Name: Jack Reddan PID: A59010543

[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.

Name: glutathione reductase

Accession: NP_191026

Species: Arabidopsis thaliana

Function: glutathione-disulfide reductase activity

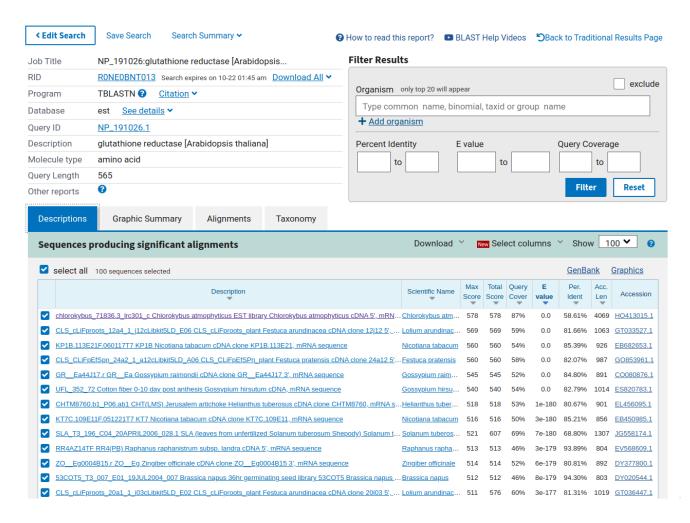
[Q2] Perform a BLAST search against a DNA database, such as a database consisting of genomic DNA or ESTs. The BLAST server can be at NCBI or elsewhere. Include details of the BLAST method used, database searched and any limits applied (e.g. Organism).

Method: NCBI TBLASTN (v2.12.0) searched against all ESTs

Database: Expressed Sequence Tags [est] database

Organism: None

Results:



Chosen match: Accession H0413015.1, 4069 bp *Chlorokybus atmophytus* mRNA sequence, highlighted in purple above. Alignment details are printed below:

Query: glutathione reductase [Arabidopsis thaliana] Query ID: NP_191026.1 Length: 565

>chlorokybus_71836.3_lrc301_c Chlorokybus atmophyticus EST library Chlorokybus atmophyticus cDNA 5', mRNA sequence Sequence ID: H0413015.1 Length: 4069 Range 1: 2451 to 3941

Score:578 bits(1489), Expect:0.0,
Method:Compositional matrix adjust.,

Identities:296/505(59%), Positives:363/505(71%), Gaps:17/505(3%)

Identities:296/505(59%), Positives:363/505(71%), Gaps:17/505(3%)						
Query	52	LRPRIALLSNHRYYHSRRFSVCASTDNGAESDRHYDFDLFTIGAGSGGVRA LRP L HSR+ S V AS++ YD+D+ TIGAGSGGVRA	102			
Sbjct	3941	LRPS-GLAQGQSRPHSRQQSRTVQRYGLRVIASSNGSGYDYDVITIGAGSGGVRA	3780			
Query	103	SRFATSFGASAAVCELPFSTISSDTAGGVGGTCVLRGCVPKKLLVYASKYSHEFEDSHGF SR A+ GA A E+PF+ ++SDT GGVGGTCVLRGCVPKKLLVY S +S+EF+DS GF	162			
Sbjct	3779	SRIASQLGAKVACVEMPFNNVASDTEGGVGGTCVLRGCVPKKLLVYGSIFSNEFDDSAGF	3600			
Query	163	GWKYETEPSHDWTTLIANKNAELQRLTGIYKNILSKANVKLIEGRGKVIDPHTVDVDGKI GWK EP W TL NKN EL RL +Y+NILSKANV+L+EGR ++D HTVD+DGK	222			
Sbjct	3599	GWKLPGEPKFTWQTLNENKNKELTRLNNVYRNILSKANVELLEGRASLVDAHTVDIDGKQ	3420			
Query	223	YTTRNILIAVGGRPFIPDIPGKEFAIDSDAALDLPSKPKKIAIVGGGYIALEFAGIFNGL T +NI++A GGR F IPG E AIDSD AL L PK+IAI GGGYIALEFA IF+G	282			
Sbjct	3419	LTAKNIILATGGRSFALPIPGAEHAIDSDKALSLDEVPKRIAIYGGGYIALEFACIFSGF	3240			
Query	283	NCEVHVFIRQKKVLRGFDEDVRDFVGEQMSLRGIEFHTEESPEAIIKAGDGSFSLKTSKG +V VF R LRGFDE++R+ + E++ +GI H + + E I K +G ++LKT+ G	342			
Sbjct	3239	GAKVDVFYRAPLPLRGFDEEIRNALVEELGKKGINLHPKCTAEEIRKEANGEYTLKTNCG	3060			
Query	343	TVEGFSHVMFATGRKPNTKNLGLENVGVKMAKNGAIEVDEYSQTSVPSIWAVGDVTDRIN + VMFATGR PNTK L L+ VGV + GAI VDEYS+T+VP+I+A+GDVT+RIN	402			
Sbjct	3059	EFKA-DLVMFATGRTPNTKYLNLDAVGVDTTEKGAIVVDEYSRTTVPNIFAIGDVTNRIN	2883			
Query	403	LTPVALMEGGALAKTLFQNEPTKPDYRAVPCAVFSQPPIGTVGLTEEQAIEQYGDVDVYT LTPVALMEG A+AKT+ Q EPTKPD+ VP AVF+QPPIGT GLTEE+A EQ+ +VDVYT	462			
Sbjct	2882	LTPVALMEGTAVAKTI-QGEPTKPDHVNVPSAVFTQPPIGTAGLTEEEAKEQFDEVDVYT	2706			
Query	463	SNFRPLKATLSGLPDRVFMKLIVCANTNKVLGVHMCGEDSPEIIQGFGVAVKAGLTKADF S+FRP+K T+SG +R MK+IV T+KVLG+HM GE SPEI+QGF VA+K G TK	522			
Sbjct	2705	SSFRPMKHTISGRDERSLMKIIVDVKTDKVLGIHMLGESSPEILQGFAVALKCGATKKQL	2526			
Query	523	DATVGVHPTAAEEFVTMRAPTRKFR 547 DAT+G+HPTAAEEFVTMR TR+ R				
Sbjct	2525	DATIGIHPTAAEEFVTMRTVTRQHR 2451				

[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 [from EMBOSS Transeq of the TBLASTN result in Q2]: >2451-3941_6 chlorokybus_71836.3_lrc301_c Chlorokybus atmophyticus EST library Chlorokybus atmophyticus cDNA 5', mRNA sequence ALLPAPRVAAVATRSSSSRRSELPGLLARPLGVSRSFRGFSGLRPSGLAQGQSRPHSRQQ SRTVQRYGLRVIASSNGSGYDYDVITIGAGSGGVRASRIASQLGAKVACVEMPFNNVASD TEGGVGGTCVLRGCVPKKLLVYGSIFSNEFDDSAGFGWKLPGEPKFTWQTLNENKNKELT RLNNVYRNILSKANVELLEGRASLVDAHTVDIDGKQLTAKNIILATGGRSFALPIPGAEH AIDSDKALSLDEVPKRIAIYGGGYIALEFACIFSGFGAKVDVFYRAPLPLRGFDEEIRNA LVEELGKKGINLHPKCTAEEIRKEANGEYTLKTNCGEFKADLVMFATGRTPNTKYLNLDA VGVDTTEKGAIVVDEYSRTTVPNIFAIGDVTNRINLTPVALMEGTAVAKTIQGEPTKPDH VNVPSAVFTQPPIGTAGLTEEEAKEQFDEVDVYTSSFRPMKHTISGRDERSLMKIIVDVK TDKVLGIHMLGESSPEILQGFAVALKCGATKKQLDATIGIHPTAAEEFVTMRTVTRQHRK EKQQQQQEEKEKVAAAK*

Name (Unofficial): Chlorokybus glutathione reductase.

Species: Chlorokybus atmophyticus: Cellular organisms;

Eukaryota; Viridiplantae; Streptophyta;

Chlorokybophyceae; Chlorokybales; Chlorokybaceae;

Chlorokybus

[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.

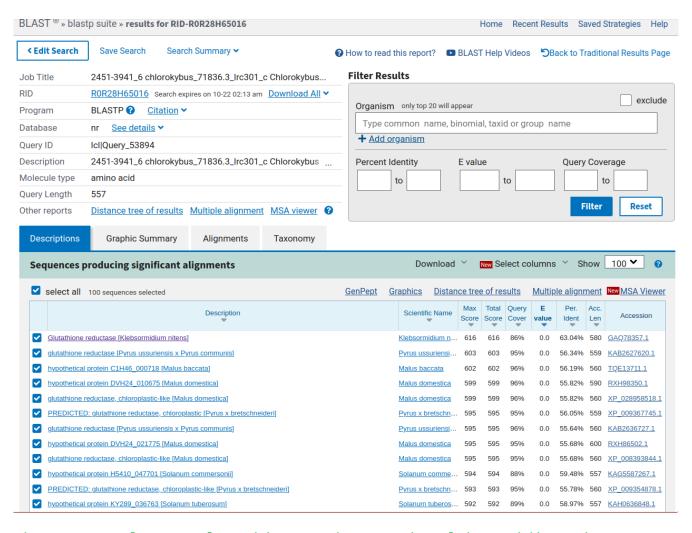
Conducted a BLASTP search against the NR database using the protein sequences listed in [Q3]:

Method: NCBI BLASTP (v2.12.0) searched all nr protein entries

Database: non-redundant protein sequences [nr] database

Organism: None

Results:



The top result was glutathione reductase in *Klebsormidium nitens* highlighted in purple above. Alignment is printed below:

Query: 2451-3941_6 chlorokybus_71836.3_lrc301_c Chlorokybus atmophyticus EST library Chlorokybus atmophyticus cDNA 5', mRNA sequence Query ID: lcl|Query_53894 Length: 557

```
>Glutathione reductase [Klebsormidium nitens]
Sequence ID: GAQ78357.1 Length: 580
Range 1: 93 to 579

Score:616 bits(1588), Expect:0.0,
Method:Compositional matrix adjust.,
Identities:307/487(63%), Positives:366/487(75%), Gaps:3/487(0%)

Query 73 ASSNGSGYDYDVITIGAGSGGVRASRIASQLGAKVACVEMPFNNVASDTEGGVGGTCVLR 132
++ +G +DYD+ TIGAGSGGVRASR ASQ GAKVA E+PF+ ASD +GGVGGTCVLR
Sbjct 93 STEDGQQFDYDLFTIGAGSGGVRASRFASQYGAKVAVCELPFSTKASDDKGGVGGTCVLR 152

Query 133 GCVPKKLLVYGSIFSNEFDDSAGFGWKLPG-EPKFTWQTLNENKNKELTRLNNVYRNILS 191
```

Sbjct	153	GCVPKKLLVYGS F++ F+DS GFGW PG EP+ W L E KNKEL RLNN Y+ L GCVPKKLLVYGSHFADYFEDSRGFGWSFPGGEPEVDWSHLIEKKNKELDRLNNAYKTTLK	212
Query	192	KANVELLEGRASLVDAHTVDIDGKQLTAKNIILATGGRSFALPIPGAEHAIDSDKALSLD A V+L+EG+ ++VD HTVD+DGK+ KNI++ATGGR F PIPGAEH I SD AL L	251
Sbjct	213	NAKVDLIEGKGTIVDRHTVDVDGKRFKVKNILIATGGRIFVPPIPGAEHVITSDDALDLT	272
Query	252	EVPKRIAIYGGGYIALEFACIFSGFGAKVDVFYRAPLPLRGFDEEIRNALVEELGKKGIN VP +IAI GGGYIALEFA IF+ GA+VD+F R LRGFD+E+R L E+L +GI	311
Sbjct	273	SVPSKIAIVGGGYIALEFAGIFNSAGAEVDIFVRGDKLLRGFDDEVREFLAEQLQAQGIR	332
Query	312	LHPKCTAEEIRKEANGEYTLKTNCGE-FKADLVMFATGRTPNTKYLNLDAVGVDTTEKGA +H EI K + TLKT G+ ++ VMFATGR PN K L L+ GVD +K A	370
Sbjct	333	IHFGAKPVEIEKRDEDQLTLKTEQGDTWQGSHVMFATGRRPNIKGLGLEEAGVDVDDKTA	392
Query	371	IVVDEYSRTTVPNIFAIGDVTNRINLTPVALMEGTAVAKT-IQGEPTKPDHVNVPSAVFT I VDEYSRT+V NI+A+GDVT+RINLTPVALMEG A AKT Q EPTKPDH NVPSAVFT	429
Sbjct	393	IKVDEYSRTSVDNIWAVGDVTDRINLTPVALMEGMAFAKTAFQDEPTKPDHTNVPSAVFT	452
Query	430	QPPIGTAGLTEEEAKEQFDEVDVYTSSFRPMKHTISGRDERSLMKIIVDVKTDKVLGIHM PPIGT GLTE EA EQ+ +VDV+TS+FRPMK TISG R+ +KI+VD TDKV+G+HM	489
Sbjct	453	NPPIGTVGLTEAEAVEQYGDVDVFTSTFRPMKSTISGNPVRTFVKILVDAATDKVIGLHM	512
Query	490	LGESSPEILQGFAVALKCGATKKQLDATIGIHPTAAEEFVTMRTVTRQHRKEKQQQQQEE GE PEI+QGFAVA++ G TKKQ+D+T+GIHPT+AEE VTMRT TRQ RKE+ Q + E	549
Sbjct	513	CGEDGPEIMQGFAVAVRMGVTKKQMDSTVGIHPTSAEELVTMRTPTRQIRKEEAQNGKGE	572
Query	550	KEKVAAA 556 KE AAA	
Sbjct	573	KEMAAAA 579	

This result matches almost all classification of "novel" for this class project:

- There is no 100% match found in the database for the protein sequence in the original species (*Chlorokybus atmophyticus*).
- The top match reported has less than 100% identity.
- There is no 100% match to a different species (redundant to above).
- x There are no database matches to the original query when no organism is specified, but when you refine the search to the original query (*Arabidopsis thaliana*) there is a database match (see below):

