



Tutorial

Aligning contigs manually using the Genome Finishing Module

March 17, 2023

Sample to Insight

Aligning contigs manually using the Genome Finishing Module

The CLC Genome Finishing Module is a collection of tools that have been developed to help finish microbial genomes.

This tutorial is an introduction to joining, splitting, and extending contigs manually using the **Align Contigs** tool, the **Join Two Contigs** option, the **Analyze Contigs** tool, and the **Extend Contigs** tool of the Genome Finishing Module.

The features demonstrated in this tutorial include:

- Aligning contigs to a reference sequence. This can be used to visualize the orientation and order of contigs by alignment to a reference sequence.
- Manually joining two contigs. This reduces the number of contigs by joining adjacent contigs.
- Splitting a contig into two. This can be used to separate sequences that mistakenly have been assembled into one contig.
- Extending a contig based on the original sequences that form the contig.

Prerequisites For this tutorial, you must be working with CLC Genomics Workbench 12.0 or higher, and have installed the Genome Finishing Module. How to install plugins is described here: <http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Install.html>.

Additionally, this tutorial was made with Genome Finishing Module 23.0. If you are working with an earlier version, some elements in this tutorial (such as contig naming) may differ slightly from your local installation.

NB! Important tips for this tutorial

- For newer CLC Genomics Workbench versions, the classic coverage view is no longer enabled per default in the Read Mapping view. To enable this, enable **Alignment info | Coverage (classic) | Graph** from the Read Mapping view **Side Panel**.
- If at any time your view doesn't line up with the figures in this tutorial, try zooming out or scrolling sideways (particularly for the Read Mapping views).
- If you are unfamiliar with Read Mapping views, using the **Help** button in the lower right corner of the view or pressing F1 while viewing a Read Mapping, will bring up the manual.

Background of the dataset and analysis The dataset available for download contains:

- *E. coli_DH10B*. Reference sequence - Escherichia coli K12 substr DH10B.
- *paired_illumina_miseq_1*. Part one of Illumina MiSeq paired-end whole genome data.
- *paired_illumina_miseq_2*. Part two of Illumina MiSeq paired-end whole genome data.

- *paired_illumina_miseq_tutorial_assembly*. Assembly of the paired read data.

In this tutorial we will improve the assembly of paired data (*paired_illumina_miseq_1*) by using the join, split, and extend contigs functionalities in Genome Finishing Module. Lastly, we will look at how to use the read data by mapping to the modified contigs.

The E. coli dataset used in this tutorial is a subset of a publicly available dataset. The E. coli strain DH10B reads are from https://emea.illumina.com/content/dam/illumina-marketing/documents/products/appnotes/appnote_miseq_ecoli.pdf and the reference sequence is from NCBI (id=NC_010473.1): http://www.ncbi.nlm.nih.gov/nuccore/NC_010473.1?report=genbank.

Importing the data

To get started, we need to download and import the sample data.

1. Download the sample data from our website and unzip it in your desired location:
http://resources.qiagenbioinformatics.com/testdata/finishing_module_tutorial.zip
2. Start the workbench.
3. To import the reference sequence go to:
File | Import (📁) | Standard Import (📁)
4. Choose the files called "E_coli_DH10B.fa" and "paired_illumina_miseq_tutorial_assembly.clc". Leave the Import type set to **Automatic import**.
5. Specify where to save the downloaded data. It is a good idea to make a folder for this, for example called "Genome Finishing tutorial". Click **Finish**.
6. Next we are going to import the paired read data. This requires simultaneous import of the two files called "paired_illumina_miseq_1.fastq" and "paired_illumina_miseq_2.fastq". Go to:
File | Import (📁) | Illumina (📁)
7. Select the two files and tick **Paired reads** under **General options**. Set **Minimum distance** at 150 and **Maximum distance** at 450. Click **Next**.
8. Specify where to save the downloaded data and click **Finish**.


De Novo Assembly


Normally, we would start our work by creating a de novo assembly. However, due to randomness introduced when assembling, we will not assemble reads and instead use the pre-constructed assembly imported in the previous step. Note: When doing your own analysis, de novo assemblies can always be created by launching **De Novo Assembly** from the **Toolbox**:

De Novo Sequencing (📁) | De Novo Assembly (📁)




Annotating potential problems

The **Analyze Contigs** tool is useful for identifying misassembled reads and can be run before aligning the contigs. The tool will only be mentioned briefly in this tutorial, but further information about the Analyze Contigs tool can be accessed at: https://resources.qiagenbioinformatics.com/manuals/clcgenomefinishing/current/index.php?manual=Analyze_Contigs.html.


1. To use the Analyze Contigs tool, double click on the tool in the Toolbox () . This opens up a wizard. Select the assembled data (paired_illumina_miseq_tutorial_assembly).
2. Proceed using the default settings and click **Next** until you reach the final window where you can select whether to **Open** or **Save** the result.
3. Choose to **Save** the results and click on the button labeled **Finish**. Save the results to the tutorial data location; for example in a subfolder titled "Analyzed Contigs". Note, that by default the tool also generates a report and an annotation table (also called *Contig analysis table*) containing all the new annotations added to the input. As we will not be using these for this tutorial, feel free to untick the respective boxes in the wizard.


Additional annotations have now been added to the read mapping list and can be assessed by clicking on **Show Annotation Table** (). We will come back to this later.

Running the Align Contigs tool

1. Next, open the Align Contigs tool from the Toolbox:
Genome Finishing Module () | **Align Contigs** ()
2. Select the assembled data (paired_illumina_miseq_tutorial_assembly) and click on the Next button.
3. This takes you to the **Select contig mapping parameters** step shown in figure 1. Select the reference sequence (E_coli_DH10B) by clicking on the folder () . Keep the default settings for the BLAST options and Match options. Click **Next**.
4. Choose to **Save** the result to a new subfolder, for example named "Aligned contigs".

The Align Contigs output

The output file of the **Align Contigs** tool is an aligned **Contig Table** () , which contains two tables:

1. The **Contig table** summarizes information about the contigs. This table opens per default when clicking on the **Contig Table** file, or can be accessed by clicking on the **Show Contig table** icon () in the bottom of the view (see red box on figure 2). Selecting a contig and clicking on **Show Contigs** will open a **Read Mapping** view of the sequences composing that contig.

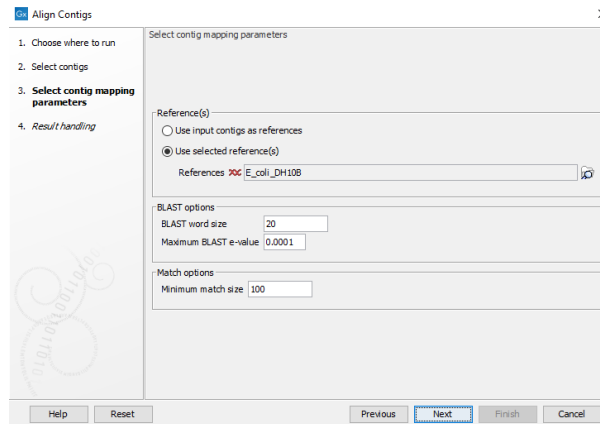


Figure 1: Select the contig mapping parameters.

2. The **Contig match table** lists the matches between the contigs and the reference sequences found by BLAST. This table is opened by clicking on the **Show Contig match table** icon (🔍) in the bottom of the view (figure 3 red box). Because this is a BLAST result, each contig may appear multiple time with different hits. A **Show Contig Matches** button in the Contig match table allows the visualization of the contigs scaffold in a **Read mapping** view. Change the *Compactness* mode in the Side Panel from *Packed* to *Low* to make the contig names visible next to the contig sequence. Under the reference sequence, a coverage track indicates with peaks the potential overlaps between contigs, and finally the different contigs aligning to the reference sequence. If the coverage graph is not showing, enable it by checking **Alignment info | Coverage (classic) | Graph** in the Side Panel.

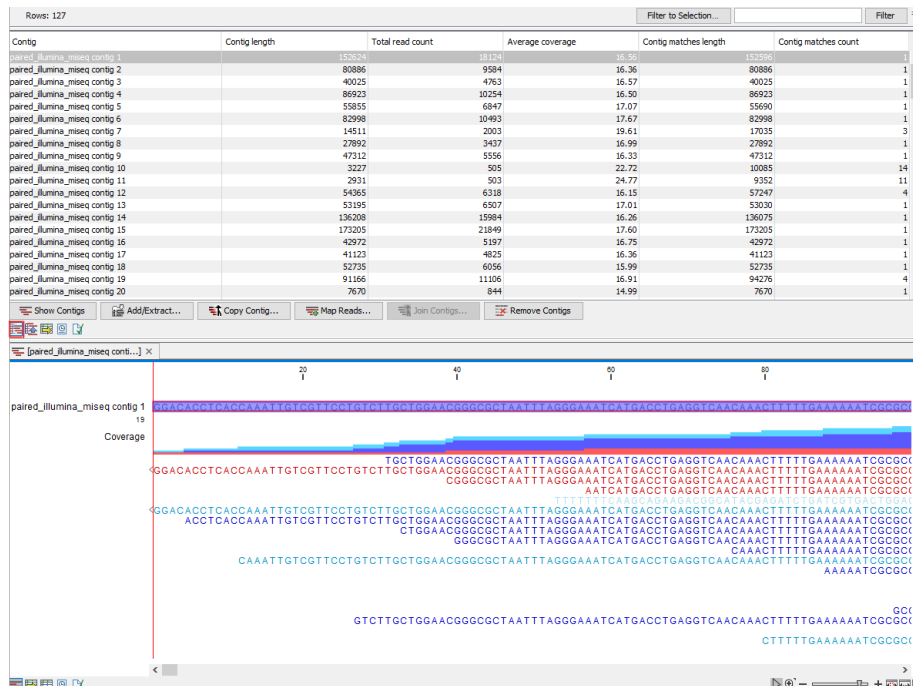


Figure 2: Top: The Contig table. Bottom: Read Mapping view of selected contig.

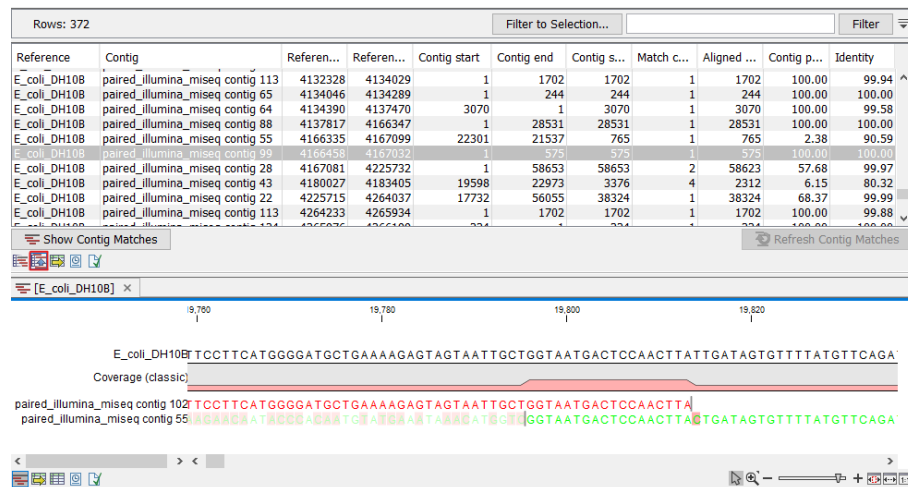


Figure 3: Top: The Contig match table. Bottom: Read Mapping view with Coverage enabled and Compactness set to Low.

These two tables are linked, which means that selecting an item in one table, will automatically select related items in the other. The tables are also linked with their respective Read Mapping views, meaning that selecting a contig in the table and clicking on the **Zoom to selection** (🔍) button in the lower left side of the view area will fill the view with the current selection. You can also zoom in and out on the regions of interest by holding down the Ctrl key (Cmd (⌘) on Mac) while scrolling with the mouse wheel. It will zoom in the region where you hold the mouse. Additionally, you can use the zoom functions at the bottom of the view.

Joining contigs

We will start with the simple operation of joining contigs using the **Contig match table**.

1. Start in the **Contig table** (📄) and select "paired_illumina_miseq contig 45" in the *Contig* column.
2. Switch to the **Contig match table** by clicking on (🔍) in the bottom left corner.
3. Contig 45 has been automatically selected in the Contig match table. Here, you will see that this contig has only one hit to the reference sequence and that the hit has 100% identity with the reference sequence and 26133 aligned nucleotides. Click on the **Show Contig Matches** button to open the **Contig match Read Mapping view**.
4. To see the full hit use **Zoom to selection** (🔍). The full alignment is now visible.
5. Remember to shift the *Compactness* mode from *Packed* to *Low* to make the contig names visible next to the contig sequence, and zoom out until contig 84 is visible on the 5' end and contig 58 is visible on the 3' end (figure 4).

In the coverage graph, we can see a number of peaks indicating overlap between these contigs. We will also notice that contig 92 appears twice in this region. By filtering for "92" on the Contig column in the Contig match table we can see that contig 92 has 15 hits with 1195 aligned

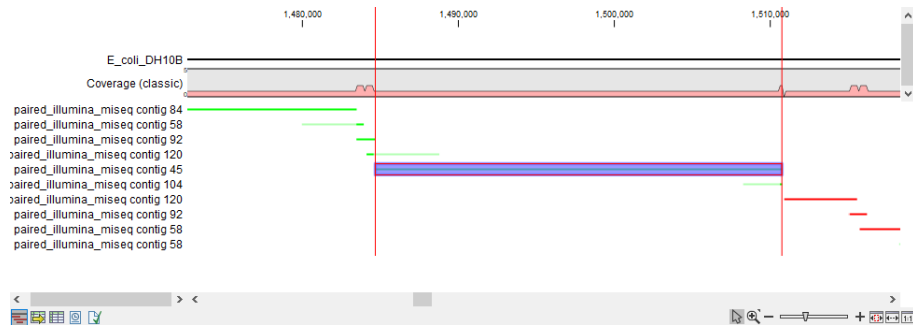


Figure 4: Show contig matches for contig 45 and zoom out to visualize the surrounding hits.

nucleotides and an identity close to 100%. This tells us this contig actually covers several regions but was only assembled into one contig due to these regions being repeats. In the next section, we will focus on connecting the region from contig 84 at the 5' end to contig 58 at the 3' end.

Joining contigs when one of the contig is a repetitive sequence

We will start out by joining contig 84 to contig 92. This can be done directly in the Contig match Read Mapping view by selecting the region you wish to join on the reference sequence.

1. Click and drag on the reference sequence to select a region containing the overlap between the contig 84 and contig 92. Note: be careful not to select too much. You only need to select nucleotides within the overlap.
2. Right click on the selected region of the reference sequence and click on **Join Two Contigs** (figure 5).

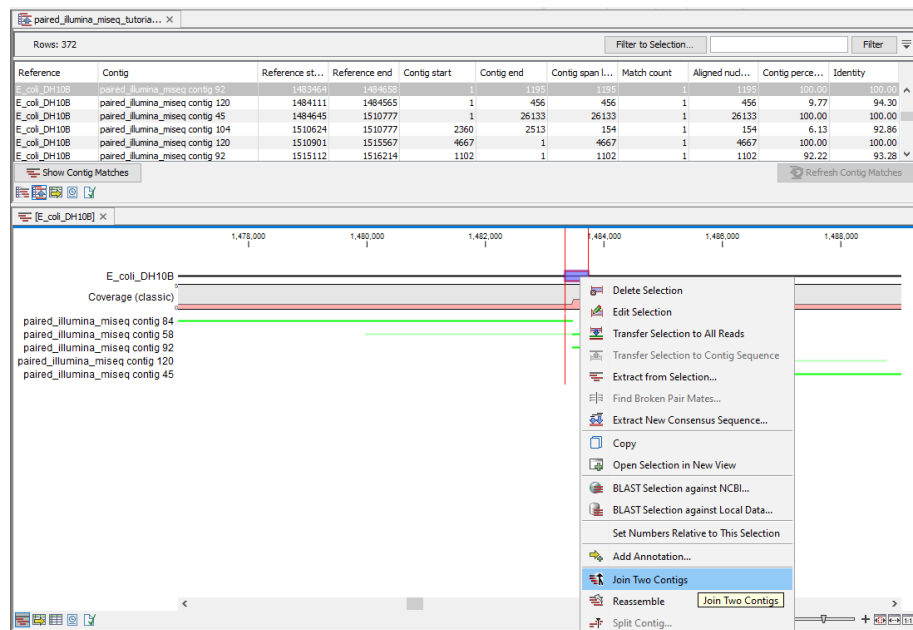


Figure 5: Join two contigs by right clicking on the reference sequence.

- This opens up the wizard shown in figure 6. When the selected region only contains two contigs, the contigs to join are selected automatically in the wizard. Otherwise it is necessary to manually select the two contigs of interest from a drop down menu in the wizard. Select contig 84 and contig 92.

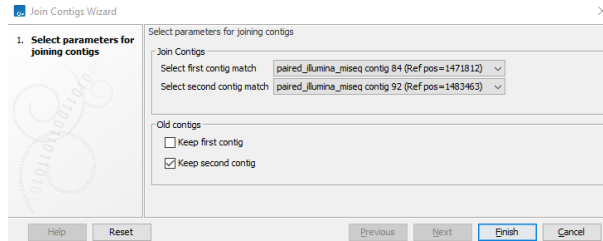


Figure 6: Join Contigs Wizard. Select contigs to join and keep contig 92.

- We wish to keep a copy of contig 92 because we, as stated earlier, need to use this contig twice in our region. Therefore check the box to keep the contig corresponding to contig 92 (in figure 6 contig 92 is selected as the second contig). Click **Finish**.
- Joined contig 1** can now be seen in figure 7. We chose to keep contig 92 but contig 84 is gone from the Contig match table as we did not keep this contig. Tip: For a fast way to compare your joined contigs to the previous state, use Ctrl + Z to undo (Cmd (⌘) + Z on Mac) and Ctrl + Y (Cmd (⌘) + Y on Mac) to redo. Because this contig was created from a repeat (contig 92), Joined contig 1 appears several times in the Contig match table (📄). However, only one match represents the full length of the contig. The rest can be ignored (figure 8).

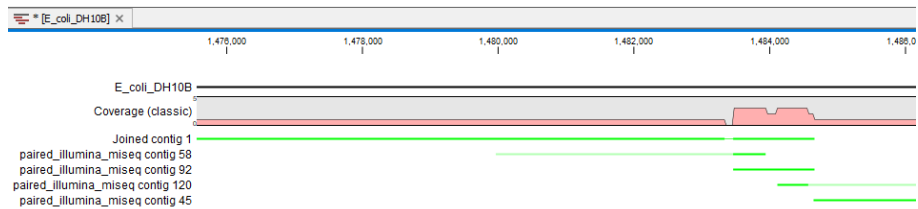


Figure 7: Joined contig 1 along with other overlapping contigs.

Reference	Contig	Reference start	Reference end	Contig start	Contig end	Contig span length	Match count	Aligned nucleotides	Contig percentage	Identity
E_coli_DH10B	Joined contig 1	247283	248477	12722	11526	1197	1	1197	9.41	99.83
E_coli_DH10B	Joined contig 1	513146	514340	12722	11526	1197	1	1197	9.41	99.83
E_coli_DH10B	Joined contig 1	626406	627600	12722	11526	1197	1	1197	9.41	99.83
E_coli_DH10B	Joined contig 1	739666	740860	12722	11526	1197	1	1197	9.41	99.83
E_coli_DH10B	Joined contig 1	1471813	1484638	1	12722	12722	2	12722	100.00	99.98
E_coli_DH10B	Joined contig 1	1515112	1516215	12629	11526	1105	1	1105	8.69	93.12
E_coli_DH10B	Joined contig 1	1521766	1522961	11525	12722	1198	1	1198	9.42	99.83
E_coli_DH10B	Joined contig 1	2067104	2068298	12722	11526	1197	1	1197	9.41	99.75
E_coli_DH10B	Joined contig 1	2155191	2156386	12722	11526	1198	1	1198	9.42	99.42
E_coli_DH10B	Joined contig 1	2190781	2191975	12722	11526	1197	1	1197	9.41	99.83
E_coli_DH10B	Joined contig 1	2377929	2379123	12722	11526	1197	1	1197	9.41	99.83
E_coli_DH10B	Joined contig 1	3170618	3171813	11525	12722	1198	1	1198	9.42	99.83
E_coli_DH10B	Joined contig 1	3229113	3227197	11526	12722	1197	1	1197	9.41	99.83
E_coli_DH10B	Joined contig 1	3461023	3462517	12722	11526	1197	1	1197	9.41	99.83
E_coli_DH10B	Joined contig 1	3747804	3748998	12722	11526	1197	1	1197	9.41	99.83

Figure 8: Contig match table shows that Joined contig 1 matches several times in the reference.

- Repeat the joining procedure by joining Joined contig 1 with contig 45. These two contigs have one true hit in the Contig match table and we therefore do not need to make copies.

Leave both 'Keep' boxes unchecked. This action removes Joined contig 1 and adds **Joined contig 2**.

7. Join contig 120 to contig 92 (figure 9). Make sure to use the better matches (in this case the reverse/red matches) for the two contigs. If any of the contigs have disappeared from view, zoom out or scroll sideways to find them. Here, it becomes clear why we chose to create a copy of contig 92. Otherwise, we would be missing this contig to bridge the gap between contig 120 and contig 58.

Normally, we would make another copy of contig 92, but for this tutorial, we are only focusing on a small region. We therefore do not need to keep contig 92.

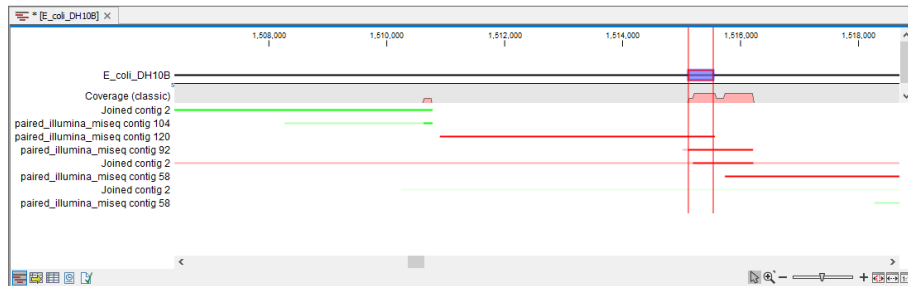


Figure 9: Join contig 120 and contig 92 by right clicking on the selected region of the reference sequence.

8. Finally, join the forward (green) **Joined contig 3** to the reverse (red) contig 58.

We have now done as much as we can by simply joining overlapping contigs. Locate **Joined contig 4** in the Contig match table (🔍) by filtering. Select the match with highest *Contig percentage* and *Identity* and zoom in to the 5' end.

Observe that there is a gap between Joined contig 4 and Joined contig 2 (figure 10). In the next section, we will cover how to join contigs separated by a gap.

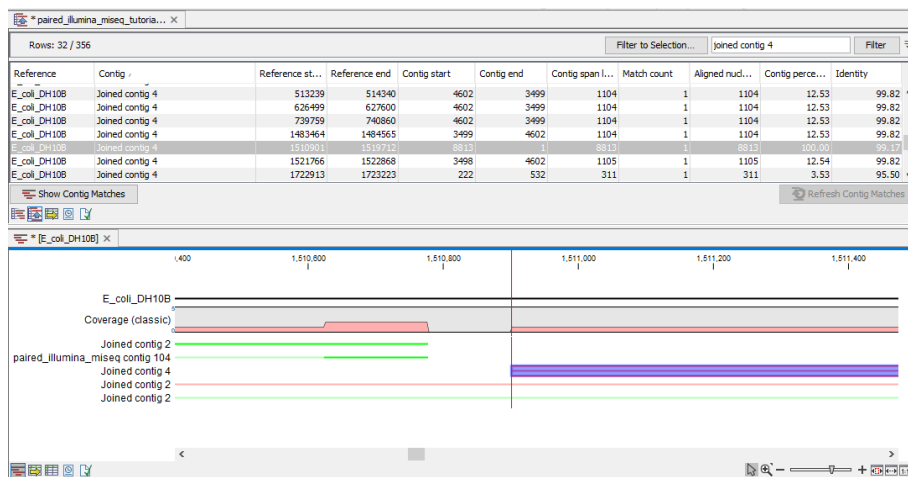


Figure 10: Joined contig 2 and Joined contig 4 are separated by a gap.

Joining contigs separated by a gap

To join two sequences separated by a gap, it is not possible to use the joining option in the right click menu. We will now go through the procedure step by step.

1. Before joining the contigs you need to take note of the gap size and the orientation of the contigs. An easy way of measuring the gap size is to select the sequence in the gap region. When mousing over the sequence, the size of the highlighted sequence is shown in the bottom right corner "size 123" (figure 11). Check also the suggested orientation of the contigs relative to each other. Here Joined contig 2 should be placed before Joined contig 4.

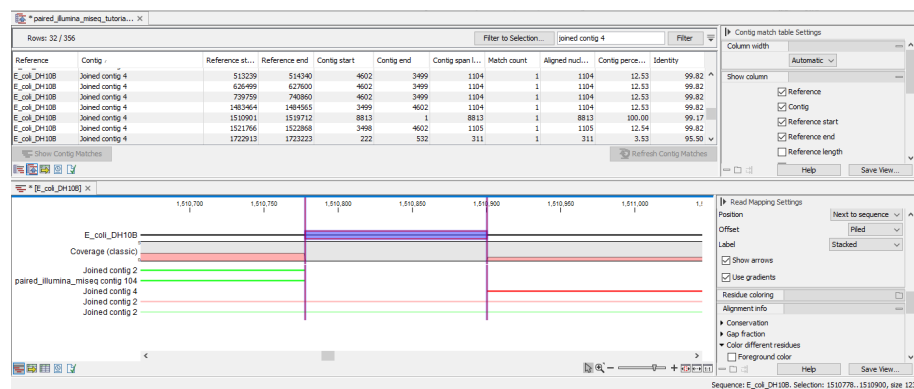


Figure 11: Measure the size of the gap and see the result in the lower right corner.

2. Go back to the **Contig table** (📄), find and then select Joined contig 2 and Joined contig 4 (figure 12). Click the **Join contigs. . .** button at the bottom of the table.

Contig	Contig length	Total read count	Average coverage	Contig matches length	Contig matches count
Joined contig 2	38286	5455	19.92	55525	16
Joined contig 4	8813	2260	35.07	24583	16

Figure 12: Find and select Joined contig 2 and Joined contig 4 and click Join Contigs.

3. In the Join Contigs wizard (figure 13) check **Manual gap** and specify the gap size. Check the option **Place contig "Joined contig 2" before "Joined contig 4"** and click **Finish**.

In figure 14 you can see the final result of the joining procedures - one long contig (**Joined contig 5**) that replaces the initial contigs 84, 92, 45, 120, and 58. We now have one large contig spanning the same region as the five original contigs. The procedure can be continued with more overlapping contigs but this is not covered in this tutorial. Note: make sure to choose the Joined contig 5 with the highest Contig percentage in the Contig match table (📄). Because of all the repetitive elements we integrated into Joined contig 5 there are multiple matches from Joined contig 5 which can be ignored.

Now that we are done with the first steps, it is a good time to save our progress. You can save the contig match table by simply clicking Ctrl + S (Cmd (⌘) + S on Mac).

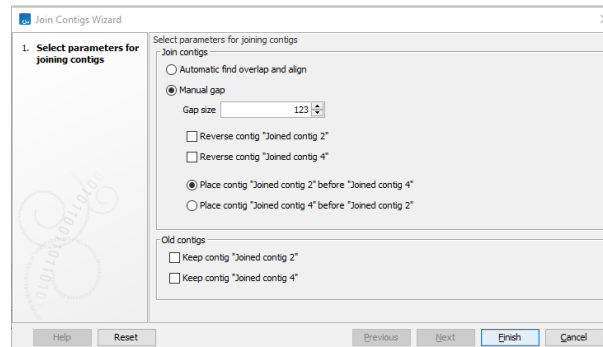


Figure 13: Join Contigs wizard window.

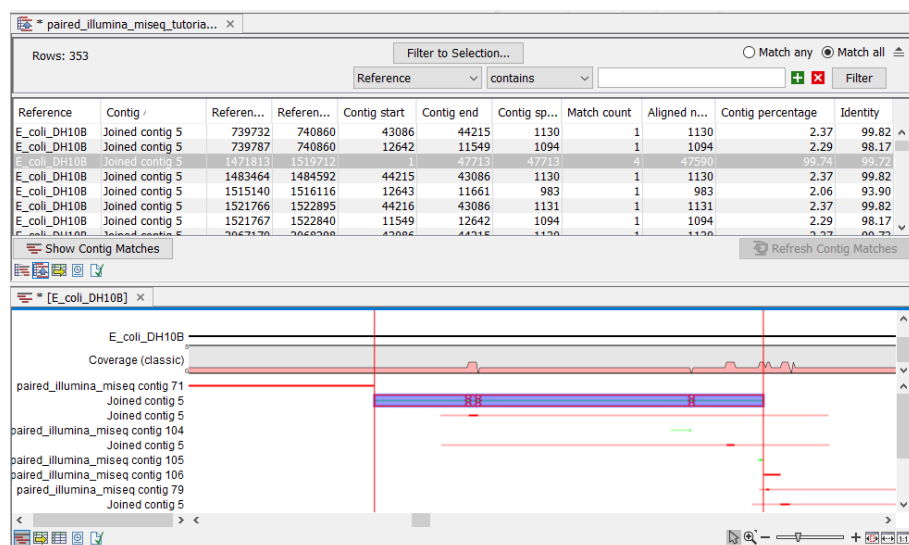


Figure 14: Joined contig 5 - the result of joining contigs 84, 92, 45 and 120, and 58

Splitting contigs

We will now look at how to split contigs. For this, we return to the annotations we added to the contigs in the first step of the tutorial. Note: If you missed this step, there is a workaround to add annotations to an aligned **Contig Table** (📄) in the bottom of this tutorial (section [Workaround for adding annotations to an aligned Contig Table](#)). Otherwise, continue on from here.

Open the annotation table by clicking **Show Annotation Table** (📄) in the bottom left corner of the aligned **Contig table** (📄), then deselect all annotation types except **Unaligned ends** in the Side Panel (figure 15). Of the 5 contigs that contain unaligned ends, we will focus on contig 68.

1. Clicking on contig 68 in the Annotation Table and then switching to the **Contig table** (📄) in the lower left corner, will keep the contig highlighted in the view. Open the Read Mapping by either double-clicking contig 68 or clicking **Show Contigs**. Under **Annotation types** in the Read Mapping Side panel tick **Unaligned ends**. We can easily locate the unaligned end by holding down Ctrl (Cmd) (⌘) on Mac and clicking on Show Annotation Table (📄) in the Read Mapping to open a linked table in split view. Selecting the unaligned end in the Annotation Table will automatically select this region in the Read Mapping (remember to tick **Unaligned ends** in the Annotation Table as well). Alternatively, zoom out to find the

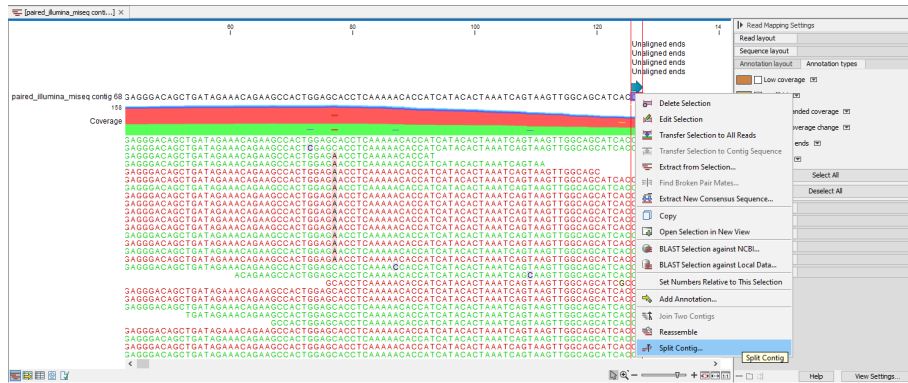


Figure 17: Splitting a contig.

Contig	Contig length	Total read count	Average coverage	Contig matches length	Contig matches count
paired_illumina_miseq contig 68 - Split a	212	147	75.00	1468	11
paired_illumina_miseq contig 68 - Split b	25059	2871	15.70	24995	1

Figure 18: Contig split a and b have different Contig match counts.

- Open the Contig match table (🏠) and find the entries for contig 68 Split a by filtering and notice that the *Contig percentage* for 10/11 matches is below 60%. We will investigate this in the Contig match Read Mapping view.
- Select the first match of Split a (position 20438 to 20563) and click **Show Contig Matches** in the bottom left. Remember to set *Compactness* to *Low* and notice that a large region of the contig is unaligned (figure 19, faded ends). There is also a small gap between contig 68 - Split a and contig 99, meaning we cannot join these directly. Split a also appears to overlap contig 55, but by switching *Compactness* to *Packed* in the Side Panel, we see that the repeat in contig 55 contains a lot of mismatches and is therefore unlikely to belong here.
- Open the Contig table Read Mapping view for contig 68 Split a. Under **Annotation types** in the Side panel tick **Split position** to enable the split annotation and observe that a region starting from position 127 (the split position) has very low coverage. This low coverage region is an artifact from the split and overlaps with contig 68 Split b and can be deleted. Select the whole region starting from position 126, right-click and choose to **Delete selection** (figure 20).
- Save the Read Mapping by clicking Ctrl + S (Cmd (⌘) + S on Mac).

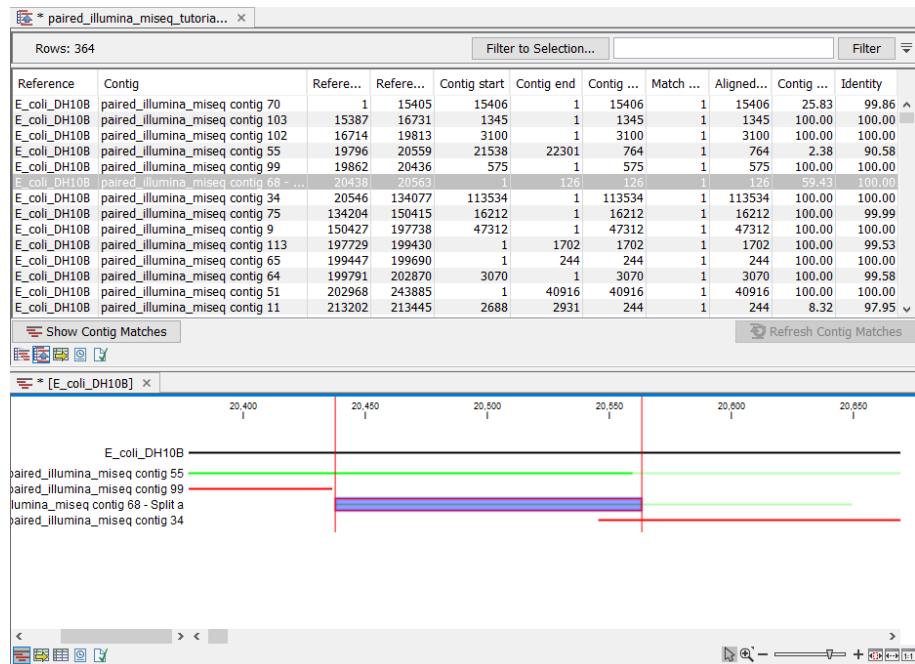


Figure 19: Contig match Read Mapping view showing the first match of contig 68 Split a.

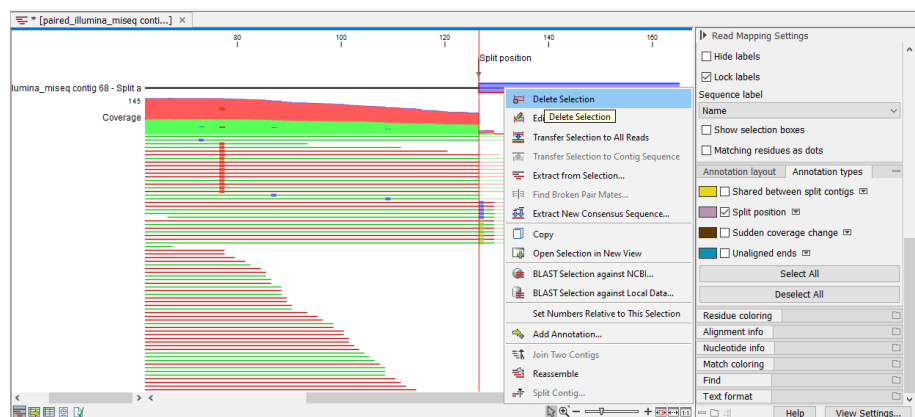


Figure 20: Deleting a low coverage region that is no longer needed.

Extending contigs

Next, we will take care of the missing fragment at the 5' end of the repeat.

1. In the Contig table Read Mapping for contig 68 - Split a scroll to the first position. Observe the arrows on the 5' end of the mapped reads (figure 21). Such reads can be used to extend the contig.
2. Ensure that the Contig table Read Mapping window for contig 68 - Split a is active by clicking anywhere within it and then select:

Genome Finishing Module (📁) | Extend contigs (🔍)

in the Toolbox. This will automatically add the contig as input to the tool. Alternatively, the contig can be exported, saved in the Navigation Area and used as input for the tool.

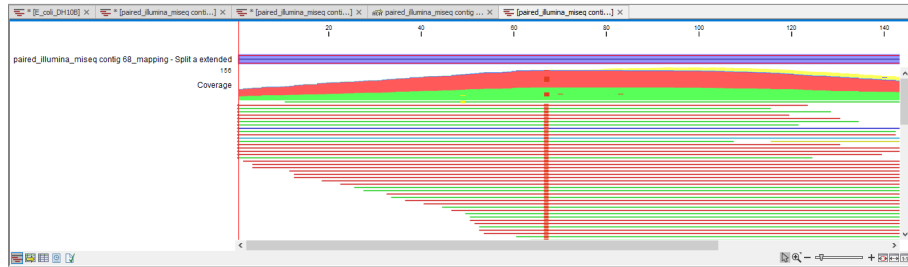


Figure 23: The extended contig after reads have been remapped.

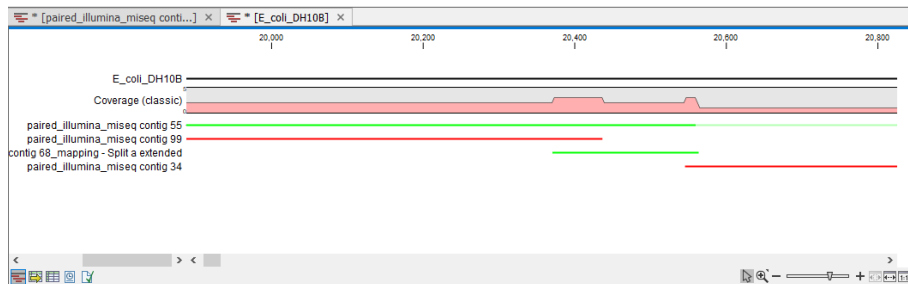


Figure 24: The Contig match table Read Mapping view confirms that the extended contig is spanning the gap between contig 99 and contig 34.

We have now gone through how to join, split, and extend contigs. If you wish, you can use these tools to further improve the tutorial assembly. A good assembly is a starting point for several downstream analysis tools found in other plugins. For example, you can use the **Microbial Genomics Module** to find antibiotic resistance: https://resources.qiagenbioinformatics.com/manuals/clcmgm/current/index.php?manual=Drug_Resistance_Analysis.html.

Another example is using the tools in the **Whole Genome Alignment** plugin to compare to similar genomes: <https://resources.qiagenbioinformatics.com/manuals/wholegenomealignment/current/index.php?manual=Introduction.html>.

Workaround for adding annotations to an aligned Contig Table

If at any point you want to use the Analyze Contigs tool, but have already aligned your contigs, there is a workaround to do so described here:

1. Extract all the contigs by saving your contigs to a Read Mapping List. To do so, click on Show contig table (☰), make sure the Filter is cleared, and click Ctrl + A (Cmd (⌘) + A on Mac) to highlight all contigs. Then use the **Add/extract. . .** button to extract and save the updated contigs.
2. Use the extracted Read Mapping List (☰) from the Navigation Area as input for **Analyze Contigs**, as described above in section **Annotating potential problems**.
3. In the aligned Contig Table (☑), select all contigs (Ctrl + A or Cmd (⌘) + A on Mac), and use the **Remove Contigs** button in the bottom of the view.
4. Lastly, use the **Add/extract. . .** button, choose Add Contigs, and select the just extracted and annotated Read Mapping List (☰).

Tutorial

When you are done and are sure everything is as you expect, save the Contig Table by clicking Ctrl + S (Cmd (⌘) + S on Mac). The annotations are now available by clicking the **Show Annotation Table** (📄➡) button in the bottom left of the view.