# WormBase Newsletter October 2003

## WormBase Version Freeze

In the spring, several bioinformaticians requested freezes for WormBase releases so that they could both work with and refer to a specific version of the *C. elegans* genome and its biological annotation. To allow doing this, we are freezing every tenth release of WormBase, e. g., WS100 (http://ws100.wormbase.org) and WS110 (http://ws110.wormbase.org). These reference releases are available for perpetuity on dedicated servers.

## Confirming gene structure in WormBase

WormBase classifies gene structures (for protein-coding genes) into the following categories:

**Predicted**: A gene that has no matching cDNA information (mRNAs, ESTs, or OSTs), the structure of which was therefore most likely suggested by a computer prediction (GeneFinder) and has probably not been altered since the initial prediction was made.

**Partially confirmed**: A gene which has partial but not complete support from cDNAs. More specifically, these are genes where the matching cDNAs match some bases of every predicted exon. The areas where there are no matching cDNAs may reflect an incorrect gene structure or an absence of available cDNAs for that part of the gene.

**Confirmed**: This is an extension of the previous category where every base in an exon of the gene prediction has cDNA support. Some gene predictions may be incorrect (i.e. missing an exon). This category usually represents our most confident gene structures, however, some gene predictions may be incorrect (i.e missing an exon). A FASTA file of confirmed genes, many of which have corresponding full-length mRNA sequences, is available at ftp://ftp.sanger.ac.uk/pub/wormbase/current_release/confirmed_genes.WS112.gz.
The numbers below, illustrate the fact that the regular addition of new cDNAs to GenBank/EMBL and the curation process of matching gene structure to available cDNA information changes the gene number in each category, e. g.

| RELEASE | CONFIRMED | PARTIAL | PREDICTED | TOTAL |
|---|---|---|---|---|
| WS111 | 4609 | 12079 | 5505 | 22193 |
| WS100 | 3889 | 8848 | 8828 | 21565 |
| WS75 | 2981 | 7103 | 10307 | 20391 |

The steady increase in the number of confirmed genes is a good indicator of the ongoing curation work. Of the ~5,000 predicted genes that remain, several will probably never have transcript support due to their extremely low level of expression, for such genes we might have other information that suggests that they are real genes (e. g. RNAi information).

## How complete is the *C. elegans* genome?

The genome sequence was essentially "finished" in 1998 (Science Vol.282, 5396) with a genome size of 97Mb. The genome sequence was derived from 2,500 cosmids, 250 YACs, 100 fosmids and 40 PCR products. At that point in time there were still 5 major clone gaps as well as many sequence gaps. The genome had been annotated with approximately 19,000 protein coding genes. Since then there has been a 2.4% increase in the overall size of the genome, mainly due to the clone and sequence gaps being closed with the final gap closure in October 2002 for WormBase WS90.

As new cDNA data becomes available, we identify discrepancies between WormBase gene models and the experimental data. Sometimes, it will be impossible to make a model that fits the experimental data and a misalignment can be seen. WormBase users play an important role in identifying sequencing prediction errors. You may be working on a gene that we have annotated as a pseudogene or vice versa and may have evidence to suggest otherwise. We will investigate such cases and will verify and adjust/correct the sequence.

The number of predictions annotated onto the genomic sequence has also steadily increased from ~19,000 in the 1998 Science publication to 22,215 as of WS112, an increase of ~17%. This has occurred mainly because of isoforms being identified through publication of scientific data and the incorporation of new EST/mRNA data but there has also been a steady increase in the number of new CDSs. These "new" predictions are also identified from ESTs and mRNAs and by individuals conducting gene family studies. It is also worth remembering that predictions can be removed, merged and split.

## Literature search engine continues to grow

You can now query across 3300 full text papers and 16,500 abstracts from the corpus of C.elegans literature using Textpresso (http://www.textpresso.org), a search engine developed by WormBase to query full text *C. elegans* literature, an early edition of which went online in February 2003. The system allows for keyword searches and identifies and marks up terms of interest, which are organized in categories of an ontology (such as gene, cell or cell group, cellular component, association, regulation, and many more). Hits are displayed as sentences that match the query and their surrounding paragraphs. Additionally, links to the full text journal article as well as to a list of related articles are provided. Citation output can be obtained in Endnote format or as a PDF document.
Example: Search across the corpus of available *C. elegans* literature to check whether a transgene expressing UNC-73::GFP exists.

## New RNAi data

WormBase release WS110 contains 24,346 RNAi objects, 4201 of which have a mutant phenotype. WS112 will include a new data set from genome-wide systematic RNAi screens based on a RNAi-hypersensitive *rrf-3* strain (Simmer et al., PLoS Biology: Vol. 1, No. 1, p77). This set includes 2,819 RNAi experiments that result in a mutant phenotype, increasing the number of genes that have an RNAi phenotype by nearly 400. We are also curating the results of

RNAi experiments from individual papers, however, this process is slower and will take us some time to complete.

## Functional Annotations in WormBase

WormBase has begun writing functional annotations appear at the top of every gene summary page, under the 'Brief I. D.' heading, the purpose of which give users a brief summary of the current understanding of a gene product's role in *C. elegans* development and/or behavior. In writing functional annotations, curators use a broad range of data, from computer-generated similarities and protein domain identifications, to published genetic and molecular studies to provide a succinct, yet thorough, annotation. For each gene, we have attempted to report several key pieces of information: 1) The molecular identity of a gene product including, where possible, identifiable domains, orthology, and information regarding human disease associations, 2) The biological processes that require the activity of this gene product, 3) A summary of known genetic and/or molecular interactions, and 4) A brief summary of the reported expression pattern. To date, we have written annotations for ~1500 genes. Our goal is to finish the annotations by the beginning of 2004 and then to continually update the annotations as new information appears in the published literature. As always, we welcome your comments and suggestions so that we can keep the annotations as accurate and up-to-date as possible.

## Persons and Authors

We continue the effort to better organize and correct mistakes in the connections between people and their papers, WormBase distinguishes Persons from Authors: Persons are unique individuals with contact information. We will continue to ask your help in making accurate connections between you and your publications. An extremely useful practice would be to use full names whenever possible in papers and abstracts. Please continue to check or update your contact information at http://minerva.caltech.edu/~azurebrd/cgi-bin/forms/person.cgi or email cecilia@minerva.caltech.edu.

## The lineage tree of *C. elegans* researchers

WormBase would like to describe the professional associations between *C. elegans* researchers that started with Sydney Brenner and has expanded to over 450 registered *C. elegans* laboratories worldwide today. Please take some time to describe both those people that trained you and the people that you have subsequently trained/collaborated with by completing our 'person lineage data submission' form at http://minerva.caltech.edu/~azurebrd/cgi-bin/forms/person_lineage.cgi. Please note that in the future, in order to be technically correct, we shall refer to these associations as a 'directed acyclic graph (DAG)' and not a tree, because a given person can be associated with several mentors.

## Coming soon:

## The *C. briggsae* genome sequence

PLoS Biology, issue #2 (November 2003) will publish "The Genome Sequence of *Caenorhabditis briggsae*: A Platform for Comparative Genomics," whose coauthors include several people from the WormBase staff. In the coming months, WormBase will feature the *C. briggsae* genome, its annotation and information on how to access such data.

## People of WormBase

New curators at the Sanger Institute: Paul Davis has joined the Sanger team as a sequence curator.
New curators at Caltech: Kimberly Van Auken (former Post-doc with Bill Wood) has joined the Caltech team.
New curators at Washington University: Aniko Sabo has joined the WashU team as a sequence curator.
New programmers at Cold Spring Harbor Laboratory: Jack Chen has joined the web team.

## WormBase Advisory Board

Our advisory board helps set priorities and goal*s* (and generally acts like a thesis committee). They ensure that the user community is best served, and that WormBase cooperates with other model organism databases.
Thomas Blumenthal, *C. elegans* molecular biologist
Martin Chalfie, *C. elegans* developmental and neuro geneticist
Jonathan Hodgkin, *C. elegans* geneticist; curator of the CGC genetic map
Leon Avery, *C. elegans* neurobiologist; CGC website
William Gelbart, PI, FlyBase
Janan Eppig, PI, Mouse Genome Database (MGD)
Michael Cherry, PI, Saccharomyces Genome Database (SGD)
Stanley Letovsky, bioinformatician.

## Funding Update

The National Human Genome Research Institute at NIH has renewed support for WormBase for an additional five years.