retrieve all partially confirmed genes, all confirmed genes, all genes with CGC names, all *C. briggsae* genes, all pseudogenes, all RNA genes, all genetically defined genes, and all genetically defined genes that have been cloned. You can also retrieve all spliced, unspliced, or translated sequences in your gene set. This option is also listed under the prefabricated query pull down menu.

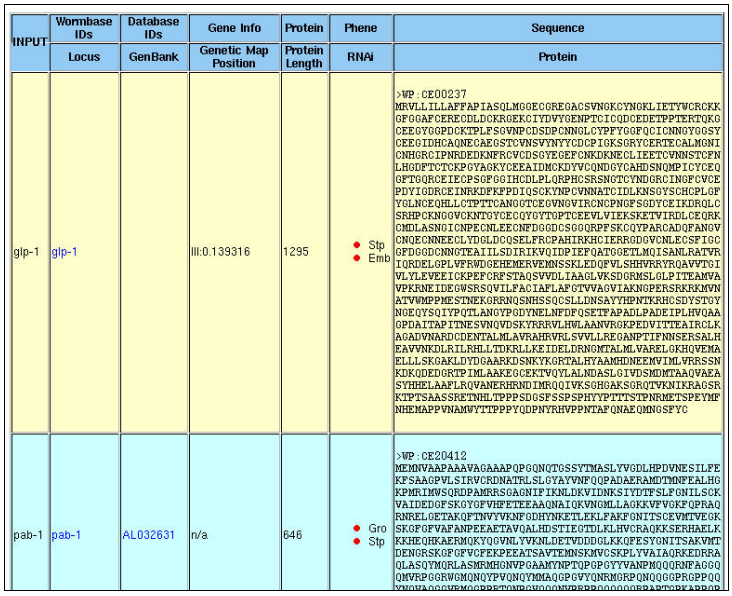


If you are interested in more information about a particular gene set, you can create your own queries. As an example let's retrieve the data about three genes glp-1, pab-1 and fog-1. Select one subfield from each of the following fields: Wormbase IDs, Database IDs, Gene Info, Protein, Phene, and Sequence.



The output with these parameters fits on one screen when in HTML format. You can vary the types of data retrieved according to your needs.

Please note that when multiple transcripts for a gene are available, one has to specify them explicitly (for example, Y54E10A.4a and Y54E10A.4b, NOT Y54E10A.4). Alternatively, one can use "" (Y54E10A.4*).

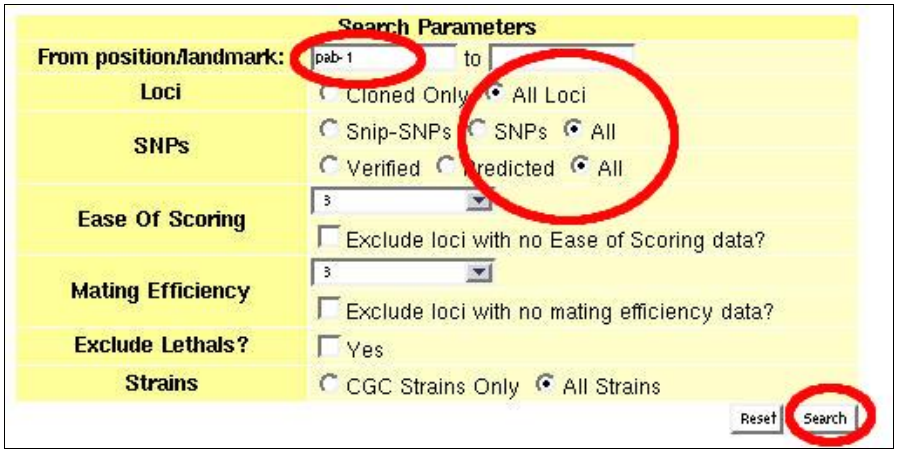


INPUT	Wormbase IDs	Database IDs	Gene Info	Protein	Phene	Sequence
	Locus	GenBank	Genetic Map Position	Protein Length	RNAi	Protein
glp-1	glp-1		III:0.139316	1295	Stp Emb	>WP:CE00237 MKVLLILLAFFAPIASQLMGGEGGREGACSVNGKCYNGKLIETYWCRCKK GFGGAFCEREGDLDCKRGEKCIYDVYGENPTGIGGDCEDETPPTERTQKG CEEGYGGPDCKTPLFSGVNPGDSDPGNNGLCYPFYGGFQGICNNGYGGSY CEEGIDHCAQNECAEGSTCVNSVYNYYCDCPIGKSGRYCERTECALMGNI CNHGRCIPNRDEDKNFRCVGDSGYEGEFGNKDKNECLIEETCVNNSTGFN LHGDFTCTCKPGYAGKYCEEAIDMCKDYVGQNDGYCAHDSNQMPICYCEQ GFTGQRCEIECPSGFGGIHGDLPLQRPHGSRSNGTGYNDGRGINGFGVCE PDYIGDRCEINRKDFKFPDIQSGKYNPGVNNATGIDLKNSGYSGHGPLGF YGLNGEQRLLCTPTTGANGGTGEGVNGVIRGNGPNGFSGDYGEIKDRQLG SRHPGKNGGVGKNTGYGEGQYGYTGPTGEEVLVIEKSKETVIRDLGEQRK CMDLASNGIGNPEGNLEEGNFDGGDGSGGRPFSKGYPARGADQFANGV GNQEGNNEEGLYDGLDGQSELFRGPAHIRKHGIERRGDGVGNLEGSFIGG GFDGGDGNNGTEAIILSDIRIKVQIDPIEFQATGGETLMQISANLRATVR IQRDELGPLVFRWDGEHEMERVEMNSSKLEDQFVLSHHVRRYRQAVVTGI VLYLEVEEIGKPEFGRFSTAQSVVDLIAAGLVKSDGRMSLGLPITEAMVA VFKENEIDEGWSRSGVILFAGIAFLAFGTVVAGVIAKNGPERSRKRKMVN ATVWMPPMESTNEKGRRNQSNHSSQCSLLDNSAYYHPNTKRHCSDYSTGY NGEQYSQIYPQTLANGYPGDYNELNFDFQSETFAPADLPADEIPLHVQAA GPDAITAPITNESVNQVDSKYRRRVLHWLAANVRGKPEDVITTEAIRCLK AGADVNARDCDENTALMLAVRAHRVRLSVVLLREGANPTIFNNSERSALH EAVVNKDLRILRHLLTDKRLLKEIDELDRNGMTALMLVARELGKHQVEMA ELLLSKGAKLDYDGAARKDSNKYKGRTALHYAAMHDNEEMVIMLVRRSSN KDKQDEDGRTPIMLAAKEGCEKTVQYLALNDASLGIVDSMDMTAAQVAEA SYHDELAAFLRQVANERHRNDIMRQQIVKSGHGAKSGRQTVKNIKRAGSR KTPTSAASSRETNHLTPPPSDGSFSSPSPHYYPTTTSTPNRMETSPEYMF NHEMAPPVNAMVYTTPPPYQDPNYRHVPPNTAFQNAEQMNGSFYC
pab-1	pab-1	AL032631	n/a	646	Gro Stp	>WP:CE20412 MEMNVAAPAAAVAGAAAPQPGQNQTGSSYTMASLYVGDLHPDVNESILFE KFSAAGPVLSIRVCRDNATRLSLGYAYVNFQQPADAERAMDTMNFEALHG KPMRIMWSQRDPAMRRSGAGNIFIKNLDKVIDNKSIYDTFSLFGNILSCK VAIDEDGFSKGYGFVHFETEEAAQNAIQKVNGMLLAGKKVFVGKFQPRAQ RNRELGETAKQFTNVYVKNFGDHYNKETLEKLFAKFGNITSCEVMTVEGK SKGFGFVAFANPEEAETAVQALHDSTIEGTDLKLHVCRAQKKSERHAELK KKHEQHKAERMQKYQGVNLYVKNLDETVDDDGLKKGFESYGNITSAKVMT DENGRSKGFGFVCFEKPEEATSAVTEMNSKMVGSKPLYVAIAQRKEDRRA QLASQYMQRLASMRMHGNVPGAAMYNPTQPGPGYYVANPMQQQRNFAGGQ QMVRPGGRWGMQNQYPVQNQYMMAQGPGVYQNRMGRPQNQQGGPRGPPQQ YNQYAQGYRMGGPRPTGNPGWQGGNMYRPRPQGGGGGPRAPTGPKAPPQP

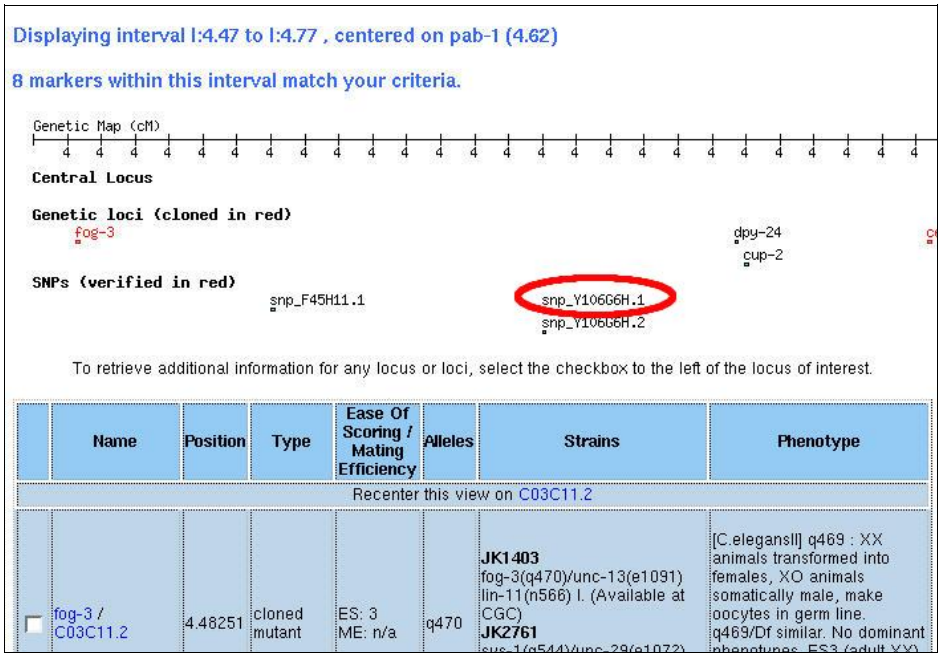
Marker/Strain Search

This page is designed to assist in the selection of SNPs and visible genetic markers for mapping experiments. The search results will yield a list of loci and details on how to score the mutant phenotype.

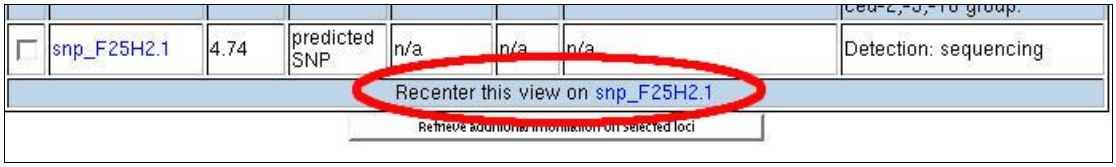
As an example we will search for genetic and physical loci around the pab-1 locus. We will search for all loci, Snip-SNPs and predicted and confirmed SNPs. Other features can be added as well (ease of scoring, mating efficiency, whether to exclude lethals, and the availability of the strains).



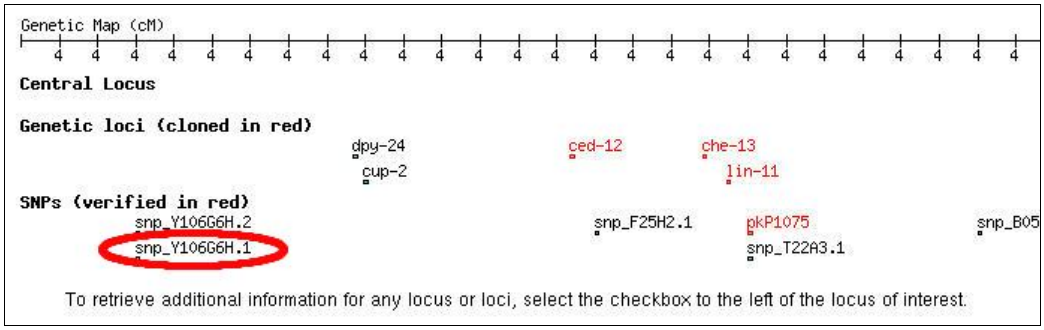
The information about resulting loci can be retrieved by clicking on the corresponding link. For example, by clicking on snp_Y106G6H.1 we can see the exact nucleotide change in this polymorphism, probability that the polymorphism exists, etc..



The resulting map can also be re-centered. This allows "one-click" search of the region neighboring the original search region.



Click on "Recenter this view on snp_F45H11.1" for example, and compare the new map display with the old one shown above:



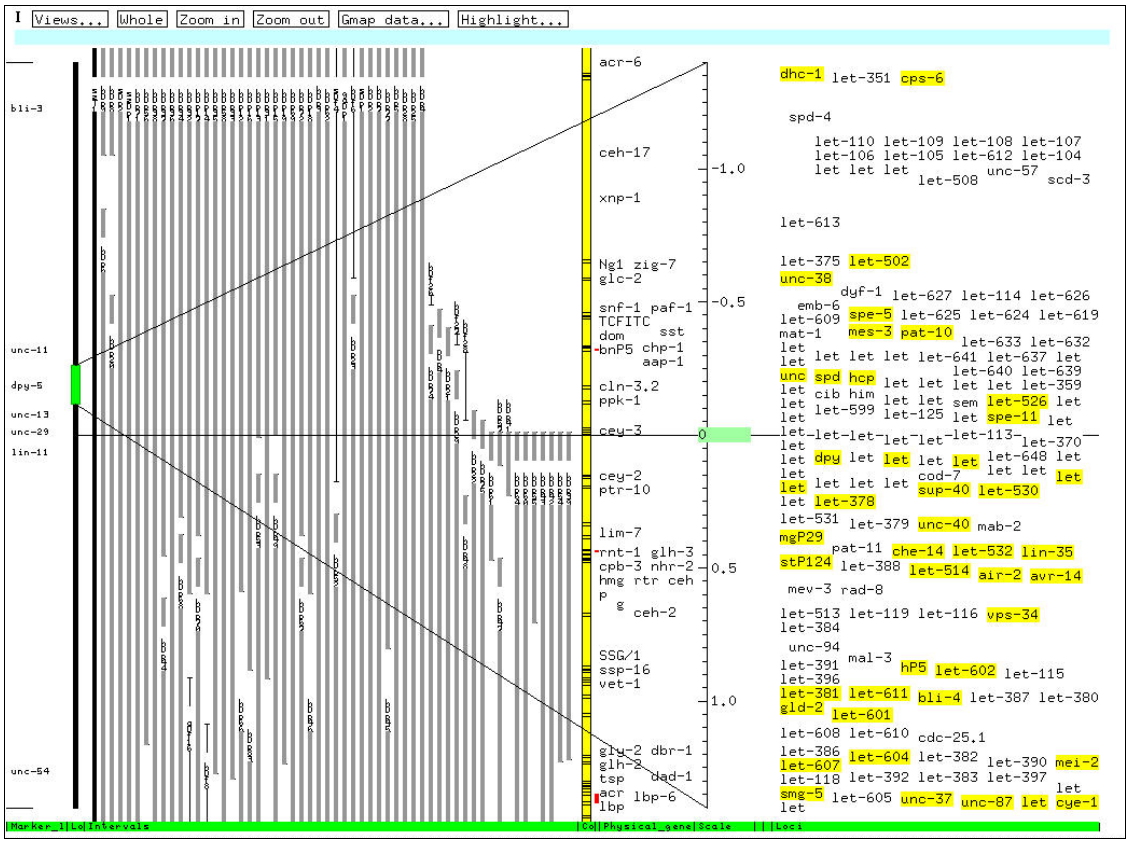
Genetic Map

The Genetic Map is a very useful resource for forward genetic mapping. The WormBase Genetic Map inherited its look from ACeDB, however because of internet interface, many features have been changed.

OVERVIEW OF DISPLAY

The Genetic Map can be accessed in two ways: either from the Genetic Map link of the Main Menu located at the top of every WormBase page, or from the Submenu located on each of the Gene Summary pages. When the Genetic Map is accessed via the Main Menu link, a whole-chromosome graphic display is first generated of chromosome I. To view a different chromosome, change the value in the pull down menu to the right of Chromosome Name and click the Change button.

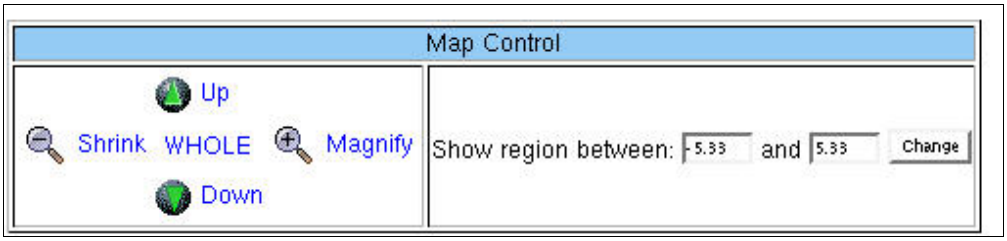
The viewable part of the chromosome with respect to the whole chromosome is indicated by the green bar that overlays a black



vertical bar. On the left are loci commonly used for mapping that have been clearly ordered relative to one another. To the right, the black bars represent rearrangements and the gray bars identify and depict the extent of deficiencies and duplications in the region. The yellow bar represents the physical map (Contigs column), and next to it, genes that are ordered on the physical map. Inconsistencies between the physical and genetic maps are indicated by red marks next to the yellow bar. Loci that have been ordered on the genetic map are indicated on the right-most side of the window. Of these, cloned loci are highlighted in yellow.

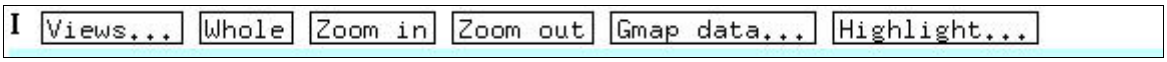
MAP CONTROL

The graphical view of any individual chromosome can be changed in two ways: either by clicking the Shrink, Magnify, Up, or Down navigation links located under Map Control. Alternatively, it is possible to specify a chromosomal region of any size by entering numerical map coordinates. More detailed information about all of the loci, including deficiencies and duplications, can be retrieved by clicking on their names. For example, the end points of deficiencies and duplications and genetic map data can be accessed by clicking on any of the mapped loci.



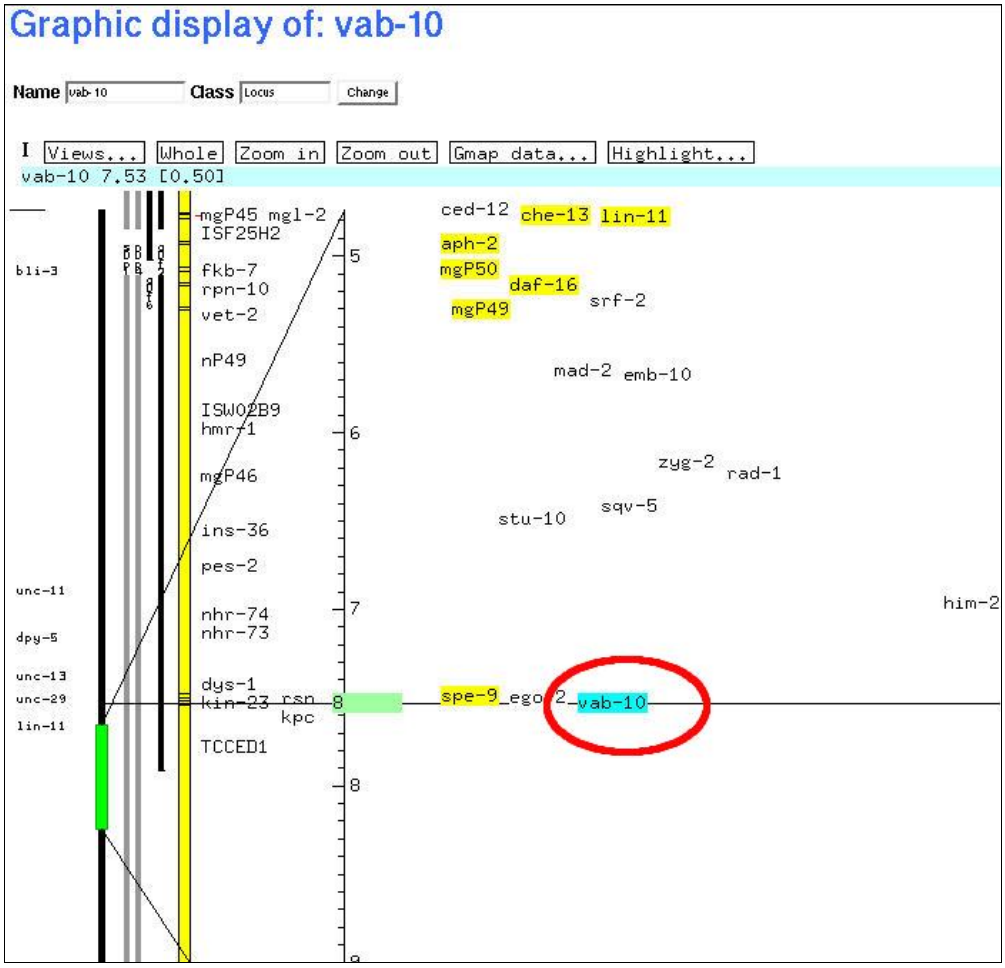
ACEDB MAP IMAGE SUBMENU

The map controls described above replace the menu links located inside of the map used originally for AceDB. These links are shown below:



The other functions, Gmap data... and Highlight also do not function from this particular page.

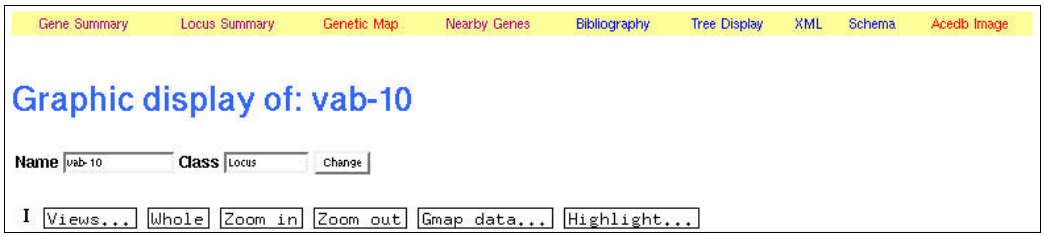
When the Genetic Map link is accessed from the submenu located on the top of each of the Gene Summary pages, a chromosome map is displayed centered upon the gene from which the map was accessed. For example, by clicking on the Genetic Map submenu link from the *vab-10* page, *vab-10* is highlighted in cyan and centered in the display. The viewable region cannot be altered from this page because there are no map controls. However, by clicking on the Gmap data, one can view a graphical representation of the results obtained from genetic mapping experiments. Rectangles indicate other markers used, and vertical lines indicate confidence information. Horizontal lines indicate where there are recombinants. The other AceDB menu buttons do not work in this window.



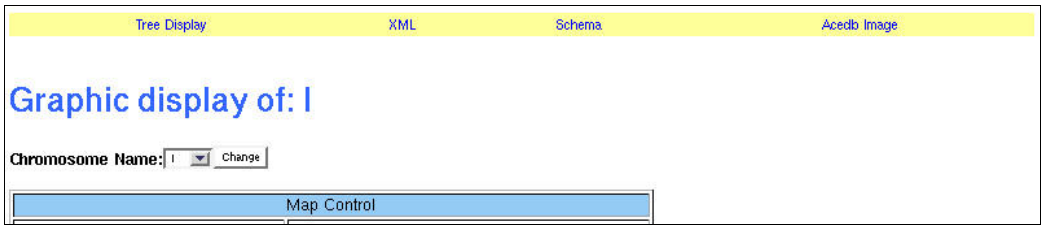
OTHER WAYS TO ACCESS THE GENETIC MAP

From a Gene Summary page, there are at least 3 submenu options provided to access the Genetic Map: "Locus Summary", "Genetic Map" and "Acedb Image".

When the Genetic Map is accessed from a Gene Summary Page, it is a graphic display of a gene, so the submenu is the same as Gene Summary Page:



When the Genetic Map is accessed from the Main Menu, it is a graphic display of a chromosome, and the submenu is different from the Gene Summary Page:



The default map display is "Acedb Image". Although the appearance is very similar to that people used to see in ACeDB, one very important thing is that users can no longer use the mouse to move or zoom the green bar on chromosomes. Instead, the "Map Control" now serves this purpose.

Advanced Search Tools

There are quite a few advanced search tools in WormBase. The "More Searches" button will bring you to an index page for all search tools available in WormBase. Besides those already listed as basic search tools, you will find:

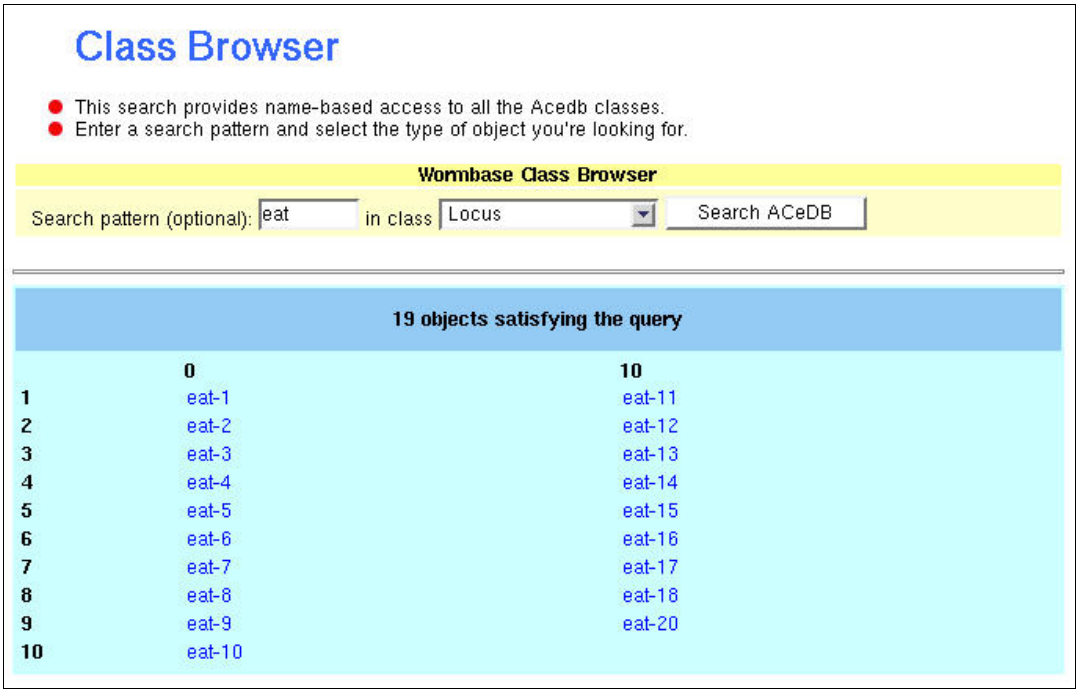
1. Advanced Search
2. ePCR
3. Gene Ontology
4. Cell Search
5. Neuron Search
6. Expression Search
7. Expr_profile Search
8. RNAi Search
9. Interval Search
10. Genome Dumper

Advanced Search

This menu includes three kinds of search tools for advanced users to browse in WormBase for more specific results.

CLASS BROWSER

The top search panel is the Class Browser. It looks similar to Quick search or Basic search, but with a longer list of classes to choose from. You can do everything you do in Basic search using the Class Browser. The advantage of the Class Browser is that there are more classes to select from, and you can view the whole list of each class, not only Cell, Clone and Strain, but also those that cannot be selected in Basic Searches, such as Allele, Expr_pattern, Gene_class, Lineage, etc. There are also classes in the Class Browser about the data structure of WormBase, such as "Model".



WORMBASE QUERY LANGUAGE SEARCH

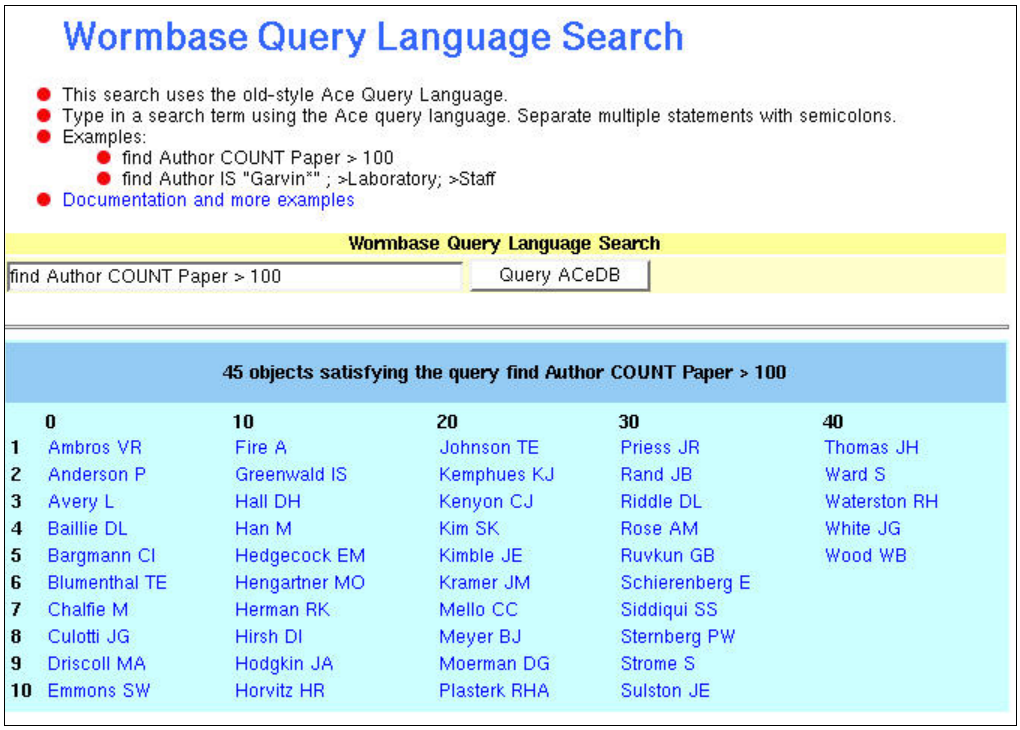
The middle search panel is for WORMBASE QUERY LANGUAGE SEARCH, also called old_ACE_query search, which requires writing queries using Ace Query Language(AQL). There are some examples given above the search box.

For more background instruction:

http://stein.cshl.org/AcePerl/.

To learn how to write queries:

http://www.acedb.org/Cornell/#query

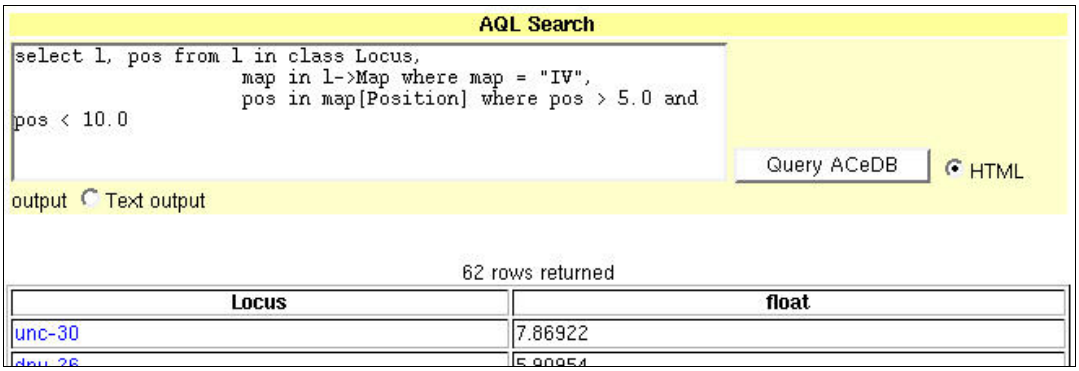


AQL SEARCH

This bottom panel is similar to the WormBase Query Language Search, however, it uses a new AQL syntax. If you are interested in learning about this query, there is a very good help page linked to WormBase:

http://www.acedb.org/Software/whelp/AQL/examples_worm.shtml

The following is an example with results:

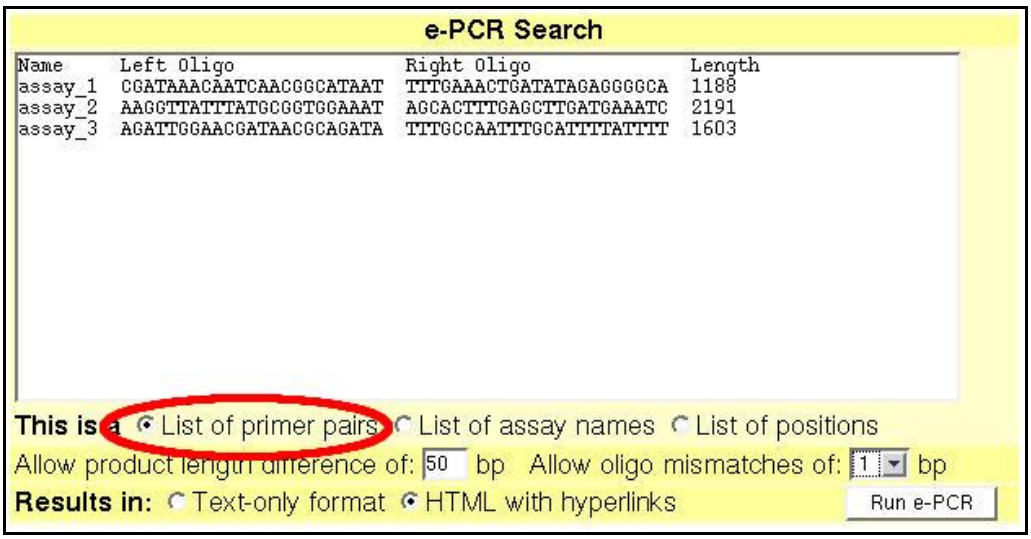


ePCR

The ePCR search page serves two functions: (1). Find coordinates in the genome that can theoretically produce a PCR product given a pair of primer sequences, and (2). Find the smallest encompassing genomic object (explained below) and intersecting genes given a genomic position.

Function (A)

Search Type I:



Step 1: The user provides a list of assays, each row with 4 columns: assay_name, sequence1, sequence2, predicted_product_length. The columns should be separated by one or more blank space. You may not know the product length a priori. Just give a generic number, say 2000, and set the tolerance liberally in Step3 and/or adjust after some trials.

Step 2: After pasting the list in the Text box, select "**This is a <>List of primer pairs**".

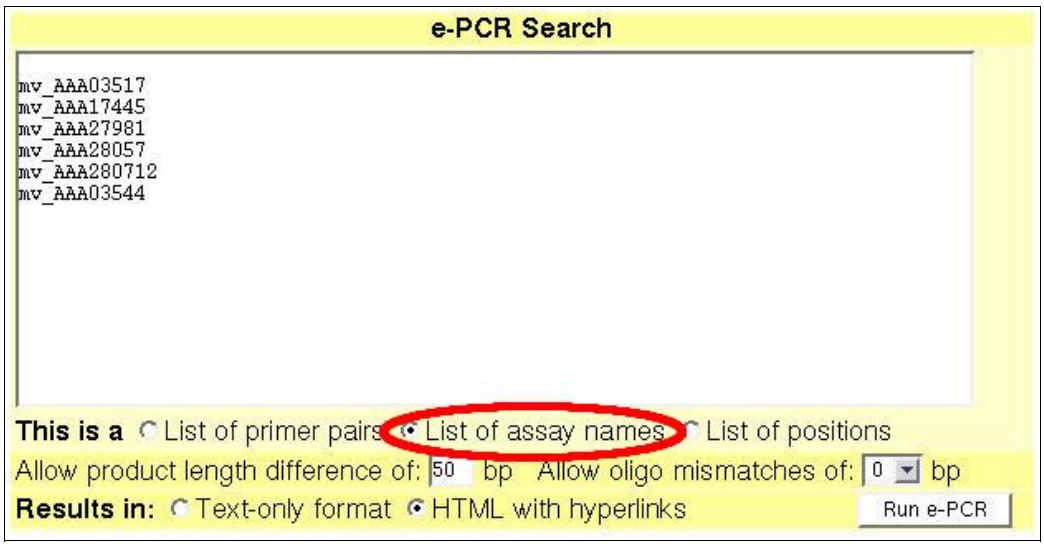
Step 3: Enter in base-pairs how much the product size can be tolerated to differ from that specified in column 4, in "**Allow product length difference of: bp**". The user may tune this parameter according to trial results.

Step 4: Choose the number (0-4) in nucleotides of primer sequence mismatch that can be tolerated with "**Allow oligo mismatches of: bp**".

Step 5: Choose the format of the results; HTML is recommended for most users.

Step 6: Click on "**Run e-PCR**".

Search Type II:



Step 1: The user provides a list of assay names (in single column format), that already exist in WormBase and can be searched as "Primer Pair".

Step 2: Select "**This is a <>List of assay names**".

Step 3-6: Same as in Search Type I.

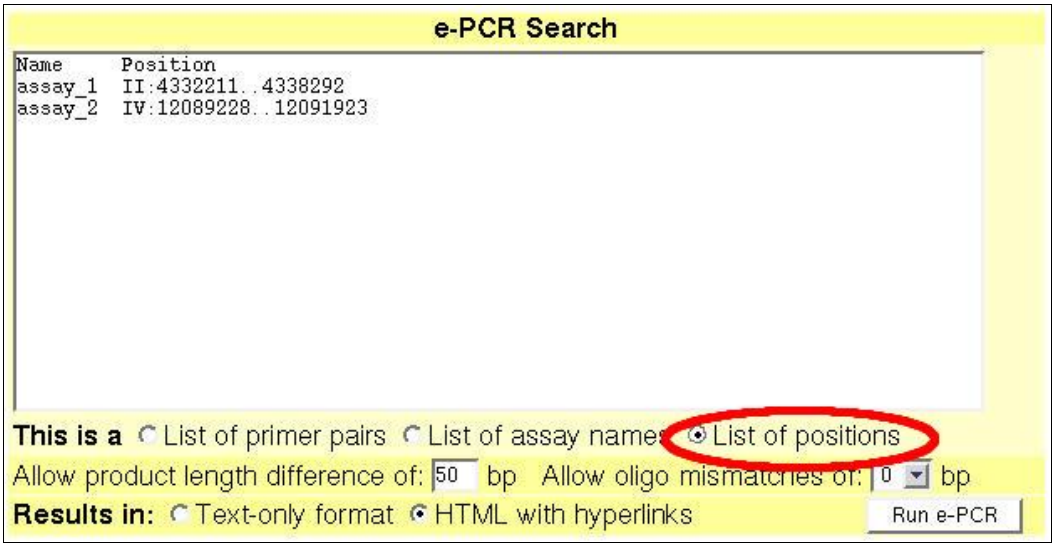
Function (B)

Search Type III:

Step 1: The User provides a 2 column list of Names and Positions (a pair of chromosomal positions).

Step 2: Click on "**This is a <>List of positions**".

Step 3-6: Same as in Search Type I.



Result Page

The result is a list of 7 columns.

Column 1: **Assay,** assay name.

Column 2: **Chromosomal Pos,** the chromosome, beginning and ending nucleotide coordinates for the predicted 'PCR' product. For each assay, there may be more than one product.

e-PCR Results

Assay	Chromosomal Pos	Genbank Pos	Cosmid/Link Pos	Genes Covered	Exons Covered	Exons in Gene
assay_3	II: 11338108..11339710	Z49907: 10874..12476	B0491: 10874..12476	B0491.3	1,2,3,4	4
assay_2	V: 10299392..10301582	Z71178: 10236..12426	B0024: 10236..12426	B0024.4	1,2,3,4,5	5
assay_1	X: 10818180..10819367	Z68108: 15392..16579	T05A10: 15392..16579	T05A10.2	1,2,3,4,5,6	6

Search Again

Column 3: **Genbank Pos,** the GenBank accession number for the smallest enclosing genomic sequence object and the corresponding nucleotide coordinates for the PCR product. The GenBank accession may not precisely correspond to the insert of a physical clone in that object. Click on the hyper link for details at NCBI.

Column 4: **Cosmid/Link Pos,** the smallest enclosing cosmid or sequence link object and the corresponding nucleotide coordinates for the PCR product in that object. Again, these may not precisely correspond to the insert of a physical clone.

Column 5: **Genes Covered,** (predicted) gene that intersects with the PCR product. There can be more than one gene per PCR product.

Column 6: **Exons Covered,** those exons in the gene that intersect with the PCR product.

Column 7: **Exons in Gene,** the total number of exons in the gene that intersects with the PCR product.

More information on the e-PCR procedure can be found in: Schuler GD. Sequence mapping by electronic PCR. Genome Res 1997 May;7(5):541-50.

Gene Ontology Search

Gene Ontology (GO) is a controlled vocabulary aimed at capturing three different aspects of a gene product in a living organism: its biological process or processes (essentially, the reason why a given gene product benefits the organism whose genome encodes them); its biochemical function or functions (i.e., what the gene products do at a basic biochemical level); and its cellular location or locations (i.e., where in a cell the gene product will normally be).

Gene Ontology is extensively documented at the official Web page of the GO Consortium: http://www.geneontology.org and has also been described in the following papers:

Ashburner et al. (2000). Gene ontology: tool for the unification of biology. Nat. Genet. vol. 25, pp. 25-29.

The Gene Ontology Consortium (2001). Creating the gene ontology resource: design and implementation. Genome Res. vol 11, pp. 1425-1433.

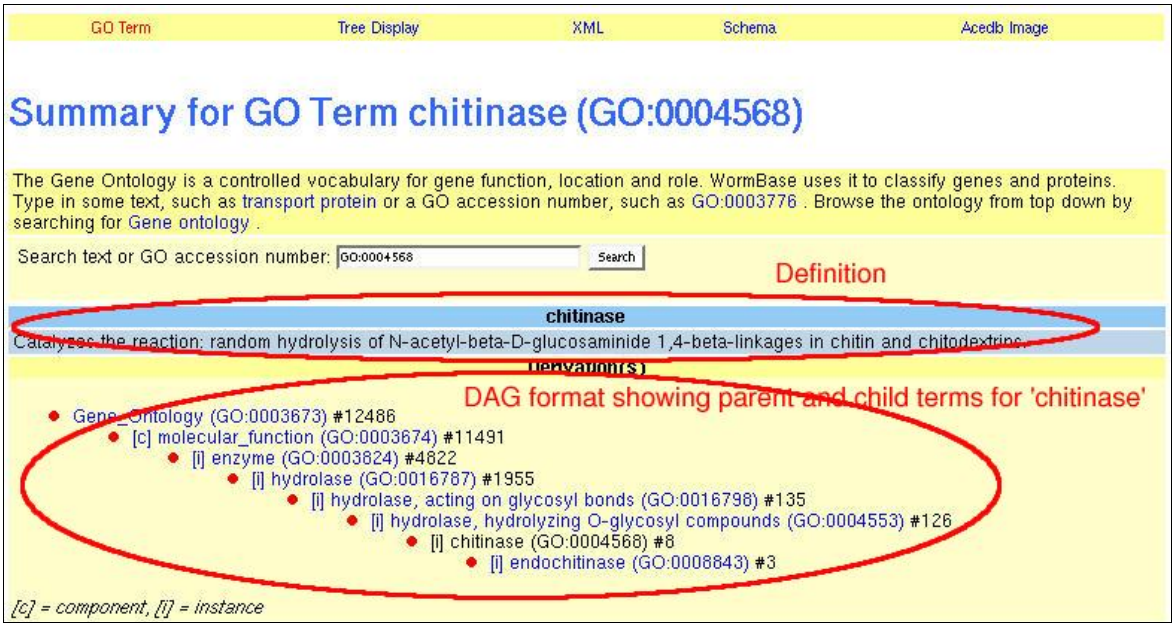
At the moment, the Gene Ontology search page for Wormbase allows the following search:

- 1. Pick a term for a biological process, biochemical function, or cellular localization that is of interest. Enter the term as text in the "Search text or GO accession number:" window.
- 2. Press "Search" button to run a search.
- 3. The result will display GO terms, and clicking on them will link to the summary reports.
- An Example Search:

Searching for 'chitinase' returns two objects: the term names 'chitinase' and 'endochitinase' along with the corresponding seven digit GO ID within parentheses.

Clicking on 'chitinase' links to the summary report page for 'chitinase'.

The summary page for the GO term 'chitinase' displays the definition for this term and the position of this term in the directed acylic graph (DAG) format so that the entire ontology structure showing the parent and children terms for 'chitinase' are displayed. Each term in this structure can be clicked on to access its summary report.



Below this, is displayed a list of all the Sequence and Motif objects in WormBase associated with the GO term 'chitinase'. Click on them to access the specific sequence and motif reports.

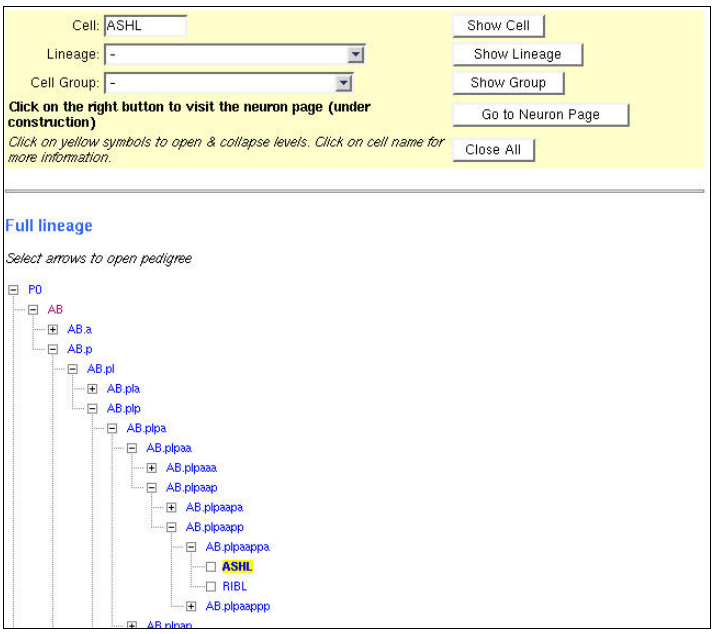
The following WormBase object(s) refer to this term	
Sequence	
Sequence associations	C08B6.4 chitinase
	Y50D4A.3
	T26F2.1 chitinase
	T05H4.7 chitinase
	R10D12.15 chitinase
Motif	
Motif associations	INTERPRO:IPR000726 Glycoside hydrolase, family 19

Cell Search

One of the much-cited advantages of *C. elegans* is that its cellular lineage and neuronal wiring are exactly known. WormBase provides several ways to search for specific cells, lineage elements, or neuronal connections.

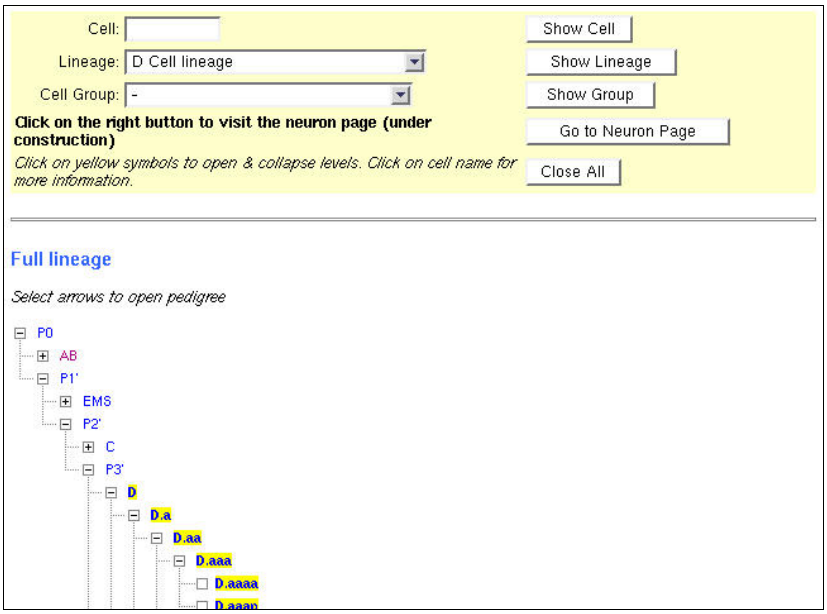
PEDIGREE BROWSER: CELL

Typing a given cell into this window, and then either typing "enter" or clicking on "Show Cell", will cause the specific lineage leading from the fertilized egg (P0) to the cell to be shown below (under 'Full lineage'). One can either search for a cell that has been exactly determined, including which side it is on (e.g., "AWAR") or for cell specified by type but not by laterality (e.g., "AWA"). If one searches for a generic cell, the search will try to interpret this intelligently by showing and highlighting all pertinent cells in their lineages (e.g. lineages to both AWAR and AWAL). See below for details of the 'Full lineage' display.



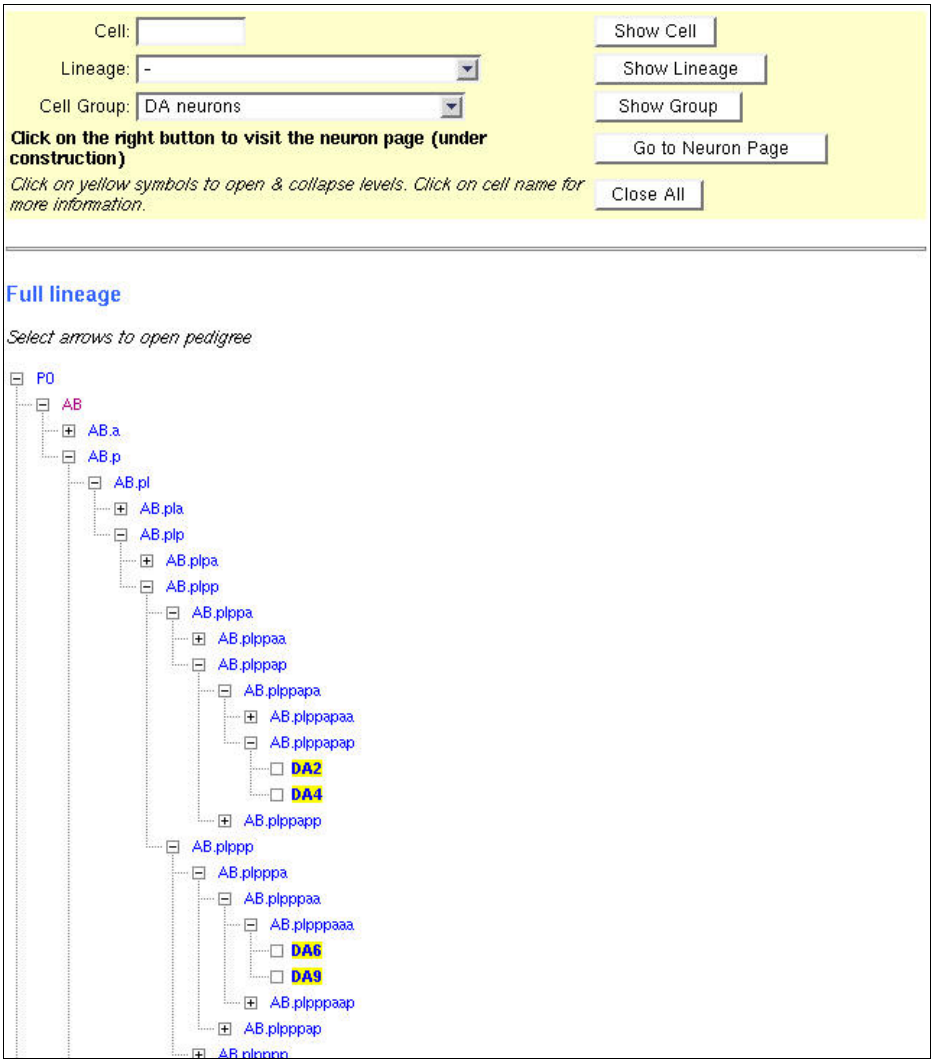
PEDIGREE BROWSER: LINEAGE

A variety of predefined lineage types are provided for both hermaphrodite and male development. As with a generic cell, all the relevant cells will be shown and highlighted in the 'Full lineage' display.

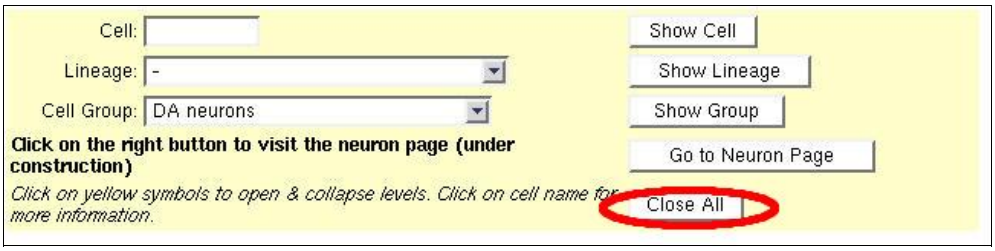


PEDIGREE BROWSER: CELL GROUP

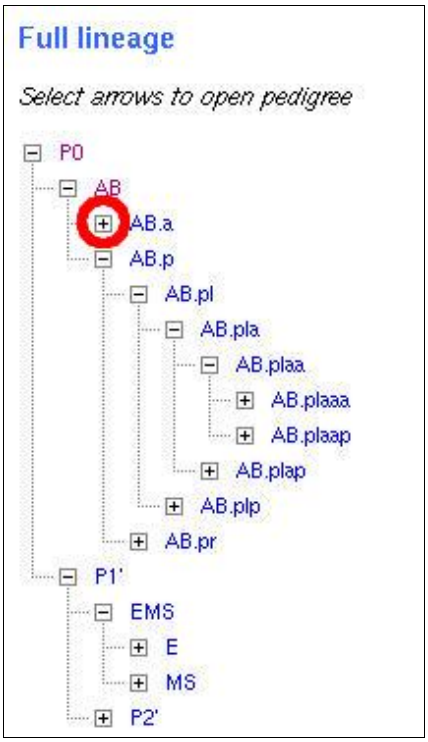
C. elegans lineages are invariant, or nearly so. The anatomy of the completed animal is also highly invariant. But one does not map onto the other in a simple-minded way: many different components of distinct lineages can coalesce to produce a single organ, tissue, or functional group. WormBase allows users to scan the developmental lineage for the pedigrees of many different predefined cell groups (e.g., "amphid neurons"). Relevant cells are shown and highlighted in 'Full lineage'.



CLOSE ALL►



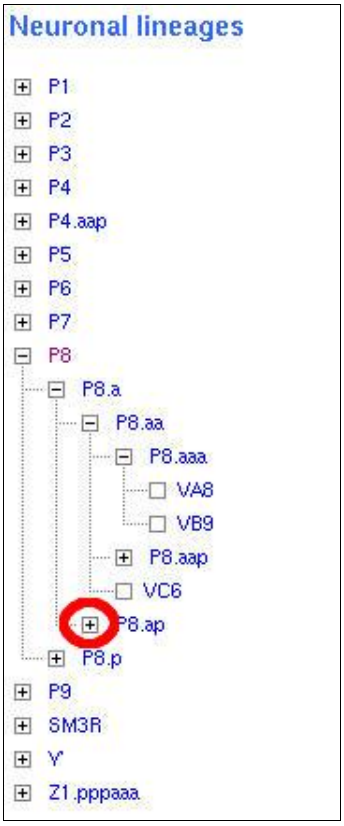
Several different forms of searching cause the 'Full lineage' display to expand into complicated forms. To quickly and cleanly reduce the Full lineage display back to its original, compressed form, one can click on the "Close All" button. This is of most use when one wants to keep the effects of a given search clearly distinct from those of a previous search.



◄FULL LINEAGE

This display begins by showing only the fertilized egg (P0), but expands when a search is done for a given cell, lineage, or group. It can also be expanded by directly browsing (by clicking on a cell marked "+" to get a full display of the cells descended from it).

Cells are given the name most likely to be intelligible to a *C. elegans* investigator; where a common short name exists for a terminally differentiated cell type (e.g., "SAAVL") that is used, instead of the full-lineage abbreviation (e.g., "ABplpaaaaap"). Where no such short name exists, a full-lineage term is used (e.g., "Earaa").

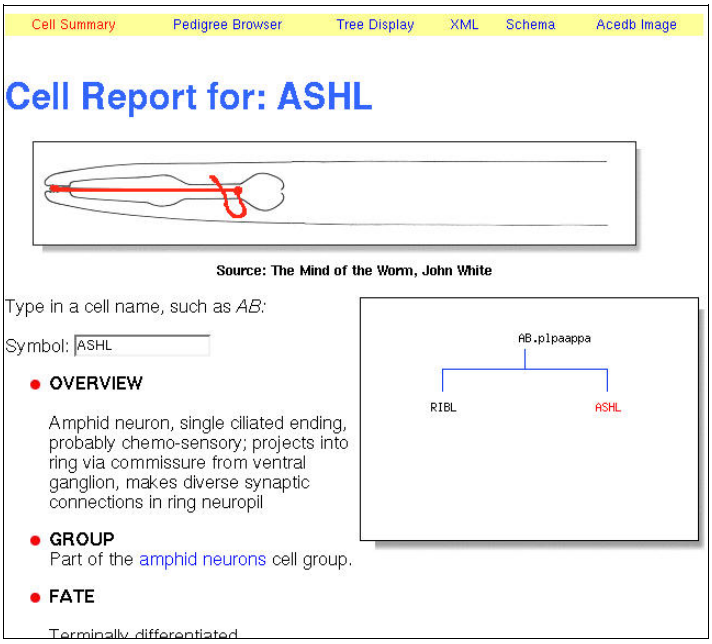


NEURONAL LINEAGES►

This display is much like the 'Full lineages' display, except that these lineages are given specifically for postembryonic blast cells that generate ventral cord motoneurons, ventral hypodermis, vulva, and the male preanal ganglion. The lineages begin with blast cells P1 through P12. Due to an oddity in the current implementation, the 'Full lineages' stop with the lineage-names of P1-P9 et al., while the 'Neuronal lineages' start with the short names of P1-P9 et al. This feature may be corrected in future releases of the WormBase site software.

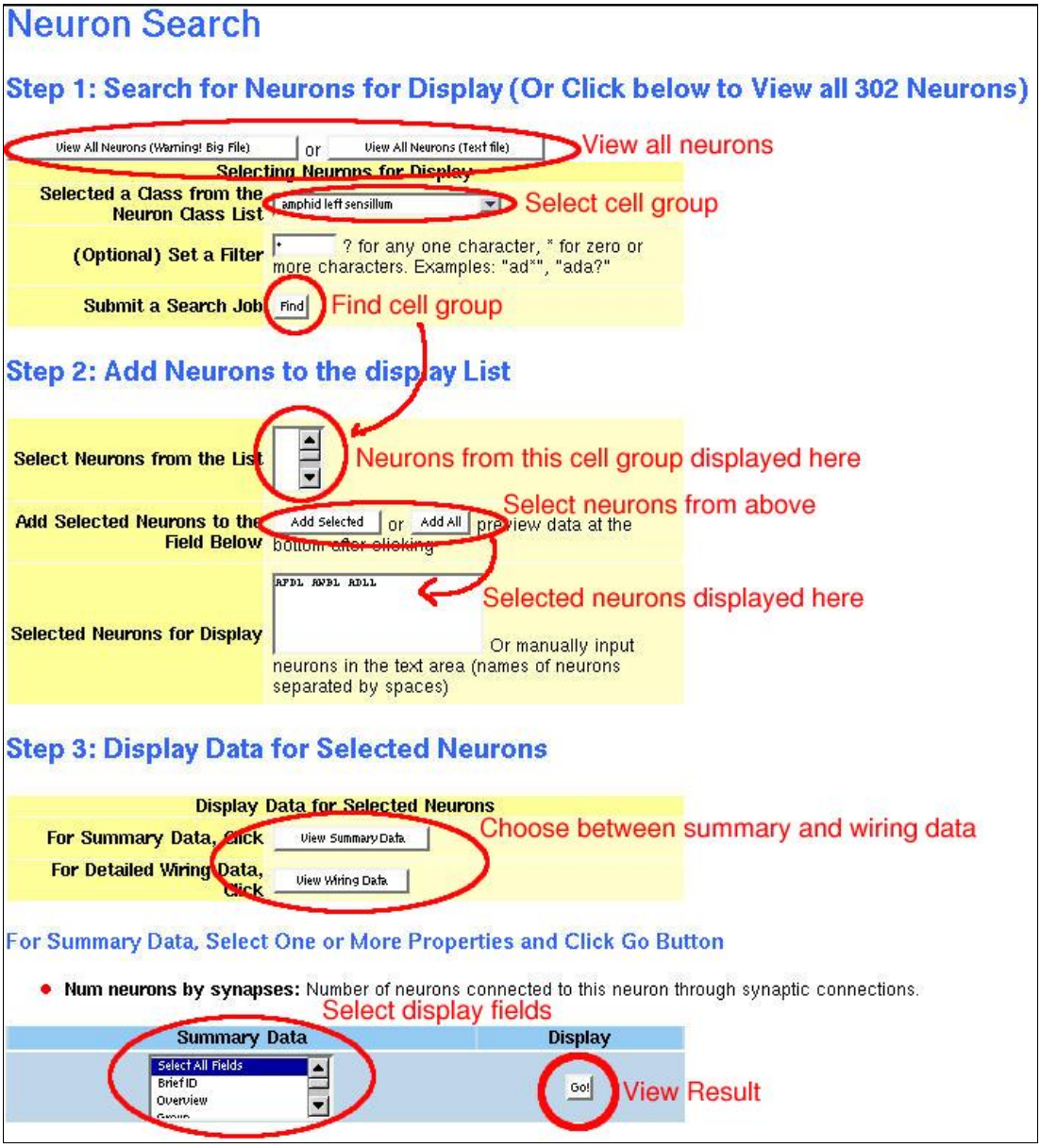
INDIVIDUAL CELLS (DESCRIBED IN CELL REPORT)►

Each individual cell is given a hypertext link; clicking on this link will take one to the appropriate Cell Report. Each such report has the following features: an overview of the cell and its role; what group of cells it belongs to; whether it is terminally differentiated or fated to undergo further division; the ancestry of the cell during development; and the expression patterns of any genes that are known or believed to overlap temporally and spatially with the cell in their activity. Where a given cell type is a neuron, a diagram of its shape in the worm body (excerpted from White et al. [1986], Philos. Trans. R. Soc. Lond. B Biol. Sci. vol. 314, pp. 1-340) is given. Finally, for each cell type, there is a graphical map of the immediate lineage that is given to the right of the text information, and that itself is browsable.



Neuron Search

This search page is (we hope) designed to be largely self-explanatory. There are some points about this page that may not be obvious, though:



"Selected a Class":

Not all classes are intuitively reasonable. For instance, 'head neurons' only includes four neurons and not (e.g.) amphid neurons! The safest rule to follow is that if a set seems not to have the neuron you want, you should scan that set with no filter (or with "*" as a filter) to see if (for whatever reason) the neuron is simply not considered a member of that class.

"Set a Filter":

This allows a class of neurons to be selected for only those neurons with a certain text string in their names. The "*" formalism for 'zero or more undefined characters' (from UNIX) is used; for example, searching for "BDU*" in the "neuron" class will give BDUL and BDUR. Another UNIX metacharacter that is used is "?", to denote 'exactly one undefined character, not zero or more than one'.

"Submit a Search Job":

You need to click the "Find" button to start a search for neurons in a given class and with given filters.

You must click on one or more cells given in the list to specify which you will have displayed in detail. To get two or more cells, you must (somewhat confusingly) hold down the "shift" key on your keyboard while clicking on the second or later cells (otherwise, each time you click on a cell, all the previous cells you have marked by clicking on them will be unmarked).

"Add Selected Neurons to the Field Below":

After you have selected one or more neurons in the "List", you can click on "Add Selected" to get these listed as neurons to be displayed. Alternatively, you can click on "Add All" and get all the listed neurons tallied at once for later display.

"Selected Neurons for Display":

If you have followed the procedures above, you should have one or more neurons listed in all-capital text, separated by individual spaces, in the empty field to the right of this step.

NOTE! One shortcut that is not mentioned but can be quite helpful: you can simply type in by hand one or more neurons into the empty box next to "Selected Neurons...". If you know exactly what neuron you want, this can save considerable clicking and so forth. It can also let you construct a list of neurons that fit your specifications rather than any list that WormBase has precompiled (e.g., "all the neurons that express tax-2", a set that includes both amphid and non-amphid sensory neurons, and is difficult to define by predefined sets but easy to type in). Also note that this is not forgiving of typographical errors or incomplete names: if you type "BDU", the WormBase server will not know that you mean "BDUL or BDUR or both".

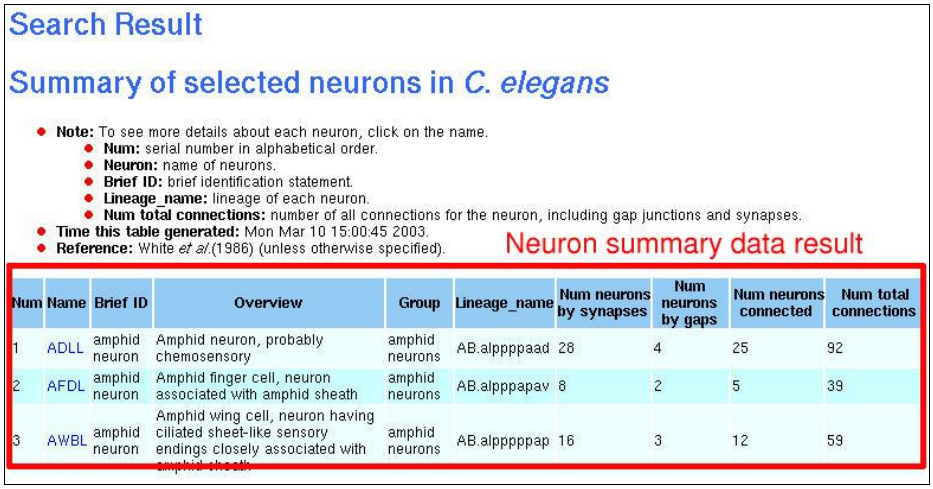
"Display Data for Selected Neurons":

This gives two different options. You can ask for "Summary Data" or "Wiring Data" by clicking on the correct button.

"For Summary Data, Select One or More Properties and Click Go Button":

This field appears after you click on the "View Summary Data" button. It gives you the choice of what fields of information you would like to display. The default ("Select All Fields") will show everything defined as "Summary Data"; other choices are more restricted. It is probably most useful to set restricted choices when there are many neurons to be compared.

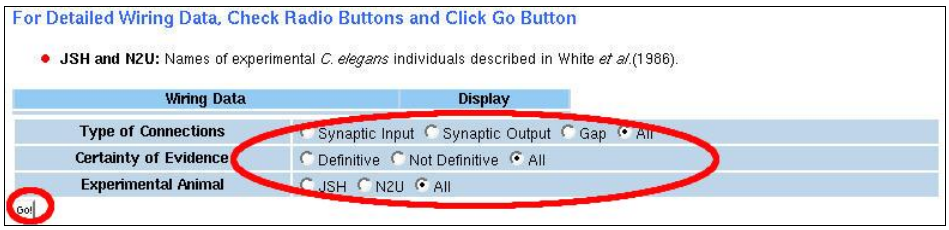
Clicking on the "Go!" button will then (finally) generate a table with the requested data fields.

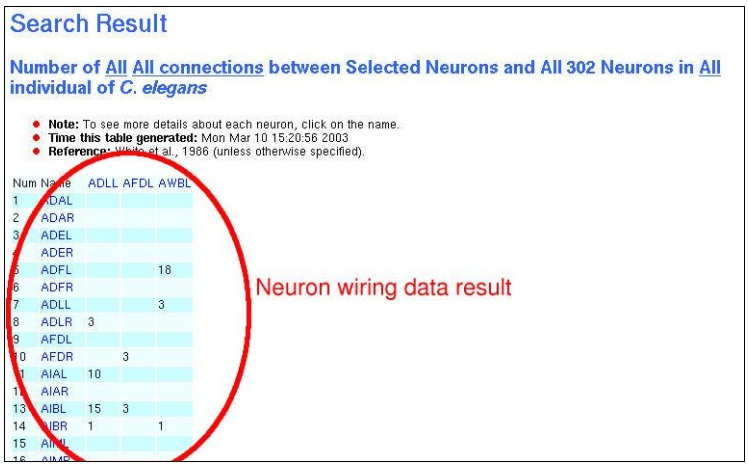


Num	Name	Brief ID	Overview	Group	Lineage_name	Num neurons by synapses	Num neurons by gaps	Num neurons connected	Num total connections
1	ADLL	amphid neuron	Amphid neuron, probably chemosensory	amphid neurons	AB.alppppaad	28	4	25	92
2	AFDL	amphid neuron	Amphid finger cell, neuron associated with amphid sheath	amphid neurons	AB.alpppapav	8	2	5	39
3	AWBL	amphid neuron	Amphid wing cell, neuron having ciliated sheet-like sensory endings closely associated with amphid sheath	amphid neurons	AB.alpppppap	16	3	12	59

"For Detailed Wiring Data, Check Radio Buttons and Click Go Button":

This gives choices of connection type, strength of evidence for connections, and the specific animal from which connections were ascertained. Again, clicking "Go!" generates the full data display.





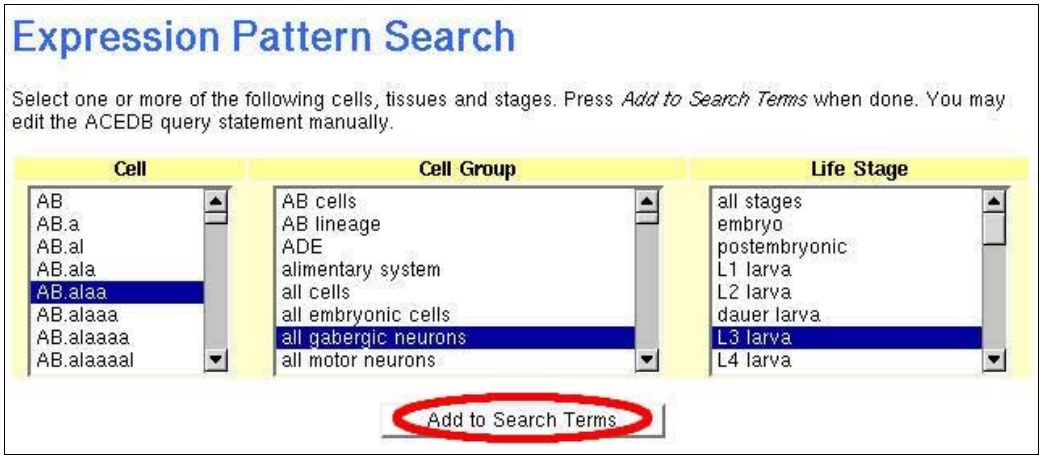
Num	Name	ADLL	AFDL	AWBL
1	ADAL			
2	ADAR			
3	ADEL			
4	ADER			
5	ADFL			18
6	ADFR			
7	ADLL			3
8	ADLR	3		
9	AFDL			
10	AFDR		3	
11	AIAL	10		
12	AIAR			
13	AIBL	15	3	
14	AIBR	1		1
15	AIML			
16	AIMR			

Expression Pattern Search

Given a cell, a Cell Group, or a Life Stage, you may wonder what kind of genes are expressed there. WormBase Expression Pattern Search provides you the tool to find this out.

SELECT SEARCH TERMS

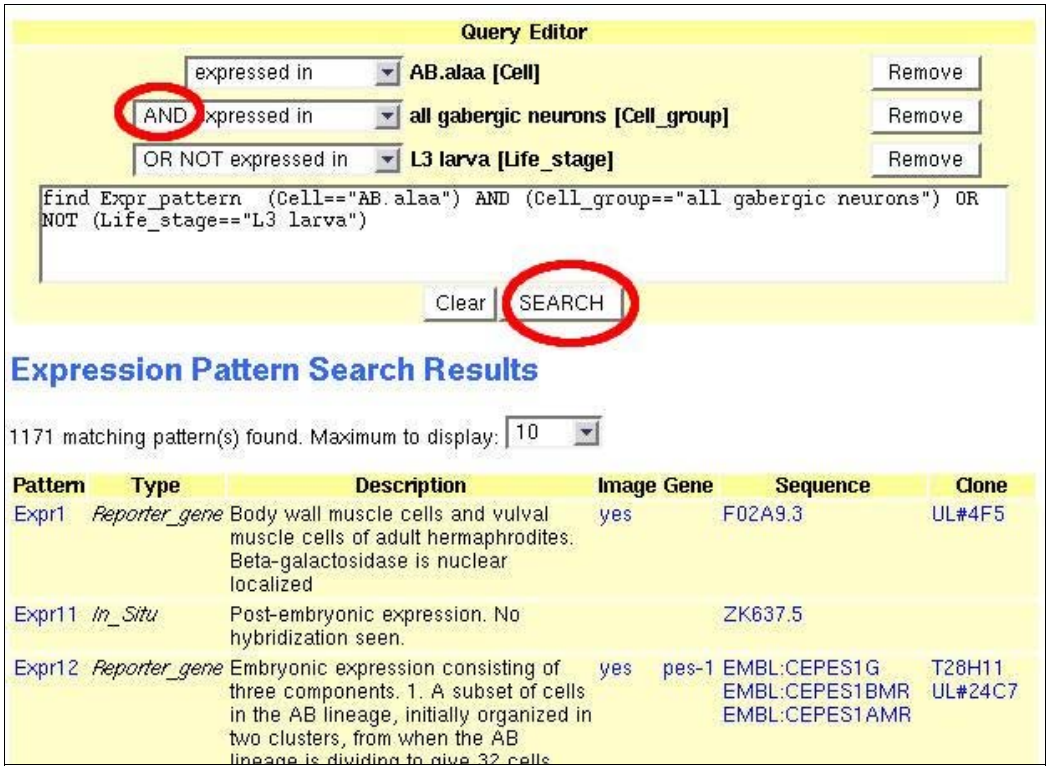
First, we need to learn how to select and remove terms. Select from Cell, Cell Group and/or Life Stage, and click the button "Add to search Terms". To select multiple terms from the same category, for example, if you want to choose both "AB.a" and "P12" from Cell, unless those terms are next to each other (use Shift key to select multiple terms in that case), you will have to add them one by one using the "Add to Search Terms" button.



CHANGE RELATIONSHIPS AMONG SEARCH TERMS

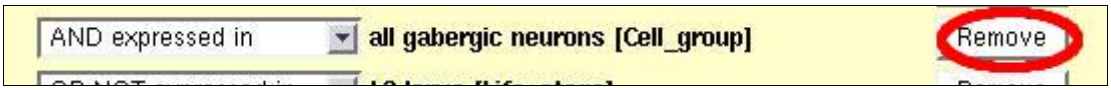
In the Query Editor, users can change the relationship among terms. There are four kinds of relationships: AND, OR, AND NOT, OR NOT.

Once the relationship is changed, the search engine will automatically start the search and list results. Or, if users do not want to change the relationship, clicking the "SEARCH" button will start the search.



REMOVE SEARCH TERMS▼

To remove a search term, click on the "Remove" button to the right of the particular search term. Clicking on it will bring you new results.



WRITE QUERY DIRECTLY

The text box at the bottom of Query Editor shows the Ace Query Language (AQL) the search engine created according to the terms and relationships users selected. Users can also directly write a query in the text box. Please check WormBase Advanced Search for instructions on how to use AQL. Or, go to http://www.acedb.org/Software/whelp/AQL/.

```
find Expr_pattern  (Cell=="AB.alaa") AND (Cell_group=="all gabergic neurons") AND
(Life_stage=="L3 larva")
```

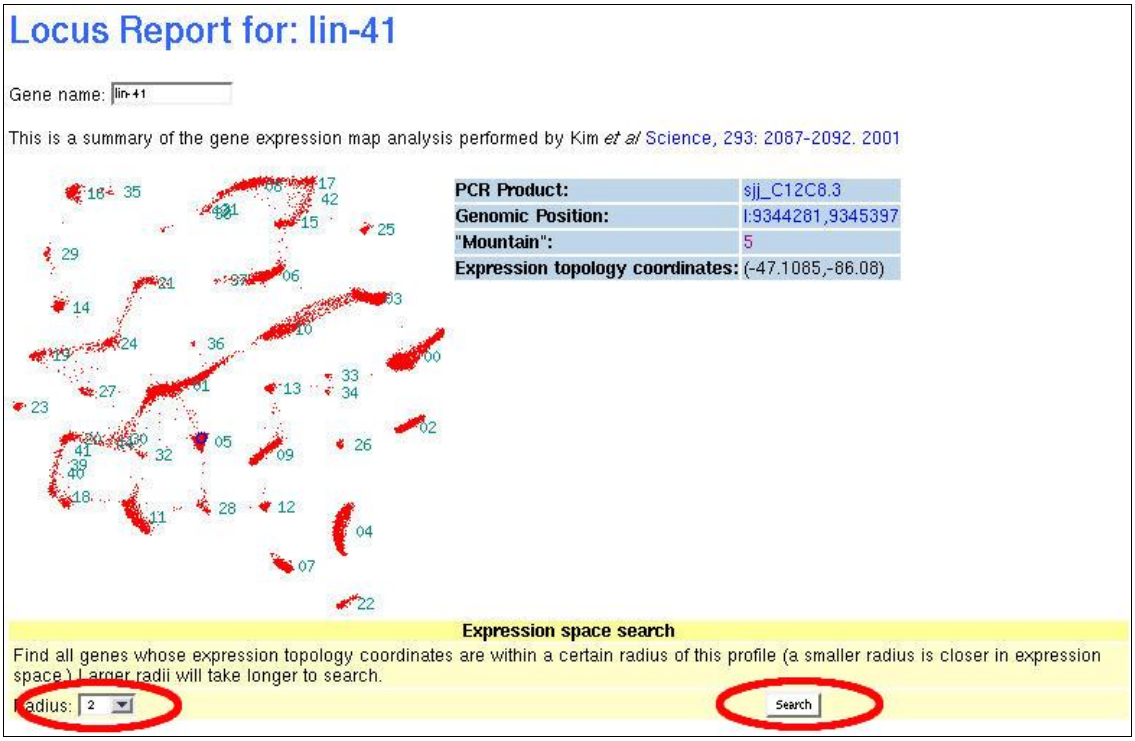
Expr_profile Search

Expression Profile is a type of gene clustering analysis based on many sets of independent and sometimes unrelated microarray data (Kim et al. Science, 293: 2087-2092. 2001). It is expressed in a topographical form.

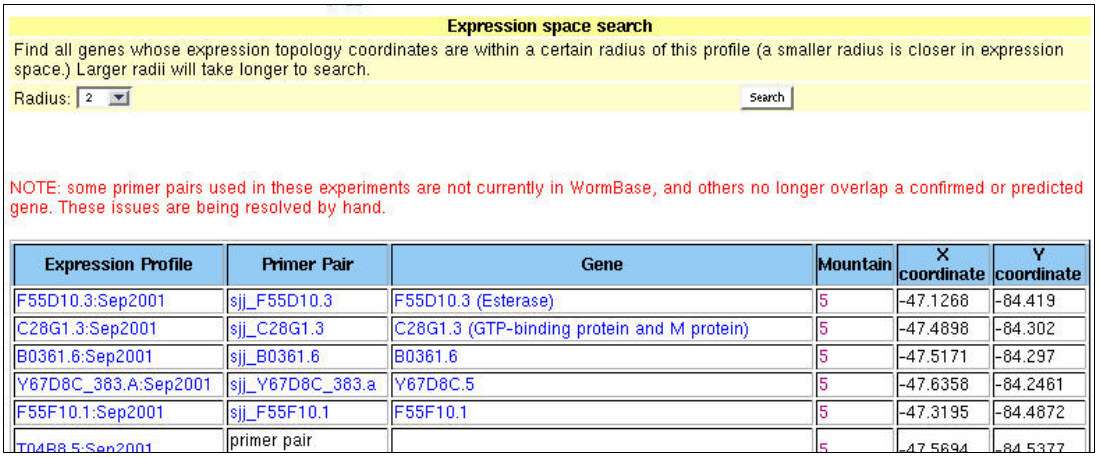
To find the location of a gene in the profile, type, for example, Gene name: lin-41. (predicted gene name such as ZK512.7 is also okay).

If the gene is found, a profile image is shown, with the gene's location marked by a blue circle. In the case of lin-41, the circle is in 'Mountain 5', a concentration of genes (red dots). The proxy for lin-41 in the microarray analysis is a PCR Product: sjj_C12C8.3. You can click on the link and get detailed information.

Clicking on "Mountain": 5 brings you to an external link at Science magazine web site where more information about this and other mountains can be found.



At the bottom of the page, you can search for genes in the neighborhood (vis-a-vis Expression Profile) by specifying your own stringency (Radius).



Expression Profile	Primer Pair	Gene	Mountain	X coordinate	Y coordinate
F55D10.3:Sep2001	sjj_F55D10.3	F55D10.3 (Esterase)	5	-47.1268	-84.419
C28G1.3:Sep2001	sjj_C28G1.3	C28G1.3 (GTP-binding protein and M protein)	5	-47.4898	-84.302
B0361.6:Sep2001	sjj_B0361.6	B0361.6	5	-47.5171	-84.297
Y67D8C_383.A:Sep2001	sjj_Y67D8C_383.a	Y67D8C.5	5	-47.6358	-84.2461
F55F10.1:Sep2001	sjj_F55F10.1	F55F10.1	5	-47.3195	-84.4872
T04B8.5:Sep2001	primer pair		5	-47.5694	-84.5377

RNAi Search

OVERVIEW: RNAi PHENOTYPE SEARCH

The ability to inactivate *C. elegans* genes by RNAi has allowed mass screens for diverse phenotypes, induced by mass RNAi-knockout of most of the genes on a chromosome or in a preselected set of functionally interesting clones. Such data is potentially of great value, but is difficult to decipher without automated searches for those RNAi results with particular characteristics. The RNAi Phenotype Search page is meant to allow such searches. It searches data that include both large-scale RNAi experiments and small-scale RNAi experiments, both of which are curated by WormBase staff as soon as possible after publication.

It is important to note that while mass RNAi screens are highly valuable, the results are -- necessarily -- frequently incomplete, and sometimes just wrong. This is unavoidable given that RNAi does not act with equal efficacy on all tissues or developmental stages in *C. elegans*, and that a mass RNAi screen can only be done on a predetermined set of well-defined phenotypes. It is thus entirely possible that an RNAi result of 'wild-type' ("WT") is misleading because the gene inactivated by RNAi acts in a tissue insusceptible to RNAi, or because there was a perfectly valid phenotype that was too subtle, specialized, or original to be seen in a mass screen. As always in science, there is no easy substitute for empirically checking the reproducibility of data oneself.

SEARCH OPTIONS

Searches of WormBase RNAi phenotypes have several different options:

- **Maternal phenotype:**

One of the most common RNAi phenotypes in mass screens is maternal sterility. If one is looking for genes with early roles in gametogenesis, it is reasonable to look for partially or completely sterile RNAi phenotypes. Conversely, to select for genes that are specific to a later stage of the life cycle, one can try screening for genes lacking such a phenotype. Sterile RNAi phenotypes are expected to identify those genes where a substantial amount of gene activity is maternal in addition to (or rather than) zygotic.

- **Embryonic phenotype:**

Another very common RNAi phenotype in mass screens is embryonic lethality. This phenotype is expected to arise without maternal sterility where an inactivated gene is zygotically active in early development. If one is interested in genes that are primarily active in postembryonic life, it would be reasonable to search for genes that are neither maternally sterile nor embryonically lethal.

- **General postembryonic phenotypes:**

A great many postembryonic phenotypes are imaginable (several hundred have been assigned gene class abbreviations as of mid 2003). However, it can sometimes be useful to combine postembryonic phenotypes into broad groups. The current groups used by WormBase are defects in growth, morphology, movement, or reproduction. In general, genes affecting metabolic control or efficiency might be expected to affect growth, those affecting development to affect morphology or reproduction, and those affecting nerve or muscles to affect movement.

- **Specific postembryonic phenotypes:**

These are listed individually, and should be reasonably self-explanatory. They are meant to be identical in character to the phenotypes that would be used for a classical genetic screen; if an animal is made hyperactive by RNAi of a given gene, that Hya phenotype should be equivalent to what would be scored as Hya if mutagenesis had been done through ethylmethane sulfonate (EMS) or through transposon mobilization.

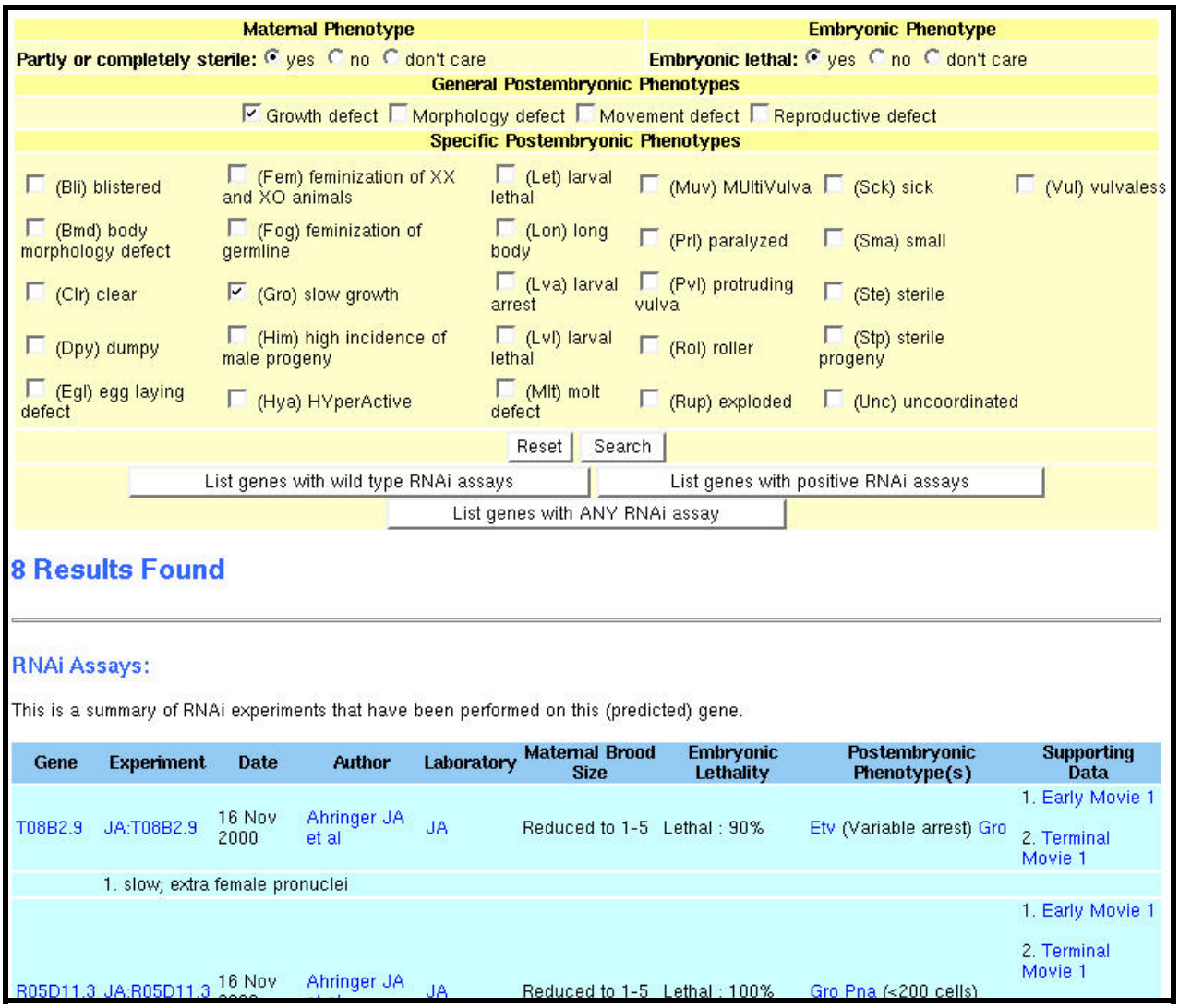
- **Combined traits:**

It is possible to search for RNAi results that combine two or more characteristics. For instance, one can search for genes

that are both Sterile and Uncoordinated when treated with RNAi; in this instance, the syndrome of joint Vul and Unc is a symptom of genes that may control postembryonic cell division, since normal proliferation of postembryonic tissues is required in adult *C. elegans* for both locomotion and fertility. Many other such combinations can be considered and used in a RNAi search.

Large lists of RNAi results:

It can be very useful to have a plain text file of all the RNAi results falling into a very broad group -- wild-type results, non-wild-type results, or any results at all. Buttons are provided on the RNAi Phenotype Search page allowing any of these lists to be displayed. Such lists can then be saved as plain (ASCII) text, and then searched by whatever means may be useful on one's own computer (e.g. through shell or Perl scripts).



Interval Search

Interval Search is a tool designed to retrieve all of the predicted genes and cosmids within a genetically defined region. This resource is extremely useful when a mutation has been mapped to a defined genetic interval. It can help identify candidate genes on the physical map that might contain the mutation, and cosmids that span the region that could be injected to assay for rescue of a mutant phenotype. To access this tool, click the More Searches link in the Main Menu at the top of any page. From there you will find the Genetic Interval link listed under the Genetics, Strains, Phenotypes header.

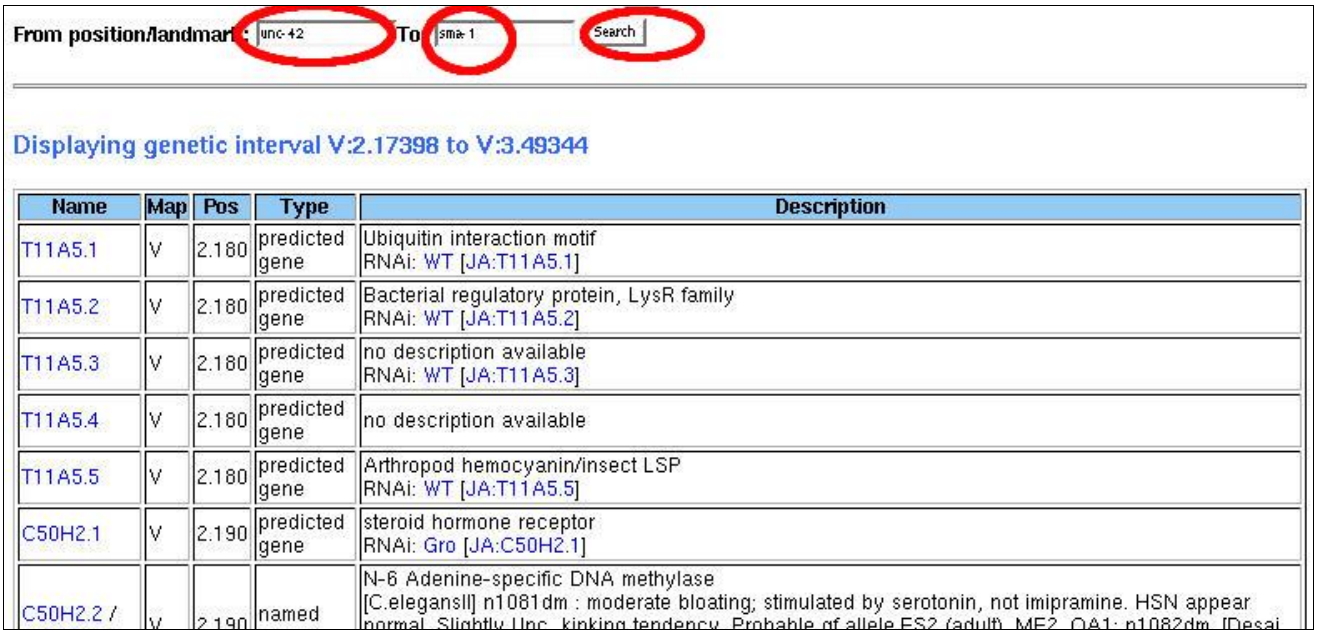
There are multiple ways to perform a search using this tool:

1) Specify a genetic interval, not more than 2 cM apart, using the names of two genetically mapped mutants, clone names, identified genes, or predicted genes (for example, evl-7 to sup-23).

2) Define endpoints using the notation chromosome:position, as in II:-4.3 to II:4.0.

3) Enter a single value to center the search on a window 0.3cM wide.

The following is an example of the search results generated using the endpoints unc-42 and sma-1:



Name	Map	Pos	Type	Description
T11A5.1	V	2.180	predicted gene	Ubiquitin interaction motif RNAi: WT [JA:T11A5.1]
T11A5.2	V	2.180	predicted gene	Bacterial regulatory protein, LysR family RNAi: WT [JA:T11A5.2]
T11A5.3	V	2.180	predicted gene	no description available RNAi: WT [JA:T11A5.3]
T11A5.4	V	2.180	predicted gene	no description available
T11A5.5	V	2.180	predicted gene	Arthropod hemocyanin/insect LSP RNAi: WT [JA:T11A5.5]
C50H2.1	V	2.190	predicted gene	steroid hormone receptor RNAi: Gro [JA:C50H2.1]
C50H2.2 /	V	2.190	named	N-6 Adenine-specific DNA methylase [C.elegansII] n1081dm : moderate bloating; stimulated by serotonin, not imipramine. HSN appear normal. Slightly Unc, kinking tendency. Probable gf allele ES2 (adult). ME2, QA1: n1082dm. [Desai

Genome Dumper

WormBase provides large sets of genomic sequence data through its Genome Dumper. This can be useful for getting data sets like the complete set of 5' flanking regions for all genes.

To run this service, do one of the following: select a pre-defined search; type in a list of sequence or chromosome names; or upload a text file with sequence or chromosome names. Note that uploaded files will need to be in plain-text to be read; Word documents are likely to fail (because they will be uploaded as the full document file with Microsoft-specific code rather than as the embedded text alone).

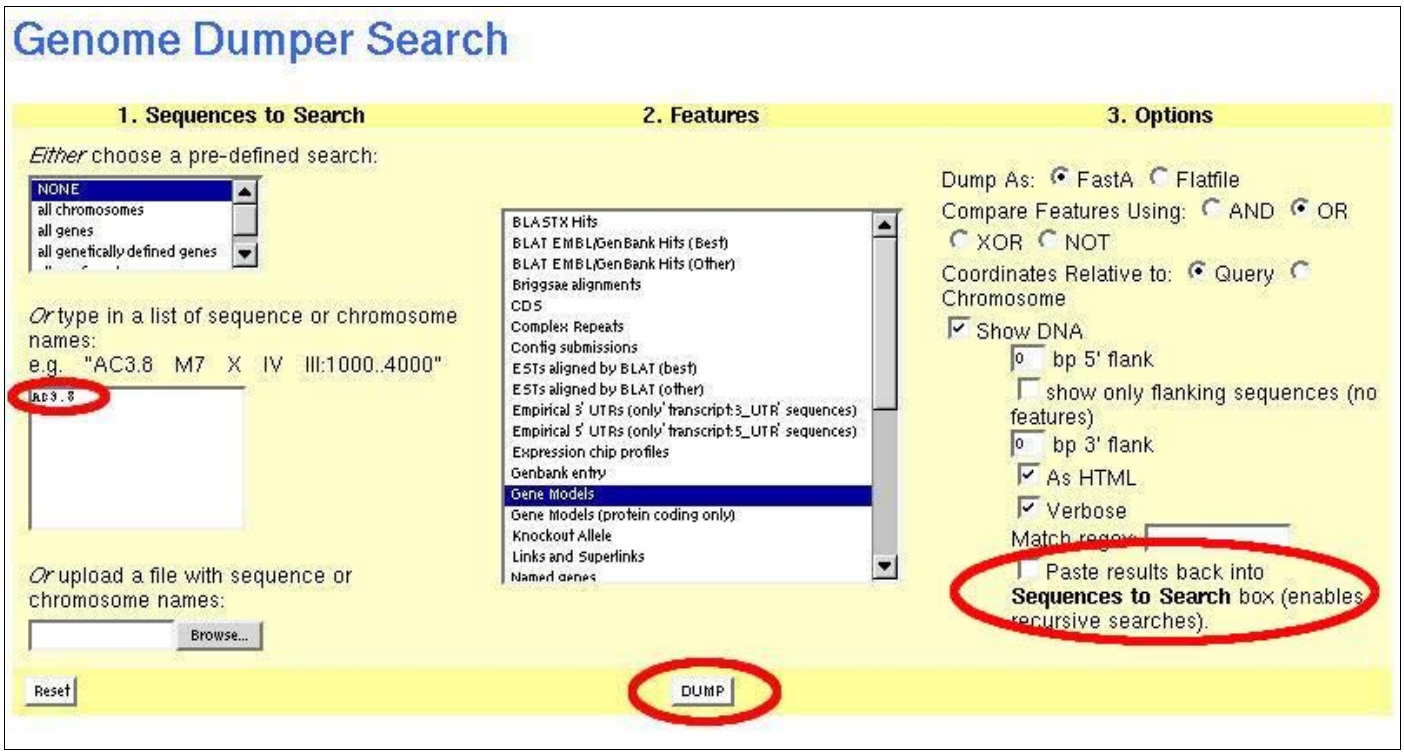
In the "Features" menu, pick a given feature that you would like to extract from WormBase. The current default is "Gene Models", which will base a search on the structures of known or predicted genes encoding either protein products (i.e., most of them) as well as non-protein-coding RNA products (e.g., miRNAs).

In the "Options" menu, use your discretion about what alternatives to select. For most normal purposes you will probably want the "FastA" default, which gives outputs in the FASTA sequence format. Unless you want to have both a flanking sequence and the gene sequence linked to it, you should check the option "show only flanking sequences". Plain (ASCII) text is often preferable to HTML, so checking *off* the "As HTML" box (on by default) may be advisable.

Then click the "DUMP" link. The resulting download will be into a window of whichever Web browser you are using. In many cases, the download will be large (for instance, in a search for flanking sequences of "all genes"). You should expect that large downloads will take some time (up to an hour); it is important not to interrupt the download by disturbing the browser's function or prematurely using a "Stop" key in the browser window.

Save the final download to a text file on your computer, and check the format of the resulting text. If you chose the "Verbose" option, you may need to reformat the headers of the sequences to get true FASTA formatting. This can be probably be done most easily with a text editor or word processor that supports text replacement across an entire file.

One important feature of the Genome Dumper is the option of recursive searches. To use this, check the "Paste results back into Sequences to Search" box in the "Options" menu before carrying out a genome dump.



Individual Pages

1. Gene Page

2. Sequence Page

3. Gene Ontology Page

4. Microarray Page

5. Person Page

Gene Summary Page

The Gene Summary page provides concise information on gene identification, genetic location, function, gene ontology, available alleles, cDNAs, protein similarities, and papers that cite the gene. Most of these classes contain multiple sub-classes with their descriptions and hypertext links to pages containing detailed information on particular subclasses, where available.

The following is a brief description of classes and subclasses on the Gene Summary page:

IDENTIFICATION

Brief ID:
Provides a concise description that allows one to rapidly assess the gene function, expression pattern, what the gene product is, etc. Clicking on the [details] link displays a more detailed gene description, if it is available.

Species:
Species of origin.

Common name:
A CGC approved gene name. Genetic nomenclature for *C. elegans* can be found here:

Other names:
Names by which the gene may also have been referred (e.g., name assigned to the gene by the sequencing consortium).

Origin:
Provides the source of information from which Locus data was extracted.

Other sequence(s):
Links to GenBank sequence entry(ies).

NCBI LocusLink:
[LocusLink] provides a single query interface to curated sequence and descriptive information about genetic loci. It presents information on official nomenclature, aliases, sequence accessions, phenotypes, EC numbers, MIM numbers, UniGene clusters, homology, map locations, and related web sites.

[AceView] shows the alignment of mRNAs and ESTs to the genome sequence, and the genes and (alternative) transcripts reconstructed from these alignments. The genes can be viewed aligned on the genome, and as spliced mRNA variants, decorated with BLASTp homologies, PFAM and PSORT motifs Stops and Met in the 3 frames, and all supporting mRNAs and ESTs. Text descriptions of the gene and its mRNAs (with useful links) are also shown.

Gene model:

Provides links to the Sequence report (Gene model) and Protein report (Protein) pages as well as SwissProt database entries, WormPep amino acid sequences and spliced/unspliced nucleotide sequences. Also contains brief information on protein similarities and domain structure.

Putative C. briggsae ortholog:
Provides a link to the putative *C. briggsae* ortholog as well as to graphical and text alignments of the corresponding *C. elegans* and *C. briggsae* genomic regions (View Synteny).

Literature citations:
Links to all papers in which the gene was cited.

Note:
Specifies whether the gene has been defined mutationally (links to locus page) and has been linked to any gene models. May contain other remarks such as corrections of gene structure or the annotation method.

LOCATION

Genetic Position:
Links to genetic map and mapping data.

Genomic Position:
Links to physical map via genome browser. Display can be customized to include different tracks.

Part of Operon:
Specifies if the gene is a part of an operon and links it to the operon report page.

Physically Mapped Clones:
Links to Clone Report pages for clones to which the gene was mapped.

Genomic Environs:
Graphical view of the physical map. Clicking on the image redirects you to the physical map page like Genomic Position link does.

FUNCTION

Mutant Phenotype:
Mutant phenotypes for studied alleles are presented. Publications from which the data were extracted are cited.

RNAi Phenotype:
Phenotype(s) produced by dsRNA interference. These could be derived from either large-scale RNAi screens or experiments that targeted the gene specifically, or both.

Anatomic Expression Pattern:
Documents gene expression based on GFP reporters, antibodies, in situ hybridization, etc. These could include subcellular localization.

Affymetrix Microarray Expression Data:
Gene expression analysis based on microarray experiments using Affymetrix chips.

SMD Microarray Expression Data:
Gene expression analysis based on microarray experiments using Stanford Microarray chips.

Microarray Topology Map:
Links to the three-dimensional expression map that displays correlations of gene expression profiles as distances in two dimensions and gene density in the third dimension. The gene expression map can be used as a gene discovery tool to identify genes that are co-regulated with known sets of genes or to uncover previously unknown genetic functions.

For more information on the topology map, visit:
http://cmgm.stanford.edu/~kimlab/topomap/c._elegans_topomap.htm

Protein Domains:
List of detected Pfam and InterPro sequence motifs.

GENE ONTOLOGY CLASSIFICATION

The goal of the Gene Ontology project is to produce a dynamic controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing. The three organizing principles of GO are biological process, cellular component and molecular function.

Biological Process:
Describes broad biological goals, such as mitosis or purine metabolism, that are accomplished by ordered assemblies of molecular functions.

Cellular Component:
Describes subcellular structures, locations, and macromolecular complexes such as nucleus, telomere, and origin recognition complex.

Molecular Function:
Describes the tasks performed by individual gene products; examples are transcription factor and DNA helicase. For more information on GO, visit:

http://geneontology.org

ALLELES

Reference Allele:
Most often the mutation that has been studied in greatest detail. Usually, though not always, the reference allele is the allele with the strongest and most obvious effects. As a result, in a few cases the reference allele is a gain-of-function allele, but most often it is the strongest known loss-of-function allele. In cases where both viable and lethal alleles are available, it will usually be the strongest viable allele. Also known as canonical allele.

Alleles:
Other known alleles of the gene.

Strains:
Genetic strains that are available from CGC.

SIMILARITIES

List of best BLAST matches using the longest protein product as a query. Links to identified sequences are provided together with brief descriptions, E-values and consensus length.

REAGENTS

Transgene Strains:
Strains carrying chromosomally integrated arrays. These could be reporter arrays, overexpression arrays, arrays carrying a mutated gene, etc.

Primer Pairs:
List of primers that amplify part of the gene or complete gene. Usually are designed by the ORFeome project or are used in SMD Microarrays.

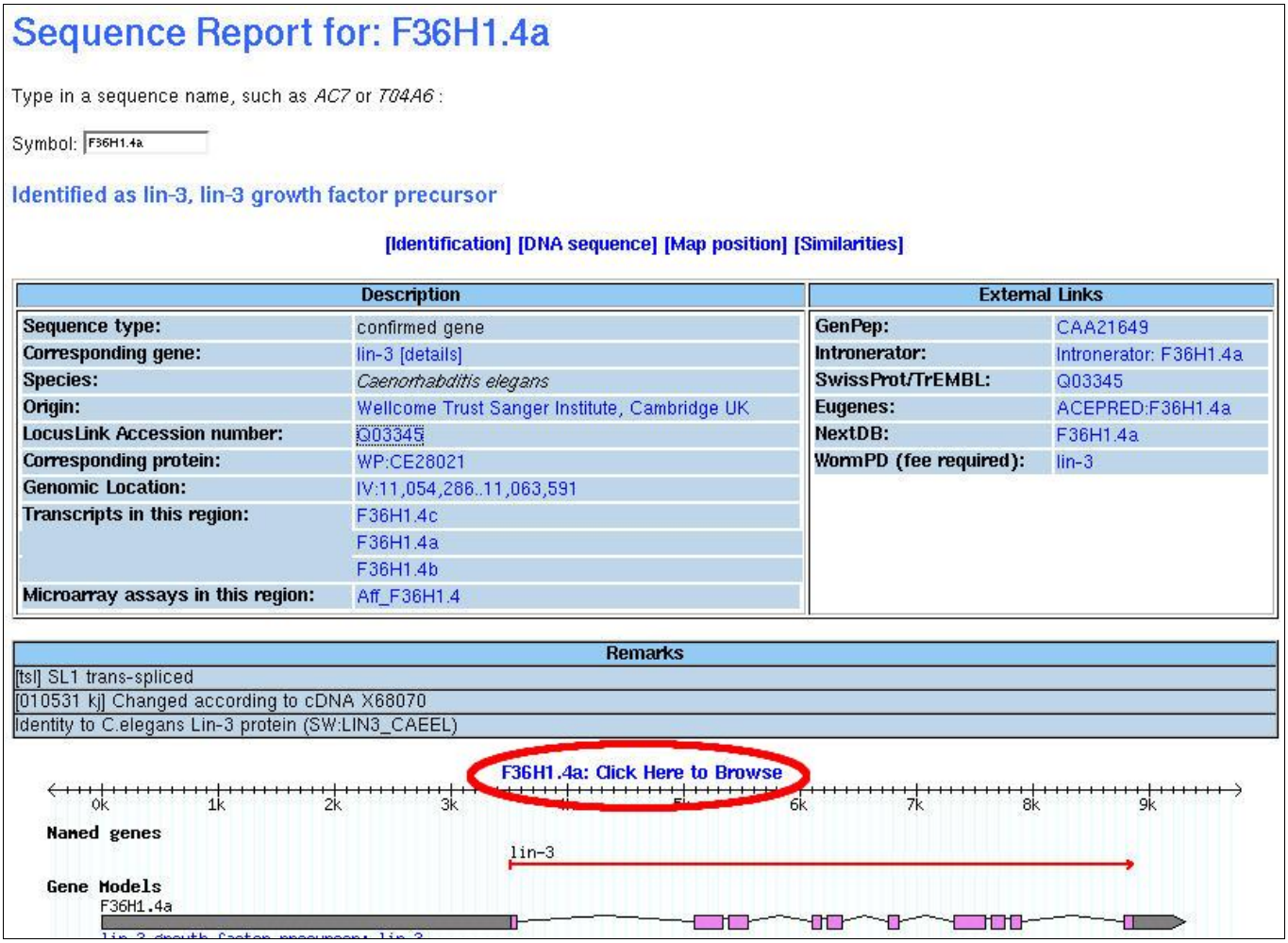
Matching cDNA:
List of available cDNA clones.

BIBLIOGRAPHY

List of literature that cited the gene. This list could contain both published papers and unpublished results such as meeting abstracts.

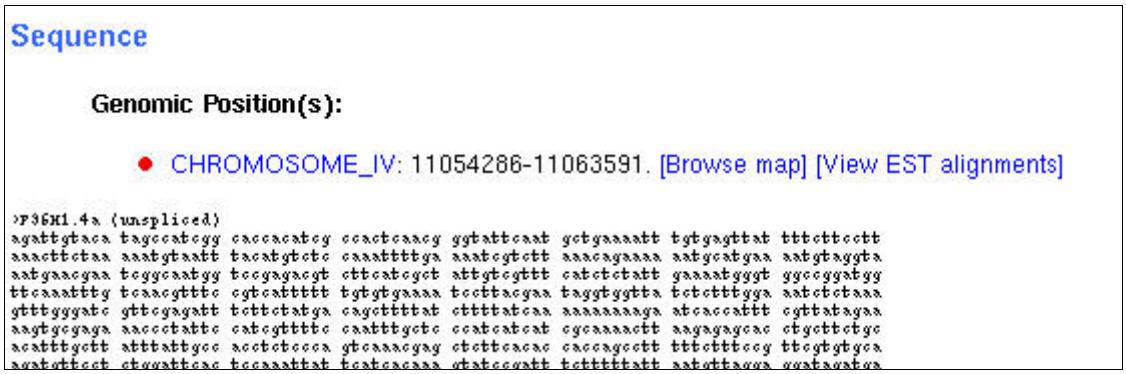
Sequence Report Page

A Sequence Report page provides those data for a given gene that are most pertinent to its genomic sequence rather than to its more general, classical biology. A schematic diagram is given for the gene's structural organization within its genomic environment; a "Click Here to Browse" link leads to a dynamic view of that environment within the Genome Browser. The schematic diagram will show classical gene names where they exist, structural models for the genes, individual frames of the coding exons in the genome, and STSes, SNPs, and transposons that can be used for genomic mapping or site-selected mutagenesis.

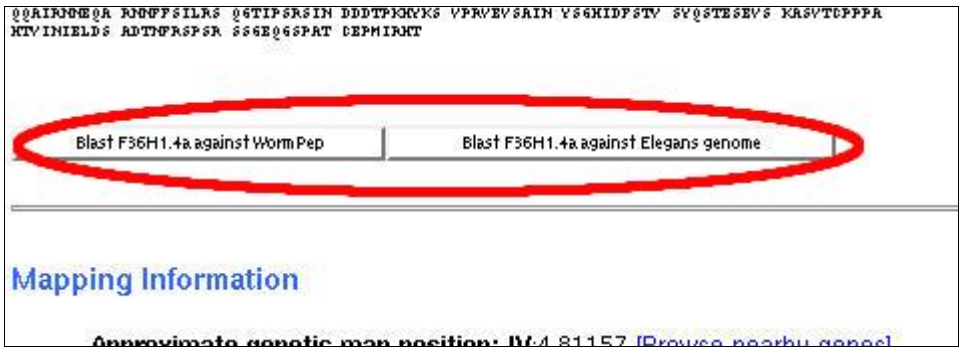


Other genomic data will include genetic identity to a classical locus (where it exists), precise numerical coordinates for the gene's transcriptional unit, identities of cDNAs matching it, links to the Protein page for its gene product (where applicable) and to its *C. briggsae* ortholog (where one is known to exist). Other data such as the source clone, sequencing center responsible for the data, microarray data, remarks on the sequence annotation, etc. are given.

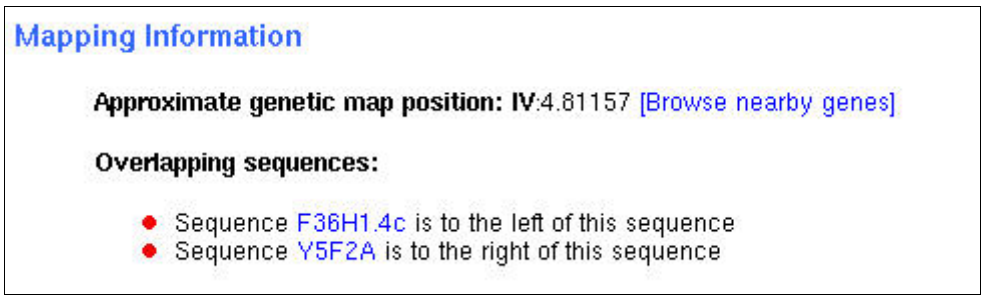
Further down the page will be sequence data in FASTA format: unspliced genomic sequence data (with exons noted in upper-case, versus introns in lower-case); the spliced coding sequence in nucleotides; and the predicted protein sequence of the gene's product. FASTA sequence formats are perhaps the most common and simple representation of sequence data, and can be handled by a great many different analytical programs.



To aid analysis of these data, a sequence page provides a preformatted Blast search with that sequence (in DNA or protein form) against the *C. elegans* genome. To set up the search, all you need to do is click on the "Blast search" link.



Moreover, each page gives a genetic map position value for its sequence, with a link to a list of nearby genes. For classically studied genes, this map function is mainly useful for designing tests of allelism. However, many genes in *C. elegans* are identified solely as protein-coding sequences (CDSes) in the genome; for genes of this type, the Mapping Information feature allows the CDS to be interpolated directly onto the classical genetic map, which, in turn, can suggest testable hypotheses of which classical mutations might reside within a given CDS.



Three useful introductory texts on practical sequence analysis of FASTA files (and other data) are:

Jambeck, P. and Gibas, C. (2001). Developing Bioinformatics Computer Skills. O'Reilly, Sebastopol, CA. ISBN: 1-56592-664-1.

Mount, D.W. (2001). Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY. ISBN: 0-87969-608-7.

Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, 2cd ed. A.D. Baxevanis and B.F.F. Ouellette, eds. Wiley-Interscience, New York. ISBN: 0-471-38391-0.

Gene Ontology Page

The Gene Ontology (GO) project (www.geneontology.org) has a central goal of building a controlled vocabulary that captures three different aspects of a gene product in a living organism: Biological Process or Processes (which is essentially, the reason why a gene product benefits the organism whose genome encodes it); Molecular Function or Functions (which is what the gene product does at a basic biochemical level) and Cellular Component (which is where in a cell the gene product is normally expressed).

The Gene Ontology project is extensively documented at the official web page of the GO consortium: http://www.geneontology.org

It has also been described in the following papers:

- Ashburner et al. (2000). Gene ontology: tool for the unification of biology. Nat. Genet. vol. 25, pp. 25-29.
- The Gene Ontology Consortium (2001). Creating the gene ontology resource: design and implementation. Genome Res. vol 11, pp. 1425-1433.

Currently GO terms for a given gene product can be found in the Gene Ontology Classification section on the Gene Summary page.

GO terms are grouped under the three different Ontology Types of Biological Process, Cellular Component and Molecular Function.

The annotation of a gene product to a GO term includes the following:

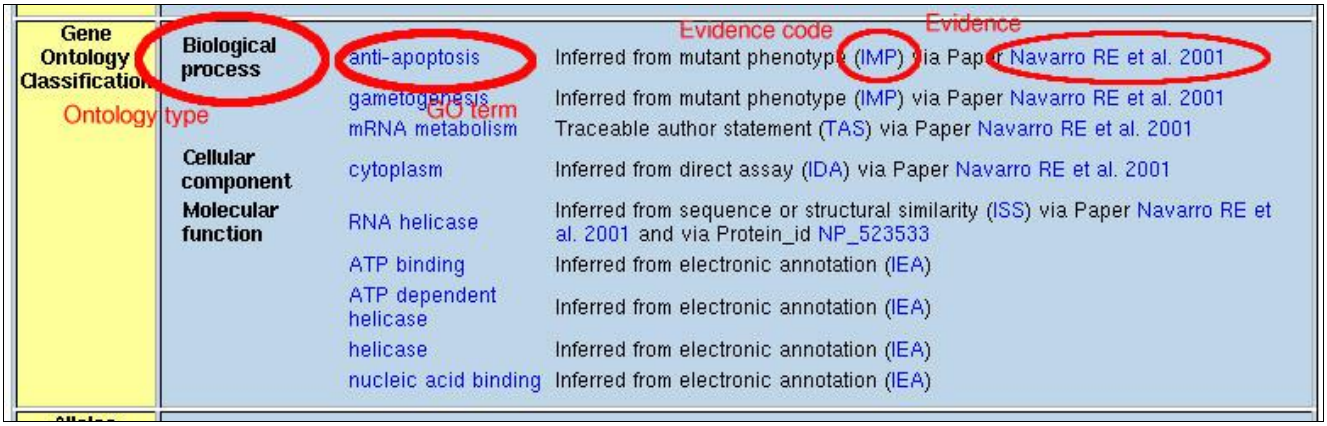
Term name: This is the primary name of the GO term assigned by the GO consortium to which the gene product is annotated. Clicking on the GO term links to the GO browser page view which displays the summary for the GO term. See below for section on 'Gene Ontology browser'.

Evidence: To the right of the GO term is the evidence on which this annotation is based, essentially the source to which the annotation is attributed. In most cases, especially in curator-performed annotations this will be a literature reference, this may also be another database or a method in computational analysis. Sometimes the annotation is based on homology to a known protein in which case it is known as 'protein_id evidence' and includes the GenBank id for the cited protein. Both curator-performed, literature-based annotations and those obtained from computational analysis-'IEA annotations' are displayed by WormBase. Clicking on the reference links to the citation display page.

Evidence code: To the immediate right of the GO term is the 'evidence code' which is the type of evidence found in the cited source to support the association of the gene product with the GO term, e.g. Inferred from mutant phenotype (IMP), Inferred from direct assay (IDA), Inferred from genetic interaction (IGI), etc. Clicking on the evidence code links to the Gene ontology consortium's documentation on evidence codes and their usage.

The annotations in each Ontology Type are displayed according to the following priority of order:

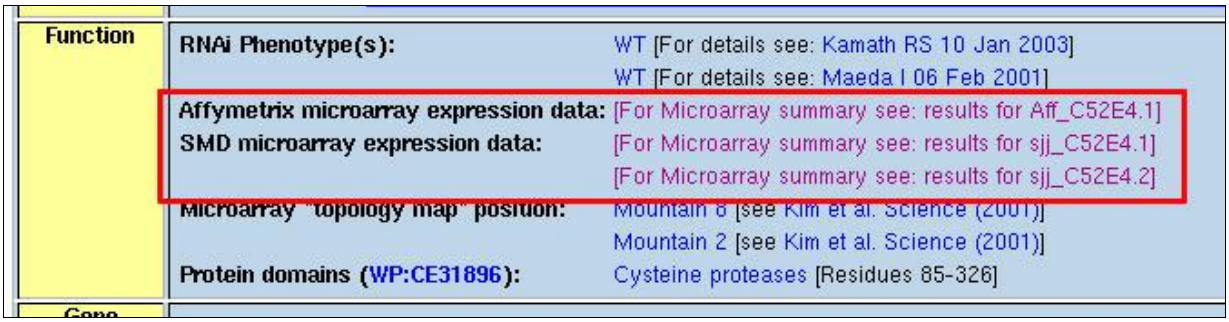
- IDA=Inferred from Direct Assay
- IMP=Inferred from Mutant Phenotype
- IGI=Inferred from Genetic Interaction
- IPI=Inferrred from Physical Interaction
- IEP=Inferred from Expression Pattern
- IEA=Inferred from Electronic Annotation
- TAS=Traceable author statement
- IC=Inferred by curator
- NAS=Non-traceable author statement
- ND=No biological data available



Microarray Page

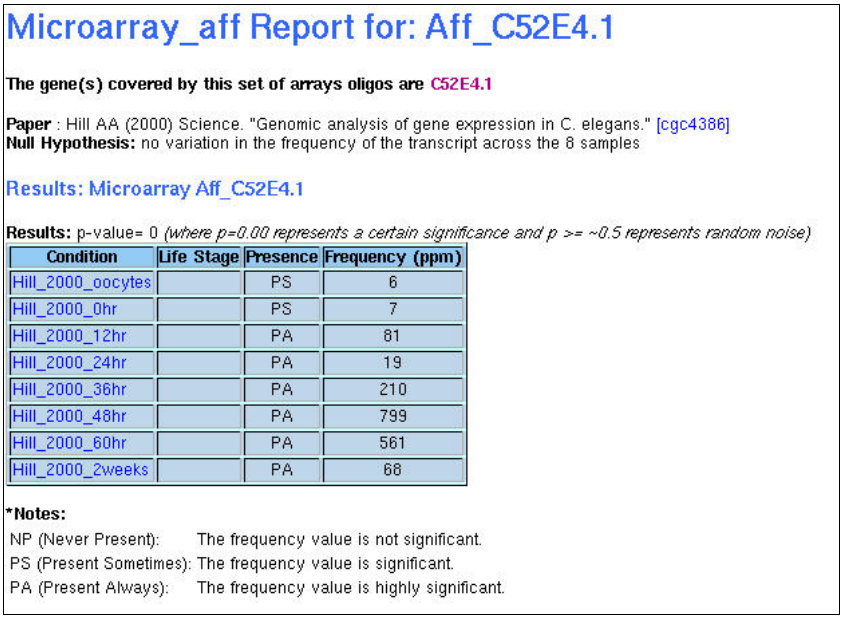
Microarray Data:

WormBase holds published microarray data performed using either Affymetrix chips or chips from the Stanford Microarray Database. These data can be accessed from the Gene Summary Page for each individual genes. Due to the completely different experimental procedures, the two kinds of microarray data are displayed differently.



Data from one set of experiments are listed as a table. Sometimes you will see two tables from one paper, because two PCR products (SMD) or target sequences (Affymetrix) were used to probe one gene. The PCR products and target sequences (Microaray_aff) are dynamically mapped to the *C. elegans* genome, so that the information will not become outdated. Please notice that the name of PCR_product or Microarray_aff may not represent the actual gene to which it corresponds.

Sample display table for Affymetrix microarray data: ►



Microarray_aff Report for: Aff_C52E4.1

The gene(s) covered by this set of arrays oligos are C52E4.1

Paper : Hill AA (2000) Science. "Genomic analysis of gene expression in C. elegans." [cgc4386]
Null Hypothesis: no variation in the frequency of the transcript across the 8 samples

Results: Microarray Aff_C52E4.1

Results: p-value= 0 *(where p=0.00 represents a certain significance and p >= ~0.5 represents random noise)*

Condition	Life Stage	Presence	Frequency (ppm)
Hill_2000_oocytes		PS	6
Hill_2000_0hr		PS	7
Hill_2000_12hr		PA	81
Hill_2000_24hr		PA	19
Hill_2000_36hr		PA	210
Hill_2000_48hr		PA	799
Hill_2000_60hr		PA	561
Hill_2000_2weeks		PA	68

***Notes:**
NP (Never Present): The frequency value is not significant.
PS (Present Sometimes): The frequency value is significant.
PA (Present Always): The frequency value is highly significant.

Sample display table for SMD microarray data:

SMD Microarray Results for PCR sjj_C52E4.1

The gene(s) covered by this set of arrays oligos are C52E4.1

Paper : Jiang M (2001) Proceedings of the National Academy of Sciences USA. "Genome-wide analysis of developmental and sex-regulated gene expression ...." [cgc4489]

Microarray_results: Microarray sjj_C52E4.1

Chip type: SMD

Microarray result	Sample A	Sample B	Life Stage	Log Ratio of test/control	Test/control SD	Cluster
cgc4489_C52E4.1_egg	cgc4489_egg	cgc4489_ref	embryo	-3.46175	0.609304	cgc4489_group1
cgc4489_C52E4.1_L1	cgc4489_L1	cgc4489_ref	L1 larva	-0.426	0.436261	cgc4489_group1
cgc4489_C52E4.1_L2	cgc4489_L2	cgc4489_ref	L2 larva	1.19425	0.614367	cgc4489_group1
cgc4489_C52E4.1_L3	cgc4489_L3	cgc4489_ref	L3 larva	0.808667	0.313857	cgc4489_group1
cgc4489_C52E4.1_L4	cgc4489_L4	cgc4489_ref	L4 larva	1.21633	1.27387	cgc4489_group1
cgc4489_C52E4.1_adult	cgc4489_adult	cgc4489_ref	adult	1.24133	0.498166	cgc4489_group1
cgc4489_C52E4.1_m_h	cgc4489_male	cgc4489_herm	adult male	-0.3655	0.629415	cgc4489_group1

Clustering Data:

WormBase also present published Microarray clustering data. The information can be accessed from the table displaying microarray data.

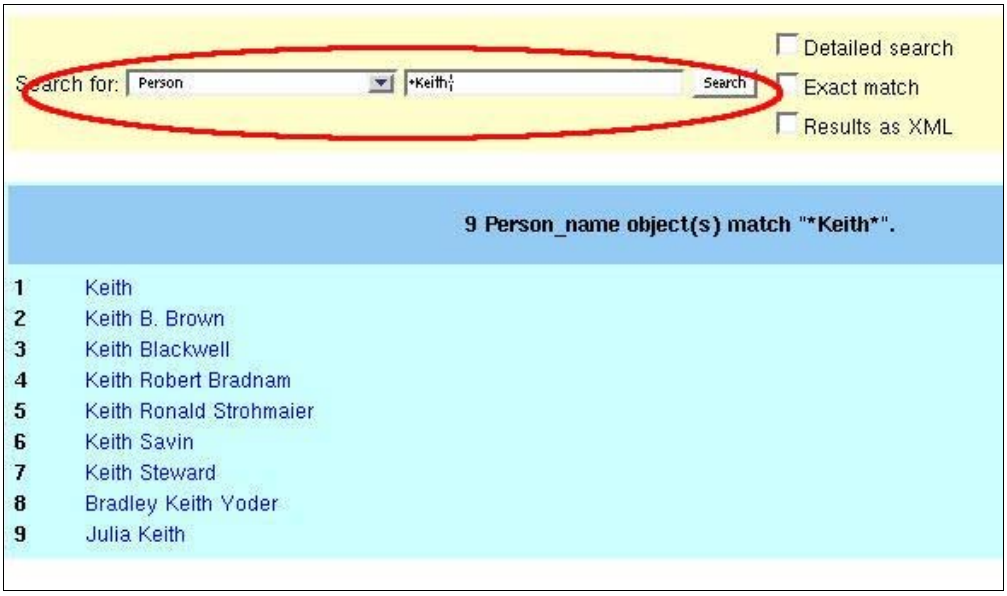
More Details:

For more details on how the Affymetrix and SMD microarray experiments were performed, please view the documents provided by the authors. Here is a page provided for Affymetrix microarray provided by Craig Hunter.

Person Page

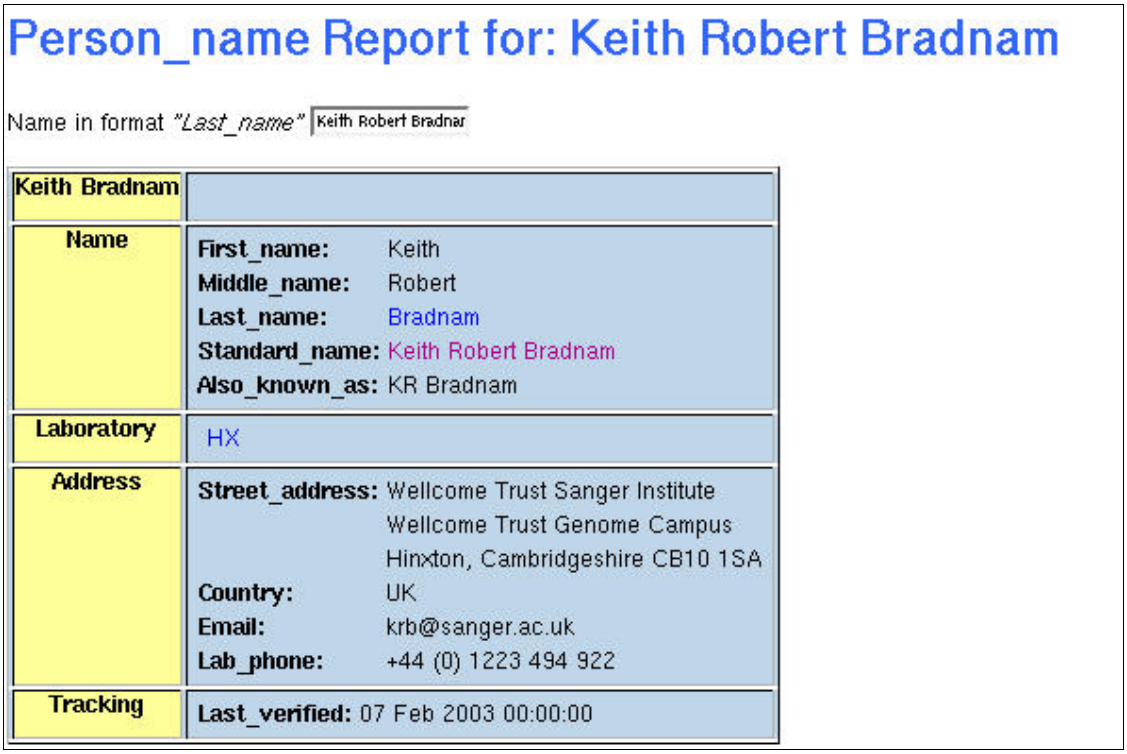
SEARCH FOR PERSON INFO

WormBase stores information about people who work in the worm field. Users can access this from the Simple Search page by specifying the Object type "Person" and typing at least part of the name of the person in the text box. Users can search with the last name and/or first name of the person.



PERSON REPORT PAGE

WormBase maintains up-to-date information about people who work in the worm field. The Person Report page displays the information which has been submitted to WormBase by the people themselves. The Person Report shows contact information and also, the last time this information was verified (see "Last_verified" field).



Person_name Report for: Keith Robert Bradnam

Name in format *"Last_name"* Keith Robert Bradnar

Keith Bradnam	
Name	**First_name:** Keith **Middle_name:** Robert **Last_name:** Bradnam **Standard_name:** Keith Robert Bradnam **Also_known_as:** KR Bradnam
Laboratory	HX
Address	**Street_address:** Wellcome Trust Sanger Institute Wellcome Trust Genome Campus Hinxton, Cambridgeshire CB10 1SA **Country:** UK **Email:** krb@sanger.ac.uk **Lab_phone:** +44 (0) 1223 494 922
Tracking	**Last_verified:** 07 Feb 2003 00:00:00

Other Resources provided by WormBase

1. How to Obtain Reagents

How to Obtain Reagents

- **How do I request a strain?**

In the Strain page, there is tag called "Location", which means the location of the strain, usually, the location is at CGC, or a two-character name standing for a laboratory. CGC (*C. elegans* Genetics Center) has most of the strains in stock and they send out/receive strains to/from *C. elegans* Labs. Please view the following page for their contact information: http://biosci.umn.edu/CGC/Strains/request.html

If the strain is not available in CGC, you have to contact directly with the Lab, whose information can be directly accessed from "Location" data field. Please remind them also to send it to CGC for the convenience of other researchers in the future.

- **How to obtain strains that contain transposon insertions? Who do I contact?**

The best way to find strains that carry transposon insertions is to first identify those strains in WormBase, then contact authors directly.

Assuming you are looking for insertions within a specific region, you can browse for insertions in your region of interest using the Genome Browser (you'll need to first check "transposons" under the display settings).

After finding a transposon insertion, clicking on it will display it's laboratory of origin, strains carrying the insertion, etc. Some of these strains may be available directly from the CGC. Alternatively, you may wish to contact the lab of origin for its availability.

--Todd Harris

- **How do I obtain C.elegans cosmids?**

As from January 1st 2003, please send all requests for *C.elegans* genomic clones to Audrey Fraser (aef@sanger.ac.uk) and cc to Ratna Shownkeen (rs2@sanger.ac.uk). Please remember to include a FedEx/DHL Account number to which we can charge the shipping cost.

On a related matter, please note that the suffixes to YAC names (e.g. Y73B3A, Y73B3B) indicate sequenced segments of the YAC (Y73B3 in this case). They do not exist as subclones. You will be sent the whole YAC.

--Alan Coulson

- **How do I get a YK cDNA clone?**

Please send emails to Dr. Yuji Kohara (ykohara@lab.nig.ac.jp).