Read Me – Full PlasmidSeq Results

Thank you for choosing Microsynth's Full PlasmidSeq service.

File List and Output Description

We have organized the analysis output of your sample into two folders.

The **main** folder contains the following files:

- ***.README.pdf**: this README document (portable document format).
- The ***.fasta** file contains the final assembled plasmid sequence in fasta format as generated de novo from the sequencing reads by our analysis pipeline. It can be opened with a text editor. **Note:** In some cases, the assembly pipeline may produce more than one contig sequence per sample. This may indicate, among other things, a low-quality DNA sample or the presence of a plasmid mixture.
- The ***.fastq** file contains the sequence and the corresponding Q scores in fastq format. It can be opened with a text editor.
- In addition, the ***.annotation.gbk** file in genbank format contains the annotated plasmid sequence as generated by pLannotate (doi: 10.1093/nar/gkab374). It can be opened with a text editor.
- A corresponding graphical map is included as ***.annotation.html** file in Hypertext Markup Language format, which can be opened by a web browser.

The additional files in the **Addendum** subfolder may be of interest to an advanced user:

- ***.raw.reads.html** in Hypertext Markup Language format as produced by NanoPlot (doi: 10.1093/bioinformatics/bty149) provides the statistics of the sequenced reads that passed our filters and were used to generate the assembly. The file can be opened with a web browser.
- ***.mapping.tsv** in tab-separated value format provides a score of all contig residues based on long read alignments. It can be opened in Excel. A derived file containing only those residues with an accumulated mismatch, insertion, and deletion rate of >10% is provided in the folder: ***.uncertain.tsv**.
- The ***.ab1** files contain the pseudo-DNA electropherograms as you might know them from Sanger sequencing. We have included sequencing quality, coverage and mismatch, insertion, and deletion rates in the peak information. Basically, all the information shown in the ***.uncertain.tsv** and ***.mapping.tsv** files is shown here. Alternative bases at a given position are shown as secondary or tertiary peaks. The relative height of the peaks represents the relative coverage within the plasmid. This allows for a more intuitive evaluation of your sequencing results, similar to Sanger sequencing. Please note, however, that unlike Sanger electropherograms, the abi1 files here are not data, but represent an interpretation of the sequencing data by assemblers, polishers, and other software based on the individual reads. The length of *.abi1 files is limited to 2'500 bp. If your contig is longer, you will receive multiple files to cover the entire sequence. Individual files will not overlap.
- ***.depth.pdf** depicts the read depth across the whole sequence
- ***.coverage.tsv** contains information about coverage, sequencing quality and the number of reads used for the assembly

- All reads (after base calling, adapter trimming, and quality filtering) are provided on request in fastq format as a gzip-compressed ***raw.reads.fastq.gz** file.

Step-by-step Guide for Full PlasmidSeq Data

- For an overview, we suggest that you open the ***.annotation.html** file in the main folder. The file shows the annotated circular map of your plasmid, including the total size, and allows you to quickly compare with your expectations. An example map is shown in Figure 1: identity.tsv

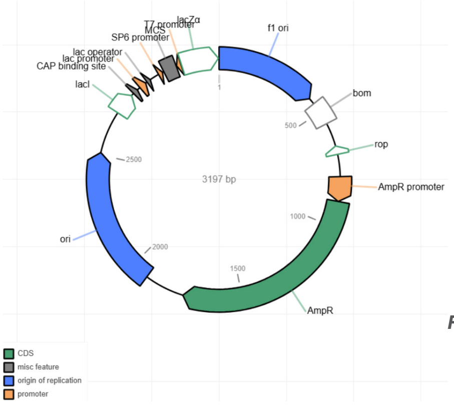

***Figure 1:** An annotated plasmid map as generated by pLannotate*

- Next, you can consult the ***.raw.reads.html** file provided in the addendum folder that contains the NanoPlot report (doi: 10.1093/bioinformatics/bty149). This report lists the lengths, Phred-scaled quality scores and number of reads that were used for the assembly. See “discussion on Phred-scaled quality scores“ for a definition of the Phred score. For good quality DNA preparations and successful sequencing runs, the read length is approximately equal to the plasmid size and the average Q score is greater than 17. The number of reads can vary, and we have observed good quality assemblies with as few as 50 reads. Typically, there are >250 reads per assembly.

- Now, you can open and check the assembled contig itself. There are two files, the first option is the naked sequence in fasta (***.fasta**) format, the second option is the annotated contig in genbank (***.gbk**) format. Any text editor will open these file types and you can also import them into your DNA software of choice.

- To identify potentially problematic regions within your construct, you can open the ***.uncertain.tsv** file in Excel. Residues with >10% of the aligned reads containing mismatches, insertions or deletions compared to the provided consensus sequence are listed. The data are sorted by position by default. In Table 1 below, we have re-sorted by mismatch [%] and selected the top three residues. The data shows that at the position 1587, 162 out of 601 reads contain a G instead of an A. The MeanBaseQ column shows the average base quality of the aligned bases at each of the positions listed. The bases at position 1587 have an average Q value of 26 (indicating a base calling accuracy of 99.75%). Since 601 reads cover this position, the overall quality at this position is much higher. The IUPAC column shows the IUPAC nucleotide code for this position. A nucleotide with >20% mismatch to the RefBase is considered when generating the IUPAC code. If a deletion makes at least 50% of the ReadDepth at a certain position, the IUPAC code ‘-’ is displayed. If this was a critical residue of your construct, you may want to consider verifying the sequence with an orthologous method such as Sanger. Please note that Sanger sequencing requires >20% of the template to have a different sequence to show up. This is because Sanger sequencing records the average signal of all templates that are sequenced in a synchronized manner, while ONT provides the sequences of each individual molecule. The assembly is then performed by the bioinformatics pipeline, which allows the statistical analysis based on all the reads. This provides information about the entire population of molecules that make up your plasmid, potentially revealing divergent subpopulations. However, the general ONT error rate is relatively high and not randomly distributed, so we advise caution in accepting SNPs in subpopulations.

Table 1: *An excerpt from a *.uncertain.tsv file*

Position	RefBase	IUPAC	A	C	G	T	Del	Ins	ReadDepth	MeanBaseQ	MeanMapQ	QScore	Match[%]	Mismatch[%]	Del[%]	Ins[%]
1587	A	R	438	0	162	0	1	12	601	25.62	60.00	5	72.88	26.96	0.17	2.00
1604	A	A	522	1	75	0	3	2	601	28.74	60.00	9	86.86	12.65	0.50	0.33
1877	T	T	0	56	5	535	7	1	603	26.98	60.00	9	88.72	10.12	1.16	0.17

- If you prefer a more graphical representation of the sequencing result you may want to use the ***.ab1** files instead.