



Tutorial

De Novo Assembly Using Long Reads and Short Read Polishing

December 3, 2024

— Sample to Insight —

De Novo Assembly Using Long Reads and Short Read Polishing

This tutorial takes you through a de novo assembly of long next-generation sequencing reads.

Here, we focus on error-prone long 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 contigs.

In this tutorial you will:

- Import data required for the analysis.
- Perform de novo assembly of long, error-prone reads from a microbial-sized genome.
- Map long reads to a reference and visualize the assembly.
- Improve 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* 25.0 or higher. If you are working with a different version, some elements in this tutorial (such as contig naming) may differ slightly.

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 <https://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 Assembly Tutorial".
4. To import the Oxford Nanopore MinION reads, go to:
File | Import | Oxford Nanopore. . .
Leave settings as default. Click **Add files** and select the following file:
S. marcescens_CAV1492-MinION.fastq
Click **Open** and then **Next**. Save to the location you created.
5. To import the Illumina reads, go to:
File | Import | Illumina. . .
Leave settings as default, including leaving **Paired reads** checked under *General options*. Click **Add files** and select the following files:
S. marcescens_CAV1492-Illumina_downsampled_R1.fastq
S. marcescens_CAV1492-Illumina_downsampled_R2.fastq
Click **Open** and then **Next**. Save to the same location as the MinION reads.

6. Lastly, to import the reference, go to:

File | Import | Standard Import. . .

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

`S.marcescens_CAV1492_genome.fa`

Click **Open** and then **Next**. Save to the same location.

Your folder should now look like that in figure 1.

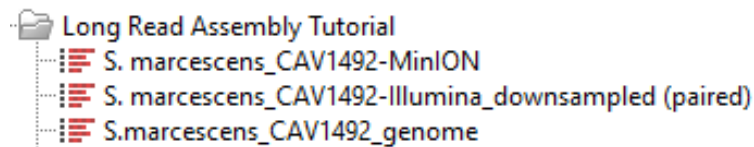


Figure 1: Tutorial folder with imported data.

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].

You will be assembling 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 over 12,000 bp, and a set of Illumina reads sequenced from the same strain. The Illumina reads have been downsampled to lower the runtime of analysis in this tutorial.

The reference genome included in the dataset was made from deep coverage PacBio and paired-end sequencing data. It is available from https://ncbi.nlm.nih.gov/datasets/genome/GCA_001022215.1/.

De novo assemble long reads

The De Novo Assemble Long Reads tool facilitates the assembly of contigs from long reads.

It is best practice to trim 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 https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Trim_Reads.html). The Oxford Nanopore reads in this tutorial have already been trimmed for adapters, so you can skip this step.

1. To launch the De Novo Assemble Long Reads tool, use the Quick Launch tool (🔍) in the toolbar, or go to:

Tools | De Novo Sequencing (📁) | De Novo Assemble Long Reads (🔧)

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

3. In the *De novo options* dialog (figure 3), define the assembly settings:

- **Polish with reads** should be checked. This will make the tool run two rounds of polishing after assembling initial contigs.

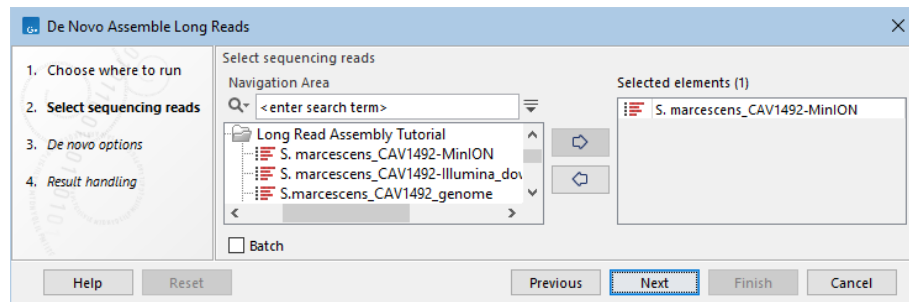


Figure 2: Selecting the MinION reads.

- Set **Minimum contig length = 1**
- Keep the remaining settings as default.

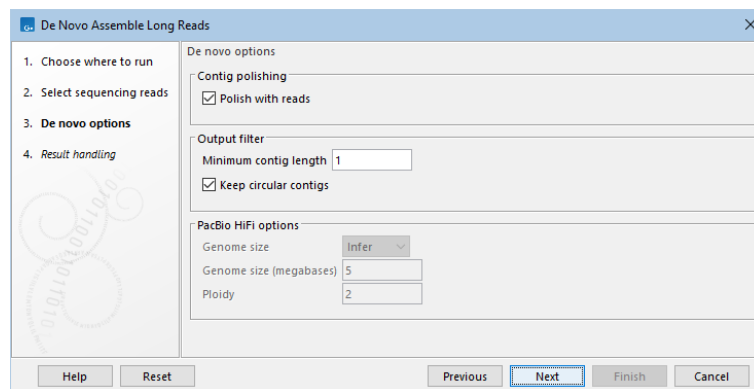


Figure 3: De Novo Assemble Long Reads options.

4. Click **Next**
5. In the *Results handling* step, select **Create report** and **Create assembly graph** and choose to save the output to a new folder, for example named "De Novo Assemble Long Reads".
6. Click on **Finish**. Depending on your setup, the tool will take a few minutes to run.

The genome of *S. marcescens* should now be assembled. The contigs list contains seven contigs.

Open the assembly report. This contains information about the contigs, including nucleotide and length distribution. You will see that the assembly produced seven contigs with a total size of approximately 5.8 Mb (figure 4).

Open the assembly graph result. From the *Graph Components* drop-down menu in the **Side Panel** you find six assembly graph components. By hovering your cursor over the individual sections of graph, you can see that the longest assembly graph component of approximately 5.5 Mb consists of the two longest contigs (Utg2126 and Utg2128), connected by an unresolved region (figure 5).

The remaining five assembly graph components are circular. Their lengths correspond to those of the plasmids in the reference genome.

1 Nucleotide distribution

Nucleotide	Count	Frequency (%)
Adenine (A)	1,193,924	20.59
Cytosine (C)	1,698,311	29.29
Guanine (G)	1,702,727	29.37
Thymine (T)	1,202,405	20.74

2 Contig measurements

Contigs	7
Minimum	3,160
Maximum	5,180,786
Average	828,195
N50	5,180,786
N90	267,361
Total	5,797,367

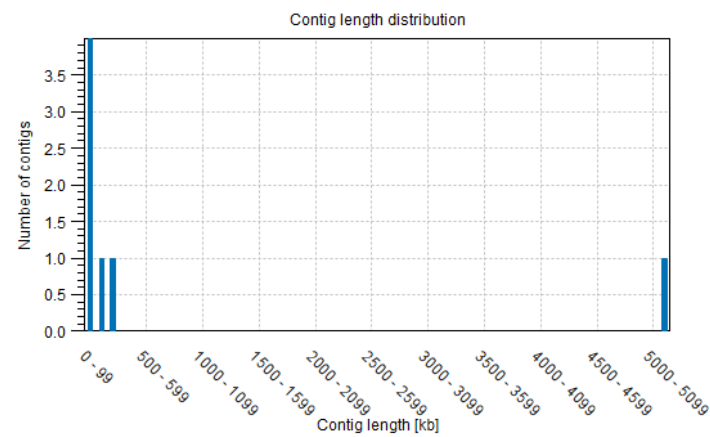


Figure 4: The contig measurements after de novo assembly.

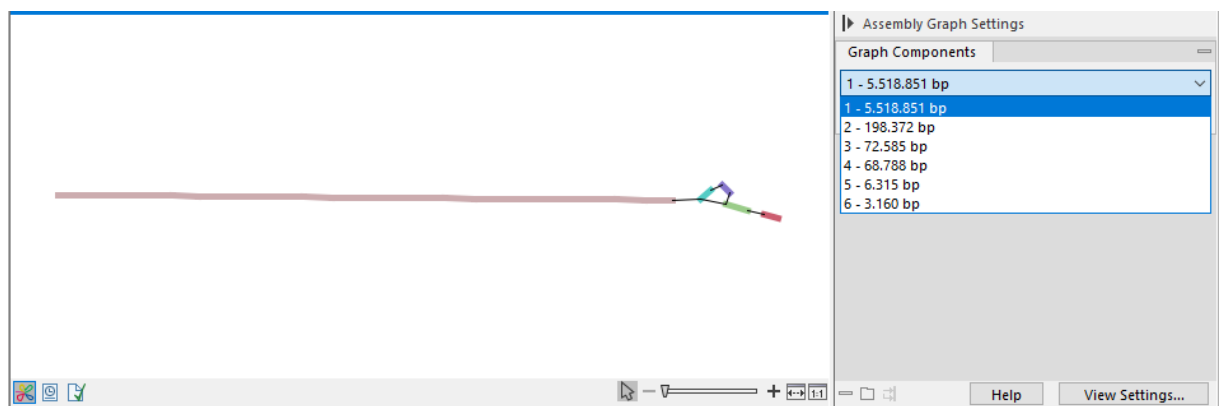


Figure 5: The long assembly graph consists of two contigs separated by an unresolved region.

Create a whole genome alignment (optional)

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

1. Launch **Create Whole Genome Alignment** by going to:

Tools | Whole Genome Alignment () | Create Whole Genome Alignment ()

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

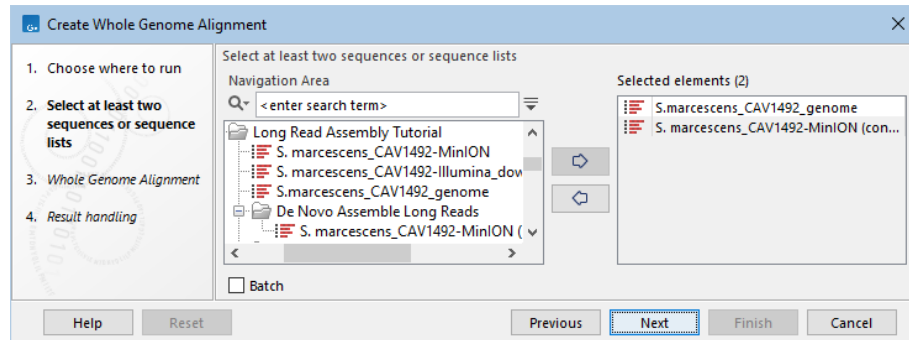


Figure 6: Select the reference genome and the list of contigs.

3. Leave the default settings as shown (figure 7). Click **Next**.

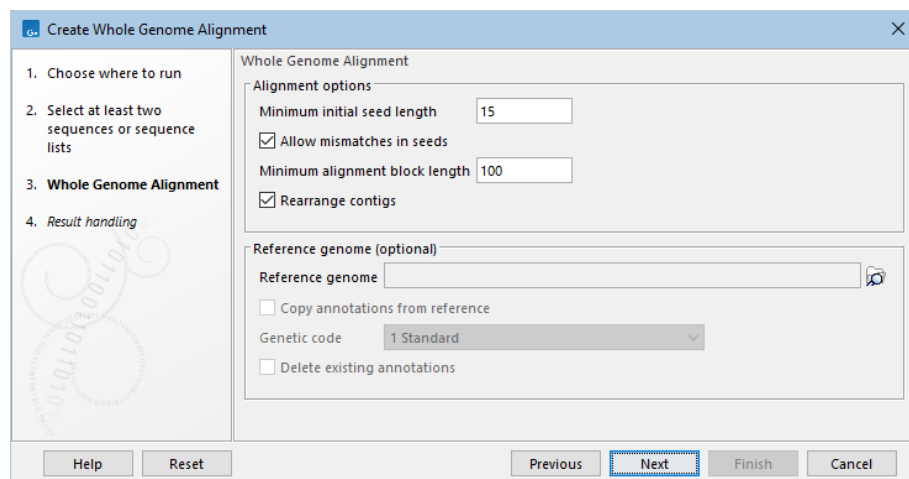


Figure 7: Whole Genome Alignment options.

4. Choose to **Save** the alignment, for example in a subfolder named "Whole Genome Alignment".
5. Open the alignment to visualize how the assembly aligns with the reference (figure 8). Aligned regions are represented by pairs of same-colored blocks along the two sequences. Hover your cursor over a block to see the sequence name, and click on a block to shift the alignment to center around this block.
In general, the sequences align well as indicated by blocks covering most of the alignment. However, if you zoom in at the end of the blocks, you will see pieces of unaligned sequence. For example, longer stretches of unaligned nucleotides are found at both ends of the large block that covers the major part of the alignment between reference sequence CP011642.1 and contig Utg2126.
6. Lastly, you can calculate the Alignment Percentage (AP) and Average Nucleotide Identity (ANI). To do so, run Create Average Nucleotide Identity Comparison by going to:

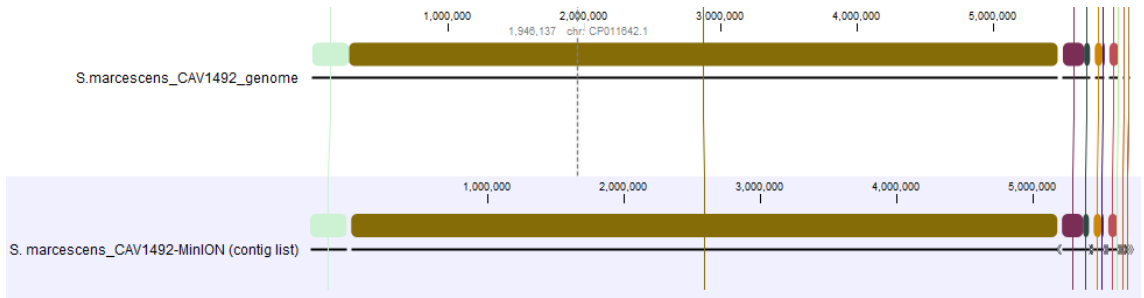




Figure 8: The whole genome alignment of the reference and assembly. Unaligned stretches of sequence can be found at the ends of the colored blocks.

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

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

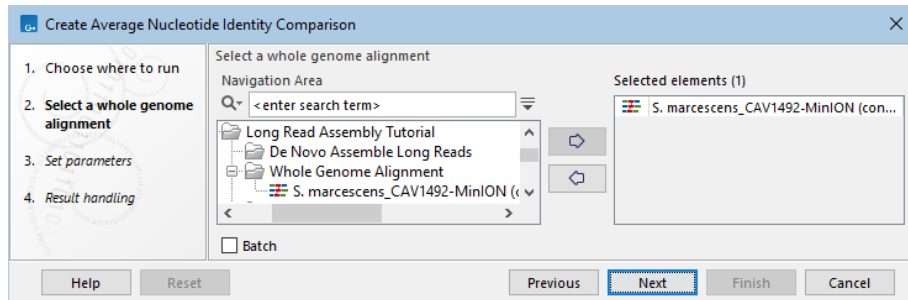


Figure 9: Select the whole genome alignment.

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

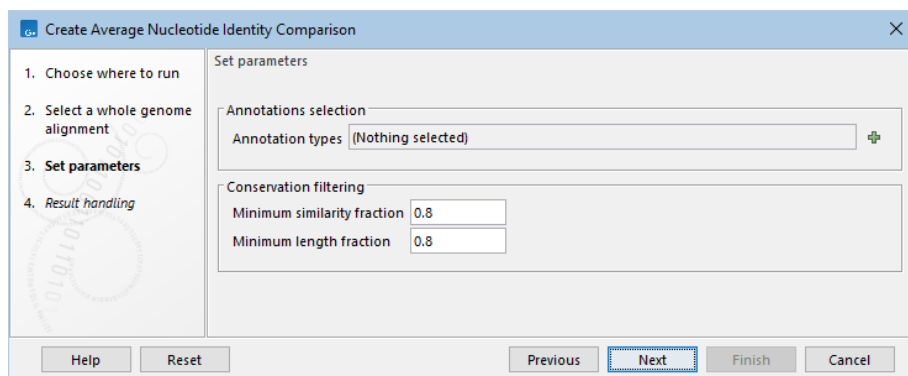


Figure 10: Create Average Nucleotide Identity Comparison options.

9. Open the results to see how well the assembly matches the reference. In this example, you should see an AP of 99.59% and an ANI of 99.84% (figure 11). This is quite high due to the **Polish with reads** option having been checked.

	1	2
S.marcescens_CAV1492_genome	1	99.84
S. marcescens_CAV1492-MinION (contig list)	2	99.59

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

Map long reads to reference

The Map Long Reads to Reference tool enables you to map long reads to contigs or to a reference. Mapping reads to your contigs can be useful for visualizing coverage and to better understand your assembly when a reference genome is not available. To show this, below we will map the reads to the assembled contigs.

1. To run **Map Long Reads to Reference**, go to:

Tools | Resequencing Analysis (📁) | **Map Long Reads to Reference** (🔍)

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

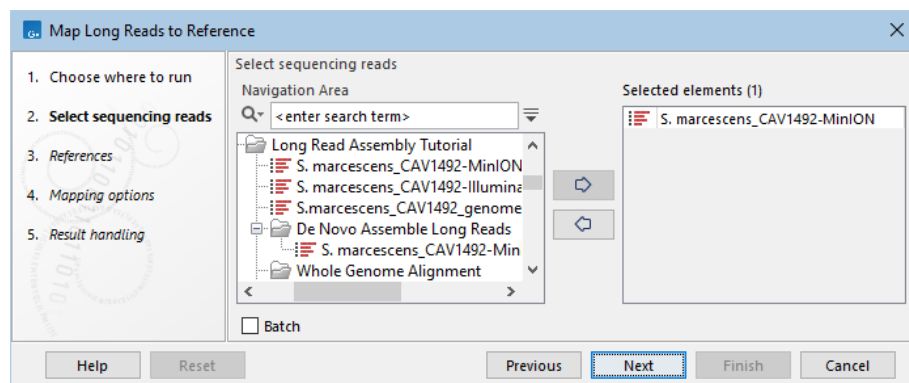


Figure 12: Select the reads to map.

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

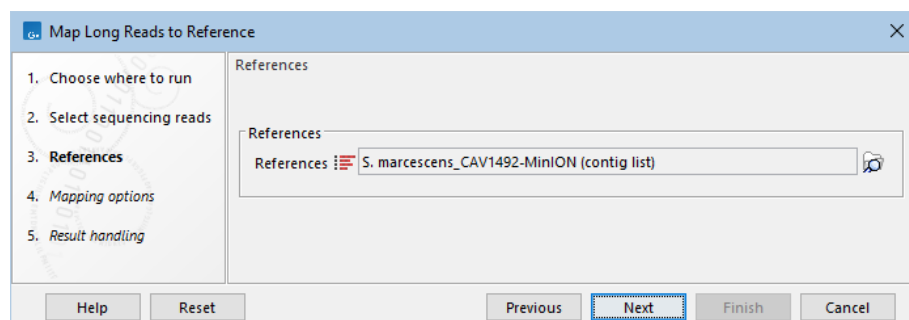


Figure 13: As reference, select the contig list.

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".

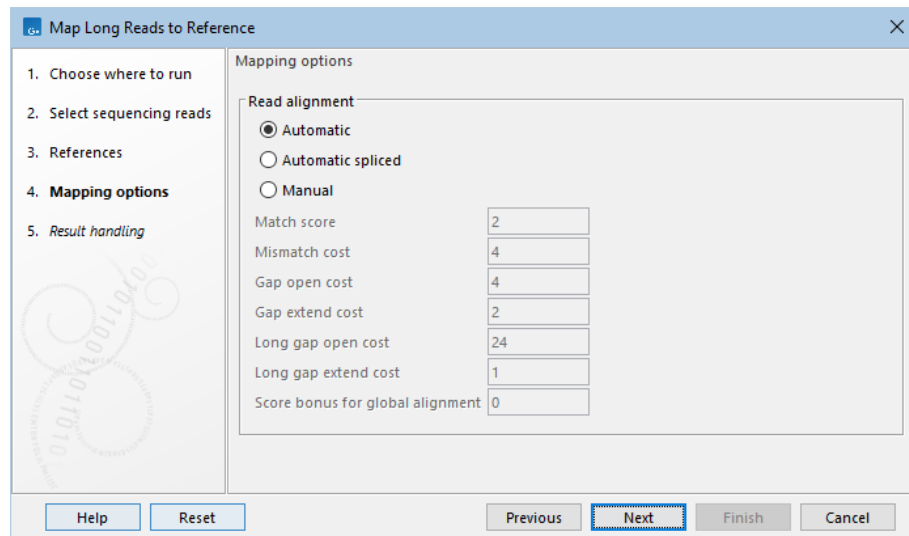


Figure 14: *Map Long Reads to Reference options.*

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 reads track to see that all contigs have coverage. For problematic assemblies, you will often see that parts of the contigs have no or very limited coverage, but that is not the case here (figure 15).

Tip: If you zoom to inspect the read mapping further, as MinION reads have many errors, a more stringent display of the read mapping gives a better viewing experience. To achieve this, change the settings in the Side Panel. Under Track layout | Reads track | Hide insertions below (%) change the value to e.g., 50.

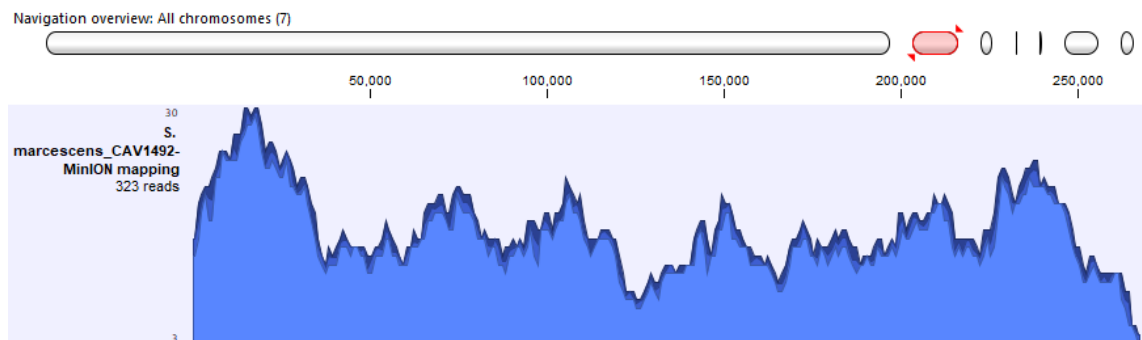


Figure 15: *Read coverage across the assembled contig Utg2128.*

Polish assembly with reads

The Polish Contigs with Reads tool makes it possible to polish de novo assemblies with high-quality reads. As a final step, you will attempt to improve the assembly by using Illumina reads to polish the contigs.

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 <https://resources.>

qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Trim_Reads.html). As with the Oxford Nanopore reads, the Illumina reads you imported in the beginning of the tutorial have already been trimmed, so you can skip this step.

1. To run **Polish Contigs with Reads**, go to:

Tools | De Novo Sequencing (🗄️) | Polish Contigs with Reads (🔍)

2. Select the "Updated assembly" you created in the previous steps and click **Next** (figure 16).

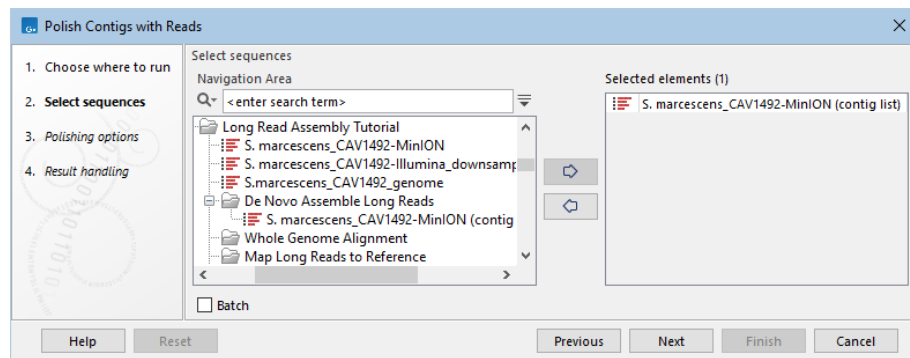


Figure 16: Select the contigs to polish.

3. Click on (🔍) and specify the Illumina reads as input and click **OK**. Leave the other settings as default (figure 17).

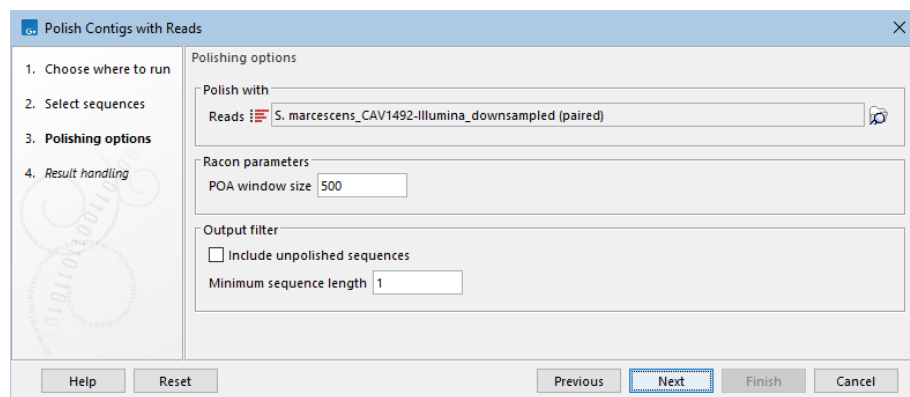


Figure 17: Select the paired-end Illumina reads.

4. Check **Create report** and choose to **Save** in a subfolder, for example named "Polish Contigs with Reads". Depending on your setup, the tool will take a few minutes to run.
5. In the polishing report, you can view a summary 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.
6. (Optional) Repeat the steps listed in **Create a whole genome alignment (optional)** using the polished contigs as input. From the nucleotide identity comparison you will see that the AP and ANI values increased to 99.84% and 99.95%, respectively (figure 18) indicating that polishing with Illumina reads improved the assembly.

		1	2
S.marcescens_CAV1492_genome	1		99.95
S. marcescens_CAV1492-MinION (contig list, polished)	2	99.84	

Figure 18: *The nucleotide identity comparison between reference and polished assembly.*

Summary

Using tools for handling long next-generation sequencing reads, you assembled a microbial genome including plasmids. You evaluated the quality of the assembly by mapping reads back to the contigs and finally, you polished the assembly and obtained an Alignment Percentage of 99.84% 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).