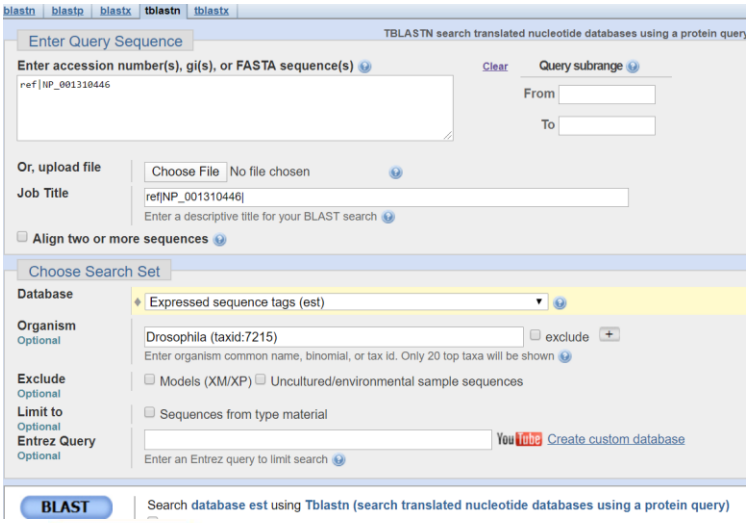
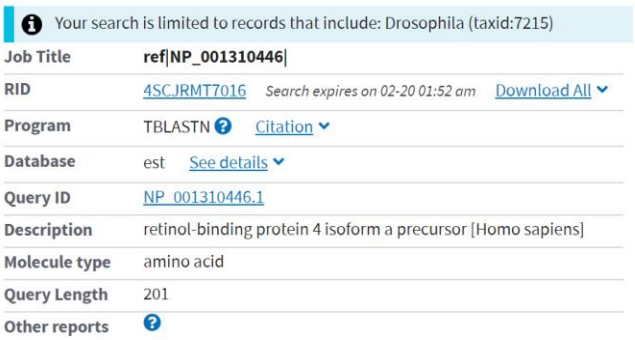
Alberto Carreno
alcarren@ucsd.edu
A10216025

Questions: [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

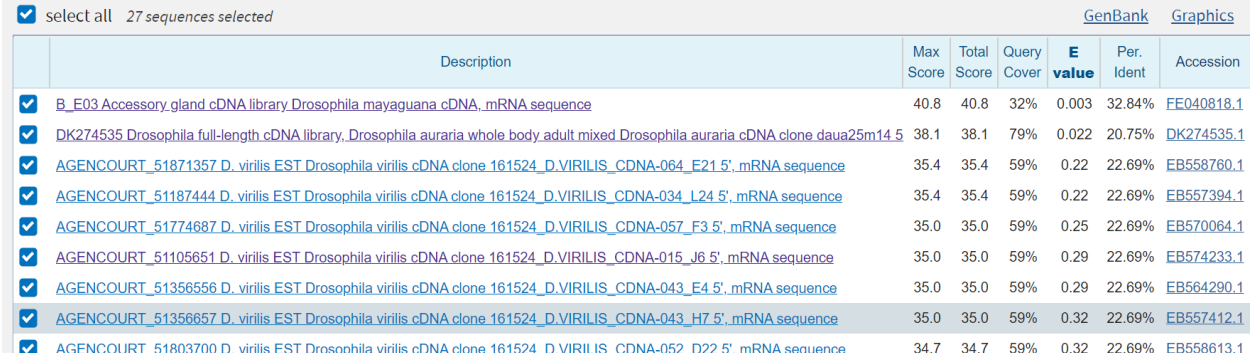
Name: RBP4
Accession: NP_001310446
Species: Homo Sapiens



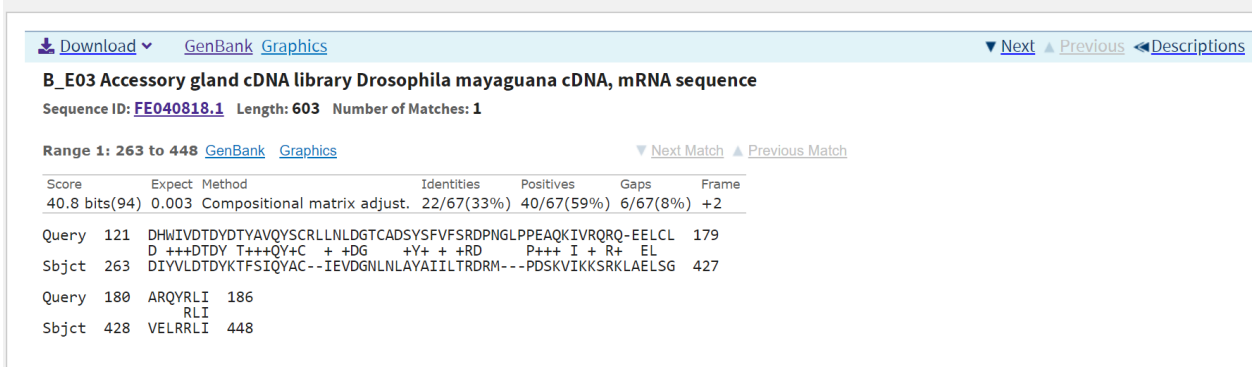


[Q2] Perform a BLAST search against a DNA database, such as a database consisting of genomic DNA or ESTs. The BLAST server can be at NCBI or elsewhere. Include details of the BLAST method used, database searched and any limits applied (e.g. Organism)

Method: TBLASTN search against drosophila.
Database: Expressed Sequence Tag
Organism: (Drosophila Taxid:7215)



Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
B_E03 Accessory gland cDNA library Drosophila mayaguana cDNA, mRNA sequence	40.8	40.8	32%	0.003	32.84%	FE040818.1
DK274535 Drosophila full-length cDNA library, Drosophila auraria whole body adult mixed Drosophila auraria cDNA clone daua25m14 5	38.1	38.1	79%	0.022	20.75%	DK274535.1
AGENCOURT_51871357 D. virilis EST Drosophila virilis cDNA clone 161524_D.VIRILIS_CDNA-064_E21 5', mRNA sequence	35.4	35.4	59%	0.22	22.69%	EB558760.1
AGENCOURT_51187444 D. virilis EST Drosophila virilis cDNA clone 161524_D.VIRILIS_CDNA-034_L24 5', mRNA sequence	35.4	35.4	59%	0.22	22.69%	EB557394.1
AGENCOURT_51774687 D. virilis EST Drosophila virilis cDNA clone 161524_D.VIRILIS_CDNA-057_F3 5', mRNA sequence	35.0	35.0	59%	0.25	22.69%	EB570064.1
AGENCOURT_51105651 D. virilis EST Drosophila virilis cDNA clone 161524_D.VIRILIS_CDNA-015_J6 5', mRNA sequence	35.0	35.0	59%	0.29	22.69%	EB574233.1
AGENCOURT_51356556 D. virilis EST Drosophila virilis cDNA clone 161524_D.VIRILIS_CDNA-043_E4 5', mRNA sequence	35.0	35.0	59%	0.29	22.69%	EB564290.1
AGENCOURT_51356657 D. virilis EST Drosophila virilis cDNA clone 161524_D.VIRILIS_CDNA-043_H7 5', mRNA sequence	35.0	35.0	59%	0.32	22.69%	EB557412.1
AGENCOURT_51803700 D. virilis EST Drosophila virilis cDNA clone 161524_D.VIRILIS_CDNA-052_D22 5', mRNA sequence	34.7	34.7	59%	0.32	22.69%	EB558613.1



Download GenBank Graphics ▼ Next ▲ Previous ◀◀ Descriptions

B_E03 Accessory gland cDNA library Drosophila mayaguana cDNA, mRNA sequence

Sequence ID: FE040818.1 **Length:** 603 **Number of Matches:** 1

Range 1: 263 to 448 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
40.8 bits(94)	0.003	Compositional matrix adjust.	22/67(33%)	40/67(59%)	6/67(8%)	+2

```
Query  121  DHWIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQ-EELCL  179
            D +++DTDY T+++QY+C  + +DG    +Y+ + +RD     P+++ I + R+  EL
Sbjct  263  DIYVLDTDYKTFSIQYAC--IEVDGNLNLAYAIILTRDRM---PDSKVIKKSRKLAELSG  427

Query  180  ARQYRLI  186
               RLI
Sbjct  428  VELRRLI  448
```

```
>FE040818.1_1 B_E03 Accessory gland cDNA library Drosophila mayaguana cDNA, mRNA sequence
HNIRWKLEYEKVSGYMVSTIVVSTLSSSFARMCKV*YC*GQKW*L*NSQIGS*RQEWSFA
SANFIDNKSG*KGWQIRFKNEKFT*RP*YICAGHRL*NLFNSICVHRSGWQSKLSVCHHI
DAGPNAGL*SD*EVTKIGRTLWS*AKKVDTHFTEELP*GCLKVVGLNMLH*VFLTKDSSL
FLFESY*NK*IRLSYQLAELV
>FE040818.1_2 B_E03 Accessory gland cDNA library Drosophila mayaguana cDNA, mRNA sequence
ITSVGNLNMKKYQGIWYPQLSYPLYLHPLPECAKFNIVKGKNGNYEIHRSDLDGKSGLLR
QRTSLITKVDKKGGKYALKMKNSHDGLDIYVLDTDYKTFSIQYACIEVDGNLNLAYAIIL
TRDRMPDSKVIKKSRKLAELSGVELRRLIPISQKSCPKDA*KLLGLICFIRYS*QKIQVC
FYSSLIRINKSDFLIN*LS*X
>FE040818.1_3 B_E03 Accessory gland cDNA library Drosophila mayaguana cDNA, mRNA sequence
*HPLET*I*KSIRVYGIHNCRIHFIFILCPNVQSLILLRAKMVIMKFTDRILTARVVFCV
SELH**QKWIKRVANTL*K*KIHMTALIYMCWTPTIKPFQFNMRA*KWMAI*T*RMPSY*
RGTECRTLK*LRSHENWPNSLELS*EG*YPFHRRAALRMPKSCWA*YASLGILNKRFKFV
SIRVLLE*INQTFLSIS*ASX
>FE040818.1_4 B_E03 Accessory gland cDNA library Drosophila mayaguana cDNA, mRNA sequence
N*LS*LIRKSDLFILIRLE*KQT*IFC*EYLMKHIKPNNF*ASLGQLFCEMGINLLSSTP
ESSANFRDFLITLESGIRSRVNMMAYAKFRLPSTSMHAY*IEKVL*SVSSTYISRPSCEF
FIFKAYLPPFLSTFVINEVR*RKRPLLPSRSDL*IS*LPFLPLTILNFAHSGKG*R*SGY
DNCGYHIP*YFFIFKFPTDVM
>FE040818.1_5 B_E03 Accessory gland cDNA library Drosophila mayaguana cDNA, mRNA sequence
LAQLIDKKV*FIYSNKTRIETNLNLLLRIPNEAY*AQQLLGILRAALL*NGYQPS*LNSR
EFGQFS*LLNHFRVRHSVPRQYDGIR*V*IAIHFYARILN*KGFIVGVQHIYIKAVM*IF
HF*SVFATLFIHFCYQ*SSLTQKTTLAVKIRSVNFIITIFALNNIKLCTFGQRMKIKWIR
QLWIPYTLILFHIQVSNGCYX
>FE040818.1_6 B_E03 Accessory gland cDNA library Drosophila mayaguana cDNA, mRNA sequence
TSSAN**ESLIYLF**DSNRNKLESFVKNT**SILSPTTFRHP*GSSSVKWVSTFLAQLQ
RVRPIFVTS*SL*SPAFGPASI*WHTLSLDCHPLLCTHIELKRFYSRCPAHIYQGRHVNF
SFLKRICHPFYPLLLSMKFADAKDHSCRQDPICEFHNYHFCP*QY*TLHIRAKDEDKVDT
TIVDTIYPDTFSYSSFQRMLX
```

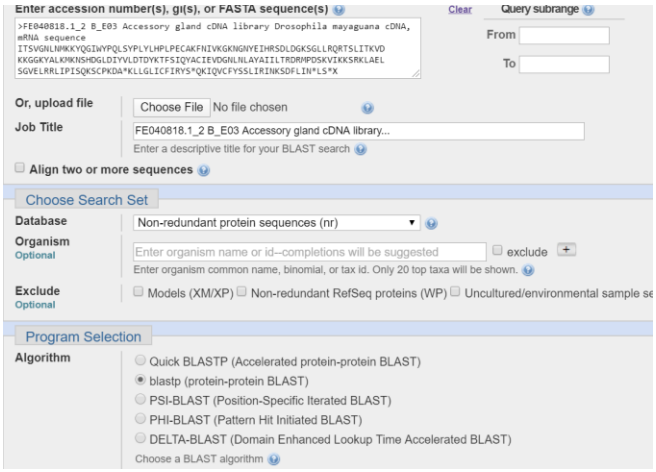
[Q3] Gather information about this “novel” protein. At a minimum, show me the protein sequence of the “novel” protein as displayed in your BLAST results from [Q2] as FASTA format (you can copy and paste the aligned sequence subject lines from your BLAST result page if necessary) or translate your novel 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.

```
>FE040818.1_2 B_E03 Accessory gland cDNA library Drosophila mayaguana cDNA, mRNA sequence
ITSVGNLNMKKYQGIWYPQLSYPLYLHPLPECAKFNIVKGKNGNYEIHRSDLDGKSGLLR
QRTSLITKVDKKGGKYALKMKNSHDGLDIYVLDTDYKTFSIQYACIEVDGNLNLAYAIIL
TRDRMPDSKVIKKSRKLAELSGVELRRLIPISQKSCPKDA*KLLGLICFIRYS*QKIQVC
FYSSLIRINKSDFLIN*LS*X
```

Taxonomy:Eukaryota;Metazoa,Ecdysozoa;Arthropoda;Hexapoda;Insecta;Pterygote;Neoptera; Holometabola;Diptera;brachycera;Muscomorpha;Ephydroidea;Drosophilidae; Drosophila; Mayaguana

[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.

Ran protein-protein BLAST of the sequence above. Top result was for D. *mojavensis*



Job Title	**FE040818.1_2 B_E03 Accessory gland cDNA library...**
RID	4SCR6MD6016 *Search expires on 02-20 01:54 am* Download All
Program	BLASTP Citation
Database	nr See details
Query ID	lcl|Query_55160
Description	FE040818.1_2 B_E03 Accessory gland cDNA library Drosophil ...
Molecule type	amino acid
Query Length	201
Other reports	Distance tree of results Multiple alignment MSA viewer

select all *100 sequences selected* GenPept Graphics Distance tree of results Multiple alignment

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
uncharacterized protein Dmoj_GI26513 [Drosophila mojavensis]	216	216	79%	3e-68	57.50%	XP_015020355.1
PREDICTED: apolipoprotein D isoform X1 [Drosophila arizonae]	215	215	79%	1e-67	57.50%	XP_017857158.1
PREDICTED: apolipoprotein D isoform X2 [Drosophila arizonae]	215	215	79%	1e-67	57.50%	XP_017857159.1
hypothetical protein AWZ03_004095 [Drosophila navojoa]	205	205	79%	8e-64	56.25%	TDG49412.1
uncharacterized protein LOC115562160 [Drosophila navojoa]	205	205	79%	9e-64	56.25%	XP_030238856.1
lazarillo protein isoform X1 [Drosophila hydei]	144	144	77%	8e-40	43.59%	XP_023165552.1
lazarillo protein isoform X2 [Drosophila hydei]	144	144	77%	1e-39	43.59%	XP_023165553.1
blast:Apolipoprotein D [Drosophila guanche]	115	115	79%	1e-28	35.80%	SPP79949.1
PREDICTED: lopap [Drosophila miranda]	113	113	79%	8e-28	37.65%	XP_017151350.1
GL14848 [Drosophila persimilis]	112	112	79%	9e-28	37.04%	EDW39695.1
lopap [Drosophila persimilis]	113	113	79%	1e-27	37.04%	XP_002020702.2