Nanopore Plasmid Sequencing Results – Quick Guide

This is a quick guide to interpreting your plasmid assembly results. The aim is to provide a brief overview of the sequencing output data and to highlight some of the most useful files. It also shows some examples of how to compare sequencing results to a plasmid reference map.

The guide briefly explains the following:

1. How the data is produced and processed
2. Overview of output data
3. The assembly report
4. The sequence assembly (FASTA file)
5. Read mapping (bam file)
6. **How the data is produced and processed**

Samples are sequenced using the Nanopore Rapid Barcoding Kit 96 V14 (SQK-RBK114.96), on MinION R10.4.1 (FLO-MIN114) flowcells, using the MinION device. This is a nice video on how Nanopore sequencing works - <https://nanoporetech.com/applications/dna-nanopore-sequencing> . Library preparation using the rapid kit involves a transposase which simultaneously cleaves the DNA (linearizing plasmids) and adds barcodes onto the ends. The samples can then be pooled together, sequencing adapters ligated, and loaded on a flowcell for sequencing.

The sequencing data is processed using this plasmid assembly pipeline - <https://github.com/zihengluo/ont-plasmid-assembly-and-reads-mapping> . The raw sequencing reads are basecalled using the Nanopore data processing toolkit Dorado and a consensus sequence is assembled using the EPI2ME wf-clone-validation pipeline. If a reference map has been provided, the sequencing reads are also aligned to the reference.

1. **Overview of output data**

When sequencing and assembly are successful, you will receive the following:

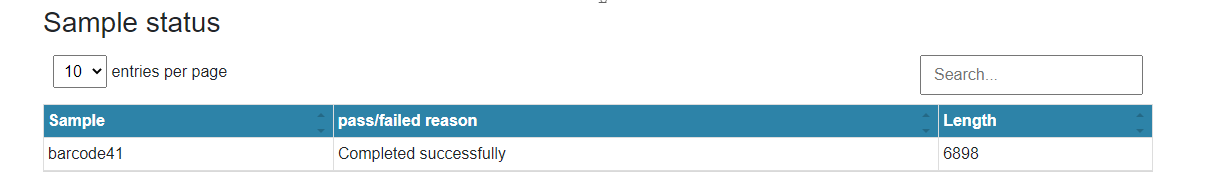
* Zipped Fastq folder (this is a standard guppy output, which contains the basecalled sequencing reads)
* An assembly report
* An assembled sequence FASTA file
* A .annotation file (produced using pLannotate, this file can be imported into programs such as Geneious, to add annotations onto the assembled sequence)

1. **The Assembly Report**

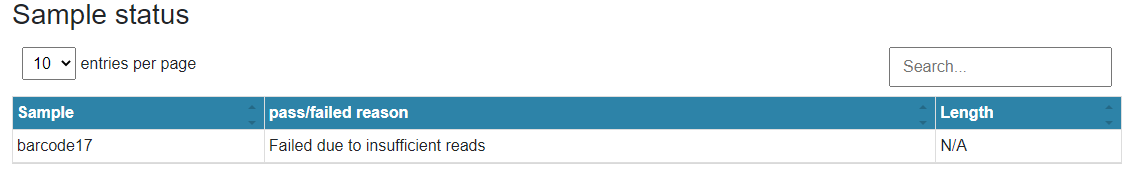
The assembly report is a html file. It shows a summary of the quality and length of sequencing reads.

The “Samples status” section shows if the reads were successfully assembled into a consensus sequence and how long the assembly is (bp). If the assembly failed, this section also displays the reason why. I have found the most common reason for failure is due to an insufficient number of reads. This usually occurs when a sample has very low concentration/high contamination (QC before submitting a sample is important!).

Example of successful assembly:

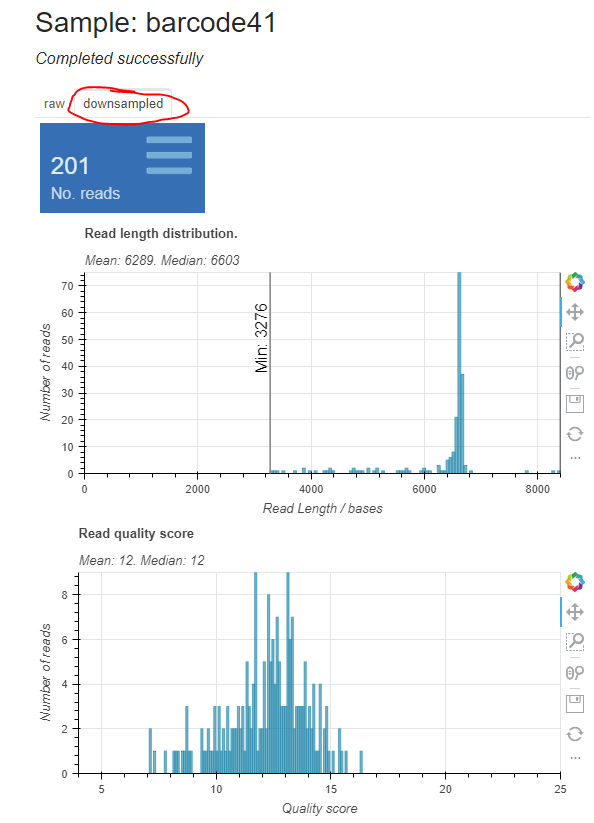
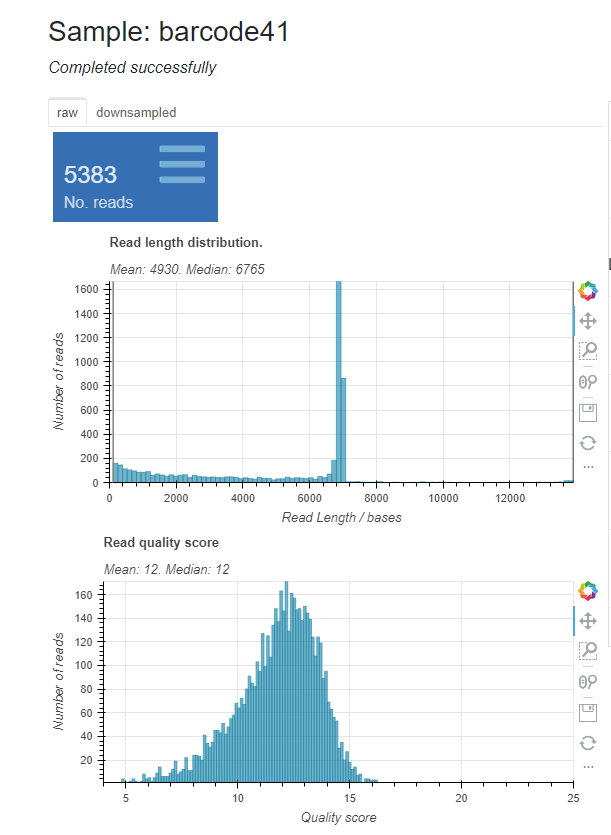


Example of assembly that failed due to insufficient reads:



Further down the report are 2 histograms displaying the sequencing read lengths and quality score of reads (the higher the average quality score, the better). If you click on the downsampled button, only the filtered reads are shown – these are the reads used by the EPI2ME pipeline for constructing the sequence assembly.

We generally expect most of the reads to be around the full length of the plasmid. You can see this in the example below, the plasmid is ~7000bp long, and there is a dominant peak around this length on the histogram. In the downsampled tab, the shortest and lowest quality reads have been filtered out.



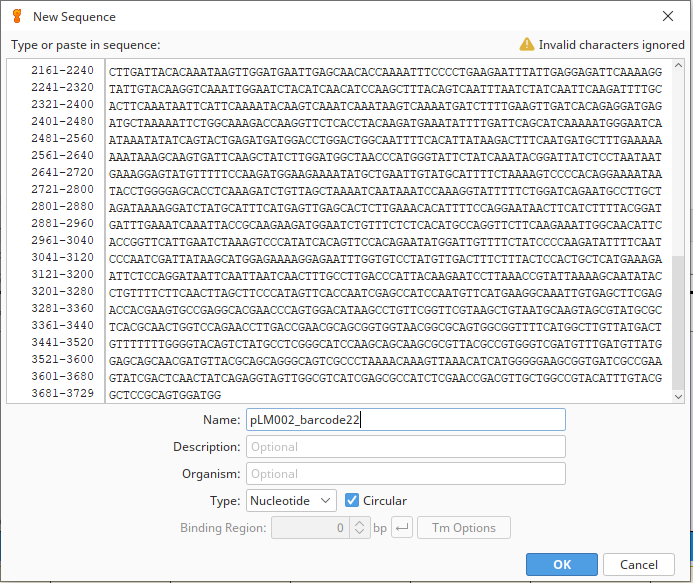
Other things you might see in read length histograms:

* **Multiple strong peaks** – if there are two or more dominant peaks in the read length histogram, this is an indication that the sample contained a mix of plasmids of different sizes. E.g. the sample may contain some full plasmids, and some empty vectors. This is not ideal, as the consensus sequence generated by the assembly pipeline may be a mix of the multiple plasmids present.
* **No strong peak, or the dominant peak is very far to the lef**t – this is likely caused by degraded/highly fragmented plasmids. If the read length histogram looks like this, it is also likely that the pipeline will have failed to generate a consensus sequence. If you have submitted a sample which has a lower than recommended concentration, results may look like this.
* **Concatemers** – dimers/multimers seem to be very common in plasmid samples, in low amounts. Many of the plasmids we have sequenced have a small peak around 2x length of plasmid. When concatemers are present in very low numbers, this does not seem to affect assembly of the consensus sequence.
* **Assembly length and read length histogram don’t seem to match –** occasionally the EPI2ME pipeline generates a consensus assembly which is much longer (2x, 3x or 4x) or shorter than what is expected. For example, the read length histogram has a dominant peak around 5000 bp, but the assembly is around 10 000bp long and it appears that large sections of the backbone may have been duplicated in the assembly. I try to keep an eye out for this when uploading the sequencing results, and re-run the assembly to try and produce a better consensus sequence when possible.

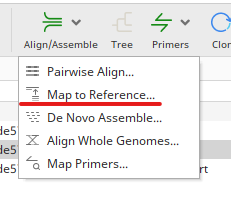
1. **The Sequence Assembly (FASTA file)**

The sequence assembly FASTA file is a text file containing the plasmid consensus sequence generated using EPI2ME.

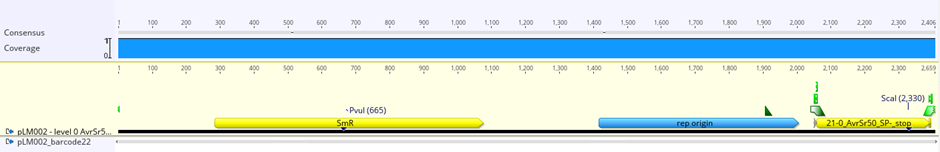
If you wish to compare the sequence assembly to a reference map, this can easily be done in the program Geneious. Firstly, add the sequence in the FASTA file as a new sequence on Geneious. Also add the corresponding reference map, if it is not in Geneious already.



Select the sequence assembly and use the “Map to Reference” function. Select the corresponding reference map, click OK and Genieous will align the assembled sequence with the reference.

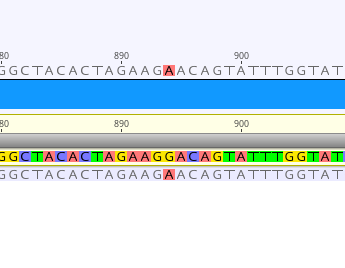
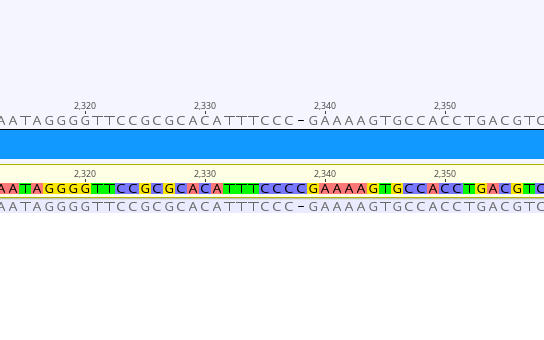


Here is an example of a sequence assembly which perfectly matches the reference sequence:



Here is an example of a plasmid sequence assembly which does not match the reference sequence. There is a SNP and multiple deletions in the sequenced plasmid.



A SNP A deletion