plasmids NG

PLASMID SEQ

Sequencing Results QC Report

Contact us

--

info@plasmidsng.com
www.plasmidsng.com

Index

Results Guide

Your results files will be named with an internal sample ID that we use to track your sample, and the reference that you supplied to us in your sample information.

Each file will have a file type suffix - here is an explanation of all of the different file types we provide to you:

Plasmid Assemblies

Your assemblies have the file suffix '**.fasta**'. Your assemblies are provided in one contig that is fully circularised, and where possible rotated to an appropriate starting gene.

Sequencing Reads

Your raw sequencing reads have the file suffix '**.fastq.gz**'. These files contain all sequencing reads that were generated for the plasmid sample you sent, including non-target sequence (e.g. host contamination). We provide all reads larger than 200 bp.

Annotation

The primary annotation files have the file suffix '**.gbk**'.

Additional Files

ReadLength_Histogram.png - A log weighted distribution of long read sizes.

VirtualGel.png - A different way to visualise the most frequent peaks within your reads.

ReadPeaks.csv - A list of approximate peak sizes found in the long read dataset.

pLann.html - An interactive plasmid map showing features identified by plAnnotate.

pLann.csv - A table of genomic features identified by plAnnotate.

bases.txt - A list of all genomic positions in your plasmid and the frequency of each nucleotide found in the raw reads. Highlights bases with an allele frequency of >70%.

Plasmid Length

The read length histogram and ReadPeaks.csv file containing a list of identified peaks are particularly useful to QC your samples - pure plasmids will show a clear peak at the plasmid length. Partially degraded or host DNA contaminated samples will appear as peaks outside of the particular plasmid size, typically at lower size ranges.

In some host backgrounds (e.g. in *recA+ E. coli*) , plasmids can form multimers, which will produce peaks at n-fold the plasmid length (e.g. a dimer would be twice as long). More information can be found here.

We provide an assembly of the peak that most closely matches the plasmid size you have declared during sample submission, however we provide access to the full readset, allowing you to filter and assemble the data as appropriate.

Lower confidence bases

We produce a consensus assembly by leveraging high-depth sequencing, and return an assembly that is highly accurate on a per-base level. However, Oxford Nanopore long read data does have some common motifs that it struggles to resolve. To address this, we identify lower confidence bases by mapping your reads against the high quality consensus assembly.

We then identify the frequency of each nucleotide at a given position. In high confidence base positions, the vast majority of raw reads will contain the assembled base. In areas where there is less confidence - for example in some of those motifs that can cause problems, like Dcm methylation sites (CC[A/T]GG) - there may be multiple nucleotides identified in the raw reads. It doesn’t mean that the assembled base is incorrect - polishing with Medaka corrects a lot of troublesome bases - but it is something to bear in mind if your assembly is different to what you are expecting.

Methodology

Sample Processing

Once your package arrives at the MicrobesNG office, our lab technicians unpack them and inspect the tubes or plates for any potential leaks or contamination events. Once everything's passed that check, your samples are marked as received and they begin progressing through our pipeline. We'll let you know via email.

Sequencing

Genomic DNA libraries are prepared using the Rapid Barcoding Kit 96 V14 (SQK-RBK114.96) and sequenced using an R10.4.1 flowcell (FLO-MIN114) on our in-house GridION (Oxford Nanopore Technologies, UK). Raw signal data is basecalled using the GridION deployment for Guppy (ont-guppy-for-gridion V 6.3.9) using model number r1041_e82_400bps_hac_v4.2.0. Barcode trimming is enabled. Reads under 200 bp are discarded.

Bioinformatics

Reads are processed with Rasusa (V 0.7.0) and SeqKit (V 2.2.0), and assembled using Canu (V 2.2.0) and Trycycler (V 0.5.3). Assemblies are annotated by pLannotate (V 1.2.0). Polishing is performed using Medaka (V 1.7.0).

Software Name	Version	Reference
Rasusa	V 0.7.1	https://doi.org/10.21105/joss.03941
SeqKit	V 2.4.0	https://doi.org/10.1371/journal.pone.0163962
Canu	V 2.2.0	https://doi.org/10.1101/gr.215087.116
Trycycler	V 0.5.4	https://doi.org/10.1186/s13059-021-02483-z
pLannotate	V 1.2.0	https://doi.org/10.1093/nar/gkab374
Medaka	V 1.8.0	https://github.com/nanoporetech/medaka

Sequencing Controls

On every sequencing run we load three internal controls to validate our strict criteria for quality and accuracy:

Positive Control

We have an internal control plasmid that is extracted using a standard miniprep kit, checked for DNA integrity, plasmid purity and correct size in a gel (Tapestation), and quantified using a Qubit. We load this plasmid at a fraction of our required concentration, and ensure that we are able to obtain more than enough long reads of the plasmid to produce a good quality assembly from the data we generate.

This means that we can be sure that even with some flexibility in sample input from our minimum requirements, we can provide a good result. If we are unable to generate sufficient amount of data to generate an assembly, it normally means that your sample was provided at a lower concentration than our positive control and/or that the integrity of your samples is low (e.g. host DNA contamination).

Negative Control

We also include a negative / no template control on every run.

Cross Contamination

Finally, we include a technical control to detect any potential cross contamination during the library preparation process. Using a proprietary synthetic oligonucleotide, we are able to track any incidents of contamination to ensure total security and accuracy of the data you receive.