Appendix. Understanding important sections of a BLAST Results page

Top Panels

The top left panel (Figure 1) of the BLAST results page shows the parameters used in the BLAST search (e.g., database name, query ID, query length). This is always a good reference to remind yourself what your search conditions were. The controls in the top right panel (Figure 1) can be used to filter the BLAST hits by organism, percent identity, and Expect value (E-value). We typically will not directly refer to these panels in our labs.

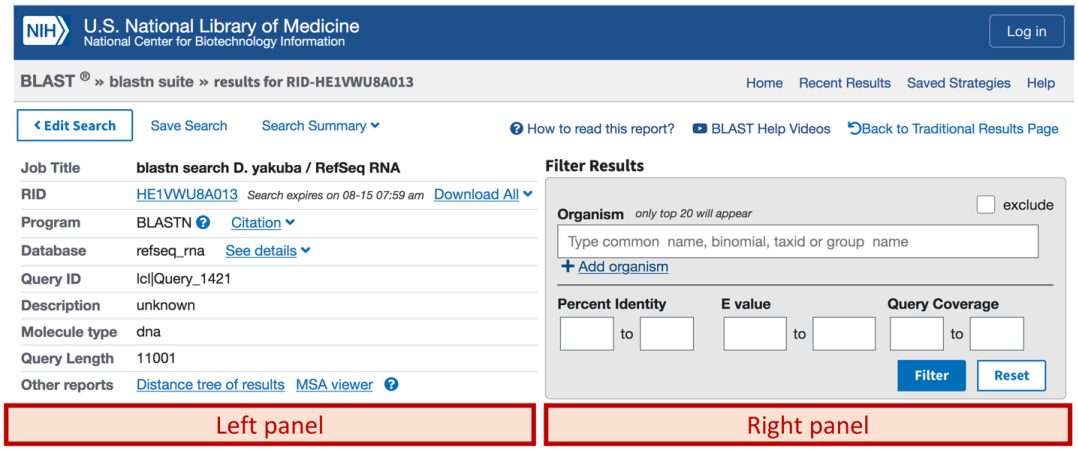


Figure 1. The parameters used by the BLAST search are listed in the top left panel of the BLAST results page. The controls for filtering the BLAST search results are available in the top right panel.

The details of the BLAST results are organized into the four tabs below these two panels: “Descriptions”, “Graphic Summary”, “Alignments”, and “Taxonomy”. We will go through the “Descriptions” and “Alignments” tabs in this appendix, as the other two tabs contain information not directly relevant to our lab project.

Descriptions tab

The ‘Descriptions’ tab shows the list of sequences in the database that have significant sequence homology with our sequence (Figure 2). By default, the results are sorted by their E-value in ascending order, where lower E-values denote more significant hits. You can click on the column headers to sort the results by the other columns. You can also use the “Select columns” drop-down menu on the main toolbar to show or hide each column.

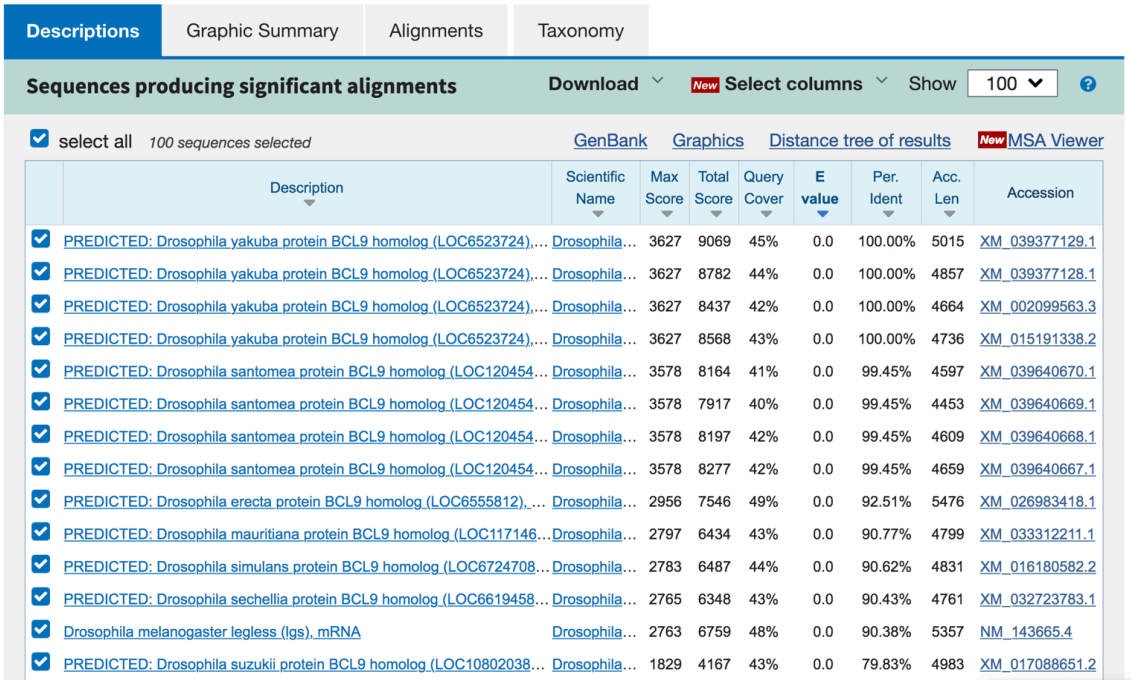
The columns indicate the name of the sequence retrieved (Description), the name of the species from which it derived (Scientific name), a score indicating the best alignment between the query sequence and the given database sequence segment (max score), the sum of alignment scores for all segments of the same database sequence that matches the query sequence, the amount of the query sequence that matches to the given database sequence segment (Query Cover), the number of hits one expects to see by chance from search in the

provided database (E value, depends on number of sequences and size of sequences in that database), the percent of bases that are the same (Per[cent] Identity), the number of bases in the sequence that was found to match (Acc[ession] Len[gth]), and the unique ID associated with the sequence that was found to match (Accession). We usually do not look at Max or Total Score, but we examine E value, Percent Identity, and Accession often.

A low E value (much smaller than one, and sometimes so small it looks like 0.0) is an important metric to tell us how much we trust the hit. An E value equal to one means that a hit like the observed one is expected to occur by random chance at least once in a search of the database from which it came. **An E value much smaller than one means there is a very low chance that the observed hit is due to random chance. Thus, the similarity is likely meaningful (and hopefully meaningful in a biological sense!). To pick the best hit, we usually look for entries with low E values and high percent identity.**

For the Accession column, where the unique accession ID associated with each sequence that was matched is located, note the structure of the IDs, either “XM_##” or “NM_##” (Figure 2). The main difference between these two prefixes is the type of information available to support the RefSeq mRNAs. The “NM_##” prefix indicates that the RefSeq mRNA record is supported by experimental evidence, whereas the “XM_##” prefix indicates that the record is based solely on computational predictions. Because we would prefer to base our inferences on a gene model that is supported by experimental evidence, we typically will focus on hits with the “NM_##” prefix rather than the “XM_##” prefix. The two types described here are specifically for mRNAs - as we do more different types of searches, you will see other accession ID structures as well.

Note that if you want to select only a single entry amongst all the potential hits, unclick ‘select all’ and then click the square next to the hit in which you are interested. Clicking on the accession number in the table will bring up a new page with the GenBank record of the sequence. Clicking on the description of the hit will bring us to the corresponding alignment in the BLAST output. Alternatively, you can click on the “Alignments” tab to jump to the first alignment.



Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download | New Select columns | Show 100

select all *100 sequences selected* GenBank | Graphics | Distance tree of results | New MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
PREDICTED: Drosophila yakuba protein BCL9 homolog (LOC6523724),...	Drosophila...	3627	9069	45%	0.0	100.00%	5015	XM_039377129.1
PREDICTED: Drosophila yakuba protein BCL9 homolog (LOC6523724),...	Drosophila...	3627	8782	44%	0.0	100.00%	4857	XM_039377128.1
PREDICTED: Drosophila yakuba protein BCL9 homolog (LOC6523724),...	Drosophila...	3627	8437	42%	0.0	100.00%	4664	XM_002099563.3
PREDICTED: Drosophila yakuba protein BCL9 homolog (LOC6523724),...	Drosophila...	3627	8568	43%	0.0	100.00%	4736	XM_015191338.2
PREDICTED: Drosophila santomea protein BCL9 homolog (LOC120454...	Drosophila...	3578	8164	41%	0.0	99.45%	4597	XM_039640670.1
PREDICTED: Drosophila santomea protein BCL9 homolog (LOC120454...	Drosophila...	3578	7917	40%	0.0	99.45%	4453	XM_039640669.1
PREDICTED: Drosophila santomea protein BCL9 homolog (LOC120454...	Drosophila...	3578	8197	42%	0.0	99.45%	4609	XM_039640668.1
PREDICTED: Drosophila santomea protein BCL9 homolog (LOC120454...	Drosophila...	3578	8277	42%	0.0	99.45%	4659	XM_039640667.1
PREDICTED: Drosophila erecta protein BCL9 homolog (LOC6555812), ...	Drosophila...	2956	7546	49%	0.0	92.51%	5476	XM_026983418.1
PREDICTED: Drosophila mauritiana protein BCL9 homolog (LOC117146...	Drosophila...	2797	6434	43%	0.0	90.77%	4799	XM_033312211.1
PREDICTED: Drosophila simulans protein BCL9 homolog (LOC6724708...	Drosophila...	2783	6487	44%	0.0	90.62%	4831	XM_016180582.2
PREDICTED: Drosophila sechellia protein BCL9 homolog (LOC6619458...	Drosophila...	2765	6348	43%	0.0	90.43%	4761	XM_032723783.1
Drosophila melanogaster legless (lgs), mRNA	Drosophila...	2763	6759	48%	0.0	90.38%	5357	NM_143665.4
PREDICTED: Drosophila suzukii protein BCL9 homolog (LOC10802038...	Drosophila...	1829	4167	43%	0.0	79.83%	4983	XM_017088651.2

Figure 2. List of *blastn* hits that produce significant alignments with our query sequence.

Alignments tab

The 'Alignments' tab contains the alignments between the selected BLAST hits in the Descriptions tab and the query sequence. The sequence alignments show us how well our query sequence matches the subject sequence in the database. Alignments to different subject sequences in the database are separated by a blue toolbar that contains options to manipulate the alignment results and to retrieve additional information for that specific BLAST hit. For example, we can use the "Download" drop-down menu on this toolbar to obtain the FASTA sequence or the GenBank record for a specific hit. We can use the navigation links at the right side of the toolbar to quickly navigate to the next or the previous BLAST hit.

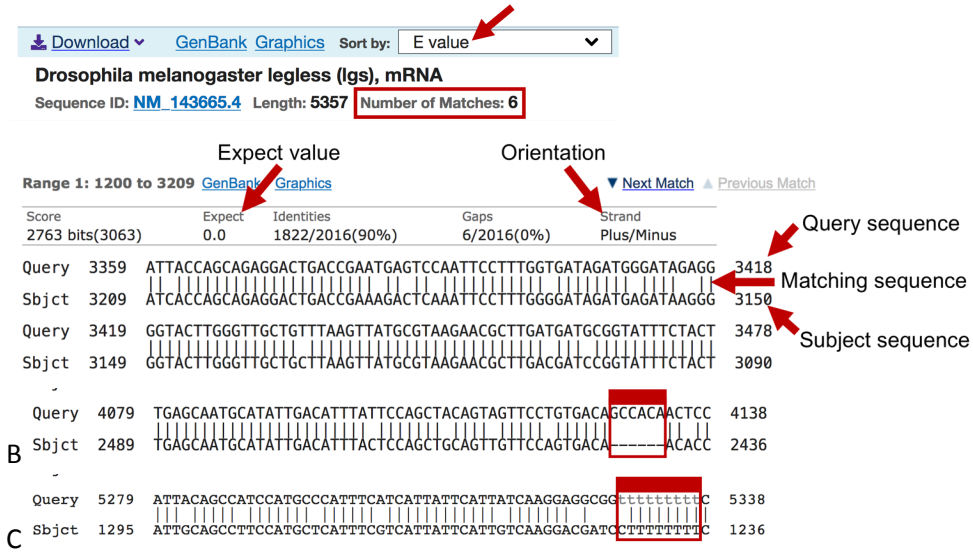
As its name suggests, BLAST is designed to identify local regions of sequence similarity. This means that BLAST might report multiple distinct regions of sequence similarity when we align a query against a subject sequence in a database. For example, if we were to align a processed mRNA sequence to a genomic sequence, we would expect to see multiple alignment blocks (many of which correspond to transcribed exons) in our BLAST output. Each alignment block demarcates a local region of similarity between the query and the subject sequences. Regions of the genomic sequence without significant alignments that fall between these alignment blocks would likely correspond to intronic sequences.

The "Number of Matches" field beneath the name of the sequence shows the number of alignment blocks identified by *blastn*. For example, the *blastn* hit for the *legless* mRNA from *D. melanogaster* contains 6 different alignment blocks to the subject sequence — (Figure 2A). Each alignment block represents a region of the *D. melanogaster legless* gene that shows sequence homology with our genomic sequence from *D. yakuba*. You can use the "Sort by" drop-down box (red arrow in Figure 2A) on the toolbar above each BLAST hit to sort the alignment blocks based on different criteria (e.g., by E-value, query start position, subject start position). Each alignment block begins with a line that has the following format: "Range #:start to end" (where # is the alignment block number). You can use the "Next

Match" and "Previous Match" links to navigate to the different alignment blocks within the same BLAST hit.

Each alignment block begins with a summary, including the Expect value (i.e., E-value, or the statistical significance of the alignment), sequence identity (number of identical bases between the query and the subject sequence), the number of gaps in the alignment, and the orientation of the query relative to the subject sequence. The alignment consists of three lines: the query sequence, the matching sequence, and the subject sequence (Figure 2A). The **-** character in either the query or the subject sequence denotes a gap in the alignment (Figure 2B). By default, NCBI BLAST automatically masks low complexity sequences in the query sequence. Depending on your BLAST search settings, these masked bases may appear as either grey lowercase letters (Figure 2C) or as X's. The matching sequence consists of a combination of **|** and empty spaces, where | denotes a matching base between the query and subject sequences and the empty space denotes a mismatched base.

A



B

C

Figure 2. (A) The key characteristics of a typical BLAST alignment. (B) Gaps in the alignment are represented by the '-' character. (C) Bases masked by the low complexity filter appear as lowercase gray letters by default. Remember that the query sequence is the one you started with, whereas the subject sequence is the one you're comparing against. The subject sequence comes from the database or sequence you are searching against.