BIMM-143: INTRODUCTION TO BIOINFORMATICS

The Find-a-Gene Project Assignment
http://thegrantlab.org/bimm143
Dr. Barry Grant

Overview:

The find-a-gene project is a required assignment for BIMM-143. You should prepare a written report in **PDF** format that has responses to each question labeled **[Q1] - [Q10]** below. You may wish to consult the scoring rubric at the end of this document and the example report provided online (note that the example report is from a previous quarter and the questions may differ).

The objective with this assignment is for you to demonstrate your grasp of database searching, sequence analysis, structure analysis and the R environment that we have covered in class.

Due Date:

Your responses to questions Q1-Q4 are due at 12pm on the **Monday of Week 5** (see the Assignments and Grading section of our website for details). Note that these first set of answers can be obtained very quickly (at best within 15 or 20 minutes), so if you don't succeed at first, just keep trying.

The complete assignment, including responses to all questions, is due at 12pm on the **Monday of Week 10**.

Submission Instructions:

Your report formatted as a **PDF document** should be uploaded to ***GradeScope***. Please make sure to include your UCSD email and PID number on the first page.

Be sure to include your UCSD email and PID number on the first page of your report.

Submit your preliminary report with answers to Q1-Q4 as soon as you can so we can determine if you have found a novel gene. Submit this preliminary report as one document with screen shots of the results inserted appropriately.

See the demonstration report linked to on the course website for an example of format. I will email you my decision; proceed with subsequent questions only after we are sure you have found a novel gene (and thus be successful in the later stages of the project).

For the final report add your results for Q5-Q10 to the preliminary report and submit the final document containing your results for all questions - **Please do not send only Q5-Q10 answers as the final report**.

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**: Titin
- **Accession Number**: Q8WZ42
- **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).

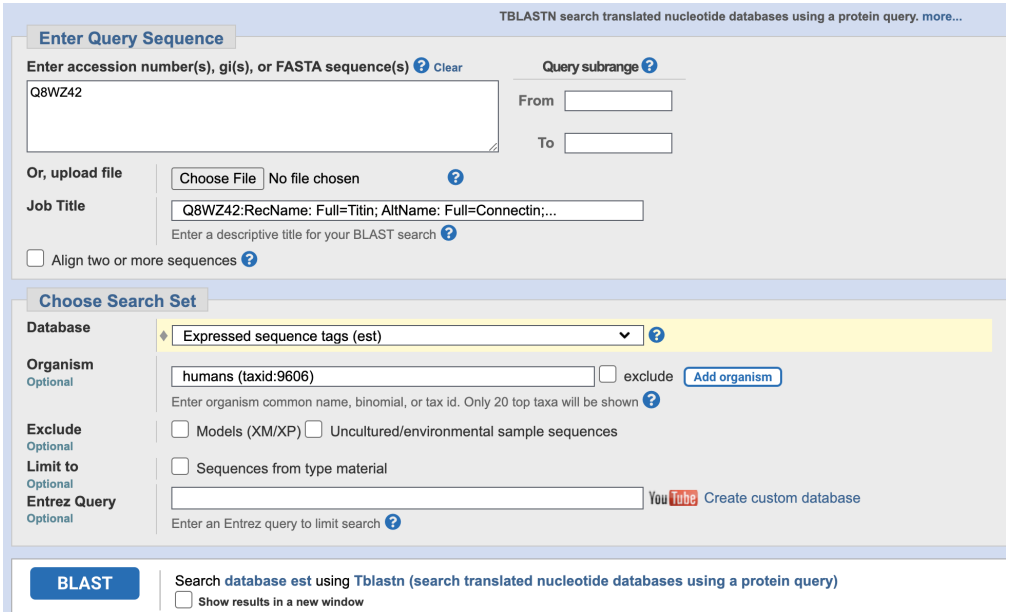
Also include the output of that BLAST search in your document. If appropriate, change the font to `Courier size 10` so that the results are displayed neatly. You can also screen capture a BLAST output (e.g. alt print screen on a PC or on a MAC press ⌘-shift-4. The pointer becomes a bulls eye. Select 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.

- **Method**: TBLASTN search against Human ESTs
- **Database**: Expressed Sequence Tags (est)
- **Organism**: Humans (Taxid: 9606)



- **Chosen Match**: Homo sapiens cDNA clone IMAGE:6143622 5', mRNA sequence, Accession Number BU170890.1

AGENCOURT_7940143 NIH_MGC_67 Homo sapiens cDNA clone IMAGE:6143622 5', mRNA sequence

Sequence ID: BU170890.1 **Length:** 872 **Number of Matches:** 12

Range 1: 2 to 850 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
548 bits(1413)	3e-175	Compositional matrix adjust.	276/283(98%)	280/283(98%)	0/283(0%)	+2

```
Query  33773  EPISSKPVIVTGLQDTTVSSDSVAKFAVKATGEPRPTAIWTKDGKAITQGGKYKLSEDKG  33832
              EPISSKPVIVTGLQDTTVSSDSVAKFAVKATGEPRPTAIWTKDGKAITQGGKYKLSEDKG
Sbjct  2      EPISSKPVIVTGLQDTTVSSDSVAKFAVKATGEPRPTAIWTKDGKAITQGGKYKLSEDKG  181

Query  33833  GFFLEIHKTDTSDSGLYTCTVKNSAGSVSSSCKLTIKAIKDTEAQKVSTQKTSEITPQKK  33892
              GFFLEIHKTDTSDSGLYTCTVKNSAGSVSSSCKLTIKAIKDTEAQKVSTQKTSEITPQKK
Sbjct  182    GFFLEIHKTDTSDSGLYTCTVKNSAGSVSSSCKLTIKAIKDTEAQKVSTQKTSEITPQKK  361

Query  33893  AVVQEEISQKALRSEEIKMSEAKSQEKLALKEEASKVLISEEVKKSAATSLEKSIVHEEI  33952
              AVVQEEISQKALRSEEIKMSEAKSQEKLALKEEASKVLISEEVKKSAATSLEKSIVHEEI
Sbjct  362    AVVQEEISQKALRSEEIKMSEAKSQEKLALKEEASKVLISEEVKKSAATSLEKSIVHEEI  541

Query  33953  TKTSQASEEVRTHAEIKAFSTQMSINEGQRLVLKANIAGATDVKWVLNGVELTNSEEYRY  34012
              TKTSQASEEVRTHAEIKAFSTQMSINEGQRLVLKANIAGATDVKWVLNGVELTNSEEYRY
Sbjct  542    TKTSQASEEVRTHAEIKAFSTQMSINEGQRLVLKANIAGATDVKWVLNGVELTNSEEYRY  721

Query  34013  GVSGSDQTLTIKQASHRDEGILTCISKTKEGIVKCQYDLTLSK  34055
              GVSGSDQTLT +QASHRDEGILTCISK +EGIV+CQYDLTL +
Sbjct  722    GVSGSDQTLTHQQASHRDEGILTCISKNQEGIVQCQYDLTLXQ  850
```

Range 2: 14 to 814 GenBank Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
63.2 bits(152)	9e-07	Compositional matrix adjust.	66/274(24%)	110/274(40%)	58/274(21%)	+2

```
Query  941   TPPTLVSGLKNVTVIEGESVTLECHISGYPSPTVTWYREDYQIESSIDFQITFQSGIARL  1000
              + P +V+GL++ TV             +G P PT  W ++   I     ++++   G   L
Sbjct  14    SKPVIVTGLQDTTVSSDSVAKFAVKATGEPRPTAIWTKDGKAITQGGKYKLSEDKGGFFL  193

Query  1001  MIREAFAEDSGRFTCSAVNEAGTVSTSCYLAVQV--SEEFEKETTAVTEKFTTEEKRFVE  1058
               I +    DSG +TC+  N AG+VS+SC L ++     E +K +T  T + T ++K  V+
Sbjct  194   EIHKTDTSDSGLYTCTVKNSAGSVSSSCKLTIKAIKDTEAQKVSTQKTSEITPQKKAVVQ  373

Query  1059  ---------SRDVVMTDTSLTEEQAGPGEPAAPYFIT----------KPVVQKLV---  1094
                     S ++ M++   ++E+    E A+   I+           K +V + +
Sbjct  374   EEISQKALRSEEIKMSEAK-SQEKLALKEEASKVLISEEVKKSAATSLEKSIVHEEITKT  550

Query  1095  ----------------------EGGSVVFGCQVGGNPKPHVYWKKSGVPLTTG--YRYK  1129
                                   EG  +V    + G     V W  +GV LT    YRY
Sbjct  551   SQASEEVRTHAEIKAFSTQMSINEGQRLVLKANIAG--ATDVKWVLNGVELTNSEEYRYG  724

Query  1130  VSYNKQTGECKLVISMTFADDAGEYTIVVRNKHG  1163
              VS + QT    L        D G  T + +N+ G
Sbjct  725   VSGSDQT----LTHQQASHRDEGILTCISKNQEG  814
```

Range 3: 110 to 826 GenBank Graphics ▼ Next Match ▲ Previous Match First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
57.8 bits(138)	6e-05	Compositional matrix adjust.	62/257(24%)	102/257(39%)	34/257(13%)	+2

```
Query  6194  SAQWFKDGKEISTSAKYRLVCHERSVSLEVNNLELEDTANYTCKVSNVAGDDACSGILTV  6253
             +A W KDGK I+   KY+L   +    LE++  +  D+  YTC V N AG  + S  LT+
Sbjct  110   TAIWTKDGKAITQGGKYKLSEDKGGFFLEIHKTDTSDSGLYTCTVKNSAGSVSSSCKLTI  289

Query  6254  KEPPSFLVKPGRQQAIPDSTVEFKAILKG--------TPPFKIKWFKDDVELV---SGPK  6302
             K       +    Q   + T + KA+++          +   K+   K   +L        K
Sbjct  290   KAIKDTEAQKVSTQKTSEITPQKKAVVQEEISQKALRSEEIKMSEAKSQEKLALKEEASK  469

Query  6303  CFIGLE----GSTSF-LNLYSVDASKTGQYTCHVTNDVGSDSCTTMLLVTEPPKFVKKLE  6357
               I  E     +TS   ++   + +KT Q +  V       + +T + + E  + V
Sbjct  470   VLISEEVKKSAATSLEKSIVHEEITKTSQASEEVRTHAEIKAFSTQMSINEGQRLV----  637

Query  6358  ASKIVKAGDSSRLECKIAGSPEIRVVWFRNEHELPASDKYRMTFIDSVAVIQMNNLSTED  6417
                        L+  IAG+ +++  W  N  EL  S++YR     S   +     S  D
Sbjct  638   ------------LKANIAGATDVK--WVLNGVELTNSEEYRYGVSGSDQTLTHQQASHRD  775

Query  6418  SGDFICEAQNPAGSTSC  6434
               G   C ++N  G   C
Sbjct  776   EGILTCISKNQEGIVQC  826
```

Range 4: 17 to 814 GenBank Graphics ▼ Next Match ▲ Previous Match First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
57.0 bits(136)	9e-05	Compositional matrix adjust.	61/272(22%)	108/272(39%)	48/272(17%)	+2

```
Query  1555  KPMFVEKLKNVNIKEGSRLEMKVRATGNPNPDIVWLKNSDIIVPHKYPKIRIEGTKGEAA  1614
             KP+ V  L++  +   S  +  V+ATG P P  +W K+   I   +  K ++   KG
Sbjct  17    KPVIVTGLQDTTVSSDSVAKFAVKATGEPRPTAIWTKDGKAIT--QGGKYKLSEDKGGFF  190

Query  1615  LKIDSTVSQDSAWYTATAINKAGRDTTRCKVNVEV----------EFAEPEPERKLI-  1661
             L+I  T + DS  YT T  N AG  ++ CK+ ++              + +E  P++K +
Sbjct  191   LEIHKTDTSDSGLYTCTVKNSAGSVSSSCKLTIKAIKDTEAQKVSTQKTSEITPQKKAVV  370

Query  1662  ---IPRGTYRAKEIAAPEL---EPLHL--------------RYGQEQWEEGDLYDkekqq  1701
                 I +   R++EI   E    E L L               +      E+  ++++  +
Sbjct  371   QEEISQKALRSEEIKMSEAKSQEKLALKEEASKVLISEEVKKSAATSLEKSIVHEEITKT  550

Query  1702  kpffkkkLTSLRLKRFGPAH---------FECRLTPIGDPTMVVEWLHDGKPLEAANRLR  1752
                ++  T   +K F               +  +    D    V+W+ +G  L  +  R
Sbjct  551   SQASEEVRTHAEIKAFSTQMSINEGQRLVLKANIAGATD----VKWVLNGVELTNSEEYR  718

Query  1753  MINEFGYCSLDYGVAYSRDSGIITCRATNKYG  1784
                      +L +  A  RD GI+TC + N+ G
Sbjct  719   YGVSGSDQTLTHQQASHRDEGILTCISKNQEG  814
```

Range 5: 20 to 295 GenBank Graphics ▼ Next Match ▲ Previous Match First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
55.8 bits(133)	2e-04	Compositional matrix adjust.	30/92(33%)	44/92(47%)	0/92(0%)	+2

```
Query  31460  PGIRKEMKDVTTKLGEAAQLSCQIVGRPLPDIKWYRFGKELIQSRKYKMSSDGRTHTLTV  31519
              P I   ++D T      A+ + +  G P P   W + GK + Q  KYK+S D     L +
Sbjct  20     PVIVTGLQDTTVSSDSVAKFAVKATGEPRPTAIWTKDGKAITQGGKYKLSEDKGGFFLEI  199

Query  31520  MTEEQEDEGVYTCIATNEVGEVETSSKLLLQA  31551
                +  D G+YTC   N  G V +S KL ++A
Sbjct  200    HKTDTSDSGLYTCTVKNSAGSVSSSCKLTIKA  295
```

Range 6: 53 to 292 GenBank Graphics ▼ Next Match ▲ Previous Match First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
55.1 bits(131)	4e-04	Compositional matrix adjust.	29/80(36%)	39/80(48%)	0/80(0%)	+2

```
Query  8244  VKQDEFTRYECKIGGSPEIKVLWYKDETEIQESSKFRMSFVDSVAVLEMHNLSVEDSGDY  8303
             V  D   ++  K  G P    +W KD   I +  K+++S       LE+H     DSG Y
Sbjct  53    VSSDSVAKFAVKATGEPRPTAIWTKDGKAITQGGKYKLSEDKGGFFLEIHKTDTSDSGLY  232

Query  8304  TCEAHNAAGSASSSTSLKVK  8323
             TC   N+AGS SSS  L +K
Sbjct  233   TCTVKNSAGSVSSSCKLTIK  292
```

Range 7: 71 to 814 GenBank Graphics ▼ Next Match ▲ Previous Match First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
51.6 bits(122)	0.005	Compositional matrix adjust.	63/268(24%)	108/268(40%)	32/268(11%)	+2

```
Query  9097  ADFECHVTGTQPIKVSWAKDSREIRSGGKYQISYLENSAHLTVLKVDKGDSGQYTCYAVN  9156
             A F    TG       W KD + I  GGKY++S  +    L + K D  DSG YTC   N
Sbjct  71    AKFAVKATGEPRPTAIWTKDGKAITQGGKYKLSEDKGGFFLEIHKTDTSDSGLYTCTVKN  250

Query  9157  EVGKDSCTAQLNIKERLIPPSFTKRLSETVEETEGNSFKLEGRVAGSQPITVAWYKNNIE  9216
               G  S + +L IK                +++TE      + + +   P   A  +  I
Sbjct  251   SAGSVSSSCKLTIK--------------AIKDTEAQKVSTQ-KTSEITPQKKAVVQEEIS  385

Query  9217  IQPTSNCEITFKNNTLVLQVRKAGMNDAG--LYTCKVSNDAGSALCTSSIVIKEPKKPPV  9274
              +   + EI  K +    Q + A   +A   L + +V   A ++L   SIV +E  K
Sbjct  386   QKALRSEEI--KMSEAKSQEKLALKEEASKVLISEEVKKSAATSL-EKSIVHEEITKTSQ  556

Query  9275  FDQHL----------TPVTVSEGEYVQLSCHVQGSEPIRIQWLKAGREIKPSDRCSFSFA  9324
               + +          T ++++EG+ + L  ++ G+  ++  W+  G E+  S+   +  +
Sbjct  557   ASEEVRTHAEIKAFSTQMSINEGQRLVLKANIAGATDVK--WVLNGVELTNSEEYRYGVS  730

Query  9325  SGTAVLELRDVAKADSGDYVCKASNVAG  9352
                   L  +  +  D G   C + N  G
Sbjct  731   GSDQTLTHQQASHRDEGILTCISKNQEG  814
```

Range 8: 23 to 448 GenBank Graphics ▼ Next Match ▲ Previous Match First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
50.8 bits(120)	0.010	Compositional matrix adjust.	44/145(30%)	63/145(43%)	7/145(4%)	+2

```
Query  9382  FFVSEPQSIRVVEKTTATFIAKVGGDPIPNVKWTKGKWRQLNQGGRVFIHQKGDEAKLEI  9441
               V+  Q   V   + A F  K  G+P P   WTK   + + QGG+  + +      LEI
Sbjct  23    VIVTGLQDTTVSSDSVAKFAVKATGEPRPTAIWTKD-GKAITQGGKYKLSEDKGGFFLEI  199

Query  9442  RDTTKTDSGLYRCVAFNEHGEIESNVNLQVDERKKQEKIEGDLRAMLKKTPILKKGAGEE  9501
               T  +DSGLY C   N  G + S+  L +   K  E  +   +  + TP  +K A  +
Sbjct  200   HKTDTSDSGLYTCTVKNSAGSVSSSCKLTIKAIKDTEAQKVSTQKTSEITP--QKKAVVQ  373

Query  9502  EEIDIMEL----LKNVDPKEYEKYA  9522
             EEI    L    +K  + K  EK A
Sbjct  374   EEISQKALRSEEIKMSEAKSQEKLA  448
```

Range 9: 119 to 826 GenBank Graphics ▼ Next Match ▲ Previous Match First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
50.4 bits(119)	0.012	Compositional matrix adjust.	61/251(24%)	95/251(37%)	22/251(8%)	+2

```
Query  9017  WYKDGKPLKDSPNVQTSFLDNTATLNIFKTDRSLAGQYSCTATNPIGSASSSARLILTEG  9076
             W KDGK +      + S       L I KTD S +G Y+CT  N  GS SSS +L +
Sbjct  119   WTKDGKAITQGGKYKLSEDKGGFFLEIHKTDTSDSGLYTCTVKNSAGSVSSSCKLTIKAI  298

Query  9077  KNPPFFDIRLAPVD-------AVVGESADFECHVTGTQPIKVSWAKDSREIRSGGKYQIS  9129
             K+     +             AVV E  +        ++ IK+S AK   ++
Sbjct  299   KDTEAQKVSTQKTSEITPQKKAVVQE--EISQKALRSEEIKMSEAKSQEKLALKE-----  457

Query  9130  YLENSAHLTVLKVDKGDSGQYTCYAVNEVGKDSCTAQLNIKERLIPPSFTKRLSETVEET  9189
               E S  L   +V K  +       V+E    +  A   ++      +F+ ++S
Sbjct  458   --EASKVLISEEVKKSAATSLEKSIVHEEITKTSQASEEVRTHAEIKAFSTQMS----IN  619

Query  9190  EGNSFKLEGRVAGSQPITVAWYKNNIEIQPTSNCEITFKNNTLVLQVRKAGMNDAGLYTC  9249
             EG    L+  +AG+    V W  N +E+  +          +   L  ++A   D G+ TC
Sbjct  620   EGQRLVLKANIAGATD--VKWVLNGVELTNSEEYRYGVSGSDQTLTHQQASHRDEGILTC  793

Query  9250  KVSNDAGSALC  9260
               N  G   C
Sbjct  794   ISKNQEGIVQC  826
```

Range 10: 20 to 850 GenBank Graphics ▼ Next Match ▲ Previous Match First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
49.3 bits(116)	0.026	Compositional matrix adjust.	64/287(22%)	110/287(38%)	18/287(6%)	+2

```
Query  4478  PTFLSRPKSLTTFVGKAAKFICTVTGTPVIETIWQKDGAALSPSPNWRISDAENKHILEL  4537
             P  ++  +  T      AKF    TG P    IW KDG A++    +++S+ +    LE+
Sbjct  20    PVIVTGLQDTTVSSDSVAKFAVKATGEPRPTAIWTKDGKAITQGGKYKLSEDKGGFFLEI  199

Query  4538  SNLTIQDRGVYSCKASNKFGADICQAELIIIDKPHFIKELEPVQSAINKKVHLECQ----  4593
                 D G+Y+C   N  G+     +L I      IK+ E  + +  K   +  Q
Sbjct  200   HKTDTSDSGLYTCTVKNSAGSVSSSCKLTI----KAIKDTEAQKVSTQKTSEITPQKKAV  367

Query  4594  VDEDRKVTVTWSKDGQKLPPGKDYKICFEDKIATLEIPLAKLKDSGTYV-CTASNEAGSS  4652
             V E+       S++ +        K+  +++ + + I     K + T +  +  +E  +
Sbjct  368   VQEEISQKALRSEEIKMSEAKSQEKLALKEEASKVLISEEVKKSAATSLEKSIVHEEITK  547

Query  4653  SCSATVTVR---EPPSFVKKVDPSYLMLPGESARLHCKLKGSPVIQVTWFKNNKELSESN  4709
             +  A+  VR   E  +F      S  M   E  RL  K   +    V W  N  EL+ S
Sbjct  548   TSQASEEVRTHAEIKAF------STQMSINEGQRLVLKANIAGATDVKWVLNGVELTNSE  709

Query  4710  TVRMYFVNSEAILDITDVKVEDSGSYSCEAVNDVGSDSCSTEIVIKE  4756
               R     S+  L        D G  +C + N  G   C  ++ + +
Sbjct  710   EYRYGVSGSDQTLTHQQASHRDEGILTCISKNQEGIVQCQYDLTLXQ  850
```

Range 11: 53 to 298 GenBank Graphics ▼ Next Match ▲ Previous Match First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
49.3 bits(116)	0.030	Compositional matrix adjust.	33/82(40%)	42/82(51%)	2/82(2%)	+2

```
Query  5517  VTQGDPATLQVKFSGTKEITAKWFKDGQELTLGSKYKISVTDTVSILKIISTEKKDSGEY  5576
             V+    A   VK +G    TA W KDG+ +T G KYK+S       L+I  T+  DSG Y
Sbjct  53    VSSDSVAKFAVKATGEPRPTAIWTKDGKAITQGGKYKLSEDKGGFFLEIHKTDTSDSGLY  232

Query  5577  TFEVQNDVG--RSSCKARINVL  5596
             T  V+N  G   SSCK  I  +
Sbjct  233   TCTVKNSAGSVSSSCKLTIKAI  298
```

Range 12: 77 to 283 GenBank Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
48.9 bits(115)	0.035	Compositional matrix adjust.	27/71(38%)	36/71(50%)	2/71(2%)	+2

```
Query  1101  FGCQVGGNPKPHVYWKKSGVPLTTGYRYKVSYNKQTGECKLVISMTFADDAGEYTIVVRN  1160
             F  +  G P+P   W K G  +T G +YK+S +K  G   L I  T   D+G YT  V+N
Sbjct  77    FAVKATGEPRPTAIWTKDGKAITQGGKYKLSEDK--GGFFLEIHKTDTSDSGLYTCTVKN  250

Query  1161  KHGETSASASL  1171
               G  S+S  L
Sbjct  251   SAGSVSSSCKL  283
```

[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 formal (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.

DNA Sequence

```
>BU170890.1 AGENCOURT_7940143 NIH_MGC_67 Homo sapiens cDNA clone
IMAGE:6143622 5', mRNA sequence
TGAACCAATTTCCTCAAAACCAGTAATTGTTACTGGGTTGCAGGATACAACTGTTTCTTCAGACAGTGTT
GCTAAATTTGCAGTTAAGGCTACTGGAGAACCCCGGCCAACTGCCATCTGGACAAAAGATGGAAAGGCCA
TTACACAAGGAGGTAAATATAAACTCTCTGAAGACAAGGGAGGGTTCTTCTTAGAAATTCATAAGACTGA
TACTTCTGACAGTGGACTTTATACTTGTACAGTAAAAAATTCAGCTGGATCTGTGTCCTCTAGCTGCAAA
TTAACAATAAAAGCTATAAAAGATACTGAGGCACAGAAAGTCTCTACACAAAAGACTTCTGAAATTACAC
CTCAGAAGAAAGCTGTTGTCCAAGAGGAAATTTCCCAAAAAGCCCTAAGGTCTGAAGAAATTAAGATGTC
AGAGGCAAAATCTCAAGAAAAGTTAGCCCTCAAAGAGGAAGCTTCAAAGGTTCTGATTTCTGAAGAAGTC
AAGAAATCAGCAGCAACCTCCCTGGAAAAATCCATTGTCCATGAGGAAATCACTAAAACATCACAGGCAT
CAGAAGAAGTCAGAACTCATGCTGAGATTAAAGCATTTTCTACTCAGATGAGCATAAACGAAGGTCAAAG
ACTGGTTTTAAAAGCCAACATTGCTGGTGCCACTGATGTGAAATGGGTACTGAATGGCGTAGAGCTTACC
AACTCTGAGGAGTACCGATATGGTGTCTCAGGCAGCGATCAGACCCTAACCCATCAGCAAGCCAGTCACA
GAGATGAAGGAATCCTCACCTGCATAAGCAAAAACCAGGAAGGAATCGTCCAGTGTCAGTATGATTTGAC
ACTGNAGCAAAAAACTCTCAGATGCGTCAGCC
```

Protein Sequence

```
>BU170890.1_1 AGENCOURT_7940143 NIH_MGC_67 Homo sapiens cDNA clone
IMAGE:6143622 5', mRNA sequence
```

```
*TNFLKTSNCYWVAGYNCFFRQCC*ICS*GYWRTPANCHLDKRWKGHYTRR*I*TL*RQG
RVLLRNS*D*YF*QWTLYLYSKKFSWICVL*LQINNKSYKRY*GTESLYTKDF*NYTSEE
SCCPRGNFPKSPKV*RN*DVRGKISRKVSPQRGSFKGSDF*RSQEISSNLPGKIHCP*GN
H*NITGIRRSQNSC*D*SIFYSDEHKRRSKTGFKSQHCWCH*CEMGTEWRRAYQL*GVPI
WCLRQRSDPNPSASQSQR*RNPHLHKQKPGRNRPVSV*FDTXAKNSQMRQP
```

- **? Name**: Homo Sapiens cDNA Clone - Titin
- **Species**: Homo sapiens
    - Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

[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.

- Running the protein sequence above gave no result in blastp. Running the sequence ID gave the below results, all of which are homo sapien titin variants.

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
PREDICTED: Homo sapiens titin (TTN), transcript variant X14, mRNA	Homo sapiens	1559	1559	100%	0.0	98.97%	82119	XM_024453099.1
PREDICTED: Homo sapiens titin (TTN), transcript variant X12, mRNA	Homo sapiens	1559	1559	100%	0.0	98.97%	100356	XM_024453098.1
PREDICTED: Homo sapiens titin (TTN), transcript variant X11, mRNA	Homo sapiens	1559	1559	100%	0.0	98.97%	100437	XM_024453097.1
PREDICTED: Homo sapiens titin (TTN), transcript variant X4, mRNA	Homo sapiens	1559	1559	100%	0.0	98.97%	103662	XM_024453095.1
PREDICTED: Homo sapiens titin (TTN), transcript variant X13, mRNA	Homo sapiens	1559	1559	100%	0.0	98.97%	82170	XM_017004823.1
PREDICTED: Homo sapiens titin (TTN), transcript variant X10, mRNA	Homo sapiens	1559	1559	100%	0.0	98.97%	100554	XM_017004822.1
PREDICTED: Homo sapiens titin (TTN), transcript variant X6, mRNA	Homo sapiens	1559	1559	100%	0.0	98.97%	103512	XM_017004821.1
PREDICTED: Homo sapiens titin (TTN), transcript variant X5, mRNA	Homo sapiens	1559	1559	100%	0.0	98.97%	103515	XM_017004820.1
PREDICTED: Homo sapiens titin (TTN), transcript variant X1, mRNA	Homo sapiens	1559	1559	100%	0.0	98.97%	108117	XM_017004819.1
PREDICTED: Homo sapiens titin (TTN), transcript variant X15, mRNA	Homo sapiens	1559	1559	100%	0.0	98.97%	71774	XM_054343668.1
PREDICTED: Homo sapiens titin (TTN), transcript variant X14, mRNA	Homo sapiens	1559	1559	100%	0.0	98.97%	82119	XM_054343667.1
PREDICTED: Homo sapiens titin (TTN), transcript variant X13, mRNA	Homo sapiens	1559	1559	100%	0.0	98.97%	82170	XM_054343666.1
PREDICTED: Homo sapiens titin (TTN), transcript variant X12, mRNA	Homo sapiens	1559	1559	100%	0.0	98.97%	100356	XM_054343665.1
PREDICTED: Homo sapiens titin (TTN), transcript variant X11, mRNA	Homo sapiens	1559	1559	100%	0.0	98.97%	100437	XM_054343664.1
PREDICTED: Homo sapiens titin (TTN), transcript variant X10, mRNA	Homo sapiens	1559	1559	100%	0.0	98.97%	100554	XM_054343663.1
PREDICTED: Homo sapiens titin (TTN), transcript variant X9, mRNA	Homo sapiens	1559	1559	100%	0.0	98.97%	100785	XM_054343662.1
PREDICTED: Homo sapiens titin (TTN), transcript variant X8, mRNA	Homo sapiens	1559	1559	100%	0.0	98.97%	101208	XM_054343661.1
PREDICTED: Homo sapiens titin (TTN), transcript variant X7, mRNA	Homo sapiens	1559	1559	100%	0.0	98.97%	102768	XM_054343660.1

[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).

HINT: You can use a single sequence from your alignment or generate a consensus sequence from your alignment using the Bio3D function consensus(). The Bio3D functions blast.pdb(), plot.blast() and pdb.annotate() are likely to be of most relevance for completing this task. Note that the results of blast.pdb() contain the hits PDB identifier (or pdb.id) as well as Evalue and identity. The results of pdb.annotate() contain the other annotation terms noted above.

Note that if your consensus sequence has lots of gap positions then it will be better to use an original sequence from the alignment for your search of the PDB. In this case you could chose the sequence with the highest identity to all others in your alignment by calculating the row-wise maximum from your sequence identity matrix.

[Q9] Using AlphaFold notebook generate a structural model using the default parameters for your novel protein sequence.

> Note that this can take some time depending upon your sequence length. If your model is taking many hours to generate or your input sequence yields a "too many amino acids" (i.e. length) error you can focus on a single domain from your sequence - identify region by searching for PFAM domain matches.

Once complete save the resulting PDB format file for your records. Finally, generate a molecular figure of your generated PDB structure using the **Mol* viewer** online (or VMD/PyMol/Chimera if you prefer). To complete your analysis you can optionally highlight *conserved residues* that are likely to be functional as **spacefill** and the protein as **cartoon** colored by local alpha fold *pLDDT quality score*. This score is contained in the B-factor column of your PDB downloaded file. Please use a white or transparent background for your figure (i.e. not the default black in PyMol/VMD/Chimera etc.).

[Q10] Perform a "Target" search of ChEMBEL (https://www.ebi.ac.uk/chembl/) with your novel sequence. Are there any **Target Associated Assays** and **ligand efficiency data** reported that

may be useful starting points for exploring potential inhibition of your novel protein? If there are no assays listed here simply list “non available as of [date]”.

Scoring Rubric: [50 total points available]

Q1 (4 points)

Protein name 1

Species 1

Accession number 1

Function known 1

Q2 (6 points)

Blast method 1

Database searched 1

Limits applied 1

Search output list (top hits) 1

Alignment of choice 1

Evalue and other alignment stats 1

Q3 (3 points)

Protein sequence of choice matches Subject above 1
Name in header 1 Species 1

Q4 (3 point)

Blastp output list with identities & Evalue 1 Top alignment shown with alignment statistics 1 Results indicates a “novel” gene found 1

Q5 (3 points)

MSA labeled with useful names 1 MSA trimmed appropriately (i.e. no gap overhangs) 1 Pasted MSA fits report page width (i.e. font, format) 1

Q6 (1 point)

Figure illustrates sequence clustering pattern 1

Q7 (10 points)
Heatmap figure included in report 5 Heatmap is legible (i.e. no labels obscured) 5

Q8 (9 points)
PDB identifiers from multiple species reported 5 Annotation of PDB source, resolution and technique 4 Annotation of Evalue and Sequence Identity 1

Q9 (10 points)
Structure figure provided 2 Uses white background for molecular figure 1 Figure of high resolution (i.e. not just snapshot) 1 Conserved residues as spacefill 3 Protein cartoon colored by pLDDT quality score 3

Q10 (1 point)
Evidence of ChEMBEL searches 1