Name_______________ Section_______ Biocore 304 Spring 2008

Bioinformatics Analysis of Deletion Allele

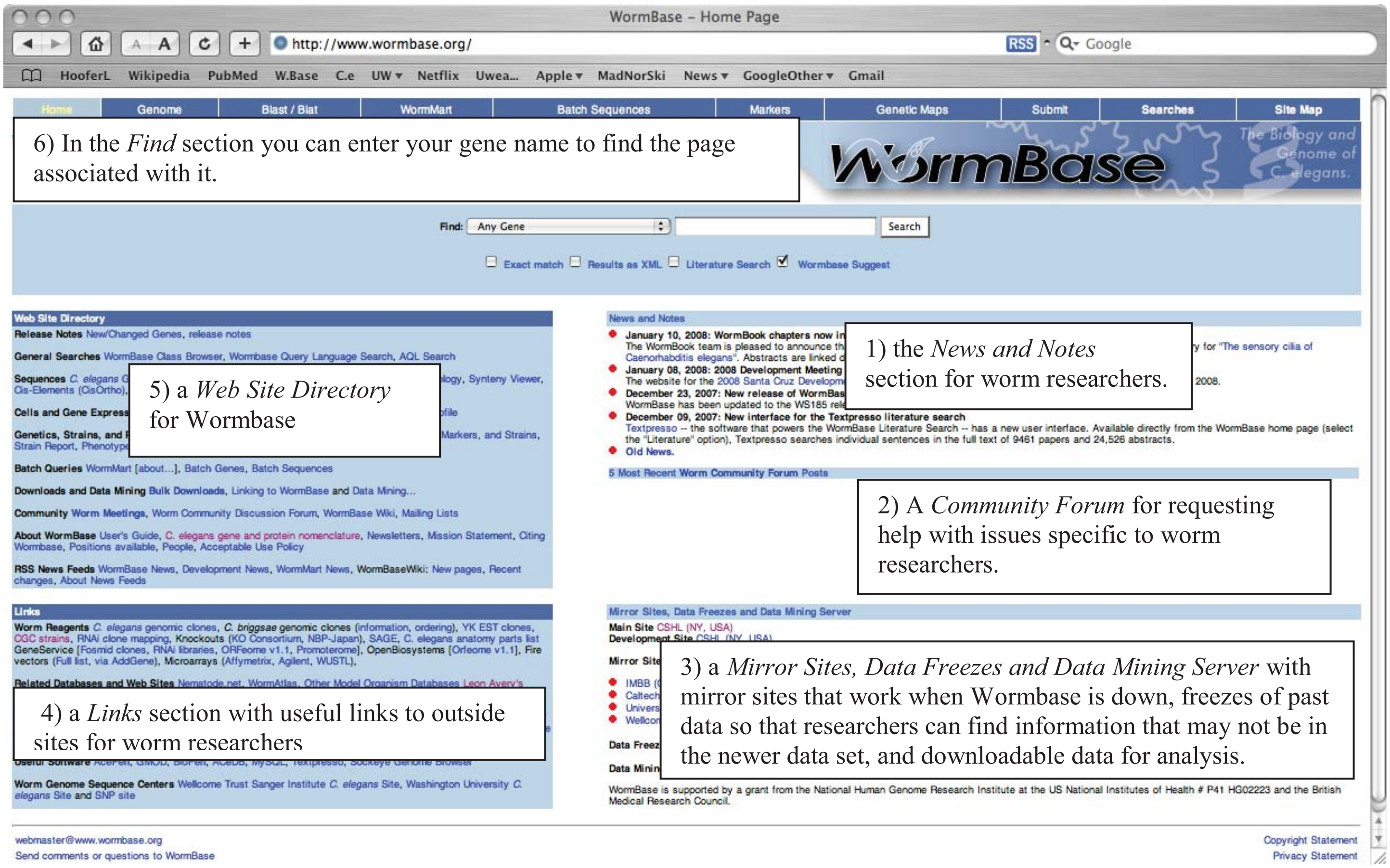
Learning Objective: You will gather information about your gene using common bioinformatics tools such as Wormbase and Geneious to examine:
1. Gene structure: exons, introns, and UTRs
2. Protein sequence
3. Protein structure for the protein encoded by your gene
4. Information on your gene in human diseases
5. Original research publications on your gene and its orthologs.

The following directions are a guided tutorial to help you gather information about your gene. You may or may not find all of the following bioinformatics tools useful in the preparation of your mini-poster. It is up to you to decide what information will best help you address the central question for your poster (*put central question here….)* The numbered questions embedded in this tutorial are meant to help you focus and organize your information, but your answers will not be graded. This is a check assignment, however, so when you finish the tutorial, be sure to provide your TA with evidence that you have obtained information relevant to your assigned gene.

I. Introduction to Wormbase (http://www.wormbase.org)

This database of the model organism *C. elegans* and related nematodes has a page for each of the 19,000 genes in *C. elegans*, with links to the genome sequence, similar genes in *C. elegans* in other organisms, all the publications mentioning the gene, as well as links to other databases.

The main page of Wormbase is filled with links to other sites and tools for worm researchers. Along the top are quick links to Wormbase tools: Home Genome, Blast/Blat, etc. The panel sections that take up most of the page include (1-6).


WormBase – Home Page
http://www.wormbase.org/
RSS
Google
HooferL Wikipedia PubMed W.Base C.e UW ▾ Netflix Uwea... Apple ▾ MadNorSki News ▾ GoogleOther ▾ Gmail
Home
Genome
Blast / Blat
WormMart
Batch Sequences
Markers
Genetic Maps
Submit
Searches
Site Map
WormBase
The Biology and Genome of C. elegans.
Find: Any Gene
Search
□ Exact match □ Results as XML □ Literature Search ☑ Wormbase Suggest
Web Site Directory
Links
News and Notes
5 Most Recent Worm Community Forum Posts
Mirror Sites, Data Freezes and Data Mining Server
webmaster@www.wormbase.org
Send comments or questions to WormBase
Copyright Statement
Privacy Statement

6) In the *Find* section you can enter your gene name to find the page associated with it.

Release Notes New/Changed Genes, release notes

General Searches WormBase Class Browser, Wormbase Query Language Search, AQL Search

5) a *Web Site Directory* for Wormbase

Batch Queries WormMart [about...], Batch Genes, Batch Sequences

Downloads and Data Mining Bulk Downloads, Linking to WormBase and Data Mining...

Community Worm Meetings, Worm Community Discussion Forum, WormBase Wiki, Mailing Lists

About WormBase User's Guide, C. elegans gene and protein nomenclature, Newsletters, Mission Statement, Citing Wormbase, Positions available, People, Acceptable Use Policy

RSS News Feeds WormBase News, Development News, WormMart News, WormBaseWiki: New pages, Recent changes, About News Feeds

Worm Reagents C. elegans genomic clones, C. briggsae genomic clones (information, ordering), YK EST clones, CGC strains, RNAi clone mapping, Knockouts (KO Consortium, NBP-Japan), SAGE, C. elegans anatomy parts list GeneService [Fosmid clones, RNAi libraries, ORFeome v1.1, Promoterome], OpenBiosystems [Orfeome v1.1], Fire vectors (Full list, via AddGene), Microarrays (Affymetrix, Agilent, WUSTL),

4) a *Links* section with useful links to outside sites for worm researchers

Worm Genome Sequence Centers Wellcome Trust Sanger Institute C. elegans Site, Washington University C. elegans Site and SNP site

1) the *News and Notes* section for worm researchers.

December 09, 2007: New interface for the Textpresso literature search
Textpresso -- the software that powers the WormBase Literature Search -- has a new user interface. Available directly from the WormBase home page (select the "Literature" option), Textpresso searches individual sentences in the full text of 9461 papers and 24,526 abstracts.

Old News.

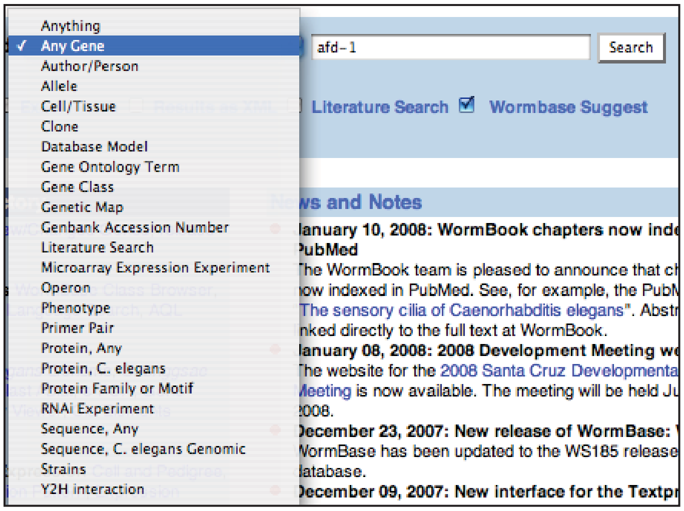
2) A *Community Forum* for requesting help with issues specific to worm researchers.

Main Site CSHL (NY, USA)

3) a *Mirror Sites, Data Freezes and Data Mining Server* with mirror sites that work when Wormbase is down, freezes of past data so that researchers can find information that may not be in the newer data set, and downloadable data for analysis.

WormBase is supported by a grant from the National Human Genome Research Institute at the US National Institutes of Health # P41 HG02223 and the British Medical Research Council.

In the *Find* section, you can search the default "Any gene" selection, but notice the other choices in the dropdown menu in the *Find* window; you may find them useful later.

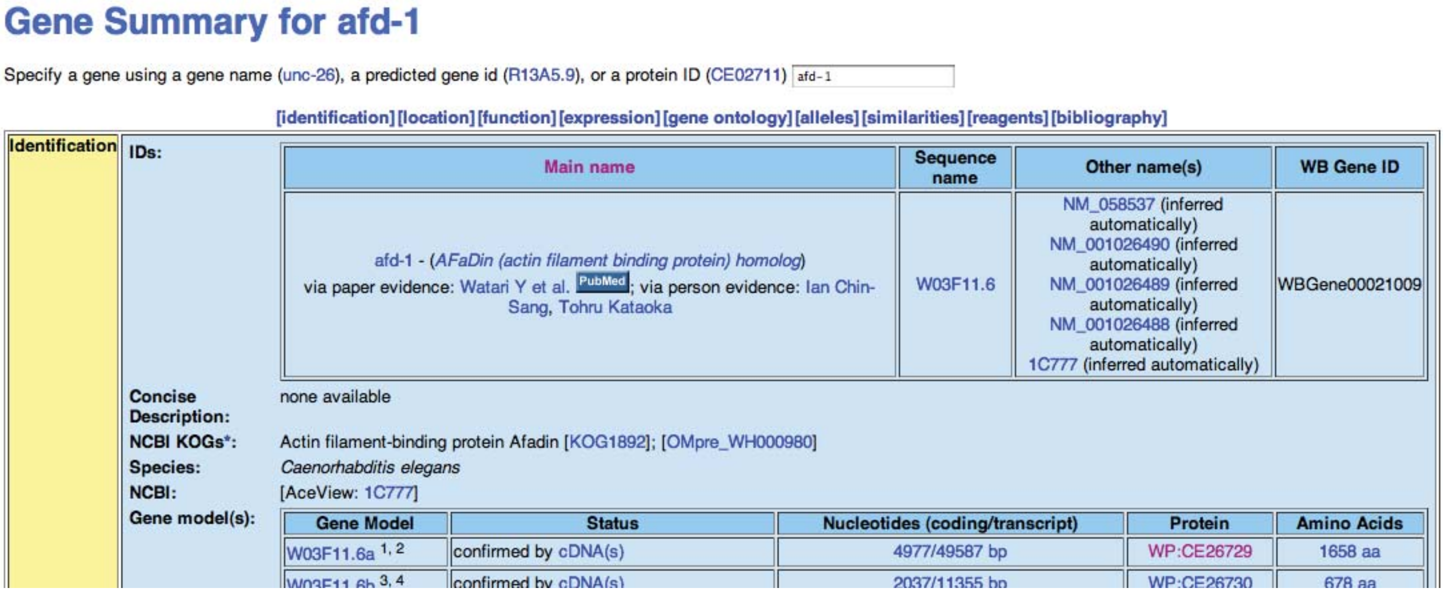


anuary 10, 2008: WormBook chapters now inde PubMed
he WormBook team is pleased to announce that ch ow indexed in PubMed. See, for example, the PubM "The sensory cilia of Caenorhabditis elegans". Abstr nked directly to the full text at WormBook.
anuary 08, 2008: 2008 Development Meeting we
he website for the 2008 Santa Cruz Developmenta Meeting is now available. The meeting will be held Ju 2008.
ecember 23, 2007: New release of WormBase: W
VormBase has been updated to the WS185 release latabase.
ecember 09, 2007: New interface for the Textpr

A note about nomenclature: The convention is to indicate genes with lower case, italic letters with a dash and number following (e.g., *unc-13*). Proteins are indicated in non-italic capitals, e.g., the protein encoded by *unc-13* can be called UNC-13. Alleles all have special designations referring back to the lab of origin. Thus, *brc-1(tm1145)* would be used to describe a deletion allele. Here, tm is the designation for the Mitani lab, which discovered it and 1145 is the 1145th mutant allele that the Mitani lab has generated. A website (http://www.wormbase.org/wiki/index.php/UserGuide:Nomenclature) describing genetic nomenclature for *C. elegans* states the following regarding worm strains: "A strain is a set of individuals of a particular genotype with the capacity to produce more individuals of the same genotype. Strains are given nonitalicized names consisting of two or three uppercase letters followed by a number. The strain letter prefixes refer to the laboratory of origin and are distinct from the mutation letter prefixes." For example, we are using strain DW102 to study the *brc-1* gene mutation.

-> Enter your team's assigned gene name in the *Search* Window and select "Search". Take a few moments to explore the page for your assigned gene. (In this tutorial we have searched for the gene *afd-1*.)

Note that each of the 19,000 gene summary pages is set up like a table. The left side of each gene page has a yellow column. This column identifies the topics for each blue section of information. The first row is "Identification," the second is "Location," etc. We will explore each of these sections in turn. The next column for "Identification" is IDs, Concise Description, NCBI KOGs, etc. Further to the right is specific information pertaining to your gene of interest.



Gene Summary for afd-1

Specify a gene using a gene name (unc-26), a predicted gene id (R13A5.9), or a protein ID (CE02711) afd-1

[identification] [location] [function] [expression] [gene ontology] [alleles] [similarities] [reagents] [bibliography]

Identification

IDs:

Main name	Sequence name	Other name(s)	WB Gene ID
afd-1 - (*AFaDin (actin filament binding protein) homolog*) via paper evidence: Watari Y et al. PubMed; via person evidence: Ian Chin-Sang, Tohru Kataoka	W03F11.6	NM_058537 (inferred automatically) NM_001026490 (inferred automatically) NM_001026489 (inferred automatically) NM_001026488 (inferred automatically) 1C777 (inferred automatically)	WBGene00021009

Concise Description: none available

NCBI KOGs*: Actin filament-binding protein Afadin [KOG1892]; [OMpre_WH000980]

Species: *Caenorhabditis elegans*

NCBI: [AceView: 1C777]

Gene model(s):

Gene Model	Status	Nucleotides (coding/transcript)	Protein	Amino Acids
W03F11.6a 1, 2	confirmed by cDNA(s)	4977/49587 bp	WP:CE26729	1658 aa
W03F11.6b 3, 4	confirmed by cDNA(s)	2037/11355 bp	WP:CE26730	678 aa

The Main name is a name given to a gene by the *Caenorhabditis* Genetics Center, which monitors gene names to keep them consistent, standardized and to prevent disputes between researchers.

1) What is the main name for your gene?

2) What are other names that have been used for your gene (if any)?

The Concise Description summarizes basic information about your gene and what is known about it in humans.

3) Briefly summarize what this section tells you about your gene:

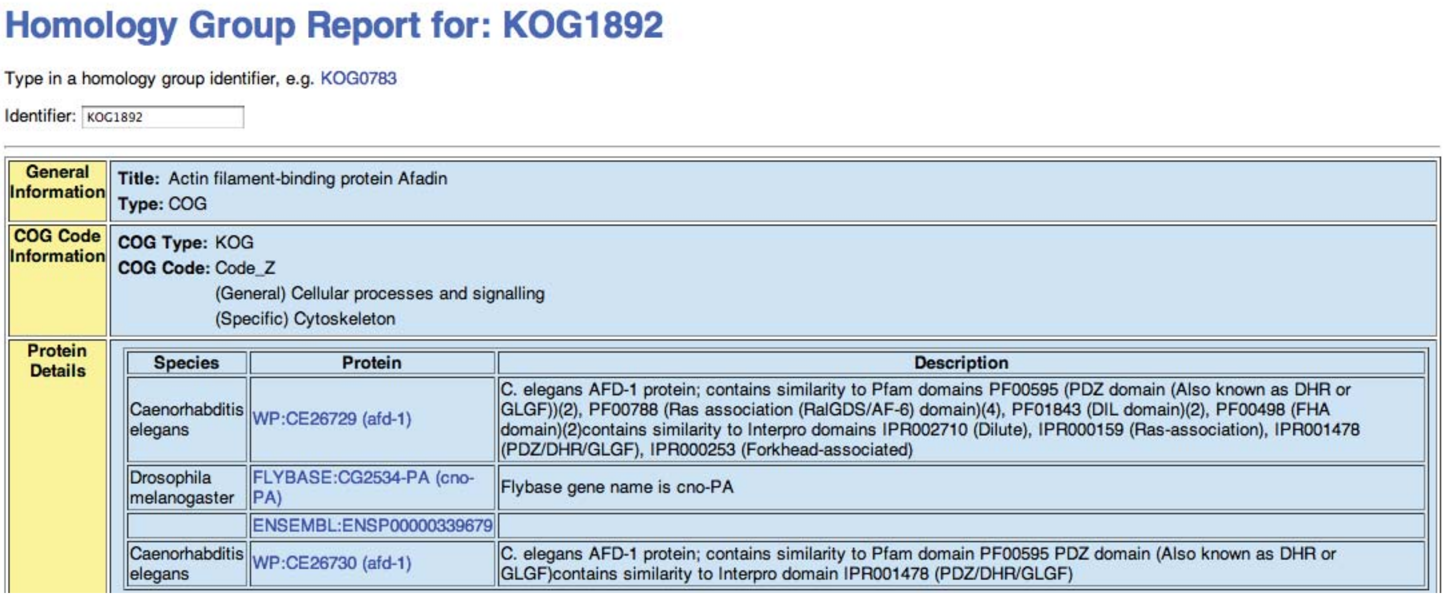
OMIM Database
In the Concise Description of your gene in the first yellow "*Identification*" row, note whether there are any links to orthologous human genes in this description, with OMIM (Online Mendelian Inheritance in Man) links. Click on one of these OMIM links. If there is an OMIM accession number without any direct link to the NCBI (National Center for Biotechnology Information) database, copy the OMIM number and go to the NCBI homepage (http://www.ncbi.nlm.nih.gov). Select the "OMIM" link in the dark blue bar near the top of the page and then copy your OMIM number into the search box at the top and search. You should see a page summarizing human disease(s) associated with the orthologous human gene, including links to PubMed articles. You should check out some of these articles as well as those listed in the Wormbase database; the information might be useful for your mini-poster.

4) Briefly describe a human disease associated with a gene that is orthologous to your worm gene.

Return to the Wormbase website. NCBI KOGs or Eurkaryotic conserved orthologous groups are predicted classifications of a gene product (protein) based on known protein functional groups in various organisms. They attempt to define the function of the protein expressed by your gene in simple terms by grouping it with similar proteins based on sequence homology.

5) What is the NCBI KOG for your gene?

-> Click on your KOG to access the Homology Group Report for your protein.



Homology Group Report for: KOG1892

Type in a homology group identifier, e.g. KOG0783

Identifier: KOG1892

General Information	**Title:** Actin filament-binding protein Afadin **Type:** COG
COG Code Information	**COG Type:** KOG **COG Code:** Code_Z (General) Cellular processes and signalling (Specific) Cytoskeleton

Protein Details

Species	Protein	Description
Caenorhabditis elegans	WP:CE26729 (afd-1)	C. elegans AFD-1 protein; contains similarity to Pfam domains PF00595 (PDZ domain (Also known as DHR or GLGF))(2), PF00788 (Ras association (RalGDS/AF-6) domain)(4), PF01843 (DIL domain)(2), PF00498 (FHA domain)(2)contains similarity to Interpro domains IPR002710 (Dilute), IPR000159 (Ras-association), IPR001478 (PDZ/DHR/GLGF), IPR000253 (Forkhead-associated)
Drosophila melanogaster	FLYBASE:CG2534-PA (cno-PA)	Flybase gene name is cno-PA
	ENSEMBL:ENSP00000339679	
Caenorhabditis elegans	WP:CE26730 (afd-1)	C. elegans AFD-1 protein; contains similarity to Pfam domain PF00595 PDZ domain (Also known as DHR or GLGF)contains similarity to Interpro domain IPR001478 (PDZ/DHR/GLGF)

6) In the *General Information* row, what is the title for your protein? Look in the COG Code Information row. **What does it tell you about the general and specific function of your protein?**

7) Look in the *Protein Details* row. **How many other species have proteins homologous to your protein? ______**
8) Look at the *Description* column. **What other proteins are also found in the same KOG group? What species are those proteins found in?**

9) Can you tell anything about the possible functions of the other proteins in that KOG group?

Go back to the main WormBase page for your gene.

Refer back to the *Identification* row. The NCBI Gene Model section has a letter/number designation for your gene that was assigned when the genome was being sequenced. In that section you can also see whether your gene is merely predicted based on genome sequencing or whether it has been confirmed by actually determining that cDNA has been made from it. cDNA stands for complementary DNA. In this case, cDNA would have been made from mRNA in using a poly-T primer (that binds to the poly-A tail of mRNA) and reverse transcriptase, in a laboratory setting. Making cDNA from mRNA is a common way to determine whether a gene is transcribed in mRNA. (Because it is made from mRNA, cDNA contains no introns.)

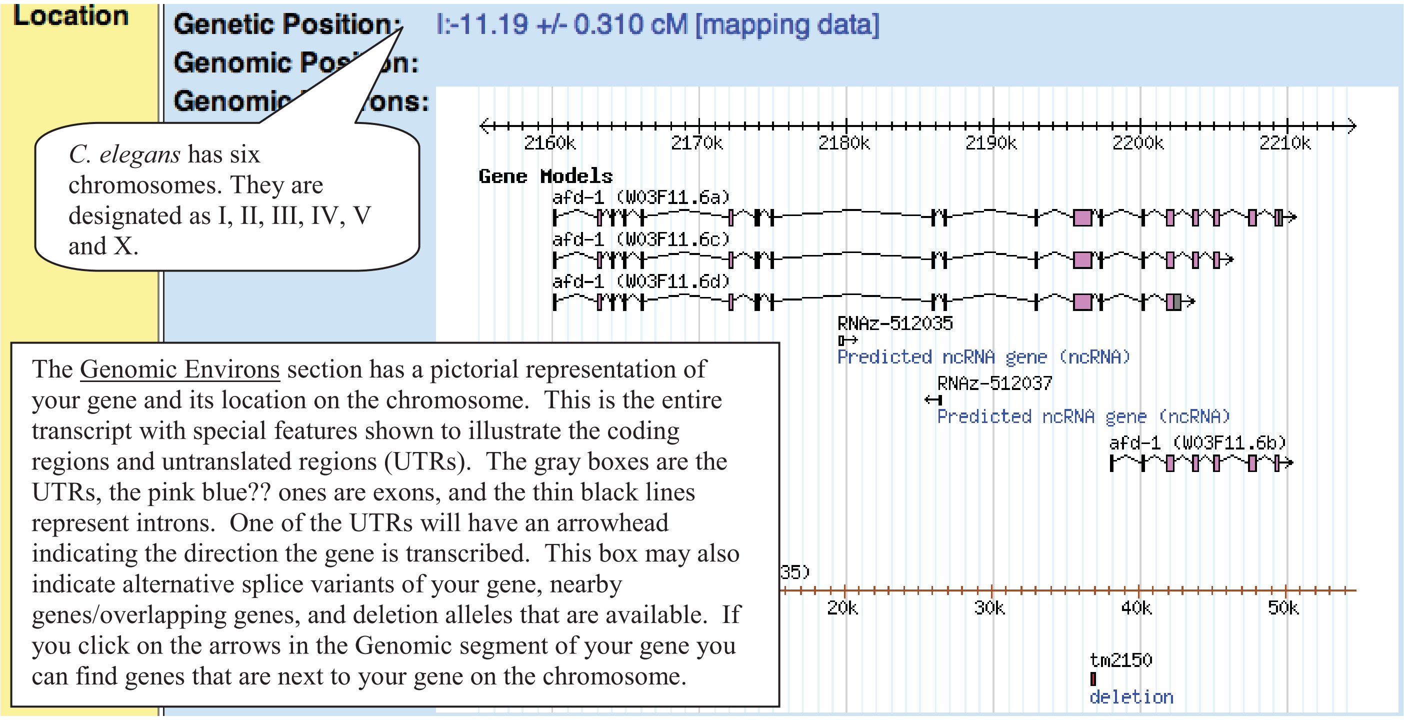
10) Has cDNA been found for your gene?

The NCBI Gene Model section also tells you how many nucleotide base pairs (bp's) are found in your gene. This includes the coding region (exons only) and the whole thing "transcript" (exons plus introns). Some genes are alternatively spliced so there may be multiple Gene Models for your gene. (Alternative splicing means that two or more similar mRNA transcripts are made from your gene by using different exons on some occasions.)

11) How long is the coding region for your gene? How long is the transcript? (For some genes only a dash is listed for the coding portion. This means that the actual coding portion has not been experimentally determined or that the information has not yet been entered in the database.)

Next, notice in the last column that the number of amino acids found in the protein encoded by your gene is given.
12) How many amino acids are in your protein of interest? Is this number what you expected given the coding number of bp's for your gene?

The next yellow row gives the *Location* of your gene, which is important for genetics and genomics studies.



Location

Genetic Position: I:-11.19 +/- 0.310 cM [mapping data]

Genomic Position:

Genomic Environs:

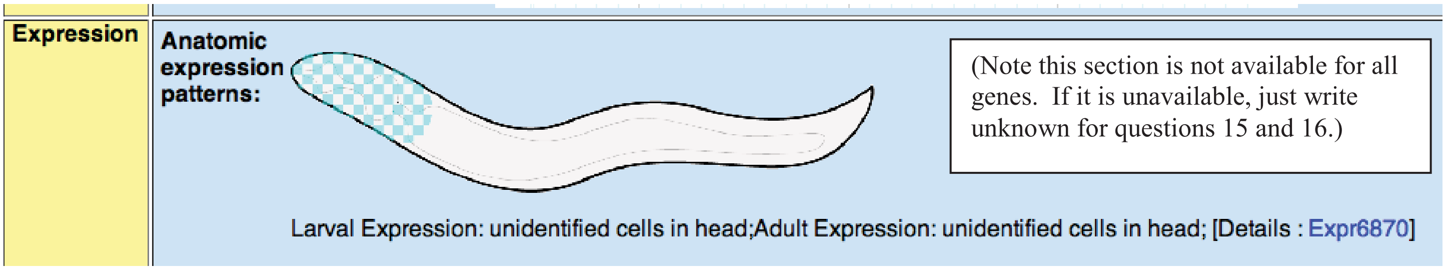
C. elegans has six chromosomes. They are designated as I, II, III, IV, V and X.

The Genomic Environs section has a pictorial representation of your gene and its location on the chromosome. This is the entire transcript with special features shown to illustrate the coding regions and untranslated regions (UTRs). The gray boxes are the UTRs, the pink blue?? ones are exons, and the thin black lines represent introns. One of the UTRs will have an arrowhead indicating the direction the gene is transcribed. This box may also indicate alternative splice variants of your gene, nearby genes/overlapping genes, and deletion alleles that are available. If you click on the arrows in the Genomic segment of your gene you can find genes that are next to your gene on the chromosome.

13) What *C. elegans* chromosome is your gene found on?

14) How many introns and exons are in your gene? (We will later explore the introns and exons further.)

You many have a yellow *Expression* row beneath the *Location* row. If you have an *Expression* row, it shows where your gene is expressed in a worm (anatomic expression pattern).



Expression | Anatomic expression patterns:

Larval Expression: unidentified cells in head;Adult Expression: unidentified cells in head; [Details : Expr6870]

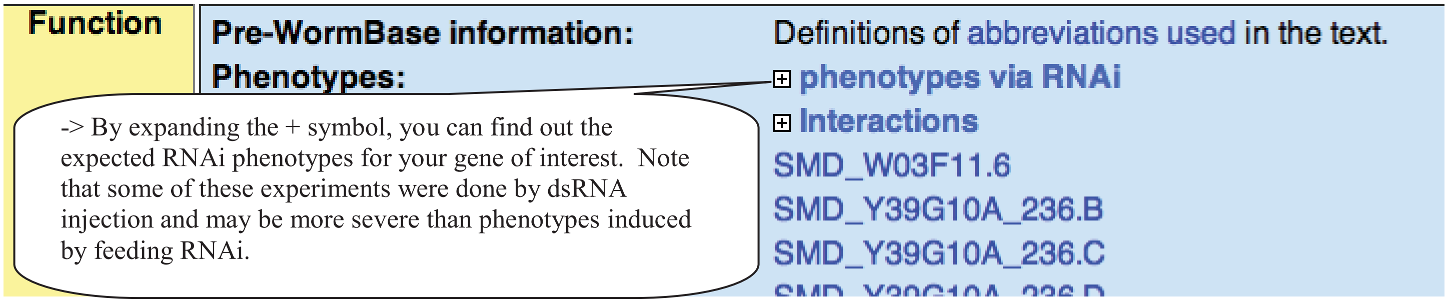
(Note this section is not available for all genes. If it is unavailable, just write unknown for questions 15 and 16.)

15) Where is your gene expressed?

16) -> Click on the link Expr..... **Which authors defined this expression?** (*Name, et. al* is sufficient)

Return to the summary Wormbase page for your gene.

The yellow *Function* section talks about gene function based on RNAi experiments that have been published and reported to the Wormbase curators.



Function | Pre-WormBase information: | Definitions of abbreviations used in the text.

Phenotypes:

⊞ phenotypes via RNAi

⊞ Interactions

SMD_W03F11.6

SMD_Y39G10A_236.B

SMD_Y39G10A_236.C

-> By expanding the + symbol, you can find out the expected RNAi phenotypes for your gene of interest. Note that some of these experiments were done by dsRNA injection and may be more severe than phenotypes induced by feeding RNAi.

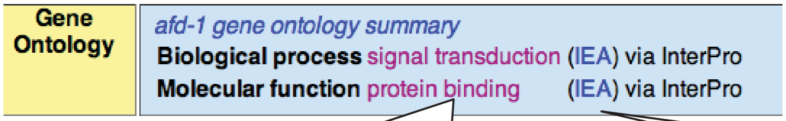
17) **What are a few of the RNAi phenotypes you expect to see for your gene?**

The *Function* row also shows microarray data. Like most microarray data, this part is a bit bewildering without careful examination so we'll just skip on to the last Protein domains section in this yellow *Function* row.

18) **What protein domains are predicted for your gene product?**

We will explore these protein domains in greater detail later.

The next yellow row is *Gene Ontology*, also known as GO. GO attempts to use a common language to describe gene function. The purpose of GO terms is to assign a probable function to genes identified by large scale sequencing projects that might not have been studied in the lab. The function assigned is based on protein sequence similarity to genes that have been studied in the lab. The actual assigned GO function may end up being wrong, but is a sufficient starting point for completely unstudied genes. Assigned GO functions use a controlled vocabulary to define a Biological process that the gene may be involved in, a Cellular component (where the protein is expected to be localized), and a Molecular function (a more specific biological function).



Gene Ontology | *afd-1 gene ontology summary*

Biological process signal transduction (IEA) via InterPro

Molecular function protein binding (IEA) via InterPro

Some genes, such as the afd-1 example shown here, are missing a GO section. In this case there is no "Cellular component" section.

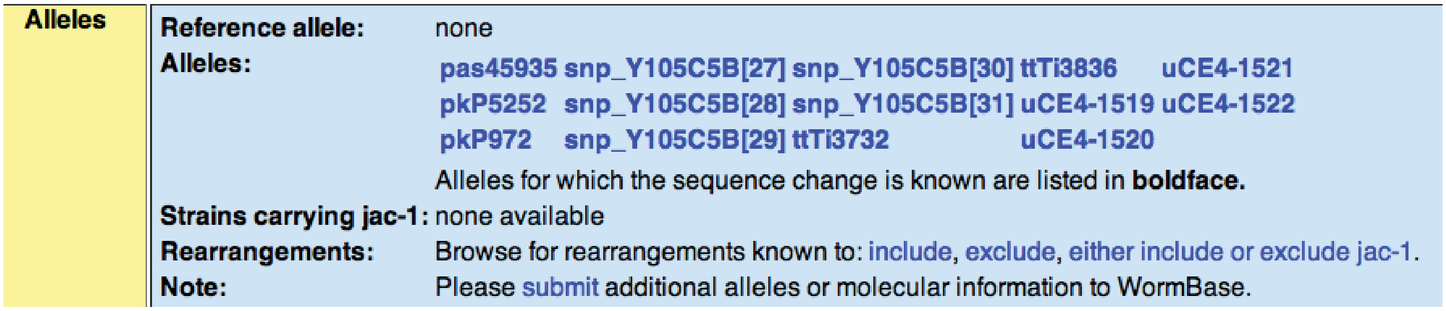
-> Click on the GO term to learn its definition.

GO terms that are assigned based on mutant phenotypes are called IMP: Inferred from Mutant Phenotype, and these are more likely to be accurate. Other sources of GO terms can be found to the right. -> You can click on each code to access a "Guide to GO Evidence Codes" for further information.

19) **For your gene/protein, what are the GO terms for Biological process, Cellular component, & Molecular function?**

20) **How were the GO terms assigned for each of the GO functions for your gene? Do you feel inclined to trust them? Why or why not?**

The next yellow *Alleles* row show mutant alleles and *C. elegans* strains available for your gene of interest. Your deletion allele and strain should be described there.



Alleles

Reference allele: none

Alleles: pas45935 snp_Y105C5B[27] snp_Y105C5B[30] ttTi3836 uCE4-1521 pkP5252 snp_Y105C5B[28] snp_Y105C5B[31] uCE4-1519 uCE4-1522 pkP972 snp_Y105C5B[29] ttTi3732 uCE4-1520

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

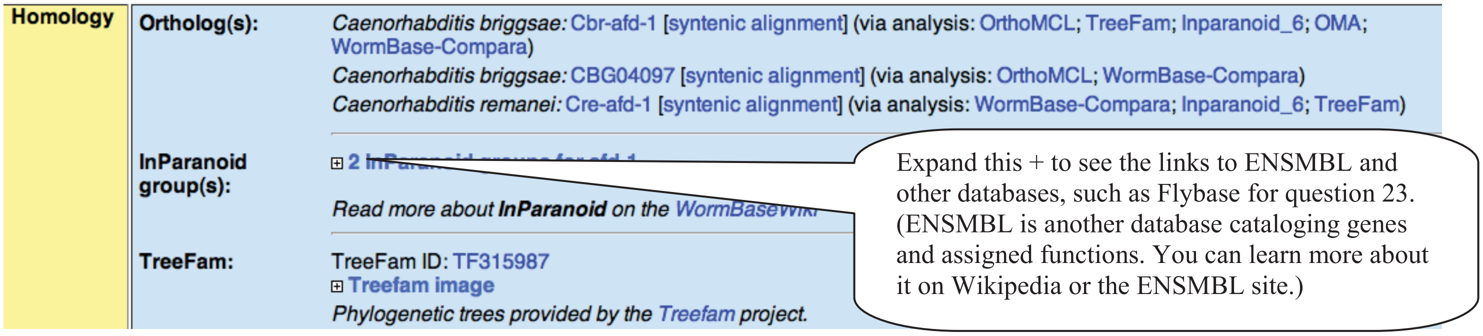
Strains carrying jac-1: none available

Rearrangements: Browse for rearrangements known to: include, exclude, either include or exclude jac-1.

Note: Please submit additional alleles or molecular information to WormBase.

21) **How many alleles are available for your gene?**

The yellow *Homology* row is useful for evolutionary studies and shows your gene's orthologs in other species with links to ENSEMBL and other protein databases. (According to the NCBI glossary of terms, "Orthology describes genes in different species that derive from a common ancestor. Orthologous genes may or may not have the same function.") We will explore these types of entries when we look at protein structure.



Homology

Ortholog(s): *Caenorhabditis briggsae:* Cbr-afd-1 [syntenic alignment] (via analysis: OrthoMCL; TreeFam; Inparanoid_6; OMA; WormBase-Compara)
Caenorhabditis briggsae: CBG04097 [syntenic alignment] (via analysis: OrthoMCL; WormBase-Compara)
Caenorhabditis remanei: Cre-afd-1 [syntenic alignment] (via analysis: WormBase-Compara; Inparanoid_6; TreeFam)

InParanoid group(s): ⊞ 2 InParanoid groups for afd-1
*Read more about **InParanoid** on the WormBase wiki*

TreeFam: TreeFam ID: TF315987
⊞ Treefam image
Phylogenetic trees provided by the Treefam project.

Expand this + to see the links to ENSMBL and other databases, such as Flybase for question 23. (ENSMBL is another database cataloging genes and assigned functions. You can learn more about it on Wikipedia or the ENSMBL site.)

22) **Pick two non-human species homologues and note their Latin name. Figure out what the species is in layman's terms and also include that name.** Are you surprised a similar protein is found in *C. elegans* worms and the species that you picked out?

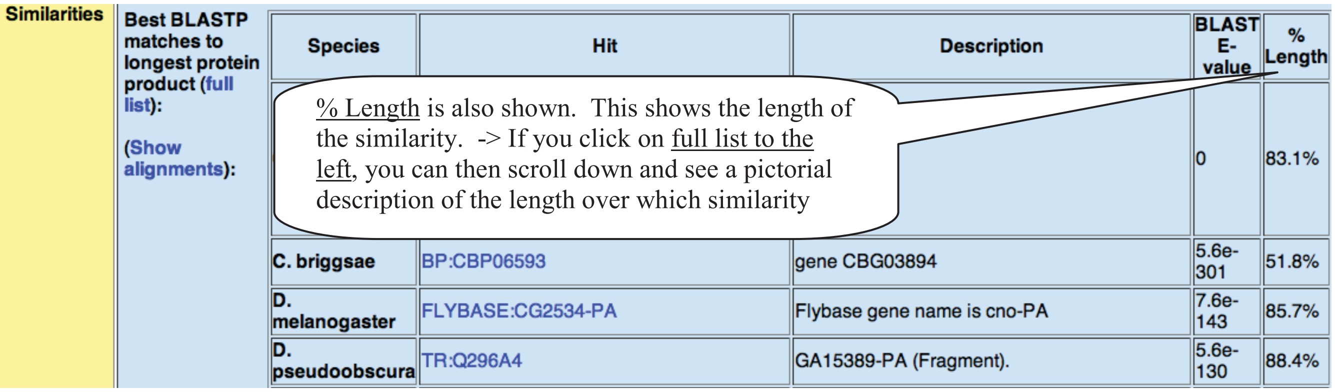
23) **Are there any human homologues of your gene? If so, what are they?** (For some genes you can find this by clicking on the ENSEMBL link and find the "Description" in the ENSEMBL Protein Report.) For other genes a link to ENSEMBL may be found in the yellow similarity section or there may not be a report available.

Return to the Wormbase summary page for your gene and find the yellow *Similarities* row. It shows BLASTP (Basic Local Alignment protein) data for your protein. BLASTP compares a protein sequence to protein databases to look for similarity. (This is what you did in the Enzyme unit's bioinformatics tutorial. See your Enzyme Catalysis 304 lab manual for more info on BLAST searches and interpreting BLAST results.)

24) **What species are listed with BLASTP hits for your gene?**

BLAST E-values are also shown. E-values show the number of matches expected to occur by random chance when you search a sequence the size of your protein against the known genomes. An E-value of 0 is perfect, meaning you'd never expect to get the match by chance alone. Other E-values can be very low and tell you that it is likely that the proteins are related evolutionarily. Higher E-values indicate the similarity may be due to chance, not an evolutionary relationship. Generally, higher E-values are not shown on Wormbase due to space restrictions so everything shown is likely to be related to your protein.

25) **What is the most similar BLASTP hit to yours?** If there is a Homo sapiens match, what is the % Length of the similarity?



Similarities

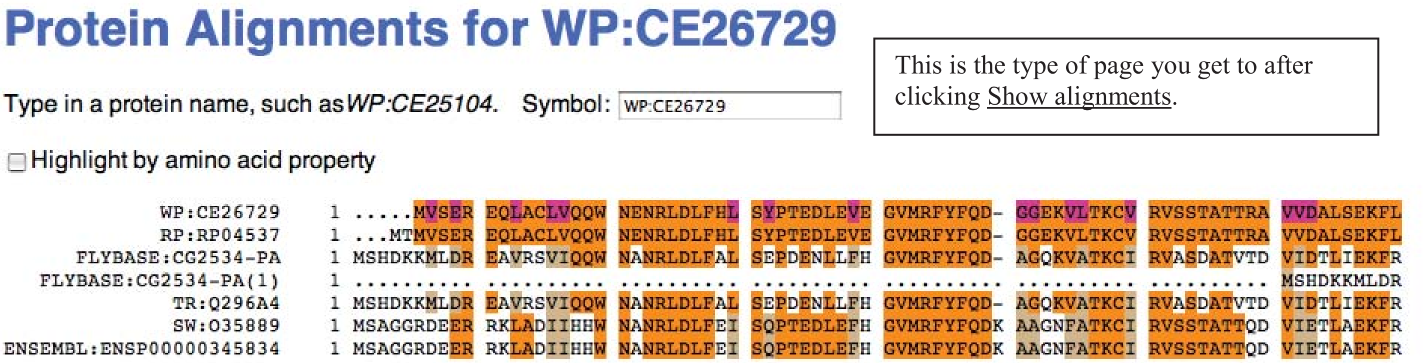
Best BLASTP matches to longest protein product (full list):

(Show alignments):

Species	Hit	Description	BLAST E-value	% Length
			0	83.1%
C. briggsae	BP:CBP06593	gene CBG03894	5.6e-301	51.8%
D. melanogaster	FLYBASE:CG2534-PA	Flybase gene name is cno-PA	7.6e-143	85.7%
D. pseudoobscura	TR:Q296A4	GA15389-PA (Fragment).	5.6e-130	88.4%

% Length is also shown. This shows the length of the similarity. -> If you click on full list to the left, you can then scroll down and see a pictorial description of the length over which similarity

-> The Show alignments link will show several related proteins and their level of similarity. The amino acids are represented by their individual amino acid symbol. (See Table 3-3 in your WOC text for these abbreviations.) The most conserved amino acids are assumed to be the most important to protein function. The meaning of the colored boxes can be found by scrolling down to the bottom of the webpage.



Protein Alignments for WP:CE26729

Type in a protein name, such as *WP:CE25104*. Symbol: WP:CE26729

□ Highlight by amino acid property

```
           WP:CE26729     1 .....MVSER EQLACLVQQW NENRLDLFHL SYPTEDLEVE GVMRFYFQD- GGEKVLTKCV RVSSTATTRA VVDALSEKFL
           RP:RP04537     1 ...MTMVSER EQLACLVQQW NENRLDLFHL SYPTEDLEVE GVMRFYFQD- GGEKVLTKCV RVSSTATTRA VVDALSEKFL
    FLYBASE:CG2534-PA     1 MSHDKKMLDR EAVRSVIQQW NANRLDLFAL SEPDENLLFH GVMRFYFQD- AGQKVATKCI RVASDATVTD VIDTLIEKFR
 FLYBASE:CG2534-PA(1)     1 .......... .......... .......... .......... .......... .......... .......... MSHDKKMLDR
            TR:Q296A4     1 MSHDKKMLDR EAVRSVIQQW NANRLDLFAL SEPDENLLFH GVMRFYFQD- AGQKVATKCI RVASDATVTD VIDTLIEKFR
            SW:O35889     1 MSAGGRDEER RKLADIIHHW NANRLDLFEI SQPTEDLEFH GVMRFYFQDK AAGNFATKCI RVSSTATTQD VIETLAEKFR
ENSEMBL:ENSP00000345834     1 MSAGGRDEER RKLADIIHHW NANRLDLFEI SQPTEDLEFH GVMRFYFQDK AAGNFATKCI RVSSTATTQD VIETLAEKFR
```

This is the type of page you get to after clicking Show alignments.

26) **Do you see any amino acids that are conserved in all species shown? Which sequence is from the orthologous human gene and which is from *C. elegans*? Where are these 2 sequences most similar?**

Return to the Wormbase gene summary page for your gene. We'll skip past the yellow Reagents section to the Bibliography section.

The Bibliography section shows citations for your gene that have been entered into Wormbase. Entry of citations by the Wormbase curators can sometimes be slow so searching PubMed would also be a good idea when you are making your poster. One advantage of the Bibliography section on Wormbase is that it includes abstracts from posters presented at Worm meetings. These abstracts may contain information that is not yet published. Sometimes the abstracts are based on preliminary data and thus may not be entirely accurate, but it can give you a good idea of what work is currently being done on your gene.

27) **Are there entries for your gene in the Bibliography section? Are they from Worm meeting abstracts, peer-reviewed publications or Wormbook?** (Wormbook is an open access collection of chapters covering various aspects of *C. elegans* biology: http://www.wormbook.org/)

II. Gene structure: exons, introns, and UTRs

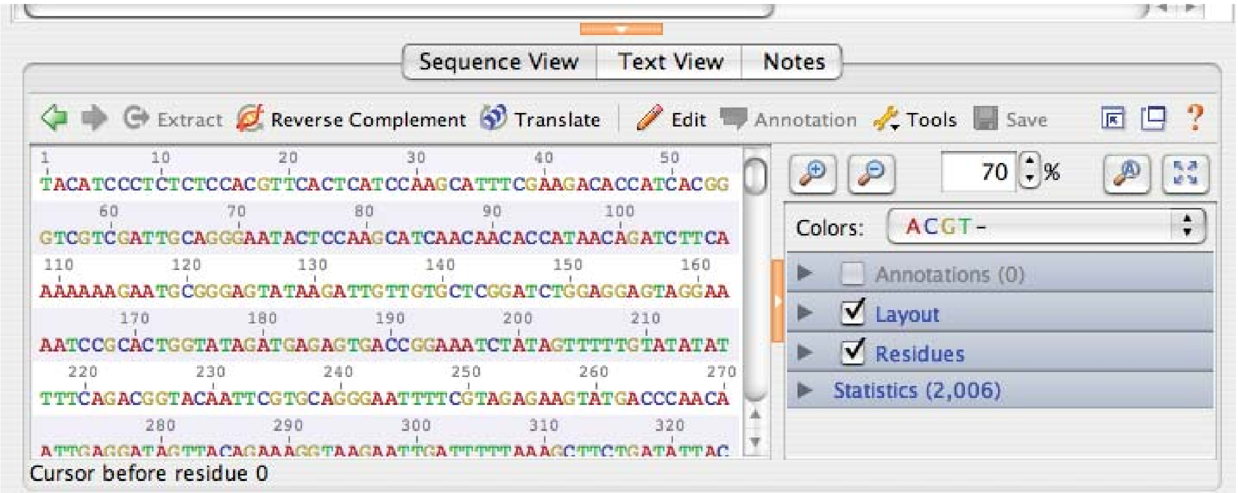
Because the worm gene you have been assigned is associated with a human disease, we want to better understand how it works. If we can figure out something about the gene in worms, we might be able to understand its role in humans better, as well. There is a wealth of resources available for worm biologists - beginning with Wormbase. Since you are already familiar with the Wormbase resources, we will now look in more detail on the structure of you gene. We'll use Wormbase along with a software package called Geneious together to examine gene and protein structure. We will first look at the wildtype genomic DNA sequence (same as unspliced RNA). Then, we'll compare this to the spliced RNA sequence - the one used to make translations. We'll also compare these to the deletion sequence and look at how the deletion affects the protein sequence.

The Geneious software is useful for gene alignment and translation. In addition, you can use Geneious to annotate your gene and protein sequences - that is you can add notes to your sequences, to make them easier to work with. This software comes in two versions: a free version and a Pro version, which can be purchased. You can also get a free one week trial of the Pro version. Because you can use the Pro version in lab, we suggest you wait until you are working on your poster to use your free one week trial.

When working in Geneious, make sure to save your Geneious documents frequently.

Eukaryotic genes contain introns. Your gene of interest will therefore likely contain introns. We will first look at the wildtype transcript sequence for your gene. Start with the Geneious file corresponding to your wildtype allele (name of file found on MyWebspace in the Primer Selection docs folder, with your gene name and containing 'genomic.doc' This is the same file that you used for your primer selection check assignment in week 8 lab.

-> In Geneious, click on Sequence -> New Sequence. Paste your sequence in and display the basepair sequence. It should look similar to this:



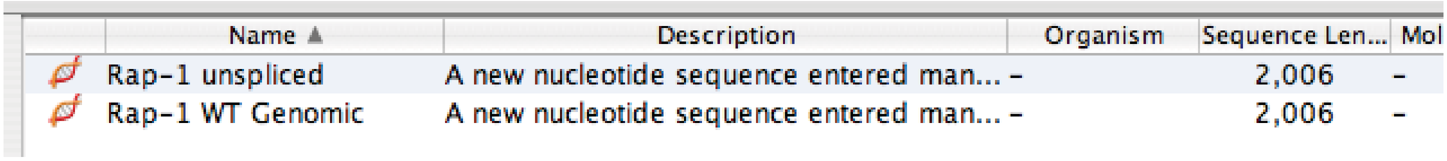
```
1        10        20        30        40        50
TACATCCCTCTCTCCACGTTCACTCATCCAAGCATTTCGAAGACACCATCACGG
     60        70        80        90        100
GTCGTCGATTGCAGGGAATACTCCAAGCATCAACAACACCATAACAGATCTTCA
110        120        130        140        150        160
AAAAAAGAATGCGGGAGTATAAGATTGTTGTGCTCGGATCTGGAGGAGTAGGAA
     170        180        190        200        210
AATCCGCACTGGTATAGATGAGAGTGACCGGAAATCTATAGTTTTTGTATATAT
  220        230        240        250        260        270
TTTCAGACGGTACAATTCGTGCAGGGAATTTTCGTAGAGAAGTATGACCCAACA
      280        290        300        310        320
ATTGAGGATAGTTACAGAAAGGTAAGAATTGATTTTTAAAGCTTCTGATATTAC
Cursor before residue 0
```

70 %

Colors: ACGT-

▶ □ Annotations (0)
▶ ☑ Layout
▶ ☑ Residues
▶ Statistics (2,006)

-> Duplicate your file in Geneious by copying and pasting the genomic file again into the New Sequence box. Rename the new file as "(your gene name) WT genomic" and under the "Description" heading indicate that it will have exons annotated. (You may already have your gene sequence from the primer design exercise. However, that may not exactly match the sequence available on Wormbase for several reasons. 1) the Wormbase sequence may be updated and made more accurate with time. 2) The sequences we provided on MyWebspace include more of the upstream and downstream non-coding regions of the gene because some of you have deletions that affect these regions, as well.

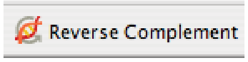


Name ▲	Description	Organism	Sequence Len...	Mol
Rap-1 unspliced	A new nucleotide sequence entered man...	-	2,006	-
Rap-1 WT Genomic	A new nucleotide sequence entered man...	-	2,006	-

-> Find your gene by searching on Wormbase: www.wormbase.org. In the yellow "*Identification*" row click on the link under the Nucleotides. This will display at least one "spliced coding region" sequence for your gene highlighted in various shades of yellow and orange. Each transition in shade of yellow represents a new exon (e.g., first yellow sequence = exon 1, following orange sequence = exon 2, etc.).

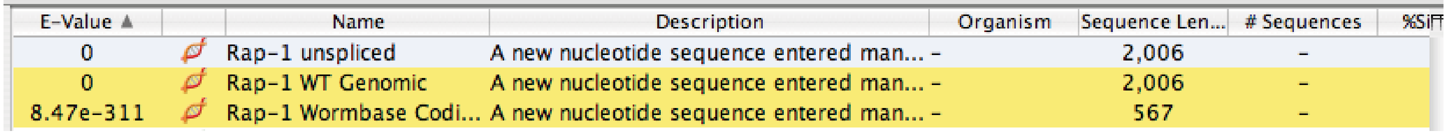
-> Copy this sequence, then create a new sequence in Geneious and paste this coding sequence in. Name it as "Wormbase coding sequence". (Note: the Explorer web browser may not allow you to select and copy these sequences. Try Firefox instead.)

-> Highlight and copy the first twenty or so amino acids in this Wormbase coding sequence. Search for these in the genomic DNA file. *Note: you may have to reverse complement your file if the matching sequence is found at the end of the gene, rather than the beginning.*

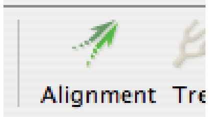


Now you can compare the sequences of the genomic DNA to that of the coding sequence. By doing so you can figure out the locations of exons, introns, and UTRs. This is important because your deletion allele may occur anywhere in the genomic sequence, but what matters most about the allele is the effect on the coding sequence.

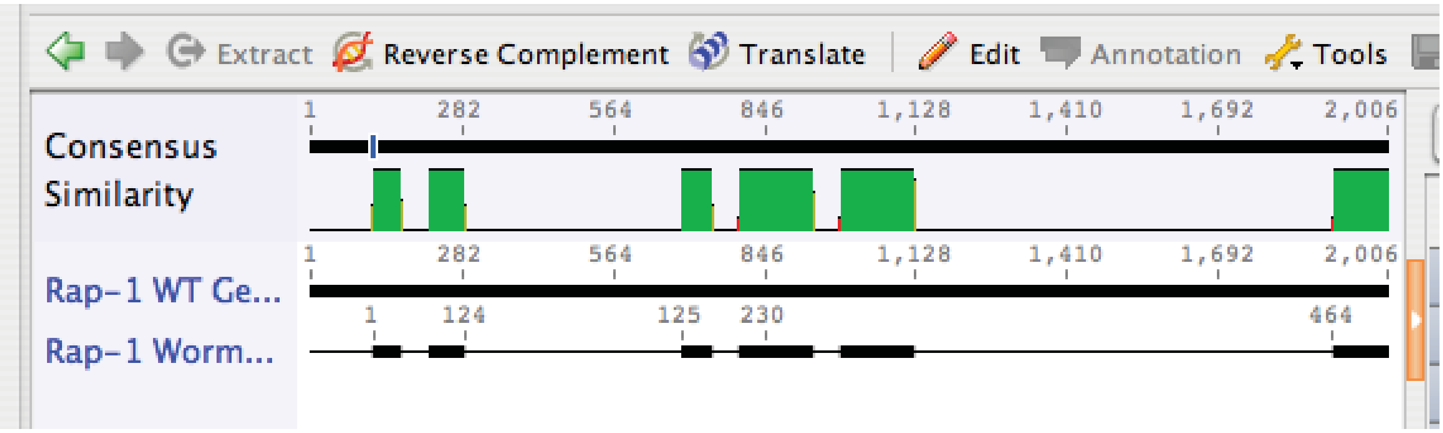
The quickest way to compare sequences is to let the software run an algorithm that lines up sequences. To do so, select the WT genomic (exons annotated) and coding sequences and click on alignment (then press OK).



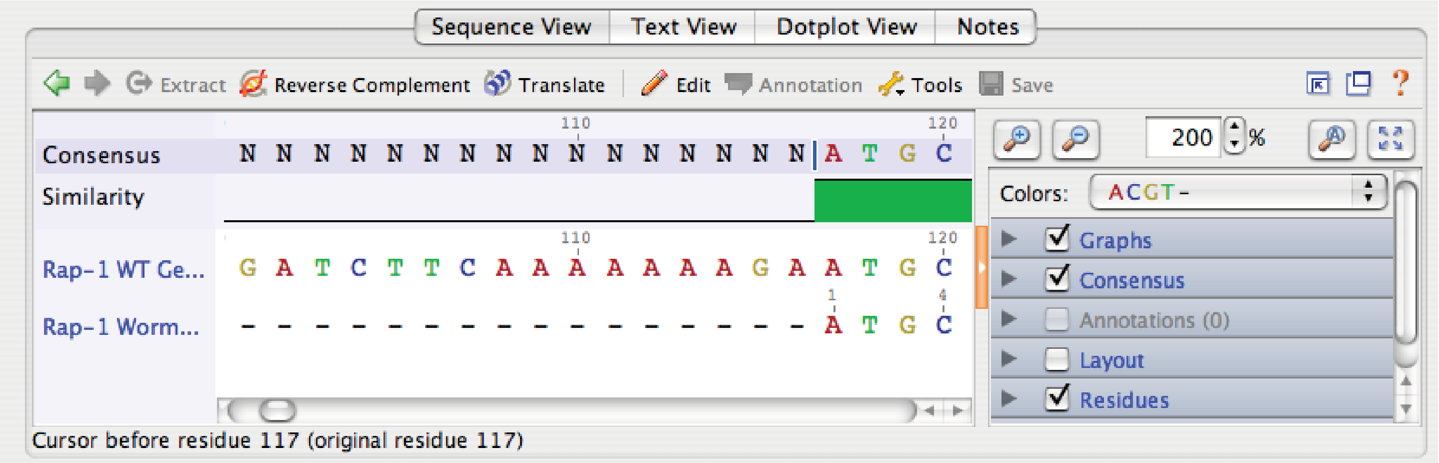
E-Value ▲		Name	Description	Organism	Sequence Len...	# Sequences	%Si
0		Rap-1 unspliced	A new nucleotide sequence entered man...	-	2,006	-	
0		Rap-1 WT Genomic	A new nucleotide sequence entered man...	-	2,006	-	
8.47e-311		Rap-1 Wormbase Codi...	A new nucleotide sequence entered man...	-	567	-	



Your alignment file should then appear.



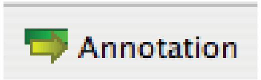
The green regions are regions that match. The represent the exons!
You can enlarge the font so that you can see individual nucleotide displayed as follows.

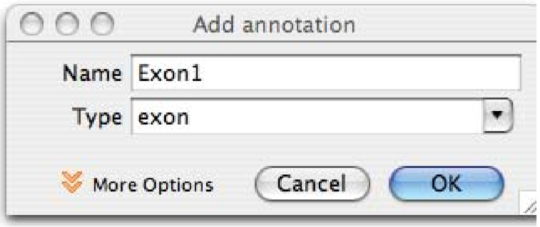


```
Consensus        N N N N N N N N N N N N N N N N|A T G C
Similarity
Rap-1 WT Ge...   G A T C T T C A A A A A A A G A A T G C
Rap-1 Worm...    - - - - - - - - - - - - - - - - A T G C
Cursor before residue 117 (original residue 117)
```

In the zoomed in view, scan to the side until you find a region that matches in the Genomic and Coding sequences. Note the first three nucleotides. What amino acid do these encode?

Now you can annotate the gene. To annotate, highlight the region you want to annotate (select all of exon 1 in the genomic file), click the annotate box and then select "exon" from the type box. The first matching region should be annotated as exon 1.





Name: Exon1
Type: exon

Annotate the region before that as the 5'UTR. The first codon in exon 1 should be an ATG. We would like you to annotate the beginning and end of each exon in the WT genomic sequence and also the 3'UTR. Go through the rest of the sequence and annotate the rest of the exons. The number of exons you end up with should match the number of exons you recorded for question #14 in section I, above.

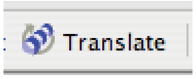
Save (apply changes to the originals) and print the annotated Geneious document you created; it may be useful in your poster.

28) **Does your deletion occur in exons, introns or in the UTRs?**

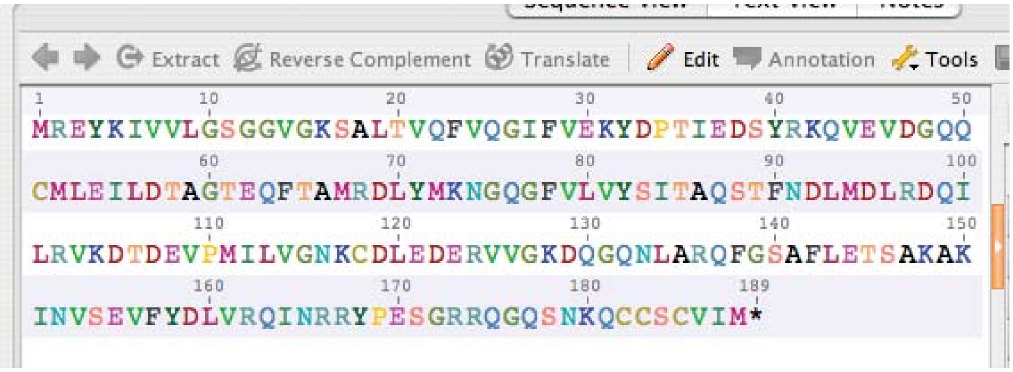
III. Protein sequence encoded by your gene

In the Geneious program, go to your Wormbase coding file for the coding region. (Remember, this has no introns.) Make sure no bases are selected.

Translate your sequence by clicking the "Translate" icon just above your sequence.



Choose the universal codon table for your translation. The amino acid sequence for your protein should appear and this should match the amino acid sequence on Wormbase. (Check this by returning to your Wormbase summary page and clicking on the Protein link in the yellow "*Identification*" row.)

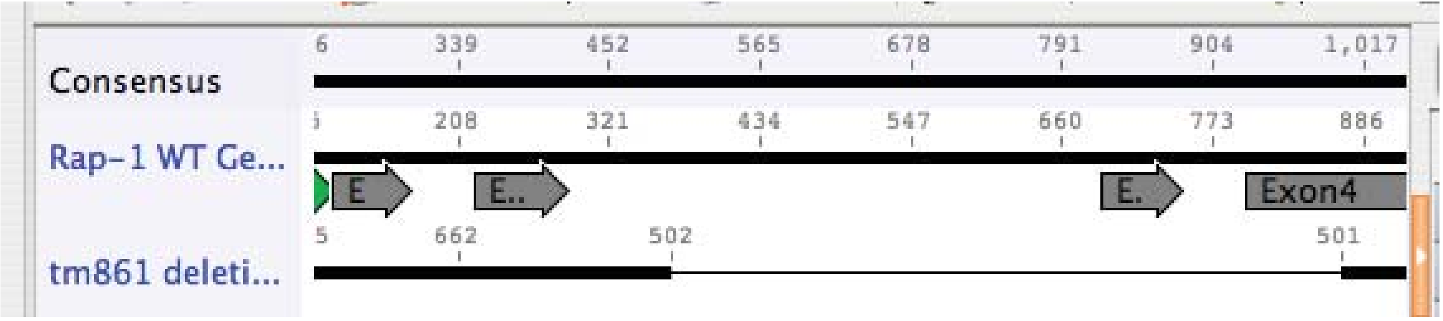


```
1        10        20        30        40        50
MREYKIVVLGSGGVGKSALTVQFVQGIFVEKYDPTIEDSYRKQVEVDGQQ
         60        70        80        90        100
CMLEILDTAGTEQFTAMRDLYMKNGQGFVLVYSITAQSTFNDLMDLRDQI
         110       120       130       140       150
LRVKDTDEVPMILVGNKCDLEDERVVGKDQGQNLARQFGSAFLETSAKAK
         160       170       180      189
INVSEVFYDLVRQINRRYPESGRRQGQSNKQCCSCVIM*
```

To make your sequence less cluttered you can go to the graphs section in the Geneious program and unselect the box for Graphs. Notice that in the nucleotide sequence file in Geneious that you can translate in frame 0, 1 or 2 (i.e., you can highlight all but the first nucleotide to translate in frame 2, or all but the first two nucleotides to translate in frame 3, etc.). Only one of these will give the correct amino acid sequence. The other translations will be out of frame and are easily spotted if because they will contain numerous *'s, which here indicate the presence of a stop codon. Observe this by highlighting various nucleotide sequences and then translating them. (You can delete these files from your Geneious folder after you look at them.)

It can be difficult to predict the size and sequence of the protein that is made from your deletion allele due to problems with splicing when parts of introns are lost. The best way to do it is experimentally, via reverse transcriptase PCR and northern blotting. However, many scientists don't go that far. Like them, you can attempt to predict the protein made by hand. As you did with the gene, make a new alignment file between your newly (exon) annotated genomic file and your deletion file (on Mywebspace), then find the deleted base pairs. If this new file has lots of gaps, instead of one big one, you may have to go back and reverse complement a sequence before aligning.

Here is a zoomed out view of my deletion.



Now find which exons are affected by the deletion. Duplicate the Wormbase coding sequence file (select name of file, right click, then select "Copy Document" then paste) and then delete those exons or parts of exons that are affected by the deletion. (Highlight deleted bases, select the Edit button, then press delete.)

(You can search for the deleted bases using edit, find OR go back to your original alignment and see where the deletion is on that.)

Rename this file. Try translating now.

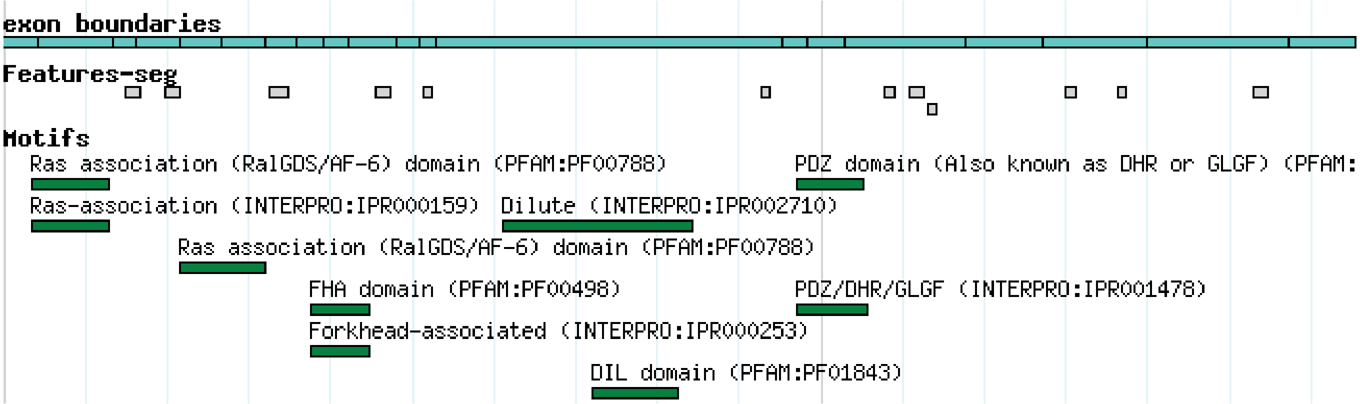
29) **Does it appear that your deletion alters splicing, the reading frame, and/or introduces a premature stop codon in the RNA transcript?**

Save and print the documents you create for use in your poster. We will further annotate your protein sequence using structural information in the next section. (You will learn how to look at the protein structure of your deletion too!)

IV. Protein structure for the protein encoded by your gene

This worksheet will help you to learn a few bioinformatics tools to examine the domains found in your protein. On the summary page for your gene in Wormbase, find the yellow *Function* row and scroll down to the section called Protein domains. In this section you can find a listing of protein domains found in your protein and that you recorded for question 18.

Return to the first yellow *Identification* row at the top of the gene summary page and click on the Protein link. This will bring you to a Protein Report with information about your protein. Scroll down beneath the amino acid sequence of your protein to the green bars, which represent regions of your protein which correspond to each of the functional domains you listed in question 18.

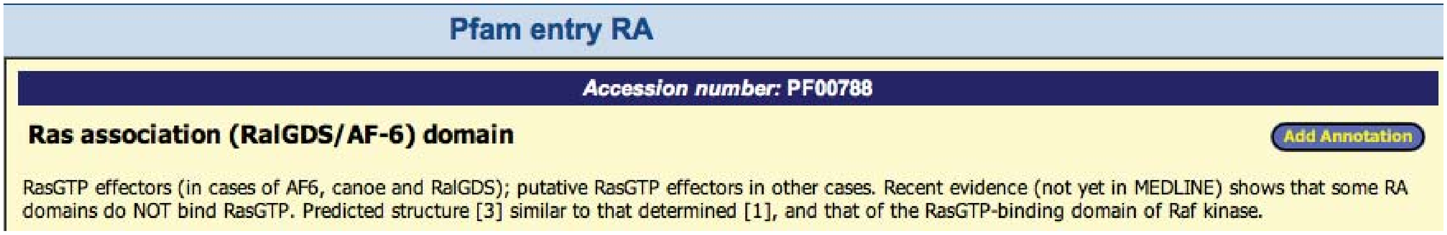

exon boundaries
Features-seg
Motifs
Ras association (RalGDS/AF-6) domain (PFAM:PF00788)
PDZ domain (Also known as DHR or GLGF) (PFAM:
Ras-association (INTERPRO:IPR000159)
Dilute (INTERPRO:IPR002710)
Ras association (RalGDS/AF-6) domain (PFAM:PF00788)
FHA domain (PFAM:PF00498)
PDZ/DHR/GLGF (INTERPRO:IPR001478)
Forkhead-associated (INTERPRO:IPR000253)
DIL domain (PFAM:PF01843)

Note that although Wormbase uses the words "domain" and "motif" interchangeably, they are distinct from one another. According to the NCBI glossary (http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/glossary2.html), a domain is "a discrete portion of a protein assumed to fold independently of the rest of the protein and possessing its own function. Motifs are frequently highly conserved parts of domains."

The exon boundaries for the coding sequence are also shown in blue, above the motifs, so that you can get an idea of where each of the protein motifs falls in relation to where your known deletion is. The labels above each green bar end with a unique Accession number which match the linked accession numbers in the "Motif Summary" table just below. The first column in this table indicates whether the InterPro or PFAM databases were used to identify any given motif in your protein.

Click on the green bars to access an InterPro or PFAM report for each motif.

Here is an example of a PFAM report:



Pfam entry RA

Accession number: PF00788

Ras association (RalGDS/AF-6) domain

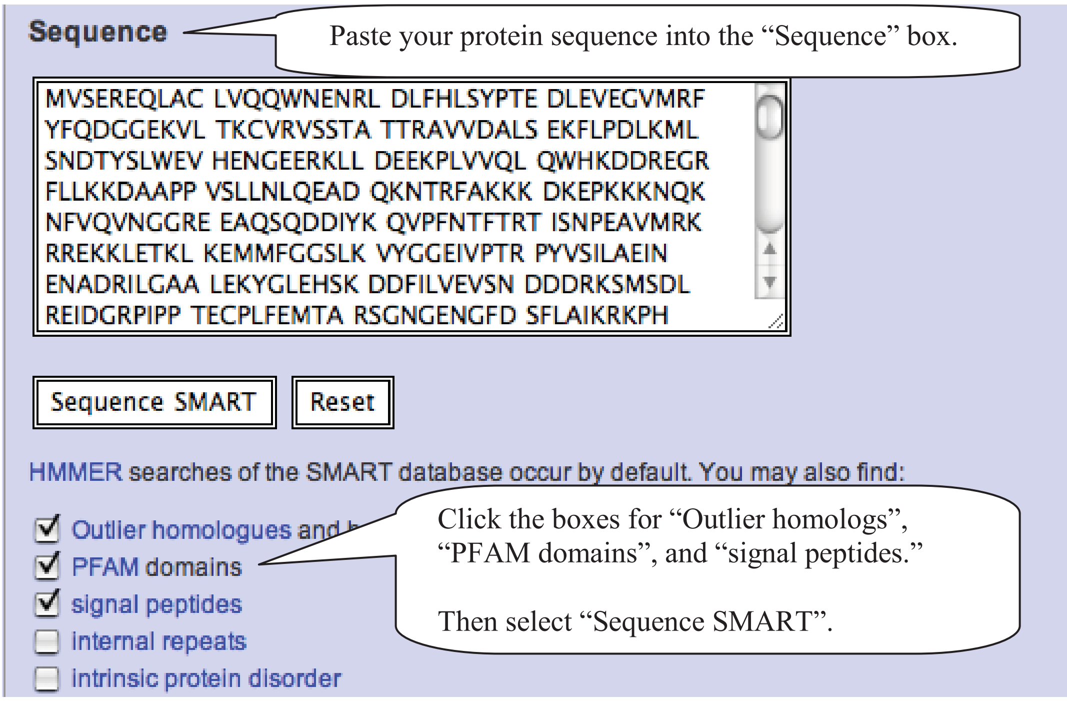
RasGTP effectors (in cases of AF6, canoe and RalGDS); putative RasGTP effectors in other cases. Recent evidence (not yet in MEDLINE) shows that some RA domains do NOT bind RasGTP. Predicted structure [3] similar to that determined [1], and that of the RasGTP-binding domain of Raf kinase.

30) Record and *briefly* summarize the predicted function of each motif.

Different databases may come up with different domains for your protein so it is useful to look at a few other databases. First, copy the amino acid sequence from the Wormbase protein report page.

SMART Database

The SMART (Simple Modular Architecture Research Tool) database allows the identification and annotation of protein domains. Go to the SMART database (http://smart.embl-heidelberg.de/). Choose the normal SMART mode.



Sequence

Paste your protein sequence into the "Sequence" box.

```
MVSEREQLAC LVQQWNENRL DLFHLSYPTE DLEVEGVMRF
YFQDGGEKVL TKCVRVSSTA TTRAVVDALS EKFLPDLKML
SNDTYSLWEV HENGEERKLL DEEKPLVVQL QWHKDDREGR
FLLKKDAAPP VSLLNLQEAD QKNTRFAKKK DKEPKKKNQK
NFVQVNGGRE EAQSQDDIYK QVPFNTFTRT ISNPEAVMRK
RREKKLETKL KEMMFGGSLK VYGGEIVPTR PYVSILAEIN
ENADRILGAA LEKYGLEHSK DDFILVEVSN DDDRKSMSDL
REIDGRPIPP TECPLFEMTA RSGNGENGFD SFLAIKRKPH
```

HMMER searches of the SMART database occur by default. You may also find:

☑ Outlier homologues and h

☑ PFAM domains

☑ signal peptides

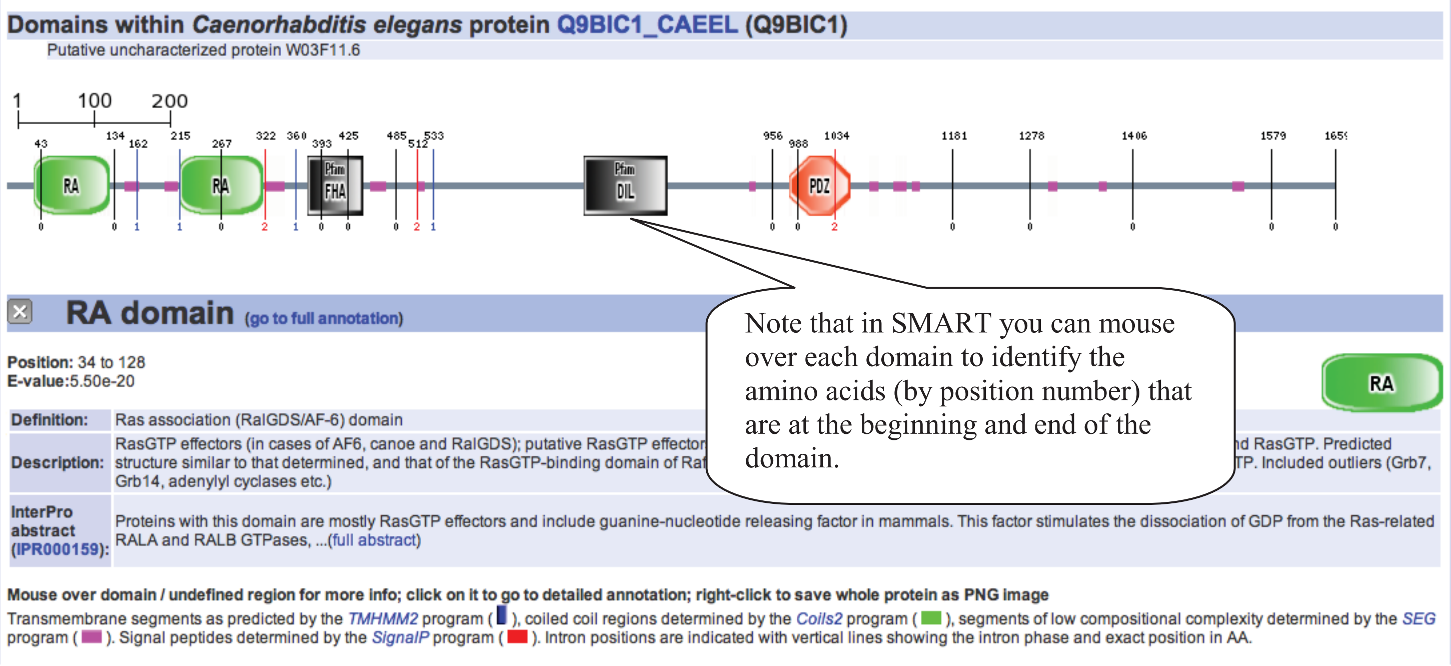
☐ internal repeats

☐ intrinsic protein disorder

Click the boxes for "Outlier homologs", "PFAM domains", and "signal peptides."

Then select "Sequence SMART".

You may have to wait a few minutes for your results. You should now see your protein represented by a grey line with large rectangular boxes representing structural domains.



Domains within *Caenorhabditis elegans* protein Q9BIC1_CAEEL (Q9BIC1)

Putative uncharacterized protein W03F11.6

Note that in SMART you can mouse over each domain to identify the amino acids (by position number) that are at the beginning and end of the domain.

RA domain (go to full annotation)

Position: 34 to 128
E-value:5.50e-20

Definition:	Ras association (RalGDS/AF-6) domain
Description:	RasGTP effectors (in cases of AF6, canoe and RalGDS); putative RasGTP effector nd RasGTP. Predicted structure similar to that determined, and that of the RasGTP-binding domain of Raf TP. Included outliers (Grb7, Grb14, adenylyl cyclases etc.)
InterPro abstract (IPR000159):	Proteins with this domain are mostly RasGTP effectors and include guanine-nucleotide releasing factor in mammals. This factor stimulates the dissociation of GDP from the Ras-related RALA and RALB GTPases, ...(full abstract)

Mouse over domain / undefined region for more info; click on it to go to detailed annotation; right-click to save whole protein as PNG image

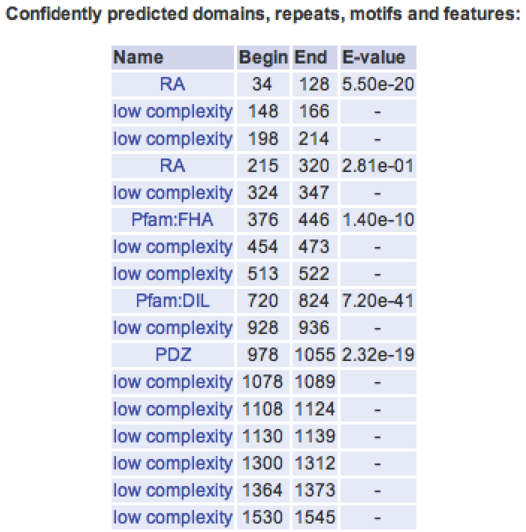
Transmembrane segments as predicted by the *TMHMM2* program (▮), coiled coil regions determined by the *Coils2* program (🟩), segments of low compositional complexity determined by the *SEG* program (🟪). Signal peptides determined by the *SignalP* program (🟥). Intron positions are indicated with vertical lines showing the intron phase and exact position in AA.

(Does the protein amino acid sequence you imported into SMART match the translated protein sequence file you made in the Genious program? Recall that you created this amino acid sequence by translating the Wormbase coding sequence. Check this.)

31) What domains did SMART identify in your protein with confidence (*i.e.* have very small E-values reported)? (Hint: scroll down on the SMART page to see a table like that shown below.)

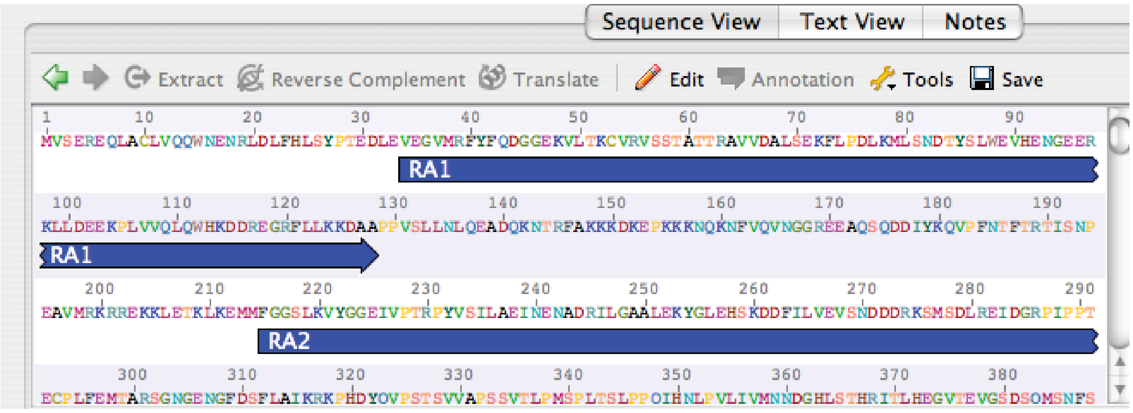
32) If additional domains are listed define their function as for #30 above.

Use the numbered amino acids from the SMART program to annotate your protein sequence in the Geneious program. Here you can choose “peptide” for the type of annotation.



Confidently predicted domains, repeats, motifs and features:

Name	Begin	End	E-value
RA	34	128	5.50e-20
low complexity	148	166	-
low complexity	198	214	-
RA	215	320	2.81e-01
low complexity	324	347	-
Pfam:FHA	376	446	1.40e-10
low complexity	454	473	-
low complexity	513	522	-
Pfam:DIL	720	824	7.20e-41
low complexity	928	936	-
PDZ	978	1055	2.32e-19
low complexity	1078	1089	-
low complexity	1108	1124	-
low complexity	1130	1139	-
low complexity	1300	1312	-
low complexity	1364	1373	-
low complexity	1530	1545	-



```
1        10        20        30        40        50        60        70        80        90
MVSEREQLACLVQQWNENRLDLFHLSYPTEDLEVEGVMRFYFQDGGEKVLTKCVRVSSTATTRAVVDALSEKFLPDLKMLSNDTYSLWEVHENGEER
 100       110       120       130       140       150       160       170       180       190
KLLDEEKPLVVQLQWHKDDREGRFLLKKDAAPPVSLLNLQEADQKNTRFAKKKDKEPKKKNQKNFVQVNGGREEAQSQDDIYKQVPFNTFTRTISNP
    200       210       220       230       240       250       260       270       280       290
EAVMRKRREKKLETKLKEMMFGGSLKVYGGEIVPTRPYVSILAEINENADRILGAALEKYGLEHSKDDFILVEVSNDDDRKSMSDLREIDGRPIPPT
       300       310       320       330       340       350       360       370       380
ECPLFEMTARSGNGENGFDSFLAIKRKPHDYQVPSTSVVAPSSVTLPMSPLTSLPPQIHNLPVLIVMNNDGHLSTHRITLHEGVTEVGSDSQMSNFS
```

Save and print the file you generate; you may want to put it in your poster.

33) From the annotated protein and knowing the location of your deletion, what protein domains are affected by the deletion? How might the deletion of these/this domain(s) affect the function of your protein?

34) Look at the answers you gave to questions 25 and 26 above, when you were asked to compare the worm protein sequence with its human orthologous sequence in Wormbase. Do the most closely aligned portions of the human and worm sequences encode any structural domains that you found in SMART? If there are similar structural domains in the worm and human sequences, are they affected by your deletion?

Now, you can choose to do section VI below (optional) or skip to **section VII (required)** to complete this check assignment.

******* Section VI below is optional and so does not have to be completed to earn your check for this assignment. It may or may not be helpful for your poster*******

VI. Exploring the structure of proteins expressed by orthologous human genes

If the human ortholog associated with your worm gene expresses a protein (or a portion of a protein) that has been crystallized, you may be able to view some of the structures predicted by the SMART program. To find out, return to the OMIM page summarizing your human ortholog. Click on the "Links" link at upper right corner of the page. You will find links to other NCBI databases, including PubMed, Entrez Protein, and perhaps Entrez Structure. If there is a "Structure" link, click on it to find the PDB (Protein Data Bank) number of the crystallized protein and the paper describing it. (Refer to the directions near the end of the Enzyme Catalysis summary in the lab manual if you can't remember how to find the publication associated with a PDB file.) Copy the PDB number and open up the FirstGlance in Jmol website (http://molvis.sdsc.edu/fgij/index.htm). Enter you PDB number in the white box and press "Submit." When your image shows up, press the Refresh button to enlarge it, and then press the "spin" button to stop the image from rotating. (Refer to the Jmol tutorial in the Enzyme Catalysis section if you can't remember all of the tools available for viewing structures in Jmol.) You will have to look at the original paper describing this structure to find out what is actually depicted in the Jmol image. If there are references to the protein product domains for your gene that are similar in humans and worms, use the paper to find them on your Jmol image.

If there are no "Structure" links from the OMIM page, find the Entrez Structure page by returning to the NCBI homepage and clicking on the blue "Structure" link near the top right of the page. Here you can type in any key words or accession numbers that you have found over the last two weeks to find crystallized protein files. Each match that appears will have a blue PDB file code linked to a MMDB (Molecular Modeling Database) Structure summary page. Write down any PDB files that may be interesting to view using Jmol and go to the Jmol FirstGlance site to view them (http://molvis.sdsc.edu/fgij/index.htm).

VII. Putting it all together

35) Based on what you've learned from your work with the Wormbase, Genious, OMIM, SMART, and (perhaps) Jmol databases/programs, logically speculate about why the deletion yields the phenotype it does.

36) Do you expect this deletion to result in complete loss of gene function?

37) Assuming that your RNAi is effective, do you expect to see a similar or different phenotype with RNAi compared to the deletion? Briefly explain your answer.