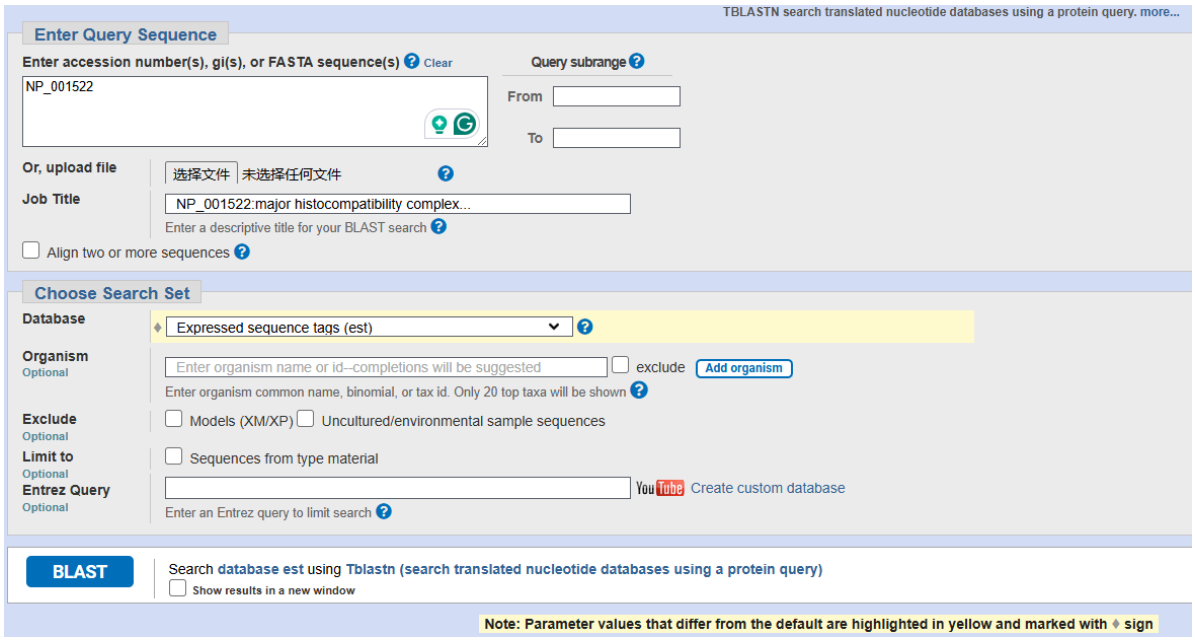
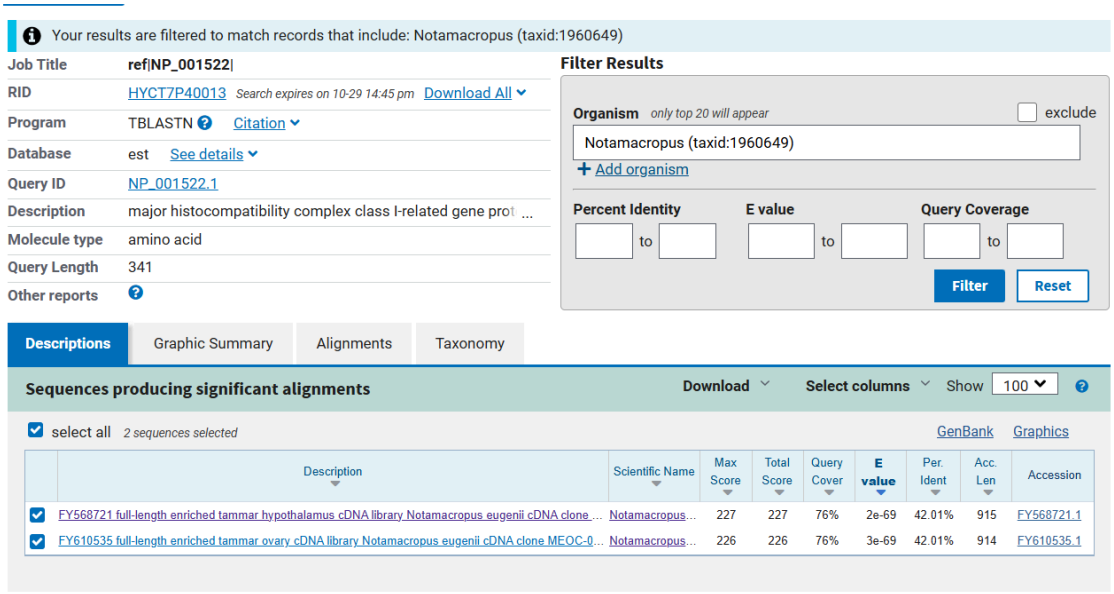
Find a gene project

Xiaoyan Wang(A16454055)

[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: MHCclass I-related gene protein Accession: NP_001522 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 search against Notamacropus ESTs Database: est Species: Notamacropus (taxid:1960649)–>(Notamacropus eugenii (taxid:9315))





Chosen match: Accession FY568721.1, a 915bp long mRNA from Notamacropus tammar hypothalamus cDNA clone

FY568721 full-length enriched tammar hypothalamus cDNA library Notamacropus eugenii cDNA clone MEHC-012G08 5', mRNA sequence

Sequence ID: FY568721.1 **Length:** 915 **Number of Matches:** 1

Range 1: 104 to 910 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
227 bits(578)	2e-69	Compositional matrix adjust.	113/269(42%)	166/269(61%)	7/269(2%)	+2

```
Query  24   THSLRYFRLGVSDPIHGVPEFISVGYVDSHPITTYDSVT--RQKEPRAPWM--AENLA
            +HS R+F   V+ P  G P F++VGYVD      +DS    ++ EPRA WM   E +
Sbjct  104  SHSRRFFDTAVTRPGLGEPRFLAVGYVDDQQFVRFDSDRPGQRLEPRAAWMEGVEQVE

Query  80   HWERYTQLLRGWQQMFKVELKRLQRHYNHS--GSHTYQRMIGCELLEDGS-TTGFLQ\
            +WER TQ++R   Q  ++V L+ L+ ++N S  G HTYQRM GCE+  D +   GF Q\
Sbjct  284  YWERSTQIMRAATQNYRVSLENLRGYFNQSAGGVHTYQRMYGCEVSPDLTFQRGFYQ\

Query  137  DGQDFLIFNKDTLSWLAVDNVAHTIKQAWEANQHELLYQKNWLEEECIAWLKRFLEYC
            DGQD++  + +TL+W A    A   K+ WEA +      K +LEE C+  +K++LE C
Sbjct  464  DGQDYIALDTETLTWTAAVPQAVNSKRTWEAERSISERHKAYLEETCVQRVKKYLEMC

Query  197  TLQRTEPPLVRVNRKETFPGVTALFCKAHGFYPPEIYMTWMKNGEEIVQEIDYGDILF
            TL RT+PP  RV R     G   L C+A  FYP EI +TW+++GEE  Q++++ +  F
Sbjct  644  TLMRTDPPKARVTRHTAPHGEVTLRCRAQDFYPKEISLTWLRDGEEQPQDMEFIETRF

Query  257  DGTYQAWASIELDPQSSNLYSCHVEHCGV  285
            DGT+Q WA++++       Y+C V+H G+
Sbjct  824  DGTFQKWAAVQMTSGQEGRYTCRVQHEGL  910
```

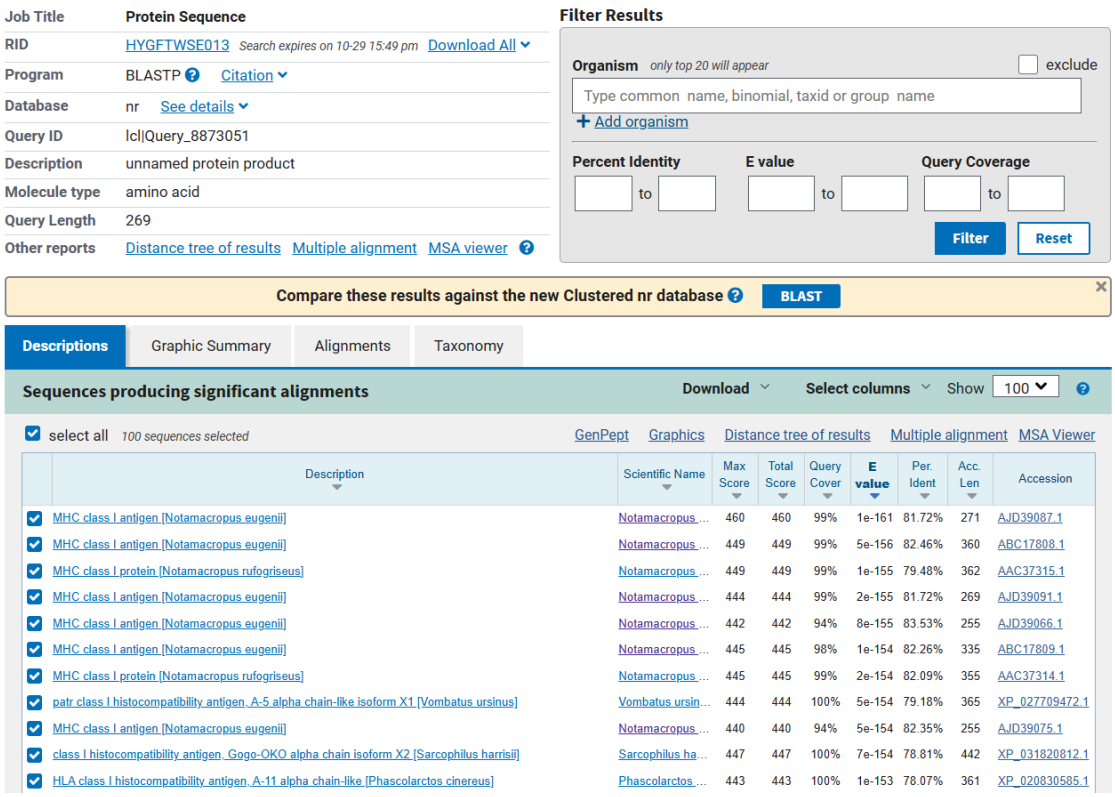
```
>FY568721.1:104-910 FY568721 full-length enriched tammar hypothalamus cDNA
library Notamacropus eugenii cDNA clone MEHC-012G08 5', mRNA sequence
TCCCACTCCAGGAGGTTTTTCGACACCGCCGTGACCCGGCCCGGGCTCGGGGAGCCGCGGTTCCTCGCCG
TGGGCTACGTGGACGATCAGCAGTTCGTGCGCTTCGACTCCGACCGCCCGGGTCAGAGGCTGGAGCCGCG
GGCGGCGTGGATGGAGGGCGTGGAGCAGGTGGAGCCGGGGTACTGGGAGCGGAGCACGCAGATCATGAGG
GCGGCCACACAGAATTACCGAGTGAGCCTGGAGAACCTGCGCGGCTACTTCAACCAGAGCGCGGGGGGCG
TCCACACCTACCAGCGCATGTACGGTTGCGAGGTGTCCCCCGACCTCACCTTCCAGCGCGGGTTTTACCA
ATACGCCTACGACGGGCAGGACTACATCGCCCTGGACACCGAGACCCTCACCTGGACGGCGGCCGTGCCT
CAGGCTGTGAACTCCAAGCGCACGTGGGAGGCGGAGAGGAGCATCTCGGAGAGACATAAAGCCTACCTGG
AGGAGACGTGCGTGCAGCGGGTGAAGAAGTACCTGGAGATGGGGAAGGAGACCCTGATGAGGACAGACCC
ACCTAAAGCCAGAGTGACCCGCCACACTGCCCCCCATGGGGAGGTGACCCTGAGGTGCCGGGCCCAGGAC
TTTTACCCTAAGGAGATCTCCCTGACTTGGCTGAGGGATGGGGAGGAACAGCCCCAGGACATGGAGTTCA
TTGAGACCAGGCCTGCAGGAGATGGCACCTTCCAGAAGTGGGCAGCTGTGCAGATGACCTCGGGCCAGGA
AGGCAGATACACCTGCCGCGTTCAGCACGAGGGGCTG
```

[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 willdiscover a new gene in a genome that is currently being sequenced, such as bacteria or plants or protozoa.

```
>FY568721.1:104-910 FY568721 full-length enriched tammar hypothalamus cDNA
library Notamacropus eugenii cDNA clone MEHC-012G08 5', mRNA sequence
SHSRRFFDTAVTRPGLGEPRFLAVGYVDDQQFVRFDSDRPGQRLEPRAAWMEGVEQVEPGYWERSTQIMRAATQNYR
VSLENLRGYFNQSAGGVHTYQRMYGCEVSPDLTFQRGFYQYAYDGQDYIALDTETLTWTAAVPQAVNSKRTWEAERS
ISERHKAYLEETCVQRVKKYLEMGKETLMRTDPPKARVTRHTAPHGEVTLRCRAQDFYPKEISLTWLRDGEEQPQDM
EFIETRPAGDGTFQKWAAVQMTSGQEGRYTCRVQHEGL
```

[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.



Job Title	Protein Sequence
RID	HYGFTWSE013 *Search expires on 10-29 15:49 pm* Download All
Program	BLASTP Citation
Database	nr See details
Query ID	lcl|Query_8873051
Description	unnamed protein product
Molecule type	amino acid
Query Length	269
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

Filter Reset

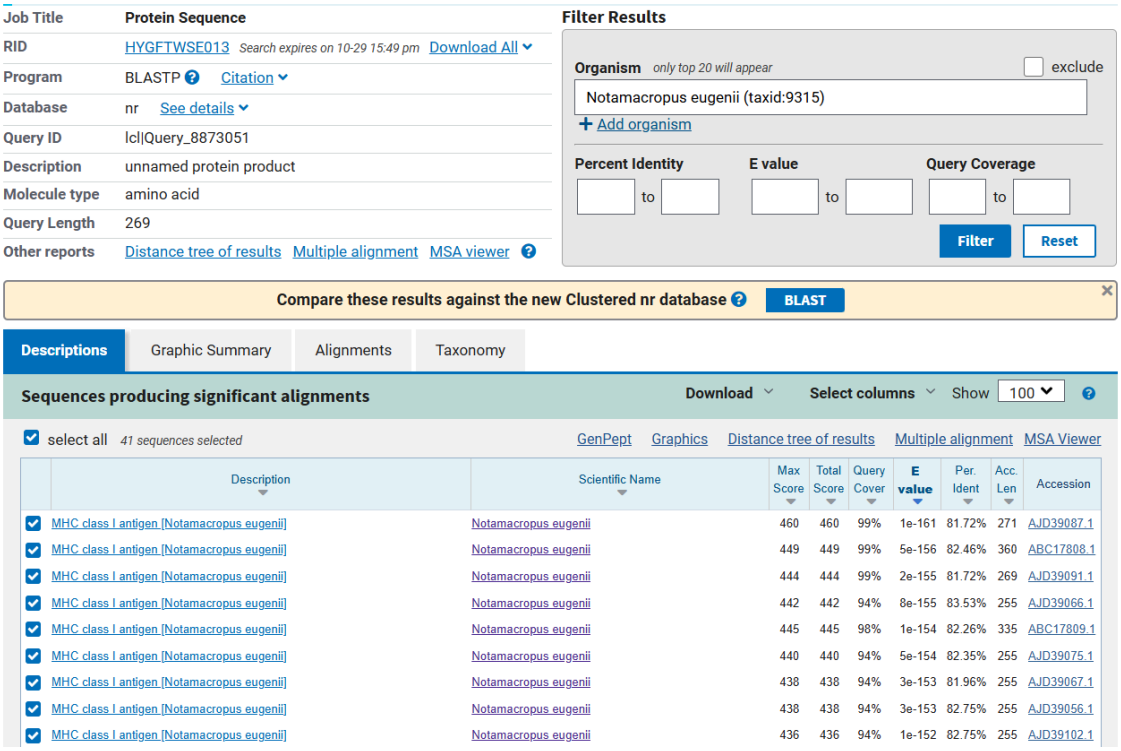
Compare these results against the new Clustered nr database BLAST

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
MHC class I antigen [Notamacropus eugenii]	Notamacropus ...	460	460	99%	1e-161	81.72%	271	AJD39087.1
MHC class I antigen [Notamacropus eugenii]	Notamacropus ...	449	449	99%	5e-156	82.46%	360	ABC17808.1
MHC class I protein [Notamacropus rufogriseus]	Notamacropus ...	449	449	99%	1e-155	79.48%	362	AAC37315.1
MHC class I antigen [Notamacropus eugenii]	Notamacropus ...	444	444	99%	2e-155	81.72%	269	AJD39091.1
MHC class I antigen [Notamacropus eugenii]	Notamacropus ...	442	442	94%	8e-155	83.53%	255	AJD39066.1
MHC class I antigen [Notamacropus eugenii]	Notamacropus ...	445	445	98%	1e-154	82.26%	335	ABC17809.1
MHC class I protein [Notamacropus rufogriseus]	Notamacropus ...	445	445	99%	2e-154	82.09%	355	AAC37314.1
patr class I histocompatibility antigen, A-5 alpha chain-like isoform X1 [Vombatus ursinus]	Vombatus ursin...	444	444	100%	5e-154	79.18%	365	XP_027709472.1
MHC class I antigen [Notamacropus eugenii]	Notamacropus ...	440	440	94%	5e-154	82.35%	255	AJD39075.1
class I histocompatibility antigen, Gogo-OKO alpha chain isoform X2 [Sarcophilus harrisii]	Sarcophilus ha...	447	447	100%	7e-154	78.81%	442	XP_031820812.1
HLA class I histocompatibility antigen, A-11 alpha chain-like [Phascolarctos cinereus]	Phascolarctos ...	443	443	100%	1e-153	78.07%	361	XP_020830585.1



Job Title	Protein Sequence
RID	HYGFTWSE013 *Search expires on 10-29 15:49 pm* Download All
Program	BLASTP Citation
Database	nr See details
Query ID	lcl|Query_8873051
Description	unnamed protein product
Molecule type	amino acid
Query Length	269
Other reports	Distance tree of results Multiple alignment MSA viewer

Filter Results

Organism *only top 20 will appear* exclude

Notamacropus eugenii (taxid:9315)

Add organism

Percent Identity to

E value to

Query Coverage to

Filter Reset

Compare these results against the new Clustered nr database BLAST

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Select columns Show 100

select all *41 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
MHC class I antigen [Notamacropus eugenii]	Notamacropus eugenii	460	460	99%	1e-161	81.72%	271	AJD39087.1
MHC class I antigen [Notamacropus eugenii]	Notamacropus eugenii	449	449	99%	5e-156	82.46%	360	ABC17808.1
MHC class I antigen [Notamacropus eugenii]	Notamacropus eugenii	444	444	99%	2e-155	81.72%	269	AJD39091.1
MHC class I antigen [Notamacropus eugenii]	Notamacropus eugenii	442	442	94%	8e-155	83.53%	255	AJD39066.1
MHC class I antigen [Notamacropus eugenii]	Notamacropus eugenii	445	445	98%	1e-154	82.26%	335	ABC17809.1
MHC class I antigen [Notamacropus eugenii]	Notamacropus eugenii	440	440	94%	5e-154	82.35%	255	AJD39075.1
MHC class I antigen [Notamacropus eugenii]	Notamacropus eugenii	438	438	94%	3e-153	81.96%	255	AJD39067.1
MHC class I antigen [Notamacropus eugenii]	Notamacropus eugenii	438	438	94%	3e-153	82.75%	255	AJD39056.1
MHC class I antigen [Notamacropus eugenii]	Notamacropus eugenii	436	436	94%	1e-152	82.75%	255	AJD39102.1

[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 t page width. Side-note: Indicate your sequence in the alignment by choosing an appropriate name for each sequence in the input unaligned sequence le (i.e. edit the sequence le 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.

novel protein:

```
> (sequence taken from BLAST result)
SHSRRFFDTAVTRPGLGEPRFLAVGYVDDQQFVRFDSDRPGQRLEPRAAWMEGVEQVEPGYWERSTQIMRAATQN
YRVSLENLRGYFNQSAGGVHTYQRMYGCEVSPDLTFQRGFYQYAYDGQDYIALDTETLTWTAAVPQAVNSKRTWE
AERSISERHKAYLEETCVQRVKKYLEMGKETLMRTDPPKARVTRHTAPHGEVTLRCRAQDFYPKEISLTWLRDGE
EQPQDM EFIETRPAGDGTFQKWAAVQMTSGQEGRYTCRVQHEGL
```

original query protein: Human MR1

```
>NP_001522.1 major histocompatibility complex class I-related gene protein
isoform 1 precursor [Homo sapiens]
MGELMAFLLPLIIVLMVKHSDSRTHSLRYFRLGVSDPIHGVPEFISVGYVDSHPITTYDSVTRQKEPRAP
WMAENLAPDHWERYTQLLRGWQQMFKVELKRLQRHYNHSGSHTYQRMIGCELLEDGSTTGFLQYAYDGQD
FLIFNKDTLSWLAVDNVAHTIKQAWEANQHELLYQKNWLEEECIAWLKRFLEYGKDTLQRTEPPLVRVNR
KETFPGVTALFCKAHGFYPPEIYMTWMKNGEEIVQEIDYGDILPSGDGTYQAWASIELDPQSSNLYSCHV
EHCGVHMVLQVPQESETIPLVMKAVSGSIVLVIVLAGVGVLVWRRRPREQNGAIYLPTPDR
```

a group of other members of this family from different species

>NP_001267784.1 class I histocompatibility antigen, Gogo-OKO alpha chain-like precursor [Sarcophilus harrisii]
MGSPARALFLLTALAALAETRAGSHSLRYFDTAVSRPGLGEPRFLSVGYVDDQQFVRFDSDSASQSEEPR
AAWMEKVKDVDPGYWEQETQIIKETAQISRVDLQTLRGYYNQSEGGAHTFQRMYGCEVSPELSFQRGFLQ
FAYDGQDYIALDTETLTWTAAQNEAVNTKRKWEAERSYAERDKAYLEETCVLWVQKYLEMGKESLQRADA
PSARVTRHSTPSGEVTLQCRAQDFYPSEISLAWLRDGEEQHQDTEFIETRPAGDGTFQKWAAVGVPSGQE
GRYTCRVQHEGLPEPLTLKWEPESSLPWIIVGVLAAVLLLTAVIAGAVVWRKKTSGGKGGDYVPAAGNDS
AQGSDVSLTAK

>AAC37315.1 MHC class I protein [Notamacropus rufogriseus]
MDRSLCALLLLGALALPDTWAGHHSLRYFHTAVTGPGLGEPRFVSVGYVDDQPFMSFDTDSPGQREEPRG
PWMDSMKHEEPEFWERQTWIHRATAQNYRVGLENLRGYFNQSAGGVHSFQRMSGCEVSPELTFQRGFDQH
AYDGRDYIALDMETLTWTAAVTPAMNTKRKWEADRSYTEGWKIYLEEECVQWLKKYLEMGKETLMRTDPP
SARVTRHTAPHGEVTLRCRAQDFYPKEISLTWLRNGEEQPQDTEFIETRPAGDGTFQKWAAVEVTSGQEG
RYTCRVQHEGLSEPLTLQWEPESSFTWYTVGGIAAALLILIAVIAGVGMWRRKHSGGKGSDYVPAADNES
AQWSDVSLTAKA

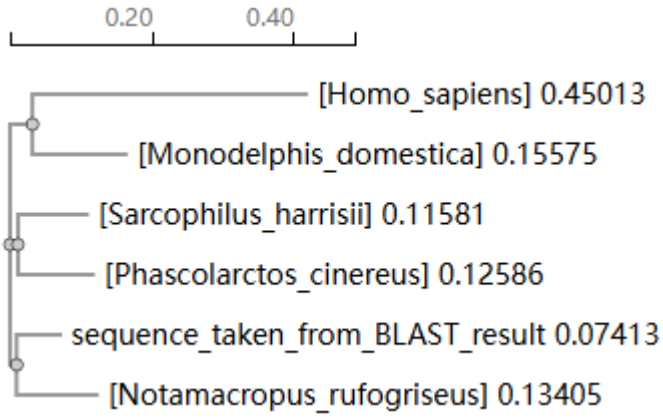
>XP_020830585.1 HLA class I histocompatibility antigen, A-11 alpha chain-like [Phascolarctos cinereus]
MEAYLRALFLLGTLALPETWAGSHSLRYFYTAVASPELAEPRFLIVGYVDDQQFVRFDSARASPRMEPRA
AWIERVQQEEPGYWDQETRNMKAVTQTYRVSLQNLRGYFNQSEGGVHTIQHMYGCEVSPELTFKRGFLQY
AYDGRDYIALDSETSTWTAEVPQALNTKRKWEAEKSYTEGQKAYLEETCVLWLKKYLEMGKETLKRTDPP
SARVTRHTGPHGEVTLRCRAQDFYPEDISLTWLRDGEEQLQDAEFIETRPAGEGTFQKWAGVDVTSGQEG
KYTCRVQHEGLPEPLTLKWEPESSSPWFIVGGIAVLLLLTAAIAGVVIWKKNTSGGKGGDYVPAAGNDSA
QGSDVSLTVKA

>XP_056673963.1 class I histocompatibility antigen, Gogo-OKO alpha chain isoform X1 [Monodelphis domestica]
MKPSLLSLFVLGVVALTETRAGSHSMRYFFTSMSRPELGDSQFISVGYVDDQQFVRFDSSSESQRMEPRA
AWMDKVDQEDPDYWEQNTQINRRNAQNDRVNLETLLGYYNQSRGGLHTIQRMYGCEIHPDGSFRKGFYQL
AYDGRDYIALDTETLTWTAADPGAENTKRKWEAERSIAERDKAYLEETCVQWVKKYLQMGKDVLLRTDPP
SARVSRHSGPDGEVSLRCRAQGFYPAEISLTWLRDGEEQLQDTEFIETRPGGDGTFQKWAAVAMAPGQED
RYSCRVQHEALAQPLSLRWEPEAPSLWVIVGVTAGVLVLVTAVVAGAVILRRRNSGGKGGAYVPAADKDS
AQGSDVSLTATA

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


[Homo_sapiens]                      MGELMAFLLPLIIVLMVKHSDSRTHSLRYFRLGVSDPIHGVPEFISVGYVDSHPITTYDS
[Monodelphis_domestica]             MKPSLLSLFVLGVV-ALTETRAGSHSMRYFFTSMSRPELGDSQFISVGYVDDQQFVRFDS
[Sarcophilus_harrisii]              MGSPARALFLLTALAALAETRAGSHSLRYFDTAVSRPGLGEPRFLSVGYVDDQQFVRFDS
[Phascolarctos_cinereus]            MEAYLRALFLLGTL-ALPETWAGSHSLRYFYTAVASPELAEPRFLIVGYVDDQQFVRFDS
sequence_taken_from_BLAST_result    ----------------------SHSRRFFDTAVTRPGLGEPRFLAVGYVDDQQFVRFDS
[Notamacropus_rufogriseus]          MDRSLCALLLLGAL-ALPDTWAGHHSLRYFHTAVTGPGLGEPRFVSVGYVDDQPFMSFDT
                                                          ** *:*  .:: *  . . *: *****.: :  :*:

[Homo_sapiens]                      --VTRQKEPRAPWM--AENLAPDHWERYTQLLRGWQQMFKVELKRLQRHYNHS--GSHTY
[Monodelphis_domestica]             SSESQRMEPRAAWMDKVDQEDPDYWEQNTQINRRNAQNDRVNLETLLGYYNQSRGGLHTI
[Sarcophilus_harrisii]              DSASQSEEPRAAWMEKVKDVDPGYWEQETQIIKETAQISRVDLQTLRGYYNQSEGGAHTF
[Phascolarctos_cinereus]            ARASPRMEPRAAWIERVQQEEPGYWDQETRNMKAVTQTYRVSLQNLRGYFNQSEGGVHTI
sequence_taken_from_BLAST_result    DRPGQRLEPRAAWMEGVEQVEPGYWERSTQIMRAATQNYRVSLENLRGYFNQSAGGVHTY
[Notamacropus_rufogriseus]          DSPGQREEPRGPWMDSMKHEEPEFWERQTWIHRATAQNYRVGLENLRGYFNQSAGGVHSF
                                           ***..*:   .   * .*:. *.  .   *  .* *: *  ::*:*  * *:

[Homo_sapiens]                      QRMIGCELLEDGS-TTGFLQYAYDGQDFLIFNKDTLSWLAVDNVAHTIKQAWEANQHELL
[Monodelphis_domestica]             QRMYGCEIHPDGSFRKGFYQLAYDGRDYIALDTETLTWTAADPGAENTKRKWEAERSIAE
[Sarcophilus_harrisii]              QRMYGCEVSPELSFQRGFLQFAYDGQDYIALDTETLTWTAAQNEAVNTKRKWEAERSYAE
[Phascolarctos_cinereus]            QHMYGCEVSPELTFKRGFLQYAYDGRDYIALDSETSTWTAEVPQALNTKRKWEAEKSYTE
sequence_taken_from_BLAST_result    QRMYGCEVSPDLTFQRGFYQYAYDGQDYIALDTETLTWTAAVPQAVNSKRTWEAERSISE
[Notamacropus_rufogriseus]          QRMSGCEVSPELTFQRGFDQHAYDGRDYIALDMETLTWTAAVTPAMNTKRKWEADRSYTE
                                    *.* ***:  : :   ** * ****.*:: :: :* :* *    * . *. ***:.

[Homo_sapiens]                      YQKNWLEEECIAWLKRFLEYGKDTLQRTEPPLVRVNRKETFPGVTALFCKAHGFYPPEIY
[Monodelphis_domestica]             RDKAYLEETCVQWVKKYLQMGKDVLLRTDPPSARVSRHSGPDGEVSLRCRAQGFYPAEIS
[Sarcophilus_harrisii]              RDKAYLEETCVLWVQKYLEMGKESLQRADAPSARVTRHSTPSGEVTLQCRAQDFYPSEIS
[Phascolarctos_cinereus]            GQKAYLEETCVLWLKKYLEMGKETLKRTDPPSARVTRHTGPHGEVTLRCRAQDFYPEDIS
sequence_taken_from_BLAST_result    RHKAYLEETCVQRVKKYLEMGKETLMRTDPPKARVTRHTAPHGEVTLRCRAQDFYPKEIS
[Notamacropus_rufogriseus]          GWKIYLEEECVQWLKKYLEMGKETLMRTDPPSARVTRHTAPHGEVTLRCRAQDFYPKEIS
                                      * :*** *: .::.:*: **: * *::.* .**.*:    * .:* *.*:.*** :*
```

```
[Homo_sapiens]                    MTWMKNGEEIVQEIDYGDILPSGDGTYQAWASIELDPQSSNLYSCHVEHCGVHMVLQVPQ
[Monodelphis_domestica]           LTWLRDGEEQLQDTEFIETRPGGDGTFQKWAAVAMAPGQEDRYSCRVQHEALAQPLSLRW
[Sarcophilus_harrisii]            LAWLRDGEEQHQDTEFIETRPAGDGTFQKWAAVGVPSGQEGRYTCRVQHEGLPEPLTLKW
[Phascolarctos_cinereus]          LTWLRDGEEQLQDAEFIETRPAGEGTFQKWAGVDVTSGQEGKYTCRVQHEGLPEPLTLKW
sequence_taken_from_BLAST_result  LTWLRDGEEQPQDMEFIETRPAGDGTFQKWAAVQMTSGQEGRYTCRVQHEGL--------
[Notamacropus_rufogriseus]        LTWLRNGEEQPQDTEFIETRPAGDGTFQKWAAVEVTSGQEGRYTCRVQHEGLSEPLTLQW
                                  ::*:.:***  *: :: :  *.*:**:* **.: : . ... *:*.*:* .:

[Homo_sapiens]                    ESET-IPLVMKAVSGSIVLVIVLAGVGVLVWRRRPREQNGAIYLPTPDR-----------
[Monodelphis_domestica]           EPEAPSLWVIVGVTAGVLVLVTAVVAGAVILRRRNSGGKGGAYVPAADKDSAQGSDVSLT
[Sarcophilus_harrisii]            EPESSLPWIIVGVL-AAVLLLTAVIAGAVVWRKKTSGGKGGDYVPAAGNDSAQGSDVSLT
[Phascolarctos_cinereus]          EPESSSPWFIVGGI-AVLLLLTAAIAGVVIWKKNTSGGKGGDYVPAAGNDSAQGSDVSLT
sequence_taken_from_BLAST_result  ------------------------------------------------------------
[Notamacropus_rufogriseus]        EPESSFTWYTVGGIAAALLILIAVIAGVGMWRRKHSGGKGSDYVPAADNESAQWSDVSLT


[Homo_sapiens]                    ---
[Monodelphis_domestica]           ATA
[Sarcophilus_harrisii]            AK-
[Phascolarctos_cinereus]          VKA
sequence_taken_from_BLAST_result  ---
[Notamacropus_rufogriseus]        AKA
```

[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 gure margins.

```
library(bio3d)
```

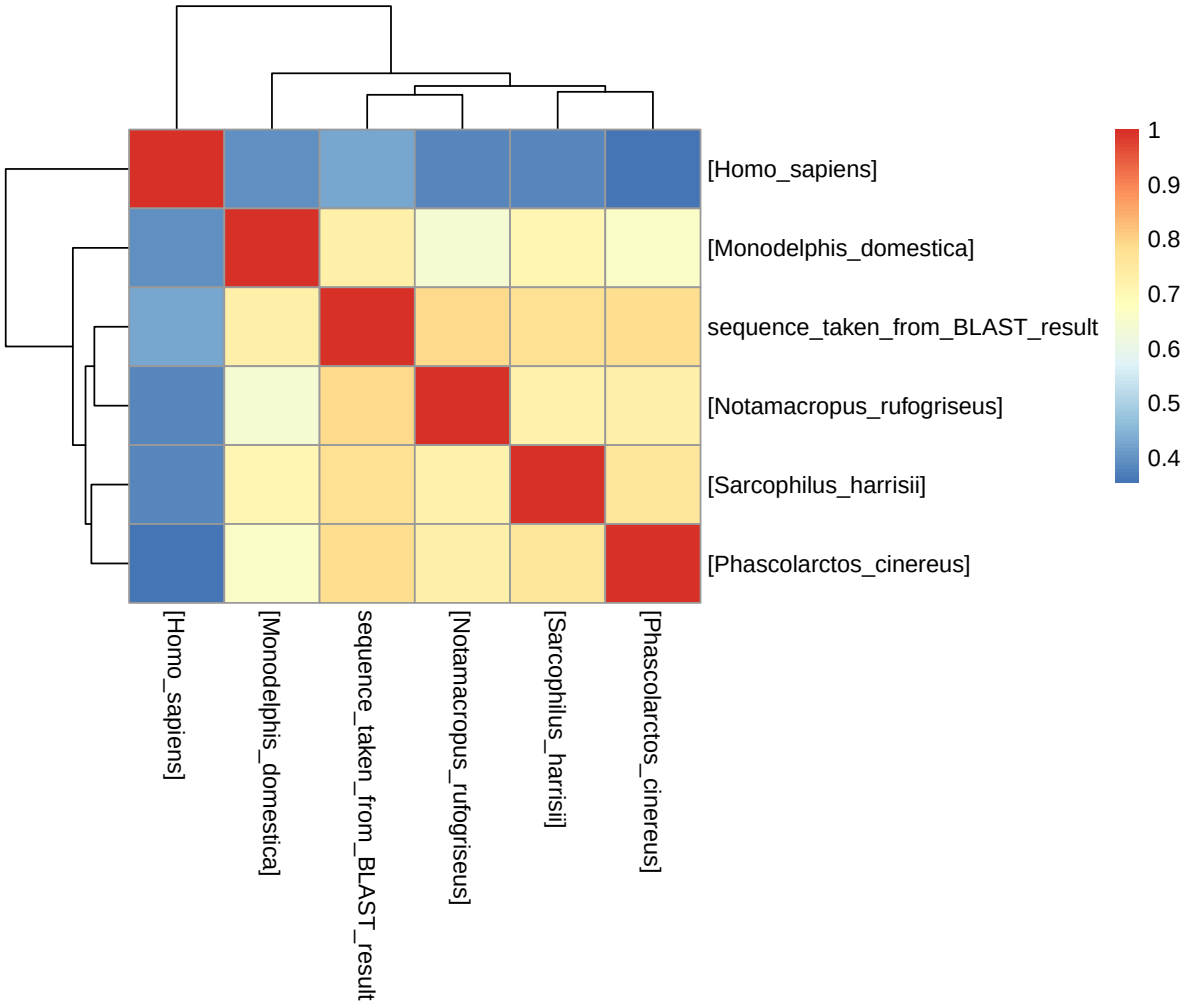
```
Warning: package 'bio3d' was built under R version 4.3.3
```

```
#install.packages("pheatmap")
library(pheatmap)
```

```
Warning: package 'pheatmap' was built under R version 4.3.3
```

```
aligned <- read.fasta("muscle-I20241203-193226-0346-30336184-p1m.fa")
ident <- seqidentity(aligned)
```

```
pheatmap(ident)
```



[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.

```
# too many "-" appeared. The Notamacropus rufogriseus sequence, which is from the same family
exp <- read.fasta("example.fasta")
top <- blast.pdb(exp)
```

```
 Searching ... please wait (updates every 5 seconds) RID = N0D57VFB013
 ..................................................................................
 Reporting 3092 hits
```

```
top3 <- top$hit.tbl[1:3,]
ann <- pdb.annotate(top3$pdb.id)
combined <- merge(top3, ann, by.x = "pdb.id", by.y = "row.names")
correct_names <- c("pdb.id", "experimentalTechnique", "resolution", "source", "evalue", "iden
matching_cols <- intersect(correct_names, colnames(combined))  # Ensure only existing columns
combined <- combined[, match(matching_cols, colnames(combined))]
combined
```

```
  pdb.id experimentalTechnique resolution                source    evalue
1 7EDO_A                X-ray       2.70 Trichosurus vulpecula  0.00e+00
2 7RTD_A                X-ray       2.05          Homo sapiens 5.22e-137
3 8RBU_A                X-ray       2.70          Homo sapiens 1.02e-131
  identity
1   71.348
2   57.983
3   57.514
```

[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?

The three targets identified from the previous step is: 7EDO, 7RTD, and 8RBU. As the 7EDO from another species have a very low E-value, I decided to look in this PDB structure.

The identity is 71% similar to the "novel" protein from Notamacropus eugenii. This is likely a type of classical MHC- class I molecule. Since MHC class I is a highly polymorphic gene, I would think the 71% similarity between Notamacropus eugenii and the 7EDO protein from Trichosurus vulpecula supports that the "novel" protein is also a type of MHC Class I molecule. Henceforce, considering the nature of MHC Class I, I think these two proteins are similar.

[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 reported that may be useful starting points for exploring potential inhibition of your novel protein?

For this sequence, there is a HLA antigen target that is similar to the novel sequence we are looking. HLA is the MHC Class I molecule in human.

https://www.ebi.ac.uk/chembl/explore/target/CHEMBL2632#NameAndClassification

There are 3 related binding assays listed here.

https://www.ebi.ac.uk/chembl/explore/assays/STATE_ID:yG22iojPssyMVh9TAuiLXQ%3D%3D

Two of them are studying the binding affinity through inhibiting a ratiolabled standard peptide(FLPSDYFPSV). The novel protein, which is also a type of MHC Class I or MHC Class I with antigen presented, could be also inhibited through similar way targeting the antigen-binding site.