Topic 5 – Sequence Alignment

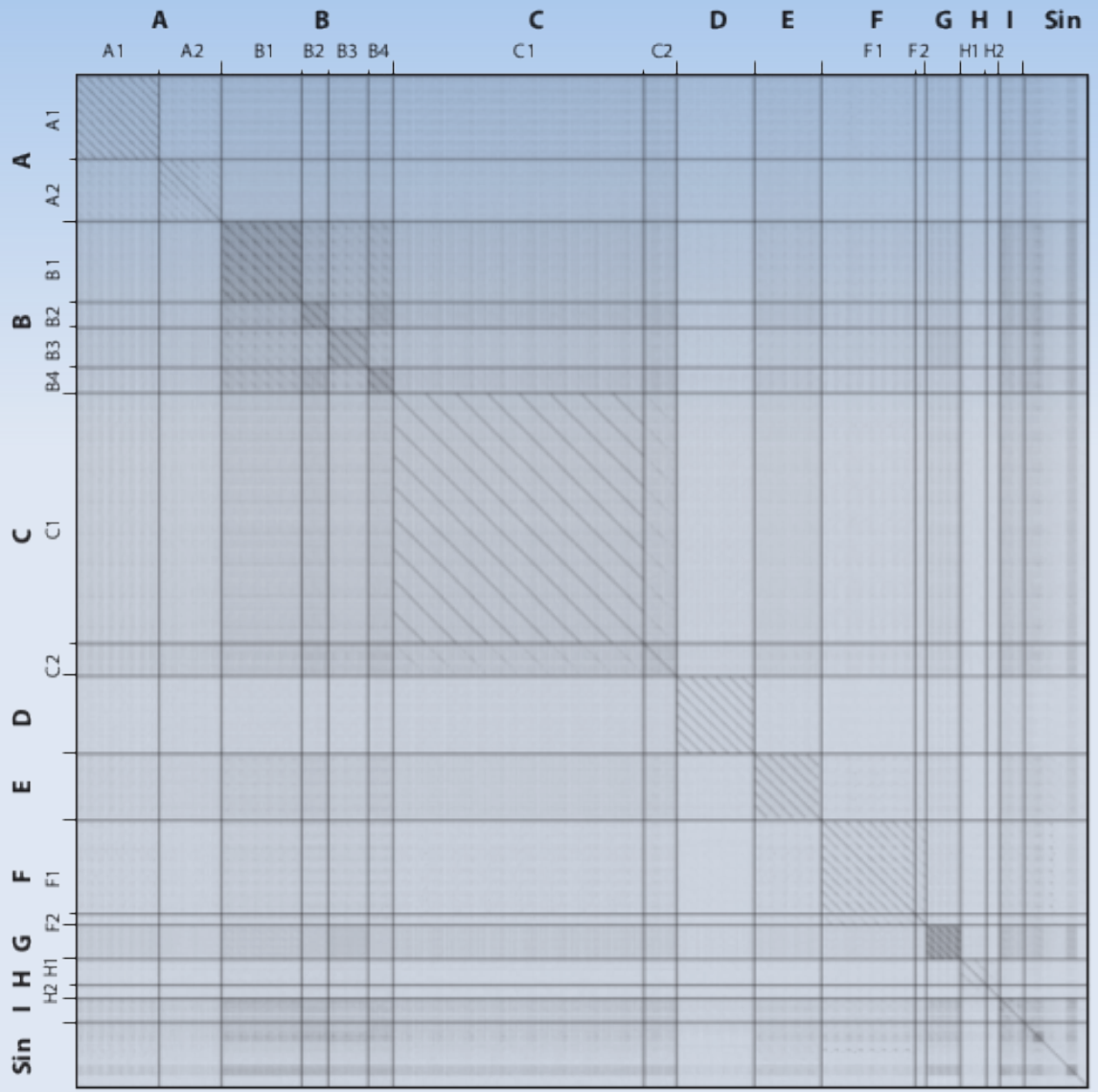


Figure 1

Dotplot comparison of 70 sequenced mycobacteriophage genomes. Each of the 70 sequenced mycobacteriophages was concatenated into a single ~5-Mbp sequence and compared with itself using Gepard (62). The genome order is the same as in **Table 1** and the Cluster and Subcluster designations are

BIOT-CSIS 373
Sequence Alignment I

The sequence alignment problem:

- you have 2 strings (sequences)
- find the ”best match” between the two

Ex:

leaf	ATCCA	MPPDNE
please	TCGA	PADQE

Quick Look at NCBI BLAST

BLAST = Basic Local Alignment Search Tool

Allows researchers (like you) to search a database of sequences – e.g. I want to find all genes in a database that have a sequence ”xyz” .

Local Alignment => a good way to compare two sequences that are very different in lengths.

Reasons for using BLAST

Identification

Examples:

You are a forensics investigator and you have a DNA sequence – is it human? How likely?

- You want an exact match.

You have a protein sequence – what is it? Is its function known?

Reasons for using BLAST

Gene Discovery

You have found a new gene – DNA and protein

Are there any known genes in a database that are similar to the new gene

Are there other organisms with similar genes?

- you don't need exact matches & there may not be any

Reasons for using BLAST

mRNA Analysis

You have an mRNA sequence

Where in the genome did it come from? What else is near the genomic location?

Could there be splice variants?

Reasons for using BLAST

Related Organisms

You have an DNA sequence

What are its closest ancestors?

What are its possible functions?

Built-in Assumptions

Protein structure determines function

Protein sequence determines structure

Evolution preserves function by preserving protein sequence.

Codon degeneracy => Evolution is more likely to preserve protein rather than DNA sequences.

3 Components of a BLAST search

- Query – usually a fully known sequence
- Database – many possibilities including nucleotide, protein, translated nucleotide
- Program – different programs are used depending on query and database and other factors

Programs

Basic BLAST

Choose a BLAST program to run.

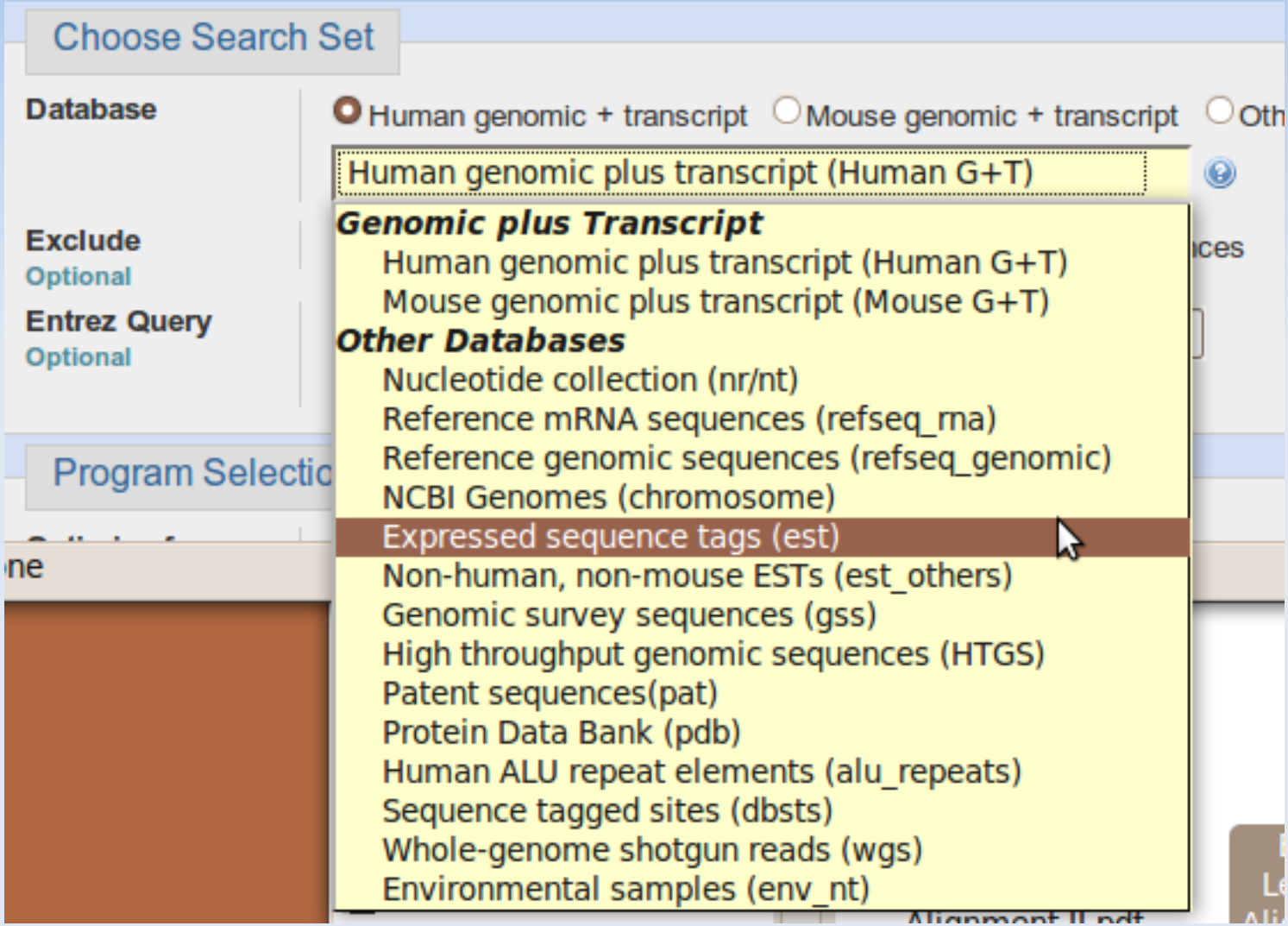
nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast, delta-blast
blastx	Search **protein** database using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

Specialized BLAST

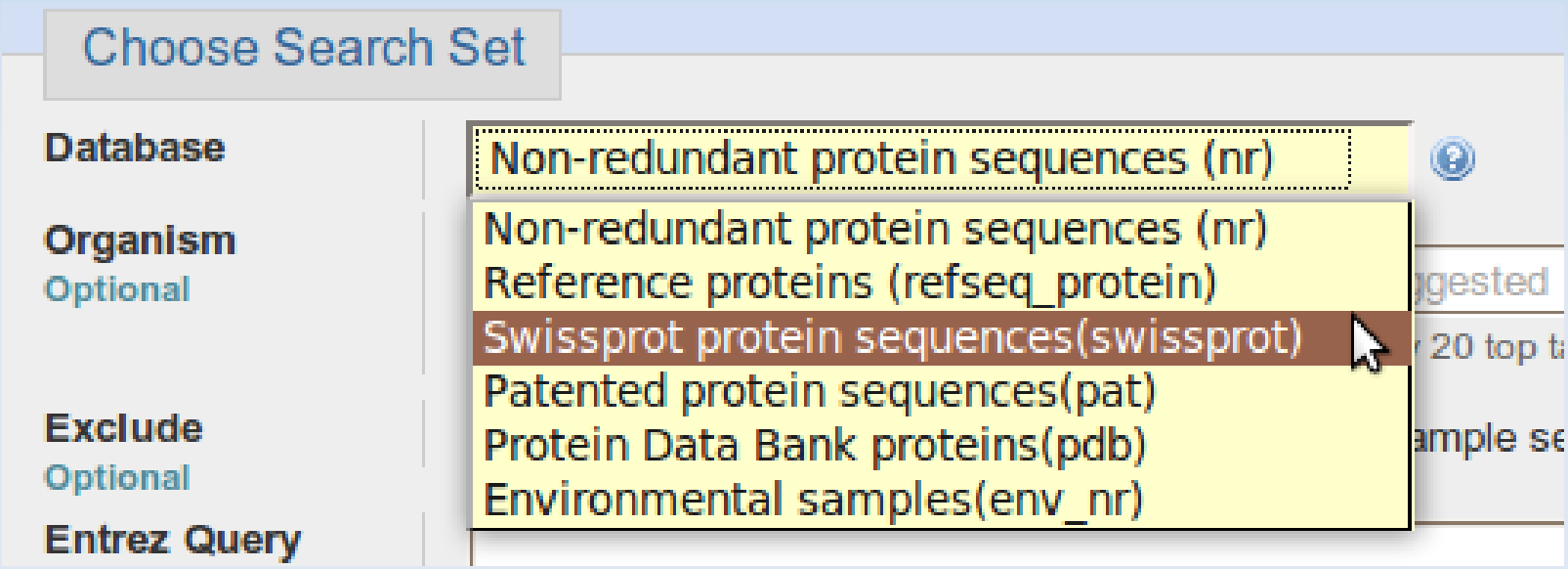
Choose a type of specialized search (or database name in parentheses.)

- Get faster protein results with a graphical view using **SmartBLAST**
- Make specific primers with **Primer-BLAST**
- Cluster multiple sequences together with their database neighbors using **MOLE-BLAST**
- Find **conserved domains** in your sequence (cds)
- Find sequences with similar **conserved domain architecture** (cdart)
- Search sequences that have **gene expression profiles** (GEO)
- Search **immunoglobulins and T cell receptor sequences** (IgBLAST)
- Screen sequence for **vector contamination** (vecscreen)
- **Align** two (or more) sequences using BLAST (bl2seq)
- Search **protein** or **nucleotide** targets in PubChem BioAssay

Nucleotide Database

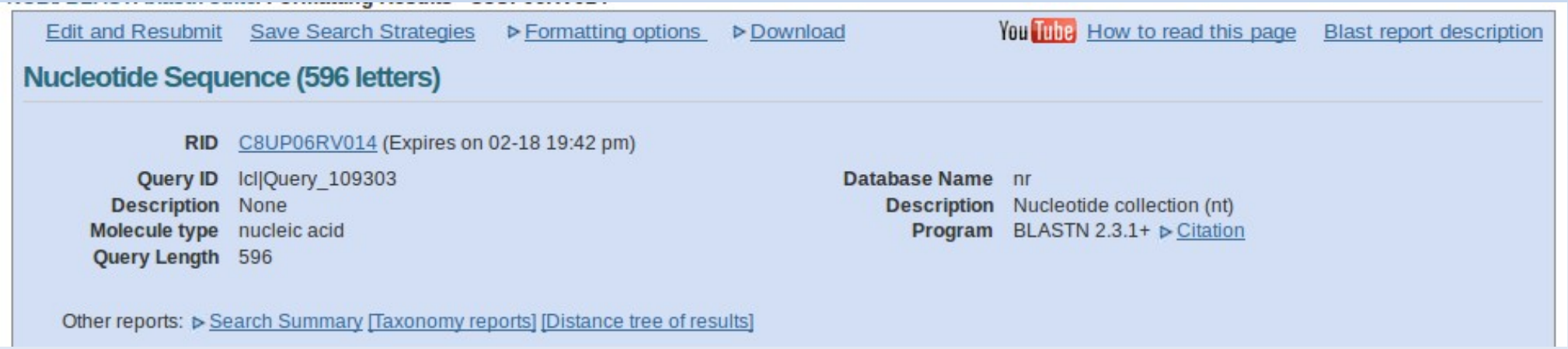


Protein Databases



BLAST Results

Part 1: Header



RID – good for 24 hours

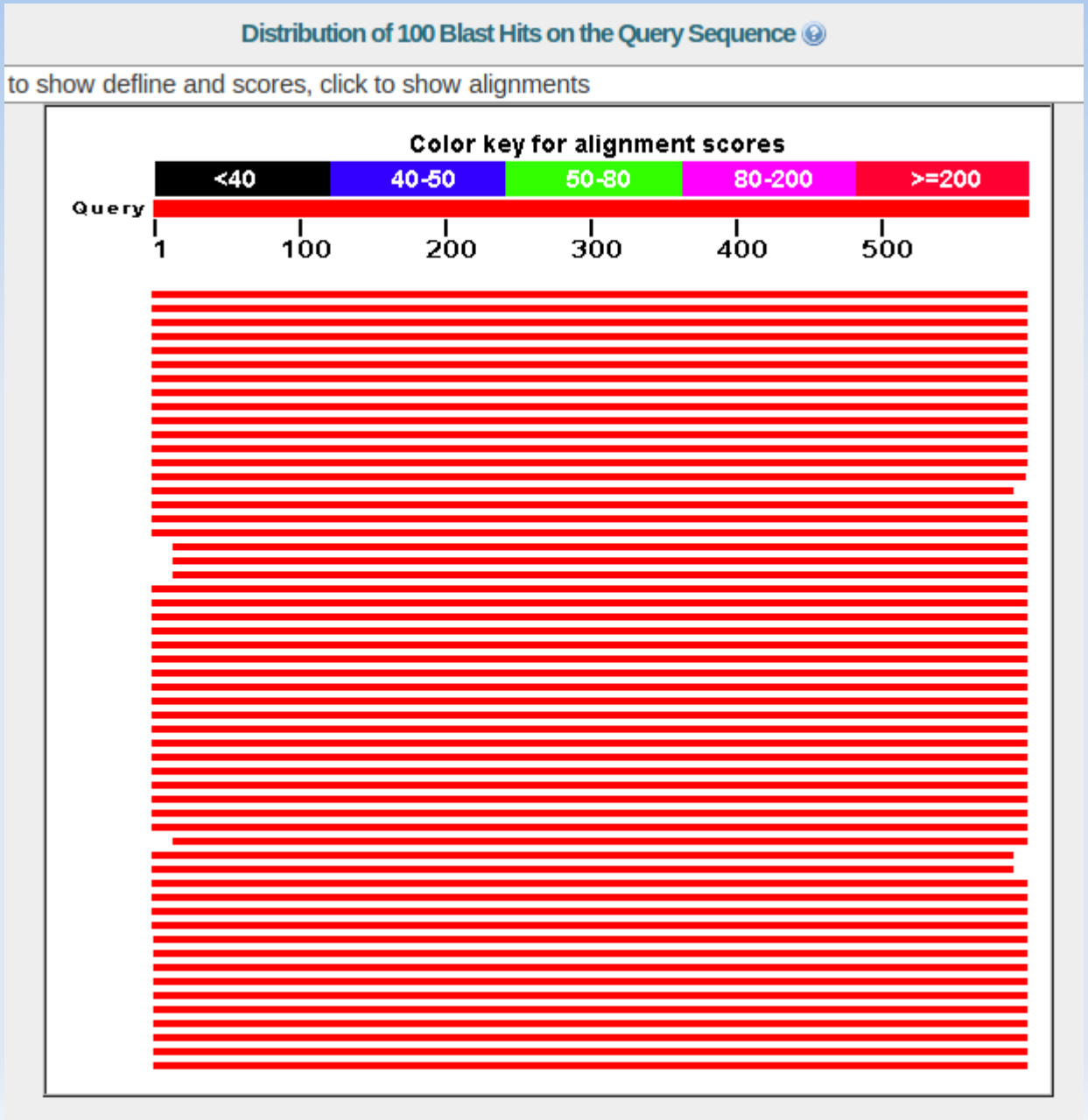
Type and length of query

DB searched

Taxonomy of results

BLAST Results Part 2

Part 2: Graphical display of ”hits” and links to alignments:



BLAST Results Part 3

Text summary of ”hits” with scores (higher is better), ”Expect” values (lower is better) and links to alignments and GenBank records:

Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Ident	Accession
PREDICTED: Homo sapiens polymerase (DNA directed), lambda (POLL), transcript varian	1101	1101	100%	0.0	100%	XM_011539655.1
PREDICTED: Homo sapiens polymerase (DNA directed), lambda (POLL), transcript varian	1101	1101	100%	0.0	100%	XM_011539654.1
PREDICTED: Homo sapiens polymerase (DNA directed), lambda (POLL), transcript varian	1101	1101	100%	0.0	100%	XM_011539653.1
PREDICTED: Homo sapiens polymerase (DNA directed), lambda (POLL), transcript varian	1101	1101	100%	0.0	100%	XM_011539652.1
PREDICTED: Homo sapiens polymerase (DNA directed), lambda (POLL), transcript varian	1101	1101	100%	0.0	100%	XM_011539651.1
PREDICTED: Homo sapiens polymerase (DNA directed), lambda (POLL), transcript varian	1101	1101	100%	0.0	100%	XM_011539650.1
Homo sapiens polymerase (DNA directed), lambda (POLL), transcript variant 2, mRNA	1101	1101	100%	0.0	100%	NM_013274.3
Homo sapiens polymerase (DNA directed), lambda (POLL), transcript variant 3, mRNA	1101	1101	100%	0.0	100%	NM_001174085.1
Homo sapiens polymerase (DNA directed), lambda (POLL), transcript variant 1, mRNA	1101	1101	100%	0.0	100%	NM_001174084.1
Homo sapiens cDNA FLJ77175 complete cds, highly similar to Homo sapiens DNA polym	1101	1101	100%	0.0	100%	AK292225.1
Homo sapiens polymerase (DNA directed), lambda, mRNA (cDNA clone MGC:87359 IMA(	1101	1101	100%	0.0	100%	BC068529.1
Homo sapiens DNA polymerase lambda (Pol lambda) mRNA, complete cds	1101	1101	100%	0.0	100%	AF161019.1
Homo sapiens mRNA for DNA polymerase lambda (POLL gene)	1101	1101	100%	0.0	100%	AJ131890.1
Homo sapiens cDNA FLJ11538 fis, clone HEMBA1002746, weakly similar to DNA POLYM	1099	1099	99%	0.0	100%	AK021600.1
PREDICTED: Homo sapiens polymerase (DNA directed), lambda (POLL), transcript varian	1085	1085	98%	0.0	100%	XM_011539656.1
PREDICTED: Gorilla gorilla gorilla polymerase (DNA directed), lambda, transcript variant 3	1085	1085	100%	0.0	99%	XM_004049985.1
PREDICTED: Gorilla gorilla gorilla polymerase (DNA directed), lambda, transcript variant 2	1085	1085	100%	0.0	99%	XM_004049984.1
PREDICTED: Gorilla gorilla gorilla polymerase (DNA directed), lambda, transcript variant 1	1085	1085	100%	0.0	99%	XM_004049983.1
Synthetic construct DNA, clone: pF1KB3002, Homo sapiens POLL gene for DNA polymera	1075	1075	97%	0.0	100%	AB384717.1
Homo sapiens DNA polymerase lamda2 mRNA, complete cds	1075	1075	97%	0.0	100%	AY302442.1
Homo sapiens DNA polymerase beta2 mRNA, complete cds	1070	1070	97%	0.0	99%	AF283478.1
PREDICTED: Pan paniscus polymerase (DNA directed), lambda (POLL), transcript variant	1068	1068	100%	0.0	99%	XM_003825495.2
PREDICTED: Pan paniscus polymerase (DNA directed), lambda (POLL), transcript variant	1068	1068	100%	0.0	99%	XM_008950830.2

BLAST Results Part 4

Alignments

Upper sequence is the one we used to search

Lower ones are the ones BLAST found

HSP – **H**igh-scoring **S**egment **P**air – local alignment

Alignments

Download GenBank Graphics ▼ Next ▲ Previous Descriptions

PREDICTED: Homo sapiens polymerase (DNA directed), lambda (POLL), transcript variant X6, mRNA

Sequence ID: ref|XM_011539655.1| Length: 2123 Number of Matches: 1

Range 1: 87 to 682 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
1101 bits(596)	0.0	596/596(100%)	0/596(0%)	Plus/Plus

```
Query  1    CCAGCCATACTTCAATGGATCCCAGGGGTATCTTGAAGGCATTTCCCAAGCGGCAGAAAA  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  87   CCAGCCATACTTCAATGGATCCCAGGGGTATCTTGAAGGCATTTCCCAAGCGGCAGAAAA  146

Query  61   TTCATGCTGATGCATCATCAAAAGTACTTGCAAAGATTCCTAGGAGGGAAGAGGGAGAAG  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  147  TTCATGCTGATGCATCATCAAAAGTACTTGCAAAGATTCCTAGGAGGGAAGAGGGAGAAG  206

Query  121  AAGCAGAAGAGTGGCTGAGCTCCCTTCGGGCCCATGTTGTGCGCACTGGCATTGGACGAG  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  207  AAGCAGAAGAGTGGCTGAGCTCCCTTCGGGCCCATGTTGTGCGCACTGGCATTGGACGAG  266

Query  181  CCCGGGCAGAACTCTTTGAGAAGCAGATTGTTCAGCATGGCGGCCAGCTATGCCCTGCCC  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  267  CCCGGGCAGAACTCTTTGAGAAGCAGATTGTTCAGCATGGCGGCCAGCTATGCCCTGCCC  326

Query  241  AGGGCCCAGGTGTCACTCACATTGTGGTGGATGAAGGCATGGACTATGAGCGAGCCCTCC  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  327  AGGGCCCAGGTGTCACTCACATTGTGGTGGATGAAGGCATGGACTATGAGCGAGCCCTCC  386

Query  301  GCCTTCTCAGACTACCCCAGCTGCCCCCGGGTGCTCAGCTGGTGAAGTCAGCCTGGCTGA  360
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  387  GCCTTCTCAGACTACCCCAGCTGCCCCCGGGTGCTCAGCTGGTGAAGTCAGCCTGGCTGA  446

Query  361  GCTTGTGCCTTCAGGAGAGGAGGCTGGTGGATGTAGCTGGATTCAGCATCTTCATCCCCA  420
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  447  GCTTGTGCCTTCAGGAGAGGAGGCTGGTGGATGTAGCTGGATTCAGCATCTTCATCCCCA  506

Query  421  GTAGGTACTTGGACCATCCACAGCCCAGCAAGGCAGAGCAGGATGCTTCTATTCCTCCTG  480
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  507  GTAGGTACTTGGACCATCCACAGCCCAGCAAGGCAGAGCAGGATGCTTCTATTCCTCCTG  566

Query  481  GCACCCATGAGGCCCTGCTTCAGACAGCCCTTTCTCCTCCTCCTCCTCCCACCAGGCCTG  540
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  567  GCACCCATGAGGCCCTGCTTCAGACAGCCCTTTCTCCTCCTCCTCCTCCCACCAGGCCTG  626

Query  541  TGTCTCCTCCCCAAAAGGCAAAAGAGGCACCAAACACCCAAGCCCAGCCCATCTCT  596
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  627  TGTCTCCTCCCCAAAAGGCAAAAGAGGCACCAAACACCCAAGCCCAGCCCATCTCT  682
```

Related Information

Gene - associated gene details

Download GenBank Graphics ▼ Next ▲ Previous Descriptions

PREDICTED: Homo sapiens polymerase (DNA directed), lambda (POLL), transcript variant X5, mRNA

Sequence ID: ref|XM_011539654.1| Length: 2144 Number of Matches: 1

Range 1: 87 to 682 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
1101 bits(596)	0.0	596/596(100%)	0/596(0%)	Plus/Plus

```
Query  1    CCAGCCATACTTCAATGGATCCCAGGGGTATCTTGAAGGCATTTCCCAAGCGGCAGAAAA  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  87   CCAGCCATACTTCAATGGATCCCAGGGGTATCTTGAAGGCATTTCCCAAGCGGCAGAAAA  146

Query  61   TTCATGCTGATGCATCATCAAAAGTACTTGCAAAGATTCCTAGGAGGGAAGAGGGAGAAG  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
```

Related Information

Gene - associated gene details

BLAST Results Summary

Search Summary:

Program used, search parameters

Match reward is 1

Mismatch penalty is -2

Database searched – 35 million sequences!

Stats: Lambda, K, & H are used to compute E value: probability of a false positive sequence similarity

Other reports: Search Summary [Taxonomy reports] [Distance tree of results]

Search Parameters	
Program	blastn
Word size	28
Expect value	10
Hitlist size	100
Match/Mismatch scores	1,-2
Gapcosts	0,0
Low Complexity Filter	Yes
Filter string	L;m;
Genetic Code	1

Database	
Posted date	Feb 8, 2016 1:49 AM
Number of letters	111,527,859,682
Number of sequences	34,665,943
Entrez query	none

Karlin-Altschul statistics		
Lambda	1.33271	1.28
K	0.620991	0.46
H	1.12409	0.85

Results Statistics	
Length adjustment	34
Effective length of query	562
Effective length of database	110349217620
Effective search space	62016260302440
Effective search space used	62016260302440

BLAST Results Taxonomy Report

Hits organized by species

Lineage Report

Organism R

Organism	Blast Name	Score	Number of Hits	Description
root			101	
. Catarrhini	primates		100	
. . Hominoidea	primates		45	
. . . Homininae	primates		41	
. . . . Homo sapiens	primates	1101	17	Homo sapiens hits
. . . . Gorilla gorilla gorilla	primates	1085	3	Gorilla gorilla gorilla hits
. . . . Pan paniscus	primates	1068	10	Pan paniscus hits
. . . . Pan troglodytes	primates	1068	10	Pan troglodytes hits
. . . . Gorilla gorilla	primates	1059	1	Gorilla gorilla hits
. . . Nomascus leucogenys	primates	1003	4	Nomascus leucogenys hits
. . Colobus angolensis palliatus	primates	1013	4	Colobus angolensis palliatus hits
. . Macaca fascicularis	primates	1005	11	Macaca fascicularis hits
. . Macaca mulatta	primates	1005	13	Macaca mulatta hits
. . Colobus guereza	primates	1003	1	Colobus guereza hits
. . Rhinopithecus roxellana	primates	1002	1	Rhinopithecus roxellana hits
. . Papio anubis	primates	1002	5	Papio anubis hits
. . Mandrillus leucophaeus	primates	996	3	Mandrillus leucophaeus hits
. . Chlorocebus sabaeus	primates	996	11	Chlorocebus sabaeus hits
. . Macaca nemestrina	primates	994	6	Macaca nemestrina hits
. synthetic construct	other sequences	1075	1	synthetic construct hits

BLAST Results Taxonomy Report

Hits organized by species

Organism Report

Lineage Report Taxonomy Rep

Description	Score	E value	Accession
Homo sapiens (human) [primates] ▼ Next ▲ Previous ⏫ First			
PREDICTED: Homo sapiens polymerase (DNA directed), lambda (POLL), transcript variant X6, mRNA	1101	0.0	XM_011539655
PREDICTED: Homo sapiens polymerase (DNA directed), lambda (POLL), transcript variant X5, mRNA	1101	0.0	XM_011539654
PREDICTED: Homo sapiens polymerase (DNA directed), lambda (POLL), transcript variant X4, mRNA	1101	0.0	XM_011539653
PREDICTED: Homo sapiens polymerase (DNA directed), lambda (POLL), transcript variant X3, mRNA	1101	0.0	XM_011539652
PREDICTED: Homo sapiens polymerase (DNA directed), lambda (POLL), transcript variant X2, mRNA	1101	0.0	XM_011539651
PREDICTED: Homo sapiens polymerase (DNA directed), lambda (POLL), transcript variant X1, mRNA	1101	0.0	XM_011539650
Homo sapiens polymerase (DNA directed), lambda (POLL), transcript variant 2, mRNA	1101	0.0	NM_013274
Homo sapiens polymerase (DNA directed), lambda (POLL), transcript variant 3, mRNA	1101	0.0	NM_001174085
Homo sapiens polymerase (DNA directed), lambda (POLL), transcript variant 1, mRNA	1101	0.0	NM_001174084
Homo sapiens cDNA FLJ77175 complete cds, highly similar to Homo sapiens DNA polymerase lamda2 mRNA	1101	0.0	AK292225
Homo sapiens polymerase (DNA directed), lambda, mRNA (cDNA clone MGC:87359 IMAGE:5267859), comple	1101	0.0	BC068529
Homo sapiens DNA polymerase lambda (Pol lambda) mRNA, complete cds	1101	0.0	AF161019
Homo sapiens mRNA for DNA polymerase lambda (POLL gene)	1101	0.0	AJ131890
Homo sapiens cDNA FLJ11538 fis, clone HEMBA1002746, weakly similar to DNA POLYMERASE BETA (EC 2.7	1099	0.0	AK021600
PREDICTED: Homo sapiens polymerase (DNA directed), lambda (POLL), transcript variant X7, mRNA	1085	0.0	XM_011539656
Homo sapiens DNA polymerase lamda2 mRNA, complete cds	1075	0.0	AY302442
Homo sapiens DNA polymerase beta2 mRNA, complete cds	1070	0.0	AF283478
Gorilla gorilla gorilla (western lowland gorilla) [primates] ▼ Next ▲ Previous ⏫ First			
PREDICTED: Gorilla gorilla gorilla polymerase (DNA directed), lambda, transcript variant 3 (POLL), mRNA	1085	0.0	XM_004049985
PREDICTED: Gorilla gorilla gorilla polymerase (DNA directed), lambda, transcript variant 2 (POLL), mRNA	1085	0.0	XM_004049984
PREDICTED: Gorilla gorilla gorilla polymerase (DNA directed), lambda, transcript variant 1 (POLL), mRNA	1085	0.0	XM_004049983
synthetic construct [other sequences] ▼ Next ▲ Previous ⏫ First			
Synthetic construct DNA, clone: pF1KB3002, Homo sapiens POLL gene for DNA polymerase lambda, complete	1075	0.0	AB384717
Pan paniscus (pygmy chimpanzee) [primates] ▼ Next ▲ Previous ⏫ First			
PREDICTED: Pan paniscus polymerase (DNA directed), lambda (POLL), transcript variant X10, mRNA	1068	0.0	XM_003825495

BLAST Results Taxonomy Report

Hits organized by taxonomy

Taxonomy Report Organism Report Lineage Report

Taxonomy	Number of hits	Number of Organisms	Description
⊟ root	101	16	
. ⊟ Catarrhini	100	15	
. . ⊟ Hominoidea	45	6	
. . . ⊟ Homininae	41	5	
. . . . Homo sapiens	17	1	Homo sapiens hits
. . . . ⊟ Gorilla	4	2	
. . . . . ⊟ Gorilla gorilla	1	2	Gorilla gorilla hits
. . . . . Gorilla gorilla gorilla	3	1	Gorilla gorilla gorilla hits
. . . . ⊟ Pan	20	2	
. . . . . Pan paniscus	10	1	Pan paniscus hits
. . . . . Pan troglodytes	10	1	Pan troglodytes hits
. . . Nomascus leucogenys	4	1	Nomascus leucogenys hits
. . ⊟ Cercopithecidae	55	9	
. . . ⊟ Colobinae	6	3	
. . . . ⊟ Colobus	5	2	
. . . . . Colobus angolensis palliatus	4	1	Colobus angolensis palliatus hits
. . . . . Colobus guereza	1	1	Colobus guereza hits
. . . . Rhinopithecus roxellana	1	1	Rhinopithecus roxellana hits
. . . ⊟ Cercopithecinae	49	6	
. . . . ⊟ Macaca	30	3	
. . . . . Macaca fascicularis	11	1	Macaca fascicularis hits
. . . . . Macaca mulatta	13	1	Macaca mulatta hits
. . . . . Macaca nemestrina	6	1	Macaca nemestrina hits
. . . . Papio anubis	5	1	Papio anubis hits
. . . . Mandrillus leucophaeus	3	1	Mandrillus leucophaeus hits
. . . . Chlorocebus sabaeus	11	1	Chlorocebus sabaeus hits
. synthetic construct	1	1	synthetic construct hits

Protein BLAST

A protein BLAST search has as input, a protein query sequence which is compared to each protein sequence in a chosen database.

Use BLOSUM62 for a wider range of sequence hits

Use BLOSUM90 for a smaller range of higher similarity hits