QIAGEN

# Tutorial

## De Novo Assembly Using Long Reads and Short Read Polishing

February 2, 2024

Sample to Insight

QIAGEN Aarhus · Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
digitalinsights.qiagen.com · ts-bioinformatics@qiagen.com

# De Novo Assembly Using Long Reads and Short Read Polishing

This tutorial is an introduction to working with the tools in the Long Read Support plugin.

Here we will focus on the tools developed for working with long, error-prone reads such as those produced by the Continuous Long Read (CLR) sequencing of Pacific Biosciences or Oxford Nanopore Technologies.

If you instead have long, high-quality reads, such as PacBio HiFi reads, you can still use **De Novo Assemble Long Reads** and **Map Long Reads to Reference**, but there is no need to polish the assembly.

The tutorial covers the following:

- Importing data required for the analysis.
- De novo assembling a microbial-sized genome using long, error-prone reads.
- Mapping long reads to a reference and visualizing an assembly.
- Improving a de novo assembly from long reads by polishing with short, high-quality reads.

## Prerequisites

For this tutorial, you must be working with *CLC Genomics Workbench* 24.0 or higher with the Long Read Support plugin installed. If you are working with a different version, some elements in this tutorial (such as contig naming) may differ slightly from your local installation.

Optional: For additional evaluation steps, you will also need the Whole Genome Alignment plugin.

How to install plugins using the Plugin Manager is described here: http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Install.html.

### Download and import data

1. Download the sample data from our website https://resources.qiagenbioinformatics.com/testdata/CAV1492_MinION_and_Illumina_tutorial_data.zip and unzip it.
2. Open the *CLC Genomics Workbench*.
3. Create a new folder for the project, for example named "Long Read Tutorial".
4. Import the Oxford Nanopore MinION reads by going to:

   **File** | **Import** | **Oxford Nanopore. . .**

   Leave settings as default. Click **Add files** and add the file:

   `S. Marcescens_CAV1492-MinION.fastq`

   Click **Next** and save it to the location you created.
5. Import the Illumina reads by going to:

   **File** | **Import** | **Illumina. . .**

Leave settings as default, including leaving **Paired reads** checked under *General options*. Click **Add files** and select:

`S. marcescens_CAV1492-Illumina_downsampled_R1.fastq`

`S. marcescens_CAV1492-Illumina_downsampled_R2.fastq`

Click **Next** and save to the same location as the MinION reads.

6. Lastly, import the reference by going to:

   **File** | **Import** | **Standard Import. . .**

   Leave the default option **Automatic import** checked. Click **Add files** and select:

   `S.marcescens_CAV1492_genome.fa`

   Click **Next** and save it to the same location.

Your folder should look like figure 1.

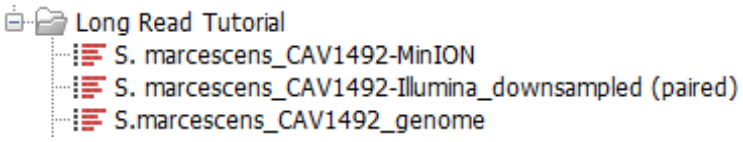


Figure 1: *Tutorial folder with imported data.*

The data for this tutorial is from a study examining the feasibility of using reads from Oxford Nanopore to fully assemble and resolve bacterial genomes including plasmids. It contains long MinION read and short Illumina read sequencing data from eight Enterobacterales isolates of six different species [George et al., 2017].

We will assemble the isolate from species Serratia marcescens strain CAV1492. This strain has one chromosome and five plasmids. The sequencing data contains 7,038 MinION reads with an average read length of 12,566 bp. We have also imported a set of Illumina reads sequenced from the same strain. These reads have been downsampled to lower the runtime of analysis in this tutorial.

Lastly, we have imported a reference standard made using deep coverage PacBio and paired-end sequencing. The reference standard can be found in BioProject PRJNA246471.

## De novo assemble long reads

The De Novo Assemble Long Reads tool makes it possible to create de novo assemblies from long reads.

It is best practice to trim any adapters from the reads before assembly, but error-prone reads should not be quality-trimmed. Trimming can be done with the Trim Reads tool (see `http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Trim_Reads.html`). The Oxford Nanopore reads we imported in the beginning of the tutorial have already been trimmed for adapters, so we will skip this step.

1. To start the tool, search and run using Launch ( ) or locate **De Novo Assemble Long Reads** in the Toolbox:

   **Long Read Support ( )** | **De Novo Assemble Long Reads ( )**

2. Select the imported MinION reads (figure 2) and click **Next**.

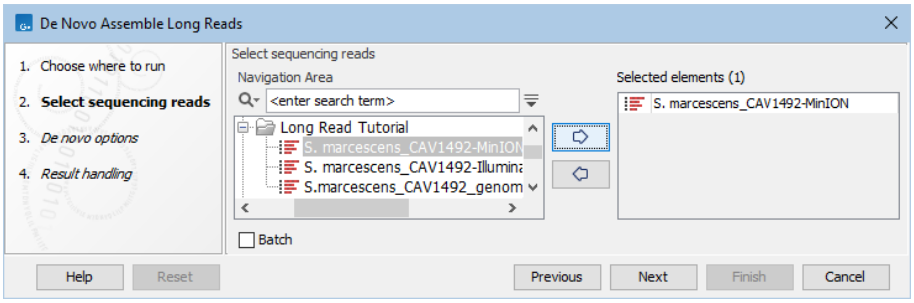


Figure 2: *Selecting the MinION reads.*

3. As we are working with a shorter genome, set **Minimum contig length** = 1 in *De novo options* as shown in figure 3. Make sure the **Polish with reads** and **Keep circular contigs** options are checked. The tool will run two rounds of polishing after assembling. Click **Next**. Note: If the input is PacBio HiFi reads, the *PacBio HiFi options* become available instead.

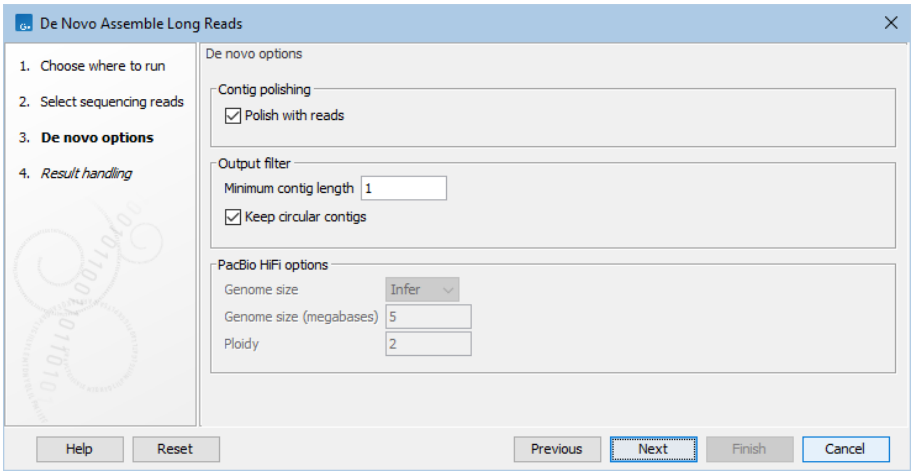


Figure 3: *De Novo Assemble Long Reads options.*

4. Make sure **Create report** and **Create assembly graph** is checked to get a quick summary of the assembly. Then choose the **Save** option and click **Next**. Save the output to a new folder, for example named "De Novo Assemble Long Reads". Depending on your setup, the tool will take a few minutes to run.

5. The genome of S. marcescens should now be assembled. Locate the output and open the assembly report. Here, you can see an overview of the assembly including nucleotide distribution and contig measurements. This dataset will assemble to seven contigs with a total size of approximately 5.8 mb (figure 4).

6. If you open the assembly graph result, you will notice that six graphs are available in the *Graph Components* drop-down menu in the **Side Panel**. The longest assembly graph of 5,518,867 bp consists of the two longest contigs (Utg 2134 and Utg2136), connected by an unresolved region (figure 5).

   Four of the other five graphs are circular and are of lengths that correlate with plasmids in the reference genome. The fifth graph (graph 2) is linear and does not match any of the

**2 Contig measurements**

| | |
|---|---|
| Contigs | 7 |
| Minimum | 3,161 |
| Maximum | 5,181,605 |
| Average | 831,328 |
| N50 | 5,181,605 |
| N90 | 267,404 |
| Total | 5,819,296 |

Figure 4: *The contig measurements after de novo assembly.*

plasmid lengths from the reference. In the bottom right corner of the view area and in the contig list, we can identify this assembly as contig Utg2140.

Figure 5: *The long assembly graph consists of two contigs separated by a short unresolved region.*

## Create a whole genome alignment (optional)

Since a reference is available for this data set, we can check the quality of the assembly. To do so, you need the Whole Genome Alignment plugin as described in the introduction.

1. Run **Create Whole Genome Alignment** from the Toolbox:

   **Whole Genome Alignment ( ) | Create Whole Genome Alignment ( )**

2. Select the imported reference genome and the contigs we just created (figure 6). Click **Next**.

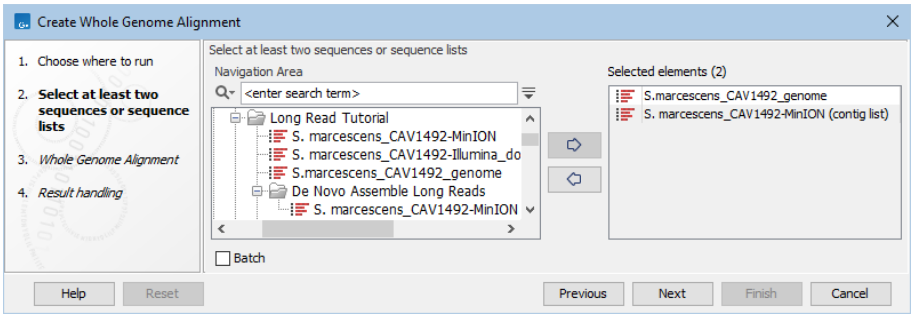


Figure 6: *Select the two genomes to align.*

3. Leave the default settings as shown (figure 7). Click **Next**.
4. Choose to **Save** the alignment, for example in a subfolder named "Whole Genome Alignment".

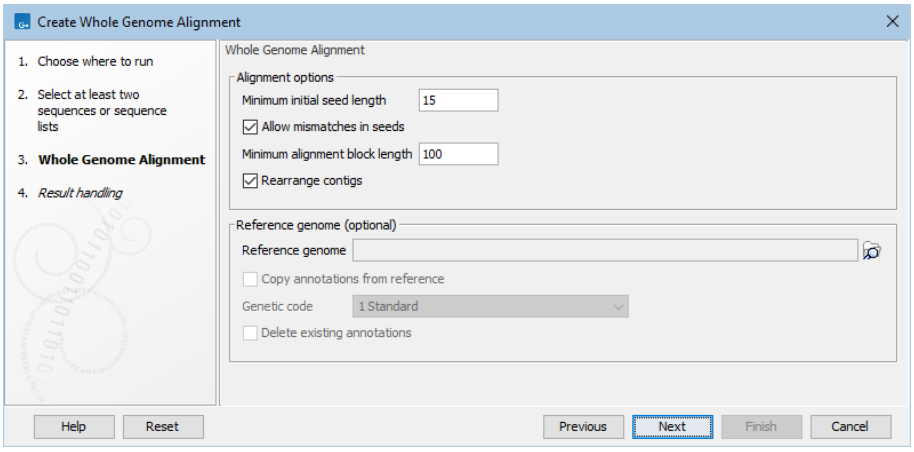


Figure 7: *Whole Genome Alignment options.*

5. Open the alignment to visualize the assembly against the reference (figure 8). Aligned regions are represented by pairs of same-colored blocks along the two sequences, and hovering over them reveals the sequence names. Clicking on a block will shift the alignment around the specific block.

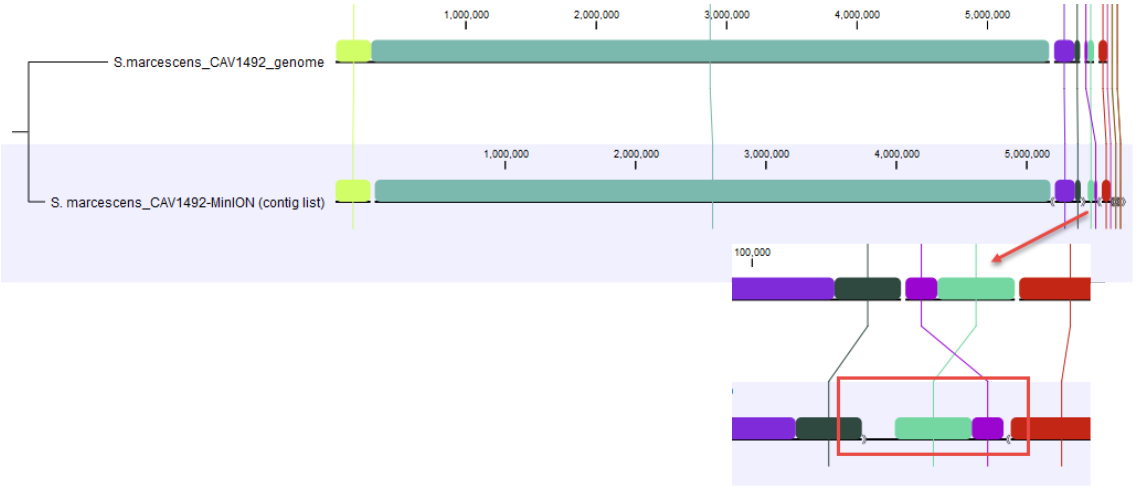


Figure 8: *The whole genome alignment of the reference and assembly, showing an unaligned end in contig Utg2140.*

Zooming in reveals generally well-aligned sequences, with blocks covering most reference/-contig sequences. Two notable unaligned regions are identified: one involving the linear contig Utg2140, aligning with plasmid CP011640.1 but extending too far, and the other with Utg2130 aligning to plasmid CP011638.1.

In reality, these regions do exhibit alignment but to regions that are duplicated in other parts of the genome. Utg2130 aligns well with plasmid CP011638.1 but also aligns with plasmid CP011640.1. Similarly, the excess length in Utg2140 aligns with plasmid CP011641.1, not CP011640.1.

We will not demonstrate it here, but for future analyses, unaligned events can be investigated by aligning sequences separately or extracting and mapping unaligned regions to the reference.

6. Lastly, we can calculate the Alignment Percentage (AP) and Average Nucleotide Identity (ANI). To do so run Create Average Nucleotide Identity Comparison from the Toolbox:

   **Whole Genome Alignment (  ) | Create Average Nucleotide Identity Comparison (  )**

7. Select the whole genome alignment we just created (figure 9) and click **Next**.

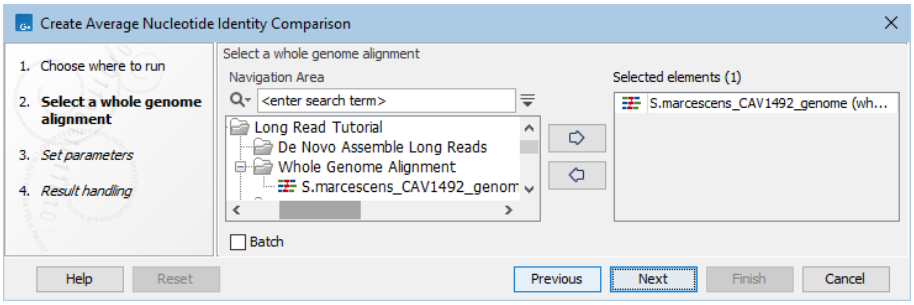


Figure 9: *Select the whole genome alignment.*

8. Leave the settings as default as shown in figure 10 and click **Next**. Choose to **Save** the comparison.

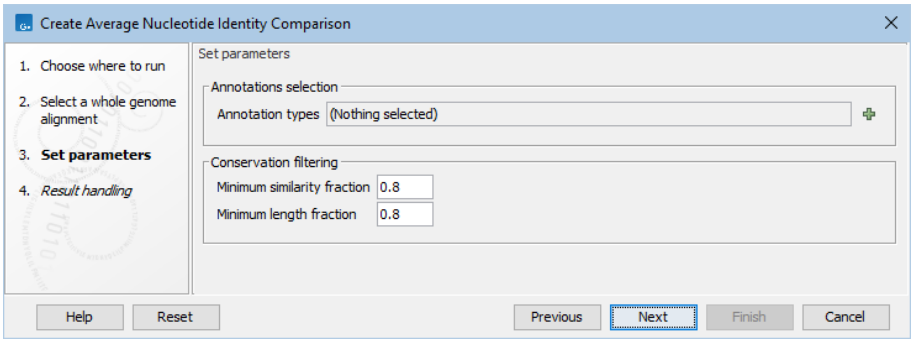


Figure 10: *Create Average Nucleotide Identity Comparison options.*

9. Open the results to see how well the assembly matches the reference. In this example, you should see an AP of 99.35% and an ANI of 99.82% (figure 11). This is quite high due to the **Polish with reads** option having been checked.

| | | 1 | 2 |
|---|---|---|---|
| S.marcescens_CAV1492_genome | 1 | | 99.82 |
| S. marcescens_CAV1492-MinION (contig list) | 2 | 99.35 | |

Figure 11: *The nucleotide identity comparison between reference and assembly.*

## Map long reads to reference

**Map Long Reads to Reference** enables you to map long reads to contigs or a reference. This is useful for visualizing coverage and to better understand your assembly even when the reference genome is not available. To show this, we will map the reads to the assembled contigs instead of the reference.

1. To start the tool, locate **Map Long Reads to Reference** in the Toolbox:

   **Long Read Support** ( ) | **Map Long Reads to Reference** ( )

2. Select the raw S. Marcescens MinION reads (figure 12) and click **Next**.

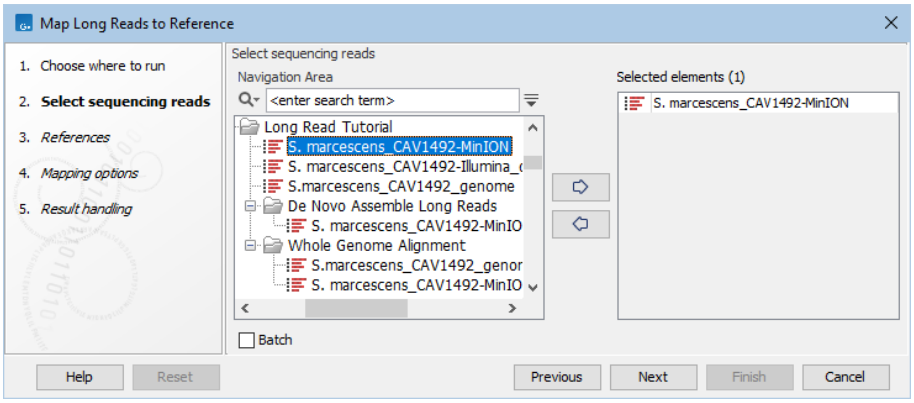


Figure 12: *Select the reads to map.*

3. In *References*, select your assembled contigs and click **OK** (figure 13). Then click **Next**.

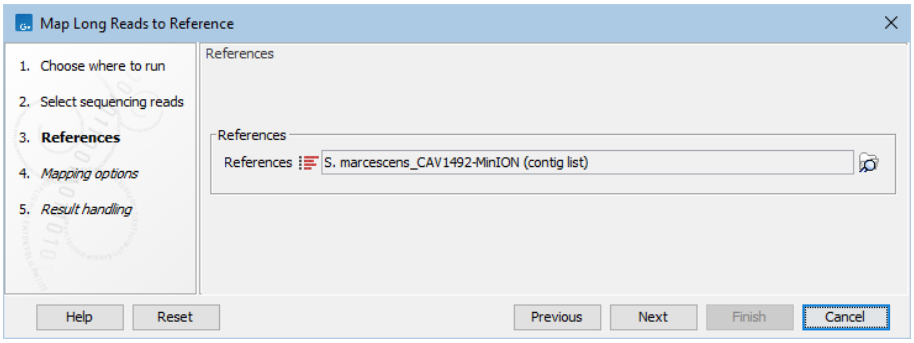


Figure 13: *As reference, select the contig list.*

4. Leave the default "Automatic" setting in the *Mapping options* (figure 14) and click **Next**.
5. Choose **Create reads track** and check **Create report**. Choose to **Save** in a subfolder, for example named "Map Long Reads to Reference".

You should have two outputs (reads track and report). Open the report and observe that >99% of the reads have mapped to the contigs.

Open the read mapping track to see that all contigs have coverage. However, the linear contig Utg2140 has almost no coverage on the 5' end (figure 15). If you did whole genome alignment, this is consistent with the unaligned end seen earlier i.e., the assembly of plasmid CP011640.1. In the following section, we will try to create a better assembly of this plasmid.

## Reassembling individual contigs

As we just saw, mapping raw reads to assembled contigs can be used to identify problematic regions in contigs. These can be a result of assembling multiple chromosomes/plasmids at once, so now we will try reassembling Utg2140 seperately.

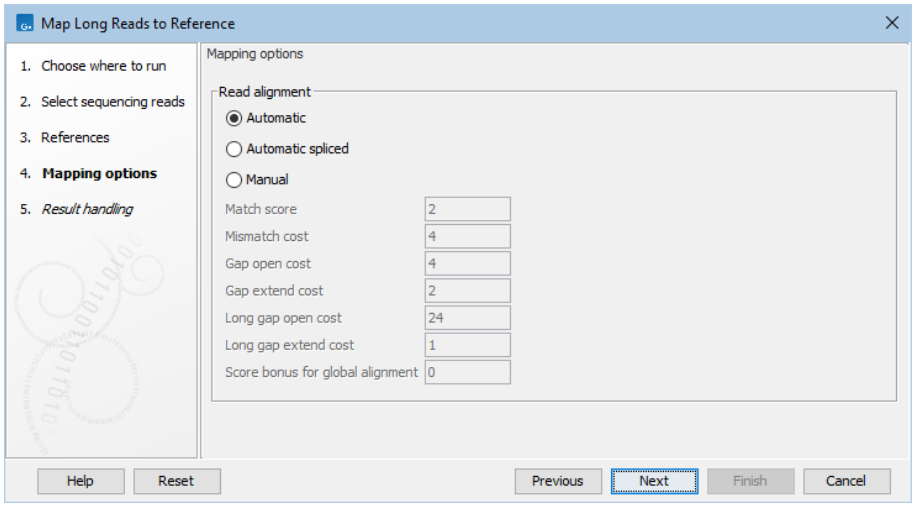


Figure 14: *Map Long Reads to Reference options.*

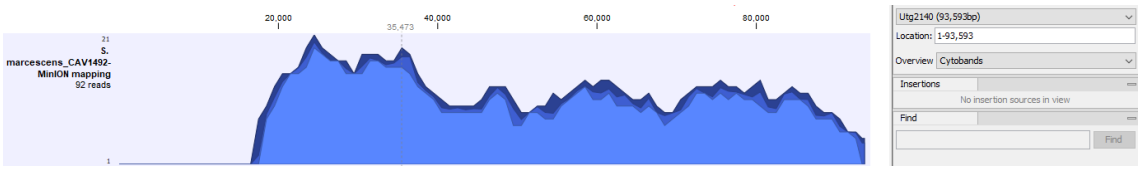


Figure 15: *Read coverage across the assembled contig Utg2140.*

First, we must isolate the reads from the target contig. Note: had we mapped the reads to the reference genome, we could also isolate reads from plasmid CP011640.1.

1. Start the tool **Convert from Tracks** from the Toolbox:

   **Utility Tools ( ) | Tracks ( ) | Track Conversion ( ) | Convert from Tracks ( )**

2. Select your read mapping (figure 16) and click **Next**.

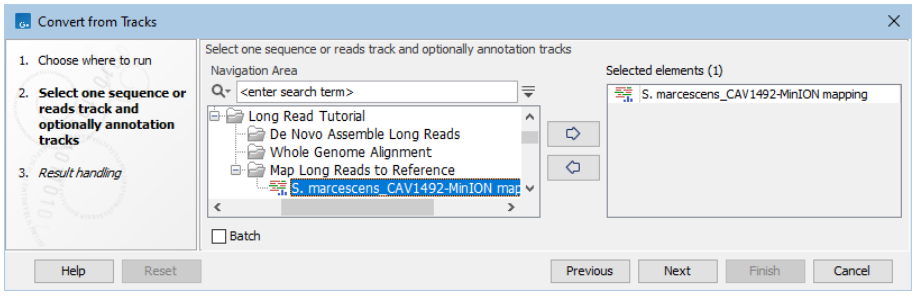


Figure 16: *Select the read mapping to convert.*

3. Save the output to your "Map Long Reads to Reference" location.
4. Open the read mapping list ( ) and select "Utg2140". It should have 92 reads mapped. Click **Extract Subset** as show in figure 17 and save the output to a new location, for example named "Reassemble contig Utg2140".
5. Start the tool **Extract Reads** from the Toolbox:

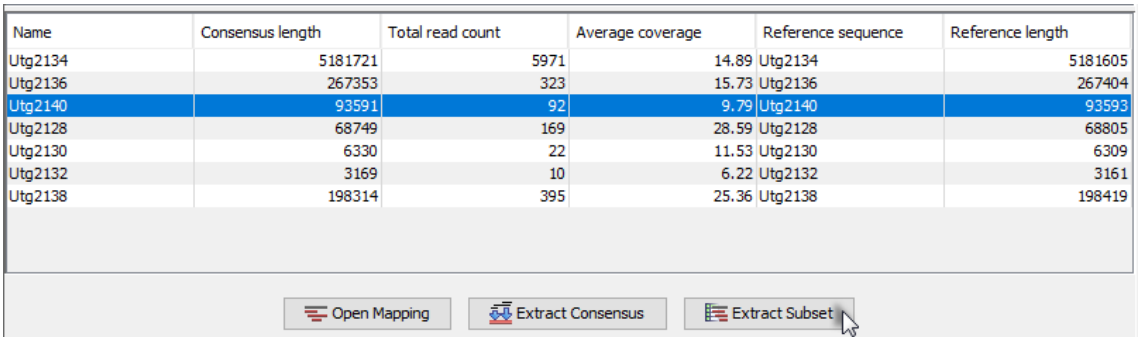


Figure 17: *Extract reads from the incorrectly assembled plasmid.*

**Utility Tools ( ) | Extract Reads ( )**

6. Select your extracted subset (figure 18) and click **Next**.

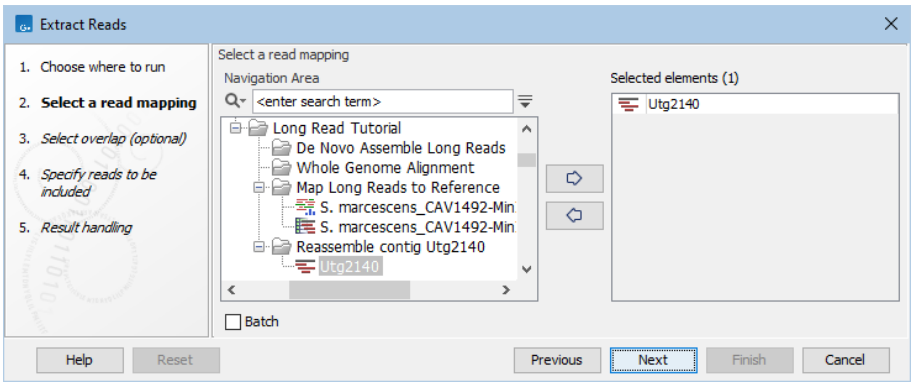


Figure 18: *Select the subset of the read mapping.*

7. Click through the wizard, leaving settings as default until you reach *Result handling*. Here, choose **Create sequence list(s)** and save the output to your ""Reassemble contig Utg2140" location.

8. Extracting reads does not retain read group information, and as the Long Read Support tools only take long reads, we must manually set the read group. Do this by opening the just created sequence list, opening the *Show Element info* ( ) view in the bottom left, and clicking "Edit" to the right of the **Read group** header. In the **Platform** drop-down menu choose "NANOPORE" and pres "OK". Finally, save the edited sequence list by pressing Ctrl + S (Cmd (⌘) + S on Mac)

9. Reads mapping to this contig have now been extracted. Run **De Novo Assemble Long Reads** on this subset just as we did previously by following the steps in De novo assemble long reads.

After reassembly, we should have a new contig Utg18, which is both closer to the expected length of plasmid CP011640.1 and correctly identified as circular. We will create a new contig list where contig Utg18 replaces Utg2140.

1. Start the tool **Create Sequence List** from the Toolbox:

**Utility Tools ( ) | Sequence Lists ( ) | Create Sequence List ( )**

2. Select the orginal assembled contigs and the new CP011640.1 contig (figure 18) and click **Next**.

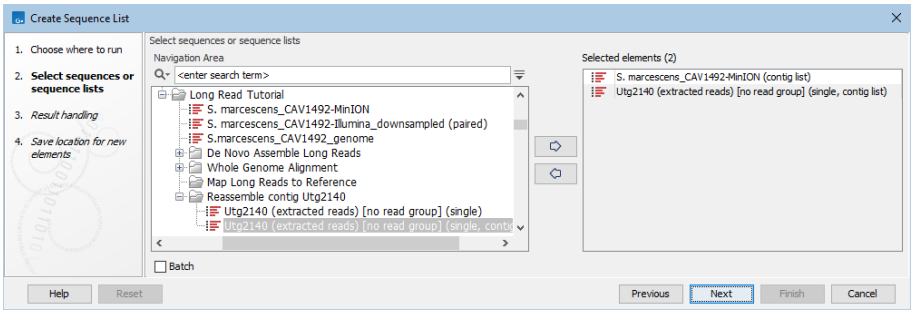


Figure 19: *Create a sequence list with the first round of assembled contigs and the new CP011640.1 contig.*

3. Choose to **Open** the output and click **Finish**. When the new sequence list opens, locate contig Utg2140, right-click the sequence, and choose **Delete Sequence** (figure 20). Then, save the edited sequence list by pressing Ctrl + S (Cmd (⌘) + S on Mac). Choose a name, for example "Updated assembly".

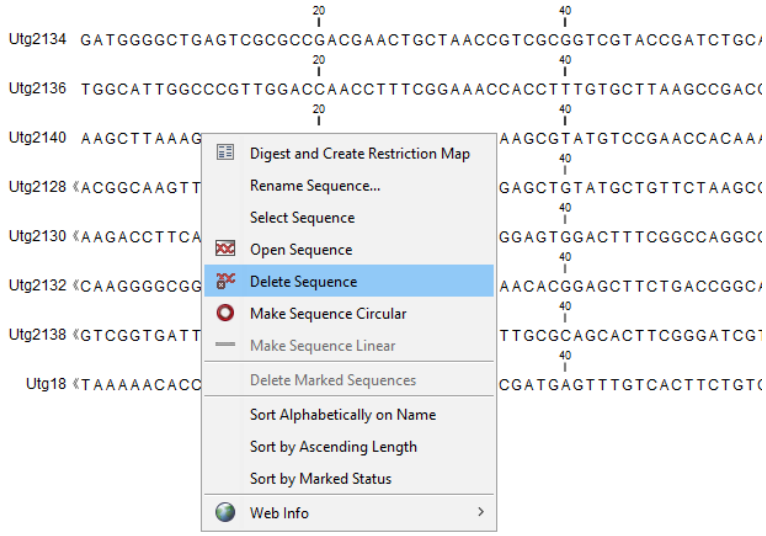


Figure 20: *Locate contig Utg2140 and delete it from the list before saving.*

4. Optionally, you can repeat the steps listed in Create a whole genome alignment (optional) using "Updated assembly" as input. Observe that the AP value has improved to 99.53% (figure 21).

In this case it was enough to isolate and reassemble the raw reads. In other cases, it may be useful to correct the isolated reads first using **Correct Long Reads** http://resources.qiagenbioinformatics.com/manuals/longreadsupport/current/index.php?manual=Correct_Long_Reads.html. However, it should be stressed that **Correct Long Reads** should generally not be run before **De Novo Assemble Long Reads** - only as a way to resolve poor assemblies due to high error rate or low coverage.

| | | 1 | 2 |
|---|---|---|---|
| S.marcescens_CAV1492_genome | 1 | | 99.81 |
| Updated assembly | 2 | 99.53 | |

Figure 21: *The nucleotide identity comparison between reference and updated assembly.*

### Polish assembly with reads

**Polish with Reads** makes it possible to polish de novo assemblies or raw long error-prone reads. As a final step, we will attempt to improve this assembly by using Illumina reads to polish our assembly.

Reads should be quality-trimmed and have adapters removed before being used to polish assemblies or error-prone reads. This can be done with the Trim Reads tool (see `http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Trim_Reads.html`). As with the Oxford Nanopore reads, the Illumina reads we imported in the beginning of the tutorial have already been trimmed, so we will skip this step.

1. To start the tool, locate **Polish with Reads** in the Toolbox:

   **Long Read Support** (  ) | **Polish with Reads** (  )

2. Select the "Updated assembly" we created in the previous steps and click **Next** (figure 22).

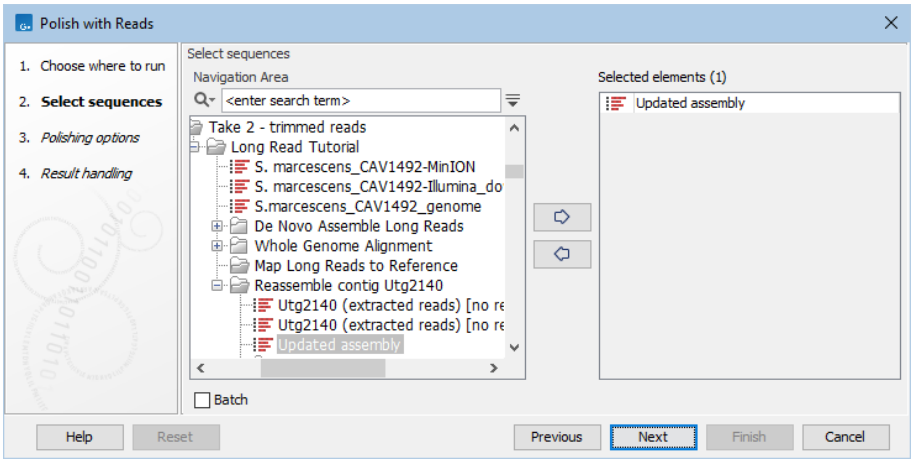


Figure 22: *Select the contigs to polish.*

3. Click on (  ) and specify the Illumina reads as input and click **OK**. Leave the other settings as default (figure 23).

4. Check **Create report** and choose to **Save** in a subfolder, for example named "Polish with Reads". Depending on your setup, the tool will take a few minutes to run.

5. In the polishing report, you can see the overview of the assembly after polishing. In this data set, there are no significant changes to the number of contigs and assembly size although incorrect bases have been polished.

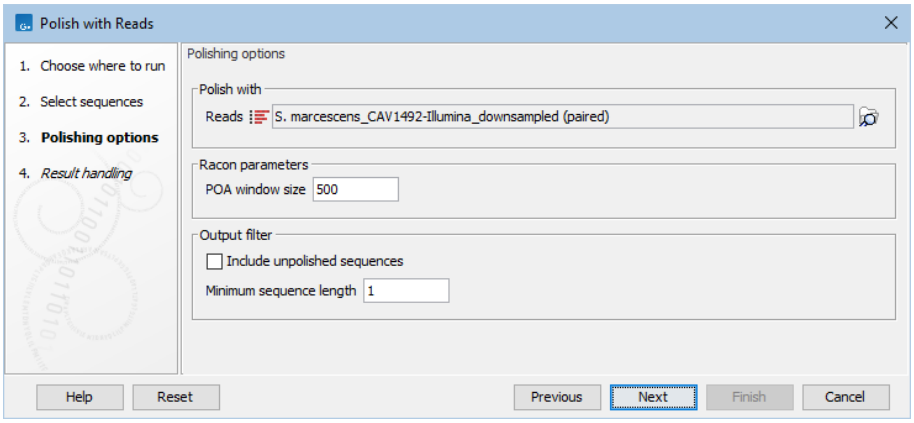


Figure 23: *Select the paired-end Illumina reads.*

6. (Optional) Repeat the steps listed in Create a whole genome alignment (optional) using the polished contigs as input. Observe that the AP and ANI values have improved to 99.80% and 99.95%, respectively (figure 24).

| | | 1 | 2 |
|---|---|---|---|
| S.marcescens_CAV1492_genome | 1 | | 99.95 |
| Updated assembly (polished) | 2 | 99.80 | |

Figure 24: *The nucleotide identity comparison between reference and polished assembly.*

## Summary

Using the Long Read Support plugin, we assembled a microbial genome, with four out of five plasmids fully resolved. We evaluated the quality of our assembly by mapping reads back to the contigs. We resolved the fifth plasmid by isolating raw reads from the unresolved plasmid and reassembling. Finally, we polished our assembly and obtained an Alignment Percentage of 99.80% and Average Nucleotide Identity of 99.95%.

# Bibliography

[George et al., 2017] George, S., Pankhurst, L., Hubbard, A., Votintseva, A., Stoesser, N., Sheppard, A. E., Mathers, A., Norris, R., Navickaite, I., Eaton, C., et al. (2017). Resolving plasmid structures in enterobacteriaceae using the minion nanopore sequencer: assessment of minion and minion/illumina hybrid data assembly approaches. *Microbial genomics*, 3(8).