**Daniel Johnson**

**A53072130**

**BGGN213 2019 Midterm “Find a Gene Project”**

**[Q1]Protein Name**

Protein Name: **CBFA2T1** (Product of *RUNX1T1* Gene)

Species: **Human**

Accession: **NX\_Q06455**

**[Q2] Blast Search Results**

**Blast Input:**

>sp|Q06455|MTG8\_HUMAN Protein **CBFA2T1** OS=**Homo sapiens** OX=9606 GN=**RUNX1T1** PE=1 SV=2

MISVKRNTWRALSLVIGDCRKKGNFEYCQDRTEKHSTMPDSPVDVKTQSRLTPPTMPPPP

TTQGAPRTSSFTPTTLTNGTSHSPTALNGAPSPPNGFSNGPSSSSSSSLANQQLPPACGA

RQLSKLKRFLTTLQQFGNDISPEIGERVRTLVLGLVNSTLTIEEFHSKLQEATNFPLRPF

VIPFLKANLPLLQRELLHCARLAKQNPAQYLAQHEQLLLDASTTSPVDSSELLLDVNENG

KRRTPDRTKENGFDREPLHSEHPSKRPCTISPGQRYSPNNGLSYQPNGLPHPTPPPPQHY

RLDDMAIAHHYRDSYRHPSHRDLRDRNRPMGLHGTRQEEMIDHRLTDREWAEEWKHLDHL

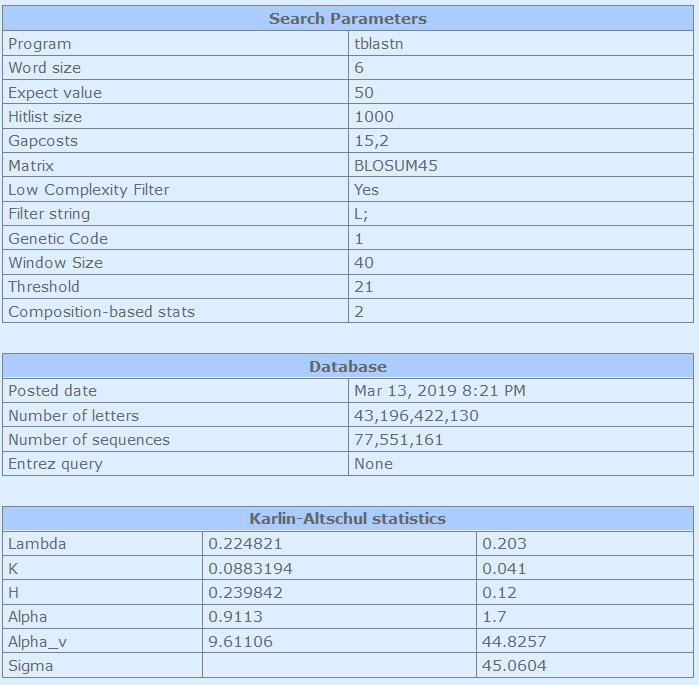
LNCIMDMVEKTRRSLTVLRRCQEADREELNYWIRRYSDAEDLKKGGGSSSSHSRQQSPVN

PDPVALDAHREFLHRPASGYVPEEIWKKAEEAVNEVKRQAMTELQKAVSEAERKAHDMIT

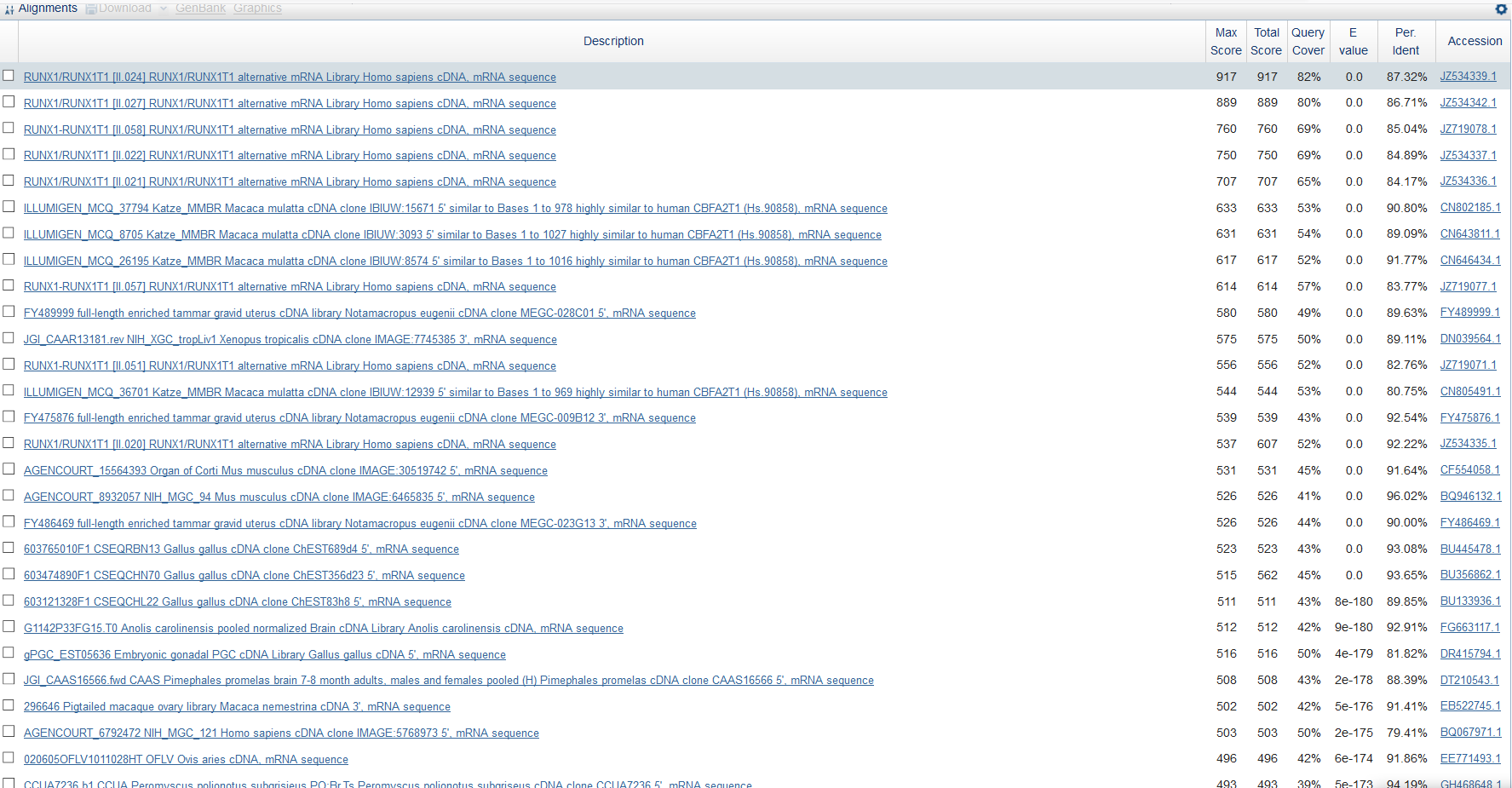
TERAKMERTVAEAKRQAAEDALAVINQQEDSSESCWNCGRKASETCSGCNTARYCGSFCQ

HKDWEKHHHICGQTLQAQQQGDTPAVSSSVTPNSGAGSPMDTPPAATPRSTTPGTPSTIE

TTPR

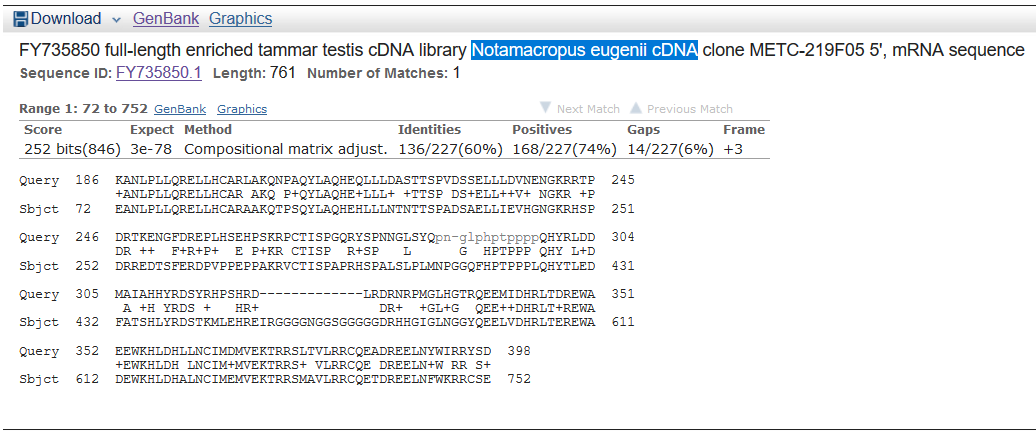


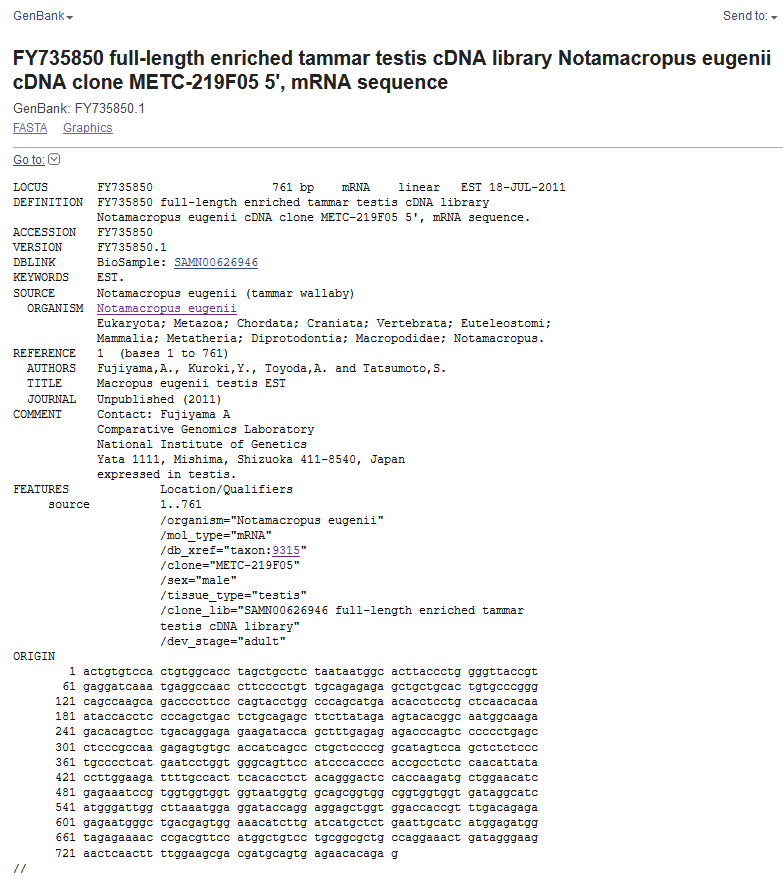
**Blast Output:**





**Selected homolog:**





**[Q3] Novel Protein Information**

Species: Notamacropus Eugenii (Tammar Wallaby)

Novel Protein Name: **CBFA2T1**

**Novel Protein Translated.**

>EMBOSS\_001\_Frame 1

TVSTVAPSCL\*\*WHLPWGYREDQMRPTFPCCRESCCTVPGQPSRPLPSTWPSMNTSCSTQ

IPPPQLTLQSFL\*KYTAMARDTVLTGEKIPALRETQSPLSLPPRECAPSALLPGIVQLSP

CPS\*ILVGSSIPPHRLSNIIPWKILPLHTSTGTPPRCWNIEKSVVVVVMVAAVAVVVIGI

MGLA\*MEDTRRSWWTTV\*QRENGLTSGNILIML\*IASWRW\*RKPDVPWLSCGAARKLIGK

NSTFGSDDAVRTQX

>EMBOSS\_001\_Frame 2

LCPLWHLAASNNGTYPGVTVRIK\*GQPSPVAERAAALCPGSQADPFPVPGPA\*TPPAQHK

YHLPS\*LCRASYRSTRQWQETQS\*QERRYQL\*ERPSPP\*ASRQESVHHQPCSPA\*SSSLP

APHESWWAVPSHPTASPTLYLGRFCHFTPLQGLHQDAGTSRNPWWWW\*WWQRWRWW\*\*AS

WDWLKWRIPGGAGGPPFDRERMG\*RVETS\*SCSELHHGDGRENPTFHGCPAALPGN\*\*GR

TQLLEATMQ\*EHRX

**>EMBOSS\_001\_Frame 3 CBFA2T1 Predicted Sequence (from EST)**

CVHCGT\*LPLIMALTLGLP\*GSNEANLPLLQRELLHCARAAKQTPSQYLAQHEHLLLNTN

TTSPADSAELLIEVHGNGKRHSPDRREDTSFERDPVPPEPPAKRVCTISPAPRHSPALSL

PLMNPGGQFHPTPPPLQHYTLEDFATSHLYRDSTKMLEHREIRGGGGNGGSGGGGGDRHH

GIGLNGGYQEELVDHRLTEREWADEWKHLDHALNCIMEMVEKTRRSMAVLRRCQETDREE

LNFWKRRCSENTE

>EMBOSS\_001\_Frame 4

LCSHCIVASKS\*VLPYQFPGSAAGQPWNVGFSLPSP\*CNSEHDQDVSTRQPILSLSNGGP

PAPPGILHLSQSHDAYHHHRHRCHHYHHHHGFLDVPASWWSPCRGVKWQNLPRYNVGEAV

GWDGTAHQDS\*GAGRELDYAGEQG\*WCTLSWREAQGGLGLSQSWYLLSCQDCVSCHCRVL

L\*EALQSQLGRWYLC\*AGGVHAGPGTGKGSAWLPGHSAAALSATGEGWPHLILTVTPG\*V

PLLEAARCHSGHS

>EMBOSS\_001\_Frame 5

SVFSLHRRFQKLSSSLSVSWQRRRTAMERRVFSTISMMQFRA\*SRCFHSSAHSLSVKRWS

TSSSWYPPFKPIP\*CLSPPPPPLPPLPPPPRISRCSSILVESL\*RCEVAKSSKV\*CWRGG

GVGWNCPPGFMRGRERAGLCRGAGLMVHTLLAGGSGGTGSLSKLVSSLLSGLCLLPLPCT

SIRSSAESAGEVVFVLSRRCSCWARYWEGVCLAARAQCSSSLCNRGRLASFDPHGNPRVS

AIIRGS\*VPQWTQX

>EMBOSS\_001\_Frame 6

LCVLTASSLPKVEFFPISFLAAPQDSHGTSGFLYHLHDAIQSMIKMFPLVSPFSLCQTVV

HQLLLVSSI\*ANPMMPITTTATAATITTTTTDFSMFQHLGGVPVEV\*SGKIFQGIMLERR

WGGMELPTRIHEGQGESWTMPGSRADGAHSLGGRLRGDWVSLKAGIFSPVRTVSLAIAVY

FYKKLCRVSWGGGICVEQEVFMLGQVLGRGLLGCPGTVQQLSLQQGKVGLI\*SSR\*PQGK

CHY\*RQLGATVDTV

**[Q4]Proof of Novelty**

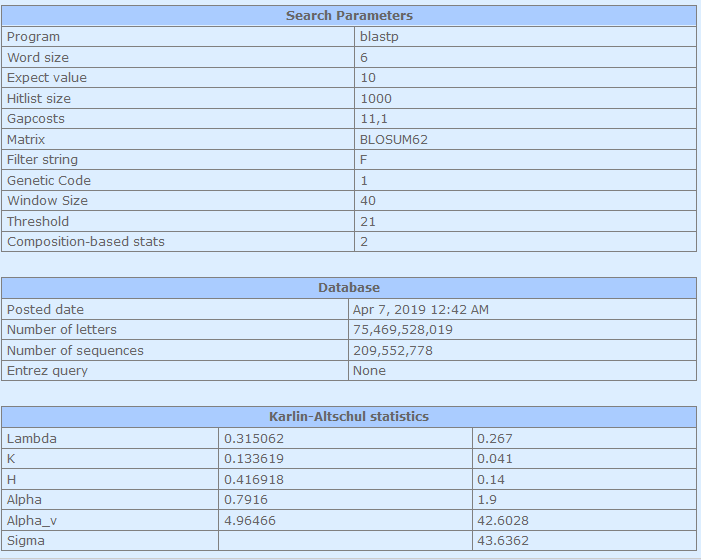
**>EMBOSS\_001\_Frame 3 CBFA2T1 Predicted Sequence (from EST)**

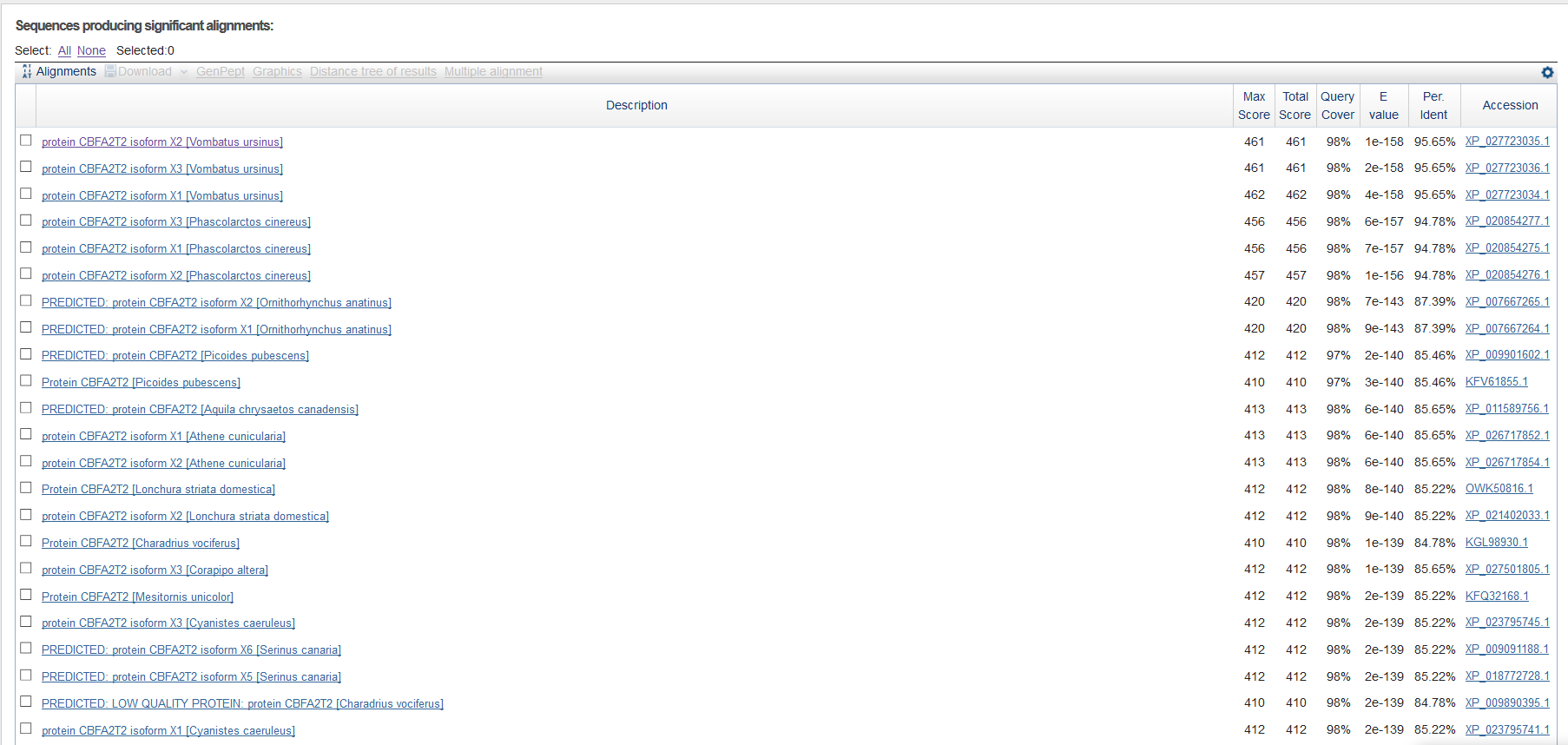
GSNEANLPLLQRELLHCARAAKQTPSQYLAQHEHLLLNTNTTSPADSAELLIEVHGNGKR

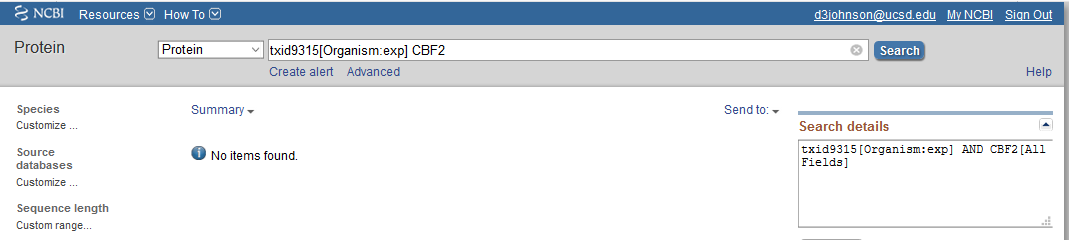
HSPDRREDTSFERDPVPPEPPAKRVCTISPAPRHSPALSLPLMNPGGQFHPTPPPLQHYT

LEDFATSHLYRDSTKMLEHREIRGGGGNGGSGGGGGDRHHGIGLNGGYQEELVDHRLTER

EWADEWKHLDHALNCIMEMVEKTRRSMAVLRRCQETDREELNFWKRRCSENTE







[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.

**CLUSTAL multiple sequence alignment by MUSCLE (3.8)**

**Human GNDISPEIGERVRTLVLGLVNSTLTIEEFHSKLQEATNFPLRPFVIPFLKANLPLLQREL**

**Orca\_Whale GNDISPEIGERVRTLVLGLVNSTLTIEEFHSKLQEATNFPLRPFVIPFLKANLPLLQREL**

**Tropical\_Clawed\_Frog GNDISPEIGEKVRTLVLALVNSTVTIEEFHCKLQEATNFPLRPFVIPFLKANLPLLQREL**

Cow GNDISPEIGEKVRTLVLALVNSTVTIEEFHCKLQEATNFPLRPFVIPFLKANLPLLQREL

Chimpanzee GNDISPEIGEKVRTLVLALVNSTVTIEEFHCKLQEATNFPLRPFVIPFLKANLPLLQREL

Mouse GNDISPEIGEKVRTLVLALVNSTVTIEEFHCKLQEATNFPLRPFVIPFLKANLPLLQREL

**Society\_Finch GNDISPEIGEKVRTLVLALVNSTVTIEEFHCKLQEATNFPLRPFVIPFLKANLPLLQREL**

**Tammar\_Wallaby GSN----------------------------------------------EANLPLLQREL**

**Common\_Wombat GNDISPEIGEKVRTLVLALVNSTVTIEEFHCKLQEATNFPLRPFVIPFLKANLPLLQREL**

**\*.: :\*\*\*\*\*\*\*\*\*\***

**Human LHCARLAKQNPAQYLAQHEQLLLDASTTSPVDSSELLLDVNENGKRRTPDRTKENGFDRE**

**Orca\_Whale LHCARLAKQNPAQYLAQHEQLLLDASTTSPVDSSELLLDVNENGKRRTPDRTKENGFDRE**

**Tropical\_Clawed\_Frog LHCARAAKQTPSQYLAQHEHILLNTSTSSPADSSELLMEVNGNGKRHSPDRREENGCERE**

Cow LHCARAAKQTPSQYLAQHEHLLLNTSIASPADSSELLMEVHGNGKRPSPERREESSFERD

Chimpanzee LHCARAAKQTPSQYLAQHEHLLLNTSIASPADSSELLMEVHGNGKRPSPERREENSFDRD

Mouse LHCARAAKQTPSQYLAQHEHLLLNTSIASPADSSELLMEVHGNGKRPSPERRDENNFERD

**Society\_Finch LHCARAAKQTPSQYLAQHEHILLNTNTTSPADSSELLIEVNGNGKRHSPDRREDSSFERE**

**Tammar\_Wallaby LHCARAAKQTPSQYLAQHEHLLLNTNTTSPADSAELLIEVHGNGKRHSPDRREDTSFERD**

**Common\_Wombat LHCARAAKQTPSQYLAQHEHLLLNTNTTSPADSAELLIEVHGNGKRHSPDRREDPSFERE**

**\*\*\*\*\* \*\*\*.\*:\*\*\*\*\*\*\*::\*\*::. :\*\*.\*\*:\*\*\*::\*: \*\*\*\* :\*:\* .: . :\*:**

**Human PLHSEHPSKRPCTISPGQRYSPNNGLSY-QPNGLPHPTPPPPQHYRLDDMAIAHHYRDS-**

**Orca\_Whale PLHSEHPSKRPCTISPGQRYSPNNGLSY-QPNGLPHPTPPPPQHYRLDDMAIAHHYRDS-**

**Tropical\_Clawed\_Frog PALPEPPVKRVCTISPAPRHSPALSLPIVNSTSHFHPTPPPLQHYSLEDIPSSQLYRDQL**

Cow AIAPEPPAKRVCTISPAPRHSPALTVPLMNPGGQFHPTPPPLQHYTLEDIATSHLYREP-

Chimpanzee TIAPEPPAKRVCTISPAPRHSPALTVPLMNPGGQFHPTPPPLQHYTLEDIATSHLYREP-

Mouse TVPPEPPAKRVCTISPAPRHSPALTVPLMNPGGQFHPTPPPLQHYTLEDIATSHLYREP-

**Society\_Finch PLPTEPPAKRVCTISPAPRHSPALTIPLMNPSGQFHPTPPPLQHYTLEDIATSHLYRDP-**

**Tammar\_Wallaby PVPPEPPAKRVCTISPAPRHSPALSLPLMNPGGQFHPTPPPLQHYTLEDFATSHLYRDS-**

**Common\_Wombat PVPPEPPAKRVCTISPAPRHSPALSLPLMNPGGQFHPTPPPLQHYTLEDFATSHLYRDS-**

**. .\* \* \*\* \*\*\*\*\*. \*:\*\* :. :. . \*\*\*\*\*\* \*\*\* \*:\*:. :: \*\*:**

**Human YRHPSHRDLR-------------DRNRPMGLHG----TRQEEMIDHRLTDREWAEEWKHL**

**Orca\_Whale YRHPSHRDLR-------------DRNRPMGLHG----TRQEEMIDHRLTDREWAEEWKHL**

**Tropical\_Clawed\_Frog SKIAEHRDIR-------------DRHHSSGVNGNLNNGYQEELVDHRLTEREWAEEWKHL**

Cow SKMIEHREVR-------------DRHHNLSLNG----GYQDELVDHRLTEREWADEWKHL

Chimpanzee NKMLEHREVR-------------DRHHSLGLNG----GYQDELVDHRLTEREWADEWKHL

Mouse NKMLEHREVR-------------ERHHNLSLNG----GYQDELVDHRLTEREWADEWKHL

**Society\_Finch SKMLEHREIR-------------DRHSGLGLNG----GYQDELVDHRLTEREWADEWKHL**

**Tammar\_Wallaby TKMLEHREIRGGGGNGGSGGGGGDRHHGIGLNG----GYQEELVDHRLTEREWADEWKHL**

**Common\_Wombat TKMLEHREIRGSA--GGGGGAGGDRHHGLGLNG----GYQEELVDHRLTEREWADEWKHL**

**. .\*\*::\* :\*: .::\* \*:\*::\*\*\*\*\*:\*\*\*\*:\*\*\*\*\***

**Human DHLLNCIMDMVEKTRRSLTVLRRCQEADREELNYWIRRYSDAED**

**Orca\_Whale DHLLNCIMDMVEKTRRSLTVLRRCQEADREELNYWIRRYSDAED**

**Tropical\_Clawed\_Frog DHALNCIMEMVEKTRRSMAVLRRCQEVDRDELNYWKRRFNESNE**

Cow DHALNCIMEMVEKTRRSMAVLRRCQESDREELNYWKRRFNENTE

Chimpanzee DHALNCIMEMVEKTRRSMAVLRRCQESDREELNYWKRRYNENTE

Mouse DHALNCIMEMVEKTRRSMAVLRRCQESDREELNYWKRRFNENTE

**Society\_Finch DHALNCIMEMVEKTRRSMAVLRRCQEADREELNYWKRRCSETAE**

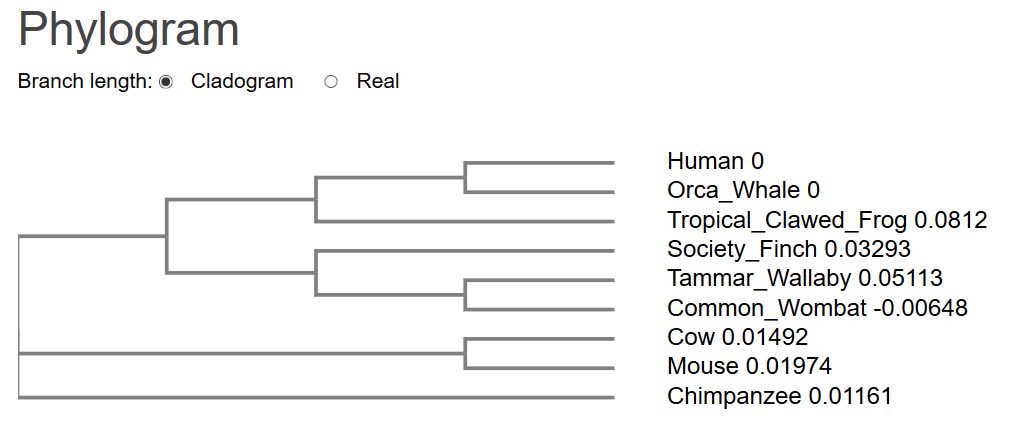
**Tammar\_Wallaby DHALNCIMEMVEKTRRSMAVLRRCQETDREELNFWKRRCSENTE**

**Common\_Wombat DHALNCIMEMVEKTRRSMAVLRRCQETDREELNFWKRRCSENTE**

\*\* \*\*\*\*\*:\*\*\*\*\*\*\*\*::\*\*\*\*\*\*\* \*\*:\*\*\*:\* \*\* .: :

**\*Note: I trimmed the alignment to match the coding region contained in the Wallaby EST data. Also, the Orca matches the Human perfectly; this is not a mistake (or at least not on my part), this is indeed the reported predicted Orca sequence. Also, I suspect the Chimp isoform numbers don’t match human, which explains the surprising phylogeny result to come!**

[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.

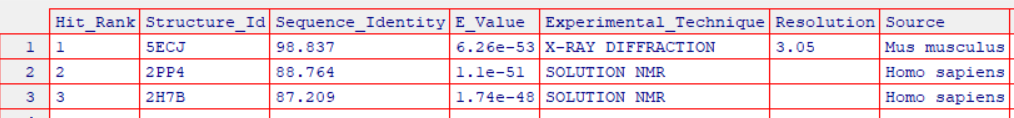


[Q7] Generate a sequence identity based heatmap of your aligned sequences using R.

A screenshot of a cell phone

Description automatically generated

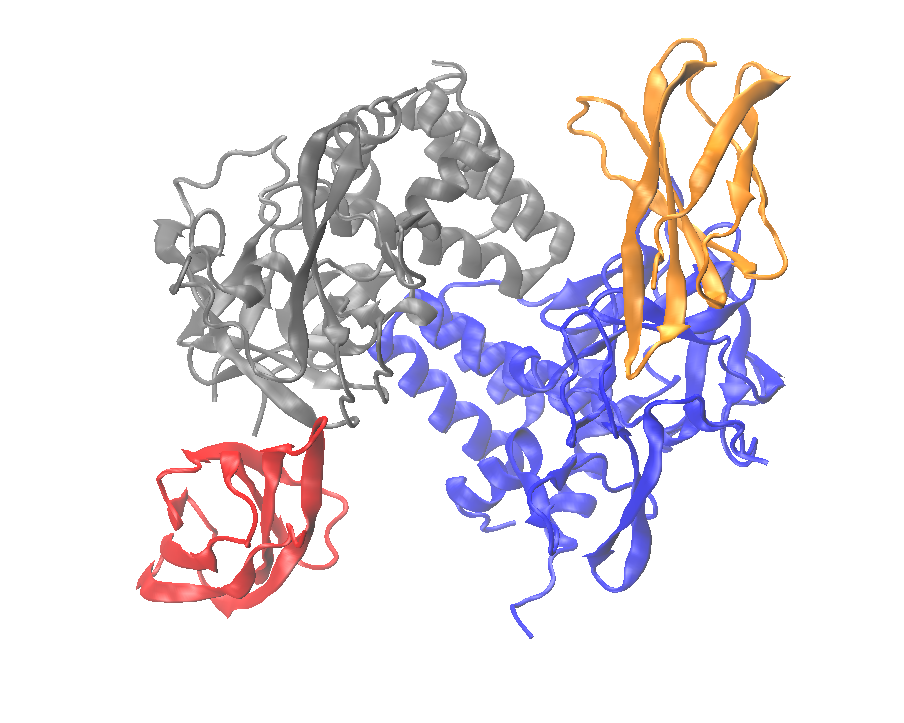
[Q8] Using R/Bio3D (or an online blast server if you prefer), search the main protein structure database for the most similar atomic resolution structures to your aligned sequences. List the top 3 unique hits (i.e. not hits representing different chains from the same structure) along with their Evalue 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).



[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?

5ECJ- Chains A,B,E,F. [Mouse CBFA2T2, bound to Prdm1]

This structure is extremely similar to my novel protein, CBFA2T1 consensus sequence (98.8% Identity Match, 86 residues!). Indeed, this is a structure of the CBFA2T1 family member, CBFA2T2.



[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 54 Targets identified by ChEMBEL search, with many Associated assays and ligand efficiency data (too much to show!).



