Eric Jordahl
ejordahl@ucsd.edu
A59019089
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: Derlin-1 (Derl1)
Accession: KAI4011837.1
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
Database: Expressed Sequence Tags
Species: All, excluding human (Taxid: 9606)
Chosen: Accession - CO735220.1 an mRNA sequence from squirrel (Callospermophilus lateralis) embryos

SILE04c11a05f1 squirrel embryo library 1 Callospermophilus lateralis cDNA clone 11a05 5', mRNA sequence

Sequence ID: CO735220.1 Length: 770 Number of Matches: 1

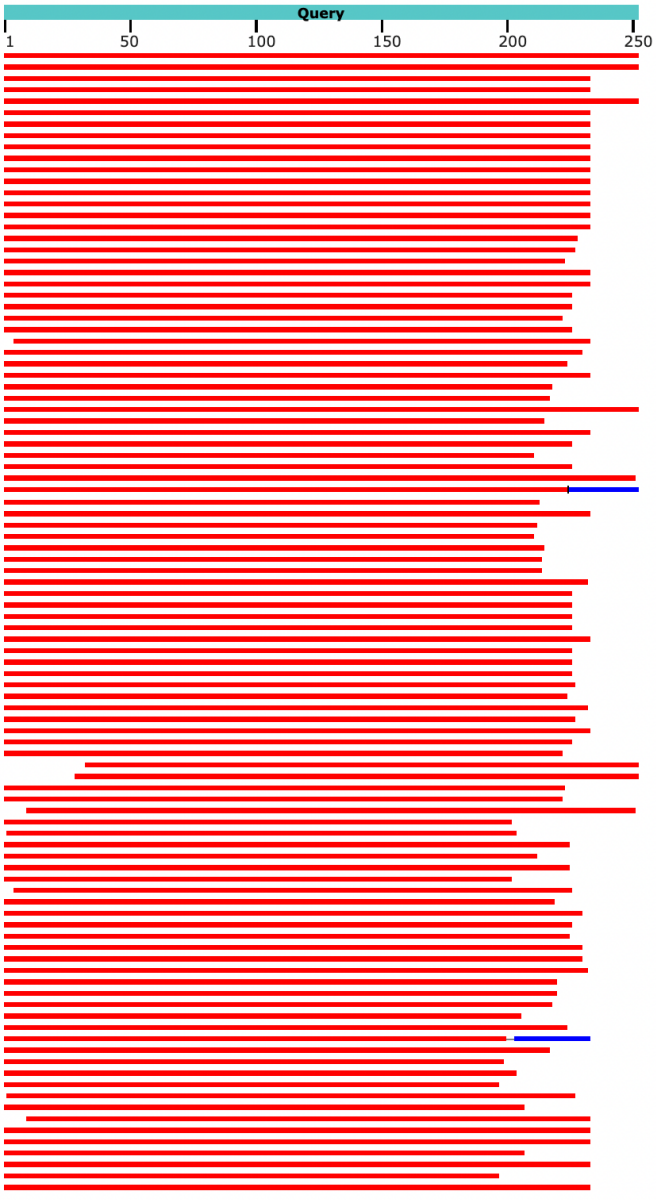
Range 1: 69 to 764 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
456 bits(1174)	5e-162	Compositional matrix adjust.	228/232(98%)	228/232(98%)	0/232(0%)	+3

```
Query  1    MSDIGDWFRSIPAITRYWFAATVAVPLVGKLGLISPAYLFLWPEAFLYRFQIWRPITATF  60
            MSDIGDWFR IP ITRYWFAATVAVPLVGKLGLISPAY FLWPEAFLYRFQIWRPITATF
Sbjct  69   MSDIGDWFRGIPIITRYWFAATVAVPLVGKLGLISPAYFFLWPEAFLYRFQIWRPITATF  248

Query  61   YFPVGPGTGFLYLVNLYFLYQYSTRLETGAFDGRPADYLFMLLFNWICIVITGLAMDMQL  120
            YFPVGPGTGFLYLVNLYFLYQYSTRLE GAFDGRPADYLFMLLFNWICIVITGLAMDMQL
Sbjct  249  YFPVGPGTGFLYLVNLYFLYQYSTRLEAGAFDGRPADYLFMLLFNWICIVITGLAMDMQL  428

Query  121  LMIPLIMSVLYVWAQLNRDMIVSFWFGTRFKACYLPWVILGFNYIIGGSVINELIGNLVG  180
            LMIPLIMSVLYVWAQLNRDMIVSFWFGTRFKACYLPWVILGFNYIIGGSVINELIGNLVG
Sbjct  429  LMIPLIMSVLYVWAQLNRDMIVSFWFGTRFKACYLPWVILGFNYIIGGSVINELIGNLVG  608

Query  181  HLYFFLMFRYPMDLGGRNFLSTPQFLYRWLPSRRGGVSGFGVPPASMRRAAD  232
            HLYFFLMFRYPMDLGGRNFLSTPQFLYRWLPSRRGGVSGFGVPPASMRRAAD
Sbjct  609  HLYFFLMFRYPMDLGGRNFLSTPQFLYRWLPSRRGGVSGFGVPPASMRRAAD  764
```

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
UMC-omix3_0A01-003-g09 Day 13 (non-pregnant) endometrium omix3 Ovis aries cDNA 5', mRNA sequence	Ovis aries	468	468	100%	3e-166	94.02%	823	GT882412.1
LB02914.CR_P01 GC_BGC-29 Bos taurus cDNA clone IMAGE:8235507 5', mRNA sequence	Bos taurus	468	468	100%	4e-166	94.02%	866	DV919076.1
LB00256.CR_M13 GC_BGC-02 Bos taurus cDNA clone IMAGE:7961271 5', mRNA sequence	Bos taurus	464	464	92%	1e-164	99.14%	826	DT828311.1
LB00227.CR_N11 GC_BGC-02 Bos taurus cDNA clone IMAGE:7950157 5', mRNA sequence	Bos taurus	462	462	92%	6e-164	99.14%	836	DT823295.1
AL574071 Homo sapiens PLACENTA COT 25-NORMALIZED Homo sapiens cDNA clone CS0DI040YI15 3-PR…	Homo sapiens	465	465	100%	6e-164	94.42%	1085	AL574071.3
BW978839 full-length enriched swine cDNA library, adult intestine Sus scrofa cDNA clone ITT010031H12 5', m…	Sus scrofa	461	461	92%	7e-164	99.57%	811	BW978839.1
AGENCOURT_6478263 NIH_MGC_72 Homo sapiens cDNA clone IMAGE:5563020 5', mRNA sequence	Homo sapiens	464	464	92%	8e-164	99.57%	993	BM471066.1
LB00248.CR_A02 GC_BGC-02 Bos taurus cDNA clone IMAGE:7957900 5', mRNA sequence	Bos taurus	461	461	92%	9e-164	99.14%	829	DT828724.1
LB02627.CR_J16 GC_BGC-26 Bos taurus cDNA clone IMAGE:8220018 5', mRNA sequence	Bos taurus	461	461	92%	1e-163	99.14%	852	DV883519.1
LB0165.CR_J22 GC_BGC-16 Bos taurus cDNA clone IMAGE:8079576 5', mRNA sequence	Bos taurus	460	460	92%	6e-163	99.14%	896	DT816914.1
SILE04c11a05f1 squirrel embryo library 1 Callospermophilus lateralis cDNA clone 11a05 5', mRNA sequence	Callospermophil…	456	456	92%	5e-162	98.28%	770	CO735220.1
LB02914.CR_F01 GC_BGC-29 Bos taurus cDNA clone IMAGE:8235267 5', mRNA sequence	Bos taurus	456	456	92%	2e-161	98.29%	851	DV918872.1

Distribution of the top 102 Blast Hits on 100 subject sequences

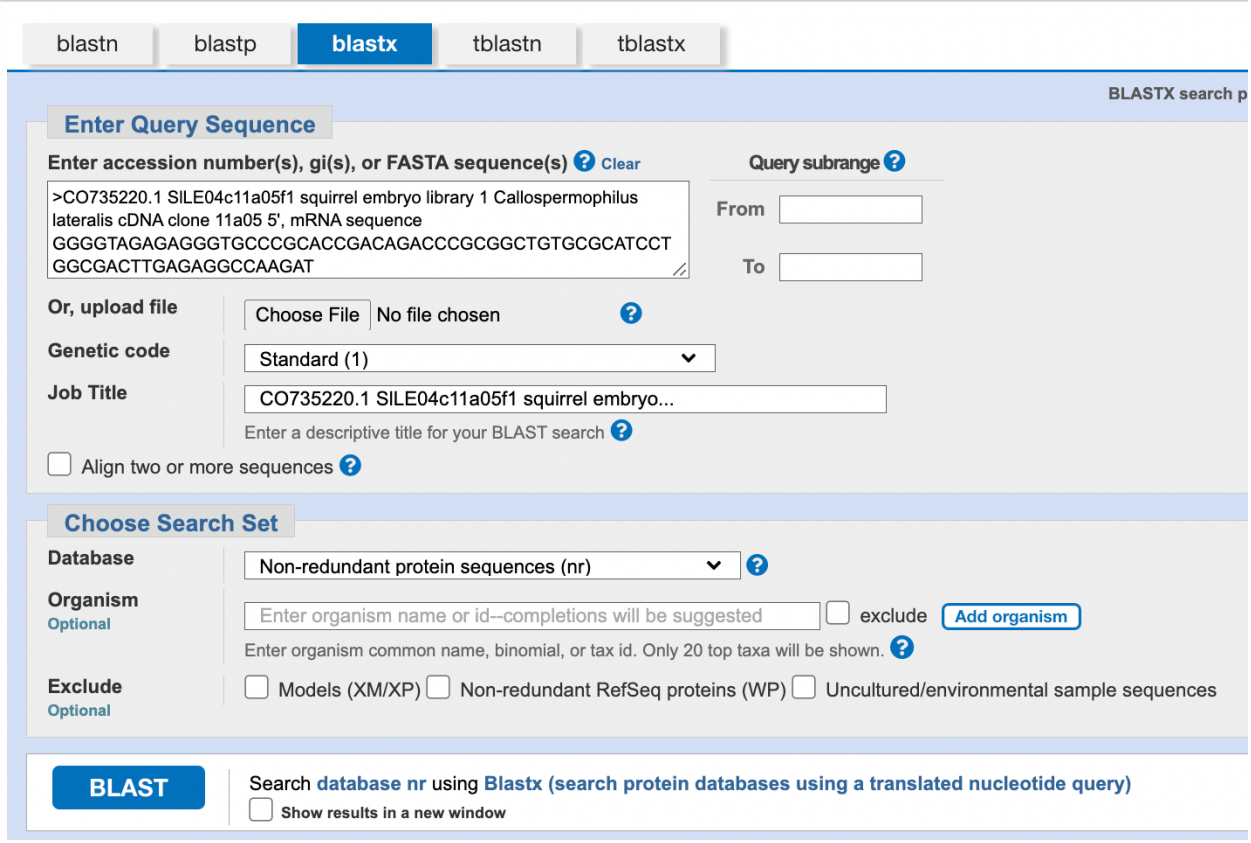


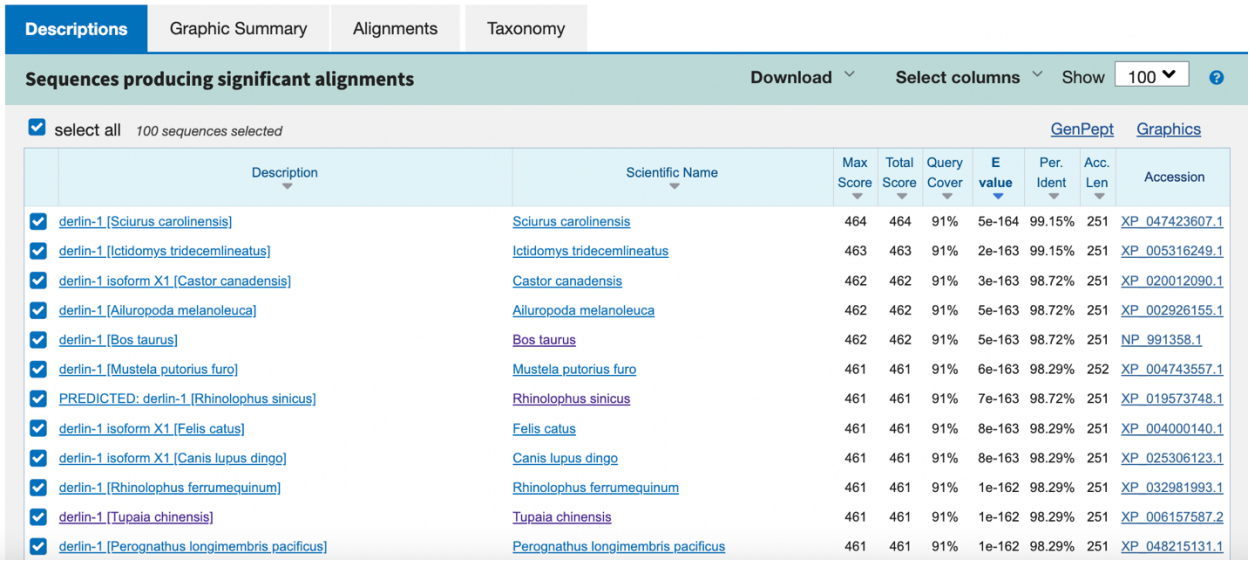
3] Gather information about this “novel” protein. At a minimum, show me the protein sequence of the “novel” protein as displayed in your BLAST results from [Q2] as FASTA format (you can copy and paste the aligned sequence subject lines from your BLAST result page if necessary) or translate your novel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

```
>CO735220.1_1 SlLE04c11a05f1 squirrel embryo library 1 Callospermophilus lateralis
cDNA clone 11a05 5', mRNA sequence
GVERVPAPTDPRLCASWRLERPRCRTSGTGSGASRSSRATGSLPPSRSPWLANSASSAQP
TSSSGPKLSSIASRFGGQSLPPFIFLWVQELDFFIWSICISYISILHDLKQEHLMGGRQT
IYSCFSLTGFAS*LLA*QWICNC**FL*SCQYFMSGPS*TET*LYHFGLEHDLRPVIYPG
LSLDSTISLEAR*SMS*LEILLDIFISS*CSDTQWTWEEEIFYPHLNFCTAGCPAGEAGC
QDSVCPLLA*GELLIKX

Name: Callospermophilus lateralis cDNA clone 11a05
Species: Callospermophilus lateralis
        Eukaryota, animalia, chordata, mammalia, rodentia, Sciuridae,
callospermophilus, lateralis
```

[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.





Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
derlin-1 [Sciurus carolinensis]	Sciurus carolinensis	464	464	91%	5e-164	99.15%	251	XP_047423607.1
derlin-1 [Ictidomys tridecemlineatus]	Ictidomys tridecemlineatus	463	463	91%	2e-163	99.15%	251	XP_005316249.1
derlin-1 isoform X1 [Castor canadensis]	Castor canadensis	462	462	91%	3e-163	98.72%	251	XP_020012090.1
derlin-1 [Ailuropoda melanoleuca]	Ailuropoda melanoleuca	462	462	91%	5e-163	98.72%	251	XP_002926155.1
derlin-1 [Bos taurus]	Bos taurus	462	462	91%	5e-163	98.72%	251	NP_991358.1
derlin-1 [Mustela putorius furo]	Mustela putorius furo	461	461	91%	6e-163	98.29%	252	XP_004743557.1
PREDICTED: derlin-1 [Rhinolophus sinicus]	Rhinolophus sinicus	461	461	91%	7e-163	98.72%	251	XP_019573748.1
derlin-1 isoform X1 [Felis catus]	Felis catus	461	461	91%	8e-163	98.29%	251	XP_004000140.1
derlin-1 isoform X1 [Canis lupus dingo]	Canis lupus dingo	461	461	91%	8e-163	98.29%	251	XP_025306123.1
derlin-1 [Rhinolophus ferrumequinum]	Rhinolophus ferrumequinum	461	461	91%	1e-162	98.29%	251	XP_032981993.1
derlin-1 [Tupaia chinensis]	Tupaia chinensis	461	461	91%	1e-162	98.29%	251	XP_006157587.2
derlin-1 [Perognathus longimembris pacificus]	Perognathus longimembris pacificus	461	461	91%	1e-162	98.29%	251	XP_048215131.1

Download GenPept Graphics

derlin-1 [Sciurus carolinensis]

Sequence ID: XP_047423607.1 **Length:** 251 **Number of Matches:** 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 234 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
464 bits(1194)	5e-164	Compositional matrix adjust.	232/234(99%)	233/234(99%)	0/234(0%)	+3

```
Query  69   MSDIGDWFRGIPIITRYWFAATVAVPLVGKLGLISPAYFFLWPEAFLYRFQIWRPITATF  248
            MSDIGDWFR IPIITRYWFAATVAVPLVGKLGLISPAYFFLWPEAFLYRFQIWRP+TATF
Sbjct  1    MSDIGDWFRSIPIITRYWFAATVAVPLVGKLGLISPAYFFLWPEAFLYRFQIWRPVTATF  60

Query  249  YFPVGPGTGFLYLVNLYFLYQYSTRLEAGAFDGRPADYLFMLLFNWICIVITGLAMDMQL  428
            YFPVGPGTGFLYLVNLYFLYQYSTRLEAGAFDGRPADYLFMLLFNWICIVITGLAMDMQL
Sbjct  61   YFPVGPGTGFLYLVNLYFLYQYSTRLEAGAFDGRPADYLFMLLFNWICIVITGLAMDMQL  120

Query  429  LMIPLIMSVLYVWAQLNRDMIVSFWFGTRFKACYLPWVILGFNYIIGGSVINELIGNLVG  608
            LMIPLIMSVLYVWAQLNRDMIVSFWFGTRFKACYLPWVILGFNYIIGGSVINELIGNLVG
Sbjct  121  LMIPLIMSVLYVWAQLNRDMIVSFWFGTRFKACYLPWVILGFNYIIGGSVINELIGNLVG  180

Query  609  HLYFFLMFRYPMDLGGRNFLSTPQFLYRWLPSRRGGVSGFGVPPASMRRAADQN  770
            HLYFFLMFRYPMDLGGRNFLSTPQFLYRWLPSRRGGVSGFGVPPASMRRAADQN
Sbjct  181  HLYFFLMFRYPMDLGGRNFLSTPQFLYRWLPSRRGGVSGFGVPPASMRRAADQN  234
```

Download GenPept Graphics

derlin-1 [Ictidomys tridecemlineatus]

Sequence ID: XP_005316249.1 **Length:** 251 **Number of Matches:** 1

See 7 more title(s) See all Identical Proteins(IPG)

Range 1: 1 to 234 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
463 bits(1191)	2e-163	Compositional matrix adjust.	232/234(99%)	232/234(99%)	0/234(0%)	+3

```
Query  69   MSDIGDWFRGIPIITRYWFAATVAVPLVGKLGLISPAYFFLWPEAFLYRFQIWRPITATF  248
            MSDIGDWFR IPIITRYWFAATVAVPLVGKLGLISPAYFFLWPEAFLYRFQIWRPITATF
Sbjct  1    MSDIGDWFRSIPIITRYWFAATVAVPLVGKLGLISPAYFFLWPEAFLYRFQIWRPITATF  60

Query  249  YFPVGPGTGFLYLVNLYFLYQYSTRLEAGAFDGRPADYLFMLLFNWICIVITGLAMDMQL  428
            YFPVGPGTGFLYLVNLYFLYQYSTRLE GAFDGRPADYLFMLLFNWICIVITGLAMDMQL
Sbjct  61   YFPVGPGTGFLYLVNLYFLYQYSTRLETGAFDGRPADYLFMLLFNWICIVITGLAMDMQL  120

Query  429  LMIPLIMSVLYVWAQLNRDMIVSFWFGTRFKACYLPWVILGFNYIIGGSVINELIGNLVG  608
            LMIPLIMSVLYVWAQLNRDMIVSFWFGTRFKACYLPWVILGFNYIIGGSVINELIGNLVG
Sbjct  121  LMIPLIMSVLYVWAQLNRDMIVSFWFGTRFKACYLPWVILGFNYIIGGSVINELIGNLVG  180

Query  609  HLYFFLMFRYPMDLGGRNFLSTPQFLYRWLPSRRGGVSGFGVPPASMRRAADQN  770
            HLYFFLMFRYPMDLGGRNFLSTPQFLYRWLPSRRGGVSGFGVPPASMRRAADQN
Sbjct  181  HLYFFLMFRYPMDLGGRNFLSTPQFLYRWLPSRRGGVSGFGVPPASMRRAADQN  234
```

[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.

Renamed Sequences

```
>Human OS=Homo sapiens OX=9606 GN=DERL1 PE=1 SV=1
MSDIGDWFRS IPAITRYWFA ATVAVPLVGK LGLISPAYLF LWPEAFLYRF QIWRPITATF
YFPVGPGTGF LYLVNLYFLY QYSTRLETGA FDGRPADYLF MLLFNWICIV ITGLAMDMQL
LMIPLIMSVL YVWAQLNRDM IVSFWFGTRF KACYLPWVIL GFNYIIGGSV INELIGNLVG
HLYFFLMFRY PMDLGGRNFL STPQFLYRWL PSRRGGVSGF GVPPASMRRA ADQNGGGGRH
NWGQGFRLGD Q

>Squirrel SlLE04c11a05f1 squirrel embryo library 1 Callospermophilus
lateralis cDNA clone 11a05 5', mRNA sequence
GVERVPAPTDPRLCASWRLERPRCRTSGTGSGASRSSRATGSLPPSRSPWLANSASSAQP
TSSSGPKLSSIASRFGGQSLPPFIFLWVQELDFFIWSICISYISILHDLKQEHLMGGRQT
IYSCFSLTGFAS*LLA*QWICNC**FL*SCQYFMSGPS*TET*LYHFGLEHDLRPVIYPG
LSLDSTISLEAR*SMS*LEILLDIFISS*CSDTQWTWEEEIFYPHLNFCTAGCPAGEAGC
QDSVCPLLA*GELLIKX

>Mouse OS=Mus musculus OX=10090 GN=Derl1 PE=1 SV=1
MSDIGDWFRS IPAITRYWFA ATVAVPLIGK LGIISPAYFF LWPEAFLYRF QIWRPFTATF
YFPVGPGTGF LYLVNLYFLY QYSTRLEAGA FDGRPADYLF MLLFNWICIV ITGLAMDMQL
LMIPLIMSVL YVWAQLNRDL IVSFWFGTRF KACYLPWVIL GFNYIIGGSV INELIGNLVG
HLYFFLMFRY PMDLGGRNFL STPQFLYRWL PSRRGGVSGF GVPPASMRRA ADQNGGGGRH
NWGQGFRLGD Q

>Cow OS=Bos taurus OX=9913 GN=DERL1 PE=2 SV=1
MSDIGDWFRS IPTITRYWFA ATVAVPLVGK LGLISPAYFF LWPEAFLYRF QIWRPITATF
YFPVGPGTGF LYLVNLYFLY QYSTRLETGA FDGRPADYLF MLLFNWICIV ITGLAMDMQL
LMIPLIMSVL YVWAQLNRDM IVSFWFGTRF KACYLPWVIL GFNYIIGGSV INELIGNLVG
HLYFFLMFRY PMDLGGRNFL STPQFLYRWL PSRRGGVSGF GVPPASMRRA ADQNGGGGRH
NWGQGFRLGD Q

>Orangutan NAB OS=Pongo abelii OX=9601 GN=DERL1 PE=2 SV=1
MSDIGDWFRS IPAITRYWFA ATVAVPLVGK LGLISPAYLF LWPEAFLYRF QIWRPITATF
YFPVGPGTGF LYLVNLYFLY HYSTRLETGA FDGRPADYLF MLLFNWICIV ITGLAMDMQL
LMIPLIMSVL YVWAQLNRDM IVSFWFGTRF KACYLPWVIL GFNYIIGGSV INELIGNLVG
HLYFFLMFRY PMDLGGRNFL STPQFLYRWL PSRRGGVSGF GVPPASMRRA ADQNGGGGRH
NWGQGFRLGD Q

>Roundworm OS=Caenorhabditis elegans OX=6239 GN=cup-2 PE=2 SV=1
MDLENFLLGI PIVTRYWFLA STIIPLLGRF GFINVQWMFL QWDLVVNKFQ FWRPLTALIY
YPVTPQTGFH WLMMCYFLYN YSKALESETY RGRSADYLFM LIFNWFFCSG LCMALDIYFL
LEPMVISVLY VWCQVNKDTI VSFWFGMRFP ARYLPWVLWG FNAVLRGGGT NELVGILVGH
AYFFVALKYP DEYGVDLIST PEFLHRLIPD EDGGIHGQDG NIRGARQQPR GHQWPGGVGA
RLGGN
```

>ThaleCress OS=Arabidopsis thaliana OX=3702 GN=DER1 PE=2 SV=1
MSSPGEFYNS LPPITKAYGT LCFFTTVATQ LGLVAPVHIA LIPELVLKQF QIWRLITNLF
FLGGFSINFG IRLLMIARYG VQLEKGPFER RTADFLWMMI FGSFTLLVLS VIPFFWTPFL
GVSLVFMLLY LWSREFPNAN ISLYGLVTLK AFYLPWAMLA LDVIFGSPIM PDLLGIIAGH
LYYFLTVLHP LATGKNYLKT PKWVNKIVAR WRIGAPVASV RQAGGVGAAG PGAGGGVGGG
GAYSSARAPP ESSNTAFRGR SYRLTD

>Rice OS=Oryza sativa subsp. japonica OX=39947 GN=DER1 PE=2 SV=2
MSSPAEYYNS LPPISKAYGT LCFFATVLCQ LQILNPPFLA LYYPFVFKKF QIWRLFTSFF
FLGKFSINFG IRLLMIARYG VQLEKGAFEK RTADFLWMMI FGAISLLALS AIPFLDIYFL
GVPMVSMLLY VWSREYPNSQ ISMYGLVQLR SFYLPWAMLG LDVIFGSEIL PGLLGILVGH
TYYFLSVLHP LATGKNYLKT PMWVHKIVAR FRIGVQANAP VRPAAANTGS GAFRGRSYRL
SQ

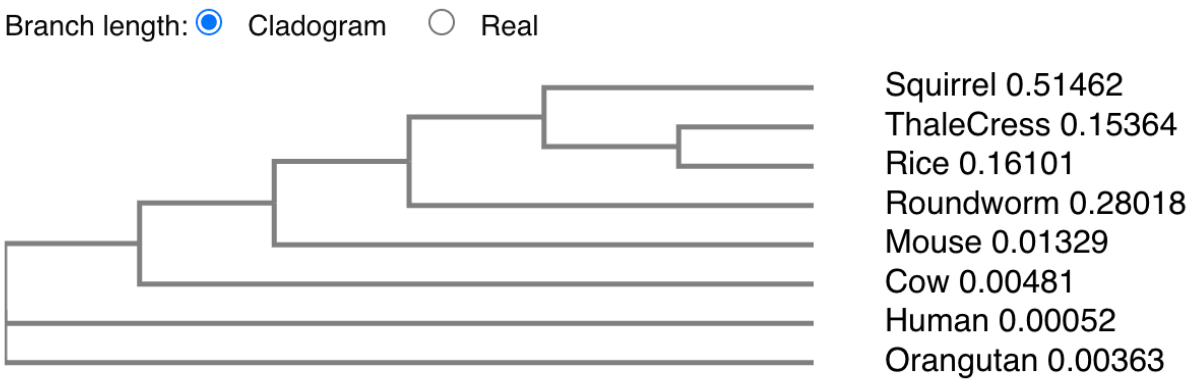
Alignment (Made using EBI ClustalW)

```
Squirrel        GVERVPAPTDPRLCASWRLERPRCRTSGTGSGASRSSRATGSLPPSRSPWLANSASSAQP 60
Thale Cress     ---------------------------------MSSPGEFYNSLPPITKAYGTLCFF---- 24
Rice            ---------------------------------MSSPAEYYNSLPPISKAYGTLCFF---- 24
Mouse           ---------------------------------MSDIGDWFRSIPAITRYWFAATVA---- 24
Cow             ---------------------------------MSDIGDWFRSIPTITRYWFAATVA---- 24
Human           ---------------------------------MSDIGDWFRSIPAITRYWFAATVA---- 24
Orangutan       ---------------------------------MSDIGDWFRSIPAITRYWFAATVA---- 24
Roundworm       ---------------------------------MDLENFLLGIPIVTRYWFLASTI---- 23
                                                      .:*     :

Squirrel        TSSSGPKLSSIASRFGGQSLPPFIFLWVQE--LDFFIWSICISYISILHDL--------- 109
Thale Cress     --------TTVATQL-GLVAPVHIALIPELVLKQFQIWRLITNLFFLGG---FSINFGIR 72
Rice            --------ATVLCQL-QILNPPFLALYYPFVFKKFQIWRLFTSFFFLGK---FSINFGIR 72
Mouse           --------VPLIGKL-GIISPAYFFLWPEAFLYRFQIWRPFTATFYFPVGPGTGFLYLVN 75
Cow             --------VPLVGKL-GLISPAYFFLWPEAFLYRFQIWRPITATFYFPVGPGTGFLYLVN 75
Human           --------VPLVGKL-GLISPAYLFLWPEAFLYRFQIWRPITATFYFPVGPGTGFLYLVN 75
Orangutan       --------VPLVGKL-GLISPAYLFLWPEAFLYRFQIWRPITATFYFPVGPGTGFLYLVN 75
Roundworm       --------IPLLGRF-GFINVQWMFLQWDLVVNKFQFWRPLTALIYYPVTPQTGFHWLMM 74
                          :  ::        : *        * :*      :

Squirrel        ----------KQEHLMGGRQTIYSCFSLTGFAS*LLA*QWICNC**F--L*SCQYFMSGP 152
Thale Cress     LLMIARYGVQLEKGPFERRTADFL---------WMMIFGSFTLLVLSVIPFFWTPFLGVS 123
Rice            LLMIARYGVQLEKGAFEKRTADFL---------WMMIFGAISLLALSAIPFLDIYFLGVP 123
Mouse           LYFLYQYSTRLEAGAFDGRPADYL---------FMLLFNWICIVITG--LAMDMQLLMIP 124
Cow             LYFLYQYSTRLETGAFDGRPADYL---------FMLLFNWICIVITG--LAMDMQLLMIP 124
Human           LYFLYQYSTRLETGAFDGRPADYL---------FMLLFNWICIVITG--LAMDMQLLMIP 124
Orangutan       LYFLYHYSTRLETGAFDGRPADYL---------FMLLFNWICIVITG--LAMDMQLLMIP 124
Roundworm       CYFLYNYSKALESETYRGRSADYL---------FMLIFNWFFCSGLC--MALDIYFLLEP 123
                          :      * : :           ::    :             ::


Squirrel        S*TET*LYHFGLEH-DLRPVIYPGLSLDSTI------SLEAR*SMS*LEILLDIFISS-- 199
Thale Cress     -LVFMLLYLWSREFPNANISLYGLVTLKAFYLPWAMLALDVIFGSPIMPDLLGIIAGHLY 182
Rice            -MVSMLLYVWSREYPNSQISMYGLVQLRSFYLPWAMLGLDVIFGSEILPGLLGILVGHTY 182
Mouse           -LIMSVLYVWAQLNRDLIVSFWFGTRFKACYLPWVILGFNYIIGGSVINELIGNLVGHLY 183
Cow             -LIMSVLYVWAQLNRDMIVSFWFGTRFKACYLPWVILGFNYIIGGSVINELIGNLVGHLY 183
Human           -LIMSVLYVWAQLNRDMIVSFWFGTRFKACYLPWVILGFNYIIGGSVINELIGNLVGHLY 183
Orangutan       -LIMSVLYVWAQLNRDMIVSFWFGTRFKACYLPWVILGFNYIIGGSVINELIGNLVGHLY 183
Roundworm       -MVISVLYVWCQVNKDTIVSFWFGMRFPARYLPWVLWGFNAVLRGGGTNELVGILVGHAY 182
                      ** :     :    ::    : :        .::          *:. : .

Squirrel        ---------------*CSDTQWTWEEEIFYPHLNFCT---AGCPAGEAGCQDS-----VC 235
Thale Cress     YFLTVLHPLATG-KNYLKTPKWVNK--IV-ARWRIGAPVASVRQAGGVGAAGPGAGGGVG 238
```

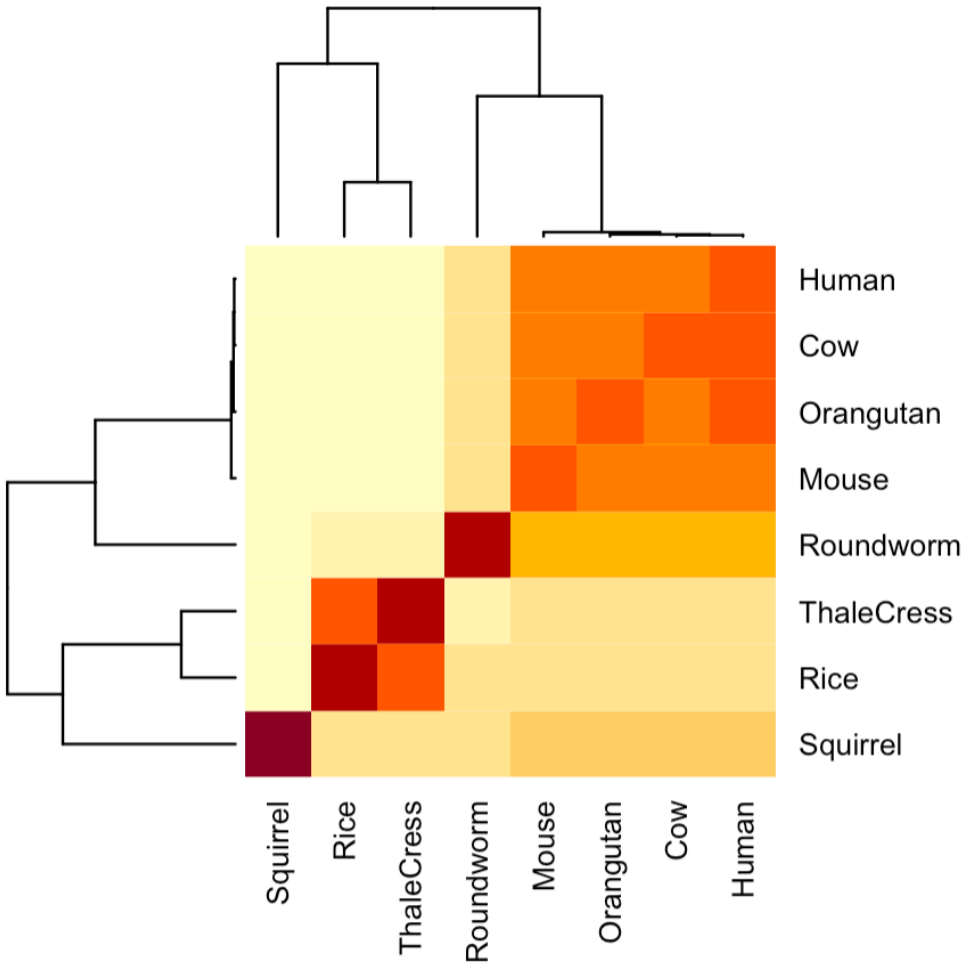
```
Rice            YFLSVLHPLATG-KNYLKTPMWVHK--IV-ARFRIGVQANAPVRPA--------------224
Mouse           FFLMFRYPMDLGGRNFLSTPQFLYR--WL-PSRRGGVSGFGVPPASMRRAADQNGGGGRH240
Cow             FFLMFRYPMDLGGRNFLSTPQFLYR--WL-PSRRGGVSGFGVPPASMRRAADQNGGGGRH240
Human           FFLMFRYPMDLGGRNFLSTPQFLYR--WL-PSRRGGVSGFGVPPASMRRAADQNGGGGRH240
Orangutan       FFLMFRYPMDLGGRNFLSTPQFLYR--WL-PSRRGGVSGFGVPPASMRRAADQNGGGGRH240
Roundworm       FFVALKYPDEYG-VDLISTPEFLHR--LI-PDEDGGIHGQDGNIRGAR----QQPRGHQW234
                                .   :  .   .               .

Squirrel        P----------L------LA*GELLIKX      246
Thale Cress     GGGAYSSARAPPESSNTAFRGRSYRLTD-      266
Rice            ----------AANTGSGAFRGRSYRLSQ-      242
Mouse           N------------------WGQGFRLGDQ      251
Cow             N------------------WGQGFRLGDQ      251
Human           N------------------WGQGFRLGDQ      251
Orangutan       N------------------WGQGFRLGDQ      251
Roundworm       P------------------GGVGARLGGN      245
                                  .  .  *
```

[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.

Cladogram made with EBI ClustalW (Clustal Omega Program)



[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.

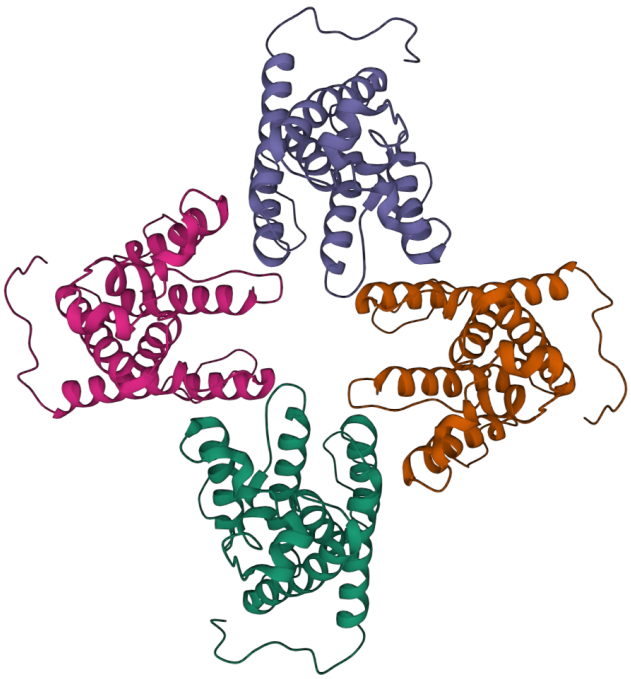
After trying this with consensus sequences and individual alignments, this was the only result, no matter which sequence/alignment was run.

ID	Technique	Resolution	Source	E-Value	Identity
7CZB	Electron Microscopy	3.8 Å	Homo Sapiens	2.9e-21	31.873

[Q9] Generate a molecular figure of one of your identified PDB structures using VMD. You can optionally highlight conserved residues that are likely to be functional. Please use a white or transparent background for your figure (i.e. not the default black). Based on sequence similarity. How likely is this structure to be similar to your "novel" protein?

Viewed in Mol* ED Viewer

I do not expect that this structure will be very similar to the novel protein structure because this structure shown is of a tetramer, so it is possible that one of the monomers could be close to the protein structure of the novel protein.



[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?
There were no significant targets or ligand efficiency data from the ChEMBEL data as shown below by my results where none of the targets have an e-value lower than 0.66, which is quite high.

E-Value	Positives %	Identities %	Score (bits)	Score	Length	ChEMBL ID	Name	UniProt Accessions	Type	Organism	Compounds
0.66	40.5	36.5	30.8018	68	4374	CHEMBL4295881	*E3 ubiquitin-protein ligase HUWE1*	Q7Z6Z7	SINGLE PROTEIN	Homo sapiens	1 
1.2	41.6	26.7	29.6462	65	412	CHEMBL2203	*Motilin receptor*	O43193	SINGLE PROTEIN	Homo sapiens	335 
1.3	52.5	45	29.6462	65	504	CHEMBL3885612	*Nuclear receptor subfamily 2 group C member 2/TGF-beta-activated kinase 1 and MAP3K7-binding protein 1*	Q15750, P49116	CHIMERIC PROTEIN	Homo sapiens	12 
1.3	52.5	45	29.6462	65	504	CHEMBL5605	*Mitogen-activated protein kinase kinase kinase 7-interacting protein 1*	Q15750	SINGLE PROTEIN	Homo sapiens	44 
1.3	52.5	45	29.6462	65	504	CHEMBL3038499	*TAK1/TAB1*	Q15750, O43318	PROTEIN COMPLEX	Homo sapiens	252 
1.9	41.7	31.9	29.261	64	1321	CHEMBL1926492	*Tau-tubulin kinase 1*	Q5TCY1	SINGLE PROTEIN	Homo sapiens	42