Tutorial: De novo assembly and BLAST

April 20, 2012

CLC bio
Finlandsgade 10-12 · 8200 Aarhus N · Denmark
Telephone: +45 70 22 55 09 · Fax: +45 70 22 55 19
www.clcbio.com · support@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. Here, we de novo assemble some reads, and then search a database at the NCBI with some of the contigs produced.

For this tutorial, we use an *E. coli* data set consisting just over 400,000 reads from a 454 sequencer. [1]

Importing the data

If you don't already have this sample data set:

1. Download the data set from our web site: `http://download.clcbio.com/testdata/raw_data/454.zip`.
2. Unzip the file somewhere on your computer (e.g. the Desktop).
3. 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

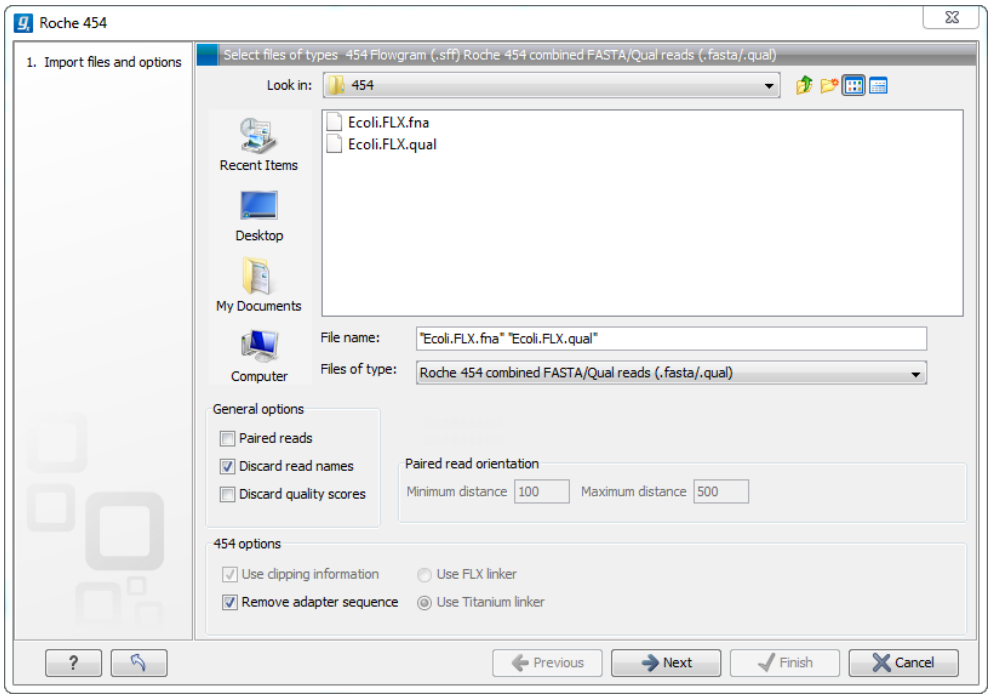


Figure 1: *Choosing the file you wish to import.*

4. 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 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. So you can leave this checked, or unchecked, as you like.

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

5. Click on the button labelled **Next**.
6. Click on the button labelled **Save** in the wizard page that appears.
7. Click on the folder you wish to save to.
8. Click on the button labelled **Finish**.

After a short while, the reads will be imported.

Assembly

The reads we are using in this tutorial are on average around 235 bp long. By doing a *de novo* assembly of the reads, we are trying to create longer, contiguous sequences from these relatively short sequences.

We recommend that you refer to the CLC Genomics Workbench manual (`http://www.clcbio.com/files/usermanuals/CLC_Genomics_Workbench_User_Manual.pdf`) for information on how the de novo assembly tool works, and the meaning of the parameters you have control over.

There are two possible types of output you can generate from the de novo assembly tool in the *CLC Genomics Workbench*: simple contigs, which is a sequence list of the contigs generated, and mapping objects, where a read mapping is carried out after the initial assembly, to map your reads back to the contigs created.

If you plan to do variant detection later, then generating mapping output can make sense. Also, there is an option when setting up your assembly, allowing you to update your contigs based on the mapping. If you enable this option, then areas of your assembled contigs that no reads map back to will be cut out of the final contigs. The idea there is that if no reads map back to a region of a contig, then there is no evidence in the data that that region exists.If this is what you want to do, then choosing to generate mapping output makes sense. You can, of course, retrieve the contig sequences from mapping objects afterwards, should you wish.

Sequence lists, like a list of contigs, can be used for different types of tasks, such as searching for patterns, or BLAST searching. As this is what we will proceed to do in this tutorial, we will choose to generate a sequence list of contig sequences in this tutorial. There is an optional section at the end of the tutorial about generating mapping output instead, and then extracting contig sequences from that for downstream use.

1. Start up a de novo assembly analysis by going to:

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

   This shows the dialog in figure 2.
2. Select the `Ecoli.FLX` ( ) sequence list and add it to the panel to the right.
3. Click on the button labelled **Next**.
4. You now have the opportunity to set some parameters for your assembly. Here, we will accept the defaults, as shown in figure 3.

   You can directly access the manual to find out more about these parameters by clicking on the **Help** (?) button in the far left hand, bottom edge of the wizard window.

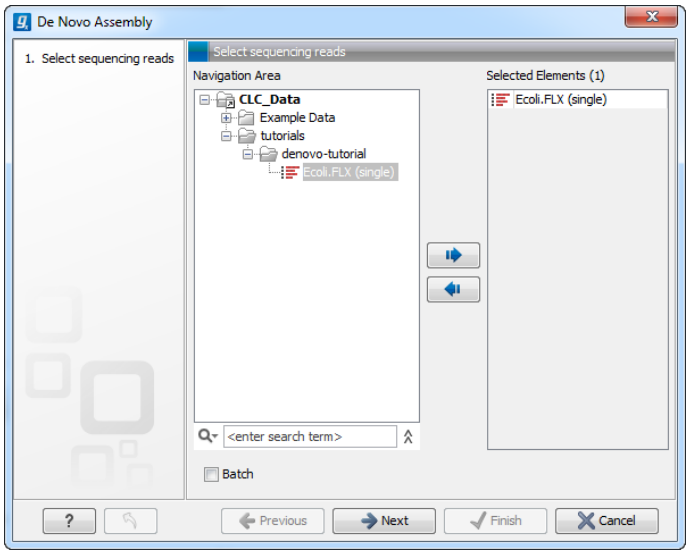


Figure 2: *Select sequence list containing the reads.*

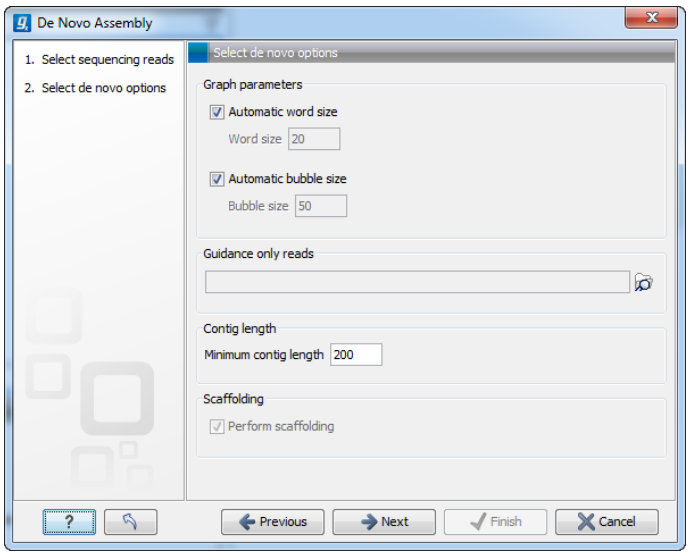


Figure 3: *De novo assembly parameter options.*

5. Click on the button labelled **Next**.
6. On this next page of the wizard, you get to choose the type of output you will generate. As mentioned earlier, we will choose to create **simple contig output**. That is, we will generate a sequence list of the contig sequences assembled during this job.

   Please click on the option **Create simple contig sequences (fast)**, as shown in figure 4.

   Note that this causes all other options on this wizard page to be greyed out. This is because all the other options pertain to mapping reads back to the contigs.
7. Click on the button labelled **Next**.
8. Choose to create a report and to Save the results.
9. Click on the button labelled **Next**.
10. Choose a place to save your results to, adn then click on the button labelled **Finish**.

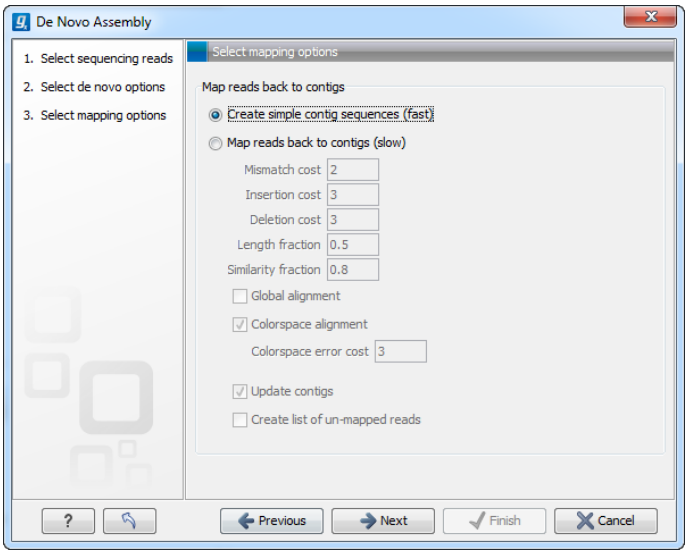


Figure 4: *De novo assembly output types. Here we choose to create simple contigs. Note that if you choose to map reads back to the contigs, a read mapping job will be launched directly after the assembly task itself is done, and the outputs generated would be mapping objects.*

The assembly should be quite quick, but this depends on how fast your computer is. You can follow the progress in the Processes pane of the Workbench.

Note that the progress bar does not progress smoothly. It gives an indicator of the stage of the assembly more than an idea of the time to finish. For large assemblies, you may notice the the progress bar may spend a lot of time at a certain percentage. This generally means it is working on a certain stage of the assembly. You can see a brief text description of the stage it is working on written in brackets above the progress bar. See figure 5.

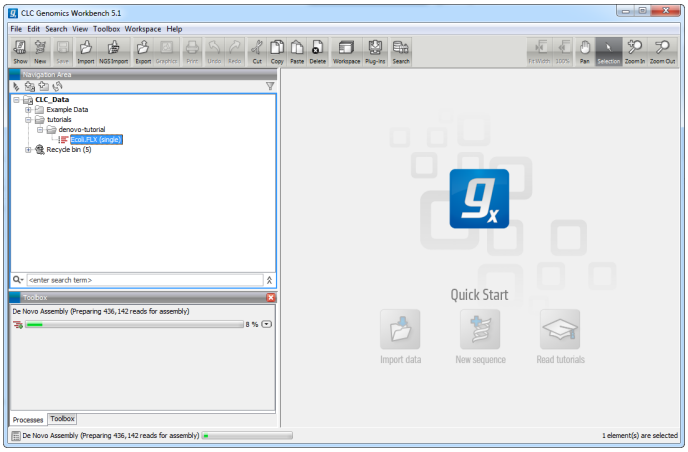


Figure 5: *You can monitor the progress of an assembly using the Processes pane in the Workbench.*

Investigate the results

The outputs of the assembly are a sequence list and a report. If you have not changed the name of the sample data, the contig list will be called **Ecoli.FLX (single) contig list**. The report will be called Summary de novo report.

1. Double click on the report object to open it. You will find general summary information in it

such as the nucleotide distribution, information on contig lengths, as well as the N25, N50 and N75 values.

2. Double click on the sequence list you have created. This opens it in the Navigation pane of the Workbench. You can zoom in using the 100% button in the top toolbar if you wish.

3. Click on the **table view icon** ( ) at the bottom of the navigation area. You will now see a table view. You can see how many contigs were generated by how many rows are reported to be in the table.

4. Click on the length column heading. This sorts the table according to the length of the contigs, as shown in figure 6.

5. Highlight the top five rows - here, the longest contigs - as shown in figure 6.

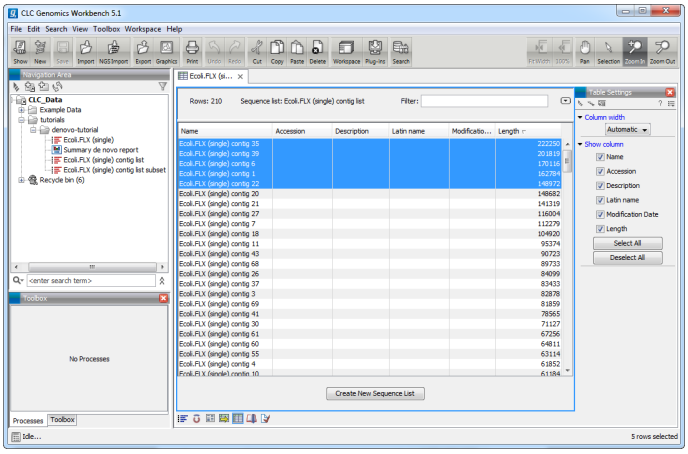


Figure 6: *The table view of the sequence list, with the top five (longest) contigs highlighted.*

6. Click on the button labelled **Create a New Sequence List**.

   This opens a new sequence list in a tab in the Viewing area of the Workbench.

7. Save this new list. You can do this in a number of ways. For example, you could:

   - Right click on the tab at the top of the new view, and choose **Save As...** from the menu that appears, or
   - Move the cursor over the tab at the top of the new view, depress the left mouse button. Keeping it depressed, drag the tab over into a folder in the Navigation area of the Workbench. The data will then be saved into that folder.

BLAST some contigs against nr

One way to investigate the contigs would be to see if there are any similar sequences in public databases. This can be done using the BLAST program suite. In this tutorial we will run a BLAST search of the five contigs we saved into a new list earlier. We will search for similar sequences in a database at the NCBI. **You need to be connected to the internet to do this tutorial.**

1. Start up the BLAST tool in the Workbench by going to:

   Toolbox | **BLAST** ( ) | **BLAST at NCBI** ( )

2. Select the sequence list with 5 contig sequences that you just saved and click on the right hand arrow to select it as the element with the query sequences that you wish to run the blast search with.
3. Click on the button lablled **Next**.
4. Choose **blastn** as the program type to run, and the **Nucleotide collection (nr)** as the database you wsih to search.
5. Click on the button lablled **Next**.
6. On this next page of the wizard, you can select the parameters you wish the blast search to be run with. There is some explanation of what these mean in the manual. You can access this by clicking on the **Help** (?) button in the far left hand, bottom edge of the wizard window.
7. Change the settings of the parameters so they match those in figure 7 . Here, we set a longer word size, and a much smaller e-value than the default. We have also limited the maximum number of matches returned to us to 10. The sequences we have sent to the NCBI are quite long, so the search will take a little while.

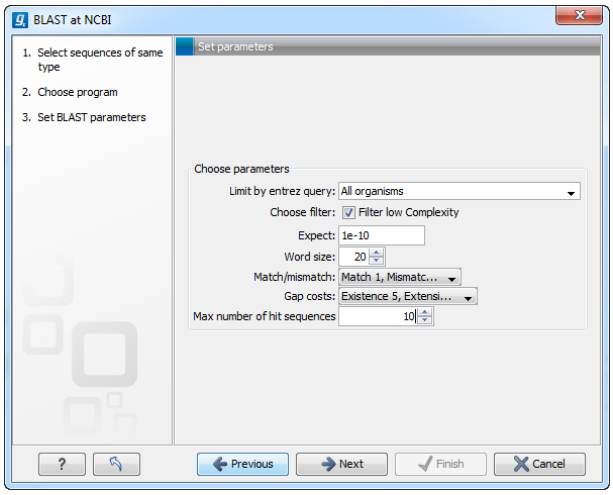


Figure 7: *Setting parameters for a BLAST search.*

8. Choose to **Save** your results.
9. Click on the button lablled **Next**.
10. Choose a place to save your results to, adn then click on the button labelled **Finish**.

Viewing the BLAST results

1. When the Blast search is finished, open the resulting data object. It should be called **Multi BLAST (5 sequences)**.

   You searched here with five query sequences. The table you just opened contains a brief overview of the results of your search for each of these query sequences.

2. Look to the right of the table, in the view settings pane. You have many choices as to what columns you wish to see in this table. In figure 8, the Description column for the hit

with the lowest evalue has been selected for display, in addition to the options checked by default. Note that you can choose to display your top hit according to e-value andor according to identity.

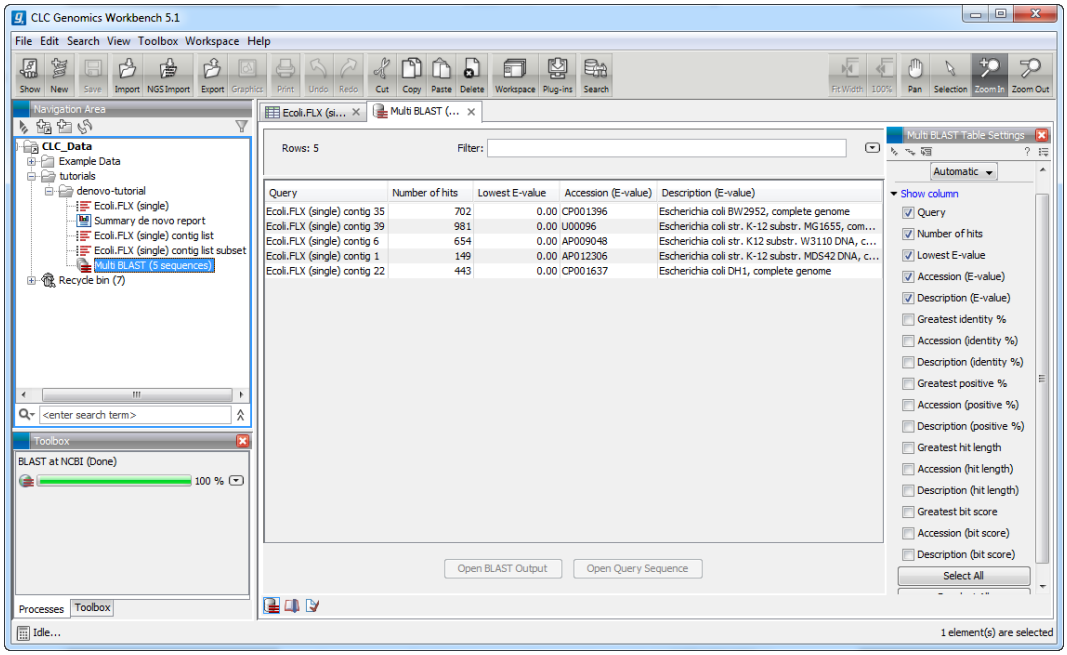


Figure 8: *The overview BLAST result table.*

For each contig, you can see the description of the top hit which is *E. coli* in eaach case, although of different strains.

You can, of course, export this summary table if you wish. Popular export formats for this sort of data would be excel, or comman separated values (csv).

3. Double click on one of the rows of the summary blast table. This opens an individual BLAST result so you can investigate it in more detail.

   By default, a visual depiction of the blast results will be shown. You can also click on the **table view icon** ( ) at the bottom of the navigation area to see a table of reults. With the table, one can filter and sort on different values. Like any table, you can export the data, with popular foramts being excel, or comman separated values (csv).

   If you are used to standard, text blast format, you may wish to click on the **Text Contents** ( ) icon at the bottom of the navigation area.

Choosing mapping as the output type from an assembly

This is an optional part of the tutorial. Here, we re-run the de novo assembly, but this time choosing to map the reads back to the contigs, and to update the reads based on the mappings.

1. Start up a de novo assembly analysis by going to:

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

2. Select the `Ecoli.FLX` ( ) sequence list as input by adding it to the panel to the right.

3. Click on the button labelled **Next**.

4. You now have the opportunity to set some parameters for your assembly. Here, we will accept the defaults, as shown in figure 3.

   Remember, you can directly access the manual to find out more about these parameters by clicking on the **Help** (?) button in the far left hand, bottom edge of the wizard window.

5. Click on the button labelled **Next**.

6. On this next page of the wizard, you get to choose the type of output you will generate. Here, choose **Map reads back to contigs (slow)**. Leave all the mapping options as the defaults. Please refer to the manual, for example, by by clicking on the **Help** (?) button in the far left hand, bottom edge of the wizard window, to find out more about these options.

7. Put a check in the box beside **Update contigs**.

8. Click on the button labelled **Next**.

9. Choose to create a report and to Save the results.

10. Click on the button labelled **Next**.

11. Choose a place to save your results to, adn then click on the button labelled **Finish**.

The outputs of the assembly are a summary table of the mappings and a report. If you have not changed the name of the sample data, the mapping output will be called **Ecoli.FLX (single) de novo assembly**. The report will be called **Ecoli.FLX (single) summary report**. There are a couple of extra graphs in this report compared to the one you generated earlier, where you requested only the contigs themselves as output.

1. Open the mapping output by double clicking on it in the Navigation area. You should see a view like that shown in figure 9.

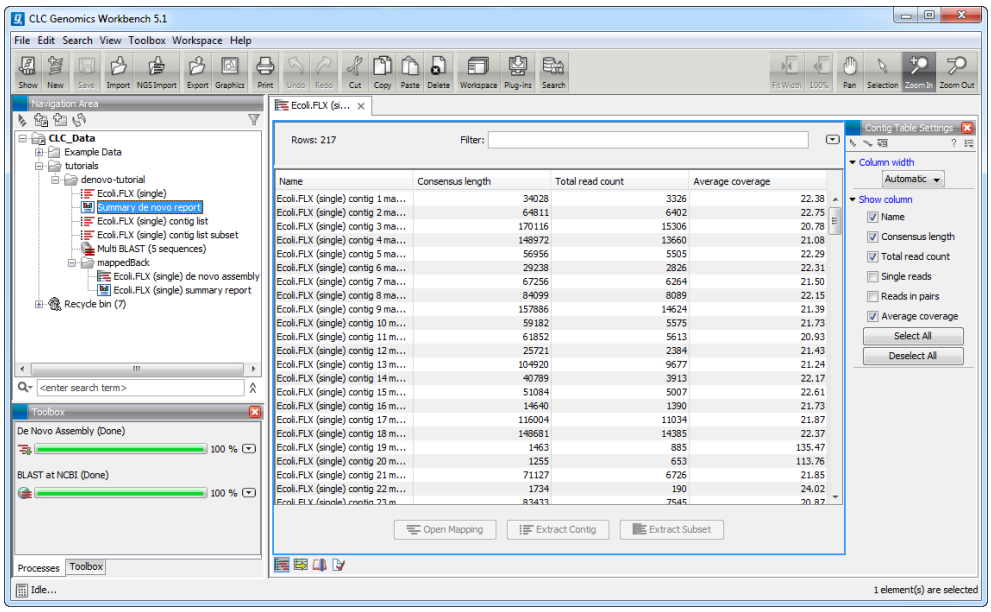


Figure 9: *A summary mapping table.*

Each row in the table represents a contig. There should be around 200 contigs (the number can vary a little from version to version). Double-clicking a row will open the contig with the mapped reads.

2. Sort the table on Average coverage by clicking on the column heading. Sort them so that the contigs with the highest average coverage appear at the top of the table.
3. Double click on the top contig listed (i.e. the one with the highest coverage) to open it in the viewing area.
4. Maximize the view of this assembly by double clicking on the tab name in the Viewing area.
5. Feel free to investigate the effects of the different viewing options on how the mapping looks, and can be searched. Please also try zooming in at different levels.
6. When you have seen enough, minimize the view again by double clicking on the tab name again.
7. Close the assembly tab - for example, by clicking on the small X at the right side of the tab.
8. Go back to viewing the mapping summary table.
9. Sort the mapping table on Consensus length.
10. Highlight the rows containing information about the five longest contigs.
11. Click on the button marked
12. Extract Contig to get a sequence list with just these contigs. Choose to **Save** these.

Now you have a sequence list for your 5 longest contigs. From here, you could run a BLAST job, or any other task that requires a sequence list as input.