



PacBio Assembly and Annotation (Prokaryote)

Assemblies were generated via Flye¹ with the `--pacbio-hifi` read input flag, a stringent read error cutoff, and additional options to initiate the assembly by first using reads longer than an estimated N50 based on a genome size of 6Mbp. Circlator² was then used to circularize eligible contigs using the long read data. Final assemblies were annotated using Bakta³ under default settings. The output assembly and annotation files are “*SampleName.fasta*”, “*SampleName.gff*”, and “*SampleName.gb*”. Finally, QUAST⁴ was used to generate assembly metrics which are reported in the “*Assembly Metrics.tsv*” file.

Kinetics data from the PacBio HiFi reads were used to determine the methylation states of the assembled microbial genomes. Kinetics data was collected from the HiFi reads using `ccs-kinetics-bystrandify`, an algorithm in the `pbbam`⁵ suite. Exported kinetics data were mapped to the generated assembly using `pbmm2` from the SMRT Tools suite in SMRT Link 11.0⁶. The `ipdSummary` algorithm putatively identified methylation states from the kinetics data as either undetermined, m4C or m6A. Methylation calls are reported in the file “*all_base_modifications.bigwig*” which can be viewed in a genome browser and “*all_base_modifications.gff3*” which can be loaded as an annotation file. The all modifications `gff3` file was subset into files for m4C and m6A `gff3` files, respectively.

Tool	Version	Parameters
flye	2.9.2	<code>--asm-coverage 50 --genome-size 6000000</code> <code>--pacbio-hifi --read-error 0.02</code>
circlator	1.5.5	<code>all</code> ; 6-hour timeout
bakta	1.8.1	default parameters; db version 5.0
quast	5.2.0	default parameters
pbbam	2.1.0	<code>ccs-kinetics-bystrandif</code> (default parameters)
smrtlink	12.0.0.177059	<code>pbmm2</code> (default parameters) <code>ipdSummary</code> (<code>--identify m6A,m4C</code>)

¹ Mikhail Kolmogorov, Jeffrey Yuan, Yu Lin and Pavel Pevzner. Assembly of Long Error-Prone Reads Using Repeat Graphs. *Nature Biotechnology* (2019)

² Hunt, M., Silva, N.D., Otto, T.D. et al. Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol* 16, 294 (2015).

³ Schwengers O., Jelonek L., Dieckmann M. A., Beyvers S., Blom J., Goesmann A. (2021). Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microbial Genomics*, 7(11).

⁴ Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi and Glenn Tesler, QUAST: quality assessment tool for genome assemblies, *Bioinformatics* (2013) 29 (8): 1072-1075.

⁵ PacificBiosciences/pbbam: Pacbio BAM C++ library. GitHub.

<https://github.com/PacificBiosciences/pbbam>

⁶ SMRT® Tools (v11.0). <https://www.pacb.com/support/software-downloads/>