Tiantai Ma
t2ma@ucsd.edu
PID: A53308792

Questions:

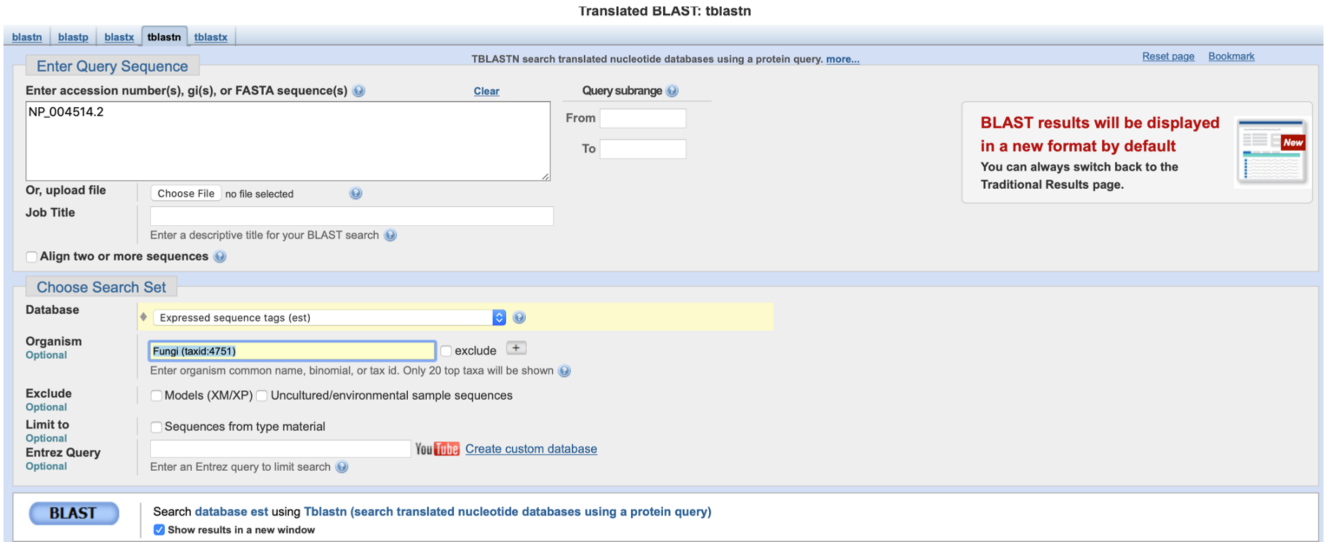
[Q1] Tell me the name of a protein you are interested in. Include the species and the accession number. This can be a human protein or a protein from any other species as long as it's function is known.

If you do not have a favorite protein, select human RBP4 or KIF11. Do not use beta globin as this is in the worked example report that I provide you with online.

KIF11 (kinesin family member 11)
Accession number: NP_004514
Species: Homo sapiens

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

Method: tblastn search against Fungi ESTs
Database: Expressed sequence tag (est)
Organism: Fungi (taxid:4751)



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): NP_004514.2

Query subrange
From
To

Or, upload file: no file selected

Job Title
Enter a descriptive title for your BLAST search

□ Align two or more sequences

BLAST results will be displayed in a new format by default
You can always switch back to the Traditional Results page.

Choose Search Set

Database: Expressed sequence tags (est)

Organism (Optional): Fungi (taxid:4751) □ exclude
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

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

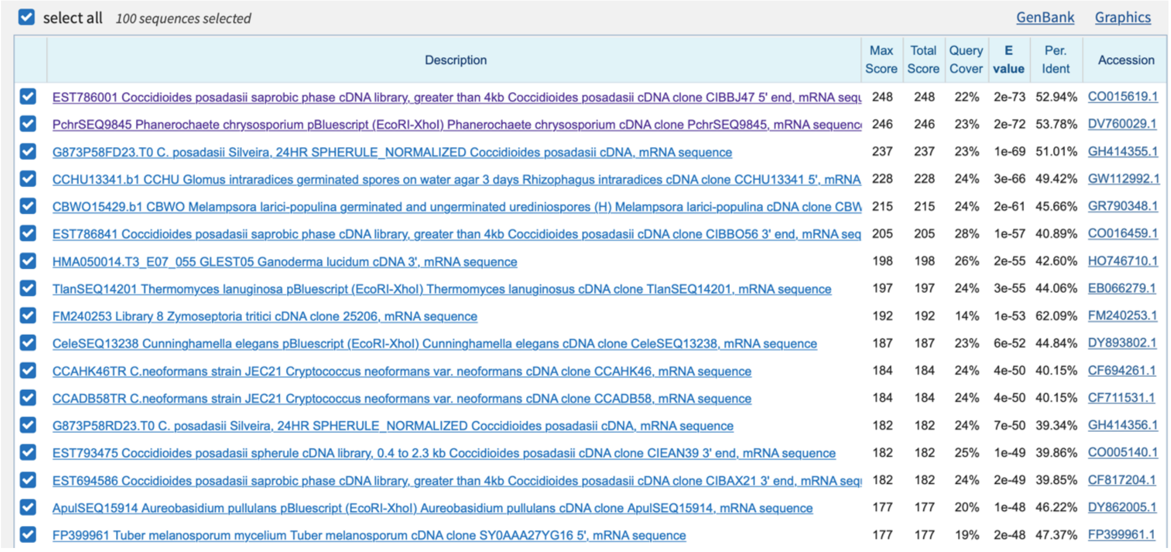
Limit to (Optional): □ Sequences from type material

Entrez Query (Optional):
Enter an Entrez query to limit search

Search database est using Tblastn (search translated nucleotide databases using a protein query)
☑ Show results in a new window

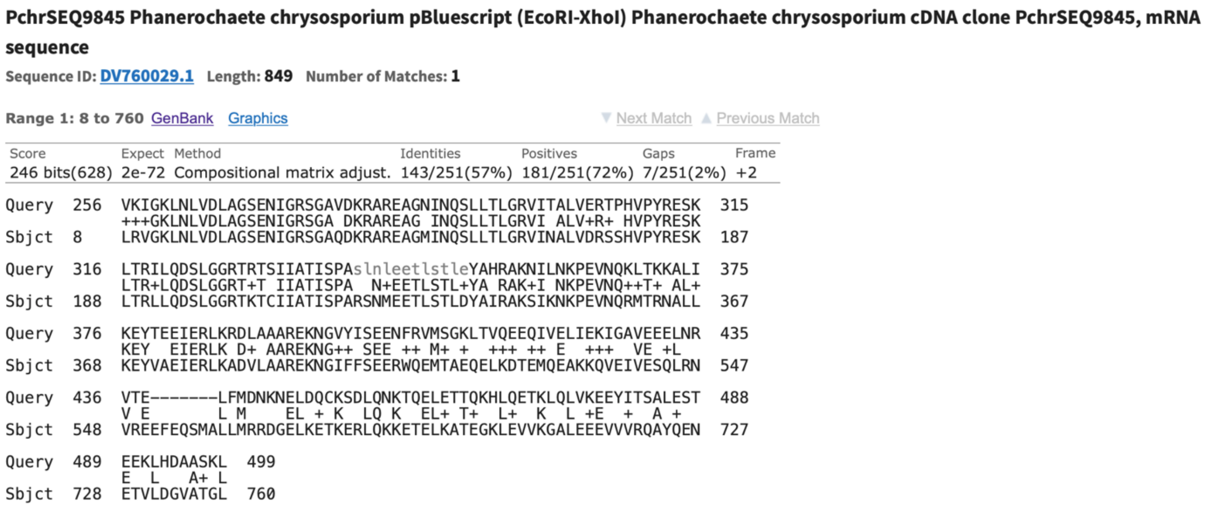
Also include the output of that BLAST search in your document. If appropriate, change the font to `Courier size 10` so that the results are displayed neatly. You can also screen capture a BLAST output (e.g. alt print screen on a PC or on a MAC press ⌘-shift-4. The pointer becomes a bulls eye. Select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

Chosen match: Accession DV760029.1**,** 849 base pairs cDNA clone from *Phanerochaete chrysosporium.*



☑ select all *100 sequences selected* GenBank Graphics

☐	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	EST786001 Coccidioides posadasii saprobic phase cDNA library, greater than 4kb Coccidioides posadasii cDNA clone CIBBJ47 5' end, mRNA sequ	248	248	22%	2e-73	52.94%	CO015619.1
☑	PchrSEQ9845 Phanerochaete chrysosporium pBluescript (EcoRI-XhoI) Phanerochaete chrysosporium cDNA clone PchrSEQ9845, mRNA sequenc	246	246	23%	2e-72	53.78%	DV760029.1
☑	G873P58FD23.T0 C. posadasii Silveira, 24HR SPHERULE_NORMALIZED Coccidioides posadasii cDNA, mRNA sequence	237	237	23%	1e-69	51.01%	GH414355.1
☑	CCHU13341.b1 CCHU Glomus intraradices germinated spores on water agar 3 days Rhizophagus intraradices cDNA clone CCHU13341 5', mRNA	228	228	24%	3e-66	49.42%	GW112992.1
☑	CBWO15429.b1 CBWO Melampsora larici-populina germinated and ungerminated urediniospores (H) Melampsora larici-populina cDNA clone CBW	215	215	24%	2e-61	45.66%	GR790348.1
☑	EST786841 Coccidioides posadasii saprobic phase cDNA library, greater than 4kb Coccidioides posadasii cDNA clone CIBBO56 3' end, mRNA seq	205	205	28%	1e-57	40.89%	CO016459.1
☑	HMA050014.T3_E07_055 GLEST05 Ganoderma lucidum cDNA 3', mRNA sequence	198	198	26%	2e-55	42.60%	HO746710.1
☑	TlanSEQ14201 Thermomyces lanuginosa pBluescript (EcoRI-XhoI) Thermomyces lanuginosus cDNA clone TlanSEQ14201, mRNA sequence	197	197	24%	3e-55	44.06%	EB066279.1
☑	FM240253 Library 8 Zymoseptoria tritici cDNA clone 25206, mRNA sequence	192	192	14%	1e-53	62.09%	FM240253.1
☑	CeleSEQ13238 Cunninghamella elegans pBluescript (EcoRI-XhoI) Cunninghamella elegans cDNA clone CeleSEQ13238, mRNA sequence	187	187	23%	6e-52	44.84%	DY893802.1
☑	CCAHK46TR C.neoformans strain JEC21 Cryptococcus neoformans var. neoformans cDNA clone CCAHK46, mRNA sequence	184	184	24%	4e-50	40.15%	CF694261.1
☑	CCADB58TR C.neoformans strain JEC21 Cryptococcus neoformans var. neoformans cDNA clone CCADB58, mRNA sequence	184	184	24%	4e-50	40.15%	CF711531.1
☑	G873P58RD23.T0 C. posadasii Silveira, 24HR SPHERULE_NORMALIZED Coccidioides posadasii cDNA, mRNA sequence	182	182	24%	7e-50	39.34%	GH414356.1
☑	EST793475 Coccidioides posadasii spherule cDNA library, 0.4 to 2.3 kb Coccidioides posadasii cDNA clone CIEAN39 3' end, mRNA sequence	182	182	25%	1e-49	39.86%	CO005140.1
☑	EST694586 Coccidioides posadasii saprobic phase cDNA library, greater than 4kb Coccidioides posadasii cDNA clone CIBAX21 3' end, mRNA seq	182	182	24%	2e-49	39.85%	CF817204.1
☑	ApulSEQ15914 Aureobasidium pullulans pBluescript (EcoRI-XhoI) Aureobasidium pullulans cDNA clone ApulSEQ15914, mRNA sequence	177	177	20%	1e-48	46.22%	DY862005.1
☑	FP399961 Tuber melanosporum mycelium Tuber melanosporum cDNA clone SY0AAA27YG16 5', mRNA sequence	177	177	19%	2e-48	47.37%	FP399961.1



PchrSEQ9845 Phanerochaete chrysosporium pBluescript (EcoRI-XhoI) Phanerochaete chrysosporium cDNA clone PchrSEQ9845, mRNA sequence

Sequence ID: DV760029.1 **Length:** 849 **Number of Matches:** 1

Range 1: 8 to 760 GenBank Graphics

```
Score           Expect  Method                        Identities     Positives      Gaps        Frame
246 bits(628)   2e-72   Compositional matrix adjust.  143/251(57%)   181/251(72%)   7/251(2%)   +2

Query  256  VKIGKLNLVDLAGSENIGRSGAVDKRAREAGNINQSLLTLGRVITALVERTPHVPYRESK  315
            +++GKLNLVDLAGSENIGRSGA DKRAREAG INQSLLTLGRVI ALV+R+ HVPYRESK
Sbjct  8    LRVGKLNLVDLAGSENIGRSGAQDKRAREAGMINQSLLTLGRVINALVDRSSHVPYRESK  187

Query  316  LTRILQDSLGGRTRTSIIATISPAslnleetlstleYAHRAKNILNKPEVNQKLTKKALI  375
            LTR+LQDSLGGRT+T IIATISPA  N+EETLSTL+YA RAK+I NKPEVNQ++T+ AL+
Sbjct  188  LTRLLQDSLGGRTKTCIIATISPARSNMEETLSTLDYAIRAKSIKNKPEVNQRMTRNALL  367

Query  376  KEYTEEIERLKRDLAAAREKNGVYISEENFRVMSGKLTVQEEQIVELIEKIGAVEEELNR  435
            KEY  EIERLK D+ AAREKNG++ SEE ++ M+ +  +++ ++ E  +++  VE +L
Sbjct  368  KEYVAEIERLKADVLAAREKNGIFFSEERWQEMTAEQELKDTEMQEAKKQVEIVESQLRN  547

Query  436  VTE-------LFMDNKNELDQCKSDLQNKTQELETTQKHLQETKLQLVKEEYITSALEST  488
            V E       L M    EL + K  LQ K  EL+ T+  L+  K  L +E  +  A +
Sbjct  548  VREEFEQSMALLMRRDGELKETKERLQKKETELKATEGKLEVVKGALEEEVVVRQAYQEN  727

Query  489  EEKLHDAASKL  499
            E  L   A+ L
Sbjct  728  ETVLDGVATGL  760
```

PchrSEQ9845 Phanerochaete chrysosporium pBluescript (EcoRI-XhoI) Phanerochaete chrysosporium cDNA clone PchrSEQ9845, mRNA sequence

Sequence ID: DV760029.1 Length: 849 Number of Matches: 1

Range 1: 8 to 760GenBankGraphicsNext MatchPrevious Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
246 bits(628)	2e-72	Compositional matrix adjust.	143/251(57%)	181/251(72%)	7/251(2%)	+2

```
Query  256   VKIGKLNLVDLAGSENIGRSGAVDKRAREAGNINQSLLTLGRVITALVERTPHVPYRESK  315
             +++GKLNLVDLAGSENIGRSGA DKRAREAG INQSLLTLGRVI ALV+R+ HVPYRESK
Sbjct  8     LRVGKLNLVDLAGSENIGRSGAQDKRAREAGMINQSLLTLGRVINALVDRSSHVPYRESK  187

Query  316   LTRILQDSLGGRTRTSIIATISPAslnleetlstleYAHRAKNILNKPEVNQKLTKKALI  375
             LTR+LQDSLGGRT+T IIATISPA  N+EETLSTL+YA RAK+I NKPEVNQ++T+ AL+
Sbjct  188   LTRLLQDSLGGRTKTCIIATISPARSNMEETLSTLDYAIRAKSIKNKPEVNQRMTRNALL  367

Query  376   KEYTEEIERLKRDLAAAREKNGVYISEENFRVMSGKLTVQEEQIVELIEKIGAVEEELNR  435
             KEY  EIERLK D+ AAREKNG++ SEE ++ M+ +  +++ ++ E  +++  VE +L
Sbjct  368   KEYVAEIERLKADVLAAREKNGIFFSEERWQEMTAEQELKDTEMQEAKKQVEIVESQLRN  547

Query  436   VTE-------LFMDNKNELDQCKSDLQNKTQELETTQKHLQETKLQLVKEEYITSALEST  488
             V E       L M    EL + K  LQ K  EL+ T+  L+  K  L +E  +  A +
Sbjct  548   VREEFEQSMALLMRRDGELKETKERLQKKETELKATEGKLEVVKGALEEEVVVRQAYQEN  727

Query  489   EEKLHDAASKL  499
             E  L   A+ L
Sbjct  728   ETVLDGVATGL  760
```

On the BLAST results, clearly indicate a match that represents a protein sequence, encoded from some DNA sequence, that is homologous to your query protein. I need to be able to inspect the pairwise alignment you have selected, including the E value and score. It should be labeled a "genomic clone" or "mRNA sequence", etc. - but include no functional annotation.

In general, [Q2] is the most difficult for students because it requires you to have a “feel” for how to interpret BLAST results. You need to distinguish between a perfect match to your query (i.e. a sequence that is not “novel”), a near match (something that might be “novel”, depending on the results of [Q4]), and a non-homologous result.

If you are having trouble finding a novel gene try restricting your search to an organism that is poorly annotated.

[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.

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.

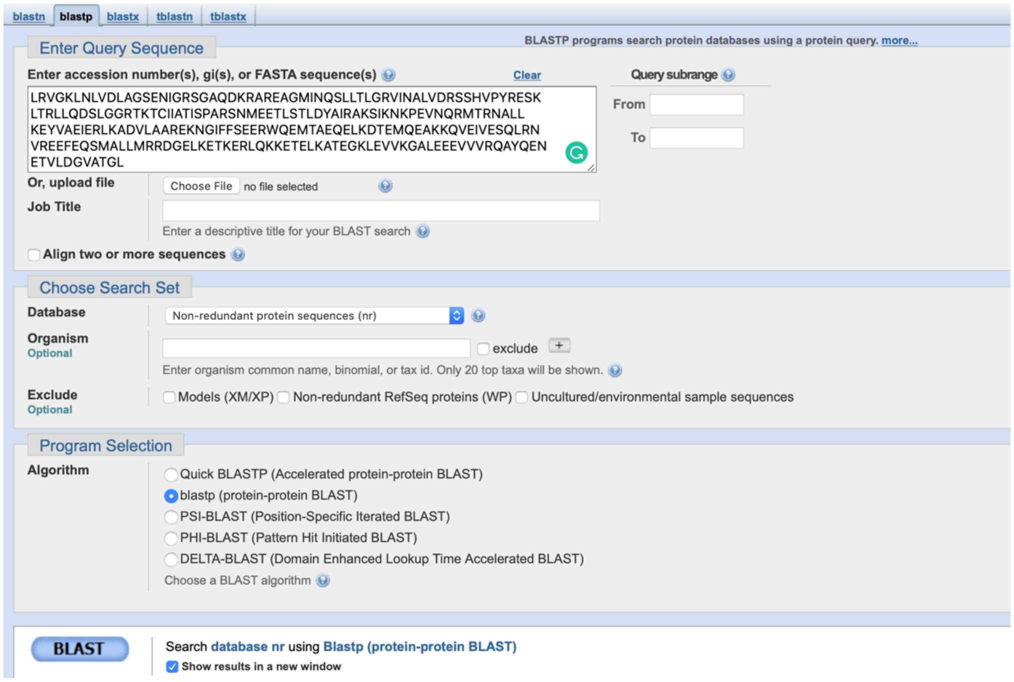
```
>8-760_1 PchrSEQ9845 Phanerochaete chrysosporium pBluescript (EcoRI-XhoI) Phanerochaete
chrysosporium cDNA clone PchrSEQ9845, mRNA sequence
LRVGKLNLVDLAGSENIGRSGAQDKRAREAGMINQSLLTLGRVINALVDRSSHVPYRESK
LTRLLQDSLGGRTKTCIIATISPARSNMEETLSTLDYAIRAKSIKNKPEVNQRMTRNALL
KEYVAEIERLKADVLAAREKNGIFFSEERWQEMTAEQELKDTEMQEAKKQVEIVESQLRN
VREEFEQSMALLMRRDGELKETKERLQKKETELKATEGKLEVVKGALEEEVVVRQAYQEN
ETVLDGVATGL
```

Name: *Phanerochaete chrysosporium*, hypothetical protein
Species: *Phanerochaete chrysosporium*
Eukaryota; Fungi; Dikarya; Basidiomycota; Agaricomycotina;
Agaricomycetes; Polyporales; Phanerochaetaceae; Phanerochaete

[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.

Blastp search against nr database:



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

Enter Query Sequence

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

```
LRVGKLNLVDLAGSENIGRSGAQDKRAREAGMINQSLLTLGRVINALVDRSSHVPYRESK
LTRLLQDSLGGRTKTCIIATISPARSNMEETLSTLDYAIRAKSIKNKPEVNQRMTRNALL
KEYVAEIERLKADVLAAREKNGIFFSEERWQEMTAEQELKDTEMQEAKKQVEIVESQLRN
VREEFEQSMALLMRRDGELKETKERLQKKETELKATEGKLEVVKGALEEEVVVRQAYQEN
ETVLDGVATGL
```

Query subrange

From

To

Or, upload file: no file selected

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism Optional: □ exclude +

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

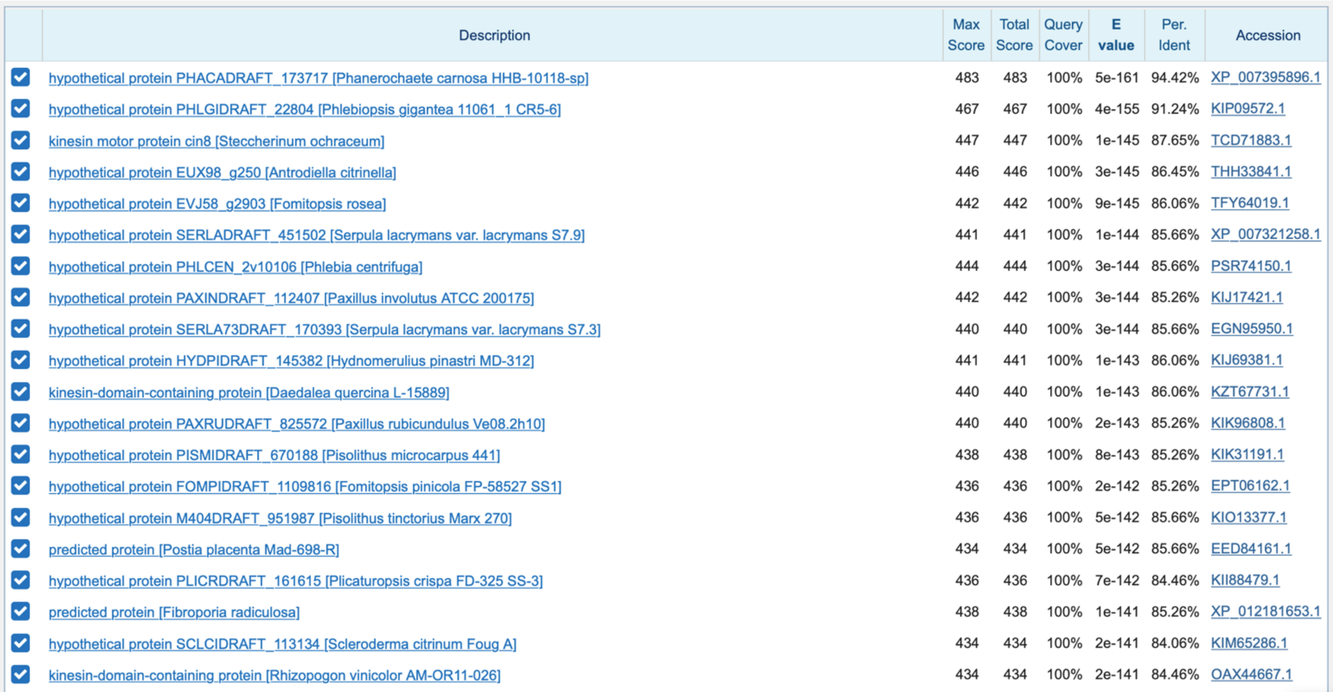
○ Quick BLASTP (Accelerated protein-protein BLAST)
● blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

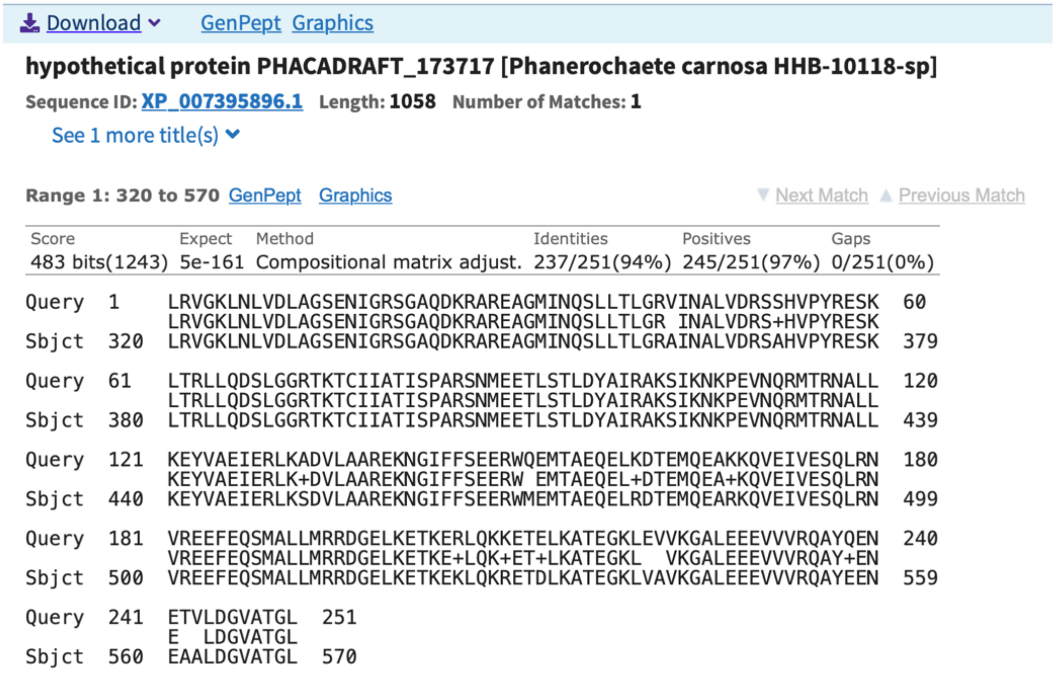
Search database nr using Blastp (protein-protein BLAST)

☑ Show results in a new window

The top result is a hypothetical protein from *Phanerochaete camosa* with identity 94.42%, some results indicate that this is a kinesin-domain-cotaining protein:



	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	hypothetical protein PHACADRAFT_173717 [Phanerochaete carnosa HHB-10118-sp]	483	483	100%	5e-161	94.42%	XP_007395896.1
☑	hypothetical protein PHLGIDRAFT_22804 [Phlebiopsis gigantea 11061_1 CR5-6]	467	467	100%	4e-155	91.24%	KIP09572.1
☑	kinesin motor protein cin8 [Steccherinum ochraceum]	447	447	100%	1e-145	87.65%	TCD71883.1
☑	hypothetical protein EUX98_g250 [Antrodiella citrinella]	446	446	100%	3e-145	86.45%	THH33841.1
☑	hypothetical protein EVJ58_g2903 [Fomitopsis rosea]	442	442	100%	9e-145	86.06%	TFY64019.1
☑	hypothetical protein SERLADRAFT_451502 [Serpula lacrymans var. lacrymans S7.9]	441	441	100%	1e-144	85.66%	XP_007321258.1
☑	hypothetical protein PHLCEN_2v10106 [Phlebia centrifuga]	444	444	100%	3e-144	85.66%	PSR74150.1
☑	hypothetical protein PAXINDRAFT_112407 [Paxillus involutus ATCC 200175]	442	442	100%	3e-144	85.26%	KIJ17421.1
☑	hypothetical protein SERLA73DRAFT_170393 [Serpula lacrymans var. lacrymans S7.3]	440	440	100%	3e-144	85.66%	EGN95950.1
☑	hypothetical protein HYDPIDRAFT_145382 [Hydnomerulius pinastri MD-312]	441	441	100%	1e-143	86.06%	KIJ69381.1
☑	kinesin-domain-containing protein [Daedalea quercina L-15889]	440	440	100%	1e-143	86.06%	KZT67731.1
☑	hypothetical protein PAXRUDRAFT_825572 [Paxillus rubicundulus Ve08.2h10]	440	440	100%	2e-143	85.26%	KIK96808.1
☑	hypothetical protein PISMIDRAFT_670188 [Pisolithus microcarpus 441]	438	438	100%	8e-143	85.26%	KIK31191.1
☑	hypothetical protein FOMPIDRAFT_1109816 [Fomitopsis pinicola FP-58527 SS1]	436	436	100%	2e-142	85.26%	EPT06162.1
☑	hypothetical protein M404DRAFT_951987 [Pisolithus tinctorius Marx 270]	436	436	100%	5e-142	85.66%	KIO13377.1
☑	predicted protein [Postia placenta Mad-698-R]	434	434	100%	5e-142	85.66%	EED84161.1
☑	hypothetical protein PLICRDRAFT_161615 [Plicaturopsis crispa FD-325 SS-3]	436	436	100%	7e-142	84.46%	KII88479.1
☑	predicted protein [Fibroporia radiculosa]	438	438	100%	1e-141	85.26%	XP_012181653.1
☑	hypothetical protein SCLCIDRAFT_113134 [Scleroderma citrinum Foug A]	434	434	100%	2e-141	84.06%	KIM65286.1
☑	kinesin-domain-containing protein [Rhizopogon vinicolor AM-OR11-026]	434	434	100%	2e-141	84.46%	OAX44667.1



hypothetical protein PHACADRAFT_173717 [Phanerochaete carnosa HHB-10118-sp]

Sequence ID: XP_007395896.1 **Length:** 1058 **Number of Matches:** 1

See 1 more title(s)

Range 1: 320 to 570 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
483 bits(1243)	5e-161	Compositional matrix adjust.	237/251(94%)	245/251(97%)	0/251(0%)

```
Query  1    LRVGKLNLVDLAGSENIGRSGAQDKRAREAGMINQSLLTLGRVINALVDRSSHVPYRESK  60
            LRVGKLNLVDLAGSENIGRSGAQDKRAREAGMINQSLLTLGR INALVDRS+HVPYRESK
Sbjct  320  LRVGKLNLVDLAGSENIGRSGAQDKRAREAGMINQSLLTLGRAINALVDRSAHVPYRESK  379

Query  61   LTRLLQDSLGGRTKTCIIATISPARSNMEETLSTLDYAIRAKSIKNKPEVNQRMTRNALL  120
            LTRLLQDSLGGRTKTCIIATISPARSNMEETLSTLDYAIRAKSIKNKPEVNQRMTRNALL
Sbjct  380  LTRLLQDSLGGRTKTCIIATISPARSNMEETLSTLDYAIRAKSIKNKPEVNQRMTRNALL  439

Query  121  KEYVAEIERLKADVLAAREKNGIFFSEERWQEMTAEQELKDTEMQEAKKQVEIVESQLRN  180
            KEYVAEIERLK+DVLAAREKNGIFFSEERW EMTAEQEL+DTEMQEA+KQVEIVESQLRN
Sbjct  440  KEYVAEIERLKSDVLAAREKNGIFFSEERWMEMTAEQELRDTEMQEARKQVEIVESQLRN  499

Query  181  VREEFEQSMALLMRRDGELKETKERLQKKETELKATEGKLEVVKGALEEEVVVRQAYQEN  240
            VREEFEQSMALLMRRDGELKETKE+LQK+ET+LKATEGKL  VKGALEEEVVVRQAY+EN
Sbjct  500  VREEFEQSMALLMRRDGELKETKEKLQKRETDLKATEGKLVAVKGALEEEVVVRQAYEEN  559

Query  241  ETVLDGVATGL  251
            E  LDGVATGL
Sbjct  560  EAALDGVATGL  570
```