Yena Lee
A12967907
BIMM 143 WI19

BIMM 143: The Find-a-Gene Project

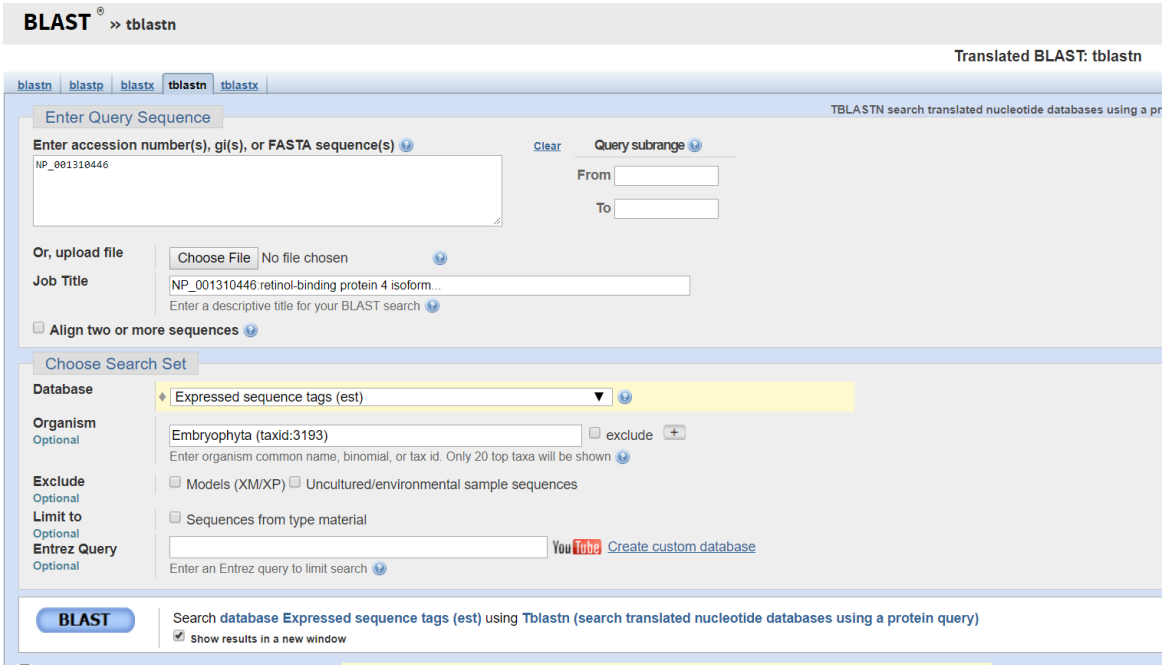
[Q1] Tell me the name of a protein you are interested in. Include the species and the accession number. This can be a human protein or a protein from any other species as long as it's function is known.
If you do not have a favorite protein, select human RBP4 or KIF11. Do not use beta globin as this is in the worked example report that I provide you with online.

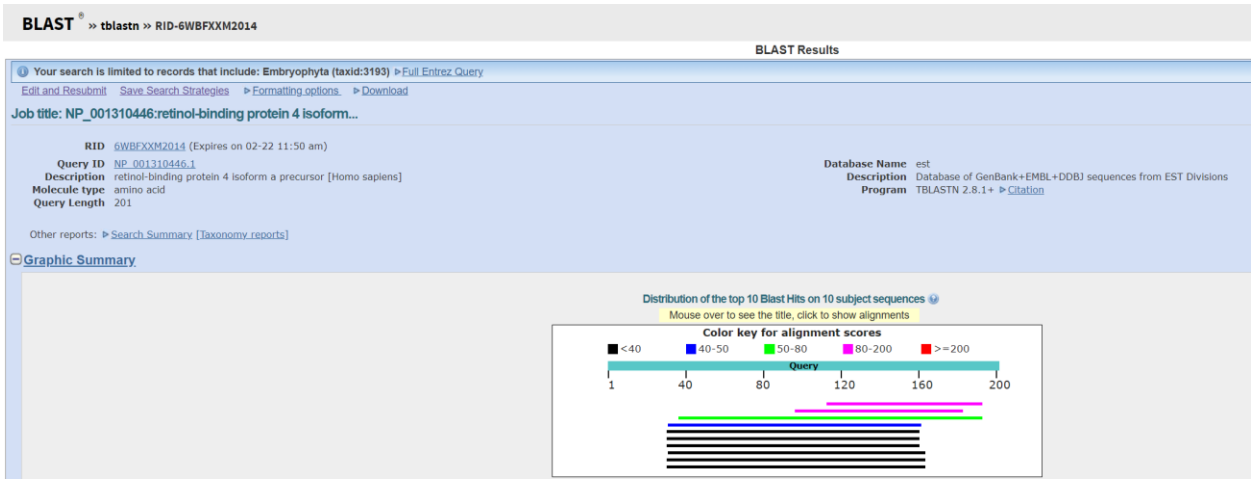
Name of a protein: Retinol Binding Protein 4
Species: Homo Sapiens
Accession number: NP_001310446

[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 bull's 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.

Method: TBLASTN (2.8.1) search against embryophyta ESTs
Organism: Embryophyta (taxid:3193)
Database: Expressed sequence tags (est)
Chosen Match: Accession FR758339.1 a 659 base pair clone from Eulalia clavigera

Yena Lee
A12967907
BIMM 143 WI19





Yena Lee
A12967907
BIMM 143 WI19

Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics

Description	Max score	Total score	Query cover	E value	Ident	Accession
pm_OSU_shoot_MH_2003-04_058D04 Douglas-fir maximum cold hardiness cDNA library 2003-2004 (MH_2003-04) Pseudotsuga menziesii cDNA clone pm_OSU_WP_MD	153	153	39%	6e-45	83.75%	ES425359.1
pm_OSU_shoot_MH_2003-04_036F01 Douglas-fir maximum cold hardiness cDNA library 2003-2004 (MH_2003-04) Pseudotsuga menziesii cDNA clone pm_OSU_WP_MD	151	151	42%	8e-44	78.16%	ES423908.1
FR758337 Eulalia clavigera whole animal adult Eulalia clavigera cDNA clone dmp080P0002I22, mRNA sequence	56.2	56.2	77%	3e-07	27.33%	FR758337.1
BJ868214 Marchantia polymorpha sexual organ E Marchantia polymorpha cDNA clone lwb40k09 5', mRNA sequence	43.9	43.9	64%	0.007	27.48%	BJ868214.1
CCIF25695.b1 CCIF Mimulus guttatus IM62 leaves (H) Erythranthe guttata cDNA clone CCIF25695 5', mRNA sequence	39.3	39.3	64%	0.22	22.30%	GR101146.1
CCIG3156.b1 CCIG Mimulus guttatus DUN10 floral buds (H) Erythranthe guttata cDNA clone CCIG3156 5', mRNA sequence	38.9	38.9	64%	0.29	22.30%	GO978961.1
CCIG3156.g1 CCIG Mimulus guttatus DUN10 floral buds (H) Erythranthe guttata cDNA clone CCIG3156 3', mRNA sequence	38.5	38.5	64%	0.40	21.05%	GO978962.1
CCPX6851.b1_F10.ab1 CCP(UWX) Globe Artichoke Cynara cardunculus var. scolymus cDNA clone CCPX6851, mRNA sequence	38.1	38.1	65%	0.64	27.10%	GE609937.1
CAZI760.fwd CAZI Artemisia annua normalized leaf library Artemisia annua cDNA clone CAZI760 5', mRNA sequence	37.4	37.4	66%	1.3	25.64%	EY111956.1
CJ703317 Y.Ogihara unpublished cDNA library Wh_V4816 Triticum aestivum cDNA clone whv16n13a18 5', mRNA sequence	35.8	35.8	66%	4.5	25.66%	CJ703317.1

Download GenBank Graphics

FR758337 Eulalia clavigera whole animal adult Eulalia clavigera cDNA clone dmp080P0002I22, mRNA sequence

Sequence ID: FR758337.1 **Length:** 659 **Number of Matches:** 1

Range 1: 7 to 477 GenBank Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
56.2 bits(134)	3e-07	Compositional matrix adjust.	44/161(27%)	73/161(45%)	9/161(5%)	+1

```
Query  37   RFSGTWYAMAKKDPEGLFLQD----NIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMV  92
            RF GTWY +    PE  F  +    + V  + +     +  T+ GR    ++   C
Sbjct  7    RFMGTWYELTWI-PENWFPPEFNFQDFVHHYEMSNDTYVHVTSSGRES--SDSPECFYGE  177

Query  93   GTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRLLNLDGTCADSYS  152
                 T+DP KF+        +  ++    D+W++ TDYD Y + + CR  +    C
Sbjct  178  DGLVVTDDPGKFRYYRLLDLTTGERFFSDYWVIVTDYDNYGLVFGCRGRDEADVCIMPDG  357

Query  153  FVFSRDPNGLPPEAQKIVRQRQEELCL-ARQYRLIVHNGYC  192
            +V+SR    L  E Q I+ ++ E+LCL + Q+ +  HN  C
Sbjct  358  WVWSR-TTTLSDEHQAIIDRKIEKLCLTSSQFMMTEHNNPC  477
```

Yena Lee
A12967907
BIMM 143 WI19

[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.

>Eulalia clavigera protein (Sequence translated by EMBOSS Transeq at the EBI).
LERFMGTWYELTWIPENWFPPEFNFQDFVHHYEMSNDTYVHVTSSGRESSDSPECFYGED
GLVVTDDPGKFRYYRLLDLTTGERFFSDYWVIVTDYDNYGLVFGCRGRDEADVCIMPDGW
VWSRTTTLSDEHQAIIDRKIEKLCLTSSQFMMTEHNNPCPLD*NGF*TMVRVPTFETKSR
PLPL*RWISDGPGN*HIVSCHTVRRPYXKEMSTAKYTVYX

Name: Eulalia clavigera retinol binding protein

Species: Eulalia clavigera

Eukaryota; Metazoa; Lophotrochozoa; Annelida; Polychaeta; Palpata;

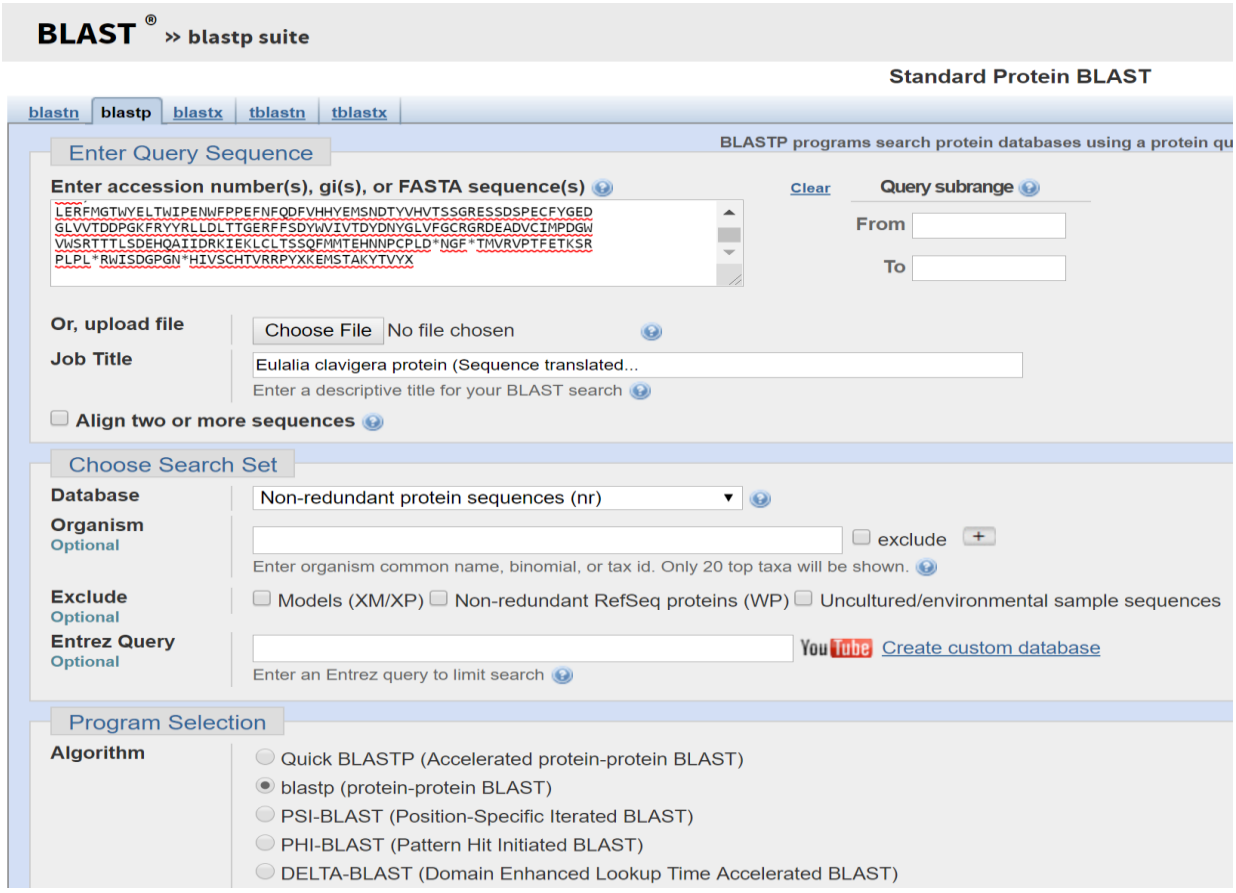
Aciculata; Phyllodocida; Phyllodocidae; Eulalia.

Yena Lee
A12967907
BIMM 143 WI19

[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.**

By using BLASTP 2.8.1 against non-redundant protein sequences database, a protein from Crassostrea gigas is yielded as a top hit result.



Yena Lee
A12967907
BIMM 143 WI19

Job title: Eulalia clavigera protein (Sequence translated...

RID 6WTEXBN7014 (Expires on 02-22 15:48 pm)
Query ID lcl|Query_156722
Description Eulalia clavigera protein (Sequence translated by EMBOSS Transeq at the EBI)
Molecule type amino acid
Query Length 220

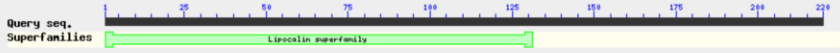
Database Name nr
Description All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Program BLASTP 2.8.1+ ▸Citation

Other reports: ▸Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment] [MSA viewer]

Graphic Summary

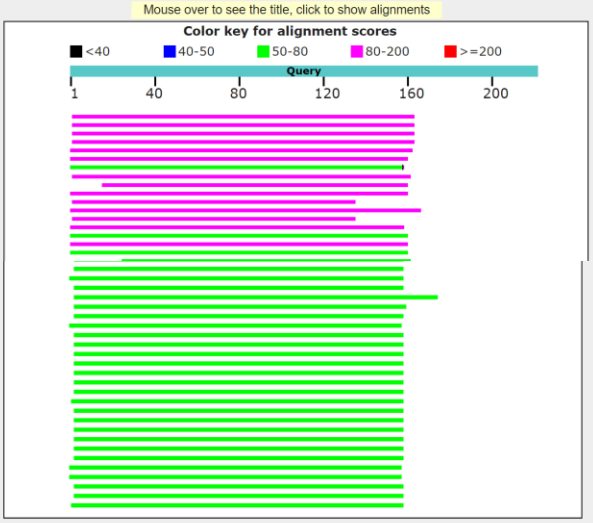
Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.



Distribution of the top 102 Blast Hits on 100 subject sequences

Mouse over to see the title, click to show alignments



Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
PREDICTED: purpurin-like isoform X1 [Crassostrea gigas]	110	110	73%	3e-26	34.16%	XP_019918638.1
PREDICTED: purpurin-like isoform X2 [Crassostrea gigas]	110	110	73%	4e-26	34.16%	XP_019918639.1
retinol-binding protein 4-like isoform X2 [Crassostrea virginica]	109	109	73%	6e-26	31.06%	XP_022334697.1
retinol-binding protein 4-like isoform X1 [Crassostrea virginica]	108	108	73%	8e-26	31.06%	XP_022334696.1
PREDICTED: retinol-binding protein 4-A-like [Crassostrea gigas]	105	105	73%	1e-24	32.30%	XP_019918637.1
apolipoprotein D [Lingula anatina]	104	104	72%	7e-24	35.62%	XP_013416484.1
Plexin-A2 [Crassostrea gigas]	107	242	71%	7e-23	33.97%	EKC20328.1
PREDICTED: retinol-binding protein 4-A-like [Crassostrea gigas]	97.1	97.1	72%	2e-21	33.33%	XP_011413056.1
PREDICTED: lazarillo protein-like [Crassostrea gigas]	93.6	93.6	65%	2e-20	35.42%	XP_011413054.1
PREDICTED: retinol-binding protein 4-A-like [Crassostrea gigas]	94.7	94.7	72%	2e-20	30.82%	XP_011413072.2

Yena Lee
A12967907
BIMM 143 WI19

Alignments

Download GenPept Graphics

PREDICTED: purpurin-like isoform X1 [Crassostrea gigas]

Sequence ID: XP_019918638.1 **Length:** 231 **Number of Matches:** 1

Range 1: 44 to 199 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
110 bits(276)	3e-26	Compositional matrix adjust.	55/161(34%)	87/161(54%)	5/161(3%)

```
Query  2    ERFMGTWYELTWIPENWFPPEFNFQDFVHHYEMSNDTYVHVTSSGRESSDSPECFYGEDG  61
            ++++G WYE+ W   E +F      FQD+ H Y       + V  +GR+  +  +CF  +
Sbjct  44   DKYLGKWYEMKWYSEVYFDESELFQDYTHEYIRKKGGNLTVLHTGRDPINLVDCFQRQST  103

Query  62   LVVTDDPGKFRYYRLLDLTTGERFFSDYWVIVTDYDNYGLVFGCRGRDEADVCIMPDGWV  121
            L +T+ PGKF    ++D    +   SD+ VI TDY NY + +GC  + +   C+    WV
Sbjct  104  LYLTETPGKF----MID-EKNQGNLSDFLVIRTDYSNYSVAYGCTTQQQDGTCLKARAWV  158

Query  122  WSRTTTLSDEHQAIIDRKIEKLCLTSSQFMMTEHNNPCPLD  162
             +SR TTL+D+     D ++EKLCL  + F++T   N C  D
Sbjct  159  FSRKTTLADDLSQEADDQLEKLCLNLTSFLVTRQTNDCTDD  199
```

Download GenPept Graphics

PREDICTED: purpurin-like isoform X2 [Crassostrea gigas]

Sequence ID: XP_019918639.1 **Length:** 219 **Number of Matches:** 1

Range 1: 32 to 187 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
110 bits(274)	4e-26	Compositional matrix adjust.	55/161(34%)	87/161(54%)	5/161(3%)

```
Query  2    ERFMGTWYELTWIPENWFPPEFNFQDFVHHYEMSNDTYVHVTSSGRESSDSPECFYGEDG  61
            ++++G WYE+ W   E +F      FQD+ H Y       + V  +GR+  +  +CF  +
Sbjct  32   DKYLGKWYEMKWYSEVYFDESELFQDYTHEYIRKKGGNLTVLHTGRDPINLVDCFQRQST  91

Query  62   LVVTDDPGKFRYYRLLDLTTGERFFSDYWVIVTDYDNYGLVFGCRGRDEADVCIMPDGWV  121
            L +T+ PGKF    ++D    +   SD+ VI TDY NY + +GC  + +   C+    WV
Sbjct  92   LYLTETPGKF----MID-EKNQGNLSDFLVIRTDYSNYSVAYGCTTQQQDGTCLKARAWV  146

Query  122  WSRTTTLSDEHQAIIDRKIEKLCLTSSQFMMTEHNNPCPLD  162
             +SR TTL+D+     D ++EKLCL  + F++T   N C  D
Sbjct  147  FSRKTTLADDLSQEADDQLEKLCLNLTSFLVTRQTNDCTDD  187
```