Tutorial

De Novo Assembly Using Long Reads and Short Read Polishing

July 9, 2021

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 (beta) plugin.

The Long Read Support (beta) plugin is a collection of tools developed for working with long, error-prone reads such as those produced by the single-molecule sequencing technologies of Pacific Biosciences or Oxford Nanopore Technologies. It is based on the open source components minimap2 [Li, 2018], miniasm [Li, 2016] and racon [Vaser et al., 2017]. The tutorial covers the following:

- Import data required for the analysis.
- De novo assemble a microbial sized genome using long, error-prone reads.
- Improve a de novo assembly from long reads by polishing with short, high-quality reads.
- Map long reads to a reference and visualizing an assembly.
- Correct raw long reads for further analysis.

Prerequisites For this tutorial, you must be working with *CLC Genomics Workbench* 20.0 or higher and have installed the Long Read Support (beta) plugin.

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

Optional: For additional evaluation steps, you will need the Whole Genome Alignment plugin. Plugins can be downloaded from: https://digitalinsights.qiagen.com/products-overview/plugins/

Download and import data

1. Download the sample data from our web site http://resources.qiagenbioinformatics.com/testdata/CAV1492_MinION_and_Illumina_example_data.zip and unzip it.
2. Open the *CLC Genomics Workbench*.
3. Create a new folder for the project. For example, titled "Long Read Tutorial".
4. Import the Oxford Nanopore MinION reads. To do so, select **File** | **Import** | **Oxford Nanopore...**. Click **Add files** and add the file

   `S. Marcescens_CAV1492-MinION.fastq`

   Leave other settings as default. Click **Next** and **Save** it to the location you created.
5. Open **Import** then click **Illumina**. Make sure to check **Paired reads** 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 `S.marcescens_CAV1492_genome.fa`

   using **Standard Import**. Keep the default option **Automatic import** checked. Click **Next** and **Save** it to the same location.

Your folder should look like figure 1.

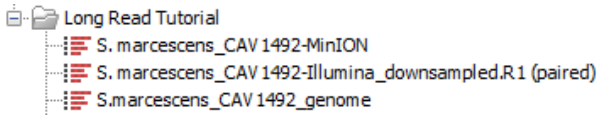

Figure 1: *Imported data list*

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 reads and Illumina short read sequencing data from 8 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 5 plasmids. The sequencing data contains 7,039 MinION reads with an average read length of 12,660bp. We have also imported a set of Illumina reads sequenced from the same strain. These reads have been downsampled from 3 million to 300,000 paired reads 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 (beta)** tool makes it possible to create de novo assemblies from long reads for microbial-sized genomes.

To start the tool, locate **De Novo Assemble Long Reads (beta)** (Figure 2) in Toolbox or search and run using Launch.

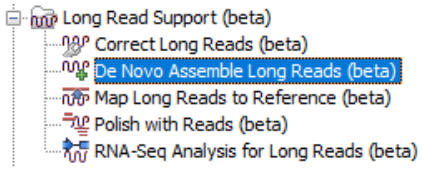

Figure 2: *Long Read Support (beta) tools*

1. Select the imported MinION reads (Figure 3) and click **Next**.
2. Run the tool using the default settings as shown in Figure 4. Make sure the **Polish with reads** option is checked. This will make the tool run two rounds of polishing after assembling. Click **Next**.
3. Make sure **Create report** is checked to get a quick overview of your assembly. Then choose to **Save** the assembly in a location of your choosing; we recommend making a subfolder

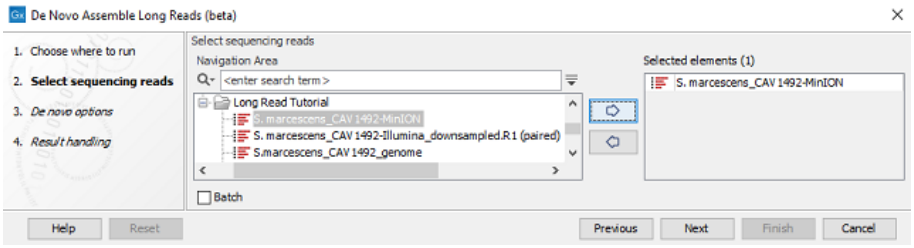

Figure 3: *Selecting the MinION reads*

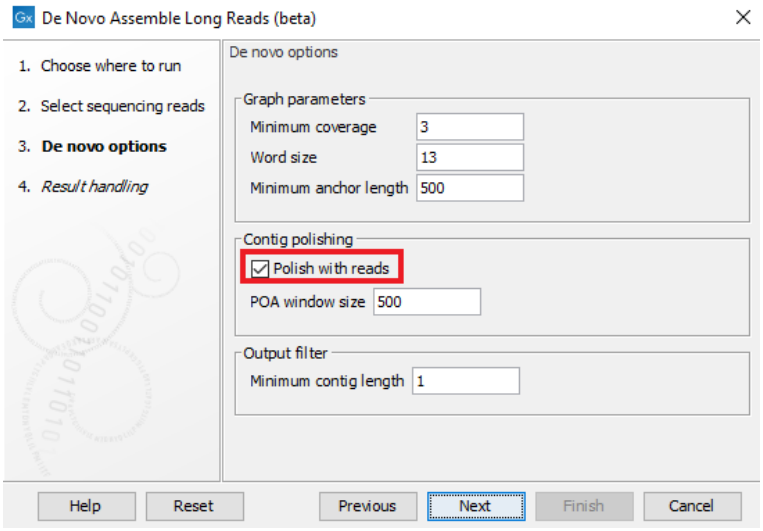

Figure 4: *De Novo Assemble Long Reads options*

titled "De Novo Assemble Long Reads". Wait for the tool to run. Depending on your setup, this will take a few minutes.

4. The genome of S. marcescens should now be assembled. You can locate the output and open the Assembly report. Here, you can see an overview of your assembly including nucleotide distribution and contig measurements. This dataset will assemble to 6 contigs with a total size of approximately 5.8mb (Figure 5).

2 Contig measurements

Contigs	6
Minimum	3,185
Maximum	5,471,189
Average	970,888
N50	5,471,189
N90	5,471,189
Total	5,825,326

Figure 5: *The contigs measurement after de novo assembly*

Open the contig list. In the assembly, you can see that 5 out of 6 contigs are circular as indicated by << next to their name. One of the contigs is linear and does not match any of

the plasmid lengths from the reference. We will return to this plasmid named CP011638 later.

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

1. Run **Create Whole Genome Alignment** from **Whole Genome Alignment** (Figure 6) in the Toolbox or use Launch to search and run.

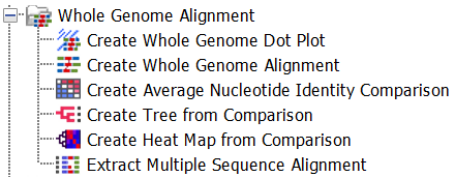

Figure 6: *Whole Genome Alignment tools*

2. Select the imported reference genome and the contigs you just created (Figure 7). Click **Next**.

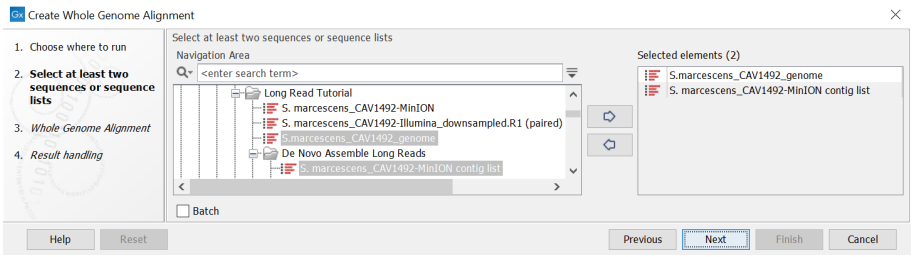

Figure 7: *Select the two genomes to align*

3. Leave settings on default as shown (Figure 8).
4. Choose to **Save** the alignment in your **De Novo Assemble Long Reads** location. Open the alignment to visualize your assembly against the reference (Figure 9).

   There is an overall good alignment except one contig where only about half aligns. If you hover over this contig, you can see that this is the same contig we identified as incorrectly being linear in the assembly.

5. Lastly, you can calculate the Alignment Percentage (AP) and Average Nucleotide Identity (ANI). To do so run **Create Average Nucleotide Identity Comparison** from Whole Genome Alignment in the Toolbox or use Launch to search and run.
6. Select the whole genome alignment you created using **Create Whole Genome Alignment** (Figure 10) and click **Next**.
7. Leave the settings on default as shown in figure 11 and click **Next**.
8. Choose to **Save** the comparison in your De Novo Assemble Long Reads location. Open it to see how well your assembly matches the reference. In this example, you should see an

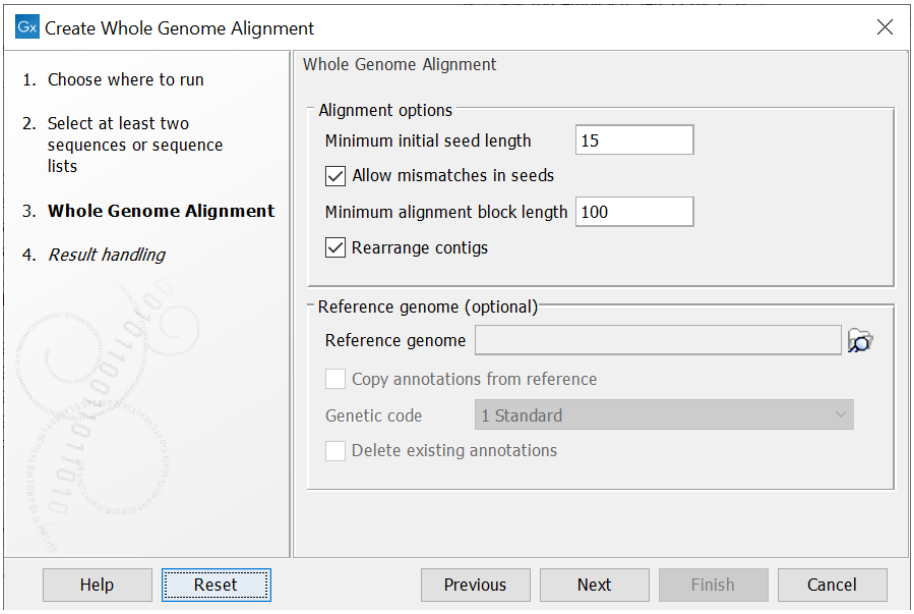

Figure 8: *Whole Genome Alignment options*

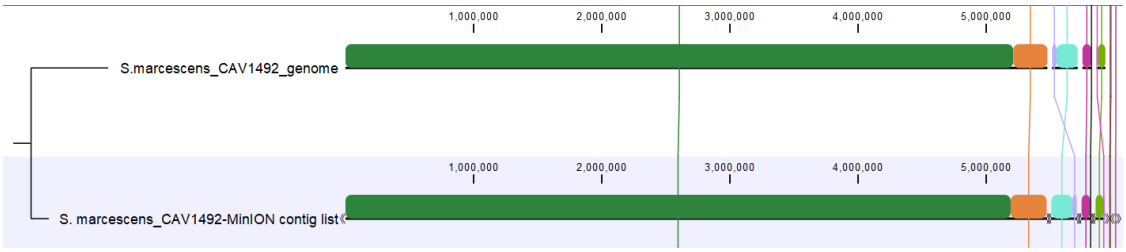

Figure 9: *The whole genome alignment of the reference and assembly*

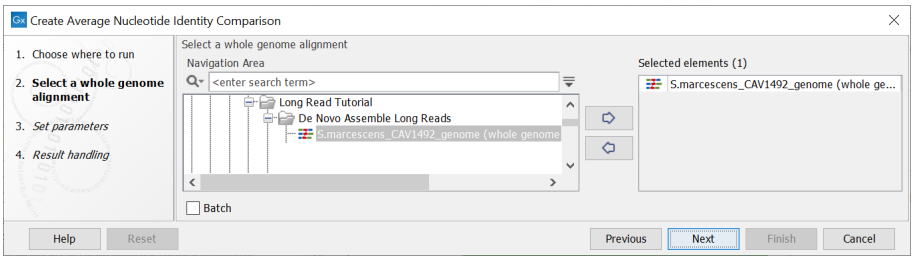

Figure 10: *Select the whole genome alignment*

AP of 99.61 and ANI of 99.92 (Figure 12). This is quite high due to the **Polish with reads** option having been checked. In the next section, you will attempt to improve this assembly by using Illumina reads to polish your assembly.

Polish with Reads

Polish with Reads (beta) makes it possible to polish de novo assemblies or raw long error-prone reads for microbial-sized genomes.

To start the tool, locate **Polish with Reads (beta)** in Long Read Support (beta) in the Toolbox or

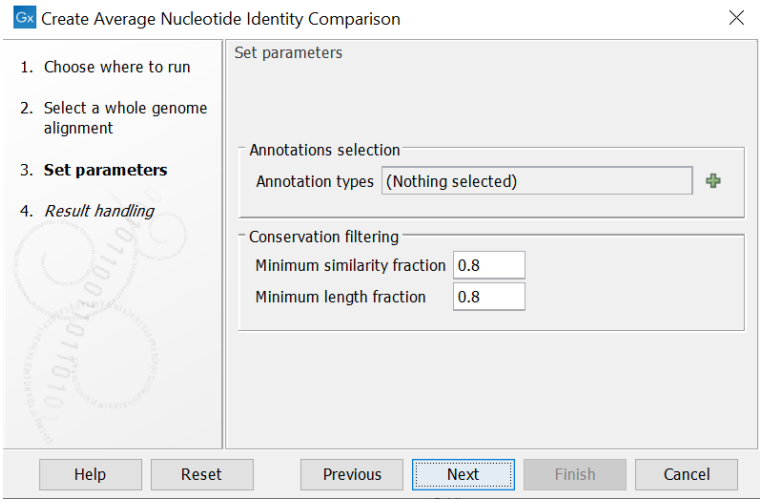

Figure 11: *Create Average Nucleotide Identity Comparison options*

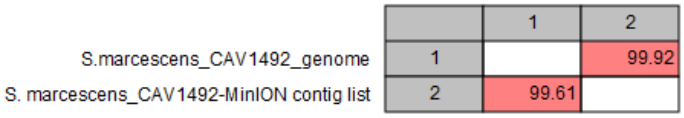

Figure 12: *The nucleotide identity comparison between reference and assembly*

search and run using Launch.

1. Select the assembly you created in the previous steps and click **Next** (Figure 13).

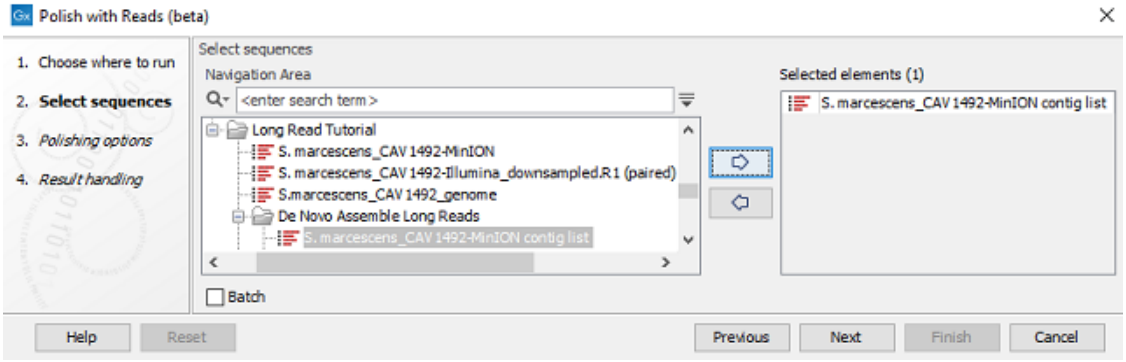

Figure 13: *Select the contigs to polish*

2. Click on ( ) and specify the Illumina reads as input and click **OK**. Leave the other settings on default (Figure 14).
3. Choose to **Save** in your **De Novo Assemble Long Reads** folder and check **Create report**. The tool will now run. Depending on your setup, this will take around 10 minutes. 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.
4. (Optional) Repeat the steps listed in "Create a Whole Genome Alignment" using the polished

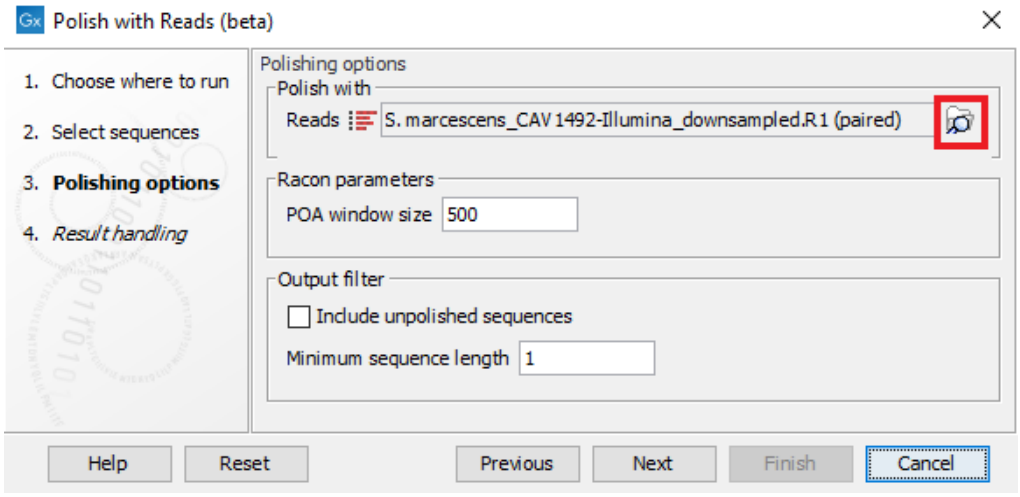

Figure 14: *Select the paired-end reads*

contigs as input. Observe that the AP and ANI values have improved (Figure 15).

		1	2
S.marcescens_CAV1492_genome	1		99.93
S. marcescens_CAV1492-MinION contig list (polished)-1	2	99.81	

Figure 15: *The nucleotide identity comparison between reference and the polished assembly*

Map Long Reads to Reference

Map Long Reads to Reference (beta) enables you to map long reads to contigs or a reference. This is useful for visualizing coverage and to better understanding your assembly.

1. To start the tool, locate **Map Long Reads to Reference (beta)** in Long Read Support (beta) in the Toolbox or search and run using Launch.
2. Select the raw S. Marcescens MinION reads (Figure 16) and click **Next**.

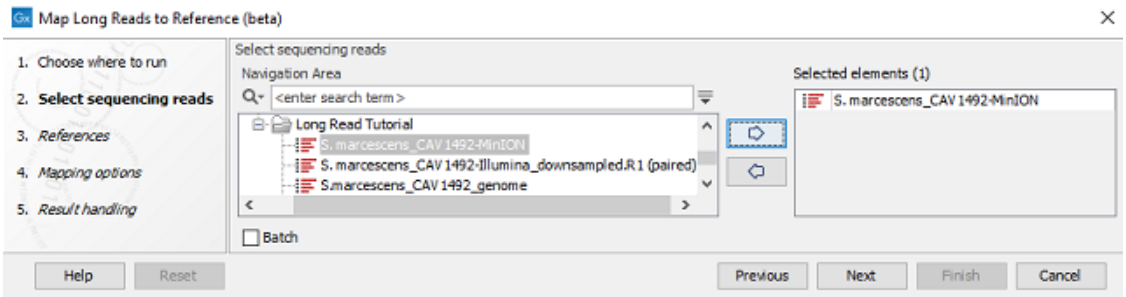

Figure 16: *Select the reads to map*

3. In **References**, select the S. marcescens reference genome and click **OK** (Figure 17). Then click **Next**.
4. Leave the Mapping options as default (Figure 18) and click **Next**.

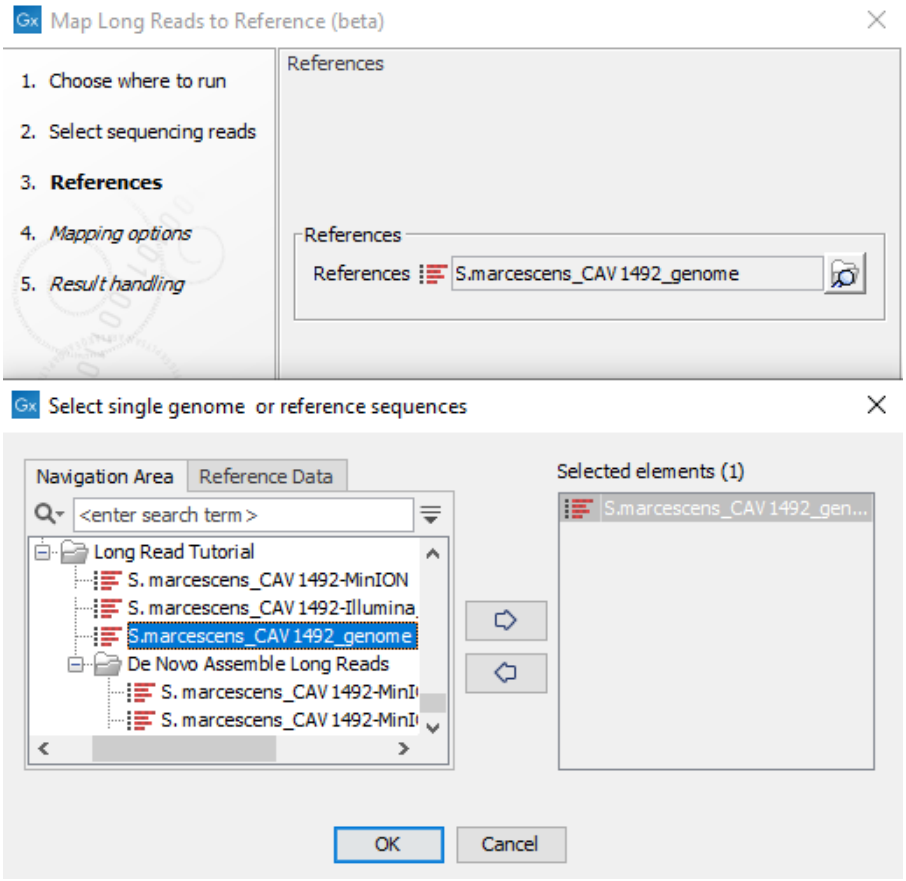

Figure 17: *Select the reference to map to*

5. Check the **Create report** option (Figure). Click **Next** and **Save** the result to your "De Novo Assemble Long Reads" folder. If you wish to work with your contigs in the Genomics Finishing Module, you should check **Create stand-alone read mapping** in the **Output options**.
6. **Save** your Read Mapping in a new folder, for example titled "Map Long Reads to Reference".

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

Open the read mapping track to see that the chromosome and all plasmids have coverage.

Correct Long Reads (Optional)

Algorithms using the CLC read mapper in downstream analysis require a reduced error rate for optimal performance. For use in these cases, it is possible to correct the raw reads. As the error-correction is based on an all-vs-all mapping, it comes at the expense of an increased time-consumption and reduced sensitivity for genetic variants as these may be considered sequencing errors.

It should therefore be stressed that **Correct Long Reads (beta)** should not be run as part of **De Novo Assemble Long Reads (beta)** as polishing is default for this tool. Another helpful use for error-correction is in finishing an assembly where smaller contigs, for example plasmids, were not

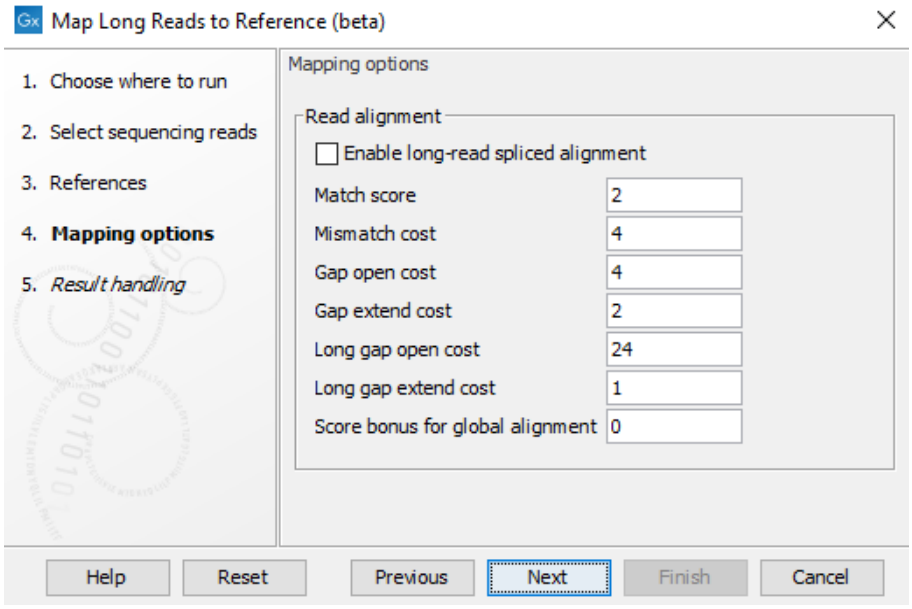

Figure 18: *Map Long Reads to Reference options*

fully resolved due to high error rate and low coverage. This is the case for plasmid CP011638. In the next steps, we will extract the reads from this plasmid in a subset. We will then run error-correction on the subset and rerun **De Novo Assemble Long Reads (beta)**. To start, you will need stand-alone read mapping output from **Map Long Reads to Reference (beta)**. The simplest way to do so without mapping the reads again, is to use the **Convert from Tracks** tool from the Track Tools, Track Conversion section of the Toolbox.

1. Start the tool **Convert from Tracks** from the **Track Tools** section of the Toolbox (Figure 19) or by using the Launch button and typing the name of the tool.

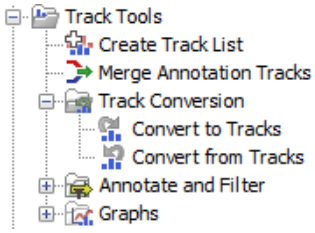

Figure 19: *Track Tools*

2. Select your read mapping (Figure 20) and click **Next**.
3. Save the output to your "Map Long Reads to Reference" location.
4. Open the output and select "CP011638.1". Click **Extract Subset** as show in figure 21 and save the output to a new location, for example titled "Plasmid CP011638.1".
5. Start the tool **Extract Sequences** from **Classical Sequence Analysis | General Sequence Analysis** from the Toolbox (Figure 22) or use Launch to search and run.
6. Select your extracted subset (Figure 23) and click **Next**.

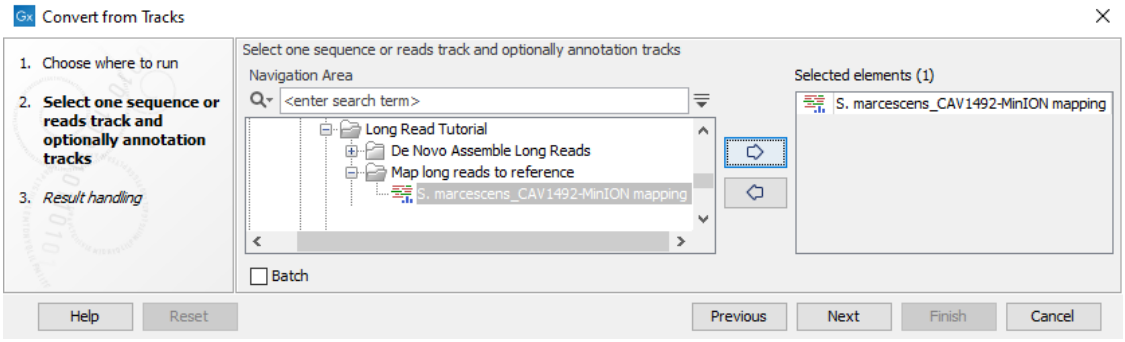

Figure 20: *Select the read mapping to convert*

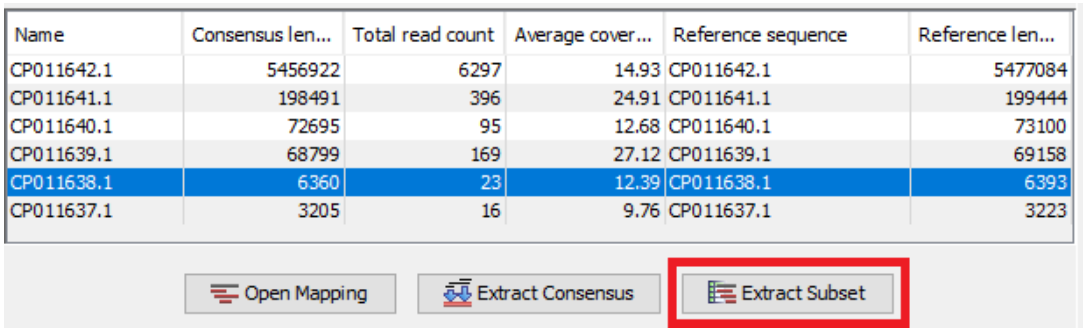

Figure 21: *Extract reads from the incorrectly assembled plasmid*

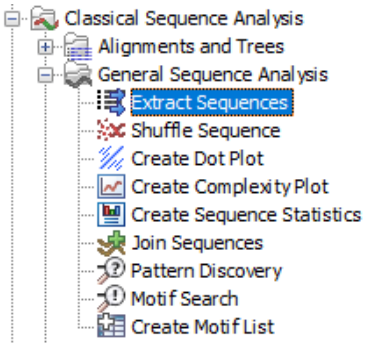

Figure 22: *Extract sequence in the Classical Sequence Analysis tools*

7. In the parameters, check **Extract to new sequence list** (Figure 24) and click **Next**. Save the output to your "Plasmid CP011638.1" location.

Reads mapping to this plasmid have now been extracted. You can now run **Correct Long Reads (beta)** on this subset.

1. To start the tool, locate **Correct Long Reads (beta)** in **Long Read Support (beta)** in the Toolbox or search and run using Launch.
2. Select the raw plasmid reads (Figure 25) and click **Next**.
3. In parameters under **Execution mode** you have the options of running Fast or Sensitive (Figure 26). Sensitive can correct additional reads at the cost of increased runtime. We will use parameters as default in this example. Click **Next**.

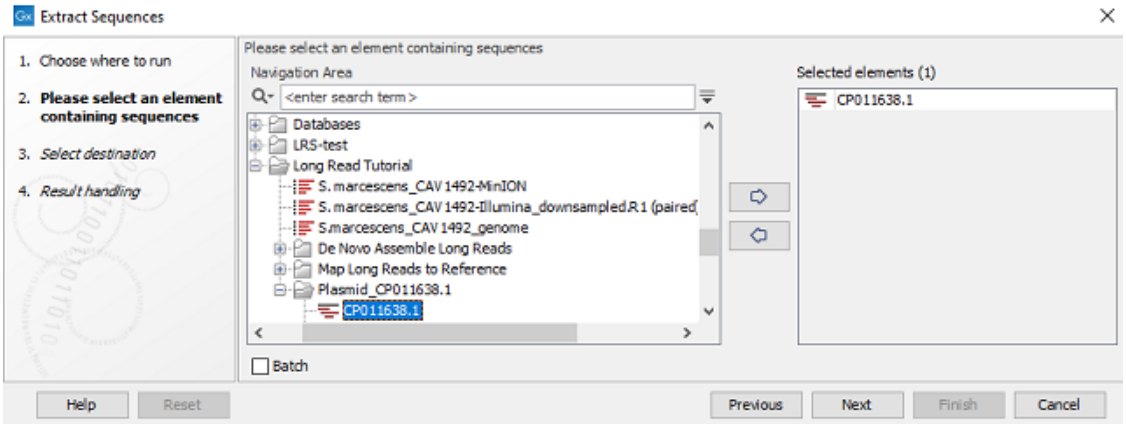

Figure 23: *Select the subset of the read mapping*

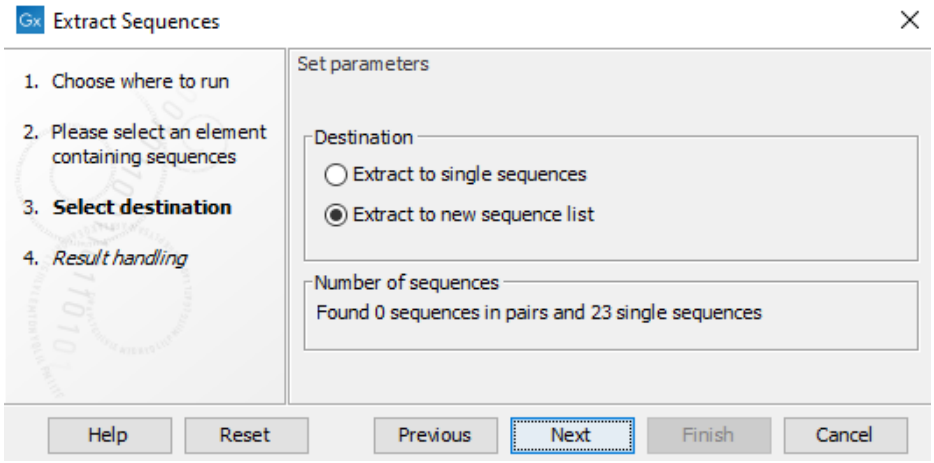

Figure 24: *Extract Sequences to a new sequence list*

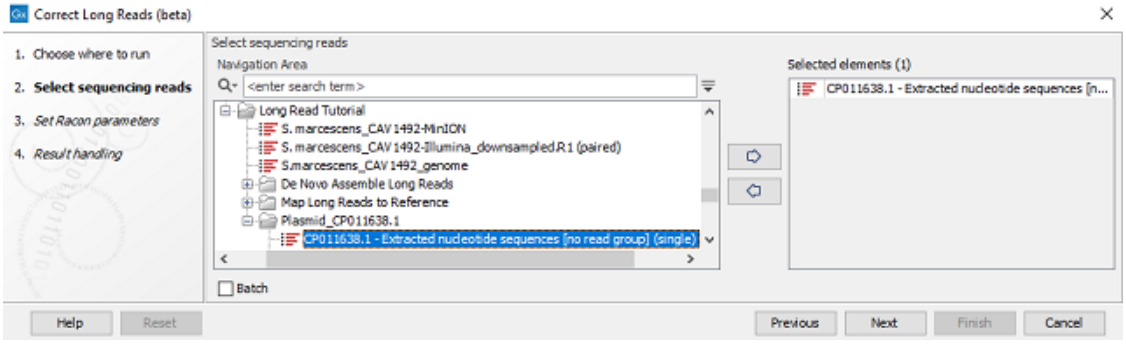

Figure 25: *Select the reads to error-correct*

4. Check **Create report** and choose to **Save** the result and click **Next**. Save the output to your "Plasmid CP011638.1" location. The error correction will now run. Since this plasmid is only covered by 23 reads, it will only take a few seconds to complete.

5. You should have two outputs; a set of corrected reads and a Read correction report. Open the report and compared the input and output statistics. Only 1 read was discarded so you have enough coverage to assemble the plasmid.

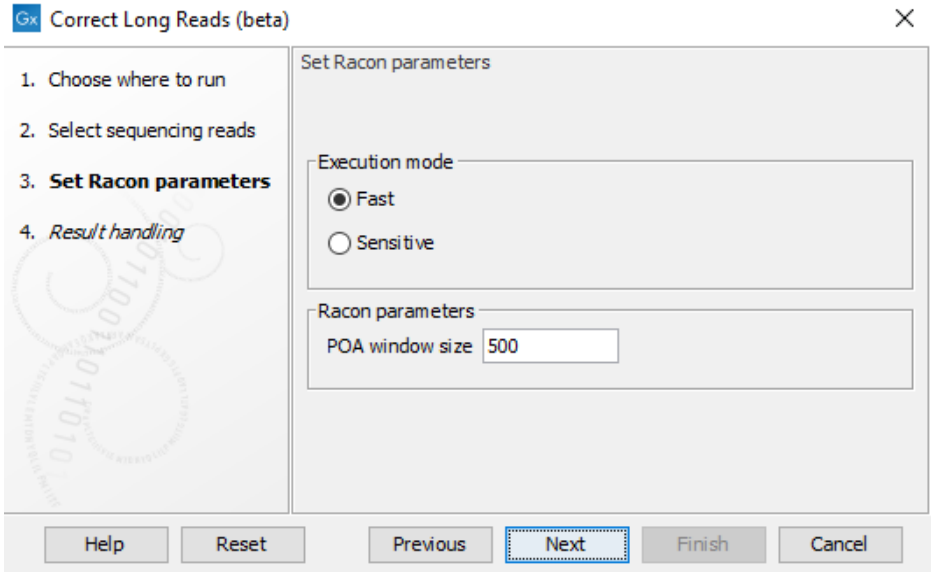

Figure 26: *Correct Long Reads (beta) options*

6. Run **De Novo Assemble Long Reads (beta)** on the uncorrected and corrected plasmid reads as described in the previous section. The uncorrected reads will assemble to a linear contig with length 12.5kb. The corrected reads, however, assemble to one circular contig with length 6.3kb which is what we expected from the reference.

7. We can visualize that the large linear plasmid consists of 2 copies of the actual plasmid. To do so, open **Create Whole Genome Dot Plot** from Whole Genome Alignment Toolbox or use Launch to search and run.

8. Select the two plasmid assemblies (Figure 27) and click **Next**.

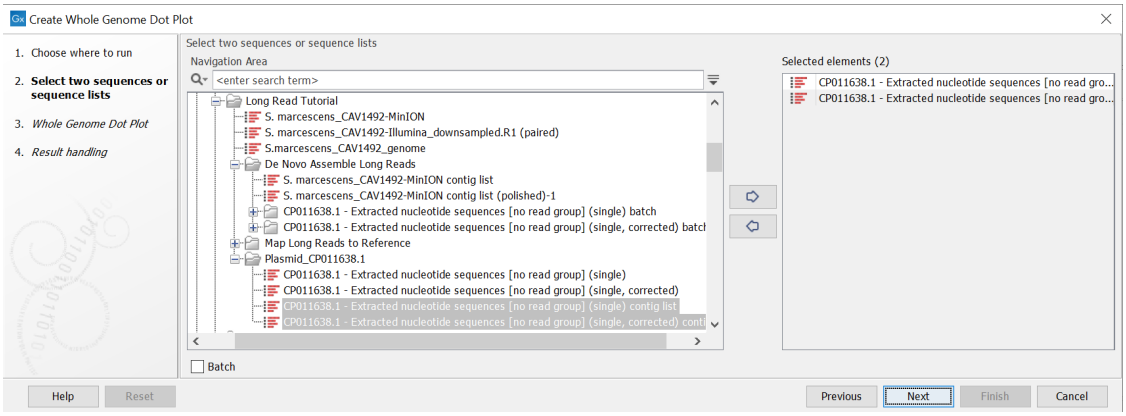

Figure 27: *Select the two assemblies*

9. Leave the settings as default (Figure 28) and click **Next**.

10. Choose to **Open** or **Save** and click **Finish**. In the dot plot on Figure 29 you see two lines showing that the smaller plasmid aligns twice to the larger assembly.

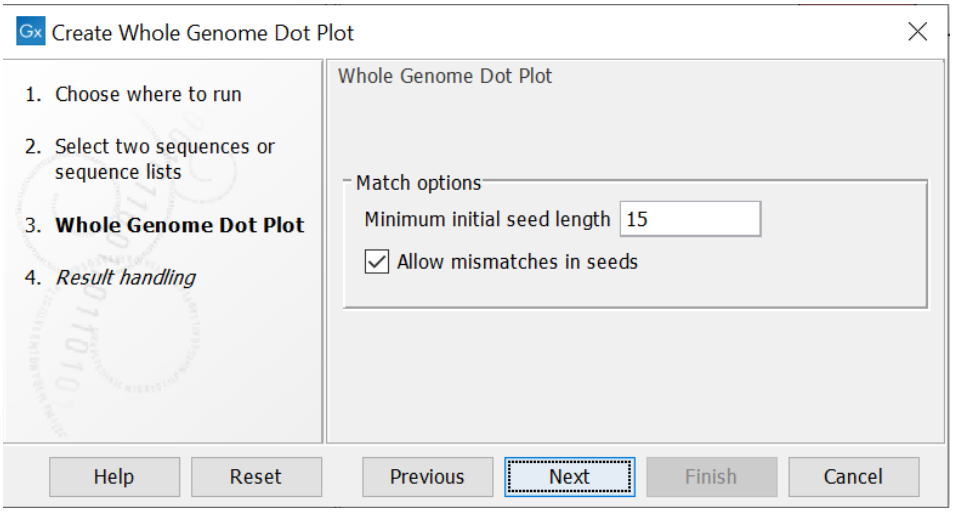

Figure 28: *Create dot plot options*

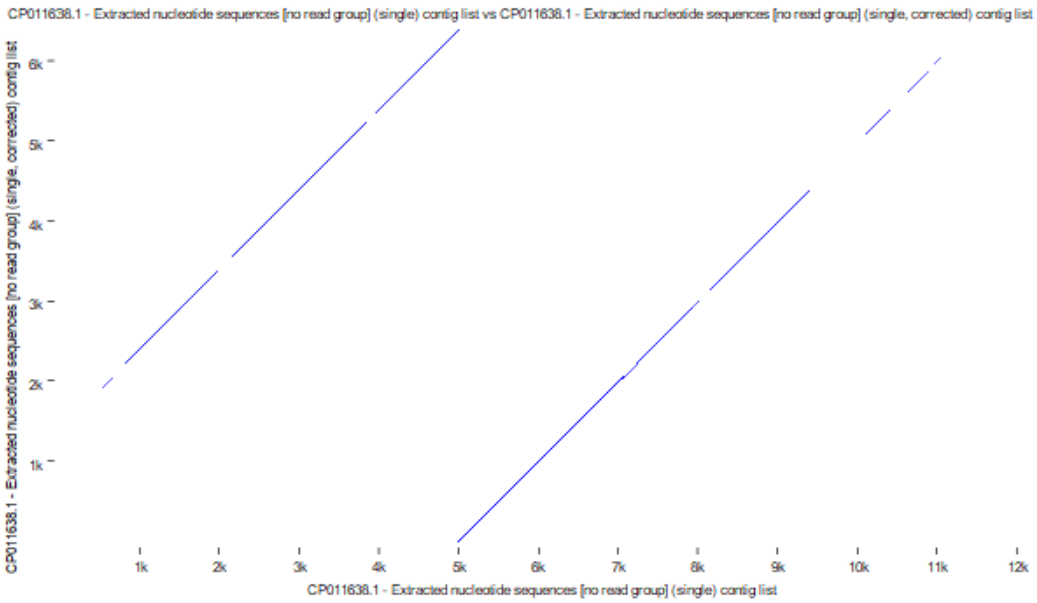

Figure 29: *Dot plot of the two assemblies*

To see the effect of error-correction, try mapping the raw and corrected Nanopore reads to your assembly and opening them together in the Track Viewer.

File | **New** | **Track list...**

View the **Track lists** help page for additional how-to. In (Figure 30), the difference in errors between the raw and corrected reads is clearly visible.

Summary

Using the Long Read Support (beta) plugin, we were able to quickly assemble a microbial genome. We were able to fully resolve 4 out of 5 plasmids. After polishing we received an Alignment percentage of 99.81 % and Average nucleotide identity of 99.93 %. We were able to resolve 5

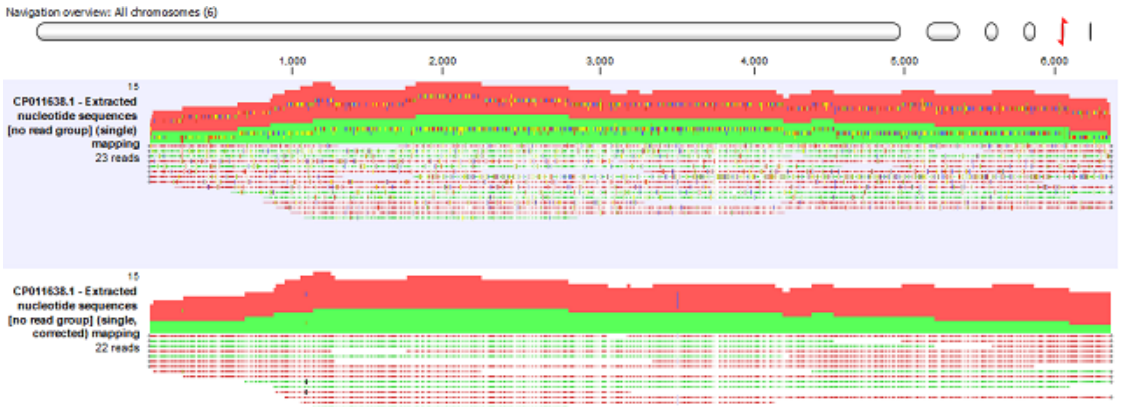

Figure 30: *The difference in errors between raw and corrected MinION reads*

out of 5 plasmids by correcting raw reads from the unresolved plasmid and reassembling.

This tutorial has demonstrated how to work with long error prone reads for creating high quality assemblies for microbial sized genomes.

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).

[Li, 2016] Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, 32(14):2103–2110.

[Li, 2018] Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100.

[Vaser et al., 2017] Vaser, R., Sovi, I., Nagarajan, N., and Šiki, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, 27(5):737–746.