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.

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 the beginning of **Week 5**. Note that these answers can be obtained very quickly (at best within 10 or 15 minutes), so if you don't succeed at first, just keep trying.

The complete assignment, including responses to all questions, is due at the beginning of **Week 10**. Late responses will not be accepted under any circumstances.

Submission instructions:

Submit your PDF document to GradeScope as directed on our class website. Please do make sure your document is in PDF format and named something like `BGGN213_F20_[yourUCSDname].pdf` for example, my document would be named `BGGN213_F20_bjgrant.pdf`

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 at the beginning of **week 5** 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. I will email you my decision; proceed with subsequent questions only after we are sure you have found a novel gene.

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

Samuel Rivera

PID: A53272335

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.

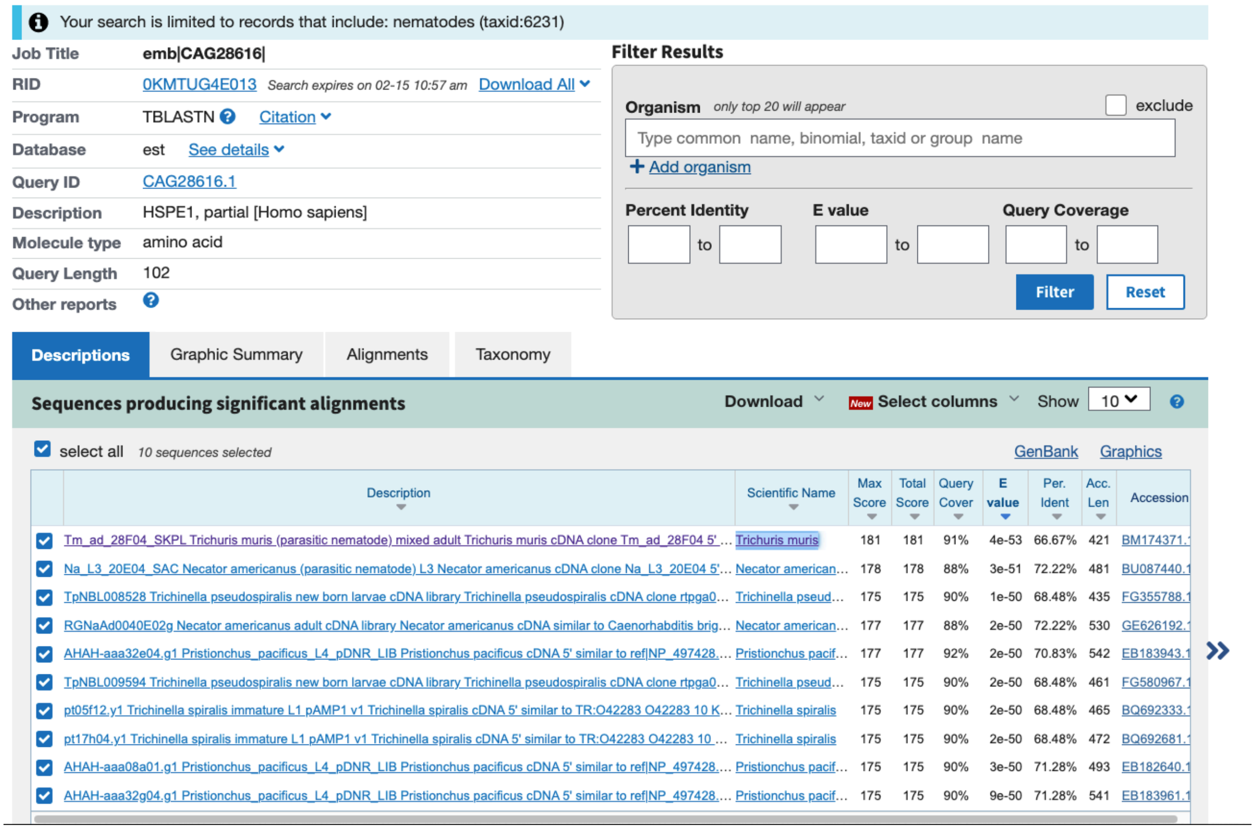
If you do not have a favorite protein, select human RBP4 or KIF11. Do not use beta globin as this is in the worked example report that I provide you with online.

```
Name: HSPE1
ACCESSION   CAG28616
ORGANISM  Homo sapiens
Function: This gene encodes a major heat shock protein which functions as a
chaperonin.
```

[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
Database: Expressed Sequence Tags
Organism: Nematodes (Taxid: 6231)

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.



Your search is limited to records that include: nematodes (taxid:6231)

Job Title	emb|CAG28616|
RID	0KMTUG4E013 *Search expires on 02-15 10:57 am* Download All
Program	TBLASTN Citation
Database	est See details
Query ID	CAG28616.1
Description	HSPE1, partial [Homo sapiens]
Molecule type	amino acid
Query Length	102
Other reports	

Filter Results

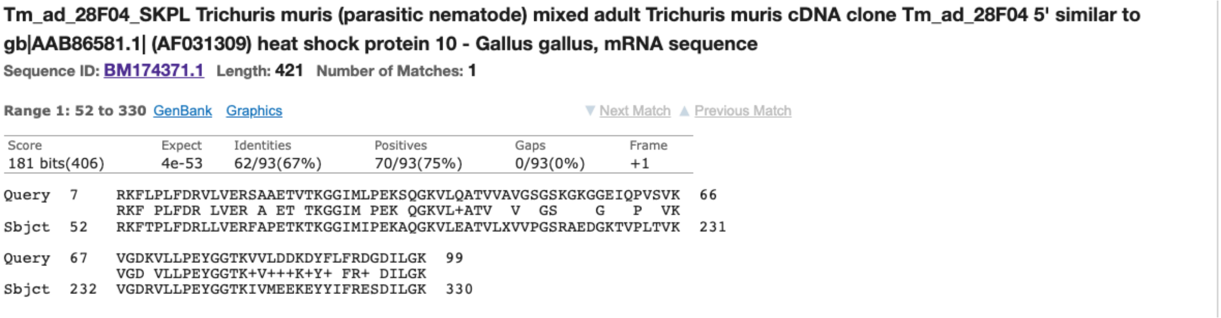
Sequences producing significant alignments

☑ select all *10 sequences selected*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Tm_ad_28F04_SKPL Trichuris muris (parasitic nematode) mixed adult Trichuris muris cDNA clone Tm_ad_28F04 5' ...	Trichuris muris	181	181	91%	4e-53	66.67%	421	BM174371.1
☑	Na_L3_20E04_SAC Necator americanus (parasitic nematode) L3 Necator americanus cDNA clone Na_L3_20E04 5'...	Necator american...	178	178	88%	3e-51	72.22%	481	BU087440.1
☑	TpNBL008528 Trichinella pseudospiralis new born larvae cDNA library Trichinella pseudospiralis cDNA clone rtpga0...	Trichinella pseud...	175	175	90%	1e-50	68.48%	435	FG355788.1
☑	RGNaAd0040E02g Necator americanus adult cDNA library Necator americanus cDNA similar to Caenorhabditis brig...	Necator american...	177	177	88%	2e-50	72.22%	530	GE626192.1
☑	AHAH-aaa32e04.g1 Pristionchus_pacificus_L4_pDNR_LIB Pristionchus pacificus cDNA 5' similar to ref|NP_497428....	Pristionchus pacif...	177	177	92%	2e-50	70.83%	542	EB183943.1
☑	TpNBL009594 Trichinella pseudospiralis new born larvae cDNA library Trichinella pseudospiralis cDNA clone rtpga0...	Trichinella pseud...	175	175	90%	2e-50	68.48%	461	FG580967.1
☑	pt05f12.y1 Trichinella spiralis immature L1 pAMP1 v1 Trichinella spiralis cDNA 5' similar to TR:O42283 O42283 10 K...	Trichinella spiralis	175	175	90%	2e-50	68.48%	465	BQ692333.1
☑	pt17h04.y1 Trichinella spiralis immature L1 pAMP1 v1 Trichinella spiralis cDNA 5' similar to TR:O42283 O42283 10 ...	Trichinella spiralis	175	175	90%	2e-50	68.48%	472	BQ692681.1
☑	AHAH-aaa08a01.g1 Pristionchus_pacificus_L4_pDNR_LIB Pristionchus pacificus cDNA 5' similar to ref|NP_497428....	Pristionchus pacif...	175	175	90%	3e-50	71.28%	493	EB182640.1
☑	AHAH-aaa32g04.g1 Pristionchus_pacificus_L4_pDNR_LIB Pristionchus pacificus cDNA 5' similar to ref|NP_497428....	Pristionchus pacif...	175	175	90%	9e-50	71.28%	541	EB183961.1

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: BM174371



Tm_ad_28F04_SKPL Trichuris muris (parasitic nematode) mixed adult Trichuris muris cDNA clone Tm_ad_28F04 5' similar to gb|AAB86581.1| (AF031309) heat shock protein 10 - Gallus gallus, mRNA sequence

Sequence ID: BM174371.1 Length: 421 Number of Matches: 1

Range 1: 52 to 330 GenBank Graphics

```
Score          Expect   Identities    Positives     Gaps       Frame
181 bits(406)  4e-53    62/93(67%)    70/93(75%)    0/93(0%)   +1

Query  7    RKFLPLFDRVLVERSAAETVTKGGIMLPEKSQGKVLQATVVAVGSGSKGKGGEIQPVSVK  66
            RKF PLFDR LVER A ET TKGGIM PEK QGKVL+ATV  V  GS    G   P  VK
Sbjct  52   RKFTPLFDRLLVERFAPETKTKGGIMIPEKAQGKVLEATVLXVVPGSRAEDGKTVPLTVK  231

Query  67   VGDKVLLPEYGGTKVVLDDKDYFLFRDGDILGK  99
            VGD VLLPEYGGTK+V+++K+Y+ FR+ DILGK
Sbjct  232  VGDRVLLPEYGGTKIVMEEKEYYIFRESDILGK  330
```

[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:

>Tm_ad_28F04_SKPL Trichuris muris (parasitic nematode) mixed adult Trichuris muris cDNA clone Tm_ad_28F04 5' similar to gb|AAB86581.1| (AF031309) heat shock protein 10 - Gallus gallus, mRNA sequence-
RKFTPLFDRLLVERFAPETKTKGGIMIPEKAQGKVLEATVLXVVPGSRAEDGKTVPLTV
KVGDRVLLPEYGGTKIVMEEKEYYIFRESDILGK

Name- Tm_ad_28F04_SKPL Trichuris muris (parasitic nematode) mixed adult Trichuris muris cDNA clone Tm_ad_28F04 5' similar to gb|AAB86581.1| (AF031309) heat shock protein 10 - Gallus gallus, mRNA sequence

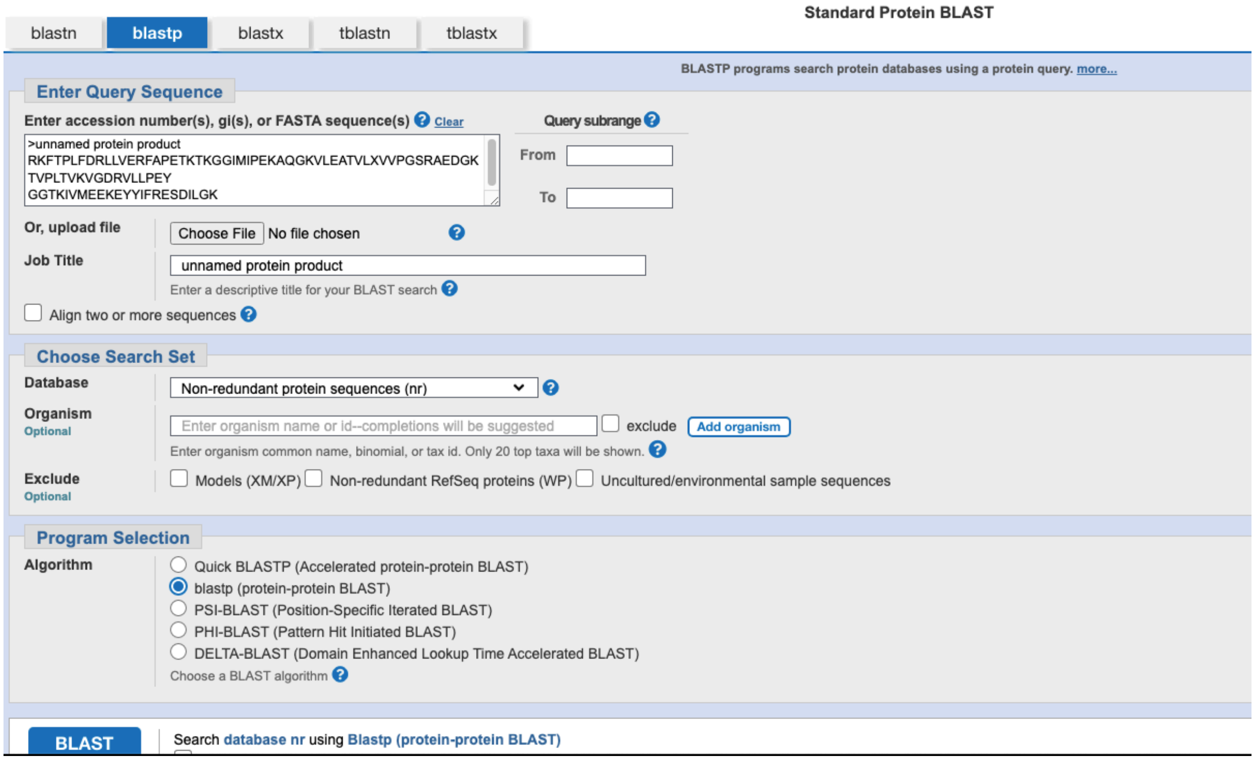
Species: Trichuris muris

[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.

BLASTP against NR database, no perfect match, top match from Trichuris suis, see alignment below.



Standard Protein BLAST

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

```
>unnamed protein product
RKFTPLFDRLLVERFAPETKTKGGIMIPEKAQGKVLEATVLXVVPGSRAEDGK
TVPLTVKVGDRVLLPEY
GGTKIVMEEKEYYIFRESDILGK
```

Query subrange

From

To

Or, upload file: Choose File No file chosen

Job Title: unnamed protein product

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism (Optional): Enter organism name or id--completions will be suggested □ exclude Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude (Optional): □ Models (XM/XP) □ Non-redundant RefSeq proteins (WP) □ Uncultured/environmental sample sequences

Program Selection

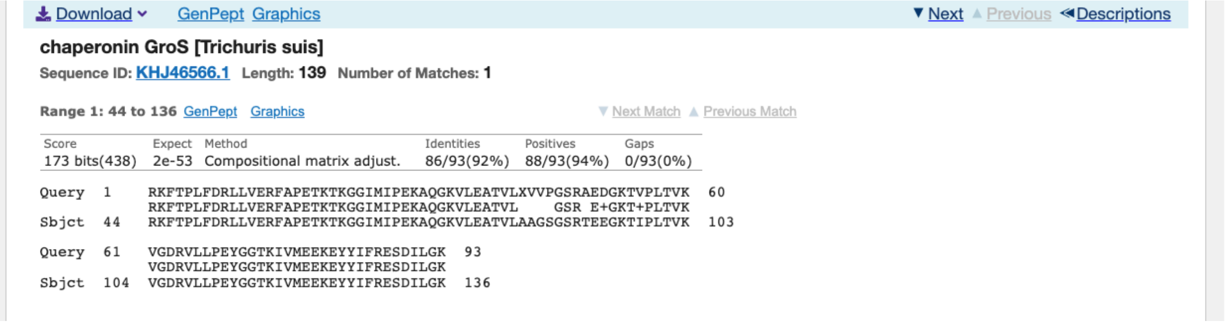
Algorithm:

○ Quick BLASTP (Accelerated protein-protein BLAST)
● blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST

Search database nr using Blastp (protein-protein BLAST)



Download
GenPept Graphics
▼ Next ▲ Previous ◀ Descriptions

chaperonin GroS [Trichuris suis]

Sequence ID: KHJ46566.1 Length: 139 Number of Matches: 1

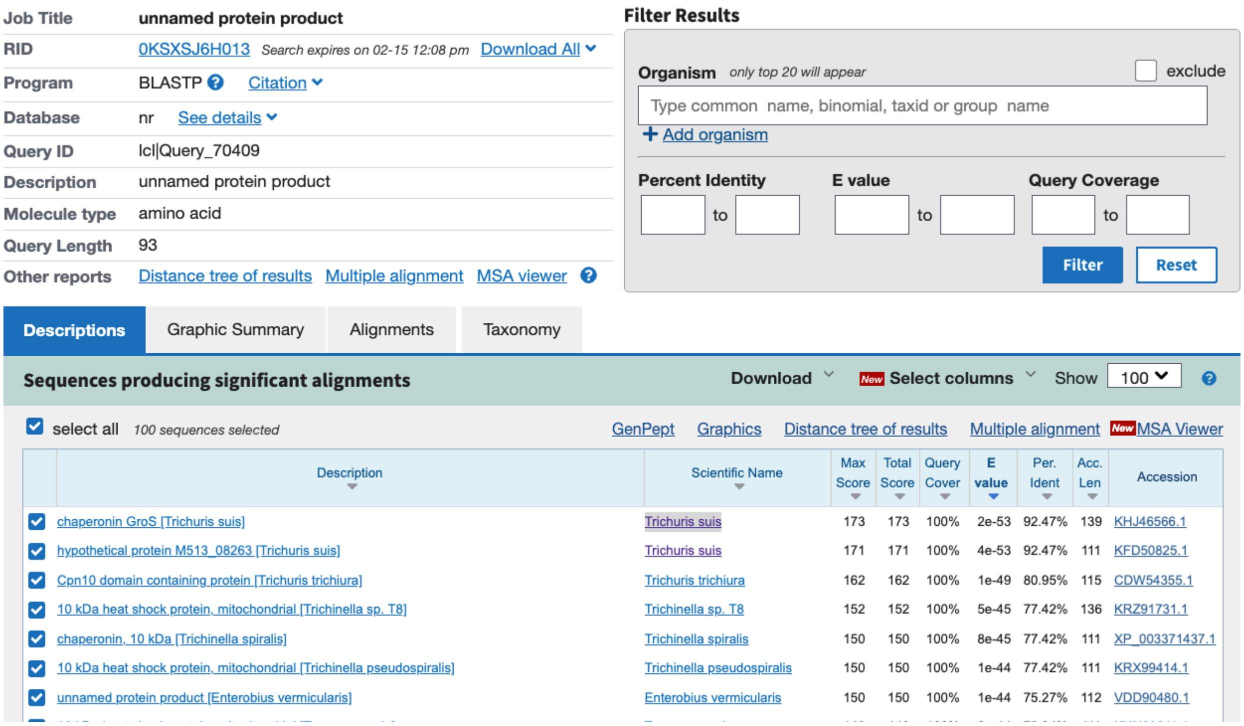
Range 1: 44 to 136 GenPept Graphics

▼ Next Match ▲ Previous Match

```
Score          Expect  Method                         Identities   Positives    Gaps
173 bits(438)  2e-53   Compositional matrix adjust.   86/93(92%)   88/93(94%)   0/93(0%)

Query  1    RKFTPLFDRLLVERFAPETKTKGGIMIPEKAQGKVLEATVLXVVPGSRAEDGKTVPLTVK  60
            RKFTPLFDRLLVERFAPETKTKGGIMIPEKAQGKVLEATVL    GSR E+GKT+PLTVK
Sbjct  44   RKFTPLFDRLLVERFAPETKTKGGIMIPEKAQGKVLEATVLAAGSGSRTEEGKTIPLTVK  103

Query  61   VGDRVLLPEYGGTKIVMEEKEYYIFRESDILGK  93
            VGDRVLLPEYGGTKIVMEEKEYYIFRESDILGK
Sbjct  104  VGDRVLLPEYGGTKIVMEEKEYYIFRESDILGK  136
```



Job Title	unnamed protein product
RID	0KSXSJ6H013 *Search expires on 02-15 12:08 pm* Download All
Program	BLASTP Citation
Database	nr See details
Query ID	lcl|Query_70409
Description	unnamed protein product
Molecule type	amino acid
Query Length	93
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 [ ]

Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	chaperonin GroS [Trichuris suis]	Trichuris suis	173	173	100%	2e-53	92.47%	139	KHJ46566.1
☑	hypothetical protein M513_08263 [Trichuris suis]	Trichuris suis	171	171	100%	4e-53	92.47%	111	KFD50825.1
☑	Cpn10 domain containing protein [Trichuris trichiura]	Trichuris trichiura	162	162	100%	1e-49	80.95%	115	CDW54355.1
☑	10 kDa heat shock protein, mitochondrial [Trichinella sp. T8]	Trichinella sp. T8	152	152	100%	5e-45	77.42%	136	KRZ91731.1
☑	chaperonin, 10 kDa [Trichinella spiralis]	Trichinella spiralis	150	150	100%	8e-45	77.42%	111	XP_003371437.1
☑	10 kDa heat shock protein, mitochondrial [Trichinella pseudospiralis]	Trichinella pseudospiralis	150	150	100%	1e-44	77.42%	111	KRX99414.1
☑	unnamed protein product [Enterobius vermicularis]	Enterobius vermicularis	150	150	100%	1e-44	75.27%	112	VDD90480.1

[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.

>Human chaperonin 10-related protein, partial [Homo sapiens]

AGQAFRKFLPLFDRVLVERSAAETVTKGGIMLPEKSQGKVLQATVVAVGSGSKGKGG
EIQPVSVKVGDKVLLPEYGGTKVVLDDKDYFLFRDGDILG

>Trichuris muris (parasitic nematode) mixed adult Trichuris muris cDNA clone Tm_ad_28F04 5' similar to gb|AAB86581.1| (AF031309) heat shock protein 10 - Gallus gallus, mRNA sequence-
RKFTPLFDRLLVERFAPETKTKGGIMIPEKAQGKVLEATVLXVVPGSRAEDGKTVPLTV
KVGDRVLLPEYGGTKIVMEEKEYYIFRESDILGK

>Mouse chaperonin 10 [Mus musculus]

MAGQAFRKFLPLFDRVLVERSAAETVTKGGIMLPEKSQGKVLQATVVAVGSGGKGKS
GEIEPVSVKVGDKVLLPEYGGTKVVLDDKDYFLFRDSDILGKYVD

>Rat heat shock 10 kDa protein 1 (chaperonin 10) [Rattus norvegicus]

MAGQAFRKFLPLFDRVLVERSAAETVTKGGIMLPEKSQGKVLQATVVAVGSGGKGKG
GEIQPVSVKVGDKVLLPEYGGTKVVLDDKDYFLFRDGDILGKYVD

>Zebra Fish chaperonin Cpn10, partial [Danio rerio]

MQAFRKFLPMFDRVLVERLAAETVSRGGIMIPEKSQAKVLQATVVAVGPG

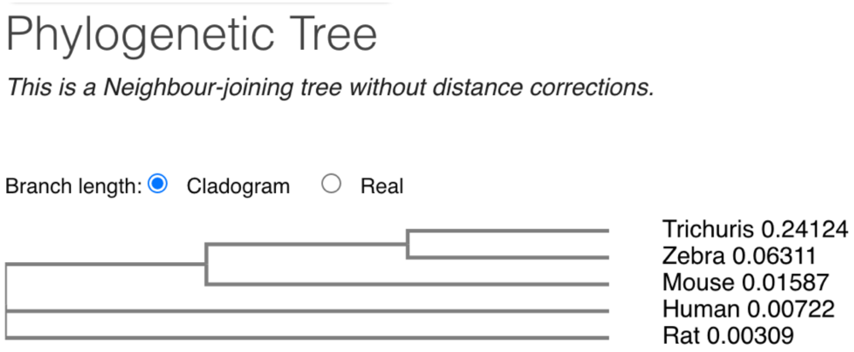
```
Alignment:
EBI MUSCLE:

CLUSTAL multiple sequence alignment by MUSCLE (3.8)


Trichuris       ------RKFTPLFDRLLVERFAPETKTKGGIMIPEKAQGKVLEATVLXVVPGSRAEDGKT
Mouse           MAGQAFRKFLPLFDRVLVERSAAETVTKGGIMLPEKSQGKVLQATVVAVGSGGKGKSGEI
Human           -AGQAFRKFLPLFDRVLVERSAAETVTKGGIMLPEKSQGKVLQATVVAVGSGSKGKGGEI
Rat             MAGQAFRKFLPLFDRVLVERSAAETVTKGGIMLPEKSQGKVLQATVVAVGSGGKGKGGEI
Zebra           --MQAFRKFLPMFDRVLVERLAAETVSRGGIMIPEKSQAKVLQATVVAVGPG--------
                      *** *:***:**** *.** :.****:***:*.***:***: * .*

Trichuris       VPLTVKVGDRVLLPEYGGTKIVMEEKEYYIFRESDILGK---
Mouse           EPVSVKVGDKVLLPEYGGTKVVLDDKDYFLFRDSDILGKYVD
Human           QPVSVKVGDKVLLPEYGGTKVVLDDKDYFLFRDGDILG----
Rat             QPVSVKVGDKVLLPEYGGTKVVLDDKDYFLFRDGDILGKYVD
Zebra           ------------------------------------------
```

[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.



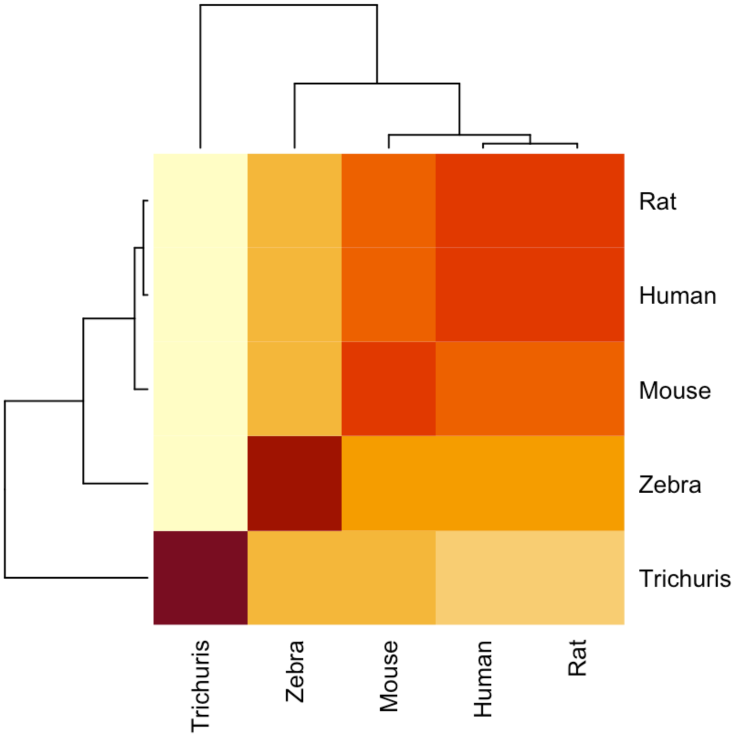
Phylogenetic Tree

This is a Neighbour-joining tree without distance corrections.

Branch length: ● Cladogram ○ Real

Trichuris 0.24124
Zebra 0.06311
Mouse 0.01587
Human 0.00722
Rat 0.00309

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



[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).

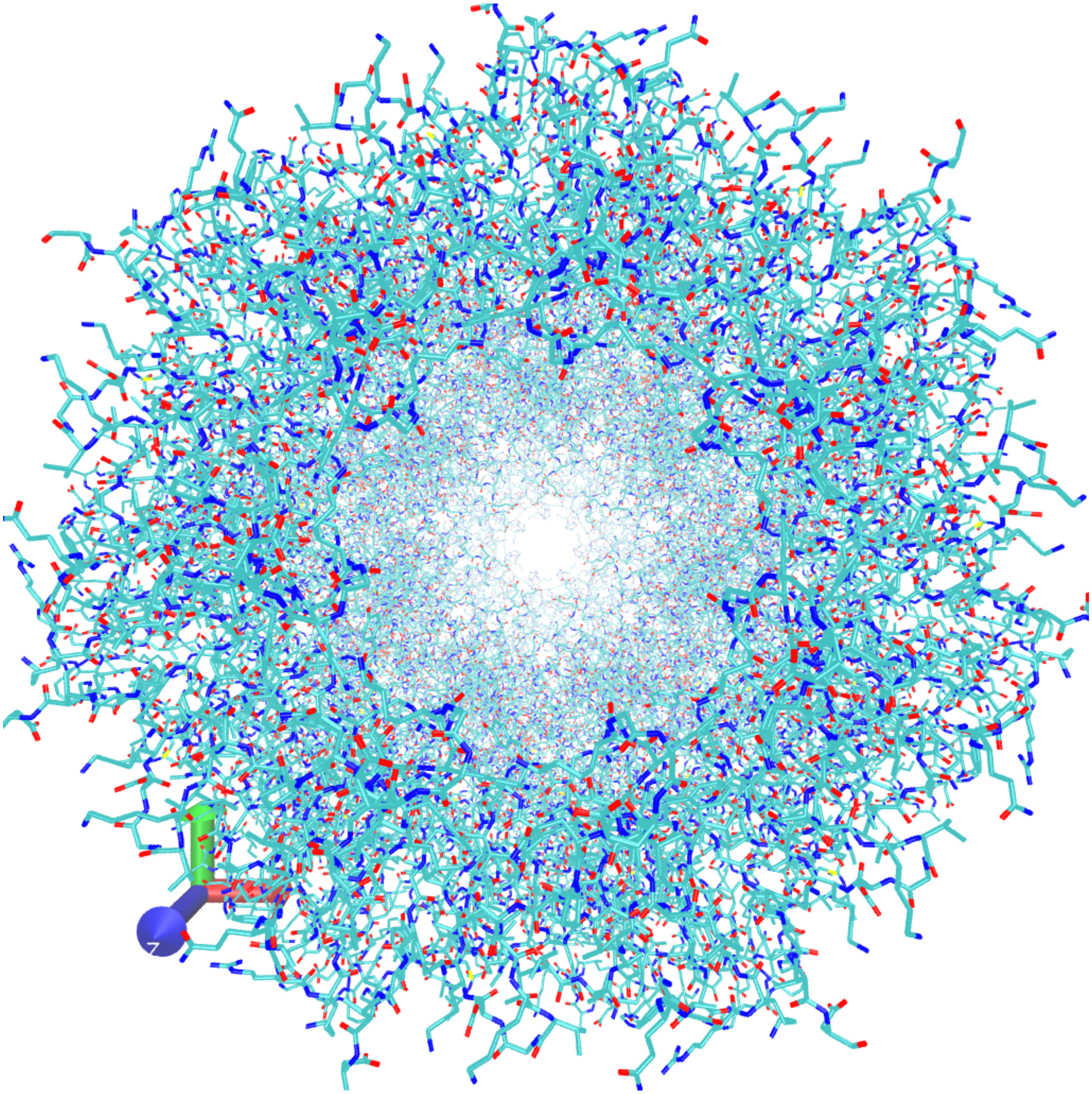
I used the Mus musculus sequence as it had the highest identity to the others.

ID	Technique	Resolution	Source	E value	Identity
Q64433	Electron Microscopy	2.9	Mus musculus	2E-56	100%
P26772	X-ray Diffraction	5.42	Mus caroli	2E-55	99.02%
Q9JI95	X-ray Diffraction	2.312	Zalophus californianus	2E-55	95.1%

[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).

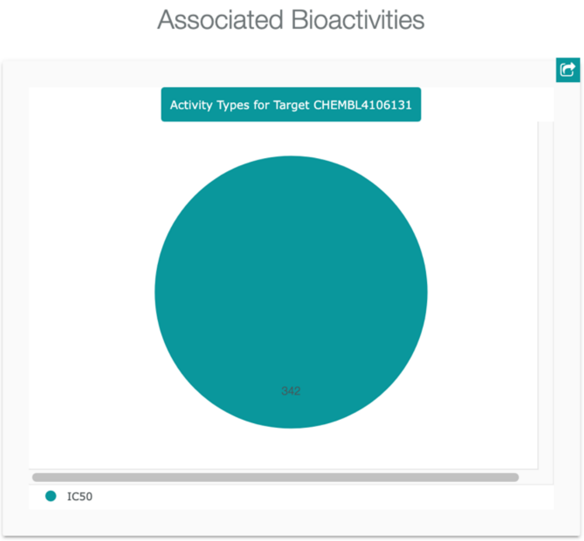
Based on sequence similarity. How likely is this structure to be similar to your "novel" protein?

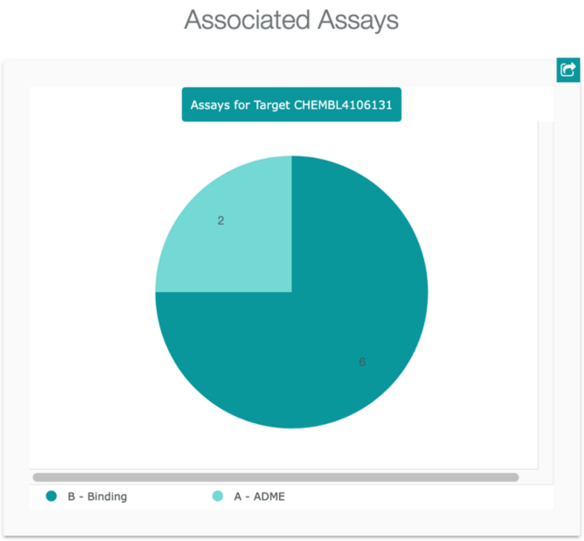
I used the human ADP-bound for this rendering. Based on my sequence alignment they certainly have significant regions of overlap but also several differences, so I would imagine that the overall 3D structure of these proteins is similar but likely these changes could make things like binding to certain molecules and activity difficult to predict.

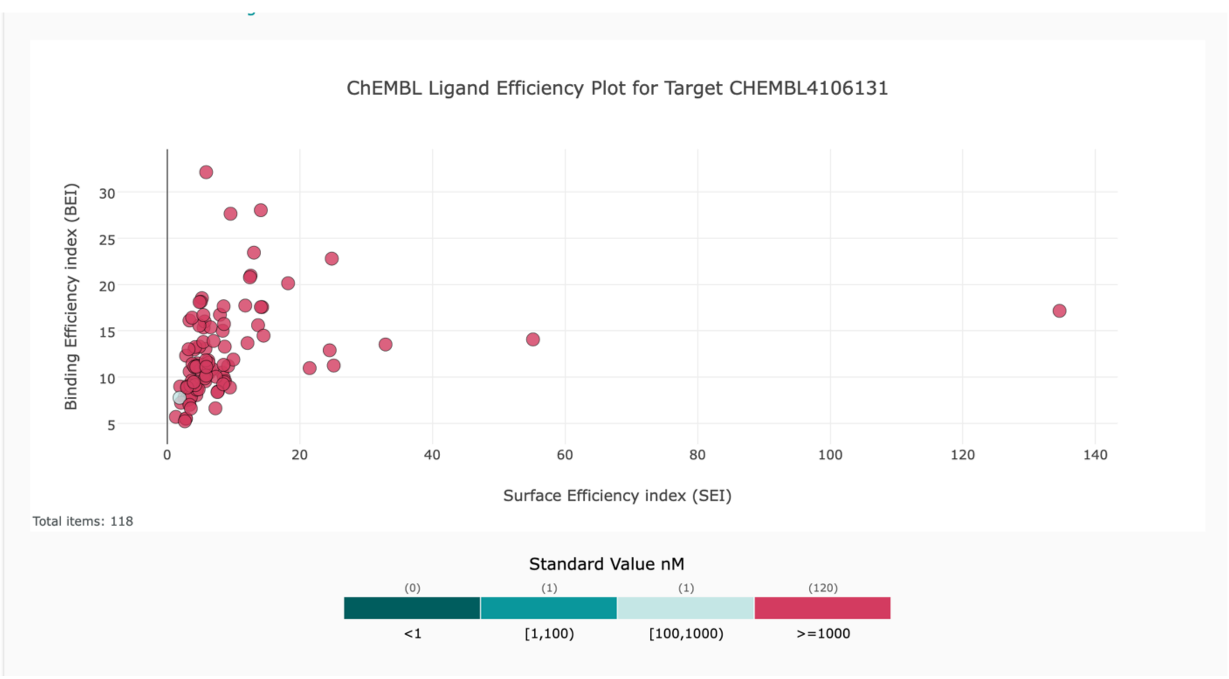


[**Q10**] Perform a “Target” search of ChEMBEL (https://www.ebi.ac.uk/chembl/) with your novel sequence. Are there any **Target Associated Assays** and **ligand efficiency data** reported that may be useful starting points for exploring potential inhibition of your novel protein?

There are 342 associated bioactivities, 2 ADME assays and 6 binding assays for the human HSPE1 version of my novel protein. Because, as we previously discussed there are significant differences between the human HSPE1 and my novel protein it is difficult to predict if these compounds would effectively interact with my novel protein, however they are a good place to start and could be promising at least as a starting point if interested in designing a molecule.







Scoring Rubric:

[45 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 (10 points)

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

Q9 (4 points)

Structure figure provided	2
Uses white background for molecular figure	1
Figure of high resolution (i.e. not just snapshot)	1

Q10 (1 point)

Evidence of ChEMBEL searches	1