Find a Gene Assignment

Alex Cagle

2023-04-30

Question 1

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.

> I am interested in studying the huntingtin (Htt) protein, specifically huntingtin isoform 1, which is involved in Huntington's Disease. This protein is found in humans (Homo sapiens) and its accession number from NCBI RefSeq is NP_001375421.1.
>
> Link: https://www.ncbi.nlm.nih.gov/protein/NP_001375421.1

Question 2

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).

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 . 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.

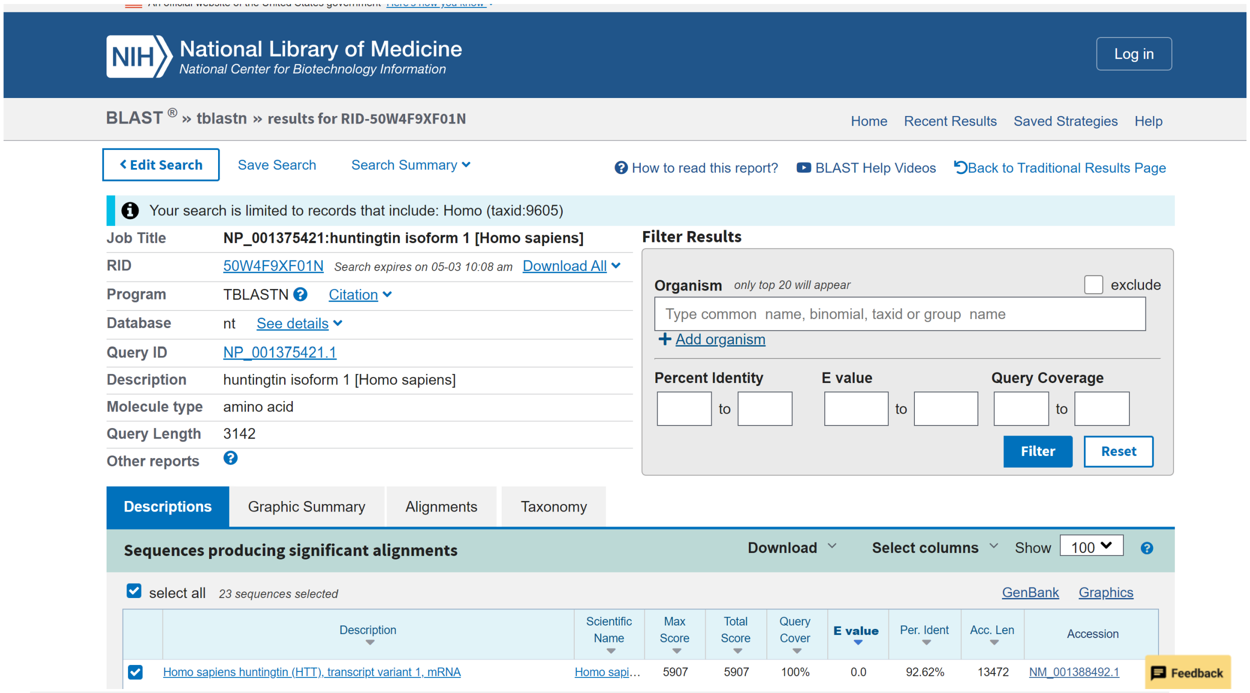
On the BLAST results, clearly indicate a match that represents a protein sequence, encoded from some DNA sequence, that is homologous to your query protein. I need to be able to inspect the pairwise alignment you have selected, including the E value and score. It should be labeled a "genomic clone" or "mRNA sequence", etc. - but include no functional annotation.

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.

> I used NCBI BLAST, specifically tblastn, to find genes relating to the huntingtin isoform 1 protein (NP_001375421.1) with the organism set to Homo (taxid: 9605). The database I searched was the nucleotide collection (nt) database. The top hit was Homo sapiens huntingtin (HTT), transcript

variant 1, mRNA, which corresponds to Homo sapiens (humans), has a query cover of 100%, an E-value of 0.0, a percent identity of 92.62%, and the accession number NM_001388492.1. Screenshot is attached with results from the BLAST search.



An official website of the United States government Here's how you know

National Library of Medicine
National Center for Biotechnology Information

BLAST ® » tblastn » results for RID-50W4F9XF01N

Your search is limited to records that include: Homo (taxid:9605)

Job Title	NP_001375421:huntingtin isoform 1 [Homo sapiens]
RID	50W4F9XF01N *Search expires on 05-03 10:08 am* Download All
Program	TBLASTN Citation
Database	nt See details
Query ID	NP_001375421.1
Description	huntingtin isoform 1 [Homo sapiens]
Molecule type	amino acid
Query Length	3142
Other reports	

Filter Results

Organism *only top 20 will appear* □ exclude

Type common name, binomial, taxid or group name

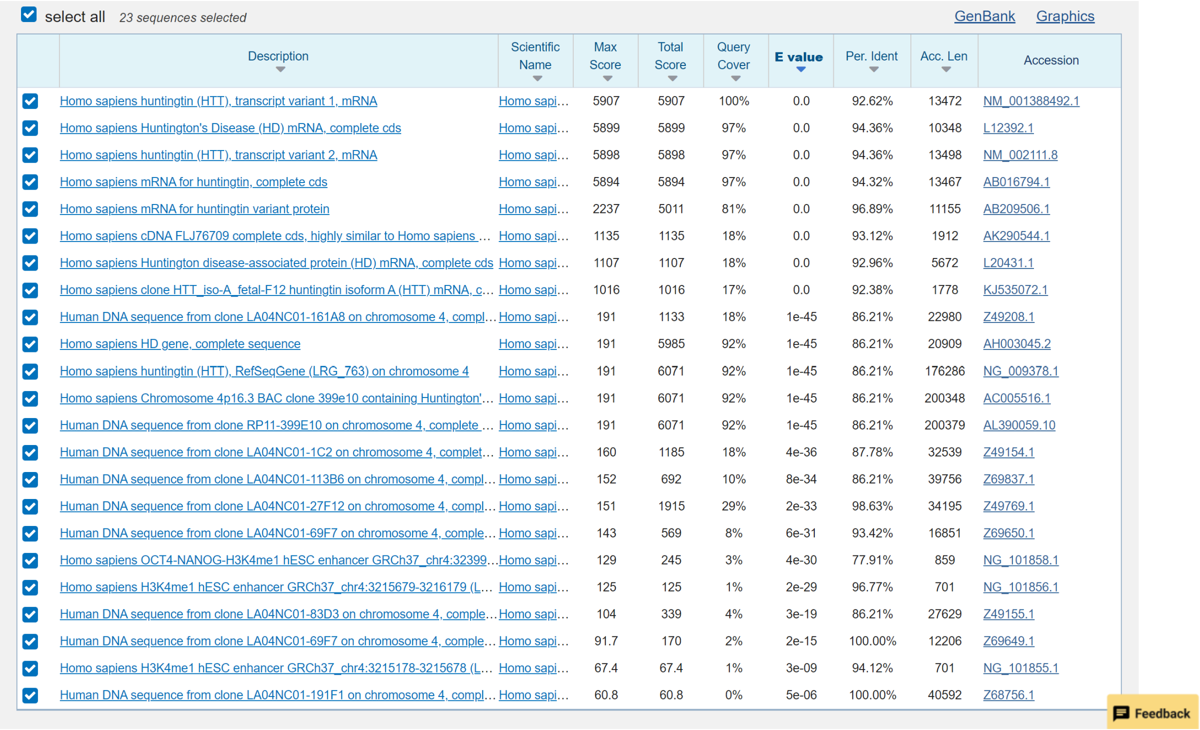
Add organism

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

Sequences producing significant alignments

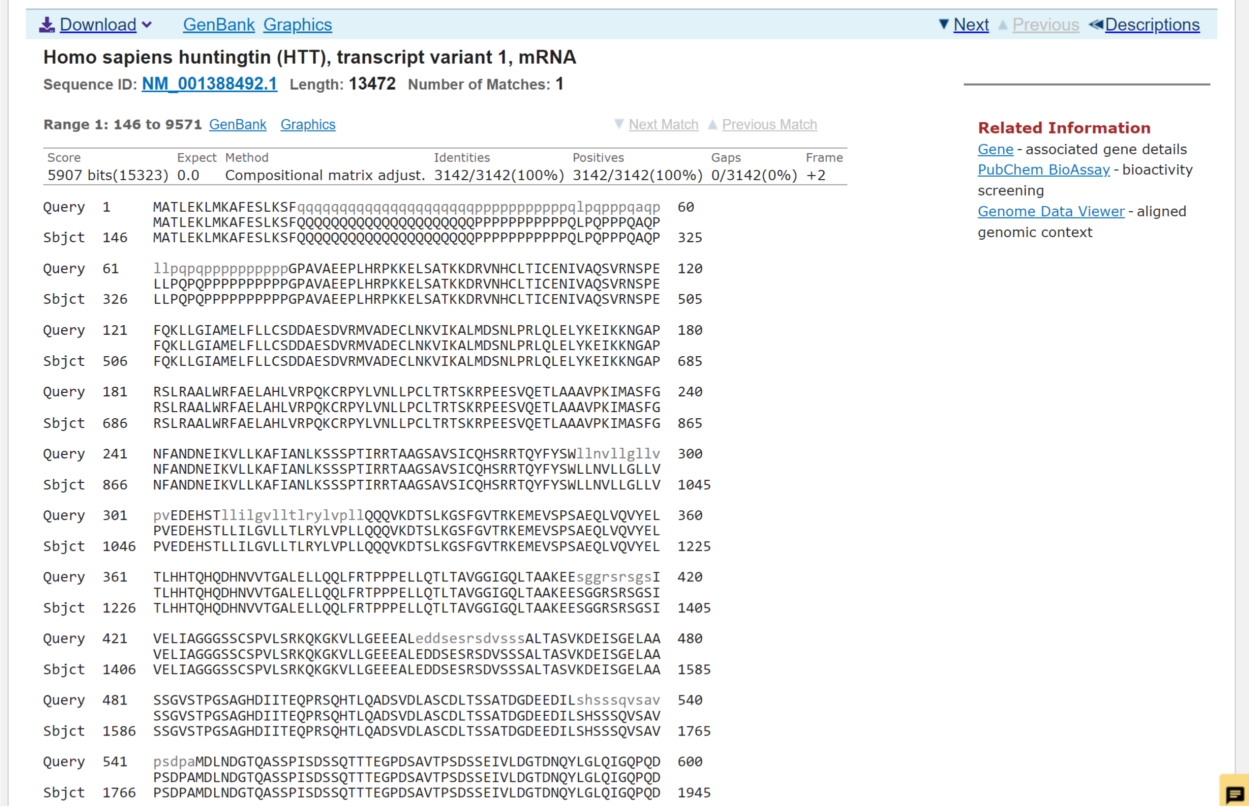
☑ select all *23 sequences selected* GenBank Graphics

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Homo sapiens huntingtin (HTT), transcript variant 1, mRNA	Homo sapi...	5907	5907	100%	0.0	92.62%	13472	NM_001388492.1



☑ select all *23 sequences selected* GenBank Graphics

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Homo sapiens huntingtin (HTT), transcript variant 1, mRNA	Homo sapi...	5907	5907	100%	0.0	92.62%	13472	NM_001388492.1
☑	Homo sapiens Huntington's Disease (HD) mRNA, complete cds	Homo sapi...	5899	5899	97%	0.0	94.36%	10348	L12392.1
☑	Homo sapiens huntingtin (HTT), transcript variant 2, mRNA	Homo sapi...	5898	5898	97%	0.0	94.36%	13498	NM_002111.8
☑	Homo sapiens mRNA for huntingtin, complete cds	Homo sapi...	5894	5894	97%	0.0	94.32%	13467	AB016794.1
☑	Homo sapiens mRNA for huntingtin variant protein	Homo sapi...	2237	5011	81%	0.0	96.89%	11155	AB209506.1
☑	Homo sapiens cDNA FLJ76709 complete cds, highly similar to Homo sapiens ...	Homo sapi...	1135	1135	18%	0.0	93.12%	1912	AK290544.1
☑	Homo sapiens Huntington disease-associated protein (HD) mRNA, complete cds	Homo sapi...	1107	1107	18%	0.0	92.96%	5672	L20431.1
☑	Homo sapiens clone HTT_iso-A_fetal-F12 huntingtin isoform A (HTT) mRNA, c...	Homo sapi...	1016	1016	17%	0.0	92.38%	1778	KJ535072.1
☑	Human DNA sequence from clone LA04NC01-161A8 on chromosome 4, compl...	Homo sapi...	191	1133	18%	1e-45	86.21%	22980	Z49208.1
☑	Homo sapiens HD gene, complete sequence	Homo sapi...	191	5985	92%	1e-45	86.21%	20909	AH003045.2
☑	Homo sapiens huntingtin (HTT), RefSeqGene (LRG_763) on chromosome 4	Homo sapi...	191	6071	92%	1e-45	86.21%	176286	NG_009378.1
☑	Homo sapiens Chromosome 4p16.3 BAC clone 399e10 containing Huntington'...	Homo sapi...	191	6071	92%	1e-45	86.21%	200348	AC005516.1
☑	Human DNA sequence from clone RP11-399E10 on chromosome 4, complete ...	Homo sapi...	191	6071	92%	1e-45	86.21%	200379	AL390059.10
☑	Human DNA sequence from clone LA04NC01-1C2 on chromosome 4, complet...	Homo sapi...	160	1185	18%	4e-36	87.78%	32539	Z49154.1
☑	Human DNA sequence from clone LA04NC01-113B6 on chromosome 4, compl...	Homo sapi...	152	692	10%	8e-34	86.21%	39756	Z69837.1
☑	Human DNA sequence from clone LA04NC01-27F12 on chromosome 4, compl...	Homo sapi...	151	1915	29%	2e-33	98.63%	34195	Z49769.1
☑	Human DNA sequence from clone LA04NC01-69F7 on chromosome 4, comple...	Homo sapi...	143	569	8%	6e-31	93.42%	16851	Z69650.1
☑	Homo sapiens OCT4-NANOG-H3K4me1 hESC enhancer GRCh37_chr4:32399...	Homo sapi...	129	245	3%	4e-30	77.91%	859	NG_101858.1
☑	Homo sapiens H3K4me1 hESC enhancer GRCh37_chr4:3215679-3216179 (L...	Homo sapi...	125	125	1%	2e-29	96.77%	701	NG_101856.1
☑	Human DNA sequence from clone LA04NC01-83D3 on chromosome 4, comple...	Homo sapi...	104	339	4%	3e-19	86.21%	27629	Z49155.1
☑	Human DNA sequence from clone LA04NC01-69F7 on chromosome 4, comple...	Homo sapi...	91.7	170	2%	2e-15	100.00%	12206	Z69649.1
☑	Homo sapiens H3K4me1 hESC enhancer GRCh37_chr4:3215178-3215678 (L...	Homo sapi...	67.4	67.4	1%	3e-09	94.12%	701	NG_101855.1
☑	Human DNA sequence from clone LA04NC01-191F1 on chromosome 4, compl...	Homo sapi...	60.8	60.8	0%	5e-06	100.00%	40592	Z68756.1



Homo sapiens huntingtin (HTT), transcript variant 1, mRNA

Sequence ID: NM_001388492.1 Length: 13472 Number of Matches: 1

Range 1: 146 to 9571 GenBank Graphics

Related Information

Gene - associated gene details

PubChem BioAssay - bioactivity screening

Genome Data Viewer - aligned genomic context

```
Score            Expect Method                       Identities      Positives       Gaps       Frame
5907 bits(15323) 0.0    Compositional matrix adjust. 3142/3142(100%) 3142/3142(100%) 0/3142(0%) +2

Query  1     MATLEKLMKAFESLKSFqqqqqqqqqqqqqqqqqqqqqppppppppppppqlpqpppqaqp  60
             MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPPQLPQPPPQAQP
Sbjct  146   MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPPQLPQPPPQAQP  325

Query  61    llpqpqppppppppppGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE  120
             LLPQPQPPPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE
Sbjct  326   LLPQPQPPPPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPE  505

Query  121   FQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAP  180
             FQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAP
Sbjct  506   FQKLLGIAMELFLLCSDDAESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAP  685

Query  181   RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMASFG  240
             RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMASFG
Sbjct  686   RSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKIMASFG  865

Query  241   NFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHSRRTQYFYSWllnvllgllv  300
             NFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHSRRTQYFYSWLLNVLLGLLV
Sbjct  866   NFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHSRRTQYFYSWLLNVLLGLLV  1045

Query  301   pvEDEHSTllilgvlltlrylvpllQQQVKDTSLKGSFGVTRKEMEVSPSAEQLVQVYEL  360
             PVEDEHSTLLILGVLLTLRYLVPLLQQQVKDTSLKGSFGVTRKEMEVSPSAEQLVQVYEL
Sbjct  1046  PVEDEHSTLLILGVLLTLRYLVPLLQQQVKDTSLKGSFGVTRKEMEVSPSAEQLVQVYEL  1225

Query  361   TLHHTQHQDHNVVTGALELLQQLFRTPPPELLQTLTAVGGIGQLTAAKEEsggrsrsgsI  420
             TLHHTQHQDHNVVTGALELLQQLFRTPPPELLQTLTAVGGIGQLTAAKEESGGRSRSGSI
Sbjct  1226  TLHHTQHQDHNVVTGALELLQQLFRTPPPELLQTLTAVGGIGQLTAAKEESGGRSRSGSI  1405

Query  421   VELIAGGGSSCSPVLSRKQKGKVLLGEEEALeddsesrsdvsssALTASVKDEISGELAA  480
             VELIAGGGSSCSPVLSRKQKGKVLLGEEEALEDDSESRSDVSSSALTASVKDEISGELAA
Sbjct  1406  VELIAGGGSSCSPVLSRKQKGKVLLGEEEALEDDSESRSDVSSSALTASVKDEISGELAA  1585

Query  481   SSGVSTPGSAGHDIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILshsssqvsav  540
             SSGVSTPGSAGHDIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAV
Sbjct  1586  SSGVSTPGSAGHDIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAV  1765

Query  541   psdpaMDLNDGTQASSPISDSSQTTTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQD  600
             PSDPAMDLNDGTQASSPISDSSQTTTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQD
Sbjct  1766  PSDPAMDLNDGTQASSPISDSSQTTTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQD  1945
```

Question 3

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. 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.

The protein I found is from humans and is called huntingtin isoform 1.

FASTA file for the protein I chose:

>NP_001375421.1 huntingtin isoform 1 [Homo sapiens]

MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPPQLPQPPPQAQPLLPQPQPPPPPP

AEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPEFQKLLGIAMELFLLCSDDAESDVRMVADECLNKVI

SNLPRLQLELYKEIKKNGAPRSLRAALWRFAELAHLVRPQKCRPYLVNLLPCLTRTSKRPEESVQETLAAAVPKI

NFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSICQHSRRTQYFYSWLLNVLLGLLVPVEDEHSTLLILGVLLT

LVPLLQQQVKDTSLKGSFGVTRKEMEVSPSAEQLVQVYELTLHHTQHQDHNVVTGALELLQQLFRTPPPELLQT

IGQLTAAKEESGGRSRSGSIVELIAGGGSSCSPVLSRKQKGKVLLGEEEALEDDSESRSDVSSSALTASVKDEISGE

SSGVSTPGSAGHDIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAVPSDPAMDLNDGTQASSP
SSQTTTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQDEDEEATGILPDEASEAFRNSSMALQQAHLLKNMSH
DSSVDKFVLRDEATEPGDQENKPCRIKGDIGQSTDDDSAPLVHCVRLLSASFLLTGGKNVLVPDRDVRVSVKALA
AAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGATAILCGTLICSILSRSRFHVGDWMGTIRTL
TFSLADCIPLLRKTLKDESSVTCKLACTAVRNCVMSLCSSSYSELGLQLIIDVLTLRNSSYWLVRTELLETLAEIDFR
SFLEAKAENLHRGAHHYTGLLKLQERVLNNVVIHLLGDEDPRVRHVAAASLIRLVPKLFYKCDQGQADPVVAVA
YLKLLMHETQPPSHFSVSTITRIYRGYNLLPSITDVTMENNLSRVIAAVSHELITSTTRALTFGCCEALCLLSTAFPV
WSLGWHCGVPPLSASDESRKSCTVGMATMILTLLSSAWFPLDLSAHQDALILAGNLLAASAPKSLRSSWASEEEA
KQEEVWPALGDRALVPMVEQLFSHLLKVINICAHVLDDVAPGPAIKAALPSLTNPPSLSPIRRKGKEKEPGEQASV
KKGSEASAASRQSDTSGPVTTSKSSSLGSFYHLPSYLKLHDVLKATHANYKVTLDLQNSTEKFGGFLRSALDVLS
ATLQDIGKCVEEILGYLKSCFSREPMMATVCVQQLLKTLFGTNLASQFDGLSSNPSKSQGRAQRLGSSSVRPGLY
APYTHFTQALADASLRNMVQAEQENDTSGWFDVLQKVSTQLKTNLTSVTKNRADKNAIHNHIRLFEPLVIKALK
CVQLQKQVLDLLAQLVQLRVNYCLLDSDQVFIGFVLKQFEYIEVGQFRESEAIIPNIFFFLVLLSYERYHSKQIIGIP
IQLCDGIMASGRKAVTHAIPALQPIVHDLFVLRGTNKADAGKELETQKEVVVSMLLRLIQYHQVLEMFILVLQQC
DKWKRLSRQIADIILPMLAKQQMHIDSHEALGVLNTLFEILAPSSLRPVDMLLRSMFVTPNTMASVSTVQLWISG
RVLISQSTEDIVLSRIQELSFSPYLISCTVINRLRDGDSTSTLEEHSEGKQIKNLPEETFSRFLLQLVGILLEDIVTKQL
KVEMSEQQHTFYCQELGTLLMCLIHIFKSGMFRRITAAATRLFRSDGCGGSFYTLDSLNLRARSMITTHPALVLLW
LLVNHTDYRWWAEVQQTPKRHSLSSTKLLSPQMSGEEEDSDLAAKLGMCNREIVRRGALILFCDYVCQNLHDSE
VNHIQDLISLSHEPPVQDFISAVHRNSAASGLFIQAIQSRCENLSTPTMLKKTLQCLEGIHLSQSGAVLTLYVDRLLC
FRVLARMVDILACRRVEMLLAANLQSSMAQLPMEELNRIQEYLQSSGLAQRHQRLYSLLDRFRLSTMQDSLSPSP
PLDGDGHVSLETVSPDKDWYVHLVKSQCWTRSDSALLEGAELVNRIPAEDMNAFMMNSEFNLSLLAPCLSLGMS
KSALFEAAREVTLARVSGTVQQLPAVHHVFQPELPAEPAAYWSKLNDLFGDAALYQSLPTLARALAQYLVVVSK
LPPEKEKDIVKFVVATLEALSWHLIHEQIPLSLDLQAGLDCCCLALQLPGLWSVVSSTEFVTHACSLIYCVHFILEA
QPGEQLLSPERRTNTPKAISEEEEEVDPNTQNPKYITAACEMVAEMVESLQSVLALGHKRNSGVPAFLTPLLRNI
RLPLVNSYTRVPPLVWKLGWSPKPGGDFGTAFPEIPVEFLQEKEVFKEFIYRINTLGWTSRTQFEETWATLLGV
VMEQEESPPEEDTERTQINVLAVQAITSLVLSAMTVPVAGNPAVSCLEQQPRNKPLKALDTRFGRKLSIIRGIVEQ
MVSKRENIATHHLYQAWDPVPSLSPATTGALISHEKLLLQINPERELGSMSYKLGQVSIHSVWLGNSITPLREEEW
EEEADAPAPSSPPTSPVNSRKHRAGVDIHSCSQFLLELYSRWILPSSSARRTPAILISEVVRSLLVVSDLFTERNQFE
YVTLTELRRVHPSEDEILAQYLVPATCKAAVLGMDKAVAEPVSRLLESTLRSSHLPSRVGALHGVLYVLECDLLD
QLIPVISDYLLSNLKGIAHCVNIHSQQHVLVMCATAFYLIENYPLDVGPEFSASIIQMCGVMLSGSEESTPSIIYHCA
GLERLLLSEQLSRLDAESLVKLSVDRVNVHSPHRAMAALGLMLTCMYTGKEKVSPGRTSDPNPAAPDSESVIVAM
LFDRIRKGFPCEARVVARILPQFLDDFFPPQDIMNKVIGEFLSNQQPYPQFMATVVYKVFQTLHSTGQSSMVRDW
SNFTQRAPVAMATWSLSCFFVSASTSPWVAAILPHVISRMGKLEQVDVNLFCLVATDFYRHQIEEELDRRAFQSV
APGSPYHRLLTCLRNVHKVTTC

FASTA file for the gene I found:

>NM_001388492.1 Homo sapiens huntingtin (HTT), transcript variant 1, mRNA
GCTGCCGGGACGGGTCCAAGATGGACGGCCGCTCAGGTTCTGCTTTTACCTGCGGCCCAGAGCCCCATTC
TGCTGAGCGGCGCCGCGAGTCGGCCCGAGGCCTCCGGGGACTGCCGTGCCGGGCGGGAGACCGCCATGG
AAGCTGATGAAGGCCTTCGAGTCCCTCAAGTCCTTCCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGC
GCAGCAGCAGCAGCAACAGCCGCCACCGCCGCCGCCGCCGCCGCCGCCTCCTCAGCTTCCTCAGCCGCCG

AGCCGCTGCTGCCTCAGCCGCAGCCGCCCCCGCCGCCGCCCCCGCCGCCACCCGGCCCGGCTGTGGCTGA

CACCGACCAAAGAAAGAACTTTCAGCTACCAAGAAAGACCGTGTGAATCATTGTCTGACAATATGTGAAAA

ACAGTCTGTCAGAAATTCTCCAGAATTTCAGAAACTTCTGGGCATCGCTATGGAACTTTTTCTGCTGTGCA

CAGAGTCAGATGTCAGGATGGTGGCTGACGAATGCCTCAACAAAGTTATCAAAGCTTTGATGGATTCTAAT

TTACAGCTCGAGCTCTATAAGGAAATTAAAAAGAATGGTGCCCCTCGGAGTTTGCGTGCTGCCCTGTGGAC

GCTGGCTCACCTGGTTCGGCCTCAGAAATGCAGGCCTTACCTGGTGAACCTTCTGCCGTGCCTGACTCGA

GACCCGAAGAATCAGTCCAGGAGACCTTGGCTGCAGCTGTTCCCAAAATTATGGCTTCTTTTGGCAATTTT

AATGAAATTAAGGTTTTGTTAAAGGCCTTCATAGCGAACCTGAAGTCAAGCTCCCCCACCATTCGGCGGAC

ATCAGCAGTGAGCATCTGCCAGCACTCAAGAAGGACACAATATTTCTATAGTTGGCTACTAAATGTGCTCTT

TCGTTCCTGTCGAGGATGAACACTCCACTCTGCTGATTCTTGGCGTGCTGCTCACCCTGAGGTATTTGGTC

CAGCAGCAGGTCAAGGACACAAGCCTGAAAGGCAGCTTCGGAGTGACAAGGAAAGAAATGGAAGTCTCTC

GCAGCTTGTCCAGGTTTATGAACTGACGTTACATCATACACAGCACCAAGACCACAATGTTGTGACCGGAG

TGTTGCAGCAGCTCTTCAGAACGCCTCCACCCGAGCTTCTGCAAACCCTGACCGCAGTCGGGGGCATTGG

GCTGCTAAGGAGGAGTCTGGTGGCCGAAGCCGTAGTGGGAGTATTGTGGAACTTATAGCTGGAGGGGGTT

CCCTGTCCTTTCAAGAAAACAAAAAGGCAAAGTGCTCTTAGGAGAAGAAGAAGCCTTGGAGGATGACTCT

CGGATGTCAGCAGCTCTGCCTTAACAGCCTCAGTGAAGGATGAGATCAGTGGAGAGCTGGCTGCTTCTTC

ACTCCAGGGTCAGCAGGTCATGACATCATCACAGAACAGCCACGGTCACAGCACACACTGCAGGCGGACTC

GGCCAGCTGTGACTTGACAAGCTCTGCCACTGATGGGGATGAGGAGGATATCTTGAGCCACAGCTCCAGC

CCGTCCCATCTGACCCTGCCATGGACCTGAATGATGGGACCCAGGCCTCGTCGCCCATCAGCGACAGCTCC

ACCGAAGGGCCTGATTCAGCTGTTACCCCTTCAGACAGTTCTGAAATTGTGTTAGACGGTACCGACAACCA

CCTGCAGATTGGACAGCCCCAGGATGAAGATGAGGAAGCCACAGGTATTCTTCCTGATGAAGCCTCGGAGC

ACTCTTCCATGGCCCTTCAACAGGCACATTTATTGAAAAACATGAGTCACTGCAGGCAGCCTTCTGACAGC

AAATTTGTGTTGAGAGATGAAGCTACTGAACCGGGTGATCAAGAAAACAAGCCTTGCCGCATCAAAGGTG

GTCCACTGATGATGACTCTGCACCTCTTGTCCATTGTGTCCGCCTTTTATCTGCTTCGTTTTTGCTAACAG

ATGTGCTGGTTCCGGACAGGGATGTGAGGGTCAGCGTGAAGGCCCTGGCCCTCAGCTGTGTGGGAGCAGC

CACCCGGAATCTTTCTTCAGCAAACTCTATAAAGTTCCTCTTGACACCACGGAATACCCTGAGGAACAGTAT

CATCTTGAACTACATCGATCATGGAGACCCACAGGTTCGAGGAGCCACTGCCATTCTCTGTGGGACCCTCA

TCCTCAGCAGGTCCCGCTTCCACGTGGGAGATTGGATGGGCACCATTAGAACCCTCACAGGAAATACATTT

GATTGCATTCCTTTGCTGCGGAAAACACTGAAGGATGAGTCTTCTGTTACTTGCAAGTTAGCTTGTACAGC

CTGTGTCATGAGTCTCTGCAGCAGCAGCTACAGTGAGTTAGGACTGCAGCTGATCATCGATGTGCTGACTC

GTTCCTATTGGCTGGTGAGGACAGAGCTTCTGGAAACCCTTGCAGAGATTGACTTCAGGCTGGTGAGCTT

AAAGCAGAAAACTTACACAGAGGGGCTCATCATTATACAGGGCTTTTAAAACTGCAAGAACGAGTGCTCAA

CATCCATTTGCTTGGAGATGAAGACCCCAGGGTGCGACATGTTGCCGCAGCATCACTAATTAGGCTTGTCC

TTTATAAATGTGACCAAGGACAAGCTGATCCAGTAGTGGCCGTGGCAAGAGATCAAAGCAGTGTTTACCTC

ATGCATGAGACGCAGCCTCCATCTCATTTCTCCGTCAGCACAATAACCAGAATATATAGAGGCTATAACCTA

CATAACAGACGTCACTATGGAAAATAACCTTTCAAGAGTTATTGCAGCAGTTTCTCATGAACTAATCACATC

GAGCACTCACATTTGGATGCTGTGAAGCTTTGTGTCTTCTTTCCACTGCCTTCCCAGTTTGCATTTGGAGT

CACTGTGGAGTGCCTCCACTGAGTGCCTCAGATGAGTCTAGGAAGAGCTGTACCGTTGGGATGGCCACAAT

CCTGCTCTCGTCAGCTTGGTTCCCATTGGATCTCTCAGCCCATCAAGATGCTTTGATTTTGGCCGGAAACT

CCAGTGCTCCCAAATCTCTGAGAAGTTCATGGGCCTCTGAAGAAGAAGCCAACCCAGCAGCCACCAAGCA

TGGCCAGCCCTGGGGGACCGGGCCCTGGTGCCCATGGTGGAGCAGCTCTTCTCTCACCTGCTGAAGGTGA

```
TGCCCACGTCCTGGATGACGTGGCTCCTGGACCCGCAATAAAGGCAGCCTTGCCTTCTCTAACAAACCCCC
GTCCCATCCGACGAAAGGGGAAGGAGAAAGAACCAGGAGAACAAGCATCTGTACCGTTGAGTCCCAAGAA
GCCAGTGCAGCTTCTAGACAATCTGATACCTCAGGTCCTGTTACAACAAGTAAATCCTCATCACTGGGGAG
TCTTCCTTCATACCTCAAACTGCATGATGTCCTGAAAGCTACACACGCTAACTACAAGGTCACGCTGGATCT
GCACGGAAAAGTTTGGAGGGTTTCTCCGCTCAGCCTTGGATGTTCTTTCTCAGATACTAGAGCTGGCCACA
ATTGGGAAGTGTGTTGAAGAGATCCTAGGATACCTGAAATCCTGCTTTAGTCGAGAACCAATGATGGCAAC
TCAACAATTGTTGAAGACTCTCTTTGGCACAAACTTGGCCTCCCAGTTTGATGGCTTATCTTCCAACCCCA
AAGGCCGAGCACAGCGCCTTGGCTCCTCCAGTGTGAGGCCAGGCTTGTACCACTACTGCTTCATGGCCCCC
TTCACCCAGGCCCTCGCTGACGCCAGCCTGAGGAACATGGTGCAGGCGGAGCAGGAGAACGACACCTCGG
TGTCCTCCAGAAAGTGTCTACCCAGTTGAAGACAAACCTCACGAGTGTCACAAAGAACCGTGCAGATAAGA
ATAATCACATTCGTTTGTTTGAACCTCTTGTTATAAAAGCTTTAAAACAGTACACGACTACAACATGTGTGC
AAGCAGGTTTTAGATTTGCTGGCGCAGCTGGTTCAGTTACGGGTTAATTACTGTCTTCTGGATTCAGATCA
TGGCTTTGTATTGAAACAGTTTGAATACATTGAAGTGGGCCAGTTCAGGGAATCAGAGGCAATCATTCCAA
TCTTCTTGGTATTACTATCTTATGAACGCTATCATTCAAAACAGATCATTGGAATTCCTAAAATCATTCAGC
GGCATCATGGCCAGTGGAAGGAAGGCTGTGACACATGCCATACCGGCTCTGCAGCCCATAGTCCACGACCT
AAGAGGAACAAATAAAGCTGATGCAGGAAAAGAGCTTGAAACCCAAAAAGAGGTGGTGGTGTCAATGTTA
TCCAGTACCATCAGGTGTTGGAGATGTTCATTCTTGTCCTGCAGCAGTGCCACAAGGAGAATGAAGACAAC
CTGTCTCGACAGATAGCTGACATCATCCTCCCAATGTTAGCCAAACAGCAGATGCACATTGACTCTCATGA
AGTGTTAAATACATTATTTGAGATTTTGGCCCCTTCCTCCCTCCGTCCGGTAGACATGCTTTTACGGAGTAT
CTCCAAACACAATGGCGTCCGTGAGCACTGTTCAACTGTGGATATCGGGAATTCTGGCCATTTTGAGGGTT
CAGTCAACTGAAGATATTGTTCTTTCTCGTATTCAGGAGCTCTCCTTCTCTCCGTATTTAATCTCCTGTACA
TAGGTTAAGAGATGGGGACAGTACTTCAACGCTAGAAGAACACAGTGAAGGGAAACAAATAAAGAATTTGC
CATTTTCAAGGTTTCTATTACAACTGGTTGGTATTCTTTTAGAAGACATTGTTACAAAACAGCTGAAGGTG
GAGCAGCAACATACTTTCTATTGCCAGGAACTAGGCACACTGCTAATGTGTCTGATCCACATCTTCAAGTCT
CCGGAGAATCACAGCAGCTGCCACTAGGCTGTTCCGCAGTGATGGCTGTGGCGGCAGTTTCTACACCCTGC
ACTTGCGGGCTCGTTCCATGATCACCACCCACCCGGCCCTGGTGCTGCTCTGGTGTCAGATACTGCTGCTT
ACCGACTACCGCTGGTGGGCAGAAGTGCAGCAGACCCCGAAAAGACACAGTCTGTCCAGCACAAAGTTAC
GATGTCTGGAGAAGAGGAGGATTCTGACTTGGCAGCCAAACTTGGAATGTGCAATAGAGAAATAGTACGA
TCATTCTCTTCTGTGATTATGTCTGTCAGAACCTCCATGACTCCGAGCACTTAACGTGGCTCATTGTAAATC
GATCTGATCAGCCTTTCCCACGAGCCTCCAGTACAGGACTTCATCAGTGCCGTTCATCGGAACTCTGCTGC
GTTCATCCAGGCAATTCAGTCTCGTTGTGAAAACCTTTCAACTCCAACCATGCTGAAGAAAACTCTTCAGT
GGATCCATCTCAGCCAGTCGGGAGCTGTGCTCACGCTGTATGTGGACAGGCTTCTGTGCACCCCTTTCCGT
CGCATGGTCGACATCCTTGCTTGTCGCCGGGTAGAAATGCTTCTGGCTGCAAATTTACAGAGCAGCATGGC
AATGGAAGAACTCAACAGAATCCAGGAATACCTTCAGAGCAGCGGGCTCGCTCAGAGACACCAAAGGCTCT
TGGACAGGTTTCGTCTCTCCACCATGCAAGACTCACTTAGTCCCTCTCCTCCAGTCTCTTCCCACCCGCTG
GGGCACGTGTCACTGGAAACAGTGAGTCCGGACAAAGACTGGTACGTTCATCTTGTCAAATCCCAGTGTT
AGATTCTGCACTGCTGGAAGGTGCAGAGCTGGTGAATCGGATTCCTGCTGAAGATATGAATGCCTTCATGA
AGTTCAACCTAAGCCTGCTAGCTCCATGCTTAAGCCTAGGGATGAGTGAAATTTCTGGTGGCCAGAAGAGT
GAAGCAGCCCGTGAGGTGACTCTGGCCCGTGTGAGCGGCACCGTGCAGCAGCTCCCTGCTGTCCATCATG
CGAGCTGCCTGCAGAGCCGGCGGCCTACTGGAGCAAGTTGAATGATCTGTTTGGGGATGCTGCACTGTAT
CCACTCTGGCCCGGGCCCTGGCACAGTACCTGGTGGTGGTCTCCAAACTGCCCAGTCATTTGCACCTTCCT
```

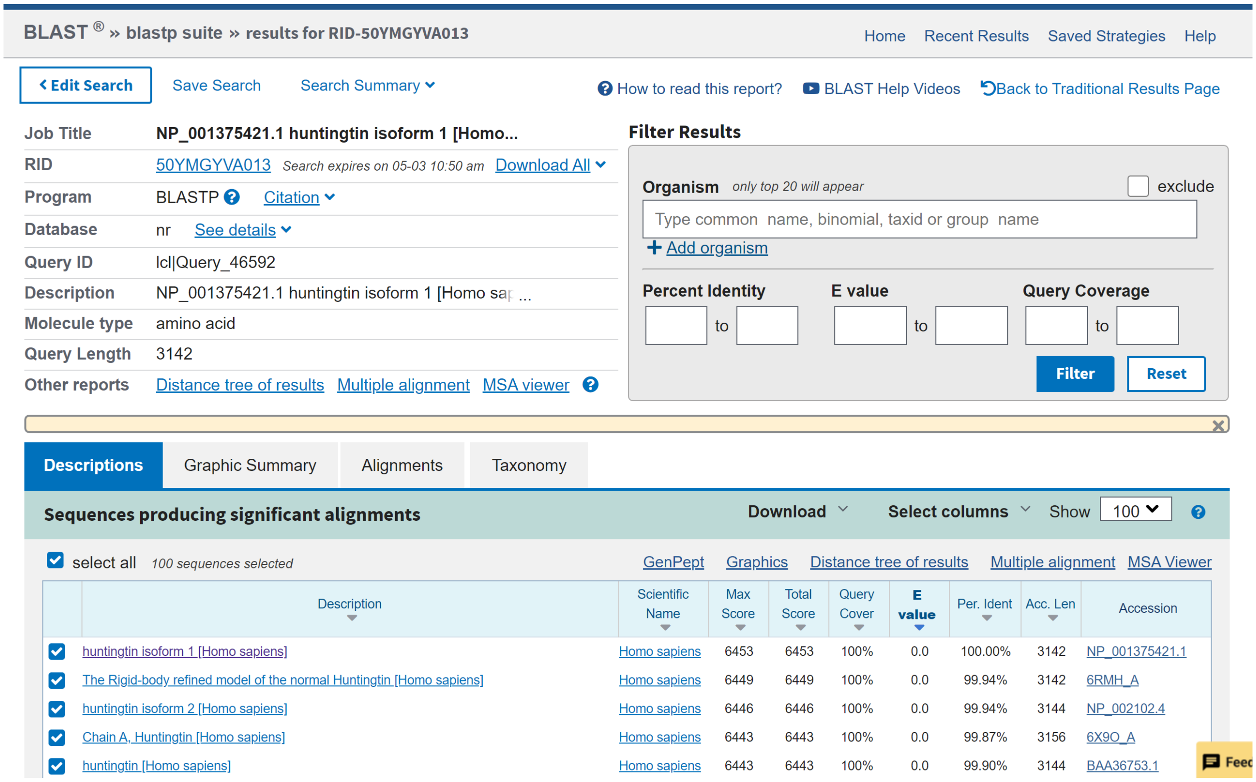
```
GAGAAGGACATTGTGAAATTCGTGGTGGCAACCCTTGAGGCCCTGTCCTGGCATTTGATCCATGAGCAGAT
TCTGGATCTCCAGGCAGGGCTGGACTGCTGCTGCCTGGCCCTGCAGCTGCCTGGCCTCTGGAGCGTGGTC
AGTTTGTGACCCACGCCTGCTCCCTCATCTACTGTGTGCACTTCATCCTGGAGGCCGTTGCAGTGCAGCCT
CTTCTTAGTCCAGAAAGAAGGACAAATACCCCAAAAGCCATCAGCGAGGAGGAGGAGGAAGTAGATCCAA
TCCTAAGTATATCACTGCAGCCTGTGAGATGGTGGCAGAAATGGTGGAGTCTCTGCAGTCGGTGTTGGCCT
AAAGGAATAGCGGCGTGCCGGCGTTTCTCACGCCATTGCTAAGGAACATCATCATCAGCCTGGCCCGCCTC
AACAGCTACACACGTGTGCCCCCACTGGTGTGGAAGCTTGGATGGTCACCCAAACCGGGAGGGGATTTTG
CCCTGAGATCCCCGTGGAGTTCCTCCAGGAAAAGGAAGTCTTTAAGGAGTTCATCTACCGCATCAACACAC
CCAGTCGTACTCAGTTTGAAGAAACTTGGGCCACCCTCCTTGGTGTCCTGGTGACGCAGCCCCTCGTGATC
GAGAGCCCACCAGAAGAAGACACAGAGAGGACCCAGATCAACGTCCTGGCCGTGCAGGCCATCACCTCAC
TGCAATGACTGTGCCTGTGGCCGGCAACCCAGCTGTAAGCTGCTTGGAGCAGCAGCCCCGGAACAAGCCT
TCGACACCAGGTTTGGGAGGAAGCTGAGCATTATCAGAGGGATTGTGGAGCAAGAGATTCAAGCAATGGT
GAGAATATTGCCACCCATCATTTATATCAGGCATGGGATCCTGTCCCTTCTCTGTCTCCGGCTACTACAGGT
CAGCCACGAGAAGCTGCTGCTACAGATCAACCCCGAGCGGGAGCTGGGGAGCATGAGCTACAAACTCGGC
TACACTCCGTGTGGCTGGGGAACAGCATCACACCCCTGAGGGAGGAGGAATGGACGAGGAAGAGGAGGA
GCCCCTGCACCTTCGTCACCACCCACGTCTCCAGTCAACTCCAGGAAACACCGGGCTGGAGTTGACATCCA
GCAGTTTTTGCTTGAGTTGTACAGCCGCTGGATCCTGCCGTCCAGCTCAGCCAGGAGGACCCCGGCCATC
AGGTGGTCAGATCCCTTCTAGTGGTCTCAGACTTGTTCACCGAGCGCAACCAGTTTGAGCTGATGTATGTC
GAACTGCGAAGGGTGCACCCTTCAGAAGACGAGATCCTCGCTCAGTACCTGGTGCCTGCCACCTGCAAGG
CCTTGGGATGGACAAGGCCGTGGCGGAGCCTGTCAGCCGCCTGCTGGAGAGCACGCTCAGGAGCAGCCAC
GGGTTGGAGCCCTGCACGGCGTCCTCTATGTGCTGGAGTGCGACCTGCTGGACGACACTGCCAAGCAGCT
ATCAGCGACTATCTCCTCTCCAACCTGAAAGGGATCGCCCACTGCGTGAACATTCACAGCCAGCAGCACGT
GTGTGCCACTGCGTTTTACCTCATTGAGAACTATCCTCTGGACGTAGGGCCGGAATTTTCAGCATCAATAA
GTGGGGTGATGCTGTCTGGAAGTGAGGAGTCCACCCCCTCCATCATTTACCACTGTGCCCTCAGAGGCCTC
CTGCTCTCTGAGCAGCTCTCCCGCCTGGATGCAGAATCGCTGGTCAAGCTGAGTGTGGACAGAGTGAACG
GCACCGGGCCATGGCGGCTCTGGGCCTGATGCTCACCTGCATGTACACAGGAAAGGAGAAAGTCAGTCCG
CAGACCCTAATCCTGCAGCCCCCGACAGCGAGTCAGTGATTGTTGCTATGGAGCGGGTATCTGTTCTTTTT
AGGAAAGGCTTTCCTTGTGAAGCCAGAGTGGTGGCCAGGATCCTGCCCCAGTTTCTAGACGACTTCTTCC
CATCATGAACAAAGTCATCGGAGAGTTTCTGTCCAACCAGCAGCCATACCCCCAGTTCATGGCCACCGTGG
TGTTTCAGACTCTGCACAGCACCGGGCAGTCGTCCATGGTCCGGGACTGGGTCATGCTGTCCCTCTCCAA
AGGGCCCCGGTCGCCATGGCCACGTGGAGCCTCTCCTGCTTCTTTGTCAGCGCGTCCACCAGCCCGTGGG
CCTCCCACATGTCATCAGCAGGATGGGCAAGCTGGAGCAGGTGGACGTGAACCTTTTCTGCCTGGTCGCC
ACAGACACCAGATAGAGGAGGAGCTCGACCGCAGGGCCTTCCAGTCTGTGCTTGAGGTGGTTGCAGCCCC
TATCACCGGCTGCTGACTTGTTTACGAAATGTCCACAAGGTCACCACCTGCTGAGCGCCATGGTGGGAGAC
CGGCAGCTGGGGCCGGAGCCTTTGGAAGTCTGCGCCCTTGTGCCCTGCCTCCACCGAGCCAGCTTGGTCC
CCGCACATGCCGCGGGCGGCCAGGCAACGTGCGTGTCTCTGCCATGTGGCAGAAGTGCTCTTTGTGGCAC
GGGAGTGTCTGCAGTCCTGGTGGGGCTGAGCCTGAGGCCTTCCAGAAAGCAGGAGCAGCTGTGCTGCACC
GACCAGGTCCTTTCTCCTGATAGTCACCTGCTGGTTGTTGCCAGGTTGCAGCTGCTCTTGCATCTGGGCCA
CCTCCTGCAGGCTGGCTGTTGGCCCCTCTGCTGTCCTGCAGTAGAAGGTGCCGTGAGCAGGCTTTGGGAA
GGTCTCCCTGGTGGGGTGTGCATGCCACGCCCCGTGTCTGGATGCACAGATGCCATGGCCTGTGCTGGGC
GGGTGCTAGACACCCGGCACCATTCTCCCTTCTCTCTTTTCTTCTCAGGATTTAAAATTTAATTATATCAGT
```

```
TAATTTTAACGTAACTCTTTCTATGCCCGTGTAAAGTATGTGAATCGCAAGGCCTGTGCTGCATGCGACAG
TGGTGGACAGGGCCCCCGGCCACGCTCCCTCTCCTGTAGCCACTGGCATAGCCCTCCTGAGCACCCGCTGA
TGTACATGTTCCTGTTTATGCATTCACAAGGTGACTGGGATGTAGAGAGGCGTTAGTGGGCAGGTGGCCAC
GAGGACAGGCCCCCATTATCCTAGGGGTGCGCTCACCTGCAGCCCCTCCTCCTCGGGCACAGACGACTGTC
CCACCAGTCAGGGACAGCAGCCTCCCTGTCACTCAGCTGAGAAGGCCAGCCCTCCCTGGCTGTGAGCAGC
TCCAGAGACATGGGCCTCCCACTCCTGTTCCTTGCTAGCCCTGGGGTGGCGTCTGCCTAGGAGCTGGCTG
GGACCTGCTGCTCCATGGATGCATGCCCTAAGAGTGTCACTGAGCTGTGTTTTGTCTGAGCCTCTCTCGGT
AGCTTGGTGTCTTGGCACTGTTAGTGACAGAGCCCAGCATCCCTTCTGCCCCCGTTCCAGCTGACATCTTC
CCCTTTTAGTCAGGAGAGTGCAGATCTGTGCTCATCGGAGACTGCCCCACGGCCCTGTCAGAGCCGCCACT
AGGCCAGGTCCCTGGACCAGCCTCCTGTTTGCAGGCCCAGAGGAGCCAAGTCATTAAAATGGAAGTGGAT
CGGGCTGCTGCTGATGTAGGAGCTGGATTTGGGAGCTCTGCTTGCCGACTGGCTGTGAGACGAGGCAGGC
CTCAGCCCTAGAGGCGAGCCAGGCAAGGTTGGCGACTGTCATGTGGCTTGGTTTGGTCATGCCCGTCGAT
TTGAATGTGGTAAGTGGAGGAAATGTTGGAACTCTGTGCAGGTGCTGCCTTGAGACCCCCAAGCTTCCAC
CCTATGTGGCAGCTGGGGAGCAGCTGAGATGTGGACTTGTATGCTGCCCACATACGTGAGGGGGAGCTGA
CTCCTCTGAGCAGCCTCTGCCAGGCCTGTATGAGGCTTTTCCCACCAGCTCCCAACAGAGGCCTCCCCCAC
CCTCGTCCTCGTGGCGGGGCAGCAGGAGCGGTAGAAAGGGGTCCGATGTTTGAGGAGGCCCTTAAGGGA
TATAACACGTAAGAAAATCACCATTCCGTATTGGTTGGGGGCTCCTGTTTCTCATCCTAGCTTTTTCCTGG
TAGAAGGTTTGGGAACGAGGGGAAAGTTCTCAGAACTGTTGGCTGCTCCCCACCCGCCTCCCGCCTCCCC
GTCAGCAGCTCTGAGACAGCAGTATCACAGGCCAGATGTTGTTCCTGGCTAGATGTTTACATTTGTAAGAA
TGAATGTAAAACAGAGCCATTCCCTTGGAATGCATATCGCTGGGCTCAACATAGAGTTTGTCTTCCTCTTG
TGATCTAAACCAGTCCTTAGCAAGGGGCTCAGAACACCCCGCTCTGGCAGTAGGTGTCCCCCACCCCCAAA
GTGTGCTCCGGAGATGAATATGAGCTCATTAGTAAAAATGACTTCACCCACGCATATACATAAAGTATCCAT
ATATAGACACATCTATAATTTTACACACACACCTCTCAAGACGGAGATGCATGGCCTCTAAGAGTGCCCGTG
TCCTGGAAGTTGACTTTCCTTAGACCCGCCAGGTCAAGTTAGCCGCGTGACGGACATCCAGGCGTGGGAC
CAGGGCTCATTCATTGCCCACTAGGATCCCACTGGCGAAGATGGTCTCCATATCAGCTCTCTGCAGAAGGG
TTATCATGTTCCTAAAAATCTGTGGCAAGCACCCATCGTATTATCCAAATTTTGTTGCAAATGTGATTAATT
AAGTTTTGGGGGTGGGCTGTGGGGAGATTGCTTTTGTTTTCCTGCTGGTAATATCGGGAAAGATTTTAATG
TAGAATTGTTTGGCAATGCACTGAAGCGTGTTTCTTTCCCAAAATGTGCCTCCCTTCCGCTGCGGGCCCAC
TGTAGGTGATGTTTCCAGCTGCCAAGTGCTCTTTGTTACTGTCCACCCTCATTTCTGCCAGCGCATGTGTC
GGAAAATGTGAAGCTGAACCCCCTCCAGACACCCAGAATGTAGCATCTGAGAAGGCCCTGTGCCCTAAAGC
GCCCCCATCTTCATGGAGGGGGTCATTTCAGAGCCCTCGGAGCCAATGAACAGCTCCTCCTCTTGGAGCTG
CCACGTGGAGCTCGGGACGGATAGTAGACAGCAATAACTCGGTGTGTGGCCGCCTGGCAGGTGGAACTTC
GGGGTGGAGTGAGGTTAGTTCTGTGTGTCTGGTGGGTGGAGTCAGGCTTCTCTTGCTACCTGTGAGCATC
GACATCCTCATCGGGCTTTGTCCCTCCCCCGCTTCCTCCCTCTGCGGGGAGGACCCGGGACCACAGCTGCT
AGACTTGGAGCTGTCCTCCAGAGGGGTCACGTGTAGGAGTGAGAAGAAGGAAGATCTTGAGAGCTGCTGA
AGAGCTCAGGATGGCTCAGACGAGGACACTCGCTTGCCGGGCCTGGGCCTCCTGGGAAGGAGGGAGCTGC
GCATGACAACTGAAGGCAACCTGGAAGGTTCAGGGGCCGCTCTTCCCCCATGTGCCTGTCACGCTCTGGT
GAACGCCTTCCCCTCAGTTGTTTCTAAGAGCAGAGTCTCCCGCTGCAATCTGGGTGGTAACTGCCAGCCTT
TGGCCAACGTGGACCTGCCTACGGAGGGTGGGCTCTGACCCAAGTGGGGCCTCCTTGTCCAGGTCTCACT
GTGGTCAGAGGGACTGTCAGCTGAGCTTGAGCTCCCCTGGAGCCAGCAGGGCTGTGATGGGCGAGTCCCC
CAGACCTGAATGCTTCTGAGAGCAAAGGGAAGGACTGACGAGAGATGTATATTTAATTTTTTAACTGCTGC
ACATCCAAATTAAAGGAAAAAAATGGAAACCA
```

Question 4

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.

> There are other proteins with percent identities less than 100%, but the top hit is 100%, since I used that exact sequence as the blastp query. Therefore, my protein is not novel?



Job Title	NP_001375421.1 huntingtin isoform 1 [Homo...
RID	50YMGYVA013 *Search expires on 05-03 10:50 am* Download All ⌄
Program	BLASTP Citation ⌄
Database	nr See details ⌄
Query ID	lcl|Query_46592
Description	NP_001375421.1 huntingtin isoform 1 [Homo sa ...
Molecule type	amino acid
Query Length	3142
Other reports	Distance tree of results Multiple alignment MSA viewer

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	huntingtin isoform 1 [Homo sapiens]	Homo sapiens	6453	6453	100%	0.0	100.00%	3142	NP_001375421.1
☑	The Rigid-body refined model of the normal Huntingtin [Homo sapiens]	Homo sapiens	6449	6449	100%	0.0	99.94%	3142	6RMH_A
☑	huntingtin isoform 2 [Homo sapiens]	Homo sapiens	6446	6446	100%	0.0	99.94%	3144	NP_002102.4
☑	Chain A, Huntingtin [Homo sapiens]	Homo sapiens	6443	6443	100%	0.0	99.87%	3156	6X9O_A
☑	huntingtin [Homo sapiens]	Homo sapiens	6443	6443	100%	0.0	99.90%	3144	BAA36753.1

Figure 1: Q4 search results