



Tutorial

De Novo Assembly of Paired Data

March 2, 2017

— Sample to Insight —

De Novo Assembly of Paired Data

A de novo assembly involves taking many short sequences and trying to assemble them into longer, contiguous sequences. We recommend that you read about how the de novo assembly tool works before running this tool on your own data. You can read more in our manual http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=De_novo_sequencing.html or in the **White Paper**: http://resources.qiagenbioinformatics.com/white-papers/White_paper_on_de_novo_assembly_4.pdf.

This tutorial takes you through a typical *de novo* sequencing work flow with paired sequence data. We will

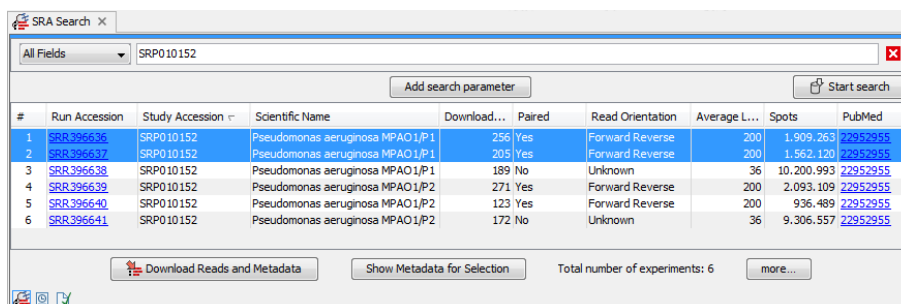
- Search and import sequence read data.
- Run a quality check on the read data.
- Trim the reads based on quality.
- Run a de novo assembly, including scaffolding.

We also include optional sections on finding broken pair mates in mapping results and exporting contigs from a set of mappings.

Import the data into the Workbench We will use two Illumina read sets from *Pseudomonas aeruginosa*, one mate-pair data, and the other paired-end data. The data is available from the Short Read Archive under the Study Accession SRP010152.

To download the data:

1. In the workbench go to **Download | Search for Reads in SRA**.
2. Enter the study accession number and click **Start search**.
3. Select SRR396636 and SRR396637. Click **Download Reads and Metadata** (figure 1).



#	Run Accession	Study Accession	Scientific Name	Download...	Paired	Read Orientation	Average L...	Spots	PubMed
1	SRR396636	SRP010152	<i>Pseudomonas aeruginosa</i> MPAO1/P1	256	Yes	Forward Reverse	200	1,909,263	22952955
2	SRR396637	SRP010152	<i>Pseudomonas aeruginosa</i> MPAO1/P1	205	Yes	Forward Reverse	200	1,562,120	22952955
3	SRR396638	SRP010152	<i>Pseudomonas aeruginosa</i> MPAO1/P1	189	No	Unknown	36	10,200,993	22952955
4	SRR396639	SRP010152	<i>Pseudomonas aeruginosa</i> MPAO1/P2	271	Yes	Forward Reverse	200	2,093,109	22952955
5	SRR396640	SRP010152	<i>Pseudomonas aeruginosa</i> MPAO1/P2	123	Yes	Forward Reverse	200	936,489	22952955
6	SRR396641	SRP010152	<i>Pseudomonas aeruginosa</i> MPAO1/P2	172	No	Unknown	36	9,306,557	22952955

Figure 1: Importing mate-pair Illumina data.

4. You can select "Discard read names" to save some time and space on your machine but keep "Discard quality scores". Click **Next**.
5. In the parameters dialog, the workbench gives some distance estimates. You can see that there is a difference between the two data sets, due to the fact that SRR396636 is Mate-pair data while SRR396637 is Paired-end.

6. Save the reads in a new folder dedicated to this tutorial.

Importing large datasets can take some time. You can take a look to see if the import is still running by clicking on the tab called **Processes** in the bottom left side of the workbench.

Once the import is completed, you should see two new files visible in the Navigation Area of the Workbench. By default, these will be called and SRR396636.sra (paired) and SRR396637.sra (paired).

Renaming the data objects

To help remember which dataset is which, please rename these data objects. In this case, we suggest you rename them to reflect the type of data is in each object.

1. Click on the name of SRR396636.sra (paired) in the **Navigation Area** so it is highlighted.
2. Click again on the name. This should put it in a mode where you can edit it. (If not, try pressing the F2 key.)
3. Edit the name, changing it to *Mate-pair*.
4. Do the same steps for SRR396637.sra (paired), but change it to *Paired-end*.


Sequencing Quality Analysis

We wish to use only high quality data in the de novo assembly. A sequencing quality analysis helps assess the quality of the datasets we are about to use. If the data contains lower quality regions, it should be trimmed, and the trimmed sequences used as input to the de novo assembly. If there are over-represented sequence motifs in the data, then it is worth checking if any adapters remain. If so, it is very important that these be trimmed away before using the data for de novo assembly.

Running a quality analysis

Here, we run the Create Sequencing QC Report tool on both our sequence data objects. We will use the batch functionality, allowing us to simultaneously launch the analysis of multiple sets of data.

1. Go to:

Toolbox | NGS Core Tools () | Create Sequencing QC Report

2. Check the box labeled **Batch** at the bottom of the Wizard window. Note that if the box labeled **Batch** is not checked, you will only be able to move data objects to the Selected Elements pane on the right, not folders, and you will generate only one report instead of one report per element.
3. Click on the folder containing your sequence data objects and then click the arrow pointing right. This will move the folder into the right hand pane as shown in figure 2. Click **Next**.
4. The next window allows you to fine tune which data objects the analysis should run on. If you click on either of the object names in the left hand side, the data object(s) to be worked on will be listed on the right hand side. Here, each folder contains only the data that we wish to work with, so we can just proceed to the next step.

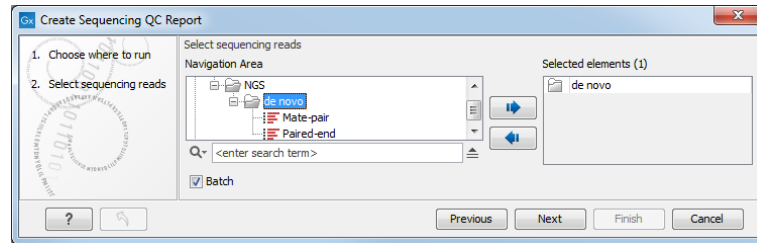


Figure 2: Set up a batch job for running a quality check analysis.

5. Check the boxes labeled **Create graphical report** and **Create supplementary report**. Uncheck the box labeled **Create duplicated sequence list**. Check the box labeled **Save in input folder** to save the results before clicking on the button labeled **Finish**.

When both the quality check analyses are finished, you should see graphical and summary reports for each of your data sets listed in the Navigation Area.

Investigating the Sequencing Quality Check Reports

Here, we wish to look at the quality distribution for the reads and the sequence duplication information.

1. Open the file called **Mate-pair - graphical QC report** and take a look at the content.
2. Look at the **Quality distribution** plots in section 2.4 and 3.5 as shown in figure 3.

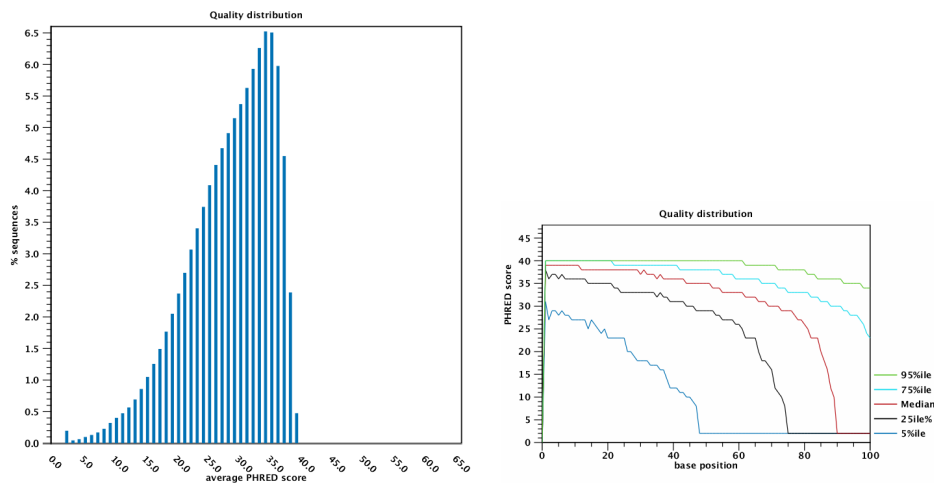


Figure 3: Quality distribution for the mate-pair dataset.

The overall distribution suggests that there is some lower quality data present, while the per-base plot shows that the data quality drops near the end of the sequences.

3. Look at the **Sequence duplication levels** plot in section 4.2. That plot does not suggest that sequence duplication is an issue here.

There is also further information in the supplementary QC report file in the table called **Duplicated sequences** in section 4.3, if you suspect that sequence duplication might be an issue in your data set.

4. Open the file called **Paired-end - graphical QC report** and take a look at the plots relating to quality. The plots will be similar to those for the mate-pair data.

If you look at the information in the supplementary reports, you will see that there are some sequences that contain only ambiguous nucleotides. Such sequences will be removed during the trimming process, which we run next.

Trimming the data

Based on what we know of the data, we will trim low quality data away as well as removing sequences that contain too many ambiguous bases.

As we will be running the same trimming task on both our sets of data, we can again take advantage of the batch functionality of the Workbench.

1. Go to:

Toolbox | NGS Core Tools (📁) | Trim Sequences (🔧)

2. Check the box labeled **Batch** at the bottom of the Wizard window. Click on the folder containing your sequence data objects and then click the arrow pointing right. Click on the button labeled **Next** twice.
3. For the "Quality trimming", use the default settings, which set a quality score of 0.05 and a maximum number of ambiguous nucleotides of 2. You can use the "Set to default" button (👈) if you are not sure whether you have previously changed the parameters for the tool. Click on the button labeled **Next**.
4. We do not need to perform "Adapter trimming". If you had run an adapter trimming previously, and an adapter trim list appears in this wizard step, please use the "Set to default" button (👈) to clear the previous settings and click fNext.
5. For Sequence filtering, check the option "Discard reads below length" and set the number 15. Click **Next**.
6. In the Output options section, choose to **Save broken pairs** and to **Create report**. **Save** the results in the input folder and click on the button labeled **Finished**.

You should see four sequence lists resulting from the trimming process. These are the ones highlighted in figure 4. There are also two reports generated. These are called Mate-paired report and Paired-end report. We will now use the four trimmed sequence lists as input to a de novo assembly.

De novo assembly

There are two general types of output you can generate from the de novo assembly tool in the *CLC Genomics Workbench*:

- **Simple contigs** Here, the output is a sequence list of the contigs generated.
- **Stand-alone mappings** Here, a read mapping is carried out after the de novo assembly, where the sequence reads used for the assembly are mapped to the contigs that were assembled.

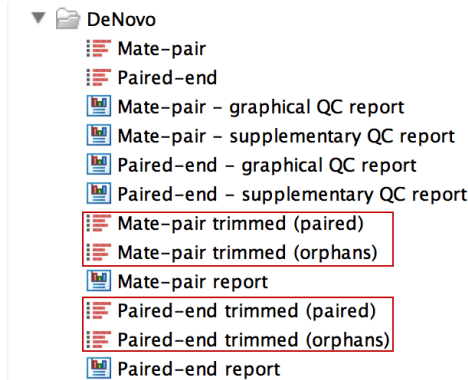


Figure 4: Four sequence list files are generated by the trimming tool, each with a name that includes the word: trimmed. The sequence lists with (paired) in their names contain paired sequences. The other trimmed lists contain single reads, where the mates of these reads were removed during the trimming process.

In this tutorial, we will choose to run a mapping of the reads to the assembled contigs when we set up the de novo assembly.

1. Go to:

Toolbox | De Novo Sequencing | De Novo Assembly (🔧)

In the Wizard that starts up:

2. Select the four sequence lists that were generated by the trimming tool: **Mate-pair trimmed (paired)**, **Mate-pair trimmed (orphans)**, **Paired-end trimmed (paired)** and **Paired-end trimmed (orphans)**. Click **Next**.
3. Set the de novo parameters as default (figure 5) and click **Next**.

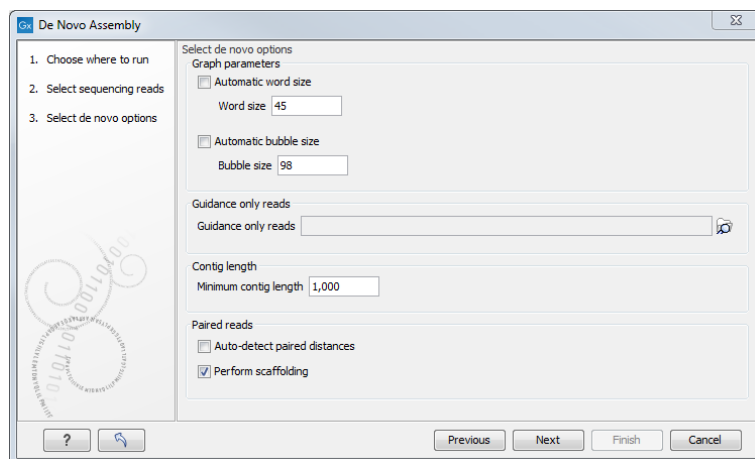


Figure 5: Configure the parameters to be used for the de novo assembly.

4. Set the mapping parameters as default, then uncheck **Update contigs**(figure 6) and click **Next**. If you checked the box labeled **Update contigs**, then areas of your contigs where no reads map will be cut out of the final contigs. This option assumes that if no reads map back to a region of a contig, then there is no evidence in the data that such a region exists. We choose not to do this in this tutorial.

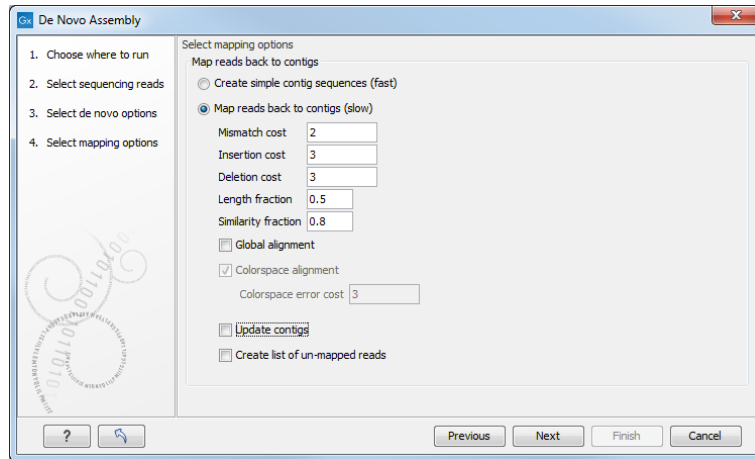


Figure 6: Configure the parameters to be used for the de mapping.

5. In this next step, you choose the type of output to produce. Choose to **Create report** and to **Save** the output. Click on the button labeled **Next**.
6. Click on the folder you wish to save the assembly outputs to before clicking on **Finish**.

Two analyses, which will run consecutively, have just been launched: a de novo assembly and a read mapping. Note that both can take some time. The length of time this assembly will take depends on the specifications of your machine. You can monitor the progress of the analysis within the Processes tab in the bottom left hand side of the Workbench. The progress in percentage points will generally be quite uneven, as the progress bar provides information on the stage the task is on, rather than being a good indicator of the relative time taken or remaining for a task.

For multi-stage jobs such as de novo analysis and mappings, the text above the progress bar is a useful indicator of what stage the task has progressed to. Figure 7 shows what you might typically see when running a de novo analysis followed by a mapping of the reads back to the contigs.

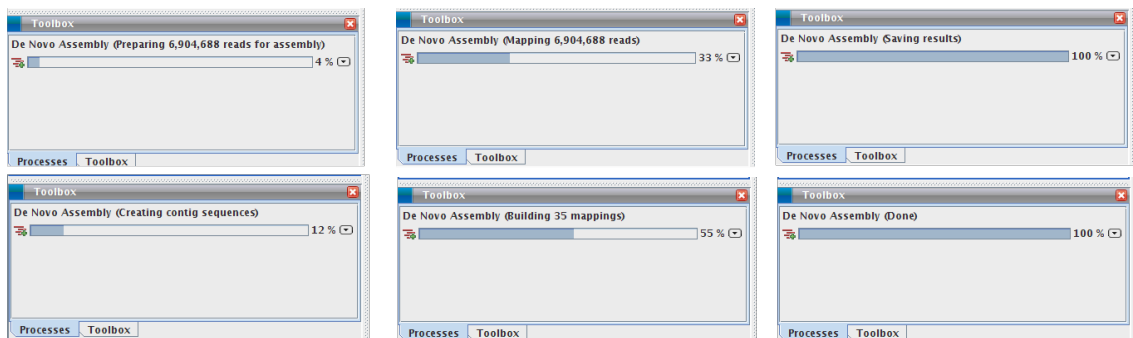



Figure 7: Phases reported in the progress tab for a de novo assembly where reads are also mapped back to the assembled contigs.


Investigating the de novo assembly results

As you chose to map the reads back to the assembled contigs, the main output generated will be a summary read mapping object, which contains read mappings against each of the scaffolds

that were generated during the de novo assembly stage of this analysis. This will be called something like **Mate-pair trimmed (paired) de novo assembly** with an icon () beside it in the Navigation Area. You should also see a corresponding report file with a name like **Mate-pair trimmed (paired) summary report**.

1. Double click on the summary report object to open it.

In the report is summary information such as the nucleotide distribution, information on contig length, as well as the N25, N50 and N75 values. You should see a table for the contig lengths with scaffolding and another table lower down for contigs without scaffolding.

2. Double click on the assembly, which has an icon beside it that looks like: ()

This opens a table view in the Navigation pane of the Workbench that lists information about the mappings that were done.

3. Click on the **Show Annotation Table** icon () at the bottom of the viewing area.

The annotations tell you about certain actions carried out when determining the output of the de novo assembly. There are three annotation types that can be reported by the de novo assembly tool:

- **Alternatives Excluded:** More than one path through the graph was possible in this region but evidence from paired data suggested the exclusion of one or more alternative routes in favour of the route chosen.
- **Contigs Joined:** More than one route was possible through the graph such that an unambiguous choice of how to traverse the graph cannot be made. However evidence from paired data supports one of these routes and on this basis, this route is followed to the exclusion of the other(s).
- **Scaffold:** The route through the graph is not clear but evidence from paired data supports the connection of two contigs. A single contig is then reported with N characters between the two connected regions. This entity is also known as a scaffold. The number of N characters represents the expected distance between the regions, based on the evidence the paired data.

Take note of a contig that contain several of these annotations.

4. Go back to the contig table and open the contig you chose. Double-click on the name of the mapping object within the tab in the viewing area.

This expands the mapping to take up all the available viewing space. (To view the full workbench again, double click on the name in the tab again.)


We will not spend time focussing on the details of viewing mappings in this tutorial. Details about this topic can be found in our manual starting at:

http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=View_settings_in_Side_Panel.html

5. Click on the **Zoom Fit** button in the right lower corner to zoom out completely and see your whole read mapping within the viewing area.


6. In the right hand pane called **Read Mapping Settings**, open the tab called **Annotation types**. You can now check (or uncheck) the annotations you wish to see on the reference sequence.

If you are interested in investigating certain regions containing particular annotations in more detail, you can zoom in by using the **Zoom in** and **Zoom out** buttons in the bottom toolbar.

However, opening the table view of annotations as a linked view is often the easiest way to work with standard mapping objects and their annotations. When you click on a row in a table linked to a graphical view, the cursor moves to the relevant position within the mapping view. To open a linked table, you just need to depress the **Ctrl** key on your keyboard (**cmd** on Mac), and simultaneously use the mouse to click on the **Show Annotation Table** icon () at the bottom of the viewing area.

Now you should have the annotation table open as a linked view for the mapping object.

7. Now, in the annotation table view, double click on a row.

You should see that the cursor jumps to the section of the mapping the row refers to. Use the Zoom to a selection icon () to see the whole annotation (as in figure 8).

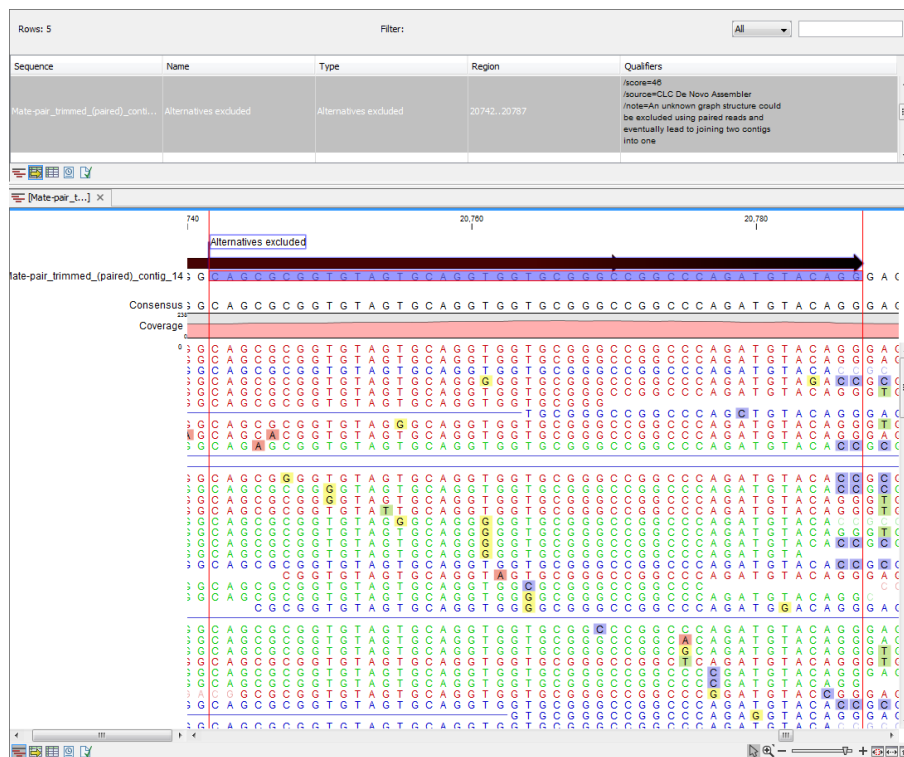



Figure 8: Double clicking on the row shown highlighted in the table above jumps the cursor to the point in the mapping view shown below. If you wish to, you can change the colors of the annotations displayed by just clicking in the coloured box next to each annotation type, and selecting a new color.

Note that the results of analyses carried out in the Workbench include history information. This can be useful if you need to recall what parameters you used or what version of the Workbench your analysis was run using. To access the history information, just click on the Show History () button at the bottom of the viewing area.

Finding broken pair mates

Reads mapped back to the contigs where both partners of a pair map in the correct relative orientation and within the expected distance range are coloured blue in the mapping and are called an intact pair. Those where only one member of the pair mapped, or those where both partners mapped, but in the wrong relative orientation or outside the expected distance range are considered members of a broken pair. Members of a broken pair that map to a unique location against the reference are colored green or red, according to whether they mapped in the forward or reverse orientation respectively.

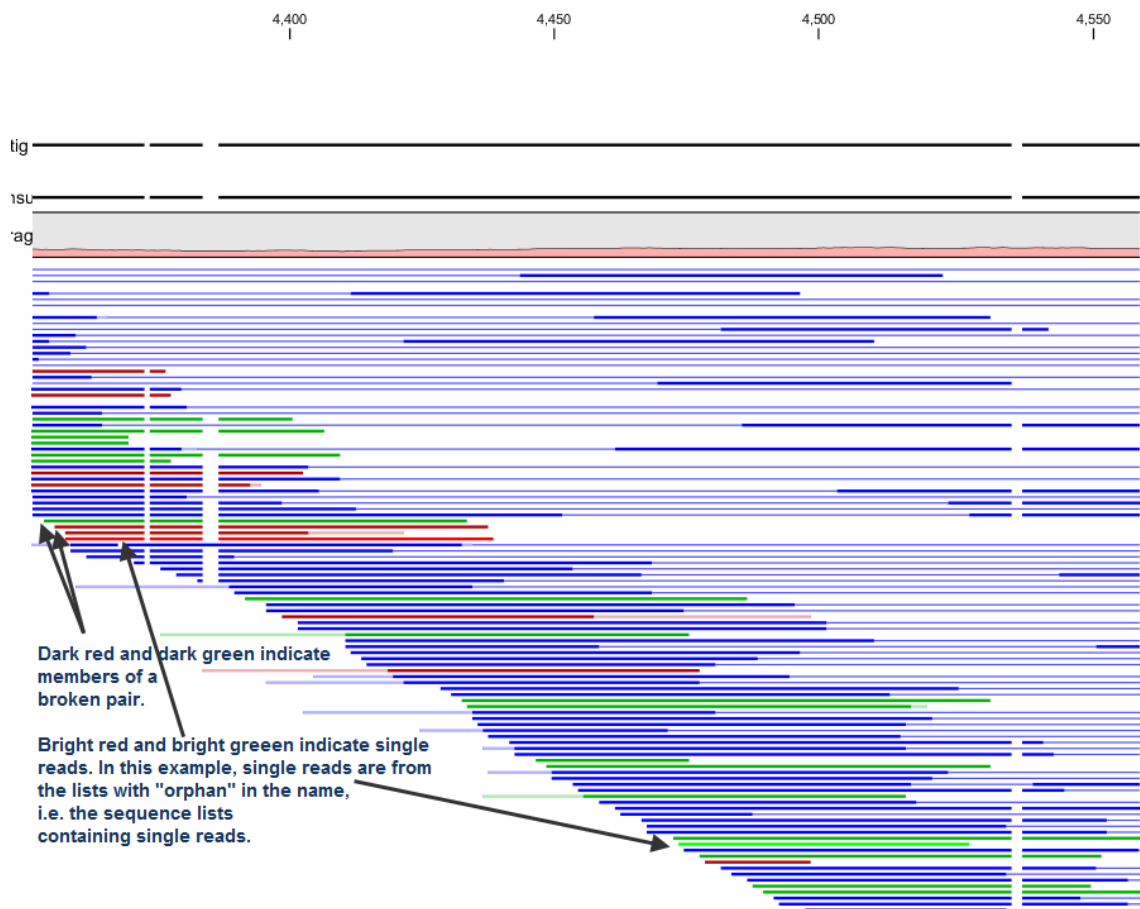


Figure 9: Read colors reveal information. Solid dark blue lines represent members of an intact pair. Light blue lines represent the connection between intact pair members. Green solid lines and red solid lines indicate single reads mapped in the forward direction and reverse direction relative to the reference respectively; dark green and dark red represent members of a broken pair, while bright red and bright green represent single reads. Pale green and pale red represent portions of a read that do not match the reference. Not shown here are reads mapping in yellow, meaning the read could map to multiple locations equally well.

To investigate where mapped mates of a broken pair are in the assembly:

1. Highlight a region of interest. For example, find a scaffold annotation where you see green and red coloured lines on either side of the gap, and highlight the region around this, as shown in figure 10.
2. Right click the mouse cursor over the highlighted region (underneath the annotation).

Choose the option **Find Broken Pair Mates** as shown in figure 10.

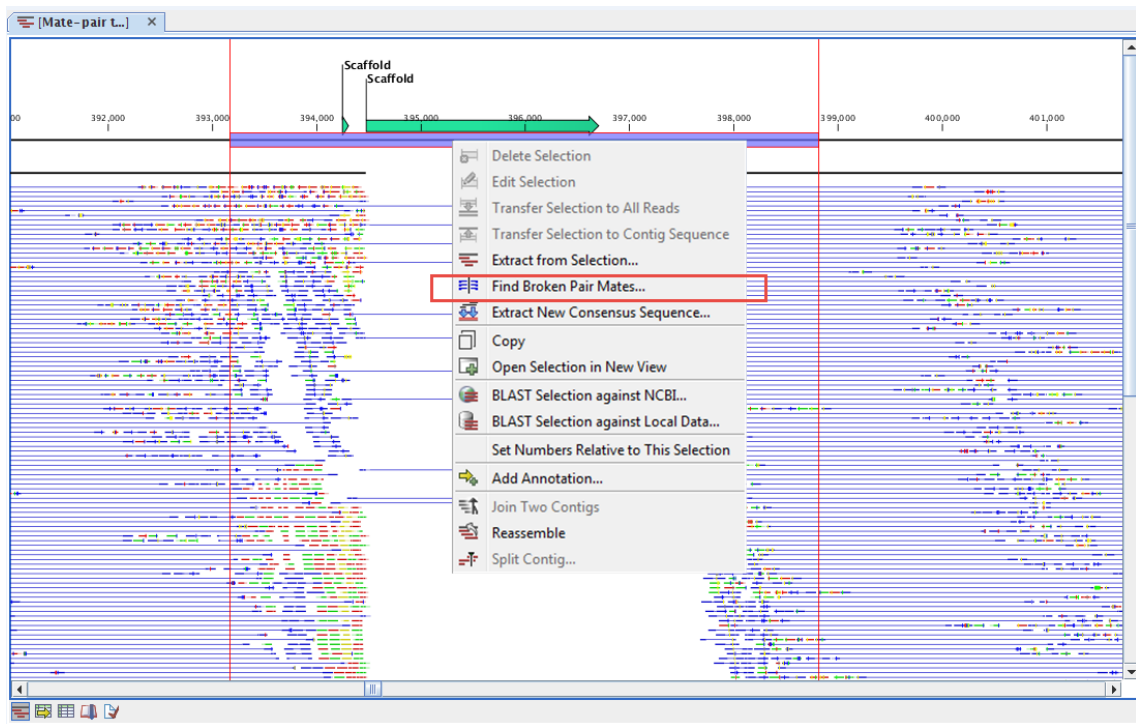


Figure 10: Finding the pair mates for broken pair reads in a particular region.

3. Check the **Show overlapping annotations** box and select all annotation types present after clicking on the yellow arrow icon in the wizard (figure 11). Note that a summary of the number of broken pairs in the region selected is also given in this dialog. Click **Next**.

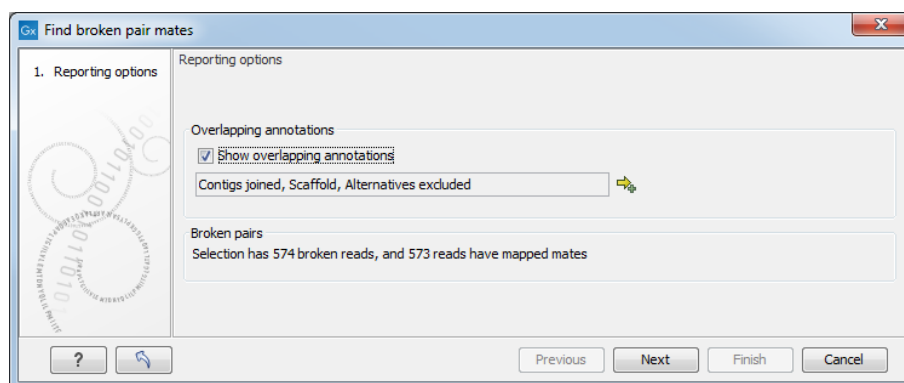


Figure 11: Choosing to include overlapping annotations.

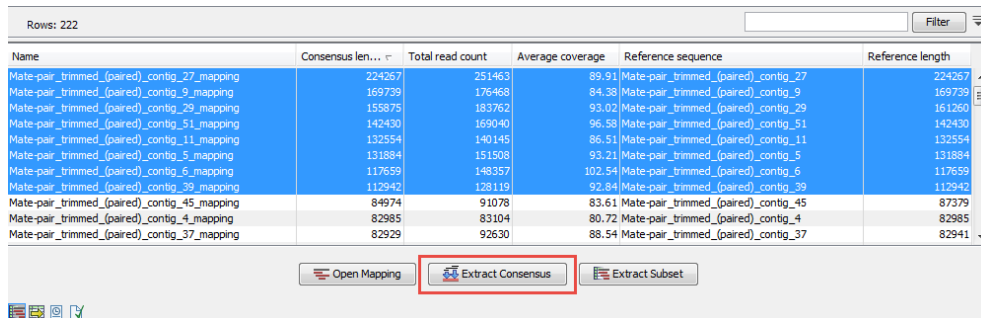
4. Choose to **Open** the results and click on the button labeled **Finish**.

The table generated is linked to your mapping object. Open both in split view. At this point, you can double click on rows in the linked table and jump to those locations where the mates of the broken pair reads from the selected region are mapped.

Exporting contigs from mapping results

You can extract the output from the de novo assembly stage as a sequence list if you wish. This can be useful when working with downstream programs that expect sequences as input, e.g., running a BLAST search, or if you wish to export just the contig sequences from the Workbench. Here, we go through the steps to export contig sequences from the mappings were generated.

1. Highlight the rows of the summary mapping table relating to the contigs you wish to extract. In figure 12, the 5 longest contigs have been selected.
2. Click on the button labeled **Extract Consensus** (figure 12).



Name	Consensus len...	Total read count	Average coverage	Reference sequence	Reference length
Mate-pair_trimmed_(paired)_contig_27_mapping	224267	251463	89.91	Mate-pair_trimmed_(paired)_contig_27	224267
Mate-pair_trimmed_(paired)_contig_9_mapping	169739	176468	84.38	Mate-pair_trimmed_(paired)_contig_9	169739
Mate-pair_trimmed_(paired)_contig_29_mapping	155875	183762	93.02	Mate-pair_trimmed_(paired)_contig_29	161260
Mate-pair_trimmed_(paired)_contig_51_mapping	142430	169040	96.58	Mate-pair_trimmed_(paired)_contig_51	142430
Mate-pair_trimmed_(paired)_contig_11_mapping	132554	140145	86.51	Mate-pair_trimmed_(paired)_contig_11	132554
Mate-pair_trimmed_(paired)_contig_5_mapping	131884	151508	93.21	Mate-pair_trimmed_(paired)_contig_5	131884
Mate-pair_trimmed_(paired)_contig_6_mapping	117659	148357	102.54	Mate-pair_trimmed_(paired)_contig_6	117659
Mate-pair_trimmed_(paired)_contig_39_mapping	112942	128119	92.84	Mate-pair_trimmed_(paired)_contig_39	112942
Mate-pair_trimmed_(paired)_contig_45_mapping	84974	91078	83.61	Mate-pair_trimmed_(paired)_contig_45	87379
Mate-pair_trimmed_(paired)_contig_4_mapping	82985	83104	80.72	Mate-pair_trimmed_(paired)_contig_4	82985
Mate-pair_trimmed_(paired)_contig_37_mapping	82929	92630	88.54	Mate-pair_trimmed_(paired)_contig_37	82941

Figure 12: Select the 5 longest contigs in the summary mapping table and then click on the button labeled *Extract Consensus*.

3. Choose to save the output.