WormBase Newsletter #1 October 2001

Introduction to the Newsletter

We are starting this Newsletter to keep you informed of changes at WormBase (http://www.wormbase.org). Many of you use WormBase frequently and thus see that improvements and additions are listed on the homepage. Others of you might only use it when some data are included or some features are added, and this newsletter will help you decide when it is time to pay more attention. We will send out Newsletters every 2-3 months.

Over the past months, there have been a number of changes, many of which are described below.

-Paul Sternberg

Release schedule

One of the most important improvements to WormBase over the past 12 months has been a drastic increase in the frequency of database updates. WormBase is built in two stages. In the first stage, a complete database (ACeDB) is built incorporating changes and additions from all WormBase sites. This is done weekly, and unless there is a major problem, the latest ACEDB can be downloaded weekly. The version of the database is listed on the WormBase homepage as WS#, e.g., WS57 for the October 12, 2001 update to WormBase site. In the second stage of construction, the database is configured to support the wormbase.org website. This is now done about every two weeks.

We are close to having mirror sites at the Sanger Centre and Caltech; see the WormBase homepage for details. WormBase can be downloaded from Cold Spring Harbor or the Sanger Centre websites.

New data

RNAi. 147 movies from RNAi experiments from the Ahringer laboratory (Zipperlen et al., 2001) are now included in WormBase.

Expression patterns from papers. We are focusing on extracting gene expression patterns from the 4630 papers in the CGC bibliography. We are almost half done, and now have with 1297 Expr patterns representing about 581 genes. In general, each experiment or cluster of related experiments is described in one Expr object. For example, if a gene's expression has been analyzed by GFP fusions and by antibodies, there will be two Expr objects in WormBase.

Global microarray clustering data. Summary expression pattern data (the "expression map") from S. Kim et al., Science, September 2001, is now available on the gene and sequence pages, and in a new expression profile page, which allows you to search for genes with similar expression profiles.

WTP genes. The regions corresponding to over 10,000 genes from the Worm Transcriptome Project's analysis of EST sequences have been added this summer. The Thierry-Miegs might

have additional information on splicing patterns of individual genes, and you should email them for more information.

WormPep. WormPep is a set of current best inferences about proteins encoded in the *C. elegans* genome. Since WormPep is now revised weekly, you can obtain the data for previous versions at http://www.sanger.ac.uk/Projects/C_elegans/wormpep/.

User Interface

The WormBase user interface is still very much evolving. Some of the changes are:

Genome Browser. The genome browser Genome Hunter has been updated to show predicted and confirmed genes, the precise endpoints of cosmids and YACs, ESTs aligned by BLAT (see below), the regions of genomic sequence corresponding to genes defined by the Worm Transcriptome Project's analysis of Y. Kohara's ESTs, regions of homology to *C. briggsae*, regions with Prosite domains, the oligos and regions they amplify that have been used in some microarray and RNAi experiments, ESTs, among other features. These features are color-coded in the display. You can check boxes to specify the features you would like to see.

BLAT is a sequence alignment program written by W.J. Kent at UC Santa Cruz. It efficiently scans a pair of DNA sequences for small regions of high identity: those 40 or more bases long with 95% identity, or perfect sequence matches down to 33 bases in length. It is highly useful for aligning cDNAs to genomic DNA, or small genomic fragments to a genome draft.

Genetic Map Viewer. The new genetic map viewer that became available this past spring is Java-based and still not compatible with Macintoshs. We therefore enhanced the classic acedb graphic map display to make it easier to navigate. A new more web-friendly viewer is under development.

Search pages

Genetic Interval Search. A new genetic Interval Search page takes advantage of the interpolation of genetic and physical maps at the resolution of individual clones. This search allows you to specify a range by map position, gene name, or clone, and returns a list of genes in that region. After determining the range, this script lists all mapped mutants within the range as well as predicted genes on clones that have been interpolated into the range. Of course, since not all genetic loci are mapped relative to one another, the order of genetic loci presented in chromosomal coordinates may not actually reflect the physical order of these genes.

RNAi Phenotype Search. An RNAi search page allows you to search for genes for which RNAi experiments have been done. Most of these are from the large scale projects published in the past year, and an increasing set from individual papers. Negative data from all but the EMBL screen are included.

Coming Soon

SNPs. The positions of the Washington University SNPs will be included in the Genome Viewer.

Deletions from the *C. elegans* Knockout Consortium will be indicated in the genome viewer and on the Gene report pages.

***C. briggsae* data.** The assembled genomic sequence from *C. briggsae* generated by 10x coverage shotgun sequencing at the Sanger Centre and at the Washington University Genome Sequence Center should be in WormBase this Fall.

Microarray data. Other published data will be added in the near future.

Gene Ontology Consortium

WormBase has joined the Gene Ontology (GO) Consortium. GO is a structured vocabulary allowing the biological functions of gene products to be described with arbitrarily high levels of detail, and compared between diverse organisms in a way independent of sequence similarity or idiosyncrasies of a given model system. More details are available at http://www.geneontology.org.

We have begun incorporating GO terms into Wormbase. The first step was to automatically generate annotations based on the Interpro repository of protein sequence motifs, which has become the standard for computational annotation of protein-encoding portions of whole genomes (e.g., the Arabidopsis and human genomes). A second step, currently underway, is to automatically map GO terms onto ~800 genes with mass-produced RNAi phenotypes. This is being done in collaboration with the WormPD database at Proteome, Inc. The longer term and most important phase is to manually annotate each gene with GO terms. During all of this, it is continually necessary to invent new GO terms specifically fitted to the biology of *C. elegans*. Another basic requirement is to develop a logical scheme (ontology) relating parts of the anatomy; this is being done in collaboration with David Hall and Zeynep Altun of the Worm Atlas Project.

WormBase People

Allen Day has left CSH for graduate school in Bioinformatics at UCLA.

Positions at WormBase

The GO Consortium's grant will fund a curator position at WormBase starting January 1, 2002 (please contact Paul Sternberg for more information).

Funding sources

WormBase is in its second year of funding from the National Human Genome Research Institute (# P41 HG02223), and receives additional support from the British Medical Research Council.