Kelly Fogelson
Find a Gene Assignment Part 1

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.

Name: ileal sodium/bile acid cotransporter

Accession #: NP_000443.2

Species: *Homo sapiens*

https://www.ncbi.nlm.nih.gov/protein/NP_000443.2

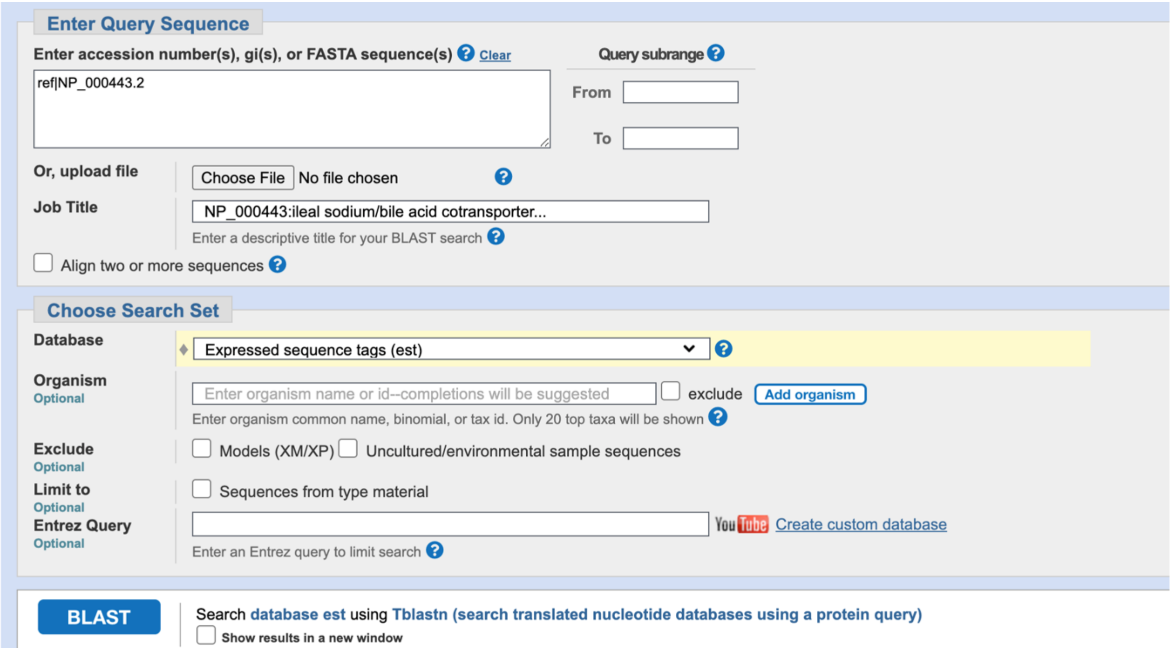
[**Q2**] Perform a BLAST search against a DNA database, such as a database consisting of genomic DNA or ESTs. The BLAST server can be at NCBI or elsewhere. Include details of the BLAST method used, database searched and any limits applied (e.g. Organism).

Method: tBLASTN

Database: Expressed Sequence Tags (ests)

Organism: no restricted to particular organism.

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. --



Enter Query Sequence

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

ref|NP_000443.2

Query subrange

From

To

Or, upload file: Choose File No file chosen

Job Title: NP_000443:ileal sodium/bile acid cotransporter...

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

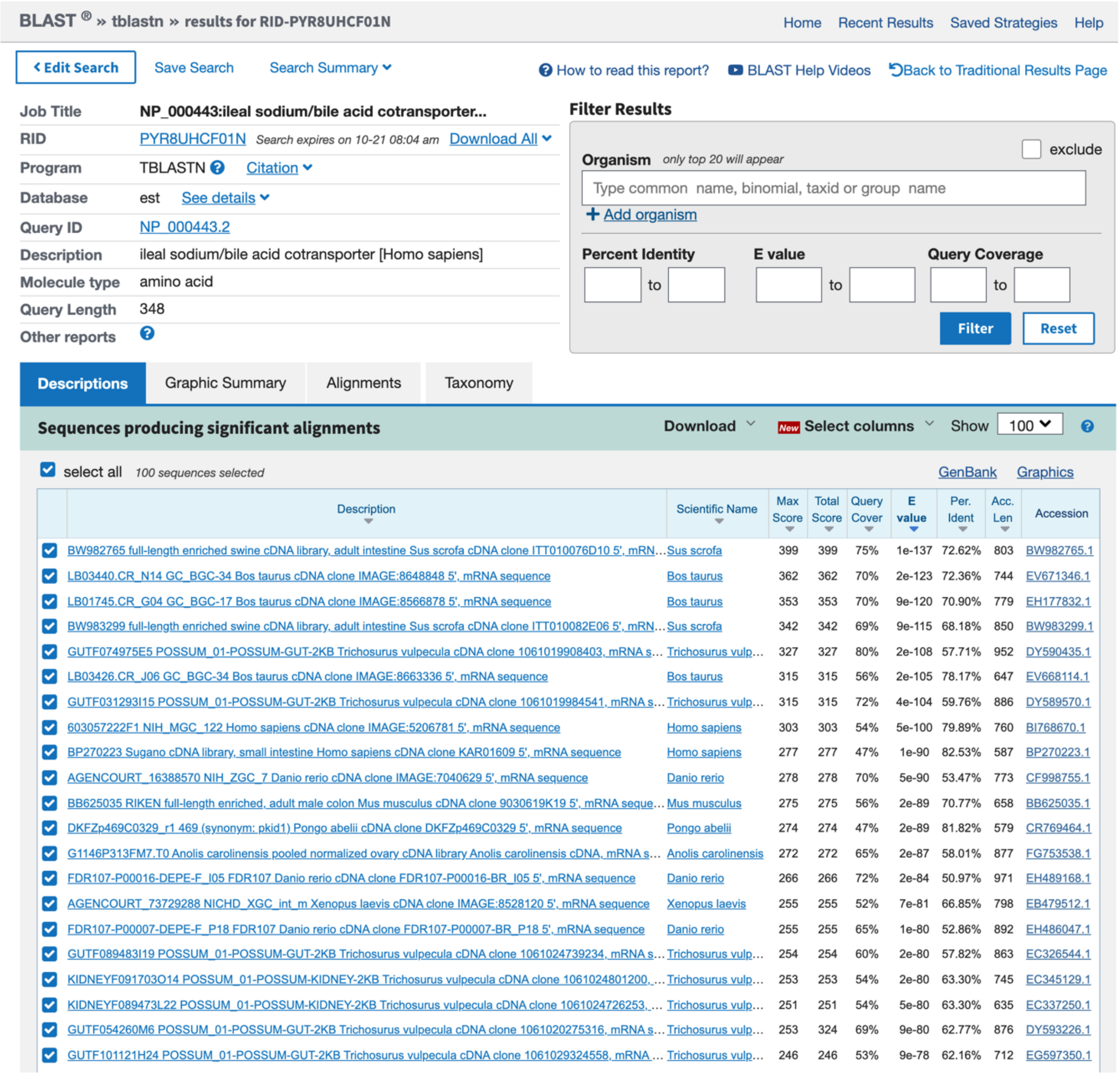
Entrez Query (Optional): Create custom database

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



BLAST ® » tblastn » results for RID-PYR8UHCF01N

Home Recent Results Saved Strategies Help

< Edit Search Save Search Search Summary

How to read this report? BLAST Help Videos Back to Traditional Results Page

Job Title	NP_000443:ileal sodium/bile acid cotransporter...
RID	PYR8UHCF01N *Search expires on 10-21 08:04 am* Download All
Program	TBLASTN Citation
Database	est See details
Query ID	NP_000443.2
Description	ileal sodium/bile acid cotransporter [Homo sapiens]
Molecule type	amino acid
Query Length	348
Other reports	

Filter Results

Organism *only top 20 will appear* □ exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity ___ to ___ E value ___ to ___ Query Coverage ___ to ___

Filter Reset

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments

Download New Select columns Show 100

☑ select all *100 sequences selected*

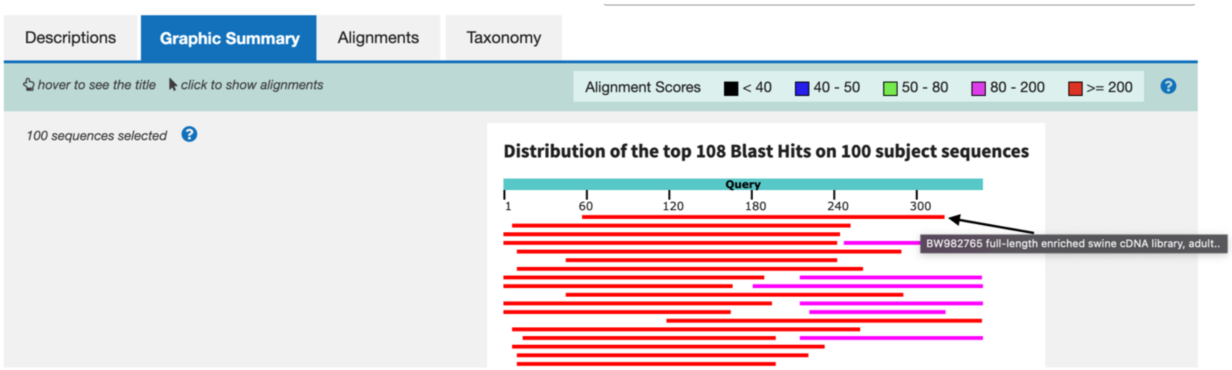
GenBank Graphics

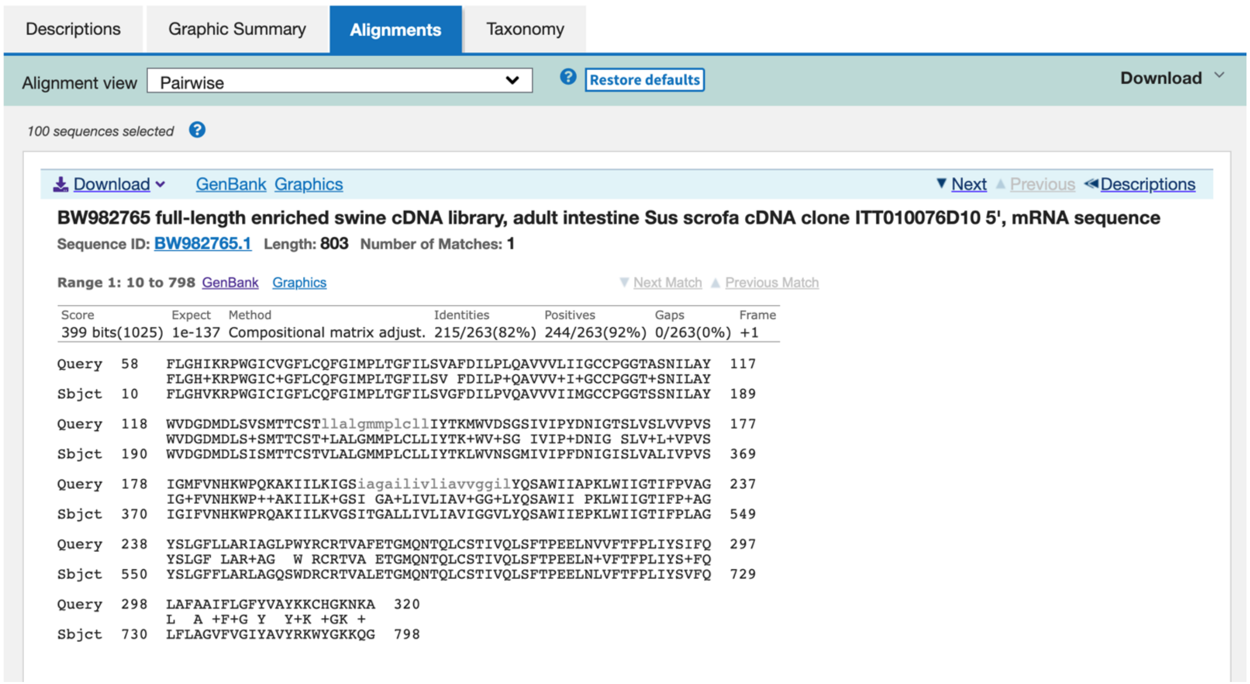
	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	BW982765 full-length enriched swine cDNA library, adult intestine Sus scrofa cDNA clone ITT010076D10 5', mRN…	Sus scrofa	399	399	75%	1e-137	72.62%	803	BW982765.1
☑	LB03440.CR_N14 GC_BGC-34 Bos taurus cDNA clone IMAGE:8648848 5', mRNA sequence	Bos taurus	362	362	70%	2e-123	72.36%	744	EV671346.1
☑	LB01745.CR_G04 GC_BGC-17 Bos taurus cDNA clone IMAGE:8566878 5', mRNA sequence	Bos taurus	353	353	70%	9e-120	70.90%	779	EH177832.1
☑	BW983299 full-length enriched swine cDNA library, adult intestine Sus scrofa cDNA clone ITT010082E06 5', mRN…	Sus scrofa	342	342	69%	9e-115	68.18%	850	BW983299.1
☑	GUTF074975E5 POSSUM_01-POSSUM-GUT-2KB Trichosurus vulpecula cDNA clone 1061019908403, mRNA s…	Trichosurus vulp…	327	327	80%	2e-108	57.71%	952	DY590435.1
☑	LB03426.CR_J06 GC_BGC-34 Bos taurus cDNA clone IMAGE:8663336 5', mRNA sequence	Bos taurus	315	315	56%	2e-105	78.17%	647	EV668114.1
☑	GUTF031293I15 POSSUM_01-POSSUM-GUT-2KB Trichosurus vulpecula cDNA clone 1061019984541, mRNA s…	Trichosurus vulp…	315	315	72%	4e-104	59.76%	886	DY589570.1
☑	603057222F1 NIH_MGC_122 Homo sapiens cDNA clone IMAGE:5206781 5', mRNA sequence	Homo sapiens	303	303	54%	5e-100	79.89%	760	BI768670.1
☑	BP270223 Sugano cDNA library, small intestine Homo sapiens cDNA clone KAR01609 5', mRNA sequence	Homo sapiens	277	277	47%	1e-90	82.53%	587	BP270223.1
☑	AGENCOURT_16388570 NIH_ZGC_7 Danio rerio cDNA clone IMAGE:7040629 5', mRNA sequence	Danio rerio	278	278	70%	5e-90	53.47%	773	CF998755.1
☑	BB625035 RIKEN full-length enriched, adult male colon Mus musculus cDNA clone 9030619K19 5', mRNA seque…	Mus musculus	275	275	56%	2e-89	70.77%	658	BB625035.1
☑	DKFZp469C0329_r1 469 (synonym: pkid1) Pongo abelii cDNA clone DKFZp469C0329 5', mRNA sequence	Pongo abelii	274	274	47%	2e-89	81.82%	579	CR769464.1
☑	G1146P313FM7.T0 Anolis carolinensis pooled normalized ovary cDNA library Anolis carolinensis cDNA, mRNA s…	Anolis carolinensis	272	272	65%	2e-87	58.01%	877	FG753538.1
☑	FDR107-P00016-DEPE-F_I05 FDR107 Danio rerio cDNA clone FDR107-P00016-BR_I05 5', mRNA sequence	Danio rerio	266	266	72%	2e-84	50.97%	971	EH489168.1
☑	AGENCOURT_73729288 NICHD_XGC_int_m Xenopus laevis cDNA clone IMAGE:8528120 5', mRNA sequence	Xenopus laevis	255	255	52%	7e-81	66.85%	798	EB479512.1
☑	FDR107-P00007-DEPE-F_P18 FDR107 Danio rerio cDNA clone FDR107-P00007-BR_P18 5', mRNA sequence	Danio rerio	255	255	65%	1e-80	52.86%	892	EH486047.1
☑	GUTF089483I19 POSSUM_01-POSSUM-GUT-2KB Trichosurus vulpecula cDNA clone 1061024739234, mRNA s…	Trichosurus vulp…	254	254	60%	2e-80	57.82%	863	EC326544.1
☑	KIDNEYF091703O14 POSSUM_01-POSSUM-KIDNEY-2KB Trichosurus vulpecula cDNA clone 1061024801200, …	Trichosurus vulp…	253	253	54%	2e-80	63.30%	745	EC345129.1
☑	KIDNEYF089473L22 POSSUM_01-POSSUM-KIDNEY-2KB Trichosurus vulpecula cDNA clone 1061024726253, …	Trichosurus vulp…	251	251	54%	5e-80	63.30%	635	EC337250.1
☑	GUTF054260M6 POSSUM_01-POSSUM-GUT-2KB Trichosurus vulpecula cDNA clone 1061020275316, mRNA s…	Trichosurus vulp…	253	324	69%	9e-80	62.77%	876	DY593226.1
☑	GUTF101121H24 POSSUM_01-POSSUM-GUT-2KB Trichosurus vulpecula cDNA clone 1061029324558, mRNA …	Trichosurus vulp…	246	246	53%	9e-78	62.16%	712	EG597350.1

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.

Chose Match: Accession #BW982765.1, a 803 base pair clone from Sus scrofa (pig). See below for alignment details:

Percent Identity: 72.62%
Query Coverage: 75%
E-val: 1e-137





Alignment view: Pairwise

BW982765 full-length enriched swine cDNA library, adult intestine Sus scrofa cDNA clone ITT010076D10 5', mRNA sequence

Sequence ID: BW982765.1 Length: 803 Number of Matches: 1

Range 1: 10 to 798 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
399 bits(1025)	1e-137	Compositional matrix adjust.	215/263(82%)	244/263(92%)	0/263(0%)	+1

```
Query  58   FLGHIKRPWGICVGFLCQFGIMPLTGFILSVAFDILPLQAVVVLIIGCCPGGTASNILAY  117
            FLGH+KRPWGIC+GFLCQFGIMPLTGFILSV FDILP+QAVVV+I+GCCPGGT+SNILAY
Sbjct  10   FLGHVKRPWGICIGFLCQFGIMPLTGFILSVGFDILPVQAVVVIIMGCCPGGTSSNILAY  189

Query  118  WVDGDMDLSVSMTTCSTllalgmmplcllIYTKMWVDSGSIVIPYDNIGTSLVSLVVPVS  177
            WVDGDMDLS+SMTTCST+LALGMMPLCLLIYTK+WV+SG IVIP+DNIG SLV+L+VPVS
Sbjct  190  WVDGDMDLSISMTTCSTVLALGMMPLCLLIYTKLWVNSGMIVIPFDNIGISLVALIVPVS  369

Query  178  IGMFVNHKWPQKAKIILKIGSiagailivliavvggilYQSAWIIAPKLWIIGTIFPVAG  237
            IG+FVNHKWP++AKIILK+GSI GA+LIVLIAV+GG+LYQSAWII PKLWIIGTIFP+AG
Sbjct  370  IGIFVNHKWPRQAKIILKVGSITGALLIVLIAVIGGVLYQSAWIIEPKLWIIGTIFPLAG  549

Query  238  YSLGFLLARIAGLPWYRCRTVAFETGMQNTQLCSTIVQLSFTPEELNVVFTFPLIYSIFQ  297
            YSLGF LAR+AG  W RCRTVA ETGMQNTQLCSTIVQLSFTPEELN+VFTFPLIYS+FQ
Sbjct  550  YSLGFFLARLAGQSWDRCRTVALETGMQNTQLCSTIVQLSFTPEELNLVFTFPLIYSVFQ  729

Query  298  LAFAAIFLGFYVAYKKCHGKNKA  320
            L  A +F+G Y  Y+K +GK +
Sbjct  730  LFLAGVFVGIYAVYRKWYGKKQG  798
```

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.

> S. scrofa protein (sequence taken from BLAST result)

```
FLGHVKRPWGICIGFLCQFGIMPLTGFILSVGFDILPVQAVVVIIMGCCPGGTSSNILAYW
VDGDMDLSISMTTCSTVLALGMMPLCLLIYTKLWVNSGMIVIPFDNIGISLVALIVPVSIG
IFVNHKWPRQAKIILKVGSITGALLIVLIAVIGGVLYQSAWIIEPKLWIIGTIFPLAGYSLGF
FLARLAGQSWDRCRTVALETGMQNTQLCSTIVQLSFTPEELNLVFTFPLIYSVFQLFLAG
VFVGIYAVYRKWYGKKQG
```

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: *Sus* bile acid transporter

Species: Sus scrofa

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;

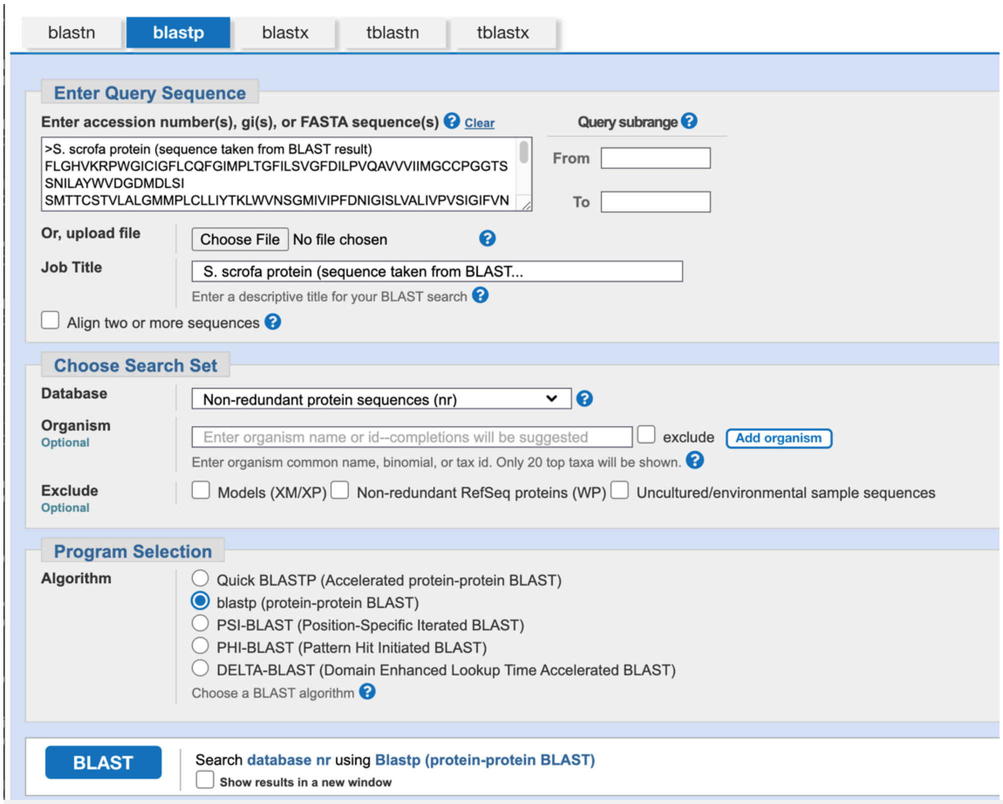
Mammalia; Eutheria; Laurasiatheria; Artiodactyla; Suina; Suidae;

Sus.

[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 the NR database (set-up shown below) resulted in top hit result to a protein from *Sus scrofa* (pig). Additional screenshots below display top hits and alignment details.



Enter Query Sequence

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

```
>S. scrofa protein (sequence taken from BLAST result)
FLGHVKRPWGICIGFLCQFGIMPLTGFILSVGFDILPVQAVVVIIMGCCPGGTS
SNILAYWVDGDMDLSI
SMTTCSTVLALGMMPLCLLIYTKLWVNSGMIVIPFDNIGISLVALIVPVSIGIFVN
```

Query subrange

From

To

Or, upload file: Choose File No file chosen

Job Title: S. scrofa protein (sequence taken from BLAST...

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

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) □ Non-redundant RefSeq proteins (WP) □ Uncultured/environmental sample sequences

Program Selection

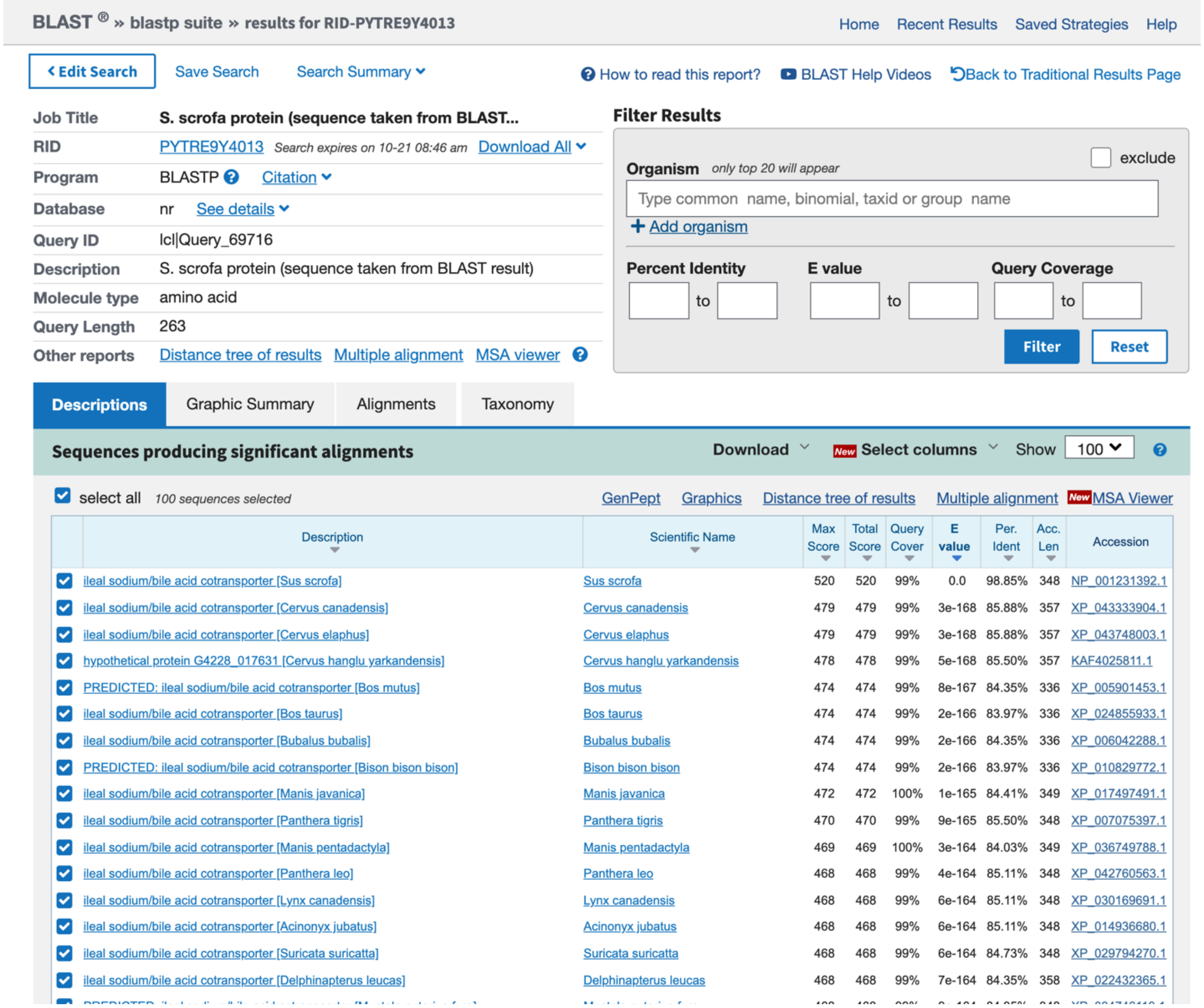
Algorithm

○ 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



Job Title	S. scrofa protein (sequence taken from BLAST...
RID	PYTRE9Y4013 *Search expires on 10-21 08:46 am* Download All
Program	BLASTP Citation
Database	nr See details
Query ID	lcl|Query_69716
Description	S. scrofa protein (sequence taken from BLAST result)
Molecule type	amino acid
Query Length	263
Other reports	Distance tree of results Multiple alignment MSA viewer

Filter Results

Organism *only top 20 will appear* □ exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity to

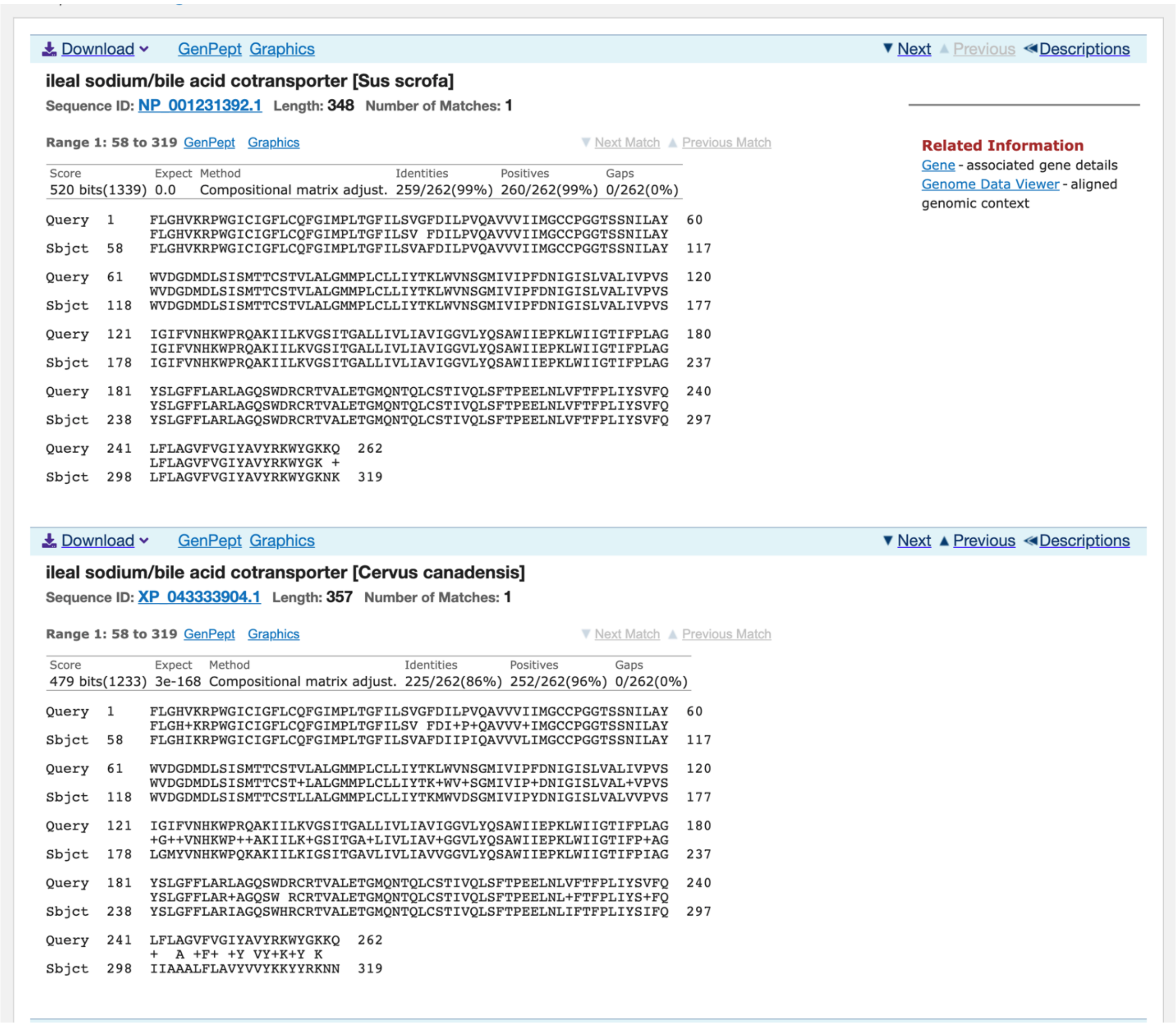
E value to

Query Coverage to

Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	ileal sodium/bile acid cotransporter [Sus scrofa]	Sus scrofa	520	520	99%	0.0	98.85%	348	NP_001231392.1
☑	ileal sodium/bile acid cotransporter [Cervus canadensis]	Cervus canadensis	479	479	99%	3e-168	85.88%	357	XP_043333904.1
☑	ileal sodium/bile acid cotransporter [Cervus elaphus]	Cervus elaphus	479	479	99%	3e-168	85.88%	357	XP_043748003.1
☑	hypothetical protein G4228_017631 [Cervus hanglu yarkandensis]	Cervus hanglu yarkandensis	478	478	99%	5e-168	85.50%	357	KAF4025811.1
☑	PREDICTED: ileal sodium/bile acid cotransporter [Bos mutus]	Bos mutus	474	474	99%	8e-167	84.35%	336	XP_005901453.1
☑	ileal sodium/bile acid cotransporter [Bos taurus]	Bos taurus	474	474	99%	2e-166	83.97%	336	XP_024855933.1
☑	ileal sodium/bile acid cotransporter [Bubalus bubalis]	Bubalus bubalis	474	474	99%	2e-166	84.35%	336	XP_006042288.1
☑	PREDICTED: ileal sodium/bile acid cotransporter [Bison bison bison]	Bison bison bison	474	474	99%	2e-166	83.97%	336	XP_010829772.1
☑	ileal sodium/bile acid cotransporter [Manis javanica]	Manis javanica	472	472	100%	1e-165	84.41%	349	XP_017497491.1
☑	ileal sodium/bile acid cotransporter [Panthera tigris]	Panthera tigris	470	470	99%	9e-165	85.50%	348	XP_007075397.1
☑	ileal sodium/bile acid cotransporter [Manis pentadactyla]	Manis pentadactyla	469	469	100%	3e-164	84.03%	349	XP_036749788.1
☑	ileal sodium/bile acid cotransporter [Panthera leo]	Panthera leo	468	468	99%	4e-164	85.11%	348	XP_042760563.1
☑	ileal sodium/bile acid cotransporter [Lynx canadensis]	Lynx canadensis	468	468	99%	6e-164	85.11%	348	XP_030169691.1
☑	ileal sodium/bile acid cotransporter [Acinonyx jubatus]	Acinonyx jubatus	468	468	99%	6e-164	85.11%	348	XP_014936680.1
☑	ileal sodium/bile acid cotransporter [Suricata suricatta]	Suricata suricatta	468	468	99%	6e-164	84.73%	348	XP_029794270.1
☑	ileal sodium/bile acid cotransporter [Delphinapterus leucas]	Delphinapterus leucas	468	468	99%	7e-164	84.35%	358	XP_022432365.1



ileal sodium/bile acid cotransporter [Sus scrofa]

Sequence ID: NP_001231392.1 Length: 348 Number of Matches: 1

Related Information
Gene - associated gene details
Genome Data Viewer - aligned genomic context

Range 1: 58 to 319 GenPept Graphics

Next Match Previous Match

```
Score           Expect  Method                        Identities     Positives      Gaps
520 bits(1339)  0.0     Compositional matrix adjust.  259/262(99%)   260/262(99%)   0/262(0%)

Query  1    FLGHVKRPWGICIGFLCQFGIMPLTGFILSVGFDILPVQAVVVIIMGCCPGGTSSNILAY  60
            FLGHVKRPWGICIGFLCQFGIMPLTGFILSV FDILPVQAVVVIIMGCCPGGTSSNILAY
Sbjct  58   FLGHVKRPWGICIGFLCQFGIMPLTGFILSVAFDILPVQAVVVIIMGCCPGGTSSNILAY  117

Query  61   WVDGDMDLSISMTTCSTVLALGMMPLCLLIYTKLWVNSGMIVIPFDNIGISLVALIVPVS  120
            WVDGDMDLSISMTTCSTVLALGMMPLCLLIYTKLWVNSGMIVIPFDNIGISLVALIVPVS
Sbjct  118  WVDGDMDLSISMTTCSTVLALGMMPLCLLIYTKLWVNSGMIVIPFDNIGISLVALIVPVS  177

Query  121  IGIFVNHKWPRQAKIILKVGSITGALLIVLIAVIGGVLYQSAWIIEPKLWIIGTIFPLAG  180
            IGIFVNHKWPRQAKIILKVGSITGALLIVLIAVIGGVLYQSAWIIEPKLWIIGTIFPLAG
Sbjct  178  IGIFVNHKWPRQAKIILKVGSITGALLIVLIAVIGGVLYQSAWIIEPKLWIIGTIFPLAG  237

Query  181  YSLGFFLARLAGQSWDRCRTVALETGMQNTQLCSTIVQLSFTPEELNLVFTFPLIYSVFQ  240
            YSLGFFLARLAGQSWDRCRTVALETGMQNTQLCSTIVQLSFTPEELNLVFTFPLIYSVFQ
Sbjct  238  YSLGFFLARLAGQSWDRCRTVALETGMQNTQLCSTIVQLSFTPEELNLVFTFPLIYSVFQ  297

Query  241  LFLAGVFVGIYAVYRKWYGKKQ  262
            LFLAGVFVGIYAVYRKWYGK +
Sbjct  298  LFLAGVFVGIYAVYRKWYGKNK  319
```

Download GenPept Graphics Next Previous Descriptions

ileal sodium/bile acid cotransporter [Cervus canadensis]

Sequence ID: XP_043333904.1 Length: 357 Number of Matches: 1

Range 1: 58 to 319 GenPept Graphics

Next Match Previous Match

```
Score           Expect  Method                        Identities     Positives      Gaps
479 bits(1233)  3e-168  Compositional matrix adjust.  225/262(86%)   252/262(96%)   0/262(0%)

Query  1    FLGHVKRPWGICIGFLCQFGIMPLTGFILSVGFDILPVQAVVVIIMGCCPGGTSSNILAY  60
            FLGH+KRPWGICIGFLCQFGIMPLTGFILSV FDI+P+QAVVV+IMGCCPGGTSSNILAY
Sbjct  58   FLGHIKRPWGICIGFLCQFGIMPLTGFILSVAFDIIPIQAVVVLIMGCCPGGTSSNILAY  117

Query  61   WVDGDMDLSISMTTCSTVLALGMMPLCLLIYTKLWVNSGMIVIPFDNIGISLVALIVPVS  120
            WVDGDMDLSISMTTCST+LALGMMPLCLLIYTK+WV+SGMIVIP+DNIGISLVAL+VPVS
Sbjct  118  WVDGDMDLSISMTTCSTLLALGMMPLCLLIYTKMWVDSGMIVIPYDNIGISLVALVVPVS  177

Query  121  IGIFVNHKWPRQAKIILKVGSITGALLIVLIAVIGGVLYQSAWIIEPKLWIIGTIFPLAG  180
            +G++VNHKWP++AKIILK+GSITGA+LIVLIAV+GGVLYQSAWIIEPKLWIIGTIFP+AG
Sbjct  178  LGMYVNHKWPQKAKIILKIGSITGAVLIVLIAVVGGVLYQSAWIIEPKLWIIGTIFPIAG  237

Query  181  YSLGFFLARLAGQSWDRCRTVALETGMQNTQLCSTIVQLSFTPEELNLVFTFPLIYSVFQ  240
            YSLGFFLAR+AGQSW RCRTVALETGMQNTQLCSTIVQLSFTPEELNL+FTFPLIYS+FQ
Sbjct  238  YSLGFFLARIAGQSWHRCRTVALETGMQNTQLCSTIVQLSFTPEELNLIFTFPLIYSIFQ  297

Query  241  LFLAGVFVGIYAVYRKWYGKKQ  262
            +  A +F+ +Y VY+K+Y K
Sbjct  298  IIAAALFLAVYVVYKKYYRKNN  319
```

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

Re-labelled sequences for alignment

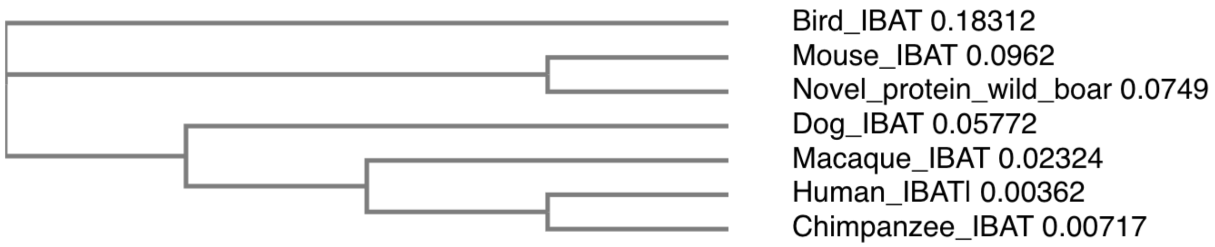
```
> Human_IBAT| NP_000443.2 | original protein | ileal sodium/bile acid
cotransporter [Homo sapiens]
MNDPNSCVDNATVCSGASCVVPESNFNNILSVVLSTVLTILLALVMFSMGCNVEIKKFLGHIKRPWGICV
GFLCQFGIMPLTGFILSVAFDILPLQAVVVLIIGCCPGGTASNILAYWVDGDMDLSVSMTTCSTLLALGM
MPLCLLIYTKMWVDSGSIVIPYDNIGTSLVSLVVPVSIGMFVNHKWPQKAKIILKIGSIAGAILIVLIAV
VGGILYQSAWIIAPKLWIIGTIFPVAGYSLGFLLARIAGLPWYRCRTVAFETGMQNTQLCSTIVQLSFTP
EELNVVFTFPLIYSIFQLAFAAIFLGFYVAYKKCHGKNKAEIPESKENGTEPESSFYKANGGFQPDEK

> Novel_protein_wild_boar | S. scrofa | (sequence taken from BLAST result)
FLGHVKRPWGICIGFLCQFGIMPLTGFILSVGFDILPVQAVVVIIMGCCPGGTSSNILAYWVDGDMDLSISMTTCST
VLALGMMPLCLLIYTKLWVNSGMIVIPFDNIGISLVALIVPVSIGIFVNHKWPRQAKIILKVGSITGALLIVLIAVI
GGVLYQSAWIIEPKLWIIGTIFPLAGYSLGFFLARLAGQSWDRCRTVALETGMQNTQLCSTIVQLSFTPEELNLVFT
FPLIYSVFQLFLAGVFVGIYAVYRKWYGKKQG

> Dog_IBAT | NP_001002968.1 | ileal sodium/bile acid cotransporter [Canis
lupus familiaris]
MNNSTGCSANATVCNGASCVVAQNNFNDILSVVLSTVLTILLAMVMFSMGCNVEIKKFLGHIKRPWGICV
GFLCQFGIMPLTGFILSVAFDILPLQAVVVLIMGCCPGGTASNILAYWVDGDMDLSISMTTCSTLLALGM
MPLCLFIYTKMWVDSGTIVIPFDNIGTSLVALVVPVSIGMLVNHKWPQKAKIILKVGSITGAILIVLIAV
VGGILYQSAWIIAPKLWIIGTLFPLAGYSLGFLLARISGQSWHRCRTVALETGMQNTQLCSTIVQLSFTQ
EELNVVFTFPLIYSIFQLAFAAIFLGIYVAYKKCYEKNNAEFPESKDNETVSESSLYKVNEGFQPDAK

> Mouse_IBAT | NP_035518.1 | ileal sodium/bile acid cotransporter [Mus
musculus]
MDNSSVCPPNATVCEGDSCVVPESNFNAILNTVMSTVLTILLAMVMFSMGCNVEVHKFLGHIKRPWGIFV
GFLCQFGIMPLTGFILSVASGILPVQAVVVLIMGCCPGGTGSNILAYWIDGDMDLSVSMTTCSTLLALGM
MPLCLFVYTKMWVDSGTIVIPYDSIGISLVALVIPVSFGMFVNHKWPQKAKIILKIGSITGVILIVLIAV
IGGILYQSAWIIEPKLWIIGTIFPIAGYSLGFFLARLAGQPWYRCRTVALETGMQNTQLCSTIVQLSFSP
EDLNLVFTFPLIYTVFQLVFAAVILGIYVTYRKCYGKNDAEFLEKTDNEMDSRPSFDETNKGFQPDEK

> Chimpanzee_IBAT | XP_522716.2 | ileal sodium/bile acid cotransporter [Pan
troglodytes]
MNDPNSCVDNATVCSGASCVVPESNFNNILSVVLSTVLTILLALVMFSMGCNVEIKKFLGHIKRPWGICV
GFLCQFGIMPLTGFILSVAFDILPLQAVVVLIIGCCPGGTASNILAYWVDGDMDLSVSMTTCSTLLALGM
MPLCLLIYTKMWVDSGSIVIPYDNIGTSLVALVVPVSIGMFVNHKWPQKAKIILKIGSIAGAILIVLIAV
VGGILYQSSWIIAPKLWIIGTIFPVAGYSLGFLLARIAGLPWYRCRTVAFETGMQNTQLCSTIVQLSFTP
EELNVVFTFPLIYSIFQLAFAAIFLGFYVAYKKCHGKNKAETPESKENGTEPESSFYKANGGFQPDEK

> Macaque_IBAT | XP_001095212.2 | ileal sodium/bile acid cotransporter
[Macaca mulatta]
MNEPNSCVDNATVCSGASCVVPDSNFNNTLSVVLSTVLTILLALVMFSMGCNVEIKKFLGHIKRPWGICV
GFLCQFGIMPLTGFVLSVAFDILPIQAVVVLIMGCCPGGTSSNILAYWVDGDMDLSVSMTTCSTLLALGM
MPLCLLIYTKMWVDSGSIVIPYDNIGTSLVALVVPVSIGMFVNHKWPQKAKIILKIGSIAGAILIVLIAV
VGGILYQSAWIIAPKLWIIGTIFPVAGYSLGFLLARIAGLPWHRCRTVAFETGMQNTQLCSTIVQLSFTL
EELNIVFTFPLIYSIFQLAFAAIFLGFYVAYKKCHGKNKAEIPESKENETEPESSFYKINGGFKPDEK

> Bird_IBAT | NP_001305956.2 | ileal sodium/bile acid cotransporter [Gallus
gallus]
MQSYLLSRHFNTKMLDNSTACPAVDNSTACPENATICSGTSCVLPEDDFNQTLSVVLSTVLTIMLALVMF
SMGCNVEIKKFLHHIKRPWGIFVGFLCQFGIMPLTAFLLSLAFDVHPIQAVVVMIMGCCPGGTASNIIAY
WVDGDMDLSISMTTCSTLLAMGMMPLCLFVYTKMWTDSDAIVLPYDSIGISLVALVVPVSVGVFVNHKWP
SKAKRILKVGSIAGAILIVITAVVGGILYKGSWVITPKLWIIGTIFPAAGYSLGFFLARLAGLSWSRCRT
VSLETGMQNTQLCSTIVQLSFSPEQLELMFTFPLIYSIFQLLFALMILAGYRVYIKRCVKTNKDVEKTEE
KDDSKSISSHAKENGGFVSDETK
```

Alignment

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


Bird_IBAT                    MQSYLLSRHFNTKMLDNSTACPAVDNSTACPENATICSGTSCVLPEDDFNQTLSVVLSTV
Mouse_IBAT                   -----------------------MDNSSVCPPNATVCEGDSCVVPESNFNAILNTVMSTV
Novel_protein_wild_boar      ------------------------------------------------------------
Dog_IBAT                     -----------------------MNNSTGCSANATVCNGASCVVAQNNFNDILSVVLSTV
Macaque_IBAT                 -----------------------MNEPNSCVDNATVCSGASCVVPDSNFNNTLSVVLSTV
Human_IBAT|                  ------------------------------------------------------------
Chimpanzee_IBAT              -----------------------MNDPNSCVDNATVCSGASCVVPESNFNNILSVVLSTV


Bird_IBAT                    LTIMLALVMFSMGCNVEIKKFLHHIKRPWGIFVGFLCQFGIMPLTAFLLSLAFDVHPIQA
Mouse_IBAT                   LTILLAMVMFSMGCNVEVHKFLGHIKRPWGIFVGFLCQFGIMPLTGFILSVASGILPVQA
Novel_protein_wild_boar      --------------------FLGHVKRPWGICIGFLCQFGIMPLTGFILSVGFDILPVQA
Dog_IBAT                     LTILLAMVMFSMGCNVEIKKFLGHIKRPWGICVGFLCQFGIMPLTGFILSVAFDILPLQA
Macaque_IBAT                 LTILLALVMFSMGCNVEIKKFLGHIKRPWGICVGFLCQFGIMPLTGFVLSVAFDILPIQA
Human_IBAT|                  ---------------------------------GFLCQFGIMPLTGFILSVAFDILPLQA
Chimpanzee_IBAT              LTILLALVMFSMGCNVEIKKFLGHIKRPWGICVGFLCQFGIMPLTGFILSVAFDILPLQA
                                                              ************.*:**:. .: *:**

Bird_IBAT                    VVVMIMGCCPGGTASNIIAYWVDGDMDLSISMTTCSTLLAMGMMPLCLFVYTKMWTDSDA
Mouse_IBAT                   VVVLIMGCCPGGTGSNILAYWIDGDMDLSVSMTTCSTLLALGMMPLCLFVYTKMWVDSGT
Novel_protein_wild_boar      VVVIIMGCCPGGTSSNILAYWVDGDMDLSISMTTCSTVLALGMMPLCLLIYTKLWVNSGM
Dog_IBAT                     VVVLIMGCCPGGTASNILAYWVDGDMDLSISMTTCSTLLALGMMPLCLFIYTKMWVDSGT
Macaque_IBAT                 VVVLIMGCCPGGTSSNILAYWVDGDMDLSVSMTTCSTLLALGMMPLCLLIYTKMWVDSGS
Human_IBAT|                  VVVLIIGCCPGGTASNILAYWVDGDMDLSVSMTTCSTLLALGMMPLCLLIYTKMWVDSGS
Chimpanzee_IBAT              VVVLIIGCCPGGTASNILAYWVDGDMDLSVSMTTCSTLLALGMMPLCLLIYTKMWVDSGS
                             ***:*:*******.***:***:*******:*******:**:*******::***:*.:*.

Bird_IBAT                    IVLPYDSIGISLVALVVPVSVGVFVNHKWPSKAKRILKVGSIAGAILIVITAVVGGILYK
Mouse_IBAT                   IVIPYDSIGISLVALVIPVSFGMFVNHKWPQKAKIILKIGSITGVILIVLIAVIGGILYQ
Novel_protein_wild_boar      IVIPFDNIGISLVALIVPVSIGIFVNHKWPRQAKIILKVGSITGALLIVLIAVIGGVLYQ
Dog_IBAT                     IVIPFDNIGTSLVALVVPVSIGMLVNHKWPQKAKIILKVGSITGAILIVLIAVVGGILYQ
Macaque_IBAT                 IVIPYDNIGTSLVALVVPVSIGMFVNHKWPQKAKIILKIGSIAGAILIVLIAVVGGILYQ
Human_IBAT|                  IVIPYDNIGTSLVSLVVPVSIGMFVNHKWPQKAKIILKIGSIAGAILIVLIAVVGGILYQ
Chimpanzee_IBAT              IVIPYDNIGTSLVALVVPVSIGMFVNHKWPQKAKIILKIGSIAGAILIVLIAVVGGILYQ
                             **:*:*.** ***:*::***.*::****** :** ***:***:*.:***: **:**:**:

Bird_IBAT                    GSWVITPKLWIIGTIFPAAGYSLGFFLARLAGLSWSRCRTVSLETGMQNTQLCSTIVQLS
Mouse_IBAT                   SAWIIEPKLWIIGTIFPIAGYSLGFFLARLAGQPWYRCRTVALETGMQNTQLCSTIVQLS
Novel_protein_wild_boar      SAWIIEPKLWIIGTIFPLAGYSLGFFLARLAGQSWDRCRTVALETGMQNTQLCSTIVQLS
Dog_IBAT                     SAWIIAPKLWIIGTLFPLAGYSLGFLLARISGQSWHRCRTVALETGMQNTQLCSTIVQLS
Macaque_IBAT                 SAWIIAPKLWIIGTIFPVAGYSLGFLLARIAGLPWHRCRTVAFETGMQNTQLCSTIVQLS
Human_IBAT|                  SAWIIAPKLWIIGTIFPVAGYSLGFLLARIAGLPWYRCRTVAFETGMQNTQLCSTIVQLS
Chimpanzee_IBAT              SSWIIAPKLWIIGTIFPVAGYSLGFLLARIAGLPWYRCRTVAFETGMQNTQLCSTIVQLS
                             .:*:* ********:** *******:***::* .* *****::*****************

Bird_IBAT                    FSPEQLELMFTFPLIYSIFQLLFALMILAGYRVYIKRCVKTNKDVEKTEEKDDSKSISSH
Mouse_IBAT                   FSPEDLNLVFTFPLIYTVFQLVFAAVILGIYVTY-RKCYGKNDAEFLEKTDNEMDSRPSF
Novel_protein_wild_boar      FTPEELNLVFTFPLIYSVFQLFLAGVFVGIYAVY-RKWYGKKQG----------------
Dog_IBAT                     FTQEELNVVFTFPLIYSIFQLAFAAIFLGIYVAY-KKCYEKNNAEFPESKDNETVSESSL
Macaque_IBAT                 FTLEELNIVFTFPLIYSIFQLAFAAIFLGFYVAY-KKCHGKNKAEIPESKENETEPESSF
Human_IBAT|                  FTPEELNVVFTFPLIYSIFQLAFAAIFLGFYVAY-KKCHGKNKAEIPESKENGTEPESSF
Chimpanzee_IBAT              FTPEELNVVFTFPLIYSIFQLAFAAIFLGFYVAY-KKCHGKNKAETPESKENGTEPESSF
                             *: *:*:::*******::*** :* :::. * .* ..   .:.

Bird_IBAT                    AKENGGFVSDETK
Mouse_IBAT                   DETNKGFQPDEK-
Novel_protein_wild_boar      -------------
Dog_IBAT                     YKVNEGFQPDAK-
Macaque_IBAT                 YKINGGFKPDEK-
Human_IBAT|                  YKANGGFQPDEK-
Chimpanzee_IBAT              YKANGGFQPDEK-
```

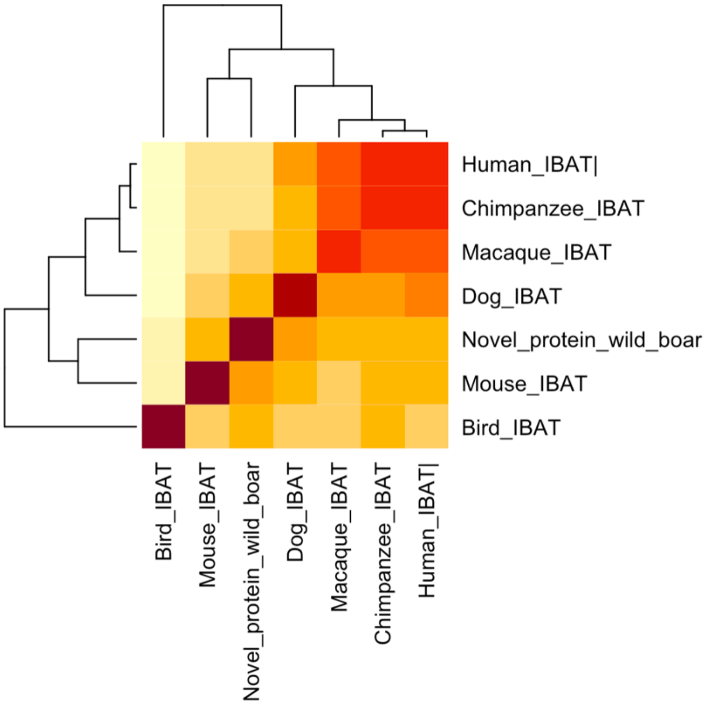
[Q6] Create a phylogenetic tree, using either a parsimony or distance-based approach. Bootstrapping and tree rooting are optional. Use “simple phylogeny” online from the EBI or any respected phylogeny program (such as MEGA, PAUP, or Phylip). Paste an image of your Cladogram or tree output in your report.

Phylogenetic tree output from EBI Muscle alignment results computed in Q5



[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.

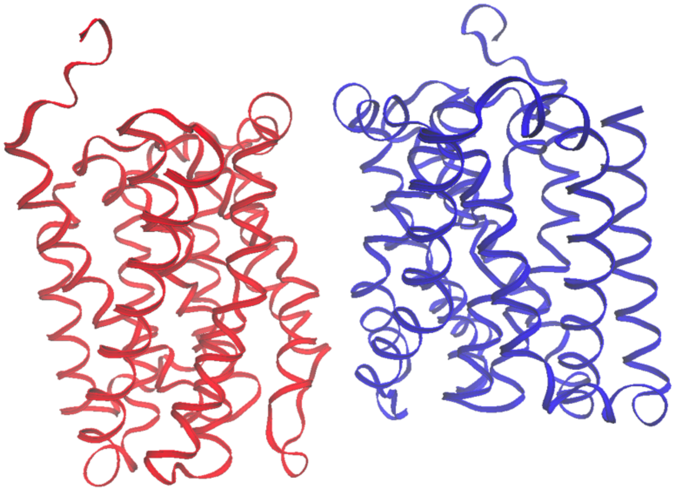
List the top 3 *unique* hits (i.e. not hits representing different chains from the same structure) along with their Evalue and sequence identity to your query. Please also add annotation details of these structures. For example include the annotation terms PDB identifier (structureId), Method used to solve the structure (experimentalTechnique), resolution (resolution), and source organism (source).

ID	Chain	Technique	Resolution	Source	E-Val	% Identity
4N7W	A	X-Ray	1.951	Yersinia frederiksenii	1e-13	27.87
3ZUY	A	X-Ray	2.200	Neisseria meningitidis	1e-11	30.26
4BER	A	X-Ray	2.600	Legionella pneumophila	0.19	32.29

[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?

Structure of 4N7W containing both chains. Somewhat unlikely to be similar in structure to my novel protein, given the low percent identity. However, the structure may provide some level of information on bile acid location and orientation during transport.



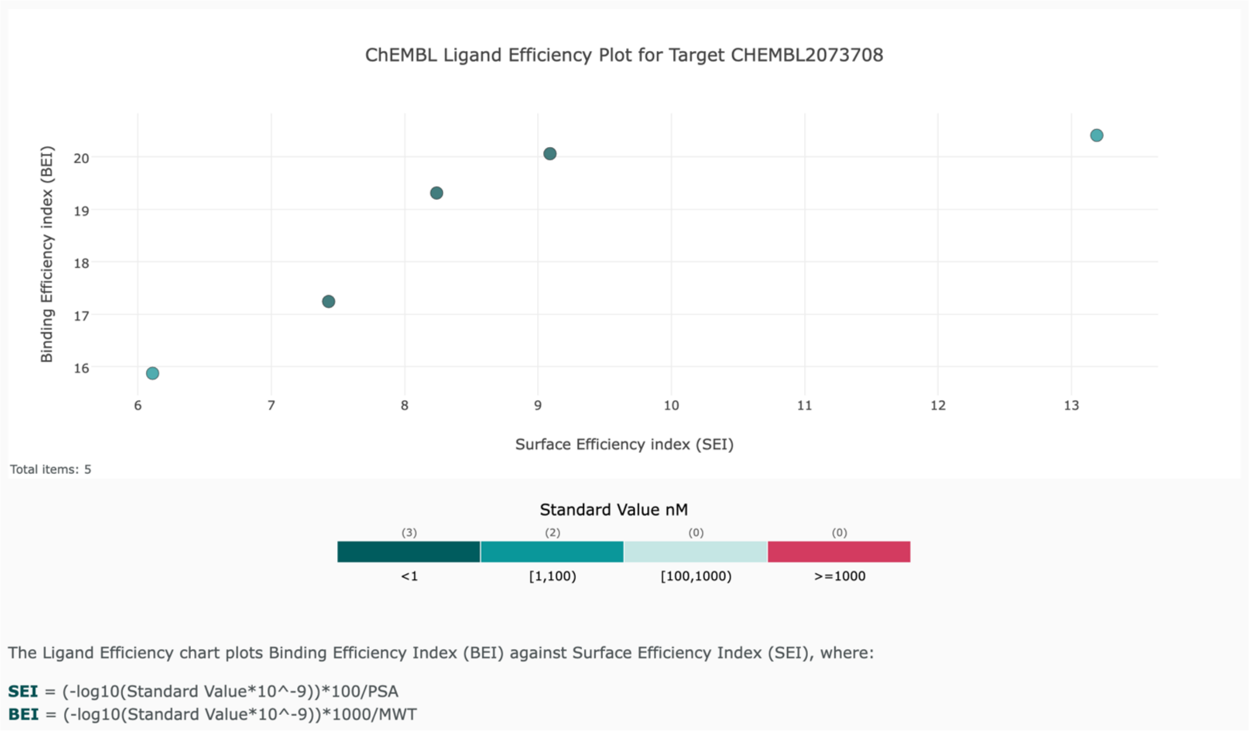
[Q10] Perform a "Target" search of ChEMBEL (https://www.ebi.ac.uk/chembl/) with your novel sequence. Are there any **Target Associated Assays** and **ligand efficiency data** reported that may be useful starting points for exploring potential inhibition of your novel protein?

The CHEMBEL target search returned 16 targets. The top target is for Ileal sodium/bile acid cotransporter in Mus musculus (CHEMBL2073708). For this target there are 2 binding assays and 8 functional assays. The binding assays provide evidence that inhibiting this protein target may provide therapeutic outcomes in individuals with hypercholesterolemia. For example, an assay (CHEMBL4200891: https://www.ebi.ac.uk/chembl/assay_report_card/CHEMBL4200891/) completed in COS cells supports that inhibition of this protein target prevents reuptake of taurocholic acid (a bile acid). The binding assay linked the following manuscript:

Design of Gut-Restricted Thiazolidine Agonists of G Protein-Coupled Bile Acid Receptor 1 (GPBAR1, TGR5)
Tao Chen, Nicholas William Reich, Noah Bell, Patricia D. Finn, David Rodriguez, Jill Kohler, Kenji Kozuka, Limin He, Andrew G. Spencer, Dominique Charmot, Marc Navre, Christopher W. Carreras, Samantha Koo-McCoy, Jocelyn Tabora, Jeremy S. Caldwell, Jeffrey W. Jacobs, and Jason Gustaf Lewis
Journal of Medicinal Chemistry ***2018*** *61 (17), 7589-7613*
DOI: 10.1021/acs.jmedchem.8b00308

In vivo, prevention of bile acid reuptake would likely result in greater synthesis and excretion of bile acids, and consequently, greater elimination of cholesterol.

The ligand efficiency graphs below:



Total items: 5

The Ligand Efficiency chart plots Binding Efficiency Index (BEI) against Surface Efficiency Index (SEI), where:

SEI = (-log10(Standard Value*10^-9))*100/PSA
BEI = (-log10(Standard Value*10^-9))*1000/MWT