

Screening Sequencing Datasets for Marker Genes in CLC Genomics

Steven Wilhelm

Abstract

Citation: Steven Wilhelm Screening Sequencing Datasets for Marker Genes in CLC Genomics. **protocols.io**
dx.doi.org/10.17504/protocols.io.g5vby66

Published: 07 Feb 2017

Before start

You will need:

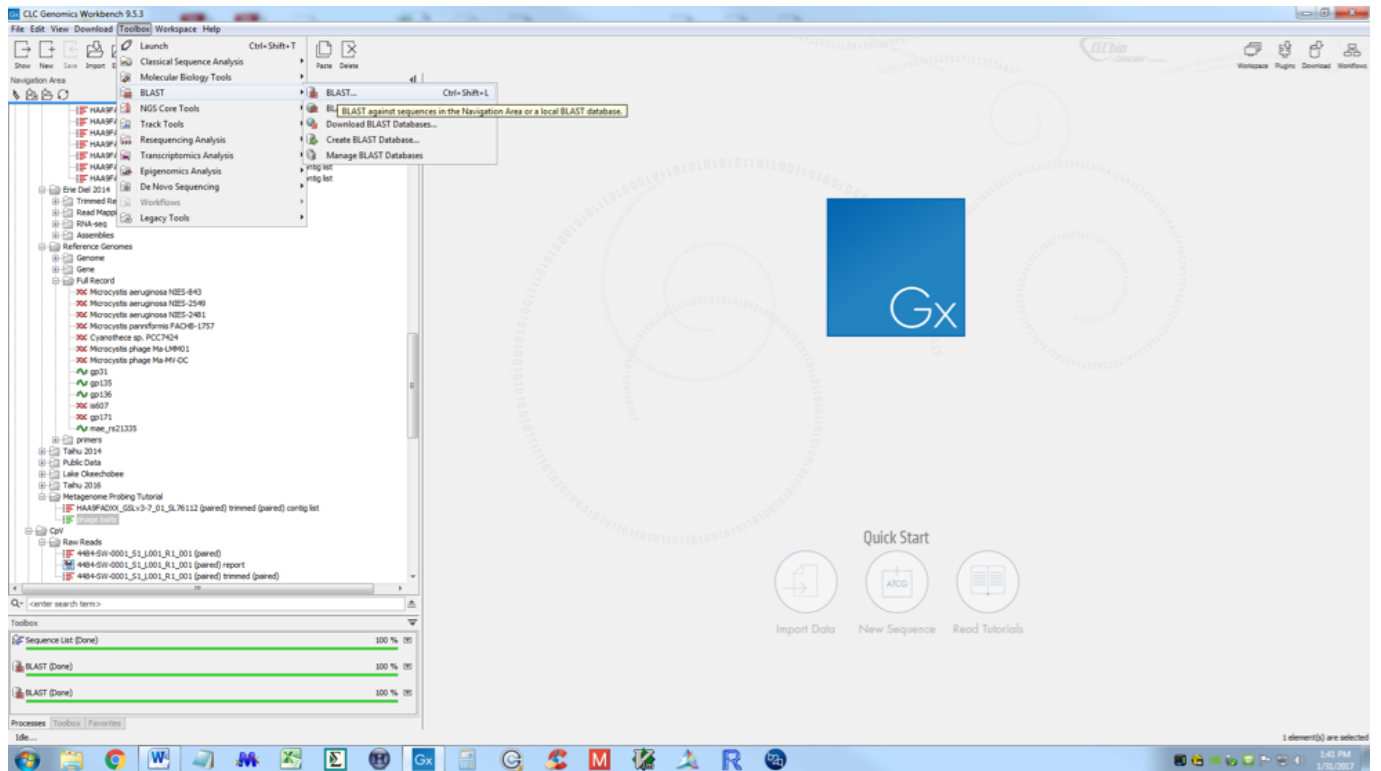
Assembled sequencing dataset--the dataset to be screened needs to be assembled and the contigs extracted. You may want to filter out contigs that are too short or have low coverage.

Marker gene sequence list--to screen the dataset, you will need a file containing the sequences of the genes that you want to screen for, whether they are DNA or protein sequences.

Protocol

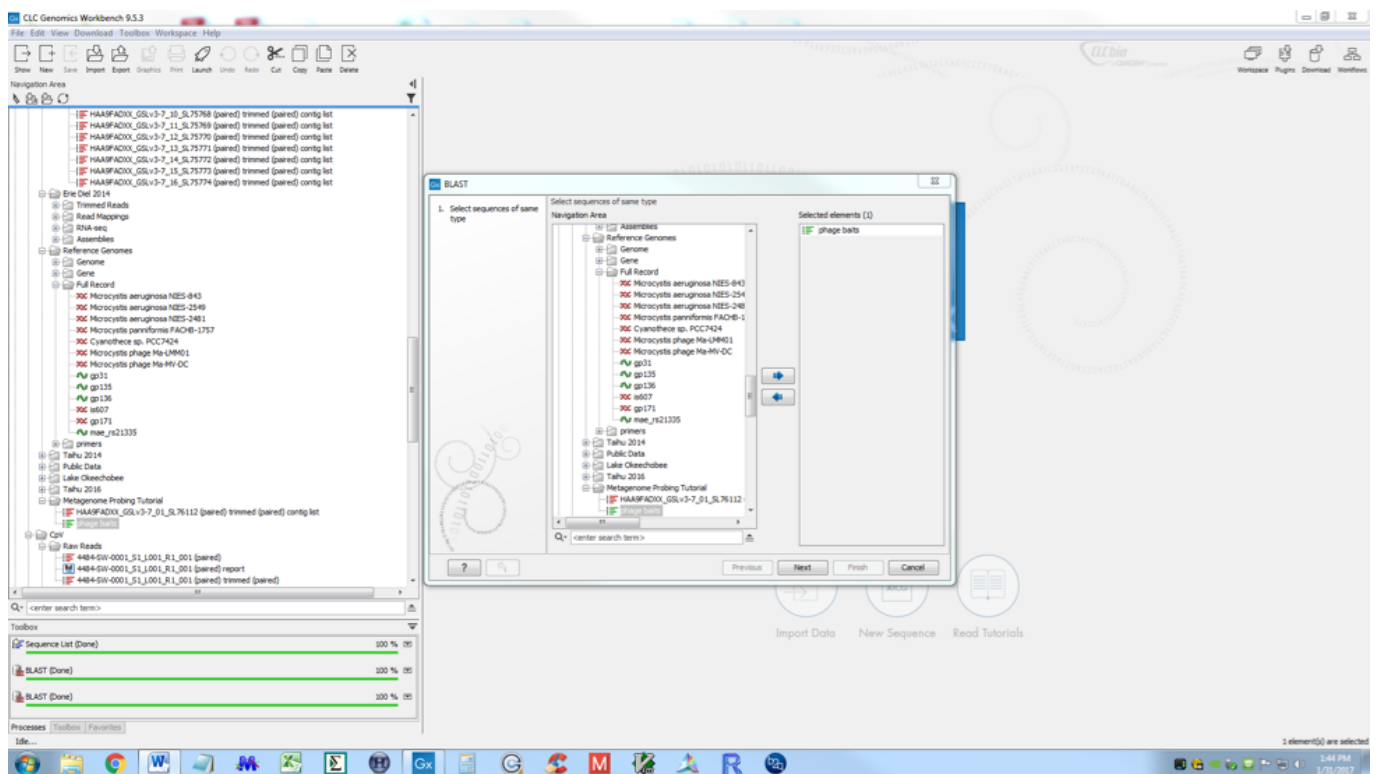
Step 1.

In the toolbar, click --> toolbox --> BLAST --> BLAST...



Step 2.

Selecting this option will bring up the dialog box, which will first prompt you to select query sequences that you want to use in the BLAST.



NOTES

Alyssa Alsante 07 Feb 2017

In this example, I want to probe a metatranscriptome assembly for a series of phage genes that I am interested in looking for, named "phage baits", which is a list of protein sequences.

Step 3.

Once you have selected the sequences that you want to BLAST, click "Next", which will bring you to the BLAST parameters. Here you want to select the type of BLAST that you will be running. The options available differ depending on whether you selected protein or DNA sequences in the previous step.

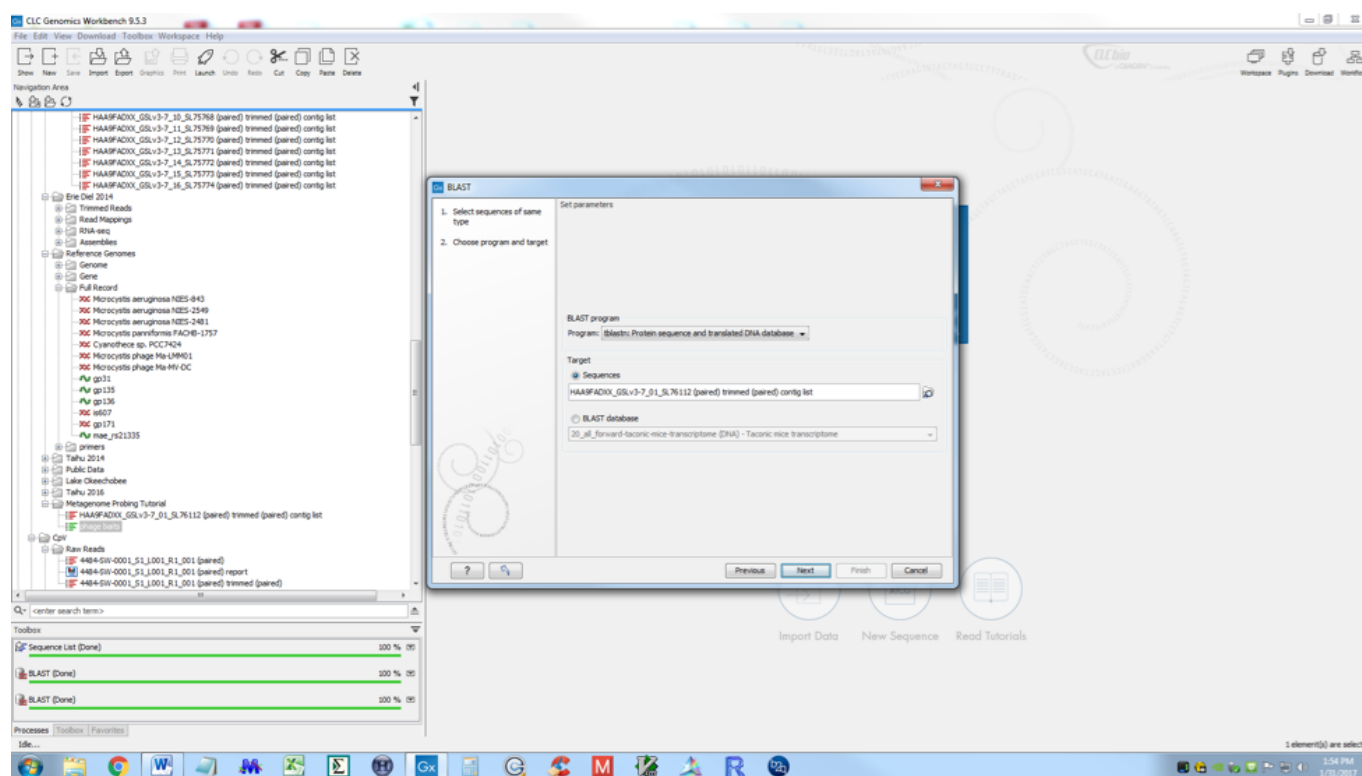
NOTES

Alyssa Alsante 07 Feb 2017

In this example, I will be using "tblastn" since I am blasting protein query sequences against the contigs from a DNA sequence assembly.

Step 4.

Select the sequence file that you want to screen for your target sequences.



NOTES

Alyssa Alsante 07 Feb 2017

In this example, I will be screening a list of contig sequences called "HAA9FADXX_GSL..." shown in the screenshot. This file was originally a metatranscriptome isolated from Lake Erie, which was

imported into CLC and assembled.

Step 5.

The sequence files available will differ depending on which BLAST program that you selected in the previous step.

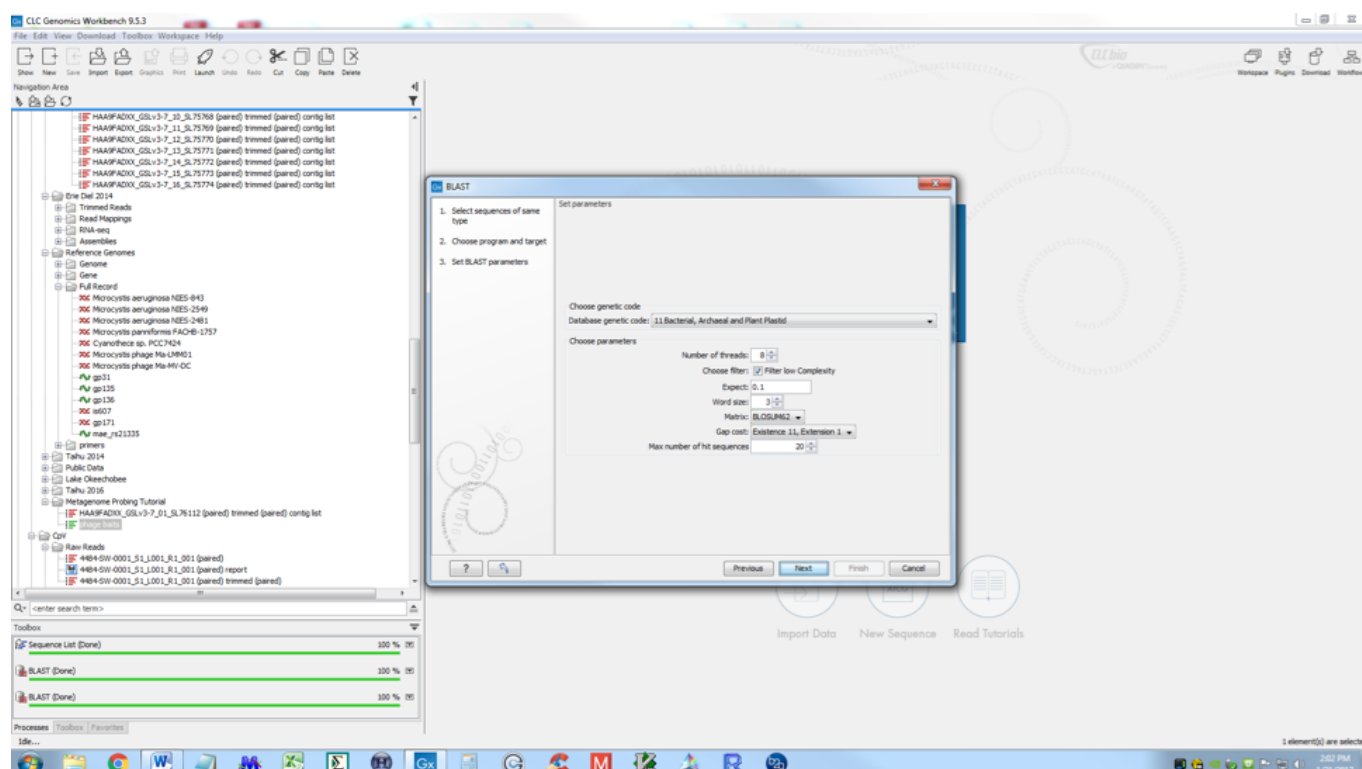
NOTES

Alyssa Alsante 07 Feb 2017

For example, selecting "blastp" requires that both the query and screened datasets are protein sequences. Since you would not be able to use DNA sequences, they will not appear when you select which dataset that you want to screen.

Step 6.

Clicking "Next" will bring you to the next parameter page, where you can tweak the BLAST settings.



NOTES

Alyssa Alsante 07 Feb 2017

In this example, since I selected "tblastn", CLC will be translating the list of contigs into protein sequences for the alignment, so I will need to select the genetic code to be used, which is Bacterial.

Remaining Settings

Step 7.

"Number of threads"--The number of BLAST processes run in parallel, as determined by the number of processors your computer has available.

Remaining Settings

Step 8.

"Choose filter"--Selecting this option filters out low complexity elements of the sequences, which may yield statistically significant hits that are biologically uninteresting or irrelevant.

Remaining Settings

Step 9.

"Expect"--The maximum e-value threshold for an alignment to count as a hit

Remaining Settings

Step 10.

"Word size"--BLAST searches for alignment between sequences by first matching a "word" or a string of characters that must match before the alignment continues. Increasing the word size increases the stringency by which BLAST identifies potential hits, whereas decreasing word size increases the sensitivity.

Remaining Settings

Step 11.

"Matrix"--The substitution matrix used to compute penalties for sequence mismatches.

Remaining Settings

Step 12.

"Gap cost"--This option determines the point system by which BLAST scores the alignment when gaps are detected. "Existence" determines the score that is penalized per gap detected in the alignment, and "Extension" refers to the increasing point penalty calculated for longer gaps.

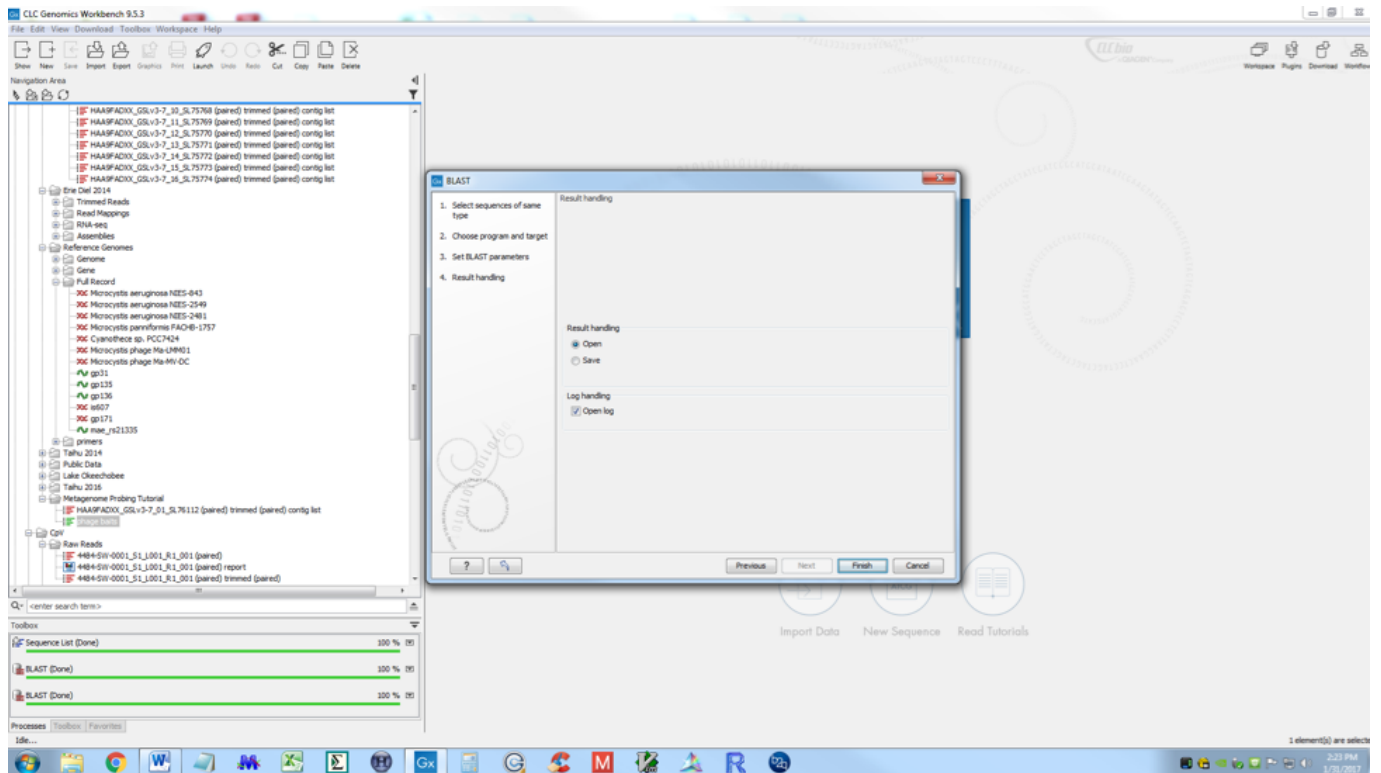
Remaining Settings

Step 13.

'Max number of hit sequences'--Maximum number of hits reported in the BLAST report.

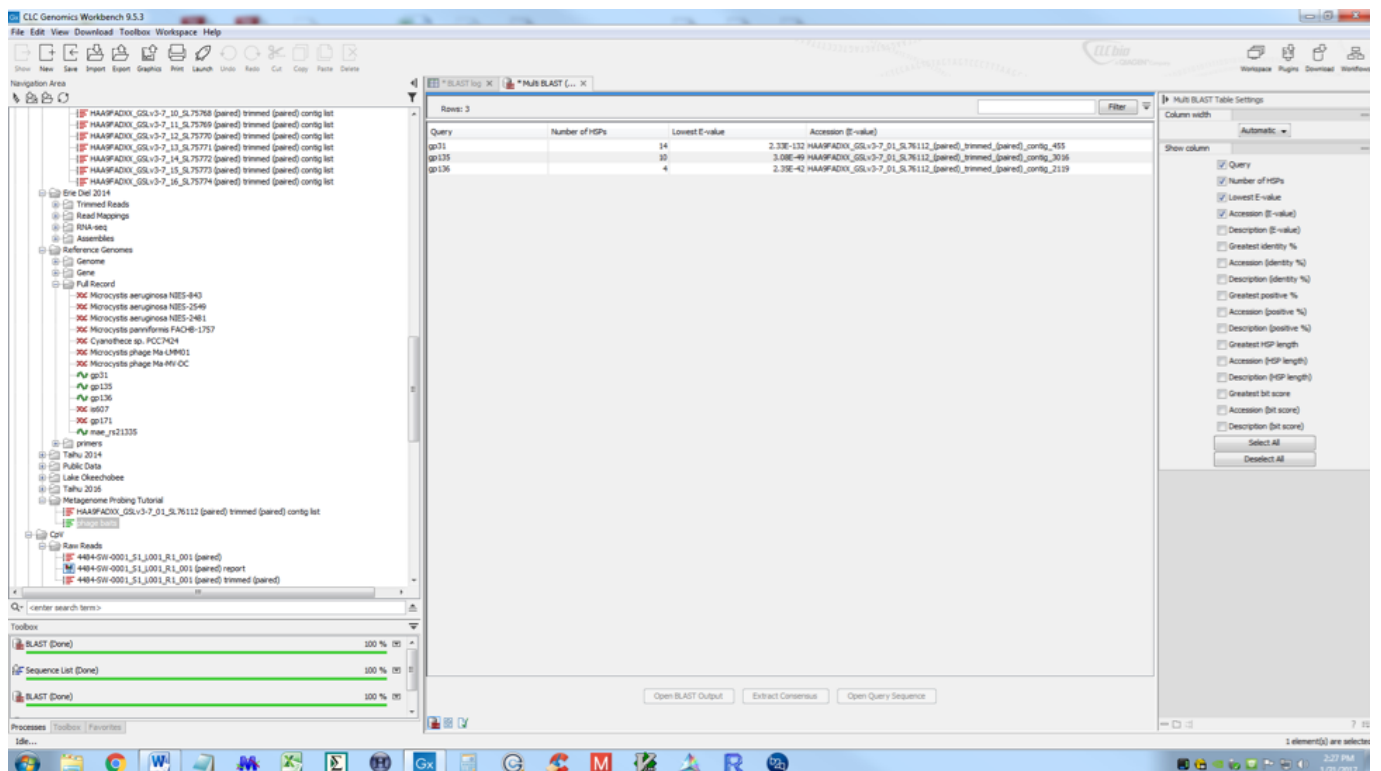
Step 14.

Clicking "Next" will bring you to the result handling options where you will select whether you want CLC to open the results upon completion or save them in a specific slot. Keep in mind that selecting "Open" will not save the data to the drive; you have to do this manually after the process is finished.



Step 15.

The results of the BLAST are similar to those on NCBI, only hits are contigs from the assembly, rather than the NCBI databases. Opening the BLAST report shows the list of query sequences in a table with their best hit by score.



Step 16.

Selecting one of the query sequences and clicking 'Open BLAST Output' will open the BLAST results for that query sequence. The first view will show the alignments of all the reported hits, color-coded by sequence % identity. Clicking on the "Table view" icon (bottom left, highlighted in red box) brings up the alignment statistics where you can select the values displayed and pick the contigs that you are interested in.

The screenshot shows the CLC Genomics Workbench 9.5.3 interface. On the left is a navigation tree with folders like 'Trimmings', 'Read Mappings', 'Reference Genomes', and 'Genome'. The main window displays BLAST results for a query sequence 'gp136'. The results are shown in a table with columns: Hit, Description, E-value, Score, %Gaps, and %Identity. The table lists four hits, with the first two having E-values of 2.79E-42 and 7.69E-19, and scores of 270.00 and 203.00 respectively. On the right, there is a 'BLAST HSP Table Settings' panel with checkboxes for various columns like 'Query sequence', 'Hit', 'Id', 'Description', 'E-value', 'Score', etc. At the bottom of the main window, there are buttons for 'Extract and Open', 'Download and Open', 'Download and Save', 'Open as NCBI', and 'Open Structure'.

Hit	Description	E-value	Score	%Gaps	%Identity
HAARFADIX_GSLv3-7_10_SL75768 (paired) trimmed (paired) contig list	No definition line	2.79E-42	270.00	0.00	6.42
HAARFADIX_GSLv3-7_11_SL75768 (paired) trimmed (paired) contig list	No definition line	7.69E-19	203.00	0.00	6.42
HAARFADIX_GSLv3-7_12_SL75770 (paired) trimmed (paired) contig list	No definition line	4.32	54.00	0.00	0.00
HAARFADIX_GSLv3-7_13_SL75771 (paired) trimmed (paired) contig list	No definition line	7.53	52.00	0.00	0.00

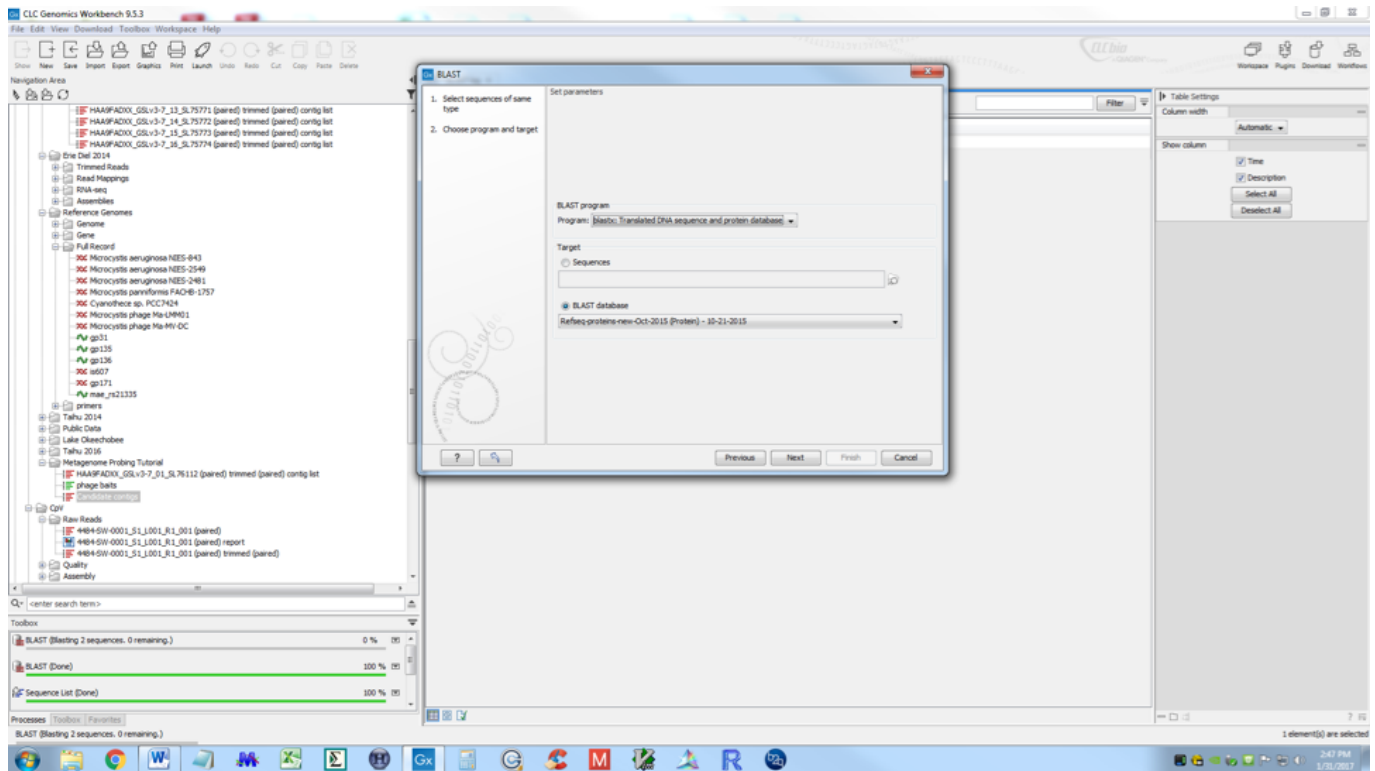
NOTES

Alyssa Alsante 07 Feb 2017

In this example, contigs 2119 and 332 exhibit decent alignments and might be worth exploring further.

Step 17.

Once you have picked the contigs that you want to examine further, you can then go back to the original assembly and extract them to a separate file. At this point, the contigs have only tentatively been identified since the blast query was biased towards your target sequences. Now you need to verify the results by blasting your contigs against a more comprehensive database, which usually is not possible given the size of assembled datasets.



⊕ NOTES

Alyssa Alsante 07 Feb 2017

In this example, I have extracted the contigs of interest to a file called "Candidate contigs" and I am going to BLAST them against the entire RefSeq protein database.

Step 18.

Since the RefSeq protein database is much more extensive than the assembly, this BLAST can take a bit longer. You may need to leave it overnight depending on the computer you use. The BLAST results are similar to what you saw previously, except the results are not biased towards your selected marker gene sequence list.

CLC Genomics Workbench 9.5.3

File Edit View Download Tools Workspace Help

Navigation pane

Tools

BLAST search results

Hit	Description	E-value	Score	%Identity	%Gaps
WP_011118	transposase [Moraxella phage No.12840]	0.00	1,881.00	98.71	0.00
WP_04555650	transposase [Moraxella aeruginosa]	0.00	1,874.00	98.71	0.00
WP_046643507	transposase [Moraxella aeruginosa]	0.00	1,866.00	97.69	0.00
WP_002764897	transposase [Moraxella aeruginosa]	0.00	1,855.00	97.44	0.00
WP_04664350	transposase [Moraxella aeruginosa]	0.00	1,853.00	97.44	0.00
WP_008277413	transposase [Carnithecus sp. CC01010]	0.00	1,607.00	84.46	0.00
WP_015955521	transposase [Carnithecus sp. PCC 7424]	0.00	1,603.00	84.46	0.00
WP_04595689	transposase, partial [Moraxella aeruginosa]	0.00	1,587.00	97.81	0.00
WP_013320475	transposase [Carnithecus sp. PCC 7822]	2.06E-176	1,333.00	72.57	0.29
WP_047155615	transposase [Trichostemon erythraeus]	4.54E-175	1,326.00	69.37	0.26
WP_04412730	transposase [Trichostemon erythraeus]	8.82E-175	1,324.00	69.37	0.26
WP_015227332	transposase [Halobacterium sp. PCC 7418]	1.27E-174	1,323.00	70.86	0.00
WP_015228617	transposase [Halobacterium sp. PCC 7418]	1.13E-173	1,315.00	69.71	0.00
WP_03809563	transposase [Halobacterium sp. PCC 7418]	8.93E-173	1,311.00	68.22	0.00
WP_01558774	transposase [Carnithecus sp. PCC 7424]	5.80E-172	1,307.00	70.43	0.29
WP_038794065	transposase [Planktobius mougeoti]	5.92E-172	1,306.00	68.37	0.00
WP_038794109	transposase [Planktobius mougeoti]	7.27E-172	1,305.00	68.37	0.00
WP_038786991	MALTOPOLYMER transposase [Planktobius]	9.03E-172	1,303.00	68.37	0.00
WP_02725539	transposase [Planktobius mougeoti]	1.17E-171	1,304.00	68.37	0.00
WP_027240243	transposase [Planktobius mougeoti]	1.17E-171	1,304.00	68.37	0.00
WP_027240218	transposase [Planktobius mougeoti]	1.79E-171	1,303.00	68.37	0.00
WP_027254878	transposase [Planktobius mougeoti]	2.10E-171	1,302.00	68.37	0.00
WP_042152157	transposase [Planktobius mougeoti]	1.09E-170	1,298.00	69.11	0.00
WP_039742346	transposase [Halobacterium sp. PCC 7418]	1.33E-169	1,288.00	70.94	0.85
WP_017621715	transposase [Halobacterium sp. PCC 7418]	3.49E-167	1,272.00	71.79	0.85
WP_017511181	transposase [Halobacterium sp. PCC 7418]	5.79E-167	1,274.00	67.64	0.00
WP_052490472	transposase [Halobacterium sp. PCC 7418]	8.63E-167	1,271.00	70.66	0.85
WP_017621856	transposase [Halobacterium sp. PCC 7418]	1.38E-166	1,264.00	71.51	0.85
WP_044450271	transposase [Halobacterium sp. PCC 7418]	1.82E-166	1,267.00	70.94	0.85
WP_048868410	transposase [Scytospora tolypochridae]	2.19E-166	1,267.00	70.66	0.85
WP_017621856	transposase [Halobacterium sp. PCC 7418]	4.02E-166	1,265.00	71.23	0.85
WP_044333801	transposase [Halobacterium sp. PCC 7418]	8.98E-166	1,264.00	68.41	4.12
WP_006161249	transposase [Halobacterium sp. PCC 7418]	1.22E-165	1,261.00	71.14	0.00
WP_006197030	transposase [Halobacterium sp. PCC 7418]	4.60E-165	1,258.00	71.23	0.85
WP_044333801	transposase [Halobacterium sp. PCC 7418]	2.13E-164	1,253.00	70.37	0.85
WP_006104684	transposase [Halobacterium sp. PCC 7418]	2.72E-164	1,253.00	69.14	0.00
WP_007306762	transposase [Halobacterium sp. PCC 7418]	2.74E-164	1,253.00	70.37	0.28
WP_014276129	transposase [Halobacterium sp. PCC 7418]	2.54E-163	1,246.00	70.29	0.00
WP_006103351	transposase [Halobacterium sp. PCC 7418]	7.45E-162	1,236.00	68.29	0.00
WP_044333801	transposase [Halobacterium sp. PCC 7418]	1.12E-161	1,238.00	65.97	8.38
WP_048868442	transposase [Scytospora tolypochridae]	4.89E-161	1,234.00	66.40	5.42
WP_015954174	transposase [Carnithecus sp. PCC 7424]	2.19E-159	1,228.00	70.36	0.57
WP_006104684	transposase [Halobacterium sp. PCC 7418]	1.09E-154	1,191.00	65.94	5.18
WP_048867418	transposase [Scytospora tolypochridae]	2.38E-154	1,188.00	67.24	0.28
WP_028122403	transposase [Scytospora tolypochridae]	5.72E-154	1,185.00	67.22	1.97
WP_015118982	transposase [Scytospora tolypochridae]	3.36E-153	1,180.00	64.56	3.85
WP_038081305	transposase [Scytospora tolypochridae]	9.15E-153	1,179.00	64.04	8.40

Extract and Open Download and Open Download and Save Open at NCBI Open Structure

Processes: Tools: Favorites

1 row selected

1/31/2017

NOTES

Alyssa Alsante 07 Feb 2017

In this example, I have identified a different contig (2727), as originating in phage, which I want to look at more closely. At this point, I can look at read mappings to the original to determine the contig's expression level in the sample or build a phylogenetic tree to determine how the gene is related to reference sequences.