Find a Gene Project - BGGN 213 AY2021 Fall
Jack Reddan (PID: A59010543)

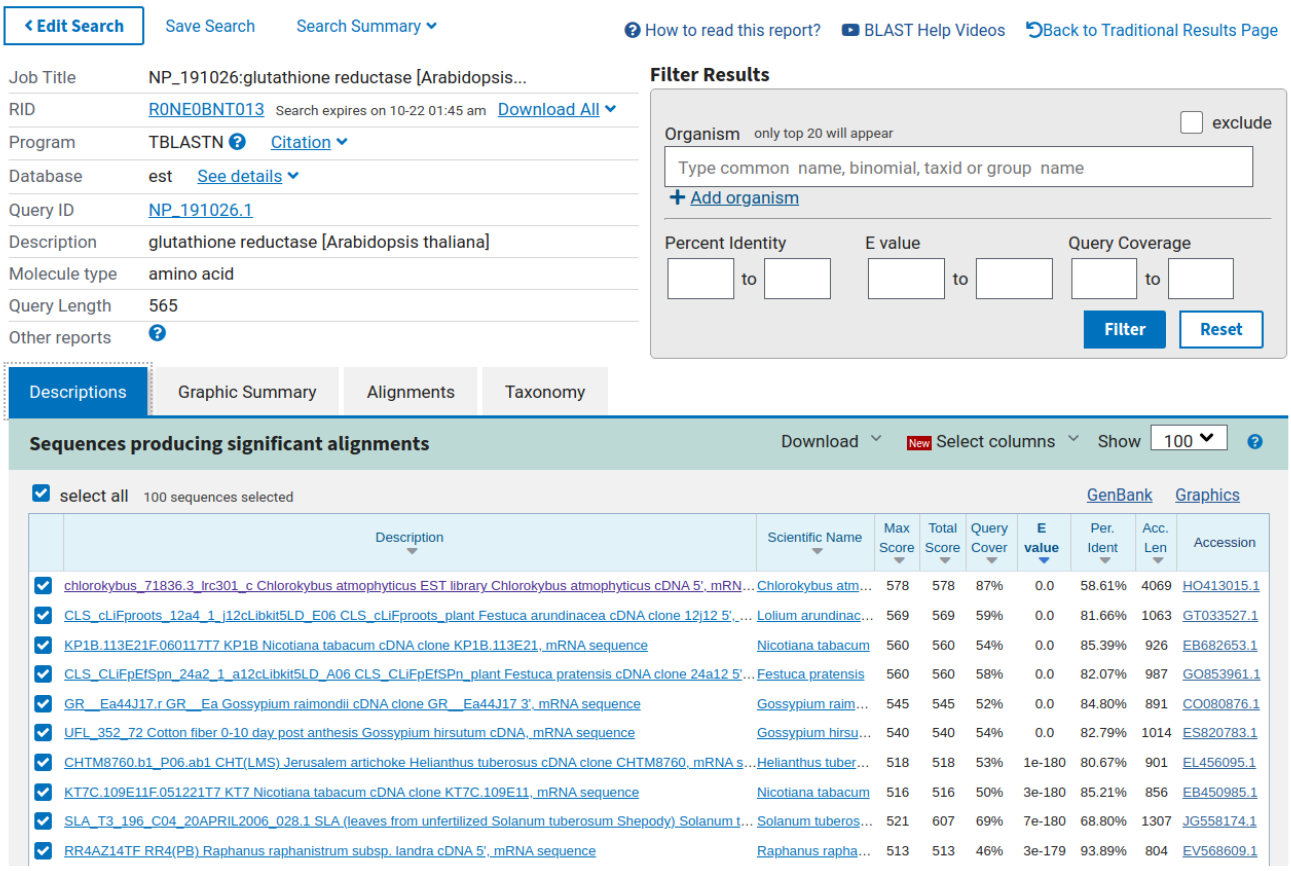
[Q01] ***Tell me the name of a protein you are interested in. Include the species and the accession number. This can be a human protein or a protein from any other species as long as it's function is known.***

Name: glutathione reductase
Accession: NP_191026
Species: *Arabidopsis thaliana*
Function: glutathione-disulfide reductase activity

Obtained from NCBI (1).

[Q02] ***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 HO413015.1, 4069 bp *Chlorokybus atmophyticus* mRNA sequence, highlighted in purple above (2). 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: HO413015.1 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%)

```
Query  52    LRPRIALLSNHRYYHSRRFS---------VCASTDNGAESDRHYDFDLFTIGAGSGGVRA  102
             LRP   L       HSR+ S         V AS++        YD+D+ TIGAGSGGVRA
Sbjct  3941  LRPS-GLAQGQSRPHSRQQSRTVQRYGLRVIASSNGSG-----YDYDVITIGAGSGGVRA  3780

Query  103   SRFATSFGASAAVCELPFSTISSDTAGGVGGTCVLRGCVPKKLLVYASKYSHEFEDSHGF  162
             SR A+  GA  A  E+PF+ ++SDT GGVGGTCVLRGCVPKKLLVY S +S+EF+DS GF
Sbjct  3779  SRIASQLGAKVACVEMPFNNVASDTEGGVGGTCVLRGCVPKKLLVYGSIFSNEFDDSAGF  3600

Query  163   GWKYETEPSHDWTTLIANKNAELQRLTGIYKNILSKANVKLIEGRGKVIDPHTVDVDGKI  222
             GWK   EP   W TL  NKN EL RL  +Y+NILSKANV+L+EGR  ++D HTVD+DGK
Sbjct  3599  GWKLPGEPKFTWQTLNENKNKELTRLNNVYRNILSKANVELLEGRASLVDAHTVDIDGKQ  3420

Query  223   YTTRNILIAVGGRPFIPDIPGKEFAIDSDAALDLPSKPKKIAIVGGGYIALEFAGIFNGL  282
              T +NI++A GGR F   IPG E AIDSD AL L   PK+IAI GGGYIALEFA IF+G
Sbjct  3419  LTAKNIILATGGRSFALPIPGAEHAIDSDKALSLDEVPKRIAIYGGGYIALEFACIFSGF  3240

Query  283   NCEVHVFIRQKKVLRGFDEDVRDFVGEQMSLRGIEFHTEESPEAIIKAGDGSFSLKTSKG  342
               +V VF R    LRGFDE++R+ + E++  +GI  H + + E I K  +G ++LKT+ G
Sbjct  3239  GAKVDVFYRAPLPLRGFDEEIRNALVEELGKKGINLHPKCTAEEIRKEANGEYTLKTNCG  3060

Query  343   TVEGFSHVMFATGRKPNTKNLGLENVGVKMAKNGAIEVDEYSQTSVPSIWAVGDVTDRIN  402
               +    VMFATGR PNTK L L+ VGV   + GAI VDEYS+T+VP+I+A+GDVT+RIN
Sbjct  3059  EFKA-DLVMFATGRTPNTKYLNLDAVGVDTTEKGAIVVDEYSRTTVPNIFAIGDVTNRIN  2883

Query  403   LTPVALMEGGALAKTLFQNEPTKPDYRAVPCAVFSQPPIGTVGLTEEQAIEQYGDVDVYT  462
             LTPVALMEG A+AKT+ Q EPTKPD+  VP AVF+QPPIGT GLTEE+A EQ+ +VDVYT
Sbjct  2882  LTPVALMEGTAVAKTI-QGEPTKPDHVNVPSAVFTQPPIGTAGLTEEEAKEQFDEVDVYT  2706

Query  463   SNFRPLKATLSGLPDRVFMKLIVCANTNKVLGVHMCGEDSPEIIQGFGVAVKAGLTKADF  522
             S+FRP+K T+SG  +R  MK+IV   T+KVLG+HM GE SPEI+QGF VA+K G TK
Sbjct  2705  SSFRPMKHTISGRDERSLMKIIVDVKTDKVLGIHMLGESSPEILQGFAVALKCGATKKQL  2526

Query  523   DATVGVHPTAAEEFVTMRAPTRKFR  547
             DAT+G+HPTAAEEFVTMR  TR+ R
Sbjct  2525  DATIGIHPTAAEEFVTMRTVTRQHR  2451
```

[Q03] ***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.***

EMBOSS Transeq for TBLASTN result in Q2 (3) –

Input:

DNA/RNA: Chlorokybus atmophyticus cDNA (Acc HO413015.1)
FRAME: 6 (All six frames)
CODON TABLE: Standard Code

Chosen sequence:

>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*

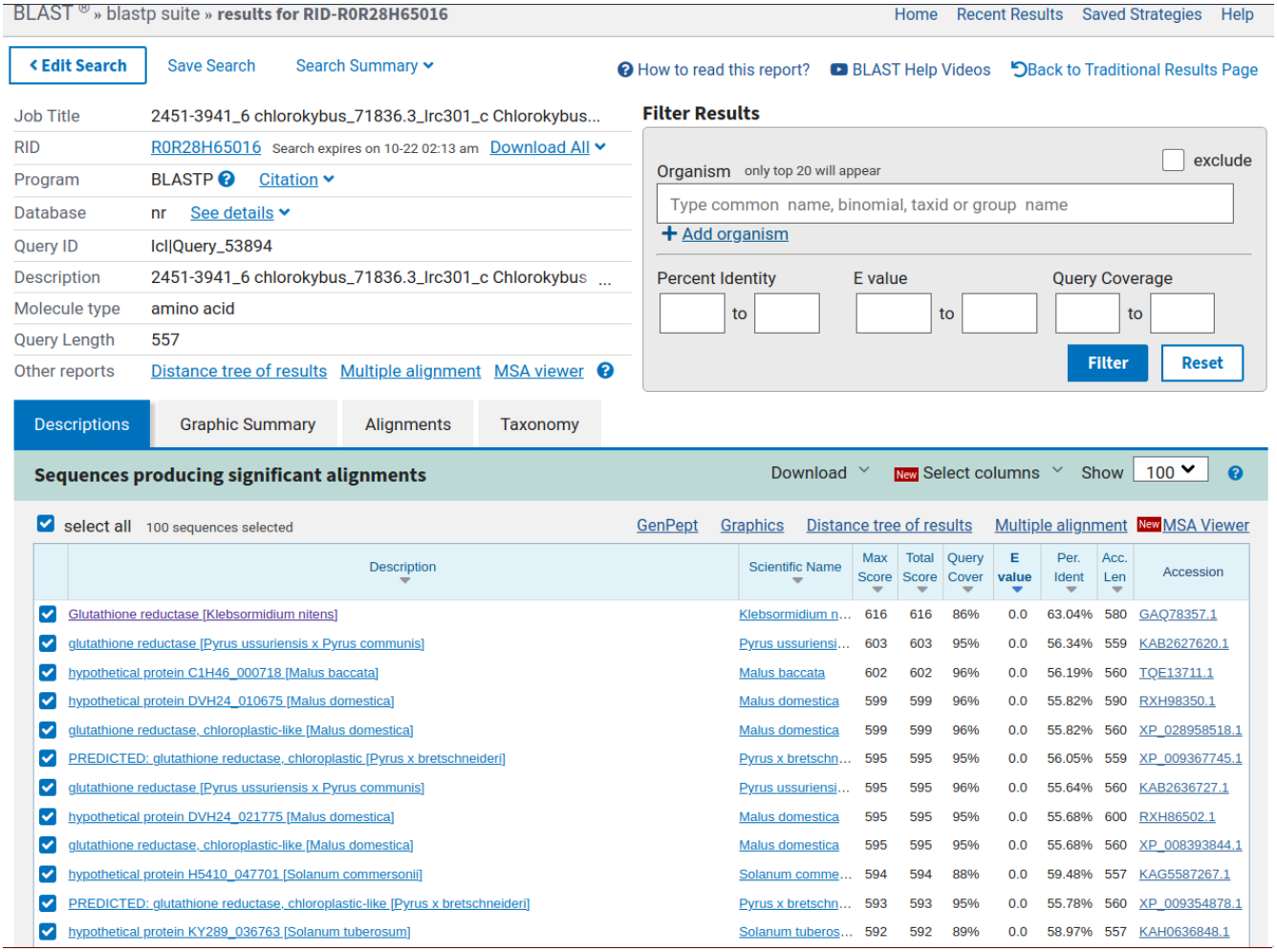
Name (Unofficial): *Chlorokybus* putative glutathione reductase.

Species (4): *Chlorokybus atmophyticus*: Cellular organisms; Eukaryota; Viridiplantae; Streptophyta; Chlorokybophyceae; Chlorokybales; Chlorokybaceae; Chlorokybus

[Q04] ***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] (2):

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
            GCVPKKLLVYGS F++ F+DS GFGW  PG EP+  W  L E KNKEL RLNN Y+  L
Sbjct  153  GCVPKKLLVYGSHFADYFEDSRGFGWSFPGGEPEVDWSHLIEKKNKELDRLNNAYKTTLK  212

Query  192  KANVELLEGRASLVDAHTVDIDGKQLTAKNIILATGGRSFALPIPGAEHAIDSDKALSLD  251
             A V+L+EG+ ++VD HTVD+DGK+   KNI++ATGGR F  PIPGAEH I SD AL L
Sbjct  213  NAKVDLIEGKGTIVDRHTVDVDGKRFKVKNILIATGGRIFVPPIPGAEHVITSDDALDLT  272

Query  252  EVPKRIAIYGGGYIALEFACIFSGFGAKVDVFYRAPLPLRGFDEEIRNALVEELGKKGIN  311
             VP +IAI GGGYIALEFA IF+  GA+VD+F R    LRGFD+E+R  L E+L  +GI
Sbjct  273  SVPSKIAIVGGGYIALEFAGIFNSAGAEVDIFVRGDKLLRGFDDEVREFLAEQLQAQGIR  332

Query  312  LHPKCTAEEIRKEANGEYTLKTNCGE-FKADLVMFATGRTPNTKYLNLDAVGVDTTEKGA  370
            +H      EI K    + TLKT  G+ ++   VMFATGR PN K L L+  GVD  +K A
Sbjct  333  IHFGAKPVEIEKRDEDQLTLKTEQGDTWQGSHVMFATGRRPNIKGLGLEEAGVDVDDKTA  392

Query  371  IVVDEYSRTTVPNIFAIGDVTNRINLTPVALMEGTAVAKT-IQGEPTKPDHVNVPSAVFT  429
            I VDEYSRT+V NI+A+GDVT+RINLTPVALMEG A AKT  Q EPTKPDH NVPSAVFT
Sbjct  393  IKVDEYSRTSVDNIWAVGDVTDRINLTPVALMEGMAFAKTAFQDEPTKPDHTNVPSAVFT  452

Query  430  QPPIGTAGLTEEEAKEQFDEVDVYTSSFRPMKHTISGRDERSLMKIIVDVKTDKVLGIHM  489
             PPIGT GLTE EA EQ+ +VDV+TS+FRPMK TISG   R+ +KI+VD  TDKV+G+HM
Sbjct  453  NPPIGTVGLTEAEAVEQYGDVDVFTSTFRPMKSTISGNPVRTFVKILVDAATDKVIGLHM  512

Query  490  LGESSPEILQGFAVALKCGATKKQLDATIGIHPTAAEEFVTMRTVTRQHRKEKQQQQQEE  549
             GE  PEI+QGFAVA++ G TKKQ+D+T+GIHPT+AEE VTMRT TRQ RKE+ Q  + E
Sbjct  513  CGEDGPEIMQGFAVAVRMGVTKKQMDSTVGIHPTSAEELVTMRTPTRQIRKEEAQNGKGE  572

Query  550  KEKVAAA  556
            KE  AAA
Sbjct  573  KEMAAAA  579
```

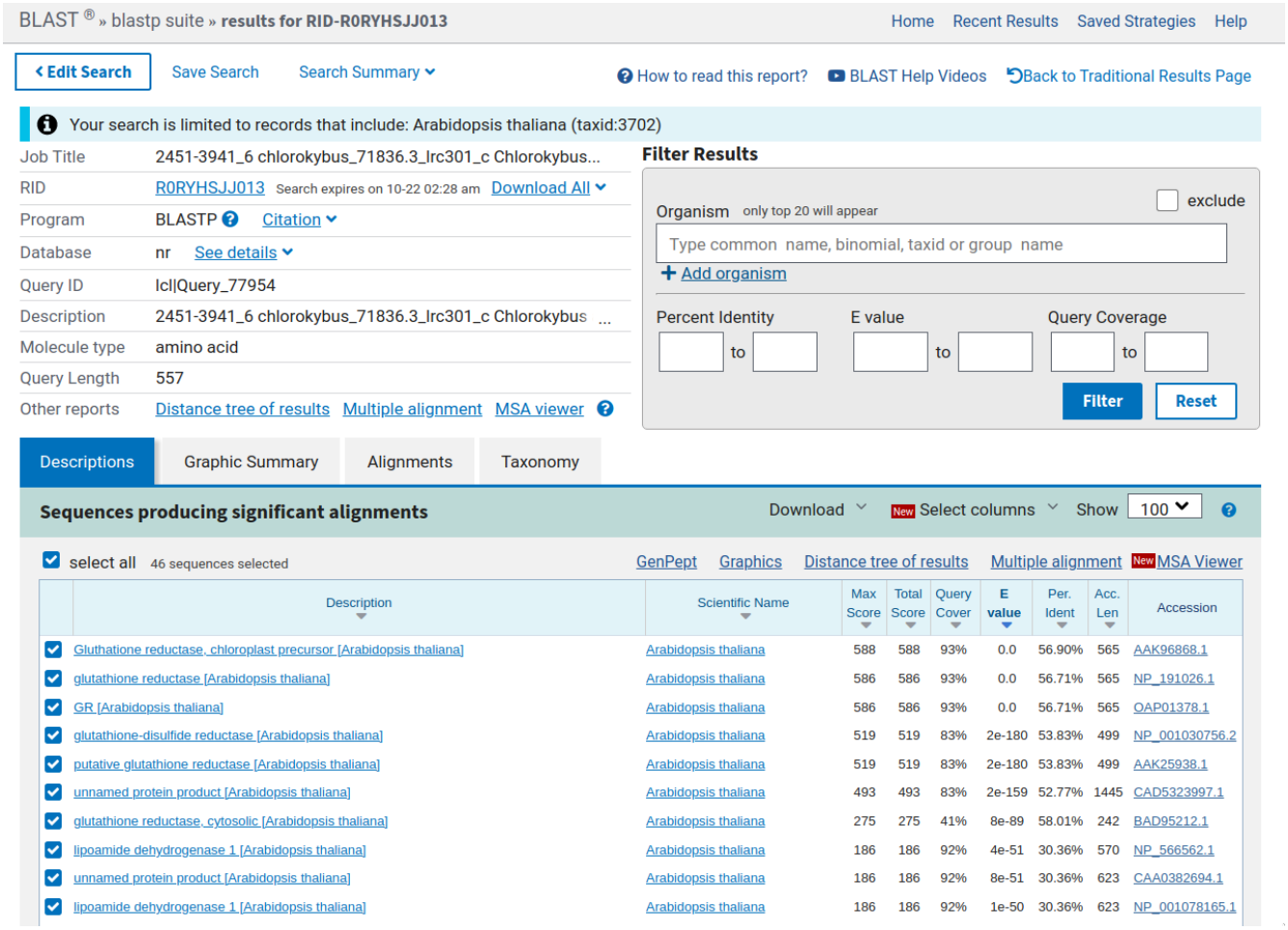
This result matches almost all classification of “novel” for this class project:

- 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 species (*Arabidopsis thaliana*) there is a database match (see below):



[Q05] ***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.***

MAFFT MSA using the GUIDANCE2 server for the novel protein [*Chlorokybus atmopyticus*], the original query [*Arabidopsis thaliana*], and the top 18 sequences (e-value) from a BLASTx using the novel protein cDNA (5, 2).

GUIDANCE2 parameters:

```
                  Sequences: All accession numbers in MSA
              Sequence Type: Amino Acids
              MSA Algorithm: MAFFT
 No. of bootstrap repeats: 100
               Output order: Aligned
               MAFFT Options
               {
                                Max-Iterate: 0
                  Pairwise alignment method: 6mer
               }
```

BLASTx parameters:

```
   Method:      NCBI BLASTX (v2.12.0) searched against all
                reference proteins using HO413015.1
 Database:      Reference proteins [refseq_protein] database
 Organism:      None
```

MSA:

```
                Malus domestica [XP_028958518.1]  ----------------------------------
        Pyrus x bretschneideri [XP_009367745.1]  --------------------------------P--
                 Pistacia vera [XP_031258822.1]  ----------------------------------
             Gossypium hirsutum [XP_016709857.2]  ----------------------------------
                 Rosa chinensis [XP_024166425.1]  ----------------------------------
    Fragaria vesca subsp. vesca [XP_004289892.1]  ----------------------------------
              Arachis hypogaea [XP_025604791.1]  ----------------------------------
            Arachis duranensis [XP_015969716.1]  ----------------------------------
              Arachis ipaensis [XP_016204709.1]  ----------------------------------
                 Prosopis alba [XP_028758428.1]  ----------------------------------
                Prunus persica [XP_007199772.1]  --------------------------------P--
            Solanum tuberosum [XP_006349748.1]  ----------------------------------
          Solanum lycopersicum [NP_001234243.2]  ----------------------------------
             Solanum pennellii [XP_015087986.1]  ----------------------------------
               Capsicum annuum [XP_016565503.1]  ----------------------------------
               Sesamum indicum [XP_011079991.1]  ----------------------------------
Arabidopsis lyrata subsp. lyrata [XP_020880886.1]  ----------------------------------
               Camelina sativa [XP_010427174.1]  ----------------------------------
           Arabidopsis thaliana [NP_191026.1]  -MASTPKLTSTISSSSPSLQFLCKKLPIAIHLPSS
       Chlorokybus atmophyticus [tHO_413015.1]  ALLPAPRVAAVATRSSSSRR---SELPGLLARP--

XP_028958518.1  -----LSLPKTLTSLSHLR----RTSTSHPHHH-----LNSRRHFSI--RASDSGNGA-DSTRHYDFDLF
XP_009367745.1  -----LSLPKTLSPLSHLR----RTSTSHPHHH-----LHSRRRFSV--RAVDSGNGA-DSTRHYDFDLF
XP_031258822.1  ------------------------------------------------------------YDFDLF
XP_016709857.2  ------------------------SRFHLHHHHPTPRFHPRRLFAV--RA-ESENGA-EPLRHYDFDLF
XP_024166425.1  ------------------------------------------------------------YDFDLF
XP_004289892.1  ---------------------------------------------------SDNGAGDPSRHYDFDLF
XP_025604791.1  ------------------------------HH----TRTRRRSFTL--SA--SANPH-----NYDFDLF
XP_015969716.1  ------------------------------HH----TRTRRRSFTL--SA--SANPH-----NYDFDLF
XP_016204709.1  ------------------------------HH----TRTRRRSFTL--SA--SANPH-----NYDFDLF
XP_028758428.1  ------------------------------------------------------------YDFDLF
```

```
XP_007199772.1  -----LSLRRTLTSLSHLH----RTSISPLHHH------SRRRSFSV--RA-DSGNGA-DSGSHYDYDLF
XP_006349748.1  ------------------------------HPR--TSSLSYGRRFTTT-RA-ESSNGA-ETPRHYDFDLF
NP_001234243.2  ------------------------------HTR--TSSLSYGRRFTTP-RA-ESSNGA-ETPRHYDFDLF
XP_015087986.1  ------------------------------HTR--TSSLSYGRRFTTP-RA-ESSNGV-ETPRHYDFDLF
XP_016565503.1  ------------------------------------------------------------YDFDLF
XP_011079991.1  --------------------------SSPHPR---------RAFTTSIRA-DSTNGS-EPPRNYDFDLF
XP_020880886.1  -----LSLPKTLTSLYSLRP---RIAVLSNHRYY------HSRRFSV--RA-STDNGA-DSERHYDFDLF
XP_010427174.1  -----LSLPKTLTSLYSLRP---RVAVLSNHRYY-----HHSRRFSV--SA-SSDNGT-DSERHYDFDLF
NP_191026.1     SSSSFLSLPKTLTSLYSLRP---RIALLSNHRYY------HSRRFSV--CA-STDNGA-ESDRHYDFDLF
tHO_413015.1    -----LGVSRSFRGFSGLRPSGLAQGQSRPHSRQ---QSRTVQRYGL--RVIASSNGS-----GYDYDVI

XP_028958518.1  TIGAGSGGVRASRFAANFGASVAVCELPFSTISSDTTGGVGGTCVLRGCVPKKLLVYASKFAHEFEESNG
XP_009367745.1  TIGAGSGGVRASRFAANFGASVAVCELPFSTISSATAGGVGGTCVLRGCVPKKLLVYASKFAHEFEESNG
XP_031258822.1  TIGAGSGGVRASRFASNFGASVAVCELPFSTISSETTGGVGGTCVLRGCVPKKLMVYASKFSHEFDESNG
XP_016709857.2  TIGAGSGGVRASRFAANFGASVAVCELPFSTISSETTGGVGGTCVLRGCVPKKLMVYASKYSHEFDESNG
XP_024166425.1  TIGAGSGGVRASRFASNFGAKVAVCELPFAAISSDTTGGVGGTCVLRGCVPKKLLVYASKYTHEFDDSIG
XP_004289892.1  TIGAGSGGVRASRFAANFGAKVAVCELPFATISSETAGGVGGTCVLRGCVPKKLMVYASKYSHEFEDSIG
XP_025604791.1  TIGAGSGGVRASRFAANYGASVAICELPFSTISSDVTGGVGGTCVLRGCVPKKLLVYSSKYSHEFEESNG
XP_015969716.1  TIGAGSGGVRASRFAANYGASVAICELPFSTISSDVTGGVGGTCVLRGCVPKKLLVYSSKYSHEFEESNG
XP_016204709.1  TIGAGSGGVRASRFAANYGASVAICELPFSTISSDVTGGVGGTCVLRGCVPKKLLVYSSKYSHEFEESNG
XP_028758428.1  TIGAGSGGVRASRFAANFGASVAICELPFSTISSDTTGGVGGTCVLRGCVPKKLLVYSSKYAHEFEESNG
XP_007199772.1  TIGAGSGGVRASRFAANFGASVAICELPFATIASDTAGGVGGTCVLRGCVPKKLLVYASQFAHEFEESNG
XP_006349748.1  TIGAGSGGVRASRFASNFGASVAVCELPFSTISSDSTGGVGGTCVLRGCVPKKLLVYASKYSHEFEESCG
NP_001234243.2  TIGAGSGGVRASRFASNFGASVAVCELPFSTISSDSTGGVGGTCVLRGCVPKKLLVYASKYSHEFEESCG
XP_015087986.1  TIGAGSGGVRASRFASNFGASVAVCELPFSTISSDSTGGVGGTCVLRGCVPKKLLVYASKYSHEFEESCG
XP_016565503.1  TIGAGSGGVRASRFASNFGASVAVCELPFSTISSDSTGGVGGTCVLRGCVPKKLLVYASKYSHEFEESCG
XP_011079991.1  TIGAGSGGVRASRFAANFGAKVAVCELPFATISSDTTGGVGGTCVLRGCVPKKLLVYASKFSHEFQESCG
XP_020880886.1  TIGAGSGGVRASRISTSFGASAAVCELPFSTISSDTAGGVGGTCVLRGCVPKKLLVYASKYSHEFEDSHG
XP_010427174.1  TIGAGSGGVRASRFATSFGASAAVCELPFSTISSDTAGGVGGTCVLRGCVPKKLLVYASKYSHEFEDSHG
NP_191026.1     TIGAGSGGVRASRFATSFGASAAVCELPFSTISSDTAGGVGGTCVLRGCVPKKLLVYASKYSHEFEDSHG
tHO_413015.1    TIGAGSGGVRASRIASQLGAKVACVEMPFNNVASDTEGGVGGTCVLRGCVPKKLLVYGSIFSNEFDDSAG

XP_028958518.1  FGWRYETEPKHDWSTLLANKNAELQRLTGIYKNVLKNANVTLIEGRGKIVDPHTVDVDGKLYSARHILVS
XP_009367745.1  FGWRYETEPKHDWSTLIANKNAELKRLTGIYKNVLSNANVTLIEGRGKIVDPHTVDVEGKLYSARHILVS
XP_031258822.1  FGWKYEVEPQHDWSTLIANKNAELQRLTGIYKNVLKNANVTLLEGRGKIVDPHTVDVDGKLYTARHILIS
XP_016709857.2  FGWKYDTEPKHDWSTLMANKNAELQRLTGIYKNILNKAGVTLIEGRGKVVDPHTVDVDGKLYTARHILIS
XP_024166425.1  YGWKYETEPKHDWSTLMANKNAELQRLTGIYKNVLKNANVTLVQGRGKVVDPHTVDVDGKLYSARHILIS
XP_004289892.1  YGWKYETEPKHDWSTLMANKNAELQRLTGIYKNVLKNANVSLLEGRGKIVDPHTVDVDGQLYSARHILIS
XP_025604791.1  FGWKYESEPKHDWSTLIANKNAELQRLTGIYKNILKNSNVQLIEGRGKIVDPHTVDVDGKLYTARHILVS
XP_015969716.1  FGWKYESEPKHDWSTLIANKNAELQRLTGIYKNILKNSNVQLIEGRGKIVDPHTVDVDGKLYTARHILVS
XP_016204709.1  FGWKYESEPKHDWSTLIANKNAELQRLTGIYKNILKNSNVQLIEGRGKIVDPHTVDVDGKLYTARHILVS
XP_028758428.1  FGWKYDSEPKHDWSTLMANKNAELQRLTGIYKNILKNSDVKLIEGRGKIVDPHTVDVDGKLYSARHIIIS
XP_007199772.1  FGWRYETEPKHDWSTLMANKNAELQRLTGIYKNVLKNAGVALIEGRGKIVDPHTVDVDGKLYSARHIIVS
XP_006349748.1  FGWNYEAEPKHDWSILIANKNAELQRLTGIYKNILKNADVTLIEGRGKVVDPHTVDVDGKLYSAKNILIS
NP_001234243.2  FGWNYEAEPKHDWSTLIANKNAELQRLTGIYKNILKNADVTLIEGRGKVVDPHTVDVDGKLYSAKNILIS
XP_015087986.1  FGWNYEAEPKHDWSTLIANKNAELQRLTGIYKNILKNADVTLIEGRGKVVDPHTVDVDGKLYSAKNILIS
XP_016565503.1  FGWNYEAEPKHDWNTLIANKNAELQRLMGIYKNILKNANVTLIEGRGKVVDPHTVDVDGKLYSAKNILIS
XP_011079991.1  FGWNYEAEPKHDWSTLIANKNAELQRLTGIYKNILKNAGVALIEGRGKIVDPHTVEVNGKLYSAKNILIS
XP_020880886.1  FGWKYETEPSHDWTTLIANKNAELQRLTGIYKNILSKANVKLIEGRGKVIDPHTVDVDGKIYTTRNILIA
XP_010427174.1  FGWKYDTEPTHDWSTLIANKNAELQRLTGIYKNILSKANVKLVEGRGKVIDPHTVDVDGKIYTTRNILIA
NP_191026.1     FGWKYETEPSHDWTTLIANKNAELQRLTGIYKNILSKANVKLIEGRGKVIDPHTVDVDGKIYTTRNILIA
tHO_413015.1    FGWKLPGEPKFTWQTLNENKNKELTRLNNVYRNILSKANVELLEGRASLVDAHTVDIDGKQLTAKNIILA

XP_028958518.1  VGGRPFIPEIPGSEYAIDSDAALDLPTKPEKIAIVGGGYIAVEFAGIFNGLTSDVHVFIRQKKVLRGFDE
XP_009367745.1  VGGRPFIPEIPGSEYAIDSDAALDLPTKPEKIAIVGGGYIAVEFAGIFNGLTSDVHVFIRQKKVLRGFDE
XP_031258822.1  VGGRPFIPDIPGSEHAIDSDAALDLPSKPKKIAIIGGGYIALEFAGIFNGLTSDVHVFIRQKKVLRGFDE
XP_016709857.2  VGGRPFIPDIPGSEYAIDSDAALDLPSKPEKVAIVGGGYIALEFAGIFNGLTSEVHVFIRQKKVLRGFDE
XP_024166425.1  VGGRPFIPDIPGSEYAIDSDAALDLPDKPGKIAIVGGGYIALEFAGIFNGLRSDVHVFIRQKQVLRGFDE
XP_004289892.1  VGGRPFIPDIPGSKYAIDSDAALDLPERPGKIAIVGGGYIALEFAGIFNGLKSDVHVFIRQKQILRGFDD
```

```
XP_025604791.1  VGGRPFIPDIPGSEHAIDSDAALDLPSKPEKIAIVGGGYIALEFAGIFNGLKSDVHVFIRQKKVLRGFDE
XP_015969716.1  VGGRPFIPDIPGSEHAIDSDAALDLPSKPEKIAIVGGGYIALEFAGIFNGLKSDVHVFIRQKKVLRGFDE
XP_016204709.1  VGGRPFIPDIPGSEHAIDSDAALDLPSKPEKIAIVGGGYIALEFAGIFNGLKSDVHVFIRQKKVLRGFDE
XP_028758428.1  VGGRPFIPDIPGREYAIDSDAALDLPSKPEKIAIVGGGYIALEFAGIFNGLASEVHVFIRQKKVLRGFDE
XP_007199772.1  VGGRPFIPEIPGIEYAIDSDAALDLPSKPKKIAIVGGGYIAVEFAGIFNGLSSDVHVFIRQKKVLRGFDE
XP_006349748.1  VGGRPFIPDIPGSEYAIDSDAALDLPTKPDKIAIVGGGYIALEFAGIFNGLKSEVHVFIRQKKVLRGFDE
NP_001234243.2  VGGRPFIPDIPGSEYAIDSDAALDLPTKPDKIAIVGGGYIALEFAGIFNGLKSEVHVFIRQKKVLRGFDE
XP_015087986.1  VGGRPFIPDIPGSEYAIDSDAALDLPTKPDKIAIVGGGYIALEFAGIFNGLKSEVHVFIRQKKVLRGFDE
XP_016565503.1  VGGRPFIPNIPGSEYAIDSDAALDLPTKPNKIAIVGGGYIALEFAGIFNGLTSEVHVFIRQKKVLRGFDE
XP_011079991.1  VGGRPFIPDIPGREYVIDSDAALDLPSKPTKIAIVGGGYIALEFAGIFNGLTSSVHVFIRQKKVLRGFDE
XP_020880886.1  VGGRPFIPDIPGKEFAIDSDAALDLPSKPKKIAIVGGGYIALEFAGIFNGLNSEVHVFIRQKKVLRGFDE
XP_010427174.1  VGGRPFIPDIPGKEFAIDSDAALDLPSKPKKIAIVGGGYIALEFAGIFNGLNSEVHVFIRQKKVLRGFDE
NP_191026.1     VGGRPFIPDIPGKEFAIDSDAALDLPSKPKKIAIVGGGYIALEFAGIFNGLNCEVHVFIRQKKVLRGFDE
tHO_413015.1    TGGRSFALPIPGAEHAIDSDKALSLDEVPKRIAIYGGGYIALEFACIFSGFGAKVDVFYRAPLPLRGFDE

XP_028958518.1  EVRDFVQEQMALRGIEFHTEESPQAIVKAADGSLSLKTNKGTIEGFSHIMFATGRRPNTKNLGLEAIGVK
XP_009367745.1  EVRDFVQEQMALRGIEFHAEESPQAIVKAADGSLSLKTNKGTIEGFSHIMFATGRRPNTKDLGLEAVGVK
XP_031258822.1  EIRDFVAEQMSVRGIEFHTEESPEAILKSADGSLSLKTNKGTVEGFSHIMFATGRRPNTKNLGLEKVGVK
XP_016709857.2  EIRDFVGEQMALRGIQFHTEESPQAIVKSADGSLSLKTNEGTIEGFSHIMFATGRRPNTKNLGLESVGVK
XP_024166425.1  EIRGFVSEQMSVRGIEFHTEESPQAILKSADGSFSLKTNKGTVGGFSHVMFATGRRPNTKNLGLEEVGVK
XP_004289892.1  EIRGFLAEQMSLRGIEFHTEESPQAILKSSDGSFSLKTNKGTVEGFSHVMFATGRRPNTKNLGLEAVGVK
XP_025604791.1  EIRDFVGEQMALRGIEFHTEESPQAIVKSADGSLSLKTNKGTVEGFSYIMFATGRKPNTKNIGLESVGVK
XP_015969716.1  EIRDFVGEQMALRGIEFHTEESPQAIVKSADGSLSLKTNKGTVEGFSHIMFATGRKPNTKNIGLESVGVK
XP_016204709.1  EIRDFVGEQMALRGIEFHTEESPQAVVKSADGSLSLKTNKGTVEGFSHIMFATGRKPNTKNIGLESVGVK
XP_028758428.1  EVRDFVSEQMAIRGIEFHVEETPQAIIKSADGSLSLKTNKGTVEGFSHIMFATGRTPNTKNLGLESVGVK
XP_007199772.1  EVRDFVQEHMSLRGIEFHAEESPQAIVKSADGSLSLKTNKGTLEGFSHIMFATGRRPNTKNLGLEEVGVK
XP_006349748.1  EIRDFVGEQMSLRGIEFHTEESPQAIVKSADGSLSLKTNRGTVEGFSHIMFATGRSPNTKNLGLDTVGVK
NP_001234243.2  EIRDFVGEQMSLRGIEFHTEESPQAIVKSADGSLSIKTNRGTVEGFSHIMFATGRSPNTKNLGLDTVGVK
XP_015087986.1  EIRDFVGEQMSLRGIEFHTEESPQAIVKSADGSLSIKTNRGTVEGFSHIMFATGRSPNTKNLGLDTVGVR
XP_016565503.1  EIRDFVGEQMSLRGIEFHTEESPQAIVKSADGSLSLKTNRGTVEGFSHIMFATGRRPNTKNLGLDTVGVK
XP_011079991.1  EIRDFVGEQMSLRGIEFHTEETPQAIVKSSDGLLSLKTNKGTVDGFSHVMFATGRRPNTKNLGLEAVGVK
XP_020880886.1  DVRDFVGEQMSLRGIEFHTEESPEAIIKAGDGSLSLKTSKGTVEGFSHVMFATGRKPNTKNLGLENVGVK
XP_010427174.1  DVRDFVGEQMSLRGIEFHTEESPEAIIKSGDGSLSLKTSKGTVEGFSHVMFATGRKPNTKNLGLENVGVK
NP_191026.1     DVRDFVGEQMSLRGIEFHTEESPEAIIKAGDGSFSLKTSKGTVEGFSHVMFATGRKPNTKNLGLENVGVK
tHO_413015.1    EIRNALVEELGKKGINLHPKCTAEEIRKEANGEYTLKTNCGEFKA-DLVMFATGRTPNTKYLNLDAVGVD

XP_028958518.1  LSKNGAIEVDEFSRTEVPSIWAIGDVTDRVNLTPIALMEGGAIAKTLFLNEPTKPDYRAVPSAVFSQPPI
XP_009367745.1  LSKNGAIEVDKFSRTAVPSIWAVGDVTDRVNLTPVALMEGGAIAKTLFLNEPTMPDYRAVPSAVFSQPPI
XP_031258822.1  MSKNGAIQVDEYSRTSVPSIWAVGDVTDRINLTPVALMEGGALAKTLFQDEPTKPDYRAVPSAVFCQPPI
XP_016709857.2  INKNGAIEVDEYSRTTVPSIWAVGDVTDRINLTPVALMEGAALAKTLFQNEPTKPDYRAVPSAVFSQPPI
XP_024166425.1  IAKNGAIEVDEFSRTSVPSIWAVGDVTDRVNLTPVALMEGGALAKTLFLNEPTKPDYRAIPSAVFSQPPI
XP_004289892.1  MANSGAIEVDEFSRTSVPSIWAVGDVTDRVNLTPVALMEGGALAKTLFLNEPTKPDYRAIPSAVFSQPPI
XP_025604791.1  IDKKGAIEVNEYSQSSVPSIWAVGDVTDRINLTPVALMEGVALAKTLFLNEPTKPDYSYVPSAVFSQPPI
XP_015969716.1  IDKKGAIEVNEYSQSSVPSIWAVGDVTDRINLTPVALMEGVALAKTLFLNEPTKPDYSYVPSAVFSQPPI
XP_016204709.1  IDKKGAIEVNEYSQSSVPSIWAVGDVTDRINLTPVALMEGVALAKTLFLNEPTKPEYSYVPSAVFSQPPI
XP_028758428.1  TAKNGAIEVDEYSQTSVPSIWAVGDVTDRMNLTPVALMEGMALAKTLFQNNPTKPDYRAVPSAVFSQPPI
XP_007199772.1  LSKTGAIEVDEFSRTSVPSIWAVGDVTDRVNLTPVALMEGGALAKTLFLNEPTKPDYRAVPSAVFSQPPI
XP_006349748.1  MTKNGAIEVDEYSRTSVPSIWAVGDVTDRINLTPVALMEGGALAKTIFAGEPTKPDYRNVPCAVFSQPPI
NP_001234243.2  MTKNGAIEVDEYSRTSVPSIWAVGDVTDRINLTPVALMEGGALAKTIFAGEPTKPDYRNVPCAVFSQPPI
XP_015087986.1  MTKNGAIEVDEYSRTSVPSIWAVGDVTDRINLTPVALMEGGALAKTIFAGEPTKPDYRNVPCAVFSQPPI
XP_016565503.1  MARNGAIEVDEYSRTSVPSIWAVGDVTDRINLTPVALMEGGALAKTIFAGEPTKPDYRNVPCAVFSQPPI
XP_011079991.1  LSKNGAVEVDEYSRTSVPSIWAVGDVTDRINLTPVALMEGGALAKTLFANEPTKPDFSNVPSAVFSQPPI
XP_020880886.1  MAKNGAIEVDEYSQTSVPSIWAVGDVTDRINLTPVALMEGGALAKTLFQNEPTKPDYRAVPCAVFSQPPI
XP_010427174.1  MAKNGAIEVDEYSQTSVPSIWAVGDVTDRINLTPVALMEGGALAKTLFQNEPTKPDYRAVPCAVFSQPPI
NP_191026.1     MAKNGAIEVDEYSQTSVPSIWAVGDVTDRINLTPVALMEGGALAKTLFQNEPTKPDYRAVPCAVFSQPPI
tHO_413015.1    TTEKGAIVVDEYSRTTVPNIFAIGDVTNRINLTPVALMEGTAVAKTI-QGEPTKPDHVNVPSAVFTQPPI

XP_028958518.1  GQVGLSEEQAVEQYGDVDIYTSNFRPLKATVSGLPDRTFMKLIVCAKTNKVLGLHMCGEDSPEIVQGFAV
XP_009367745.1  GQVGLSEEQAVEQYGDVDIYTSNFRPLKATLSGLPDRTFMKLIVCAKTNKVLGLHMCGEDSPEIVQGFAV
```

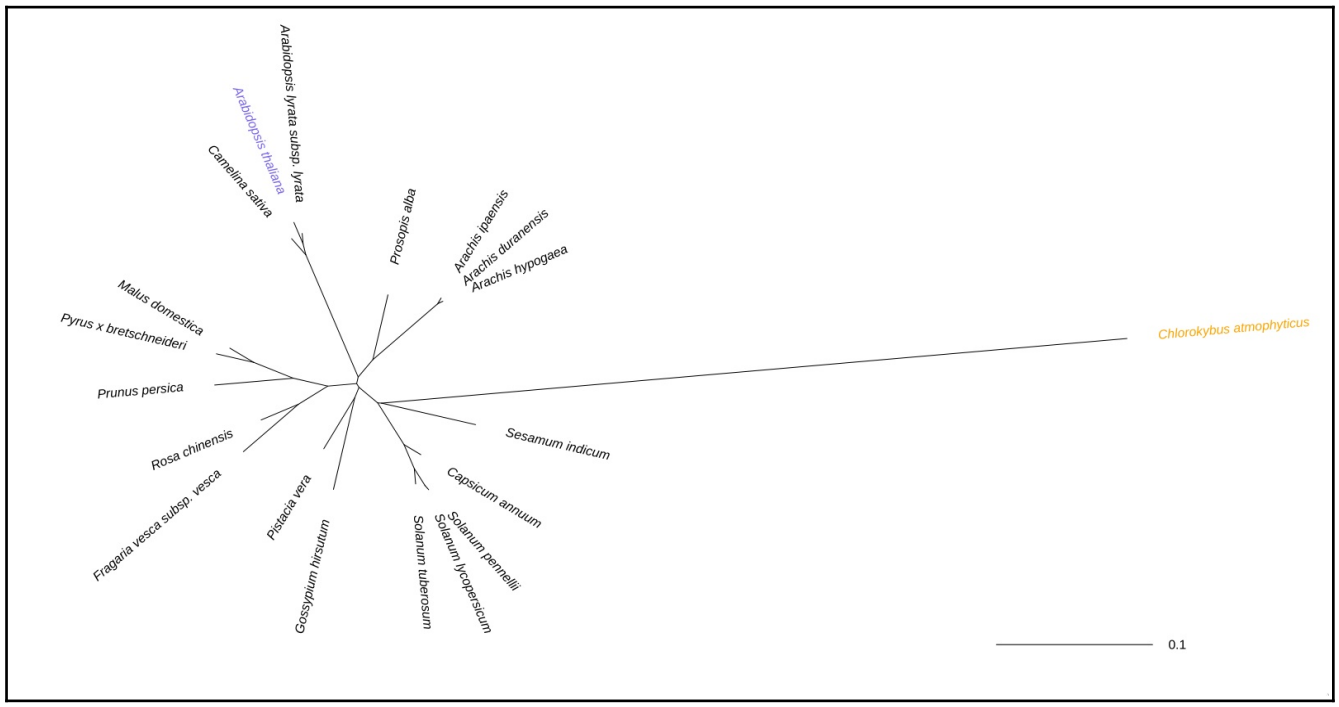
```
XP_031258822.1  GQVGLSEEQAIQEYGDIDVFTANFRPLKATLSGLPDRVFMKLIVCAKTNKVLGLHMCGEDAPEIVQGFAV
XP_016709857.2  GQVGLTEEQARKEYGDIDVYTANFRPLKATLSGLPDRVFMKLIVCAKTNKVIGLHMCGEDSAEIAQGFAV
XP_024166425.1  GQVGLSEEQATEQYGDVDIYTSNFRPMKATLSGLPDRVFMKLVVCAKTNKLLGLHMCGEDSPEIVQGFAV
XP_004289892.1  GQVGLSEEQATEQYGDVDIYTSNFKPMKATLSGLPDRVFMKLIVCAKTNKILGLHMCGDDSPEIVQGFAV
XP_025604791.1  GQVGLTEEQAVEQYGDVDIFTSNFRPLKATLSGLPDRTFMKLIVCAKTNKVLGLHMCGEDSPEITQGFAV
XP_015969716.1  GQVGLTEEQAVEQYGDVDIFTSNFRPLKATLSGLPDRTFMKLIVCAKTNKVLGLHMCGEDSPEITQGFAV
XP_016204709.1  GQVGLTEEQAVEQYGDVDIFTSNFRPLKATLSGLPDRTFMKLIVCAKTNKVLGLHMCGEDSPEITQGFAV
XP_028758428.1  GQVGLTEEQAVQQYGNVDIFTANFKPLKATLSGLPDRAFMKLIVCAKTNKVLGLHMCGEDSPEIVQGFAV
XP_007199772.1  GQVGLTEEQAIEQYGDVDIYTSNFRPLKATLSGLPDRVFMKLLVCAKTNKVLGLHMCGEDSAEIVQGFAV
XP_006349748.1  GLVGLTEEEAIKEYGDVDVYTANFRPLKATLSGLPDRAFMKLIVCSKTSKVLGLHMCGEDAPEIVQGFAV
NP_001234243.2  GLVGLTEEEAIKEYGDVDVYTANFRPLKATLSGLPDRVFMKLVVCAKSSKVLGLHMCGDDAPEIVQGFAV
XP_015087986.1  GLVGLTEEEAIKEYGDVDVYTANFRPLKATLSGLPDRVFMKLVVCAKSSKVLGLHMCGDDAPEIVQGFAV
XP_016565503.1  GIVGLTEEQAINEYGDIDVYTTNFRPLKATLSGLPDRVFMKLIVCAKSSKVLGLHMCGEDAPEIVQGFAV
XP_011079991.1  GQVGLTEEQAIKEYGDIDVYTANFRPMKATLSGLPDRVFMKLIVCAKTNKVLGVHMCGEDSPEIIQGFAV
XP_020880886.1  GTVGLTEEQAIEQYGDVDVFTSNFRPLKATLSGLPDRVFMKLIVCANTNKVLGVHMCGEDSPEIIQGFGV
XP_010427174.1  GTVGLTEEQAIEQYGDVDVYTSNFRPLKATLSGLPDRVFMKLIVCANTNKVVGVHMCGEDSPEIIQGFGV
NP_191026.1     GTVGLTEEQAIEQYGDVDVYTSNFRPLKATLSGLPDRVFMKLIVCANTNKVLGVHMCGEDSPEIIQGFGV
tHO_413015.1    GTAGLTEEEAKEQFDEVDVYTSSFRPMKHTISGRDERSLMKIIVDVKTDKVLGIHMLGESSPEILQGFAV

XP_028958518.1  AVKAGLTKADLDSTIGIHPTAAEEFVTMRTPTRKIR------------------
XP_009367745.1  AVKAGLTKADLDSTIGIHPTAAEEFVTMRTPTRKIR------------------
XP_031258822.1  AVKAGLTKADFDTTVGIHPTAAEEFVTMRTPTRKIR------------------
XP_016709857.2  AVKAGLTKADFDATVGIHPTSAEEFVTMRTPTRKIR------------------
XP_024166425.1  AVKAGLTKADLDATIGIHPTAAEEFVTMRTPTRKIR------------------
XP_004289892.1  AVKAGLTKADLDATIGIHPTAAEELVTMRTPTRKIR------------------
XP_025604791.1  AIKAGLTKGDFDATVGIHPTAAEEFVTMRTPTRKIR------------------
XP_015969716.1  AIKAGLTKGDFDATVGIHPTAAEEFVTMRTPTRKIR------------------
XP_016204709.1  AIKAGLTKADFDATVGIHPTAAEEFVTMRTPTRKIR------------------
XP_028758428.1  AVKAGLTKADFDATVGVHPTAAEEFVTMRTTTRKIR------------------
XP_007199772.1  VVKAGLTKADLDATIGIHPTAAEEFVTMRTPTRKIR------------------
XP_006349748.1  AVKAGLTKADFDATVGIHPTAAEEFVTMRTPTRKVR------------------
NP_001234243.2  AVKAGLTKADFDTTVGIHPTAAEEFVTMRTPTRKIR------------------
XP_015087986.1  AVKAGLTKADFDTTVGIHPTAAEEFVTMRTPTRKIR------------------
XP_016565503.1  AVKAGLTKADFDATVGIHPTAAEEFVTMRTPTRKVR------------------
XP_011079991.1  AVKAGLTKADFDATVGIHPTAAEELVTMRTPTRKIR------------------
XP_020880886.1  AVKAGLTKADFDATVGVHPTAAEEFVTMRTPTRKIR------------------
XP_010427174.1  AVKAGLTKADFDATVGVHPTASEEFVTMRTPTRKIR------------------
NP_191026.1     AVKAGLTKADFDATVGVHPTAAEEFVTMRAPTRKFRKDSSEGKASPEAKTAAGV
tHO_413015.1    ALKCGATKKQLDATIGIHPTAAEEFVTMRTVTRQHRKEKQQQQQEEKEKVAAAK
```

[Q06] ***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.***

Phylogenetic tree parameters for 'Simple Phylogeny' on the EBI (3):

- MSA: See Q05 above
- TREE FORMAT: Default
- DISTANCE CORRECTION: on
- EXCLUDE GAPS: off
- CLUSTERING METHOD: Neighbour-joining
- P.I.M.: off

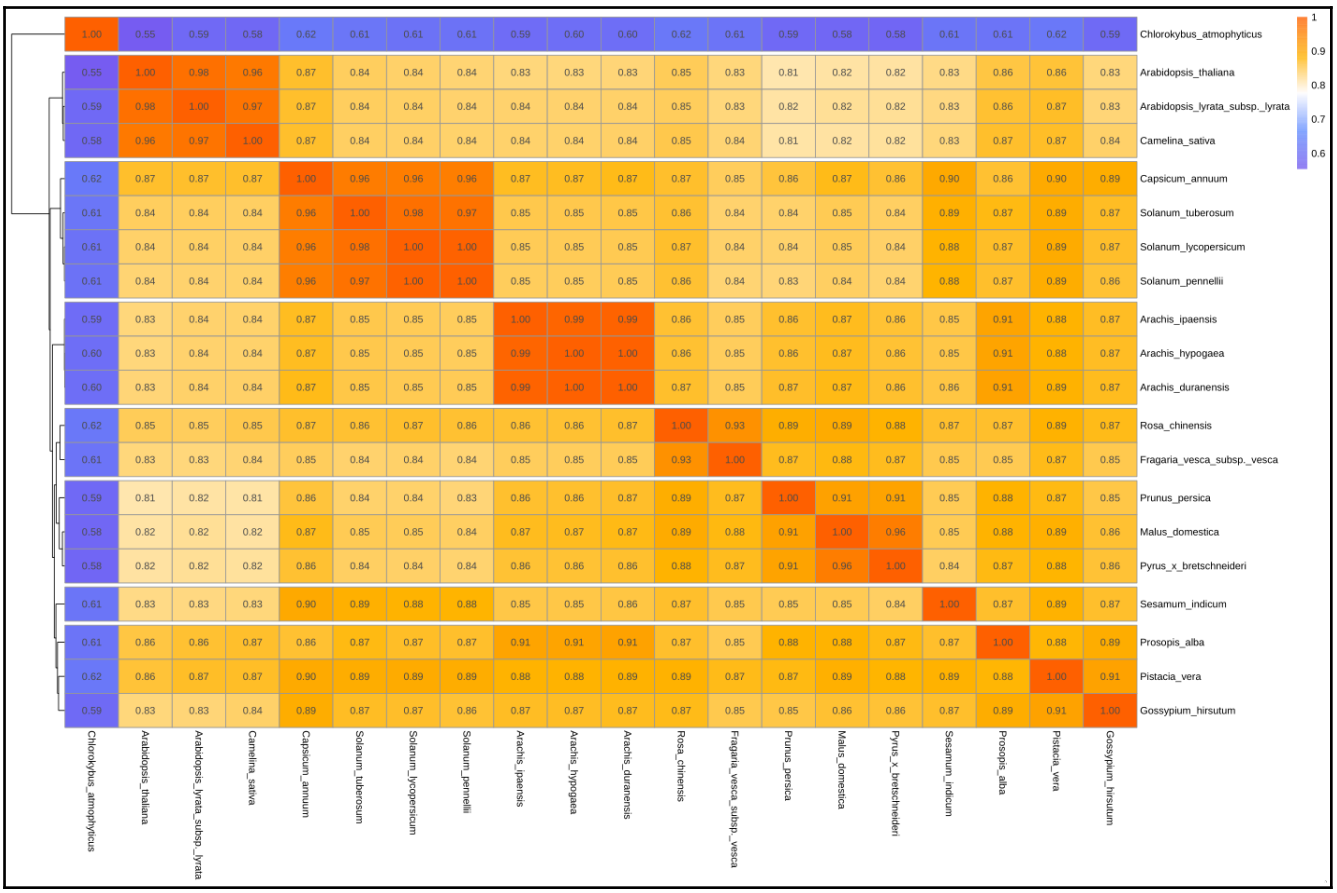
Neighbour-joining unrooted tree for alignment.
Original query protein: *Arabidopsis thaliana*
Novel protein: *Chlorokybus atmophyticus*



Made using APE in R [Appendix I] (6, 7).

[Q07] ***Generate a sequence identity based heatmap of your aligned sequences using R.***

Percent sequence identity matrix for the MSA (Q05).



Sequence identities obtained using Bio3D in R [Appendix II] (8, 7). Heatmap generated using pheatmap in R [Appendix III] (9, 7).

[Q08] ***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).***

Queried the PDB using the Bio3D package in R [Appendix IV] (10, 8, 7).

==========Hits for novel protein==========
**********BEGIN**********
Hit 01 - 3O0H_A (glutathione reductase) (11):
E-Value [1.61e-124]
Sequence ID [42.763%]
Annotations
{
PDB identifier [3O0H]
Method [X-RAY DIFFRACTION]
Resolution [1.90 Å]
Source Organism [*Bartonella henselae* str. Houston-1]
}

Hit 02 - 4DNA_A (putative glutathione reductase) (12):
E-Value [3.31e-121]
Sequence ID [45.633%]
Annotations
{
PDB identifier [4DNA]
Method [X-RAY DIFFRACTION]
Resolution [2.80 Å]
Source Organism [*Escherichia coli* BL21(DE3)]
}

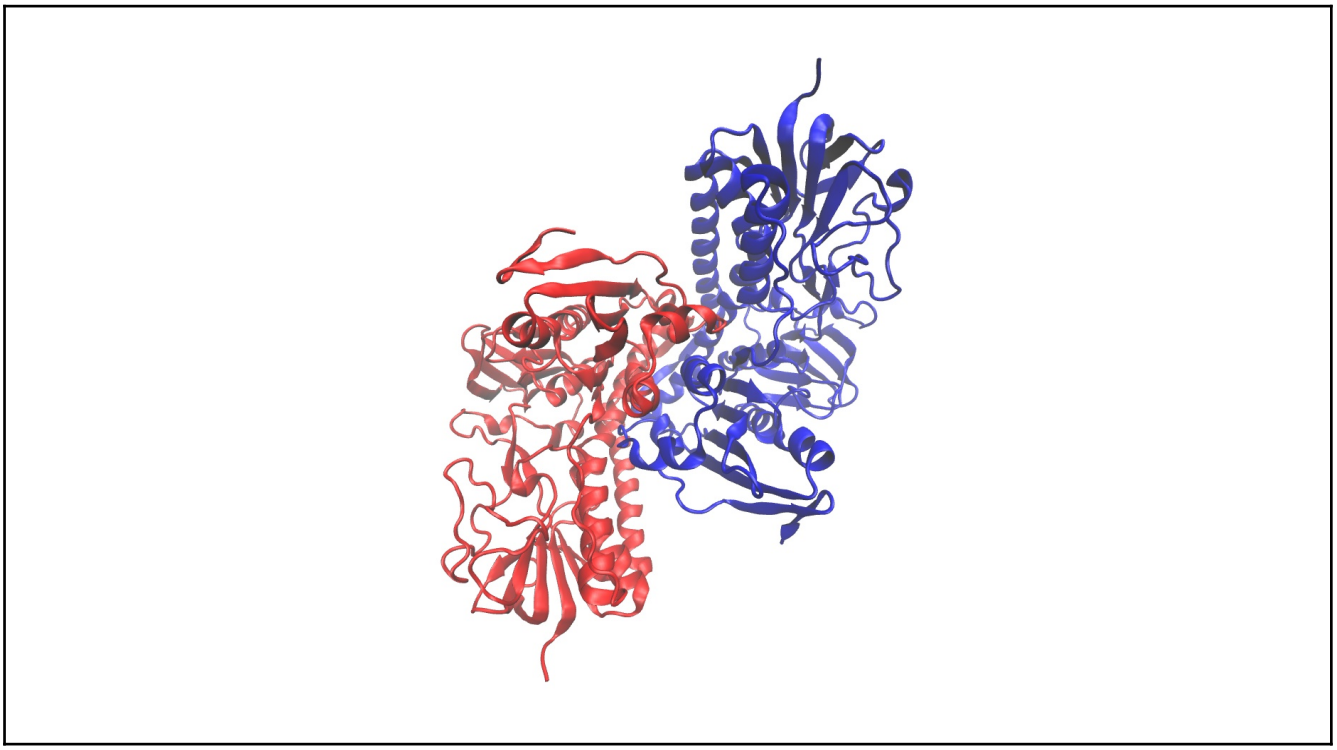
Hit 03 - 6ER5_A (trypanothione reductase) (13):
E-Value [7.59e-117]
Sequence ID [42.766%]
Annotations
{
PDB identifier [6ER5]
Method [X-RAY DIFFRACTION]
Resolution [3.37 Å]
Source Organism [*Leishmania infantum*]
}
**********END**********
=Hits for highest intersequence similarity protein (*Capsicum annuum*)=
**********BEGIN**********
Hit 01 - 3O0H_A (glutathione reductase) (11):
E-Value [5.61e-131]
Sequence ID [44.805%]
Annotations
{
PDB identifier [3O0H]
Method [X-RAY DIFFRACTION]
Resolution [1.90 Å]
Source Organism [*Bartonella henselae* str. Houston-1]
}

Hit 02 - 4DNA_A (putative glutathione reductase) (12):
E-Value [3.09e-124]
Sequence ID [44.516%]
Annotations
{
PDB identifier [4DNA]
Method [X-RAY DIFFRACTION]
Resolution [2.80 Å]
Source Organism [*Escherichia coli* BL21(DE3)]
}

Hit 03 - 5VDN_A (glutathione reductase) (14):
E-Value [3.17e-120]
Sequence ID [45.652%]
Annotations
{
PDB identifier [5VDN]
Method [X-RAY DIFFRACTION]
Resolution [1.55 Å]
Source Organism [*Yersinia pestis* KIM10+]
}
**********END**********

[Q09] ***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?***

Structure for 4DNA, Q08:Hit 02 for novel protein (12).



Generated using VMD (15).

Based on the sequence similarity, percent identity of 45.633%, this structure is likely to be similar to the novel protein. This further implies that the function is likely to be similar between the novel protein and that of this hit, glutathione reductase activity.

[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?***

Yes, hits were found for both recombinant trypanothione reductase inhibitors and glutathione reductase inhibitors (16, 17). Both of which would be helpful in exploring potential inhibition of the novel protein. Additionally, this could help differentiate between whether the novel protein within *Chlorokybus atmophyticus* is closer in

function to glutathione reductase or trypanothione reductase, which is of particular importance for trypanosomal infections (18).

Input:

Sequence: Novel protein sequence (Q03)
BLASTp Parameters: Default

Results:

Recombinant trypanothione reductase (CHEMBL1944501) –
Inhibitors of oxidoreductase activity (top 5 pCHEMBL shown)
{

CHEMBL ID	IC50	pCHEMBL
CHEMBL4169040	[IC50 = 820.0 nM]	(6.09)
CHEMBL4105040	[IC50 = 900.0 nM]	(6.05)
CHEMBL4205871	[IC50 = 1200.0 nM]	(5.92)
CHEMBL4076896	[IC50 = 1200.0 nM]	(5.92)
CHEMBL4172675	[IC50 = 1200.0 nM]	(5.92)

}
Glutathione reductase (CHEMBL2755) –
Inhibitors of glutathione activity (top 5 pCHEMBL shown)
{

CHEMBL ID	IC50	pCHEMBL
CHEMBL2068507	[IC50 = 1.0 nM]	(9.00)
CHEMBL120147	[IC50 = 4.1 nM]	(8.39)
CHEMBL1824793	[IC50 = 344.0 nM]	(6.46)
CHEMBL135536	[IC50 = 750.0 nM]	(6.12)
CHEMBL39225	[IC50 = 1000.0 nM]	(6.00)

}

References:

1. Protein [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2021 Dec 03]. Available from: https://www.ncbi.nlm.nih.gov/protein/

2. BLAST [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2021 Dec 03]. Available from: https://blast.ncbi.nlm.nih.gov/Blast.cgi

3. Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R., Potter, S.C., Finn, R.D. and Lopez, R., 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic acids research*, *47*(W1), pp.W636-W641.

4. Taxonomy [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2021 Dec 03]. Available from: https://www.ncbi.nlm.nih.gov/taxonomy/

5. Sela, I., Ashkenazy, H., Katoh, K. and Pupko, T., 2015. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic acids research*, *43*(W1), pp.W7-W14.

6. Paradis, E. and Schliep, K., 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), pp.526-528.

7. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

8. Grant, B.J., Rodrigues, A.P., ElSawy, K.M., McCammon, J.A. and Caves, L.S., 2006. Bio3d: an R package for the comparative analysis of protein structures. *Bioinformatics*, *22*(21), pp.2695-2696.

9. Raivo Kolde (2019). pheatmap: Pretty Heatmaps. R package version 1.0.12. https://CRAN.R-project.org/package=pheatmap

10. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E., 2000. The protein data bank. *Nucleic acids research*, *28*(1), pp.235-242.

11. Seattle Structural Genomics Center for Infectious Disease (SSGCID) (2010) Crystal structure of glutathione reductase from *Bartonella henselae* doi: 10.221/pdb3o0h/pdb

12. Malashkevich, V.N., Bhosle, R., Toro, R., Seidel, R., Almo, S.C., New York Structural Genomics Research Consortium (NYSGRC) (2012) Crystal structure of putative glutathione reductase from *Sinorhizobium meliloti* 1021 doi: 10.2210/pdb4dna/pdb

13. Ilari, A., Fiorillo, A. (2018) X-ray structure of Trypanothione Reductase from *Leishmania infantum* in complex with 2-(diethylamino)ethyl 4-((3-(4-nitrophenyl)-3-oxopropyl)amino)benzoate doi: 10.2210/pdb6er5/pdb

14. Minasov, G., Shuvalova, L., Dubrovska, I., Cardona-Correa, A., Grimshaw, S., Kwon, K., Anderson, W.F., Center for Structural Genomics of Infectious Diseases (CSGID)(2017) 1.55 Angstrom Resolution Crystal Structure of Glutathione Reductase from *Yersinia pestis* in Complex with FAD doi: 10.2210/pdb5vdn/pdb

15. Humphrey, W., Dalke, A. and Schulten, K., 1996. VMD: visual molecular dynamics. *Journal of molecular graphics*, *14*(1), pp.33-38.

16. Mendez, D., Gaulton, A., Bento, A.P., Chambers, J., De Veij, M., Félix, E., Magariños, M.P., Mosquera, J.F., Mutowo, P., Nowotka, M. and Gordillo-Marañón, M., 2019. ChEMBL: towards direct deposition of bioassay data. *Nucleic acids research*, *47*(D1), pp.D930-D940.

17. Davies, M., Nowotka, M., Papadatos, G., Dedman, N., Gaulton, A., Atkinson, F., Bellis, L. and Overington, J.P., 2015. ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic acids research*, *43*(W1), pp.W612-W620.

18. Krauth-Siegel, R.L. and Inhoff, O., 2003. Parasite-specific trypanothione reductase as a drug target molecule. *Parasitology research*, *90*(2), pp.S77-S85.

APPENDIX

I:

```
library(ape)

# gor-MAFFT-20-NJ_tree.tre available upon request
NJTree <- read.tree(file = "gor-MAFFT-20-NJ_tree.tre")
col_vect <- rep("black", length(NJTree$tip.label))
col_vect[grep("Chlorokybus", NJTree$tip.label)] <- "orange"
col_vect[grep("thaliana", NJTree$tip.label)] <- "slateblue2"

plot(NJTree,
     edge.width = 1,
     label.offset = 0.02,
     no.margin = TRUE,
     underscore = FALSE,
     lab4ut = "axial",
     align.tip.label = TRUE,
     type = "u",
     cex = 1,
     lwd = 2,
     rotate.tree = -40,
     font = 3,
     tip.color = col_vect)
add.scale.bar(x = 0.5, y = -0.1, lwd = 1)
```

II:

```
library(bio3d)

# gor-MAFFT_aln-20_short.faa available upon request
aln <- read.fasta("gor-MAFFT_aln-20_short.faa")
seq_id <- seqidentity(aln)
```

III:

```
library(pheatmap)

IBMColorBlindSafe = c("#785ef0", "#648fff", "#ffffff", "#ffb000",
"#fe6100")
colBlindScale <- colorRampPalette(IBMColorBlindSafe)
hmColors <- colBlindScale(512)

pheatmap(seq_id,
         color = hmColors,
         display_numbers = TRUE,
         fontsize_number = 10,
         cutree_rows = 8,
         treeheight_col = 0)
```

IV:

```
library(bio3d)

# chlorokybus_gor_seq.faa available upon request
hits <- blast.pdb(read.fasta("chlorokybus_gor_seq.faa"))
blast_hits <- plot.blast(hits)

pdb <- get.pdb(blast_hits$pdb.id, path="pdbs", split=TRUE, gzip=TRUE)

pdbs <- pdbaln(pdb[1:3], fit = TRUE)
top_3 <- hits$hit.tbl[1:3,]
```