BIOL312
Bioinformatics and Computational Biology

Practical 2

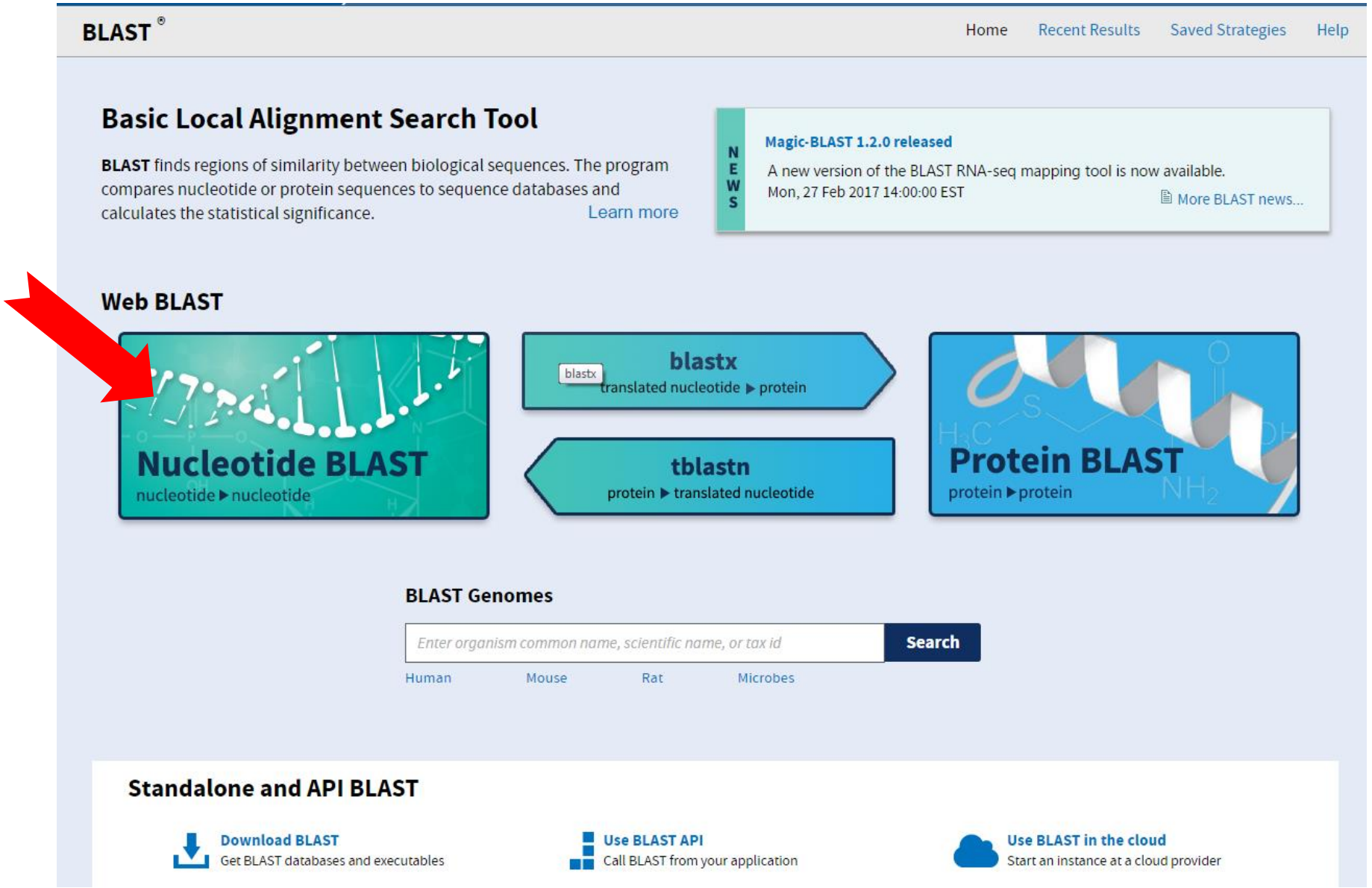
Identifying an Unknown DNA Sequence

To search an unknown DNA sequence;

- Directly DNA database
- Find the amino acid sequence and search via protein database

BLASTn is used for DNA database search. Searching a DNA query gives you the most similar DNA sequence from this database.

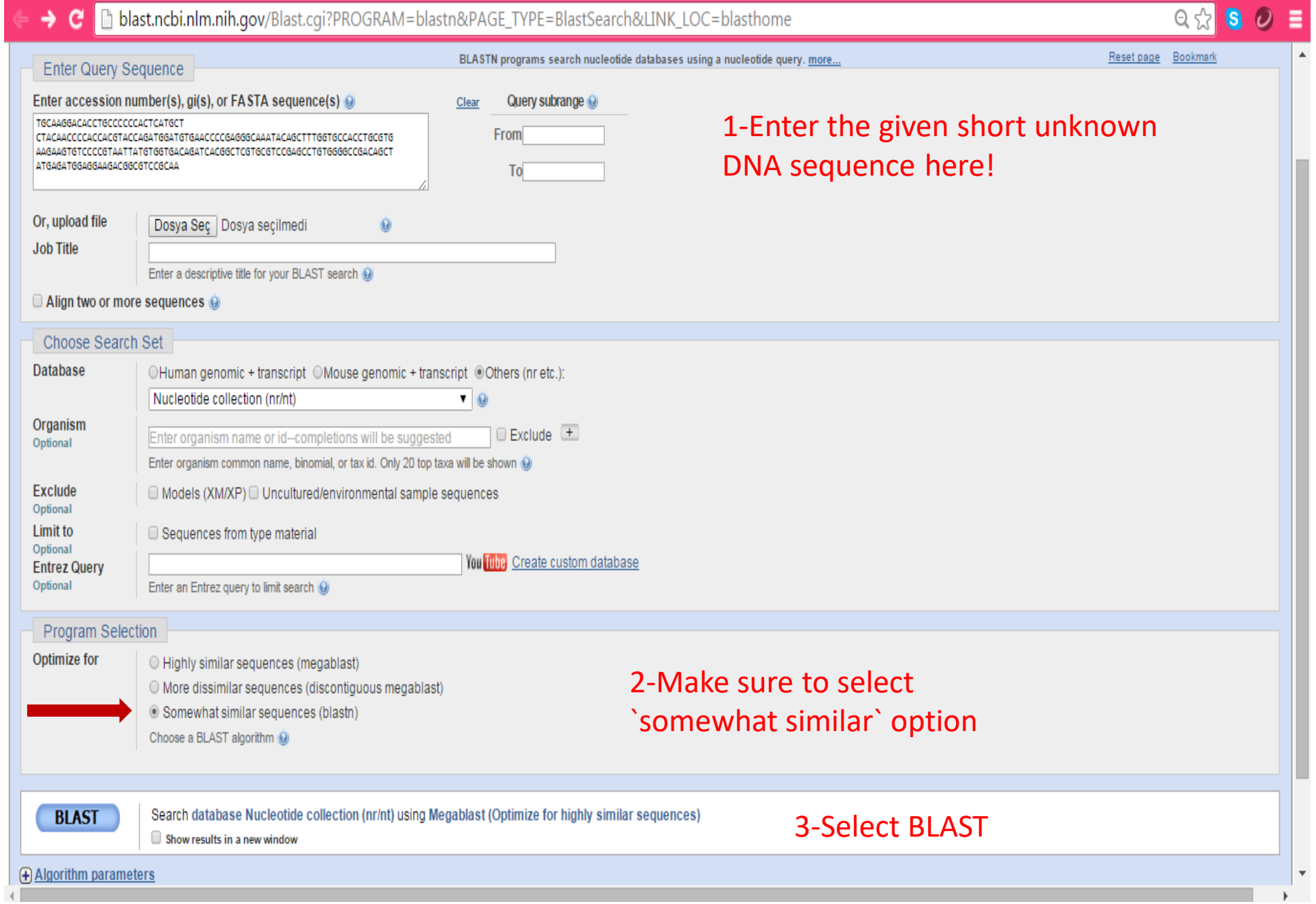
I. Identifying the Unknown DNA Sequence by using BLASTn



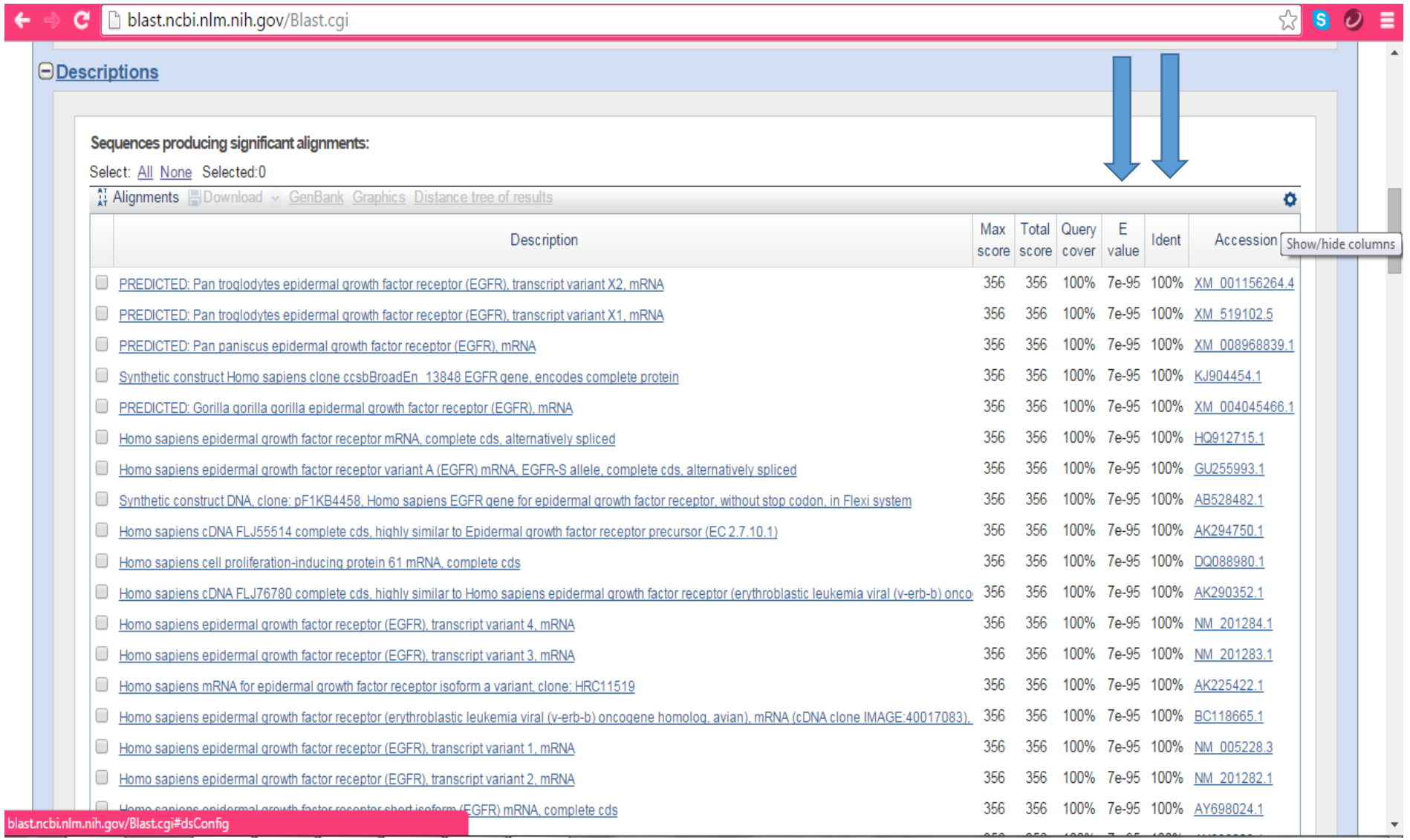
Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast, delta-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query


blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome
Enter Query Sequence
BLASTN programs search nucleotide databases using a nucleotide query. more...
Reset page
Bookmark
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
TGCAAGGACACCTGCCCCCCACTCATGCT
CTACAACCCCACCACGTACCAGATGGATGTGAACCCCGAGGGCAAATACAGCTTTGGTGCCACCTGCGTG
AAGAAGTGTCCCCGTAATTATGTGGTGACAGATCACGGCTCGTGCGTCCGAGCCTGTGGGGCCGACAGCT
ATGAGATGGAGGAAGACGGCGTCCGCAA
1-Enter the given short unknown DNA sequence here!
Or, upload file
Dosya Seç
Dosya seçilmedi
Job Title
Enter a descriptive title for your BLAST search
Align two or more sequences
Choose Search Set
Database
Human genomic + transcript
Mouse genomic + transcript
Others (nr etc.):
Nucleotide collection (nr/nt)
Organism
Optional
Enter organism name or id--completions will be suggested
Exclude
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown
Exclude
Optional
Models (XM/XP)
Uncultured/environmental sample sequences
Limit to
Optional
Sequences from type material
Entrez Query
Optional
Create custom database
Enter an Entrez query to limit search
Program Selection
Optimize for
Highly similar sequences (megablast)
More dissimilar sequences (discontiguous megablast)
Somewhat similar sequences (blastn)
Choose a BLAST algorithm
2-Make sure to select `somewhat similar` option
BLAST
Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)
Show results in a new window
3-Select BLAST
Algorithm parameters

Analyze the outcome by studying %identity and e-value score.



Description	Max score	Total score	Query cover	E value	Ident	Accession
PREDICTED: Pan troglodytes epidermal growth factor receptor (EGFR), transcript variant X2, mRNA	356	356	100%	7e-95	100%	XM_001156264.4
PREDICTED: Pan troglodytes epidermal growth factor receptor (EGFR), transcript variant X1, mRNA	356	356	100%	7e-95	100%	XM_519102.5
PREDICTED: Pan paniscus epidermal growth factor receptor (EGFR), mRNA	356	356	100%	7e-95	100%	XM_008968839.1
Synthetic construct Homo sapiens clone ccsbBroadEn_13848 EGFR gene, encodes complete protein	356	356	100%	7e-95	100%	KJ904454.1
PREDICTED: Gorilla gorilla gorilla epidermal growth factor receptor (EGFR), mRNA	356	356	100%	7e-95	100%	XM_004045466.1
Homo sapiens epidermal growth factor receptor mRNA, complete cds, alternatively spliced	356	356	100%	7e-95	100%	HQ912715.1
Homo sapiens epidermal growth factor receptor variant A (EGFR) mRNA, EGFR-S allele, complete cds, alternatively spliced	356	356	100%	7e-95	100%	GU255993.1
Synthetic construct DNA, clone: pF1KB4458, Homo sapiens EGFR gene for epidermal growth factor receptor, without stop codon, in Flexi system	356	356	100%	7e-95	100%	AB528482.1
Homo sapiens cDNA FLJ55514 complete cds, highly similar to Epidermal growth factor receptor precursor (EC 2.7.10.1)	356	356	100%	7e-95	100%	AK294750.1
Homo sapiens cell proliferation-inducing protein 61 mRNA, complete cds	356	356	100%	7e-95	100%	DQ088980.1
Homo sapiens cDNA FLJ76780 complete cds, highly similar to Homo sapiens epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) onco	356	356	100%	7e-95	100%	AK290352.1
Homo sapiens epidermal growth factor receptor (EGFR), transcript variant 4, mRNA	356	356	100%	7e-95	100%	NM_201284.1
Homo sapiens epidermal growth factor receptor (EGFR), transcript variant 3, mRNA	356	356	100%	7e-95	100%	NM_201283.1
Homo sapiens mRNA for epidermal growth factor receptor isoform a variant, clone: HRC11519	356	356	100%	7e-95	100%	AK225422.1
Homo sapiens epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian), mRNA (cDNA clone IMAGE:40017083),	356	356	100%	7e-95	100%	BC118665.1
Homo sapiens epidermal growth factor receptor (EGFR), transcript variant 1, mRNA	356	356	100%	7e-95	100%	NM_005228.3
Homo sapiens epidermal growth factor receptor (EGFR), transcript variant 2, mRNA	356	356	100%	7e-95	100%	NM_201282.1
Homo sapiens epidermal growth factor receptor short isoform (EGFR) mRNA, complete cds	356	356	100%	7e-95	100%	AY698024.1

E-Value

- E-value: Expectation value that indicates the number of alignments with a score that one can expect to find by chance in a database.
- E-value depends on the database size & query length.
- The closer the E-value to 0, the better the alignment is.
- E.g.: E=1e-2 (= $1 \times 10^{-2} = 0.01$)
- The lower the E-value (close to zero), more significant match.
- Statistics of sequence similarity score