Finding data

BLAST

Answer key

Contents

1. BLAST basics

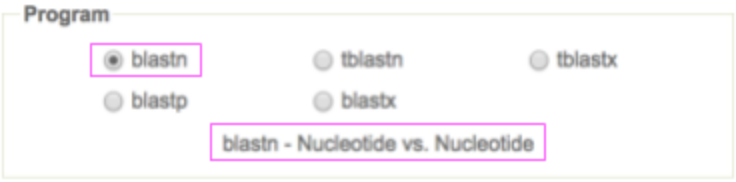
The Basic Local Alignment Search Tool, or BLAST, is a way to compare sequences. It performs pairwise alignment between the *query* and the *targets* in the database to find a specific or similar sequences. Alignments with the best-matching sequences are shown and scored. BLAST similarity searches can be used to find homologous genes[1], identify potential functions of novel genes, predict the size of PCR products, identify transcript or peptide evidence (from RNAseq or mass spectrometry experiments), and assign VectorBase and/or GenBank accession to genes among others.

BLAST is actually a family of programs and these are:

Program	Query	Database
blastn	nucleotide	nucleotide
blastp	peptide	peptide
blastx	nucleotide (translated to peptide)	peptide
tblastn	peptide	nucleotide (translated to peptide)
tblastx	nucleotide (translated to peptide)	nucleotide (translated to peptide)

[1] The task of homolog finding is better accomplished using HMMER, https://www.vectorbase.org/tutorials/tools-and-resources-tutorials/hmmer

BLAST will *not* run if you choose an inappropriate program for the selected database, *e.g.*, blastn for a protein query. Hover with your mouse over each program and short description will describe you the type of query and database used in each one.



BLAST provides statistical evaluations that serve as a guide for the evaluation of the alignment scores.

BLAST algorithm is heuristic, thus BLASTn results are not always optimal due to the degeneracy of codons. Its strongly recommended to use BLASTp for homology searches, because protein searches are more specific than DNA searches. Translated BLAST, *i.e.*, tblastn, tblastx[2] and blastx, translates in all six reading frames. These programs are useful when the reading frame is unknown or the query is unknown. The image below shows the six potential reading frames of a sequence.

ATG →Met
TGA → Stop
GAC → Asp
5' - **ATGAC**GAGAGAGCAGCCATTTTAG – 3'

3' – TACTGCTCTCTCGTCGGTA**AAATC** – 5'

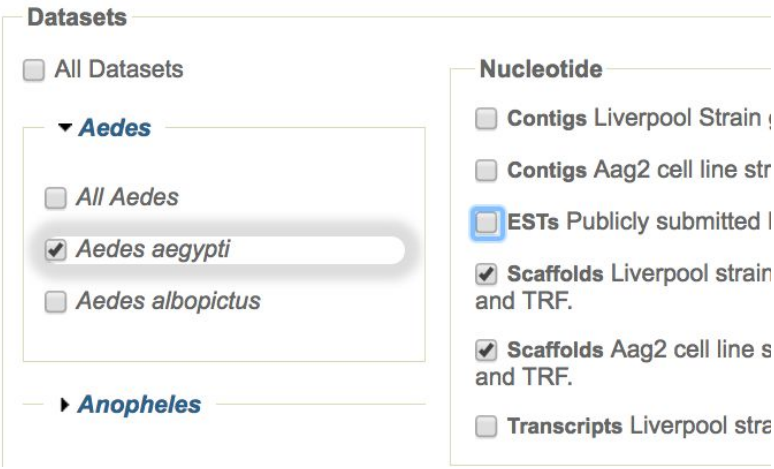
Leu ←ATC
Stop ←AAT
Lys ←AAA

In the forward strand the are three reading frames, and its translations can start at position one, two or three and read in codons (in groups of three nucleotides at the time). Similarly happens in the reverse strand. Which one is the real reading frame? Possibly the one with no stop codons in the open reading frame (ORF). All six protein sequences, from the six reading frames, are used as as query against the target database.

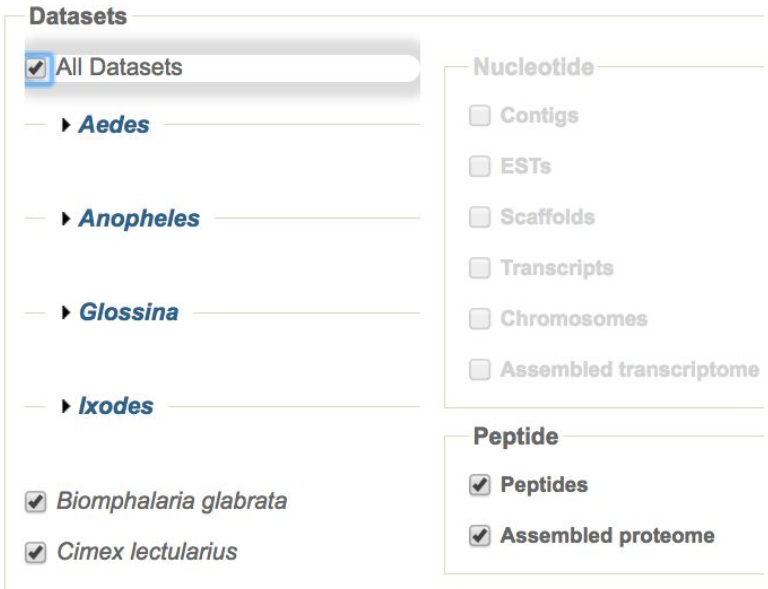
BLASTx takes a DNA query, translates it into a protein sequence and then searches a protein database. tBLASTn translates and entire nucleotide database, in all six reading frames, to create a temporary protein database to search with the protein query. tBLASTx takes a translated DNA query to search a translated DNA database.

[2] tblastx is computationally intensive and should be used only as a last resource. Searching with large genomic queries is not recommended.

To perform a BLAST you can use a single or multiple query sequences, likewise, your target can be single, multiple or all databases, from one or all species. What you *cannot* do is to search using a mixture of nucleotide and peptide queries and/or databases. The allowed searches take place independently and the results will be reported for each database separately.



Target single or multiple datasets in a single species



Target single or multiple datasets in multiple species

There are different default options for each BLAST program. The following is a brief description of each one.

E-value:
Is the number of unrelated sequences in a similarity search of a sequence database that are expected to achieve a local alignment score as high or higher than the one obtained between the query sequence and the matching database sequence. Like with a p-value, the lower the better. Recently, new options for higher e-values were added. *Previously*, our highest e-value allowed was 10. Current options include 100, 1000, 10000 and 100000. You can test this new feature when using protein queries against scaffolds in tBLASTn for the search of (new) highly derived/rapidly evolving genes.

Word size:
The query will be divided up into smaller chunks which will then be compared to the databases(s) separately. Default is 3 for DNA and 11 for protein. A larger words is more efficient proving an exact match.

Scoring Matrix:
This option is available for proteins searches, *i.e.*, all programs except blastn, and are used to evaluate the quality of the alignment with a score, in which the higher the score the better. By default BLOSUM62 is selected (and its calculated from comparisons of sequences were 62% of the similar ones are grouped), but it can be changed depending on the type of sequences you are searching with:

- *BLOSUM* is based on local multiple sequence alignment (MSA) of distantly related proteins. Higher numbers mean smaller evolutionary distance (*i.e.*, higher sequence identity). BLOSUM62 is used for closer sequences than BLOSUM45
- *PAM* is based on global MSA of closely related proteins. Higher numbers mean larger evolutionary distance. PAM70 is used for more distant sequences than PAM30

Complexity Masking:
It removes segments of the query sequence that have low compositional complexity. Filtering leaves the more biologically interesting regions of the query sequence available for specific matching against database sequences. Filtering low-complexity regions from the query sequence helps to reduce the number of false positives.

Number of results:
Limits the output of BLAST to a maximum of the specified numbers of hits and can potentially speed up a search. Default is 10.

2. How to use the tool and interpret its output?

Quick start

Optional: Login to VectorBase. Your BLAST (ClustalW and Hmmer) jobs will be saved and viewable in your user page.

Go to BLAST (from the Tools navigation menu or from the organism pages)

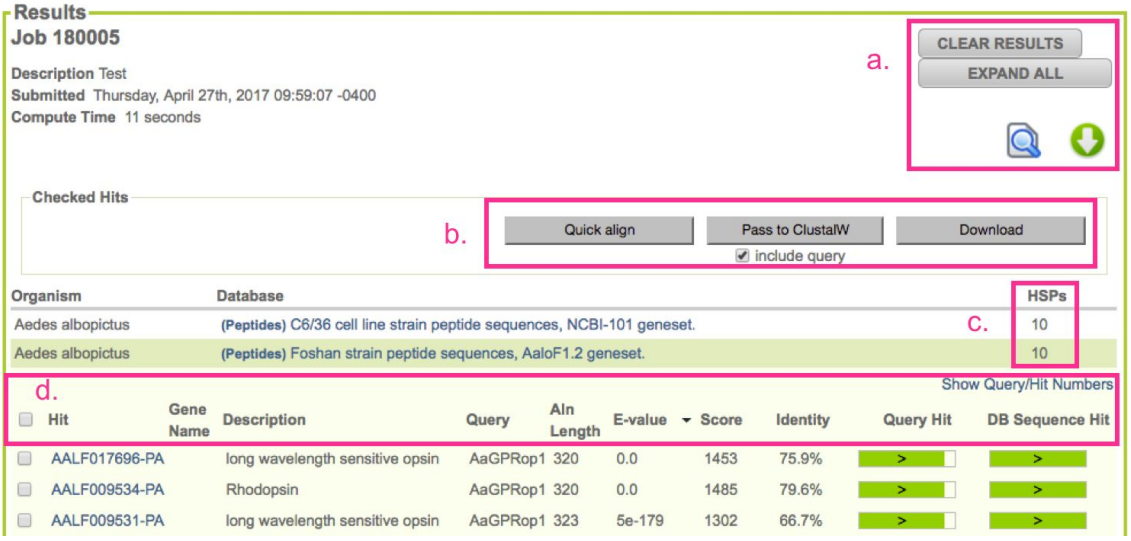
Paste or upload the query sequence(s)

Perform Search of target database(s)

Are matching sequences found with reasonable alignments and significant E values for alignment scores? Depending on your query 100% identical matches or sequences with mismatches but clearly related are of interest.

Output interpretation

The alignments should be examined for a small value of E, absence of low-complexity regions that falsely give high alignment scores, and quality of the alignment (*i.e.*, presence of long stretches of aligned regions that do not depend on gaps to produce an alignment). The results are divided into three sections, shown below as a, b, c and d.



Section a:

- '*Clear Results*' keeps input sequence and parameters, unlike Reset, which also deletes the query.
- 'Expand All (/Sort all)': Allows sorting across all organism hits
- '*View Raw Results*' shows data in plain format on a new tab.
- '*Download Results*' downloads raw results in a text file.

Section b: It only works when hits are selected using the tick box next to them

- '*Quick align*' in the same page a small box shows a Clustal alignment.
- '*Pass to ClustalW*' a new tab is open with the sequences in FASTA format on VectorBase ClustalW.
- '*Download*' text file with sequences in FASTA format.

Section c:

- '*HSPs*' or high-scoring segment pair, is a high-scoring word or alignment between the query and the hits. In protein searches, the HSP is usually three amino acids long. The word will be enlarged if sequence positions neighboring the word also match and provide a higher-scoring alignments.

Section d:
Results are presented in a table format, with each column providing different lines of evidence to select the best hit(s).

Use only high confidence hits, i.e., genomic regions identified with high similarity with your query, for further analysis. The confidence in a hit can be judged by the:

E-value The smaller the better

Score
Identity } The bigger the better
Length

which should be considered jointly.

$$\text{Identity} = \frac{\text{\# of identical letters}}{\text{alignment length (\# matching letters + \# gaps)}}$$

When looking for homologous sequences, only hits with close to 100% identity and approximately the same length as the query should be retained.

For example, hit A is more indicative of common ancestry than hit B, even though the E-value is similar in the two hits.

Hit	E-value	Identity	Alignment Length
A	0.005	50%	100% (whole query)
B	0.006	80%	10% of the query

For example, a 50% identity of a match that extends for the whole protein sequence is more indicative of common ancestry than a 80% identity over a region = 10% of the length of one of the two proteins, without significant similarity in the other 90% region.

In summary. In the BLAST output there are exact matches and sequences with mismatches. For all the hits that are not exact matches, there are calculations that generate statistics that help evaluate which hits are significant.

3. Questions and practice exercises

Question 1. 1

What is BLAST?

	True	False
A program to construct sequence alignments		X
The BLAST algorithm is a way to perform a sequence similarity search	X	

A program to do small and big scale data mining queries such as the ones done with VectorBase Search or Biomart		X

Question 1. 2

What are the five types of VectorBase BLAST programs?

BLASTn, tBLASTn, tBLASTX, BLASTp and BLASTx

Question 1.3

When planning a PCR experiment, it is necessary to know the size of the amplicon products for the gDNA and cDNA, to interpret the obtained results in the agarose gel. This requires a knowledge of the gene structure to localize where the primers will bind.

Using the primer sequences (in sample file provided in the tutorial page), run a BLASTn job for *Ixodes scapularis* 'Scaffolds' and 'Transcripts'.

BLAST

Basic Local Alignment Search Tool

```
>Serp2Ae1-6_F
TTACGCTCCCGACGTTATTC


>Serp2Ae1-6_R
TTCGAGGGATCAAACAGGTC
```

In the BLAST results click on the "Transcripts" database. Are the best hits binding to the gene in the table below?

Table 2. PCR primer sequences used in this study.

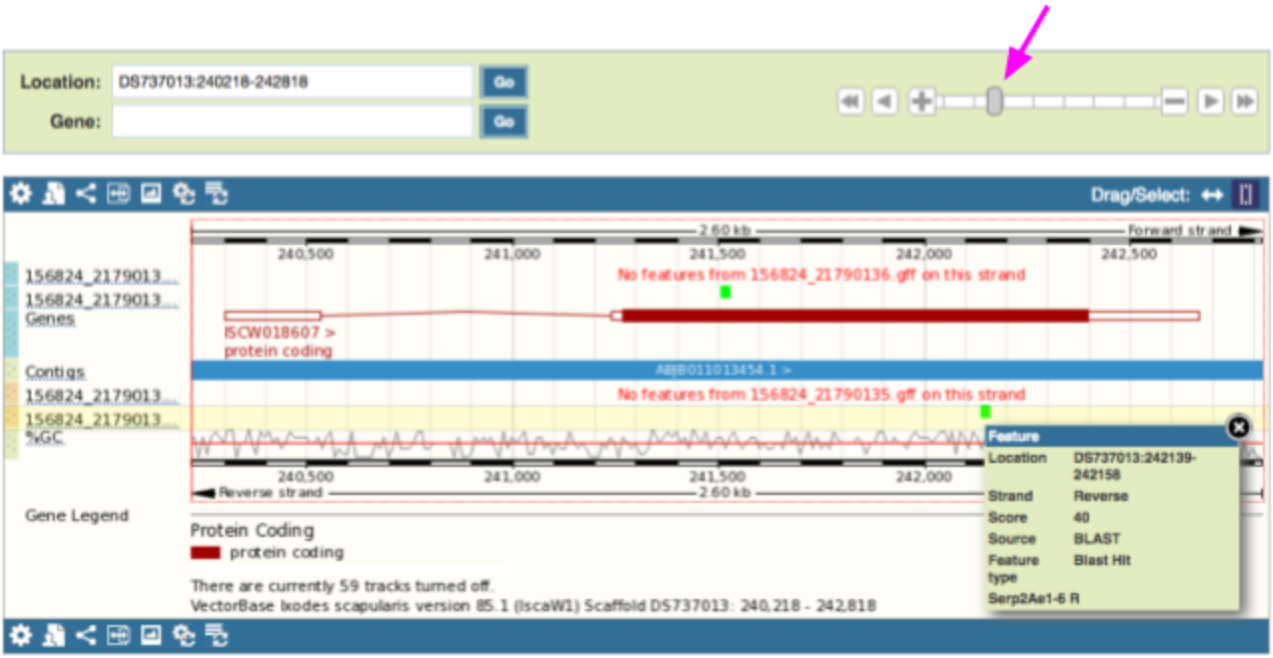
Gene	Primer	5'→3'	Amplicon size	Genbank /Vectorbase
Serpin 2	Serp2Ae1-6_F	TTACGCTCCCGACGTTATTC	651	NW_002630218.1 IscW_ISCW018607
	Serp2Ae1-6_R	TTCGAGGGATCAAACAGGTC		

Yes __X__ No _____

In the BLAST results click on the "Scaffold" database. Click on the first hit. In the popup window click on "Browse Genome". What is the green bar in the bottom display?

	True	False
The gene sequence		X
The scaffold sequence		X
The primer (query) sequence	X	

Repeat previous step but for the second hit. Move the zoom slide bar to 2,600 bp. Click on the green squares to obtain popup windows with details.



Question 1. 4

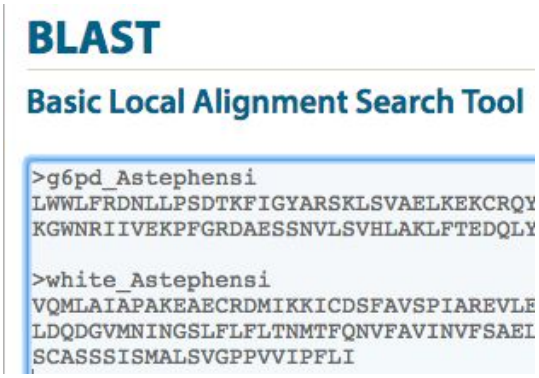
Anopheline mosquitoes are of prime medical importance as vectors of human *Plasmodium*, a parasite that cause malaria, yet *Anopheles* phylogeny is poorly known and there are many species complexes. In 2001, Krzywinski and colleagues published a study where they analyze sequences of two protein-coding single copy genes from Anophelines, *glucose-6-phosphate dehydrogenase* (G_6pd) and *white*. These genes sequences were used in 16 species to determine their phylogenetic relationships.

Syst Biol. 2001 Aug;50(4):540-56.

Toward understanding Anophelinae (Diptera, Culicidae) phylogeny: insights from nuclear single-copy genes and the weight of evidence.

Krzywinski J[1], Wilkerson RC, Besansky NJ.

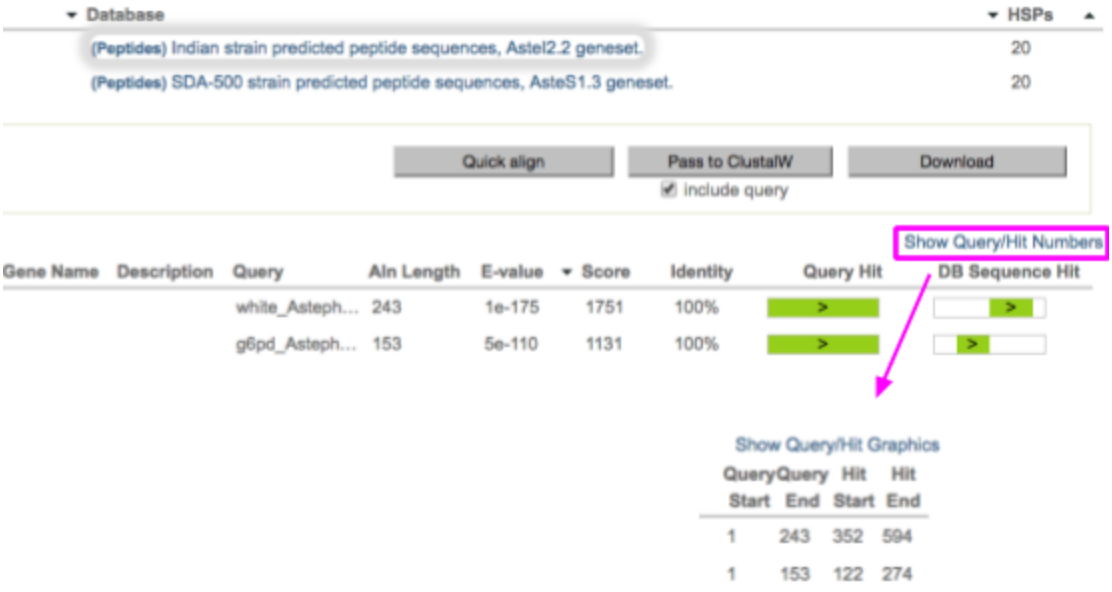
In this exercise, you are going to use BLASTp to identify their VectorBase gene IDs in *An. stephensi*, using the GenBank accession numbers provided in the paper 'Material and Methods', AF317808 and AF318208 for *G_6pd* and *white,* respectively. There is a file with the GenBank sequences in the tutorial page. **Hint**: there is no need to run BLAST twice, you can run a single job with both sequences



	Anopheles stephensi	
	SDA-500 strain	Indian strain
white	ASTE001213	ASTEI04427
G_6pd	ASTE009404	ASTEI04175

Do not close this job results, you need them for the next question.

Are *An. stephensi G_6pd* and *white* sequences from GenBank the same length when compared to VectorBase gene models? **Hint**: look at query and hit graphics (and/or numbers).



No they are not, the ones from Genbank are shorter/incomplete gene sequences.

Question 1.5

A file with the VectorBase gene sequences, AALB010162 and AALB006905, is provided in the tutorial page. Rund a separate BLASTn job for each query, changing to '50' the 'Results per query per database' and select Assembled transcriptome (RNAseq) as the target database. Is there transcript evidence for *An. albimanus, G_6pd* and *white* genes? **Hint**: Remember to check alignment length, e-value, score and identity.

There is transcript evidence, with statistical confidence, only for *G_6pd*.

What the blue and green colors mean in the bars of the query and hit graphics?

The direction of the transcript, forward is green and reverse is blue.

What is the difference between 'assembled transcriptomes' and 'transcripts'?

Transcripts are *in silico* predictions from the genome sequences. The 'assembled transcriptomes' is experimental evidence, in the case of *An. albimanus* from ESTs and RNAseq experiments.

Question 1.6

Similarity or identification searches can also be done using VectorBase data (https://www.vectorbase.org/downloads). Commercial parties have developed enhanced BLAST applications, but its use is not free.

To study the effect of Dengue-2 virus infection in the salivary glands of *Ae. aegypti* mosquitoes, scientist performed an experiment to identify differential expression in this organ. Complete the experimental strategy shown below with VectorBase data.

- Treatment groups: mosquitoes feed with (1) blood and (2) blood + DEN-2
- After a 10-day extrinsic incubation period, the salivary glands were dissected
- Proteins were extracted from the sample tissue and run on a two-dimensional gel electrophoresis
- Differentially expressed proteins (spots) were cut from the gel and used for mass spectrometry
- Protein identification was performed using the commercial software x. Using BLAST the protein spectra were compared against

Aedes-aegypti-Liverpool_PEPTIDES_AaegL3.3	(target database)

  downloaded from VectorBase.
- Results are reported in the paper using as stable accession

VectorBase gene IDs

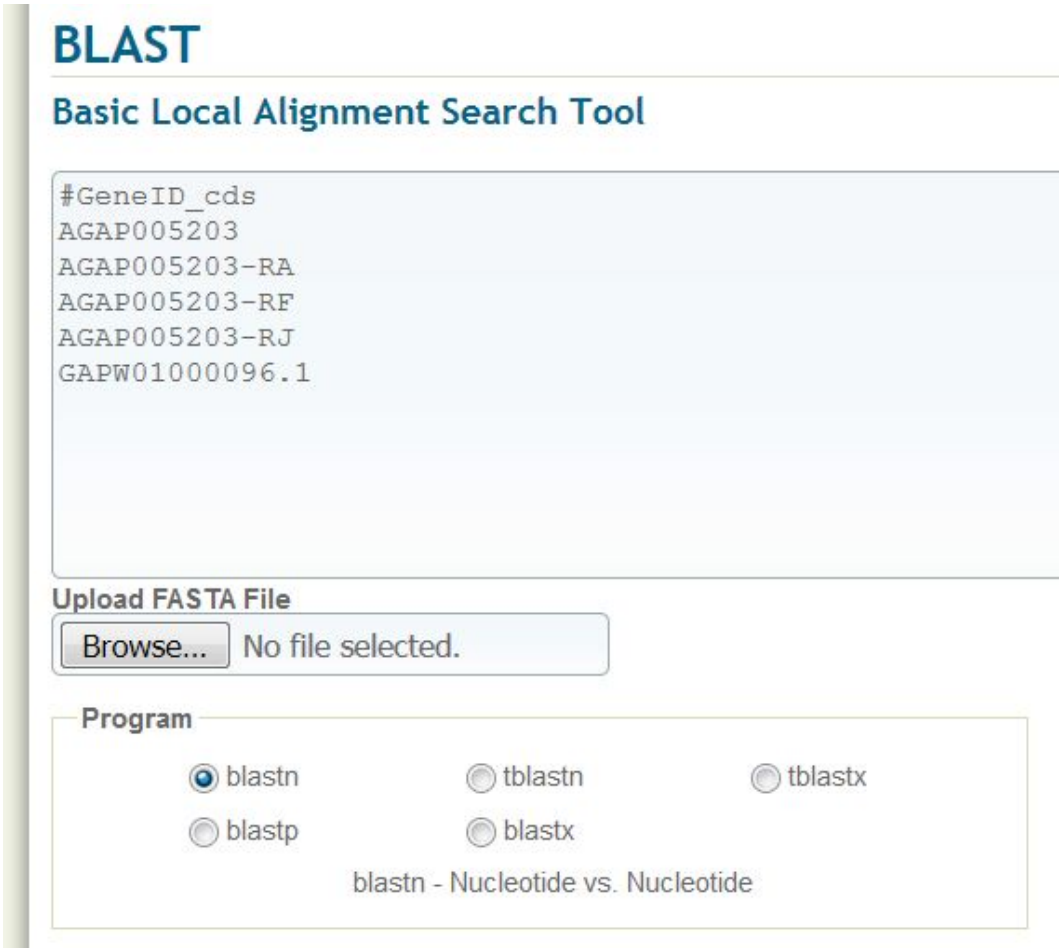
- For genes with unknown symbol or description GenBank non-redundant protein sequences (nr) was used for a second run of similarity searches.

Question 1.7 **Gene ID lookup** new!

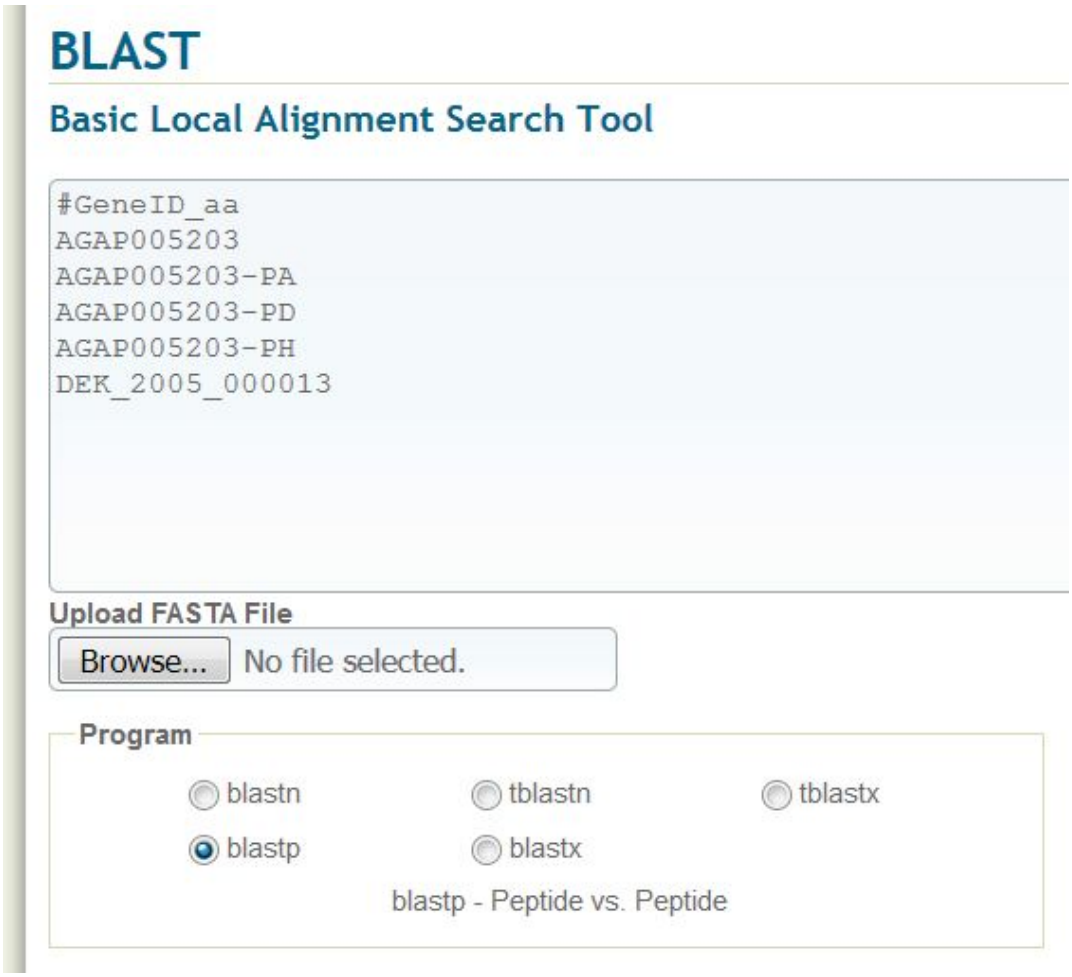
In conjunction with inputting sequence(s) through either the text area or file upload in Blast, a new feature called **Gene ID lookup** in Blast is introduced. In lieu of inputting sequences, one can input Gene ID(s) through either Textarea or File upload. To distinguish between Nucleotide (cds type) and Amino-Acid (aa type) sequences, the first line of the input should be either **#GeneID_cds** (for Nucleotide sequences) or **#GeneID_aa** (for Amino-Acid

sequences): please refer to below figures. Note that one should select correct Program to run Blast depending on _cds or _aa. After inputting Gene IDs and submitting Blast job, sequences corresponding to Gene ID input(s) are retrieved and Blast will be automatically executed based on the sequences.

Example #1) _cds input



Example #2) _aa input



If you need help with any question and its answer contact us at info@vectorbase.org. Because VectorBase data, tools and resources are updated every two months (6 release cycles per year), answers to these exercises will change too.