Name: Thi Hanh Nguyen Ly (0772489)

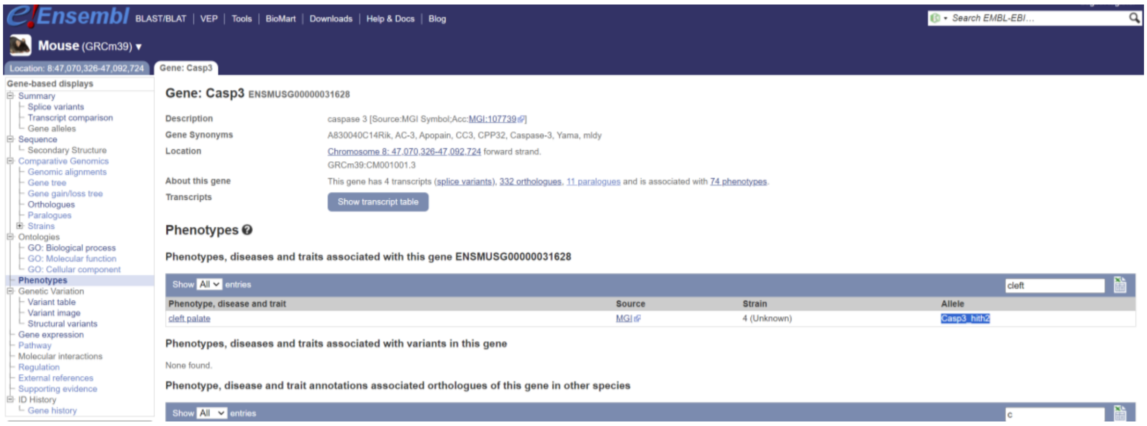
Practical Computing of Bioinformatics – Assignment 1

Question 1.

1. The IDs from **ID_list.txt** is from UniProt database of protein sequences.
2. The IDs are derived from Caspase-3 protein sequence from **CASP3** gene of human (**Homo Sapiens**) lying on chromosome 4.
3. The protein resides in the cytoplasm, nucleus, nucleoplasm, cytosol, plasma membrane of the cell.

   It associates mostly apoptosis processes. More specifically, the protein links to immune system process, autophagy, cell adhesion, programmed cell death, reproductive process, signalling, cell differentiation, protein catabolic process, anatomical structure development, nervous system process, protein maturation, …
4. Information for caspase 3 gene in mouse (Mus musculus) can be found under **EBI Search** in ebi.ac.uk website. Further information about this gene can be reviewed when clicking "View in Ensembl".

   Under "gene-based display" panel, when clicking to the option **Phenotypes,** a list of all phenotypes related to Casp3 gene of mouse can be found. Each phenotype will be associated with its corresponding allele. Thus, we can search for the **cleft palate** phenotype and its corresponding allele. In this case, allele **Casp3_hith2** can be found (Figure 1).



Gene: Casp3 ENSMUSG00000031628

Description	caspase 3 [Source:MGI Symbol;Acc:MGI:107739]
Gene Synonyms	A830040C14Rik, AC-3, Apopain, CC3, CPP32, Caspase-3, Yama, mldy
Location	Chromosome 8: 47,070,326-47,092,724 forward strand. GRCm39:CM001001.3
About this gene	This gene has 4 transcripts (splice variants), 332 orthologues, 11 paralogues and is associated with 74 phenotypes.
Transcripts	

Phenotypes

Phenotypes, diseases and traits associated with this gene ENSMUSG00000031628

Phenotype, disease and trait	Source	Strain	Allele
cleft palate	MGI	4 (Unknown)	Casp3_hith2

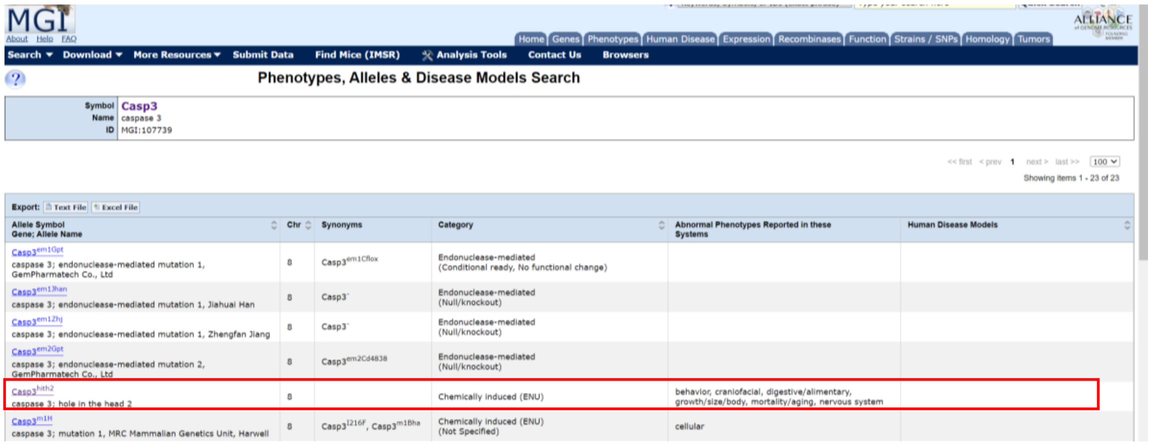
Phenotypes, diseases and traits associated with variants in this gene

None found.

Phenotype, disease and trait annotations associated orthologues of this gene in other species

Figure 1. Allele for cleft palate development of Casp3 gene in mouse.

Information in Figure 1 also includes a link to **MGI source**, an online database of mouse genome informatics where all mutations and alleles information can be found. When choosing the **Casp3^{hith2} allele** (Figure 2), an information of how the allele was introduced in the mouse strain can be found.


MGI
About Help FAQ
Home Genes Phenotypes Human Disease Expression Recombinases Function Strains / SNPs Homology Tumors
ALLIANCE of GENOME RESOURCES
Search Download More Resources Submit Data Find Mice (IMSR) Analysis Tools Contact Us Browsers
?
<< first < prev 1 next > last >>
100
Export: Text File Excel File

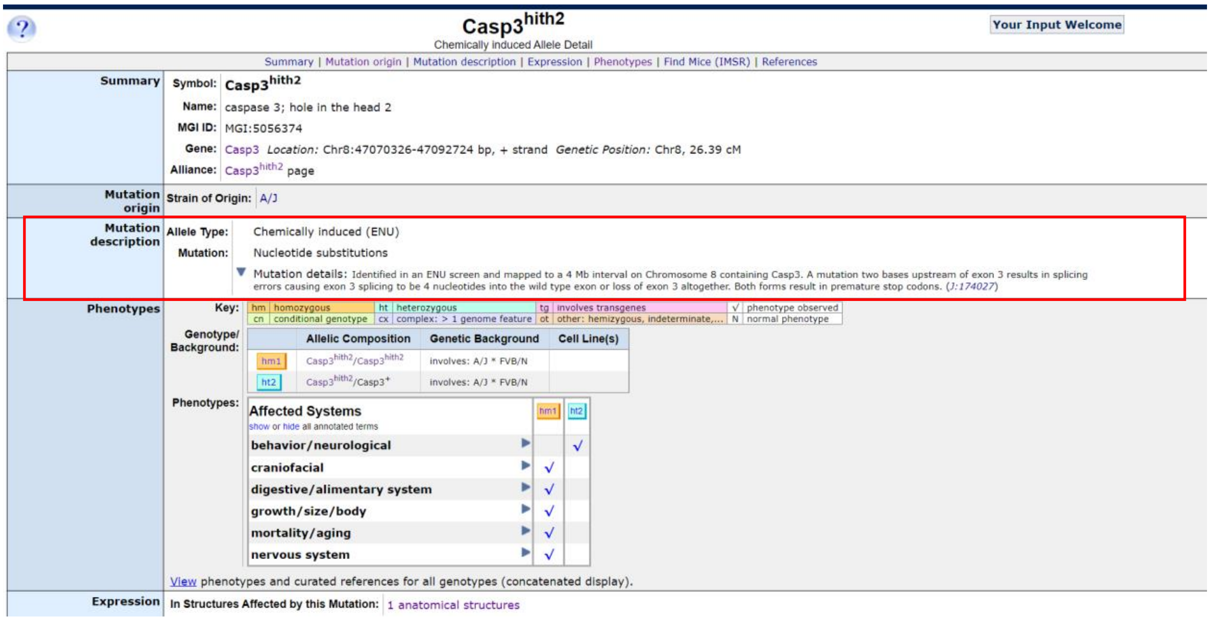
Phenotypes, Alleles & Disease Models Search

Symbol	Casp3
Name	caspase 3
ID	MGI:107739

Showing items 1 - 23 of 23

Allele Symbol Gene; Allele Name	Chr	Synonyms	Category	Abnormal Phenotypes Reported in these Systems	Human Disease Models
Casp3^{em1Gpt} caspase 3; endonuclease-mediated mutation 1, GemPharmatech Co., Ltd	8	Casp3^{em1Cflox}	Endonuclease-mediated (Conditional ready, No functional change)		
Casp3^{em1Jhan} caspase 3; endonuclease-mediated mutation 1, Jiahuai Han	8	Casp3^{-}	Endonuclease-mediated (Null/knockout)		
Casp3^{em1Zhj} caspase 3; endonuclease-mediated mutation 1, Zhengfan Jiang	8	Casp3^{-}	Endonuclease-mediated (Null/knockout)		
Casp3^{em2Gpt} caspase 3; endonuclease-mediated mutation 2, GemPharmatech Co., Ltd	8	Casp3^{em2Cd4838}	Endonuclease-mediated (Null/knockout)		
Casp3^{hith2} caspase 3; hole in the head 2	8		Chemically induced (ENU)	behavior, craniofacial, digestive/alimentary, growth/size/body, mortality/aging, nervous system	
Casp3^{m1H} caspase 3; mutation 1, MRC Mammalian Genetics Unit, Harwell	8	Casp3^{I216F}, Casp3^{m1Bha}	Chemically induced (ENU) (Not Specified)	cellular	

In figure 3, it is said in the mutation details that **Casp3^{hith2} allele** is “identified in an ENU screen and mapped to a 4 Mb interval on Chromosome 8 containing Casp3. A mutation two bases upstream of exon 3 results in splicing errors causing exon 3 splicing to be 4 nucleotides into the wild type exon or loss of exon 3 altogether. Both forms result in premature stop codons.”



Casp3^{hith2}

Chemically induced Allele Detail

Summary | Mutation origin | Mutation description | Expression | Phenotypes | Find Mice (IMSR) | References

Summary

Symbol: **Casp3^{hith2}**

Name: caspase 3; hole in the head 2

MGI ID: MGI:5056374

Gene: Casp3 *Location:* Chr8:47070326-47092724 bp, + strand *Genetic Position:* Chr8, 26.39 cM

Alliance: Casp3^{hith2} page

Mutation origin

Strain of Origin: A/J

Mutation description

Allele Type: Chemically induced (ENU)

Mutation: Nucleotide substitutions

▼ Mutation details: Identified in an ENU screen and mapped to a 4 Mb interval on Chromosome 8 containing Casp3. A mutation two bases upstream of exon 3 results in splicing errors causing exon 3 splicing to be 4 nucleotides into the wild type exon or loss of exon 3 altogether. Both forms result in premature stop codons. (*J:174027*)

Phenotypes

Key:

hm	homozygous	ht	heterozygous	tg	involves transgenes	√	phenotype observed
cn	conditional genotype	cx	complex: > 1 genome feature	ot	other: hemizygous, indeterminate,...	N	normal phenotype

Genotype/ Background:

	Allelic Composition	Genetic Background	Cell Line(s)
hm1	Casp3^{hith2}/Casp3^{hith2}	involves: A/J * FVB/N	
ht2	Casp3^{hith2}/Casp3^{+}	involves: A/J * FVB/N	

Phenotypes:

Affected Systems show or hide all annotated terms	hm1	ht2
behavior/neurological ▶		√
craniofacial ▶	√	
digestive/alimentary system ▶	√	
growth/size/body ▶	√	
mortality/aging ▶	√	
nervous system ▶	√	

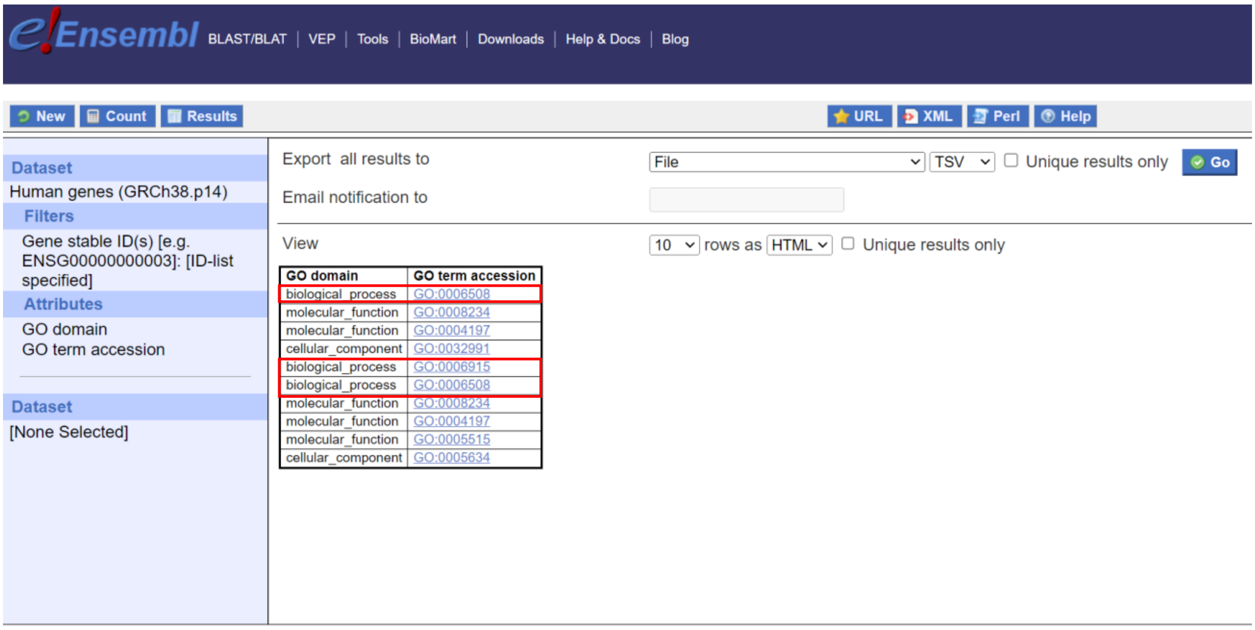
View phenotypes and curated references for all genotypes (concatenated display).

Expression

In Structures Affected by this Mutation: 1 anatomical structures

Figure 3. Mutation description of Casp3^{hith2} allele.

5. Retrieving GO term in BioMart for Casp3 gene of Human:
Casp3 gene of Human has reference gene id of ENSG00000164305 (Ensembl) (Figure 4)



e!Ensembl
BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog
New
Count
Results
URL
XML
Perl
Help
Go

Dataset
Human genes (GRCh38.p14)
Filters
Gene stable ID(s) [e.g. ENSG00000000003]: [ID-list specified]
Attributes
GO domain
GO term accession

Dataset
[None Selected]

Export all results to File TSV □ Unique results only

Email notification to

View 10 rows as HTML □ Unique results only

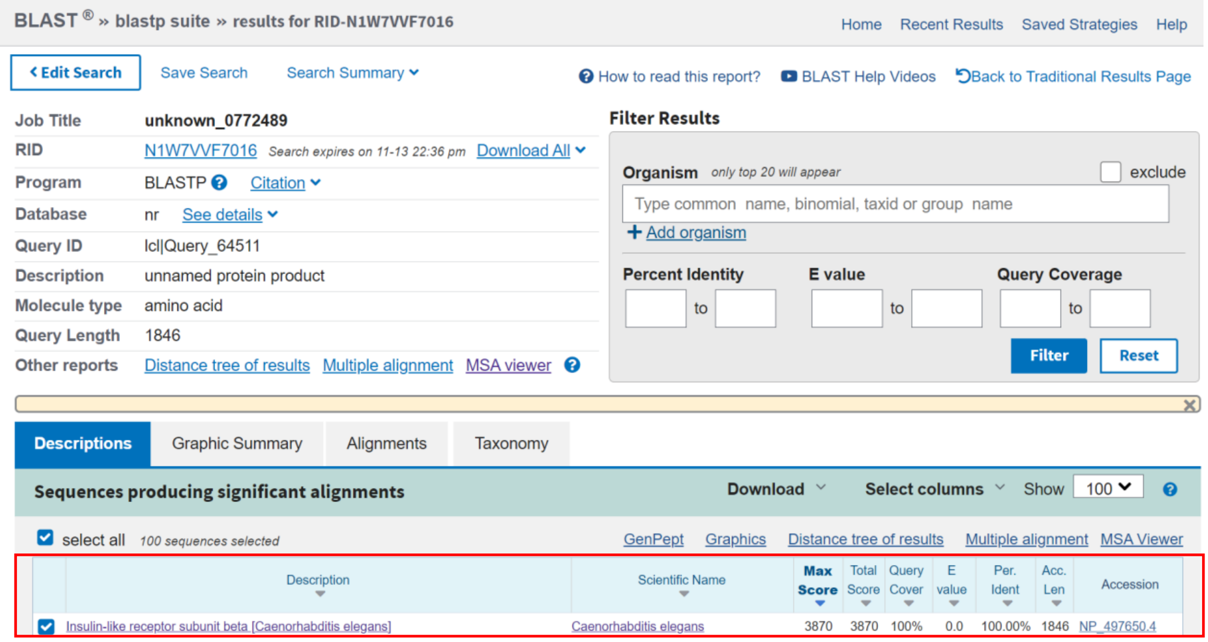
GO domain	GO term accession
biological_process	GO:0006508
molecular_function	GO:0008234
molecular_function	GO:0004197
cellular_component	GO:0032991
biological_process	GO:0006915
biological_process	GO:0006508
molecular_function	GO:0008234
molecular_function	GO:0004197
molecular_function	GO:0005515
cellular_component	GO:0005634

Figure 4. Retrieving GO-terms for ENSG00000164305

The reference gene id of Casp3 gene is input in the **Filters** part when searching in BioMart. GO-term accession and GO-domain are chosen for the output attributes. In this list, the GO-terms for GO-domain “biological process” are: GO:0006508, GO:0006915, GO:0006508.

Question 2.

1. BLAST algorithm from https://blast.ncbi.nlm.nih.gov/ (BLOSUM-62 matrix) is used to analyse the sequence in the file **unknown_sequence.fsa.** The result is shown to support that the sequence belongs to the **Insulin-like receptor protein** of **Caenorhabditis elegans** (Roundworm species) (Figure 5).



BLAST ® » blastp suite » results for RID-N1W7VVF7016
Home Recent Results Saved Strategies Help
< Edit Search Save Search Search Summary
How to read this report? BLAST Help Videos Back to Traditional Results Page

Job Title unknown_0772489
RID N1W7VVF7016 *Search expires on 11-13 22:36 pm* Download All
Program BLASTP Citation
Database nr See details
Query ID lcl|Query_64511
Description unnamed protein product
Molecule type amino acid
Query Length 1846
Other reports Distance tree of results Multiple alignment MSA viewer

Filter Results

Organism *only top 20 will appear* □ exclude
Type common name, binomial, taxid or group name
+ Add organism
Percent Identity to E value to Query Coverage to
Filter Reset

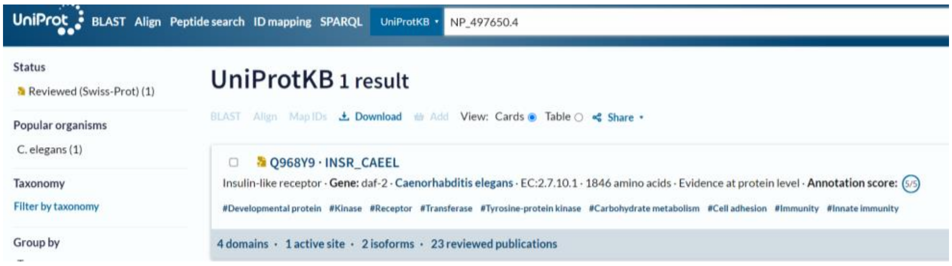
Sequences producing significant alignments Download Select columns Show 100

☑ select all *100 sequences selected* GenPept Graphics Distance tree of results Multiple alignment MSA Viewer

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Insulin-like receptor subunit beta [Caenorhabditis elegans]	Caenorhabditis elegans	3870	3870	100%	0.0	100.00%	1846	NP_497650.4

Figure 5. BLAST analysis of unknown_sequence.fsa

2. When searching for the **Accession** “NP_497650.4” in UniProtKB, the result for this protein sequence is Q968Y9 - INSR_CAEEL. It is shown that this protein sequence is derived from gene **daf-2 (**Figure 6**)**



UniProt BLAST Align Peptide search ID mapping SPARQL UniProtKB
NP_497650.4
Status
Reviewed (Swiss-Prot) (1)
Popular organisms
C. elegans (1)
Taxonomy
Filter by taxonomy
Group by
BLAST Align Map IDs Download Add View: Cards ● Table ○ Share

UniProtKB 1 result

□ Q968Y9 · INSR_CAEEL
Insulin-like receptor · **Gene:** daf-2 · **Caenorhabditis elegans** · EC:2.7.10.1 · 1846 amino acids · Evidence at protein level · **Annotation score:** 5/5
#Developmental protein #Kinase #Receptor #Transferase #Tyrosine-protein kinase #Carbohydrate metabolism #Cell adhesion #Immunity #Innate immunity
4 domains · 1 active site · 2 isoforms · 23 reviewed publications

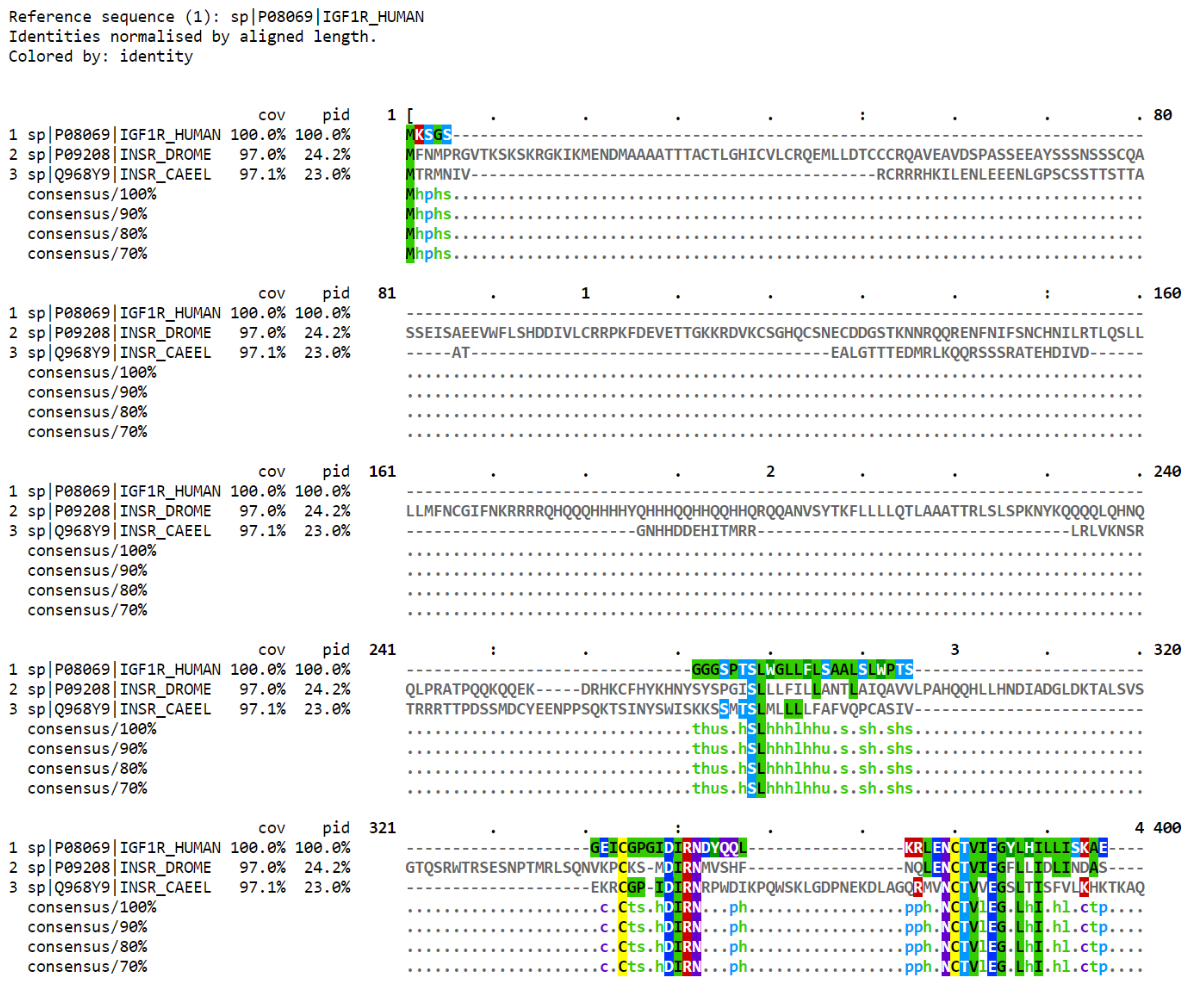
Figure 6. Genomic information of NP_497650.4

Searching for Homologous genes in NCBI for “Caenorhabditis elegans daf-2” (HomoloGene database) results in **IGF1R (insulin-like growth factor 1 receptor) gene** for Homo Sapiens (Human) and **InR (insulin-like receptor) gene** for D. melanogaster (Fruit fly).

3. The three protein sequences extracted from UniProt are aligned with **t-coffee algorithm** in ebi.ac.uk website. The result is then viewed in MView (Figure 7)

t-coffee result: https://www.ebi.ac.uk/Tools/services/web/toolresult.ebi?jobId=tcoffee-I20231112-172932-0756-99822783-p1m&analysis=alignments

MView result: https://www.ebi.ac.uk/Tools/services/rest/mview/result/mview-I20231112-173008-0791-33417150-p1m/aln-html



```
Reference sequence (1): sp|P08069|IGF1R_HUMAN
Identities normalised by aligned length.
Colored by: identity

                           cov    pid   1 [        .         .         .         .         :         .         .         . 80
1 sp|P08069|IGF1R_HUMAN 100.0% 100.0%     MKSGS---------------------------------------------------------------------------
2 sp|P09208|INSR_DROME   97.0%  24.2%     MFNMPRGVTKSKSKRGKIKMENDMAAAATTTACTLGHICVLCRQEMLLDTCCCRQAVEAVDSPASSEEAYSSSNSSSCQA
3 sp|Q968Y9|INSR_CAEEL   97.1%  23.0%     MTRMNIV--------------------------------------------RCRRRHKILENLEEENLGPSCSSTTSTTA
  consensus/100%                          Mhphs...........................................................................
  consensus/90%                           Mhphs...........................................................................
  consensus/80%                           Mhphs...........................................................................
  consensus/70%                           Mhphs...........................................................................

                           cov    pid  81          .         1         .         .         .         .         :         . 160
1 sp|P08069|IGF1R_HUMAN 100.0% 100.0%     --------------------------------------------------------------------------------
2 sp|P09208|INSR_DROME   97.0%  24.2%     SSEISAEEVWFLSHDDIVLCRRPKFDEVETTGKKRDVKCSGHQCSNECDDGSTKNNRQQRENFNIFSNCHNILRTLQSLL
3 sp|Q968Y9|INSR_CAEEL   97.1%  23.0%     -----AT---------------------------------------EALGTTTEDMRLKQQRSSSRATEHDIVD------
  consensus/100%                          ................................................................................
  consensus/90%                           ................................................................................
  consensus/80%                           ................................................................................
  consensus/70%                           ................................................................................

                           cov    pid 161          .         .         .         2         .         .         .         . 240
1 sp|P08069|IGF1R_HUMAN 100.0% 100.0%     --------------------------------------------------------------------------------
2 sp|P09208|INSR_DROME   97.0%  24.2%     LLMFNCGIFNKRRRRQHQQQHHHHYQHHHQQHHQQHHQRQQANVSYTKFLLLLQTLAAATTRLSLSPKNYKQQQQLQHNQ
3 sp|Q968Y9|INSR_CAEEL   97.1%  23.0%     -------------------------GNHHDDEHITMRR----------------------------------LRLVKNSR
  consensus/100%                          ................................................................................
  consensus/90%                           ................................................................................
  consensus/80%                           ................................................................................
  consensus/70%                           ................................................................................

                           cov    pid 241          :         .         .         .         .         3         .         . 320
1 sp|P08069|IGF1R_HUMAN 100.0% 100.0%     -------------------------------GGGSPTSLWGLLFLSAALSLWPTS-------------------------
2 sp|P09208|INSR_DROME   97.0%  24.2%     QLPRATPQQKQQEK-----DRHKCFHYKHNYSYSPGISLLLFILLANTLAIQAVVLPAHQQHLLHNDIADGLDKTALSVS
3 sp|Q968Y9|INSR_CAEEL   97.1%  23.0%     TRRRTTPDSSMDCYEENPPSQKTSINYSWISKKSSMTSLMLLLLFAFVQPCASIV-------------------------
  consensus/100%                          ...............................thus.hSLhhhlhhu.s.sh.shs.........................
  consensus/90%                           ...............................thus.hSLhhhlhhu.s.sh.shs.........................
  consensus/80%                           ...............................thus.hSLhhhlhhu.s.sh.shs.........................
  consensus/70%                           ...............................thus.hSLhhhlhhu.s.sh.shs.........................

                           cov    pid 321          .         .         :         .         .         .         .         4 400
1 sp|P08069|IGF1R_HUMAN 100.0% 100.0%     --------------------GEICGPGIDIRNDYQQL-----------------KRLENCTVIEGYLHILLISKAE----
2 sp|P09208|INSR_DROME   97.0%  24.2%     GTQSRWTRSESNPTMRLSQNVKPCKS-MDIRNMVSHF-----------------NQLENCTVIEGFLLIDLINDAS----
3 sp|Q968Y9|INSR_CAEEL   97.1%  23.0%     --------------------EKRCGP-IDIRNRPWDIKPQWSKLGDPNEKDLAGQRMVNCTVVEGSLTISFVLKHKTKAQ
  consensus/100%                          .....................c.Cts.hDIRN...ph.................pph.NCTVlEG.LhI.hl.ctp....
  consensus/90%                           .....................c.Cts.hDIRN...ph.................pph.NCTVlEG.LhI.hl.ctp....
  consensus/80%                           .....................c.Cts.hDIRN...ph.................pph.NCTVlEG.LhI.hl.ctp....
  consensus/70%                           .....................c.Cts.hDIRN...ph.................pph.NCTVlEG.LhI.hl.ctp....
```

Figure 7. Partial MView result from multiple alignment.

According to Figure 7, the conservation percentage reflects the degree of homology in the three protein sequences. Thus, if we take the sequence of human as a reference, then fruit fly and roundworm have around 97% similarity in their protein sequence.

Question 3 - Linux

The result from Linux assignment is store in Linux folder. The folder contains analyze_GFF_features.sh, analyzeGFF_job.slurm, the two error, output log files of the slurm job and the gff3 file of chromosome 1. There is also another folder call “text files result” containing the text file output from running the code of the bash file.