Tutorial

Aligning contigs manually using the Genome Finishing Module

November 21, 2017

Sample to Insight

QIAGEN Aarhus · Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
Telephone: +45 70 22 32 44 · www.qiagenbioinformatics.com ·
AdvancedGenomicsSupport@qiagen.com

Tutorial

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. One of the most important tools in the CLC Genome Finishing Module toolbox is the **Join Contigs** tool. The Join Contigs tool align contigs - using or not a reference sequence - and automatically join them based on shared overlap regions. Using the tool on the data provided in this tutorial will join all 127 initial contigs in one single contig and 8 smaller contigs that could not be joined within a short period of time. But you may want to join contigs manually if an overlap was not considered by the Join Contigs tool because of a low coverage in the overlap region, or if the overlap was too small or even nonexistent but the contigs are clearly close to each other when align to a reference sequence. This tutorial is an introduction to joining, splitting and extending contigs manually using the **Align Contigs** tool and the **Join Two Contigs** option, the **Analyze Contigs** tool and the **Extend Contigs** tool of the CLC 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.
- Joining manually 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 a contig.
- Extending a contig based on the original sequences that form the contig.

Prerequisites To run this tutorial, you must be working with the *CLC Genomics Workbench*, version 6.0 or higher.

Minimum recommended machine specifications can be found at http://www.qiagenbioinformatics.com/system-requirements/.

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 and *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 use the assembled data (paired_illumina_miseq_tutorial assembly) for a demonstration of how to join and split two contigs. The read data will be used to demonstrate the usefulness of remapping reads after contigs have been modified.

The E. coli dataset used in this tutorial is a subset of a publicly available dataset. The E. coli strain DH10B reads are from http://www.illumina.com/systems/miseq/scientific_data.ilmn 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 web site and unzip it on your desktop: http://resources.qiagenbioinformatics.com/testdata/finishing_module_tutorial.zip
2. Start the *CLC Genomics Workbench*.
3. To import the reference sequence and the assembly data 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**.
5. Specify where to save the downloaded data and click **Finish**.
6. Next we are going to import 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 on the button labeled Next
8. Specify where to save the downloaded data and click **Finish**.

Running the Align Contigs tool

1. Once the data have been downloaded, open the Align Contigs tool from the toolbox:
   Toolbox | **Genome Finishing Module** ( ) | **Align Contigs tool** ( )
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, which are **BLAST word size**: 20 and, **Maximum BLAST e-value** 0.0001, and **Minimum match size**: 100. The arrow button in the lower left corner of the window can take you back to the default settings if desired. Click **Next**.
4. Choose to **Save** the result. Click on the button labeled Next and specify where to before clicking **Finish**.

The Align Contigs output

The output file of the **Align contigs** tool has the suffix **contig match table**, and contains two tables:

1. **The Contig table** summarizes information about the contigs. This table opens per default when clicking on the **contig match table** file, or can be accessed by clicking on the **Show contig table** icon ( ) at the bottom of the **View Area** (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.

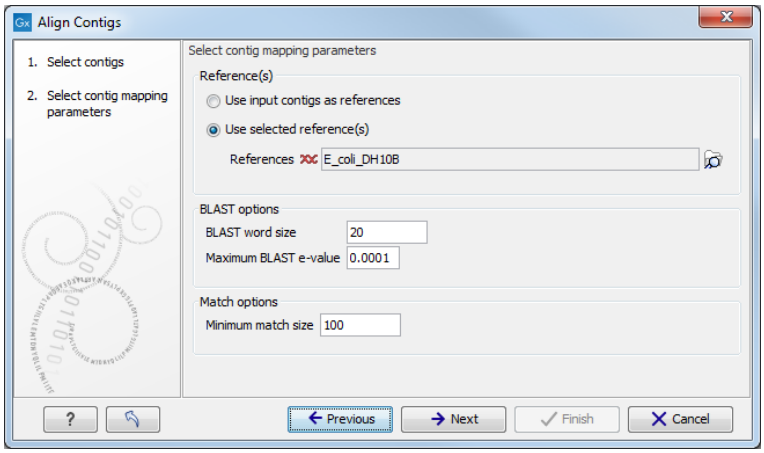


Figure 1: *Select the contig mapping parameters.*

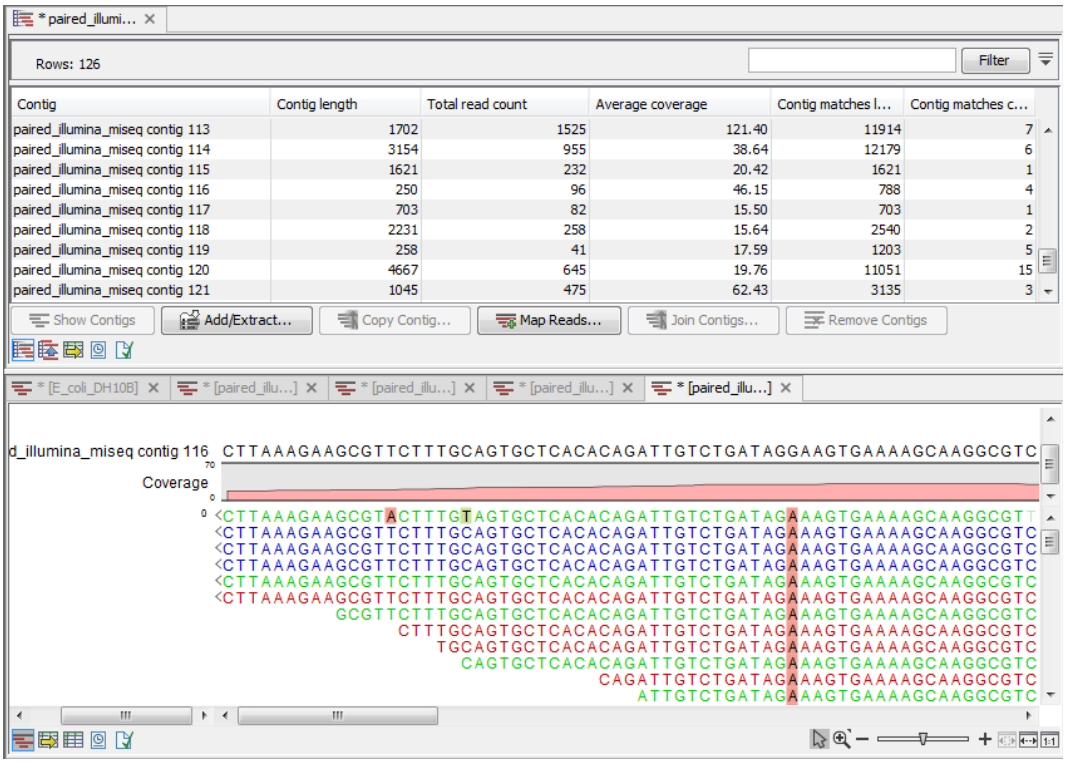


Figure 2: *The Contig table and its read mapping view.*

2. **The Contig match table** lists the matches found by BLAST between the contigs and the reference sequences (figure 3). This table is opened by clicking on the show **Contig match table** ( ) in the bottom left corner of the **View Area**. A **Show contig matches** button in the **Contig match table** allows the visualization of the contigs scaffold in a **Read mapping view**. 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.Shift the **Compactness mode** in the right side panel of the **Read mapping view** from **Packed** to **Low** to make the contig names visible next to the contig sequence.

These two tables are linked, which means that when you select an item in one table, related items will automatically be selected in the other table. The Contig match table ( ) and the

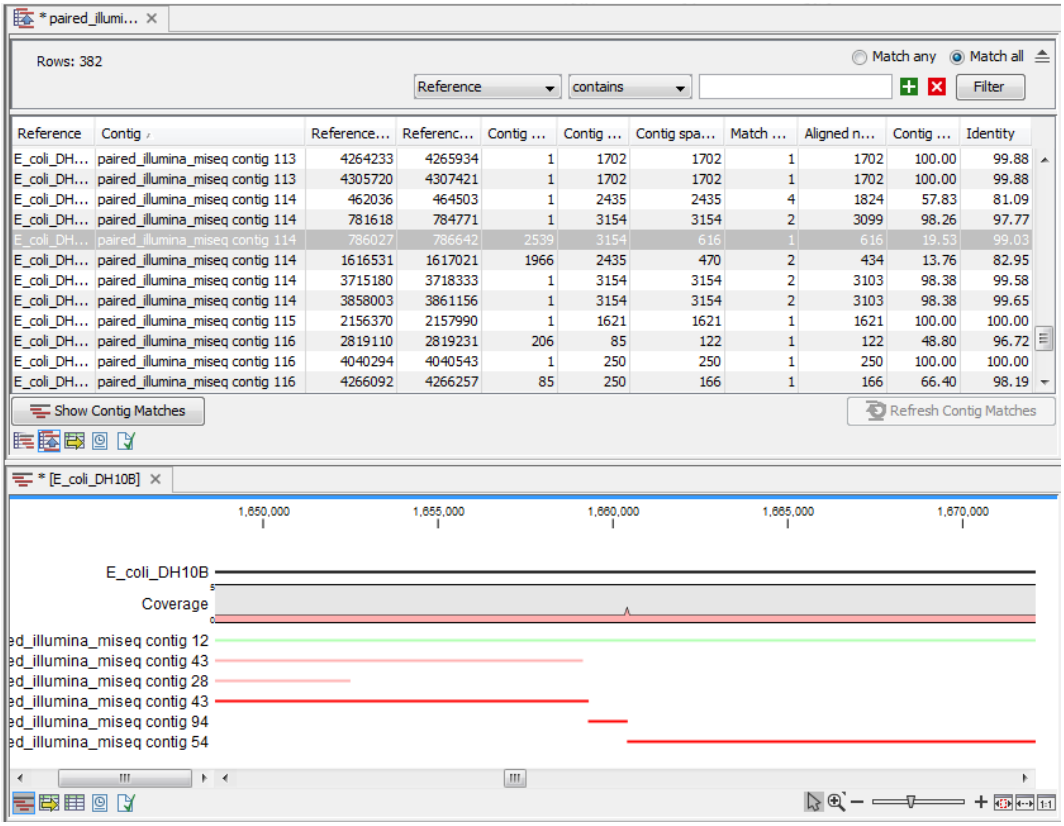


Figure 3: *The Contig match table and the Contig match view.*

Contig match view are also linked, meaning that selecting a contig in the table and clicking on the **Fit Width** ( ) button in the lower left side of the view area will zoom in on the contig in the read mapping view. You can also zoom in directly in the read mapping view window by selecting the region of interest in the reference sequence and clicking on the **Fit Width** ( ) button. You can also zoom in and out on the regions of interest by holding down the Ctrl key while scrolling with the mouse wheel. It will zoom in the region where you hold the mouse. Finally you can use the zoom functions in right side of the toolbar.

1. Start in the **Contig table** ( ) and select contig 67 in the Contig column. In the **Contig matches count** column you will see that this contig has two hits to the reference sequence.
2. Switch to the **Contig match table** by clicking on ( ) in the bottom left corner.
3. Click on the **Contig** column header or use the filter function to find both occurrences of contig 67 (see figure 4).
4. You can see that the first hit has 100% identity with the reference sequence and 12767 aligned nucleotides, while the second hit has only 81% identity and 185 aligned nucleotides. Select both contig 67 hits in the **Contig match table** ( ) and click on the **Show Contig Matches** button to open the **Contig match view**.
5. To see both hits use the **Fit Width** ( ) function in the toolbar. The two hits will now be visible. Try to zoom in and out on the two regions by holding the ctrl key while scrolling up and down your mouse wheel. You can now for example visualize that the second hit is only a small part of contig 67 that aligns to the reference sequence.

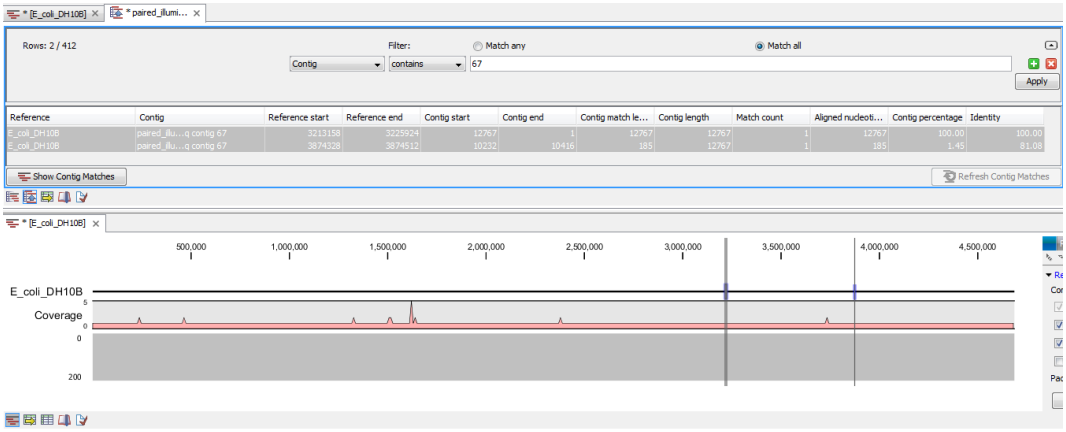


Figure 4: *The Contig match table. Top part of figure: Use the filter function to identify the two contig 67 hits. Lower part of figure: Click "Show contig matches" to visualize the two hits.*

6. Take now a closer look at the first hit from contig 67. Remember to shift the **Compactness mode** from **Packed** to **Low** to make the contig names visible next to the contig sequence, and zoom out a few times to see the other contigs surrounding 67 (figure 5).

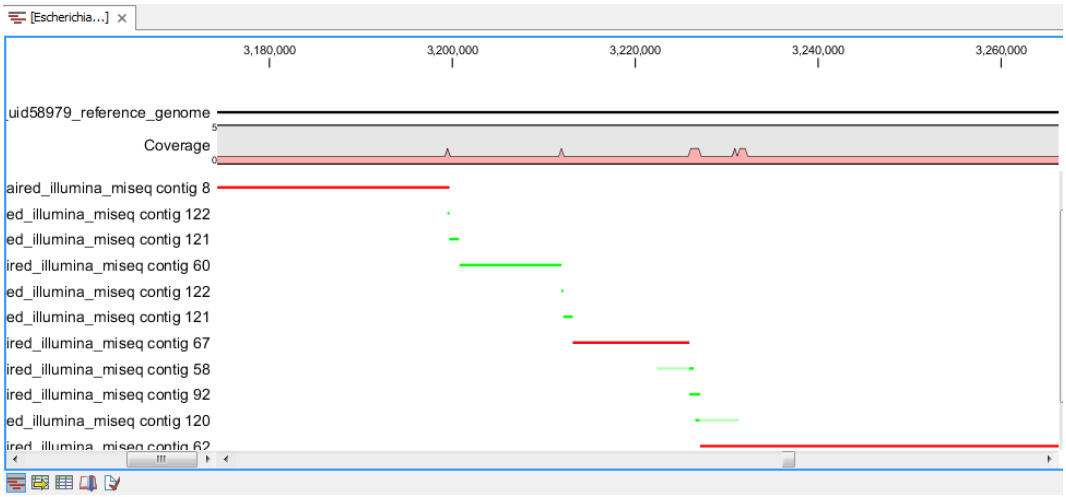


Figure 5: *Multiple sequential contigs with potential overlaps.*

In this tutorial we will focus on joining contig 8 from the 5' side to contig 67 at the 3' end. The coverage track shows small peaks indicating overlap between contigs, for example between contig 67 and contig 121, between contig 121 and contig 122, and between contig 122 and contig 60. These individual regions where there is overlapping contigs can be investigated to look for possible points where contigs may be joined. Additionally, we can see that contig 121 and contig 122 are represented twice in this region.

Joining contigs manually

Joining contigs when one of the contig is a repetitive sequence

We will start out by joining contig 67 and contig 121. This can be done directly in the Contig match view by selecting in the reference sequence the region you wish to join.

1. Click on the reference sequence and select a region containing the overlap between the two contigs (figure 6).

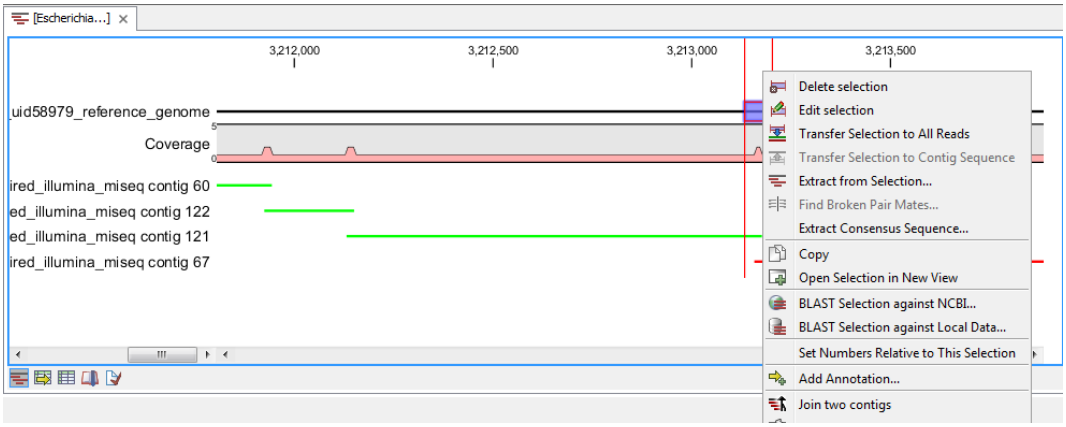


Figure 6: *Join two contigs by right clicking on the reference sequence.*

2. Right click on the selected region of the reference sequence and click on **Join Two Contigs**.

3. This opens up the wizard shown in figure 7. When the selected region only contains two contigs, the contigs to join are selected automatically in the wizard. Otherwise it would be necessary to manually select the two contigs of interest from a drop down menu in the wizard.

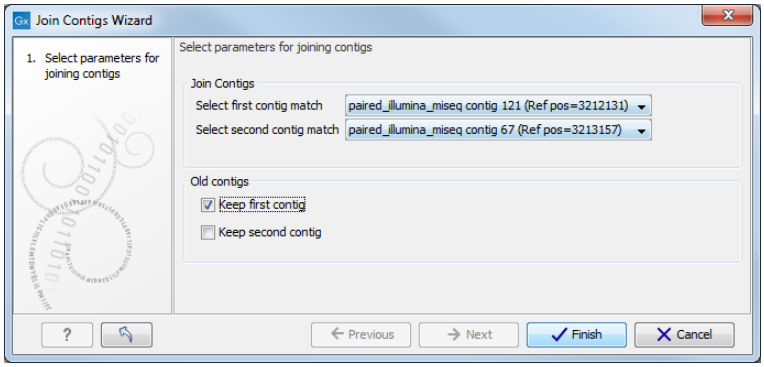


Figure 7: *Join Contigs Wizard. Select contigs to join.*

4. Check **Keep first contig** or **Keep second contig** to keep a copy of contig 121 as it was earlier identified as a a repetitive region that needs to be joined to another contig. Click **Finish**.

5. The **joined contig 1** can now be seen in figure 8 along with the old contig, including contig 121 which will be useful for future joins.

6. Repeat the joining procedure by joining **Joined contig 1** with contig 122 while creating a copy of contig 122. A copy of contig 122 can be created in the wizard by checking the relevant box. Indeed, the **Contig table** ( ) indicates that contig 122 is repeated three times (figure 9) so potentially we need to use contig 122 for one more join.

7. Because we copied two repeats (contig 121 and contig 122), the number of matches for **Joined contig 2** is shown as four in the Contig match table ( ). However, 3 out of the

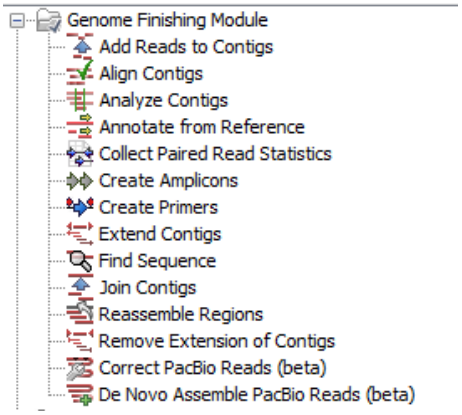


Figure 8: *The joined contig 1 along with other overlapping contigs.*

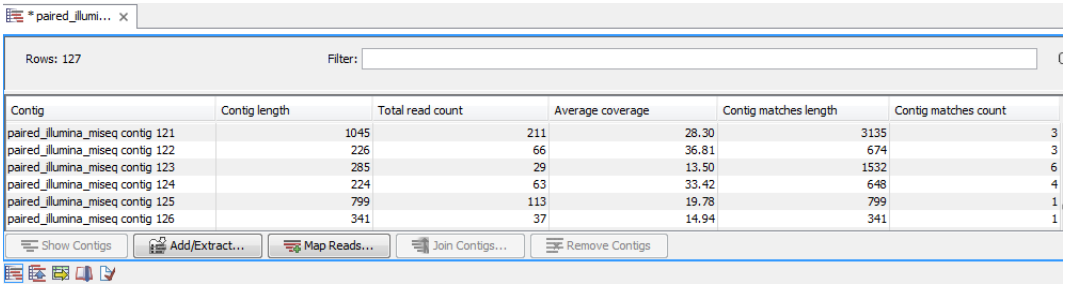


Figure 9: *Contig table where the column "Contig matches count" shows that contig 122 occurs three times in the reference.*

4 matches from Joined contig 2 only represent a small fraction of the contig and can be ignored (figure 10).

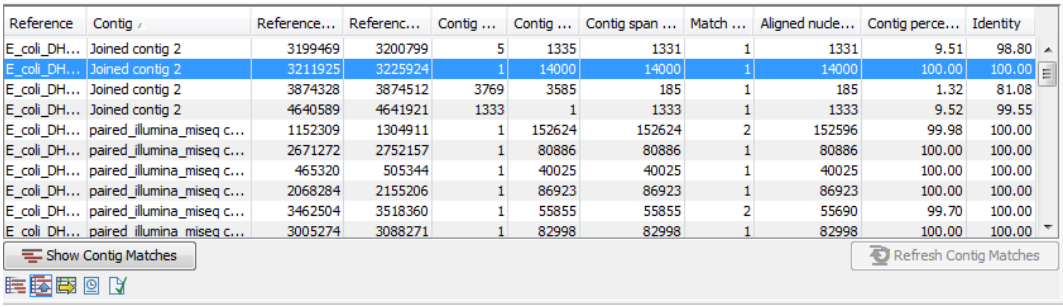


Figure 10: *Contig table where the column "Contig matches count" shows that Joined contig 2 occurs four times in the reference.*

8. Now join contig 8 and contig 121 by right clicking on the reference sequence for the relevant region. Select contig 8 and contig 121 and click **Finish**.

   Note that it now becomes clear why it was relevant to keep contig 121. If contig 121 had been discarded after the first join, we would be missing this sequence to join contig 8 and contig 60. **Joined contig 2** spans the same region but the long unaligned end of the match in this region indicate that it does not belong here (it belongs at the 5' end of contig 60). The original contig 121 is therefore necessary to close the gap between contig 8 and contig 60.

   Please also note that a copy of contig 121 is needed in a different region but for simplicity we will ignore this in this tutorial.

9. Next, join contig 60 and Joined contig 2 (use the match of "Joined contig 2" without the long unaligned end) (figure 11).

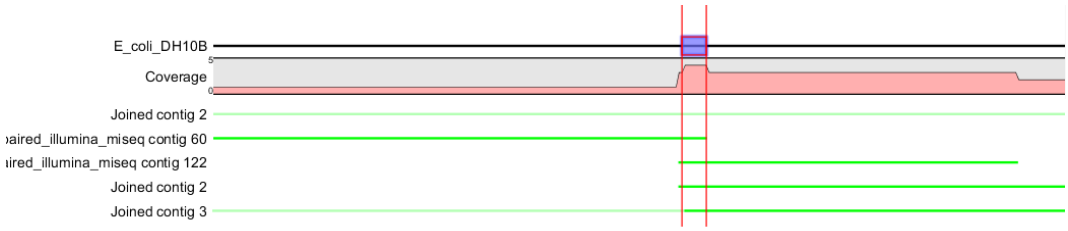


Figure 11: *Join contig 60 and Joined contig 2 by right clicking on the selected region of the reference sequence.*

Joining contigs separated by a gap

The last thing we need to do is to join "Joined contig 3" to "Joined contig 4". However, if you zoom in on the region between the two contigs (we do not take the versions of the joined contigs with the unaligned ends into account), you will discover a gap between contig 3 and contig 4 (figure 12).

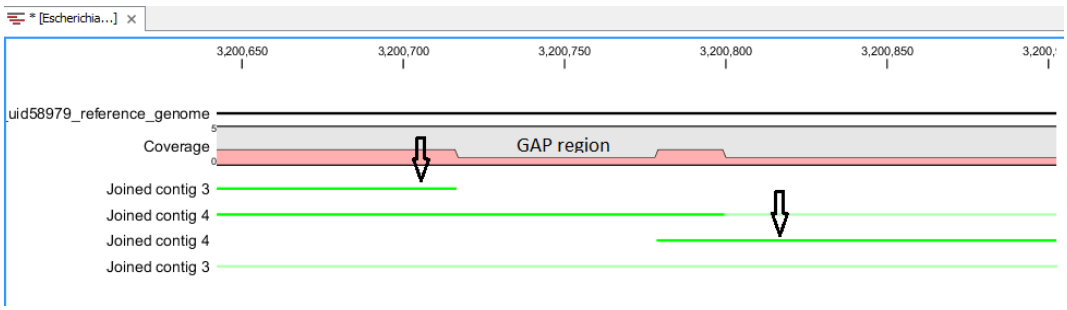


Figure 12: *Joined contig 3 and Joined contig 4 are separated by a gap.*

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 describe the procedure step by step.

1. Before joining the contigs separated by a gap you need to take note of the gap size and the contigs orientation. 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 62" (figure 13). Check also the suggested orientation of the contigs relative to each other. Here Joined contig 3 should be placed before Joined contig 4.

2. Go back to the **Contig table** ( ) and select **Joined contig 3** and **Joined contig 4** from the table.

3. Click the **Join contigs** button at the bottom of the table. In the Join Contigs wizard (figure 14) check **Manual gap** and specify the gap size. Check the option **"Joined contig 3" placed before "Joined contig 4"** and click **Finish**.

In figure 15 you can see the final result of the joining procedures - one long contig (Joined contig 5) that replaces contigs 8, 122, 121, 60, and 67. Because of all the repetitive elements we integrated into "Joined contig 5" there is multiple matches from "Joined contig 5" which can be

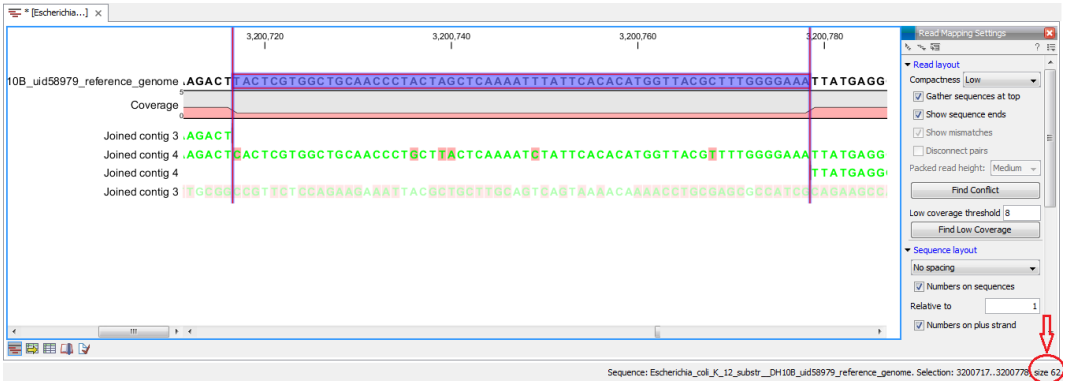


Figure 13: *Measure the size of the gap and see the result in the lower right corner.*

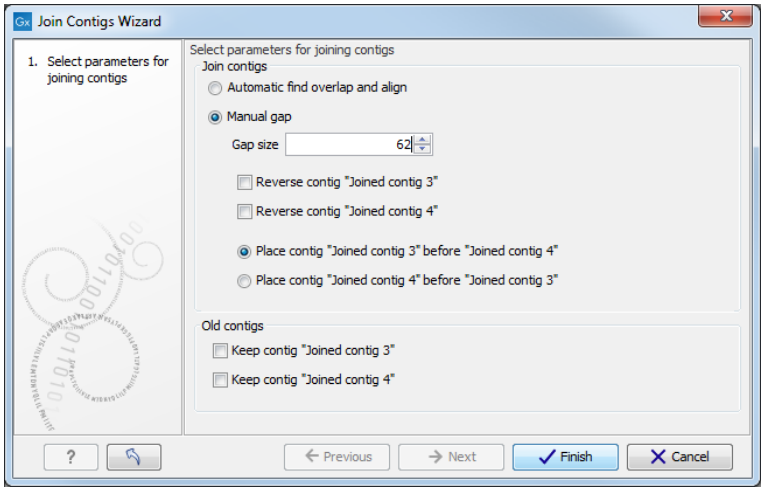


Figure 14: *Join Contigs wizard window.*

ignored. The important thing to note is that we now have one large contig spanning the same region as the five original contigs. The procedure can be continued with more overlapping contigs.

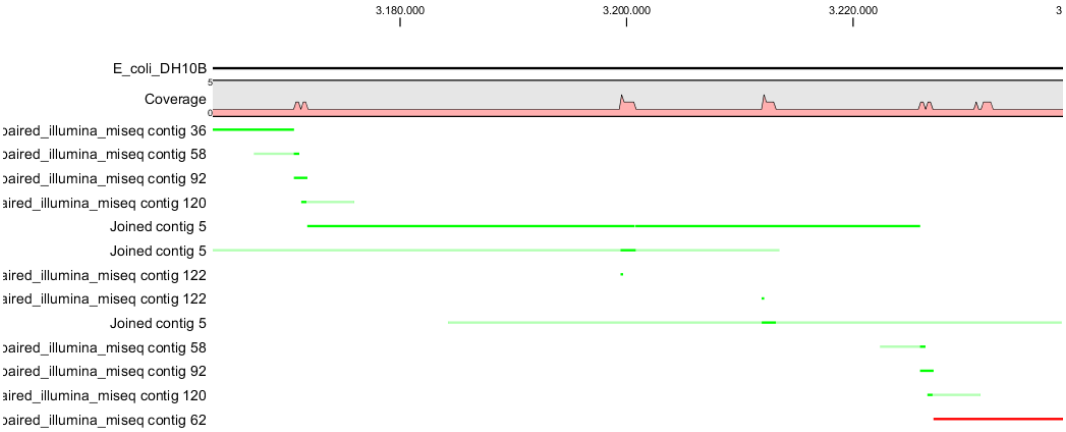


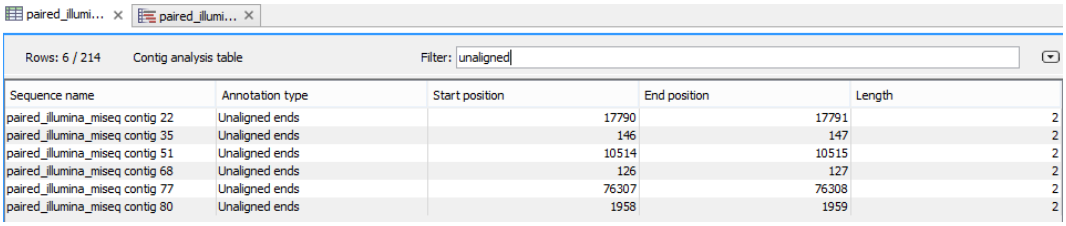
Figure 15: *Joined contig 5 - the result of joining contigs 8, 122, 121, 60, and 67.*

Note that although we have joined several contigs, we still have to remember that a gap was included when joining Joined contig 3 and 4. Filling out the gap can be done by designing primers to the region around the gap and resequence this area. Primer design can be performed using the "Create primers" tool (not described in this tutorial).

Splitting contigs

We will now look at how to split contigs. This is useful whenever you suspect that reads have been misassembled. To identify such regions, you can use the **Analyze Contigs** tool, which will only be mentioned briefly in this tutorial. Further information about the Analyze Contigs tool can be accessed at: http://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 relevant assembly (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. Chose to **Save** the results and click on the button labeled **Finish**. This generates a **Contig analysis table**.
4. Open the table and use the filter function to identify **Unaligned ends** by typing "unaligned" in the search field as in figure 16.



Sequence name	Annotation type	Start position	End position	Length
paired_illumina_miseq contig 22	Unaligned ends	17790	17791	2
paired_illumina_miseq contig 35	Unaligned ends	146	147	2
paired_illumina_miseq contig 51	Unaligned ends	10514	10515	2
paired_illumina_miseq contig 68	Unaligned ends	126	127	2
paired_illumina_miseq contig 77	Unaligned ends	76307	76308	2
paired_illumina_miseq contig 80	Unaligned ends	1958	1959	2

Figure 16: *Result of contig analysis. The filter has been used to identify contigs with unaligned ends.*

Of the six contigs that contain unaligned ends, we will focus on contig 68. As contigs cannot be opened from the **Contig analysis table** we must open it from the **Contig table** ( ).

5. Open the **Contig table** ( ) and double-click on contig 68 to open the read mapping view. Under **Annotation types** in the right side panel tick **Unaligned ends** and find the region containing the unaligned ends. You can hold down the Ctrl key while scrolling with the mouse wheel to zoom out until you find the annotation, and zoom in holding your mouse on the annotation (figure 17). The relative high read coverage of the region to the left of the unaligned ends indicates that this is a repetitive region which has not been duplicated the correct number of times by the de novo assembler. To resolve this error we need to split the contig and hereby isolate the repetitive region.
6. To split the contig select the two nucleotides covered by the **unaligned ends** annotation in the read mapping view.
7. Right-click on the two selected nucleotides and select **Split contig...** (figure 18). This opens up a dialog where reads intersecting the split can be distributed between the resulting two contigs. The contigs will be automatically expanded to preserve the alignment of the reads, which means that after a split where one or more reads intersect the split, the resulting two contigs will have an overlap. Each read is automatically placed on the contig where it has the best alignment.

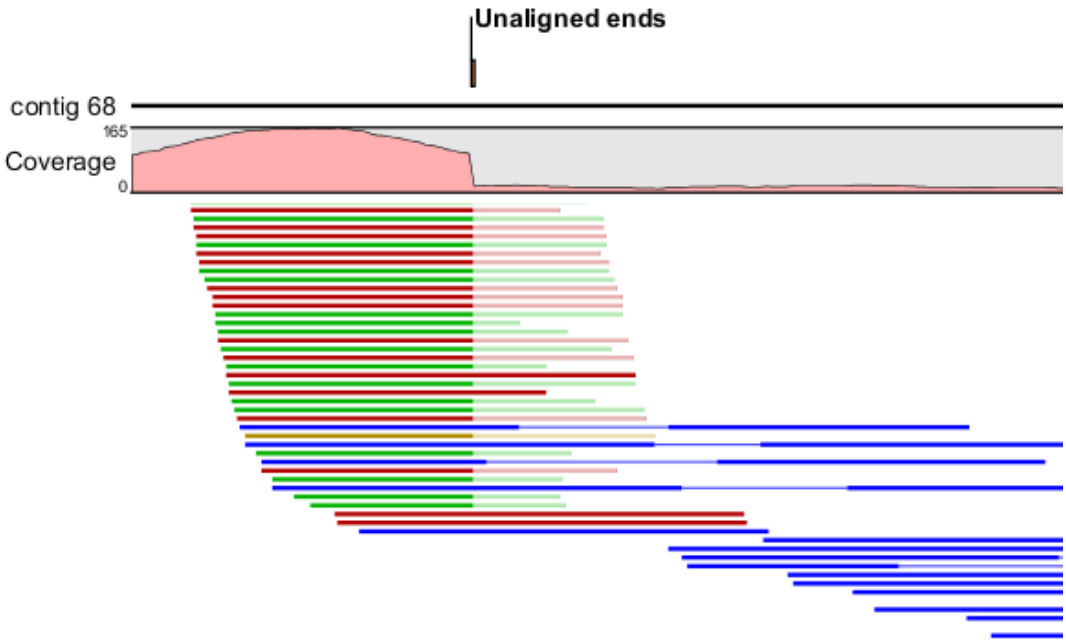


Figure 17: *Contig 68 contains unaligned ends as a result of a repeat region which has not been duplicated the correct number of times.*

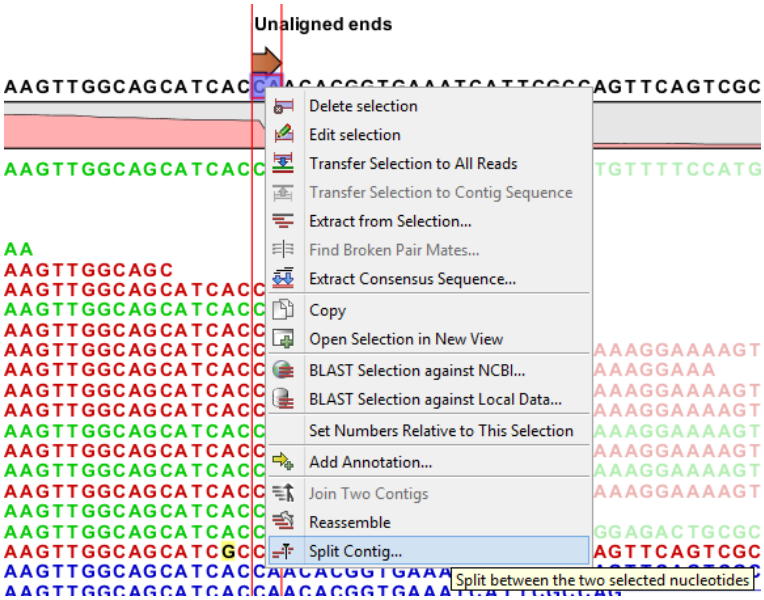
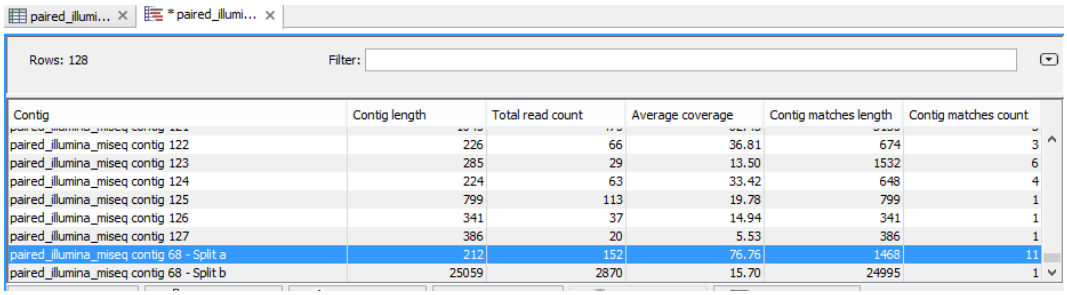


Figure 18: *Splitting a contig.*

8. Leave the reads in the default position and click **Finish**. The contig has now been split into two contigs (**contig 68 - split a** and **split b**) that have replaced contig 68 in the Contig table ( ). The workbench will automatically close the read mapping for contig 68 after the split has been performed as it no longer exists.

9. Open the **Contig table** ( ) and scroll down to contig contig 68 split a and split b in the bottom of the list. The column **Contig matches count** indicates the number of times a contig matches the reference (figure 19). Contig **68 split b** has only one match while contig **68 split a** has 11 matches, another indicator of a repetitive element.



Contig	Contig length	Total read count	Average coverage	Contig matches length	Contig matches count
paired_illumina_miseq contig 122	226	66	36.81	674	3
paired_illumina_miseq contig 123	285	29	13.50	1532	6
paired_illumina_miseq contig 124	224	63	33.42	648	4
paired_illumina_miseq contig 125	799	113	19.78	799	1
paired_illumina_miseq contig 126	341	37	14.94	341	1
paired_illumina_miseq contig 127	386	20	5.53	386	1
paired_illumina_miseq contig 68 - Split a	212	152	76.76	1468	11
paired_illumina_miseq contig 68 - Split b	25059	2870	15.70	24995	1

Figure 19: *Contig table after contig 68 has been split in two. Split b is a repetitive element that match at 11 different position in the reference.*

10. Open the **Contig match table** ( ) and sort the entries by contig name (click on the header for the second column). Find the entries for contig 68 split a and notice that only 126 out of 212 nucleotides match the reference in most cases. We will investigate this in the Contig match view.

11. Open the **Contig match view** and scroll to the first match for contig 68 split a (position 20438 to 20563). Notice that a large region of the contig does not align to the reference (figure 20). There is also a small gap between the contig 68 split a and contig 99 which makes impossible to join the two. Furthermore, the repeat seem to be present in contig 55, but by switching to **packed view** in the right side panel, we see that the repeat in contig 55 contains a lot of mismatches and is therefore unlikely to belong here.

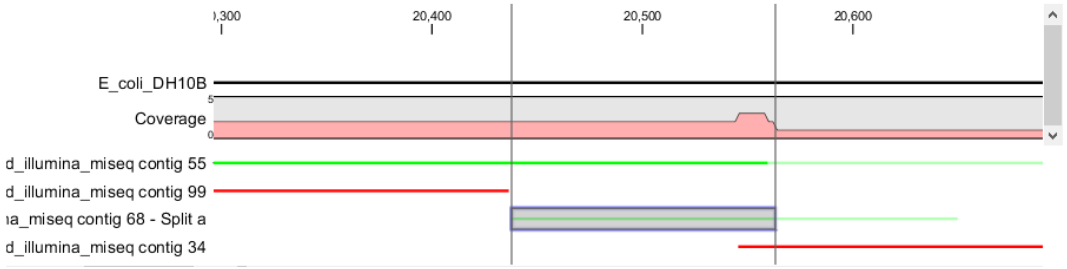


Figure 20: *Contig match view showing the first match of contig 68 split a.*

12. Open the **Contig view** for contig 68 split a, turn on the **split position** annotation and observe that a region starting from position 126 (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 21).

Extending contigs

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

1. Open the read mapping for **contig 68 split a** and scroll to the first position. Arrows pointing out the 5' end of the contig indicate reads that stick out the contig end (figure 22).

   Such reads can be used to extend the contig and hereby close the gap to contig 99.

2. Open the read mapping for **contig 68 split a**, ensure that the window is active and then select: **Toolbox** | **Genome Finishing Module** ( ) | **Extend contigs** ( )

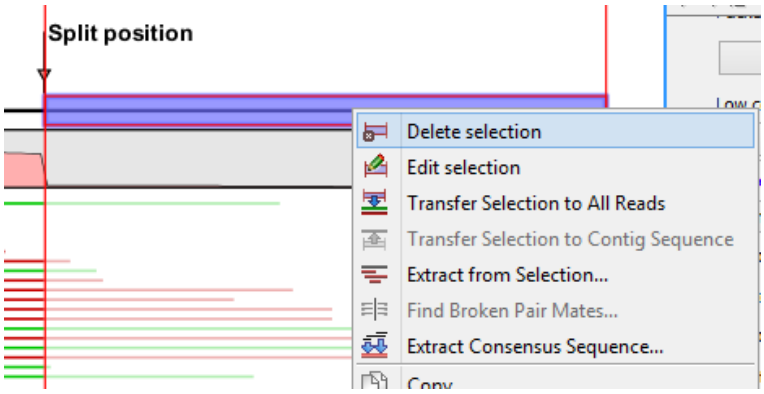


Figure 21: *Deleting a low coverage region that is no longer needed.*

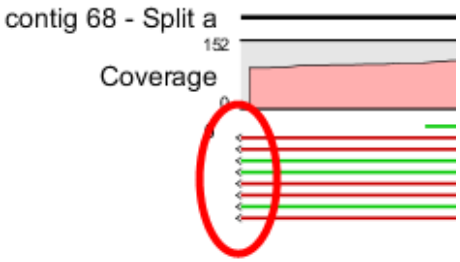


Figure 22: *Reads pointing out of a contig end that could be used to extend the contig.*

This will start the tool and 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.

3. Ensure that the **Minimum coverage to extend** is set to 3 and click **Next**. In the next step of the wizard choose to **Save** the results and choose a suitable location to store them.

4. Open the extended contig you just saved and enable the **Extended region** annotation. The contig should now have a 74bp extension at the 5' end (figure 23).

Figure 23: *Contig 68 split a after extension.*

5. Open the **Contig table** ( ) and use the **Add/Extract** button to add the extended contig.

6. **Delete** the old **contig 68 split a** as this is no longer needed. This is done through the Contig table ( ) by selecting the contig in the list and clicking **Remove contigs**.

7. The extended contig does not have any reads mapped to it. To change that, click the **Map reads** button in the Contig table ( ) and select **Replace all read**". This will remove all reads from all contigs and map new reads.

8. In the wizard select the paired Illumina reads which were imported at the start of this tutorial **paired_illumina_miseq_1 (paired)**. These reads are the same set of reads that were mapped to the contigs when the tutorial started. Mapping them again will update the read mapping on all contigs. Note that this may take some time.

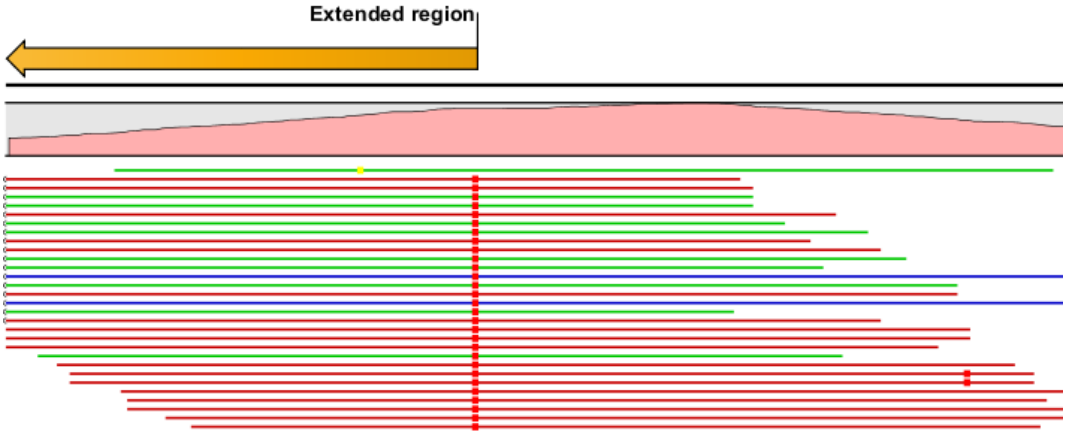


Figure 24: *The extended contig after reads have been remapped.*

9. Open the extended contig to confirm that the extension has an even coverage (figure 24)
10. The extended contig now completely spans the gap between contig 99 and contig 34 (figure 25). Use the join functionality as described in section to close the gap but remember to make copies of the contig as it is needed for closing up to 11 gaps.

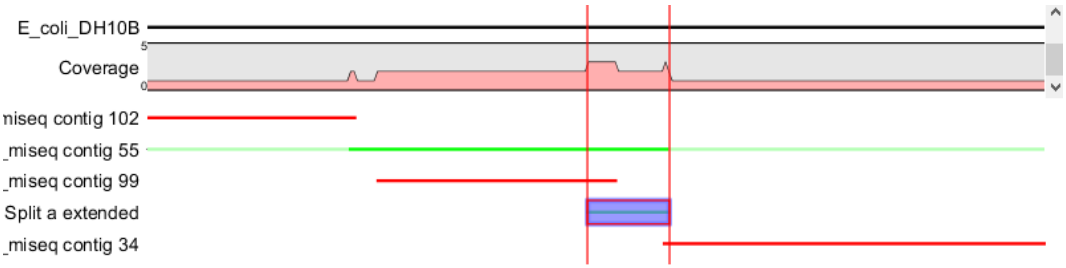


Figure 25: *The contig match view confirm that the extended contig is spanning the gap between contig 99 and contig 34.*