

## Find a Gene Project Assignment

1. 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:** tumor necrosis factor receptor superfamily, member 6B, decoy

**Accession:** NP\_003814

**Species:** Homo sapiens

2. 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 searched against Ornithorhynchus

**Database:** Expressed Sequence Tags (est)

**Organism:** Ornithorhynchus (Taxid: 9257)

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  $\text{⌘}$ -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.

See search setup in screen-shot below:

The screenshot shows the NCBI BLAST search setup interface. The 'Enter Query Sequence' section has a text box containing 'NP\_003814' and a 'Query subrange' section with 'From' and 'To' fields. Below this is an 'Or, upload file' section with a 'Choose File' button and a 'Job Title' field containing 'NP\_003814:tumor necrosis factor receptor superfamily...'. There is also an 'Align two or more sequences' checkbox. The 'Choose Search Set' section has a 'Database' dropdown set to 'Expressed sequence tags (est)', an 'Organism' field set to 'Ornithorhynchus', and an 'Exclude' section with checkboxes for 'Models (XM/XP)' and 'Uncultured/environmental sample sequences'. There is also a 'Limit to' section with a checkbox for 'Sequences from type material'. At the bottom, there is an 'Entrez Query' field and a 'BLAST' button. The 'BLAST' button is highlighted in blue. Below the button, there is a checkbox for 'Show results in a new window'.

**Enter Query Sequence**

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

NP\_003814

Query subrange [?](#)

From

To

Or, upload file [?](#)

[Choose File](#) No file chosen

**Job Title**

NP\_003814:tumor necrosis factor receptor superfamily...

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

☐ Align two or more sequences [?](#)

**Choose Search Set**

**Database**

Expressed sequence tags (est) [?](#)

**Organism**

Optional

Ornithorhynchus ☐ 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 [?](#)

[You Tube](#) [Create custom database](#)

**BLAST**

Search database est using Tblastn (search translated nucleotide databases using a protein query)

☐ Show results in a new window

The search yielded 5 results, a screen shot of the results is shown below:

Descriptions

Graphic Summary

Alignments

Taxonomy

Sequences producing significant alignments

Download 

New

Select columns 

Show

100

select all

5 sequences selected

GenBank

Graphics

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<div></div>	KAAN-aaa12e11.b1 Platypus_EST_Cell_line_1.0-4.0kb Ornithorhynchus anatinus cDNA similar to ref NP_003814.1 ...	Ornithorhynchus...	268	268	70%	3e-91	57.75%	715	EG339348.1
<div></div>	KAAN-aaa13f10.b1 Platypus_EST_Cell_line_1.0-4.0kb Ornithorhynchus anatinus cDNA similar to gb AAP03889.1 ...	Ornithorhynchus...	167	167	45%	4e-52	54.41%	565	EG339448.1
<div></div>	KAAN-aab22a09.b1 Platypus_EST_Cell_line_1.0-4.0kb Ornithorhynchus anatinus cDNA similar to gb AAP03889.1 ...	Ornithorhynchus...	152	152	35%	3e-46	59.81%	651	EH001086.1
<div></div>	KAAN-aaa53g01.b1 Platypus_EST_Cell_line_1.0-4.0kb Ornithorhynchus anatinus cDNA mRNA sequence	Ornithorhynchus...	118	118	32%	4e-34	55.67%	350	EH001579.1
<div></div>	KAAN-aaa25b06.b1 Platypus_EST_Cell_line_1.0-4.0kb Ornithorhynchus anatinus cDNA similar to gb AAP03889.1 ...	Ornithorhynchus...	100	151	33%	6e-26	65.15%	659	EG339750.1

**Chosen match:** Accession EG339348.1, a 715 base pair cDNA clone from *Ornithorhynchus anatinus*. See below for alignment details.

### Alignment Details:

Query: tumor necrosis factor receptor superfamily member 6B precursor [Homo sapiens] Query ID: NP\_003814.1 Length: 300

>KAA-aaa12e11.b1 Platypus\_EST\_Cell\_line\_1.0-4.0kb Ornithorhynchus anatinus cDNA similar to ref|NP\_003814.1| tumor necrosis factor receptor superfamily, member 6b; decoy receptor 3 [Homo sapiens] ref|NP\_116563.1| tumor necrosis factor receptor superfamily, member 6b; decoy receptor 3 [Homo sapiens] sp|O95407|TR6B\_HUMAN Tumor necrosis fact, mRNA sequence  
Sequence ID: EG339348.1 Length: 715  
Range 1: 63 to 701

Score: 268 bits(686), Expect:3e-91,  
Method: Compositional matrix adjust.,  
Identities: 126/213(59%), Positives: 162/213(76%), Gaps: 1/213(0%)

Query	34	PTYPWRDAETGERLVCAQCPPGTFVQRPCRRDSPTTCGPCPPRHYTQFWNYLERCRYCNV	93
		PTY W+D+ T ERL C QCPPGT+V + C R SPT C PCP HYTQ+WNYL++CRYCNV	
Sbjct	63	PTYSWKDSTTQERLQCQCQCPPGTYSVQHCSTRTSPTQCQPCPTLHYTQYWNYLDKCRYCNV	242
Query	94	LCGEREEEEARACHATHNRACRCRTGFFAHAGFCLEHASCPPGAGVIAPGTPSNTQCQPC	153
		CG +EEE C ATHNR C+C+ G++A+ FC+EH++CP G+GV++ GTP++NT+CQ C	
Sbjct	243	FCGAQEEEVHPCSATHNRVCQCQPGYYAYMDFCIEHSTCPLGSGVVSQGTPTKNTCEQEC	422
Query	154	PPGTFsassssssEQCQPHRNCTALGLALNVPGSSSHDTLCTSTCTGFPLSTRVPGAEECER	213
		P GTFS +SS +E CQ H+NCT LG+ +NVPG+ HDTLCT C F L++ PG ++CE+	
Sbjct	423	PRGTFSDNSSRTEPCQSHQNCTLLGMKVNVPGNRFHDTLCTRCDFQLNSSEPGNKDCEQ	602
Query	214	AVIDFVAFQDISIKRLQRLQLAL-EAPEGWGPT	245
		A+IDFVA+QDI +KRL RL Q L EAP G T	
Sbjct	603	ALIDFVAYQDIPLKRLRLQLQVVLGEAPGAAGQT	701

3. 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

```
> O. anatinus protein (sequence taken from BLAST results)
TKKLGTFVLAVTFPMGSNNPPTYSWKDSTTQERLQCQCPCPGTYVSQHCSRTSPTQCQPCPTLHYTQYWNLYLDKCRYCN
VFCGAQEEEVHPCSAATHNRVCQCQPGYYAYMDFCIEHSTCPLGSGVVSQGTPTKNTECQECPRGTFSDNSSRTEPCQSH
QNCTLLGMKVNVPGNRFHDTLCTRCDNFQLNSSEPGNKDCEQALIDFVAYQDIPLKRLRLRQQVLGEAPGAAGQTRGFQ
V
```

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.

**Name:** Ornithorhynchus tumor necrosis factor receptor superfamily, member 6B, decoy

**Species:** Ornithorhynchus anatinus

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata;  
Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii;  
Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Prototheria; Monotremata;  
Ornithorhynchidae; Ornithorhynchus

4. 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.

## Details:

A BLASTP search against NR database was used (see setup in first screen-shot below).

This yielded a top hit result is to a protein from *Ornithorhynchus anatinus* (Platypus)

See additional screen shots below for top hits and selected alignment details:

The first hit has a 95.74% identity with our query, thus since the percent identity reported is less than 100% it is likely that our gene is novel as defined in the question. So we have succeeded as required.

Enter Query Sequence

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

> O. anatinus protein (sequence taken from BLAST results)  
TKKLGTFVLATFPMGSSNPPTYSWKDSTTQERLQCQQCPPTGYVSQHS  
RTSPQTCQPCPTLHYTYWNYLDKRCYCNVFCGAQEEVHPCATHNRVC  
QCQPGYAYMDFCIHSTCPLSGVWSQGTPTKNTCEQECPRGTFSDNSS  
RTEPCQSHQNCITLLGMKVNVPGNRFHDTLCRCDNFQLNSSEPGNKDCEQ  
ALIDFVAYQDIPLKRLRLQVLGEAPGAAGQTRGFQV

Query subrange

From

To

Or, upload file

No file chosen [?](#)

Job Title

O. anatinus protein (sequence taken from BLAST...)

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

☐ Align two or more sequences [?](#)

Choose Search Set

Database

Non-redundant protein sequences (nr) [?](#)

Organism

Optional

Enter organism name or id--completions will be suggested ☐ 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

Program Selection

Algorithm

☐ Quick BLASTP (Accelerated protein-protein BLAST)

☒ blastp (protein-protein BLAST)

☐ PSI-BLAST (Position-Specific Iterated BLAST)

☐ PHI-BLAST (Pattern Hit Initiated BLAST)

☐ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm [?](#)

Search database nr using Blastp (protein-protein BLAST)

☐ Show results in a new window

Descriptions	Graphic Summary	Alignments	Taxonomy					
Sequences producing significant alignments								
Download <span>▼</span> <span>New</span> Select columns <span>▼</span> Show <span>100</span> <span>▼</span> <span>?</span>								
<input checked="" type="checkbox"/> select all 100 sequences selected								
GenPept Graphics Distance tree of results Multiple alignment <span>New</span> MSA Viewer								
Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/> tumor necrosis factor receptor superfamily member 6B [Ornithorhynchus anatinus]	Ornithorhyn...	470	470	98%	1e-165	95.74%	315	XP_028926265.1
<input checked="" type="checkbox"/> tumor necrosis factor receptor superfamily member 6B [Tachyglossus aculeatus]	Tachyglossu...	458	458	98%	4e-161	93.19%	314	XP_038606231.1
<input checked="" type="checkbox"/> LOW QUALITY PROTEIN: tumor necrosis factor receptor superfamily member 6B [Ph...	Phascolarct...	329	329	94%	3e-109	67.26%	371	XP_020837056.1
<input checked="" type="checkbox"/> tumor necrosis factor receptor superfamily member 6B [Vombatus ursinus]	Vombatus ur...	324	324	94%	2e-108	65.93%	302	XP_027703820.1
<input checked="" type="checkbox"/> tumor necrosis factor receptor superfamily member 6B [Sarcophilus harrisii]	Sarcophilus...	324	324	92%	2e-108	67.27%	302	XP_003757643.1
<input checked="" type="checkbox"/> tumor necrosis factor receptor superfamily member 6B [Trichosurus vulpecula]	Trichosurus...	323	323	95%	1e-107	65.35%	302	XP_036604199.1
<input checked="" type="checkbox"/> PREDICTED: tumor necrosis factor receptor superfamily member 6B isoform X2 [Mon...	Monodelphi...	314	314	91%	5e-104	66.51%	324	XP_007475604.1
<input checked="" type="checkbox"/> PREDICTED: tumor necrosis factor receptor superfamily member 6B isoform X1 [Mon...	Monodelphi...	315	315	91%	7e-104	66.51%	345	XP_007475603.1
<input checked="" type="checkbox"/> tumor necrosis factor receptor superfamily member 6B [Gracilinanus agilis]	Gracilinanus...	317	317	93%	2e-102	65.32%	510	XP_044515587.1
<input checked="" type="checkbox"/> tumor necrosis factor receptor superfamily member 6B [Dromiciops gliroides]	Dromiciops...	306	306	94%	2e-101	62.39%	302	XP_043836134.1
<input checked="" type="checkbox"/> tumor necrosis factor receptor superfamily member 6B isoform X2 [Alligator sinensis]	Alligator sin...	293	293	91%	3e-95	62.39%	355	XP_025057489.1
<input checked="" type="checkbox"/> PREDICTED: tumor necrosis factor receptor superfamily member 6B isoform X2 [Croc...	Crocodylus...	292	292	91%	3e-95	61.93%	335	XP_019403867.1
<input checked="" type="checkbox"/> PREDICTED: tumor necrosis factor receptor superfamily member 6B isoform X2 [Gavi...	Gavialis gan...	292	292	91%	4e-95	61.93%	335	XP_019365590.1
<input checked="" type="checkbox"/> PREDICTED: tumor necrosis factor receptor superfamily member 6B isoform X1 [Croc...	Crocodylus...	291	291	91%	7e-95	61.93%	336	XP_019403866.1

[Download](#) [GenPept](#) [Graphics](#)

## tumor necrosis factor receptor superfamily member 6B [Ornithorhynchus anatinus]

Sequence ID: [XP\\_028926265.1](#) Length: 315 Number of Matches: 1

Range 1: 17 to 250 [GenPept](#) [Graphics](#)

[▼ Next Match](#) [▲ Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps
470 bits(1209)	1e-165	Compositional matrix adjust.	225/235(96%)	227/235(96%)	1/235(0%)
Query 4	LGTFVLAVTFPMGSNNPPTYSWKDSTTQERLQCQCPCPGTYVSQHCSRTSPTQCQPCPTL				63
	GTFVLAVTFPMGSNNPPTYSWKDSTTQERLQCQCPCPGTYVSQHCSRTSPTQCQPCPTL				
Sbjct 17	FGTFVLAVTFPMGSNNPPTYSWKDSTTQERLQCQCPCPGTYVSQHCSRTSPTQCQPCPTL				76
Query 64	HYTQYWNYLDKCRYCNVFCGAQEEVHPCSATHNRVCQCQPGYYAYMDFCIEHSTCPLGS				123
	HYTQYWNYLDKCRYCNVFCGAQEEVHPCSATHNRVCQCQPGYYAYMDFCIEHSTCPLGS				
Sbjct 77	HYTQYWNYLDKCRYCNVFCGAQEEVHPCSATHNRVCQCQPGYYAYMDFCIEHSTCPLGS				136
Query 124	GVVSQGTPTKNTCEQCPRGTFSDNSSRTEPCQSHQNTLLGMKVNPVGNRFHDTLCITRC				183
	GVVSQGTPTKNTCEQCPRGTFSDNSSRTEPCQSHQNTLLGMKVNPVGNRFHDTLCITRC				
Sbjct 137	GVVSQGTPTKNTCEQCPRGTFSDNSSRTEPCQSHQNTLLGMKVNPVGNRFHDTLCITRC				196
Query 184	DNFQLNSSEPGNKDCEQALIDFVAYQDIPLKRLRLQQVLGEAPGAAGQTRGFQV				238
	DNFQLNSSEPGNKDCEQALIDFVAYQDIPLKRLRLQQVLG+ G +GFQV				
Sbjct 197	DNFQLNSSEPGNKDCEQALIDFVAYQDIPLKRLRLQQVLGKR-GGRRSDQGFQV				250