Erica Sanchez
ersanche@ucsd.edu
PID: A15787505

[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: retinol binding protein 4
Accession:  NM_001323518.2
Species: Human
Function: enables protein binding, retinol binding, and retinol transmembrane
transporter activity
```

[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).

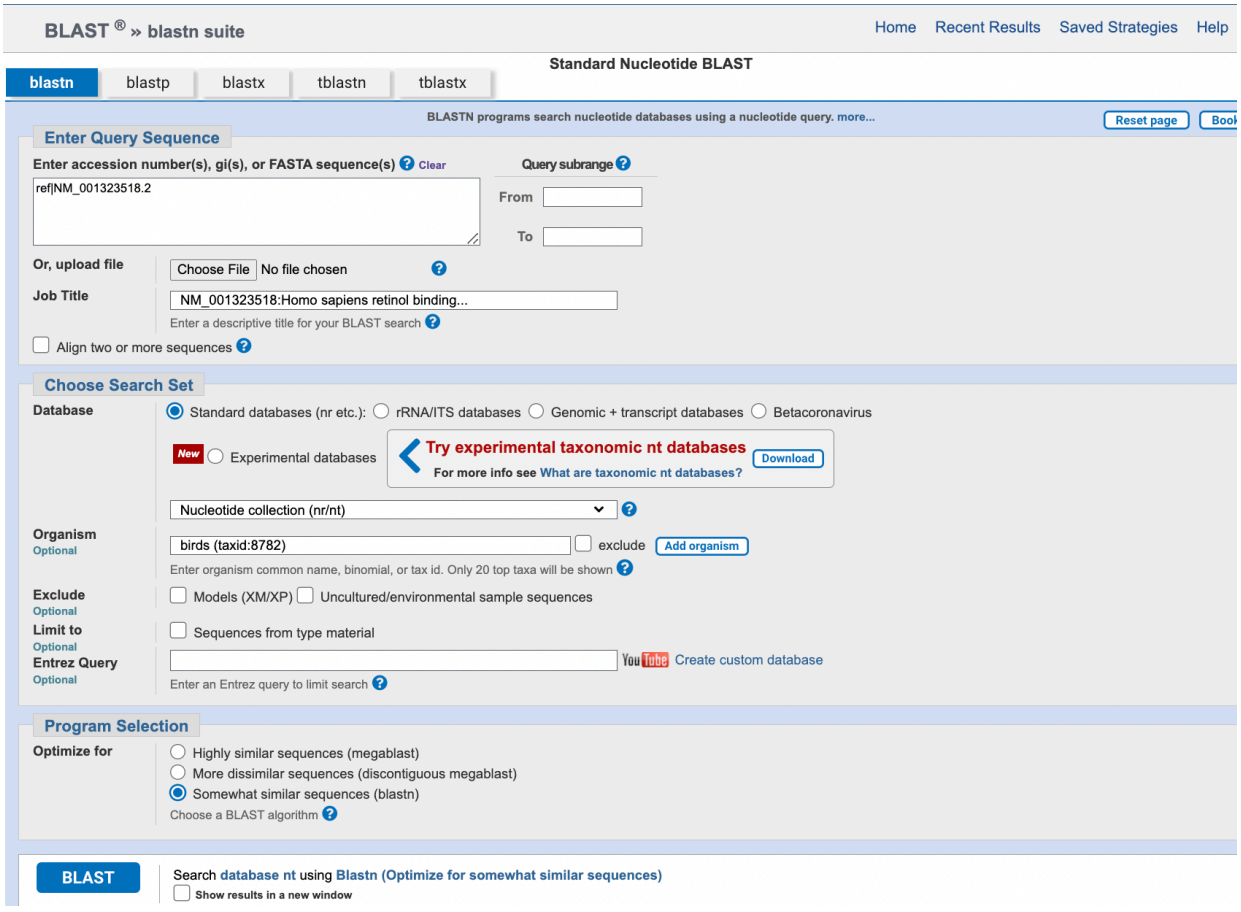
Also include the output of that BLAST search in your document. If appropriate, change the font to Courier size 10 so that the results are displayed neatly. You can also screen capture a BLAST output (e.g. alt print screen on a PC or on a MAC press ⌘-shift-4. The pointer becomes a bulls eye. Select the 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.

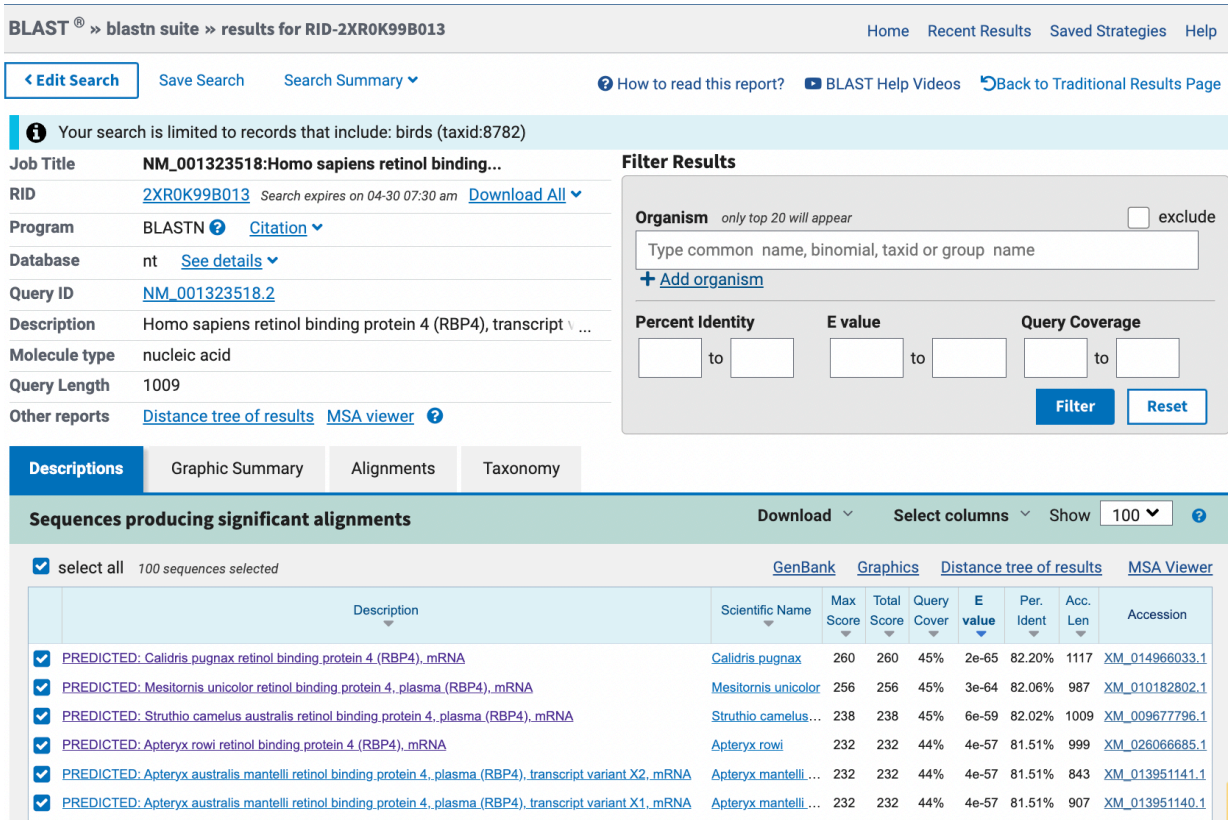
On the BLAST results, clearly indicate a match that represents a protein sequence, encoded from some DNA sequence, that is homologous to your query protein. I need to be able to inspect the pairwise alignment you have selected, including the E value and score. It should be labeled a "genomic clone" or "mRNA sequence", etc. - but include no functional annotation.

In general, [Q2] is the most difficult for students because it requires you to have a “feel” for how to interpret BLAST results. You need to distinguish between a perfect match to your query (i.e. a sequence that is not “novel”), a near match (something that might be “novel”, depending on the results of [Q4]), and a non-homologous result.
If you are having trouble finding a novel gene try restricting your search to an organism that is poorly annotated.

```
Blast method: Nucleotide Blast
Database searched: National Library of Medicine
Limits applied: birds (taxid:8782)
Chosen gene: XM_026094147, 932 bp, Dromaius novaehollandiae retinol binding
protein 4 (RBP4), mRNA
```

Search output lists (top hits):





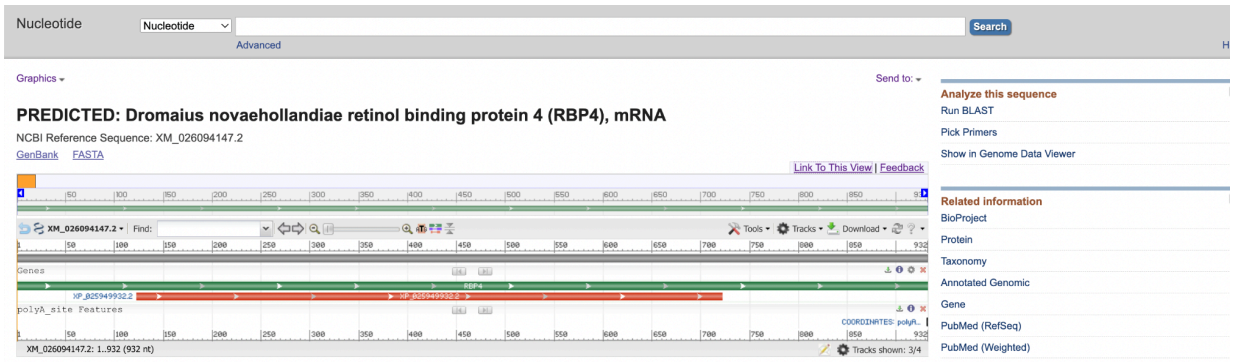
PREDICTED: Dromaius novaehollandiae retinol binding protein 4 (RBP4), mRNA

Sequence ID: XM_026094147.2 **Length:** 932 **Number of Matches:** 1

Range 1: 201 to 652 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
214 bits(108)	9e-52	366/452(81%)	0/452(0%)	Plus/Plus

```
Query  83   GACTGCCGAGTGAGCAGCTTCCGAGTCAAGGAGAACTTCGACAAGGCTCGCTTCTCTGGG  142
            ||||||||||||||||||||  ||||||||||||||||||||||||     | |    |||
Sbjct  201  GACTGCCGAGTGAGCAGCTTCAAAGTCAAGGAGAACTTCGACAAGAACAGGTATAGTGGC  260

Query  143  ACCTGGTACGCCATGGCCAAGAAGGACCCCGAGGGCCTCTTTCTGCAGGACAACATCGTC  202
            ||||||||  ||||||||  |||||  ||  ||  |||||  ||  ||||||||||||| | ||
Sbjct  261  ACCTGGTATGCCATGGCAAAGAAAGATCCTGAGGGGCTGTTTCTGCAGGACAACGTGGTA  320

Query  203  GCGGAGTTCTCCGTGGACGAGACCGGCCAGATGAGCGCCACAGCCAAGGGCCGAGTCCGT  262
            ||  ||||  | ||  ||  ||||  ||  |||||||  |||||  ||  ||  |||  ||||| |
Sbjct  321  GCCCAGTTTACAGTAGATGAGAATGGAGAGATGAGTGCCACGGCAAAAGGCAGAGTCAGA  380

Query  263  CTTTTGAATAACTGGGACGTGTGCGCAGACATGGTGGGCACCTTCACAGACACCGAGGAC  322
            ||  ||  ||||||||||  ||  ||  |||||||||  |  |||  |  ||||  |||||  ||||||
Sbjct  381  CTCTTTAATAACTGGGATGTCTGTGCAGACATGATTGGCTCTTTCAAGGACACAGAGGAT  440

Query  323  CCTGCCAAGTTCAAGATGAAGTACTGGGGCGTAGCCTCCTTTCTCCAGAAAGGAAATGAT  382
            ||||||||  ||||||||||||||||||||||||||  ||  ||  |||||  ||||||||||||||||
Sbjct  441  CCTGCCAAATTCAAGATGAAGTACTGGGGCGTTGCTTCTTTTCTGCAGAAAGGAAATGAT  500

Query  383  GACCACTGGATCGTCGACACAGACTACGACACGTATGCCGTGCAGTACTCCTGCCGCCTC  442
            ||  ||||||  |  ||  ||||||||  |||||  ||  |||||  |  ||  ||||||||||||||
Sbjct  501  GATCACTGGGTAGTGGACACAGATTACGATACTTATGCTCTTCATTACTCCTGCCGCCAA  560

Query  443  CTGAACCTCGATGGCACCTGTGCTGACAGCTACTCCTTCGTGTTTTCCCGGGACCCCAAC  502
            ||  ||||  ||||||||||||||||||  |||||  |||||  ||||||||||||||||
Sbjct  561  CTAAACGAAGATGGCACCTGTGCTGATAGCTATTCCTTTGTGTTCTCCCGGGACCCCAAA  620

Query  503  GGCCTGCCCCCAGAAGCGCAGAAGATTGTAAG  534
            ||  ||||  ||||||  ||  ||||||  ||||||||
Sbjct  621  GGATTGCCTCCAGAGGCACAGAAAATTGTAAG  652
```



E-value and other alignment stats: 44% Coverage, 9e-52 E-value, 80.97% Identity

Alignment Details:

>XM_026094147.2 PREDICTED: Dromaius novaehollandiae retinol binding protein 4 (RBP4), mRNA

Length = 932 bp

Score = 214 bits(108), Expect = 9e-52

Identities = 366/452(81%), Gaps = 0/452(0%)

```
Query  83   GACTGCCGAGTGAGCAGCTTCCGAGTCAAGGAGAACTTCGACAAGGCTCGCTTCTCTGGG  142
           ||||||||||||||||||||  ||||||||||||||||||||||||     | |    |||
Sbjct  201  GACTGCCGAGTGAGCAGCTTCAAAGTCAAGGAGAACTTCGACAAGAACAGGTATAGTGGC  260

Query  143  ACCTGGTACGCCATGGCCAAGAAGGACCCCGAGGGCCTCTTTCTGCAGGACAACATCGTC  202
           ||||||||  ||||||||  |||||  ||  ||  |||||  ||  ||||||||||||| | ||
Sbjct  261  ACCTGGTATGCCATGGCAAAGAAAGATCCTGAGGGGCTGTTTCTGCAGGACAACGTGGTA  320
```

```
Query  203  GCGGAGTTCTCCGTGGACGAGACCGGCCAGATGAGCGCCACAGCCAAGGGCCGAGTCCGT  262
            ||  ||||  | || || ||||  ||  ||||||| ||||| || || ||| ||||| |
Sbjct  321  GCCCAGTTTACAGTAGATGAGAATGGAGAGATGAGTGCCACGGCAAAAGGCAGAGTCAGA  380

Query  263  CTTTTGAATAACTGGGACGTGTGCGCAGACATGGTGGGCACCTTCACAGACACCGAGGAC  322
            || || |||||||||||  || || |||||||||  | ||| | ||||  ||||| |||||
Sbjct  381  CTCTTTAATAACTGGGATGTCTGTGCAGACATGATTGGCTCTTTCAAGGACACAGAGGAT  440

Query  323  CCTGCCAAGTTCAAGATGAAGTACTGGGGCGTAGCCTCCTTTCTCCAGAAAGGAAATGAT  382
            ||||||||  ||||||||||||||||||||||  || || |||||  ||||||||||||||
Sbjct  441  CCTGCCAAATTCAAGATGAAGTACTGGGGCGTTGCTTCTTTTCTGCAGAAAGGAAATGAT  500

Query  383  GACCACTGGATCGTCGACACAGACTACGACACGTATGCCGTGCAGTACTCCTGCCGCCTC  442
            || ||||||  | || ||||||||  |||||  || |||||  | || ||||||||||||
Sbjct  501  GATCACTGGGTAGTGGACACAGATTACGATACTTATGCTCTTCATTACTCCTGCCGCCAA  560

Query  443  CTGAACCTCGATGGCACCTGTGCTGACAGCTACTCCTTCGTGTTTTCCCGGGACCCCAAC  502
            || |||   ||||||||||||||||  |||||  |||||  |||||  ||||||||||||||
Sbjct  561  CTAAACGAAGATGGCACCTGTGCTGATAGCTATTCCTTTGTGTTCTCCCGGGACCCCAAA  620

Query  503  GGCCTGCCCCCAGAAGCGCAGAAGATTGTAAG  534
            ||  |||| |||||  || |||||  ||||||||
Sbjct  621  GGATTGCCTCCAGAGGCACAGAAAATTGTAAG  652
```

[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.

Protein sequence of choice matches Subject above:
EMBOSS TRANSEQ FASTA:
>XM_026094147.2_1 PREDICTED: Dromaius novaehollandiae retinol binding protein 4 (RBP4), mRNA

YPKGLRKPQGYIKHCLAGWPSCCHQHLGS*EAAGTGTVWTRWPTRGEPRPGCCYWCWPCW
VAAQQNGTAE*AASKSRRTSTRTGIVAPGMPWQRKILRGCFCRTTW*PSLQ*MRMER*VP
RQKAESDSLITGMSVQT*LALSRTQRILPNSR*STGALLLFCRKEMMITG*WTQITILML
FITPAAN*TKMAPVLIAIPLCSPGTPKDCLQRHRKL*DKGR*TSAWTENTELSFIMDSAL
KEIQNYGRKSCNSIMDVMLTH*LSVLLKSF*MALFRS*IYLSELCSSPCQ*TNEMF**TS
ITGDANCLYAX

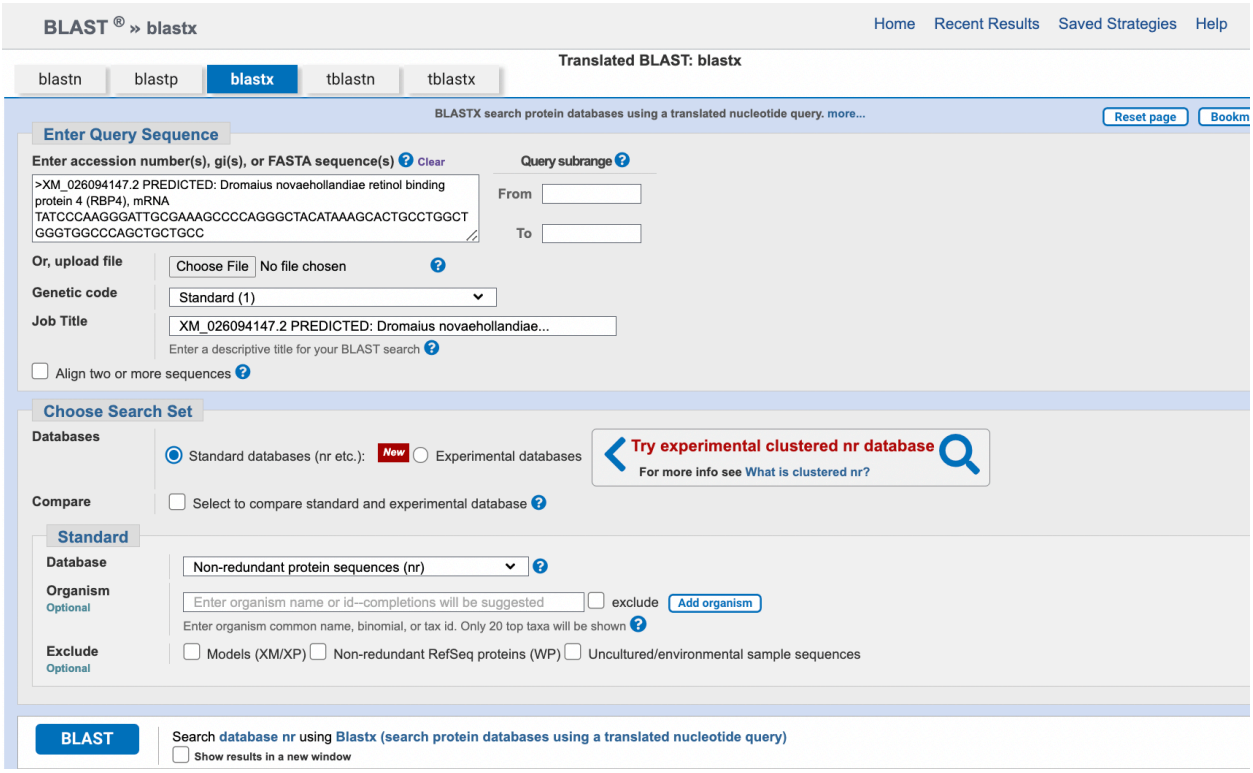
Name in header: Dromaius novaehollandiae (emu)
Species: Dromaius novaehollandiae
Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Sauropsida; Sauria; Archelosauria; Archosauria; Dinosauria; Saurischia; Theropoda; Coelurosauria; Aves; Palaeognathae; Casuariiformes; Dromaiidae; Dromaius

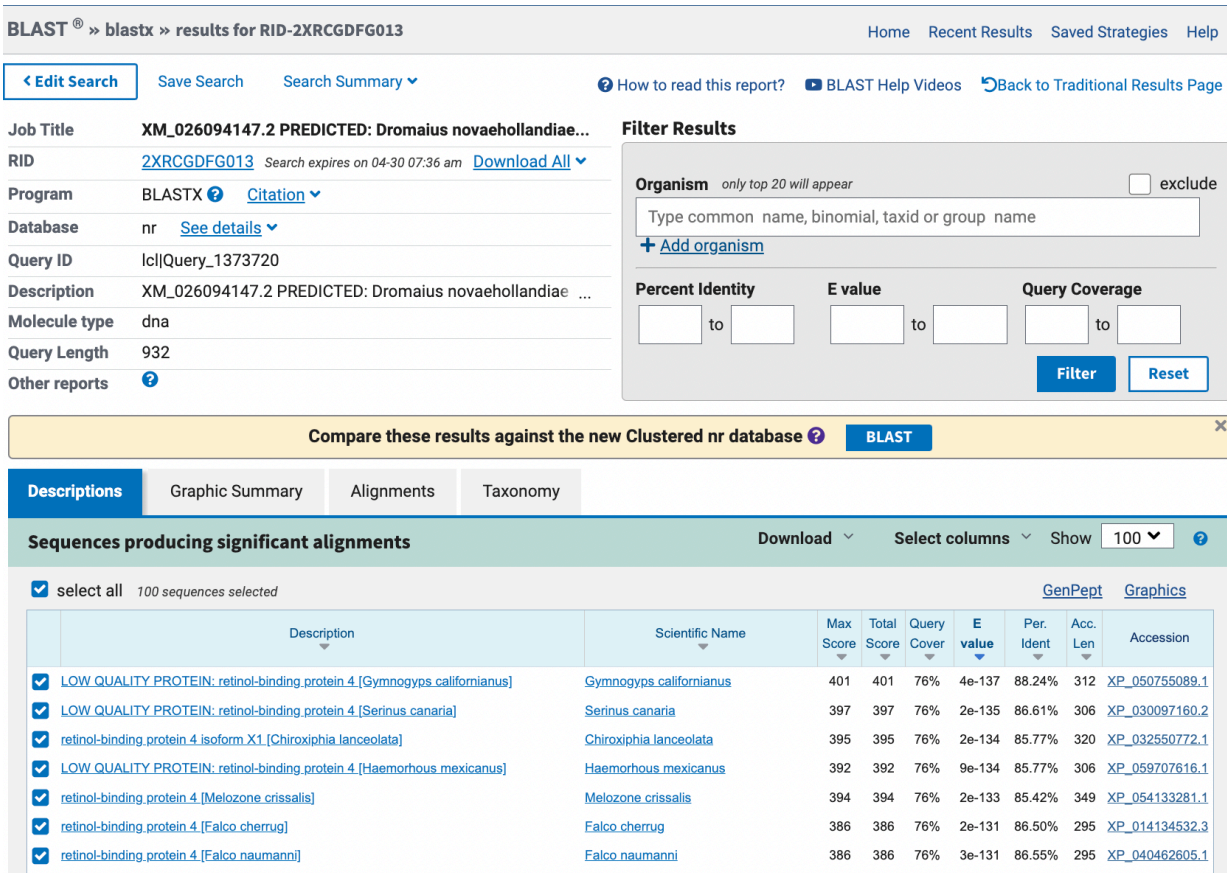
[Q4] Prove that this gene, and its corresponding protein, are novel. For the purposes of this project, “novel” is defined as follows. Take the protein sequence (your answer to [Q3]), and use it as a query in a blastp search of the nr database at NCBI.

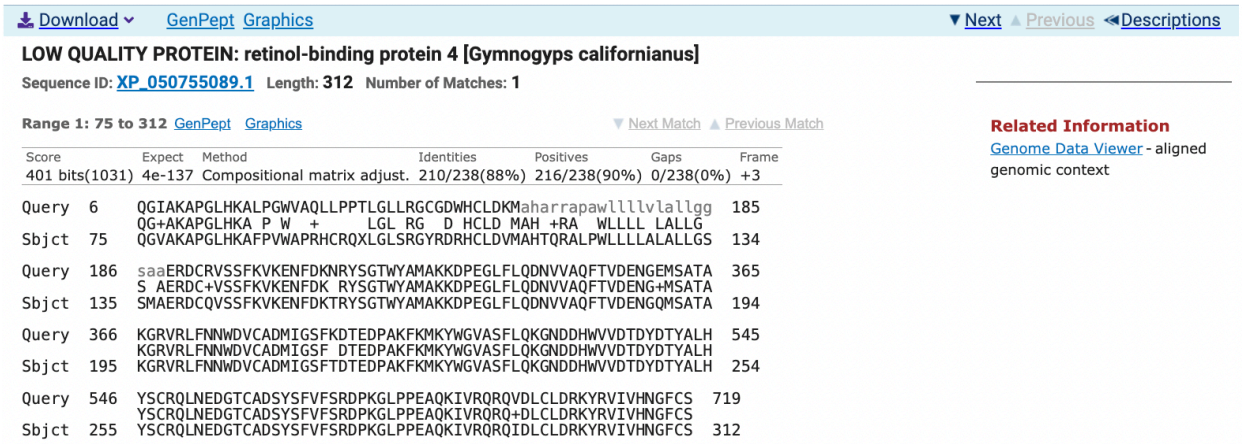
- If there is a match with 100% amino acid identity to a protein in the database, from the same species, then your protein is NOT novel (even if the match is to a protein with a name such as “unknown”). Someone has already found and annotated this sequence, and assigned it an accession number. There is not a match with 100% amino acid identity ✓
- If the top match reported has less than 100% identity, then it is likely that your protein is novel, and you have succeeded. The top match reported has less than 100% identity ✓
- 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.

Protein sequence of choice matches Subject above:
>XP_050755089.1 LOW QUALITY PROTEIN: retinol-binding protein 4 [Gymnogyps californianus]
MMFATVYQGKRGKEFGEEVEETSLHLVTGLSTLGQGQRPPEPMGVMQEAVMKGPQPAACISPPPYSPSNA
GSVPQGVAKAPGLHKAFPVWAPRHCRQXLGLSRGYRDRHCLDVMAHTQRALPWLLLLALALLGSSMAERD
CQVSSFKVKENFDKTRYSGTWYAMAKKDPEGLFLQDNVVAQFTVDENGQMSATAKGRVRLFNNWDVCADM
IGSFTDTEDPAKFKMKYWGVASFLQKGNDDHWVVDTDYDTYALHYSCRQLNEDGTCADSYSFVFSRDPKG
LPPEAQKIVRQRQIDLCLDRKYRVIVHNGFCS

Top alignment shown with alignment statistics: LOW QUALITY PROTEIN: retinol-binding protein 4 [Gymnogyps californianus; 76% coverage, 4e-137 E-value, 88.24% identity







Download GenPept Graphics ▼Next ▲Previous ◀◀Descriptions

LOW QUALITY PROTEIN: retinol-binding protein 4 [Gymnogyps californianus]

Sequence ID: XP_050755089.1 Length: 312 Number of Matches: 1

Range 1: 75 to 312 GenPept Graphics ▼Next Match ▲Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
401 bits(1031)	4e-137	Compositional matrix adjust.	210/238(88%)	216/238(90%)	0/238(0%)	+3

```
Query  6    QGIAKAPGLHKALPGWVAQLLPPTLGLLRGCGDWHCLDKMaharrapawllllvlallgg  185
            QG+AKAPGLHKA P W  +     LGL RG  D HCLD MAH +RA  WLLLL LALLG
Sbjct  75   QGVAKAPGLHKAFPVWAPRHCRQXLGLSRGYRDRHCLDVMAHTQRALPWLLLLALALLGS  134

Query  186  saaERDCRVSSFKVKENFDKNRYSGTWYAMAKKDPEGLFLQDNVVAQFTVDENGEMSATA  365
            S AERDC+VSSFKVKENFDK RYSGTWYAMAKKDPEGLFLQDNVVAQFTVDENG+MSATA
Sbjct  135  SMAERDCQVSSFKVKENFDKTRYSGTWYAMAKKDPEGLFLQDNVVAQFTVDENGQMSATA  194

Query  366  KGRVRLFNNWDVCADMIGSFKDTEDPAKFKMKYWGVASFLQKGNDDHWVVDTDYDTYALH  545
            KGRVRLFNNWDVCADMIGSF DTEDPAKFKMKYWGVASFLQKGNDDHWVVDTDYDTYALH
Sbjct  195  KGRVRLFNNWDVCADMIGSFTDTEDPAKFKMKYWGVASFLQKGNDDHWVVDTDYDTYALH  254

Query  546  YSCRQLNEDGTCADSYSFVFSRDPKGLPPEAQKIVRQRQVDLCLDRKYRVIVHNGFCS  719
            YSCRQLNEDGTCADSYSFVFSRDPKGLPPEAQKIVRQRQ+DLCLDRKYRVIVHNGFCS
Sbjct  255  YSCRQLNEDGTCADSYSFVFSRDPKGLPPEAQKIVRQRQIDLCLDRKYRVIVHNGFCS  312
```

Related Information

Genome Data Viewer - aligned genomic context

[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.

[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).