



An Introduction to NCBI BLAST

Wilson Leung

Prerequisites

- [Detecting and Interpreting Genetic Homology: Lecture Notes on Alignment](#)

Resources & Tools

- BLAST web server: <https://blast.ncbi.nlm.nih.gov/Blast.cgi>
- Gene Record Finder: <https://thegep.org/finder>
- The [package](#) containing the files for this walkthrough are available through the “[An Introduction to NCBI BLAST](#)” page on the GEP website.

Table of Contents

Introduction.....	2
The NCBI BLAST web interface.....	2
Detecting sequence homology to mRNA using <i>blastn</i>	4
I. Descriptions.....	7
II. Graphic Summary.....	8
III. Alignments.....	9
IV. Taxonomy.....	13
Interpreting the <i>blastn</i> search result.....	15
Detecting Coding Regions Using <i>blastx</i>	16
Define the Intron-Exon Boundaries with <i>Gene Record Finder</i> and <i>bl2seq</i>	20
Define the 5' UTR of <i>legless</i> Using <i>blastn</i>	24
Conclusion	28

Introduction

The Basic Local Alignment Search Tool (BLAST) is a program that can detect sequence similarity between a query sequence and sequences within a database. The ability to detect sequence homology allows us to identify putative genes in a novel sequence. It also allows us to determine if a gene or a protein is related to other known genes or proteins.

BLAST is popular because it can quickly identify regions of *local similarity* between two sequences. More importantly, BLAST uses a robust statistical framework that can determine if the alignment between two sequences is statistically significant. In this walkthrough, we will use the National Center for Biotechnology Information (NCBI) BLAST service to help us annotate a sequence from the *Drosophila yakuba* genome (*unknown.fna* in the walkthrough package).

The NCBI BLAST web interface

Before we begin the analysis, we should first familiarize ourselves with the NCBI BLAST web interface. Open a new web browser window and navigate to the [NCBI BLAST main page](#). In this walkthrough, we will only use a few of the tools available on the NCBI BLAST website. To learn about the more advanced options available (such as setting up My NCBI accounts), click on the “Help” link on the main navigation bar to access the documentations for NCBI BLAST (Figure 1).

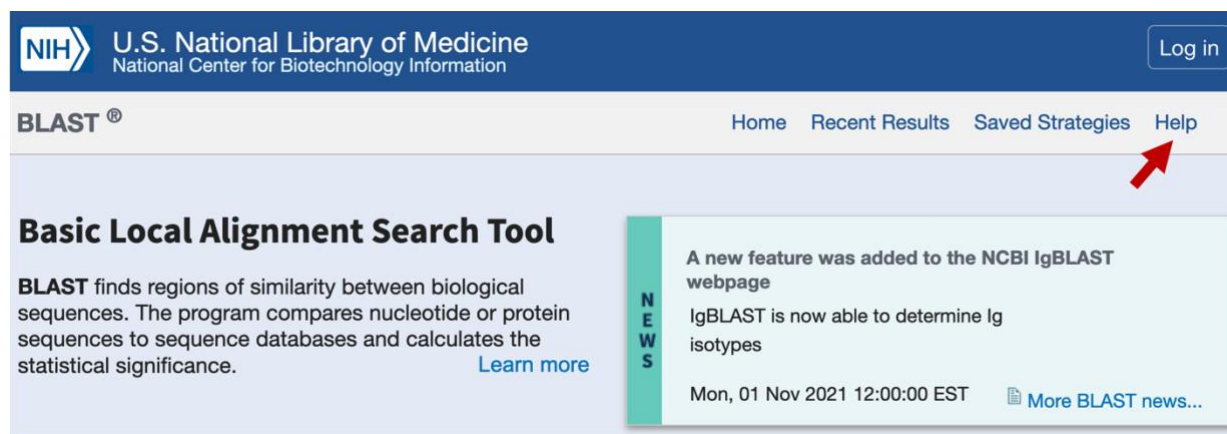


Figure 1. Click on the “Help” link to learn more about the NCBI BLAST web interface.

All the NCBI BLAST pages have the same header with four links:

Links	Explanation
Home	Link to the NCBI BLAST home page
Recent Results	Link to results of the BLAST searches you have run previously
Saved Strategies	NCBI BLAST search parameters you have previously saved to your My NCBI account
Help	Documentation for NCBI BLAST

Besides the main toolbar, there are two other sections of the NCBI BLAST web interface that are of interests: the “Web BLAST” section contains links to the common BLAST programs and the “Specialized searches” section contains links to additional tools for performing sequence searches (e.g., use CD-search to identify conserved domains within a query sequence). The type of BLAST search you need to use will depend primarily on the type of query sequence and the database you would like to search.

Four of the five common BLAST programs are available through the “Web BLAST” section of the NCBI BLAST home page (Figure 2, top). The program *tblastx*, which translates the nucleotide query and nucleotide database when it performs the sequence comparisons, is not listed under the “Web BLAST” section. However, you can access this program by clicking on any of the BLAST programs in the “Web BLAST” section and then click on the “*tblastx*” tab in the NCBI BLAST search form (Figure 2, bottom).

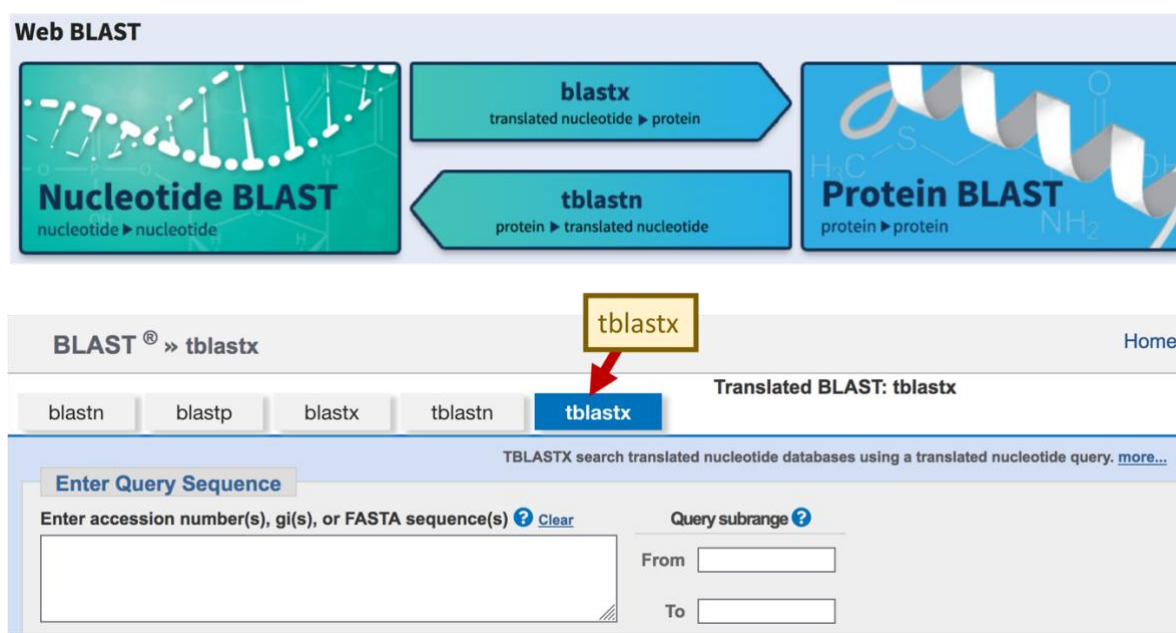


Figure 2. The different BLAST programs available through the NCBI web server home page (top). The *tblastx* program is available through the “*tblastx*” tab in the NCBI BLAST search form (bottom).

The basic BLAST programs are summarized below:

BLAST program	Query	Database
Nucleotide BLAST (<i>blastn</i>)	Nucleotide	Nucleotide
Protein BLAST (<i>blastp</i>)	Protein	Protein
<i>blastx</i>	Translated Nucleotide	Protein
<i>tblastn</i>	Protein	Translated Nucleotide
<i>tblastx</i>	Translated Nucleotide	Translated Nucleotide

Instead of searching a query sequence against sequences in a database, you can also align two (or more) sequences by selecting the “Align two or more sequences” checkbox at the bottom of the “Enter Query Sequence” section (Figure 3). This feature is also known as *BLAST 2 Sequences* (*bl2seq*).

The screenshot shows the NCBI BLAST tblastx interface. At the top, the NIH logo and 'U.S. National Library of Medicine National Center for Biotechnology Information' are displayed. Below this, the 'BLAST' logo and 'tblastx' are shown. The 'Align Sequences Translated BLAST: tblastx' tab is selected. The 'Enter Query Sequence' section includes a text input for 'Enter accession number(s), gi(s), or FASTA sequence(s)', a 'Query subrange' section with 'From' and 'To' inputs, and a 'Browse...' button for uploading a file. The 'Genetic code' is set to 'Standard (1)'. The 'Job Title' field is empty. A red arrow points to the 'Align two or more sequences' checkbox, which is checked. The 'Enter Subject Sequence' section has similar input fields. At the bottom, the 'BLAST' button is visible, along with a checkbox for 'Show results in a new window'.

Figure 3. Select the “Align two or more sequences” checkbox to compare a query sequence against a subject sequence instead of a BLAST database.

Detecting sequence homology to mRNA using *blastn*

In this walkthrough, we will characterize an unknown genomic sequence (*unknown.fna*) and determine if it has sequence similarity to any known genes. One strategy we can use is to search for sequence similarity to mRNA sequences in the NCBI Reference Sequence (RefSeq) database.

When we set up a BLAST search, there are three basic decisions we must make: the BLAST program we want to use, the query sequence we want to annotate, and the database we want to search. In addition, we can change several optional parameters (such as the Expect threshold and low complexity filters) in order to modify the behavior of BLAST.

In this case, we will set up our BLAST search using mostly default parameters. We will use the *blastn* program to search our sequence (query) against the NCBI Reference Sequence (RefSeq) RNA database (Figure 4).

1. Navigate to the NCBI BLAST home page and click on the “Nucleotide BLAST” image under the “Web BLAST” section
2. Under the “Enter Query Sequence” section, click on the “Browse” or the “Choose File” button and select the file with the unknown sequence (unknown.fna)
3. Enter the Job Title “*blastn* search *D. yakuba* / RefSeq RNA”
4. In the “Choose Search Set” section, change the database to “Reference RNA sequences (refseq_rna)”
5. Under “Program Selection”, select “Somewhat similar sequences (*blastn*)”
6. Check the box “Show results in a new window” next to the “BLAST” button
7. Click “BLAST”

Standard Nucleotide BLAST

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [?](#) [Clear](#)

unknown.fna

Query subrange [?](#)

From

To

Or, upload file [Browse...](#) **unknown.fna** [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#)

☐ Align two or more sequences [?](#)

Choose Search Set

Database ☒ Standard databases (nr etc.): ☐ rRNA/ITS databases ☐ Genomic + transcript databases ☐ Betacoronavirus

Reference RNA sequences (refseq_rna) [?](#) **refseq_rna database**

Organism [Optional](#) ☐ exclude [Add organism](#)

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](#) [?](#) [YouTube](#) [Create custom database](#)

Program Selection

Optimize for ☐ Highly similar sequences (megablast) ☐ More dissimilar sequences (discontiguous megablast) ☒ Somewhat similar sequences (blastn)

Choose a BLAST algorithm [?](#)

BLAST Search database **Reference RNA sequences (refseq_rna)** using **Blastn** (Optimize for somewhat similar sequences)

☒ Show results in a new window

Figure 4. Setting up our *blastn* search of the unknown sequence against the NCBI RefSeq RNA database.

Note: the *blastn* search may take a few minutes to complete if the NCBI web server is busy (Figure 5).

NIH U.S. National Library of Medicine
National Center for Biotechnology Information

BLAST® » **blastn suite** » RID-VZG3U7ZE013

Home Recent Results Saved Strategies Help

Format Request Status

[\[Formatting options\]](#)

Job Title: **blastn search D. yakuba / RefSeq RNA**

Request ID	VZG3U7ZE013
Status	Searching
Submitted at	Sun Dec 19 18:30:02 2021
Current time	Sun Dec 19 18:30:51 2021
Time since submission	00:00:49

This page will be automatically updated in 22 seconds

Figure 5. Waiting for the *blastn* search results

Once the search is complete, a new web page will appear with the BLAST report. For teaching purposes, the BLAST output (*blastnInitial.txt*) is available in the package for this walkthrough.

The top left panel (Figure 6) of the BLAST results page shows the parameters used in the BLAST search (e.g., database name, query ID, query length). The controls in the top right panel (Figure 6) can be used to filter the BLAST hits by organism, percent identity, and Expect value (E-value).

NIH U.S. National Library of Medicine
National Center for Biotechnology Information

BLAST® » **blastn suite** » results for RID-VZG3U7ZE013

Home Recent Results Saved Strategies Help

[< Edit Search](#) [Save Search](#) [Search Summary](#) [How to read this report?](#) [BLAST Help Videos](#) [Back to Traditional Results Page](#)

Job Title: **blastn search D. yakuba / RefSeq RNA**

RID: [VZG3U7ZE013](#) Search expires on 12-21 06:30 am [Download All](#)

Program: BLASTN [Citation](#)

Database: refseq_rna [See details](#)

Query ID: lcl|Query_64179

Description: unknown

Molecule type: dna

Query Length: 11001

Other reports: [Distance tree of results](#) [MSA viewer](#)

Filter Results

Organism only top 20 will appear ☐ exclude

Type common name, binomial, taxid or group name

[+ Add organism](#)

Percent Identity: to E value: to Query Coverage: to

[Filter](#) [Reset](#)

Left panel Right panel

Figure 6. 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 each of these sections in order to interpret our *blastn* output.

I. Descriptions

This tab shows the list of sequences in the database that have significant sequence homology with our sequence (Figure 7). 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.

Descriptions		Graphic Summary		Alignments		Taxonomy																	
Sequences producing significant alignments						Download		New Select columns		Show		100											
<input checked="" type="checkbox"/> 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														
<input checked="" type="checkbox"/>	PREDICTED: <i>Drosophila yakuba</i> protein BCL9 homolog (LOC6523724)...	Drosophila ...	3627	9070	45%	0.0	100.00%	5016	XM_039377129.2														
<input checked="" type="checkbox"/>	PREDICTED: <i>Drosophila yakuba</i> protein BCL9 homolog (LOC6523724)...	Drosophila ...	3627	9226	46%	0.0	100.00%	5102	XM_002099563.4														
<input checked="" type="checkbox"/>	PREDICTED: <i>Drosophila yakuba</i> protein BCL9 homolog (LOC6523724)...	Drosophila ...	3627	9358	47%	0.0	100.00%	5174	XM_015191338.3														
<input checked="" type="checkbox"/>	PREDICTED: <i>Drosophila santomea</i> protein BCL9 homolog (LOC120454...	Drosophila ...	3578	8895	45%	0.0	99.45%	5011	XM_039640670.2														
<input checked="" type="checkbox"/>	PREDICTED: <i>Drosophila santomea</i> protein BCL9 homolog (LOC120454...	Drosophila ...	3578	8972	46%	0.0	99.45%	5048	XM_039640668.2														
<input checked="" type="checkbox"/>	PREDICTED: <i>Drosophila santomea</i> protein BCL9 homolog (LOC120454...	Drosophila ...	3578	9051	46%	0.0	99.45%	5097	XM_039640667.2														
<input checked="" type="checkbox"/>	PREDICTED: <i>Drosophila santomea</i> protein BCL9 homolog (LOC120454...	Drosophila ...	3578	9122	46%	0.0	99.45%	5156	XM_039640669.2														
<input checked="" type="checkbox"/>	PREDICTED: <i>Drosophila teissieri</i> protein BCL9 homolog (LOC12262284...	Drosophila ...	3315	8220	46%	0.0	96.48%	5070	XM_043801420.1														
<input checked="" type="checkbox"/>	PREDICTED: <i>Drosophila teissieri</i> protein BCL9 homolog (LOC12262284...	Drosophila ...	3315	8378	48%	0.0	96.48%	5241	XM_043801419.1														
<input checked="" type="checkbox"/>	PREDICTED: <i>Drosophila teissieri</i> protein BCL9 homolog (LOC12262284...	Drosophila ...	3315	8384	48%	0.0	96.48%	5244	XM_043801418.1														
<input checked="" type="checkbox"/>	PREDICTED: <i>Drosophila erecta</i> protein BCL9 homolog (LOC6555812)...	Drosophila ...	2956	7546	49%	0.0	92.51%	5476	XM_026983418.1														
<input checked="" type="checkbox"/>	PREDICTED: <i>Drosophila mauritiana</i> protein BCL9 homolog (LOC117146...	Drosophila ...	2797	6434	43%	0.0	90.77%	4799	XM_033312211.1														
<input checked="" type="checkbox"/>	PREDICTED: <i>Drosophila simulans</i> protein BCL9 homolog (LOC6724708...	Drosophila ...	2783	6788	49%	0.0	90.62%	5332	XM_016180582.3														
<input checked="" type="checkbox"/>	PREDICTED: <i>Drosophila sechellia</i> protein BCL9 homolog (LOC6619458...	Drosophila ...	2765	6348	43%	0.0	90.43%	4761	XM_032723783.1														
<input checked="" type="checkbox"/>	Drosophila melanogaster legless (lgs), mRNA	Drosophila ...	2763	6759	48%	0.0	90.38%	5357	NM_143665.4														
<input checked="" type="checkbox"/>	PREDICTED: <i>Drosophila suzukii</i> protein BCL9 homolog (LOC108020387...	Drosophila ...	1829	4167	43%	0.0	79.83%	4983	XM_017088651.2														

Figure 7. List of *blastn* hits that produce significant alignments with our query sequence.

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.

In addition to reviewing the records for individual sequences, you can also review multiple sequence records by selecting the checkbox next to each match. The contents of the other tabs will update automatically based on your selection. You can use the “Download” drop-down menu on the main toolbar to download the selected hits in multiple formats (e.g., FASTA, GenBank, Hit Table). For example, we can use the following steps to retrieve the GenBank records for the first five *blastn* hits in the Descriptions table (Figure 8).

1. Uncheck the “select all” checkbox above the BLAST hit table
2. Select the checkboxes for the first five *blastn* hits
3. Click on the “Download” drop-down menu on the main toolbar, and then select the “GenBank (complete sequence)” option

The screenshot shows the NCBI BLAST interface with the 'Descriptions' tab selected. The 'Sequences producing significant alignments' section shows a list of sequences. The first five sequences are checked, and the 'Download' dropdown menu is open, highlighting 'GenBank (complete sequence)'. The table below shows the details of these sequences.

Description	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/> PREDICTED: <i>Drosophila yakuba</i> protein BCL9 homolog (LOC6523724).	0.0	100.00%	5016	XM_039377129.2
<input checked="" type="checkbox"/> PREDICTED: <i>Drosophila yakuba</i> protein BCL9 homolog (LOC6523724).	0.0	100.00%	5102	XM_002099563.4
<input checked="" type="checkbox"/> PREDICTED: <i>Drosophila yakuba</i> protein BCL9 homolog (LOC6523724).	0.0	100.00%	5174	XM_015191338.3
<input checked="" type="checkbox"/> PREDICTED: <i>Drosophila santomea</i> protein BCL9 homolog (LOC120454).	0.0	99.45%	5011	XM_039640670.2
<input checked="" type="checkbox"/> PREDICTED: <i>Drosophila santomea</i> protein BCL9 homolog (LOC120454).	0.0	99.45%	5048	XM_039640668.2
<input type="checkbox"/> PREDICTED: <i>Drosophila santomea</i> protein BCL9 homolog (LOC120454).	0.0	99.45%	5097	XM_039640667.2
<input type="checkbox"/> PREDICTED: <i>Drosophila santomea</i> protein BCL9 homolog (LOC120454).	0.0	99.45%	5156	XM_039640669.2
<input type="checkbox"/> PREDICTED: <i>Drosophila teissieri</i> protein BCL9 homolog (LOC12262284).	0.0	96.48%	5070	XM_043801420.1
<input type="checkbox"/> PREDICTED: <i>Drosophila teissieri</i> protein BCL9 homolog (LOC12262284)...	0.0	96.48%	5241	XM_043801419.1
<input type="checkbox"/> PREDICTED: <i>Drosophila teissieri</i> protein BCL9 homolog (LOC12262284)...	0.0	96.48%	5244	XM_043801418.1
<input type="checkbox"/> PREDICTED: <i>Drosophila erecta</i> protein BCL9 homolog (LOC6555812)...	0.0	92.51%	5476	XM_026983418.1
<input type="checkbox"/> PREDICTED: <i>Drosophila mauritiana</i> protein BCL9 homolog (LOC117146)...	0.0	90.77%	4799	XM_033312211.1
<input type="checkbox"/> PREDICTED: <i>Drosophila simulans</i> protein BCL9 homolog (LOC6724708)...	0.0	90.62%	5332	XM_016180582.3
<input type="checkbox"/> PREDICTED: <i>Drosophila sechellia</i> protein BCL9 homolog (LOC6619458)...	0.0	90.43%	4761	XM_032723783.1
<input type="checkbox"/> <i>Drosophila melanogaster</i> legless (lgs), mRNA	0.0	90.38%	5357	NM_143665.4
<input type="checkbox"/> PREDICTED: <i>Drosophila suzukii</i> protein BCL9 homolog (LOC108020387)...	0.0	79.83%	4983	XM_017088651.2

Figure 8. Click on the “GenBank (complete sequence)” link under the “Download” drop-down menu to retrieve the GenBank records for the five selected mRNA sequences.

II. Graphic Summary

This tab provides a graphical overview of the alignments between the selected BLAST hits in the Descriptions tab and the query sequence. The boxes correspond to regions in the query that have sequence similarity to the sequences in the database. The color of the box corresponds to the score, where hits with higher scores are more significant. When you move your mouse over a BLAST hit, the title of the subject sequence will appear in a tooltip. Click on the color box and then click on the “Alignment” link to jump to the alignments associated with that BLAST hit.

To examine the graphical overview for all the *blastn* hits, go back to the “Descriptions” tab and then select the “select all” checkbox. Click on the “Graphic Summary” tab to view the updated graphical overview (Figure 9).



Figure 9. The “Graphic Summary” tab shows the graphical overview for the selected BLAST hits in the “Descriptions” tab. Select the “select all” checkbox in the “Descriptions” tab and then navigate back to the “Graphic Summary” tab to view the graphical overview for all the BLAST hits.

III. Alignments

This 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 match the subject sequence in the database. Because we will rely on sequence alignments heavily in our annotation efforts, we will examine this Alignment tab more closely.

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 (Figure 10). 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.

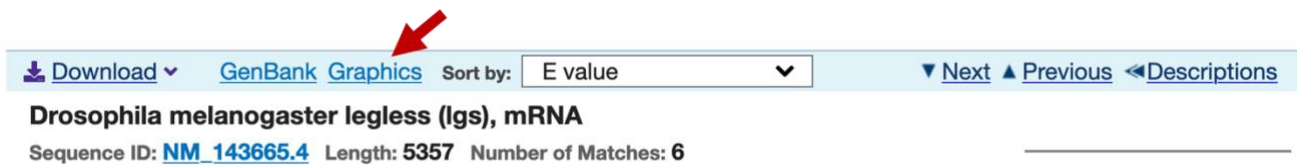


Figure 10. Alignments to different subject sequences in the database are separated by a blue toolbar with options to manipulate and download the alignment results (e.g., the *D. melanogaster legless* mRNA).

In addition, we can click on the “Graphics” link to examine the location of each alignment block relative to the subject sequence (Figure 11).

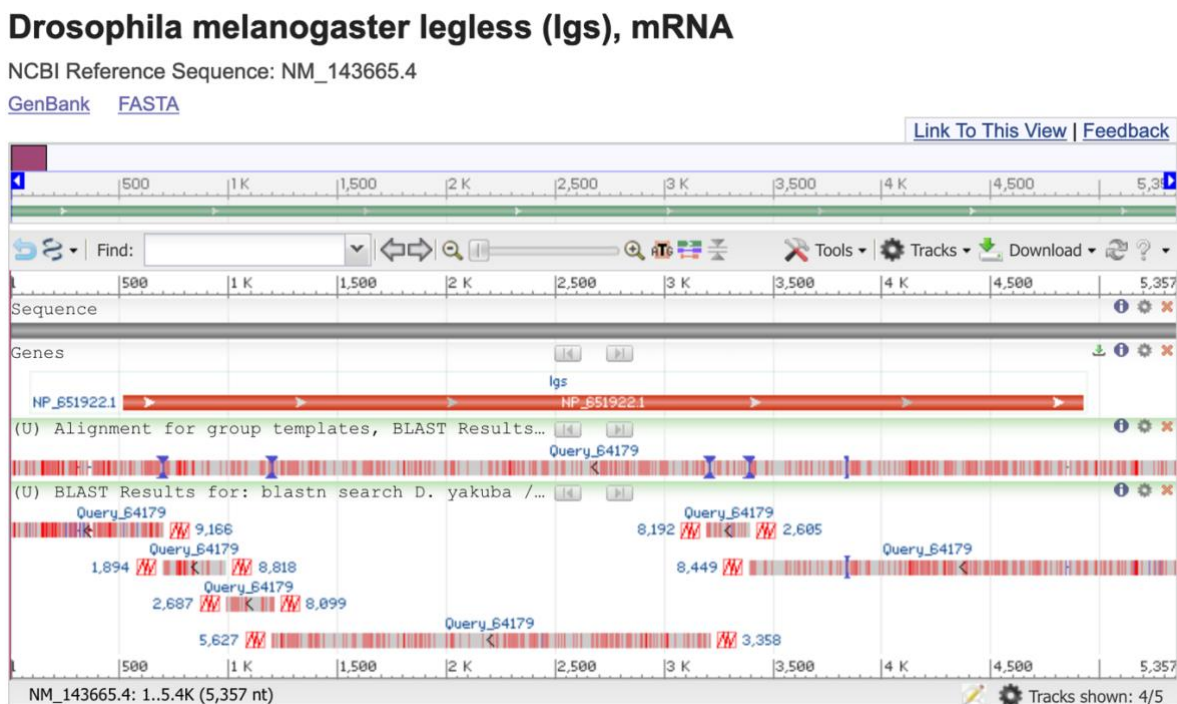
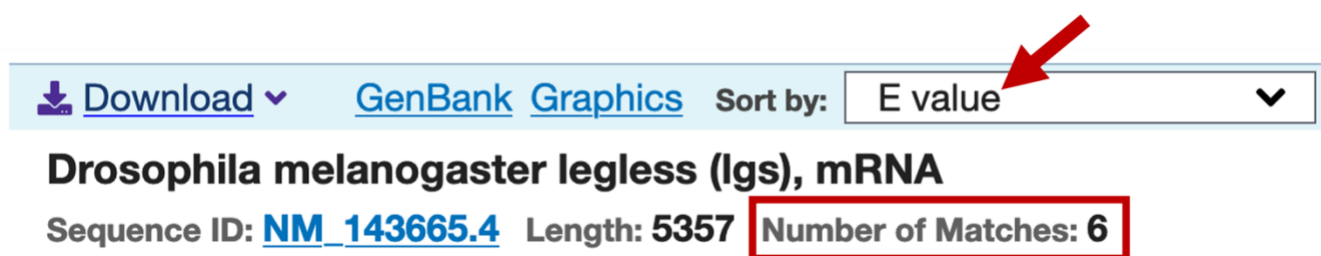


Figure 11. The “Graphics” link allows us to see a graphical view of the alignment blocks relative to the subject sequence (e.g., the *D. melanogaster legless* mRNA).

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 12). Each alignment block represents a region of the *D. melanogaster legless* gene that shows sequence homology with our genomic sequence from *D. yakuba*.



Download ▾ GenBank Graphics Sort by: E value ▾

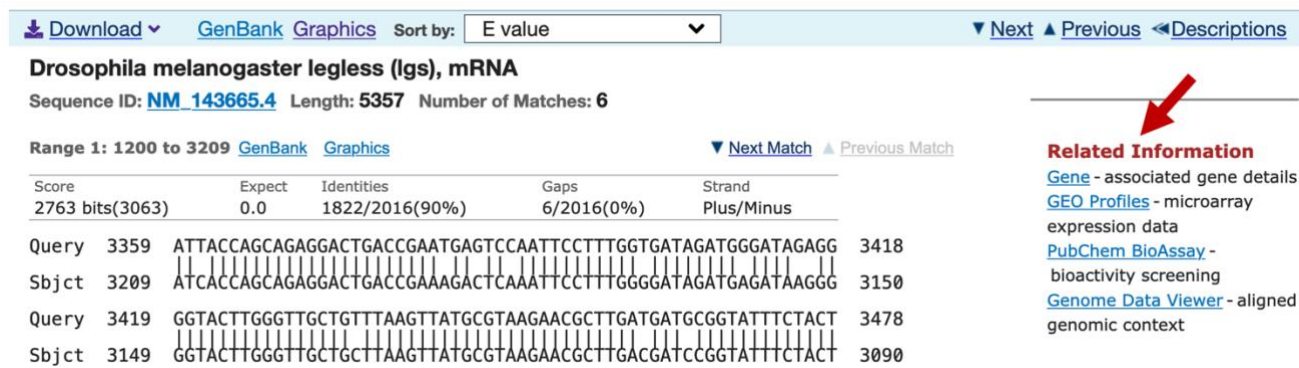
Drosophila melanogaster legless (lgs), mRNA

Sequence ID: [NM_143665.4](#) Length: 5357 Number of Matches: 6

Figure 12. *blastn* detected 6 distinct alignment blocks between the *D. melanogaster legless* mRNA and the *D. yakuba* genomic sequence.

You can use the “Sort by” drop-down box (red arrow in Figure 12) 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.

Depending on the database you use, there might be additional links to other parts of NCBI listed under the “Related Information” panel next to the sequence alignments. For example, there are links to Entrez Gene, GEO Profiles, PubChem BioAssay, and the Genome Data Viewer for the “*Drosophila melanogaster legless (lgs), mRNA*” (NM_143665.4; Figure 13). Entrez Gene provides us with an overview of the gene and links to literature references. GEO Profiles allow us to access expression data associated with the gene. PubChem BioAssay contains bioactivity and toxicity data derived from small-molecule and RNAi screens. The Genome Data Viewer allow us to view the BLAST alignments in a genome browser with other evidence tracks (e.g., gene annotations, RNA-Seq data, repeats).



Download ▾ GenBank Graphics Sort by: E value ▾ ▼ Next ▲ Previous ◀ Descriptions

Drosophila melanogaster legless (lgs), mRNA

Sequence ID: [NM_143665.4](#) Length: 5357 Number of Matches: 6

Range 1: 1200 to 3209 [GenBank](#) [Graphics](#) ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
2763 bits(3063)	0.0	1822/2016(90%)	6/2016(0%)	Plus/Minus

Query 3359 ATTACCAGCAGAGGACTGACCGAATGAGTCCAATTCCTTTGGTGATAGATGGGATAGAGG 3418

Sbjct 3209 ATCACCAGCAGAGGACTGACCGAAGACTCAAATTCCTTTGGGGATAGATGAGATAAGGG 3150

Query 3419 GGTACTTGGGTTGCTGTTTAAGTTATGCGTAAGAAGCGCTTGATGATGCGGTATTCTACT 3478

Sbjct 3149 GGTACTTGGGTTGCTGCTTAAGTTATGCGTAAGAAGCGCTTGACGATCCGGTATTCTACT 3090

Related Information

- [Gene](#) - associated gene details
- [GEO Profiles](#) - microarray expression data
- [PubChem BioAssay](#) - bioactivity screening
- [Genome Data Viewer](#) - aligned genomic context

Figure 13. You can learn more about the *blastn* match using the links under the “Related Information” section.

What about the alignments themselves? 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 14).

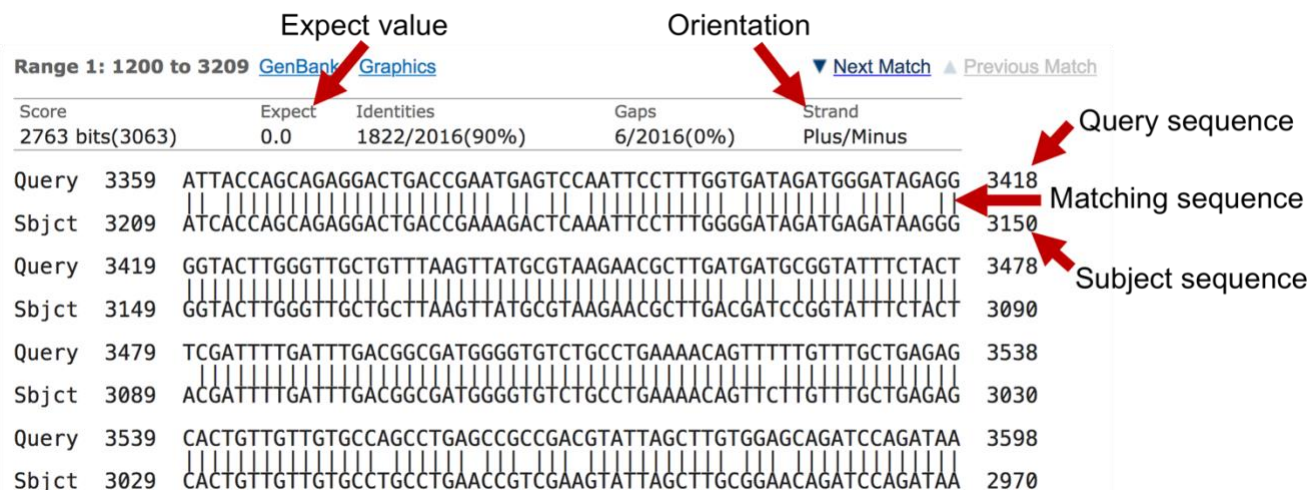


Figure 14. The key characteristics of a typical BLAST alignment.

The - character in either the query or the subject sequence denotes a gap in the alignment (Figure 15).

Query 4019 AAACGAAGCCAGCATATCCGTTGAGCTTCCCATCATGTTGCCATTCGGAGCGCCGGAGCT 4078

Sbjct 2549 AAACGAGGCTAGCATATCCGTAGAGCTTCCCATCATATTGCCATTCGGGGCGCCGGAGCT 2490

Query 4079 TGAGCAATGCATATTGACATTTATTCCAGCTACAGTAGTTCCTGTGACAGCCACA ACTCC 4138

Sbjct 2489 TGAGCAATGCATATTGACATTTACTCCAGCTGCAGTTGTTCCAGTGACA-----ACACC 2436

Query 4139 TACTCCAGATCCACATTGCACGCTGTTTTTCTGACTGTTATTACATCCATTACAGCACC 4198

Sbjct 2435 TACTCCAGATCCACATTGCACACTGGTTTTTTGATTATTATTACATCCTATTACGGCACC 2376

Figure 15. Gaps in the alignment are represented by the '-' character.

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 16) 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.

Query 5219 TAATATGCTCGAAATTGGAGGATTATTTAAAGATTGACTAATAAAATCGGGGTTCAAATG 5278

Sbjct 1355 TAATATGCTCGAAATTGCAGGATTATTTAAAGATTCATTGATAAAATCGGGATTCAACTG 1296

Query 5279 ATTACAGCCATCCATGCCATTTTCATCATTATTCAAGGAGGCGGttttttttttC 5338

Sbjct 1295 ATTGCAGCCTTCCATGCTCATTTCGTCAATTATTCAAGGACGATCCTTTTTTTTTC 1236

Query 5339 CGTGGTACTGTCGCTTGAAATTCTTTCTATTTTGAC 5374

Sbjct 1235 CGTGGTACTGTCGTTTGAAATTCTTTCAATTTTAAC 1200

Figure 16. Bases masked by the low complexity filter appear as lowercase grey letters by default.

IV. Taxonomy

This tab shows the taxonomy of the selected BLAST hits in the Descriptions tab. The Taxonomy tab organizes the selected BLAST hits in three different report formats: Lineage, Organism, and Taxonomy (Figure 17).

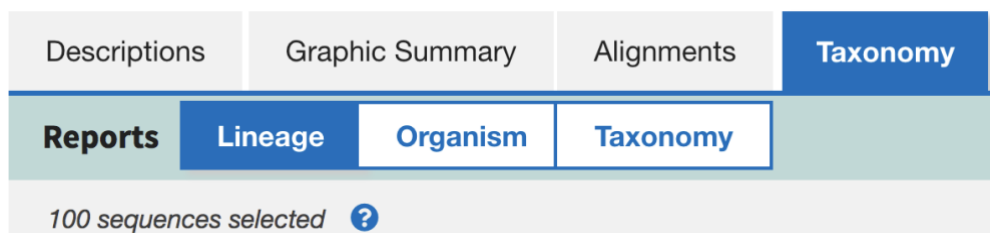


Figure 17. Use the buttons next to the “Reports” label on the main toolbar of the Taxonomy tab to view the Lineage, Organism, and Taxonomy reports for the selected BLAST hits.

The Lineage report provides an overview of the number of selected BLAST hits that are at each taxonomic level. The level of indentation in the “Organism” column corresponds to the taxonomic level. The value in the “Score” column corresponds to the maximum score for the BLAST hits of a terminal node. The value in the “Number of Hits” column shows the number of selected hits that are at the corresponding taxonomic level (Figure 18).

Organism	Blast Name	Score	Number of Hits	Description
Sophophora	flies		100	
• melanogaster group	flies		86	
• • melanogaster subgroup	flies		60	
• • • Drosophila yakuba	flies	3627	15	Drosophila yakuba hits
• • • Drosophila santomea	flies	3578	7	Drosophila santomea hits
• • • Drosophila teissieri	flies	3315	3	Drosophila teissieri hits
• • • Drosophila erecta	flies	2956	1	Drosophila erecta hits
• • • Drosophila mauritiana	flies	2797	1	Drosophila mauritiana hits
• • • Drosophila simulans	flies	2783	20	Drosophila simulans hits
• • • Drosophila sechellia	flies	2765	6	Drosophila sechellia hits
• • • Drosophila melanogaster	flies	2763	7	Drosophila melanogaster hits
• • Drosophila suzukii	flies	1829	1	Drosophila suzukii hits
• • Drosophila subpulchrella	flies	1826	1	Drosophila subpulchrella hits

Figure 18. The Lineage report under the Taxonomy tab shows that 60 of the 100 selected *blastn* hits are in the melanogaster subgroup.

The Organism report groups the selected BLAST hits by organism. The BLAST hits for the different species are separated by a blue header. Within each species, the BLAST hits are sorted by E-value in ascending order (Figure 19).

Descriptions	Graphic Summary	Alignments	Taxonomy
Reports	Lineage	Organism	Taxonomy
100 sequences selected ?			
Description	Score	E value	Accession
Drosophila yakuba [flies]			
PREDICTED: Drosophila yakuba protein BCL9 homolog (LOC6523724), transcript variant X3, mRNA	3627	0.0	XM_039377129
PREDICTED: Drosophila yakuba protein BCL9 homolog (LOC6523724), transcript variant X2, mRNA	3627	0.0	XM_002099563
PREDICTED: Drosophila yakuba protein BCL9 homolog (LOC6523724), transcript variant X1, mRNA	3627	0.0	XM_015191338
PREDICTED: Drosophila yakuba uncharacterized LOC120322138 (LOC120322138), ncRNA	724	0.0	XR_005561904
PREDICTED: Drosophila yakuba uncharacterized LOC120320762 (LOC120320762), transcript variant X2, ncRNA	456	7e-123	XR_005560305
PREDICTED: Drosophila yakuba uncharacterized LOC120320762 (LOC120320762), transcript variant X1, ncRNA	456	7e-123	XR_005560304
PREDICTED: Drosophila yakuba iron-sulfur protein NUBPL (LOC6529157), transcript variant X2, mRNA	451	3e-121	XM_039372352
PREDICTED: Drosophila yakuba uncharacterized LOC26534944 (LOC26534944), mRNA	292	2e-73	XM_039375417
PREDICTED: Drosophila yakuba uncharacterized LOC26535349 (LOC26535349), transcript variant X4, ncRNA	286	7e-72	XR_005562019
PREDICTED: Drosophila yakuba uncharacterized LOC26535349 (LOC26535349), transcript variant X3, ncRNA	286	7e-72	XR_005562018
PREDICTED: Drosophila yakuba uncharacterized LOC26535349 (LOC26535349), transcript variant X1, ncRNA	286	7e-72	XR_001453626
PREDICTED: Drosophila yakuba uncharacterized LOC26535349 (LOC26535349), transcript variant X2, ncRNA	286	7e-72	XR_005562017
PREDICTED: Drosophila yakuba uncharacterized LOC120321400 (LOC120321400), ncRNA	249	2e-60	XR_005561002
PREDICTED: Drosophila yakuba uncharacterized LOC26534685 (LOC26534685), mRNA	244	2e-59	XM_015189272
PREDICTED: Drosophila yakuba hairy/enhancer-of-split related with YRPW motif protein (LOC6528517), mRNA	229	2e-54	XM_002089527
Drosophila santomea [flies]			
PREDICTED: Drosophila santomea protein BCL9 homolog (LOC120454922), transcript variant X4, mRNA	3578	0.0	XM_039640670
PREDICTED: Drosophila santomea protein BCL9 homolog (LOC120454922), transcript variant X3, mRNA	3578	0.0	XM_039640668
PREDICTED: Drosophila santomea protein BCL9 homolog (LOC120454922), transcript variant X2, mRNA	3578	0.0	XM_039640667
PREDICTED: Drosophila santomea protein BCL9 homolog (LOC120454922), transcript variant X1, mRNA	3578	0.0	XM_039640669
PREDICTED: Drosophila santomea serine/threonine-protein kinase 32A (LOC120451155), transcript variant X3, mRNA	896	0.0	XM_039634588
PREDICTED: Drosophila santomea acetylcholine receptor subunit alpha-like (LOC120447845), mRNA	223	7e-53	XM_044005969
PREDICTED: Drosophila santomea uncharacterized LOC120452845 (LOC120452845), transcript variant X2, mRNA	220	9e-52	XM_039637282
Drosophila teissieri [flies]			

BLAST hits in
D. yakuba
(sorted by
E-values)

BLAST hits in
D. santomea
(sorted by
E-values)

Figure 19. The Organism report under the Taxonomy tab allows one to quickly identify the best match in each species.

The Taxonomy report has a similar layout compared to the Lineage report. However, the Taxonomy report provides additional controls (the +/- icons under the "Taxonomy" column) to expand or collapse the non-leaf nodes (Figure 20). It also includes the number of organisms with BLAST hits at each taxonomic level.

Descriptions	Graphic Summary	Alignments	Taxonomy
Reports	Lineage	Organism	Taxonomy
100 sequences selected ?			
Taxonomy	Number of hits	Number of Organisms	Description
[-] Sophophora	100	27	
[-] melanogaster group	86	20	
[-] melanogaster subgroup	60	8	
[-] Drosophila yakuba	15	1	Drosophila yakuba hits
[-] Drosophila santomea	7	1	Drosophila santomea hits
[-] Drosophila teissieri	3	1	Drosophila teissieri hits
[-] Drosophila erecta	1	1	Drosophila erecta hits
[-] Drosophila mauritiana	1	1	Drosophila mauritiana hits
[-] Drosophila simulans	20	1	Drosophila simulans hits
[-] Drosophila sechellia	6	1	Drosophila sechellia hits
[-] Drosophila melanogaster	7	1	Drosophila melanogaster hits
[-] suzukii subgroup	3	3	
[-] Drosophila rhopaloea	2	1	Drosophila rhopaloea hits

Figure 20. Click on the "-" icon next to the taxonomic level under the "Taxonomy" column to collapse a non-leaf node (red arrow). Click on the "+" icon next to the taxonomic level to expand the non-leaf node (purple arrow).

Interpreting the *blastn* search result

Now that we have a better understanding of how the BLAST report is organized, we are ready to interpret the *blastn* results. The “Descriptions” and the “Graphic Summary” tabs (Figure 7 and Figure 9) show that many of the top hits are much more significant (with E-values of 0.0) than the rest of the *blastn* hits. Most of these top hits contain regions of sequence similarity that span the entire length of the query sequence (Figure 9). Looking at the descriptions and the corresponding GenBank records, it appears that these *blastn* hits correspond to the gene *legless* (also known as *BCL9*) in different *Drosophila* species.

Among these significant matches, only the *D. melanogaster* hit has an accession number that begins with the prefix “NM_” (Figure 21). The accession numbers for the matches to the other *Drosophila* species all have the prefix “XM_”. 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 will use the *D. melanogaster* model in this analysis.

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
<input checked="" type="checkbox"/> PREDICTED: Drosophila yakuba protein BCL9 homolog (LOC6523724), transcript variant X3, mRNA	Drosophila yak...	3627	9070	45%	0.0	100.00%	5016	XM_039377129.2
<input checked="" type="checkbox"/> PREDICTED: Drosophila yakuba protein BCL9 homolog (LOC6523724), transcript variant X2, mRNA	Drosophila yak...	3627	9226	46%	0.0	100.00%	5102	XM_002099563.4
<input checked="" type="checkbox"/> PREDICTED: Drosophila yakuba protein BCL9 homolog (LOC6523724), transcript variant X1, mRNA	Drosophila yak...	3627	9358	47%	0.0	100.00%	5174	XM_015191338.3
...								
<input checked="" type="checkbox"/> PREDICTED: Drosophila sechellia protein BCL9 homolog (LOC6619458), mRNA	Drosophila sec...	2765	6348	43%	0.0	90.43%	4761	XM_032723783.1
<input checked="" type="checkbox"/> Drosophila melanogaster legless (lgs), mRNA	Drosophila mel...	2763	6759	48%	0.0	90.38%	5357	NM_143665.4
<input checked="" type="checkbox"/> PREDICTED: Drosophila suzukii protein BCL9 homolog (LOC108020387), mRNA	Drosophila suz...	1829	4167	43%	0.0	79.83%	4983	XM_017088651.2

Figure 21. The best manually curated RefSeq match to the query sequence is the *D. melanogaster legless (lgs)* mRNA (with the accession number NM_143665.4).

From the *blastn* hit list, click on the description that corresponds to the *D. melanogaster* hit to jump to the alignment section. Our analysis above has shown that there are six alignment blocks (Figure 12). We also notice that the *D. melanogaster* mRNA has a total length of 5357 bases, so the first question we would like to address is whether the entire mRNA aligns to our sequence.

To address this question, we will examine the subject coordinates of the alignment blocks from the *D. melanogaster legless* mRNA. We find that these blocks span from 3209-1200, 5357-3394, 699-2, 987-699, 1204-990, and 3395-3198. Re-ordering the coordinates of the alignment blocks with respect to our subject sequence (i.e., Sort by: Subject start position) produces the following list of alignments (coordinates of the query sequence are in parenthesis): 2-699 (9853-9167), 699-987 (9107-8819), 990-1204 (8314-8100), 1200-3209 (5374-3359), 3198-3395 (2809-2606), and 3394-5357 (2552-586).

Despite some minor overlaps and missing bases, we can account for most of the mRNA sequence in this collection of alignments. Note that all the alignment blocks are collinear with respect to our query sequence (i.e., all the alignment blocks are in the reverse orientation relative to the subject mRNA) and show a high degree of sequence similarity (with sequence identity that ranges from 75–92% at the nucleotide level).

Detecting Coding Regions Using *blastx*

Because the RefSeq mRNA sequence consists of both translated and untranslated regions (i.e., 5' and 3' UTRs), the next step in our analysis is to identify the coding region in our sequence. We will set up a *blastx* search in order to compare a nucleotide genomic sequence against a protein database. Because every mRNA in the RefSeq RNA database has a corresponding sequence in the RefSeq Protein database, we will search our *D. yakuba* sequence against the RefSeq Protein (refseq_protein) database. We now have all the information we need to setup the *blastx* search.

1. Navigate to the NCBI BLAST home page and click on the “*blastx*” image
2. Under the “Enter Query Sequence” section, click on the “Browse” or the “Choose File” button and select our sequence (unknown.fna).
3. Enter the Job Title “*blastx* search *D. yakuba* / RefSeq Protein”
4. In the “Choose Search Set” section, change the database to “Reference proteins (refseq_protein)”.
5. Check the box “Show results in a new window” next to the “BLAST” button
6. Click “BLAST” (Figure 22)

The screenshot shows the NCBI BLAST interface for a *blastx* search. At the top, the 'blastx' tab is selected under the heading 'Translated BLAST: blastx'. The 'Enter Query Sequence' section has a text box containing 'unknown.fna' and a 'Browse...' button next to it. Below this, the 'Job Title' is set to 'blastx search D. yakuba / RefSeq Protein'. In the 'Choose Search Set' section, the 'Database' dropdown is set to 'Reference proteins (refseq_protein)'. The 'BLAST' button is highlighted, and the checkbox 'Show results in a new window' is checked.

Figure 22. Configure our *blastx* search of the unknown sequence against the NCBI RefSeq Protein database.

For teaching purposes, the *blastx* search result (*blastxRefSeqProtein.txt*) is available in the package for this walkthrough. (Note that this *blastx* search can take several minutes to complete.)

The *blastx* report is similar to the *blastn* report. It consists of the “Descriptions”, “Graphic Summary”, “Alignments”, and “Taxonomy” tabs. The “Graphic Summary” tab shows the highly significant hits to the legless protein in *D. melanogaster* and the homologous protein in the other *Drosophila* species. It also shows a few significant hits to transposases in the region between 6000-8000 bp of our sequence (Figure 23).

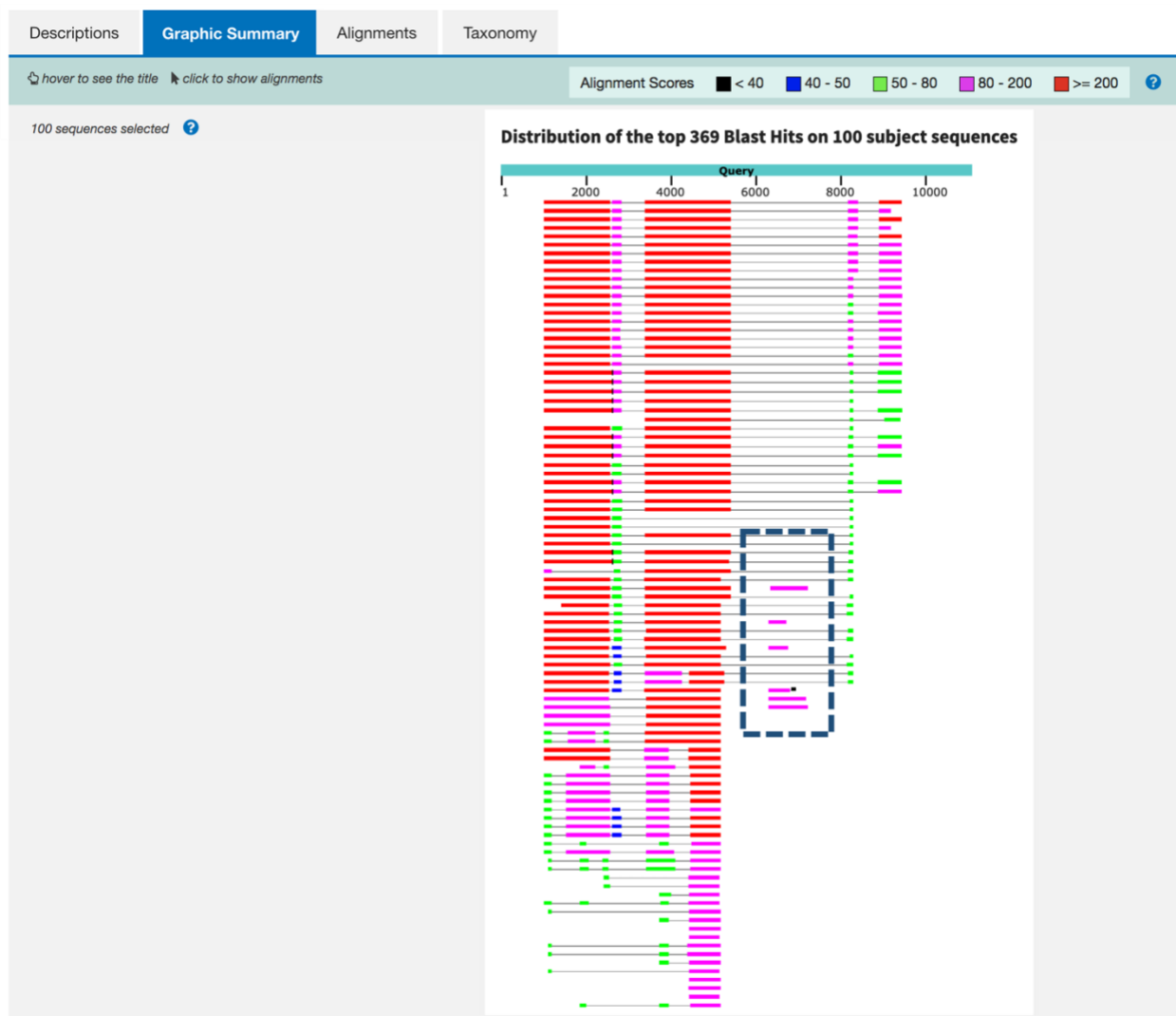


Figure 23. Multiple *blastx* hits in the region between 6000-8000 bp in our sequence.

These hits suggest our sequence contains a type of repetitious element called a transposable element. In future walkthroughs, we will learn how we can reduce the number of spurious hits in our BLAST reports by masking these elements prior to performing the BLAST search. For now, we will ignore these additional matches and focus on the best manually curated RefSeq hit — the *D. melanogaster* legless protein (NP_651922.1; Figure 24).

Similar to the *blastn* alignment, each alignment block in our *blastx* report also consists of three lines: the query sequence, the matching sequence, and the subject sequence. Note that the query sequence has been translated into the corresponding amino acid sequence in the reading frame specified by the “Frame” field. However, the coordinates of the query sequence are still relative to the original nucleotide sequence. Like our *blastn* alignment, the grey lowercase residues in the query sequence correspond to low complexity sequences that were masked by BLAST.

There are some minor differences in the matching sequence of the *blastn* and *blastx* outputs. Residues in the matching sequence represent amino acids that are identical between the query and subject sequences. The “+” character denotes amino acids that are different between the query and subject, but these different amino acids have similar chemical properties. A space indicates that the two aligned amino acid in the query and subject are different, and they have different chemical properties.

When investigating the *blastx* alignment with the *D. melanogaster* legless protein, the first question is whether there are matches to the entire legless protein. We see from the “Length” field underneath the sequence name that the *D. melanogaster* legless protein has 1469 residues. Sorting the alignment blocks by the subject start position, we see matches to the protein sequence at 1-158 (9344-8814), 148-229 (8332-8099), 228-897 (5374-3359), 888-959 (2827-2606), and 959-1469 (2553-1018). In addition, the coordinates relative to our query sequence (in parenthesis) are consistent with the results from our previous *blastn* search. Based on both the *blastn* and *blastx* results, we can determine the approximate coordinates of the UTRs and the coding regions in our *D. yakuba* sequence. Hence it appears that our *D. yakuba* sequence contains an ortholog of the *D. melanogaster* legless gene.

While the alignment generally looks good, there are a few problems with some of the *blastx* alignment blocks. Looking at the alignment block that corresponds to the first 158 amino acids of the protein sequence (9344-8814 in our query sequence), we noticed a large gap beginning at residue 61 (9167 in our query sequence) (Figure 26). Furthermore, the translation of the query in this region contains a stop codon (the * character). One possible explanation for the stop codon is that *blastx* might have combined two separate exons into the same alignment. If that were the case, the intron between the two exons would also be translated by *blastx*.

legless [Drosophila melanogaster]

Sequence ID: [NP_651922.1](#) Length: 1469 Number of Matches: 5

Range 1: 1 to 158 [GenPept](#) [Graphics](#)

[▼ Next Match](#) [▲ Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps	Frame
177 bits(449)	4e-40	Compositional matrix adjust.	124/178(70%)	132/178(74%)	21/178(11%)	-2
Query 9344	MLSTTMPRSPQAQPQNSDAS	TSASGSNPGVIGNGISATNISSPKNLKNELFSTMSP				9168
Sbjct 1	MLSTTMPRSP Q QPQ NSDAS	TSASGSNPG IGNG SA + SSPK L +E FST+SP				60
Query 9167	MLSTTMPRSP TQQQ P NSDAS	TSASGSNPGAAIGNGDSAASRSSPKTLNSEPFSTLSP				
Query 9167	GKCYVLIFHCAEI*QLSMFT	DQIKVTPDEGTEKSGSLSTSDKaggvavgggGNISSEGPTM				8988
Sbjct 61	-----	DQIK+TP+EGTEKSGSLSTDKA G GN EG TM				100
Query 8987	LRQNSSSSINSCLVAspqnsssehsnssnv	SGTVGLTQMVDCDEQSKKKKCSVKDEEGK				8814
Sbjct 101	LRQNS+S+INSCLVASPQNSSEHSNSSNV	TVGLTQMVDCDEQSKK KCSVKDEE +				158

Figure 26. The *blastx* alignment between the unknown sequence (query) and the *D. melanogaster* legless protein (subject) shows a large gap and an in-frame stop codon (*).

Another problem with the alignments is the substantial amount of overlap between two adjacent alignment blocks; this occurs with blocks 1-158 and 148-229, and with blocks 228-897 and 888-959. However, examination of the beginning of the alignment block that spans from 148-229 shows that the first 10 residues in the alignment block have much weaker sequence similarity than the rest of the residues in the alignment (Figure 27). We also see a similar pattern in the alignment block beginning at 888 compared to the block that ends at residue 897. Hence our observations suggest that *blastx* might have over-extended the alignments in both cases.

Range 2: 148 to 229 [GenPept](#) [Graphics](#) [▼ Next Match](#) [▲ Previous Match](#) [▲ First Match](#)

Score	Expect	Method	Identities	Positives	Gaps	Frame
92.0 bits(227)	4e-14	Compositional matrix adjust.	69/82(84%)	72/82(87%)	4/82(4%)	-3
Query 8332	HKCPI-----S	ICSNkakglaagggcgtgstssl	TVKEEPTDVLGSLVNMKKEERENHSP	8165		
	+KC + + +	EI SNKAKG AAGGGC TGSTSSL	TVKEEPTDVLGSLVNMKKEERENHSP			
Sbjct 148	NKCSVKDEEA	ISSNKAKGQAAGGGCETGSTSSL	TVKEEPTDVLGSLVNMKKEERENHSP	207		
Query 8164	TMSPVGFGSIGNAQDLSATPGK	8099				
	TMSPVGFGSIGNAQD SATP K					
Sbjct 208	TMSPVGFGSIGNAQDNSATPVK	229				

Figure 27. The beginning of the alignment shows a much lower degree of sequence homology.

Define the Intron-Exon Boundaries with *Gene Record Finder* and *bl2seq*

Based on our previous *blastn* and *blastx* analyses, our current hypothesis is that we have identified the putative ortholog of the *legless* gene in our *D. yakuba* sequence. However, in order to construct a complete gene model, we must resolve the discrepancies in the alignments of our *blastn* and *blastx* output. Because the coding region is under strong selective pressure and is likely to be more conserved than other regions of the genome, our first step is to identify the coding regions of our putative gene.

To begin the more detailed analysis, we will perform a series of BLAST searches using the amino acid sequence of each exon in the *D. melanogaster* version of the *legless* gene. It will be helpful to our annotation efforts if we can obtain the amino acid sequence that corresponds to each exon individually. Fortunately, we can easily obtain the individual exon sequences using the [Gene Record Finder](#).

1. Navigate to the [F Element Project page](#) on the GEP website
2. Click on the “Gene Record Finder” link under “Resources & Tools”
3. Type “lgs” (the official FlyBase symbol for the *legless* gene) in the textbox and click on the “Find Record” button (Figure 28)

Expansion of the *Drosophila* Muller F Element

Resources & Tools

- Annotation Files Merger
- BLAST Viewer Generator
- Core Promoter Motifs
- Gene Model Checker
- Gene Record Finder

Faculty Resources

- Demo Systems
- GEP Data Repository
- Project Management System
- Project Trello Board
- Quick Check of Student Annotations

Contact Information

Project Leaders:
Cindy J. Arrigo
Sally C. R. Elgin

Lab Website:
The Elgin Lab

Gene Record Finder FlyBase Release 6.43 - (Last Update: 12/31/2021)

Search *D. melanogaster* Gene Records:

FlyBase Gene Symbol

[GEP Home Page](#) | [User Guide](#)

Figure 28. Access the *Gene Record Finder* from the F Element Project page on the GEP website

In the “mRNA Details” section of the gene report, we notice that there is only one isoform of the *legless* gene in *D. melanogaster* (lgs-RA, the A isoform of *lgs*). We can access all the transcribed exons through the “Transcript Details” tab and all the coding exons through the “Polypeptide Details” tab (Figure 29).

Gene Record Finder Search *D. melanogaster* Gene Records:

FlyBase Release 6.43 - (Last Update: 12/31/2021)

FlyBase Gene Symbol

Gene Details

FlyBase ID	FlyBase Name	Chr	5' Start	3' End	Strand	Graphical Viewer
FBgn0039907	lgs	4	443,911	436,957	-	View in GBrowse

mRNA Details

Window Position: D. melanogaster Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6) chr4:436,957-443,911 (6,955 bp)

Scale: 2 kb

chr4: 437,500 | 438,000 | 438,500 | 439,000 | 439,500 | 440,000 | 440,500 | 441,000 | 441,500 | 442,000 | 442,500 | 443,000 | 443,500

Test Track: lgs-RA

Select a row to display the corresponding transcript and peptide details:

FlyBase ID	FlyBase Name	Chr	5' Start	3' End	Strand	Protein ID	Graphical Viewer
FBtr0089111	lgs-RA	4	443,911	436,957	-	FBpp0088180	View in GBrowse

Transcript Details **Polypeptide Details**

Options:

CDS usage map:

Isoform	1_9485_0	2_9485_2	3_9485_2	4_9485_2	5_9485_2	6_9485_0
lgs-PA	1	2	3	4	5	6

Select a row to display the corresponding CDS sequence:

FlyBase ID	5' Start	3' End	Strand	Phase	Size (aa)
1_9485_0	443,393	443,213	-	0	60
2_9485_2	443,154	442,864	-	2	96
3_9485_2	442,389	442,180	-	2	69
4_9485_2	441,451	439,448	-	2	667
5_9485_2	439,165	438,975	-	2	63
6_9485_0	438,918	437,386	-	0	511

Figure 29. Coding exons for the selected isoform of *lgs* is listed under the “Polypeptide Details” section

To retrieve the amino acid sequence for each coding exon (CDS), click on the row that corresponds to the coding exon in the CDS table (Figure 30).

The screenshot shows the 'Polypeptide Details' tab with options to export CDS data. Below the options is a 'CDS usage map' table and a 'Select a row to display the corresponding CDS sequence:' table. A red arrow points to the first row of the second table. To the right, a 'Sequence viewer' window displays the amino acid sequence for lgs:1_9485_0.

CDS usage map:

Isoform	1_9485_0	2_9485_2	3_9485_2	4_9485_2	5_9485_2	6_9485_0
lgs-PA	1	2	3	4	5	6

Select a row to display the corresponding CDS sequence:

FlyBase ID	5' Start	3' End	Strand	Phase	Size (aa)
1_9485_0	443,393	443,213	-	0	60
2_9485_2	443,154	442,864	-	2	96
3_9485_2	442,389	442,180	-	2	69
4_9485_2	441,451	439,448	-	2	667
5_9485_2	439,165	438,975	-	2	63
6_9485_0	438,918	437,386	-	0	511

Sequence viewer for lgs: lgs:1_9485_0

```
>lgs:1_9485_0
MLSTTMPRSPTQQPQPNDSASSTSASGSNPGAAIGNGDSAASRSPKTL
NSEPFSTLSP
```

Figure 30. Click on a row in the CDS table to retrieve the amino acid sequence for the corresponding coding exon.

The first problem in our *blastx* results is the stop codon in the alignment block that spans from 1-158 of the translated protein sequence. To determine the locations of the coding exons, we will perform BLAST searches to compare the individual exons with our sequence. Because we are comparing a protein sequence against a nucleotide sequence, we will use the *tblastn* program for our search. In order to prevent BLAST from masking low complexity regions in our protein, we will turn off the low complexity filter. In addition, because we are only comparing two sequences, we will also turn off compositional adjustments under scoring parameters.

1. Select the first CDS (1_9485_0) from the *Gene Record Finder* CDS table and copy the sequence to the clipboard
2. Open a new web browser tab and navigate to the NCBI BLAST home page; click on the “*tblastn*” image under the “Web BLAST” section
3. Select the checkbox “Align two or more sequences” under the “Enter Query Sequence” section
4. Paste the CDS sequence for 1_9485_0 into the “Enter Query Sequence” field
5. For the “Subject Sequence”, click on the “Browse” or the “Choose File” button and select our unknown sequence (unknown.fna)
6. Click on the “Algorithm parameters” link to expand this section. **Verify that the “Word size” parameter is set to 3.**
7. Change the “Compositional adjustments” field to “No adjustment” under the “Scoring Parameters” section
8. Uncheck the box “Low complexity regions” under “Filters and Masking”
9. Click “BLAST” (Figure 31).

For teaching purposes, the BLAST output (*bl2lgsExon1_tblastn.txt*) is available in the package for this walkthrough.

Align Sequences Translated BLAST: tblastn

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)

>igs:1_9485_0
MLSTTMPRSPQOQPQNSDASSTASGSGNPGAAIGNGDSAASRSSPKTL
NSEPFSTLSP

Query subrange

From

To

Or, upload file No file selected.

Job Title

Enter a descriptive title for your BLAST search

☒ Align two or more sequences

Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)

unknown.fna

Subject subrange

From

To

Or, upload file unknown.fna

BLAST Search nucleotide sequence using Tblastn (search translated nucleotide subjects using a protein query)

☒ Show results in a new window

Note: Parameter values that differ from the default are highlighted in yellow and marked with ?

Algorithm parameters

General Parameters

Max target sequences

Expect threshold

Word size Word size = 3

Max matches in a query range

Scoring Parameters

Matrix

Gap Costs

Compositional adjustments No adjustment

Filters and Masking

Filter ☐ Low complexity regions Turn off low complexity filter

Figure 31. Use the “Align two or more sequences” feature with *tblastn* to align the first coding exon of *lgs* against our sequence with the low complexity filter turned off and no compositional adjustment.

From the “Alignments” tab of the *tblastn* output, we see that the first coding exon has a length of 60 amino acids and corresponds to 9344-9168 of the query sequence when it is translated in frame -2 (Figure 32).

Descriptions

Graphic Summary

Alignments

Dot Plot

Alignment view

Pairwise

?

Restore defaults

Download

1 sequences selected

?

Download

Graphics

Next

Previous

Descriptions

unknown

Sequence ID: **Query_1249** Length: **11001** Number of Matches: **1**

Range 1: **9168 to 9344**

Graphics

Next Match

Previous Match

Score	Expect	Identities	Positives	Gaps	Frame
84.7 bits(208)	8e-25	45/60(75%)	48/60(80%)	1/60(1%)	-2
Query 1	MLSTTMPRSPQTQQPQNSDASSTASGSGNPGAAIGNGDSAASRSSPKTLNSEPFSTLSP	60			
Sbjct 9344	MLSTTMPRSPQAQPQNSDAS-TASGSGNPGVGIGNGISATNISSPKNLKNELFSTMSP	9168			

Query ID	lcl Query_1247 (amino acid)
Query Descr	lgs:1_9485_0
Query Length	60
Subject ID	lcl Query_1249 (dna)
Subject Descr	unknown
Subject	11001
Length	

Figure 32. *bl2seq* results showing the *tblastn* alignment of the first coding exon with our unknown sequence

We can use the same strategy to map the rest of the coding exons. The table below is generated from the *tblastn* searches of the second through the sixth CDSs of the *D. melanogaster legless* gene (query) against the unknown sequence (subject). For teaching purposes, the *tblastn* output for all six CDSs of the *legless* gene is available in the package for this walkthrough (*tblastn_lgs_all_CDS.txt*).

Instead of using *tblastn*, the alignment information in the table below could also have been generated by performing a *blastx* search of the unknown sequence (query) against the six CDSs of the *legless* gene (subject). For teaching purposes, the *blastx* output for all the CDSs of *legless* is available in the package for this walkthrough (*blastx_lgs_all_CDS.txt*).

Exon # (Number of complete codons)	Protein Alignment (Start-End)	Our Sequence Alignment (Start-End)	Frame
1 (60)	1-60	9344-9168	-2
2 (96)	1-95	9104-8820	-2
3 (69)	1-69	8311-8105	-3
4 (667)	1-667	5371-3365	-3
5 (63)	1-63	2800-2606	-3
6 (511)	1-511	2550-1015	-1

The results of our exon-by-exon *bl2seq* analyses suggest we can account for all of the coding exons of the *D. melanogaster legless* gene in our sequence. Furthermore, we were able to resolve the problem with the first exon in our initial *blastx* search: the alignment block that spans from 9344-8814 actually consists of two separate exons, one that spans approximately from 9344-9168 and the other from 9104-8820.

For your own annotation projects, it will be advantageous to save the *bl2seq* outputs as you construct your gene model so that you can revisit the results later (e.g., via the “Download All” drop-down menu in the RID field of the BLAST results page). Note that we have yet to generate a complete gene model for this putative gene in our *D. yakuba* sequence. In future walkthroughs, we will learn how we can use the *UCSC Genome Browser* to identify intron splice sites and to define the exact exon boundaries.

Define the 5' UTR of *legless* Using *blastn*

We could apply the same strategy to map the locations of the putative untranslated regions (UTRs) using *bl2seq* with the *blastn* program. Go back to the *Gene Record Finder* web browser window. The genome browser image in the “mRNA Details” panel shows that the first CDS of *legless* (1_9485_0) is part of a larger exon (lgs:1) that includes the 5' UTR (Figure 33).

To estimate the location of the 5' UTR in our unknown sequence, we will perform a *blastn* search of the exon lgs:1 against the unknown sequence, and then compare the alignment with the *blastx* search result for CDS 1_9485_0.

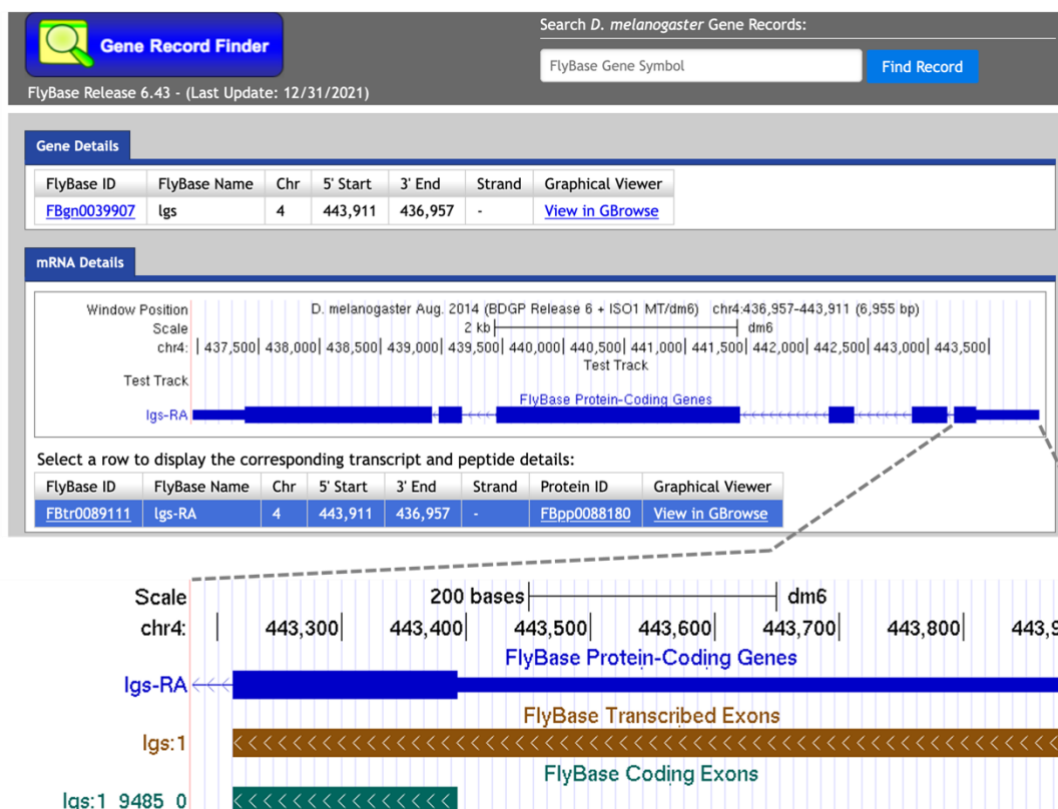


Figure 33. The mRNA Details panel of the *Gene Record Finder* for *legless (lgs)* shows that the first CDS of *legless* (CDS 1_9485_0) is part of a larger transcribed exon *lgs:1*. (Top) The “Strand” column in the “mRNA Details” panel of the *Gene Record Finder* shows that *legless* is on the minus strand in chromosome 4 of *D. melanogaster*. (Bottom) The thick box in the “FlyBase Protein-Coding Genes” evidence track corresponds to the coding exon, the thinner box corresponds to the untranslated region, and the line with the arrows corresponds to the intron.

To retrieve the exon sequence for the first transcribed exon, select the “Transcript Details” tab in the *Gene Record Finder* and then click on the first row in the exon table (Figure 34).

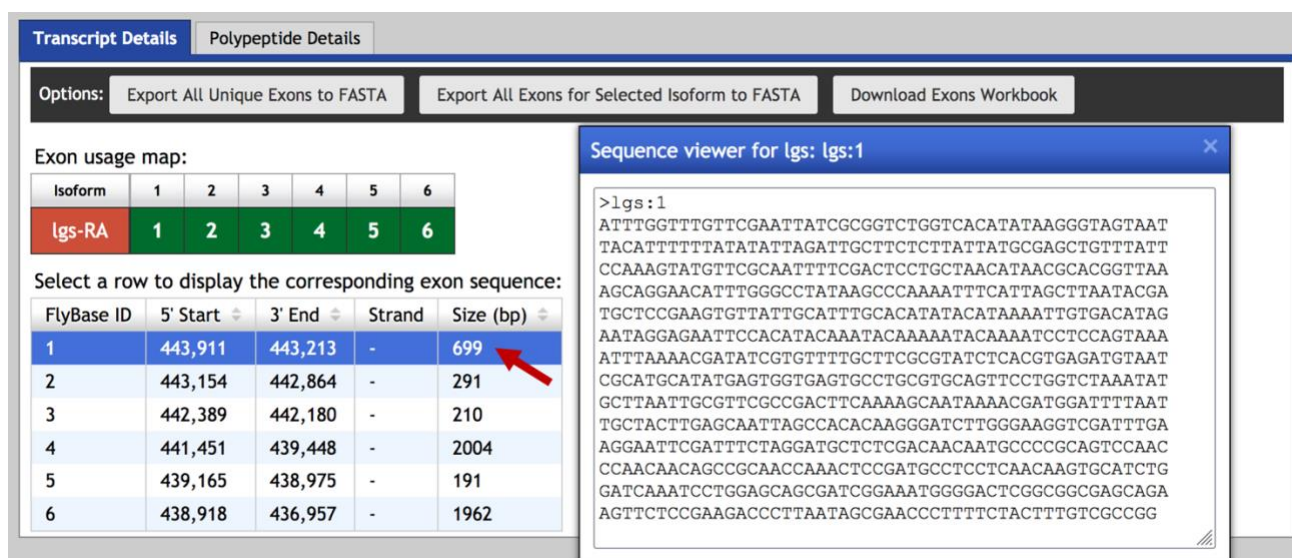


Figure 34. The “Transcript Details” tab of the *Gene Record Finder* shows the list of transcribed exons for the *D. melanogaster* gene *legless*. Click on the first row to retrieve the nucleotide sequence for the first exon of the *legless* gene (*lgs:1*).

We can then use the following steps to perform the *blastn* search:

1. Select the first exon (lgs:1) from the *Gene Record Finder* exon table and copy the sequence to the clipboard
2. Open a new web browser tab and navigate to the NCBI BLAST home page; click on the “Nucleotide BLAST” image under the “Web BLAST” section
3. Select the checkbox “Align two or more sequences” under the “Enter Query Sequence” section
4. Paste the exon sequence for lgs:1 into the “Enter Query Sequence” field
5. For the “Subject Sequence”, click on the “Browse” or the “Choose File” button and select our unknown sequence (unknown.fna)
6. Select the “Somewhat similar sequences (*blastn*)” option under “Program Selection”
7. Click on the “Algorithm parameters” link to expand this section.
8. Uncheck the box “Low complexity regions” under “Filters and Masking”
9. Click “BLAST” (Figure 35).

Align Sequences Nucleotide BLAST

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) [Query subrange](#)

>lgs:1
ATTGGTTTGTTCGAATTATCGCGGTCTGGTCACATATAAGGGTAGTAAT
TACATTTTATATATTAGATTGCTTCTCTTATATGCGAGCTGTTTATCCAA
AGTATGTTCCGAATTTTCGACTCCTGCTAACATAACGCACGTTAAAGCAG

Or, upload file [Browse...](#) No file selected. [?](#)

Job Title [Enter a descriptive title for your BLAST search](#) [?](#)

☒ Align two or more sequences [?](#)

Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) [Subject subrange](#)

unknown.fna

Or, upload file [Browse...](#) unknown.fna [?](#)

Program Selection

Optimize for

☐ Highly similar sequences (megablast)

☐ More dissimilar sequences (discontiguous megablast)

☒ Somewhat similar sequences (*blastn*) [Choose a BLAST algorithm](#) [?](#)

blastn

BLAST Search nucleotide sequence using Blastn (Optimize for somewhat similar sequences)

☒ Show results in a new window

Note: Parameter values that differ from the default are highlighted in yellow and marked with [?](#) sign

Algorithm parameters

General Parameters

Max target sequences [Select the maximum number of aligned sequences to display](#) [?](#)

Short queries ☒ Automatically adjust parameters for short input sequences [?](#)

Expect threshold [?](#)

Word size [?](#)

Max matches in a query range [?](#)

Scoring Parameters

Match/Mismatch Scores [?](#)

Gap Costs [?](#)

Filters and Masking

Filter

☒ Low complexity regions [?](#)

☐ Species-specific repeats for: [?](#)

Turn off low complexity filter

Figure 35. Configure the *blastn* search of exon lgs:1 (query) against the unknown sequence (subject).

For teaching purposes, the BLAST output (*bl2lgsExon1_blastn.txt*) is available in the package for this walkthrough.

The “Query Length” field in the top left panel of the *blastn* results page shows that the exon *lgs:1* has a total length of 699 bp. Examination of the *blastn* alignment under the “Alignments” tab of the BLAST output shows a significant alignment (E-value = 6e-149; 75% percent identity) between *lgs:1* and the unknown sequence at 9853-9167. The query coordinates in the *blastn* alignment shows that the alignment includes almost the entire length of the exon *lgs:1* (only the first nucleotide of *lgs:1* is missing from the *blastn* alignment). Since the *tblastn* search of CDS 1_9485_0 against the unknown sequence placed the start codon at 9344-9342 (Figure 32), we can infer from the *blastn* search result for *lgs:1* that the 5' UTR for *lgs* is located approximately at 9853-9345 of the unknown sequence (Figure 36).

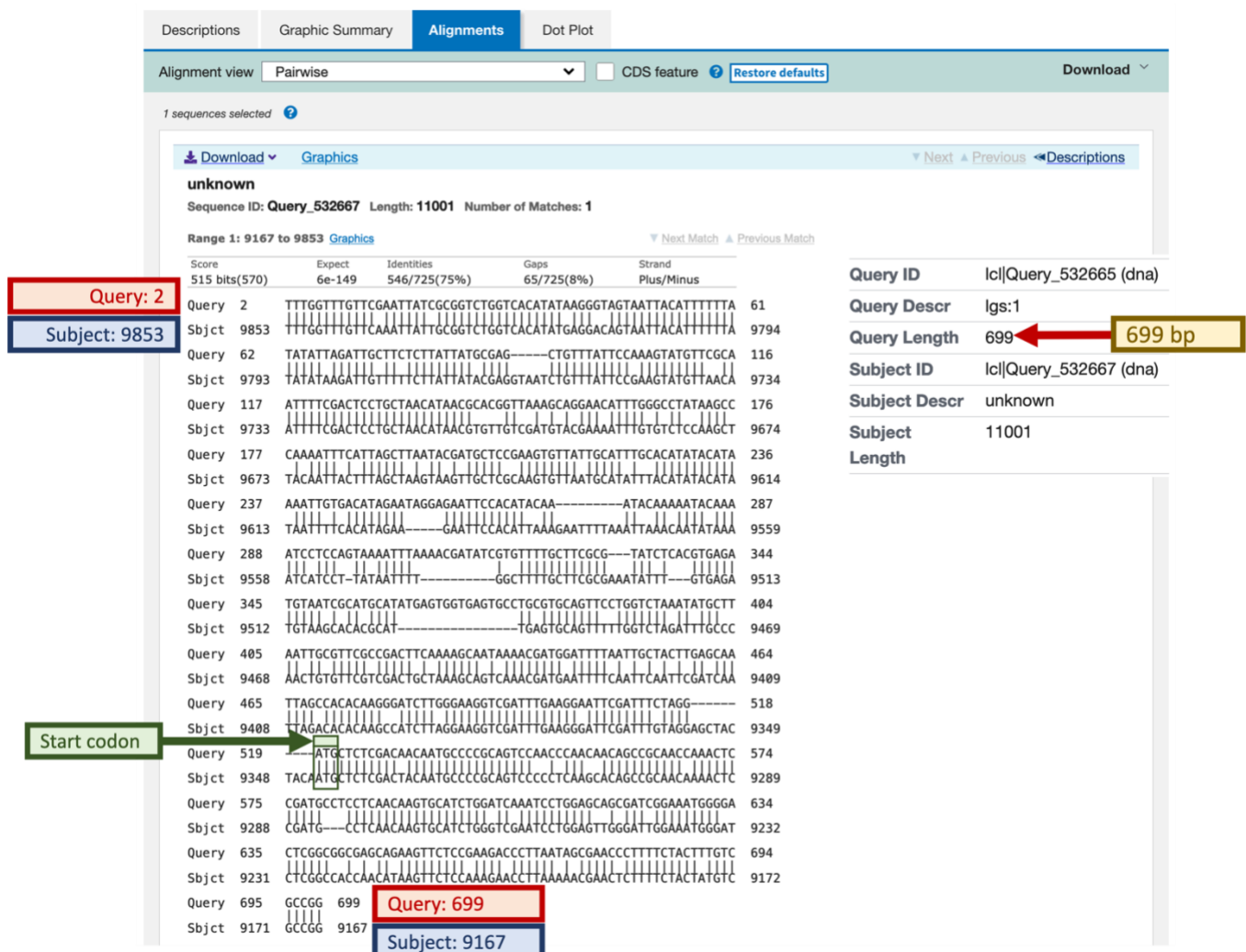


Figure 36. Inferring the location for the 5' UTR of *legless* from the *blastn* alignment. The *blastn* search shows that bases 2-699 of the *D. melanogaster* exon *lgs:1* (query) have significant sequence similarity with the 9853-9167 region of the unknown sequence (subject). The location of the start codon within the *blastn* alignment can be determined from the *tblastn* search result of CDS 1_9485_0 against the unknown sequence (Figure 32), which placed the start codon at 9344-9342 (green box). Hence the 5' UTR for *legless* is located at approximately 9853-9345 of the unknown sequence.

Conclusion

In this walkthrough, we have used multiple BLAST programs to identify and characterize a putative gene in a genomic sequence from *D. yakuba*. You are now ready to tackle some of the more challenging BLAST exercises on the [GEP website](#):

- [Detecting and Interpreting Genetic Homology](#)
- [Using mRNA and EST Evidence in Annotation](#)