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.

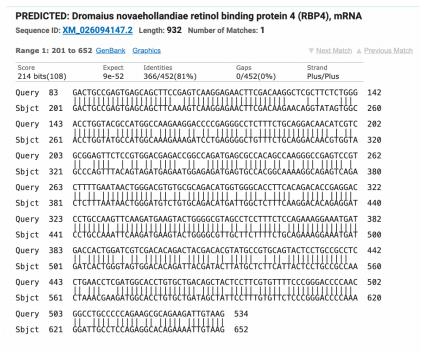
In general, [Q2] is the most difficult for students because it requires you to have a "feel" for how to interpret BLAST results. You need to distinguish between a perfect match to your query (i.e. a sequence that is not "novel"), a near match (something that might be "novel", depending on the results of [Q4]), and a non-homologous result.

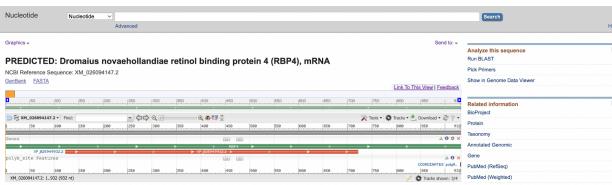
If you are having trouble finding a novel gene try restricting your search to an organism that is poorly annotated.

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):

BLAST®»	blastn suite	Home Recent Results Saved Strategies Help
bloots. Itl		Nucleotide BLAST
blastn bla	stp blastx tblastn tblastx	
Enter Query		tide databases using a nucleotide query. more Reset page Boo
	number(s), gi(s), or FASTA sequence(s) 😯 Clear Query subrange	0
ref NM_001323518.	From	
	// То	
Or, upload file	Choose File No file chosen	
Job Title	NM_001323518:Homo sapiens retinol binding	
☐ Align two or m	Enter a descriptive title for your BLAST search ?	
Choose Sear		
Database	Standard databases (nr etc.): rRNA/ITS databases Genomic -	+ transcript databases O Betacoronavirus
	New Experimental databases Try experimental taxono For more info see What are tax	Download
	Nucleotide collection (nr/nt)	· · · · · · · · · · · · · · · · · · ·
Organism Optional	birds (taxid:8782)	clude Add organism
	Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown	0
Exclude Optional	☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences	
Limit to Optional	Sequences from type material	Tw. COM 0
Entrez Query Optional	Enter an Entrez query to limit search ?	You Tibb Create custom database
Program Sele	ection	
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 seq	uences)
BLAST® » blas	stn suite » results for RID-2XR0K99B013	Home Recent Results Saved Strategies Help
< Edit Search	Save Search Search Summary ▼	Property How to read this report? ■ BLAST Help Videos → Back to Traditional Results Page
Your search	ch is limited to records that include: birds (taxid:8782)	
Job Title	NM_001323518:Homo sapiens retinol binding	Filter Results
RID	2XR0K99B013 Search expires on 04-30 07:30 am Download All ▼	
Program	BLASTN ② <u>Citation</u> ▽	Organism only top 20 will appear exclude
Database	nt <u>See details</u> ♥	Type common name, binomial, taxid or group name + Add organism
Query ID	NM_001323518.2	
Description	Homo sapiens retinol binding protein 4 (RBP4), transcript v	Percent Identity E value Query Coverage
Molecule type Query Length	nucleic acid 1009	to to to
Other reports	Distance tree of results MSA viewer	Filter Reset
Descriptions	Graphic Summary Alignments Taxonomy	
Sequences p	roducing significant alignments	Download ∨ Select columns ∨ Show 100 ♥
✓ select all	100 sequences selected	GenBank Graphics Distance tree of results MSA Viewer
Select all	Description	Scientific Name Max Total Query E Per. Acc.
✓ PREDICTED	2: Calidris pugnax retinol binding protein 4 (RBP4), mRNA	<u>Calidris pugnax</u> 260 260 45% 2e-65 82.20% 1117 <u>XM 014966033.1</u>
	D: Mesitornis unicolor retinol binding protein 4, plasma (RBP4), mRNA	Mesitornis unicolor 256 256 45% 3e-64 82.06% 987 XM_010182802.1
PREDICTED	2: Struthio camelus australis retinol binding protein 4, plasma (RBP4), mRNA	<u>Struthio camelus</u> 238 238 45% 6e-59 82.02% 1009 <u>XM_009677796.1</u>
	2: Apteryx rowi retinol binding protein 4 (RBP4), mRNA	<u>Apteryx rowi</u> 232 232 44% 4e-57 81.51% 999 <u>XM_026066685.1</u>
	2: Apteryx australis mantelli retinol binding protein 4. plasma (RBP4), transcript va 2: Apteryx australis mantelli retinol binding protein 4. plasma (RBP4), transcript va	





```
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
          GACTGCCGAGTGAGCAGCTTCCGAGTCAAGGAGAACTTCGACAAGGCTCGCTTCTCTGGG
     83
          GACTGCCGAGTGAGCAGCTTCAAAGTCAAGGAGAACTTCGACAAGAACAGGTATAGTGGC
Sbjct
      201
                                                                260
          ACCTGGTACGCCATGGCCAAGAAGGACCCCGAGGGCCTCTTTCTGCAGGACAACATCGTC
Ouery
     143
          ACCTGGTATGCCATGGCAAAGAAGATCCTGAGGGGCTGTTTCTGCAGGACAACGTGGTA
Sbjct
     261
```

```
Query
         GCGGAGTTCTCCGTGGACGAGACCGGCCAGATGAGCGCCACAGCCAAGGGCCGAGTCCGT
     203
                                                         262
                GCCCAGTTTACAGTAGATGAGAATGGAGAGATGAGTGCCACGGCAAAAGGCAGAGTCAGA
Sbjct
     321
                                                         380
         \tt CTTTTGAATAACTGGGACGTGTGCGCAGACATGGTGGGCACCTTCACAGACACCGAGGAC
                                                         322
Query
     263
         \tt CTCTTTAATAACTGGGATGTCTGTGCAGACATGATTGGCTCTTTCAAGGACACAGAGGAT
Sbjct
                                                         440
Query
     323
         CCTGCCAAGTTCAAGATGAAGTACTGGGGCGTAGCCTCCTTTCTCCAGAAAGGAAATGAT
                                                         382
         Sbjct
         \tt CCTGCCAAATTCAAGATGAAGTACTGGGGCGTTGCTTCTTTTCTGCAGAAAGGAAATGAT
                                                         500
         {\tt GACCACTGGATCGTCGACACAGACTACGACACGTATGCCGTGCAGTACTCCTGCCGCCTC}
Query
     383
                                                         442
         GATCACTGGGTAGTGGACACAGATTACGATACTTATGCTCTTCATTACTCCTGCCGCCAA
Sbjct
     501
                                                         560
Query
     443
         \tt CTGAACCTCGATGGCACCTGTGCTGACAGCTACTCCTTCGTGTTTTCCCGGGACCCCAAC
                                                         502
                Sbjct
     561
         \tt CTAAACGAAGATGGCACCTGTGCTGATAGCTATTCCTTTGTGTTCTCCCGGGACCCCAAA
                                                         620
Query
     503
         GGCCTGCCCCAGAAGCGCAGAAGATTGTAAG
            Sbjct
     621 GGATTGCCTCCAGAGGCACAGAAAATTGTAAG
                                  652
```

[Q3] Gather information about this "novel" <u>protein</u>. 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.

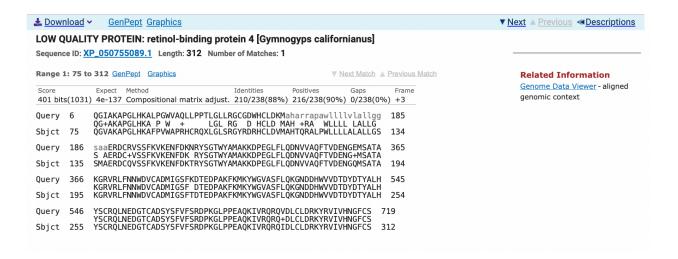
Protein sequence of choice matches Subject above:

>XP_050755089.1 LOW QUALITY PROTEIN: retinol-binding protein 4 [Gymnogyps californianus]

MMFATVYQGKRGKEFGEEVEETSLHLVTGLSTLGQGQRPPEPMGVMQEAVMKGPQPAACISPPPYSPSNA
GSVPQGVAKAPGLHKAFPVWAPRHCRQXLGLSRGYRDRHCLDVMAHTQRALPWLLLLALALLGSSMAERD
CQVSSFKVKENFDKTRYSGTWYAMAKKDPEGLFLQDNVVAQFTVDENGQMSATAKGRVRLFNNWDVCADM
IGSFTDTEDPAKFKMKYWGVASFLQKGNDDHWVVDTDYDTYALHYSCRQLNEDGTCADSYSFVFSRDPKG
LPPEAOKIVROROIDLCLDRKYRVIVHNGFCS

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

Enter accession mayber(s, pile), or FASTA sequence(s) © clare Compare The accession mayber(s, pile), or FASTA sequence(s) © clare Compare Compare Compare Standard And protein registry, experimental clustered in database Genetic code Similator (P) Compare Standard And protein code (Samurdor (P) Compare Standard Compare Standard Databases Standard	rotein databases using a translated nucleotide query, more Reset page Jery subrange	Cuery sub	BLA quence(s) ? Clear andiae retinol binding	Sequence number(s), gi(s), or FASTA se	Enter Query S
Enter Query Sequence Enter accession number(s), gic), or FASTA sequence(s) © care "Suggested (2 PREDCTED braneau recentral reason in bridge "Our plead file "Suggested (1) "Job Title "No. (26094147 2 PREDCTED: Dramatus movembridates	uery subrange 😯	Query sub	BLA quence(s) ? Clear andiae retinol binding	Sequence number(s), gi(s), or FASTA se	Enter Query S
Enter accession mathers), gift, or FASTA sequence(s) © core "MAL (1994)H17.2 PREDICTED Dromalus novaeholandare rend binding "Try experimental clustered in database "Try experimental clustered in dat	uery subrange ②	From	andiae retinol binding	number(s), gi(s), or FASTA se	
Promiser 4 (1994), miles of the control of the cont		From	andiae retinol binding		
Compare Standard databases (or sec.) Description Standard Description Standard databases (or sec.) Description Descript	h	т	_	PREDICTED: Dromaius novaeholi	
Genetic code Standard (1) Wild			AAAGCACTGCCTGGC	RNA FGCGAAAGCCCCAGGGCTACA	orotein 4 (RBP4), ml TATCCCAAGGGAT
Job Title XML 026094147.2 PREDICTED: Dromatus novaehollandiae Elitar a decription Elitar a decription to the your BLAST search Chabbases	A		en 😯	Choose File No file chos	r, upload file
Align have or more advantages with the your bLAST easure Choose Search Set Databases Select to compare standed and experimental databases Select databases (in els.): Select databases (in els.): Select databases (in els.): Select databases (in els.): Select databases Non-reducdand protein sequences (in) Ty experimental clustered in database For more into see What is clustered in database Per more into see What is clustered in database Select database (in els.): Select database (in els.)	a	~		Standard (1)	Senetic code
Algo has or more sequences Choose Search Set Databases Standard databases (nr.dc.): Experimental databases Standard Databases Standard Databases Standard Databases Standard Databases Standard Databases Interroganism name or id-completions will be suggested Organism Enter organism permit or its vide (nr.dx): Permit of the vide of the completion will be suggested Organism Enter organism name or id-completions will be suggested Organism Enter organism name or id-completions will be suggested Organism Enter organism name or id-completions will be suggested Organism Enter organism name or id-completions will be suggested Organism Enter organism name or id-completions will be suggested Organism Enter organism name or id-completions will be suggested Organism Enter organism name or id-completions will be suggested Organism Enter organism name or id-completions will be suggested Organism Enter organism name or id-completions will be suggested Organism Enter organism name or id-completions will be suggested Organism Enter organism name or id-completions will be suggested Organism Enter organism name or id-completions will be suggested Organism Enter organism name or id-completions will be suggested Organism Enter organism name or id-completions will be suggested Organism only incompletions Organism Organ		aehollandiae			ob Title
Delabases © Standard databases (n° etc.):					Align two or mo
Standard Select to compare standard and experimental databases Standard Database Organism Operation Enter organism name or IR—completions will be suggested Enter organism name or IR—completions and IR suggested Enter organism				ch Set	Choose Search
Database Organism Org	tabases	erimental databases	etc.): New Exp	Standard databases (n	atabases
Database Organism Organism Coptional Enter organism name or id -completions will be suggested exclude Add organism	9	al database 😯	dard and experiment	Select to compare star	ompare
Organism Optional Technic organism name of id-completions will be auggested exclude Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences Enter organism name of id-completions will be auggested exclude Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences Exclude Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences Exclude Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences Exclude Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences Exclude Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences Exclude Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences Exclude Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences Exclude Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences Exclude Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences Exclude Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences Exclude Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences Exclude Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences Exclude Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences Exclude Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences Exclude Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences Exclude Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences Exclude Models (XM/XP) Uncultured/environmental sample sequences Exclu					Standard
Exclude	0	v •	sequences (nr)	Non-redundant protein	Database
Exclude Models (XMXP) Non-redundant RerSeq 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 Search Sea	exclude Add organism	be suggested	or idcompletions will	Enter organism name	
BLAST Search database nr using Blastx (search protein databases using a translated nucleotide query) AST® >> blastx >> results for RID-2XRCGDFG013 Home Recent Results Saved Strategies Edit Search Save Search Search Search Summary Phone Recent Results Saved Strategies Edit Search Save Search Search Search Search summary Phone Recent Results Edit Search Save Search Search Search summary Phone Recent Results Edit Search Save Search Search Search Search summary Phone Recent Results Edit Search Save Search Search summary Phone Recent Results Filter Results Organism only top 20 will appear Type common name, binomial, taxid or group name Add organism Percent Identity Evalue Query Coverage Type common name, binomial, taxid or group name Add organism Percent Identity Evalue Query Coverage Type common name, binomial, taxid or group name Add organism Filter Results Compare these results against the new Clustered nr database BLAST Descriptions Graphic Summary Alignments Taxonomy Sequences producing significant alignments Download Select columns Show 100			•		Optional
BLAST Search database nr using Blastx (search protein databases using a translated nucleotide query) AST ** blastx ** results for RID-2XRCGDFG013 Home Recent Results Saved Strategies Edit Search Save Search Search Summary **	WP) Uncultured/environmental sample sequences	eq proteins (WP)	Non-redundant RefS	☐ Models (XM/XP) ☐	
AST® » blastx » results for RID-2XRCGDF6013 Home Recent Results Saved Strategier Edit Search Search Search Search Search Summary Phow to read this report? BLAST Help Videos Back to Traditional Results 2XRCGDF6013 Search expires on 04-30 07:36 am Download All gram BLASTX © Citation abase nr See details arry ID Icl[Query_1373720 Idecule type dna arry Length 932 grer reports Compare these results against the new Clustered nr database Compare these results against the new Clustered nr database Compare these results against the new Clustered nr database Seedet columns Select columns Show 100 Sequences producing significant atignments Download Select columns Source Cover Scientific Name Score Score Cover Source Cover Scientific Name Score Cover Score Score Cover Cow QUALITY PROTEIN: retinol-binding protein 4 [Gymnogypa californianus] Serious scanaria 397 397 76% 4e-137 88.24% 312 XP_0507. Serious Score Score Cover Cow QUALITY PROTEIN: retinol-binding protein 4 [Serious canaria] Serious scanaria 397 397 76% 2e-135 86.81% 306 XP_0300.					Ориона
AST® » blastx » results for RID-2XRCGDFG013 Home Recent Results Saved Strategies Edit Search Save Search Search Search Summary Phow to read this report? BLAST Help Videos Back to Traditional Results XM_026094147.2 PREDICTED: Dromaius novaehollandiae Filter Results Filter Results Filter Results Organism only top 20 will appear Type common name, binomial, taxid or group name Add organism Percent Identity E value Query Coverage Identity E value Query Coverage Type common name, binomial, taxid or group name Add organism Percent Identity E value Query Coverage Type common name, binomial, taxid or group name Add organism Percent Identity E value Query Coverage Type common name, binomial, taxid or group name Add organism Percent Identity E value Query Coverage Type common name, binomial, taxid or group name Add organism Percent Identity E value Query Coverage Type common name, binomial, taxid or group name Add organism Percent Identity E value Query Coverage Type common name, binomial, taxid or group name Add organism Percent Identity E value Query Coverage Type common name, binomial, taxid or group name Add organism Percent Identity E value Query Coverage Type common name, binomial, taxid or group name Add organism Percent Identity E value Query Coverage Type common name, binomial, taxid or group name Add organism Percent Identity E value Query Coverage Type common name, binomial, taxid or group name Add organism Percent Identity E value Query Coverage Type common name, binomial, taxid or group name Add organism Type common name, binomial, taxid or group name Add organism Type common name, binomial, taxid or group name Add organism Type common name, binomial, taxid or group name Add organism Type common name, binomial, taxid or group name Add organism Type common name, binomial, taxid or group name Add organism Type common name, binomial, taxid or group name Add organism Type common name, binomial, taxid or group name Add organism Type common name, binomi				1	
Edit Search Save Search Save Search Search Summary How to read this report? BLAST Help Videos Back to Traditional Results Filter Results Graphic Summary Alignments Taxonomy Compare these results against the new Clustered nr database Graphic Summary Alignments Taxonomy Compare these results against the new Clustered nr database BLAST Compare these results against the new Clustered nr database Compare these results against the new Clustered nr database BLAST Compare these results against the new Clustered nr database Compare the new Clustered nr database Compare the new Clustered nr database Compare the new Cluster	es using a translated nucleotide query)	ein databases usin			BLAST
gram BLASTX ② Citation ▼ abase nr See details ▼ ary ID Icl Query_1373720 Scription XIM_026094147.2 PREDICTED: Dromaius novaehollandiae Recule type dna ary Length 932 er reports ③ Compare these results against the new Clustered nr database ② BLAST Compare these results against the new Clustered nr database ② BLAST Compare these results against the new Clustered nr database ③ BLAST Compare these results against the new Clustered nr database ③ BLAST Compare these results against the new Clustered nr database ③ BLAST Compare these results against the new Clustered nr database ③ BLAST Compare these results against the new Clustered nr database ③ BLAST Compare these results against the new Clustered nr database ③ BLAST Compare these results against the new Clustered nr database ③ BLAST Compare these results against the new Clustered nr database ③ BLAST Compare these results against the new Clustered nr database ③ BLAST Compare these results against the new Clustered nr database ③ BLAST Compare these results against the new Clustered nr database ② BLAST Compare these results against the new Clustered nr database ② BLAST Compare these results against the new Clustered nr database ② BLAST Compare these results against the new Clustered nr database ② BLAST Compare these results against the new Clustered nr database ② BLAST Compare these results against the new Clustered nr database ② BLAST Compare these results against the new Clustered nr database ② BLAST Compare these results against the new Clustered nr database ② BLAST Compare these results against the new Clustered nr database ② BLAST Compare these results against the new Clustered nr database ② BLAST Compare these results against the new Clustered nr database ② BLAST Compare these results against the new Clustered nr database ② BLAST Compare these results against the new Clustered nr database ② BLAST Compare these results against the new Clustered nr database ② BLAST Compare these results against the new Clustered nr database ② BL	ndiae Filter Results		CTED: Dromaius n	(M_026094147.2 PREDI	
Type common name, binomial, taxid or group name + Add organism Percent Identity E value Query Coverage to to to to to Tope common name, binomial, taxid or group name + Add organism Percent Identity E value Query Coverage Tope common name, binomial, taxid or group name + Add organism Percent Identity E value Query Coverage Tope common name, binomial, taxid or group name + Add organism Percent Identity E value Query Coverage Filter Re 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 Select all 100 s		DOWINGED AII	165 011 04-30 07.30 all		
Percent Identity E value Query Coverage To to The compare these results against the new Clustered nr database Compare these results against the new Cl	Type common name, binomial, taxid or group name				
Percent Identity E value Query Coverage	+ Add organism				
Compare these results against the new Clustered nr database BLAST	Depart Identity Evalue Overvoor		TED: D	•	•
Compare these results against the new Clustered nr database Compare these results against the new Clustered nr database BLAST Descriptions Graphic Summary Alignments Taxonomy Sequences producing significant alignments Download Select columns Show 100 Sequences selected GenPept Graph Description Scientific Name Score Score Cover value Ident Len Access Score Score Cover value Ident Len Access LOW QUALITY PROTEIN: retinol-binding protein 4 [Gymnogyps californianus] LOW QUALITY PROTEIN: retinol-binding protein 4 [Gymnogyps californianus] Serinus canaria		ovaenoliandiae ,	TED: Dromaius no		•
Compare these results against the new Clustered nr database BLAST Descriptions Graphic Summary Alignments Taxonomy Sequences producing significant alignments Download Select columns Show 100 Sequences producing significant alignments Download Select columns Show 100 Low QUALITY PROTEIN: retinol-binding protein 4 [Gymnogyps californianus] Gymnogyps californianus Serinus canaria Serinus canaria Serinus canaria Serinus canaria Serinus canaria	to to to				
Compare these results against the new Clustered nr database	Filter Reset				
Descriptions Graphic Summary Alignments Taxonomy Sequences producing significant alignments Download ∨ Select columns ∨ Show 100 ∨ Select all 100 sequences selected GenPept Graph Description Scientific Name Max Total Query E Score Cover value Ident Len Access Score Cover value Ident Len Access Low QUALITY PROTEIN: retinol-binding protein 4 [Gymnogyps californianus] Gymnogyps californianus 401 401 76% 4e-137 88.24% 312 XP_05078 LOW QUALITY PROTEIN: retinol-binding protein 4 [Serinus canaria] Serinus canaria 397 397 76% 2e-135 86.61% 306 XP_03008	ast the new Clustered nr database ® BLAST	ults against the	mpare these res	Co	
✓ select all 100 sequences selected GenPept Scription GenPept Scription GenPept Scription Scientific Name Max Score	my	Taxonomy	Alignments	Graphic Summary	escriptions
✓ select all 100 sequences selected GenPept Scription GenPept Scription GenPept Scription Scientific Name Max Score	Doubled V. Calest salumne V. Chou				
Description Scientific Name Score Sc	Download Select Columns Show 100 F		gnments	oducing significant at	equences pro
LOW QUALITY PROTEIN: retinol-binding protein 4 [Serinus canaria] Scrientlic Name Score Score Score Cover value Ident Len Access Access Len Access Len Access Len Access Len Len Access Len Len Access Len				00 sequences selected	select all 10
LOW QUALITY PROTEIN: retinol-binding protein 4 [Serinus canaria] Serinus canaria 397 397 76% 2e-135 86.61% 306 XP_03009	GenPept Graphics		ition	Descri	
	Scientific Name Max Total Query E Per. Acc.		in 4 [Gymnogyps calif	PROTEIN: retinol-binding protein	LOW QUALITY
reting Linding protein A isoform V1 [Chirovinhia Janosolata] Chirovinhia Janosolata 205 205 700/ 0-434 05 770/ 000 VD 0005	Scientific Name Max Total Query E Per. Acc. Score Score Cover value Ident Len Accession	ornianus]			I OW OUALITY
retinol-binding protein 4 isoform X1 [Chiroxiphia lanceolata] Chiroxiphia lanceolata 395 395 76% 2e-134 85.77% 320 XP_0325	Scientific Name Max Score Score Score Score Cover Value E Ident Ident Ident Acc. Ident Ident Accession Gymnogyps californianus 401 401 76% 4e-137 88.24% 312 XP_05075508	ornianus]	in 4 [Serinus canaria]	PROTEIN: retinol-binding proti	LOW QUALITI
	Scientific Name Max Score Score Score Score Cover Value E Ident Ident Ident Acc. Ident Ident Accession Gymnogyps californianus 401 401 76% 4e-137 88.24% 312 XP_05075508	ornianus]			
	Scientific Name Max Score Total Score Query Cover E Value Per. Ident Acc. Len Accession Gymnogyps californianus 401 401 76% 4e-137 88.24% 312 XP_05075508 Serinus canaria 397 397 76% 2e-135 86.61% 306 XP_03099718 Chiroxiphia lanceolata 395 395 76% 2e-134 85.77% 320 XP_03255075		a lanceolata]	protein 4 isoform X1 [Chiroxiphi	retinol-binding.
✓ retinol-binding protein 4 [Falco cherrug] Falco cherrug 386 386 76% 2e-131 86.50% 295 XP_0141: ✓ retinol-binding protein 4 [Falco naumanni] Falco naumanni 386 386 76% 3e-131 86.55% 295 XP_0404!	Scientific Name Max Score Score Total Score Score Query Low Value E Ident Len Acc. Len Accession Gymnogyps californianus 401 401 76% 4e-137 88.24% 312 XP_05075508 Serinus canaria 397 397 76% 2e-135 86.61% 306 XP_0309716 Chiroxiphia lanceolata 395 395 76% 2e-134 85.77% 320 XP_0325507 Haemorhous mexicanus 392 392 76% 9e-134 85.77% 306 XP_0597076 Melozone crissalis 394 394 76% 2e-133 85.42% 349 XP_05413328		a lanceolata]	protein 4 isoform X1 [Chiroxiphi PROTEIN: retinol-binding protein 4 [Melozone crissalis]	retinol-binding. LOW QUALITY retinol-binding.



[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).