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.

Protein: 5HT2_A

Accession number: P28223

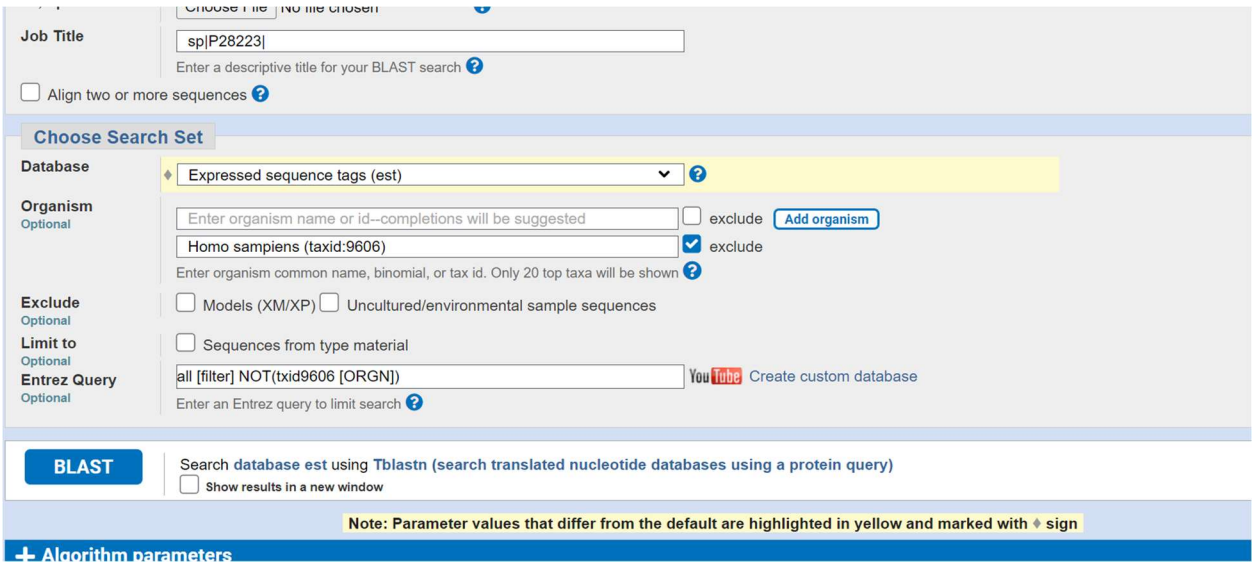
Species: *Homo sapiens*

[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 (est)

Organism: exclude *Homo sapiens*



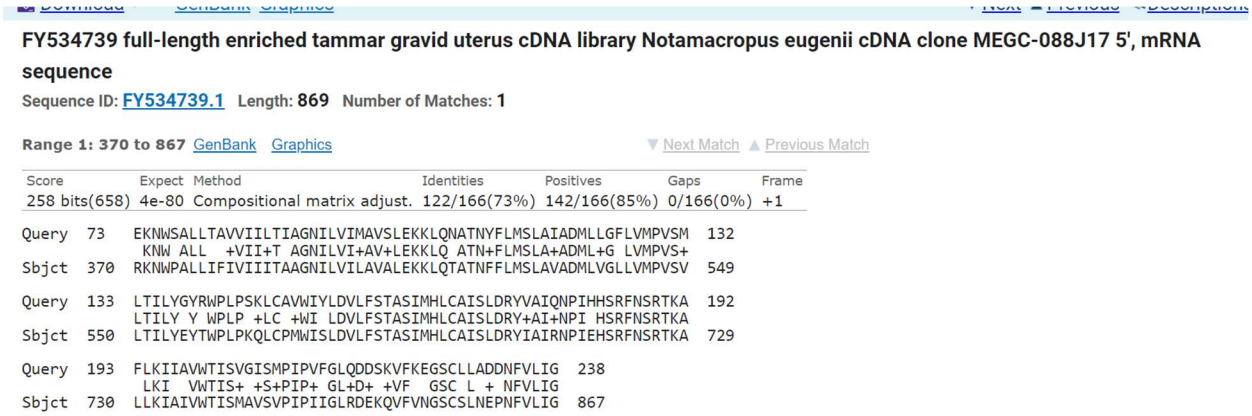
DKFZp459G1365_r1 459 (synonym: pcor1) Pongo abelii cDNA clone DKFZp459G1365 5', mRNA sequence	Pongo abelii	480	480	51%	5e-168	94.67%	736	CR789453.1
CJ446032 macaque brain cDNA library QflA Macaca fascicularis cDNA clone QflA-11203 5', mRNA sequence	Macaca fascicul…	446	446	52%	4e-154	89.80%	830	CJ446032.1
CJ459537 macaque brain cDNA library QmoA Macaca fascicularis cDNA clone QmoA-11767 5', mRNA seq…	Macaca fascicul…	395	395	45%	2e-134	90.23%	766	CJ459537.1
hb18a12.g1 Canis cDNAs from testes cells Canis lupus familiaris cDNA clone hb18a12 5', mRNA sequence	Canis lupus fam…	329	329	37%	3e-109	91.57%	594	BM540101.1
DRDBCa0014C01r DRDBCa Danio rerio cDNA, mRNA sequence	Danio rerio	288	288	48%	3e-92	59.57%	744	GO937196.1
4154902 BARC_3GAL chicken mixed tissue Gallus gallus cDNA clone 3GAL_52H08 5', mRNA sequence	Gallus gallus	278	278	38%	5e-89	75.42%	622	DN929261.1
35213 MARC 1BOV Bos taurus cDNA 5', mRNA sequence	Bos taurus	265	265	28%	4e-85	96.99%	400	AW352999.1
DKFZp459F0616_r1 459 (synonym: pcor1) Pongo abelii cDNA clone DKFZp459F0616 5', mRNA sequence	Pongo abelii	252	252	28%	4e-80	90.91%	410	CR542424.1
FY534739 full-length enriched tammar gravid uterus cDNA library Notamacropus eugenii cDNA clone MEG…	Notamacropus …	258	258	35%	4e-80	73.49%	869	FY534739.1
SB06009B1H01.f1 Normalized subtracted Keck-Tagu Library SB06 Taeniopygia guttata cDNA clone SB060…	Taeniopygia gutt…	253	253	51%	2e-78	49.43%	781	FE732340.1
AL919835 PJR-Z1+Z2 Danio rerio cDNA clone 069-F07-2, mRNA sequence	Danio rerio	241	241	32%	2e-75	75.50%	506	AL919835.1
AB595034 rockbream gills cDNA Oplegnathus fasciatus cDNA clone 07-B10, mRNA sequence	Oplegnathus fas…	242	242	43%	1e-73	56.40%	906	AB595034.1
JGI_XZT45146.fwd NIH_XGC_tropTad5 Xenopus tropicalis cDNA clone IMAGE:7620034 5', mRNA sequence	Xenopus tropicalis	227	227	49%	2e-68	46.06%	820	CX347999.2
12 Capra hircus (breed Majorera goat) cDNA library Capra hircus cDNA 5' similar to 5-hydroxytryptamine 2A…	Capra hircus	221	221	22%	2e-68	100.00%	322	JZ845012.1
11 Ovis aries (breed Canarian sheep) cDNA library Ovis aries cDNA 5' similar to 5-hydroxytryptamine 2A re…	Ovis aries	219	219	22%	1e-67	99.06%	322	JZ844999.1
UI-M-FY0-cgp-p-05-0-UI.r1 NIH_BMAP_FY0 Mus musculus cDNA clone IMAGE:30356044 5', mRNA seque…	Mus musculus	193	193	19%	9e-58	100.00%	292	CF531304.1
JGI_ANNO42327.fwd ANNO Pimephales promelas Whole (M) Pimephales promelas cDNA clone ANNO423…	Pimephales pro…	190	190	29%	8e-56	63.04%	423	DT170865.1

match:

Accession
FY534739.1

Species: *Notamacropus eugenii*

Base pairs: 869



FY534739 full-length enriched tammar gravid uterus cDNA library Notamacropus eugenii cDNA clone MEGC-088J17 5', mRNA sequence

Sequence ID: FY534739.1 Length: 869 Number of Matches: 1

Range 1: 370 to 867 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
258 bits(658)	4e-80	Compositional matrix adjust.	122/166(73%)	142/166(85%)	0/166(0%)	+1

```
Query  73   EKNWSALLTAVVIILTIAGNILVIMAVSLEKKLQNATNYFLMSLAIADMLLGFLVMPVSM  132
            KNW ALL  +VII+T AGNILVI+AV+LEKKLQ ATN+FLMSLA+ADML+G LVMPVS+
Sbjct  370  RKNWPALLIFIVIIITAAGNILVILAVALEKKLQTATNFFLMSLAVADMLVGLLVMPVSV  549

Query  133  LTILYGYRWPLPSKLCAVWIYLDVLFSTASIMHLCAISLDRYVAIQNPIHHSRFNSRTKA  192
            LTILY Y WPLP +LC +WI LDVLFSTASIMHLCAISLDRY+AI+NPI HSRFNSRTKA
Sbjct  550  LTILYEYTWPLPKQLCPMWISLDVLFSTASIMHLCAISLDRYIAIRNPIEHSRFNSRTKA  729

Query  193  FLKIIAVWTISVGISMPIPVFGLQDDSKVFKEGSCLLADDNFVLIG  238
             LKI  VWTIS+ +S+PIP+ GL+D+ +VF  GSC L + NFVLIG
Sbjct  730  LLKIAIVWTISMAVSVPIPIIGLRDEKQVFVNGSCSLNEPNFVLIG  867
```

[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. 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.

Translated sequence:

```
>370-867_1 full-length enriched tammar gravid uterus
cDNA library Notamacropus eugenii cDNA clone MEGC-
088J17 5', mRNA sequence
RKNWPALLIFIVIIITAAGNILVILAVALEKKLQTATNFFLMSLAVADMLVGLLV
MPVSV
LTILYEYTWPLPKQLCPMWISLDVLFSTASIMHLCAISLDRYIAIRNPIEHSRFN
SRTKA
LLKIAIVWTISMAVSVPIPIIGLRDEKQVFVNGSCSLNEPNFVLIG
>370-867_2 full-length enriched tammar gravid uterus
cDNA library Notamacropus eugenii cDNA clone MEGC-
088J17 5', mRNA sequence
ARTGLPS*SS*SSSSLPLAISW*SWPWPWRRNCRLRPTSS*CLWPWQTC*SGY**
CQCLS
SPSSTNTPGLFQNNSAQCGSPLMCSSRLHPSCTSVLSPWIATLLSGTQLSIAAST
LALRP
F*RLPLSGPSRWLCPCPSRSSA*GMRSRSL*MEAAA*TNPTLCSLA
>370-867_3 full-length enriched tammar gravid uterus
cDNA library Notamacropus eugenii cDNA clone MEGC-
088J17 5', mRNA sequence
QELACPPDLHSHHHHCRWQYPGDLGRGPGEETADCDQLLLDVSGRGRHVSRVISD
ASVCP
HHPLRIHLASSKTTLPNVDLP*CALLDCIHHAPLCYLPGSLHCYPEPN*A*PLQL
SH*GP
SEDCHCLDHLDGCVRAHPDHRPEG*EAGLCEWKLQPERTQLCAHWX
```

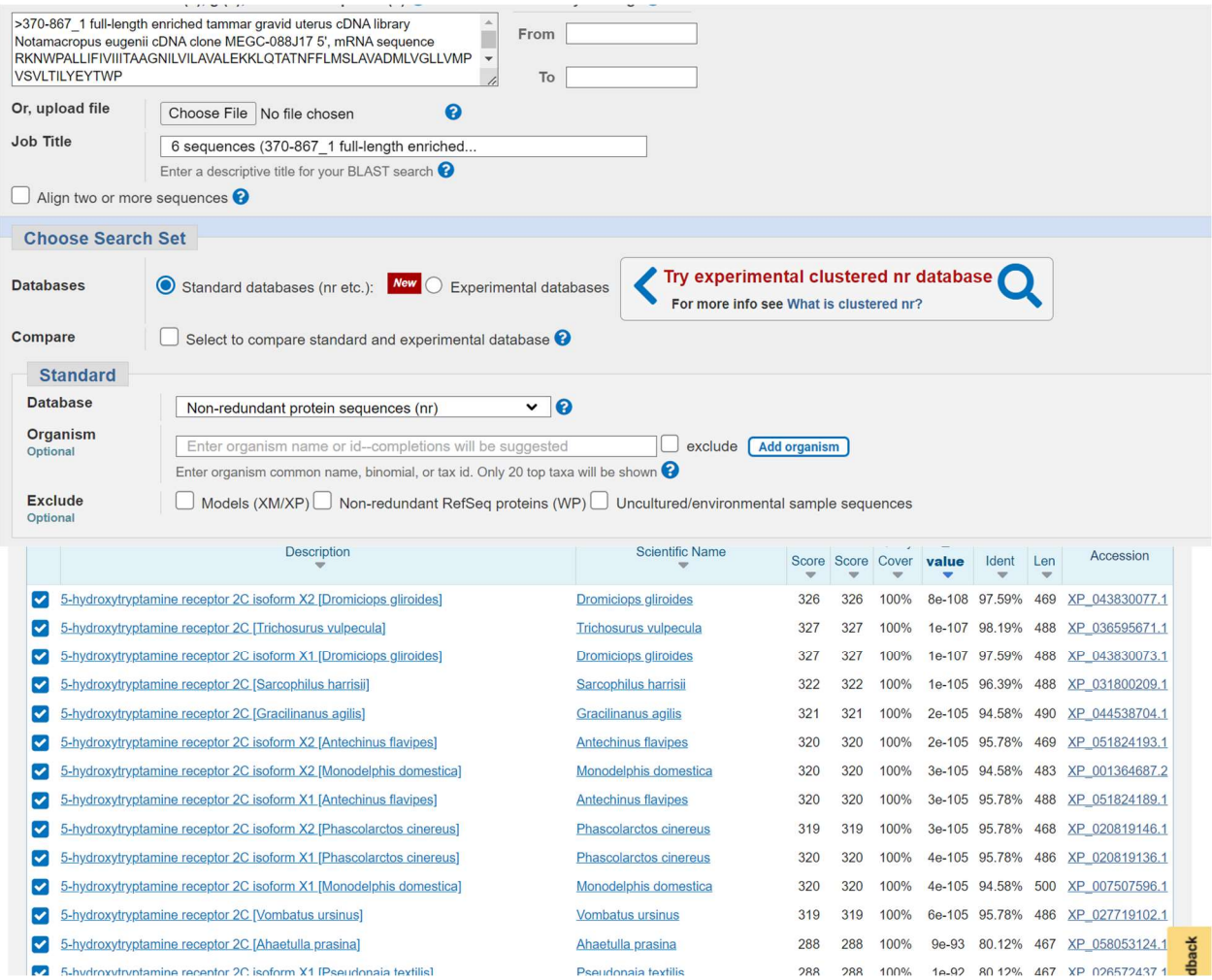
>370-867_4 full-length enriched tammar gravid uterus cDNA library Notamacropus eugenii cDNA clone MEGC-088J17 5', mRNA sequence
ANEHKVGFVQAAASIHKDLLLIPQADDRDGHGHSHRDGPDNGNLQKGLSARVEAA
MLNWV
PDSNVAIQGDSTEVHDGCSREEHIKGDPHWAELFWKRPGVFVEDGEDRHWHH**P
D*HVC
HGQRHQEEVGRSLQFLLQGHGQDHQDIASGSDDDDYEDQEGRPVLA
>370-867_5 full-length enriched tammar gravid uterus cDNA library Notamacropus eugenii cDNA clone MEGC-088J17 5', mRNA sequence
Q*AQSWVRSGCSFHSQRPASHPSGR*SGWARTQPSRWSRQWQSSEGP*CES*SGY
AQLGS
G*QCSDPGR*HRGA*WMQSRRAHQGRSTLGRVVLEEARCIRRGW*GQTLASLITR
LTCLP
RPETSRRSWSQSAVSSPGPRPRSPGYCQRQ****L*RSGGQASSCX
>370-867_6 full-length enriched tammar gravid uterus cDNA library Notamacropus eugenii cDNA clone MEGC-088J17 5', mRNA sequence
PMSTKLGSFRLQLPFTKTCFSSLRPMIGMGTDTAIEMVQTMAIFRRALVRELKRL
CSIGF
RIAM*RSREIAQRCMMDAVEKSTSREIHIGQSCFGRGQVYS*RMVRTDTGITNNP
TNMSA
TARDIKKKLVAVCSFFSRATAKITRILPAAVMMMTMKIRRAGQFLR

[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.

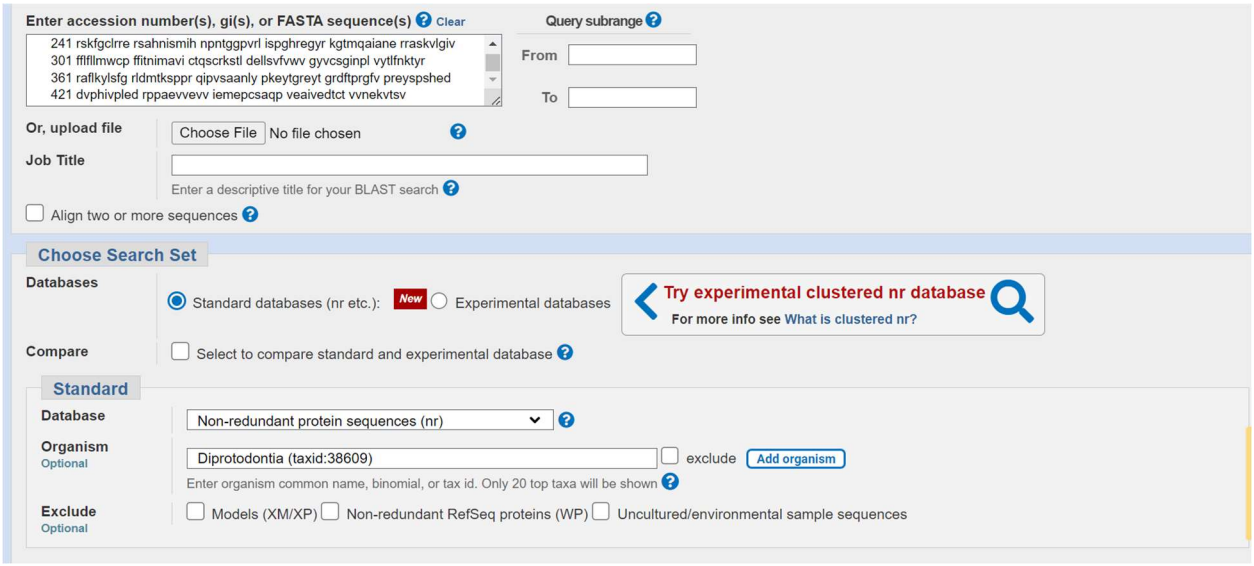
• If there are no database matches to the original query from [Q1], this indicates that you have partially succeeded: yes, you may have found a new gene, but no, it is not actually homologous to the original query. You should probably start over.



Description	Scientific Name	Score	Score	Cover	value	Ident	Len	Accession
5-hydroxytryptamine receptor 2C isoform X2 [Dromiciops gliroides]	Dromiciops gliroides	326	326	100%	8e-108	97.59%	469	XP_043830077.1
5-hydroxytryptamine receptor 2C [Trichosurus vulpecula]	Trichosurus vulpecula	327	327	100%	1e-107	98.19%	488	XP_036595671.1
5-hydroxytryptamine receptor 2C isoform X1 [Dromiciops gliroides]	Dromiciops gliroides	327	327	100%	1e-107	97.59%	488	XP_043830073.1
5-hydroxytryptamine receptor 2C [Sarcophilus harrisii]	Sarcophilus harrisii	322	322	100%	1e-105	96.39%	488	XP_031800209.1
5-hydroxytryptamine receptor 2C [Gracilinanus agilis]	Gracilinanus agilis	321	321	100%	2e-105	94.58%	490	XP_044538704.1
5-hydroxytryptamine receptor 2C isoform X2 [Antechinus flavipes]	Antechinus flavipes	320	320	100%	2e-105	95.78%	469	XP_051824193.1
5-hydroxytryptamine receptor 2C isoform X2 [Monodelphis domestica]	Monodelphis domestica	320	320	100%	3e-105	94.58%	483	XP_001364687.2
5-hydroxytryptamine receptor 2C isoform X1 [Antechinus flavipes]	Antechinus flavipes	320	320	100%	3e-105	95.78%	488	XP_051824189.1
5-hydroxytryptamine receptor 2C isoform X2 [Phascolarctos cinereus]	Phascolarctos cinereus	319	319	100%	3e-105	95.78%	468	XP_020819146.1
5-hydroxytryptamine receptor 2C isoform X1 [Phascolarctos cinereus]	Phascolarctos cinereus	320	320	100%	4e-105	95.78%	486	XP_020819136.1
5-hydroxytryptamine receptor 2C isoform X1 [Monodelphis domestica]	Monodelphis domestica	320	320	100%	4e-105	94.58%	500	XP_007507596.1
5-hydroxytryptamine receptor 2C [Vombatus ursinus]	Vombatus ursinus	319	319	100%	6e-105	95.78%	486	XP_027719102.1
5-hydroxytryptamine receptor 2C [Ahaetulla prasina]	Ahaetulla prasina	288	288	100%	9e-93	80.12%	467	XP_058053124.1
5-hydroxytryptamine receptor 2C isoform X1 [Pseudonaja textilis]	Pseudonaja textilis	288	288	100%	1e-92	80.12%	467	XP_026572437.1

[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.



>370-867_1 full-length enriched tammar gravid uterus cDNA library Notamacropus eugenii cDNA clone MEGC-088J17 5', mRNA sequence
RKNWPALLIFIVIIITAAGNILVILAVALEKKLQTATNFFLMSLAVADMLVGLLV
MPVSV
LTILYEYTWPLPKQLCPMWISLDVLFSTASIMHLCAISLDRYIAIRNPIEHSRFN
SRTKA
LLKIAIVWTISMAVSVPIPIIGLRDEKQVFVNGSCSLNEPNFVLIG

> Trichosurus_vulpecula XP_036595671.1 5-hydroxytryptamine receptor 2C
MSALLPRLRTLSSVIQNGIMALQLSRNLTDTGLNEYVNGTDNSTTPAPTSAPGPI
RKNWPALLIFIVIII
TVAGNILVILAVALEKKLQTATNFFLMSLAVADMLVGLLVMPVSVLTILYEYTWP
LPKQLCPMWISLDVL

FSTASIMHLCAISLDRYIAIRNPIEHSRFNSRTKALLKIAIVWTISMAVSVPIPI
IGLRDESKVFVNGSC
SLNEPNFVLIGSFVAFFIPLFIMVITYCLTIQVLQGQSNVFGPGERRRRRSKFGC
LRRERSAHNIAVIHN
PTTGGPVRLISPGHREGYRKGTMQAIANERRASKVLGIVFFLFLLMWCPFFITNI
MAVICTQSCRKSTLD
ELLSVFVWVGYVCSGVNPLVYTLFNKTYRRAFLKYLSFGWLGKTKSPPRQIPVSA
TNLYPREYTGPEYTG
RDFAPRGYVPREYSPSHDEDVPQIVPLEERPNTEVVEVVIEMEPCSPQPVEARVE
DTCTMVNEKVSSV

>[Vombatus_ursinus] XP_027719102.1 5-hydroxytryptamine
receptor 2C
MSLLIPRLRILSSVLQNSIMALQLSRNMTDDGMDEYTNSTDNSTAPTSAPGHIRK
NWPALLISIVIIITV
AGNILVILAVALEKKLQTATNFFLMSLAVADMLVGLLVMPVSVVTIFYEYTWPLP
KQLCPMWISLDVLFS
TASIMHLCAISLDRYIAIRNPIEHSRFNSRTKALLKIAIVWTISMAVSLPIPIIG
LRDESKVFVNGSCSL
NEPNFVLIGSFVAFFIPLFIMVITYCLTIQVLQGQSSVFGPGERRRRRSRFGCLR
RERSAHNISMIHNPN
TGVPVRLISPGHREGYRKGTMQAIANERRASKVLGIVFFLFLLMWCPFFITNIMA
VICTQSCRKSTLDEL
LSVFVWVGYVCSGINPLVYTLFNKTYRRAFLKYLSFGWLGRTKSPPRQLPVSAAN
LYPREYTGRQYTGRD
FTPRVFVPREYSPSHDEDVPQIVPLEDRPNTEVVEVVIEMEPCSPQPVEARVEDT
CTMVNEKVTSV

>[Phascolarctos_cinereus] XP_020819136.1 5-
hydroxytryptamine receptor 2C isoform X1
MSLLIPRLKILSSILQNSMMALQLSRNVSDDGMDEYTNSTDNSTAPTSAPGHIRK
NWPALLISIVIIITV
AGNILVILAVALEKKLQTATNFFLMSLAVADMLVGLLVMPVSVVTIFYEYTWPLP
KQLCPMWISLDVLFS
TASIMHLCAISLDRYIAIRNPIEHSRFNSRTKALLKIAIVWTISMAVSLPIPIIG
LRDESKVFVNGSCSL
NEPNFVLIGSFVAFFIPLFIMVITYCLTIQVLQGQSSVFGPGERRRRRSRFGCLR
RERSAHNISMIHNPN
TGVPVRLISPGHREGYRKGTMQAIANERRASKVLGIVFFLFLLMWCPFFITNIMA
VICTQSCRKSTMDEL

LSVFVWVGYVCSGINPLVYTLFNKTYRRAFLKYLSFGWLGKTKSPPRQLPVSAAN
LYPREYTGRQYTGRD
FTPRVFVPREYSPSHDEDVPQIVPFEERPNTEVVEVVIEMEPCSPQPAEAIVEDT
CTMVNEKVTSV

>[Dromiciops_gliroides] XP_043830077.1 5-
hydroxytryptamine receptor 2C isoform X2 (from BLAST
results)
MALQLSRNLTDTGLDEYMNGTDNSTKPEPTSAPGPIRKNWPALLIFIVIVITVAG
NILVILAVALEKKLQ
TATNFFLMSLAVADMLVGLLVMPVSVLTILYEYTWPLPKQLCPMWISLDVLFSTA
SIMHLCAISLDRYIA
IRNPIEHSRFNSRTKALLKIAIVWTISMAVSVPIPIIGLRDESKVFVNGSCSLNE
PNFVLIGSFVAFFIP
LFIMVITYCLTIQVLQGQSSVFGPGERRRKRSKFGCLRRERSAHNISMIHNPNTG
GPVRLISPGHREGYR
KGTMQAIANERRASKVLGIVFFLFLLMWCPFFITNIMAVICTQSCRKSTLDELLS
VFVWVGYVCSGINPL
VYTLFNKTYRRAFLKYLSFGRLDMTKSPPRQIPVSAANLYPKEYTGREYTGRDFT
PRGFVPREYSPSHED
DVPHIVPLEDRPPAEVVEVVIEMEPCSAQPVEAIVEDTCTVVNEKVTSV

>sp|P28223|5HT2A_HUMAN 5-hydroxytryptamine receptor 2A
OS=Homo sapiens OX=9606 GN=HTR2A PE=1 SV=2
MDILCEENTSLSSTTNSLMQLNDDTRLYSNDFNSGEANTSDAFNWTVDSENRTNL
SCEGC
LSPSCLSLLHLQEKNWSALLTAVVIILTIAGNILVIMAVSLEKKLQNATNYFLMS
LAIAD
MLLGFLVMPVSMLTILYGYRWPLPSKLCAVWIYLDVLFSTASIMHLCAISLDRYV
AIQNP
IHHSRFNSRTKAFLKIIAVWTISVGISMPIPVFGLQDDSKVFKEGSCLLADDNFV
LIGSF
VSFFIPLTIMVITYFLTIKSLQKEATLCVSDLGTRAKLASFSFLPQSSLSSEKLF
QRSIH
REPGSYTGRRTMQSISNEQKACKVLGIVFFLFVVMWCPFFITNIMAVICKESCNE
DVIGA
LLNVFVWIGYLSSAVNPLVYTLFNKTYRSAFSRYIQCQYKENKKPLQLILVNTIP
ALAYK
SSQLQMGQKKNSKQDAKTTDNDCSMVALGKQHSEEASKDNSDGVNEKVSCV

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


sp|P28223|5HT2A_HUMAN
MDILCEENTSLSSTTNSLMQLNDDTRLYSNDFNSGEANTSDAFNWTVDSENRTNL
SCEGC
[Vombatus_ursinus]
MSLLIPRLRILSSVLQNSIMALQLSR------NMTDDGMDEYTNSTDNST--
APTSAPGH
[Phascolarctos_cinereus]
MSLLIPRLKILSSILQNSMMALQLSR------NVSDDGMDEYTNSTDNST--
APTSAPGH
[Dromiciops_gliroides]        ------------------
MALQLSR------NLTDTGLDEYMNGTDNSTKPEPTSAPGP
Trichosurus_vulpecula
MSALLPRLRTLSSVIQNGIMALQLSR------
NLTDTGLNEYVNGTDNSTTPAPTSAPGP
[Notamacropus eugenii]        -------------------------
-----------------------------------


sp|P28223|5HT2A_HUMAN
LSPSCLSLLHLQEKNWSALLTAVVIILTIAGNILVIMAVSLEKKLQNATNYFLMS
LAIAD
[Vombatus_ursinus]            IR-----------
KNWPALLISIVIIITVAGNILVILAVALEKKLQTATNFFLMSLAVAD
[Phascolarctos_cinereus]      IR-----------
KNWPALLISIVIIITVAGNILVILAVALEKKLQTATNFFLMSLAVAD
[Dromiciops_gliroides]        IR-----------
KNWPALLIFIVIVITVAGNILVILAVALEKKLQTATNFFLMSLAVAD
Trichosurus_vulpecula         IR-----------
KNWPALLIFIVIIITVAGNILVILAVALEKKLQTATNFFLMSLAVAD
[Notamacropus eugenii]        -R-----------
KNWPALLIFIVIIITAAGNILVILAVALEKKLQTATNFFLMSLAVAD
```

```
                                                        ***.***
:**::* *******:**:******.***:******:**

sp|P28223|5HT2A_HUMAN
MLLGFLVMPVSMLTILYGYRWPLPSKLCAVWIYLDVLFSTASIMHLCAISLDRYV
AIQNP
[Vombatus_ursinus]
MLVGLLVMPVSVVTIFYEYTWPLPKQLCPMWISLDVLFSTASIMHLCAISLDRYI
AIRNP
[Phascolarctos_cinereus]
MLVGLLVMPVSVVTIFYEYTWPLPKQLCPMWISLDVLFSTASIMHLCAISLDRYI
AIRNP
[Dromiciops_gliroides]
MLVGLLVMPVSVLTILYEYTWPLPKQLCPMWISLDVLFSTASIMHLCAISLDRYI
AIRNP
Trichosurus_vulpecula
MLVGLLVMPVSVLTILYEYTWPLPKQLCPMWISLDVLFSTASIMHLCAISLDRYI
AIRNP
[Notamacropus eugenii]
MLVGLLVMPVSVLTILYEYTWPLPKQLCPMWISLDVLFSTASIMHLCAISLDRYI
AIRNP
                                    **:*:******::**:* *
****.:**.:** **********************:**.**

sp|P28223|5HT2A_HUMAN
IHHSRFNSRTKAFLKIIAVWTISVGISMPIPVFGLQDDSKVFKEGSCLLADDNFV
LIGSF
[Vombatus_ursinus]
IEHSRFNSRTKALLKIAIVWTISMAVSLPIPIIGLRDESKVFVNGSCSLNEPNFV
LIGSF
[Phascolarctos_cinereus]
IEHSRFNSRTKALLKIAIVWTISMAVSLPIPIIGLRDESKVFVNGSCSLNEPNFV
LIGSF
```

```
[Dromiciops_gliroides]
IEHSRFNSRTKALLKIAIVWTISMAVSVPIPIIGLRDESKVFVNGSCSLNEPNFV
LIGSF
Trichosurus_vulpecula
IEHSRFNSRTKALLKIAIVWTISMAVSVPIPIIGLRDESKVFVNGSCSLNEPNFV
LIGSF
[Notamacropus eugenii]
IEHSRFNSRTKALLKIAIVWTISMAVSVPIPIIGLRDEKQVFVNGSCSLNEPNFV
LIG--
                                  * **********:***
*****:.:*:***::**.*:.:** :*** * : ******

sp|P28223|5HT2A_HUMAN
VSFFIPLTIMVITYFLTIKSLQKEATLCVSDLGTRAKLASFSFLPQSSLSSEKLF
QRSIH
[Vombatus_ursinus]
VAFFIPLFIMVITYCLTIQVLQGQSSVFGPGERRRRRSRFGCLRRERSAHNISMI
HNPNT
[Phascolarctos_cinereus]
VAFFIPLFIMVITYCLTIQVLQGQSSVFGPGERRRRRSRFGCLRRERSAHNISMI
HNPNT
[Dromiciops_gliroides]
VAFFIPLFIMVITYCLTIQVLQGQSSVFGPGERRRKRSKFGCLRRERSAHNISMI
HNPNT
[Notamacropus eugenii]
VAFFIPLFIMVITYCLTIQVLQGQSNVFGPGERRRRRSKFGCLRRERSAHNIAVI
HNPTT
370-867_1                         -------------------------
-----------------------------------


sp|P28223|5HT2A_HUMAN             R------EPGSYTGRR--
TMQSISNEQKACKVLGIVFFLFVVMWCPFFITNIMAVICKES
```

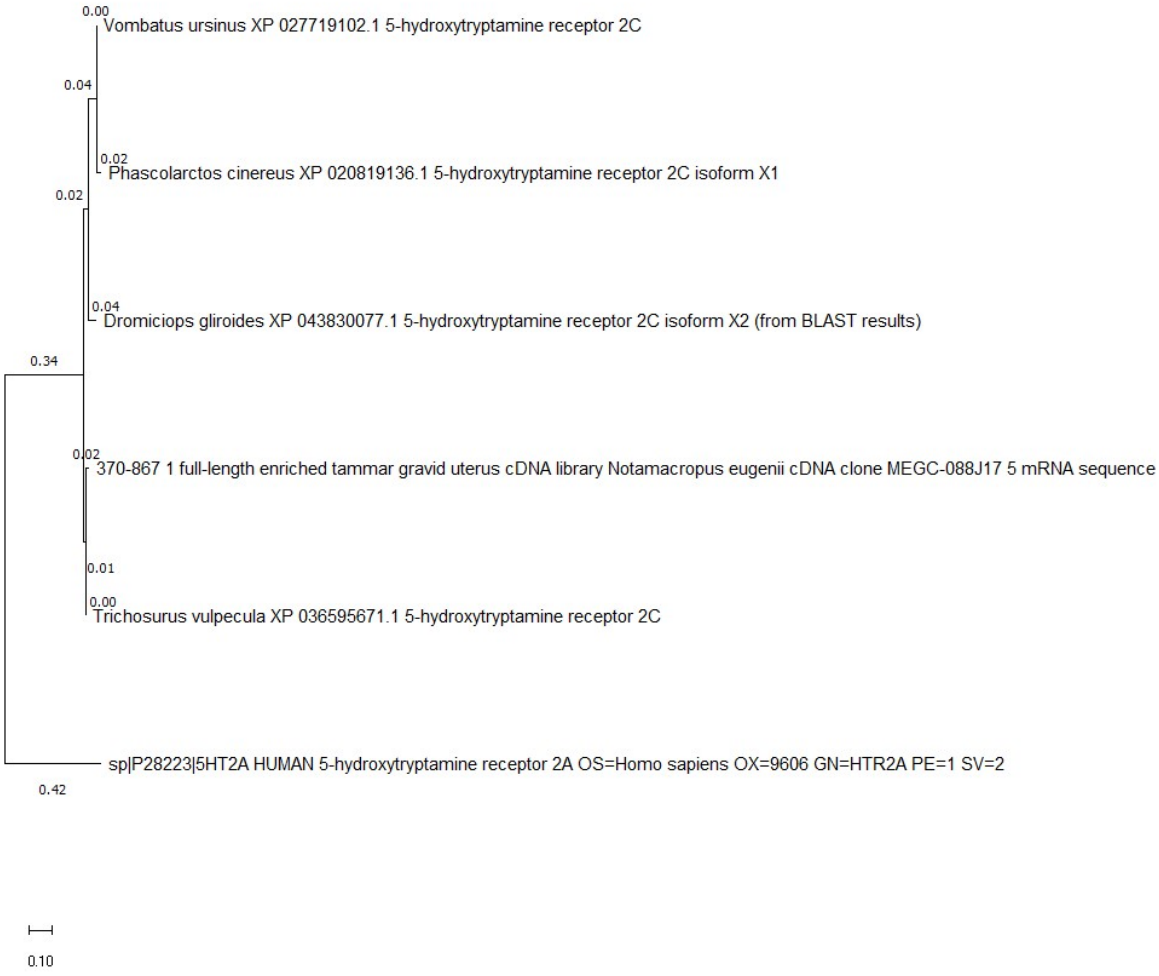
[Vombatus_ursinus]
GVPVRLISPGHREGYRKGTMQAIANERRASKVLGIVFFLFLLMWCPFFITNIMAV
ICTQS
[Phascolarctos_cinereus]
GVPVRLISPGHREGYRKGTMQAIANERRASKVLGIVFFLFLLMWCPFFITNIMAV
ICTQS
[Dromiciops_gliroides]
GGPVRLISPGHREGYRKGTMQAIANERRASKVLGIVFFLFLLMWCPFFITNIMAV
ICTQS
Trichosurus_vulpecula
GGPVRLISPGHREGYRKGTMQAIANERRASKVLGIVFFLFLLMWCPFFITNIMAV
ICTQS
[Notamacropus eugenii] ----------
----

sp|P28223|5HT2A_HUMAN
CNEDVIGALLNVFVWIGYLSSAVNPLVYTLFNKTYRSAFSRYIQCQY-
KENKKPLQLILV
[Vombatus_ursinus]
CRKSTLDELLSVFVWVGYVCSGINPLVYTLFNKTYRRAFLKYLSFGWLGRTKSPP
RQLPV
[Phascolarctos_cinereus]
CRKSTMDELLSVFVWVGYVCSGINPLVYTLFNKTYRRAFLKYLSFGWLGKTKSPP
RQLPV
[Dromiciops_gliroides]
CRKSTLDELLSVFVWVGYVCSGINPLVYTLFNKTYRRAFLKYLSFGRLDMTKSPP
RQIPV
Trichosurus_vulpecula
CRKSTLDELLSVFVWVGYVCSGVNPLVYTLFNKTYRRAFLKYLSFGWLGKTKSPP
RQIPV
[Notamacropus eugenii] ----------

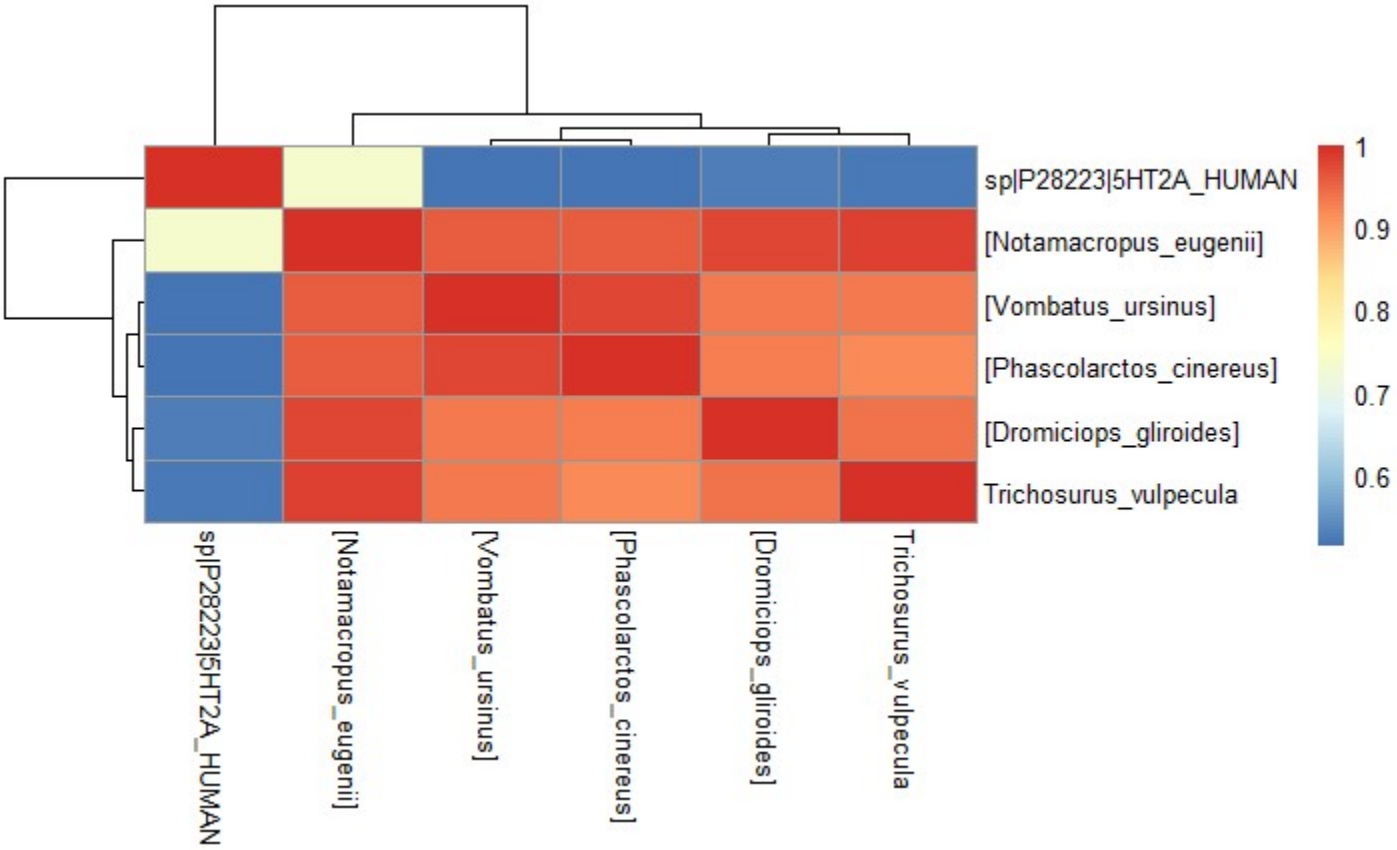
```
sp|P28223|5HT2A_HUMAN          NTIPALAYKSSQLQMGQKK------
-------------------NSKQDAKTTDNDCSMV
[Vombatus_ursinus]
SAANLYPREYTGRQYTGRDFTPRVFVPREYSPSHDEDVPQIVPLEDRPNTEVVEV
VIEME
[Phascolarctos_cinereus]
SAANLYPREYTGRQYTGRDFTPRVFVPREYSPSHDEDVPQIVPFEERPNTEVVEV
VIEME
[Dromiciops_gliroides]
SAANLYPKEYTGREYTGRDFTPRGFVPREYSPSHEDDVPHIVPLEDRPPAEVVEV
VIEME
Trichosurus_vulpecula
SATNLYPREYTGPEYTGRDFAPRGYVPREYSPSHDEDVPQIVPLEERPNTEVVEV
VIEME
[Notamacropus eugenii]         -------------------------
-----------------------------------


sp|P28223|5HT2A_HUMAN          ALGKQHSEEASKDNSDGVNEKVSCV
[Vombatus_ursinus]             PCSPQPVEARVEDTCTMVNEKVTSV
[Phascolarctos_cinereus]       PCSPQPAEAIVEDTCTMVNEKVTSV
[Dromiciops_gliroides]         PCSAQPVEAIVEDTCTVVNEKVTSV
Trichosurus_vulpecula          PCSPQPVEARVEDTCTMVNEKVSSV
[Notamacropus eugenii]         -------------------------
```

[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.



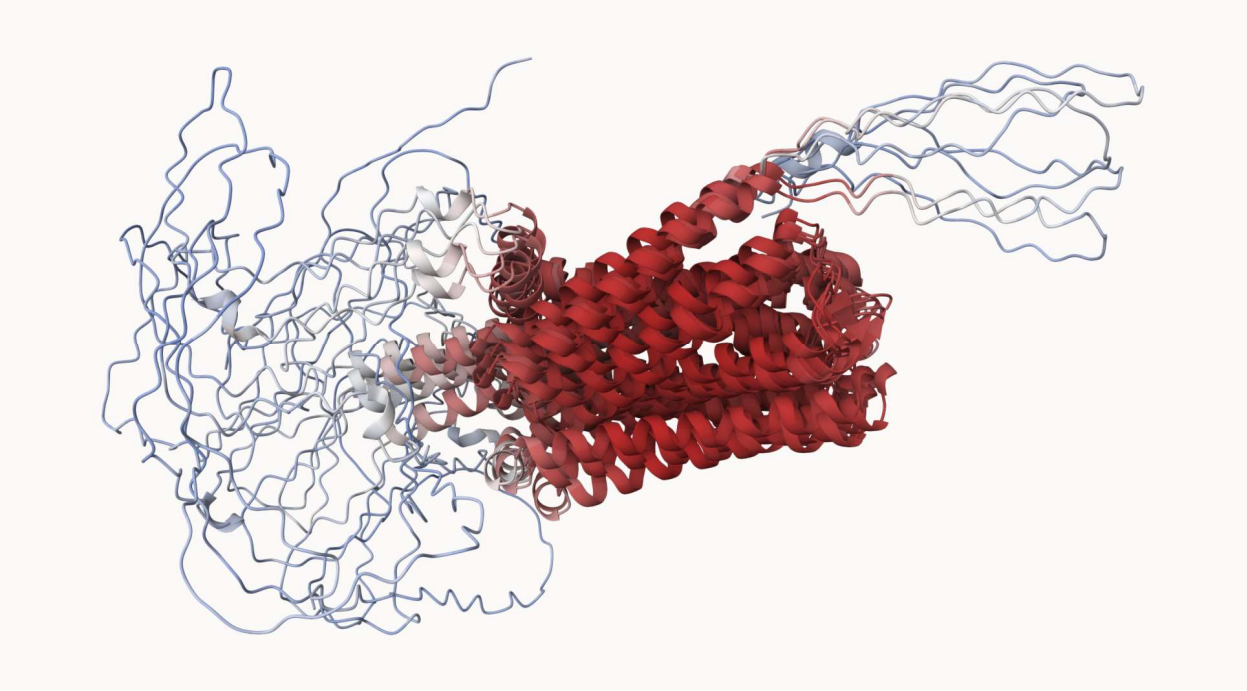
[Q7] Generate a sequence identity based heatmap of your aligned sequences using R. If necessary convert your sequence alignment to the ubiquitous FASTA format (Seaview can read in clustal format and “Save as” FASTA format for example). Read this FASTA format alignment into R with the help of functions 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.

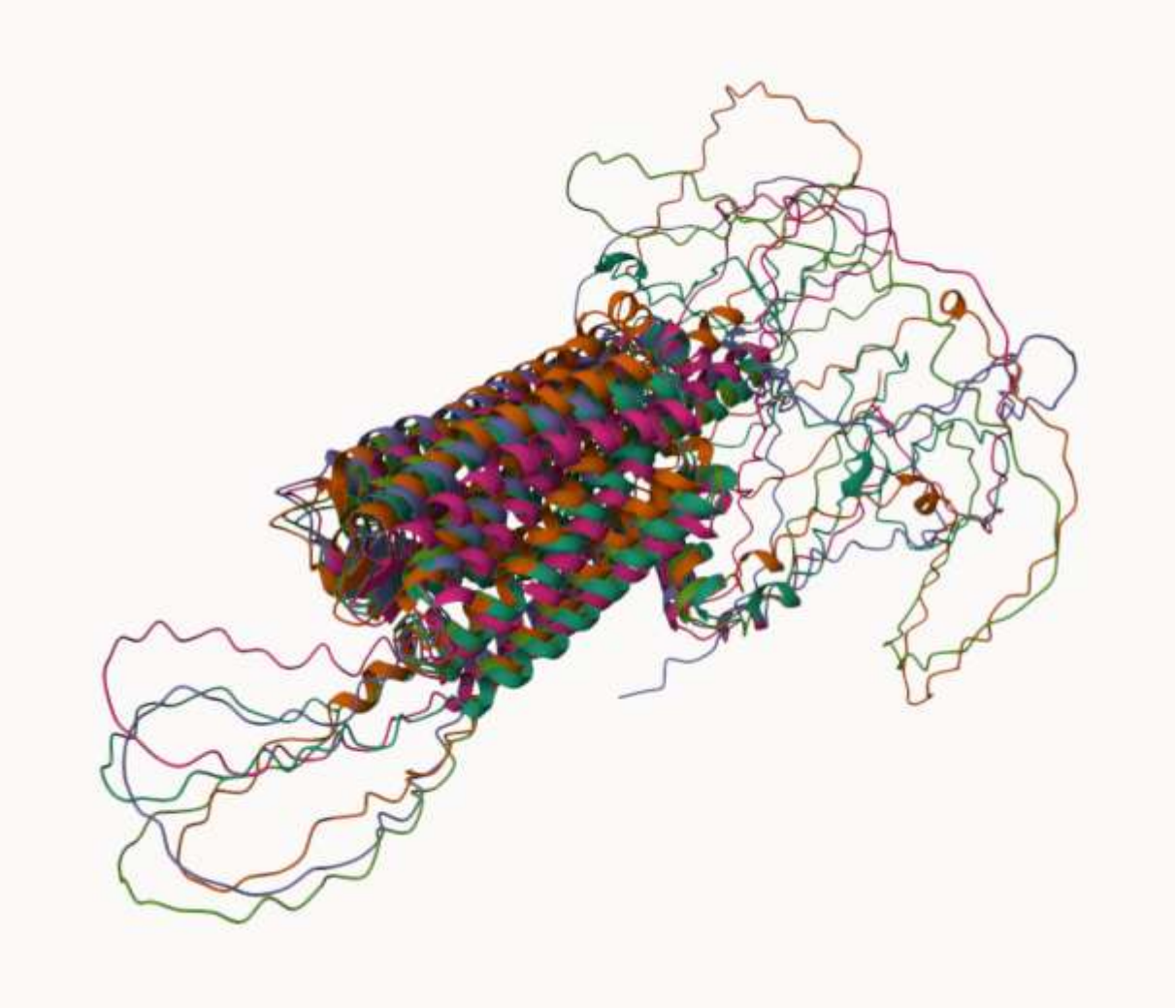


[Q8] Using R/Bio3D (or an online blast server if you prefer), search the main protein structure database for the most similar atomic resolution structures to your aligned sequences. List the top 3 unique hits (i.e. not hits representing different chains from the same structure) along with their Evalue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

PDB_id	Technique	Resolution	Source	Evalue	Identity
7RAN	EM	3.45	Homo sapiens	8e-86	73.49%
6WHA	EM	3.36	E. coli	2e-84	73.49%
6VMS	EM	3.80	Tequatrovirus T4	5e-31	40.85%

[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.).





[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]”.

There are 452, 541 items in this target search.

An example of an inhibitor is ID:

https://www.ebi.ac.uk/chembl/web_components/explore/activities/STATE_ID:M2BNaMeTmLJ_FyXThlOydg==

Which does this:

	Inverse agonist activity at alpha1A adrenoceptor in guinea pig thoracic aorta assessed as inhibition of ca2+-induced IRT at 10 nM after 30 mins

https://www.ebi.ac.uk/chembl/web_components/explore/activities/STATE_ID:iC_JHnDfluD6Np6NkaDLZw==