Screening Sequencing Datasets for Marker Genes in CLC Genomics Version 2

Dr. Steven Wilhelm, Josh Stough

Abstract

Contact Dr. Steven Wilhelm (wilhelm@utk.edu) or Josh Stough (jstough@vols.utk.edu) for additional information regarding this protocol.

Citation: Dr. Steven Wilhelm, Josh Stough Screening Sequencing Datasets for Marker Genes in CLC Genomics. **protocols.io**
dx.doi.org/10.17504/protocols.io.hgsb3we
Published: 29 Mar 2017

Before start

You will need:

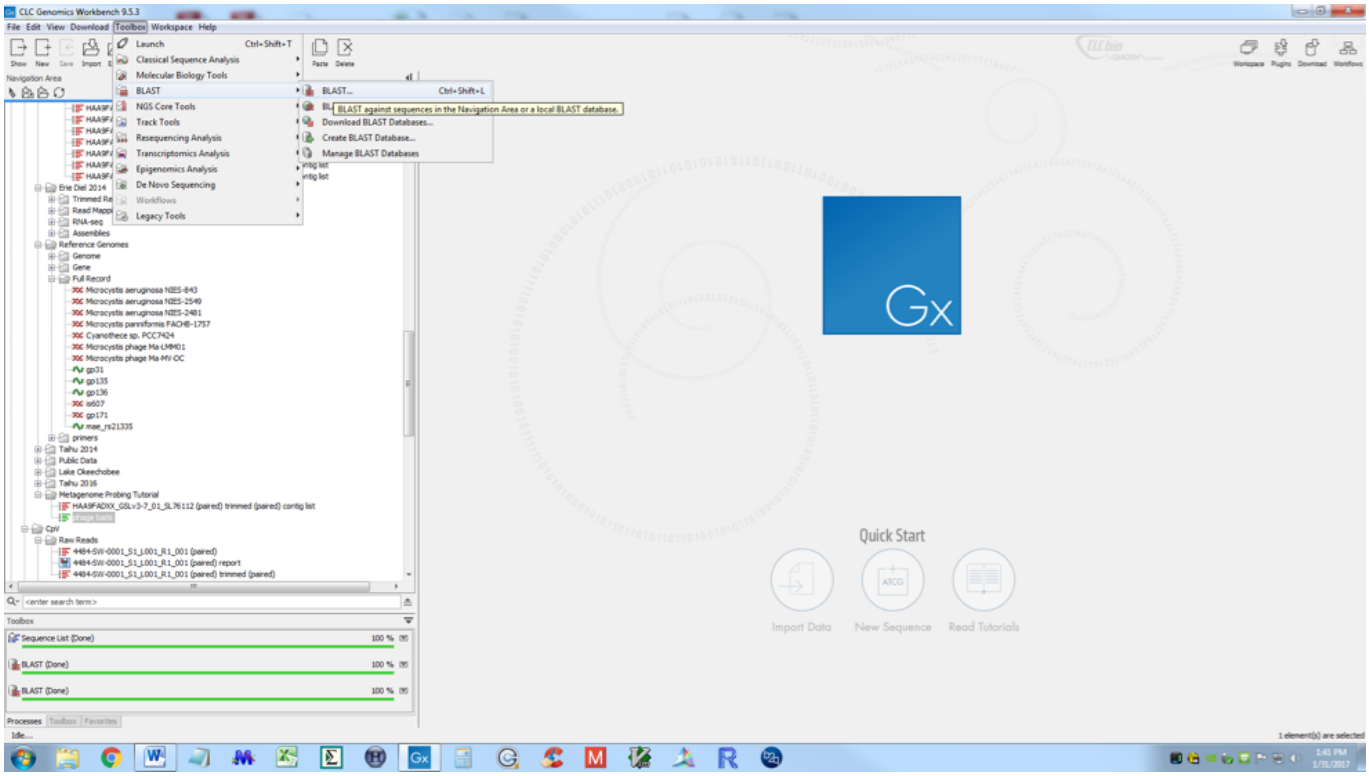
Assembled sequencing dataset--the dataset to be screened needs to be assembled and the contigs extracted. You may want to filter out contigs that are too short or have low coverage.

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

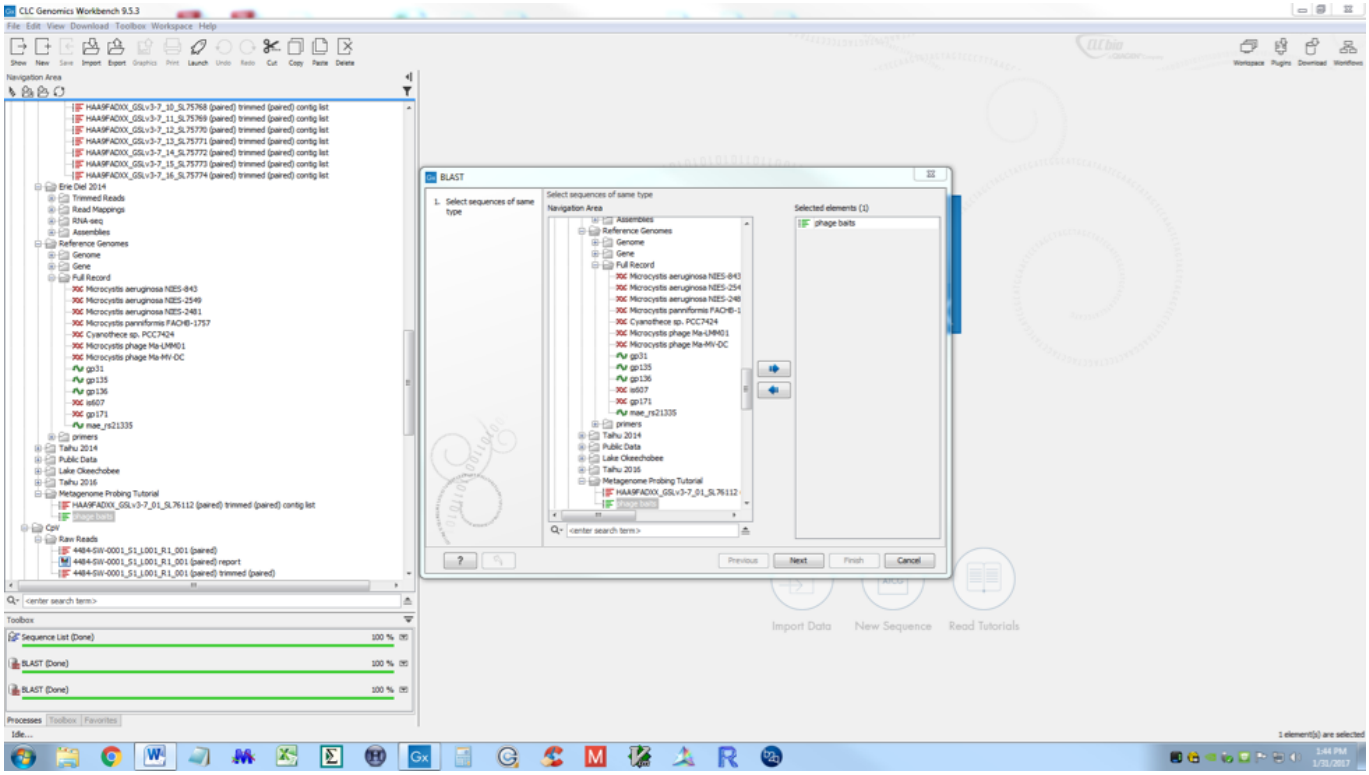
Step 1.

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 29 Mar 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.

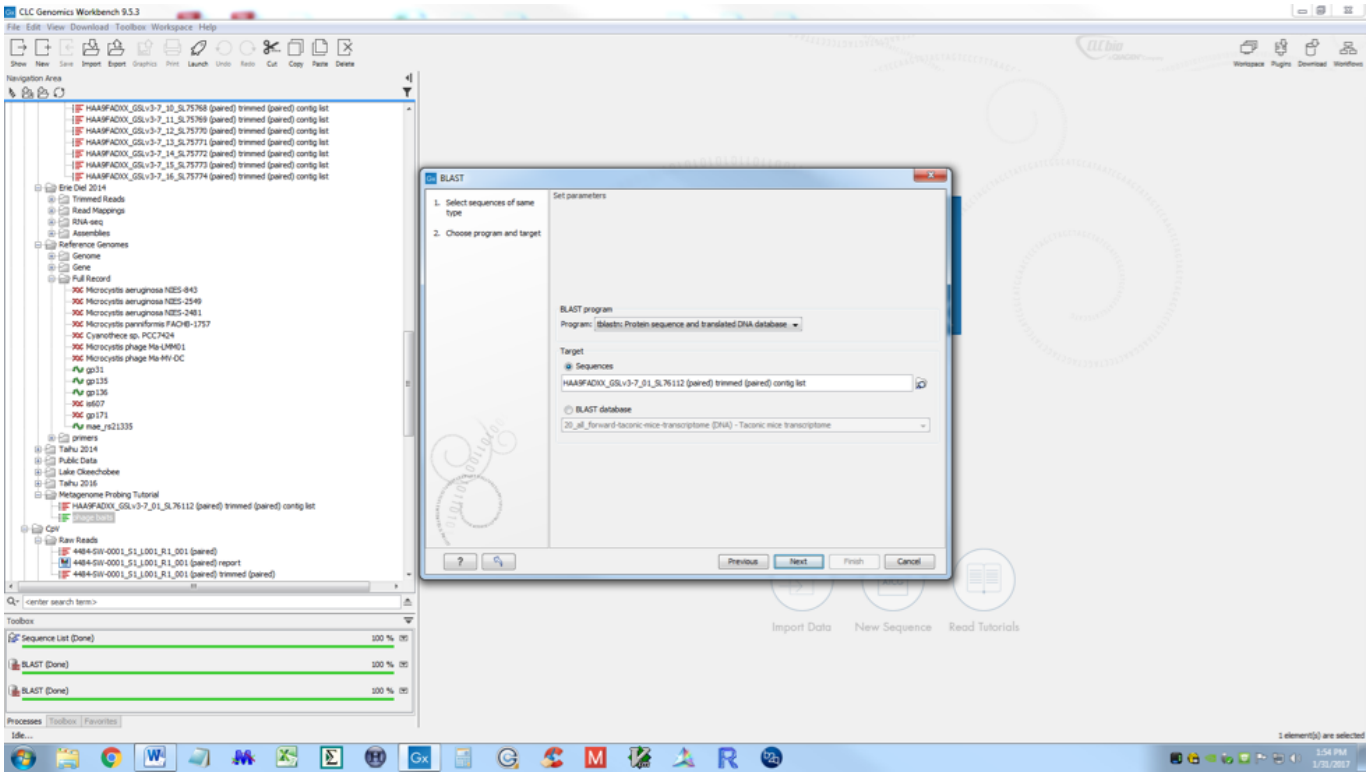
NOTES

Alyssa Alsante 29 Mar 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 29 Mar 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.

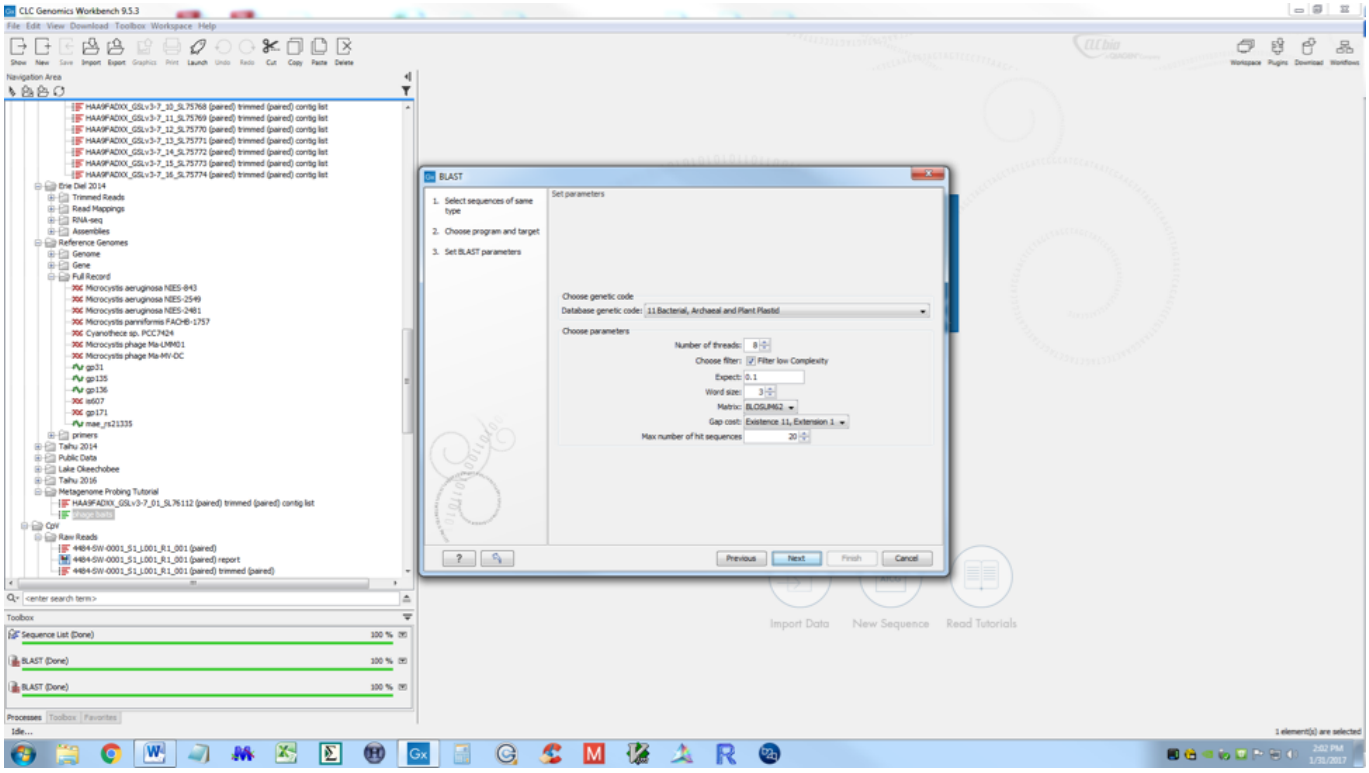
The sequence files available will differ depending on which BLAST program that you selected in the previous step.

NOTES
Alyssa Alsante 29 Mar 2017

For example, selecting "blastp" requires that both the query and sceened 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 29 Mar 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.

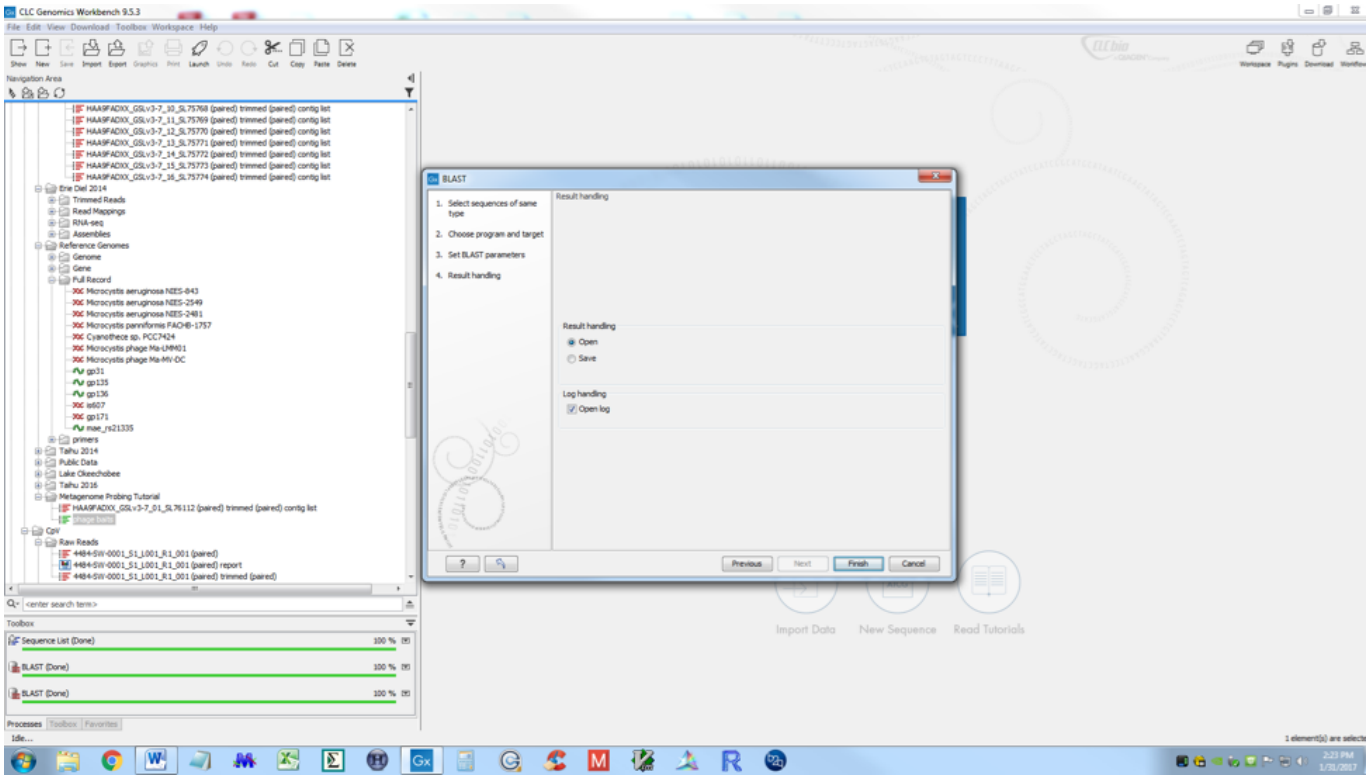
Remaining Settings

Step 13.

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

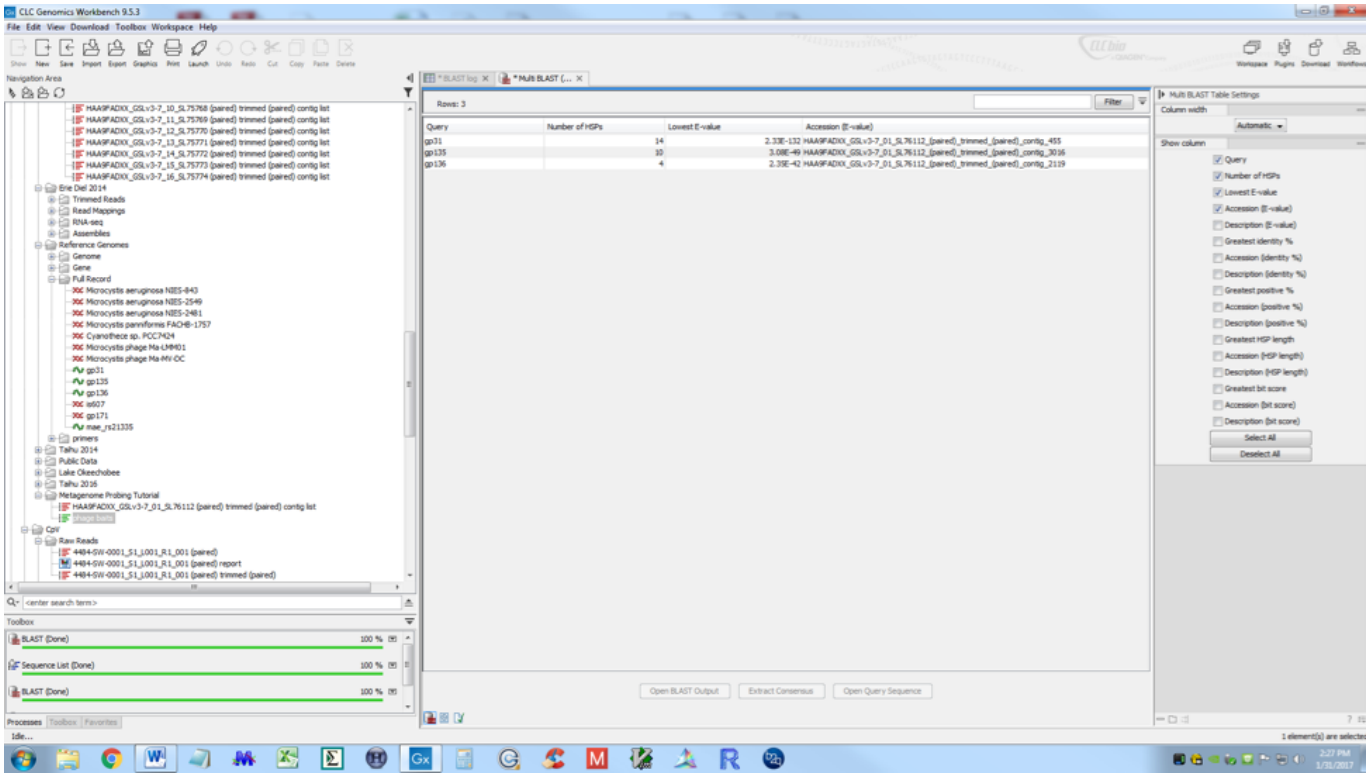
Step 14.

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.



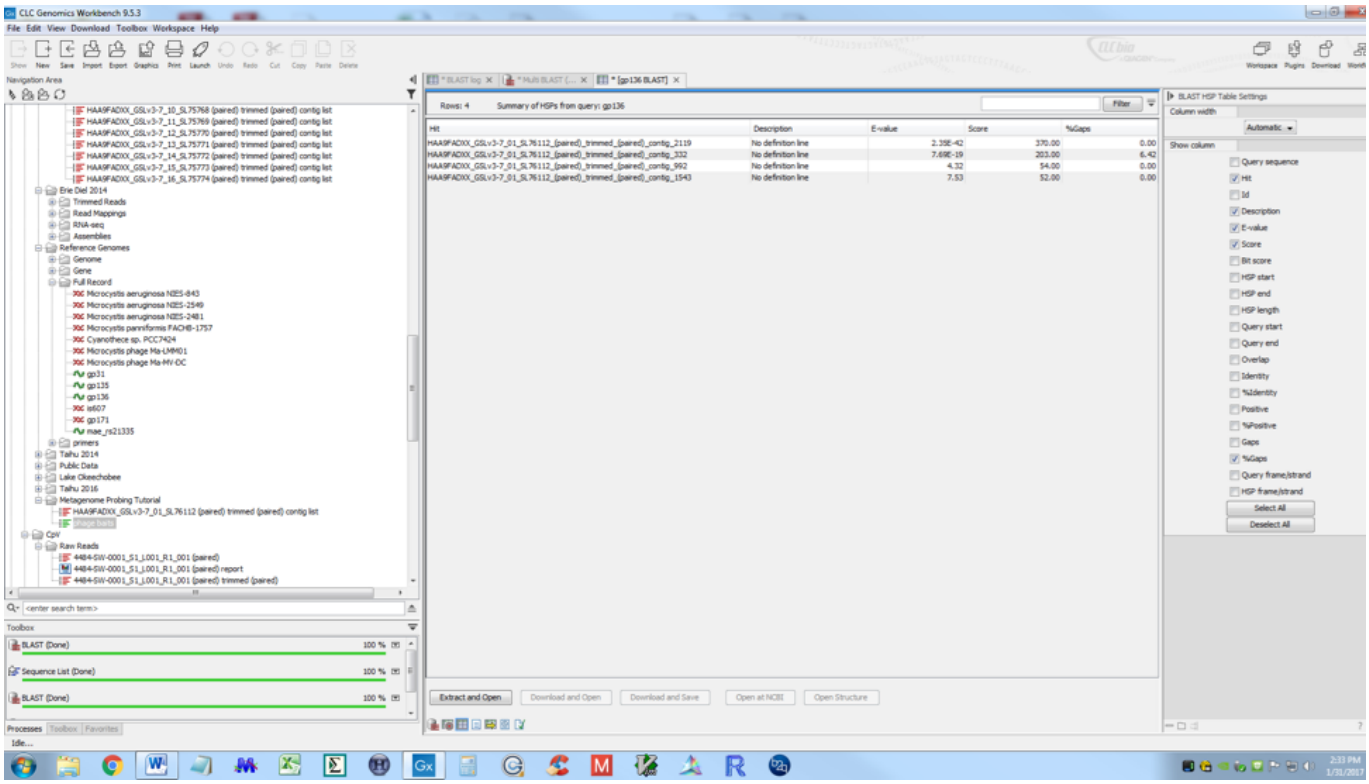
Step 15.

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.



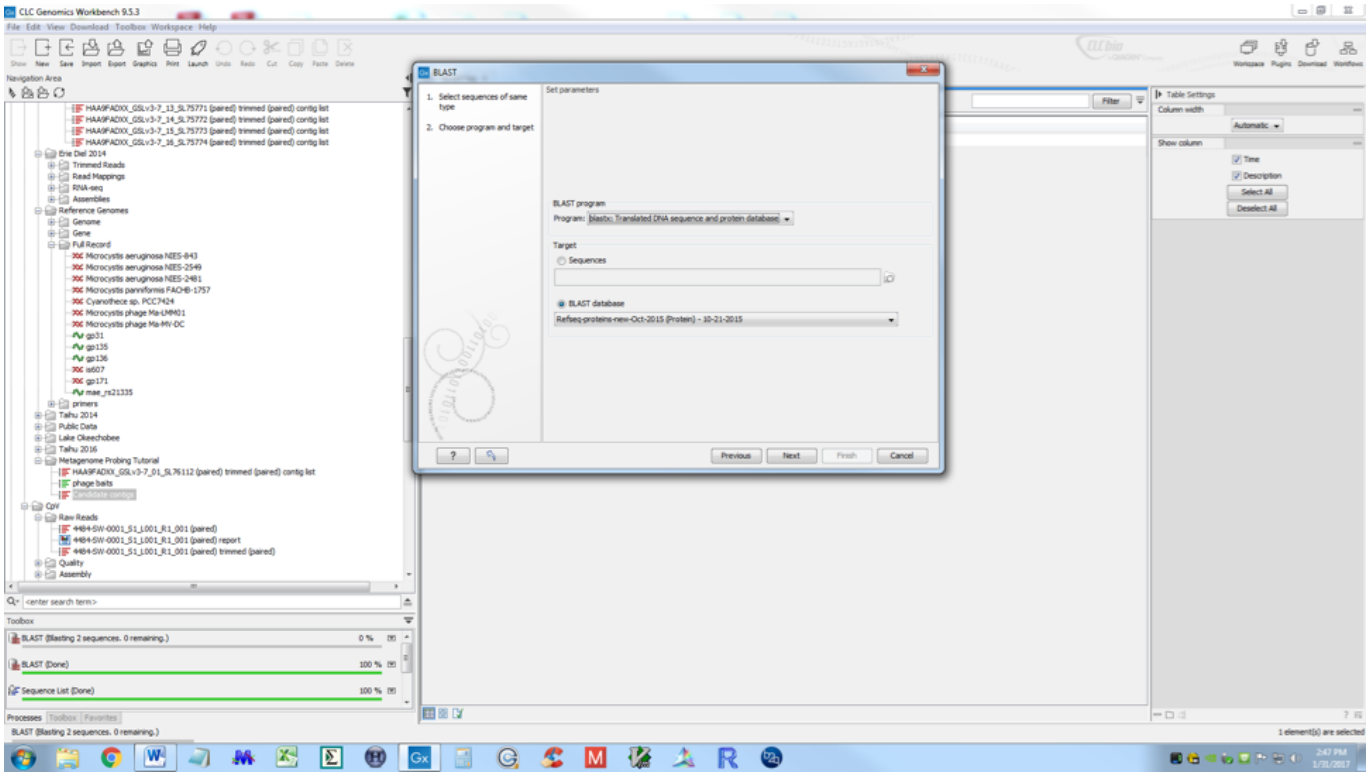
NOTES

Alyssa Alsante 29 Mar 2017

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

Step 17.

Once you have picked the contigs that you want to examine further, you can then go back to the originial 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.



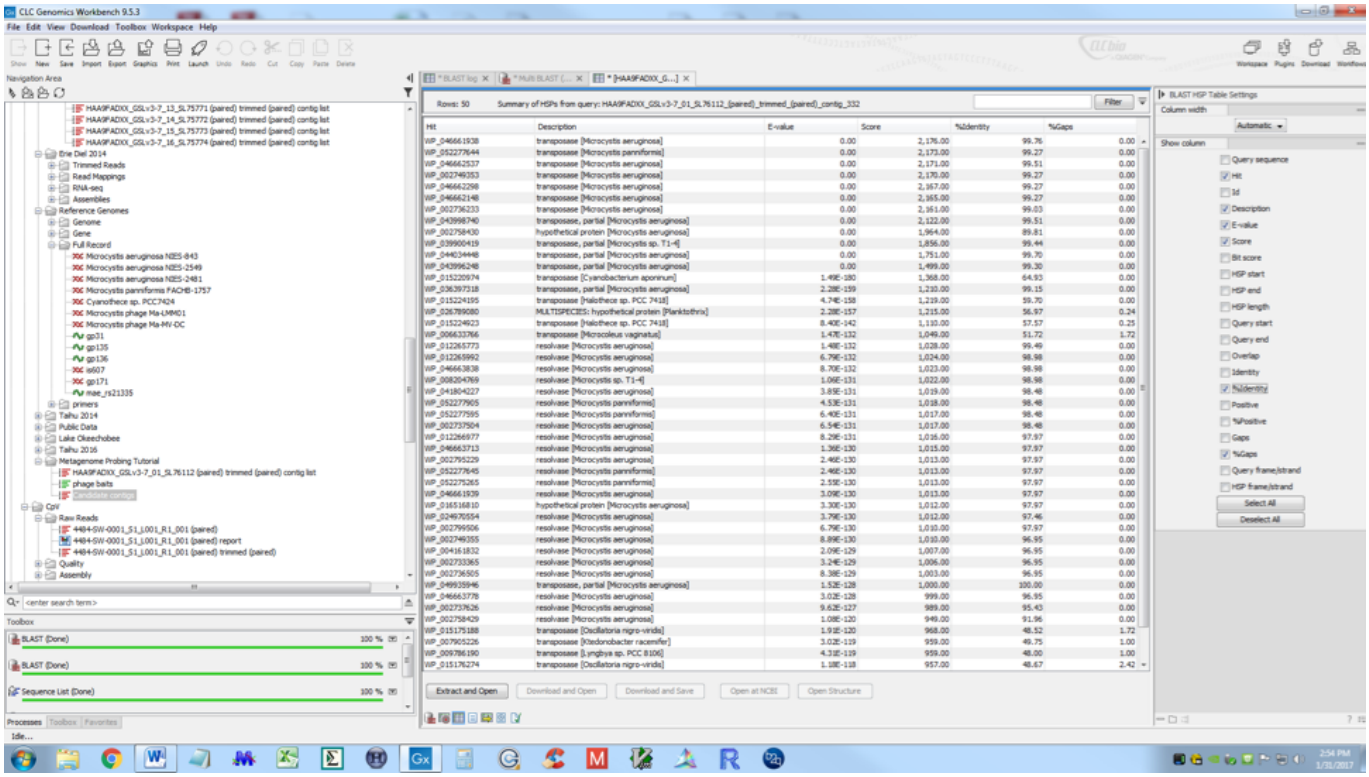
NOTES

Alyssa Alsante 29 Mar 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.



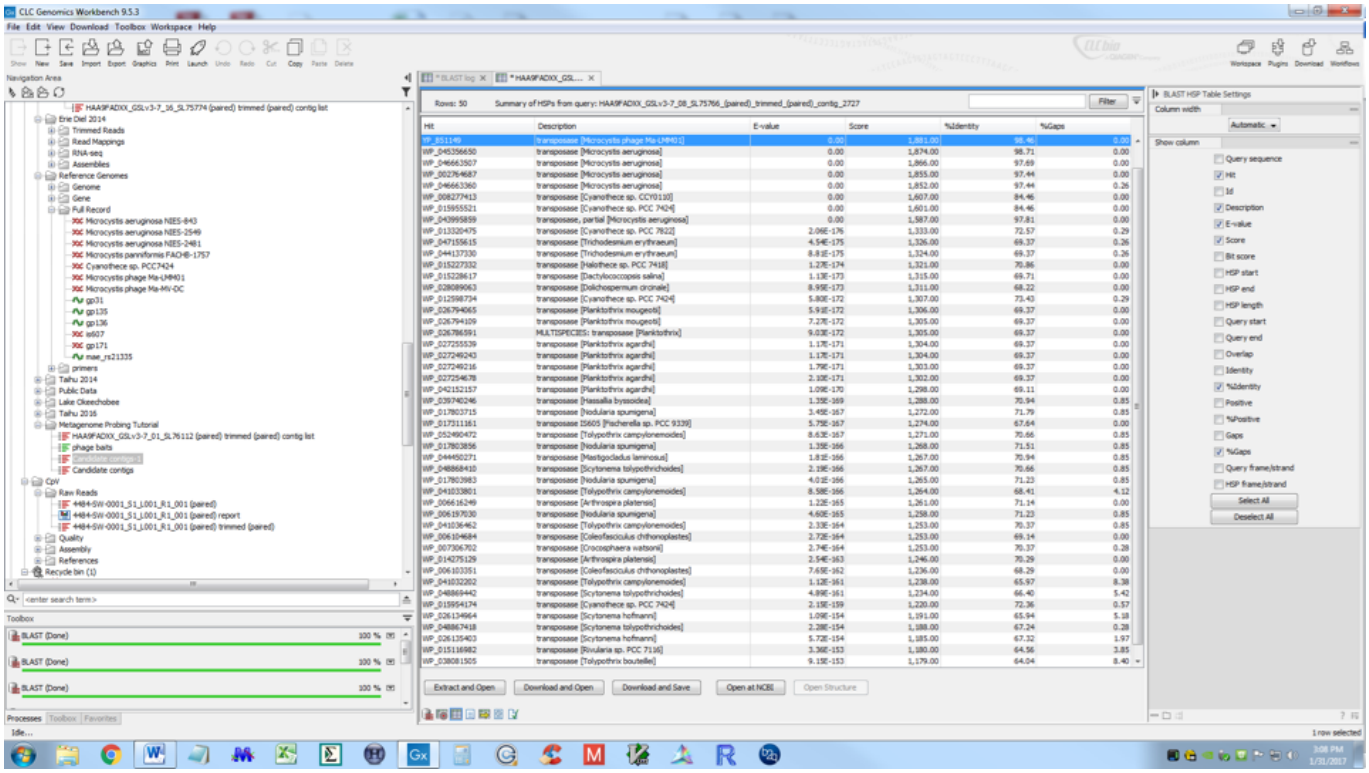
NOTES

Alyssa Alsante 29 Mar 2017

In this example, the two contigs that I pulled from the assembly (332 and 2119) most likely originated in the cyanobacterium, *Microcystis aeruginosa*, rather than the phage that I was looking for. At this point, I would likely conclude that the phage gene that I was looking for was not expressed in the sample that I checked.

Step 19.

If the BLAST results show exactly what you are looking for, you can move forward with some confidence that the contigs you extracted from the dataset are worth examining more closely.



NOTES

Alyssa Alsante 29 Mar 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.