Joshua Cheung
PID: A15441585
jcheung@ucsd.edu

Find a Gene Project Assignment

1. 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 its 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.

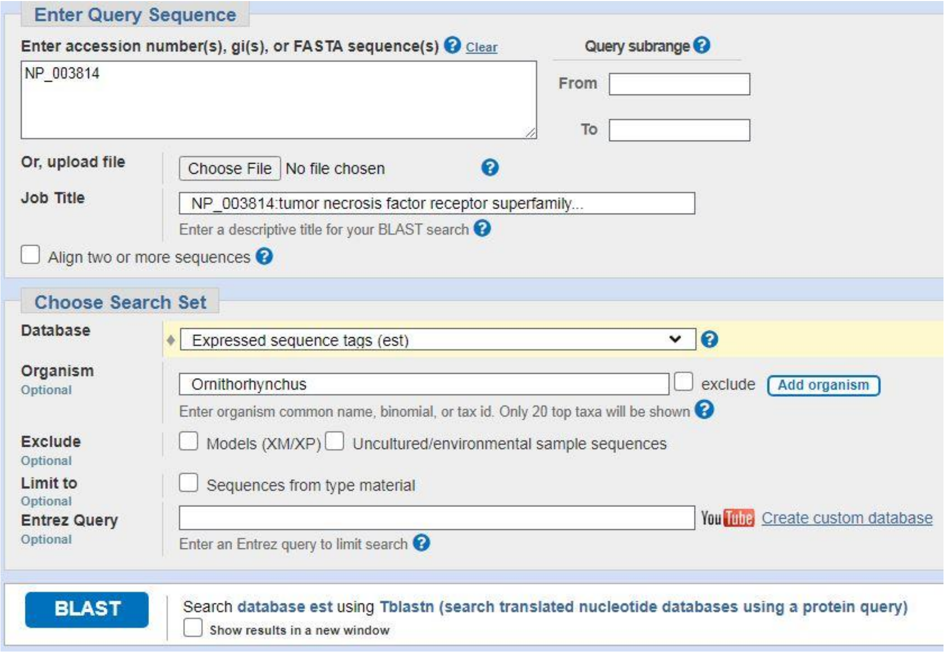
   Name: tumor necrosis factor receptor superfamily, member 6B, decoy
   Accession: NP_003814
   Species: Homo sapiens

2. 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 searched against Ornithorhynchus
   Database: Expressed Sequence Tags (est)
   Organism: Ornithorhynchus (Taxid: 9257)

   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.

   See search setup in screen-shot below:



Enter Query Sequence

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

NP_003814

Query subrange
From
To

Or, upload file: Choose File No file chosen

Job Title: NP_003814:tumor necrosis factor receptor superfamily...

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Expressed sequence tags (est)

Organism (Optional): Ornithorhynchus □ exclude Add organism

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

Exclude (Optional): □ Models (XM/XP) □ Uncultured/environmental sample sequences

Limit to (Optional): □ Sequences from type material

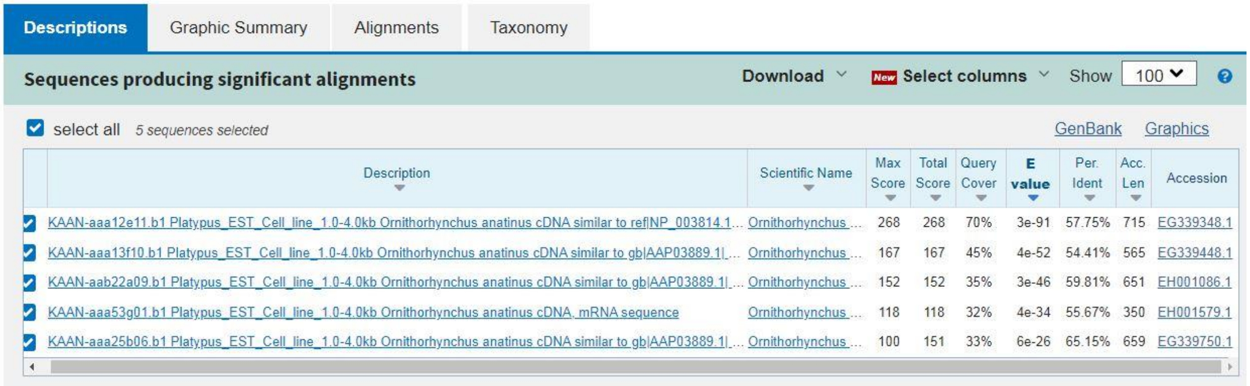
Entrez Query (Optional): YouTube Create custom database

Enter an Entrez query to limit search

BLAST Search database est using Tblastn (search translated nucleotide databases using a protein query)

□ Show results in a new window

The search yielded 5 results, a screen shot of the results is shown below:



Sequences producing significant alignments

☑ select all 5 sequences selected

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	KAAN-aaa12e11.b1 Platypus_EST_Cell_line_1.0-4.0kb Ornithorhynchus anatinus cDNA similar to ref|NP_003814.1...	Ornithorhynchus ...	268	268	70%	3e-91	57.75%	715	EG339348.1
☑	KAAN-aaa13f10.b1 Platypus_EST_Cell_line_1.0-4.0kb Ornithorhynchus anatinus cDNA similar to gb|AAP03889.1|...	Ornithorhynchus ...	167	167	45%	4e-52	54.41%	565	EG339448.1
☑	KAAN-aab22a09.b1 Platypus_EST_Cell_line_1.0-4.0kb Ornithorhynchus anatinus cDNA similar to gb|AAP03889.1|...	Ornithorhynchus ...	152	152	35%	3e-46	59.81%	651	EH001086.1
☑	KAAN-aaa53g01.b1 Platypus_EST_Cell_line_1.0-4.0kb Ornithorhynchus anatinus cDNA, mRNA sequence	Ornithorhynchus ...	118	118	32%	4e-34	55.67%	350	EH001579.1
☑	KAAN-aaa25b06.b1 Platypus_EST_Cell_line_1.0-4.0kb Ornithorhynchus anatinus cDNA similar to gb|AAP03889.1|...	Ornithorhynchus ...	100	151	33%	6e-26	65.15%	659	EG339750.1

Chosen match: Accession EG339348.1, a 715 base pair cDNA clone from *Ornithorhynchus anatinus*. See below for alignment details.

Alignment Details:

```
Query: tumor necrosis factor receptor superfamily member 6B precursor [Homo
sapiens] Query ID: NP_003814.1 Length: 300

>KAAN-aaa12e11.b1 Platypus_EST_Cell_line_1.0-4.0kb Ornithorhynchus anatinus
cDNA similar to ref|NP_003814.1| tumor necrosis factor receptor superfamily,
member 6b; decoy receptor 3 [Homo sapiens] ref|NP_116563.1| tumor necrosis
factor receptor superfamily, member 6b; decoy receptor 3 [Homo sapiens]
sp|O95407|TR6B_HUMAN Tumor necrosis fact, mRNA sequence
Sequence ID: EG339348.1 Length: 715
Range 1: 63 to 701

Score: 268 bits(686), Expect:3e-91,
Method: Compositional matrix adjust.,
Identities: 126/213(59%), Positives: 162/213(76%), Gaps: 1/213(0%)

Query  34   PTYPWRDAETGERLVCAQCPPGTFVQRPCRRDSPTTCGPCPPRHYTQFWNYLERCRYCNV  93
            PTY W+D+ T ERL C QCPPGT+V + C R SPT C PCP  HYTQ+WNYL++CRYCNV
Sbjct  63   PTYSWKDSTTQERLQCQQCPPGTYVSQHCSRTSPTQCQPCPTLHYTQYWNYLDKCRYCNV  242

Query  94   LCGEREEEARACHATHNRACRCRTGFFAHAGFCLEHASCPPGAGVIAPGTPSQNTQCQPC  153
             CG +EEE   C ATHNR C+C+ G++A+  FC+EH++CP G+GV++ GTP++NT+CQ C
Sbjct  243  FCGAQEEEVHPCSATHNRVCQCQPGYYAYMDFCIEHSTCPLGSGVVSQGTPTKNTECQEC  422

Query  154  PPGTFsasssssEQCQPHRNCTALGLALNVPGSSSHDTLCTSCTGFPLSTRVPGAEECER  213
            P GTFS +SS +E CQ H+NCT LG+ +NVPG+  HDTLCT C  F L++  PG ++CE+
Sbjct  423  PRGTFSDNSSRTEPCQSHQNCTLLGMKVNVPGNRFHDTLCTRCDNFQLNSSEPGNKDCEQ  602

Query  214  AVIDFVAFQDISIKRLQRLLQAL-EAPEGWGPT  245
            A+IDFVA+QDI +KRL RL Q L EAP   G T
Sbjct  603  ALIDFVAYQDIPLKRLLRLQQVLGEAPGAAGQT  701
```

3. 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

```
> O. anatinus protein (sequence taken from BLAST results)
TKKLGTFVLAVTFPMGSNNPPTYSWKDSTTQERLQCQQCPPGTYVSQHCSRTSPTQCQPCPTLHYTQYWNYLDKCRYCN
VFCGAQEEEVHPCSATHNRVCQCQPGYYAYMDFCIEHSTCPLGSGVVSQGTPTKNTECQECPRGTFSDNSSRTEPCQSH
QNCTLLGMKVNVPGNRFHDTLCTRCDNFQLNSSEPGNKDCEQALIDFVAYQDIPLKRLLRLQQVLGEAPGAAGQTRGFQ
V
```

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: Ornithorhynchus tumor necrosis factor receptor superfamily, member 6B, decoy
Species: Ornithorhynchus anatinus

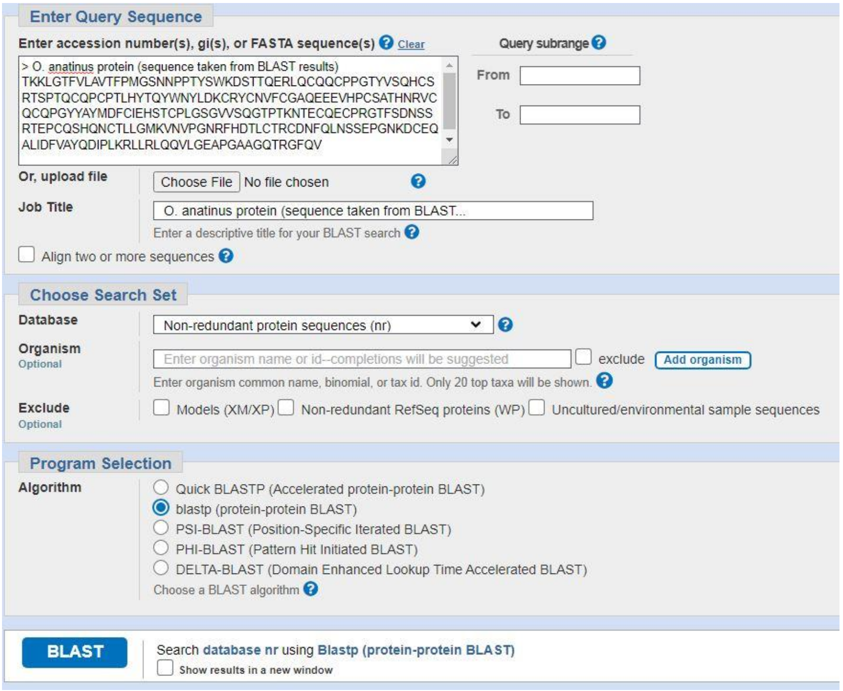
Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Prototheria; Monotremata; Ornithorhynchidae; Ornithorhynchus

4. 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.

Details:

A BLASTP search against NR database was used (see setup in first screen-shot below).
This yielded a top hit result is to a protein from Ornithorhynchus anatinus (Platypus)
See additional screen shots below for top hits and selected alignment details:

The first hit has a 95.74% identity with our query, thus since the percent identify reported is less than 100% it is likely that our gene is novel as defined in the question. So we have succeeded as required.



Enter Query Sequence

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

```
> O. anatinus protein (sequence taken from BLAST results)
TKKLGTFVLAVTFPMGSNNPPTYSWKDSTTQERLQCQQCPPGTYVSQHCS
RTSPTQCQPCPTLHYTQYWNYLDKCRYCNVFCGAQEEEVHPCSATHNRVC
QCQPGYYAYMDFCIEHSTCPLGSGVVSQGTPTKNTECQECPRGTFSDNSS
RTEPCQSHQNCTLLGMKVNVPGNRFHDTLCTRCDNFQLNSSEPGNKDCEQ
ALIDFVAYQDIPLKRLLRLQQVLGEAPGAAGQTRGFQV
```

Query subrange ?
From
To

Or, upload file: Choose File No file chosen ?

Job Title: O. anatinus protein (sequence taken from BLAST...
Enter a descriptive title for your BLAST search ?

☐ Align two or more sequences ?

Choose Search Set

Database: Non-redundant protein sequences (nr) ?

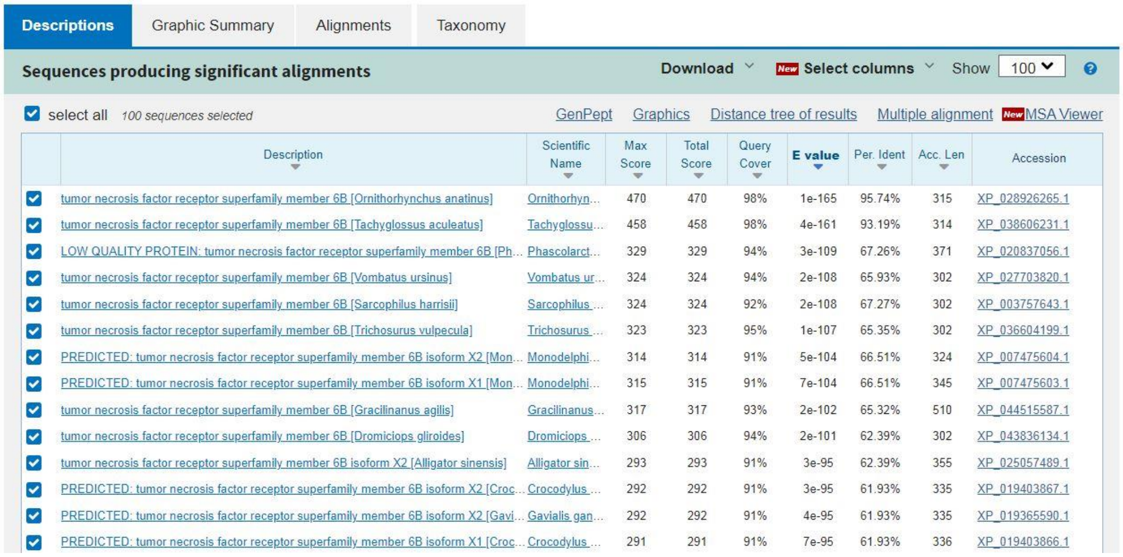
Organism (Optional): Enter organism name or id--completions will be suggested ☐ exclude Add organism
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. ?

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)
☐ Show results in a new window

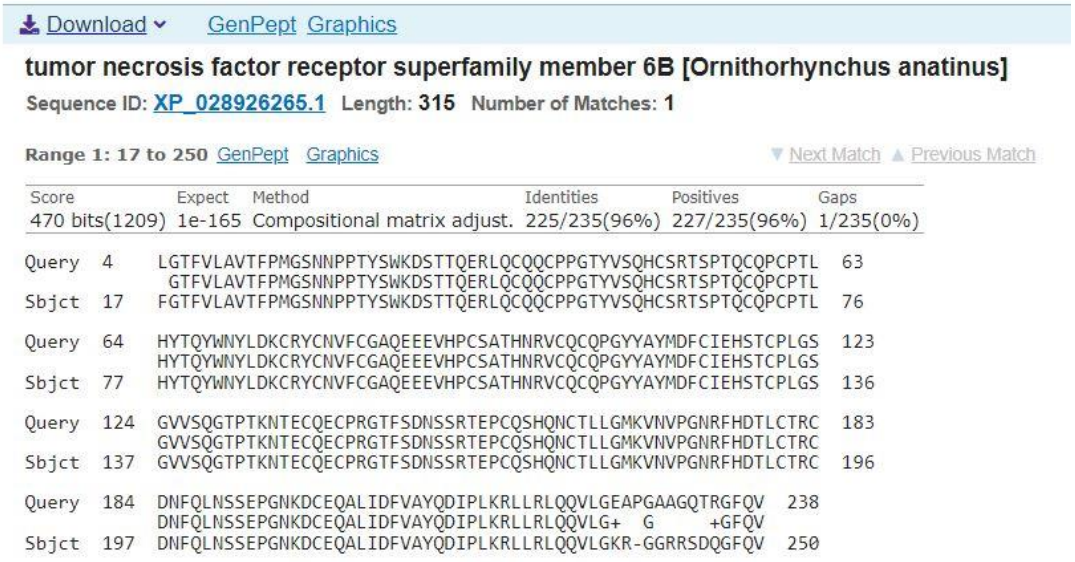


Descriptions | Graphic Summary | Alignments | Taxonomy

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
☑	tumor necrosis factor receptor superfamily member 6B [Ornithorhynchus anatinus]	Ornithorhyn...	470	470	98%	1e-165	95.74%	315	XP_028926265.1
☑	tumor necrosis factor receptor superfamily member 6B [Tachyglossus aculeatus]	Tachyglossu...	458	458	98%	4e-161	93.19%	314	XP_038606231.1
☑	LOW QUALITY PROTEIN: tumor necrosis factor receptor superfamily member 6B [Ph...	Phascolarct...	329	329	94%	3e-109	67.26%	371	XP_020837056.1
☑	tumor necrosis factor receptor superfamily member 6B [Vombatus ursinus]	Vombatus ur...	324	324	94%	2e-108	65.93%	302	XP_027703820.1
☑	tumor necrosis factor receptor superfamily member 6B [Sarcophilus harrisii]	Sarcophilus ...	324	324	92%	2e-108	67.27%	302	XP_003757643.1
☑	tumor necrosis factor receptor superfamily member 6B [Trichosurus vulpecula]	Trichosurus ...	323	323	95%	1e-107	65.35%	302	XP_036604199.1
☑	PREDICTED: tumor necrosis factor receptor superfamily member 6B isoform X2 [Mon...	Monodelphi...	314	314	91%	5e-104	66.51%	324	XP_007475604.1
☑	PREDICTED: tumor necrosis factor receptor superfamily member 6B isoform X1 [Mon...	Monodelphi...	315	315	91%	7e-104	66.51%	345	XP_007475603.1
☑	tumor necrosis factor receptor superfamily member 6B [Gracilinanus agilis]	Gracilinanus...	317	317	93%	2e-102	65.32%	510	XP_044515587.1
☑	tumor necrosis factor receptor superfamily member 6B [Dromiciops gliroides]	Dromiciops ...	306	306	94%	2e-101	62.39%	302	XP_043836134.1
☑	tumor necrosis factor receptor superfamily member 6B isoform X2 [Alligator sinensis]	Alligator sin...	293	293	91%	3e-95	62.39%	355	XP_025057489.1
☑	PREDICTED: tumor necrosis factor receptor superfamily member 6B isoform X2 [Croc...	Crocodylus ...	292	292	91%	3e-95	61.93%	335	XP_019403867.1
☑	PREDICTED: tumor necrosis factor receptor superfamily member 6B isoform X2 [Gavi...	Gavialis gan...	292	292	91%	4e-95	61.93%	335	XP_019365590.1
☑	PREDICTED: tumor necrosis factor receptor superfamily member 6B isoform X1 [Croc...	Crocodylus ...	291	291	91%	7e-95	61.93%	336	XP_019403866.1



tumor necrosis factor receptor superfamily member 6B [Ornithorhynchus anatinus]

Sequence ID: XP_028926265.1 Length: 315 Number of Matches: 1

Range 1: 17 to 250 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
470 bits(1209)	1e-165	Compositional matrix adjust.	225/235(96%)	227/235(96%)	1/235(0%)

```
Query  4    LGTFVLAVTFPMGSNNPPTYSWKDSTTQERLQCQQCPPGTYVSQHCSRTSPTQCQPCPTL  63
             GTFVLAVTFPMGSNNPPTYSWKDSTTQERLQCQQCPPGTYVSQHCSRTSPTQCQPCPTL
Sbjct  17   FGTFVLAVTFPMGSNNPPTYSWKDSTTQERLQCQQCPPGTYVSQHCSRTSPTQCQPCPTL  76

Query  64   HYTQYWNYLDKCRYCNVFCGAQEEEVHPCSATHNRVCQCQPGYYAYMDFCIEHSTCPLGS  123
            HYTQYWNYLDKCRYCNVFCGAQEEEVHPCSATHNRVCQCQPGYYAYMDFCIEHSTCPLGS
Sbjct  77   HYTQYWNYLDKCRYCNVFCGAQEEEVHPCSATHNRVCQCQPGYYAYMDFCIEHSTCPLGS  136

Query  124  GVVSQGTPTKNTECQECPRGTFSDNSSRTEPCQSHQNCTLLGMKVNVPGNRFHDTLCTRC  183
            GVVSQGTPTKNTECQECPRGTFSDNSSRTEPCQSHQNCTLLGMKVNVPGNRFHDTLCTRC
Sbjct  137  GVVSQGTPTKNTECQECPRGTFSDNSSRTEPCQSHQNCTLLGMKVNVPGNRFHDTLCTRC  196

Query  184  DNFQLNSSEPGNKDCEQALIDFVAYQDIPLKRLLRLQQVLGEAPGAAGQTRGFQV  238
            DNFQLNSSEPGNKDCEQALIDFVAYQDIPLKRLLRLQQVLG+  G     +GFQV
Sbjct  197  DNFQLNSSEPGNKDCEQALIDFVAYQDIPLKRLLRLQQVLGKR-GGRRSDQGFQV  250
```

5. 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.

Re-labeled sequences for alignment

```
>Human|ref|NP_000509.1|Human tumor necrosis factor decoy receptor 6B [Homo
sapiens]
MRALEGPGLSLLCLVLALPALLPVPAVRGVAETPTYPWRDAETGERLVCAQCPPGTFVQRPCRRDSPTTCGPCPPRHYT
QFWNYLERCRYCNVLCGEREEEARACHATHNRACRCRTGFFAHAGFCLEHASCPPGAGVIAPGTPSQNTQCQPCPPGTF
SASSSSSEQCQPHRNCTALGLALNVPGSSSHDTLCTSCTGFPLSTRVPGAEECERAVIDFVAFQDISIKRLQRLLQALE
APEGWGPTPRAGRAALQLKLRRRLTELLGAQDGALLVRLLQALRVARMPGLERSVRERFLPVH

> Ornithorhynchus tumor necrosis factor receptor superfamily, member 6B, decoy
(sequence taken from BLAST results); AltName: Ornithorhynchus TNFRSF6B
TKKLGTFVLAVTFPMGSNNPPTYSWKDSTTQERLQCQQCPPGTYVSQHCSRTSPTQCQPCPTLHYTQYWNYLDKCRYCN
VFCGAQEEEVHPCSATHNRVCQCQPGYYAYMDFCIEHSTCPLGSGVVSQGTPTKNTECQECPRGTFSDNSSRTEPCQSH
QNCTLLGMKVNVPGNRFHDTLCTRCDNFQLNSSEPGNKDCEQALIDFVAYQDIPLKRLLRLQQVLGEAPGAAGQTRGFQ
V

>Platypus|ref|XP_028926265.1| tumor necrosis factor receptor superfamily member
6B [Ornithorhynchus anatinus]
```

```
MPTQCPKLRSCLSRWAFGTFVLAVTFPMGSNNPPTYSWKDSTTQERLQCQQCPPGTYVSQHCSRTSPTQCQPCPTLHYT
QYWNYLDKCRYCNVFCGAQEEEVHPCSATHNRVCQCQPGYYAYMDFCIEHSTCPLGSGVVSQGTPTKNTECQECPRGTF
SDNSSRTEPCQSHQNCTLLGMKVNVPGNRFHDTLCTRCDNFQLNSSEPGNKDCEQALIDFVAYQDIPLKRLLRLQQVLG
KRGGRRSDQGFQVVMQKKLLQQLMEKKEAQTSDALITELLQALRTVKLYGLIEKIQKHFSLHSDNLTSTSAPWTYLVL


>Short-beaked echidna|ref|XP_038606231.1| tumor necrosis factor receptor
superfamily member 6B [Tachyglossus aculeatus]
MPTQCPKLRSCLSRWAFGTFVLAVTFPMGSNNPPTYPWKDSVTQERLQCQQCPPGTYVSQHCSRTSPTQCQPCPTLHYT
QYWNYLDKCRYCNVFCGAQEEEVHPCSATHNRVCQCQSGYYTYMDFCIEHSTCPLGSGVVSQGTPTKNTQCQECPRGTF
SDNSSRTEPCQSHQNCTLLGMKVNVPGNRFHDTLCTRCDNFQLNSSEPGNRDCEQALIDFVAYQDIPLKRLLRLQQVLG
RRGGRRTDQGFQVVMQKKLLQQLMEKREAQTSDALITELLQALRTVKLYGLIEKIQKHFSLHMDNLTSTAAPWTLVL

>Common wombat|ref|XP_027703820.1|tumor necrosis factor receptor superfamily
member 6B [Vombatus ursinus]
MDLPTQNIKFLWIVSTLLLLLVMPGDAGNFPTYPWRDAETQEWLLCDQCPPGTFVKHHCSYKSPTVCQPCPSLHYTQYW
NYLEKCRYCNVFCGEREEEAQACNATHNRACRCQLGYYAHADFCIEHSACPPGSGVVTLGTPNQNTQCQPCPKGTFSDN
SSSTEKCQPHRNCSTLGMFLNVPGTSFHDAICTRCSGFLSSTPEPGDKECEQAVIDFVAFQNISLKRLMRLQQALEAPG
SWHRQWPEPESRAAVQKELLHRLTELSETQGSSGLLLQVLQALRKAKLTTLERNIQKHFVLDQKD

>Tasmanian devil|ref|XP_003757643.1| tumor necrosis factor receptor superfamily
member 6B [Sarcophilus harrisii]
MDLPVQNVKFSWLVSTLLPLVSMPGDAGNAPTYSWRDAETQEWLVCNQCPPGTFVKQHCSHRSPTNCQPCPSLHYTQYW
NYLEKCRYCNVFCGEHEEEAQACNATHNRACRCQLGYYAHADFCIEHSACPPGSGVVTLGTPTQNTQCQPCPKGTFSDN
SSSTERCQPHRNCTAFGMFLNVPGTSFHDTMCTRCASFLSSTPEPGNKECEKAVIDFVAFQNISLKRLRKLQQALETPD
SWQREWPEPENRAAVQKELLHRLTELSDPQESSIFVLKLLQALRKAKLTTLEKNLRKRFLLALKD

>Common brushtail possum|ref|XP_036604199.1| tumor necrosis factor receptor
superfamily member 6B [Trichosurus vulpecula]
MDLTAQNIKFLRIVSTLLLLMVMPRDAGNFPTYSWRDAETQEWLLCDQCPPGTFVKHHCSYRSRTVCQPCPSLHYTQYW
NYLEKCRYCNVFCGEHEEEAQACNATHNRACRCQLGYYAHADFCIEHSACPPGSGVVTLGTPNQNTQCQPCPKGTFSDN
SSSTEKCQPHRNCTTLGMFLNVPGTSFHDAICTRCAGFLSSTPEPGDKECEQAVIDFVAFQNISLKRLMRLQQALEGPG
SWHRQWPEPESRAAVQKELLHRLTELSETQGSSALLLQLLQALRKAKLTTLERNIQKHFSLDIKD
```

Alignment

Obtained using MUSCLE (version 3.8) at EBI

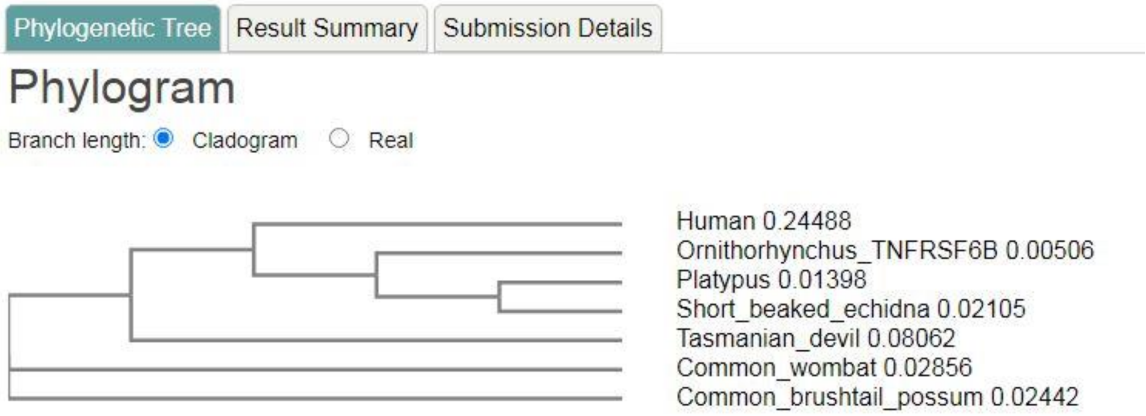
```
Human                      RPCRRDSPTTCGPCPPRHYTQFWNYLERCRYCNVLCGEREEEARACHATHNRACRCRTGF
Ornithorhynchus_TNFRSF6B   QHCSRTSPTQCQPCPTLHYTQYWNYLDKCRYCNVFCGAQEEEVHPCSATHNRVCQCQPGY
Platypus                   QHCSRTSPTQCQPCPTLHYTQYWNYLDKCRYCNVFCGAQEEEVHPCSATHNRVCQCQPGY
Short_beaked_echidna       QHCSRTSPTQCQPCPTLHYTQYWNYLDKCRYCNVFCGAQEEEVHPCSATHNRVCQCQSGY
Tasmanian_devil            QHCSHRSPTNCQPCPSLHYTQYWNYLEKCRYCNVFCGEHEEEAQACNATHNRACRCQLGY
Common_wombat              HHCSYKSPTVCQPCPSLHYTQYWNYLEKCRYCNVFCGEREEEAQACNATHNRACRCQLGY
Common_brushtail_possum    HHCSYRSRTVCQPCPSLHYTQYWNYLEKCRYCNVFCGEHEEEAQACNATHNRACRCQLGY
                           . *   * * * ***. ****:****:.******:** .***...* *****.*.*. *:

Human                      FAHAGFCLEHASCPPGAGVIAPGTPSQNTQCQPCPPGTFSASSSSSEQCQPHRNCTALGL
Ornithorhynchus_TNFRSF6B   YAYMDFCIEHSTCPLGSGVVSQGTPTKNTECQECPRGTFSDNSSRTEPCQSHQNCTLLGM
Platypus                   YAYMDFCIEHSTCPLGSGVVSQGTPTKNTECQECPRGTFSDNSSRTEPCQSHQNCTLLGM
Short_beaked_echidna       YTYMDFCIEHSTCPLGSGVVSQGTPTKNTQCQECPRGTFSDNSSRTEPCQSHQNCTLLGM
Tasmanian_devil            YAHADFCIEHSACPPGSGVVTLGTPTQNTQCQPCPKGTFSDNSSSTERCQPHRNCTAFGM
Common_wombat              YAHADFCIEHSACPPGSGVVTLGTPNQNTQCQPCPKGTFSDNSSSTEKCQPHRNCSTLGM
Common_brushtail_possum    YAHADFCIEHSACPPGSGVVTLGTPNQNTQCQPCPKGTFSDNSSSTEKCQPHRNCTTLGM
                           ::: .**:**::** *:**:: ***.:**:** ** **** .** :* **.*.**: :*:

Human                      ALNVPGSSSHDTLCTSCTGFPLSTRVPGAEECERAVIDFVAFQDISIKRLQRLLQAL-EA
Ornithorhynchus_TNFRSF6B   KVNVPGNRFHDTLCTRCDNFQLNSSEPGNKDCEQALIDFVAYQDIPLKRLLRLQQVLGEA
Platypus                   KVNVPGNRFHDTLCTRCDNFQLNSSEPGNKDCEQALIDFVAYQDIPLKRLLRLQQVL-GK
Short_beaked_echidna       KVNVPGNRFHDTLCTRCDNFQLNSSEPGNRDCEQALIDFVAYQDIPLKRLLRLQQVL-GR
Tasmanian_devil            FLNVPGTSFHDTMCTRCASFLSSTPEPGNKECEKAVIDFVAFQNISLKRLRKLQQAL-ET
Common_wombat              FLNVPGTSFHDAICTRCSGFLSSTPEPGDKECEQAVIDFVAFQNISLKRLMRLQQAL-EA
Common_brushtail_possum    FLNVPGTSFHDAICTRCAGFLSSTPEPGDKECEQAVIDFVAFQNISLKRLMRLQQAL-EG
                            :****.  **::** * .*  .:  **  :**.*:*****:*:*.:*** .* *.*
```

6. 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.

   The multiple sequence alignments generated in question 6 were imported into the Simple Phylogeny tool at EBI to create a neighbor-joining tree.

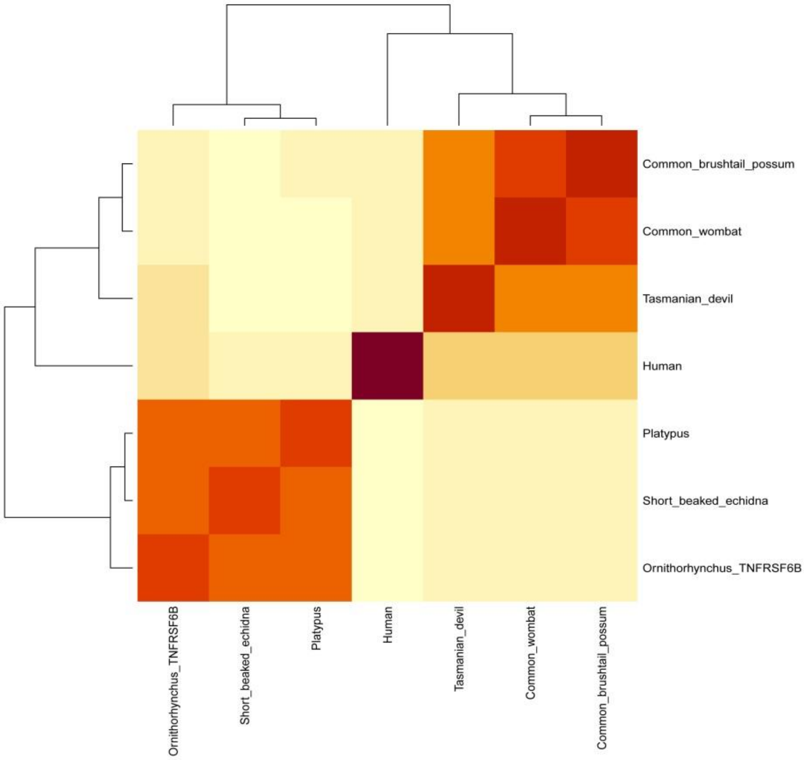


Phylogram

Branch length: ● Cladogram ○ Real

```
Human 0.24488
Ornithorhynchus_TNFRSF6B 0.00506
Platypus 0.01398
Short_beaked_echidna 0.02105
Tasmanian_devil 0.08062
Common_wombat 0.02856
Common_brushtail_possum 0.02442
```

7. 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.



8. 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 (experimental Technique), resolution (resolution), and source organism (source).

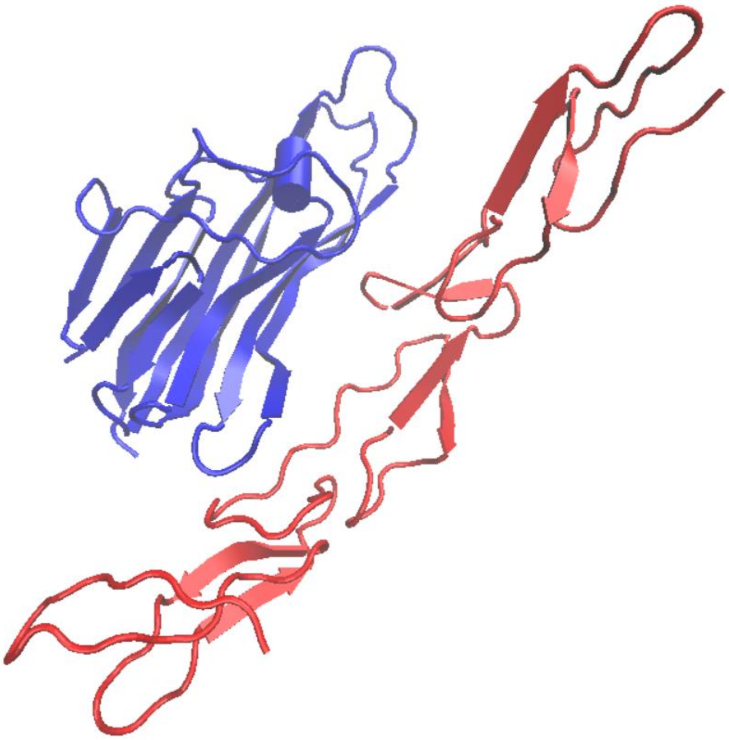
HINT: You can use a single sequence from your alignment or generate a consensus sequence from your alignment using the Bio3D function consensus(). The Bio3D functions blast.pdb(), plot.blast() and pdb.annotate() are likely to be of most relevance for completing this task. Note that the results of blast.pdb() 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.

ID	Technique	Resolution	Source	E-value	Identity
4E4D	X-ray Diffraction	2.7	Mus musculus	5E-38	46.30
3WVT	X-ray Diffraction	1.6	Equus Caballus	1E-10	34.84
3IJ2	X-ray Diffraction	3.75	Rattus Norvegicus	2E-07	32.56

9. 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?
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?

A molecular figure of the 4E4D PDB structure was generated. While it is possible that this structure may share some large similarities with the Ornithorhynchus tumor necrosis factor receptor superfamily, member 6B, decoy, it is unlikely that the two proteins will be similar in their entirety given that their two respective sequence do not have a relatively high sequence similarity (>50%). In the figure below the red chain is a conserved region that is similar to a protein region in the Ornithorhynchus tumor necrosis factor receptor superfamily, member 6B, decoy subject of this report.



10. 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?

CHEMBL details 405131 binding assays, and 675927 functional assays, and no ligand efficeny data.

A functional assay and SAR analysis tested a set Human immunodeficiency virus 1 HXB2 containing reverse transcriptase V179D with the results suggesting “a potential Antiviral activity against Human immunodeficiency virus 1 HXB2 containing reverse transcriptase V179D, Y181C mutant relative to Human immunodeficiency virus 1 3B”

Citation:
Azijn, H., Tirry, I., Vingerhoets, J., de Béthune, M. P., Kraus, G., Boven, K., Jochmans, D., Van Craenenbroeck, E., Picchio, G., & Rimsky, L. T. (2010). TMC278, a next-generation nonnucleoside reverse transcriptase inhibitor (NNRTI), active against wild-type and NNRTI-resistant HIV-1. Antimicrobial agents and chemotherapy, 54(2), 718–727. https://doi.org/10.1128/AAC.00986-09

https://pubmed.ncbi.nlm.nih.gov/19933797/