

BGGN-213: FOUNDATIONS OF BIOINFORMATICS

The find-a-gene project assignment

UCSD email: ptvo@ucsd.edu

PID: A59010610

[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 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: Leucine-rich PPR motif-containing protein (LRPPRC)
Accession: NP_573566
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 against Cynops pyrrhogaster ESTs
Database: Expressed Sequence Tags (est)
Organism: Cynops pyrrhogaster (Taxid: 8330)

The screenshot shows the NCBI BLAST search interface. The top navigation bar has tabs for 'blastn', 'blastp', 'blastx', 'tblastn' (which is selected), and 'tblastx'. Below the tabs, it says 'Translated BLAST: tblastn'. A sub-header reads 'TBLASTN search translated nucleotide databases using a protein query. [more...](#)'. The main form area is titled 'Enter Query Sequence' and contains a text input field with 'NP_573566.2' and a 'Clear' button. To the right are 'Query subrange' fields for 'From' and 'To'. Below the input field are options for 'Or, upload file' (with a 'Choose File' button showing 'No file chosen') and 'Job Title' (with a text input field containing 'NP_573566:leucine-rich PPR motif-containing...'). There is also a checkbox for 'Align two or more sequences'. The next section, 'Choose Search Set', includes a 'Database' dropdown set to 'Expressed sequence tags (est)', which is highlighted with a yellow background. Under 'Organism' (optional), 'Cynops pyrrhogaster (taxid:8330)' is listed with a 'Create custom database' link. 'Exclude' and 'Limit to' sections are present with checkboxes for 'Models (XM/XP)', 'Uncultured/environmental sample sequences', and 'Sequences from type material'. An 'Entrez Query' section with a text input field and a 'Create custom database' link is also shown. At the bottom left is a large blue 'BLAST' button, and at the bottom right is a note: 'Search database est using Tblastn (search translated nucleotide databases using a protein query) Show results in a new window'.

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.

BLAST® » tblastn » results for RID-ZF6YBBDE016

Job Title: NP_573566:leucine-rich PPR motif-containing...

RID: ZF6YBBDE016 Search expires on 02-01 15:20 pm
Download All ▾

Program: TBLASTN ⓘ Citation ▾

Database: est See details ▾

Query ID: NP_573566.2

Description: leucine-rich PPR motif-containing protein, mitochondrial ...

Molecule type: amino acid

Query Length: 1394

Other reports: ⓘ

Filter Results

Organism: only top 20 will appear exclude
Type common name, binomial, taxid or group name
+ Add organism

Percent Identity: [] to [] E value: [] to [] Query Coverage: [] to []
Filter Reset

Descriptions Graphic Summary Alignments Taxonomy

Sequences producing significant alignments Download New Select columns Show 100 ⓘ

Description		Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	FS293294_Cp_al Cynops pyrrhogaster cDNA clone Cp_al_009_G17 3'. mRNA sequence	Cynops pyrrhogaster	468	468	21%	4e-154	71.38%	985	FS293294.1
<input checked="" type="checkbox"/>	FS299916_Cp_al Cynops pyrrhogaster cDNA clone Cp_al_027_E08 3'. mRNA sequence	Cynops pyrrhogaster	382	430	22%	5e-122	67.53%	1022	FS299916.1
<input checked="" type="checkbox"/>	FS300508_Cp_al Cynops pyrrhogaster cDNA clone Cp_al_028_N20 3'. mRNA sequence	Cynops pyrrhogaster	215	215	12%	1e-62	59.65%	896	FS300508.1
<input checked="" type="checkbox"/>	FS293338_Cp_al Cynops pyrrhogaster cDNA clone Cp_al_009_I13 3'. mRNA sequence	Cynops pyrrhogaster	85.1	85.1	6%	3e-18	52.27%	708	FS293338.1

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.

Chosen match: Accession FS293294.1, a 985 base pair clone from *Cynops pyrrhogaster*. See below for alignment details.

FS293294 Cp_al Cynops pyrrhogaster cDNA clone Cp_al_009_G17 3', mRNA sequence						
Sequence ID: FS293294.1 Length: 985 Number of Matches: 1						
Range 1: 1 to 912 GenBank Graphics				▼ Next Match ▲ Previous Match		
Score	Expect	Method	Identities	Positives	Gaps	Frame
468 bits(1203)	4e-154	Compositional matrix adjust.	217/304(71%)	256/304(84%)	0/304(0%)	+1
Query 699	LELKAKYYESDMVTGGYAALINLCCRHDKVEDALNLKEEFDRLDSSAVIDTGKYVGLVRVL		758			
Sbjct 1	LE+E+K Y+E+DMV GGYAALIN CCRHD VE+ALNLK E R DSS LDT KY+ LV+V					
Query 759	AKHGKLQDAINILKEMKEKDVLIKDTTALSFFHMLNGAALRGEIETVKQLHEAIVTLGLA		818			
Sbjct 181	AKHG+L DAINILKEMKEKDVLIKDTT SFFH+LNG A+RGE+ETV +L E IVTLGLA					
Query 819	EPSTNISFPPLVTVHLEKGDLSTALEVADCYEKYKVLPRIHDVLCKLVEGETDLIQKAM		878			
Sbjct 361	+P N+ P+VTVHLEK D ALE +IDCY+KY LPR+HDVLCKLVE+G+T+L+QKAM					
Query 879	KPVANLCSPVVTVHLEKDDAPAALEASIDCYKKYNCLPRLHDVLCKLVERGDTELLQKAM		540			
Sbjct 541	DFVSQEQQEMVMLYDLFFAFLQTGNYKEAKKIIETPGIRARSARLQWFCDRCVANNQVET		938			
Query 939	DFVSQE+GEM MLYDLFFAFLQT YKEAKKIIETPG+RAR RLQWF ++C+ NQ+ET					
Sbjct 721	DFVSQERGEMTMLYDLFFAFLQTAKYKEAKKIIETPGLRARPGRQLQWFAEKCITGNQMET		720			
Query 999	LEKLVELTQKLFECDRDQMYYNLLKLYKINGDWQRADAVWNKIQEENVIPREKTLRLLAE		998			
Sbjct 901	LE LENFVEMTSKLFECRDEMIFYYLLKLCKENNEWQKADAIWTKMQUEENLIPRERTLKLAD		900			
Query 1002	ILRE 1002					
Sbjct 912	+ +E					

[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.

Chosen sequence:

```
> Cynops pyrrhogaster protein (sequence taken from BLAST result)
LEVKGKYEAADMVVGGAALINACCRHDNVVEALNLKREVHRKDSSVALDTNKYLSLVKCAKHGRLLDRAINILKEMK
EKDVLIKDTTLGSFFFHVNLNGVAMRGEVETVNRLEVITVTLGLAKPVANLCSPVVTVHLEKDDAPAALEASIDCYKKY
NCLPRLHDVLCKLVERGDTELLQKAMDFVSQERGEMTMLYDLFFAFLQTAKYKEAKKIIETPGLRARPGRQLQWFAEK
CITGNQMETLENFVEMTSKLFECRDEMIFYYLLKLCKENNEWQKADAIWTKMQUEENLIPRERTLKLADLFKEMVRK
FHLMFLRIGMKKLQHQK*
```

Name: *Cynops pyrrhogaster* LRPPRC

Species: *Cynops pyrrhogaster*

[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.

Details:

A BLASTP search against NR database (see setup in first screen-shot below) yielded a top hit result is to a protein from *Cynops orientalis* (Chinese fire belly newt). See additional screen shots below for top hits and selected alignment details:

blastn **blastp** blastx tblastn tblastx

BLASTP programs search protein databases using a protein query. [more...](#)

Enter Query Sequence

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

```
LEVKKGKYEADMVVGGAALINACCRHDNVEEALNLKREVHRKDSSVALDTNKY
LSLVKVCAKHGRLDDAINILKEMKEKDVLIKDTTLGSFFHVNLNGVAMRGEVETVN
RLLEVIVTGLAKPVANLCSPVTVHLEKDDAAPAALEASIDCYKKYNCLPRLHDV
LCKLVERGDTELLQKAMDFVSQERGEMTMLYDLFFAFLQTAKYKEAKKIIETPGL
```

Query subrange [?](#)

From To

Or, upload file Choose File No file chosen [?](#)

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

Align two or more sequences [?](#)

Choose Search Set

Database [?](#)

Organism Enter organism name or id--completions will be suggested exclude [Add organism](#)

Optional
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. [?](#)

Exclude Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences
Optional

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 [?](#)

BLAST Search database nr using Blastp (protein-protein BLAST)
 Show results in a new window

The top hit result is to a protein from *Cynops orientalis* (Chinese fire belly newt). See additional screen shots below for selected alignment details:

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

leucine-rich PPR motif-containing protein LRPPRC [Cynops orientalis]

Sequence ID: [QIS93427.1](#) Length: 1407 Number of Matches: 1

Range 1: 705 to 1007 [GenPept](#) [Graphics](#)

▼ Next Match ▲ Previous Match

Score 606 bits(1562)	Expect 0.0	Identities 301/303(99%)	Positives 301/303(99%)	Gaps 0/303(0%)
Query 1	LEVKGKYEADMVVGGAALINACCRHDNVEEALNLKREVHRKDSSVALDTNKYLSLVKVC		60	
Sbjct 705	LEVKGKYEADMVVGGAALINACCRHDNVEEALNLKRE RDSSVALDTNKYLSLVKVC			764
Query 61	AKHGRLLDDAINILKEMKEKDVLIKDTTLGSFFFHVNLNGVAMRGEVETVNRLLLEVIVTGLA		120	
Sbjct 765	AKHGRLLDDAINILKEMKEKDVLIKDTTLGSFFFHVNLNGVAMRGEVETVNRLLLEVIVTGLA			824
Query 121	KPVANLCSPVVTVHLEKDDAPAALAEASIDCYKKYNCLPRLHDVLCKLVERGDTELLQKAM		180	
Sbjct 825	KPVANLCSPVVTVHLEKDDAPAALAEASIDCYKKYNCLPRLHDVLCKLVERGDTELLQKAM			884
Query 181	DFVSQERGEMTMLYDLFFAFLQTAKYKEAKKIIETPGLRARPGRQLQWFAEKCITGNQMET		240	
Sbjct 885	DFVSQERGEMTMLYDLFFAFLQTAKYKEAKKIIETPGLRARPGRQLQWFAEKCITGNQMET			944
Query 241	LENFVEMTSKLFECDRDEMYFYLLKLCKENNEWQKADAIWTKMQUEENLIPRERTLKLLAD		300	
Sbjct 945	LENFVEMTSKLFECDRDEMYFYLLKLCKENNEWQKADAIWTKMQUEENLIPRERTLKLLAD			1004
Query 301	LFK 303			
Sbjct 1005	LFK 1007			

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

hypothetical protein KIL84_016365, partial [Mauremys mutica]

Sequence ID: [KAH1172526.1](#) Length: 694 Number of Matches: 1

Range 1: 277 to 579 [GenPept](#) [Graphics](#)

▼ Next Match ▲ Previous Match

Score 510 bits(1313)	Expect 2e-174	Identities 247/303(82%)	Positives 274/303(90%)	Gaps 0/303(0%)
Query 1	LEVKGKYEADMVVGGAALINACCRHDNVEEALNLKREVHRKDSSVALDTNKYLSLVKVC		60	
Sbjct 277	LEV KYE DMVVGGYAALIN CCRHDNVE+A+NLK EV RKDSSVALDT+KYL+LVKV			336
Query 61	AKHGRLLDDAINILKEMKEKDVLIKDTTLGSFFFHVNLNGVAMRGEVETVNRLLLEVIVTGLA		120	
Sbjct 337	GHGRLEDAINILKEMKEKD+IPIKDTTVTSFFFHILNAAAMRGEVETVNLHESILTLGLA			396
Query 121	KPVANLCSPVVTVHLEKDDAPAALAEASIDCYKKYNCLPRLHDVLCKLVERGDTELLQKAM		180	
Sbjct 397	KPSANLCSPPLITVHLEKDDVPAALEATIDCCKYKGKIPRLHDVLCLRLEQGNTDLLQKAM			456
Query 181	DFVSQERGEMTMLYDLFFAFLQTAKYKEAKKIIETPGLRARPGRQLQWFAEKCITGNQMET		240	
Sbjct 457	DFVSQERGEMTMLYDIFFAFLNTGKYKEAKKIIETPGLRARPGRQLQWFAEKCIATNQMET			516
Query 241	LENFVEMTSKLFECDRDEMYFYLLKLCKENNEWQKADAIWTKMQUEENLIPRERTLKLLAD		300	
Sbjct 517	LENM VEMT QKLFECDR DQMYYLLKLCKISNDWRKADATWTKMQUEENVIPRETTLRLAD			576
Query 301	LFK 303			
Sbjct 577	TLK 579			

[List of all hits:](#)

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	leucine-rich PPR motif-containing_protein_LRPPRC [Cynops orientalis]	Cynops ori...	628	628	92%	0.0	99.34%	1407	QIS93427.1
<input checked="" type="checkbox"/>	hypothetical protein_KIL84_016365 [Mauremys mutica]	Mauremys...	523	523	92%	1e-179	81.52%	694	KAH117256.1
<input checked="" type="checkbox"/>	LOW QUALITY PROTEIN: leucine-rich PPR motif-containing_protein,...	Rhinatrem...	538	538	93%	2e-177	82.24%	1407	XP_029449192.1
<input checked="" type="checkbox"/>	leucine-rich PPR motif-containing_protein_mitochondrial-like [Pelodisc...	Pelodiscu...	523	523	92%	3e-177	79.21%	865	XP_006138247.1
<input checked="" type="checkbox"/>	leucine-rich PPR motif-containing_protein_mitochondrial [Chelonoidis ...	Chelonoidi...	523	523	92%	9e-174	80.53%	1188	XP_032630883.1
<input checked="" type="checkbox"/>	leucine-rich PPR motif-containing_protein_mitochondrial [Chelonia my...	Chelonia ...	527	527	92%	4e-173	81.19%	1399	XP_037751766.1
<input checked="" type="checkbox"/>	leucine-rich PPR motif-containing_protein_mitochondrial isoform X2 [...	Mauremys...	523	523	92%	8e-173	81.19%	1254	XP_039385622.1
<input checked="" type="checkbox"/>	leucine-rich PPR motif-containing_protein_mitochondrial [Mauremys ...	Mauremys...	526	526	92%	1e-172	81.52%	1398	XP_044867041.1
<input checked="" type="checkbox"/>	leucine-rich PPR motif-containing_protein_mitochondrial [Dermochely...	Dermochel...	525	525	92%	5e-172	80.86%	1415	XP_038251504.2
<input checked="" type="checkbox"/>	leucine rich pentatricopeptide repeat containing [Chelydra serpentina]	Chelydra s...	523	523	92%	1e-171	80.20%	1397	KAG6933563.1
<input checked="" type="checkbox"/>	leucine-rich PPR motif-containing_protein_mitochondrial [Gopherus e...	Gopherus ...	523	523	92%	1e-171	80.53%	1399	XP_030413525.1
<input checked="" type="checkbox"/>	leucine-rich PPR motif-containing_protein_mitochondrial [Terrapene c...	Terrapene ...	523	523	92%	1e-171	80.20%	1397	XP_026508293.1
<input checked="" type="checkbox"/>	leucine-rich PPR motif-containing_protein_mitochondrial isoform X1 [...	Mauremys...	522	522	92%	3e-171	81.19%	1398	XP_039385621.1
<input checked="" type="checkbox"/>	leucine-rich PPR motif-containing_protein_mitochondrial isoform X3 [...	Sceloporu...	513	513	92%	1e-167	79.87%	1409	XP_042312263.1
<input checked="" type="checkbox"/>	leucine-rich PPR motif-containing_protein_mitochondrial isoform X2 [...	Sceloporu...	513	513	92%	1e-167	79.87%	1410	XP_042312254.1
<input checked="" type="checkbox"/>	leucine-rich PPR motif-containing_protein_mitochondrial isoform X1 [...	Sceloporu...	513	513	92%	1e-167	79.87%	1411	XP_042312243.1
<input checked="" type="checkbox"/>	leucine-rich PPR motif-containing_protein_mitochondrial-like [Rhincod...	Rhincodon...	491	542	92%	1e-166	73.93%	719	XP_020390257.1
<input checked="" type="checkbox"/>	LRPPRC [Latimeria menadoensis]	Latimeria ...	509	509	92%	7e-166	79.54%	1431	AWT24668.1
<input checked="" type="checkbox"/>	PREDICTED: leucine-rich PPR motif-containing_protein_mitochondria...	Latimeria ...	508	508	92%	1e-165	79.21%	1431	XP_005999623.1
<input checked="" type="checkbox"/>	PREDICTED: leucine-rich PPR motif-containing_protein_mitochondria...	Anolis car...	500	500	92%	3e-165	76.90%	1147	XP_016850993.1
<input checked="" type="checkbox"/>	LRPPRC [Protopterus annectens]	Protoptero...	506	506	92%	1e-164	77.56%	1435	AWT24643.1
<input checked="" type="checkbox"/>	leucine-rich PPR motif-containing_protein_mitochondrial [Trachemys ...	Trachemy...	504	504	92%	3e-164	77.89%	1398	XP_034620325.1
<input checked="" type="checkbox"/>	leucine-rich PPR motif-containing_protein_mitochondrial [Chrysemys ...	Chrysemy...	503	503	92%	8e-164	77.89%	1397	XP_005296166.1
<input checked="" type="checkbox"/>	leucine-rich PPR motif-containing_protein_mitochondrial [Polyodon sp...	Polyodon ...	503	503	92%	8e-164	75.58%	1430	XP_041107358.1
<input checked="" type="checkbox"/>	PREDICTED: leucine-rich PPR motif-containing_protein_mitochondria...	Anolis car...	502	502	92%	2e-163	76.90%	1409	XP_008113921.1
<input checked="" type="checkbox"/>	leucine-rich PPR motif-containing_protein_mitochondrial isoform X1 [...	Mus musc...	480	480	92%	1e-162	73.60%	699	XP_011244945.1
<input checked="" type="checkbox"/>	unnamed protein product [Mus musculus]	Mus musc...	481	481	92%	1e-162	73.60%	712	BAB29082.2
<input checked="" type="checkbox"/>	leucine-rich PPR motif-containing_protein_mitochondrial isoform X2 [...	Microcaeci...	498	498	94%	9e-162	77.92%	1407	XP_030051658.1
<input checked="" type="checkbox"/>	leucine-rich PPR motif-containing_protein_mitochondrial isoform X1 [...	Microcaeci...	498	498	94%	9e-162	77.92%	1408	XP_030051657.1
<input checked="" type="checkbox"/>	leucine-rich PPR motif-containing_protein_mitochondrial [Amblyraja ra...	Amblyraja ...	495	557	92%	1e-160	73.93%	1406	XP_032881396.1
<input checked="" type="checkbox"/>	leucine-rich PPR motif-containing_protein_mitochondrial [Acipenser ru...	Acipenser ...	495	495	92%	1e-160	74.92%	1432	XP_034777299.1
<input checked="" type="checkbox"/>	leucine-rich PPR motif-containing_protein_mitochondrial isoform X2 [...	Scyliorhin...	494	494	92%	1e-160	74.59%	1417	XP_038669480.1
<input checked="" type="checkbox"/>	leucine-rich PPR motif-containing_protein_mitochondrial isoform X1 [...	Scyliorhin...	494	494	92%	2e-160	74.59%	1420	XP_038669472.1
<input checked="" type="checkbox"/>	putative sodium-coupled neutral amino acid transporter 7 [Platysterno...	Platystern...	493	545	93%	3e-160	77.23%	1395	TFK01649.1
<input checked="" type="checkbox"/>	leucine-rich PPR motif-containing_protein_mitochondrial [Podarcis mu...	Podarcis ...	491	491	92%	2e-159	75.25%	1395	XP_028577588.1
<input checked="" type="checkbox"/>	PREDICTED: leucine-rich PPR motif-containing_protein_mitochondria...	Lepisoste...	491	491	92%	2e-159	72.94%	1421	XP_015218688.1

[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.

Re-labeled sequences for alignment:

> Homo_sapiens | NP_573566.2

MAALLRSARWLLRAGAAPRLPLSLRLLPGGPGRLHAASYLPAARAGPVAGGLSPARLYAIAAKEKDQIESTFSSRKISNQFDWALMRLDLSVRRTGRIPKKLLQKVFNNDTCRGGLGGSHALLLRSRCSGSLPELKLEERTEFAHRIWDTLQKLGAVYDVSHYNALLKVLQNEYKFSPDFLAKMEEANIOPNRVTYQRЛИASYCNVGDIEGASKILGFMDKLPVT EAVFSALVTGHARAGDMENAENILTVMRDAGIEPGPDTYLALLNAYAEKGDIHVVKQTLKVEKSELHLMRDLLQIIFSFSKAGYPQYVSEILEKVTCCERRYIPDAMNLILLLVTEKLEDVALQILLACPVSKEDGPSVFGSFFLQHCVMNT PVEKLTDYCKKLKEVQMHSFPLQFTLHCALLANKTDLAKALMKAVKEEGFPIRPHYFWPLLGVGRKEKNVQGIIIEL KGMQELGVHPDQETYTDYVIPCFDSVNSARAILQENGCLSDDMFSQAGLRSEAANGNLDVFVLSFLKSNTLPISLQS IRSSLIGFRRSMNINLWSEITELLYKDGRYCQEPRGPTEAVGYFLYNLIDSMSDSEVQAKEEHLRQYFHQLEKMNVKIPENIYRGIRNLLESYHVPÉLIKDAHLLVESKNLDFQKTVQLTSELESTLETLKAENQPIRDVLKQLLIVLCSEENMQKALELKAKYESDMVTGGYAALINLCCRHDKVEDALNLKEEFDRLDSSAVLDTGKVYGLVRVLAKHGKLQDAINILKEMKEKDVLIKDTTALSFFHMLNGAALRGEIETVKQLHEAIVTGLAEPSTNISFPLTVHLEKGDLSALEVAID CYEKYKVLPRIHDVLCKLVEKGETDLIQKAMDFVSQEQQEMVMLYDLFFAFLQTGNYKEAKKIIETPGIRARSARLQWFCDRCVANNQVETLEKVLVELTQKLFECDRDQMYNNLLKLYKINGDWQRADAVWNKIQEENVIPREKTLRLAEILREGNQEVPDFVPELWYEDEKHSLNSSASTTEPDFQKDILIAICRLNQKKGAYDIFLNAKEQNIVFNAETYSNLIKLLMSEDYFTQAMEVKAFATHIKGFTLNDAAANSRLIITQVRRDYLKEAVTTLKTVLDQQQTPSRLAVTRVIQALAMKGDVENIEVVQKMLNGLEDSIGLSKMFVINNIALAQIKNNNIDAIIENIENMLTSENKVIIEPQYFGLAYLFRKVIEEQLEPAVEKISIMAERLANQFAIYKPVTDFFLQLVDAGVDDARALLQRCGAIAEQTPILLFLRNSRKQGKASTVKSLELIPELNEKEEAYNSLMKSYVSEKDVTSAKALYEHLTAKNTKLDDLFLKRYASLLKYAGEPVFIEPPESFEFYAQQLRKLRENSSS

> Newt

LEVKGKYEADMVVGGAALINACCRHDNVEEALNLKREVHRKDSSVALDTNKYLSQLVKVCAKHGRDDAINILKEMKEKDVLIKDTLGSFFFHVLNGVAMRGEVETVNRLLLEVINTLGLAKPVANLCSPVTVHLEKDDAPAALEASIDCYKKYNCLPRLHDVLCKLVERGDTELLQKAMDFVSQERGEMTMLYDLFFAFLQTAKYKEAKKIIETPGLRARPGRLQWFAEK CITGNQMETLENFVEMTSKLFECRDEMYFYLLKLCKENNEWQKADAIWTKMQEENLIPRERTLKLADLFKEMVRKFHLMFLRIGMKKLQHQK

> Turtle | XP_006138247

MHSVDFHFKGCLISGFRRSKNVDLWSKITELLYKDGRYCQTPPGPSEAVGYFLYNLIDSMSDSEVQAKEEHLRQYFHQLKKMNIVIPPNIYRGIRNVLDSDYHVPÉLIKDIMLVVDSEKLSAADIPKNTELEVLSSEELEKLKAEKQPIGNVLIKOLVALCAEENMQKALEVKAKYEPMVVGGYAALINLCCRHDNAEDAMNLKEEVYRKDSSVALDTNKYLALVKVLGKHGRLEDAINILKEMKEKDVPIRDTTVTSFHLNAAAMQGEVETVNRHESILTGLVKPSTNLCSPLITVHLEKD DVPAALEAAIDCFKKYGNIPRLHDILCRLIEKGNTDLLQKAMDFVSQERGEMTMLYDLFFAFLNTGKYKEAKKIIETPGLRARPGRLQWFAEKCIANNQMDILENMVEMTEKLFECRDRQMYYYILQLCKISSDWRRADATWTKMQEENVIPERRTLRLADIKNNGQEVPDFVPEIWIYEDCVESAPVSENNPEKKILILCKKGNIQEAYNILLEQKKDIMFPSFTSYI LIKALLAEGCLEKAINVNIAETHIKGFTLNDAAASSLIITQVRRDYLKDAISTLKAVLENDMVPTRLAVTRLIQALAMKGDVESVRTVEKVENLAWSIGLSRMLFVNNTVLAHKNNNLDTAVEYIESLIISGMQNPDSPITSISYVFRVIEEKLESALEKLSAMAERLVNQFGIYKPATDLFLQYVSEGRIDDARLLLQRCGGIAEQKKIMMAFIAKSSQKPGQGQKIKMlldlvpdfPEMEVVSYLLKCHVLDNDIVSAKTLFEKAKAENIRTDEIFLKRLAFLLKSAGEPVFTEPPESFH FYVNKLKRQELSSHVD

> Tortoise | XP_032630883

MAALLSWACRLRPLCSHLLRLPRGVAAHPALRAAGLVGAPGSTHЛИQARLFAIAPHQKGKVQEEPVLVQNKQAQHFDWALNKLDSVRRTGRTTRTLLRIFHDMCRTGPSSNQALLLRSRCSGSLPEVPLCERTELAHMIWGKLQDLGAVYDASHYNALLKVLQNEHKFSPTEFLSKMEEANVQPNRVTYQRЛИAAYCNEGDIEGASKILGFMDKMKELPITEAVFSSLVTGHARTGDMENAENIFSVMRDAGIEPGPDTYLALLNAYAEKGDIINSVKQILEKIEKMEGYLMRDLLQVIYSLAKAGYPQYVPDFLSQLMRYERYGIYDAMNLSLNLITQGLEDTAFQVFKSFPTLPSENHSETSMYGNFFLQHCVHMDKPLSKLKQFCDELKEANMHSSPLQFTLYCALESKKAALAINLMKTMKEEGLPLRPHYSWPLLGVYQKEKNVQGTIVG

LKAMHELGVEPDVETYTNVLTNFDDIQTFRALLQENGCPFETEGLSVAALRHEAIYGTLENVLSSLSSPSMPSVNL
SHFKGCLIFGFKRSNDIDLWSKITELLYKDGRYCQTPPGPAEAVYFLYNLIDSMSDSEVQAKEEHLRQYFHQLKKM
NVVIPTNIYRGIRNLLAAYHVPÉLIKDIILVTDREKLSADIPKNTEVSALLEEELEKLKAEKQPIGNVLKQLIVALC
AEENMQKALEVKAKYEPDMVVGGYAALINLCCRHDNVEDAMNLKEEVFRKDSSVALDTNKYLALVKVLGKHGRLEDA
INILKEIKEKDIPIKDVTTSFFILNAAAMRGEVETVNRLHESITLGLAKPSTNLCTPLITVHLEKDDVPSALEA
TIDCYKKYKGMPRLHDILCRLIEQGNTELLQKAMDFVSQERGEMMLYDIFFAFLNTGKYKEAKKIETPGLRARPG
SLQWFAEKCIATNQMETLENMVEMTQKLFECRDEMYYLLKLCKISNDWRKAEATWTKMQEENVI PRARTLRLAD
TLKNNGQEVPFDVPEIWYEDHVESASVLENNPEKKVLMCKGSVQEAYNILLEAQKKDIMFPASAYSTLIKALLAE
GCLEEALKVKNIAETHIKGFTLNDAASSLLITQVRDYLKDAISTLKKVLESMDVPMPLAVIRLIQALAMGDLES
ICTVEKMVENLATSIGLSRMLFVNNTVLAHIKK

> Zebrafish | NP_001136064 (sequence from BLAST)

MAALLRSARLLKTSSASLIQVIGNSKHAPAAGRILYSGAIGTLRVGVCGRQTVPGRLNANSLSYRPALPCVSCRQYA
VVPEQSGQVKDEASLA VRSKQAQQFDWALS KLDSSV RRTGRVTKLLLHIFHDICRTG PSGNQALLLRSCGSLLP
EVPLAERTELVKRIWDKLLELGVS YDVSHYNALLKTYLQNEFRSPTD FLAKMEAANVQPNRVTYQRLIAAYCEGN
IEGASAILGF MKNKDLPITEAVFNSLVVGHARAGDITSSEGILSVMKSAGIEPGPDTYLSLLNMYAEKGDI D KIKQT
LDV VENADF FLM DRDL MQLV S LART GHEQHV PEIVS RMR HERYV P DAINL CLN LITH GHEKTA FSVL KS LTGMLD
TH TGDTPDFGNFFL RHC NVMDK S AEDIVGFC KDLK DLGH STPL QFTL QCALEG KKT S I GLM KRM KA E S L PIK PH
YFLPLFAHHHKDKNIPAIIEVLRGMQEMS VSPDVDAFSFYILPSFPSLDNAK ASLKEAGV DVNTDGLIV AELRVQAY
SGNLAKLLSLMSSPALPTIDLSVFRAGLIA GF KRF QNVENMAK I ELLYKDARFAADGHDASENVGYFLYNLIDSMS
ETELQSNEEKIREYFG LLKSMN NI ISVNIFRGIRNILESHHVPELVKEAL TLVDKT DDMTEVMMFRSSEGRISALVK
TLAEQKAEGPAH TLK KLIN VLSIEEKLEQALDLKSKYEDDMTPAAYATL INLCRHDNAEEALKLKIEMARKDSE
VALDAQKYIALVRVLSKHGKLEE ALDILKEMKEKNIMIRDN LIGFTTMNSIAMKG DAD AIRRLQETIFTLGLAKP
SNGLCSP LVTCY LE SGDHAGAF DAVMECHKQY NQLPKIH DLMCSLVEKG D ALLLQKVMEFLTLERGEMMLYDLFFA
FLQTGRNREARKIIETPGLRARSNRLQWF A EK C IAA QQME PLENL VDMTVKLF ECDRDEM YH YLFR LCK ETNDWRKA
EAIWMKMQEENLA PRERTL RL LAEILKSNNQEV PFEV PKW FEDDG KGQSE VPEKEESSA VAKT DDRS RL NATIRL
KLQSLCKKG EA QEA FD IL KEV DSKG I VPG PAI YD AI I KALLAKG NIE DA ISVKD I AVGHIPS FILSDV ANN LISH
VKCQMKDSVQVL RDMLKADQMP SQLA I TRLVQGLADEGNLK D I QEV EAM TKA FG SFN LS NM LFVN NTAL ALL NNGD
VDSA VDMLQTF YTENTERQ NNSIAHVFRKVLNANN DAAMD KLSAMA ER LCN QF ASYRA ATDLF LYV VDT GRTEE AKF
LLQRCAAVGEQKD LLSYV L RAS Q QPG QAA KVMS LMELI P DIREK EDI YSQL MKCH GLD QD LASAK AL YERM QVEGV
RIDE TLK RL A KLYR DS GEP VP FEE PAES F RF YAD KL K D Q RT Q STAS IDN

> Drosophila_mojavensis | XP_002002110

MASILRTGKLLRYFAGFTRNVVVNSVRD CESNNLLQ SAPCMCGQFQNGFAS NAAASKA E VT LDRQIRRLQDARRMG
RISRRD LEEV LDEIRTH TATSSQSL LVI RCCGNL VPEELPEV R TALVQ EIW KTLN ALN VPMDI SHYN ALL RVY LEN
EHQFAP TD FLAEIEAKGIEP NRVTYQRLIARYCQ QGDIEGATR ILEYMRG KKL PVN ENVF NSL I L GH SQ ANDLES AR
GILGVMKQAGLEPTADTYTTLLCAFARHG DIE ALQ ST LAECEPKEI I LLDK D LLDV AYT LAV HGN GEHV DAVLSK LR
ISP GFNQ DAVN I I LRL VNKG QED VAL KLLR LMP RNS RVNG E PVD V GAFFI RQL V KAN RP VEK I L SIC RTL QSE GLNP
KALT IATEA GLT NGV MNNA LPLL HEMK NLGLP IRQHYFWPLFC SVDSN QV LDV VRRM Q QDF ALNP NSET VRD YV IPN
LKEKNWDRVTVL RDAG V PN STAV S A VY A ALT TH H IADA A KIME QN RAY YMPLI FR QPLI L AHT ND FAS FIRC V
RQVYEGLQL RAGKE ATE E AAE AAT DGE A PAT PER Q P DVG QIV DAAI Y FRR RER VPT L EK I L QGL V N QGL SSSG K
AT AL SELL GSE MTPKIAE QL GK L TS G E L E P I PL PNS GKR S L D T I D E L E R F I K N V E A K G E N S N N I K R Q L L N A C F RS
QN LEK TL QV IER LEK DT FQ I P I G I Y A QL V D L Y T H K R T S D A L A Q Y N K L R T A D A S F K L D N F K T V R L A D L L L Q E E R V D E
A L Q L L K E N Q K E A A V E A G E G S F N Y V S T V W R I L N S I A E T G N A E R L K V V F D A L V A G N Y V V P T N V L L G P L I K V H L S R D D I P
K A I D A F E Q I C Q Q Y K A T P W K N E L A C R L I Q K E D A A N L Q R L T D L S T S I H G E V N S L Y D L V F S V E C G R V R Q A R K I L E T P G L
RTR P Q R I S N A C D R Y K N E G M L Q P L E G L I E A T K D L G H I D R N K I Y T L L L S Y D K A D E T E K A L G L W T K M Q E E G V T P T D A F L
L K L A E L L K R K N I D V P F V V P E T Q Q P K S K R N K P A K T E A N V A E Q V V Q S P P E K V E P K A K A V P A A K P I S N I A G F R R A I Q A N
D P D A A I S F K E R V L S G D K F N V L D T S R L I E L L V R A D R L S E A T K Y V E E L L A E K Q H P Q P K I F K F Y L N K I A A S G D L E V M Q R I
G Q Q L N D E Q K R L V S F D N R Y C H A Y I V A G K A E Q F L K Q L S T E I E A V K S S E E A G K L A E K F P R G G A V G I L E K H P E L I A Q Y E T L
A E K Y A A H N Q L G P M N V L W M H L I S N G Q E V A S K Q I W D K H L S S A P R L M F Q R V L Q T A R E Q Q D E K L A S T V I S Q L R G S K I S E G A
I G N A Y S C L I D I Q T T K G N T D K A L E V L A N A I K D V S L E N I N R T A L L R L K Q A V E E K S Q Q F P Y T V P E K R A K A E D S S S S S S S S S
S D D D V T P K R P E T A P T R P E R V

Alignment:

CLUSTAL multiple sequence alignment by MUSCLE (3.8)

Drosophila_mojavensis	MASILRTGKLLRYFA-----GFTRNVVVNSVRDCESNNLQLQAPCMCGOFQNGFASNA
Zebrafish	MAALLRSARLLKTSSASLIQVIGNSKHAPAAGRRLYSGAIGTLRVGVCGRQTVPGRLNANS
Homo_sapiens	MAALLRSARWLLRA-----GAAPRLPLSLRLLPGGPGRHLHAASYLPAARAGPVAGGL
Newt	-----
Tortoise	MAALLSWACRLR-----PLCSHLLLRLPRGVAAHRPALRAAGLVGAPGSTHL
Turtle	-----
 Drosophila_mojavensis	 AA-----SKEVTLDQIRRLDQDARRMGRIS
Zebrafish	LFSYRPALPCVSCRQYAVVPEQSGQVKDEASLAVERSQAAQFDWALSKLDSSVRRTGRVT
Homo_sapiens	LS-----PARLYAIAAKEKDQIQEESTFSS--RKISNQFDWALMRLDLSVRRTGRIP
Newt	-----
Tortoise	IY-----QARLFAIAPHQKGKVQEEPVLSVQNQQAQHFDWALNKLDSSVRRTGRIT
Turtle	-----
 Drosophila_mojavensis	 RRDLEEVLDIERTHTATSSQSLLVIRCCGNLVPEELPEVRTALVQEIWKTLNALNVPMD
Zebrafish	KTLLLHIFHDICRTGYPSGNQALLLRSCGSLLPEVPLAERTELVKRIWDKLLELGVSYD
Homo_sapiens	KKLLQKVFNNDTCRGGLGGSHALLLRSCGSLLPELKLEERTEFAHRIWDTLQKLGAVYD
Newt	-----
Tortoise	RTLLLRIFHDMCRTGYPSSNQALLLRSCGSLLPEVPLCERTELAHMIWGKLQDLGAVYD
Turtle	-----
 Drosophila_mojavensis	 ISHYNALLRVYLENEHQFAPTDLAEIEAKGIEPNRVTYQRLIARYCQQGDIEGATRILE
Zebrafish	VSHYNALLKTYLQNEFRSPTDFLAKMEAANVQPNRVTYQRLIAAYCEEGNIEGASAILE
Homo_sapiens	VSHYNALLKVYLQNEYKFSPTDFLAKMEEANIOPNMRVTYQRLIASYCNGDIEGASKILG
Newt	-----
Tortoise	ASHYNALLKVYLQNEHKFSPTEFLSKMEEANVQPNRVTYQRLIAAYCNEG DIEGASKILG
Turtle	-----
 Drosophila_mojavensis	 YMRGKKLPVNENVFNSLILGHQSQANDLESARGILGVMKQAGLEPTADTTLLCAFARHG
Zebrafish	FMKNKDLPITEAVFNSLUVGHARAGDITSSEGILSVMKSAGIEPGPDTYLSLLNMYAEKG
Homo_sapiens	FMKTKDLPVTEAVFSALVTGHARAGDMENAENILTVMRDAGIEPGPDTYLLALLNAYAEKG
Newt	-----
Tortoise	FMKMKELPITEAVFSSLVTGHARTGDMENAENIFSVMRDAGIEPGPDTYLLALLNAYAEKG
Turtle	-----
 Drosophila_mojavensis	 DIEALQSTLAECEPKIELLLDKLVDVAYTLAVHGNGEHVDAVLSKLRISPFGNQDAVNI
Zebrafish	DIDKIKQTLDVVENADFFLMDRDLMLQVSSLARTGHEQHVPEIVSRMRHERGYVPDAINL
Homo_sapiens	DIDHVKQTLEKVEKSELHLMRDLLQIIFSFSKAGYPQVYSEILEKVTERRYIPDAMNL
Newt	-----
Tortoise	DINSVKQILEKIEKMEGYLMDRDLQVIYSLAKAGYPQVYVDILSQMRYERYGIPDAMNL
Turtle	-----
 Drosophila_mojavensis	 ILRLVNKGQEDVALKLLRMP--RNSRVNGEPDVGAFFIRQLVKANRPVEKILSICRTL
Zebrafish	CLNLITHGHEKTAFSVLKSL-TGMLDTHTGDTPDFGNFLRHCVNMDSKAEDIVGFCCKDL
Homo_sapiens	ILLLVTEKLEDVALQILLACPVSKED---GPSVFGSFFLQHCVTMNTPVEKLTDYCKKL
Newt	-----
Tortoise	SLNLITQGLEDTAFQVFKSFTLPSHENHSETSMYGNFLQHCVHMDKPLSKLKQFCDEL
Turtle	-----
 Drosophila_mojavensis	 QSEGLNPKALTIAATEAGLTNGVMNNALPLLHEMKNLGLPIRQHYFWPLFCSDSNQ---
Zebrafish	KDLGLHSTPLQFTLQCALEGKKTSLSIGLMKRMKAESLPIKPHYFLPLFAHHHKDKNIPA
Homo_sapiens	KEVQMHSFPLQFTLHCALLANKTDLAKALMKAKEEGFPIRPHYFWPLLGVGRREKVNQG
Newt	-----
Tortoise	KEANMHSSPLQFTLYCALESKKAALAINLMKTMKEEGLPLRPHYSWPLLGVGYQKEKNVQG
Turtle	-----

Drosophila_mojavensis	VLDVVRRM QODFALNP NSETVRD YVIPNLKEKNWDRV VTILRDAGVPN STAVSS AVYA AL
Zebrafish	IIEVLRGM-QEMSVSPDVDAFSFYI LPSFPSL DNAK ASLKEA GVDVN T DGLIV AELRVQA
Homo_sapiens	IIEI LKG M-QELGVH PDQETYTDYV I PCFD SVNSARAILQENGCL SDM FSQAGL RSEA
Newt	-----
Tortoise	TIGVLKAM-HELGVEPD VETY TNVLTNFDDI QTFR ALLQENGCP FETE GLSVA ALRHEA
Turtle	-----
 Drosophila_mojavensis	 TTHHIADA AKIMEQN RAYMPL-I FRQPL LALA HTND FAS FIR CVRQV YEG LQL RAG KE
Zebrafish	YSGNLAKL LSLM SSSPA LPTIDLSV FRAGL IAG FKRF QNV ENMA KIT ELLY K-----
Homo_sapiens	ANGN LDFV LSFL KSNT LP- ISLQSIRSS LLLG FRRS MNIN LWSE IT ELLY K-----
Newt	-----
Tortoise	IYGTLEN VLSL SS PSMPSV NL SHFK GCLIF GF KR SNDI DLW SKI T ELLY K-----
Turtle	MHSV DLFH FK GCLIS G FRRS KNV DLW SKI T ELLY K-----
 Drosophila_mojavensis	 ATEEEAAEA ATDGEAPAT PERQTP DVVGQ IYDAAI YF RRER VPTE KILO GLV NQ-----G
Zebrafish	-----DARFAADG-HDASEN VGYFL YNLID SMSET ELSNEEKIREYFG LKSMN
Homo_sapiens	-----DGRY CQEP-RGP TEAVG YFL YNLID SMSD SEVQ AKEEHL RQYFH QLK MN
Newt	-----DGRY CQTP-PGP AEAVG YFL YNLID SMSD SEVQ AKEEHL RQYFH QLK MN
Tortoise	-----DGRY CQTP-PGP SEAVG YFL YNLID SMSD SEVQ AKEEHL RQYFH QLK MN
Turtle	-----
 Drosophila_mojavensis	 LSLSSG KATAL SELL GSE MTPKIAEQLGKL TSG--ELEPIPLPNS GKRS LD TL TIDE LER
Zebrafish	INISVN IFRG IRNILE SHHV PELV KEAL TLV DKT DDMT EVMM FRS SE-----GRIS ALVK
Homo_sapiens	VKI PENI YRG IRN ILLES YHV PELIK DAH LLV ES---KNL DFQ KTV Q-----L TSSE LEES
Newt	-----
Tortoise	VVI PTNI YRG IRN LLAA YHV PELIK DI I LVT DR-EKLS VADIP KNT E-----VSA LEES
Turtle	IVIPP NI YRG IRN VLDS YHV PELIK D IML LVVD S-EKLS AADIP KNT E-----LEV LSLEE
 Drosophila_mojavensis	 FIKNVEAKGENSNNIKRQLLNACFRSQNLEKTLQVIERLEKDTFQIPIG IY AQLV DLY TH
Zebrafish	T LAE QKAEGK P A HLT KKL INVL SIEEK LQ A D LK SKY EDD---MTPA AYAT LIN LCCR
Homo_sapiens	TLET LKAEN QPI RDV LK LIL VLC SEEN MQ KALE LK AKE YES D---MVTGGY AALIN LCCR
Newt	-----LEV KGK YEAD---MV VGGY AALIN ACCR
Tortoise	ELEKLKA EKQPI GNVL KQLI VALCA EEN MQ KALE V KA KY EPD---MV VGGY AALIN LCCR
Turtle	ELEKLKA EKQPI GNVL KQLI VALCA EEN MQ KALE V KA KY EPD---MV VGGY AALIN LCCR
	* :: . * * : . * * * :: * : .
 Drosophila_mojavensis	 HKRTSD ALAQYNKL RTAD ASFKL DNFK TVRL AD LLLQ EER VDE AL QLL K-E-QKEA AVE A
Zebrafish	HDNAEEA LKL KI EMARK DSEVAL DAQ KYI AL V RVLS KH G KLE AL DIL KEM KE KNIM IRD
Homo_sapiens	HD KV EDAL NLKEE FD RL DSSA VLD TGK YV GL V RL VLA HG K LQ DAIN IL KEM KE KDV LIKD
Newt	HD NV EEAL NLK REV H VRK DSSVAL DTN KYL SLV K VCA HK GR L DAIN IL KEM KE KDV LIKD
Tortoise	HD NV EDAM NLK EEV YRK DSSVAL DTN KYL ALV K VLG K HG R LE DAIN IL KEM KE KDV PI RD
Turtle	HD NAED AM NLK EEV YRK DSSVAL DTN KYL ALV K VLG K HG R LE DAIN IL KEM KE KDV PI RD
	* . . . * : . . * * * : * . : . : . : . : . : . : . : . : . : . : . :
 Drosophila_mojavensis	 GEGSF NYVST VWR ILNSIA ETGNA ER LKV VF DAL VAGN YV VPTN VLL GPLI KV H LSR DD
Zebrafish	----NLIGFL TFTM NSIA MKGD ADAIR RLQ ETI FTGLA KPSN GLC SP LVTC YLES GDH
Homo_sapiens	----TT AL SF FHML NG AAL RG EIE TVKOL HEA I VTL GLA E P STN ISFP LVTV HLE KGD L
Newt	----TT LG SFF H VLN G VAM RG E VET VNR LLEV I VTL GLA KPVAN LCSP VV T V HLE KDD
Tortoise	----TT VTS FF HIL NAA AM RG E VET VNR LHE SI LT GLA K P STN LCT PLI T V HLE KDD
Turtle	----TT VTS FF HIL NAA AM QG E VET VNR LHE SI LT GLV K P STN LC SPLIT V HLE KDD
	. . : * . * * : : . : . : . : . : . : . : . : . : . : . : . : . : . : . :
 Drosophila_mojavensis	 PKA IDA FEQ IC QQY KAT PWK NEL ACRLI Q KEDA ANL QRL TD LST SI HGE VN SLY DLV SF
Zebrafish	AGAF DAV MECH KQ YN QLP KI HDL MCS LVE KG D ALL LQ KV MF LT LER GE MM LY DLF FAF
Homo_sapiens	STA LEVA ID CYK YK VLP RI HDV LCK LVE K GET DL IQ KAM DF VS Q E Q GEM V MLY DLF FAF
Newt	PA ALEA IS DCY K KYK VGP RL HDI CL CRL IE QGN TELL Q KAM DF VS Q E R GEM MLY DIF FAF
Tortoise	PA ALEA IS DCY K KYK VGP RL HDI CL CRL IE QGN TELL Q KAM DF VS Q E R GEM MLY DIF FAF
Turtle	PA ALEA IS DCY K KYK VGP RL HDI CL CRL IE QGN TELL Q KAM DF VS Q E R GEM MLY DIF FAF
	. * :: . : : * . : : * . : : * . : : * . : . : . : * . : . : . : . : . : . :

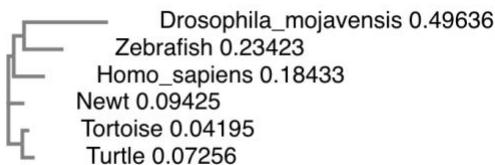
Drosophila_mojavensis	VECGRVRQARKILETPGLRTRPQRISNACDRYKNEGMLQPLEGLIEATKDLGHIDRNKIY
Zebrafish	LQTGRNREREAKIIETPGLRARSNLQWFAEKCIAAQOMEPLENLVDMTVKLFECDRDEMY
Homo_sapiens	LQTGNYKEAKKIIETPGIRARSARLQWFCDCRVANNQVETLEKVELTQQLFECDRDQMY
Newt	LQTAKYKEAKKIIETPGLRARPGRLOWFAEKCITGNQMETLENFVEMTSKLFECDRDEMY
Tortoise	LNTGKYKEAKKIIETPGLRARPGLQWFAEKCIATNQMETLENMVEMTQQLFECDRDQMY
Turtle	LNTGKYKEAKKIIETPGLRARPGRLOWFAEKCIANNQMDILENMVEMTEKLFECDRDQMY :: .. .*:***:****:*. : .. : .. :: ** ::: * .* ***:*
 Drosophila_mojavensis	 YTLLLSYDKADETEKALGLWTKMQEEGVPTDAFLLKLAELLKRKNIDVPFVVPEI---
Zebrafish	HYLFRLCKETNDWRKAEAIWMKMQEENLAPRERTLRLLAELIKSNNQEVFVPKVWFED
Homo_sapiens	YNLLKLKYKINGDWQRADAVWNKIQEENVIPREKTLRLLAELILREGNQEVPDFVPEIWYED
Newt	FYLLKLCKENNEWQKADAIIWTKMQUEENLIPRERTLKLADLFKE-----
Tortoise	YYLLKLCKISNDWRKAEATWTKMQUEENVIPRARTLRLLAIDLKNNGQEVPDFVPEIWYED
Turtle	YYILQLCKISSDWRRADATWTKMQUEENVIPRERTLRLLAIDLKNNGQEVPDFVPEIWYED . : . . : * . * * :***:.* * * **: :
 Drosophila_mojavensis	 -----QQPKSKRNKPAK-----TEANVAEQVQSPPEPKVE
Zebrafish	DKGQSEVVPKEEESAVAKTDDRSRLNATIRLKLQSLCKGEAQEAFDILKEVDSKGIV
Homo_sapiens	EKHSLNS----SSASTTEPDFQK-----DILIACRNLQKKGAYDIFLNAKEQNIV
Newt	-----HV-----ESASVLENNPEK-----KVLMCKGSVQEAYNILLEAQKKDIM
Tortoise	-----CV-----ESAPVSENNPEK-----KILILCKGNIQEAYNILLEEQKKDIM
Turtle	
 Drosophila_mojavensis	 PKAKAVPA-AKPISNIAGFRRAIQANDPDAIASFKERVLSGDKFNVLDTSRILLVRAD
Zebrafish	PGPAIYDAIIKALLAKGNIEDAISVKD-----IAVGHIPSFILSDVANNLLISHVKKC
Homo_sapiens	FNAETYSNLIKLLMSDYFTQAMEVKA-----FAETHIKGFTLNDAAANSRLIITQVRD
Newt	-----MVRKFHL-----
Tortoise	FPASAYSTLIKALLAEGCLEEALKVKN-----IAETHIKGFTLNDAAASSLLIITQVRD
Turtle	FPSFTYSILIKALLAEGCLEKAINVKN-----IAETHIKGFTLNDAAASSLLIITQVRD : :
 Drosophila_mojavensis	 RLSEATKYVEELLAEKQHPQPKIFKFYLNKIAASGDLEVQMRIGQQLNDEQKRL---VS
Zebrafish	QMKDSVQLRDMKLADQMPSQLAITRLVQGLADEGNLKD1QVEVEAMTK-AFGSFNLSNML
Homo_sapiens	YLKEAVTTLKTVLDQQQTPSRLAVTRVIQALAMKGDVENIEVVKMLNGLEDSIGLSKVM
Newt	-----MFLRIGMKK--
Tortoise	YLKDAISTLKKVLESDMVPMSLAVIRLIQALAMKGDLESICTVEKMVENLATSIGLSRML
Turtle	YLKDAISTLKAVALENDMVPTRLAVTRLIQALAMKGDVESVRTVEKMVENLAVSIGLSRML :
 Drosophila_mojavensis	 FDNRYCHAYIVAGKAEQFLKQLSTEIEAVSSEEAG--KLAEKPRGGAVGILEKHPELI
Zebrafish	FVNNTALALLNNNGDVDSAVIDMLQTFYT---ENTERQNNSTIAHFRKVLNANNDAAMDKLS
Homo_sapiens	FINNIALAQIKNNNIDAIAENIENMLSENKVIEPQYFGLAYLFRKVIEEQLEPAVEKIS
Newt	-----LQHQK-----
Tortoise	FVNNTVLAHIKK-----
Turtle	FVNNTVLAHIKNNNLDATAVEYIESLIISGMQNPDSPITSISYVFRKVIEEKLESALEKLS
 Drosophila_mojavensis	 AQYETLAEKYAAHNQLGPMNVLWMHLISNGQEVAASKQIWDK--HLSSAPRLMFQRVLQTA
Zebrafish	AMAERLCNQFASYR---AATDLFLVYVDTGRTEEAKFLLQRCAAVGEQKDLLFSYVLRAS
Homo_sapiens	IMAERLANQFAIYK---PVTDFFQLVDAGKVDDARALLQRCGAIAEQTPILLFLLRNS
Newt	-----
Tortoise	-----
Turtle	AMAERLVNQFGIYK---PATDLFLQYVSEGRIDDARLLLQRCGGIAEQQKIMMAFIAKSS
 Drosophila_mojavensis	 REQQDEKLASTVISQLRGSKISEGAIGNAYSCLIDIOTTKGNTDKALEVLANAICKDVSLE
Zebrafish	QOPGQAQKVMSLMELIPDIREKE---DIYSQLMKCHGLDQDLASAKALYERMQVEGV--
Homo_sapiens	RKQGKASTVKSYLEIPELNEKE---EAYNSLMKSYVSEKDVTSAKALYEHLTAKNT--
Newt	-----
Tortoise	-----
Turtle	OKPGQGQKIKMLLDLVPDFPEME---VVYSYLLKCHVLDNDIVSAKTLFEKAKAENI--

Drosophila_mojavensis	NINRTALLRLKQAVEEKSQQFPYTVPEKRAKAEDSSSSSSSSDDVTPKRPEAPTRPE
Zebrafish	RIDELTLKRLAKLYRDSGEPVFEEPAE-----SFRFYADKLKDQRTQSTASIDN
Homo_sapiens	KLDDLFLKRYASLLKYAGEPVPFIEPPE-----SFEFYAQQLRKLRENSS-----
Newt	-----
Tortoise	-----
Turtle	RTDEIFLKRLAFLLKSAAGEPVPFTEPPE-----SFHFYVNKLRKERQELSSHVD-

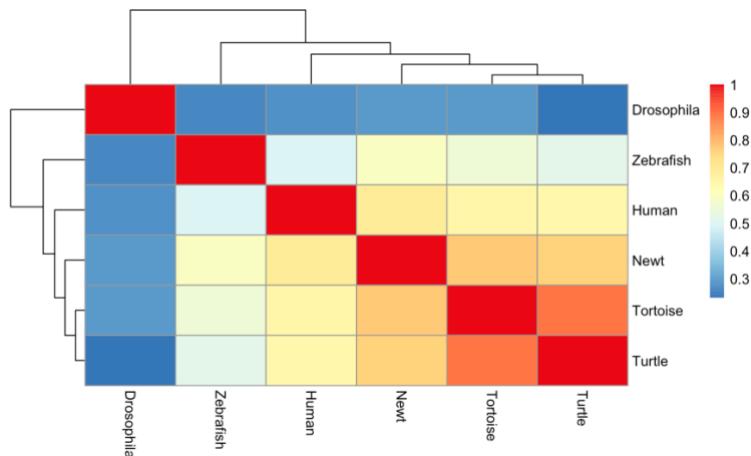
Drosophila_mojavensis	RV
Zebrafish	--
Homo_sapiens	--
Newt	--
Tortoise	--
Turtle	--

[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 Phyliip). Paste an image of your Cladogram or tree output in your report.

Obtained from EBI Simple Phylogeny website:



[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.



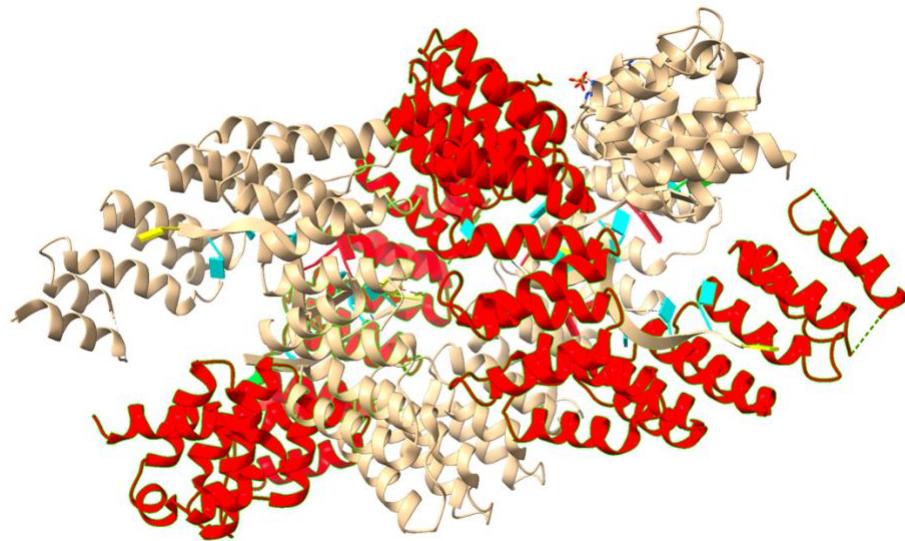
[Q8] Using R/Bio3D (or an online blast server if you prefer), search the main protein structure database for the most similar atomic resolution structures to your aligned sequences.

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).

ID	Technique	Resolution	Source	Evalue	Identity
6EEN	X-RAY DIFFRACTION	2.01	Zea mays	2.12e-14	25.338
5I9H	X-RAY DIFFRACTION	2.50	Escherichia coli BL21 (DE3)	8.80e-13	22.321
4M59	X-RAY DIFFRACTION	2.46	Zea mays	2.53e-11	25.188

[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?

It is unlikely to be similar in structure to *Zea mays* LRPPRC protein given the low sequence similarity (~25%). In the figure below LRPPRC chain A is colored red.



[Q10] Perform a “Target” search of ChEMBL (<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?

	E-Value	Positives %	Identities %	Score (bits)	Score	Length	ChEMBL ID	Name	UniProt Accessions	Type	Organism	Compounds	Activities
<input type="checkbox"/>	1.3e-150	84.2	71.4	456.833	1174	1394	CHEMBL4295762	Leucine-rich PPR motif-containing protein, mitochondrial	P42704	SINGLE PROTEIN	Homo sapiens	1 By Mol. Wt.	1 By Std. Type:

Target Report Card

Name And Classification

ID:	CHEMBL4295762
Type:	SINGLE PROTEIN
Preferred Name:	Leucine-rich PPR motif-containing protein, mitochondrial
Synonyms:	130 kDa leucine-rich protein GP130 Leucine-rich PPR motif-containing protein, mitochondrial LRP130 LRP 130 LRPPRC
Organism:	Homo sapiens
Species Group:	No
Protein Target Classification:	- Unclassified protein

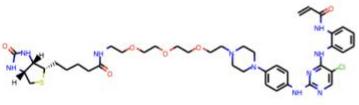
ChEMBL ID: CHEMBL4295762

(https://www.ebi.ac.uk/chembl/target_report_card/CHEMBL4295762/)

No ligand efficiency data.

Binding assay linked to a small molecule compound (CHEMBL4129274) that was showed to act as a multi-targeting TAK1-centered inhibitors for cancer and other diseases.

DOI: <http://dx.doi.org/10.1016%2Fj.bmc.2016.11.034>

	ID: CHEMBL4129274 Name: Undefined Max Phase: 0 Research ⓘ Molecular Formula: C41H55ClN10O6S Molecular Weight: 851.48 Molecule Type: Small molecule
---	---