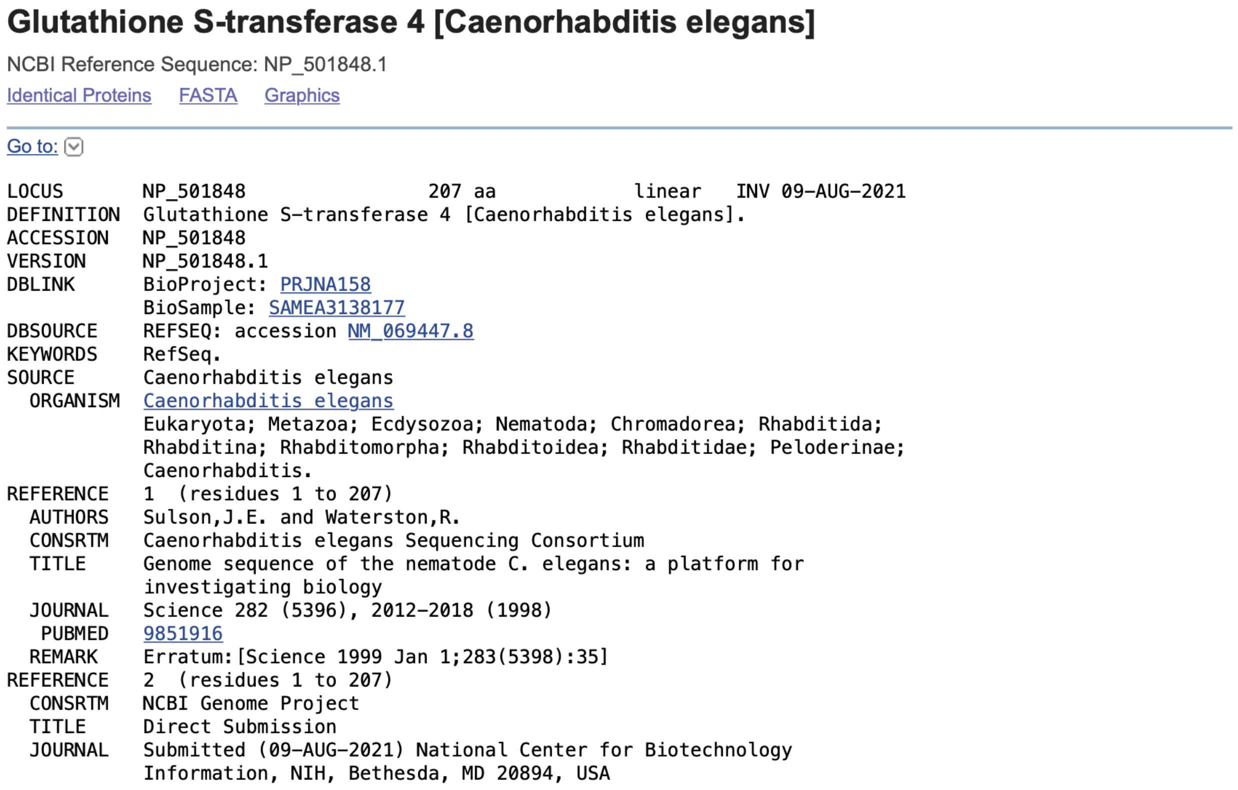
San Luc
PID: A50910657

Questions:

[Q1] Tell me the name of a protein you are interested in. Include the species and the accession number. This can be a human protein or a protein from any other species as long as it's function is known.



Glutathione S-transferase 4 [Caenorhabditis elegans]

NCBI Reference Sequence: NP_501848.1

Identical Proteins FASTA Graphics

Go to:

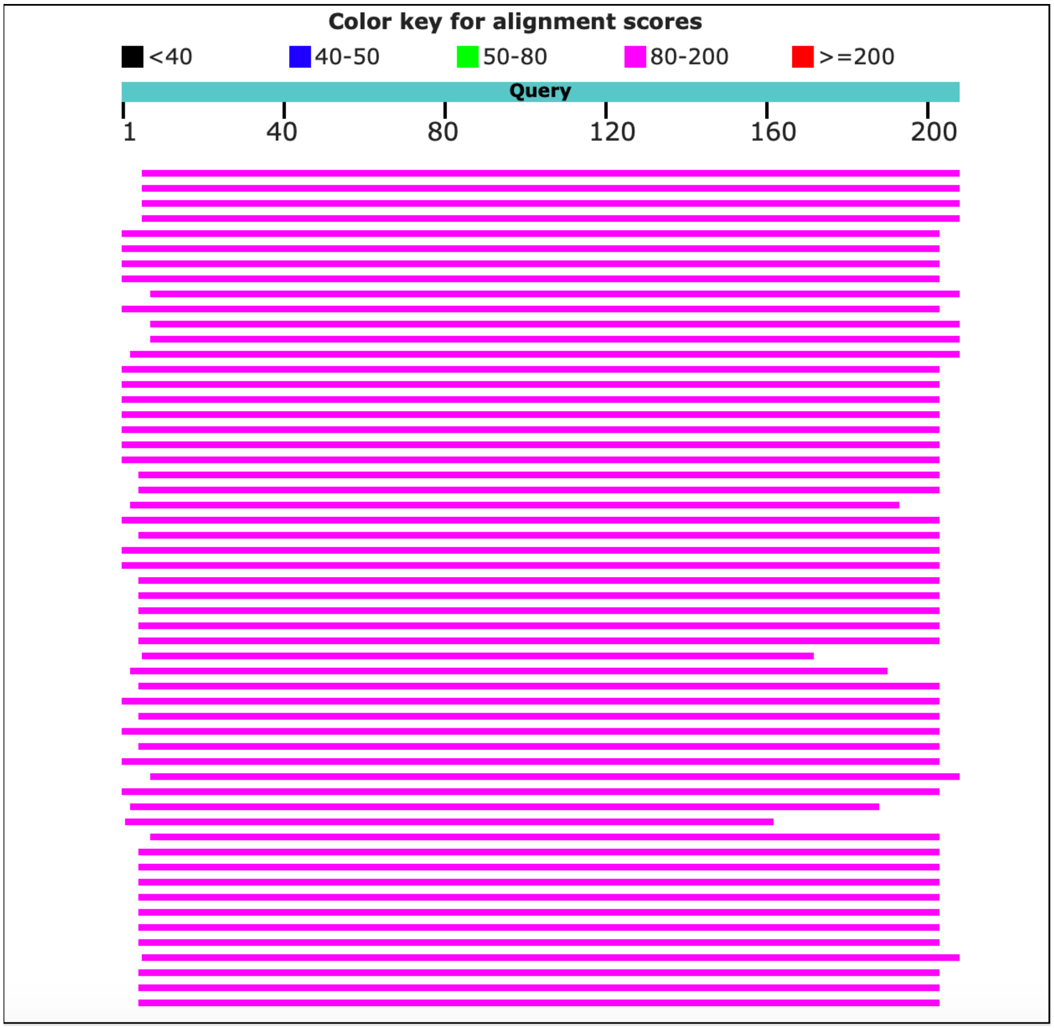
```
LOCUS       NP_501848                207 aa            linear   INV 09-AUG-2021
DEFINITION  Glutathione S-transferase 4 [Caenorhabditis elegans].
ACCESSION   NP_501848
VERSION     NP_501848.1
DBLINK      BioProject: PRJNA158
            BioSample: SAMEA3138177
DBSOURCE    REFSEQ: accession NM_069447.8
KEYWORDS    RefSeq.
SOURCE      Caenorhabditis elegans
  ORGANISM  Caenorhabditis elegans
            Eukaryota; Metazoa; Ecdysozoa; Nematoda; Chromadorea; Rhabditida;
            Rhabditina; Rhabditomorpha; Rhabditoidea; Rhabditidae; Peloderinae;
            Caenorhabditis.
REFERENCE   1  (residues 1 to 207)
  AUTHORS   Sulson,J.E. and Waterston,R.
  CONSRTM   Caenorhabditis elegans Sequencing Consortium
  TITLE     Genome sequence of the nematode C. elegans: a platform for
            investigating biology
  JOURNAL   Science 282 (5396), 2012-2018 (1998)
   PUBMED   9851916
  REMARK    Erratum:[Science 1999 Jan 1;283(5398):35]
REFERENCE   2  (residues 1 to 207)
  CONSRTM   NCBI Genome Project
  TITLE     Direct Submission
  JOURNAL   Submitted (09-AUG-2021) National Center for Biotechnology
            Information, NIH, Bethesda, MD 20894, USA
```

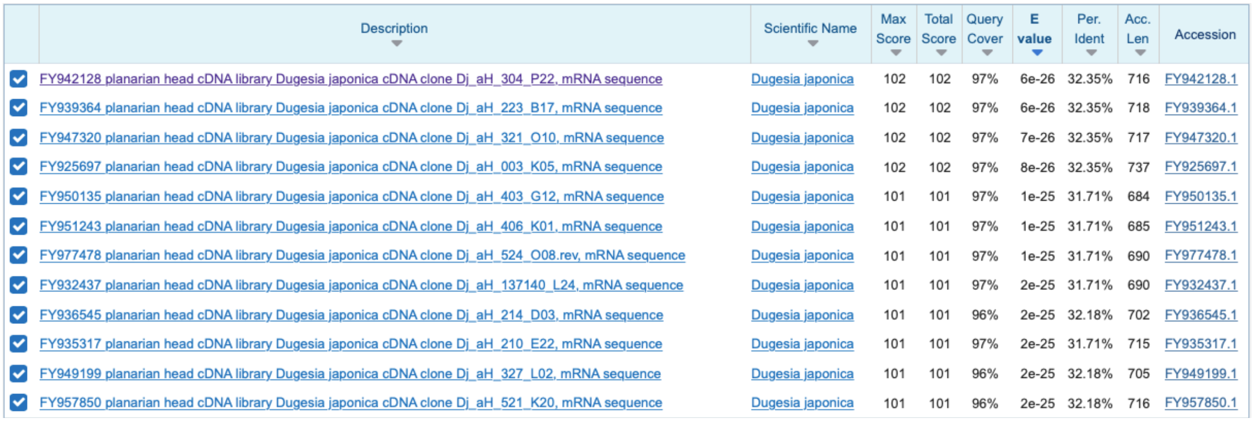
Name: Glutathione S-Transferase
Species: *C. elegans*
Accession:NP_501848.1 (protein), NM_069447.8 (mRNA)
Function: this protein enables glutathione transferase activity. It is involved in the glutathione metabolic process.

[Q2] Perform a BLAST search against a DNA database, such as a database consisting of genomic DNA or ESTs. The BLAST server can be at NCBI or elsewhere. Include details of the BLAST method used, database searched, and any limits applied (e.g. Organism).

Method: TBLASTN (2.7.1) search against flatworms ESTs

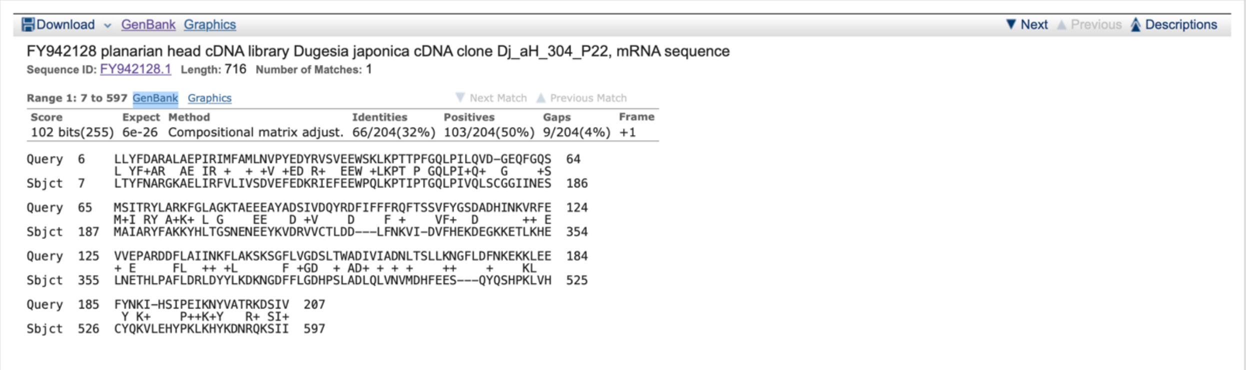
Database: Expressed Sequence Tags (est)
Organism: flatworms (taxid:6157)





	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	FY942128 planarian head cDNA library Dugesia japonica cDNA clone Dj_aH_304_P22, mRNA sequence	Dugesia japonica	102	102	97%	6e-26	32.35%	716	FY942128.1
☑	FY939364 planarian head cDNA library Dugesia japonica cDNA clone Dj_aH_223_B17, mRNA sequence	Dugesia japonica	102	102	97%	6e-26	32.35%	718	FY939364.1
☑	FY947320 planarian head cDNA library Dugesia japonica cDNA clone Dj_aH_321_O10, mRNA sequence	Dugesia japonica	102	102	97%	7e-26	32.35%	717	FY947320.1
☑	FY925697 planarian head cDNA library Dugesia japonica cDNA clone Dj_aH_003_K05, mRNA sequence	Dugesia japonica	102	102	97%	8e-26	32.35%	737	FY925697.1
☑	FY950135 planarian head cDNA library Dugesia japonica cDNA clone Dj_aH_403_G12, mRNA sequence	Dugesia japonica	101	101	97%	1e-25	31.71%	684	FY950135.1
☑	FY951243 planarian head cDNA library Dugesia japonica cDNA clone Dj_aH_406_K01, mRNA sequence	Dugesia japonica	101	101	97%	1e-25	31.71%	685	FY951243.1
☑	FY977478 planarian head cDNA library Dugesia japonica cDNA clone Dj_aH_524_O08.rev, mRNA sequence	Dugesia japonica	101	101	97%	1e-25	31.71%	690	FY977478.1
☑	FY932437 planarian head cDNA library Dugesia japonica cDNA clone Dj_aH_137140_L24, mRNA sequence	Dugesia japonica	101	101	97%	2e-25	31.71%	690	FY932437.1
☑	FY936545 planarian head cDNA library Dugesia japonica cDNA clone Dj_aH_214_D03, mRNA sequence	Dugesia japonica	101	101	96%	2e-25	32.18%	702	FY936545.1
☑	FY935317 planarian head cDNA library Dugesia japonica cDNA clone Dj_aH_210_E22, mRNA sequence	Dugesia japonica	101	101	97%	2e-25	31.71%	715	FY935317.1
☑	FY949199 planarian head cDNA library Dugesia japonica cDNA clone Dj_aH_327_L02, mRNA sequence	Dugesia japonica	101	101	96%	2e-25	32.18%	705	FY949199.1
☑	FY957850 planarian head cDNA library Dugesia japonica cDNA clone Dj_aH_521_K20, mRNA sequence	Dugesia japonica	101	101	96%	2e-25	32.18%	716	FY957850.1

Chosen sequence: FY942128 planarian head cDNA library Dugesia japonica cDNA clone Dj_aH_304_P22, mRNA sequence.



FY942128 planarian head cDNA library Dugesia japonica cDNA clone Dj_aH_304_P22, mRNA sequence

Sequence ID: FY942128.1 Length: 716 Number of Matches: 1

Range 1: 7 to 597 GenBank Graphics

Score	Expect	Method	Identities	Positives	Gaps	Frame
102 bits(255)	6e-26	Compositional matrix adjust.	66/204(32%)	103/204(50%)	9/204(4%)	+1

```
Query  6    LLYFDARALAEPIRIMFAMLNVPYEDYRVSVEEWSKLKPTTPFGQLPILQVD-GEQFGQS  64
            L YF+AR  AE IR +  + +V +ED R+  EEW +LKPT P GQLPI+Q+  G    +S
Sbjct  7    LTYFNARGKAELIRFVLIVSDVEFEDKRIEFEEWPQLKPTIPTGQLPIVQLSCGGIINES  186

Query  65   MSITRYLARKFGLAGKTAEEEAYADSIVDQYRDFIFFFRQFTSSVFYGSDADHINKVRFE  124
            M+I RY A+K+ L G    EE   D +V    D    F +    VF+  D      ++ E
Sbjct  187  MAIARYFAKKYHLTGSNENEEYKVDRVVCTLDD---LFNKVI-DVFHEKDEGKKETLKHE  354

Query  125  VVEPARDDFLAIINKFLAKSKSGFLVGDSLTWADIVIADNLTSLLKNGFLDFNKEKKLEE  184
            + E     FL  ++ +L      F +GD  + AD+ + + +    ++    +    KL
Sbjct  355  LNETHLPAFLDRLDYYLKDKNGDFFLGDHPSLADLQLVNVMDHFEES---QYQSHPKLVH  525

Query  185  FYNKI-HSIPEIKNYVATRKDSIV  207
             Y K+    P++K+Y   R+ SI+
Sbjct  526  CYQKVLEHYPKLKHYKDNRQKSII  597
```

Score	Expect	Identities	Positives	Gaps
102 bits(255)	6e-26	66/204(32%)	103/204(50%)	9/204(4%)

```
Query  6    LLYFDARALAEPIRIMFAMLNVPYEDYRVSVEEWSKLKPTTPFGQLPILQVD-GEQFGQS  64
            L YF+AR  AE IR +  + +V +ED R+  EEW +LKPT P GQLPI+Q+  G    +S
Sbjct  7    LTYFNARGKAELIRFVLIVSDVEFEDKRIEFEEWPQLKPTIPTGQLPIVQLSCGGIINES  186

Query  65   MSITRYLARKFGLAGKTAEEEAYADSIVDQYRDFIFFFRQFTSSVFYGSDADHINKVRFE  124
            M+I RY A+K+ L G    EE   D +V    D    F +    VF+  D      ++ E
Sbjct  187  MAIARYFAKKYHLTGSNENEEYKVDRVVCTLDD---LFNKVI-DVFHEKDEGKKETLKHE  354

Query  125  VVEPARDDFLAIINKFLAKSKSGFLVGDSLTWADIVIADNLTSLLKNGFLDFNKEKKLEE  184
            + E     FL  ++ +L      F +GD  + AD+ + + +    ++    +    KL
Sbjct  355  LNETHLPAFLDRLDYYLKDKNGDFFLGDHPSLADLQLVNVMDHFEES---QYQSHPKLVH  525

Query  185  FYNKI-HSIPEIKNYVATRKDSIV  207
             Y K+    P++K+Y   R+ SI+
Sbjct  526  CYQKVLEHYPKLKHYKDNRQKSII  597
```

[Q3] Gather information about this “novel” protein. At a minimum, show me the protein sequence of the “novel” protein as displayed in your BLAST results from [Q2] as FASTA format (you can copy and paste the aligned sequence subject lines from your BLAST result page if necessary) or translate your novel DNA sequence using a tool called EMBOSS Transeq at the EBI. Don’t forget to translate all six reading frames; the ORF (open reading frame) is likely to be the longest sequence without a stop codon. It may not start with a methionine if you don’t have the complete coding region. Make sure the sequence you provide includes a header/subject line and is in traditional FASTA format.

Used EMBOSS transeq to translate the protein sequence above.

```
>FY942128.1_1 planarian head cDNA library Dugesia japonica cDNA clone Dj_aH_304_P22,
mRNA sequence
IILTYFNARGKAELIRFVLIVSDVEFEDKRIEFEEWPQLKPTIPTGQLPIVQLSCGGIIN
ESMAIARYFAKKYHLTGSNENEEYKVDRVVCTLDDLFNKVIDVFHEKDEGKKETLKHELN
ETHLPAFLDRLDYYLKDKNGDFFLGDHPSLADLQLVNVMDHFEESQYQSHPKLVHCYQKV
LEHYPKLKHYKDNRQKSII*KNSFTVSEYL*KLMKLF*LFQKLMIINLLLIVEKKKKKX

>FY942128.1_2 planarian head cDNA library Dugesia japonica cDNA clone Dj_aH_304_P22,
mRNA sequence
LY*HILMHEEKLN*FDLF*S*AMLNLKIKELNSKNGHN*NQQFQQVSCQLFNFLVEELSM
KAWQ*RDILQRNTI*PDRMKTKNIKLIELCVHSMICLIKLSTCSTRKMKGKRKH*NMN*M
KLICLHFLIDSITI*KIKMAISSSAIILHLLIYNW*MLWIILKNLNTRAIRN*YIVIKRY
WNIIQNSSITKIIGKNQ*SKKIHLLFQNIYKS**NCFNYFKN****IYY*SLKKKKKKX

>FY942128.1_3 planarian head cDNA library Dugesia japonica cDNA clone Dj_aH_304_P22,
mRNA sequence
YINIF*CTRKS*IDSICFDRKRC*I*R*KN*IRRMATIETNNSNRSVANCSTFLWRNYQ*
KHGNSEIFCKEIPFNRIE*KRRI*S*SSCVYTR*FV**SYRRVPRER*REKGNIKT*IK*
NSFACIS**TRLLFKR*KWRFLPRRSSFTC*FTIGKCYGSF*RISIPEPSEISTLLSKGI
GTLSKTQALQR*SAKINNLKKFIYCFRIFIKVDEIVLIISKINDNKFIINR*KKKKKK

>FY942128.1_4 planarian head cDNA library Dugesia japonica cDNA clone Dj_aH_304_P22,
mRNA sequence
FFFFFFND***IYYH*FLK*LKQFHQLL*IF*NSK*IFLDY*FLPIIFVMLEFWIMFQYL
LITMY*FRMALVLRFFKMIHNIYQL*ISK*RMIAEEEIAIFIF*IVIESIKKCRQMSFI*
FMF*CFLFPFIFLVEHVDNFIKQIIECTHNSINFIFFVFIRSG*MVFLCKISRYCHAFID
NSSTRKLNNWQLTCWNCWFQLWPFFEFNSFIFKFNIAYDQNKSNQFSFSSCIKIC*Y

>FY942128.1_5 planarian head cDNA library Dugesia japonica cDNA clone Dj_aH_304_P22,
mRNA sequence
FFFFFFQRLIINLLSLIFEIIKTISSTFINILKQ*MNFFRLLIFADYLCNA*VLDNVPIP
FDNNVLISDGSGIEILQNDP*HLPIVNQQVKDDRRGRNRHFYLLNSNRVYQEMQANEFHL
IHVLMFPFSLHLSRGTRR*LY*TNHRVYTQLDQLYILRFHSIRLNGISLQNISLLPCFH*
*FLHKKVEQLATDLLELLVSIVAILRIQFFYLQIQHRLRSKQIESIQLFLVH*NMLI*X

>FY942128.1_6 planarian head cDNA library Dugesia japonica cDNA clone Dj_aH_304_P22,
mRNA sequence
FFFFFFSTINNKFIIINF*NN*NNFINFYKYSETVNEFF*IIDFCRLSL*CLSFG*CSNT
F**QCTNFGWLWY*DSSK*SITFTNCKSASEG*SPRKKSPFLSFK**SSLSRNAGK*VSF
NSCFNVSFFPSSFSWNTSITLLNKSSSVHTTRSTLYSSFSFDPVKWYFFAKYLAIAMLSL
IIPPQES*TIGN*PVGIVGFNCGHSSNSILLSSNSTSLTIKTNRINSAFPRALKYVNIX
```

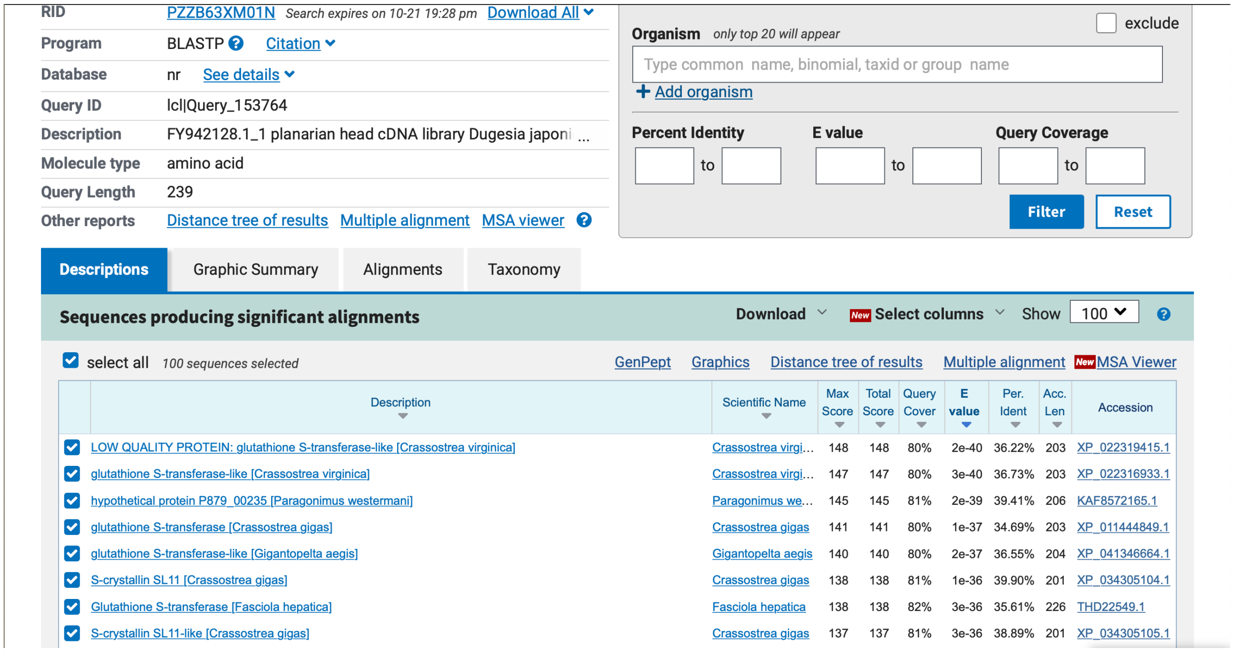
```
Eukaryota; Metazoa; Spiralia; Lophotrochozoa; Platyhelminthes;
             Rhabditophora; Seriata; Tricladida; Continenticola; Geoplanoidea;
             Dugesiidae; Dugesia.
```

Species: *Dugesia japonica*

[Q4] Prove that this gene, and its corresponding protein, are novel. For the purposes of this project, “novel” is defined as follows. Take the protein sequence (your answer to [Q3]), and use it as a query in a blastp search of the nr database at NCBI.

- If there is a match with 100% amino acid identity to a protein in the database, from the same species, then your protein is NOT novel (even if the match is to a protein with a name such as “unknown”). Someone has already found and annotated this sequence, and assigned it an accession number.

- If the top match reported has less than 100% identity, then it is likely that your protein is novel, and you have succeeded.
- If there is a match with 100% identity, but to a different species than the one you started with, then you have likely succeeded in finding a novel gene.
- If there are no database matches to the original query from [Q1], this indicates that you have partially succeeded: yes, you may have found a new gene, but no, it is not actually homologous to the original query. You should probably start over.

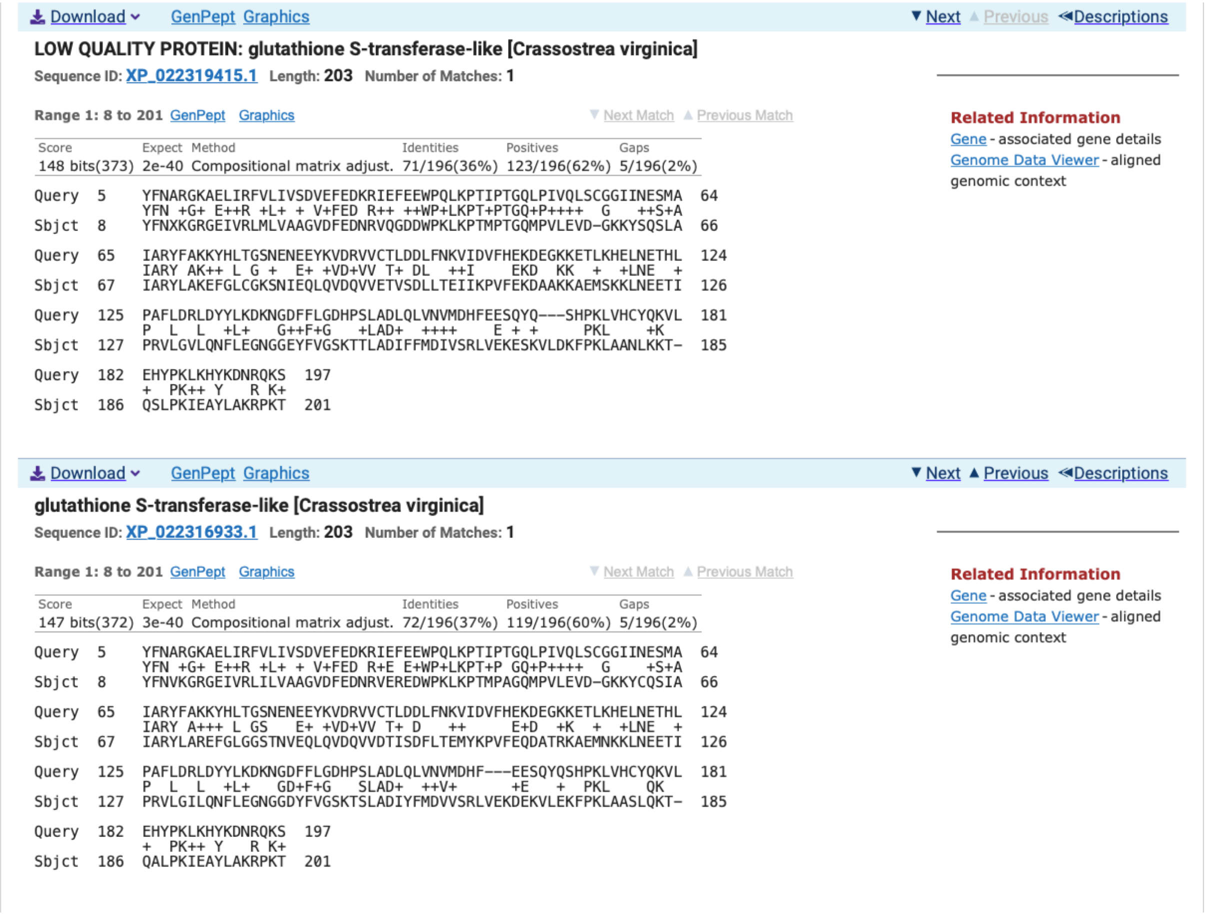


RID	PZZB63XM01N *Search expires on 10-21 19:28 pm* Download All
Program	BLASTP Citation
Database	nr See details
Query ID	lcl|Query_153764
Description	FY942128.1_1 planarian head cDNA library Dugesia japoni ...
Molecule type	amino acid
Query Length	239
Other reports	Distance tree of results Multiple alignment MSA viewer

Sequences producing significant alignments

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

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	LOW QUALITY PROTEIN: glutathione S-transferase-like [Crassostrea virginica]	Crassostrea virgi...	148	148	80%	2e-40	36.22%	203	XP_022319415.1
☑	glutathione S-transferase-like [Crassostrea virginica]	Crassostrea virgi...	147	147	80%	3e-40	36.73%	203	XP_022316933.1
☑	hypothetical protein P879_00235 [Paragonimus westermani]	Paragonimus we...	145	145	81%	2e-39	39.41%	206	KAF8572165.1
☑	glutathione S-transferase [Crassostrea gigas]	Crassostrea gigas	141	141	80%	1e-37	34.69%	203	XP_011444849.1
☑	glutathione S-transferase-like [Gigantopelta aegis]	Gigantopelta aegis	140	140	80%	2e-37	36.55%	204	XP_041346664.1
☑	S-crystallin SL11 [Crassostrea gigas]	Crassostrea gigas	138	138	81%	1e-36	39.90%	201	XP_034305104.1
☑	Glutathione S-transferase [Fasciola hepatica]	Fasciola hepatica	138	138	82%	3e-36	35.61%	226	THD22549.1
☑	S-crystallin SL11-like [Crassostrea gigas]	Crassostrea gigas	137	137	81%	3e-36	38.89%	201	XP_034305105.1



Download GenPept Graphics ▼Next ▲Previous ◀Descriptions

LOW QUALITY PROTEIN: glutathione S-transferase-like [Crassostrea virginica]

Sequence ID: XP_022319415.1 Length: 203 Number of Matches: 1

Range 1: 8 to 201 GenPept Graphics ▼ Next Match ▲ Previous Match

```
Score          Expect  Method                        Identities    Positives      Gaps
148 bits(373)  2e-40   Compositional matrix adjust.  71/196(36%)   123/196(62%)   5/196(2%)

Query  5    YFNARGKAELIRFVLIVSDVEFEDKRIEFEEWPQLKPTIPTGQLPIVQLSCGGIINESMA  64
            YFN +G+ E++R +L+ + V+FED R++ ++WP+LKPT+PTGQ+P++++  G   ++S+A
Sbjct  8    YFNXKGRGEIVRLMLVAAGVDFEDNRVQGDDWPKLKPTMPTGQMPVLEVD-GKKYSQSLA  66

Query  65   IARYFAKKYHLTGSNENEEYKVDRVVCTLDDLFNKVIDVFHEKDEGKKETLKHELNETHL  124
            IARY AK++ L G +  E+ +VD+VV T+ DL  ++I     EKD  KK  +  +LNE  +
Sbjct  67   IARYLAKEFGLCGKSNIEQLQVDQVVETVSDLLTEIIKPVFEKDAAKKAEMSKKLNEETI  126

Query  125  PAFLDRLDYYLKDKNGDFFLGDHPSLADLQLVNVMDHFEESQYQ---SHPKLVHCYQKVL  181
            P  L  L  +L+   G++F+G   +LAD+  ++++     E + +     PKL    +K
Sbjct  127  PRVLGVLQNFLEGNGGEYFVGSKTTLADIFFMDIVSRLVEKESKVLDKFPKLAANLKKT-  185

Query  182  EHYPKLKHYKDNRQKS  197
            +  PK++ Y   R K+
Sbjct  186  QSLPKIEAYLAKRPKT  201
```

Related Information
Gene - associated gene details
Genome Data Viewer - aligned genomic context

Download GenPept Graphics ▼Next ▲Previous ◀Descriptions

glutathione S-transferase-like [Crassostrea virginica]

Sequence ID: XP_022316933.1 Length: 203 Number of Matches: 1

Range 1: 8 to 201 GenPept Graphics ▼ Next Match ▲ Previous Match

```
Score          Expect  Method                        Identities    Positives      Gaps
147 bits(372)  3e-40   Compositional matrix adjust.  72/196(37%)   119/196(60%)   5/196(2%)

Query  5    YFNARGKAELIRFVLIVSDVEFEDKRIEFEEWPQLKPTIPTGQLPIVQLSCGGIINESMA  64
            YFN +G+ E++R +L+ + V+FED R+E E+WP+LKPT+P GQ+P++++  G    +S+A
Sbjct  8    YFNVKGRGEIVRLILVAAGVDFEDNRVEREDWPKLKPTMPAGQMPVLEVD-GKKYCQSIA  66

Query  65   IARYFAKKYHLTGSNENEEYKVDRVVCTLDDLFNKVIDVFHEKDEGKKETLKHELNETHL  124
            IARY A+++ L GS   E+ +VD+VV T+ D   ++      E+D  +K  +  +LNE  +
Sbjct  67   IARYLAREFGLGGSTNVEQLQVDQVVDTISDFLTEMYKPVFEQDATRKAEMNKKLNEETI  126

Query  125  PAFLDRLDYYLKDKNGDFFLGDHPSLADLQLVNVMDHF---EESQYQSHPKLVHCYQKVL  181
            P  L  L  +L+   GD+F+G   SLAD+  ++V+      +E   +  PKL    QK
Sbjct  127  PRVLGILQNFLEGNGGDYFVGSKTSLADIYFMDVVSRLVEKDEKVLEKFPKLAASLQKT-  185

Query  182  EHYPKLKHYKDNRQKS  197
            +  PK++ Y   R K+
Sbjct  186  QALPKIEAYLAKRPKT  201
```

Related Information
Gene - associated gene details
Genome Data Viewer - aligned genomic context

[Q5] Generate a multiple sequence alignment with your novel protein, your original query protein, and a group of other members of this family from different species. A typical number of proteins to use in a multiple sequence alignment for this assignment purpose is a minimum of 5 and a maximum of 20 - although the exact number is up to you. Include the multiple sequence alignment in your report. Use Courier font with a size appropriate to fit page width.

Side-note: Indicate your sequence in the alignment by choosing an appropriate name for each sequence in the input unaligned sequence file (i.e. edit the sequence file so that the species, or short common, names (rather than accession numbers) display in the output alignment and in the subsequent answers below). The goal in this step is to create an interesting an alignment for building a phylogenetic tree that illustrates species divergence.

```
>Dugesia japonica cDNA clone
IILTYFNARGKAELIRFVLIVSDVEFEDKRIEFEEWPQLKPTIPTGQLPIVQLSCGGIIN
ESMAIARYFAKKYHLTGSNENEEYKVDRVVCTLDDLFNKVIDVFHEKDEGKKETLKHELN
ETHLPAFLDRLDYYLKDKNGDFFLGDHPSLADLQLVNVMDHFEESQYQSHPKLVHCYQKV
LEHYPKLKHYKDNRQKSII*KNSFTVSEYL*KLMKLF*LFQKLMIINLLLIVEKKKKKX

>Glutathione S-transferase 4 [Caenorhabditis elegans]
MPNYKLLYFDARALAEPIRIMFAMLNVPYEDYRVSVEEWSKLKPTTPFGQLPILQVDGEQFGQSMSITRY
LARKFGLAGKTAEEEAYADSIVDQYRDFIFFFRQFTSSVFYGSDADHINKVRFEVVEPARDDFLAIINKF
```

LAKSKSGFLVGDSLTWADIVIADNLTSLLKNGFLDFNKEKKLEEFYNKIHSIPEIKNYVATRKDSIV

> glutathione S transferase-1 [Schmidtea mediterranea]
MSTVKVTYFDARGRAELIRLVLKASKIEFEDVRITKDKWPEVKPTTPTGKLPVVEYEGKQLTQSMAIARV
VARKHGFMGEDDKEYYLVERAIGQMVDVLEGLYKIYFAPEEKKEELRAEYVATSGRDNLKALEGFIKETG
FFAGEKITLAELFFLVVSDYLVKLPQLYDDFPKLKELRERILKANTDVEEWVNTRPVTEM

> glutathione S-transferase-like [Crassostrea virginica]
MTKYTVHYFNVKGRGEIVRLILVAAGVDFEDNRVEREDWPKLKPTMPAGQMPVLEVDGKKYCQSIAIARY
LAREFGLGGSTNVEQLQVDQVVDTISDFLTEMYKPVFEQDATRKAEMNKKLNEETIPRVLGILQNFLEGN
GGDYFVGSKTSLADIYFMDVVSRLVEKDEKVLEKFPKLAASLQKTQALPKIEAYLAKRPKTEL

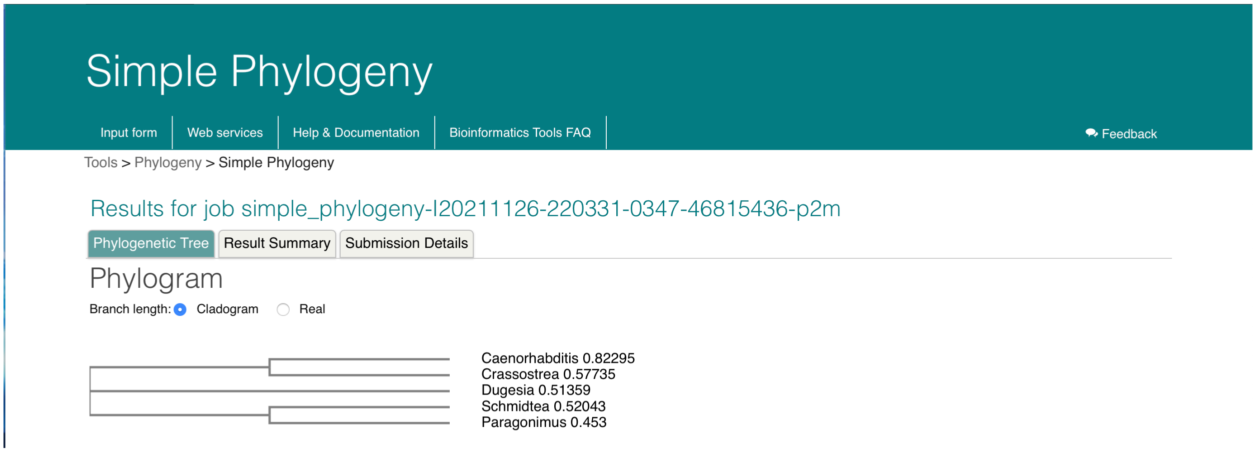
>hypothetical protein P879_00235 [Paragonimus westermani]
LTYFNGRGRAEYIRMVLHAADLEFEDHRIEMNDWPTIKPTIAGGQLPVLDVTTCCGKSKQMNESMAIARW
FARKHHMMGSNDEEYYEVERVIGQCSDIYQDVYRIFRATGEEKQKLLKQFTEGNGPRLLKVISKHLEASP
TGLVVGDKPTLADFCILCAIDQVEVTVPGLSKDKFPIFERHRETVLKKHAKLAAYMETRPTT

CLUSTAL multiple sequence alignment by MUSCLE (3.8)

```
Caenorhabditis      MPNYKLLYFDARALAEPIRIMFAMLNVPYEDYRVSVEEWSKLKPTTPFGQLPILQV---D
Crassostrea         MTKYTVHYFNVKGRGEIVRLILVAAGVDFEDNRVEREDWPKLKPTMPAGQMPVLEV---D
Dugesia             ---IILTYFNARGKAELIRFVLIVSDVEFEDKRIEFEEWPQLKPTIPTGQLPIVQLS--C
Schmidtea           MSTVKVTYFDARGRAELIRLVLKASKIEFEDVRITKDKWPEVKPTTPTGKLPVVEY---E
Paragonimus         -----LTYFNGRGRAEYIRMVLHAADLEFEDHRIEMNDWPTIKPTIAGGQLPVLDVTTCC
                         : **: .. .* :*:::    : :** *:  :.*. :*** . *::*:::

Caenorhabditis      G--EQFGQSMSITRYLARKFGLAGKTAEEEAYADSIVDQYRDFIFFFRQFTSSVFYGSDA
Crassostrea         G--KKYCQSIAIARYLAREFGLGGSTNVEQLQVDQVVDTISDFL----TEMYKPVFEQDA
Dugesia             G--GIINESMAIARYFAKKYHLTGSNENEEYKVDRVVCTLDDLF----NKVIDVFHEKDE
Schmidtea           G--KQLTQSMAIARVVARKHGFMGEDDKEYYLVERAIGQMVDVL----EGLYKIYFAPEE
Paragonimus         GKSKQMNESMAIARWFARKHHMMGSNDEEYYEVERVIGQCSDIY----QDVYRIFRATGE
                    *      :*::*:* .*.:. : *.   *   .:  :    *.

Caenorhabditis      DHINKVRFEVVEPARDDFLAIINKFLAKSKSGFLVGDSLTWADIVIADNLTSLLKNGFLD
Crassostrea         TRKAEMNKKLNEETIPRVLGILQNFLEGNGGDYFVGSKTSLADIYFMDVVSRLVEKDEKV
Dugesia             GKKETLKHELNETHLPAFLDRLDYYLKDKNGDFFLGDHPSLADLQLVNVMDHFE---ESQ
Schmidtea           -KKEELRAEYVATSGRDNLKALEGFIKE--TGFFAGEKITLAELFFLVVSDYLV-KLPQL
Paragonimus         -EKQKLLKQFTEGNGPRLLKVISKHLEASPTGLVVGDKPTLADFCILCAIDQVEVTVPGL
                         :  :         *  :. .:     . . *.  : *:: :      .

Caenorhabditis      F-NKEKKLEEFYNKI-HSIPEIKNYVATRKDSIV
Crassostrea         L-EKFPKLAASLQKT-QALPKIEAYLAKRPKTEL
Dugesia             Y-QSHPKLVHCYQKVLEHYPKLKHYKDNRQKSII
Schmidtea           Y-DDFPKLKELRERILKANTDVEEWVNTRPVTEM
Paragonimus         SKDKFPIFERHRETVLKKHAKLAAYMETRPTT--
                      :.   :    :      ..:  :  .*  :
```

NOTE: I added sequences from the planarian class too due to the lack sequences from the same family.

[Q6] Create a phylogenetic tree, using either a parsimony or distance-based approach. Bootstrapping and tree rooting are optional. Use “simple phylogeny” online from the EBI or any respected phylogeny program (such as MEGA, PAUP, or Phylip). Paste an image of your Cladogram or tree output in your report.



Simple Phylogeny

Input form | Web services | Help & Documentation | Bioinformatics Tools FAQ | Feedback

Tools > Phylogeny > Simple Phylogeny

Results for job simple_phylogeny-I20211126-220331-0347-46815436-p2m

Phylogenetic Tree | Result Summary | Submission Details

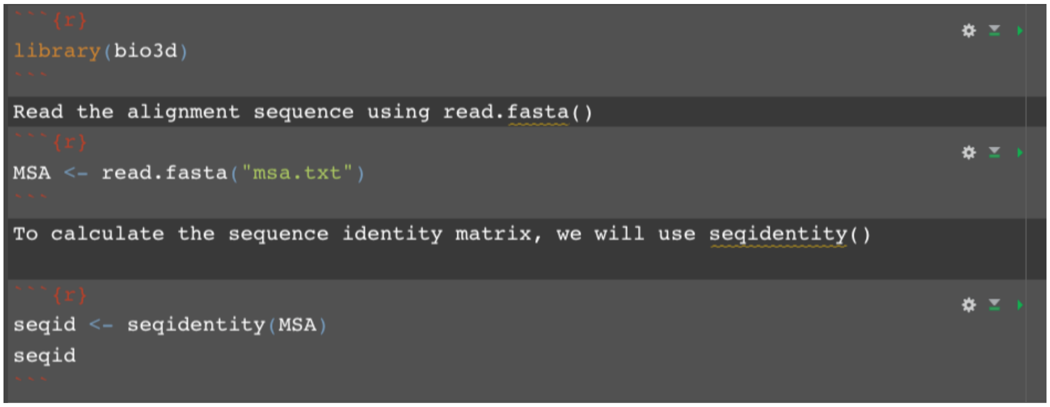
Phylogram

Branch length: ● Cladogram ○ Real

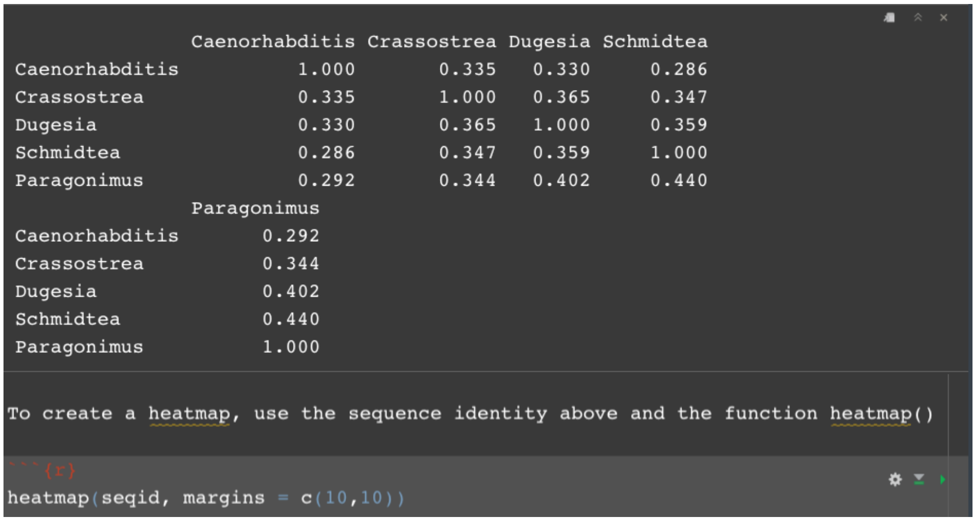
Caenorhabditis 0.82295
Crassostrea 0.57735
Dugesia 0.51359
Schmidtea 0.52043
Paragonimus 0.453

[Q7] Generate a sequence identity based heatmap of your aligned sequences using R.

If necessary convert your sequence alignment to the ubiquitous FASTA format (Seaview can read in clustal format and "Save as" FASTA format for example). Read this FASTA format alignment into R with the help of functions in the Bio3D package. Calculate a sequence identity matrix (again using a function within the Bio3D package). Then generate a heatmap plot and add to your report. Do make sure your labels are visible and not cut at the figure margins.



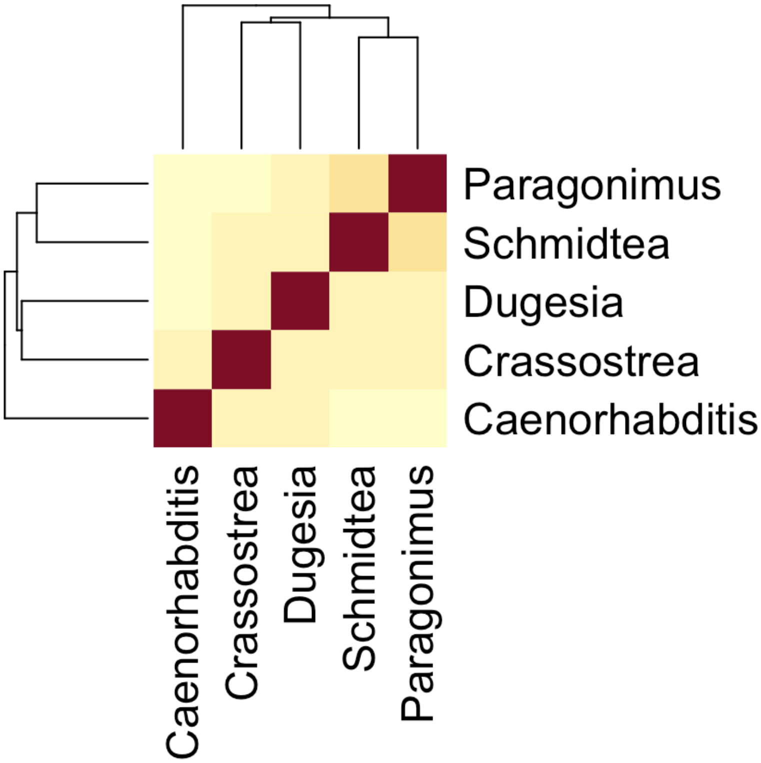
~~~
```{r}
library(bio3d)
```
Read the alignment sequence using read.fasta()
```{r}
MSA <- read.fasta("msa.txt")
```
To calculate the sequence identity matrix, we will use seqidentity()

```{r}
seqid <- seqidentity(MSA)
seqid
```
~~~



```
               Caenorhabditis Crassostrea Dugesia Schmidtea
Caenorhabditis          1.000       0.335   0.330     0.286
Crassostrea             0.335       1.000   0.365     0.347
Dugesia                 0.330       0.365   1.000     0.359
Schmidtea               0.286       0.347   0.359     1.000
Paragonimus             0.292       0.344   0.402     0.440
               Paragonimus
Caenorhabditis       0.292
Crassostrea          0.344
Dugesia              0.402
Schmidtea            0.440
Paragonimus          1.000
```

To create a heatmap, use the sequence identity above and the function heatmap()

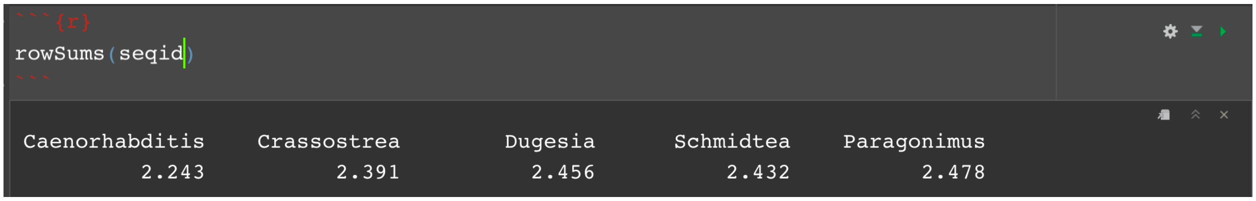
~~~
```{r}
heatmap(seqid, margins = c(10,10))
~~~



[Q8] Using R/Bio3D (or an online blast server if you prefer), search the main protein structure database for the most similar atomic resolution structures to your aligned sequences.

List the top 3 *unique* hits (i.e. not hits representing different chains from the same structure) along with their Evalue and sequence identity to your query. Please also add annotation details of these structures. For example include the annotation terms PDB identifier (structureId), Method used to solve the structure (experimentalTechnique), resolution (resolution), and source organism (source).

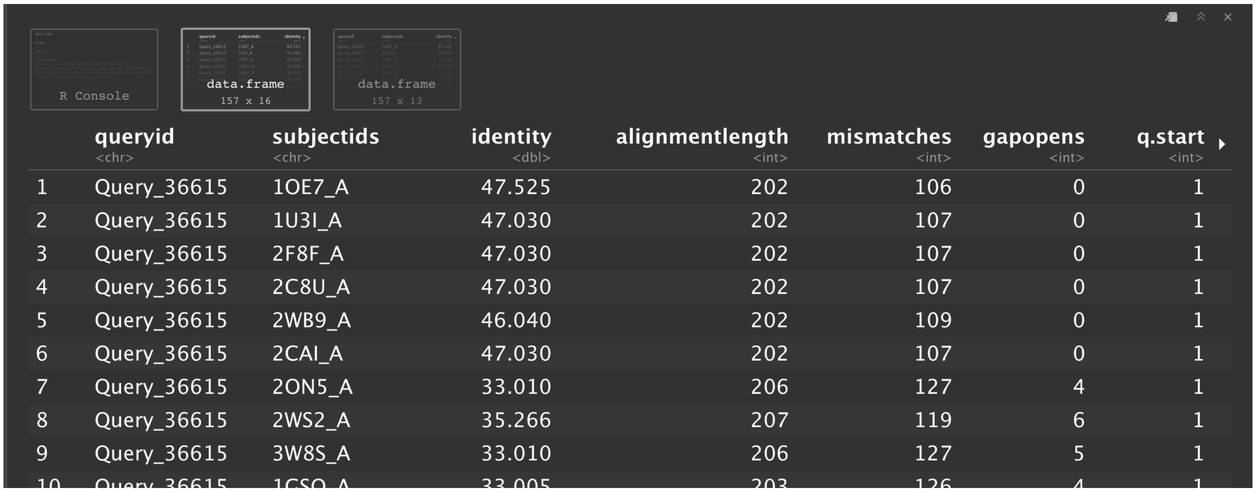
In R, using the bio3d package, I calculated a consensus between all five sequences using consensus () function, however, there are too many gaps, so I calculated the rowSums of the sequence identities.



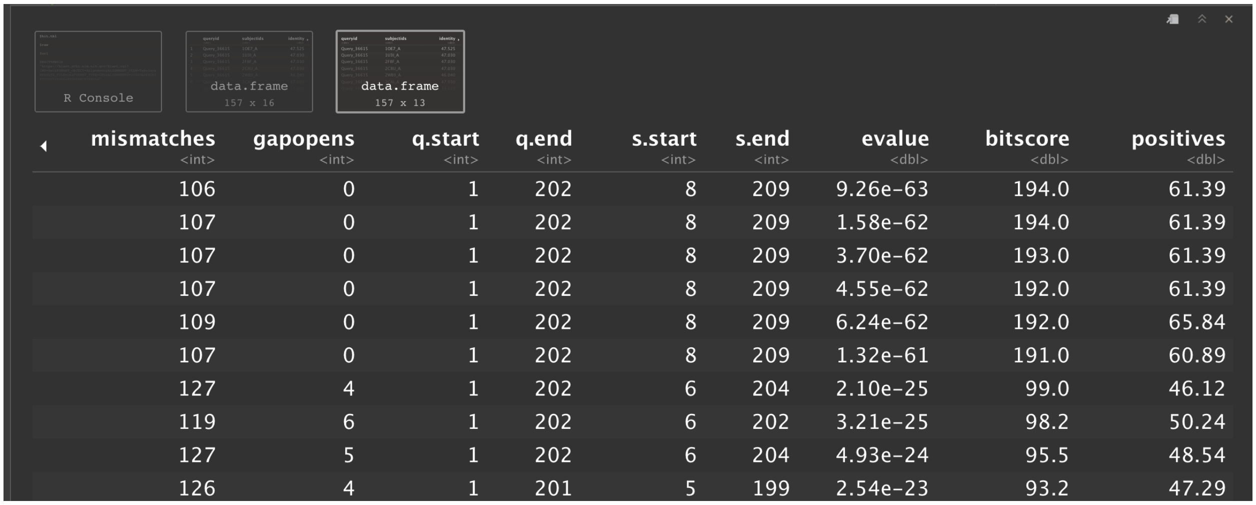
~~~
```{r}
rowSums(seqid)
```
~~~

```
Caenorhabditis    Crassostrea       Dugesia     Schmidtea    Paragonimus
         2.243          2.391         2.456         2.432          2.478
```

Since hypothetical protein P879_00235 [Paragonimus westermani] has the highest sequence id calculation, it was chosen to blast for a structure on pdb. Blast using blast.pdb(). Here is the result.

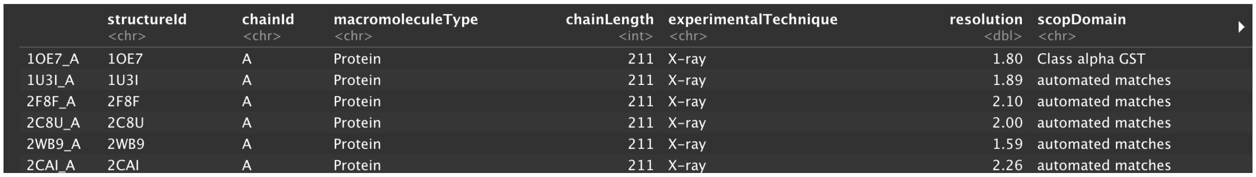


	queryid <chr>	subjectids <chr>	identity <dbl>	alignmentlength <int>	mismatches <int>	gapopens <int>	q.start <int>
1	Query_36615	1OE7_A	47.525	202	106	0	1
2	Query_36615	1U3I_A	47.030	202	107	0	1
3	Query_36615	2F8F_A	47.030	202	107	0	1
4	Query_36615	2C8U_A	47.030	202	107	0	1
5	Query_36615	2WB9_A	46.040	202	109	0	1
6	Query_36615	2CAI_A	47.030	202	107	0	1
7	Query_36615	2ON5_A	33.010	206	127	4	1
8	Query_36615	2WS2_A	35.266	207	119	6	1
9	Query_36615	3W8S_A	33.010	206	127	5	1
10	Query_36615	1CSO_A	33.005	203	126	4	1

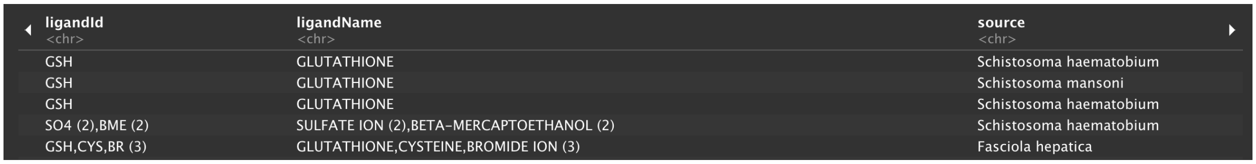


mismatches <int>	gapopens <int>	q.start <int>	q.end <int>	s.start <int>	s.end <int>	evalue <dbl>	bitscore <dbl>	positives <dbl>
106	0	1	202	8	209	9.26e-63	194.0	61.39
107	0	1	202	8	209	1.58e-62	194.0	61.39
107	0	1	202	8	209	3.70e-62	193.0	61.39
107	0	1	202	8	209	4.55e-62	192.0	61.39
109	0	1	202	8	209	6.24e-62	192.0	65.84
107	0	1	202	8	209	1.32e-61	191.0	60.89
127	4	1	202	6	204	2.10e-25	99.0	46.12
119	6	1	202	6	202	3.21e-25	98.2	50.24
127	5	1	202	6	204	4.93e-24	95.5	48.54
126	4	1	201	5	199	2.54e-23	93.2	47.29

Analyze the blast data using plot.blast() and annotate them using pdb.annotate()



```
        structureId  chainId  macromoleculeType  chainLength  experimentalTechnique  resolution  scopDomain
        <chr>        <chr>    <chr>                    <int>  <chr>                       <dbl>  <chr>
1OE7_A  1OE7         A        Protein                    211  X-ray                        1.80  Class alpha GST
1U3I_A  1U3I         A        Protein                    211  X-ray                        1.89  automated matches
2F8F_A  2F8F         A        Protein                    211  X-ray                        2.10  automated matches
2C8U_A  2C8U         A        Protein                    211  X-ray                        2.00  automated matches
2WB9_A  2WB9         A        Protein                    211  X-ray                        1.59  automated matches
2CAI_A  2CAI         A        Protein                    211  X-ray                        2.26  automated matches
```

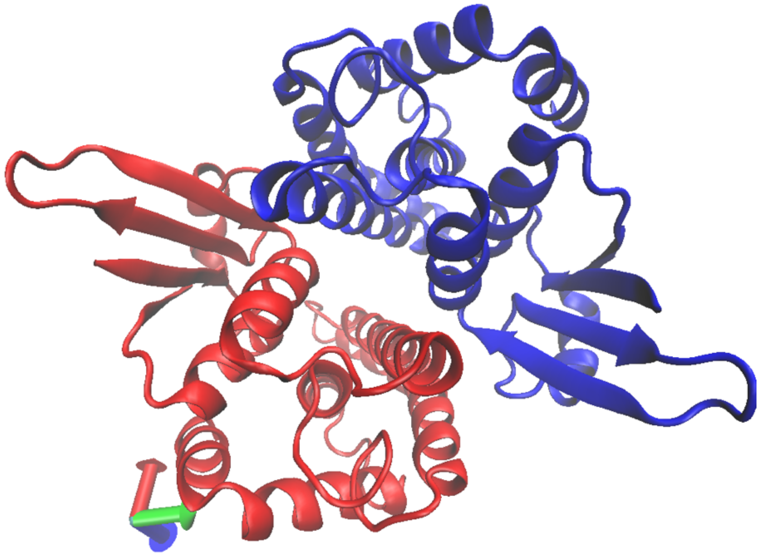


```
ligandId             ligandName                                      source
<chr>                <chr>                                           <chr>
GSH                  GLUTATHIONE                                     Schistosoma haematobium
GSH                  GLUTATHIONE                                     Schistosoma mansoni
GSH                  GLUTATHIONE                                     Schistosoma haematobium
SO4 (2),BME (2)      SULFATE ION (2),BETA-MERCAPTOETHANOL (2)        Schistosoma haematobium
GSH,CYS,BR (3)       GLUTATHIONE,CYSTEINE,BROMIDE ION (3)            Fasciola hepatica
```

ID	Technique	Resolution	Source	Evalue	Identity
1U3I	X-RAY DIFFRACTION	1.89	*Schistosoma mansoni*	1.58e-62	47.030
1OE7	X-RAY DIFFRACTION	1.80	*Schistosoma haematobium*	9.26e-63	47.525
2WB9	X-RAY DIFFRACTION	1.59	*Fasciola hepatica*	6.24e-62	46.040

[Q9] Generate a molecular figure of one of your identified PDB structures using **VMD**. You can optionally highlight conserved residues that are likely to be functional. Please use a white or transparent background for your figure (i.e. not the default black).

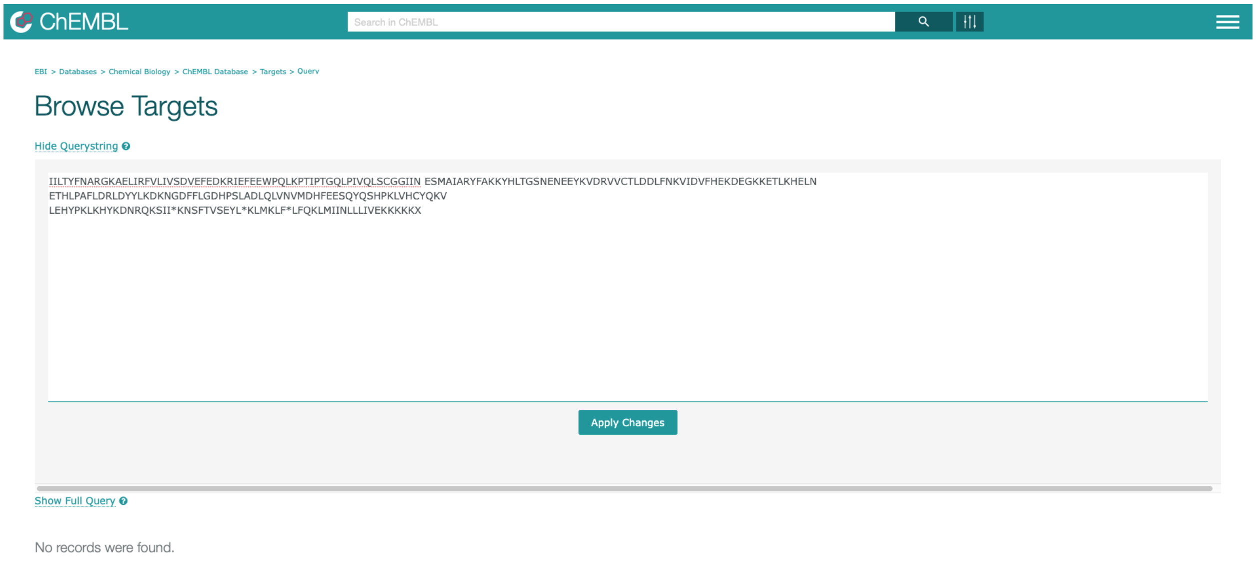
Based on sequence similarity. How likely is this structure to be similar to your "novel" protein?



This structure is 1OE7 from *Schistosoma haematobium.* The ligand (glutathione) interacts with chain A (in red) at Lys45 (not shown on structure). Based on sequence similarity, the novel protein and this structure might not have very similar structure, since it only has an identity score of 47.5.

[Q10] Perform a "Target" search of ChEMBEL (https://www.ebi.ac.uk/chembl/) with your novel sequence. Are there any **Target Associated Assays** and **ligand efficiency data** reported that may be useful starting points for exploring potential inhibition of your novel protein?

Using chEMBEL target, I changed querystring with my novel sequence, but that do not yield any result.



EBI > Databases > Chemical Biology > ChEMBL Database > Targets > Query

Browse Targets

Hide Querystring

```
IILTYFNARGKAELIRFVLIVSDVEFEDKRIEFEEWPQLKPTIPTGQLPIVQLSCGGIIN ESMAIARYFAKKYHLTGSNENEEYKVDRVVCTLDDLFNKVIDVFHEKDEGKKETLKHELN
ETHLPAFLDRLDYYLKDKNGDFFLGDHPSLADLQLVNVMDHFEESQYQSHPKLVHCYQKV
LEHYPKLKHYKDNRQKSII*KNSFTVSEYL*KLMKLF*LFQKLMIINLLLIVEKKKKKX
```

Show Full Query

No records were found.