
Master Biomedizin 2016

Genomics
Homology

1

- a. Get the fasta sequence of the human (*Homo sapiens*) protein P53 from UniProt (<http://www.uniprot.org/>). Which one of all the isoforms should you download?
- b. Find the P53 protein from mouse (*Mus musculus*). As you see, there are more than one entry for mouse. Which UniProt entry should you select?
- c. BLAT the human P53 using “hg19” as database (in UCSC, <http://genome.ucsc.edu/>), and answer:
How many amino acids has the query sequence?
And how many nucleotides?
Is it a perfect alignment?
Which is the genomic locus of the target?
- d. Visualize and navigate through the P53 genome region, and answer:
Which genes are around?
How many exons does it have?
- e. BLAT the mouse P53 against the human genome “hg19”. What do you observe?



Human
(*Homo sapiens*)



Mouse
(*Mus musculus*)

1

UniProtKB p53 human

Advanced Search

Help Contact

About UniProtKB Basket

BLAST Align Download Add to basket Columns

1 to 25 of 5,151 Show 25

Entry	Entry name	Gene names	Organism	Length
<input checked="" type="checkbox"/>	P04637	P53_HUMAN	Homo sapiens (Human)	393
<input type="checkbox"/>	P02340	P53_MOUSE	Mus musculus (Mouse)	387
<input type="checkbox"/>	Q00987	MDM2_HUMAN	Homo sapiens (Human)	491

Download selected (1)
Download all (5151)

Format: FASTA (canonical)
Compressed Uncompressed

Preview first 10

Go

a. P04637 (P53_HUMAN). The canonical.

UniProtKB p53 mouse

Advanced Search

Help Contact

About UniProtKB Basket

BLAST Align Download Add to basket Columns

1 to 25 of 839 Show 25

Entry	Entry name	Gene names	Organism	Length
<input checked="" type="checkbox"/>	P02340	P53_MOUSE	Mus musculus (Mouse)	387
<input type="checkbox"/>	P23804	MDM2_MOUSE	Mus musculus (Mouse)	489

Download selected (1)
Download all (839)

Format: FASTA (canonical)
Compressed Uncompressed

Preview first 10

Go

b. P02340 (P53_MOUSE).

1

BLAT Search Genome

Genome: Assembly: Query type: Sort output: Output type:

```
>sp|P04637|P53_HUMAN Cellular tumor antigen p53 OS=Homo sapiens GN=TP53 PE=1 SV=4
MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGP
DEAPRMPEAAPVAPAPAAPTAAAPAPAPSWPLSSVPSQKTYQGSYGFRLGFLHSGTAK
SVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGRVRAMAIYKQSQHMTVEVRRCPHHE
RCSDSDGLAPPQHLIRVEGNLRVEYLDLDRNTFRHSVVPYEPPEVGSDDCTTIHYNMNCNS
SCMGGMNRRPILTIITLEDSSGNLLGRNSFEVRCACPGRRDRTEENLRKKGEPHELPL
PGSTKRALPNNTSSSPQPKKPLDGEYFTLQIRGRERFEMFRELNEALELKDAQAGKEPG
GSAHSHSLKSKKGQSTRHKKLMFKTEGPDSD
```

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	P53_HUMAN	1149	1	393	393	100.0%	17	+-	7572930	7579912	6983

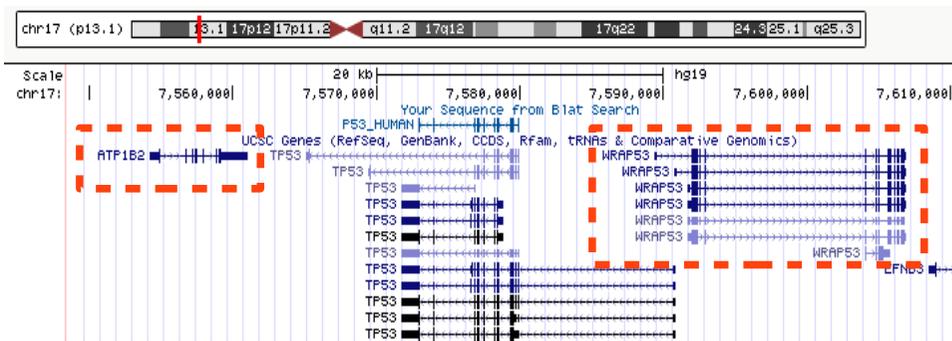
c. 393 amino acids
393*3 = 1179 nucleotides

Not a perfect alignment
("lpennvl")

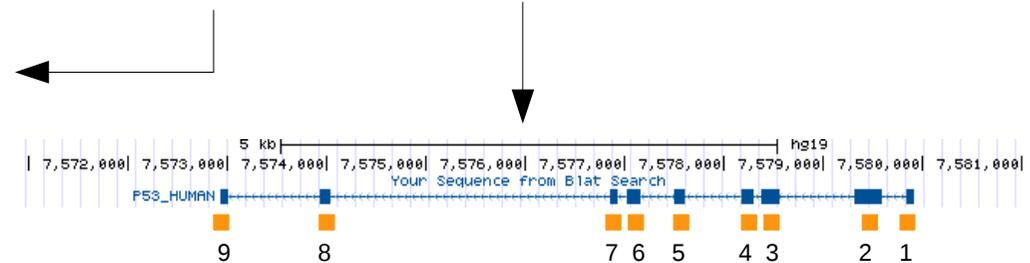
chr17
7572930-7579912

P53_HUMAN

```
MEEPQSDPSV EPPLSQETFSDLWKLLpenn vLSPLPSQAM DDLMLSPDDI EQWFTEDPGP 60
DEAPRMPEAA PPVAPAPAAP TPAAPAPAPS WPLSSVPSQ KTYQGSYGFRLGFLHSGTAK 120
SVTCTYSPAL NKMFCQLAKT CPVQLWVDST PPPGRVRAM AIYKQSQHMT EVVRRCPHHE 180
RCSDSDGLAP PQHLIRVEGN LRVEYLDLDRN TFRHSVVPY EPPEVGSDDCT TIHYNMNCNS 240
SCMGGMNRRP ILTIITLEDSSGNLLGRNSFEVRCACPGRRDRTEENLRKKGEPHELPL 300
PGSTKRALPN NTSSSPQPKK PLDGEYFTLQIRGRERFEM FRELNEALEL KDAQAGKEPG 360
GSAHSHSLK SKKGQSTRH KKLMFKTEGP DSD
```



d. ATP1B2 and WRAP53. 9 exons (9 blocks).



BLAT Search Genome

Genome: Assembly: Query type: Sort output: Output type:

```
>sp|P02340|P53_MOUSE Cellular tumor antigen p53 OS=Mus musculus GN=TP53 PE=1 SV=3
MEESQSDISLELPLSQETFSGLWKLLPPEDILPSPHCMDLLLPQDVEEFEGPSEALRV
SGAPAAQDPVTEPGVPAPAPATPWPPLSSVPSQKTYQGNVGFHGLGFLQSGTAKSVMTCTY
SPPLNKLFCQLAKTCPVQLWVSAATPPAGSRVRAMAIYKQSQHMTVEVRRCPHHERCSDGD
GLAPPQHLIRVEGNLRYPEYLEDROTFRHSVVPYEPPEAGSEYTTIHYKMCNNSCMGGM
NRRPILTIITLEDSSGNLLGRDSFEVRCACPGRRDRTEENFRKKEVLCPELPPGSAKR
ALPTCTASPPQPKKPLDGEYFTLQIRGRERFEMFRELNEALELKAHAHTEESGDSRAHS
SYLTKKKGQSTRHKKTMVKVGPDS
```

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	P53_MOUSE	592	74	387	387	84.3%	17	+-	7572930	7579449	6520

e. The result is worse (84.3%).

P53_MOUSE

```
meesqsdisl elplsqtetfs glwklpped ilpsphcmdd lllpqdveef fegpsealrv 60
sgapaaqdpv tetPqVAPA PATpWPLSSf VPSQKTYQGN YGFhLGLqS GTAKSVmCTY 120
SPPLNKLFCQ LAKTCPVQLW vsaTPPaGsR VRAMAIYkks QHMTVEVRRCPHHERCSDGD 180
GLAPPQHLIR VEGNlypEYL eDRqTFRHSV VVPYEPPEaG SeyTTIHYKy MCNNSCMGGM 240
NRRPILTIIT LEDSSGNLLG RdsFEVRCAC PGRRDRTEE ENFRKKEVlc pELPPGSAKR 300
alptctasap pqkkkpldge yftLkIRGrk RFEMFRELNE ALELKAHAh eESGdsRAHS 360
SyLkTKKKGQs TSRHKKTMvk kvGPDS
```

2

- a. How many “Apoptosis inhibitor 5” (api5) proteins are there in human (*Homo sapiens*)? Use UniProt.
- b. And how many UniProt entries?

2

UniProtKB name:"apoptosis inhibitor 5" organism:human

UniProtKB results

Filter by:

- Reviewed (1) Swiss-Prot
- Unreviewed (4) TrEMBL

Popular organisms: Human (5)

Search terms: Filter "human" as: organism

Entry	Entry name	Protein names	Gene names	Organism	Length
Q9BZZ5	API5_HUMAN	Apoptosis inhibitor 5	API5 MIG8	Homo sapiens (Human)	524
G3V1C3	G3V1C3_HUMAN	Apoptosis inhibitor 5	API5	Homo sapiens (Human)	510
E9PQK6	E9PQK6_HUMAN	Apoptosis inhibitor 5	API5	Homo sapiens (Human)	123
H0YER7	H0YER7_HUMAN	Apoptosis inhibitor 5	API5	Homo sapiens (Human)	294
B4DDR3	B4DDR3_HUMAN	cDNA FLJ52148, highly similar to Ap...		Homo sapiens (Human)	331

a. One protein (SwissProt): Q9BZZ5.

b. At least four UniProt entries.

```

Q9BZZ5 API5 HUMAN 1 _____ 524
G3V1C3 G3V1C3 HUMAN 1 _____ 510
H0YER7 H0YER7 HUMAN 1 _____ 294
E9PQK6 E9PQK6_HUMAN 1 _____ 123
    
```

3

- a. Using the human protein “P21741”, find its orthologous proteins in frog (*Xenopus laevis*) and get their UniProt AC.
- b. Check the identity between the orthologs (human – frog proteins).
- c. Check the identity between the paralogs (frog – frog proteins).



Human
(*Homo sapiens*)



Frog
(*Xenopus laevis*)

3

View by

Taxonomy

Text version

XML version

Taxonomy view

Search:

Xenopus laevis (African clawed frog) (16876 hits)

Entry	Alignment overview	Info	Status
<input type="checkbox"/> Query: sp P21741 MK_HUMAN B201603049BY1TV3S27			
<input type="checkbox"/> P48530 MKA_XENLA - Midkine-A - Xenopus laevis ... - View alignment		E-value: 430E-66 Score: 510 Ident.: 61.1%	
<input type="checkbox"/> P48531 MKB_XENLA - Midkine-B - Xenopus laevis ... - View alignment		E-value: 600E-66 Score: 509 Ident.: 60.4%	

a. Query: P21741.
Ortholog1: P48530.
Ortholog2: P48531.

b. P21741-P48530 = 61.1%
P21741-P48531 = 60.4%

c. P48530-P48531 = 97.9%
Note: may also be done with “alignments”.

Entry	Protein names	Match hit	Identity
P48530	Midkine-A (Xenopus laevis)		100.0%
Q6P8F3	Midkine (Xenopus tropicalis)		98.6%
P48531	Midkine-B (Xenopus laevis)		97.9%

Homology

4

- a. Using the protein “Q90486”, find its paralog/s using NCBI BLAST.
- b. Check the identity between each paralog and the query.

4

a. Query + four paralogs.



	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input checked="" type="checkbox"/>	hemoglobin subunit beta-1 [Danio rerio]	312	312	100%	6e-113	100%	NP_571095.1
<input type="checkbox"/>	Ba1 globin [Danio rerio]	311	311	100%	2e-112	99%	AAI15159.1
<input checked="" type="checkbox"/>	hemoglobin subunit beta-2 [Danio rerio]	310	310	100%	4e-112	98%	NP_001005403.1
<input type="checkbox"/>	Ba2 globin [Danio rerio]	308	308	100%	2e-111	97%	AAH53176.1
<input type="checkbox"/>	beta globin [Danio rerio]	224	224	70%	1e-78	100%	AAB05405.1
<input type="checkbox"/>	novel protein similar to zebrafish ba2 globin (ba2) [Danio rerio]	222	222	70%	5e-78	98%	CAE30439.1
<input checked="" type="checkbox"/>	hemoglobin beta embryonic-1.1 [Danio rerio]	216	216	100%	3e-75	69%	NP_932339.1
<input checked="" type="checkbox"/>	hemoglobin beta embryonic-2 [Danio rerio]	214	214	100%	3e-74	68%	NP_998011.1
<input type="checkbox"/>	PREDICTED: uncharacterized protein LOC445037 [Danio rerio]	204	204	99%	2e-70	61%	XP_005164394.1
<input type="checkbox"/>	uncharacterized protein LOC445037 [Danio rerio]	203	203	99%	4e-70	61%	NP_001003431.1
<input checked="" type="checkbox"/>	hemoglobin beta embryonic-3 [Danio rerio]	194	194	99%	2e-66	61%	NP_001015058.1

b. Hemoglobin subunit beta-1 (query) to:
 Subunit beta-2 = 98%
 Beta embryonic-1.1 = 69%
 Beta embryonic-2 = 68%
 Beta embryonic-3 = 61%

We consider only the RefSeq protein entries "NP".

5

- a. Based on the sequence of the “ATP synthase subunit a” protein from the extinct mammoth (*Mammuthus primigenius*), was the mammoth closer to the asian elephant (*Elephas maximus*) or to the african elephant (*Loxodonta africana*)? Use only SwissProt proteins.
- b. Is there evidence enough to conclude if they are / are not closer?
- c. Could you check with the “cytochrome b” protein too? Use only SwissProt proteins.



Woolly mammoth
(*Mammuthus primigenius*)



Asian elephant
(*Elephas maximus*)



African elephant
(*Loxodonta africana*)

5

UniProtKB

BLAST Align Retrieve/ID mapping

BLAST

```
>sp|Q38PR7|ATP6_MAMPR ATP synthase subunit a OS=Mammuthus primigenius GN=MT-ATP6 PE=3 SV=1
MNEELSAFFDVPVGTMMMLAIAFPAILLPTPNRLITNRWITIQWLKLVKLIKQKLLSIHNTK
GLSWSMLITLTLFIGLTLNLLGLLPYFAPTAQLTVNLSMAIPLWTGTIVLGFYKTKIS
LAHLLPQGTPTFLIPMIIIIETISLLIRPVTLAVRLTANITAGHLLIHLTGTAALLLSI
HSMITITVTFITVWVLTILELAVALIQAYVFALLISLYLHESA
```

Target database¹ UniProtKB/Swiss-Prot E-Threshold¹ 10 Matrix¹ Auto Filtering¹ None Gapped¹ yes

Run Blast in a separate window.

Clear Run BLAST

UniProtKB

BLAST Align Retrieve/ID mapping

BLAST

```
>sp|P92658|CYB_MAMPR Cytochrome b OS=Mammuthus primigenius GN=MT-CYB PE=3 SV=3
MTHIRKSHPLLKILNKSFIDLPTPSNISTWNNFGSLLGACLITQILTGLFLAMHYTPDTM
TAFSSMSHICRDVNYGWIIRQLHNSGASIFFLCLYTHIGRNIYYGSYLYSETWNTGIMLL
LITMATAFMGYVLPWQMSFWGATVITNLFSAIPYIGTDLVEWIIWGGFSVDKATLNRFFA
LHFILPFTMIALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLGLLILILFLL
LLALLSPDMLGDPDNYMPADPLNTPHLHIKPEWYFLFAYAILRSVPNKLGGVLALLLSIL
LGIMPLLHTSKHRSMMLRPLSQVLFWTLATDMLLMTWIGSQPVEYPYIIIGQMASILYFS
IILAFLPIAGMIENYLK
```

Target database¹ UniProtKB/Swiss-Prot E-Threshold¹ 10 Matrix¹ Auto Filtering¹ None

Run Blast in a separate window.

Clear Run BLAST

Entry	Protein names	Match hit	Identity
		50 100 150 200	
Q38PR7	ATP synthase subunit a (Mammuthus primigenius)		100.0%
Q2I3G9	ATP synthase subunit a (Elephas maximus)		95.5%
Q9TA24	ATP synthase subunit a (Loxodonta africana)		93.2%

- a. *M.primigenius* (Q38PR7) – *E.maximus* (Q2I3G9) = 95.5%
M.primigenius (Q38PR7) – *L.africana* (Q9TA24) = 93.2%

b. Just this sequence similarity is not evidence enough for claiming the Mammoth is closer to the asian elephant than to the african elephant,
BUT

the last genome sequencing works on the Woolly Mammoth (PMID: 19020620), in 2008, provides evidence enough to determine that it is really closer to the asian elephant; corroborating the similarity shown in exercise 5a.

- c. Different results! (read “b” again...)
M.primigenius (P92658) – *E.maximus* (O47885) = 96.3%
M.primigenius (P92658) – *L.africana* (P24958) = 97.9%

Entry	Protein names	Match hit	Identity
		100 200 300 400	
P92658	Cytochrome b (Mammuthus primigenius)		100.0%
P24958	Cytochrome b (Loxodonta africana)		97.9%
O47885	Cytochrome b (Elephas maximus)		96.3%

6

- a. Based solely on the sequence of the “Cytochrome b” protein from the extinct dodo (*Raphus cucullatus*), was the dodo closer to the Nicobar pigeon “*Caloenas nicobarica*” or to the chicken (*Gallus gallus*)? Use NCBI Blast.
- b. There are more than 300 species of pigeons. Do the results differ if you consider the street pigeon (*Columba livia*)?



Dodo
(*Raphus cucullatus*)



Nicobar pigeon
(*Caloenas nicobarica*)



Chicken (rooster)
(*Gallus gallus*)



Pigeon
(*Columba livia*)

6

a. It seems that the dodo was closer to the pigeon than to the chicken.

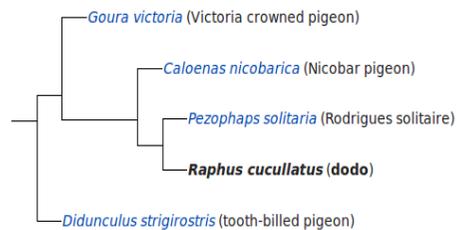
R.cucullatus – *C.nicobarica* = 99%

R.cucullatus – *G.gallus* = 93%

b. Same results for different pigeons.

R.cucullatus – *C.livia* = 96%

R.cucullatus – *G.gallus* = 93%



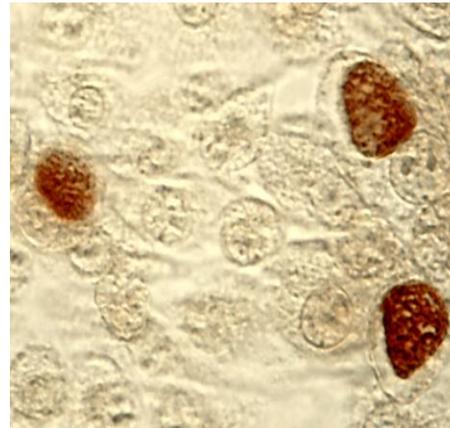
	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input checked="" type="checkbox"/>	cytochrome b [Caloenas nicobarica]	535	535	100%	0.0	99%	AAM19503.1
<input checked="" type="checkbox"/>	cytochrome b [Columba livia]	526	526	100%	0.0	96%	YP_003540719.1
<input type="checkbox"/>	cytochrome b [Columba livia]	522	522	100%	0.0	95%	AJK30555.1
<input type="checkbox"/>	cytochrome b [Columba livia]	521	521	100%	0.0	95%	AKB93366.1
<input checked="" type="checkbox"/>	cytochrome b [Gallus gallus]	509	509	100%	0.0	93%	ADB06697.1

7

- a. The UniProt entry “P04585” contains the Gag-Pol polyprotein from the virus HV1H2. Do you think it would resemble any protein in the proteome of the Zebra finch (*Taeniopygia guttata*)? Check it using NCBI Blast.
- b. Discuss the results. What is the query coverage telling us?
- c. The Gag-Pol polyprotein is composed of many proteins. Using only protein entries from the bacteria “*Chlamydia trachomatis*”, can you identify some of the individual proteins of the Gag-Pol polyprotein?



Zebra finch
(*Taeniopygia guttata*)



Chlamydia trachomatis

Homology

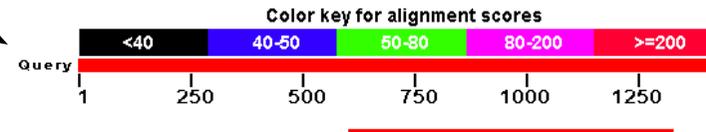
*Images from: NCBI

7

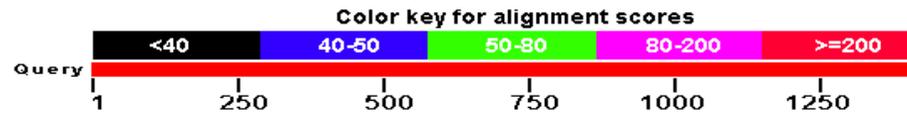
Description	Max score	Total score	Query cover	E value	Ident	Accession
PREDICTED: endogenous retrovirus group K member 18 Pol protein-like [Taeniopygia guttata]	240	240	50%	2e-65	27%	XP_012432209.1

a. XP_012432209.1. It has 27% identity with an endogenous retrovirus in *T.guttata*'s genome.

b. The query coverage is 50%, meaning that the viral "pol" protein (C-terminal) is integrated, while the "gag" protein (N-terminal) is not.



c. Database: protein entries from "*Chlamydia trachomatis*".



P24, 282-426

gag gene protein p24 (core nucleocapsid protein) [Chlamydia trachomatis]
Sequence ID: [embjCRH58881.1](#) Length: 382 Number of Matches: 1

Range 1: 176 to 314 [GenPept](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps
74.3 bits(181)	2e-13	Compositional matrix adjust.	52/152(34%)	73/152(48%)	20/152(13%)

Reverse transcriptase, 608-899 + 1018-1416

Reverse transcriptase (RNA-dependent DNA polymerase) [Chlamydia trachomatis]
Sequence ID: [embjCRH45814.1](#) Length: 867 Number of Matches: 2

Range 1: 3 to 292 [GenPept](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps
164 bits(414)	6e-41	Compositional matrix adjust.	105/295(36%)	157/295(53%)	8/295(2%)

Range 2: 423 to 828 [GenPept](#) [Graphics](#) [Next Match](#) [Previous Match](#) [First Match](#)

Score	Expect	Method	Identities	Positives	Gaps
97.1 bits(240)	4e-20	Compositional matrix adjust.	105/431(24%)	178/431(41%)	57/431(13%)

Ribonuclease H, 1109-1306

ribonuclease H [Chlamydia trachomatis]
Sequence ID: [embjCRH57958.1](#) Length: 831 Number of Matches: 1

Range 1: 368 to 589 [GenPept](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps
40.8 bits(94)	0.007	Compositional matrix adjust.	57/222(26%)	92/222(41%)	24/222(10%)

8

Using the protein “P38398”, perform a “tblastn” search in NCBI against human entries.

- a. What would be this search used for?
- b. Is there any difference between the first and the second result?

8

a. Query: protein. Database: nucleotide. To look for the gene encoding the query protein.

	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	Homo sapiens breast cancer 1 (BRCA1), transcript variant 1, mRNA	3576	3576	100%	0.0	94%	NM_007294.3
<input type="checkbox"/>	Homo sapiens breast and ovarian cancer susceptibility (BRCA1) mRNA, complete cds	3576	3576	100%	0.0	94%	U14680.1

Homo sapiens breast cancer 1 (BRCA1), transcript variant 1, mRNA

Sequence ID: [ref|NM_007294.3|](#) Length: 7224 Number of Matches: 1

Range 1: 233 to 5821 [GenBank](#) [Graphics](#) ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
3576 bits(9273)	0.0	Compositional matrix adjust.	1863/1863(100%)	1863/1863(100%)	0/1863(0%)	+2
Query 1		MDLSALRVEEVQNVINAMQKILECPICLLELIKEPVSTKCDHIFCKFCMLKLLNQKKGPSQ				60
Sbjct 233		MDLSALRVEEVQNVINAMQKILECPICLLELIKEPVSTKCDHIFCKFCMLKLLNQKKGPSQ				412



b. First result: NM_007294.3
7224 bp, transcript variant 1, mRNA

Query: 1 _____ 1863
Subject: 233 _____ 5821

Homo sapiens breast and ovarian cancer susceptibility (BRCA1) mRNA, complete cds

Sequence ID: [gb|U14680.1|HSU14680](#) Length: 5711 Number of Matches: 1

Range 1: 120 to 5708 [GenBank](#) [Graphics](#) ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
3576 bits(9273)	0.0	Compositional matrix adjust.	1863/1863(100%)	1863/1863(100%)	0/1863(0%)	+3
Query 1		MDLSALRVEEVQNVINAMQKILECPICLLELIKEPVSTKCDHIFCKFCMLKLLNQKKGPSQ				60
Sbjct 120		MDLSALRVEEVQNVINAMQKILECPICLLELIKEPVSTKCDHIFCKFCMLKLLNQKKGPSQ				299



Second result: U14680.1
5711bp, complete CDS

Query: 1 _____ 1863
Subject: 120 _____ 5708