BV-BRC Test Report

A1. Service – Genome Assembly - Bacteria

Item to test	Genome Assembly Service using bacterial read files and SRA accessions
URL	https://www.bv-brc.org/app/Annotation
Prerequisites	Bacterial Fasta contig files in Workspace
References	https://www.bv-brc.org/docs/quick_references/services/genome_assembly_service.html https://www.bv-brc.org/docs/tutorial/genome_assembly/assembly.html
Tester(s)	Rebecca Wattam, Maulik Shukla
Test date	21-Apr-2022 (follow-up from original test)
Test result	**Passed**

Overview

- Test the Genome Assembly Service using exemplar reads sets for bacterial genomes.
- Test input options, i.e., single-end and paired-end read sets using files uploaded to the workspace and using an SRA run accession as input.
- Test the assembly strategies, i.e., Auto, Unicycler, SPAdes, Canu, MetaSPAdes, PlasmidSPAdes, and MDA.
- For each job submitted, verify successful completion of the job, presence of output files, and quality of the assembled contigs by comparing them with the same or closely related public genome.

Test Data

Dataset	Rational	Input Format	Input
Escherichia coli - SRR3584989	Workshop example	SRA accession	SRR3584989
Escherichia coli - SRR3584989 - read files	Workshop example	Read files	SRR3584989_1.fastq, SRR3584989_2.fastq
Buchnera aphidicola - SRR4240359	Workshop example	SRA accession	SRR4240359
Mycobacterium tuberculosis H37Rv	Reference genome	SRA accessions	SRR974841, SRR974842, SRR974843

- All test datasets and corresponding job results are available in the following public workspace: https://www.bv-brc.org/workspace/BVBRC@patricbrc.org/BVBRC%20Tests/Genome%20Assembly/Bacteria

Test Results

- All assembly jobs completed successfully, without any errors.

- All jobs resulted in expected output files in corresponding job output directory, providing assembly report in HTML format and assembled contigs in fasta format.
- The assembly report was informative and provided a concise summary of the input reads, assembly process and parameters used, iterative refining steps, filtering of the contigs based on minimum length and coverage, Quast report, and the reads used in the assembly. It also provided an assembly graph.
- For each of the genomes, the total length of the assembled contigs were as expected when compared to those in corresponding public genomes in PATRIC.
- All test datasets and corresponding job results are available in the following public workspace: https://www.bv-brc.org/workspace/BVBRC@patricbrc.org/BVBRC%20Tests/Genome%20Assembly/Bacteria
- Below are a series of screenshots showing successful completion of the jobs, availability of the result files in the workspace, excerpts of the assembly repport and summary of assembled contigs and total length for each of the jobs.

Status	ID	Service	Output Name	Submit	Start	Completed
completed	7419886	GenomeAssembly2	Escherichia_coli_SRR3584989	4/20/22, 4:19 PM	4/21/22, 2:35 AM	4/21/22, 3:15 AM
completed	7419887	GenomeAssembly2	Buchnera_aphidicola_SRR4240359	4/20/22, 4:20 PM	4/21/22, 2:37 AM	4/21/22, 3:37 AM
completed	7419890	GenomeAssembly2	Mycobacterium_tuberculosis_H37Rv	4/20/22, 4:27 PM	4/21/22, 3:07 AM	4/21/22, 4:02 AM
completed	7419922	GenomeAssembly2	Escherichia_coli_SRR3584989 - read files	4/20/22, 5:00 PM	4/21/22, 3:36 AM	4/21/22, 3:55 AM
running	7419923	GenomeAssembly2	Buchnera aphidicola SRR7796591 - read files	4/20/22, 5:01 PM	4/21/22, 3:38 AM	
completed	7419935	GenomeAssembly2	SRR4240359 - Auto	4/20/22, 5:11 PM	4/21/22, 3:39 AM	4/21/22, 4:07 AM
completed	7419936	GenomeAssembly2	SRR4240359 - Unicycler	4/20/22, 5:11 PM	4/21/22, 3:47 AM	4/21/22, 4:15 AM
completed	7419937	GenomeAssembly2	SRR4240359 - Spades	4/20/22, 5:11 PM	4/21/22, 3:54 AM	4/21/22, 4:14 AM
completed	7419938	GenomeAssembly2	SRR4240359 - Canu	4/20/22, 5:11 PM	4/21/22, 3:55 AM	4/21/22, 3:56 AM
completed	7419939	GenomeAssembly2	SRR4240359 - metaSpades	4/20/22, 5:12 PM	4/21/22, 3:56 AM	4/21/22, 3:58 AM
completed	7419943	GenomeAssembly2	SRR4240359 - PlasmidSpades	4/20/22, 5:12 PM	4/21/22, 3:58 AM	4/21/22, 4:16 AM

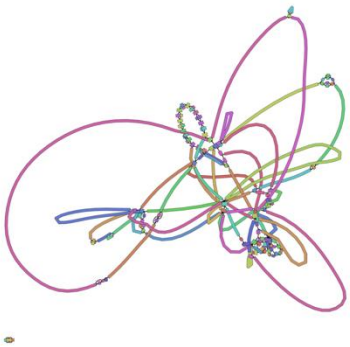
- Assembly results for Escherichia coli - SRR3584989

	Name	Size	Owner	Members	Created
	Parent folder			-	
	Escherichia_coli_SRR3584989_assembly_report.html	491.0 kB	me	Public	4/21/22, 3:15 AM
	Escherichia_coli_SRR3584989_contigs.fasta	5.3 MB	me	Public	4/21/22, 3:15 AM
	details		me	Public	4/21/22, 3:15 AM

Genome Assembly Report

Assembly Plot

Bandage Plot:



Bandage Version: 0.8.1

Assembly

Assembly Process	
assembler:	unicycler
assembly_seconds:	1603.90538597
assembly_time:	0.45 hours
command_line:	unicycler -t 12 -o . --min_fasta_length 300 --keep 2 --no_pilon --short1 SRR3584989_1.fastq --short2 SRR3584989_2.fastq
contigs.fasta file size:	5260434
version:	Unicycler v0.4.8

Polishing

Polishing Rounds	
Round:	1
input_contigs:	contigs.fasta
num_changes:	33
output:	contigs_pilon_1
program:	pilon
reads:	SRR3584989_1.fastq:SRR3584989_2.fastq
seconds:	26.957379818
Round:	2
input_contigs:	contigs_pilon_1.fasta
num_changes:	4
output:	contigs_pilon_2
program:	pilon
reads:	SRR3584989_1.fastq:SRR3584989_2.fastq
seconds:	30.7328560352

Filtering Contigs on Length and Coverage

Contig Filtering	
average depth (short reads):	81.1042569245
average short read coverage:	81.104
min_contig_coverage_threshold:	5.0
min_contig_length_threshold:	300
num contigs above thresholds:	100
num contigs below thresholds:	0
total length of good contigs:	5183531
total_short_read_bases:	438266197

Input Reads

SRR3584989_1.fastq:SRR3584989_2.fastq	
read file:	SRR3584989_1.fastq:SRR3584989_2.fastq
platform:	illumina
layout:	paired-end
num_reads:	746149
num_bases:	438266197
avg_len:	293
max_read_len:	301
sample_read_id:	@SRR3584989.1

Tools Used:

Tools Used	
Bandage:	Version: 0.8.1
pilon:	Pilon version 1.23 Mon Nov 26 16:04:05 2018 -0500
quast:	QUAST v5.0.2, fb0b821
samtools:	Version: 1.3 (using htslib 1.3)
unicycler:	Unicycler v0.4.8

- Summary of assembly results for test jobs:

Dataset	Input Format	Recipe	Contigs	Total Length	Expected Length
Mycobacterium tuberculosis H37Rv	SRA accessions	Auto	106	4,345,818	4,372,083
Escherichia coli - SRR3584989	SRA accession	Auto	100	5,183,531	5,211,994
Escherichia coli - SRR3584989 - read files	Read files	Auto	100	5,183,531	5,211,994
Buchnera aphidicola - SRR4240359	SRA accession	Auto	5	655,300	641,895
Buchnera aphidicola - SRR4240359	SRA accession	Unicycler	5	655,300	641,895
Buchnera aphidicola - SRR4240359	SRA accession	Spades	515	915,748	641,895
Buchnera aphidicola - SRR4240359	SRA accession	Canu *	-	-	641,895
Buchnera aphidicola - SRR4240359	SRA accession	PlasmidSpades	1	3,691	-

- As shown in the table above, the Auto recipe chose the best assembler for the input dataset (Unicycler) and provided optimal assembly. Also note that the Canu assembler did not produce contigs as it is designed for the long reads and the input dataset contained short reads from Illumina.

References

- Genome Annotation Protocol
- *RASTtk:* A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes
- Genome Annotation Service
- Genome Annotation Service Quick Reference Guide
- Genome Annotation Service Tutorial