Tutorial: De novo assembly and BLAST

December 15, 2009

CLC bio
Finlandsgade 10-12 · 8200 Aarhus N · Denmark
Telephone: +45 70 22 55 09 · Fax: +45 70 22 55 19
www.clcbio.com · info@clcbio.com

Tutorial: De novo assembly and BLAST

This tutorial takes you through some of the tools for a typical *de novo* sequencing work flow with a data set from a high-throughput sequencing machine. As an example we use an *E. coli* data set consisting of a little more than 400,000 reads from a 454 sequencer. [1]

Importing the data

First, download the data set from our web site: `http://download.clcbio.com/testdata/raw_data/454.zip`. Unzip the file somewhere on your computer (e.g. the Desktop).

Start the *CLC Genomics Workbench* and import the data:

File | **Import High-throughput Sequencing** | **Roche 454**

This will bring up the dialog shown in figure 1

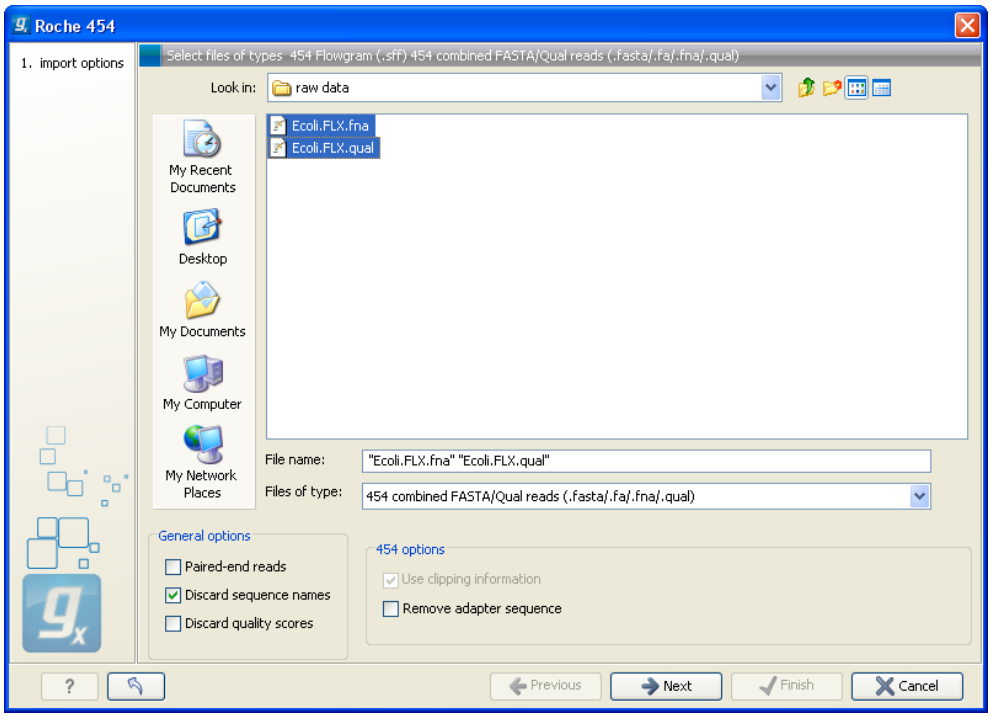


Figure 1: *Choosing the file you wish to import.*

Select the `Ecoli.FLX.fna` and `Ecoli.FLX.qual` files that come from the downloaded zip file. Make sure the **Remove adapter sequence** checkbox is checked and that the **Paired-end reads** checkbox is NOT checked. The option to discard read names is not significant in this context because of the relatively small amount of reads. Click **Next**, choose to **Save** and click **Finish**.

After a short while, the reads will be imported.

[1] Note that there are special system requirements for *CLC Genomics Workbench*, see `http://www.clcbio.com/index.php?id=72`

Assembly

The first step is to do a *de novo* assembly of the reads. Briefly explained, this means that we want the *CLC Genomics Workbench* to create longer contiguous sequences out of the relatively short reads that are on average 235 bp long:

Toolbox | **High-throughput Sequencing** ( ) | **De Novo Assembly** ( )

This shows the dialog in figure 2.

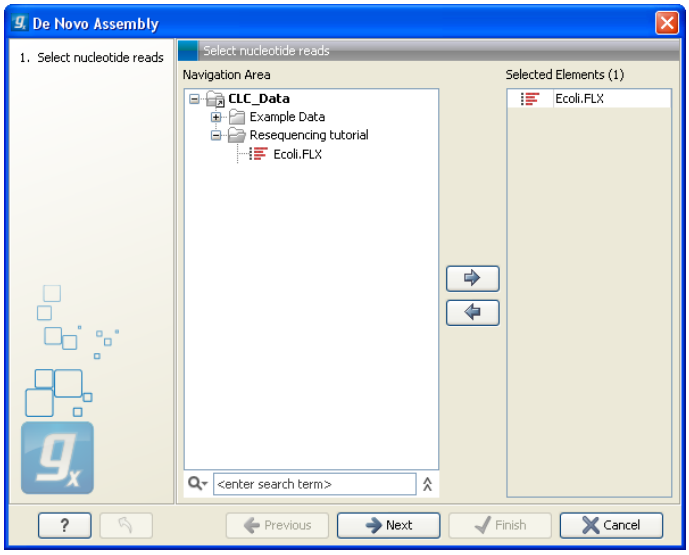


Figure 2: *Select sequence list containing the reads.*

Select the `Ecoli.FLX` ( ) sequence list and add it to the panel to the right. Click **Next**, **Next** and **Finish**.

For information about the settings, click the **Help** (?) button.

The assembly will take several minutes. There are two steps in the assembly process. In the first step, contig sequences are created by assembling all the reads. Second, all the reads are assembled using the contig sequences as references. This is done to create what we call full contigs, i.e. including alignment of all the aligned reads (if you only wish to see the contiguous sequences without the reads, you should choose **Create simple contig sequence** before you clicked Finish).

The result of the assembly is a contig table as shown in figure 3.

Each row in the table represents a contig. There should be around 142 contigs (the number can vary a little from version to version). Double-clicking a row will open the contig.

BLAST the results

One way to investigate the contigs would be to BLAST them against public databases. In this tutorial we will take out five of the contigs and BLAST them. Note that you need to be connected to the internet to do this.

First, sort the table according to the contig length by clicking the **Length of consensus sequence** column header. Now find the contigs that are just above 10,000 nucleotides. Select the five first

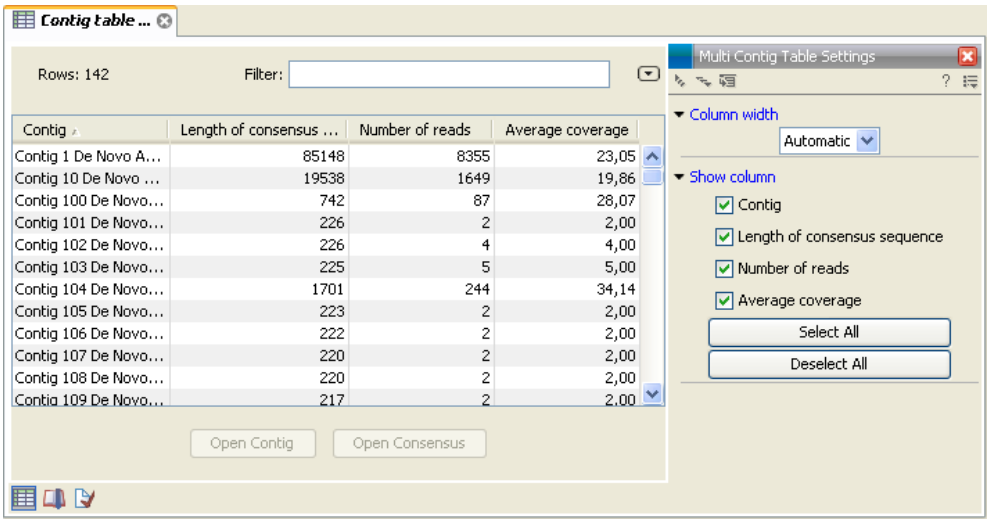


Figure 3: *A contig table.*

as shown in figure 4.

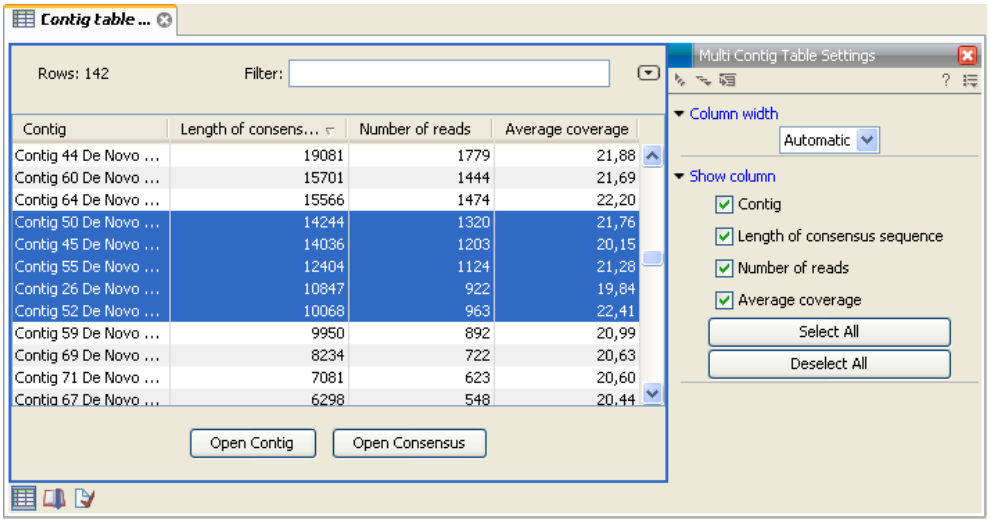


Figure 4: *Selecting five contigs that are just above 10,000 nucleotides long.*

Then click the **Open Consensus** button at the bottom of the table. That will create a sequence list with the five contig (consensus) sequences that you have selected.

To BLAST these sequences, first make sure the sequence list with the five contigs has been **Saved** (). Then:

Toolbox | **BLAST Search ()** | **NCBI BLAST ()**

Select the sequence list that you have saved and click **Next**. To be completely open in terms of what kind of data there is in the sequences, choose the **Nucleotide collection (nr)** database (see figure 5).

Click **Next** and leave the settings at their default (to make sure the settings are back to default, click the **Reset to CLC Standard Settings** () button). Click **Next** again. This will allow you to specify how the result of the BLAST search should be delivered (see figure 6).

For high-throughput purposes, the **Create overview BLAST table** option is preferable. Instead of

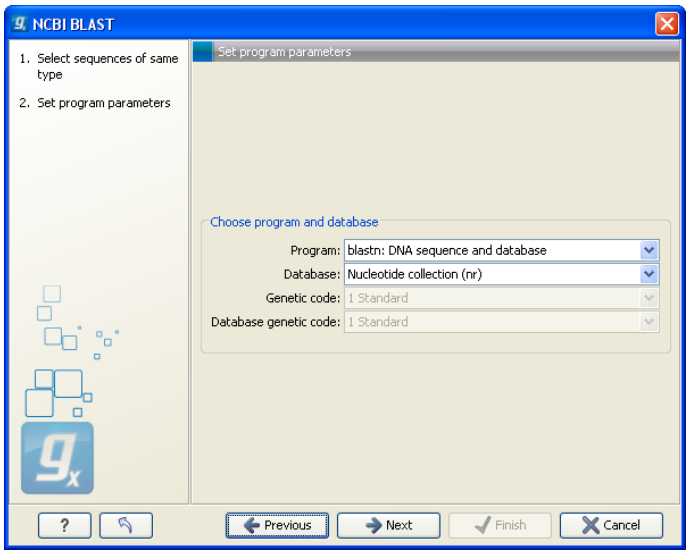


Figure 5: *Selecting the Nucleotide collection (nr) database.*

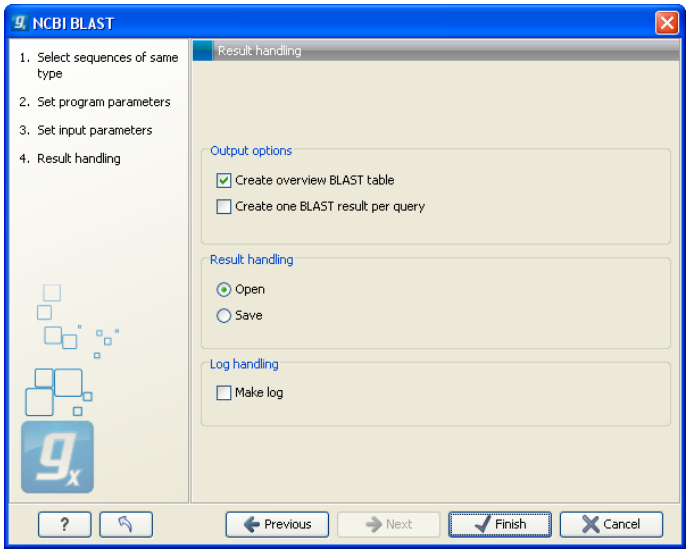


Figure 6: *Selecting the Create overview BLAST table option.*

creating five separate BLAST results, you will get one overview table summarizing the results. Click **Finish** to run the BLAST search.

The result is shown in figure 7.

For each contig, you can see the description of the top hit which is *E. coli* str. K12 in all cases. If you wish to inspect the individual BLAST results in more detail, simply double-click the row and the BLAST result will open. If you wish to use the result for reporting, e.g. in a spread-sheet, you can **Copy** ( ) the contents of the table and paste into e.g. a spreadsheet for further processing. To select all the rows in the table, press Ctrl + A (⌘ +A on Mac).

You can also use the filter at the top to sort the table (e.g. to only show BLAST results with an E-value lower than a certain threshold).

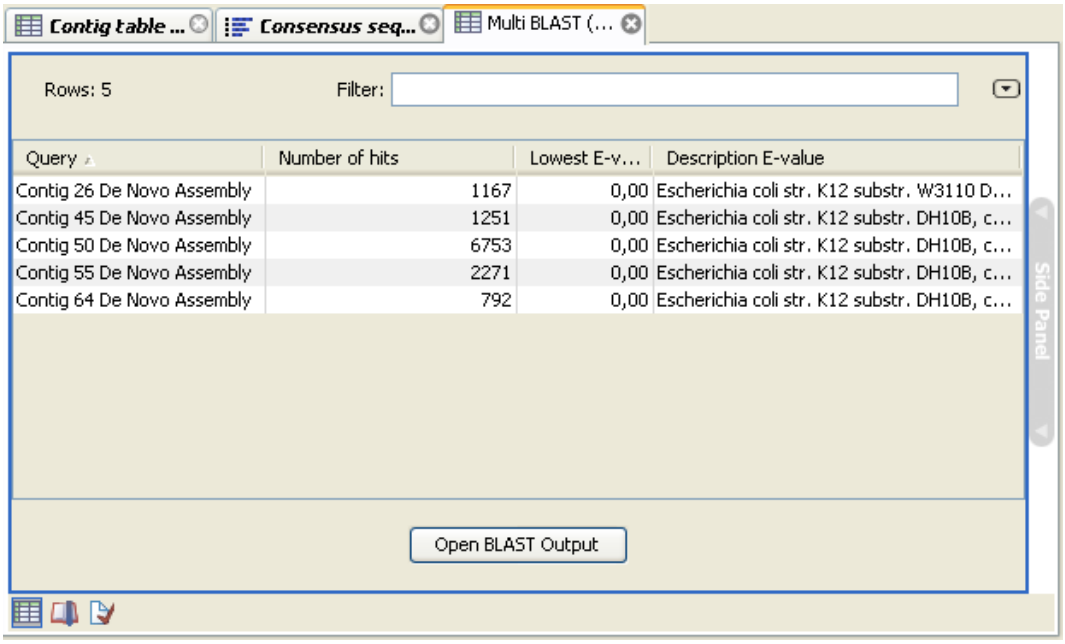


Figure 7: *The overview BLAST table.*