Joel Kosareff

Find a Gene Project Assignment

jkosaref@ucsd.edu

A15866978

[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: p53

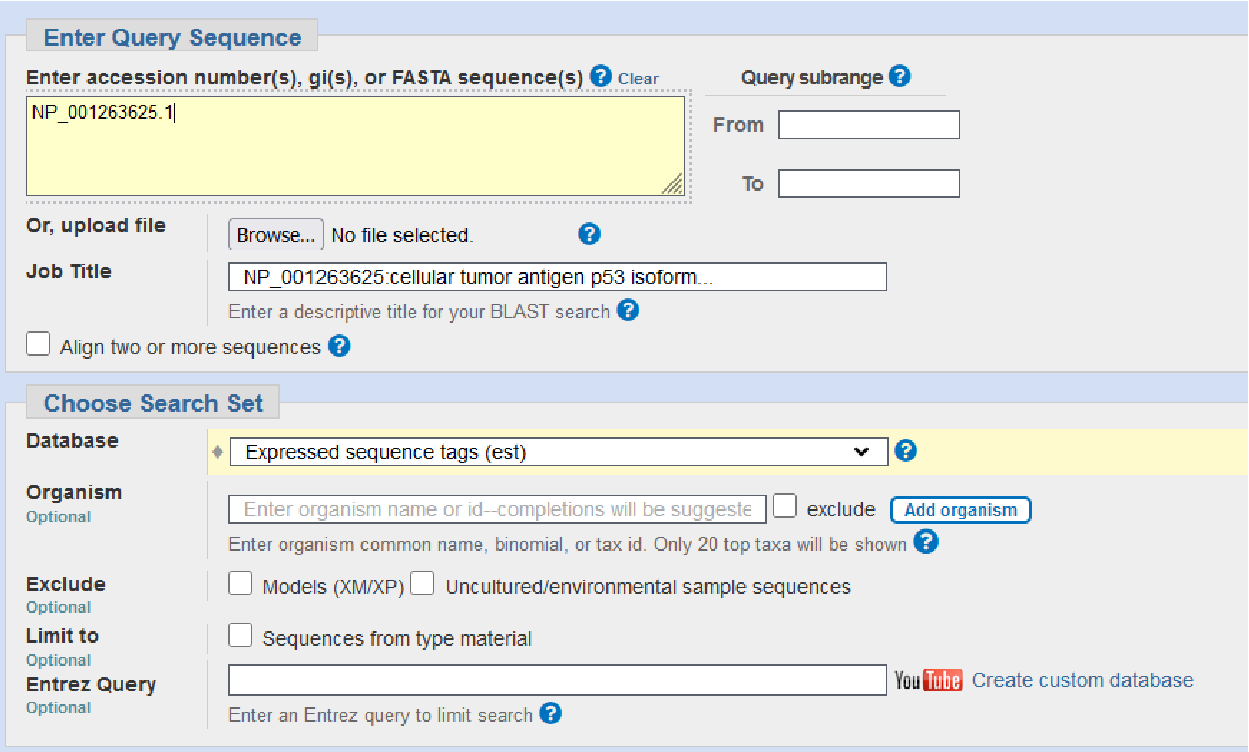
Accession: NP_001263625.1

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

Database: Expressed Sequence Tags (est)



Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s): NP_001263625.1

Query subrange: From □ To □

Or, upload file: No file selected.

Job Title: NP_001263625:cellular tumor antigen p53 isoform...

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 suggeste □ exclude

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

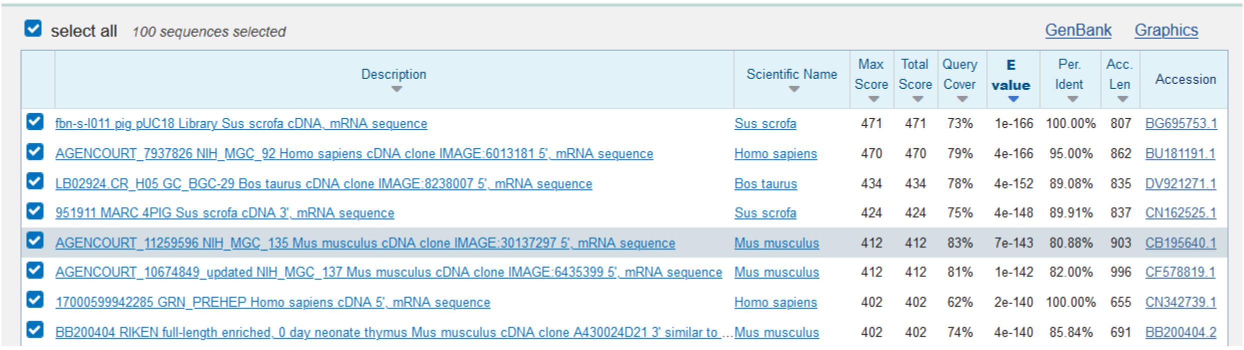
Exclude Optional: □ Models (XM/XP) □ Uncultured/environmental sample sequences

Limit to Optional: □ Sequences from type material

Entrez Query Optional:

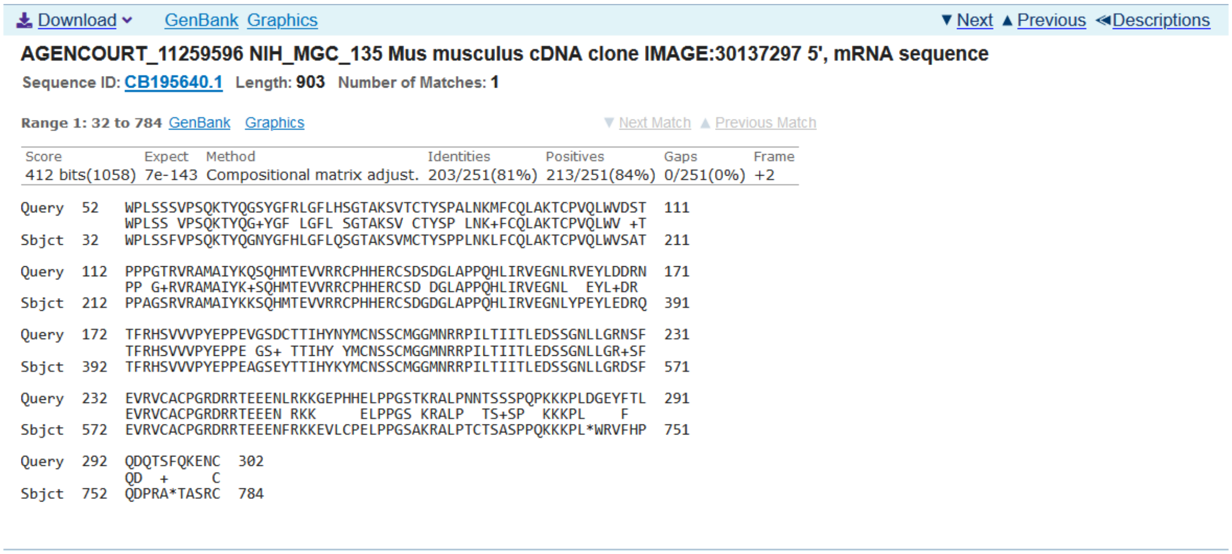
Enter an Entrez query to limit search

Match: Accession Number: CB195640.1



☑ select all *100 sequences selected*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	fbn-s-l011 pig pUC18 Library Sus scrofa cDNA, mRNA sequence	Sus scrofa	471	471	73%	1e-166	100.00%	807	BG695753.1
☑	AGENCOURT_7937826 NIH_MGC_92 Homo sapiens cDNA clone IMAGE:6013181 5', mRNA sequence	Homo sapiens	470	470	79%	4e-166	95.00%	862	BU181191.1
☑	LB02924.CR_H05 GC_BGC-29 Bos taurus cDNA clone IMAGE:8238007 5', mRNA sequence	Bos taurus	434	434	78%	4e-152	89.08%	835	DV921271.1
☑	951911 MARC 4PIG Sus scrofa cDNA 3', mRNA sequence	Sus scrofa	424	424	75%	4e-148	89.91%	837	CN162525.1
☑	AGENCOURT_11259596 NIH_MGC_135 Mus musculus cDNA clone IMAGE:30137297 5', mRNA sequence	Mus musculus	412	412	83%	7e-143	80.88%	903	CB195640.1
☑	AGENCOURT_10674849_updated NIH_MGC_137 Mus musculus cDNA clone IMAGE:6435399 5', mRNA sequence	Mus musculus	412	412	81%	1e-142	82.00%	996	CF578819.1
☑	17000599942285 GRN_PREHEP Homo sapiens cDNA 5', mRNA sequence	Homo sapiens	402	402	62%	2e-140	100.00%	655	CN342739.1
☑	BB200404 RIKEN full-length enriched, 0 day neonate thymus Mus musculus cDNA clone A430024D21 3' similar to ...	Mus musculus	402	402	74%	4e-140	85.84%	691	BB200404.2



AGENCOURT_11259596 NIH_MGC_135 Mus musculus cDNA clone IMAGE:30137297 5', mRNA sequence

Sequence ID: CB195640.1 Length: 903 Number of Matches: 1

Range 1: 32 to 784 GenBank Graphics ▼ Next Match ▲ Previous Match

```
Score           Expect   Method                        Identities     Positives      Gaps       Frame
412 bits(1058)  7e-143  Compositional matrix adjust.  203/251(81%)  213/251(84%)  0/251(0%)  +2

Query  52   WPLSSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKTCPVQLWVDST  111
            WPLSS VPSQKTYQG+YGF LGFL SGTAKSV CTYSP LNK+FCQLAKTCPVQLWV +T
Sbjct  32   WPLSSFVPSQKTYQGNYGFHLGFLQSGTAKSVMCTYSPPLNKLFCQLAKTCPVQLWVSAT  211

Query  112  PPPGTRVRAMAIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQHLIRVEGNLRVEYLDDRN  171
            PP G+RVRAMAIYK+SQHMTEVVRRCPHHERCSD DGLAPPQHLIRVEGNL  EYL+DR
Sbjct  212  PPAGSRVRAMAIYKKSQHMTEVVRRCPHHERCSDGDGLAPPQHLIRVEGNLYPEYLEDRQ  391

Query  172  TFRHSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRNSF  231
            TFRHSVVVPYEPPE GS+ TTIHY YMCNSSCMGGMNRRPILTIITLEDSSGNLLGR+SF
Sbjct  392  TFRHSVVVPYEPPEAGSEYTTIHYKYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRDSF  571

Query  232  EVRVCACPGRDRRTEEENLRKKGEPHHELPPGSTKRALPNNTSSSPQPKKKPLDGEYFTL  291
            EVRVCACPGRDRRTEEEN RKK      ELPPGS KRALP  TS+SP  KKKPL     F
Sbjct  572  EVRVCACPGRDRRTEEENFRKKEVLCPELPPGSAKRALPTCTSASPPQKKKPL*WRVFHP  751

Query  292  QDQTSFQKENC  302
            QD  +     C
Sbjct  752  QDPRA*TASRC  784
```

[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

Sequence from EMBOSS:

```
>CB195640.1_1 AGENCOURT_11259596 NIH_MGC_135 Mus musculus cDNA clone
IMAGE:30137297 5', mRNA sequence
WASGPCPSHSMAPVIFCPFSKNLPGQLWLPPGLPAVWDSQVCYVHVLSSPQ*AILPAGED
VPCAVVGQRHTSSWEPCPRHGHLQEVTAHDGGRETLPPP*ALLRW*WPGSSPASYPGGRK
FVSRVSGRQADFSPQRGGTL*ATRGRL*VYHHPLQVHV**LLHGGHEPPTYPYHHHTGRL
QWEPSGTGQL*GSCLCLPWERPPYRRRKFPQKGSPLP*TAPRERKESAAHLHKRLSPAKE
KTTLMESISPSRSAGVNGFEMFRGAEMRPLELKGMPMLQKESGRQQGASLQLTWKTKEKG
Q
```

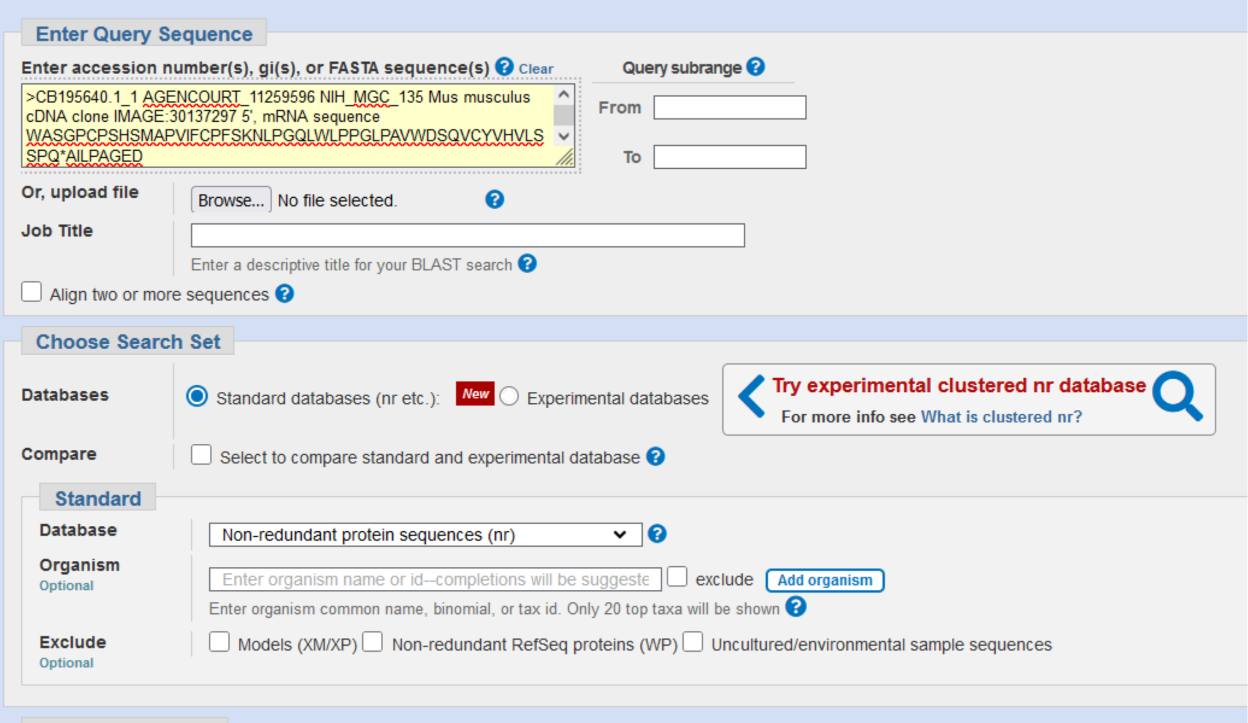
Name: cellular tumor antigen p53 isoform a [Mus musculus]

Species: Mus Musculus

```
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
Myomorpha;
            Muroidea; Muridae; Murinae; Mus; Mus.
```

[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

A BlastP search against the NR database provided a result from mus musculus with 17% query cover and 84.91% identity



Enter Query Sequence

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

```
>CB195640.1_1 AGENCOURT_11259596 NIH_MGC_135 Mus musculus
cDNA clone IMAGE:30137297 5', mRNA sequence
WASGPCPSHSMAPVIFCPFSKNLPGQLWLPPGLPAVWDSQVCYVHVLS
SPQ*AILPAGED
```

Query subrange

From:

To:

Or, upload file: Browse... No file selected.

Job Title:

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

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

Try experimental clustered nr database
For more info see What is clustered nr?

Compare: □ Select to compare standard and experimental database

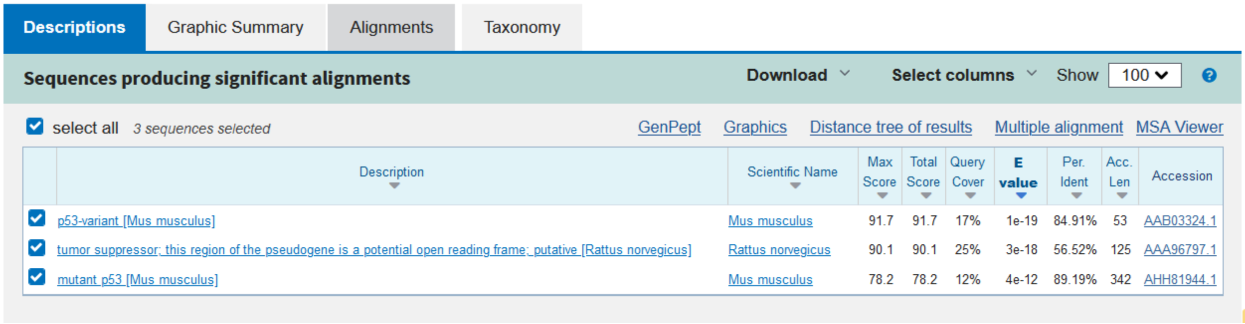
Standard

Database: Non-redundant protein sequences (nr)

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

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

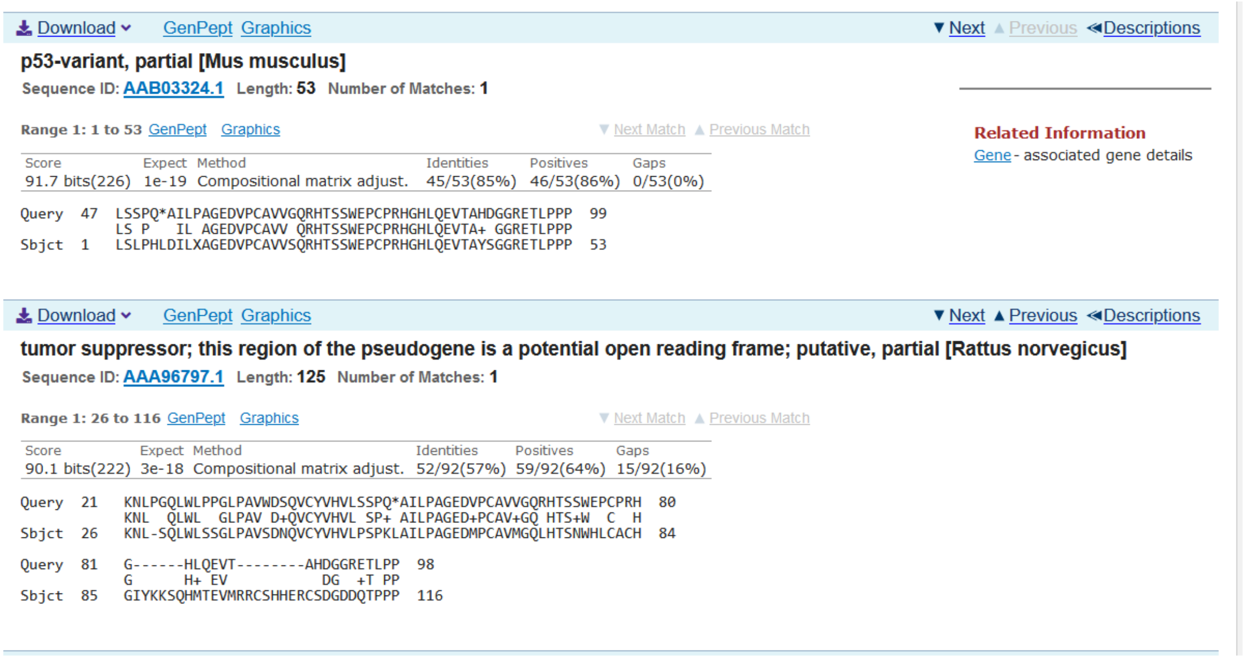
Exclude (Optional): □ Models (XM/XP) □ Non-redundant RefSeq proteins (WP) □ Uncultured/environmental sample sequences



Sequences producing significant alignments

☑ select all 3 sequences selected

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	p53-variant [Mus musculus]	Mus musculus	91.7	91.7	17%	1e-19	84.91%	53	AAB03324.1
☑	tumor suppressor; this region of the pseudogene is a potential open reading frame; putative [Rattus norvegicus]	Rattus norvegicus	90.1	90.1	25%	3e-18	56.52%	125	AAA96797.1
☑	mutant p53 [Mus musculus]	Mus musculus	78.2	78.2	12%	4e-12	89.19%	342	AHH81944.1



p53-variant, partial [Mus musculus]

Sequence ID: AAB03324.1 Length: 53 Number of Matches: 1

Related Information

Gene - associated gene details

Range 1: 1 to 53 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
91.7 bits(226)	1e-19	Compositional matrix adjust.	45/53(85%)	46/53(86%)	0/53(0%)

```
Query  47   LSSPQ*AILPAGEDVPCAVVGQRHTSSWEPCPRHGHLQEVTAHDGGRETLPPP  99
            LS P   IL AGEDVPCAVV QRHTSSWEPCPRHGHLQEVTA+ GGRETLPPP
Sbjct  1    LSLPHLDILXAGEDVPCAVVSQRHTSSWEPCPRHGHLQEVTAYSGGRETLPPP  53
```

tumor suppressor; this region of the pseudogene is a potential open reading frame; putative, partial [Rattus norvegicus]

Sequence ID: AAA96797.1 Length: 125 Number of Matches: 1

Range 1: 26 to 116 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
90.1 bits(222)	3e-18	Compositional matrix adjust.	52/92(57%)	59/92(64%)	15/92(16%)

```
Query  21   KNLPGQLWLPPGLPAVWDSQVCYVHVLSSPQ*AILPAGEDVPCAVVGQRHTSSWEPCPRH  80
            KNL  QLWL  GLPAV D+QVCYVHVL SP+ AILPAGED+PCAV+GQ HTS+W  C  H
Sbjct  26   KNL-SQLWLSSGLPAVSDNQVCYVHVLPSPKLAILPAGEDMPCAVMGQLHTSNWHLCACH  84

Query  81   G------HLQEVT--------AHDGGRETLPP  98
            G      H+ EV           DG  +T PP
Sbjct  85   GIYKKSQHMTEVMRRCSHHERCSDGDDQTPPP  116
```

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.

Labeled Sequences for Alignment

```
>Human P53 NP_001263625.1 cellular tumor antigen p53 isoform i [Homo sapiens]
MDDLMLSPDDIEQWFTEDPGPDEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGF
RLGFLHSGTAKSVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHH
ERCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRR
PILTIITLEDSSGNLLGRNSFEVRVCACPGRDRRTEEENLRKKGEPHHELPPGSTKRALPNNTSSSPQPK
KKPLDGEYFTLQDQTSFQKENC


>House Mouse P53 Sequence from blast result CB195640.1_1 AGENCOURT_11259596
NIH_MGC_135 Mus musculus cDNA clone IMAGE:30137297 5', mRNA sequence
WASGPCPSHSMAPVIFCPFSKNLPGQLWLPPGLPAVWDSQVCYVHVLSSPQAILPAGED
VPCAVVGQRHTSSWEPCPRHGHLQEVTAHDGGRETLPPPALLRWWPGSSPASYPGGRK
FVSRVSGRQADFSPQRGGTLATRGRLVYHHPLQVHVLLHGGHEPPTYPYHHHTGRL
QWEPSGTGQLGSCLCLPWERPPYRRRKFPQKGSPLPTAPRERKESAAHLHKRLSPAKE
KTTLMESISPSRSAGVNGFEMFRGAEMRPLELKGMPMLQKESGRQQGASLQLTWKTKEKG
Q

>Western Lowland Gorilla P53 XP_018868682.2 cellular tumor antigen p53
isoform X2 [Gorilla gorilla gorilla]
MGSSQTAFRVTAMEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDP
GPDEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYSP
```

```
ALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQHLIRVE
GNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRN
SFEVRVCACPGRDRRTEEENLRKKGEPHHELPPGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERF
EMFRELNEALELKDAQAGKEPGGSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD

>Chimpanzee P53 XP_016786959.2 cellular tumor antigen p53 isoform X1 [Pan
troglodytes]
METVSGSIGKAGPPPPHPNPSPLVETCGKRKFHGTDFLLLSFRLPENNVLSPLPSQAMDDLMLSPDDIEQ
WFTEDPGPDEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAKSV
TCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQ
HLIRVEGNLRVEYLDDRNTFRHSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSG
NLLGRNSFEVRVCACPGRDRRTEEENLRKKGEPHHELPPGSTKRALPNNTSSSPQPKKKPLDGEYFTLQI
RGRERFEMFRELNEALELKDAQAGKEPGGSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD

>Bonobo P53 XP_003810114.2 cellular tumor antigen p53 [Pan paniscus]
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPDEAPRMPEAA
PRVAPAPAAPTPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKT
CPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQHLIRVEGNLRVEYLDDRN
TFRHSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGR
DRRTEEENLRKKGEPHHELPPGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALEL
KDAQAGKEPGGSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD

>Northern White-Cheeked Gibbon P53 XP_030656345.1 cellular tumor antigen p53
[Nomascus leucogenys]
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPEDIAQWFTEDPGPHEAPRMSEAA
PPMAPAPAAPTLAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKT
CPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQHLIRVEGNLRVEYLDDRN
TFRHSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGR
DRRTEEENFHKKGEPHHELPPGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALEL
KDAQAGKEPGGSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD

>Sumatran Orangutan P53 XP_002827020.2 cellular tumor antigen p53 [Pongo
abelii]
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAVDDLLLSPDDIAQWFIEDPGPDEAPRMSEAA
SPVDPAPAAPIPAAPAPAPSWPLSSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKT
CPVQLWVDSTPPPGTRVRAMAIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQHLIRVEGNLRVEYLDDRN
TFRHSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRVCACPGR
DRRTEEENFRKKGEPHHELPPGSTKRALPNNTSSSPQPKKKPLDGEYFTLQIRGRERFEMFRELNEALEL
KDAQAGKEPGGSRAHSSHLKSKKGQSTSRHKKLMFKTEGPDSD
```

Results obtained from MUSCLE

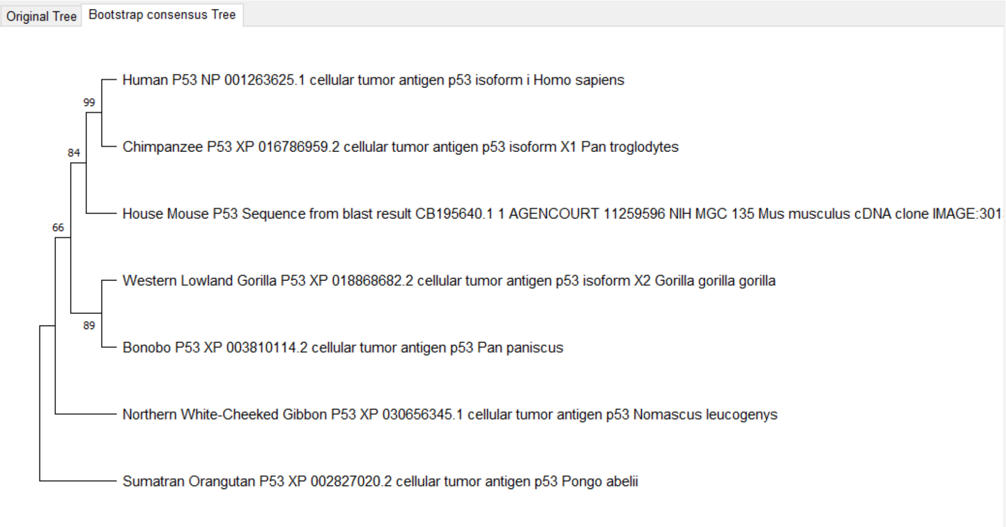
CLUSTAL multiple sequence alignment by MUSCLE (3.8)

```
House           ----------------------------------------W-------ASGPCPSHSMAP
Sumatran        ------------MEEPQSDPSV------EPPLSQETFSDLWKLLPENNVLSPLPSQAVDD
Western         MGSSQTAFRVTAMEEPQSDPSV------EPPLSQETFSDLWKLLPENNVLSPLPSQAMDD
Bonobo          ------------MEEPQSDPSV------EPPLSQETFSDLWKLLPENNVLSPLPSQAMDD
Northern        ------------MEEPQSDPSV------EPPLSQETFSDLWKLLPENNVLSPLPSQAMDD
Chimpanzee      METVSGSIGKAGPPPPHPNPSPLVETCGKRKFHGTDFLLLSFRLPENNVLSPLPSQAMDD
Human           ---------------------------------------------------------MDD
                                                                         :

House           VIFCP------FSKNLPGQ---LWLPPGLPAVWDSQVCYVHVLSSPQAILPAGEDVPCAV
Sumatran        LLLSPDDIAQWFIED-PGPDEAPRMSEAASPVDPAPAAPIPAAPAPAPSWPLSSSVPSQK
```

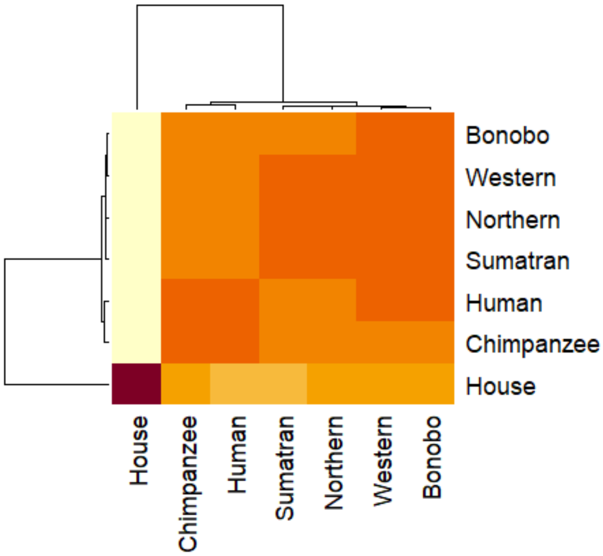
```
Western         LMLSPDDIEQWFTED-PGPDEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQK
Bonobo          LMLSPDDIEQWFTED-PGPDEAPRMPEAAPRVAPAPAAPTPAAPAPAPSWPLSSSVPSQK
Northern        LMLSPEDIAQWFTED-PGPHEAPRMSEAAPPMAPAPAAPTLAAPAPAPSWPLSSSVPSQK
Chimpanzee      LMLSPDDIEQWFTED-PGPDEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQK
Human           LMLSPDDIEQWFTED-PGPDEAPRMPEAAPPVAPAPAAPTPAAPAPAPSWPLSSSVPSQK
                :::.*      * :: **     .:. . . :  : ..   . .:* .  * ...**.

House           ---------VGQRHTSSWEP--CPRHGHLQEVTAHDGGRETLPPPALLRWWPGSSPASYP
Sumatran        TYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQ------LAKTCPVQLWVDSTPP--P
Western         TYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQ------LAKTCPVQLWVDSTPP--P
Bonobo          TYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQ------LAKTCPVQLWVDSTPP--P
Northern        TYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQ------LAKTCPVQLWVDSTPP--P
Chimpanzee      TYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQ------LAKTCPVQLWVDSTPP--P
Human           TYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQ------LAKTCPVQLWVDSTPP--P
                         :*  *:.: :.  *.    *::: .:      *. .. :. * .*:*.  *

House           GGRKFVSRVSGRQADFS------PQRGGTLATRGRLVYHHPLQVHVLLHGGHEPPTYPYH
Sumatran        GTRVRAMAIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFR
Western         GTRVRAMAIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFR
Bonobo          GTRVRAMAIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFR
Northern        GTRVRAMAIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFR
Chimpanzee      GTRVRAMAIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFR
Human           GTRVRAMAIYKQSQHMTEVVRRCPHHERCSDSDGLAPPQHLIRVEGNLRVEYLDDRNTFR
                * *  .  :  ..  ::      *:.     : *    :* :.*   *.  :     .:.

House           HHTGRLQWEPSGTGQ-----------LGSCLCLPWERP----------------------
Sumatran        HSVVV-PYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRNSFEV
Western         HSVVV-PYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRNSFEV
Bonobo          HSVVV-PYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRNSFEV
Northern        HSVVV-PYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRNSFEV
Chimpanzee      HSVVV-PYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRNSFEV
Human           HSVVV-PYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRNSFEV
                * .    :**. .*.            .**:   . **

House           -----PYRRRK-----FPQKGSPLPTAPRERKESAAHLHKRLSPAKEKTTLMESISPSRS
Sumatran        RVCACPGRDRRTEEENFRKKGEPHHELPPGSTKRALPNNTSSSPQPKKKPLDGEYFTLQI
Western         RVCACPGRDRRTEEENLRKKGEPHHELPPGSTKRALPNNTSSSPQPKKKPLDGEYFTLQI
Bonobo          RVCACPGRDRRTEEENLRKKGEPHHELPPGSTKRALPNNTSSSPQPKKKPLDGEYFTLQI
Northern        RVCACPGRDRRTEEENFHKKGEPHHELPPGSTKRALPNNTSSSPQPKKKPLDGEYFTLQI
Chimpanzee      RVCACPGRDRRTEEENLRKKGEPHHELPPGSTKRALPNNTSSSPQPKKKPLDGEYFTLQI
Human           RVCACPGRDRRTEEENLRKKGEPHHELPPGSTKRALPNNTSSSPQPKKKPLDGEYFTLQ-
                     * * *.     : :**.*    *   .: *   :.  **  :*..*  .  . .

House           AGVNGFEMFRGAEMRPLELKGMPM-------------LQKESGRQQGASLQLTWKTKEKG
Sumatran        RGRERFEMFRELN-EALELKDAQAGKEPGGSRAHSSHLKSKKGQSTSRHKKLMFKTEGPD
Western         RGRERFEMFRELN-EALELKDAQAGKEPGGSRAHSSHLKSKKGQSTSRHKKLMFKTEGPD
Bonobo          RGRERFEMFRELN-EALELKDAQAGKEPGGSRAHSSHLKSKKGQSTSRHKKLMFKTEGPD
Northern        RGRERFEMFRELN-EALELKDAQAGKEPGGSRAHSSHLKSKKGQSTSRHKKLMFKTEGPD
Chimpanzee      RGRERFEMFRELN-EALELKDAQAGKEPGGSRAHSSHLKSKKGQSTSRHKKLMFKTEGPD
Human           --------------------------------------------------DQTSFQKENC-
                                                                  :  ::.:

House           Q-
Sumatran        SD
Western         SD
Bonobo          SD
Northern        SD
```

```
Chimpanzee       SD
Human            --
```

[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.


Original Tree
Bootstrap consensus Tree
Human P53 NP 001263625.1 cellular tumor antigen p53 isoform i Homo sapiens
99
Chimpanzee P53 XP 016786959.2 cellular tumor antigen p53 isoform X1 Pan troglodytes
84
House Mouse P53 Sequence from blast result CB195640.1 1 AGENCOURT 11259596 NIH MGC 135 Mus musculus cDNA clone IMAGE:301
66
Western Lowland Gorilla P53 XP 018868682.2 cellular tumor antigen p53 isoform X2 Gorilla gorilla gorilla
89
Bonobo P53 XP 003810114.2 cellular tumor antigen p53 Pan paniscus
Northern White-Cheeked Gibbon P53 XP 030656345.1 cellular tumor antigen p53 Nomascus leucogenys
Sumatran Orangutan P53 XP 002827020.2 cellular tumor antigen p53 Pongo abelii

[Q7] Generate a sequence identity based heatmap of your aligned sequences using R. If necessary convert your sequence alignment to the ubiquitous FASTA format (Seaview can read in clustal format and “Save as” FASTA format for example). Read this FASTA format alignment into R with the help of functions in the Bio3D package. Calculate a sequence identity matrix (again using a function within the Bio3D package). Then generate a heatmap plot and add to your report. Do make sure your labels are visible and not cut at the figure margins.

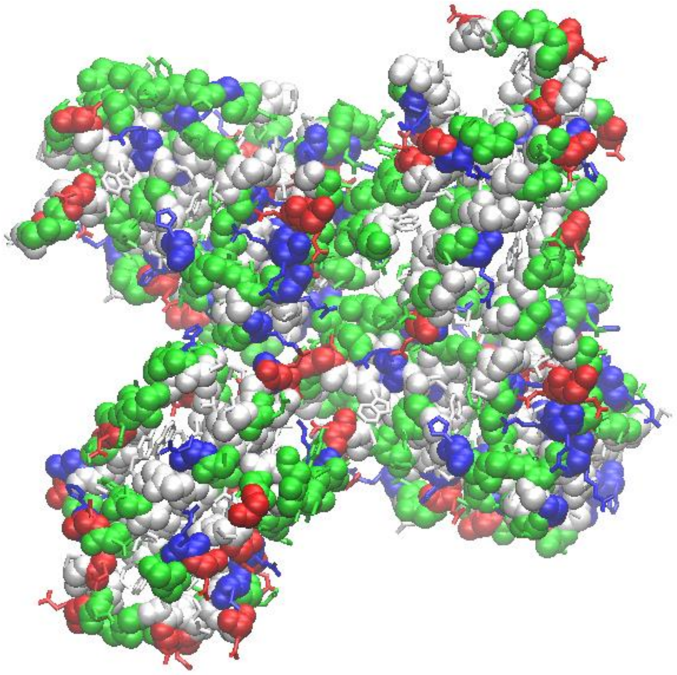


[Q8] Using R/Bio3D (or an online blast server if you prefer), search the main protein structure database for the most similar atomic resolution structures to your aligned sequences.
List the top 3 unique hits (i.e. not hits representing different chains from the same structure) along with their Evalue and sequence identity to your query. Please also add annotation details of these structures. For example include the annotation terms PDB identifier (structureId), Method used to solve the structure (experimentalTechnique), resolution (resolution), and source organism (source).

ID	Technique	Resolution	Source	e-value	Identity
6XRE_M	EM	4.6	Homo Sapiens	0.00e00	98.663
6LHD_A	X-ray	2.499	Homo Sapiens	1.70e-147	84.937
2XWC_A	X-ray	1.82	Homo Sapiens	6.19e-80	57.143

[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?

As the percent identity is almost 85% this structure is likely to be similar to our novel protein in some form.



[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?

ChEMBEL displays one inhibition assay and no others. There is no ligand efficiency data.

https://www.ebi.ac.uk/chembl/g/#search_results/all/query=%3ECB195640.1_1%20AGENCOURT_11259596%20NIH_MGC_135%20Mus%20musculus%20cDNA%20clone%20IMAGE%3A30137297%205%26%23x27%3B%2C%20mRNA%20sequence%20WASGPCPSHSMAPVIFCPFSKNLPGQLWLPPGLPAVWDSQVCYVHVLSSPQ*AILPAGED%20VPCAVVGQRHTSSWEPCPRHGHLQEVTAHDGGRETLPPP*ALLRW*WPGSSPASYPGGRK%20FVSRVSGRQADFSPQRGGTL*ATRGRL*VYHHPLQVHV**LLHGGHEPPTYPYHHHTGRL%20QWEPSGTGQL*GSCLCLPWERPPYRRRKFPQKGSPLP*TAPRERKESAAHLHKRLSPAKE%20KTTLMESISPSRSAGVNGFEMFRGAEMRPLELKGMPMLQKESGRQQGASLQLTWKTKEKG%20Q

The inhibition assay utilized CHO-K1 cells expressing human Kv1.5 sequences in order to test the expression of Kv1.5 across mediums.

Note: ChEMBL does not provide a reference journal for this assay.