

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 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: 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):

BLAST® » blastn suite

HomeRecent ResultsSaved StrategiesHelp

blastnblastptblastxtblastntblastx

Standard Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide query. more...

Reset pageBookmarks

Enter Query Sequence

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

ref|NM\_001323518.2

Query subrange ?

From

To

Or, upload file

Choose FileNo file chosen ?

Job Title

NM\_001323518:Homo sapiens retinol binding...

Enter a descriptive title for your BLAST search ?

☐ Align two or more sequences ?

Choose Search Set

Database

☒ Standard databases (nr etc.):☐ rRNA/ITS databases☐ Genomic + transcript databases☐ Betacoronavirus

New

☐ Experimental databases

Try experimental taxonomic nt databases

For more info see What are taxonomic nt databases?

Download

Nucleotide collection (nr/nt)

Organism

Optional

birds (taxid:8782)

☐ 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

Enter an Entrez query to limit search ?

YouTube

Create custom database

Program Selection

Optimize for

☐ Highly similar sequences (megablast)

☐ More dissimilar sequences (discontiguous megablast)

☒ Somewhat similar sequences (blastn)

Choose a BLAST algorithm ?

BLAST

Search database nt using Blastn (Optimize for somewhat similar sequences)

☐ Show results in a new window

BLAST® » blastn suite » results for RID-2XR0K99B013

HomeRecent ResultsSaved StrategiesHelp

< Edit Search

Save Search

Search Summary ▼

How to read this report?

BLAST Help Videos

Back to Traditional Results Page

📘

Your search is limited to records that include: birds (taxid:8782)

Job Title

NM\_001323518:Homo sapiens retinol binding...

RID

2XR0K99B013

Search expires on 04-30 07:30 am

Download All ▼

Program

BLASTN ?

Citation ▼

Database

nt

See details ▼

Query ID

NM\_001323518.2

Description

Homo sapiens retinol binding protein 4 (RBP4), transcript \ ...

Molecule type

nucleic acid

Query Length

1009

Other reports

Distance tree of results

MSA viewer ?

Filter Results

Organism

only top 20 will appear

☐ exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity

E value

Query Coverage

to

to

to

Filter

Reset

Descriptions

Graphic Summary

Alignments

Taxonomy

Sequences producing significant alignments

Download ▼

Select columns ▼

Show

100 ▼

?

☒ select all

100 sequences selected

GenBank

Graphics

Distance tree of results

MSA Viewer

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	PREDICTED: Calidris pugnax retinol binding protein 4 (RBP4), mRNA	Calidris pugnax	260	260	45%	2e-65	82.20%	1117	XM_014966033.1
<input checked="" type="checkbox"/>	PREDICTED: Mesitornis unicolor retinol binding protein 4, plasma (RBP4), mRNA	Mesitornis unicolor	256	256	45%	3e-64	82.06%	987	XM_010182802.1
<input checked="" type="checkbox"/>	PREDICTED: Struthio camelus australis retinol binding protein 4, plasma (RBP4), mRNA	Struthio camelus...	238	238	45%	6e-59	82.02%	1009	XM_009677796.1
<input checked="" type="checkbox"/>	PREDICTED: Apteryx rowi retinol binding protein 4 (RBP4), mRNA	Apteryx rowi	232	232	44%	4e-57	81.51%	999	XM_026066685.1
<input checked="" type="checkbox"/>	PREDICTED: Apteryx australis mantelli retinol binding protein 4, plasma (RBP4), transcript variant X2, mRNA	Apteryx mantelli...	232	232	44%	4e-57	81.51%	843	XM_013951141.1
<input checked="" type="checkbox"/>	PREDICTED: Apteryx australis mantelli retinol binding protein 4, plasma (RBP4), transcript variant X1, mRNA	Apteryx mantelli...	232	232	44%	4e-57	81.51%	907	XM_013951140.1

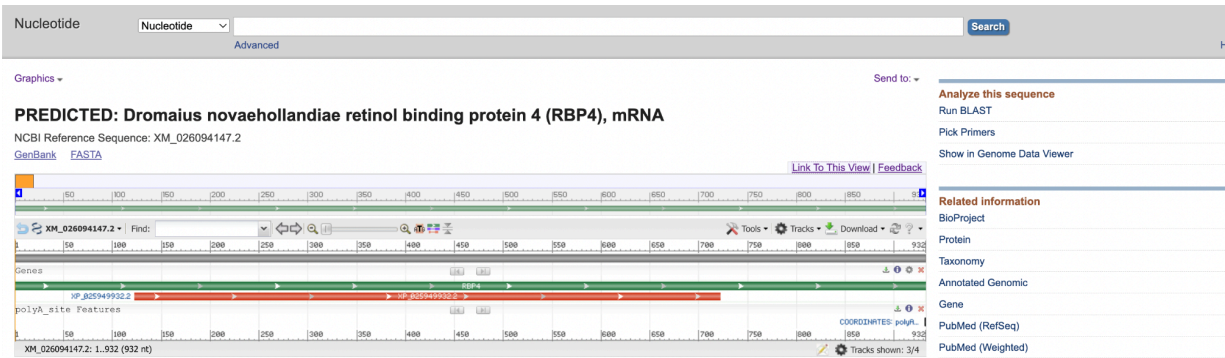
**PREDICTED: Dromaius novaehollandiae retinol binding protein 4 (RBP4), mRNA**

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	GCCCAGTTTACAGTAGATGAGATGGAGAGATGAGTGCCACGGCAAAAGGCAGAGTCAGA	380		
Query 263	CTTTTGAATAACTGGGACGTGTGCGCAGACATGGTGGGCACCTTCACAGACACCGAGGAC	322		
Sbjct 381	CTCTTTAATAACTGGGATGCTGTGCGCAGACATGATTGGCTCTTCAAGGACACAGAGGAT	440		
Query 323	CCTGCCAAGTTCAAGATGAAGTACTGGGGCGTAGCCTCCTTTCTCCAGAAAGGAAATGAT	382		
Sbjct 441	CCTGCCAAATTCAAGATGAAGTACTGGGGCGTGTCTCTTTCTGCAGAAAGGAAATGAT	500		
Query 383	GACCACTGGATCGTCGACACAGACTACGACAGTATGCCGTGCAGTACTCTGCCGCCCTC	442		
Sbjct 501	GATCACTGGGTAGTGGACACAGATTACGATACTTATGCTCTTCATTACTCCTGCCGCCAA	560		
Query 443	CTGAACCTCGATGGCACCTGTGCTGACAGCTACTCCTTCGTGTTTTCCCGGGACCCCAAC	502		
Sbjct 561	CTAAACGAAGATGGCACCTGTGCTGATAGCTATTCCTTGTGTTCTCCCGGGACCCCAAA	620		
Query 503	GGCCTGCCCCCAGAAGCGCAGAAGATTGTAAG	534		
Sbjct 621	GGATTGCCTCCAGAGGCACAGAAATTGTAAG	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



YPKGLRKPQGYIKHCLAGWPSCCHQLGS\*EAAGTGTWTRWPTRGEPRPGCCYWCWPCW  
VAAQQNGTAE\*AASKSRRTSTRTGIVAPGMPWQRKILRGCF CRTTW\*PSLQ\*MRMER\*VP  
RQKAESDSLITGMSVQT\*LALSRTQRILPNSR\*STGALLLFCRKEMMITG\*WTQITILML  
FITPAAN\*TKMAPVLIAIPLCSPGTPKDCLQRHRKL\*DKGR\*TSAWTENTELSFIMDSAL  
KEIQNYGRKSCNSIMDVMLTH\*LSVLLKSF\*MALFRS\*IYLSELCSPPCQ\*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*]

MMFATVYQGKRGKEFGEEVEETSLHLVTGLSTLGQGQRPEPEMGVMQEAVMKGPPAACISPPPYSPSNA  
GSVPQGVAKAPGLHKAFFVWAPRHCQRXLGLSRGYRDRHCLDVMHTQRALPWLLLLLALALLGSSMAERD  
CQVSSFVKVKNFDKTRYSGTWYAMAKKDPEGLFLQDNVVAQFTVDENGQMSATAKGRVRLFNNWDVCDM  
IGSFDTDTEDPAKFVKMYWGVASFLLQKGNDDHWVVDTDYDTYALHYSCRQLNEDGTCADSYSFVFSRDPKG  
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

BLAST® » blastx

HomeRecent ResultsSaved StrategiesHelp

blastntblastpblastxtblastxtblastx

Translated BLAST: blastx

BLASTX search protein databases using a translated nucleotide query. more...

Reset pageBookm

Enter Query Sequence

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

>XM\_026094147.2 PREDICTED: Dromaius novaehollandiae retinol binding protein 4 (RBP4), mRNA  
TATCCCAAGGGATTGCGAAAGCCCCAGGGCTACATAAGCACTGCCTGGCT  
GGGTGGCCCGAGTCTGCTGCC

Query subrange [?](#)  
From   
To

Or, upload file  No file chosen [?](#)

Genetic code

Job Title   
Enter a descriptive title for your BLAST search [?](#)

☐ Align two or more sequences [?](#)

Choose Search Set

Databases ☒ Standard databases (nr etc.): New ☐ Experimental databases [Try experimental clustered nr database](#) [?](#)  
For more info see [What is clustered nr?](#)

Compare ☐ Select to compare standard and experimental database [?](#)

Standard

Database  [?](#)

Organism Optional  ☐ exclude   
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

BLAST

Search database nr using BLASTx (search protein databases using a translated nucleotide query)  
☐ Show results in a new window

BLAST® » blastx » results for RID-2XRCGDFG013

HomeRecent ResultsSaved StrategiesHelp

[< Edit Search](#)

[Save Search](#)

[Search Summary](#) [?](#)

[How to read this report?](#)

[BLAST Help Videos](#)

[Back to Traditional Results Page](#)

Job Title **XM\_026094147.2 PREDICTED: Dromaius novaehollandiae...**

RID [2XRCGDFG013](#) Search expires on 04-30 07:36 am [Download All](#) [?](#)

Program BLASTX [?](#) [Citation](#) [?](#)

Database nr [See details](#) [?](#)

Query ID lcl|Query\_1373720

Description XM\_026094147.2 PREDICTED: Dromaius novaehollandiae ...

Molecule type dna

Query Length 932

Other reports [?](#)

Filter Results

Organism only top 20 will appear ☐ exclude  
  
[+ Add organism](#)

Percent Identity

E value

Query Coverage

to   to   to

Compare these results against the new Clustered nr database [?](#)

Descriptions

Graphic Summary

Alignments

Taxonomy

Sequences producing significant alignments

Download [?](#) Select columns [?](#) Show  [?](#)

☒ select all 100 sequences selected [GenPept](#) [Graphics](#)

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	LOW QUALITY PROTEIN: retinol-binding protein 4 [Gymnogyps californianus]	Gymnogyps californianus	401	401	76%	4e-137	88.24%	312	XP_050755089.1
<input checked="" type="checkbox"/>	LOW QUALITY PROTEIN: retinol-binding protein 4 [Serinus canaria]	Serinus canaria	397	397	76%	2e-135	86.61%	306	XP_030097160.2
<input checked="" type="checkbox"/>	retinol-binding protein 4 isoform X1 [Chiroxiphia lanceolata]	Chiroxiphia lanceolata	395	395	76%	2e-134	85.77%	320	XP_032550772.1
<input checked="" type="checkbox"/>	LOW QUALITY PROTEIN: retinol-binding protein 4 [Haemorrhous mexicanus]	Haemorrhous mexicanus	392	392	76%	9e-134	85.77%	306	XP_059707616.1
<input checked="" type="checkbox"/>	retinol-binding protein 4 [Melozone crissalis]	Melozone crissalis	394	394	76%	2e-133	85.42%	349	XP_054133281.1
<input checked="" type="checkbox"/>	retinol-binding protein 4 [Falco cherrug]	Falco cherrug	386	386	76%	2e-131	86.50%	295	XP_014134532.3
<input checked="" type="checkbox"/>	retinol-binding protein 4 [Falco naumanni]	Falco naumanni	386	386	76%	3e-131	86.55%	295	XP_040462605.1

# 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](#)

## Related Information

[Genome Data Viewer](#) - aligned genomic context

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	QGIAPGLHKA	LPGVAAQLLPPTLGLLRGCGDWHCLDKMaharapawllllviallgg				185
Sbjct 75	QG+AKAPGLHKA P W +	LGL RG D HCLD MAH +RA WLLLL LALLG				134
Query 186	saaERDCRVSSFKVKENFDKNRYS	GTWYAMAKKDPEGLFLQDNVVAQFTVDENGMSATA				365
Sbjct 135	S AERDC+VSSFKVKENFDK RYSGTWYAMAKKDPEGLFLQDNVVAQFTVDENG+MSATA	SMAERDCQVSSFKVKENFDKTRYSGTWYAMAKKDPEGLFLQDNVVAQFTVDENGQMSATA				194
Query 366	KGRVRLFNNWVDCADMIGSF	KDTEPAKFKMKYWGVSFLQKGNDDHWVVDTDYDTYALH				545
Sbjct 195	KGRVRLFNNWVDCADMIGSF	DTEDPAKFKMKYWGVSFLQKGNDDHWVVDTDYDTYALH				254
Query 546	YSCRQLNEDGTCADSYSFVFSRDPKGLPPEAQKIVRQRQV	DLCLDRKYRVIVHNGFCS				719
Sbjct 255	YSCRQLNEDGTCADSYSFVFSRDPKGLPPEAQKIVRQRQ	IDLCLDRKYRVIVHNGFCS				312

**[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).