Gaeun Jun
A16814573
gajun@ucsd.edu

The Find-a-Gene Project

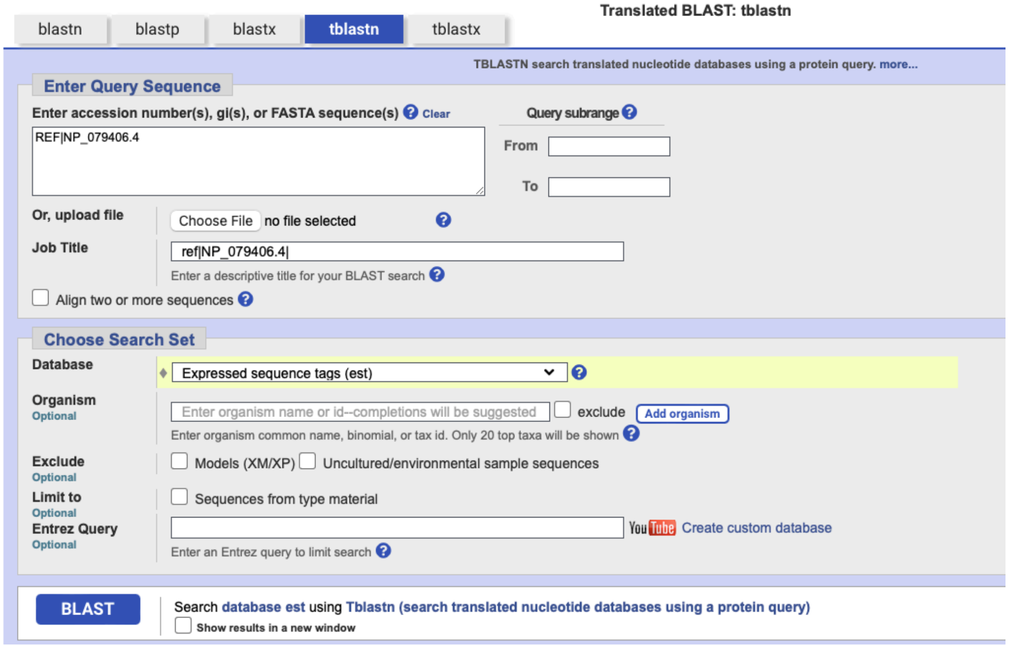
[Q1] Tell me the name of a protein you are interested in. Include the species and the accession number. This can be a human protein or a protein from any other species as long as its function is known.

Name: Hexokinase HKDC1
Accession: NP_079406.4
Species: Homo Sapiens
Function: Enzyme that catalyzes the phosphorylation reaction from glucose to glucose 6-phosphate in anaerobic glycolysis.

[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 (est)

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.



Translated BLAST: tblastn

TBLASTN search translated nucleotide databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

REF|NP_079406.4

Query subrange

From

To

Or, upload file: Choose File no file selected

Job Title: ref|NP_079406.4|

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Expressed sequence tags (est)

Organism (Optional): Enter organism name or id--completions will be suggested □ exclude Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude (Optional): □ Models (XM/XP) □ Uncultured/environmental sample sequences

Limit to (Optional): □ Sequences from type material

Entrez Query (Optional): YouTube Create custom database

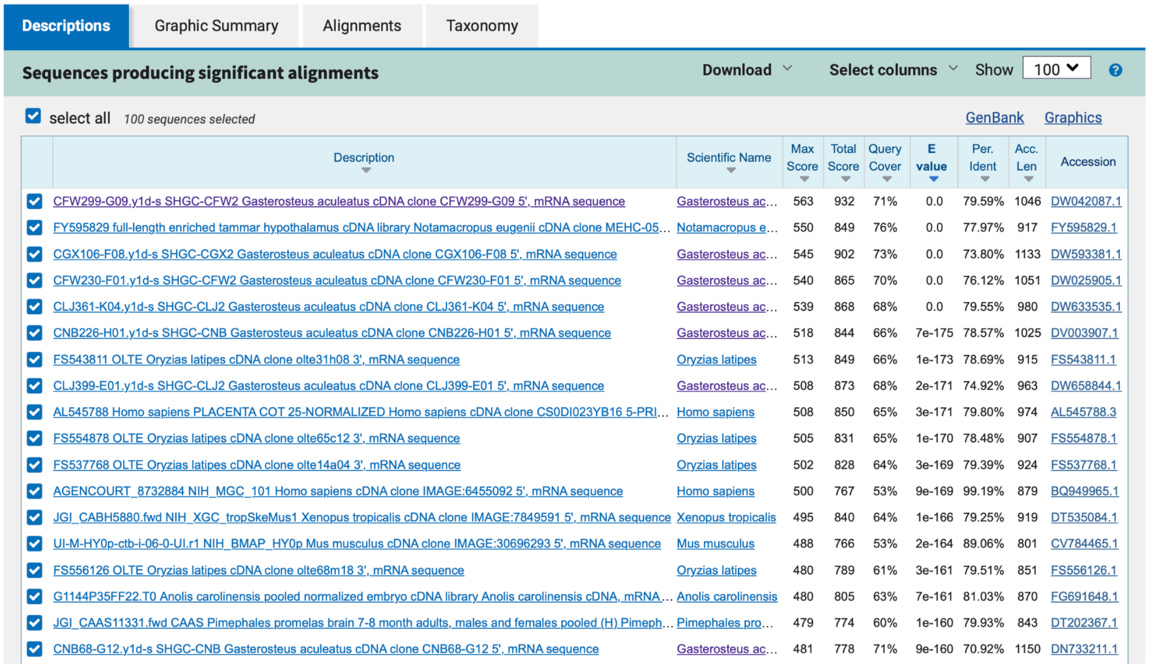
Enter an Entrez query to limit search

BLAST | Search database est using Tblastn (search translated nucleotide databases using a protein query)

□ Show results in a new window

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 DW042087.1, a 1046 base pair clone from *Gasterosteus aculeatus*. See below for alignment details.

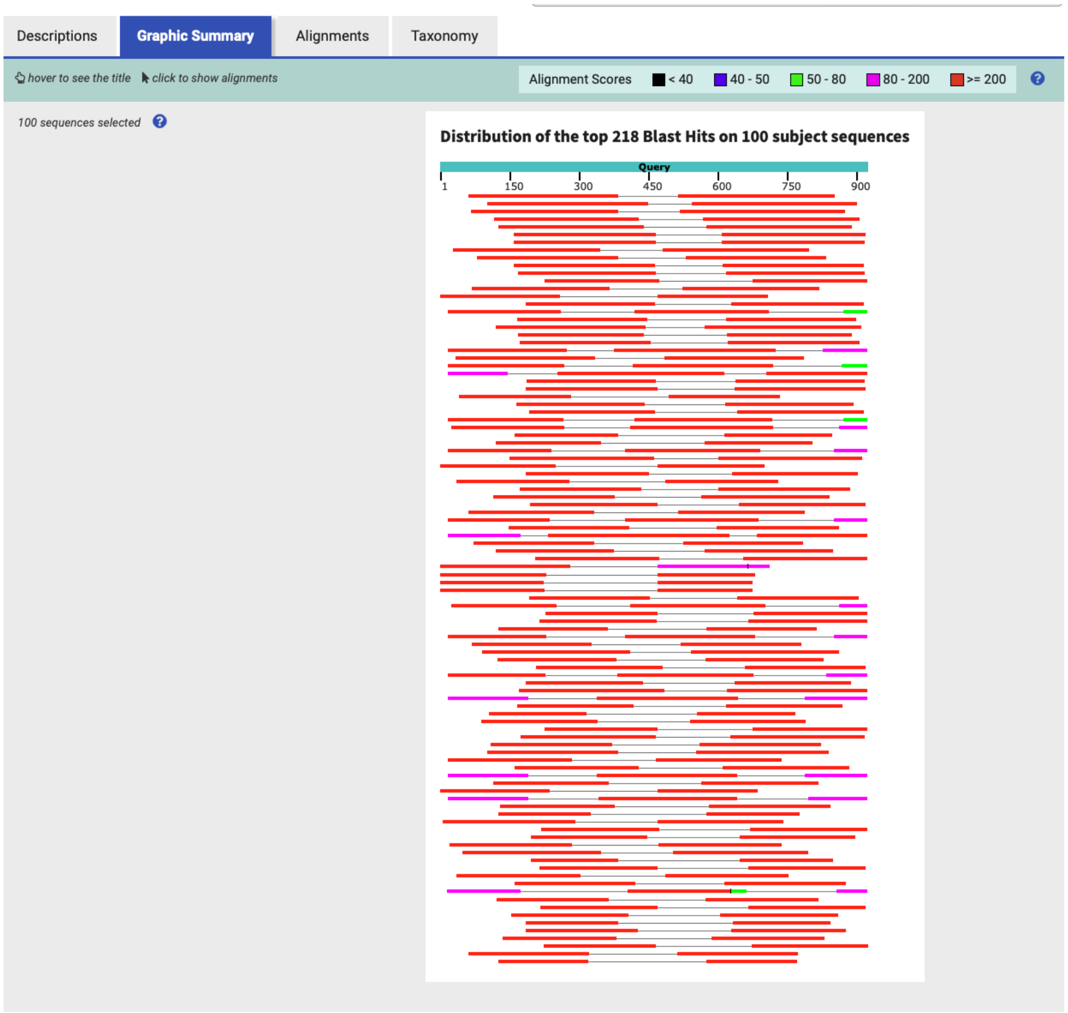


Sequences producing significant alignments

Show 100

☑ select all 100 sequences selected

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	CFW299-G09.y1d-s SHGC-CFW2 Gasterosteus aculeatus cDNA clone CFW299-G09 5', mRNA sequence	Gasterosteus ac...	563	932	71%	0.0	79.59%	1046	DW042087.1
☑	FY595829 full-length enriched tammar hypothalamus cDNA library Notamacropus eugenii cDNA clone MEHC-05...	Notamacropus e...	550	849	76%	0.0	77.97%	917	FY595829.1
☑	CGX106-F08.y1d-s SHGC-CGX2 Gasterosteus aculeatus cDNA clone CGX106-F08 5', mRNA sequence	Gasterosteus ac...	545	902	73%	0.0	73.80%	1133	DW593381.1
☑	CFW230-F01.y1d-s SHGC-CFW2 Gasterosteus aculeatus cDNA clone CFW230-F01 5', mRNA sequence	Gasterosteus ac...	540	865	70%	0.0	76.12%	1051	DW025905.1
☑	CLJ361-K04.y1d-s SHGC-CLJ2 Gasterosteus aculeatus cDNA clone CLJ361-K04 5', mRNA sequence	Gasterosteus ac...	539	868	68%	0.0	79.55%	980	DW633535.1
☑	CNB226-H01.y1d-s SHGC-CNB Gasterosteus aculeatus cDNA clone CNB226-H01 5', mRNA sequence	Gasterosteus ac...	518	844	66%	7e-175	78.57%	1025	DV003907.1
☑	FS543811 OLTE Oryzias latipes cDNA clone olte31h08 3', mRNA sequence	Oryzias latipes	513	849	66%	1e-173	78.69%	915	FS543811.1
☑	CLJ399-E01.y1d-s SHGC-CLJ2 Gasterosteus aculeatus cDNA clone CLJ399-E01 5', mRNA sequence	Gasterosteus ac...	508	873	68%	2e-171	74.92%	963	DW658844.1
☑	AL545788 Homo sapiens PLACENTA COT 25-NORMALIZED Homo sapiens cDNA clone CS0DI023YB16 5-PRI...	Homo sapiens	508	850	65%	3e-171	79.80%	974	AL545788.3
☑	FS554878 OLTE Oryzias latipes cDNA clone olte65c12 3', mRNA sequence	Oryzias latipes	505	831	65%	1e-170	78.48%	907	FS554878.1
☑	FS537768 OLTE Oryzias latipes cDNA clone olte14a04 3', mRNA sequence	Oryzias latipes	502	828	64%	3e-169	79.39%	924	FS537768.1
☑	AGENCOURT_8732884 NIH_MGC_101 Homo sapiens cDNA clone IMAGE:6455092 5', mRNA sequence	Homo sapiens	500	767	53%	9e-169	99.19%	879	BQ949965.1
☑	JGI_CABH5880.fwd NIH_XGC_tropSkeMus1 Xenopus tropicalis cDNA clone IMAGE:7849591 5', mRNA sequence	Xenopus tropicalis	495	840	64%	1e-166	79.25%	919	DT535084.1
☑	UI-M-HY0p-ctb-i-06-0-UI.r1 NIH_BMAP_HY0p Mus musculus cDNA clone IMAGE:30696293 5', mRNA sequence	Mus musculus	488	766	53%	2e-164	89.06%	801	CV784465.1
☑	FS556126 OLTE Oryzias latipes cDNA clone olte68m18 3', mRNA sequence	Oryzias latipes	480	789	61%	3e-161	79.51%	851	FS556126.1
☑	G1144P35FF22.T0 Anolis carolinensis pooled normalized embryo cDNA library Anolis carolinensis cDNA, mRNA...	Anolis carolinensis	480	805	63%	7e-161	81.03%	870	FG691648.1
☑	JGI_CAAS11331.fwd CAAS Pimephales promelas brain 7-8 month adults, males and females pooled (H) Pimeph...	Pimephales pro...	479	774	60%	1e-160	79.93%	843	DT202367.1
☑	CNB68-G12.y1d-s SHGC-CNB Gasterosteus aculeatus cDNA clone CNB68-G12 5', mRNA sequence	Gasterosteus ac...	481	778	71%	9e-160	70.92%	1150	DN733211.1



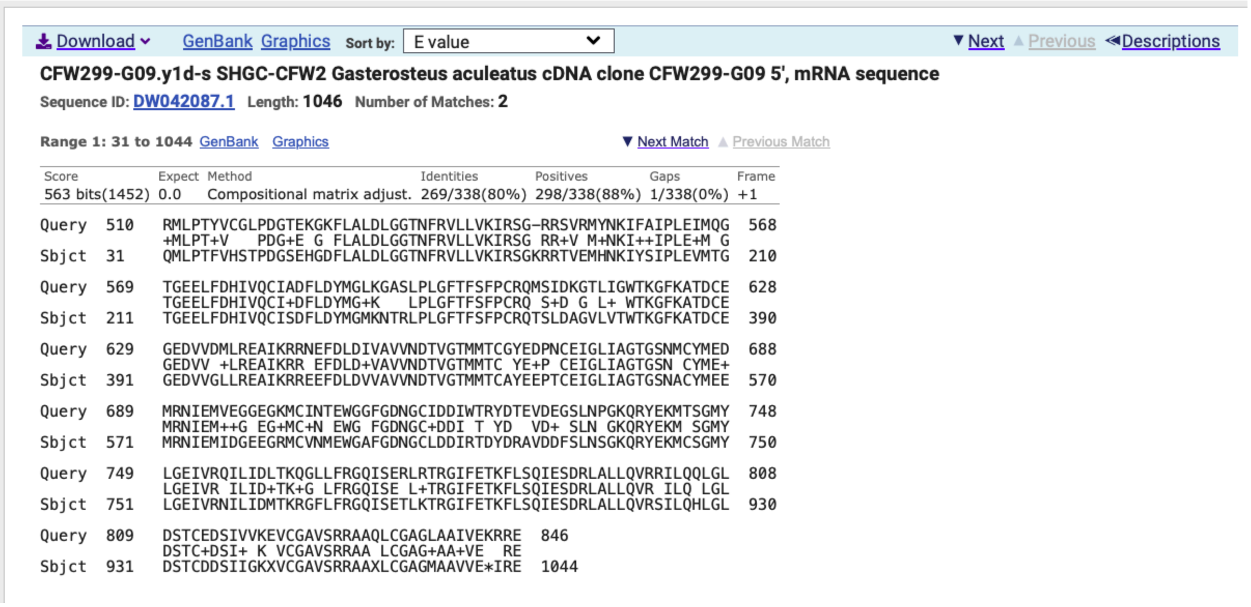
Descriptions | Graphic Summary | Alignments | Taxonomy

hover to see the title click to show alignments

Alignment Scores ■ < 40 ■ 40 - 50 🟩 50 - 80 ■ 80 - 200 🟥 >= 200

100 sequences selected

Distribution of the top 218 Blast Hits on 100 subject sequences



Download GenBank Graphics Sort by: E value ▼ Next ▲ Previous ◀Descriptions

CFW299-G09.y1d-s SHGC-CFW2 Gasterosteus aculeatus cDNA clone CFW299-G09 5', mRNA sequence

Sequence ID: DW042087.1 Length: 1046 Number of Matches: 2

Range 1: 31 to 1044 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
563 bits(1452)	0.0	Compositional matrix adjust.	269/338(80%)	298/338(88%)	1/338(0%)	+1

```
Query  510    RMLPTYVCGLPDGTEKGKFLALDLGGTNFRVLLVKIRSG-RRSVRMYNKIFAIPLEIMQG  568
              +MLPT+V   PDG+E G FLALDLGGTNFRVLLVKIRSG RR+V M+NKI++IPLE+M G
Sbjct  31     QMLPTFVHSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTVEMHNKIYSIPLEVMTG  210

Query  569    TGEELFDHIVQCIADFLDYMGLKGASLPLGFTFSFPCRQMSIDKGTLIGWTKGFKATDCE  628
              TGEELFDHIVQCI+DFLDYMG+K   LPLGFTFSFPCRQ S+D G L+ WTKGFKATDCE
Sbjct  211    TGEELFDHIVQCISDFLDYMGMKNTRLPLGFTFSFPCRQTSLDAGVLVTWTKGFKATDCE  390

Query  629    GEDVVDMLREAIKRRNEFDLDIVAVVNDTVGTMMTCGYEDPNCEIGLIAGTGSNMCYMED  688
              GEDVV +LREAIKRR EFDLD+VAVVNDTVGTMMTC YE+P CEIGLIAGTGSN CYME+
Sbjct  391    GEDVVGLLREAIKRREEFDLDVVAVVNDTVGTMMTCAYEEPTCEIGLIAGTGSNACYMEE  570

Query  689    MRNIEMVEGGEGKMCINTEWGGFGDNGCIDDIWTRYDTEVDEGSLNPGKQRYEKMTSGMY  748
              MRNIEM++G EG+MC+N EWG FGDNGC+DDI T YD  VD+ SLN GKQRYEKM SGMY
Sbjct  571    MRNIEMIDGEEGRMCVNMEWGAFGDNGCLDDIRTDYDRAVDDFSLNSGKQRYEKMCSGMY  750

Query  749    LGEIVRQILIDLTKQGLLFRGQISERLRTRGIFETKFLSQIESDRLALLQVRRILQQLGL  808
              LGEIVR ILID+TK+G LFRGQISE L+TRGIFETKFLSQIESDRLALLQVR ILQ LGL
Sbjct  751    LGEIVRNILIDMTKRGFLFRGQISETLKTRGIFETKFLSQIESDRLALLQVRSILQHLGL  930

Query  809    DSTCEDSIVVKEVCGAVSRRAAQLCGAGLAAIVEKRRE  846
              DSTC+DSI+ K VCGAVSRRAA LCGAG+AA+VE  RE
Sbjct  931    DSTCDDSIIGKXVCGAVSRRAAXLCGAGMAAVVE*IRE  1044
```

Alignment details:

```
>gb|DW042087.1| CFW299-G09.y1d-s SHGC-CFW2 Gasterosteus aculeatus
cDNA clone CFW299-G09 5', mRNA sequence
Length: 1046

Score = 563 bits (1452), Expect = 0.0, Method: compositional matrix adjust. Identities
= 269/338 (80%), Positives = 298/338 (88%), Gaps = 1/338 (0%)
Frame = +1

Query  510   RMLPTYVCGLPDGTEKGKFLALDLGGTNFRVLLVKIRSG-RRSVRMYNKIFAIPLEIMQG  568
             +MLPT+V   PDG+E G FLALDLGGTNFRVLLVKIRSG RR+V M+NKI++IPLE+M G
Sbjct  31    QMLPTFVHSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTVEMHNKIYSIPLEVMTG  210

Query  569   TGEELFDHIVQCIADFLDYMGLKGASLPLGFTFSFPCRQMSIDKGTLIGWTKGFKATDCE  628
             TGEELFDHIVQCI+DFLDYMG+K   LPLGFTFSFPCRQ S+D G L+ WTKGFKATDCE
Sbjct  211   TGEELFDHIVQCISDFLDYMGMKNTRLPLGFTFSFPCRQTSLDAGVLVTWTKGFKATDCE  390

Query  629   GEDVVDMLREAIKRRNEFDLDIVAVVNDTVGTMMTCGYEDPNCEIGLIAGTGSNMCYMED  688
             GEDVV +LREAIKRR EFDLD+VAVVNDTVGTMMTC YE+P CEIGLIAGTGSN CYME+
Sbjct  391   GEDVVGLLREAIKRREEFDLDVVAVVNDTVGTMMTCAYEEPTCEIGLIAGTGSNACYMEE  570

Query  689   MRNIEMVEGGEGKMCINTEWGGFGDNGCIDDIWTRYDTEVDEGSLNPGKQRYEKMTSGMY  748
             MRNIEM++G EG+MC+N EWG FGDNGC+DDI T YD  VD+ SLN GKQRYEKM SGMY
Sbjct  571   MRNIEMIDGEEGRMCVNMEWGAFGDNGCLDDIRTDYDRAVDDFSLNSGKQRYEKMCSGMY  750

Query  749   LGEIVRQILIDLTKQGLLFRGQISERLRTRGIFETKFLSQIESDRLALLQVRRILQQLGL  808
             LGEIVR ILID+TK+G LFRGQISE L+TRGIFETKFLSQIESDRLALLQVR ILQ LGL
Sbjct  751   LGEIVRNILIDMTKRGFLFRGQISETLKTRGIFETKFLSQIESDRLALLQVRSILQHLGL  930

Query  809   DSTCEDSIVVKEVCGAVSRRAAQLCGAGLAAIVEKRRE  846
             DSTC+DSI+ K VCGAVSRRAA LCGAG+AA+VE  RE

Sbjct  931   DSTCDDSIIGKXVCGAVSRRAAXLCGAGMAAVVE*IRE  1044
```

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.

[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: > three-spined stickleback | EMBOSS_001_1 (sequence taken from EMBOSS Transeq at the EBI)

```
RSGFPG*LPCQMLPTFVHSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTVEMHNKI
YSIPLEVMTGTGEELFDHIVQCISDFLDYMGMKNTRLPLGFTFSFPCRQTSLDAGVLVTW
TKGFKATDCEGEDVVGLLREAIKRREEFDLDVVAVVNDTVGTMMTCAYEEPTCEIGLIAG
TGSNACYMEEMRNIEMIDGEEGRMCVNMEWGAFGDNGCLDDIRTDYDRAVDDFSLNSGKQ
RYEKMCSGMYLGEIVRNILIDMTKRGFLFRGQISETLKTRGIFETKFLSQIESDRLALLQ
VRSILQHLGLDSTCDDSIIGKXVCGAVSRRAAXLCGAGMAAVVE*IRET
```

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.

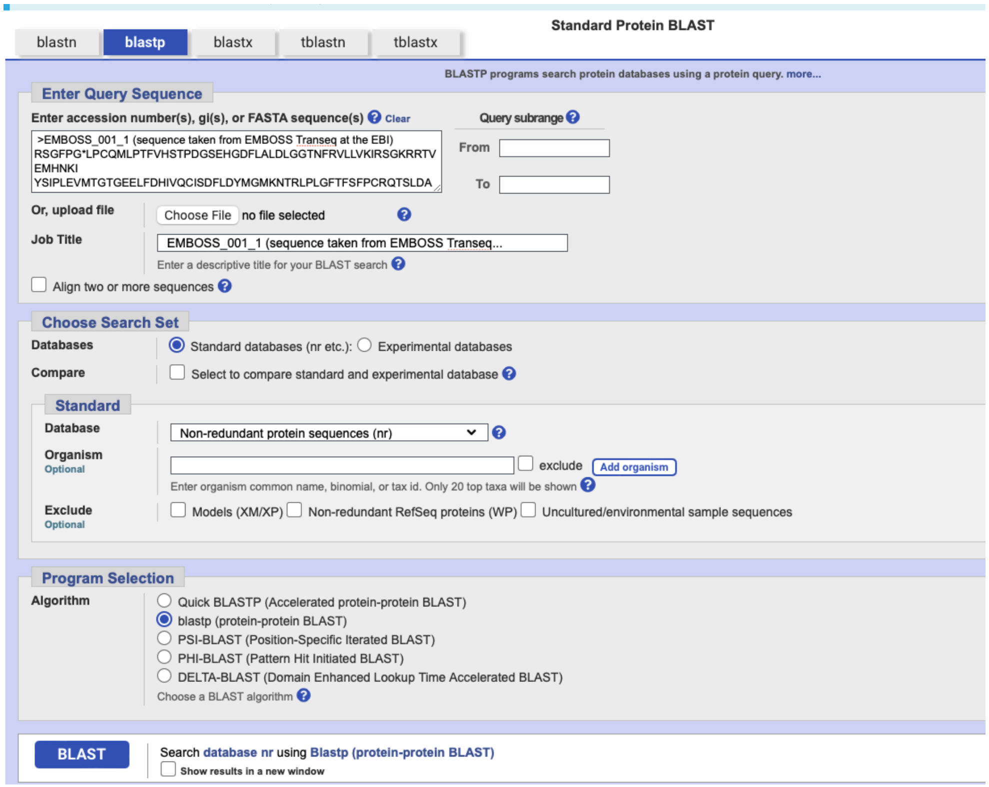
Name: Three-spined stickleback
Species: Gasterosteus aculeatus

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Actinopterygii; Neopterygii; Teleostei; Neoteleostei;
Acanthomorphata; Eupercaria; Perciformes; Cottioidei;
Gasterosteales; Gasterosteidae; Gasterosteus.

[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 yielded a top hit result to a protein from *Gasterosteus aculeatus aculeatus* (three-spined stickleback).
See additional screen shots below for top hits and selected alignment details:



Standard Protein BLAST

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

```
>EMBOSS_001_1 (sequence taken from EMBOSS Transeq at the EBI)
RSGFPG*LPCQMLPTFVHSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTV
EMHNKI
YSIPLEVMTGTGEELFDHIVQCISDFLDYMGMKNTRLPLGFTFSFPCRQTSLDA
```

Query subrange

From

To

Or, upload file: Choose File no file selected

Job Title: EMBOSS_001_1 (sequence taken from EMBOSS Transeq...

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Databases: ● Standard databases (nr etc.): ○ Experimental databases

Compare: □ Select to compare standard and experimental database

Standard

Database: Non-redundant protein sequences (nr)

Organism Optional: □ exclude Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

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

Program Selection

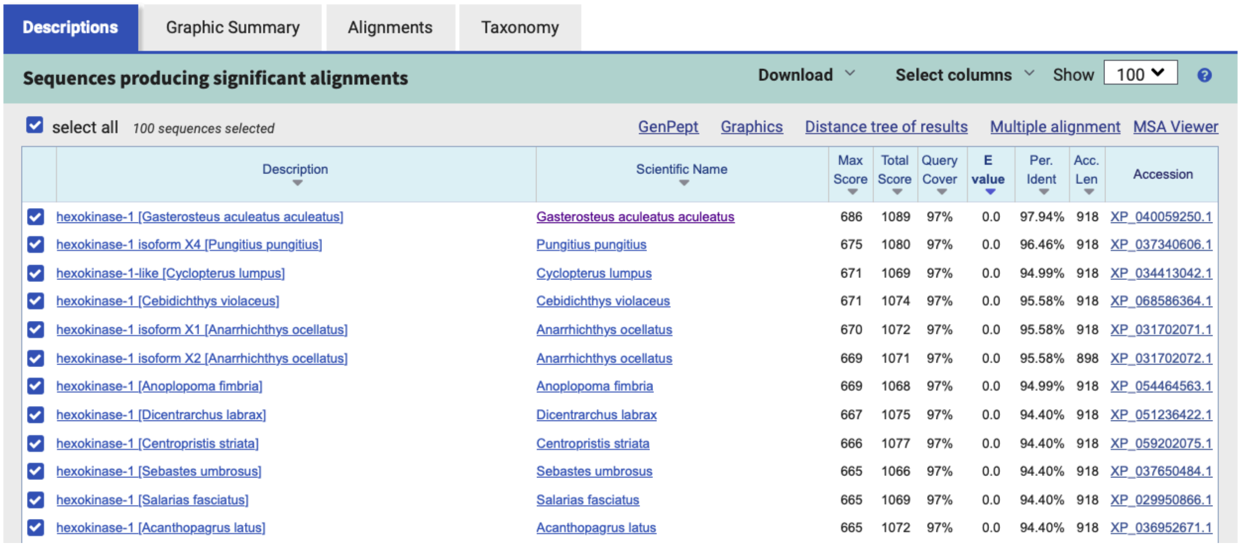
Algorithm:
○ Quick BLASTP (Accelerated protein-protein BLAST)
● blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST — Search database nr using Blastp (protein-protein BLAST)

□ Show results in a new window

The top result is to a protein from *Gasterosteus aculeatus aculeatus* (three-spined stickleback), see second screen shot below for alignment details:

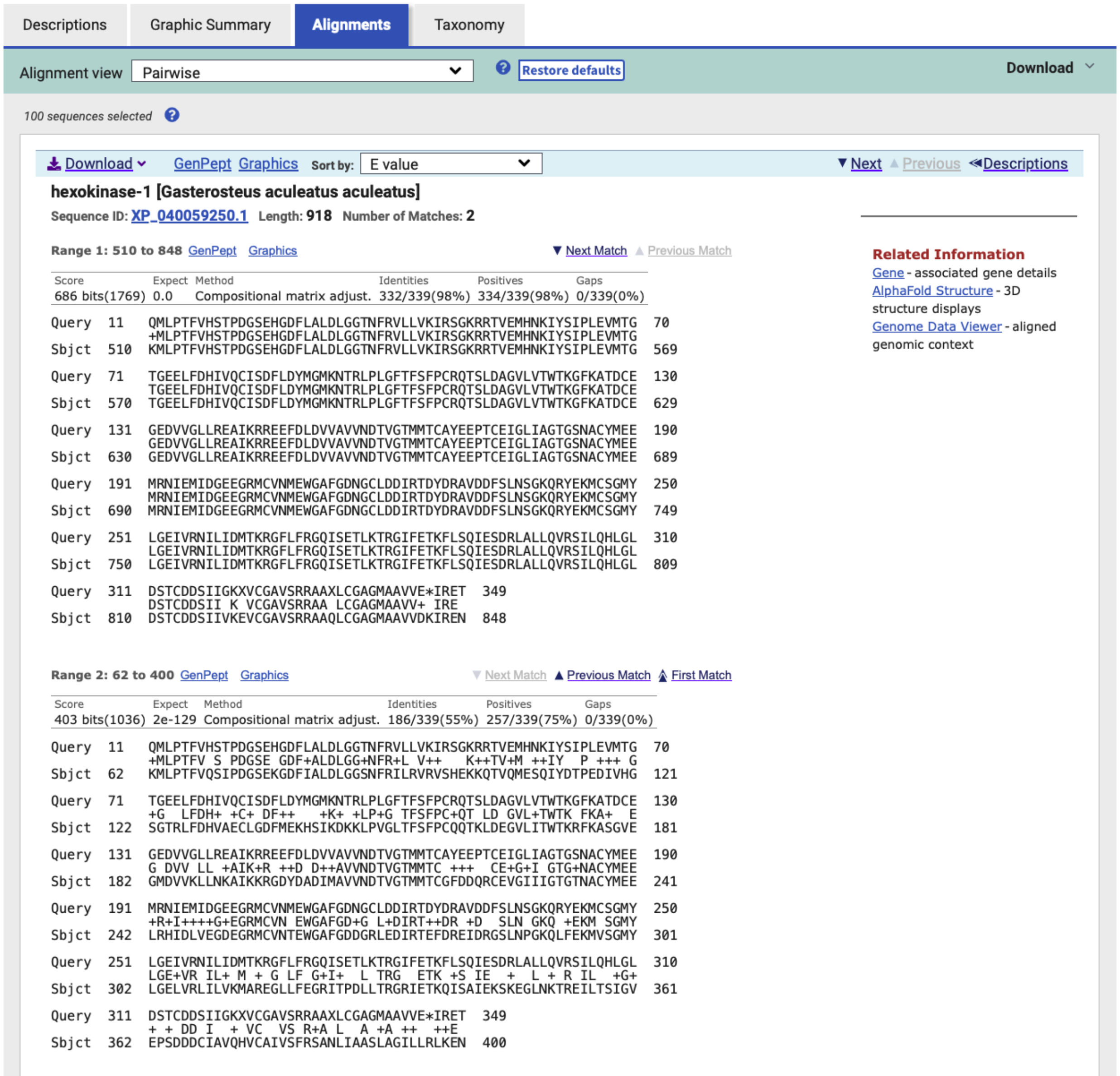


Sequences producing significant alignments

☑ select all 100 sequences selected

GenPept Graphics Distance tree of results Multiple alignment MSA Viewer

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	hexokinase-1 [Gasterosteus aculeatus aculeatus]	Gasterosteus aculeatus aculeatus	686	1089	97%	0.0	97.94%	918	XP_040059250.1
☑	hexokinase-1 isoform X4 [Pungitius pungitius]	Pungitius pungitius	675	1080	97%	0.0	96.46%	918	XP_037340606.1
☑	hexokinase-1-like [Cyclopterus lumpus]	Cyclopterus lumpus	671	1069	97%	0.0	94.99%	918	XP_034413042.1
☑	hexokinase-1 [Cebidichthys violaceus]	Cebidichthys violaceus	671	1074	97%	0.0	95.58%	918	XP_068586364.1
☑	hexokinase-1 isoform X1 [Anarrhichthys ocellatus]	Anarrhichthys ocellatus	670	1072	97%	0.0	95.58%	918	XP_031702071.1
☑	hexokinase-1 isoform X2 [Anarrhichthys ocellatus]	Anarrhichthys ocellatus	669	1071	97%	0.0	95.58%	898	XP_031702072.1
☑	hexokinase-1 [Anoplopoma fimbria]	Anoplopoma fimbria	669	1068	97%	0.0	94.99%	918	XP_054464563.1
☑	hexokinase-1 [Dicentrarchus labrax]	Dicentrarchus labrax	667	1075	97%	0.0	94.40%	918	XP_051236422.1
☑	hexokinase-1 [Centropristis striata]	Centropristis striata	666	1077	97%	0.0	94.40%	918	XP_059202075.1
☑	hexokinase-1 [Sebastes umbrosus]	Sebastes umbrosus	665	1066	97%	0.0	94.40%	918	XP_037650484.1
☑	hexokinase-1 [Salarias fasciatus]	Salarias fasciatus	665	1069	97%	0.0	94.40%	918	XP_029950866.1
☑	hexokinase-1 [Acanthopagrus latus]	Acanthopagrus latus	665	1072	97%	0.0	94.40%	918	XP_036952671.1



100 sequences selected

hexokinase-1 [Gasterosteus aculeatus aculeatus]

Sequence ID: XP_040059250.1 Length: 918 Number of Matches: 2

Range 1: 510 to 848 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
686 bits(1769)	0.0	Compositional matrix adjust.	332/339(98%)	334/339(98%)	0/339(0%)

```
Query  11   QMLPTFVHSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTVEMHNKIYSIPLEVMTG  70
            +MLPTFVHSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTVEMHNKIYSIPLEVMTG
Sbjct  510  KMLPTFVHSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTVEMHNKIYSIPLEVMTG  569

Query  71   TGEELFDHIVQCISDFLDYMGMKNTRLPLGFTFSFPCRQTSLDAGVLVTWTKGFKATDCE  130
            TGEELFDHIVQCISDFLDYMGMKNTRLPLGFTFSFPCRQTSLDAGVLVTWTKGFKATDCE
Sbjct  570  TGEELFDHIVQCISDFLDYMGMKNTRLPLGFTFSFPCRQTSLDAGVLVTWTKGFKATDCE  629

Query  131  GEDVVGLLREAIKRREEFDLDVVAVVNDTVGTMMTCAYEEPTCEIGLIAGTGSNACYMEE  190
            GEDVVGLLREAIKRREEFDLDVVAVVNDTVGTMMTCAYEEPTCEIGLIAGTGSNACYMEE
Sbjct  630  GEDVVGLLREAIKRREEFDLDVVAVVNDTVGTMMTCAYEEPTCEIGLIAGTGSNACYMEE  689

Query  191  MRNIEMIDGEEGRMCVNMEWGAFGDNGCLDDIRTDYDRAVDDFSLNSGKQRYEKMCSGMY  250
            MRNIEMIDGEEGRMCVNMEWGAFGDNGCLDDIRTDYDRAVDDFSLNSGKQRYEKMCSGMY
Sbjct  690  MRNIEMIDGEEGRMCVNMEWGAFGDNGCLDDIRTDYDRAVDDFSLNSGKQRYEKMCSGMY  749

Query  251  LGEIVRNILIDMTKRGFLFRGQISETLKTRGIFETKFLSQIESDRLALLQVRSILQHLGL  310
            LGEIVRNILIDMTKRGFLFRGQISETLKTRGIFETKFLSQIESDRLALLQVRSILQHLGL
Sbjct  750  LGEIVRNILIDMTKRGFLFRGQISETLKTRGIFETKFLSQIESDRLALLQVRSILQHLGL  809

Query  311  DSTCDDSIIGKXVCGAVSRRAAXLCGAGMAAVVE*IRET  349
            DSTCDDSII K VCGAVSRRAA LCGAGMAAVV+ IRE
Sbjct  810  DSTCDDSIIVKEVCGAVSRRAAQLCGAGMAAVVDKIREN  848
```

Range 2: 62 to 400 GenPept Graphics ▼ Next Match ▲ Previous Match ▲ First Match

Score	Expect	Method	Identities	Positives	Gaps
403 bits(1036)	2e-129	Compositional matrix adjust.	186/339(55%)	257/339(75%)	0/339(0%)

```
Query  11   QMLPTFVHSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTVEMHNKIYSIPLEVMTG  70
            +MLPTFV S PDGSE GDF+ALDLGG+NFR+L V++   K++TV+M ++IY  P +++ G
Sbjct  62   KMLPTFVQSIPDGSEKGDFIALDLGGSNFRILRVRVSHEKKQTVQMESQIYDTPEDIVHG  121

Query  71   TGEELFDHIVQCISDFLDYMGMKNTRLPLGFTFSFPCRQTSLDAGVLVTWTKGFKATDCE  130
            +G  LFDH+ +C+ DF++   +K+ +LP+G TFSFPC+QT LD GVL+TWTK FKA+  E
Sbjct  122  SGTRLFDHVAECLGDFMEKHSIKDKKLPVGLTFSFPCQQTKLDEGVLITWTKRFKASGVE  181

Query  131  GEDVVGLLREAIKRREEFDLDVVAVVNDTVGTMMTCAYEEPTCEIGLIAGTGSNACYMEE  190
            G DVV LL +AIK+R ++D D++AVVNDTVGTMMTC +++  CE+G+I GTG+NACYMEE
Sbjct  182  GMDVVKLLNKAIKKRGDYDADIMAVVNDTVGTMMTCGFDDQRCEVGIIIGTGTNACYMEE  241

Query  191  MRNIEMIDGEEGRMCVNMEWGAFGDNGCLDDIRTDYDRAVDDFSLNSGKQRYEKMCSGMY  250
            +R+I++++G+EGRMCVN EWGAFGD+G L+DIRT++DR +D  SLN GKQ +EKM SGMY
Sbjct  242  LRHIDLVEGDEGRMCVNTEWGAFGDDGRLEDIRTEFDREIDRGSLNPGKQLFEKMVSGMY  301

Query  251  LGEIVRNILIDMTKRGFLFRGQISETLKTRGIFETKFLSQIESDRLALLQVRSILQHLGL  310
            LGE+VR IL+ M + G LF G+I+  L TRG  ETK +S IE  +  L + R IL  +G+
Sbjct  302  LGELVRLILVKMAREGLLFEGRITPDLLTRGRIETKQISAIEKSKEGLNKTREILTSIGV  361

Query  311  DSTCDDSIIGKXVCGAVSRRAAXLCGAGMAAVVE*IRET  349
            + + DD I  + VC  VS R+A L  A +A ++  ++E
Sbjct  362  EPSDDDCIAVQHVCAIVSFRSANLIAASLAGILLRLKEN  400
```

[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 alignment for building a phylogenetic tree that illustrates species divergence.

```
>Human_hexokinase|NP_079406.4:510-846 hexokinase HKDC1 [Homo sapiens]
RMLPTYVCGLPDGTEKGKFLALDLGGTNFRVLLVKIRSGRRSVRMYNKIFAIPLEIMQGTGEELFDHIVQCIADFLDYM
GLKGASLPLGFTFSFPCRQMSIDKGTLIGWTKGFKATDCEGEDVVDMLREAIKRRNEFDLDIVAVVNDTVGTMMTCGYE
DPNCEIGLIAGTGSNMCYMEDMRNIEMVEGGEGKMCINTEWGGFGDNGCIDIWTRYDTEVDEGSLNPGKQRYEKMTSGM
YLGEIVRQILIDLTKQGLLFRGQISERLRTRGIFETKFLSQIESDRLALLQVRRILQQLGLDSTCEDSIVVKEVCGAVS
RRAAQLCGAGLAAIVEKRR

>Three_spined_stickleback|XP_040059250.1:510-848 hexokinase-1 [Gasterosteus
aculeatus aculeatus]
KMLPTFVHSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTVEMHNKIYSIPLEVMTGTGEELFDHIVQCISDFLDY
MGMKNTRLPLGFTFSFPCRQTSLDAGVLVTWTKGFKATDCEGEDVVGLLREAIKRREEFDLDVVAVVNDTVGTMMTCAY
EEPTCEIGLIAGTGSNACYMEEMRNIEMIDGEEGRMCVNMEWGAFGDNGCLDDIRTDYDRAVDDFSLNSGKQRYEKMCS
GMYLGEIVRNILIDMTKRGFLFRGQISETLKTRGIFETKFLSQIESDRLALLQVRSILQHLGLDSTCDDSIIVKEVCGA
VSRRAAQLCGAGMAAVVDKIREN

>Ninespine_stickleback|XP_037340606.1:510-848 hexokinase-1 isoform X4
[Pungitius pungitius]
KMLPTFVYSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTVEMHNKIYSIPLEVMTGTGEELFDHIVQCISDFLDY
MGMKNTRLPLGFTFSFPCRQTSLDAGILVTWTKGFKATDCEGEDVVGLLREAIKRREEFDLDVVAVVNDTVGTLMTCAY
EEPTCEIGLIAGTGSNACYMEEMRNIEMIDGEEGRMCVNMEWGAFGDNGCLDDIRTDYDRAVDDLSLNSGKQRYEKMCS
GMYLGEIVRNILIDMTKRGFLFRGQISETLKTRGIFETKFLSQIESDRLALLQVRSILQHLGLDSTCDDSIIVKEVCGA
VSCRAAQLCGAGMAAVVDKIREN

>Cyclopterus|XP_034413042.1:510-848 hexokinase-1-like [Cyclopterus lumpus]
KMLPTFVHSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTVEMHNKIYSIPQEVMQGTGEELFDHIVQCISDFLDY
MGMKNTRLPLGFTFSFPCRQTSLDAGILVTWTKGFKATDCEGEDVVGLLREAIKRREEFDLDVVAIVNDTVGTMITCAY
EEPTCEIGLIAGTGSNACYMEEMRNIEMIDGDEGQMCVNMEWGAFGDNGCLDDIRTEYDRAVDDFSLNSGKQRYEKMCS
GMYLGEIVRNILIDMTKRGFLFRGQISETLKTRGIFETKFLSHIESDRLALLQVRSILQHLGLDSTCEDSIIVKEVCGA
VSRRAAQLCGAGMAAVVDKIREN

>Monkeyface_prickleback|XP_068586364.1:510-848 hexokinase-1 [Cebidichthys
violaceus]
KMLPTFVYSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTVEMHNKIYSIPLEVMQGTGEELFDHIVQCISDFLDY
MGMKNTRLPLGFTFSFPCRQTSLDAGILVTWTKGFKATDCEGEDVVGLLREAIKRREEFELDVVAVVNDTVGTMMTCAY
EEPTCEIGLIAGTGSNACYIEEMRNIEMIDGDEGRMCVNMEWGAFGDNGCLDDIRTEYDRAVDDFSLNSGKQRYEKMCS
GMYLGEIVRNILIDMTKRGFLFRGQISETLKTRGIFETKFLSQIESDRLALLQVRSILQHLGLDSTCDDSIIVKEVCGA
VSCRAAQLCGAGMAAVVDKIREN

>Wolf_eel|XP_031702071.1:510-848 hexokinase-1 isoform X1 [Anarrhichthys
ocellatus]
KMLPTFVYSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTVEMHNKIYSIPLEVMQGTGEELFDHIVQCISDFLDY
MGMKNTRLPLGFTFSFPCRQTSLDAGILVTWTKGFKATDCEGEDVVGLLREAIKRREEFELDVVAVVNDTVGTMMTCAY
EEPTCEIGLIAGTGSNACYIEEMRNIEMIDGDEGRMCVNMEWGAFGDNGCLDDIRTEYDRAVDDFSLNSGKQRYEKMCS
GMYLGEIVRNILIDMTKRGFLFRGQISETLKTRGIFETKFLSQIESDRLALLQVRSILQHLGLDSTCDDSIIVKEVCGA
VSCRAAQLCGAGMAAVVDKIREN

>Sablefish|XP_054464563.1:510-848 hexokinase-1 [Anoplopoma fimbria]
KMLPTFVNSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTVEMHNKIYSIPLEVMQGTGEELFDHIVQCISDFLDY
MGMKNTRLPLGFTFSFPCRQTSLDAGILMTWTKGFKATDCEGEDVVGLLREAIKRREEFDLDVVAVVNDTVGTMMTCAY
```

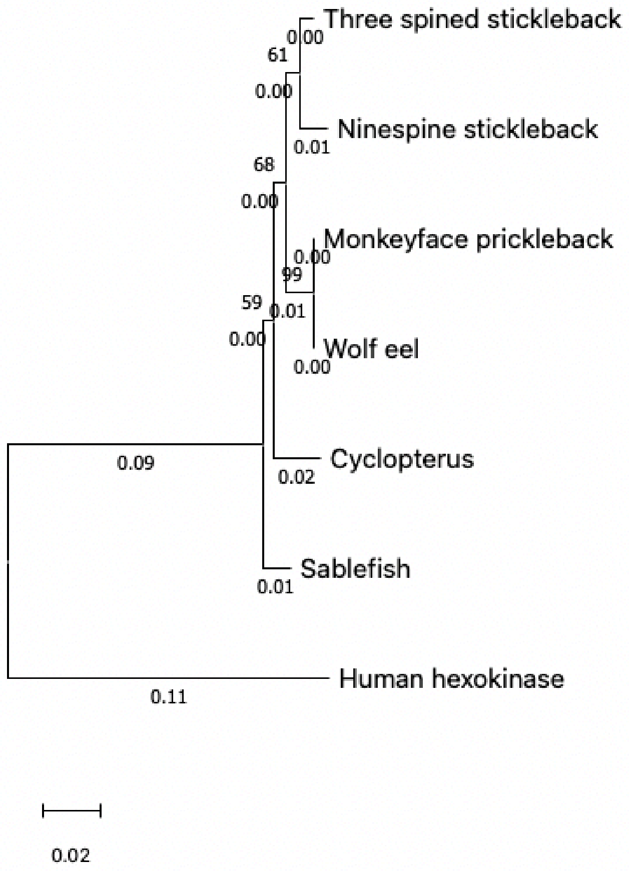
```
EEPTCEIGLIAGTGSNACYMEEMRNIEMIDGDEGQMCVNMEWGAFGDNGCLDDIRTEYDRAVDDFSLNPGKQRYEKMCS
GMYLGEIVRNILIDMTKRGFLFRGQISETLKTRGIFETKFLSQIESDRLALLQVRAILQHLGLDSTCDDSIIVKEVCGV
VSRRAAQLCGAGMAAVVDKIREN
```

Alignment:
Obtained using MUSCLE (version 3.8) at EBI:

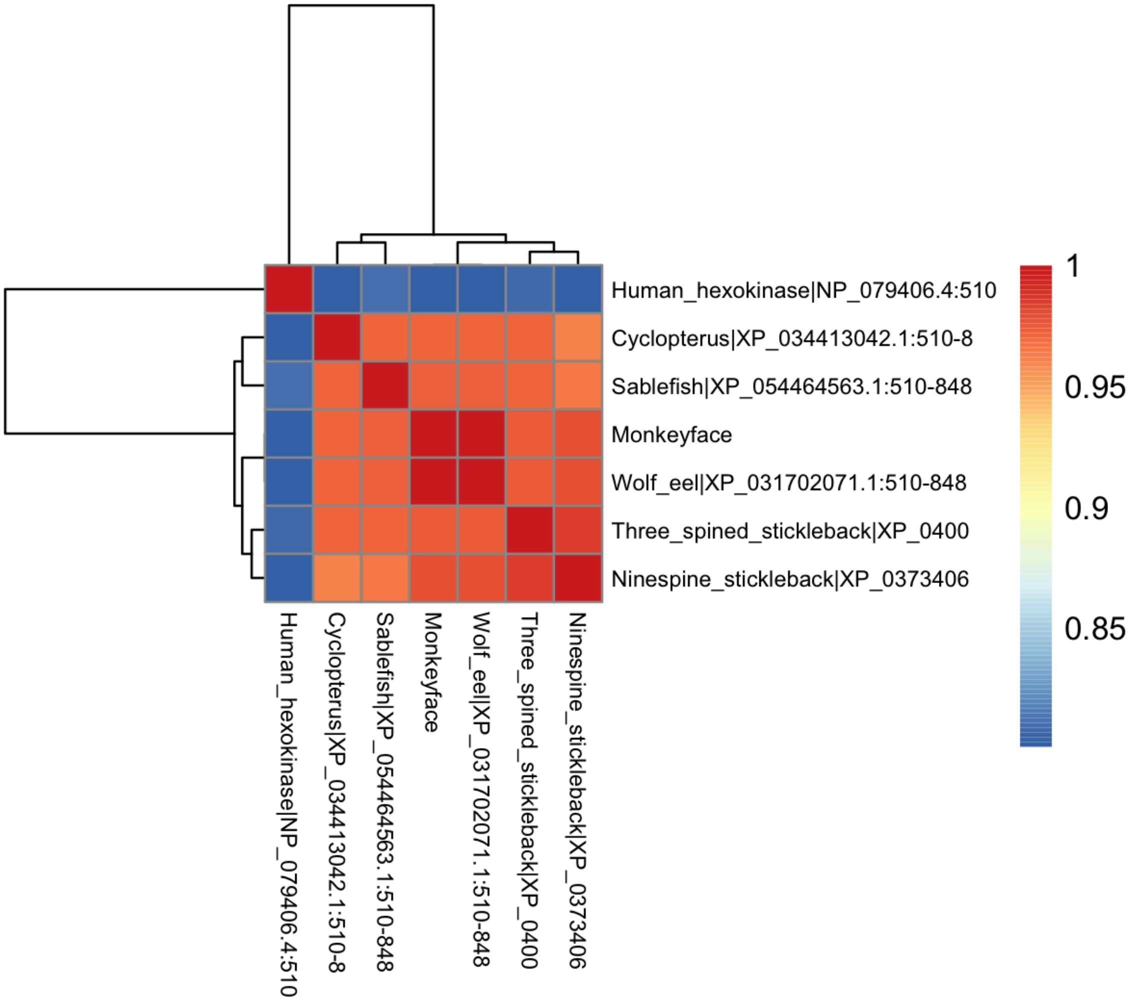
```
CLUSTAL multiple sequence alignment by MUSCLE (3.8)


Human_hexokinase         RMLPTYVCGLPDGTEKGKFLALDLGGTNFRVLLVKIRSG-RRSVRMYNKIFAIPLEIMQG
Three_spined_stickleback KMLPTFVHSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTVEMHNKIYSIPQEVMQG
Ninespine_stickleback    KMLPTFVNSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTVEMHNKIYSIPLEVMQG
Cyclopterus              KMLPTFVYSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTVEMHNKIYSIPLEVMQG
Monkeyface_prickleback   KMLPTFVYSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTVEMHNKIYSIPLEVMQG
Wolf_eel                 KMLPTFVHSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTVEMHNKIYSIPLEVMTG
Sablefish                KMLPTFVYSTPDGSEHGDFLALDLGGTNFRVLLVKIRSGKRRTVEMHNKIYSIPLEVMTG
                         .****:* . ***:*:*.********************* **:* *:***::** *:* *

Human_hexokinase         TGEELFDHIVQCIADFLDYMGLKGASLPLGFTFSFPCRQMSIDKGTLIGWTKGFKATDCE
Three_spined_stickleback TGEELFDHIVQCISDFLDYMGMKNTRLPLGFTFSFPCRQTSLDAGILVTWTKGFKATDCE
Ninespine_stickleback    TGEELFDHIVQCISDFLDYMGMKNTRLPLGFTFSFPCRQTSLDAGILMTWTKGFKATDCE
Cyclopterus              TGEELFDHIVQCISDFLDYMGMKNTRLPLGFTFSFPCRQTSLDAGILVTWTKGFKATDCE
Monkeyface_prickleback   TGEELFDHIVQCISDFLDYMGMKNTRLPLGFTFSFPCRQTSLDAGILVTWTKGFKATDCE
Wolf_eel                 TGEELFDHIVQCISDFLDYMGMKNTRLPLGFTFSFPCRQTSLDAGVLVTWTKGFKATDCE
Sablefish                TGEELFDHIVQCISDFLDYMGMKNTRLPLGFTFSFPCRQTSLDAGILVTWTKGFKATDCE
                         *************:*******:*.: ************* *:* * *: ***********

Human_hexokinase         GEDVVDMLREAIKRRNEFDLDIVAVVNDTVGTMMTCGYEDPNCEIGLIAGTGSNMCYMED
Three_spined_stickleback GEDVVGLLREAIKRREEFDLDVVAIVNDTVGTMITCAYEEPTCEIGLIAGTGSNACYMEE
Ninespine_stickleback    GEDVVGLLREAIKRREEFDLDVVAVVNDTVGTMMTCAYEEPTCEIGLIAGTGSNACYMEE
Cyclopterus              GEDVVGLLREAIKRREEFELDVVAVVNDTVGTMMTCAYEEPTCEIGLIAGTGSNACYIEE
Monkeyface_prickleback   GEDVVGLLREAIKRREEFELDVVAVVNDTVGTMMTCAYEEPTCEIGLIAGTGSNACYIEE
Wolf_eel                 GEDVVGLLREAIKRREEFDLDVVAVVNDTVGTMMTCAYEEPTCEIGLIAGTGSNACYMEE
Sablefish                GEDVVGLLREAIKRREEFDLDVVAVVNDTVGTLMTCAYEEPTCEIGLIAGTGSNACYMEE
                         *****.:********:**:**:**:*******::**.**:*.************ **:*:

Human_hexokinase         MRNIEMVEGGEGKMCINTEWGGFGDNGCIDDIWTRYDTEVDEGSLNPGKQRYEKMTSGMY
Three_spined_stickleback MRNIEMIDGDEGQMCVNMEWGAFGDNGCLDDIRTEYDRAVDDFSLNSGKQRYEKMCSGMY
Ninespine_stickleback    MRNIEMIDGDEGQMCVNMEWGAFGDNGCLDDIRTEYDRAVDDFSLNPGKQRYEKMCSGMY
Cyclopterus              MRNIEMIDGDEGRMCVNMEWGAFGDNGCLDDIRTEYDRAVDDFSLNSGKQRYEKMCSGMY
Monkeyface_prickleback   MRNIEMIDGDEGRMCVNMEWGAFGDNGCLDDIRTEYDRAVDDFSLNSGKQRYEKMCSGMY
Wolf_eel                 MRNIEMIDGEEGRMCVNMEWGAFGDNGCLDDIRTDYDRAVDDFSLNSGKQRYEKMCSGMY
Sablefish                MRNIEMIDGEEGRMCVNMEWGAFGDNGCLDDIRTDYDRAVDDLSLNSGKQRYEKMCSGMY
                         ******::* **.**:* ***.******:***.* **  **: ***.******** ****

Human_hexokinase         LGEIVRQILIDLTKQGLLFRGQISERLRTRGIFETKFLSQIESDRLALLQVRRILQQLGL
Three_spined_stickleback LGEIVRNILIDMTKRGFLFRGQISETLKTRGIFETKFLSHIESDRLALLQVRSILQHLGL
Ninespine_stickleback    LGEIVRNILIDMTKRGFLFRGQISETLKTRGIFETKFLSQIESDRLALLQVRAILQHLGL
Cyclopterus              LGEIVRNILIDMTKRGFLFRGQISETLKTRGIFETKFLSQIESDRLALLQVRSILQHLGL
Monkeyface_prickleback   LGEIVRNILIDMTKRGFLFRGQISETLKTRGIFETKFLSQIESDRLALLQVRSILQHLGL
Wolf_eel                 LGEIVRNILIDMTKRGFLFRGQISETLKTRGIFETKFLSQIESDRLALLQVRSILQHLGL
Sablefish                LGEIVRNILIDMTKRGFLFRGQISETLKTRGIFETKFLSQIESDRLALLQVRSILQHLGL
                         ******:****:**.*:******** *.***********:************ ***:***
```

```
Human_hexokinase         DSTCEDSIVVKEVCGAVSRRAAQLCGAGLAAIVEKRR--
Three_spined_stickleback DSTCEDSIIVKEVCGAVSRRAAQLCGAGMAAVVDKIREN
Ninespine_stickleback    DSTCDDSIIVKEVCGVVSRRAAQLCGAGMAAVVDKIREN
Cyclopterus              DSTCDDSIIVKEVCGAVSCRAAQLCGAGMAAVVDKIREN
Monkeyface_prickleback   DSTCDDSIIVKEVCGAVSCRAAQLCGAGMAAVVDKIREN
Wolf_eel                 DSTCDDSIIVKEVCGAVSRRAAQLCGAGMAAVVDKIREN
Sablefish                DSTCDDSIIVKEVCGAVSCRAAQLCGAGMAAVVDKIREN
                         ****:***:******.** *********:**:*:* *
```

[**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 it to your report. Do make sure your labels are visible and not cut at the figure margins.



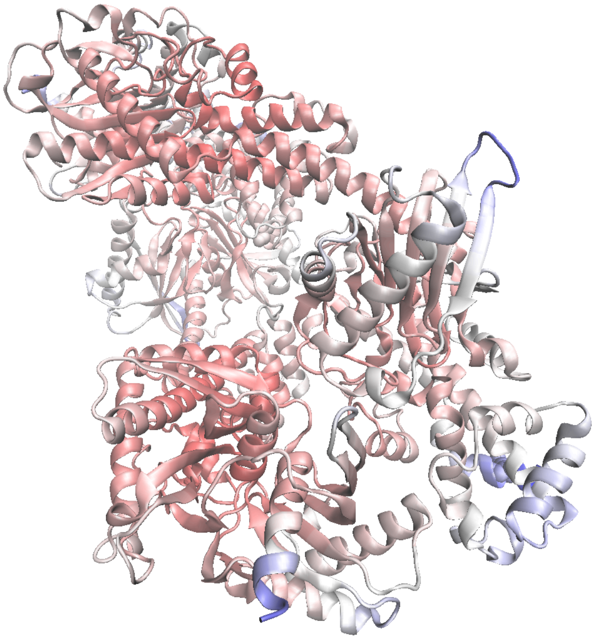
[Q8] Using R/Bio3D (or an online blast server if you prefer), search the main protein structure database for the most similar atomic resolution structures to your aligned sequences. List the top 3 unique hits (i.e. not hits representing different chains from the same structure) along with their E-value and sequence identity to your query. Please also add annotation details of these structures. For example include the annotation terms PDB identifier (structure Id), method used to solve the structure (experimental technique), 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 E-value 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 choose the sequence with the highest identity to all others in your alignment by calculating the row-wise maximum from your sequence identity matrix.

ID	Technique	Resolution	Source	E-value	Identity
4FOI	X-ray diffraction	2.40	Homo Sapiens (Mammalian type)	0.00	78.338
1CZA	X-ray diffraction	1.90	Homo Sapiens (Mammalian type)	0.00	78.338
4F9O	X-ray diffraction	2.65	Homo Sapiens (Mammalian type)	0.00	78.338

[**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?



There’s a high chance that it is similar in structure to the three-spined stickleback due to the E-value of ~0 and a sequence identity of ~80%.

[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]”.

CHEMBL details 1 Functional Assay (CHEMBL3374607); no binding assays or ligand efficiency data.

https://www.ebi.ac.uk/chembl/search_results/Gasterosteus%20aculeatus

Functional assay linked “Antimicrobial activity against Aspergillus aculeatus incubated for 48 hrs by disk diffusion method.”