**Updating the GExplore gene and mutation databases**

This file contains instructions for downloading and processing of Wormbase data to build the sql databases that correspond to the "gene" and "mutation" databases in GExplore.

**Requirements:**

Ability to run perl scripts.

**Notes:**

* change the Wormbase release number to whatever release you are processing. The example is for WS296.
* collect all files generated in the "Data collection" section in the folder that contains the perl scripts and the additional files (see "Data processing" section)

**Data collection**

**Download data from the Wormbase FTP server for the gene and mutation databases**

connect to

ftp://ftp.wormbase.org/pub/wormbase/releases/WS294/species/c\_elegans/PRJNA13758

download file: c\_elegans.PRJNA13758.WS294.protein.fa

note: this file is needed for comparison with the new version (WS296 in this example). SMART domain predictions for all WS294 protein are included (folder "SMART\_results\_all\_WS294" so that only proteins that are new are processed for SMART domain prediction.

connect to

ftp://ftp.wormbase.org/pub/wormbase/releases/WS296/species/c\_elegans/PRJNA13758

download files

c\_elegans.PRJNA13758.WS296.annotations.gff3

c\_elegans.PRJNA13758.WS296.protein.fa

goto the 'annotation' folder and download

c\_elegans.PRJNA13758.WS296.orthologs.txt.gz

unpack gz files

**Download data from WormMine for the mutation database**

* Goto the WormMine query builder http://intermine.wormbase.org/tools/wormmine/importQueries.do
* paste the content of the file "query\_protein\_alleles\_v4.xml" into the search field and hit submit
* hit "show results"
* you should get a result page saying something like "Showing 1 to 25 of 2,806,564 rows"
* hit the "export" button and the "download file button" accepting the defaults (TSV file). Ignore any messages saying "Error The offset is greater than the total number of results." The download will take a while to complete.
* rename the result file to “WormMine\_protein\_to\_alleles\_WS296.tsv”

**Download SimpleMine data**

Run the following queries on SimpleMine

https://wormbase.org/tools/mine/simplemine.cgi

Settings for all queries

Select "Caenorhabditis elegans" as species

Select " download results as a tab-delimited file"

Set checkboxes "Public Name" and "Sequence Name"

Query "all genes in this species"

Perform individual queries for

**Map position**

[x] Genetic Map Position

[x] Chromosome Coordinates

**Phenotype (Allele)**

[x] Allele Phenotype Observed

**Phenotype RNAi**

[x] RNAi Phenotype Observed

**Expression**

[x] Expr\_pattern Tissue

[x] Expr\_pattern LifeStage

**Expression - genomic studies**

[x] Genomic Study Tissue

[x] Genomic Study LifeStage

**Interacting genes**

[x]Interacting Gene

**Disease association**

[x] Disease Info

**GO terms**

[x] Gene Ontology Association

**Description**

[x] Concise Description

[x] Automated Description

rename result files to match the following names

simplemine\_results\_description.txt

simplemine\_results\_disease\_association.txt

simplemine\_results\_expression\_genomic\_study.txt

simplemine\_results\_expression.txt

simplemine\_results\_GO\_terms.txt

simplemine\_results\_interacting\_genes.txt

simplemine\_results\_location.txt

simplemine\_results\_phenotype\_allele.txt

simplemine\_results\_phenotype\_RNAi.txt

**Download CGC data**

* download CGC strain list from https://cgc.umn.edu  
  Link "Strain List (text file)"; convert line break types to unix  
  name file: cgc-strains\_unix.txt
* on the CGC website, search for 'elegans', download csv file, convert line break types to unix; name file: ce-strains\_all\_unix.csv

**Data processing**

run the following command "perl create\_GExplore\_db\_tables.pl 294 296" with 296 being the wormbase release number you are processing. All files listed below need to be in the same folder.  
  
**Required input files**

All downloaded files (blue filenames from the data collection section)

Translation tables

SMART\_name\_to\_ID.txt

PfamA\_name\_to\_ID.txt

translation\_table\_v4.txt

meta\_domain\_table\_v4.txt

Perl scripts

perl get\_IDs\_and\_seqs.pl  
process\_orthologs.pl  
process\_simplemine\_data.pl  
filter\_proteins\_for\_SMART.pl  
process\_SMART\_files.pl   
create\_sql\_protein\_domain\_table.pl  
process\_gff\_file.pl   
filter\_variations.pl   
process\_alleles.pl   
collect\_KO\_alleles\_from\_CGC\_data.pl  
create\_sql\_mutation\_table.pl   
SMART\_batch2.pl

**Output files**

|  |  |
| --- | --- |
| for gene database | for mutation database |
| WS296\_all\_IDs\_sql.txt WS296\_protein\_domains\_seq\_sql.txt WS296\_location\_sql.txt WS296\_description\_sql.txt WS296\_expression\_sql.txt WS296\_expression\_genomic\_study\_sql.txt WS296\_phenotype\_allele\_sql.txt WS296\_phenotype\_RNAi\_sql.txt WS296\_GO\_terms\_sql.txt WS296\_interacting\_genes\_sql.txt WS296\_disease\_association\_sql.txt WS296\_orthologs\_sql.txt | WS296\_protTable\_mutDB\_SQL.txt WS296\_mutation\_table\_SQL.txt |

**Build SQL databases**

The files are designed to build the GExplore databases. If you are interested in only some of the data, a breakdown of the file content is listed below.

**Build gene database**

* within mysql use "source create\_gexplore\_db\_WS296\_genes.sql" to create the database and load basic data  
  requires files  
  WS296\_all\_IDs\_sql.txt  
  WS296\_protein\_domains\_seq\_sql.txt

use perl script: fill\_geneSQLtable\_WS296.pl to add additional data  
requires files   
WS296\_location\_sql.txt  
WS296\_description\_sql.txt  
WS296\_expression\_sql.txt  
WS296\_expression\_genomic\_study\_sql.txt  
WS296\_phenotype\_allele\_sql.txt  
WS296\_phenotype\_RNAi\_sql.txt  
WS296\_GO\_terms\_sql.txt  
WS296\_interacting\_genes\_sql.txt  
WS296\_disease\_association\_sql.txt  
WS296\_orthologs\_sql.txt

In all these files the first row contains the sql database column names.

**Database files**

WS296\_all\_IDs\_sql.txt

The file contains basic information about all proteins. Columns are:

|  |  |
| --- | --- |
| **column name in sql database** | **content** |
| geneWBID | Wormbase Gene ID (WBGene00006745) |
| CDS | name of splice variant (B0273.4a) |
| gene | 'sequence' name of the gene (B0273.4) |
| CGC | 'official' gene name (unc-5) |
| wormpepID | wormpepID (CE16790) |
| protsize | number of amino acids in the protein (919) |
| uniprot | Uniprot ID, see https://www.uniprot.org - as of May 2025 |
| longest | is this the longest splice variant (yes/no)? |

Sample entries

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **geneWBID** | **CDS** | **gene** | **CGC** | **wormpepID** | **protsize** | **uniprot** | **longest** |
| WBGene00007063 | 2L52.1a | 2L52.1 | 2L52.1 | CE32090 | 427 | A4F336 | yes |
| WBGene00007063 | 2L52.1b | 2L52.1 | 2L52.1 | CE50569 | 220 | A0A0K3AWR5 | no |
| WBGene00006745 | B0273.4a | B0273.4 | unc-5 | CE16790 | 919 | Q26261 | no |
| WBGene00006745 | B0273.4b | B0273.4 | unc-5 | CE37693 | 567 | Q26261 | no |
| WBGene00006745 | B0273.4c | B0273.4 | unc-5 | CE16791 | 947 | Q26261 | yes |
| WBGene00006745 | B0273.4d | B0273.4 | unc-5 | CE49241 | 811 | Q26261 | no |
| WBGene00006745 | B0273.4e | B0273.4 | unc-5 | CE49455 | 752 | Q26261 | no |
| WBGene00006745 | B0273.4f | B0273.4 | unc-5 | CE49300 | 657 | Q26261 | no |

WS296\_protein\_domains\_seq\_sql.txt

The file contains information about the domain organization of proteins. Columns are:

|  |  |
| --- | --- |
| **column name in sql database** | **content** |
| gene | 'sequence' name of the gene (2L52.1) |
| CDS | name of splice variant (2L52.1a) |
| domains | all domains found in the protein |
| pattern | the 'arrangement' of domain from N to C terminus Znf,17,45 means a Znf domain between amino acids 17 and 45 51,46,96 means no domain between amino acids 46 and 96 |
| sequence | the protein sequence |

Sample entries

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **gene** | **CDS** | **domains** | **pattern** | **sequence** |
| 2L52.1 | 2L52.1a | .SM00355.  .Znf. | 16,1,16  Znf,17,45  51,46,96  Znf,97,122  305,123,427 | MSMVRNVSNQSEKLEILSCKWVGCLKSTEVFKTVEKLLDHVTADHIPEVIVNDDGSEEVVCQWDCCEMGASRGNLQKKKEWMENHFKTRHVRKAKIFKCLIEDCPVVKSSSQEIETHLRISHPINPKKERLKEFKSSTDHIEPTQANRVWTIVNGEVQWKTPPRVKKKTVIYYDDGPRYVFPTGCARCNYDSDESELESDEFWSATEMSDNEEVYVNFRGMNCISTGKSASMVPSKRRNWPKRVKKRLSTQRNNQKTIRPPELNKNNIEIKDMNSNNLEERNREECIQPVSVEKNILHFEKFKSNQICIVRENNKFREGTRRRRKNSGESEDLKIHENFTEKRRPIRSCKQNISFYEMDGDIEEFEVFFDTPTKSKKVLLDIYSAKKMPKIEVEDSLVNKFHSKRPSRACRVLGSMEEVPFDVEIGY |
| 2L52.1 | 2L52.1b | .no\_domain. | 220,1,220 | MSDNEEVYVNFRGMNCISTGKSASMVPSKRRNWPKRVKKRLSTQRNNQKTIRPPELNKNNIEIKDMNSNNLEERNREECIQPVSVEKNILHFEKFKSNQICIVRENNKFREGTRRRRKNSGESEDLKIHENFTEKRRPIRSCKQNISFYEMDGDIEEFEVFFDTPTKSKKVLLDIYSAKKMPKIEVEDSLVNKFHSKRPSRACRVLGSMEEVPFDVEIGY |

**Build mutation database**

* within mysql use "source create\_gexplore\_db\_WS296\_mutations.sql" to create the database and load data  
  requires files  
  WS296\_protTable\_mutDB\_SQL.txt  
  WS296\_mutation\_table\_SQL.txt

**Database files**

WS296\_protTable\_mutDB\_SQL.txt

The file contains basic information about proteins including the domain organization. Columns are:

|  |  |
| --- | --- |
| **column name in sql database** | **content** |
| protein | name of splice variant (B0273.4a) |
| gene | 'sequence' name of the gene (B0273.4) |
| locus | 'official' gene name (unc-5) |
| UniID | Uniprot ID, see https://www.uniprot.org - as of May 2025 |
| longest | is this the longest splice variant (yes/no)? |
| psize | number of amino acids in the protein (427) |
| domains | all domains found in the protein |
| pattern | the 'arrangement' of domain from N to C terminus Znf,17,45 means a Znf domain between amino acids 17 and 45 51,46,96 means no domain between amino acids 46 and 96 |
| psequence | the protein sequence |

Sample entries

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **protein** | **gene** | **locus** | **UniID** | **longest** | **psize** | **domains** | **pattern** | **psequence** |
| 2L52.1a | 2L52.1 | 2L52.1 | A4F336 | yes | 427 | .SM00355..Znf. | 16,1,16 Znf,17,45 51,46,96 Znf,97,122 305,123,427 | amino acid sequence |
| 2L52.1b | 2L52.1 | 2L52.1 | A0A0K3AWR5 | no | 220 | .no\_domain. | 220,1,220 | amino acid sequence |

WS296\_mutation\_table\_SQL.txt

The file contains basic information about the individual mutations. Columns are:

|  |  |
| --- | --- |
| **column name in sql database** | **content** |
| no | record number |
| protein | name of splice variant (B0273.4a) |
| gene | 'sequence' name of the gene (B0273.4) |
| locus | 'official' gene name (unc-5) |
| psize | number of amino acids in the protein (427) |
| longest | is this the longest splice variant (yes/no)? |
| allele | allele name (ev400) |
| chromosome | chromosome (I, II, etc) |
| nt\_start | nucleotide position of gene start |
| nt\_end | nucleotide position of gene end |
| type | type of mutation (missense, nonsense, deletion, etc) |
| source | e.g. "million\_mutation" or "KO\_consortium" |
| wt\_aa | wild type amino acid (missense or nonsense alleles only) |
| mut\_aa | mutant amino acid |
| aa\_start | amino acid position of mutation (or start of deletion) |
| aa\_end | amino acid position of mutation (or end of deletion) |
| single | does the mutation affect only one gene? |
| mut\_domain | protein domain affected by the mutation |
| domains | all domains found in the protein |
| pattern | the 'arrangement' of domain from N to C terminus Znf,17,45 means a Znf domain between amino acids 17 and 45 51,46,96 means no domain between amino acids 46 and 96 |

sample entries

1 2L52.1a 2L52.1 2L52.1 427 yes gk130648 II 2750 2750 missense million\_mutation single 83 83 E K .SM00355..Znf. 16,1,16 Znf,17,45 51,46,96 Znf,97,122 305,123,427

15 2L52.1a 2L52.1 2L52.1 427 yes tm417 II 3606 3981 int\_ex\_deletion KO\_consortium single 213 273 .SM00355..Znf. 16,1,16 Znf,17,45 51,46,96 Znf,97,122 305,123,427