WORMBASE – NEMATODE BIOLOGY AND GENOMES

Paul Davis, Wellcome Trust Genome Campus.
WormBase Consortium, (PI List) Lincoln Stein – OICR, Paul Sternberg – CALTECH,
John Spieth - GSC, Richard Durbin – WTSI

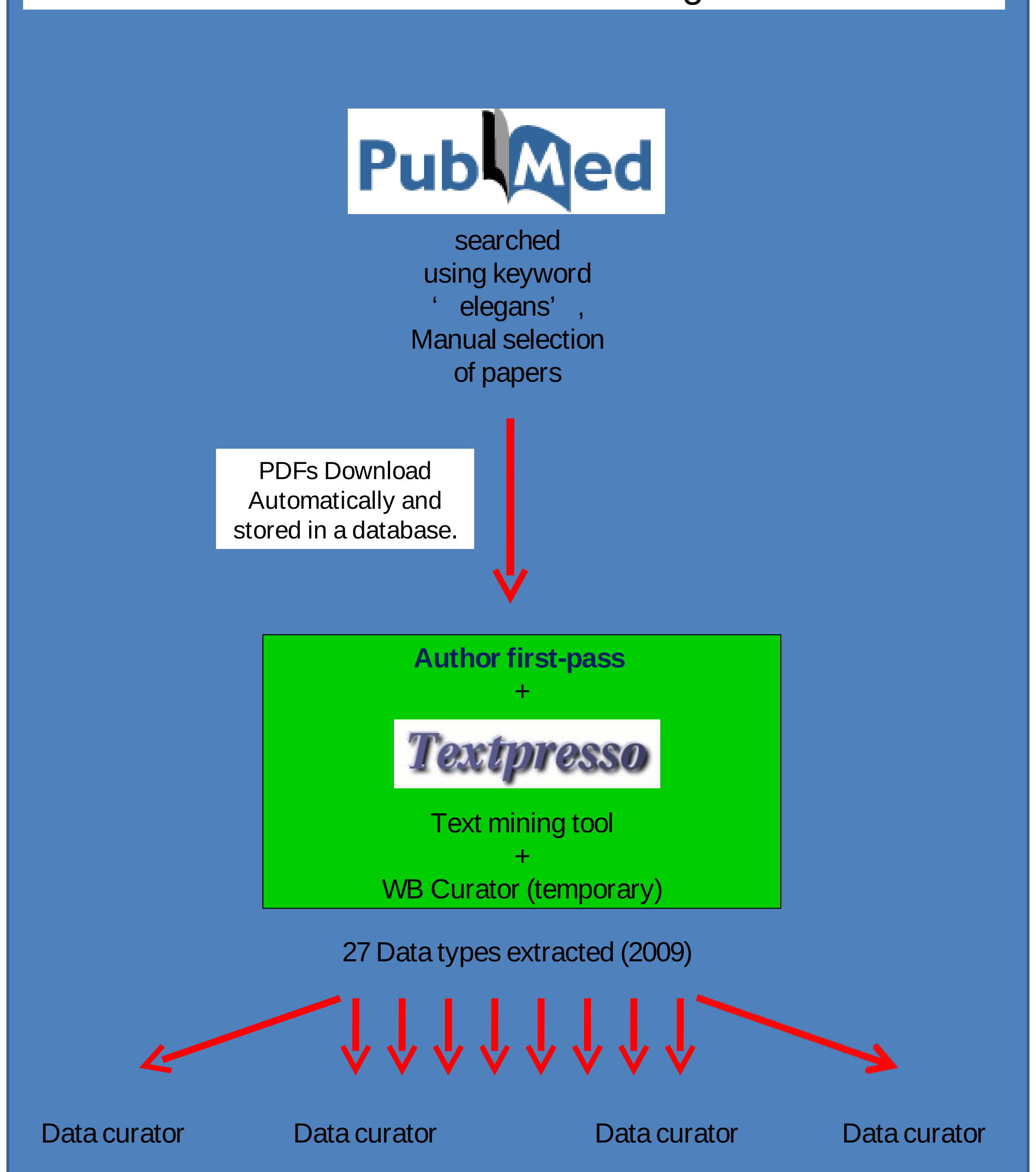
WormBase www.wormbase.org

WormBase is the major public online database resource for the Caenorhabditis research community. The database was developed primarily for the nematode C. elegans but expanded to host genomes and biological data from other closely related nematode species including C. briggsae, C. remanei, C. brenneri , C. japonica and Pristionchus pacificus . WormBase has developed tools to mine the data held within the database and compare the hosted species. Over the years we have developed a variety of curation pipelines which often begin in a "first-pass" literature curation step. This involves a brief overview of the literature before directing it to specialised data curators who extract all relevant information. Curators focus on particular data types or experimental techniques such as gene structure changes (see the Sequence curation poster), variations, phenotypes or RNAi and their expertise in these fields make curation efficient. WormBase works with many other groups and consortiums to validate, process and integrate both large and small scale data resources. WormBase also provides data that will be of interest to the wider biomedical and bioinformatics communities allowing researchers to utilise the information and techniques offered by nematodes to study wider aspects including medicine and disease.

Nature Precedings : doi:10.1038/npre.2009.3127.1 : Posted 22 Apr 2009

Automated First Pass Paper Curation

Paper and Data Type Identification:
There is a project underway to move from a manual approach to a semi automated pipeline with author input. Currently we are in a transitional phase moving from WB Curators towards Authors and Text Mining.



PubMed searched using keyword ‘ elegans’ , Manual selection of papers

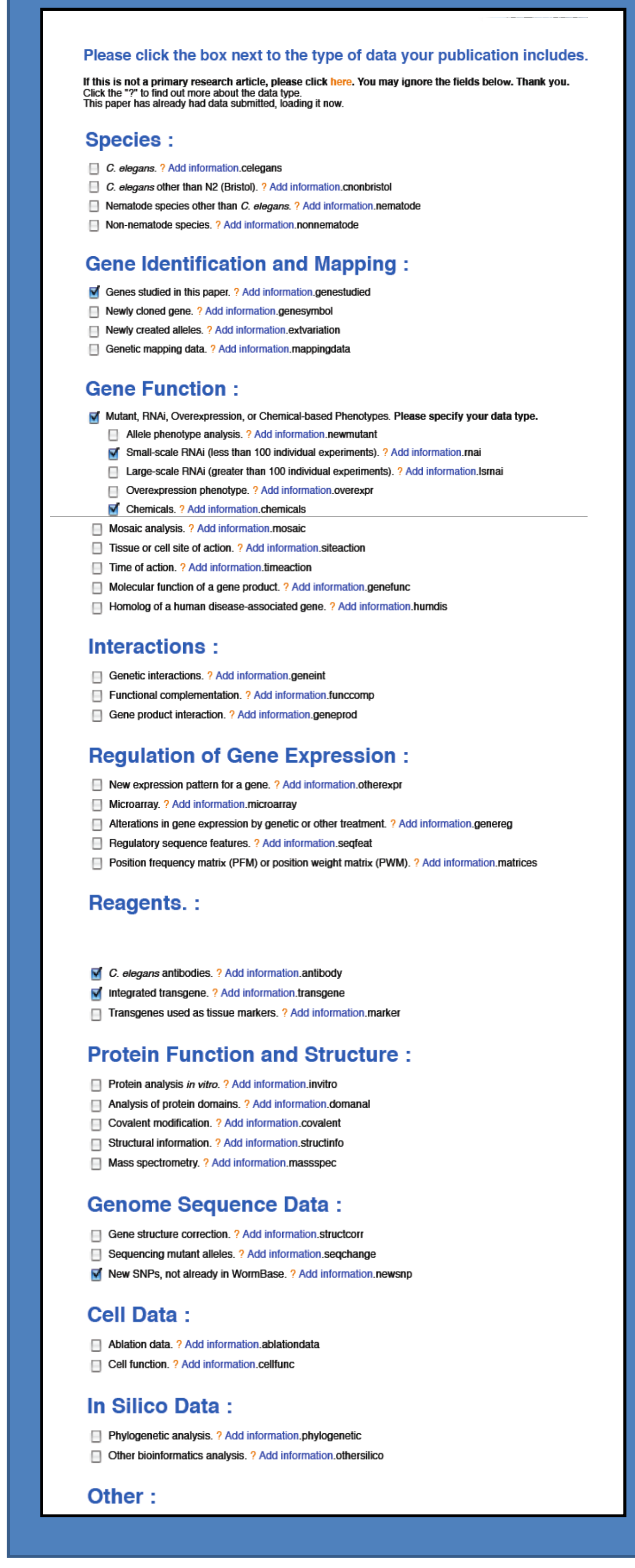
PDFs Download Automatically and stored in a database.

Author first-pass
+
Textpresso
Text mining tool
+
WB Curator (temporary)

27 Data types extracted (2009)

Data curator | Data curator | Data curator | Data curator

Author First-Pass Form



Please click the box next to the type of data your publication includes.

If this is not a primary research article, please click here. You may ignore the fields below. Thank you.
Click the "?" to find out more about the data type.
This paper has already had data submitted, loading it now.

Species :

- ☐ C. elegans. ? Add information.celegans
- ☐ C. elegans other than N2 (Bristol). ? Add information.cnonbristol
- ☐ Nematode species other than C. elegans. ? Add information.nematode
- ☐ Non-nematode species. ? Add information.nonnematode

Gene Identification and Mapping :

- ☑ Genes studied in this paper. ? Add information.genestudied
- ☐ Newly cloned gene. ? Add information.genesymbol
- ☐ Newly created alleles. ? Add information.extvariation
- ☐ Genetic mapping data. ? Add information.mappingdata

Gene Function :

- ☑ Mutant, RNAi, Overexpression, or Chemical-based Phenotypes. Please specify your data type.
  - ☐ Allele phenotype analysis. ? Add information.newmutant
  - ☑ Small-scale RNAi (less than 100 individual experiments). ? Add information.rnai
  - ☐ Large-scale RNAi (greater than 100 individual experiments). ? Add information.lsrnai
  - ☐ Overexpression phenotype. ? Add information.overexpr
  - ☑ Chemicals. ? Add information.chemicals
- ☐ Mosaic analysis. ? Add information.mosaic
- ☐ Tissue or cell site of action. ? Add information.siteaction
- ☐ Time of action. ? Add information.timeaction
- ☐ Molecular function of a gene product. ? Add information.genefunc
- ☐ Homolog of a human disease-associated gene. ? Add information.humdis

Interactions :

- ☐ Genetic interactions. ? Add information.geneint
- ☐ Functional complementation. ? Add information.funccomp
- ☐ Gene product interaction. ? Add information.geneprod

Regulation of Gene Expression :

- ☐ New expression pattern for a gene. ? Add information.otherexpr
- ☐ Microarray. ? Add information.microarray
- ☐ Alterations in gene expression by genetic or other treatment. ? Add information.genereg
- ☐ Regulatory sequence features. ? Add information.seqfeat
- ☐ Position frequency matrix (PFM) or position weight matrix (PWM). ? Add information.matrices

Reagents. :

- ☑ C. elegans antibodies. ? Add information.antibody
- ☑ Integrated transgene. ? Add information.transgene
- ☐ Transgenes used as tissue markers. ? Add information.marker

Protein Function and Structure :

- ☐ Protein analysis in vitro. ? Add information.invitro
- ☐ Analysis of protein domains. ? Add information.domanal
- ☐ Covalent modification. ? Add information.covalent
- ☐ Structural information. ? Add information.structinfo
- ☐ Mass spectrometry. ? Add information.massspec

Genome Sequence Data :

- ☐ Gene structure correction. ? Add information.structcorr
- ☐ Sequencing mutant alleles. ? Add information.seqchange
- ☑ New SNPs, not already in WormBase. ? Add information.newsnp

Cell Data :

- ☐ Ablation data. ? Add information.ablationdata
- ☐ Cell function. ? Add information.cellfunc

In Silico Data :

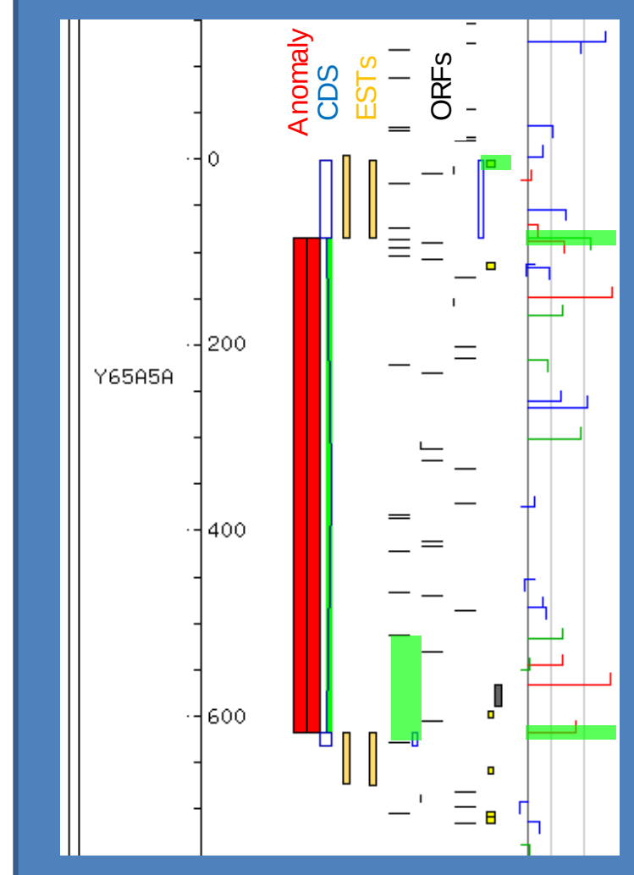
- ☐ Phylogenetic analysis. ? Add information.phylogenetic
- ☐ Other bioinformatics analysis. ? Add information.othersilico

Other :

Authors will be automatically contacted once their paper is downloaded, at this point the paper will not be visible to the curators. Once a set period of time has elapsed the text mining will be conducted and work distributed within the consortium.

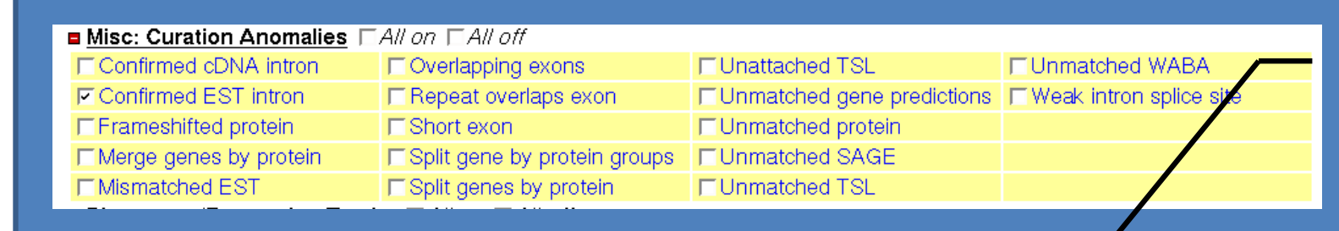
Curation Anomaly display

(see sequence curation poster for more details)



Curators use the Fmap within ACeDB to add/modify genes that go into the WormBase dataset. Curators use pre-computed anomalies to identify gene models that require attention as well as missing genes. Here we have an Fmap display with a Confirmed intron not in gene model anomaly. The curator created this two exon gene model based on the 2 EST sequences which contain a single intron.

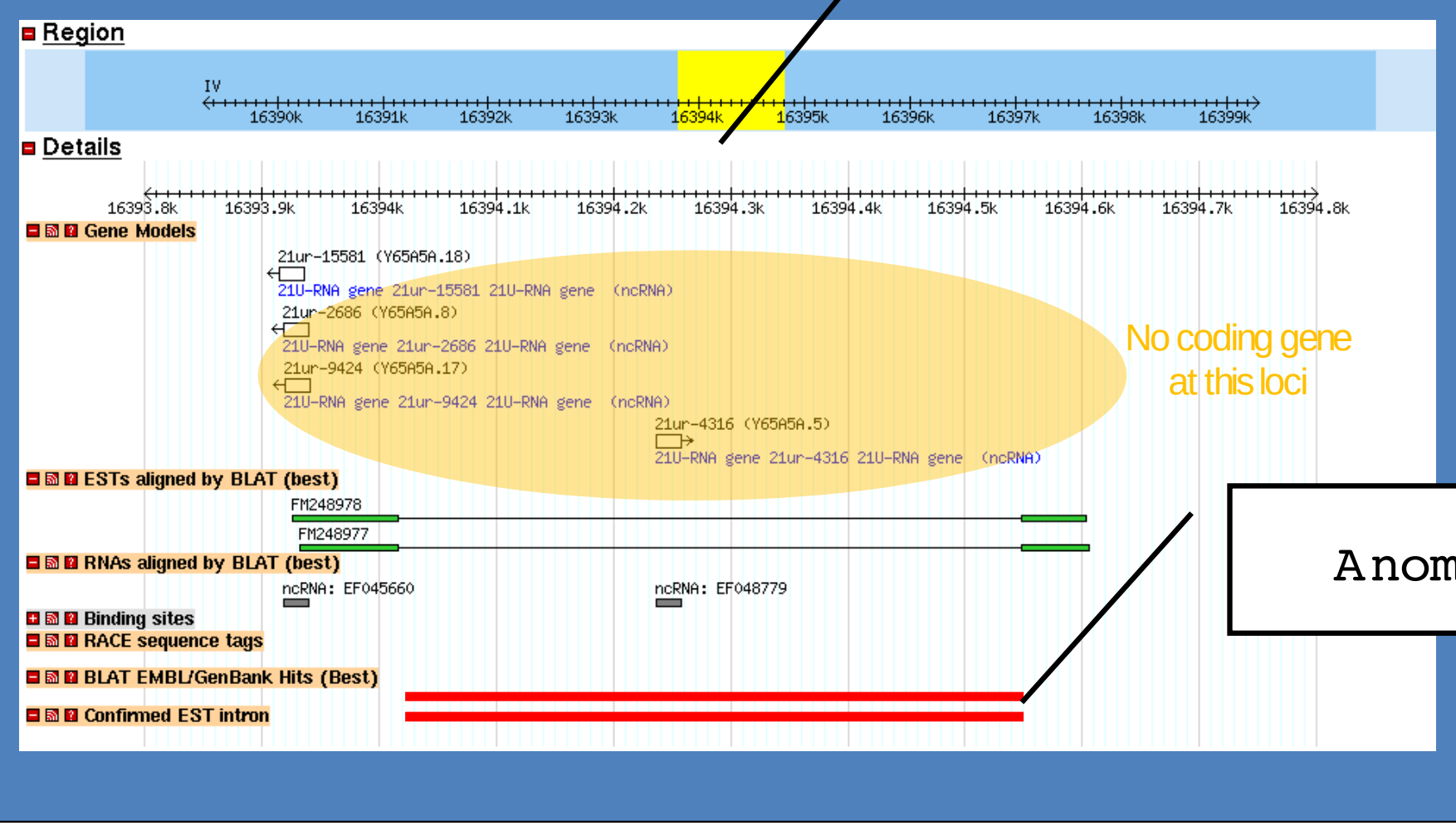
Users can choose to see these anomalies/possible errors in the genome browser by selecting from these tracks.



Misc: Curation Anomalies ☐ All on ☐ All off

☐ Confirmed cDNA intron	☐ Overlapping exons	☐ Unattached TSL	☐ Unmatched WABA
☑ Confirmed EST intron	☐ Repeat overlaps exon	☐ Unmatched gene predictions	☐ Weak intron splice site
☐ Frameshifted protein	☐ Short exon	☐ Unmatched protein	
☐ Merge genes by protein	☐ Split gene by protein groups	☐ Unmatched SAGE	
☐ Mismatched EST	☐ Split genes by protein	☐ Unmatched TSL	

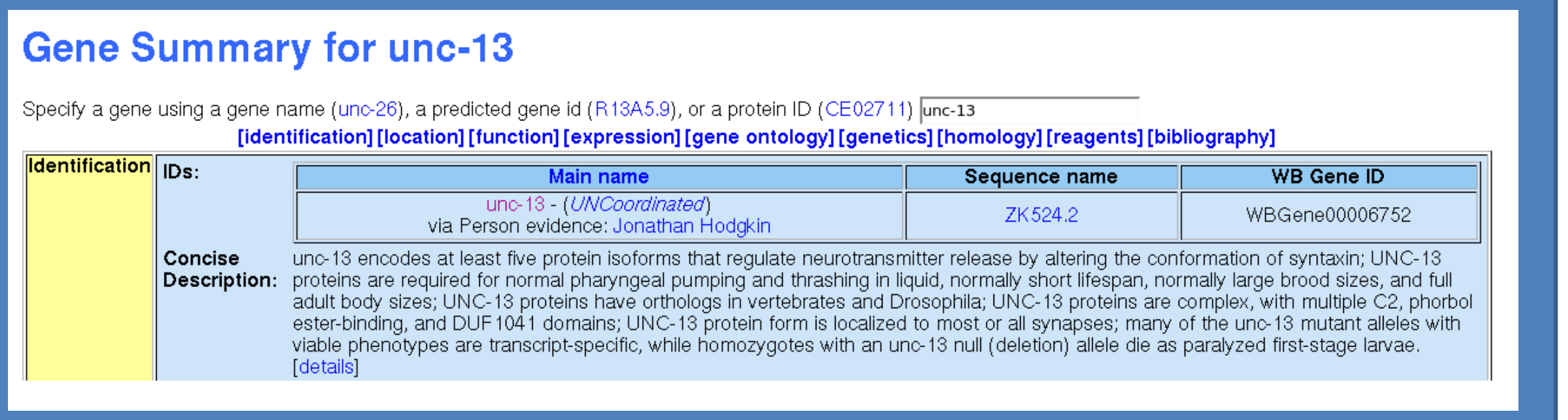
Genome browser prior to new gene annotation.



Anomalies

Gene Summary Page: Concise descriptions

Curators are working to produce a manually generated concise description for all *C. elegans* genes (extending to include tier II nematodes). The aim of this is to produce a abstract like summary of the gene and it’ s function so that WormBase users get a good understanding of the gene, with a minimum amount of effort.



Gene Summary for unc-13

Specify a gene using a gene name (unc-26), a predicted gene id (R13A5.9), or a protein ID (CE02711) unc-13

[identification] [location] [function] [expression] [gene ontology] [genetics] [homology] [reagents] [bibliography]

Identification

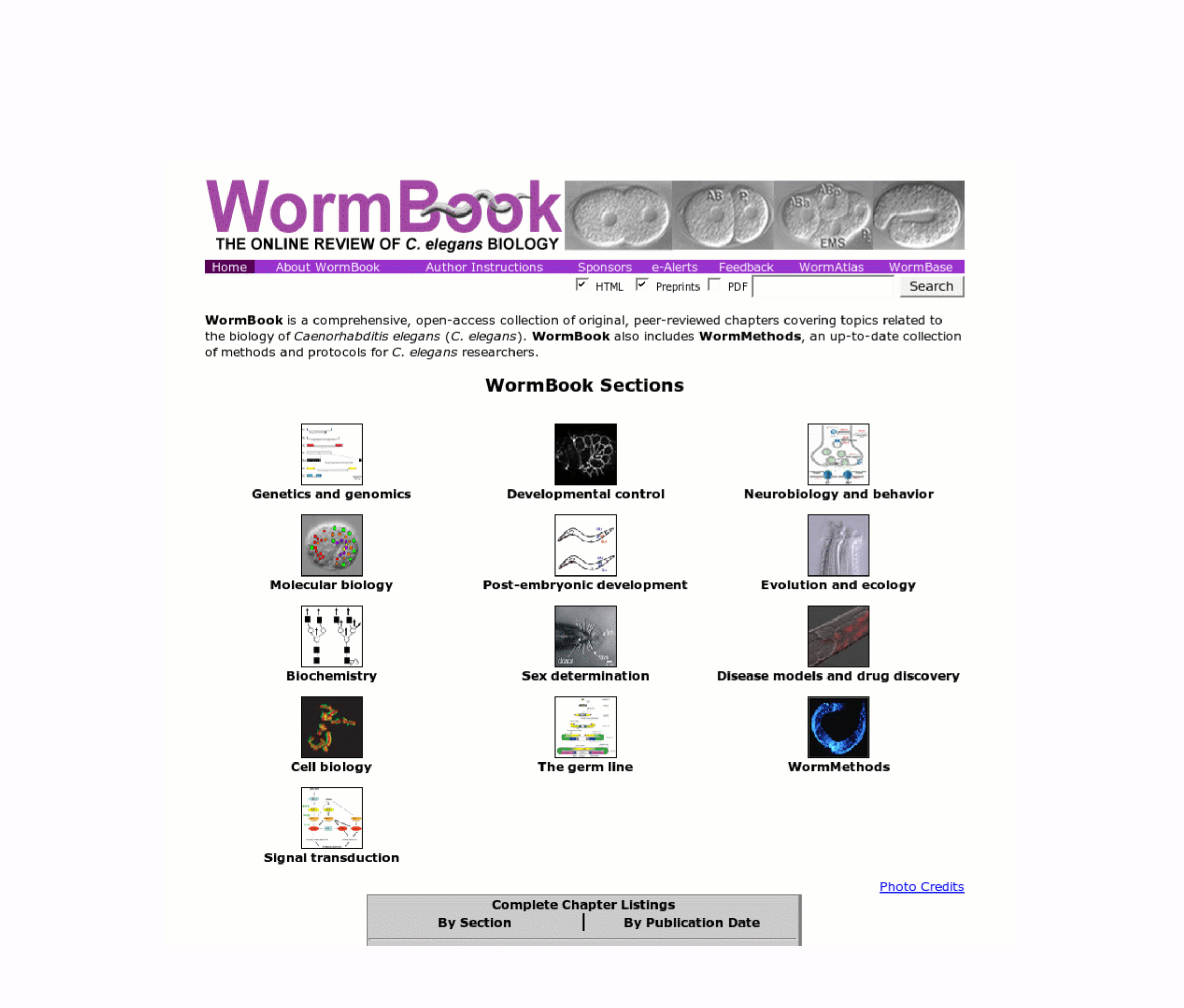
IDs:

Main name	Sequence name	WB Gene ID
unc-13 - (*UNCoordinated*) via Person evidence: Jonathan Hodgkin	ZK524.2	WBGene00006752

Concise Description: unc-13 encodes at least five protein isoforms that regulate neurotransmitter release by altering the conformation of syntaxin; UNC-13 proteins are required for normal pharyngeal pumping and thrashing in liquid, normally short lifespan, normally large brood sizes, and full adult body sizes; UNC-13 proteins have orthologs in vertebrates and Drosophila; UNC-13 proteins are complex, with multiple C2, phorbol ester-binding, and DUF1041 domains; UNC-13 protein form is localized to most or all synapses; many of the unc-13 mutant alleles with viable phenotypes are transcript-specific, while homozygotes with an unc-13 null (deletion) allele die as paralyzed first-stage larvae. [details]

WormBook

WormBook is the online text companion to WormBase, the *C. elegans* model organism database. WormBook contains original reviews on all aspects of *C. elegans* biology and up-to-date descriptions of technical procedures used to study this animal.



WormBook
THE ONLINE REVIEW OF *C. elegans* BIOLOGY

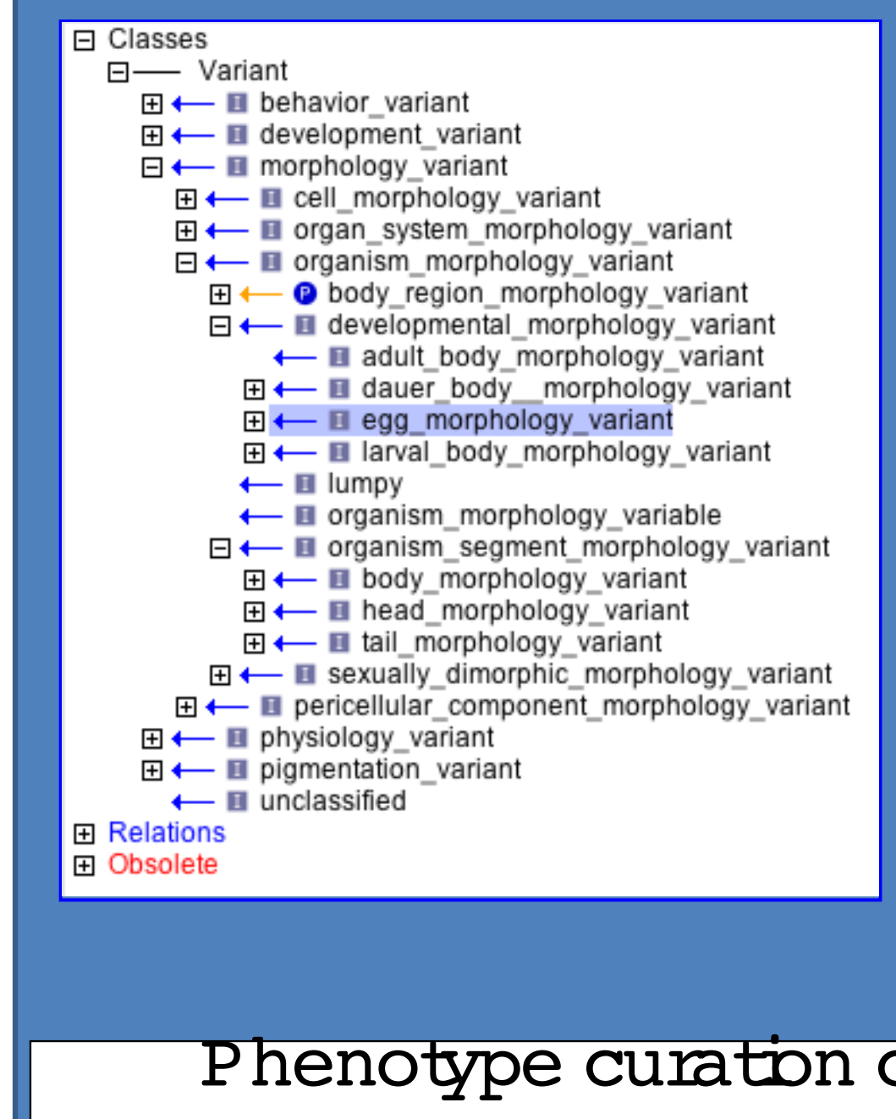
WormBook is a comprehensive, open-access collection of original, peer-reviewed chapters covering topics related to the biology of *Caenorhabditis elegans* (*C. elegans*). **WormBook** also includes **WormMethods**, an up-to-date collection of methods and protocols for *C. elegans* researchers.

WormBook Sections

Phenotype ontology

Our Phenotype Ontology has been modified to curate nematode species other than *C. elegans* “N2” strain

Phenotype Ontology
A hierarchy-based ontology
1823 terms, 66% defined,
55% associated with a variation



- Classes
  - Variant
    - behavior_variant
    - development_variant
    - morphology_variant
      - cell_morphology_variant
      - organ_system_morphology_variant
      - organism_morphology_variant
        - body_region_morphology_variant
        - developmental_morphology_variant
          - adult_body_morphology_variant
          - dauer_body__morphology_variant
          - egg_morphology_variant
          - larval_body_morphology_variant
        - lumpy
        - organism_morphology_variable
        - organism_segment_morphology_variant
          - body_morphology_variant
          - head_morphology_variant
          - tail_morphology_variant
        - sexually_dimorphic_morphology_variant
      - pericellular_component_morphology_variant
    - physiology_variant
    - pigmentation_variant
    - unclassified
- Relations
- Obsolete

Modifications required

Changes to Term Names
from “ _abnormal” to “ _variant” .

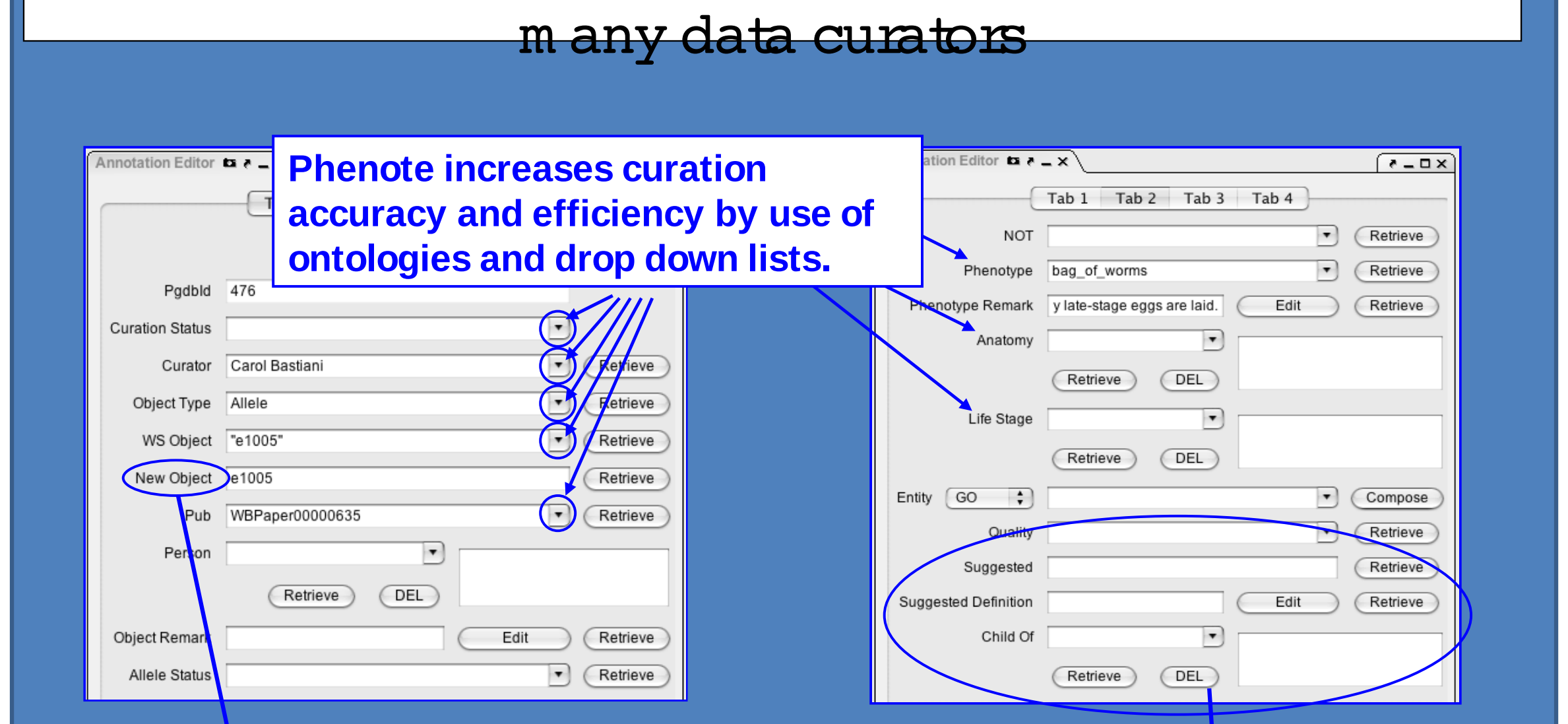
Changes to Definitions
Use of “ control animals” rather than “ wild-type” or “ N2” (*C. elegans* strain).

C. elegans -specific terminology, e.g., “ hermaphrodite” , were removed from definitions when possible.

Example:
WBPhenotype:0000037: **egg_morphology_abnormal**
Def: "Any deviation in the overall structure or appearance of fertilized oocytes that are deposited by adult hermaphrodites.

Changed to:
WBPhenotype:0000037: **egg_morphology_variant**
Def: "Any variation in the overall structure or appearance of fertilized oocytes that are laid compared to those laid by control animals.
"synonym: "egg_morphology_abnormal
(The “ _abnormal” version of the term is kept as a synonym so people used to these terms will still be able to find them.)

Phenotype curation captures multiple attributes reported by authors and requires the efforts of many data curators



Phenote increases curation accuracy and efficiency by use of ontologies and drop down lists.

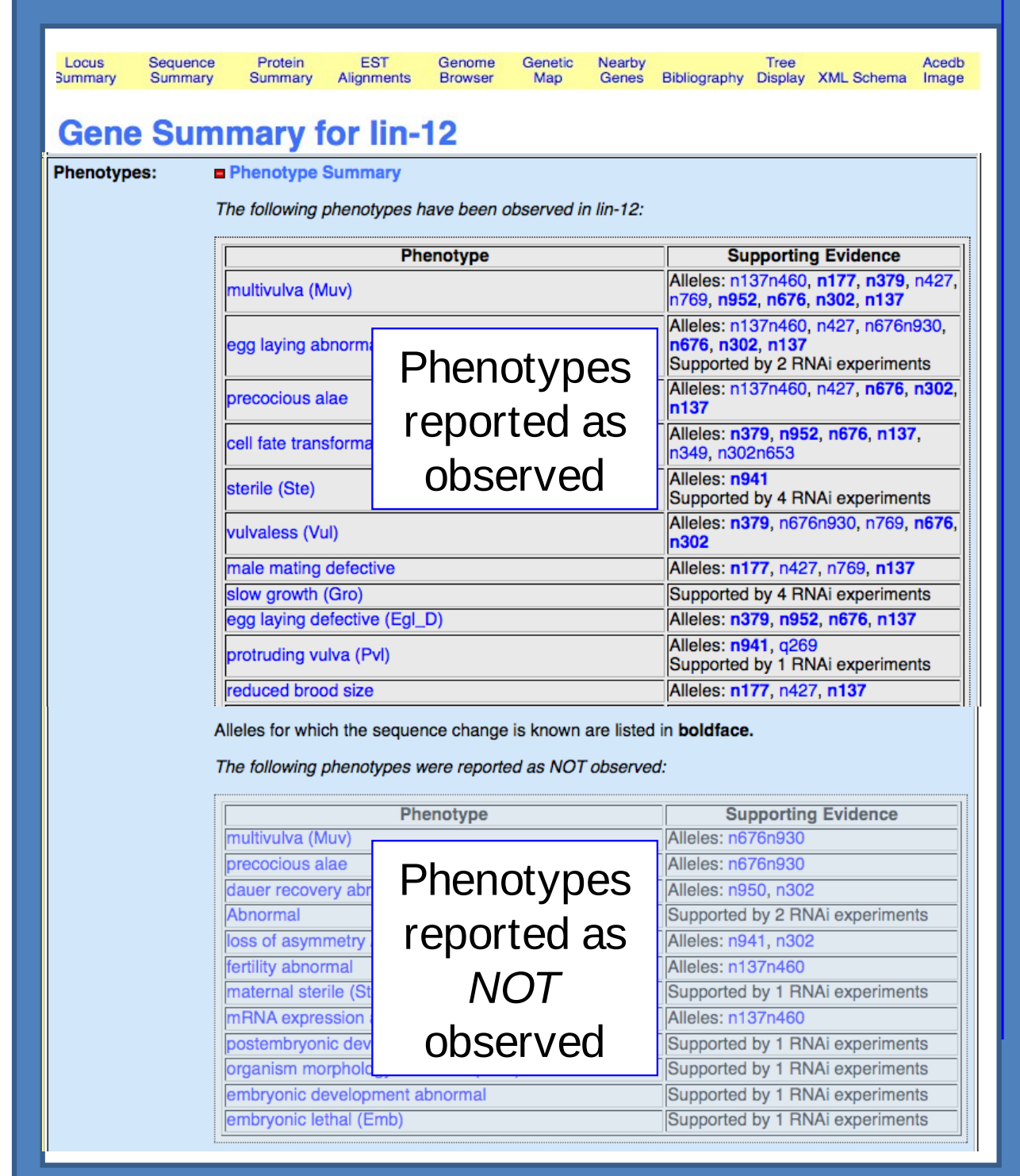
COORDINATED WITH OBJECT CURATORS:
If object (allele or transgene) does not exist in the latest release of the database, an e-mail is automatically sent to the curator responsible for creating those objects.

COORDINATED WITH ONTOLOGY CURATOR:
Phenotype curators can request a term, send a suggested definition and hierarchy placement through the Phenote interface. New terms are automatically assigned to the record when they are approved.

OTHER ATTRIBUTES CAPTURED INCLUDE:
Genotype, Treatment, Nature of allele (recessive, semi-dominant, dominant), Penetrance (incomplete, low, high, complete), Maternal effect (strictly maternal, with maternal effect), Paternal effect, Temperature sensitivity, Haploinsufficiency,
Allele type (amorph, hypomorph, etc.).

Phenotypes are linked to genes through allele or RNAi curation

3626 / 23709* genes with alleles were annotated with phenotype data (includes *NOT* annotations)
**as of WS200*



Gene Summary for lin-12

Phenotypes: Phenotype Summary

The following phenotypes have been observed in lin-12:

Phenotype	Supporting Evidence
multivulva (Muv)	Alleles: n137n460, **n177**, **n379**, n427, n769, **n952**, **n676**, **n302**, **n137**
egg laying abnorm[illegible]	Alleles: n137n460, n427, n676n930, **n676**, **n302**, **n137** Supported by 2 RNAi experiments
precocious alae	Alleles: n137n460, n427, **n676**, **n302**, **n137**
cell fate transforma[illegible]	Alleles: **n379**, **n952**, **n676**, **n137**, n349, n302n653
sterile (Ste)	Alleles: **n941** Supported by 4 RNAi experiments
vulvaless (Vul)	Alleles: **n379**, n676n930, n769, **n676**, **n302**
male mating defective	Alleles: **n177**, n427, n769, **n137**
slow growth (Gro)	Supported by 4 RNAi experiments
egg laying defective (Egl_D)	Alleles: **n379**, **n952**, **n676**, **n137**
protruding vulva (Pvl)	Alleles: **n941**, q269 Supported by 1 RNAi experiments
reduced brood size	Alleles: **n177**, n427, **n137**

Phenotypes reported as observed

Alleles for which the sequence change is known are listed in **boldface**.

The following phenotypes were reported as NOT observed:

Phenotype	Supporting Evidence
multivulva (Muv)	Alleles: n676n930
precocious alae	Alleles: n676n930
dauer recovery abr[illegible]	Alleles: n950, n302
Abnormal	Supported by 2 RNAi experiments
loss of asymmetry [illegible]	Alleles: n941, n302
fertility abnormal	Alleles: n137n460
maternal sterile (St[illegible]	Supported by 1 RNAi experiments
mRNA expression [illegible]	Alleles: n137n460
postembryonic dev[illegible]	Supported by 1 RNAi experiments
organism morpholo[illegible]	Supported by 1 RNAi experiments
embryonic development abnormal	Supported by 1 RNAi experiments
embryonic lethal (Emb)	Supported by 1 RNAi experiments

Phenotypes reported as *NOT* observed

	May 2008 WS188	March 2009 WS200
Allele-phenotype connections	**9771**	**15951**
Alleles Curated (total # alleles)	**28%** (15326)	**34%** (17448)
Papers curated (total papers flagged)	NA	**23%** (+125 unflagged papers)