Improving Long-Read Assembly of Microbial Genomes and Plasmids

Ivan Sović[1], Zev Kronenberg[1], Christopher Dunn[1], James Drake[1], Joan Wong[1,2], **Dan Browne[1]**, Meredith Ashby[1]
[1]PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025
[2]Assembly Bio, 331 Oyster Point Blvd, South San Francisco, CA 94080

Contact:
Dan Browne, PhD
(650) 785-1513
dbrowne@pacb.com

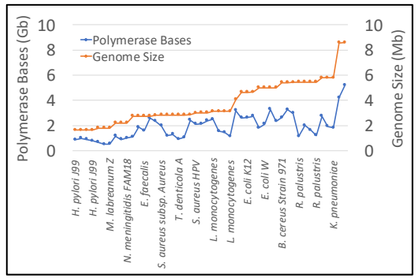
Background and Motivation

Complete, high-quality microbial genomes are very valuable across a broad array of fields, from environmental studies, to human microbiome health, food pathogen surveillance, etc. Long-read sequencing enables accurate resolution of complex microbial genomes and is becoming the new standard. Here we report our novel Microbial Assembly pipeline to facilitate rapid, large-scale analysis of microbial genomes. We sequenced a 48-plex library with one SMRT Cell 8M on the Sequel II System, demultiplexed, then analyzed the data with Microbial Assembly.

Strains, Sequencing, and Demultiplexing

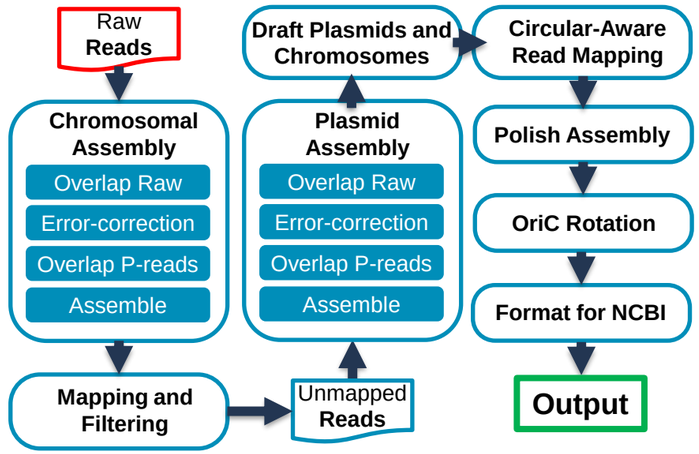
Strain	ATCC ID	Replicates	Genome	Plasmids
Bacillus cereus 971	14579	3	5.4 Mb	15 kb
Bacillus subtilis W23	in-house	1	4.0 Mb	n/a
*Burkholderia cepacia**	25416	2	8.4 Mb	209 kb
Enterococcus faecalis OG1RF	47077D-5	4	2.7 Mb	n/a
Escherichia coli K12	in-house	3	4.6 Mb	n/a
Escherichia coli W	9637	4	4.9 Mb	5, 102 kb
Helicobacter pylori J99	700824	4	1.7 Mb	n/a
Klebsiella pneumoniae	BAA-2146	3	5.4 Mb	2, 85, 118, 141 kb
Listeria monocytogenes	19117	4	2.9 Mb	n/a
Methanocorpusculum labreanum Z	43576	3	1.8 Mb	n/a
Neisseria meningitidis FAM18	700532	3	2.2 Mb	n/a
Rhodopseudomonas palustris	in-house	4	5.5 Mb	n/a
Staphylococcus aureus HPV	BAA-44	3	2.9 Mb	24 kb
Staphylococcus aureus subsp. aureus	25923	3	2.8 Mb	27 kb
Treponema denticola A	35405	4	2.8 Mb	n/a

Value	Analysis Metric
104,515,141,915	Polymerase Read Bases
2,764,838	Polymerase Reads
37,802	Polymerase Read Length (mean)
87,502	Polymerase Read N50
9,544	Subread Length (mean)
12,192	Subread N50
10,735	Insert Length (mean)
14,812	Insert N50
28,944,687,104	Unique Molecular Yield

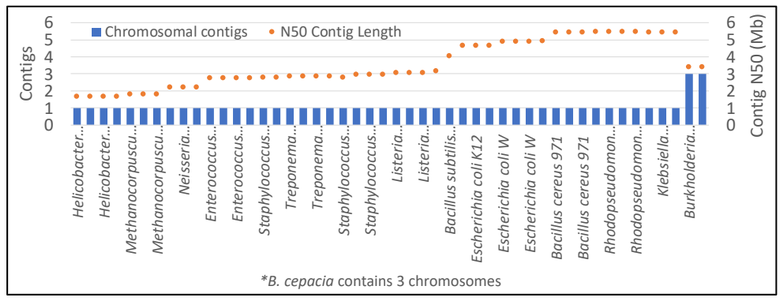


Results of sequencing and demultiplexing barcodes. The microbial pooling calculator helps balance coverage across species with different genome sizes, such that each has enough data to be processed with Microbial Assembly. Our lima demultiplex tool is available as a free bioconda package or in SMRT Link.

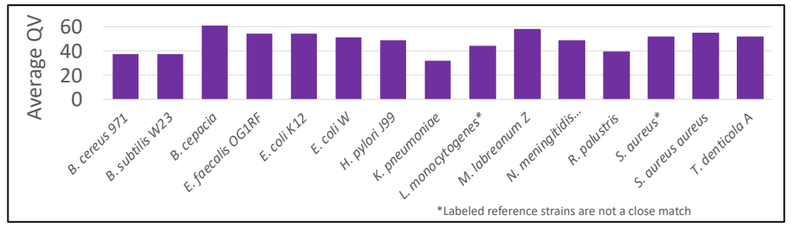
Brand New Microbial Assembly Tool



Performance of Chromosome Recovery



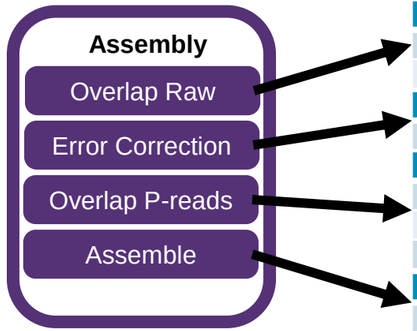
Microbial Assembly results in recovery of chromosomes in complete contigs. The chromosomes of each bacteria were consistently assembled into complete, circular contigs. This figure does not account for the plasmids assembled by the pipeline and included in the output.



Analysis of Microbial Assembly results compared with references. The average nucleotide identity (ANI) was calculated for the Microbial Assembly results and respective reference assemblies from the NCBI database. These data show that our results agree strongly with reference assemblies.

Implementation of New Features

- Key differences between HGAP4 and Microbial Assembly
  - No repeat masking in Microbial Assembly workflow
  - DALIGNER tool stack is replaced with Raptor tool stack
  - **Raptor – New graph-based mapping and alignment tool**
  - **New method to detect and filter out chimeric reads**
- String graph and error-correction reused from FALCON



HGAP4	Microbial Assembly
DALIGNER	Raptor
DAZZ_DB	RaptorDB
HGAP4	**Microbial Assembly**
fc_consensus	fc_consensus
HGAP4	**Microbial Assembly**
DALIGNER	Raptor
DAZZ_DB	RaptorDB
-	Chimera Detection
HGAP4	**Microbial Assembly**
Falcon SG	Falcon SG

Performance of Plasmid Recovery

E. coli W plasmids	Replicate 1	Replicate 2	Replicate 3	Replicate 4
5 kb	X	✓	X	✓
102 kb	✓	✓	✓	✓

K. pneumoniae plasmids	Replicate 1	Replicate 2	Replicate 3
2 kb	X	X	X
85 kb	✓	✓	✓
118 kb	✓	✓	✓
141 kb	✓	✓	✓

Microbial Assembly results in successful recovery for most plasmids. The notable exception is a 2 kb plasmid in *K. pneumoniae* and a 5 kb plasmid in *E. coli W*. However, this is likely due to removal of shorter DNA fragments during the library preparation. By default, the pipeline assumes all contigs <300 kb are plasmids.

Bacteria with single plasmid	*B. cereus* 971 (15 kb)	*B. cepacia* (209 kb)	*S. aureus* HPV (24 kb)	*S. aureus* aureus (27 kb)	*L. monocytogenes* (no plasmid in NCBI)
Replicate 1	✓	✓ (204 kb)	✓ (32 kb)	✓	*New Plasmid* (94 kb)
Replicate 2	X	✓ (204 kb)	✓ (32 kb)	✓	*New Plasmid* (94 kb)
Replicate 3	X	N/A*	✓ (32 kb)	✓	*New Plasmid* (94 kb)
Replicate 4	N/A*	N/A*	N/A*	N/A*	X

*N/A indicates lack of replicates for strain (see strain table for number of replicates per strain)

Overall performance of Microbial Assembly is very good. With the recovery of complete chromosomes and a majority of plasmids, our tool will help advance the study of microbes in a wide variety of fields. Further developments in the sample preparation protocol and data analysis software will continue to improve results.