BGGN-213: FOUNDATIONS OF BIOINFORMATICS

The find-a-gene project assignment
http://thegrantlab.org/bggn213
Dr. Barry Grant

Overview:

The find-a-gene project is a required assignment for BGGN-213. You should prepare a written report in **PDF** format that has responses to each question labeled **[Q1] - [Q10]** below. You may wish to consult the scoring rubric at the end of this document and the example report provided online (note that the example report is from a previous quarter and the questions may differ).

The objective with this assignment is for you to demonstrate your grasp of database searching, sequence analysis, structure analysis and the R environment that we have covered in class.

Due Date:

Your responses to questions Q1-Q4 are due at 12pm on the **Monday of Week 5** (see the Assignments and Grading section of our website for details). Note that these first set of answers can be obtained very quickly (at best within 15 or 20 minutes), so if you don't succeed at first, just keep trying.

The complete assignment, including responses to all questions, is due at 12pm on the **Monday of Week 10**.

Submission instructions:

Your report formatted as a **PDF document** should be uploaded to ***GradeScope***. Please make sure to include your UCSD email and PID number on the first page.

Be sure to include your UCSD email and PID number on the first page of your report.

Submit your preliminary report with answers to Q1-Q4 as soon as you can so we can determine if you have found a novel gene. Submit this preliminary report as one document with screen shots of the results inserted appropriately.

See the demonstration report linked to on the course website for an example of format but note that questions there may differ as it is from a previous quarter. I will email you my decision; proceed with subsequent questions only after we are sure you have found a novel gene (and thus be successful in the later stages of the project).

For the final report add your results for Q5-Q10 to the preliminary report and submit the final document containing your results for all questions - **Please do not send only Q5-Q10 answers as the final report**.

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.

If you do not have a favorite protein, select human RBP4 or KIF11. Do not use beta globin as this is in the worked example report that I provide you with online.

Name: ECT-2

Function: activates Rho GTPases and controls cytokinesis and many other cellular processes

Accession number: NP_001245244

Species: Human

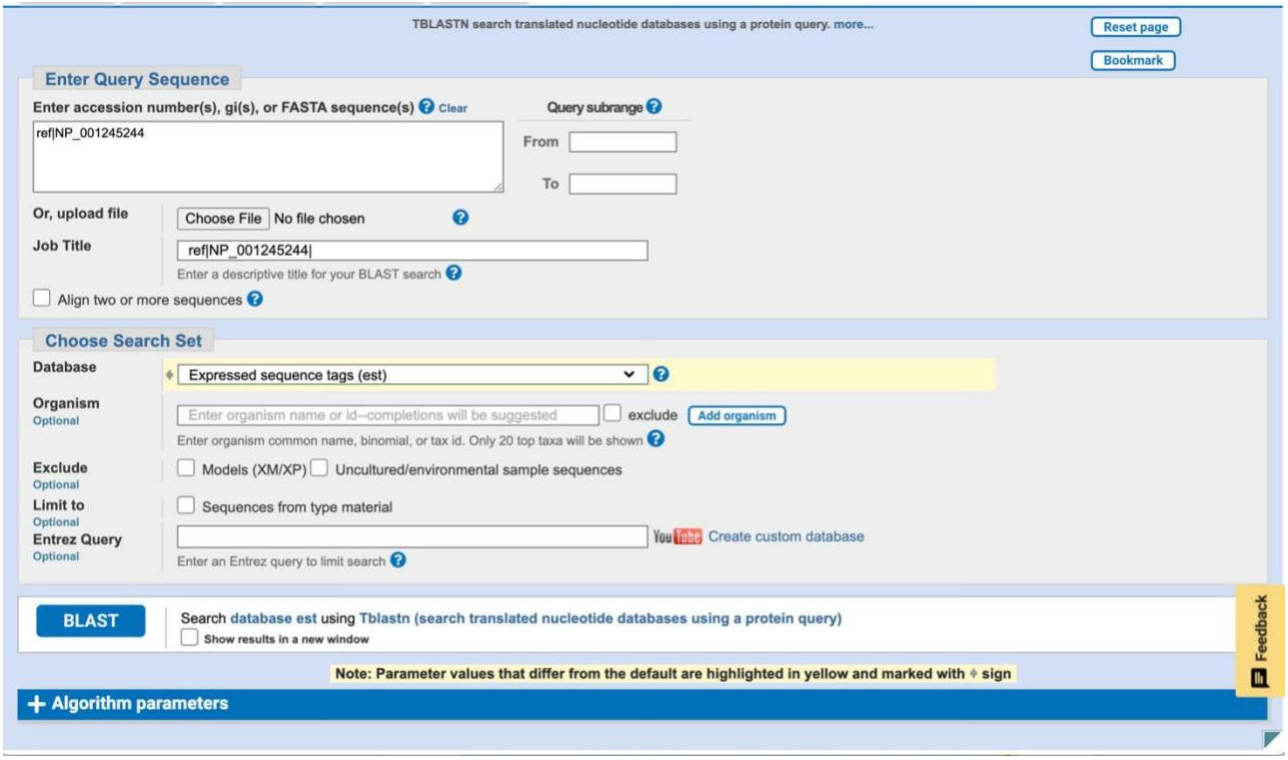
[Q2] Perform a BLAST search against a DNA database, such as a database consisting of genomic DNA or ESTs. The BLAST server can be at NCBI or elsewhere. Include details of the BLAST method used, database searched and any limits applied (e.g. Organism).

Method: tBLASTN general search

Database : Expressed sequence tags (EST)

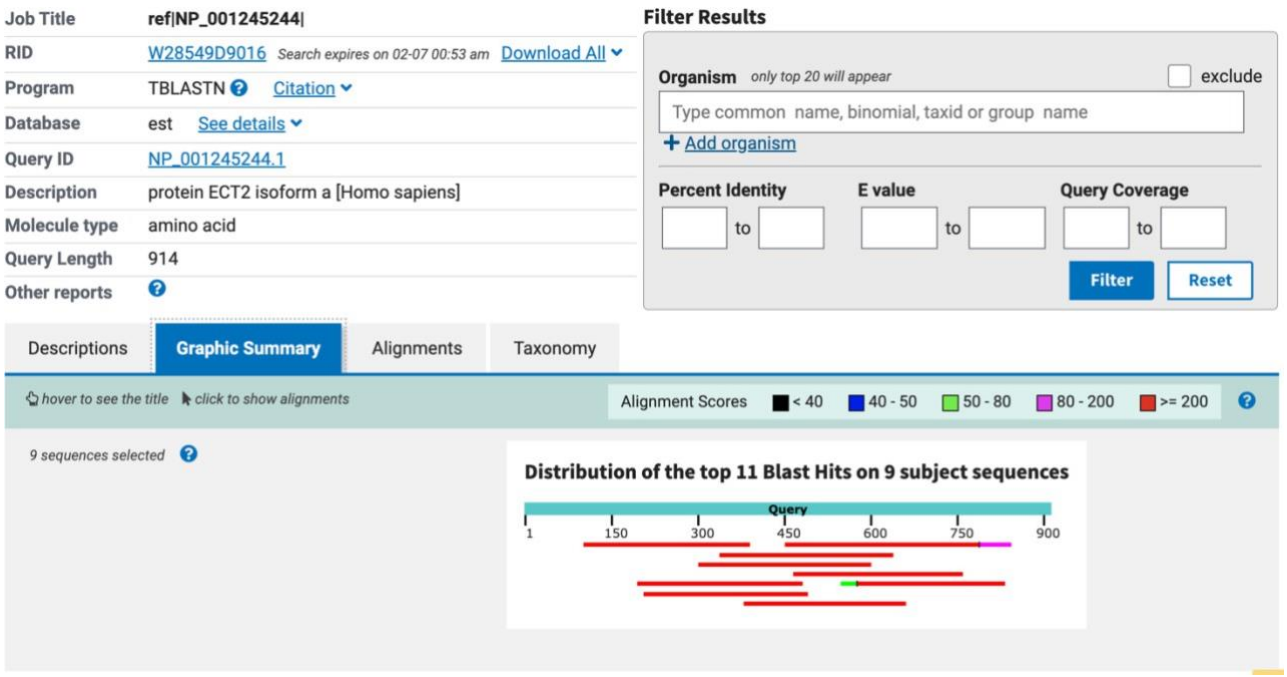
Organism: None

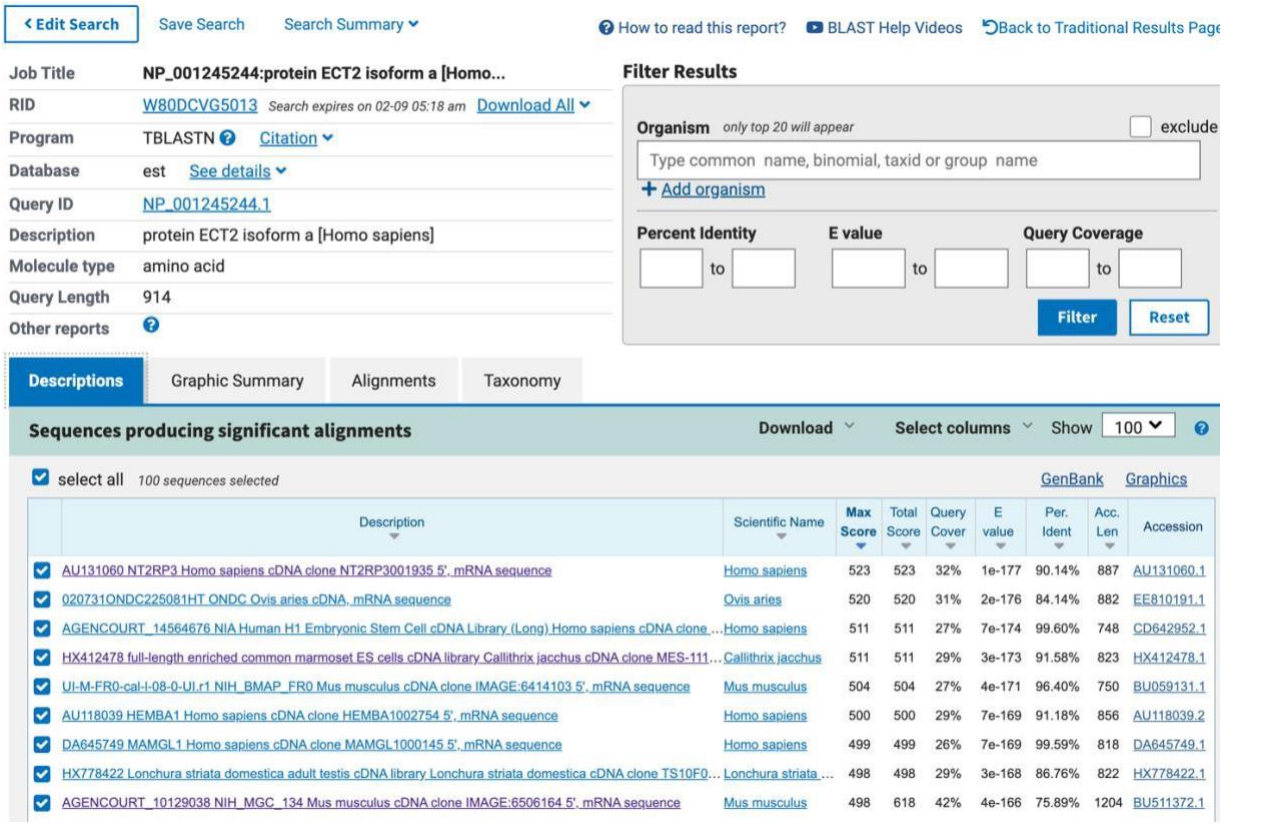
Also include the output of that BLAST search in your document. If appropriate, change the font to Courier size 10 so that the results are displayed neatly. You can also screen capture a BLAST output (e.g. alt print screen on a PC or on a MAC press ⌘-shift-4. The pointer becomes a bulls eye. Select the area you wish to capture and release. The image is saved as a file called Screen Shot [].png in your Desktop directory). It is **not** necessary to print out all of the blast results if there are many pages.



On the BLAST results, clearly indicate a match that represents a protein sequence, encoded from some DNA sequence, that is homologous to your query protein. I need to be able to inspect the pairwise alignment you have selected, including the E value and score. It should be labeled a "genomic clone" or "mRNA sequence", etc. - but include no functional annotation.

Chosen match: Accession number: BU511372.1 1204 base pair mRNA sequence from Mus Musculus





Refiltered with highest query cover

Job Title	NP_001245244:protein ECT2 isoform a [Homo...
RID	W80DCVG5013 *Search expires on 02-09 05:18 am* Download All
Program	TBLASTN Citation
Database	est See details
Query ID	NP_001245244.1
Description	protein ECT2 isoform a [Homo sapiens]
Molecule type	amino acid
Query Length	914
Other reports	

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

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Select columns Show 100

select all *100 sequences selected* GenBank Graphics

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
AGENCOURT_10129038 NIH_MGC_134 Mus musculus cDNA clone IMAGE:6506164 5', mRNA sequence	Mus musculus	498	618	42%	4e-166	75.89%	1204	BU511372.1
JGI_CABE6575.fwd NIH_XGC_tropOva1 Xenopus tropicalis cDNA clone IMAGE:7825333 5', mRNA sequence	Xenopus tropicalis	484	484	32%	2e-162	78.67%	896	DN095848.1
JGI_CAAT7914.fwd CAAT Pimephales promelas brain 7-8 month adults, males and females pooled (M) Pimeph...	Pimephales pro...	473	473	32%	3e-158	73.33%	897	DT233970.1
AU131060 NT2RP3 Homo sapiens cDNA clone NT2RP3001935 5', mRNA sequence	Homo sapiens	523	523	32%	1e-177	90.14%	887	AU131060.1
020731ONDC225081HT ONDC Ovis aries cDNA, mRNA sequence	Ovis aries	520	520	31%	2e-176	84.14%	882	EE810191.1
JGI_CABE10925.fwd NIH_XGC_tropOva1 Xenopus tropicalis cDNA clone IMAGE:7829493 5', mRNA sequence	Xenopus tropicalis	427	427	31%	9e-141	70.63%	854	DR846473.1
JGI_CAAO8013.fwd NIH_XGC_tropTe5 Xenopus tropicalis cDNA clone CAAO8013 5', mRNA sequence	Xenopus tropicalis	463	522	31%	2e-154	85.17%	868	CX942055.1
AU142499 Y79AA1 Homo sapiens cDNA clone Y79AA1000433 5', mRNA sequence	Homo sapiens	449	449	31%	2e-149	77.24%	848	AU142499.1
AU133302 NT2RP4 Homo sapiens cDNA clone NT2RP4001760 5', mRNA sequence	Homo sapiens	478	478	30%	3e-160	86.57%	869	AU133302.1
UI-M-HN0-cng-k-23-0-UI.r1 NIH_BMAP_HN0 Mus musculus cDNA clone IMAGE:30639910 5', mRNA sequence	Mus musculus	439	439	30%	1e-145	74.38%	753	CK634092.1

Download GenBank Graphics Sort by: E value ▼ Next ▲ Previous ◀ Descriptions

AGENCOURT_10129038 NIH_MGC_134 Mus musculus cDNA clone IMAGE:6506164 5', mRNA sequence

Sequence ID: BU511372.1 Length: 1204 Number of Matches: 2

Range 1: 37 to 1041 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
498 bits(1281)	4e-166	Compositional matrix adjust.	255/336(76%)	268/336(79%)	1/336(0%)	+1

```
Query  452   ARWQVAKELYQTESNYVNILATIIQLFQVPLEEEGQRGGPILAPEEIKTIFGSIPDIFDV  511
             RWQVAKELYQTESNYVNILATIIQLFQVPLEEEGQRGGPILAPEEIKTIFGSIPDIFDV
Sbjct  37    VRWQVAKELYQTESNYVNILATIIQLFQVPLEEEGQRGGPILAPEEIKTIFGSIPDIFDV  216

Query  512   HTKIKDDLEDLIVNWDESKSIGDIFLKYSKDLVKTYPPFVNFFEMSKETIIKCEKQKPRF  571
             H KIKDDLEDLI NWDES+SIGDIFLKY+KDLVKTYPPFVNFFEMSKE IIKCEKQKPRF
Sbjct  217   HMKIKDDLEDLIANWDESRSIGDIFLKYAKDLVKTYPPFVNFFEMSKEMIIKCEKQKPRF  396

Query  572   HAFLKINQAKPECGRQSLVELLIRPVQRLPSVALLLNDLKKHTADENPDKSTLEKAIGSL  631
             HAFLKINQAKPECGRQSLVELLIRPVQRLPSVALLLNDLKKHTADENPDKSTLEKAIGSL
Sbjct  397   HAFLKINQAKPECGRQSLVELLIRPVQRLPSVALLLNDLKKHTADENPDKSTLEKAIGSL  576

Query  632   KEVMTHINEDKRKTEAQKQIFDVVYEVDGCPANLLSSHRSLVQRVETISLGEHPCDRGEQ  691
             KEVMTHINEDKRKTEAQKQIFDVVYEVDGCPANLLSSHRSLVQRVET+SLGEHPCDRGEQ
Sbjct  577   KEVMTHINEDKRKTEAQKQIFDVVYEVDGCPANLLSSHRSLVQRVETVSLGEHPCDRGEQ  756

Query  692   VTLFLFNDCLEIARKRHKVIGTFRSPHGQTRPPASLKHIHLMPLSQIKKVLDIRETEDCH  751
             VTLF          K++   S    PP          ++K+    R+ +
Sbjct  757   VTLFPLMTASR*QESGTKLLALLESLTNAPGPPLL*STFISCLFLRLKRAGHPRD-KRLS  933

Query  752   NAFALLVRPPTEQANVLLSFQMTSDELPKENWLKML  787
                   +P TEQA    +F+      PKE  LK L
Sbjct  934   PCLCPACKPXTEQAMYCSTFK*RPRTFPKETALKSL  1041
```

AGENCOURT_10129038 NIH_MGC_134 Mus musculus cDNA clone IMAGE:6506164 5', mRNA sequence

Sequence ID: BU511372.1 Length: 1204 Number of Matches: 2

Range 1: 37 to 1041 GenBank Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
498 bits(1281)	4e-166	Compositional matrix adjust.	255/336(76%)	268/336(79%)	1/336(0%)

```
Query  452   ARWQVAKELYQTESNYVNILATIIQLFQVPLEEEGQRGGPILAPEEIKTIFGSIPDIFDV  511
             RWQVAKELYQTESNYVNILATIIQLFQVPLEEEGQRGGPILAPEEIKTIFGSIPDIFDV
Sbjct  37    VRWQVAKELYQTESNYVNILATIIQLFQVPLEEEGQRGGPILAPEEIKTIFGSIPDIFDV  216

Query  512   HTKIKDDLEDLIVNWDESKSIGDIFLKYSKDLVKTYPPFVNFFEMSKETIIKCEKQKPRF  571
             H KIKDDLEDLI NWDES+SIGDIFLKY+KDLVKTYPPFVNFFEMSKE IIKCEKQKPRF
Sbjct  217   HMKIKDDLEDLIANWDESRSIGDIFLKYAKDLVKTYPPFVNFFEMSKEMIIKCEKQKPRF  396

Query  572   HAFLKINQAKPECGRQSLVELLIRPVQRLPSVALLLNDLKKHTADENPDKSTLEKAIGSL  631
             HAFLKINQAKPECGRQSLVELLIRPVQRLPSVALLLNDLKKHTADENPDKSTLEKAIGSL
Sbjct  397   HAFLKINQAKPECGRQSLVELLIRPVQRLPSVALLLNDLKKHTADENPDKSTLEKAIGSL  576

Query  632   KEVMTHINEDKRKTEAQKQIFDVVYEVDGCPANLLSSHRSLVQRVETISLGEHPCDRGEQ  691
             KEVMTHINEDKRKTEAQKQIFDVVYEVDGCPANLLSSHRSLVQRVET+SLGEHPCDRGEQ
Sbjct  577   KEVMTHINEDKRKTEAQKQIFDVVYEVDGCPANLLSSHRSLVQRVETVSLGEHPCDRGEQ  756

Query  692   VTLFLFNDCLEIARKRHKVIGTFRSPHGQTRPPASLKHIHLMPLSQIKKVLDIRETEDCH  751
             VTLF              K++    S      PP             ++K+    R+ +
Sbjct  757   VTLFPLMTASR*QESGTKLLALLESLTNAPGPPLL*STFISCLFLRLKRAGHPRD-KRLS  933

Query  752   NAFALLVRPPTEQANVLLSFQMTSDELPKENWLKML  787
                    +P TEQA    +F+      PKE  LK L
Sbjct  934   PCLCPACKPXTEQAMYCSTFK*RPRTFPKETALKSL  1041
```

Range 2: 774 to 1202GenBankGraphicsNext MatchPrevious MatchFirst Match

Score	Expect	Method	Identities	Positives	Gaps
120 bits(300)	9e-26	Compositional matrix adjust.	76/147(52%)	89/147(60%)	4/147(2%)

```
Query  698   NDCLEIARKRHKVIGTFRSPHGQTRPPASLKHIHLMPLSQIKKVLDIRETEDCHNAFALL  757
             +DCLEIARKRHKVIGTFR  H +TRPPASLKHIHLMPLSQIKK       +      + L
Sbjct  774   HDCLEIARKRHKVIGTFRKSHERTRPPASLKHIHLMPLSQIKKGWTSERQKIVTMPLSCL  953

Query  758   VRPPTEQANVLLSFQMTSDELPKENWLKMLCRHVANTICKADAENLIYTADPESFEVNTK  817
                     +VLL+FQMTS +LPK N LK+    ++  T+CK  A NL+    DP  F+ N K
Sbjct  954    -*AXNRTGHVLLNFQMTSKDLPKGNCLKIFA-NIYPTLCKGKAGNLLDGLDPNPFKKN-K  1124

Query  818   DMDSTLSRASRAIKKTSKKVTRAFSFS  844
               DS    A     K SKK TR F FS
Sbjct  1125  RWDSHWG-ALLNH*KNSKKGTRGFLFS  1202
```

In general, [Q2] is the most difficult for students because it requires you to have a “feel” for how to interpret BLAST results. You need to distinguish between a perfect match to your query (i.e. a sequence that is not “novel”), a near match (something that might be “novel”, depending on the results of [Q4]), and a non-homologous result.

If you are having trouble finding a novel gene try restricting your search to an organism that is poorly annotated.

[**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.

Chosen Sequence: protein sequence translated using EBI Transeq of cDNA sequence from NCBI.

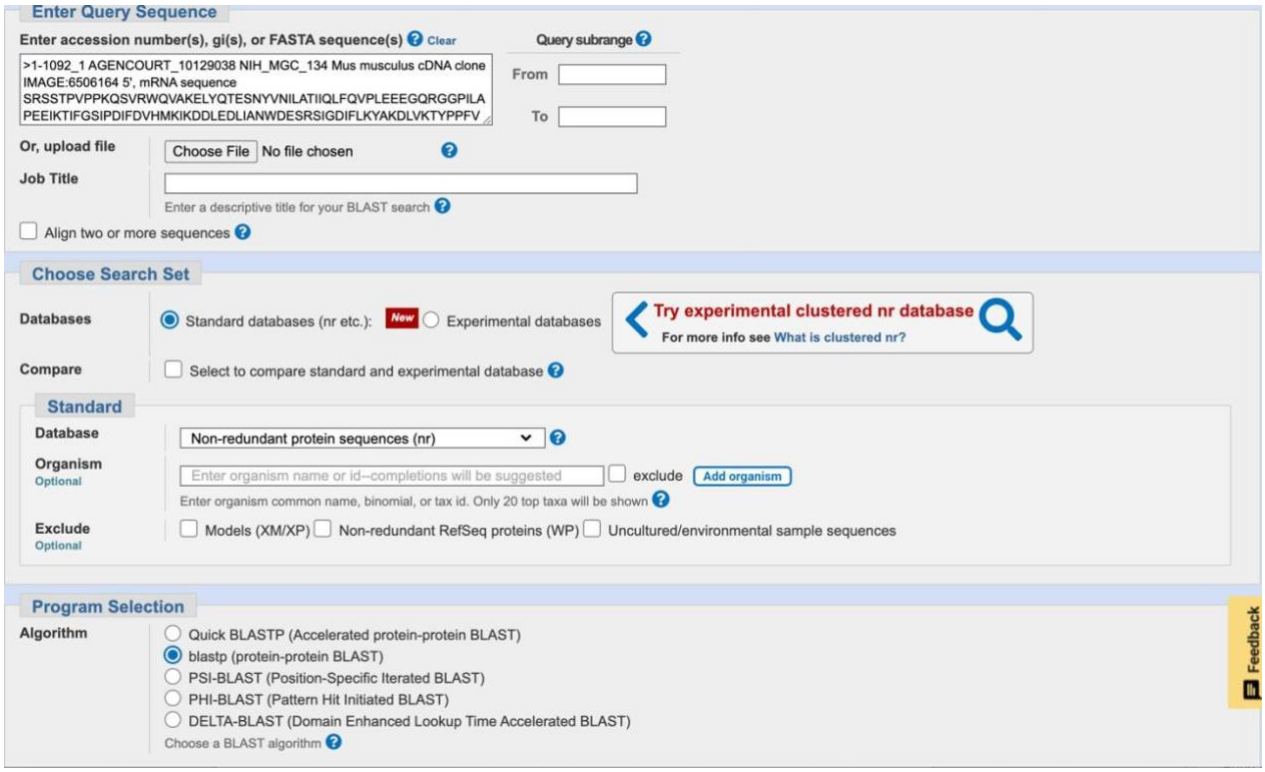
>1-1092_1 AGENCOURT_10129038 NIH_MGC_134 Mus musculus cDNA clone IMAGE:6506164 5', mRNA sequence
SRSSTPVPPKQSVRWQVAKELYQTESNYVNILATIIQLFQVPLEEEGQRGGPILAPEEIK
TIFGSIPDIFDVHMKIKDDLEDLIANWDESRSIGDIFLKYAKDLVKTYPPFVNFFEMSKE
MIIKCEKQKPRFHAFLKINQAKPECGRQSLVELLIRPVQRLPSVALLLNDLKKHTADENP
DKSTLEKAIGSLKEVMTHINEDKRKTEAQKQIFDVVYEVDGCPANLLSSHRSLVQRVETV
SLGEHPCDRGEQVTLFPLMTASR*QESGTKLLALLESLTNAPGPPLL*STFISCLFLRLK
RAGHPRDKRLSPCLCPACKPXTEQAMYCSTFK*RPRTFPKETALKSLPTFTQPFVRAKLE IFWM

Here, tell me the name of the novel protein, and the species from which it derives. It is very unlikely (but still definitely possible) that you will find a novel gene from an organism such as *S. cerevisiae*, human or mouse, because those genomes have already been thoroughly annotated. It is more likely that you will discover a new gene in a genome that is currently being sequenced, such as bacteria or plants or protozoa.

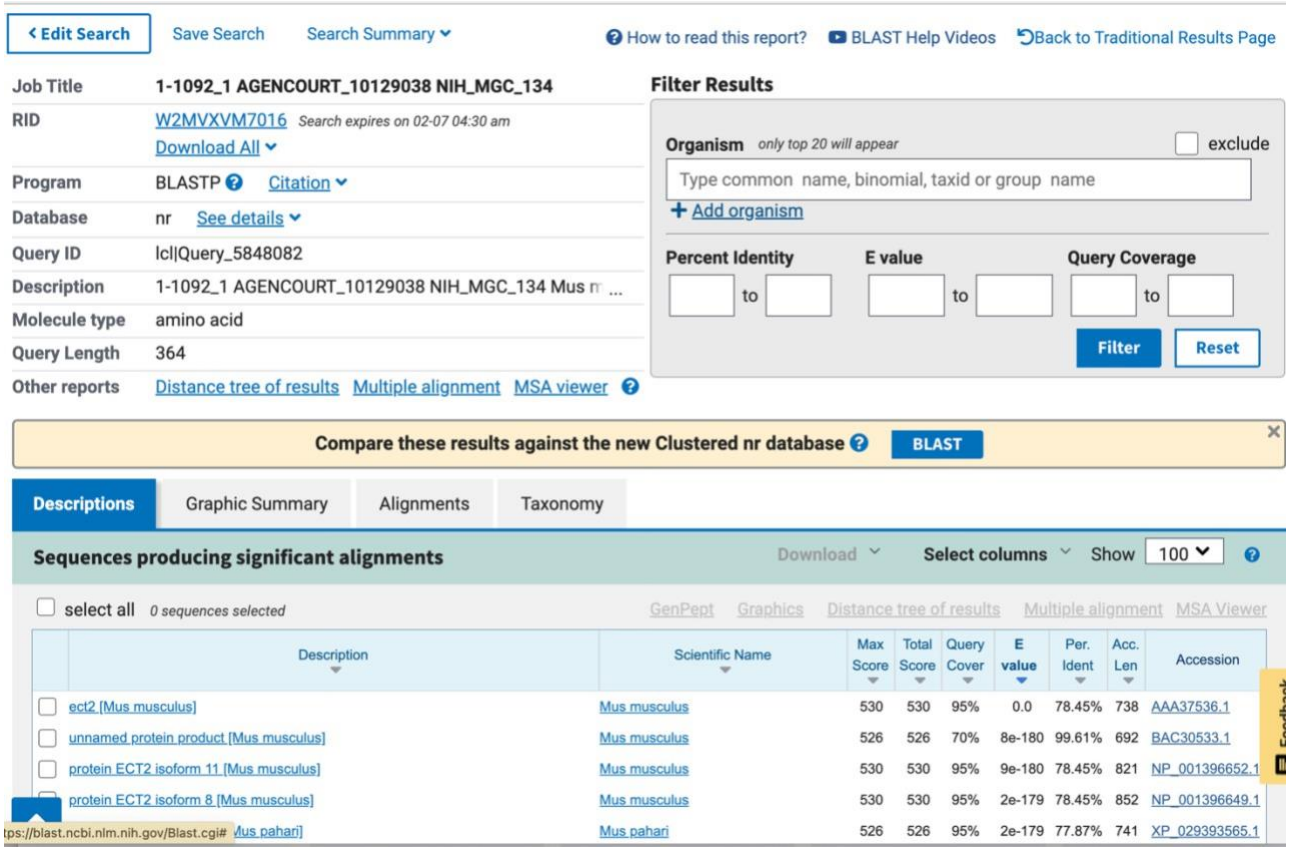
Name: Mus ECT-2 protein
Species: Mus musculus
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Myomorpha; Muroidea; Muridae; Murinae; Mus; Mus.

[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.



Top hits are from Mus musculus



[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.

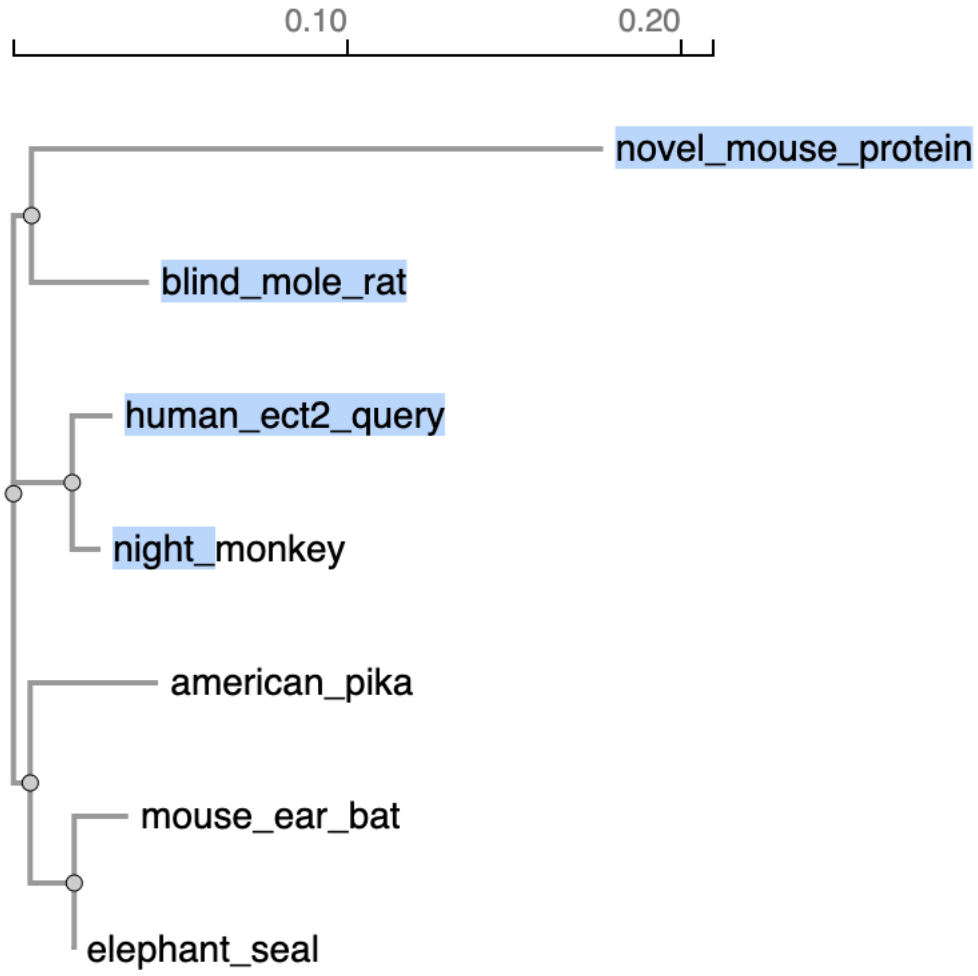
```
1     .     .     .     .     .     60
novel_mouse_ECT2  ------------------------------------------------------------
HUMAN_query_ECT2  MAENSVLTSTTGRTSLADSSIFDSKVTEISKENLLIGSTSYVEEEMPQIETRVILVQEAG
blind_mole_rat    MADSSVLATATGTASLVDSSIFDSKVTETSKENLCTESTSYVDEEMPQVQTRVVLVQEAG
night_monkey      ------------------------------------------------------------
american_pika     ------------------------------------------------------------
mouse_ear_bat     ------------------------------------------------------------
elephant_seal     MADSSVLTSTTGRTSLADSSIFDSKVTEISKENVVIRSASYVEEEMPQIETRLILVQEAG

          1     .     .     .     .     .     60

         61     .     .     .     .     .     120
```

```
novel_mouse_ECT2  ------------------------------------------------------------
HUMAN_query_ECT2  KQEELIKALKTIKIMEVPVIKIKESCPGKSDEKLIKSVINMDIKVGFVKMESVEEFEGLD
blind_mole_rat    SREELLKAAK-------------------------------EIKAPCVKMDSMEEFGGLD
night_monkey      ----------MFKVFT-------------------------DIKVCFVKMESVEEFEGLD
american_pika     ------------------------------------------------------------
mouse_ear_bat     ------------------------------------------------------------
elephant_seal     KQEELIKALKTIKIMEVPVIKIRESCPGKSDEK--------LIKSVVNMMESLEEFESLD

           61     .      .      .      .      120

          121     .      .      .      .      180
novel_mouse_ECT2  ------------------------------------------------------------
HUMAN_query_ECT2  SPEFENVFVVTDFQDSVFNDLYKADCRVIGPPVVLNCSQKGEPLPFSCRPLYCTSMMNLV
blind_mole_rat    SPEFENVFVVMDFQDPIFDDLYKADCRIIGPPVLLNCARMGEPVPFSCRPLYCASMVGLV
night_monkey      SPEFENVFVVMDFQDSVFNDLYKADCRVIGPPVVLNCSQKGEPLPFSCRPLYCTSMMNLV
american_pika     ------------------------------------------------------------
mouse_ear_bat     ------------------------------------------------------------
elephant_seal     SPEFENVFVVMDFQDSVFNELHKTDYRIIGPPVILNCAQKGE------------------

          121     .      .      .      .      180

          181     .      .      .      .      240
novel_mouse_ECT2  ------------------------------------------------------------
HUMAN_query_ECT2  LCFTGFRKKEELVRLVTLVHHMGGVIRKDFNSKVTHLVANCTQGEKFRVAVSLGTPIMKP
blind_mole_rat    LCFTGFRRKEELVKLVTLVHHMGGVIRKDFNSRVTHLVANCTQGEKFRVAVSLGTPIMKP
night_monkey      LCFTGFRKKEELVRLVTLVHHMGGVIRKDFNSKVTHLVANSTQGEKFRVAVSLGTPIMKP
american_pika     ----------------------MGGVIRKDFNSKVTHLVANNTQGEKFRVAVSLGTPIMKT
mouse_ear_bat     ----------------------MGGVIRKDFNSKVTHLVANCTQGEKFRIAVSLGTPIMKP
elephant_seal     ------------VRLVTLVHHMGGVIRKDFNSKVTHLVANCTQGEKFR------------

          181     .      .      .      .      240

          241     .      .      .      .      300
novel_mouse_ECT2  ------------------------------------------------------------
HUMAN_query_ECT2  EWIYKAWERRNEQDFYAAVDDFRNEFKVPPFQDCILSFLGFSDEEKTNMEEMTEMQGGKY
blind_mole_rat    EWIYKAWEKRNEQDFCAAVDDFRNEFKIPPFQDCILSFLGFSDDERTNMEEMTEMQGGSY
night_monkey      EWIYKAWARRNEQDFCAAVDDFRNEFKVPPFQDCVLSFLGFSDEEKTNMEEMTEMQGGKY
american_pika     EWIYKAWDRRNEQGFCAAVDDFRNEFKVPPFQDCVLSFLGFSEEDKTNMEEMTEMQGGKY
mouse_ear_bat     EWIYKAWERRNEQDFCASADDFRNEFKVPPFQDCILSFLGFSDEEKTNMEEMTEMQGGSY
elephant_seal     ------------------------------------------------------------

          241     .      .      .      .      300

          301     .      .      .      .      360
novel_mouse_ECT2  ------------------------------------------------------------
HUMAN_query_ECT2  LPLGDERCTHLVVEENIVKDLPFEPSKKLYVVKQEWFWGSIQMDARAGETMYLYEKANTP
blind_mole_rat    LAVGDERCTHLVVEENTVKDLPFEPSKKLYVVKQEWFWGSIQMDARAGESMYLYEKANTP
night_monkey      LPLGDERCTHLVVEENIVKDLPFEPSKKLYVVKQEWFWGSIQMDARAGETMYLYEKANTP
american_pika     LPVGDERCTHLIVEENIVKELPFEPAKKLYVVKQEWFWGSIQMDARAGETMYLYEKASTP
mouse_ear_bat     LQVGDERCTHLIVEENTVKELPFEPSKKLYVVKQEWFWGSIQMDARAGETMYLYEKANTP
elephant_seal     ----------------------------------WFWGSIQMDARAGETMYLYEKANTP

          301     .      .      .      .      360

          361     .      .      .      .      420
novel_mouse_ECT2  ------------------------------------------------------------
HUMAN_query_ECT2  ELKKSVSMLSLNTPNSNRKRRRLKETLAQLSRETDVSPFPPRKRPSAEHSLSIGSLLDIS
blind_mole_rat    ELKKSVSLLSLSTPNSNRKRRRLKETLAQLSRETDLSPFPPRKRPSAEHSLSIGSLLDIS
night_monkey      ELKKSVSMLSLNTPNSNRKRRRLKETLAQLSRETDLSPFPPRKRASAEHSLSIGSLLDIS
american_pika     ELKKSVSLLSLSTPNSNRKRRRLKESLAQLSRETDLSPFPPRKRPSAEHSLSIGSLLDIS
mouse_ear_bat     ELKKSVSLLSLSTPNSNRKRRRLKETLAQLSRETDLSPFPPRKRPSAEHSLSIGSLLDIS
elephant_seal     ELKKSVSLLSLNTPNSNRKRRRLKETLAQLSRETDMSPFPPRKRPSAEHSLSIGSLLDIS

          361     .      .      .      .      420

          421     .      .      .      .      480
novel_mouse_ECT2  ----------SRSSTPVPPKQSVRWQVAKELYQTESNYVNILATIIQLFQV
HUMAN_query_ECT2  NTPESSINYGDTPKSCTKSSKSSTPVPSKQSARWQVAKELYQTESNYVNILATIIQLFQV
blind_mole_rat    NTPESSINYGETPKSCTKSSRNSTPVPPKQSARWQVAKELYQTESNYVNILATIIQLFQV
night_monkey      NTPESSINYGETPKSCTKSSKNSTPVPSKQSARWQVAKELYQTESNYVNILATIIQLFQV
```

```
american_pika     NTPDSSINYGETPKSCTKSSKNSTPVPLKQSARWQVAKELYQTESNYVNILATIIQLFQV
mouse_ear_bat     NTPESSINYGETPKSCTKSSKNSTPVPSKQSARWQVAKELYQTESNYVNILATIIQLFQV
elephant_seal     NTPESSINYGETPKSCTKSSKNSTPVPSKQSARWQVAKELYQTESNYVNILATIIQLFQV
                    *^ ***** *** ****************************
          421     .     .     .     .     480

          481     .     .     .     .     540
novel_mouse_ECT2  PLEEEGQRGGPILAPEEIKTIFGSIPDIFDVHMKIKDDLEDLIANWDESRSIGDIFLKYA
HUMAN_query_ECT2  PLEEEGQRGGPILAPEEIKTIFGSIPDIFDVHTKIKDDLEDLIVNWDESKSIGDIFLKYS
blind_mole_rat    PLEEEGQRGGPILAPEEIKTIFGSIPDIFDVHMKIKDDLEDLIVNWDENKSIGDIFLKYS
night_monkey      PLEEEGQRGGPILAPEEIKTIFGSIPDIFDVHTKIKDDLEDLIVNWDESKSIGDIFLKYS
american_pika     PLEEEGQRGGPILAPEEIKTIFGSIPDIFDVHTKIKDDLEDLIVNWDESKSIGDIFLKYS
mouse_ear_bat     PLEEEGQRGGPILAPEEIKTIFGSIPDIFDVHTKIKDDLEDLIVNWDESKSIGDIFLKYS
elephant_seal     PLEEEGQRGGPILAPEEIKTIFGSIPDIFDVHTKIKDDLEDLIVNWDESKSIGDIFLKYS
                   ******************************** ********** **** ^*********
          481     .     .     .     .     540

          541     .     .     .     .     600
novel_mouse_ECT2  KDLVKTYPPFVNFFEMSKEMIIKCEKQKPRFHAFLKINQAKPECGRQSLVELLIRPVQRL
HUMAN_query_ECT2  KDLVKTYPPFVNFFEMSKETIIKCEKQKPRFHAFLKINQAKPECGRQSLVELLIRPVQRL
blind_mole_rat    KDLVKTYPPFVNFFEMSKETIIKCEKQKPRFHAFLKINQAKPECGRQSLVELLIRPVQRL
night_monkey      KDLVKTYPPFVNFFEMSKETIIKCEKQKPRFHAFLKINQAKPECGRQSLVELLIRPVQRL
american_pika     KDLIKTYPPFVNFFEMSKETIIKCEKQKPRFHAFLKINQAKPECGRQSLVELLIRPVQRL
mouse_ear_bat     KDLVKTYPPFVNFFEMSKETIIKCEKQKPRFHAFLKINQAKPECGRQSLVELLIRPVQRL
elephant_seal     KDLVKTYPPFVNFFEMSKETIVKCEKQKPRFHAFLKINQAKPECGRQSLVELLIRPVQRL
                   ***^*************** *^************************************
          541     .     .     .     .     600

          601     .     .     .     .     660
novel_mouse_ECT2  PSVALLLNDLKKHTADENPDKSTLEKAIGSLKEVMTHINEDKRKTEAQKQIFDVVYEVDG
HUMAN_query_ECT2  PSVALLLNDLKKHTADENPDKSTLEKAIGSLKEVMTHINEDKRKTEAQKQIFDVVYEVDG
blind_mole_rat    PSVALLLNDLKKHTADENPDKSTLEKAIGSLKEVMTHINEDKRKTEAQKQIFDVVYEVDG
night_monkey      PSVALLLNDLKKHTADENPDKSTLEKAIGSLKEVMTHINEDKRKTEAQKQIFDVVYEVDG
american_pika     PSVALLLNDLKKHTVDENPDKSTLEKAIGSLKEVMTHINEDKRKTEAQKQIFDVVYEVDG
mouse_ear_bat     PSVALLLNDLKKHTAEENPDKSTLEKAIGSLKEVMTHINEDKRKTEAQKQIFDVVYEVDG
elephant_seal     PSVALLLNDLKKHTAEENPDKSTLEKAIGSLKEVMTHINEDKRKTEAQKQIFDVVYEVDG
                   ************** ^*********************************************
          601     .     .     .     .     660

          661     .     .     .     .     720
novel_mouse_ECT2  CPANLLSSHRSLVQRVETVSLGEHPCDRGEQVTLFPL---MTASRQESGTKLLALLESLT
HUMAN_query_ECT2  CPANLLSSHRSLVQRVETISLGEHPCDRGEQVTLFLFNDCLEIARKRH--KVIGTFRSPH
blind_mole_rat    CPANLLSSHRSLVQRVETVSLGEQPCDRGEQVTLFLFNDCLEIARKRH--KVIGTFRSPH
night_monkey      CPANLLSSHRSLVQRVETISLGEHPCDRGEQVTLFLFNDCLEIARKRH--KVIGTFRSPH
american_pika     CPANLLSSHRSLIQRVETISLGEHPCDRGEQVTLFLFNDCLEIARKRH--KVIGTFRSPH
mouse_ear_bat     CPANLLSSHRSLIQRVETISLGEHPCDRGEQVTLFLFNDCLEIARKRH--KVIGTFRSPH
elephant_seal     CPANLLSSHRSLVQRVETVSLGEHPCDRGEQVTLFLFNDCLEIARKRH--KVIGTFRSPP
                   ***********^******^**** ***********    ^  *    *^^^  *
          661     .     .     .     .     720

          721     .     .     .     .     780
novel_mouse_ECT2  NAPGPPLLSTFISCLFL-RLKRAGHPRD-KRLSPCLCPACKPXTEQAMYCSTFK-RPRTF
HUMAN_query_ECT2  GQTRPPASLKHIHLMPLSQIKKVLDIRETEDCHNAFALLVRPPTEQANVLLSFQMTSDEL
blind_mole_rat    GHTRPPASLKHIHLMPLSQIKKVLDIRETEDCRNAFALLVRPPTEQANVLLSFQMTSDEL
night_monkey      GQTRPPASLKHIHLMPLSQIKKVLDIRETEDCHNAFALLVRPPTEQANVLLSFQMTSDEL
american_pika     GHTRPPASLKHVRLMPLSQIKKVLDIRETEDCHNAFALLVRPPTEQANVLLSFQMTSYDL
mouse_ear_bat     GQTRPPASLKHIHLMPLSQIKKVLDIRETEDCHNAFALLVRPPTEQANVLLSFQMTSEEL
elephant_seal     GHTRPPASLKHIHLMPLSQIKKVLDIRETEDCHNAFALLVRPPTERANVLLSFQMTSEEL
                    **   ^ ^* ^*^   *^       ^* ** *   ^*
          721     .     .     .     .     780

          781     .     .     .     .     840
novel_mouse_ECT2  PKETALKSL-PTFTQPFVRAKLEIFWMGLIP----IPLRKTKDGIAIGAR----------
HUMAN_query_ECT2  PKENWLKMLCRHVANTICKADAENLIYTADPESFEVNTKDMDSTLSRASRAIKKTSKKVT
blind_mole_rat    PKENWLKMLCRHVANTICKADAENLIYTADPESFEVNTKDMDSTLSRASRAIKKTSKKVT
night_monkey      PKENWLKMLCRHVANTICKADAENLIYTADPETFEVNTKDMDSTLSRASRAIKKTSKKVT
american_pika     PKENWLKMLCRHVANTICKADAENLMYPADPESFEVNTKDMDSTLSRASRAIKKTSKKVT
mouse_ear_bat     PKENWLKMLCRHVANTICKADAENLIYTADPETFEVNTKDMDSTLSRASRAIKKTSKRVT
elephant_seal     PKENWLKMLCRHVANTICKAD----------------------------------------
                   *** ** *   ^  ^*
```

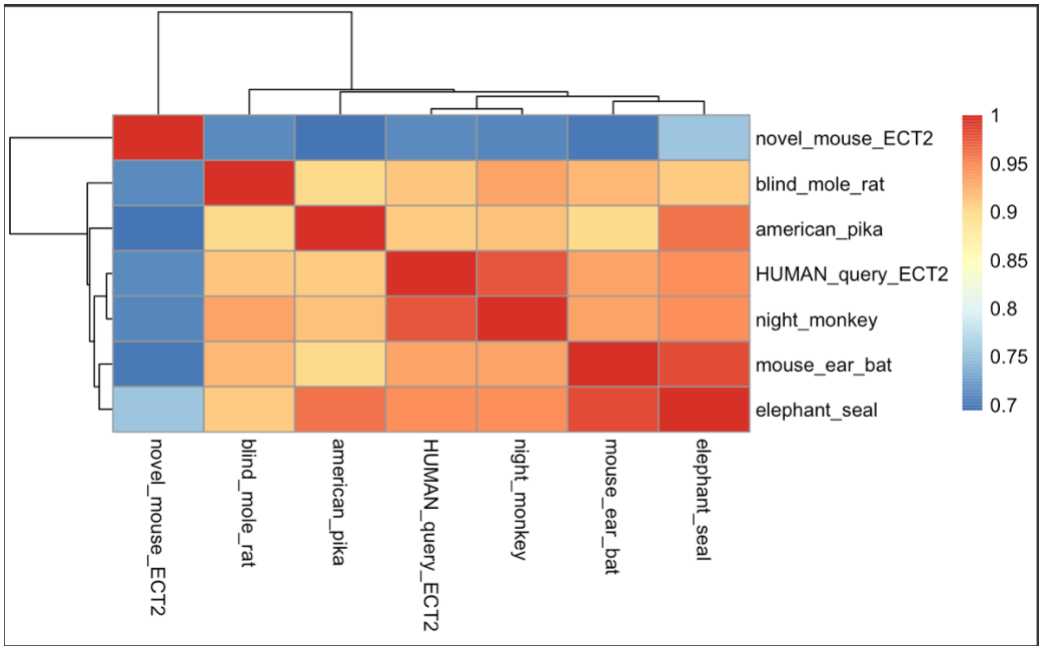
```
          781      .       .       .       .       840

          841      .       .       .       .       900
novel_mouse_ECT2   --FTIKKTPKRVRG----------------------------------------------
HUMAN_query_ECT2   RAFSFSKTPKRALRRALMTSHGSVEGRSPSSNDKHVMSRLSSTSSLAGIPSPSLVSL---
blind_mole_rat     RAFSFSKTPKRALRLALMSSHSSSEARSPPSSDKHAVSRLSSTSSLAGISSPSLVSL---
night_monkey       RAFSFSKTPKRALRRALMTSHSSGEGRSPSSSDKHGMSRLSSTSSLAGIPSPSLVSL---
american_pika      RAFSFSKTPKRAFRMTLTTSHSSAEGRSPTSSDKLAVSRLPSTSSLAITHSVSTSNIIGF
mouse_ear_bat      RAFSFSKTPRRVLRRALMTSQSSVEGRSPLSSDKHVMSRMSSTSSLAIIHSVSTSSTIGF
elephant_seal      ------------------------------------------------------------

          841      .       .       .       .       900

          901      .       .       .       .       960
novel_mouse_ECT2   -----------------VSFFQX-----------------------------------
HUMAN_query_ECT2   -----------------PSFFERRSHTLSRSTTHLI-----------------------
blind_mole_rat     -----------------PSFFERRSHTLSRSTTHLI-----------------------
night_monkey       -----------------PSFFERRSHTLSRSTTHLI-----------------------
american_pika      TRHSYGQRSNSTRGHSRSSWFRSIHHSSSQASFSEILEGNTDFQISQKFYPTHL------
mouse_ear_bat      TKHVYAQCSHSTGGRSQNSWFPSIRHSASRVSFSETLKENIDFSNFKKSSIQVIFGICEE
elephant_seal      ------------------------------------------------------------

          901      .       .       .       .       960

          961      .  974
novel_mouse_ECT2   --------------
HUMAN_query_ECT2   --------------
blind_mole_rat     --------------
night_monkey       --------------
american_pika      --------------
mouse_ear_bat      LRDMPTSNGNKPQV
elephant_seal      --------------

          961      .  974
```

[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.



[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.



[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).

HINT: You can use a single sequence from your alignment or generate a consensus sequence from your alignment using the Bio3D function consensus(). The Bio3D functions blast.pdb(), plot.blast() and pdb.annotate() are likely to be of most relevance for completing this task. Note that the results of blast.pdb() contain the hits PDB identifier (or pdb.id) as well as Evalue and identity. The results of pdb.annotate() contain the other annotation terms noted above.

Note that if your consensus sequence has lots of gap positions then it will be better to use an original sequence from the alignment for your search of the PDB. In this case you could chose the sequence with the highest identity to all others in your alignment by calculating the row-wise maximum from your sequence identity matrix.

I used the human query sequence to do the blast.pdb() function as my consensus sequence

and novel sequence had lots of gaps and could not connect to blast when using the function. So, I used the human sequence for the blast, with the top 3 results displayed below.

queryid <chr>	subjectids <chr>	identity <dbl>	evalue <dbl>	macromoleculeType <chr>	chainLength <int>	experimentalTechnique <chr>	resolution <dbl>	pfam <chr>	source <chr>
Query_7842035	6L30_A	100.000	0.00e+00	Protein	707	X-ray	2.800	BRCA1 C Terminus (BRCT) domain (BRCT)	Homo sapiens
Query_7842035	4N40_A	86.538	0.00e+00	Protein	281	X-ray	3.106	BRCA1 C Terminus (BRCT) domain (BRCT)	Homo sapiens
Query_7842035	3L46_A	100.000	1.73e-61	Protein	112	X-ray	1.482	BRCA1 C Terminus (BRCT) domain (BRCT)	Homo sapiens

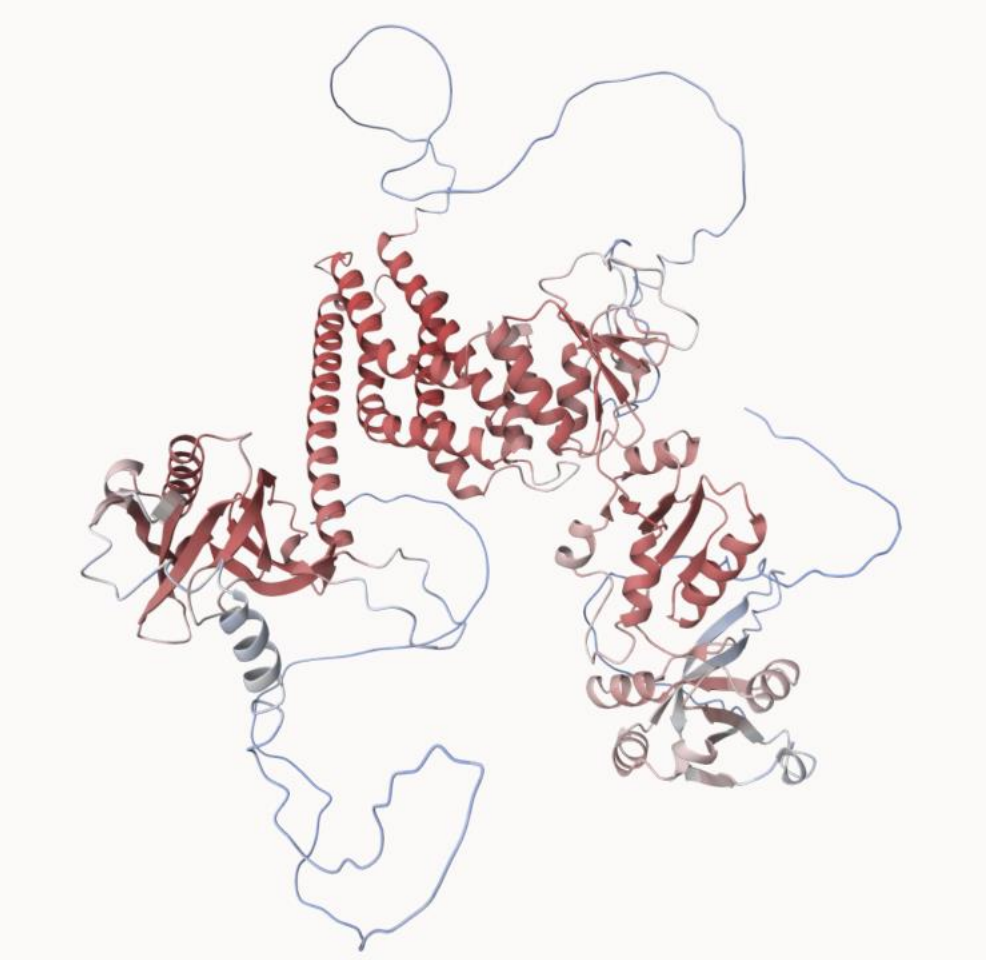
3 rows

[**Q9**] Using AlphaFold notebook generate a structural model using the default parameters for your novel protein sequence.

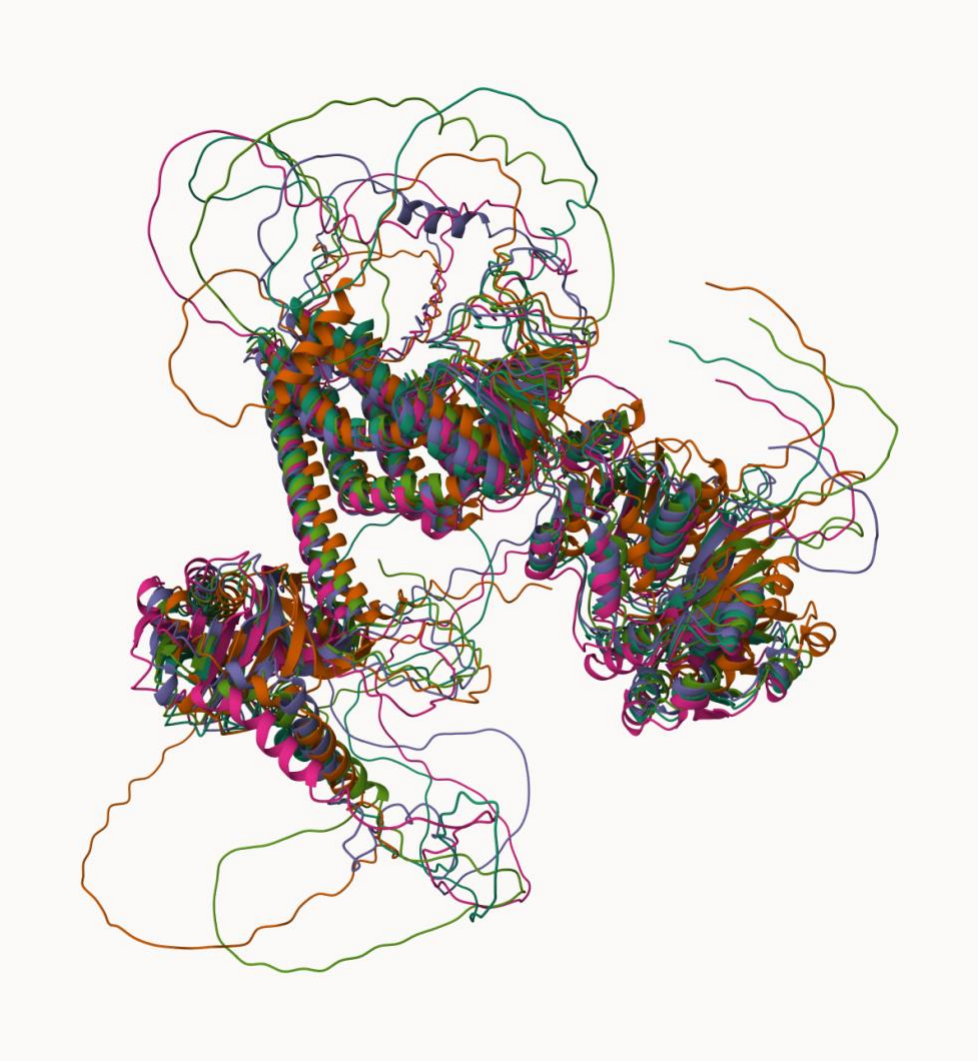
Note that this can take some time depending upon your sequence length. If your model is taking many hours to generate or your input sequence yields a “too

many amino acids" (i.e. length) error you can focus on a single domain from your sequence - identify region by searching for PFAM domain matches.

Once complete save the resulting PDB format file for your records. Finally, generate a molecular figure of your generated PDB structure using the **Mol* viewer** online (or VMD/PyMol/Chimera if you prefer). To complete your analysis you can optionally highlight *conserved residues* that are likely to be functional as **spacefill** and the protein as **cartoon** colored by local alpha fold *pLDDT quality score*. This score is contained in the B-factor column of your PDB downloaded file. Please use a white or transparent background for your figure (i.e. not the default black in PyMol/VMD/Chimera etc.).



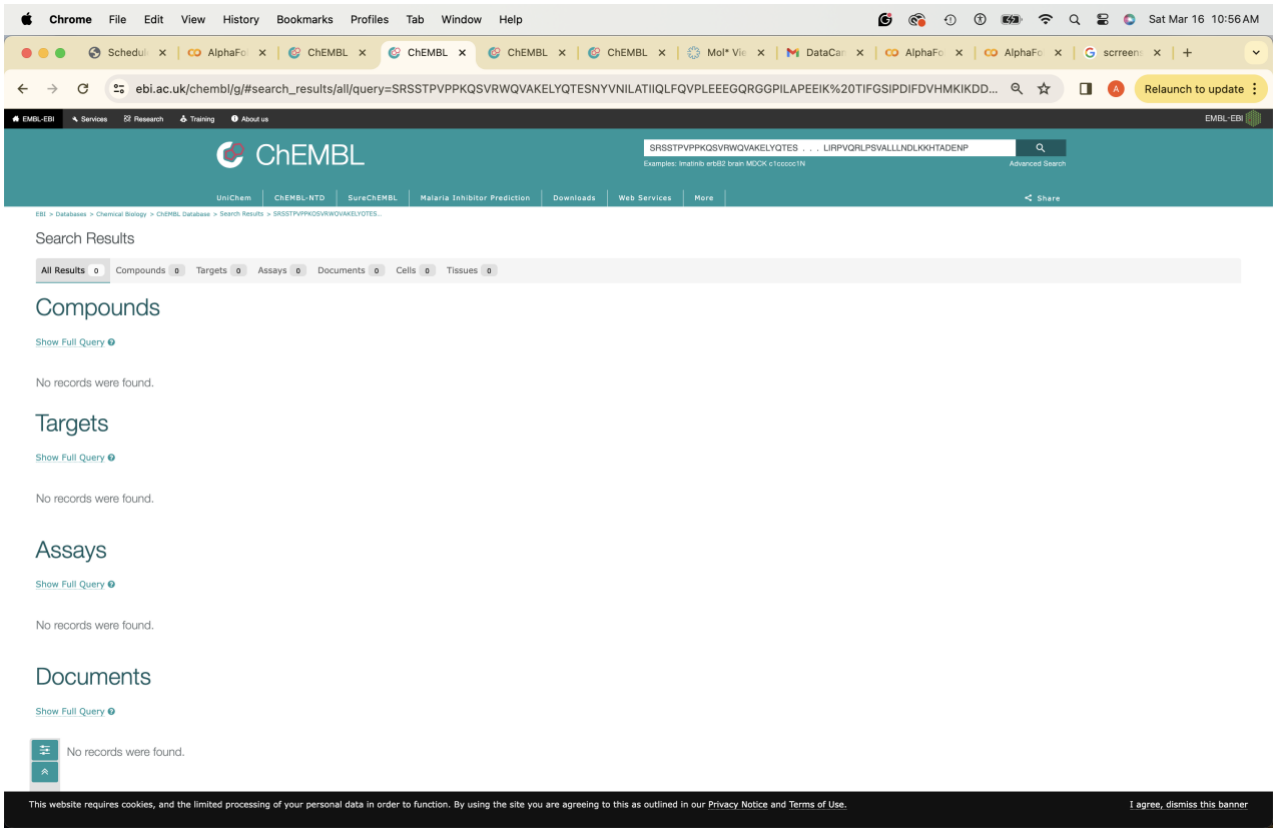
Colored by residue uncertainty



Top 5 prediction superimposed on top of each other.

[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? If there are no assays listed here simply list “non available as of [date]”.



None were found.

Scoring Rubric: [50 total points available]

Q1 (4 points)

Protein name	1
Species	1
Accession number	1
Function known	1

Q2 (6 points)

Blast method	1
Database searched	1
Limits applied	1
Search output list (top hits)	1
Alignment of choice	1
Evalue and other alignment stats	1

Q3 (3 points)

Protein sequence of choice matches Subject above 1

Name in header	1
Species	1

Q4 (3 point)

Blastp output list with identities & Evalue	1
Top alignment shown with alignment statistics	1
Results indicates a “novel” gene found	1

Q5 (3 points)

MSA labeled with useful names	1
MSA trimmed appropriately (i.e. no gap overhangs)	1
Pasted MSA fits report page width (i.e. font, format)	1

Q6 (1 point)

Figure illustrates sequence clustering pattern	1

Q7 (10 points)

Heatmap figure included in report	5
Heatmap is legible (i.e. no labels obscured)	5

Q8 (9 points)

PDB identifiers from multiple species reported	5
Annotation of PDB source, resolution and technique	4
Annotation of Evalue and Sequence Identity	1

Q9 (10 points)

Structure figure provided	2
Uses white background for molecular figure	1
Figure of high resolution (i.e. not just snapshot)	1
Conserved residues as spacefill	3
Protein cartoon colored by pLDDT quality score	3

Q10 (1 point)

Evidence of ChEMBEL searches	1