Jimmi Nguyen A15976796

BIMM 143

Professor Grant

February 6th, 2023

Find a Gene Assignment Part 1

[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: Tumor protein p53

Accession: NP_000537

Species: Homo Sapiens

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

Method: TBLASTN search against chicken ESTs

Database: Expressed Sequence Tags (est)

Organism: Chicken (Taxid:9031)

Chosen Match: Accession BM440292.1, a 568 base pair clone from Gallus gallus. See below for alignment details

ℹ Your search is limited to records that include: chicken (taxid:9031)

Job Title	NP_000537.3 cellular tumor antigen p53 isoform...
RID	Y2HVKW3K016 *Search expires on 02-08 01:31 am* Download All ▾
Program	TBLASTN ❓ Citation ▾
Database	est See details ▾
Query ID	lcl|Query_20167
Description	NP_000537.3 cellular tumor antigen p53 isoform a [Homo ...
Molecule type	amino acid
Query Length	393
Other reports	❓

Filter Results

Organism *only top 20 will appear*   ☐ exclude

Type common name, binomial, taxid or group name

➕ Add organism

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

Filter   **Reset**

Descriptions | Graphic Summary | Alignments | Taxonomy

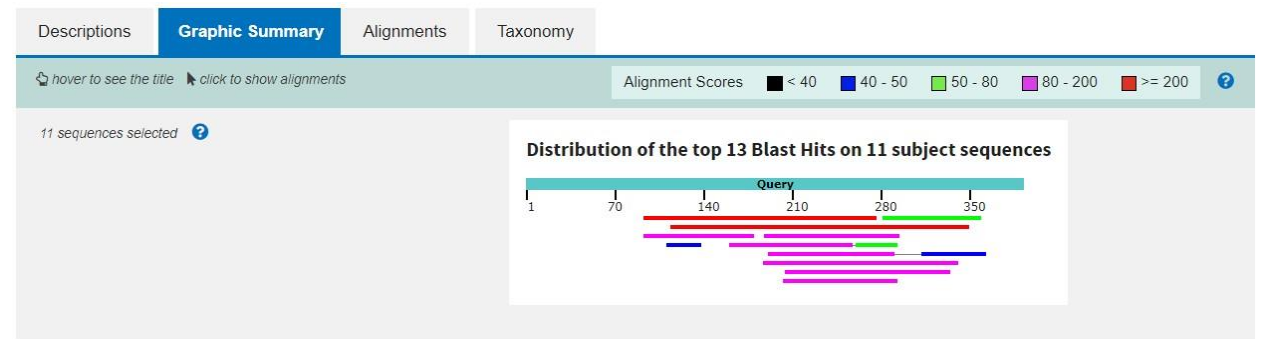**Sequences producing significant alignments**   Download ▾   Select columns ▾   Show 100 ▾ ❓

☑ select all *11 sequences selected*   GenBank   Graphics

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑ pgr1n.pk002.l17 Normalized Chicken Reproductive Tract cDNA Library (pgr1n) Gallus gallus cDNA clone pgr1n.pk002.l...	Gallus gallus	268	268	46%	1e-88	68.11%	568	BM440292.1
☑ 603101519F1 CSEQCHN03 Gallus gallus cDNA clone ChEST34e5 5', mRNA sequence	Gallus gallus	246	246	60%	9e-79	50.21%	854	BU203323.1
☑ pat.pk0040.d2.f chicken activated T cell cDNA Gallus gallus cDNA clone pat.pk0040.d2.f 5' similar to cellular tumor anti...	Gallus gallus	174	174	27%	6e-52	76.64%	547	AI981042.1
☑ pgr1n.UA001.559 Normalized chicken reproductive tract (pgr1n) Gallus gallus cDNA clone UAGgRT21_D09_03061705...	Gallus gallus	166	217	33%	1e-49	77.78%	408	DT659470.1
☑ pgp2n.pk008.m15 Normalized chicken pituitary/hypothalamus/pineal cDNA library (pgp2n) Gallus gallus cDNA clone pg...	Gallus gallus	169	216	38%	1e-49	79.00%	639	CD215735.1

Descriptions | **Graphic Summary** | Alignments | Taxonomy

🔍 *hover to see the title* ▶ *click to show alignments*   Alignment Scores ■ < 40 ■ 40 - 50 ■ 50 - 80 ■ 80 - 200 ■ >= 200 ❓

11 sequences selected ❓

Distribution of the top 13 Blast Hits on 11 subject sequences

---

⬇ Download ▾   GenBank Graphics   ▼ Next ▲ Previous ◀Descriptions

pgr1n.pk002.l17 Normalized Chicken Reproductive Tract cDNA Library (pgr1n) Gallus gallus cDNA clone pgr1n.pk002.l17 5' similar to sp|P10360|P53_CHICK CELLULAR TUMOR ANTIGEN P53 (TUMOR SUPPRESSOR P53) pir|S02193 cellular tumor antigen p53 - chicken emb|CAA31456.1| (X13057) nuclear protein p53 (AA 1 - 367) [Gallus gallus], mRNA sequence

Sequence ID: BM440292.1 Length: 568 Number of Matches: 1

Range 1: 13 to 567 GenBank Graphics   ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
268 bits(685)	1e-88	Compositional matrix adjust.	126/185(68%)	147/185(79%)	1/185(0%)	+1

```
Query  94   SSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPP   153
            S  VPS + Y G + FR+GF+ +GTAKSVTCTYSP LNK++C+LAK CPVQ+ V   PPP
Sbjct  13   SPVVPSTEDYGGDFDFRVGFVEAGTAKSVTCTYSPVLNKVYCRLAKPCPVQVRVGVAPPP   192

Query  154  GTRVRAMAIYKQSQHMTEVVRRCPHHERCSD-SDGLAPPQHLIRVEGNLRVEYLDDRNTF   212
            G+ +RA+A+YK+S+H+ EVVRRCPHHERC   +DGLAP QHLIRVEGN +  Y DD  T
Sbjct  193  GSSLRAVAVYKKSEHVAEVVRRCPHHERCGGGTDGLAPAQHLIRVEGNPQARYHDDETTK   372

Query  213  RHSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRNSFEV   272
            RHSVVVPYEPPEVGSDCTT+ YN+MCNSSCMGGMNRRPILTI+TL    G LLGR
Sbjct  373  RHSVVVPYEPPEVGSDCTTVLYNFMCNSSCMGGMNRRPILTILTLXXPGGQLLGRRCXXX   552

Query  273  RVCAC   277
            RVCAC
Sbjct  553  RVCAC   567
```

Alignment Details:

pgr1n.pk002.l17 Normalized Chicken Reproductive Tract cDNA Library (pgr1n) Gallus gallus cDNA clone pgr1n.pk002.l17 5' similar to sp|P10360|P53_CHICK CELLULAR TUMOR ANTIGEN P53 (TUMOR SUPPRESSOR P53) pir|S02193 cellular tumor antigen p53 - chicken emb|CAA31456.1| (X13057) nuclear protein p53 (AA 1 - 367) [Gallus gallus], mRNA sequence
Sequence ID: BM440292.1Length: 568Number of Matches: 1
Range 1: 13 to 567GenBankGraphicsNext MatchPrevious Match

 Alignment statistics for match #1

Score	Expect	Method	Identities	Positives	Gaps
268 bits(685)	1e-88	Compositional matrix adjust.	126/185(68%)	147/185(79%)	1/185(0%)

```
Query  94   SSSVPSQKTYQGSYGFRLGFLHSGTAKSVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPP  153
            S  VPS + Y G + FR+GF+ +GTAKSVTCTYSP LNK++C+LAK CPVQ+ V  PPP
Sbjct  13   SPVVPSTEDYGGDFDFRVGFVEAGTAKSVTCTYSPVLNKVYCRLAKPCPVQVRVGVAPPP  192


Query  154  GTRVRAMAIYKQSQHMTEVVRRCPHHERCSD-SDGLAPPQHLIRVEGNLRVEYLDDRNTF  212
            G+ +RA+A+YK+S+H+ EVVRRCPHHERC  +DGLAP QHLIRVEGN + Y DD  T
Sbjct  193  GSSLRAVAVYKKSEHVAEVVRRCPHHERCGGGTDGLAPAQHLIRVEGNPQARYHDDETTK  372


Query  213  RHSVVVPYEPPEVGSDCTTIHYNYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRNSFEV  272
            RHSVVVPYEPPEVGSDCTT+ YN+MCNSSCMGGMNRRPILTI+TL   G LLGR
Sbjct  373  RHSVVVPYEPPEVGSDCTTVLYNFMCNSSCMGGMNRRPILTILTLXXPGGQLLGRRCXXX  552


Query  273  RVCAC  277
            RVCAC
Sbjct  553  RVCAC  567
```

[Q3] Gather information about this "novel" protein. At a minimum, show me the protein sequence of the "novel" protein as displayed in your BLAST results from [Q2] as FASTA format (you can copy and paste the aligned sequence subject lines from your BLAST result page if necessary) or translate your novel DNA sequence using a tool called EMBOSS Transeq at the EBI. Don't forget to translate all six reading frames; the ORF (open reading frame) is likely to be the longest sequence without a stop codon. It may not start with a methionine if you don't have the complete coding region. Make sure the sequence you provide includes a header/subject line and is in traditional FASTA format.

Chosen Sequence:

>BM440292.1_1 pgr1n.pk002.l17 Normalized Chicken Reproductive Tract cDNA Library (pgr1n) Gallus gallus cDNA clone pgr1n.pk002.l17 5' similar to sp|P10360|P53_CHICK CELLULAR TUMOR ANTIGEN P53 (TUMOR SUPPRESSOR P53) pir|S02193 cellular tumor antigen p53 - chicken emb|CAA31456.1| (X13057) nuclear protein p53 (AA 1 - 367) [Gallus gallus], mRNA sequence
RAAPSPVVPSTEDYGGDFDFRVGFVEAGTAKSVTCTYSPVLNKVYCRLAKPCPVQVRV
GV
APPPGSSLRAVAVYKKSEHVAEVVRRCPHHERCGGGTDGLAPAQHLIRVEGNPQARYH
DD
ETTKRHSVVVPYEPPEVGSDCTTVLYNFMCNSSCMGGMNRRPILTILTLXXPGGQLLGR
R
CXXXRVCACX

Name: Cellular Tumor Antigen P53

Species: Gallus gallus

> Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
> Archelosauria; Archosauria; Dinosauria; Saurischia; Theropoda;
> Coelurosauria; Aves; Neognathae; Galloanserae; Galliformes;
> Phasianidae; Phasianinae; Gallus.

[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 NR database yielded a top hit result to a protein from Gallus gallus. See additional screen shots below for top hits and selected alignment details:

Enter Query Sequence

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

```
GNPQARYHDD
ETTKRHSVVPYEPPEVGSDCTTVLYNFMCNSSCMGGMNRRPILTILTLXX
PGGQLLGRR
CXXXRVCACX
```

Query subrange ❓
From [   ]
To [   ]

Or, upload file [Choose File] No file chosen ❓

Job Title [BM440292.1_1 pgr1n.pk002.I17 Normalized Chicken...]
Enter a descriptive title for your BLAST search ❓

☐ Align two or more sequences ❓

Choose Search Set

Databases ⦿ Standard databases (nr etc.): 🆕 ◯ 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 [   ] ☐ 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

The top result is to a protein from Gallus gallus (chicken), see second screen shot below for alignment details:

Sequences producing significant alignments

Download ∨   Select columns ∨   Show [100 ▽] ❓

☑ select all 100 sequences selected   GenPept   Graphics   Distance tree of results   Multiple alignment   MSA Viewer

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	cellular tumor antigen p53 [Gallus gallus]	Gallus gallus	378	378	99%	2e-129	97.35%	367	NP_990595.1
☑	cellular tumor antigen p53 isoform X3 [Anas platyrhynchos]	Anas platyrhyn...	335	335	96%	1e-112	87.98%	352	XP_038027747.1
☑	cellular tumor antigen p53 isoform X1 [Anas platyrhynchos]	Anas platyrhyn...	335	335	96%	1e-112	87.98%	365	XP_038027745.1
☑	cellular tumor antigen p53 isoform X2 [Anas platyrhynchos]	Anas platyrhyn...	335	335	96%	1e-112	87.98%	364	XP_038027746.1
☑	cellular tumor antigen p53 isoform X2 [Camarhynchus parvulus]	Camarhynchus...	331	331	98%	7e-112	84.57%	303	XP_030825917.1
☑	cellular tumor antigen p53 isoform X1 [Accipiter gentilis]	Accipiter gentilis	331	331	94%	2e-111	88.89%	319	XP_049648596.1
☑	cellular tumor antigen p53 isoform X2 [Accipiter gentilis]	Accipiter gentilis	330	330	94%	3e-111	88.89%	318	XP_049648597.1

cellular tumor antigen p53 [Gallus gallus]

Sequence ID: NP_990595.1 Length: 367 Number of Matches: 1

See 2 more title(s) ⌄   See all Identical Proteins(IPG)

Range 1: 75 to 263 GenPept Graphics   ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
378 bits(971)	2e-129	Compositional matrix adjust.	184/189(97%)	184/189(97%)	0/189(0%)

```
Query    1    RAAPSPVVPSTEDYGGDFDFRVGFVEAGTAKSVTCTYSPVLNKVYCRLAKPCPVQVRVGV   60
              RAAPSPVVPSTEDYGGDFDFRVGFVEAGTAKSVTCTYSPVLNKVYCRLAKPCPVQVRVGV
Sbjct   75    RAAPSPVVPSTEDYGGDFDFRVGFVEAGTAKSVTCTYSPVLNKVYCRLAKPCPVQVRVGV  134

Query   61    APPPGSSLRAVAVYKKSEHVAEVVRRCPHHERCGGGTDGLAPAQHLIRVEGNPQARYHDD  120
              APPPGSSLRAVAVYKKSEHVAEVVRRCPHHERCGGGTDGLAPAQHLIRVEGNPQARYHDD
Sbjct  135    APPPGSSLRAVAVYKKSEHVAEVVRRCPHHERCGGGTDGLAPAQHLIRVEGNPQARYHDD  194

Query  121    ETTKRHSVVVPYEPPEVGSDCTTVLYNFMCNSSCMGGMNRRPILTILTLXXPGGQLLGRR  180
              ETTKRHSVVVPYEPPEVGSDCTTVLYNFMCNSSCMGGMNRRPILTILTL  PGGQLLGRR
Sbjct  195    ETTKRHSVVVPYEPPEVGSDCTTVLYNFMCNSSCMGGMNRRPILTILTLEGPGGQLLGRR  254

Query  181    CXXXRVCAC  189
              C   RVCAC
Sbjct  255    CFEVRVCAC  263
```

cellular tumor antigen p53 isoform X3 [Anas platyrhynchos]

Sequence ID: XP_038027747.1 Length: 352 Number of Matches: 1

Range 1: 75 to 257 GenPept Graphics   ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
335 bits(859)	1e-112	Compositional matrix adjust.	161/183(88%)	169/183(92%)	0/183(0%)

```
Query    7    VVPSTEDYGGDFDFRVGFVEAGTAKSVTCTYSPVLNKVYCRLAKPCPVQVRVGVAPPPGS   66
              +VPSTEDYGG +DF++GF E GTAKSVTCTYSPVLNK+YCRLAKPCPVQVRVG APPPG+
Sbjct   75    MVPSTEDYGGHYDFQLGFQETGTAKSVTCTYSPVLNKLYCRLAKPCPVQVRVGAAPPPGA  134

Query   67    SLRAVAVYKKSEHVAEVVRRCPHHERCGGGTDGLAPAQHLIRVEGNPQARYHDDETTKRH  126
               LRAVAVYKKSEHVAEVVRRCPHHER G GTDGLAPAQHLIRVEGNPQARYHDDETTKRH
Sbjct  135    VLRAVAVYKKSEHVAEVVRRCPHHERNGEGTDGLAPAQHLIRVEGNPQARYHDDETTKRH  194

Query  127    SVVVPYEPPEVGSDCTTVLYNFMCNSSCMGGMNRRPILTILTLXXPGGQLLGRRCXXXRV  186
```