Beta Test Report for BigBacter

Summary:

The beta test of the BigBacter pipeline involved five tests, assessing its efficiency and accuracy in bacterial genomic surveillance. With processing times of 42 to 47 seconds per sample using 20 CPUs and 32GB of memory, BigBacter demonstrated high accuracy in isolate clustering. Minor issues were noted with large file downloads from PopPUNK databases. Key feedback includes improving documentation clarity and adding demo data. Overall, BigBacter performed well but would benefit from these enhancements.

Introduction:

BigBacter is a pipeline designed for bacterial genomic surveillance. It simplifies the process by pre-clustering isolates into related subtypes before phylogenetic analysis, automatically selecting and archiving cluster-specific reference genomes for SNP analysis, excluding low-quality samples, and reusing archived alignment files to speed up SNP analysis. Additionally, it automatically generates necessary figures, such as phylogenetic trees and SNP matrices.

The beta test was conducted by the Office of Scientific Innovation and Integration of the Clinical and Environmental Microbiology Branch at the CDC. The test utilized data from the study by Stanton RA, McAllister G, Daniels JB, Breaker E, Vlachos N, Gable P, Moulton-Meissner H, and Halpin AL (2020), titled "Development and Application of a Core Genome Multilocus Sequence Typing Scheme for the Health Care-Associated Pathogen *Pseudomonas aeruginosa*," published in J Clin Microbiol, 58:10.1128/jcm.00214-20.

Test Objectives:

- Verify installation and setup
- Assess functionality and performance
- Identify bugs and issues

Test Environment:

- OS: CentOS Linux 7
- Nextflow version: 23.10.0
- Singularity version: 3.8.7
- BigBacter version: beta version (6b08a87)

Test Data:

- Data Source: "Development and Application of a Core Genome Multilocus Sequence Typing Scheme for the Health Care-Associated Pathogen Pseudomonas aeruginosa," published in J Clin Microbiol, 58:10.1128/jcm.00214-20.
- BioProject: https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA288601
- Number of samples: 25 samples from outbreak 2; 8 samples from outbreak 3; 6 samples from NCBI SRA database
- Sample List:

Outbreak 2	**Outbreak 3**	**Assembly**	**SRA sample**
2017-15-03	2017-40-01	GCA_040789945.1	SRR29830251
2017-15-08	2017-40-02	GCF_040513475.1	SRR29506595
2017-15-11	2017-40-03	GCF_040513845.1	SRR29506596
2017-15-15	2017-40-04	GCF_040513855.1	SRR29506597
2017-15-18	2017-40-08-02	GCA_030410115.2	SRR10012085
2017-15-217	2017-40-08-04	GCF_030284605.1	SRR21721505
2017-15-220	2017-40-17		
2017-15-01	2017-40-20		
2017-15-02			
2017-15-218			
2017-15-219			
2017-15-221			
2017-15-161-01			
2017-15-222			
2017-15-49-01			
2017-15-51-02			
2017-15-64-03			
2017-15-69-01			
2017-15-42-01			
2017-15-45-01			
2017-15-96-01			
2017-15-67-02			
2017-15-223			
2017-15-224			
2017-15-225			

Test Cases:

1. Installation & Setup:

- Successfully cloned the repository

```
git clone https://github.com/DOH-JDJ0303/bigbacter-nf.wiki.git
```

- Configured PopPUNK databases

```
nextflow run $pipeline/bigbacter-nf \
    -profile singularity,all_dbs \
    -entry PREPARE_DB \
    --db $PWD/db \
    --max_cpus 4 \
    --max_memory 8.GB
```

- Verified database file in db folder

2. Functionality Tests:

1) Test 1 with Default PopPUNK Database:

Sample Preparation:

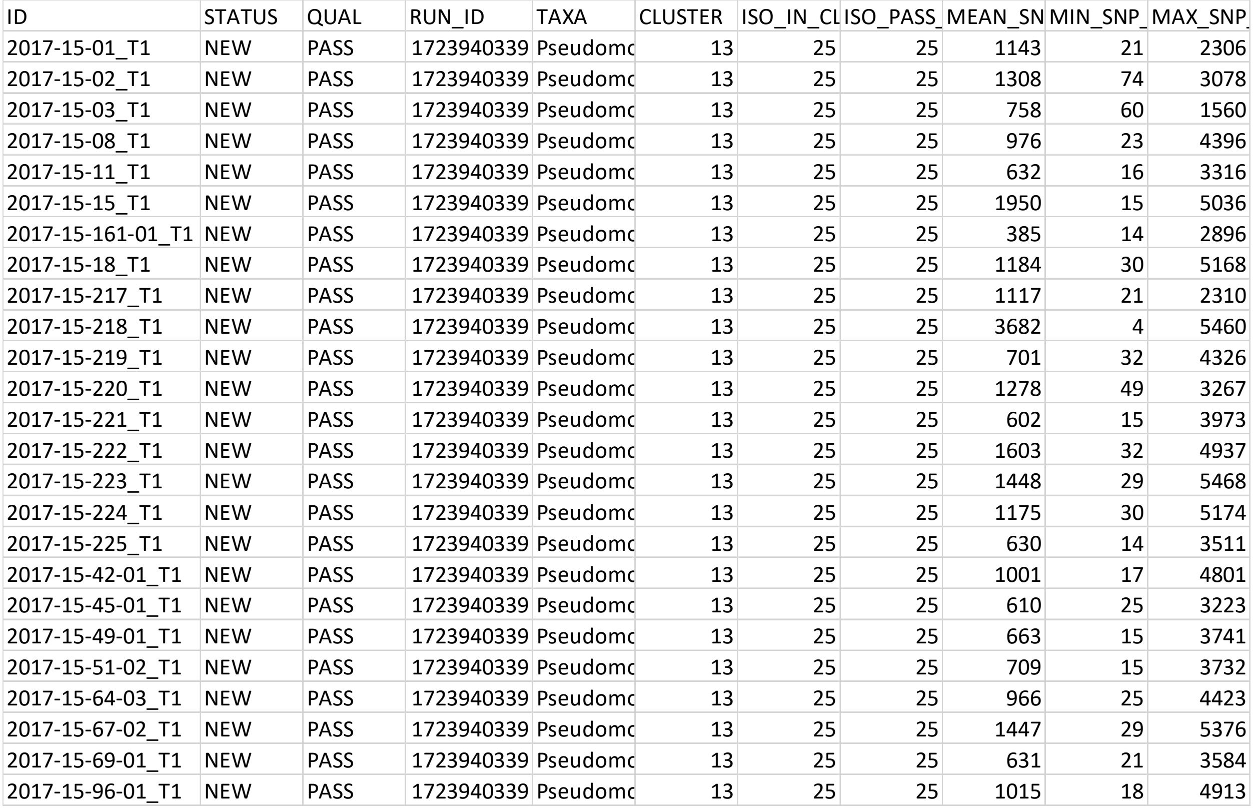
- Species: *Pseudomonas aeruginosa*
- Sample number: 25
- Assembly: Phoenix filtered scaffolds
- Input data: fastq files

Pipeline Execution:

- Run the pipeline with the command

```
nextflow run $pipeline/bigbacter-nf \
   -profile singularity \
   --input ${PWD}/samplesheet.csv \
   --db $db_dir/db \
   --outdir $PWD/results/ \
   --max_cpus 20 \
   --max_memory '32.GB'
```

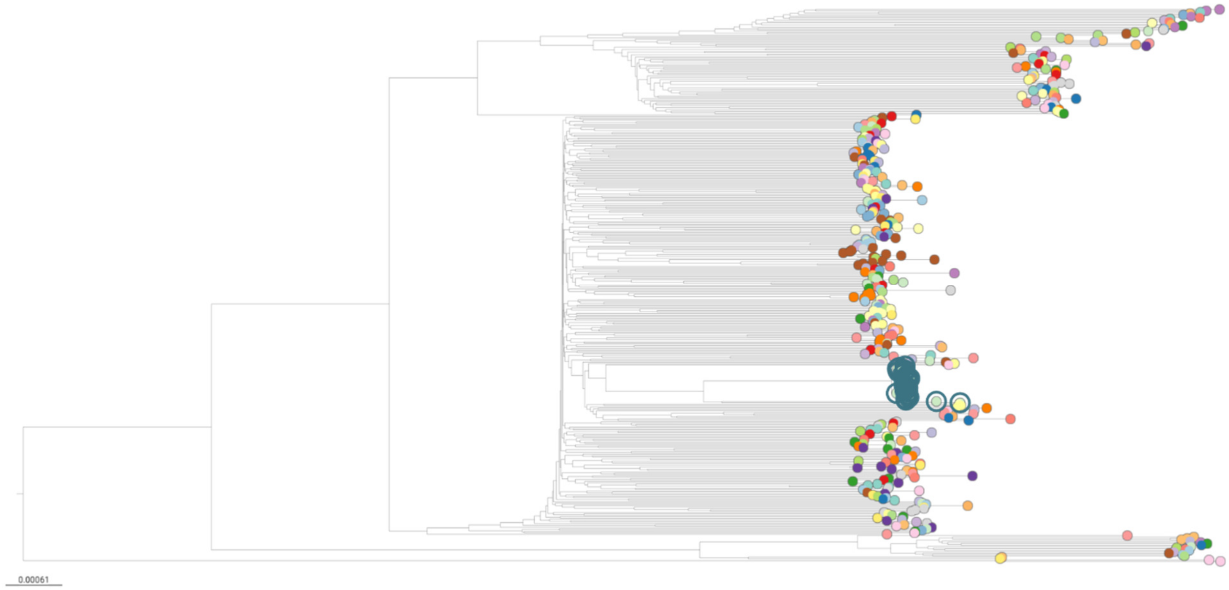
- Observed samples clustering and phylogenetic tree generation.



ID	STATUS	QUAL	RUN_ID	TAXA	CLUSTER	ISO_IN_CL	ISO_PASS_	MEAN_SN	MIN_SNP_	MAX_SNP_
2017-15-01_T1	NEW	PASS	1723940339	Pseudomc	13	25	25	1143	21	2306
2017-15-02_T1	NEW	PASS	1723940339	Pseudomc	13	25	25	1308	74	3078
2017-15-03_T1	NEW	PASS	1723940339	Pseudomc	13	25	25	758	60	1560
2017-15-08_T1	NEW	PASS	1723940339	Pseudomc	13	25	25	976	23	4396
2017-15-11_T1	NEW	PASS	1723940339	Pseudomc	13	25	25	632	16	3316
2017-15-15_T1	NEW	PASS	1723940339	Pseudomc	13	25	25	1950	15	5036
2017-15-161-01_T1	NEW	PASS	1723940339	Pseudomc	13	25	25	385	14	2896
2017-15-18_T1	NEW	PASS	1723940339	Pseudomc	13	25	25	1184	30	5168
2017-15-217_T1	NEW	PASS	1723940339	Pseudomc	13	25	25	1117	21	2310
2017-15-218_T1	NEW	PASS	1723940339	Pseudomc	13	25	25	3682	4	5460
2017-15-219_T1	NEW	PASS	1723940339	Pseudomc	13	25	25	701	32	4326
2017-15-220_T1	NEW	PASS	1723940339	Pseudomc	13	25	25	1278	49	3267
2017-15-221_T1	NEW	PASS	1723940339	Pseudomc	13	25	25	602	15	3973
2017-15-222_T1	NEW	PASS	1723940339	Pseudomc	13	25	25	1603	32	4937
2017-15-223_T1	NEW	PASS	1723940339	Pseudomc	13	25	25	1448	29	5468
2017-15-224_T1	NEW	PASS	1723940339	Pseudomc	13	25	25	1175	30	5174
2017-15-225_T1	NEW	PASS	1723940339	Pseudomc	13	25	25	630	14	3511
2017-15-42-01_T1	NEW	PASS	1723940339	Pseudomc	13	25	25	1001	17	4801
2017-15-45-01_T1	NEW	PASS	1723940339	Pseudomc	13	25	25	610	25	3223
2017-15-49-01_T1	NEW	PASS	1723940339	Pseudomc	13	25	25	663	15	3741
2017-15-51-02_T1	NEW	PASS	1723940339	Pseudomc	13	25	25	709	15	3732
2017-15-64-03_T1	NEW	PASS	1723940339	Pseudomc	13	25	25	966	25	4423
2017-15-67-02_T1	NEW	PASS	1723940339	Pseudomc	13	25	25	1447	29	5376
2017-15-69-01_T1	NEW	PASS	1723940339	Pseudomc	13	25	25	631	21	3584
2017-15-96-01_T1	NEW	PASS	1723940339	Pseudomc	13	25	25	1015	18	4913

ID: Sample identified with a suffix of T1.
STATUS: "New" samples refer to those recently analyzed, while "old" samples are historical isolates already in BigBacter database.
QUAL: Indicates whether the sample meets the BigBacter QC thresholds.
RUN_ID: Unique identifier for the test run.
TAXA: The species name associated with the sample.
CLUSTER: The cluster assignment for the sample, determined via PopPUNK.

ISO_IN_CLUSTER: Total number of isolates within the cluster.
ISO_PASS_QC: Total number of isolates that passed QC within the cluster.
MEAN_SNP_DIST_SNIPPY, MIN_SNP_DIST_SNIPPY, MAX_SNP_DIST_SNIPPY: Basic statistics regarding the genetic relatedness of these isolates.
STRONG_LINKAGE_SNIPPY, INTER_LINKAGE_SNIPPY: Summary of "strong" and "intermediate" genetic linkages based on pairwise SNP distance thresholds.

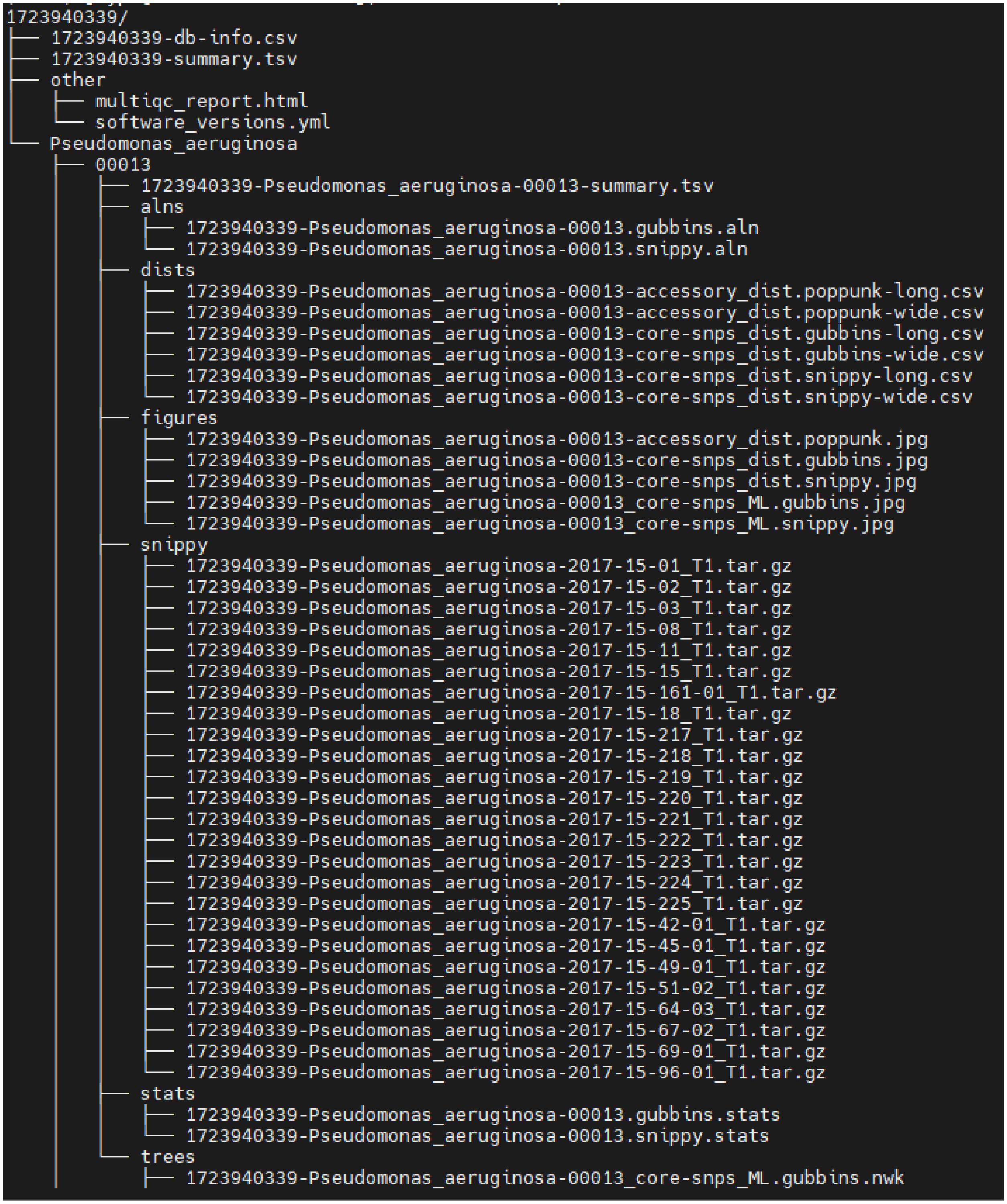


Execution Time:

- Duration: 17m 46s
- CPU hours: 6.7

Output Files:

Verified output files were generated correctly



```
1723940339/
├── 1723940339-db-info.csv
├── 1723940339-summary.tsv
├── other
│   ├── multiqc_report.html
│   └── software_versions.yml
└── Pseudomonas_aeruginosa
    ├── 00013
    │   ├── 1723940339-Pseudomonas_aeruginosa-00013-summary.tsv
    │   ├── alns
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-00013.gubbins.aln
    │   │   └── 1723940339-Pseudomonas_aeruginosa-00013.snippy.aln
    │   ├── dists
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-00013-accessory_dist.poppunk-long.csv
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-00013-accessory_dist.poppunk-wide.csv
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-00013-core-snps_dist.gubbins-long.csv
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-00013-core-snps_dist.gubbins-wide.csv
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-00013-core-snps_dist.snippy-long.csv
    │   │   └── 1723940339-Pseudomonas_aeruginosa-00013-core-snps_dist.snippy-wide.csv
    │   ├── figures
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-00013-accessory_dist.poppunk.jpg
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-00013-core-snps_dist.gubbins.jpg
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-00013-core-snps_dist.snippy.jpg
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-00013_core-snps_ML.gubbins.jpg
    │   │   └── 1723940339-Pseudomonas_aeruginosa-00013_core-snps_ML.snippy.jpg
    │   ├── snippy
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-2017-15-01_T1.tar.gz
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-2017-15-02_T1.tar.gz
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-2017-15-03_T1.tar.gz
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-2017-15-08_T1.tar.gz
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-2017-15-11_T1.tar.gz
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-2017-15-15_T1.tar.gz
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-2017-15-161-01_T1.tar.gz
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-2017-15-18_T1.tar.gz
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-2017-15-217_T1.tar.gz
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-2017-15-218_T1.tar.gz
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-2017-15-219_T1.tar.gz
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-2017-15-220_T1.tar.gz
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-2017-15-221_T1.tar.gz
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-2017-15-222_T1.tar.gz
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-2017-15-223_T1.tar.gz
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-2017-15-224_T1.tar.gz
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-2017-15-225_T1.tar.gz
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-2017-15-42-01_T1.tar.gz
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-2017-15-45-01_T1.tar.gz
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-2017-15-49-01_T1.tar.gz
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-2017-15-51-02_T1.tar.gz
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-2017-15-64-03_T1.tar.gz
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-2017-15-67-02_T1.tar.gz
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-2017-15-69-01_T1.tar.gz
    │   │   └── 1723940339-Pseudomonas_aeruginosa-2017-15-96-01_T1.tar.gz
    │   ├── stats
    │   │   ├── 1723940339-Pseudomonas_aeruginosa-00013.gubbins.stats
    │   │   └── 1723940339-Pseudomonas_aeruginosa-00013.snippy.stats
    │   └── trees
    │       ├── 1723940339-Pseudomonas_aeruginosa-00013_core-snps_ML.gubbins.nwk
```

2) Test 2 with Default PopPUNK Database

Sample Preparation:

- Species: *Pseudomonas aeruginosa*
- Sample number: 8
- Assembly: Phoenix filtered scaffolds

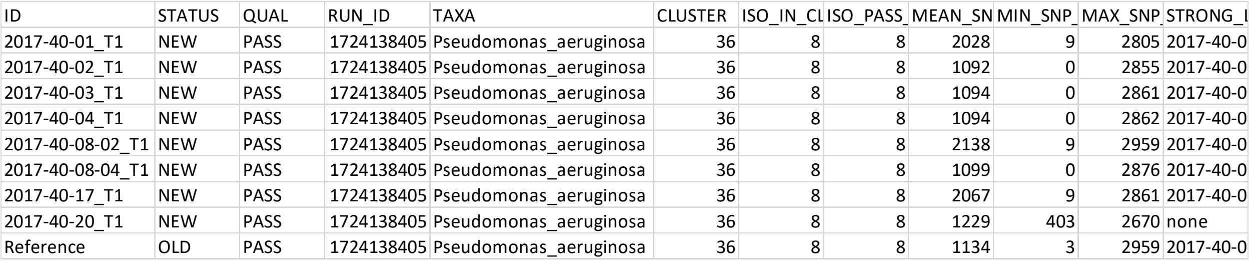
- Input data: SRA data/fastq files

Pipeline Execution:

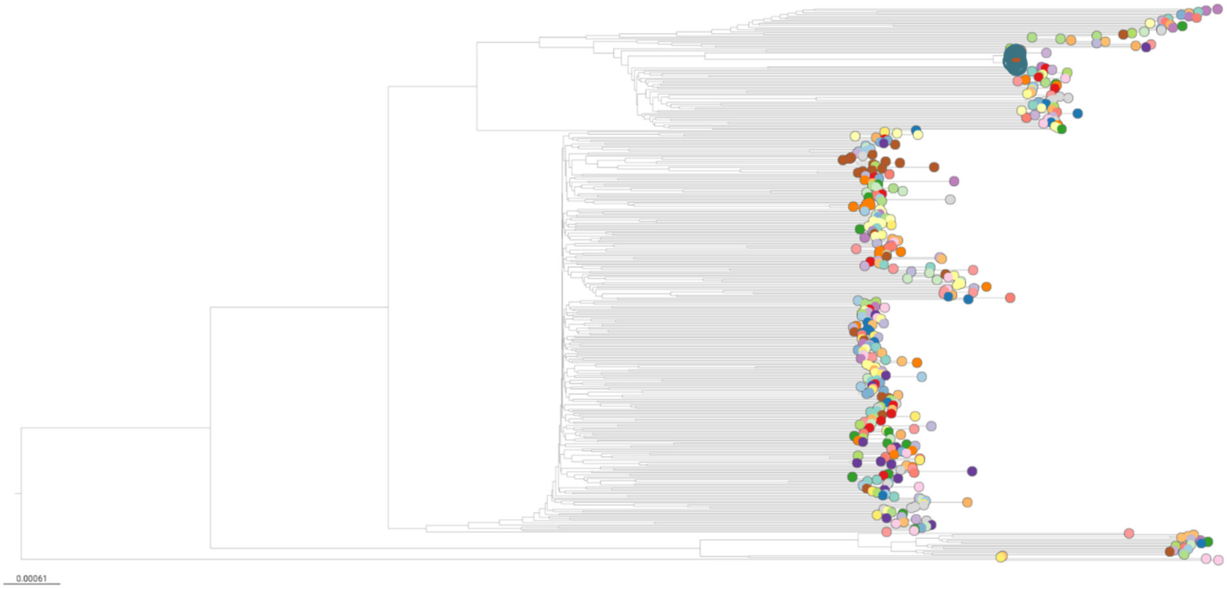
- Run the pipeline with the command

```
nextflow run $pipeline/bigbacter-nf \
   -profile singularity \
   --input ${PWD}/samplesheet.csv \
   --db $db_dir/db \
   --outdir $PWD/results/ \
   --max_cpus 20 \
   --max_memory '32.GB'
```

- Observed samples clustering and phylogenetic tree generation.



ID	STATUS	QUAL	RUN_ID	TAXA	CLUSTER	ISO_IN_Cl	ISO_PASS_	MEAN_SN	MIN_SNP_	MAX_SNP_	STRONG_l
2017-40-01_T1	NEW	PASS	1724138405	Pseudomonas_aeruginosa	36	8	8	2028	9	2805	2017-40-0
2017-40-02_T1	NEW	PASS	1724138405	Pseudomonas_aeruginosa	36	8	8	1092	0	2855	2017-40-0
2017-40-03_T1	NEW	PASS	1724138405	Pseudomonas_aeruginosa	36	8	8	1094	0	2861	2017-40-0
2017-40-04_T1	NEW	PASS	1724138405	Pseudomonas_aeruginosa	36	8	8	1094	0	2862	2017-40-0
2017-40-08-02_T1	NEW	PASS	1724138405	Pseudomonas_aeruginosa	36	8	8	2138	9	2959	2017-40-0
2017-40-08-04_T1	NEW	PASS	1724138405	Pseudomonas_aeruginosa	36	8	8	1099	0	2876	2017-40-0
2017-40-17_T1	NEW	PASS	1724138405	Pseudomonas_aeruginosa	36	8	8	2067	9	2861	2017-40-0
2017-40-20_T1	NEW	PASS	1724138405	Pseudomonas_aeruginosa	36	8	8	1229	403	2670	none
Reference	OLD	PASS	1724138405	Pseudomonas_aeruginosa	36	8	8	1134	3	2959	2017-40-0

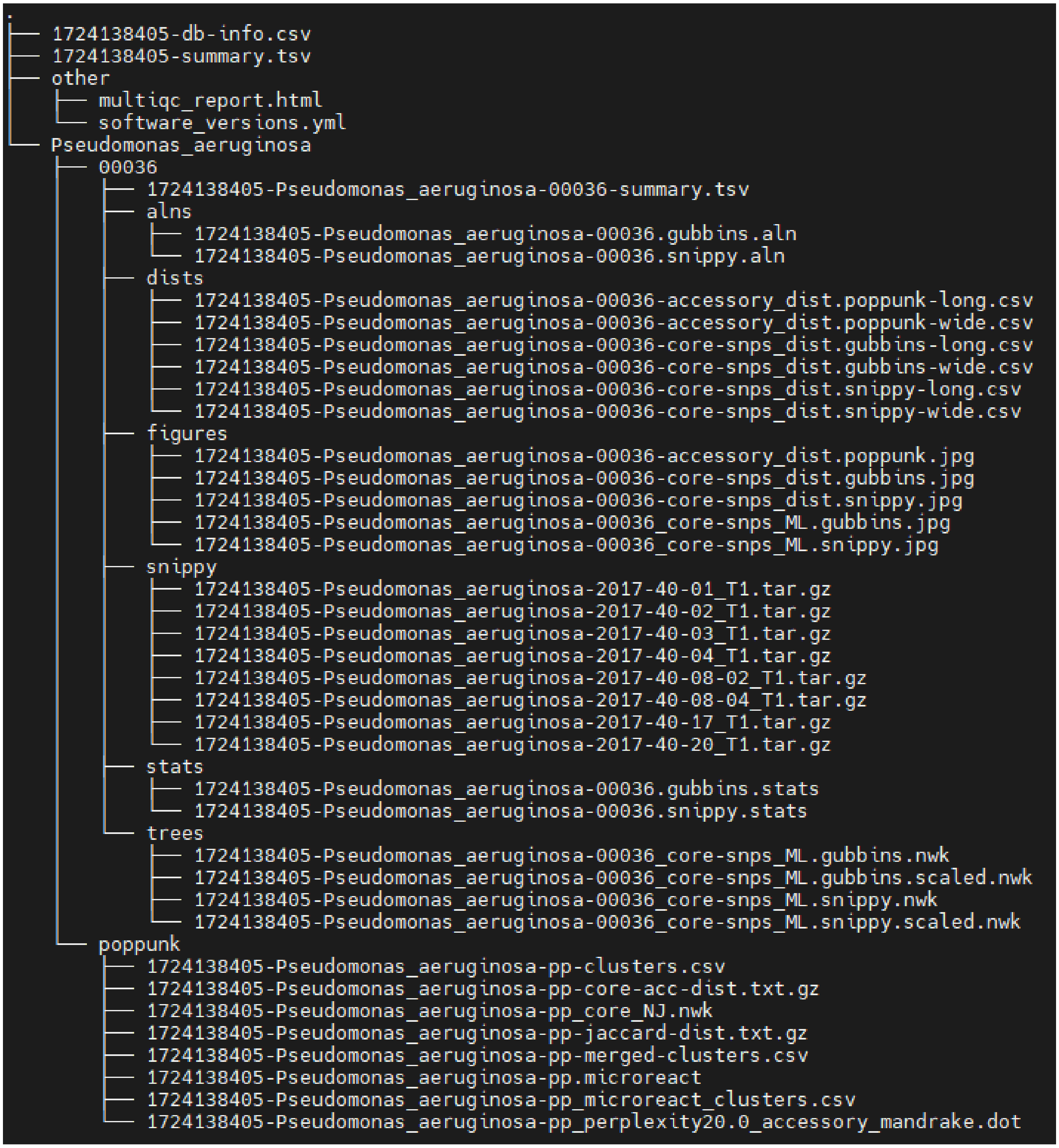


Execution Time:

- Duration: 6m 18s
- CPU hours: 2.3

Output Files:

Verified output files were generated correctly



```
.
├── 1724138405-db-info.csv
├── 1724138405-summary.tsv
├── other
│   ├── multiqc_report.html
│   └── software_versions.yml
└── Pseudomonas_aeruginosa
    ├── 00036
    │   ├── 1724138405-Pseudomonas_aeruginosa-00036-summary.tsv
    │   ├── alns
    │   │   ├── 1724138405-Pseudomonas_aeruginosa-00036.gubbins.aln
    │   │   └── 1724138405-Pseudomonas_aeruginosa-00036.snippy.aln
    │   ├── dists
    │   │   ├── 1724138405-Pseudomonas_aeruginosa-00036-accessory_dist.poppunk-long.csv
    │   │   ├── 1724138405-Pseudomonas_aeruginosa-00036-accessory_dist.poppunk-wide.csv
    │   │   ├── 1724138405-Pseudomonas_aeruginosa-00036-core-snps_dist.gubbins-long.csv
    │   │   ├── 1724138405-Pseudomonas_aeruginosa-00036-core-snps_dist.gubbins-wide.csv
    │   │   ├── 1724138405-Pseudomonas_aeruginosa-00036-core-snps_dist.snippy-long.csv
    │   │   └── 1724138405-Pseudomonas_aeruginosa-00036-core-snps_dist.snippy-wide.csv
    │   ├── figures
    │   │   ├── 1724138405-Pseudomonas_aeruginosa-00036-accessory_dist.poppunk.jpg
    │   │   ├── 1724138405-Pseudomonas_aeruginosa-00036-core-snps_dist.gubbins.jpg
    │   │   ├── 1724138405-Pseudomonas_aeruginosa-00036-core-snps_dist.snippy.jpg
    │   │   ├── 1724138405-Pseudomonas_aeruginosa-00036_core-snps_ML.gubbins.jpg
    │   │   └── 1724138405-Pseudomonas_aeruginosa-00036_core-snps_ML.snippy.jpg
    │   ├── snippy
    │   │   ├── 1724138405-Pseudomonas_aeruginosa-2017-40-01_T1.tar.gz
    │   │   ├── 1724138405-Pseudomonas_aeruginosa-2017-40-02_T1.tar.gz
    │   │   ├── 1724138405-Pseudomonas_aeruginosa-2017-40-03_T1.tar.gz
    │   │   ├── 1724138405-Pseudomonas_aeruginosa-2017-40-04_T1.tar.gz
    │   │   ├── 1724138405-Pseudomonas_aeruginosa-2017-40-08-02_T1.tar.gz
    │   │   ├── 1724138405-Pseudomonas_aeruginosa-2017-40-08-04_T1.tar.gz
    │   │   ├── 1724138405-Pseudomonas_aeruginosa-2017-40-17_T1.tar.gz
    │   │   └── 1724138405-Pseudomonas_aeruginosa-2017-40-20_T1.tar.gz
    │   ├── stats
    │   │   ├── 1724138405-Pseudomonas_aeruginosa-00036.gubbins.stats
    │   │   └── 1724138405-Pseudomonas_aeruginosa-00036.snippy.stats
    │   └── trees
    │       ├── 1724138405-Pseudomonas_aeruginosa-00036_core-snps_ML.gubbins.nwk
    │       ├── 1724138405-Pseudomonas_aeruginosa-00036_core-snps_ML.gubbins.scaled.nwk
    │       ├── 1724138405-Pseudomonas_aeruginosa-00036_core-snps_ML.snippy.nwk
    │       └── 1724138405-Pseudomonas_aeruginosa-00036_core-snps_ML.snippy.scaled.nwk
    └── poppunk
        ├── 1724138405-Pseudomonas_aeruginosa-pp-clusters.csv
        ├── 1724138405-Pseudomonas_aeruginosa-pp-core-acc-dist.txt.gz
        ├── 1724138405-Pseudomonas_aeruginosa-pp_core_NJ.nwk
        ├── 1724138405-Pseudomonas_aeruginosa-pp-jaccard-dist.txt.gz
        ├── 1724138405-Pseudomonas_aeruginosa-pp-merged-clusters.csv
        ├── 1724138405-Pseudomonas_aeruginosa-pp.microreact
        ├── 1724138405-Pseudomonas_aeruginosa-pp_microreact_clusters.csv
        └── 1724138405-Pseudomonas_aeruginosa-pp_perplexity20.0_accessory_mandrake.dot
```

3) Test 3: Create PopPUNK databases

Preparing the PopPUNK Database:

- Followed instructions to create a new database for a non-default species.

```
nextflow run $pipeline/bigbacter-nf \
  -profile singularity \
  -entry PREPARE_DB \
  --input ${PWD}/pp_db_list.csv \
  --db $db_dir/db2 \
```

- Verified database integration into the pipeline

Sample Preparation:

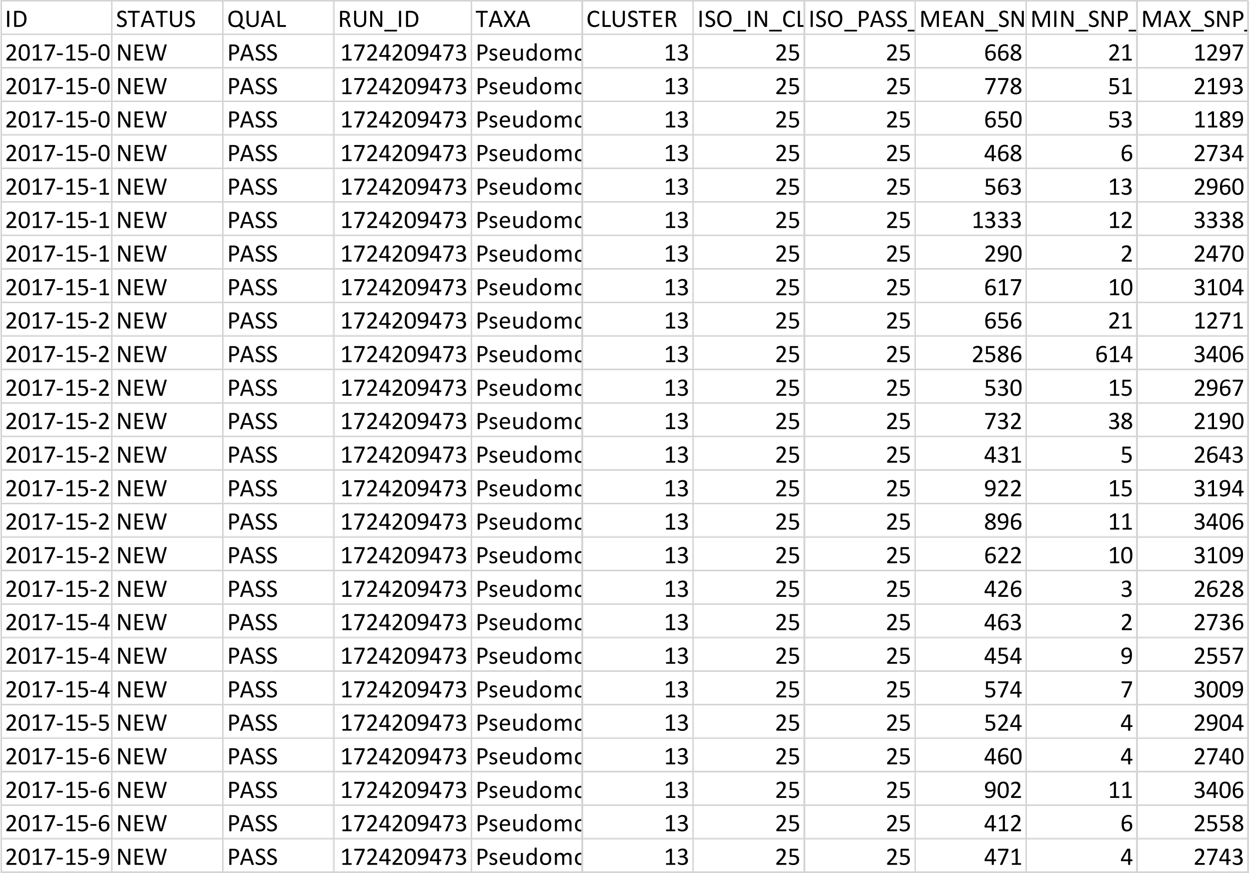
- Species: *Pseudomonas aeruginosa*
- Sample number: 25
- Assembly: Phoenix filtered scaffolds
- Input data: fastq files

Pipeline Execution:

- Run the pipeline with the command

```
nextflow run $pipeline/bigbacter-nf \
  -profile singularity \
  --input ${PWD}/samplesheet.csv \
  --db $db_dir/db2 \
  --outdir $PWD/results/ \
  --max_cpus 20 \
  --max_memory '32.GB'
```

- Observed samples clustering

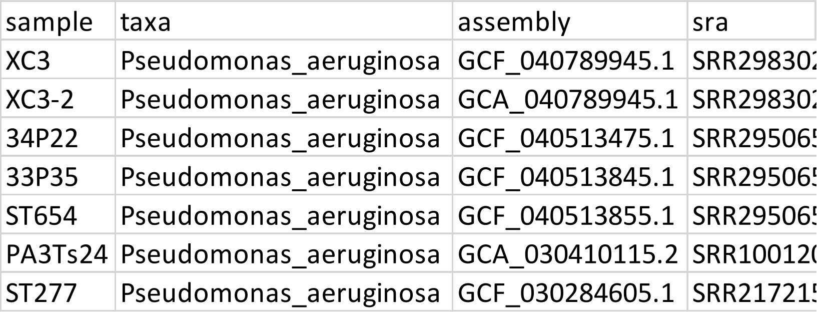


ID	STATUS	QUAL	RUN_ID	TAXA	CLUSTER	ISO_IN_Cl	ISO_PASS_	MEAN_SN	MIN_SNP_	MAX_SNP_
2017-15-0	NEW	PASS	1724209473	Pseudomc	13	25	25	668	21	1297
2017-15-0	NEW	PASS	1724209473	Pseudomc	13	25	25	778	51	2193
2017-15-0	NEW	PASS	1724209473	Pseudomc	13	25	25	650	53	1189
2017-15-0	NEW	PASS	1724209473	Pseudomc	13	25	25	468	6	2734
2017-15-1	NEW	PASS	1724209473	Pseudomc	13	25	25	563	13	2960
2017-15-1	NEW	PASS	1724209473	Pseudomc	13	25	25	1333	12	3338
2017-15-1	NEW	PASS	1724209473	Pseudomc	13	25	25	290	2	2470
2017-15-1	NEW	PASS	1724209473	Pseudomc	13	25	25	617	10	3104
2017-15-2	NEW	PASS	1724209473	Pseudomc	13	25	25	656	21	1271
2017-15-2	NEW	PASS	1724209473	Pseudomc	13	25	25	2586	614	3406
2017-15-2	NEW	PASS	1724209473	Pseudomc	13	25	25	530	15	2967
2017-15-2	NEW	PASS	1724209473	Pseudomc	13	25	25	732	38	2190
2017-15-2	NEW	PASS	1724209473	Pseudomc	13	25	25	431	5	2643
2017-15-2	NEW	PASS	1724209473	Pseudomc	13	25	25	922	15	3194
2017-15-2	NEW	PASS	1724209473	Pseudomc	13	25	25	896	11	3406
2017-15-2	NEW	PASS	1724209473	Pseudomc	13	25	25	622	10	3109
2017-15-2	NEW	PASS	1724209473	Pseudomc	13	25	25	426	3	2628
2017-15-4	NEW	PASS	1724209473	Pseudomc	13	25	25	463	2	2736
2017-15-4	NEW	PASS	1724209473	Pseudomc	13	25	25	454	9	2557
2017-15-4	NEW	PASS	1724209473	Pseudomc	13	25	25	574	7	3009
2017-15-5	NEW	PASS	1724209473	Pseudomc	13	25	25	524	4	2904
2017-15-6	NEW	PASS	1724209473	Pseudomc	13	25	25	460	4	2740
2017-15-6	NEW	PASS	1724209473	Pseudomc	13	25	25	902	11	3406
2017-15-6	NEW	PASS	1724209473	Pseudomc	13	25	25	412	6	2558
2017-15-9	NEW	PASS	1724209473	Pseudomc	13	25	25	471	4	2743

4) Test 4: Run samples with GenBank and SRA accession

Preparing sample file:

- Followed instructions to create a file with accession number



sample	taxa	assembly	sra
XC3	Pseudomonas_aeruginosa	GCF_040789945.1	SRR298302
XC3-2	Pseudomonas_aeruginosa	GCA_040789945.1	SRR298302
34P22	Pseudomonas_aeruginosa	GCF_040513475.1	SRR295065
33P35	Pseudomonas_aeruginosa	GCF_040513845.1	SRR295065
ST654	Pseudomonas_aeruginosa	GCF_040513855.1	SRR295065
PA3Ts24	Pseudomonas_aeruginosa	GCA_030410115.2	SRR100120
ST277	Pseudomonas_aeruginosa	GCF_030284605.1	SRR217215

Sample Preparation:

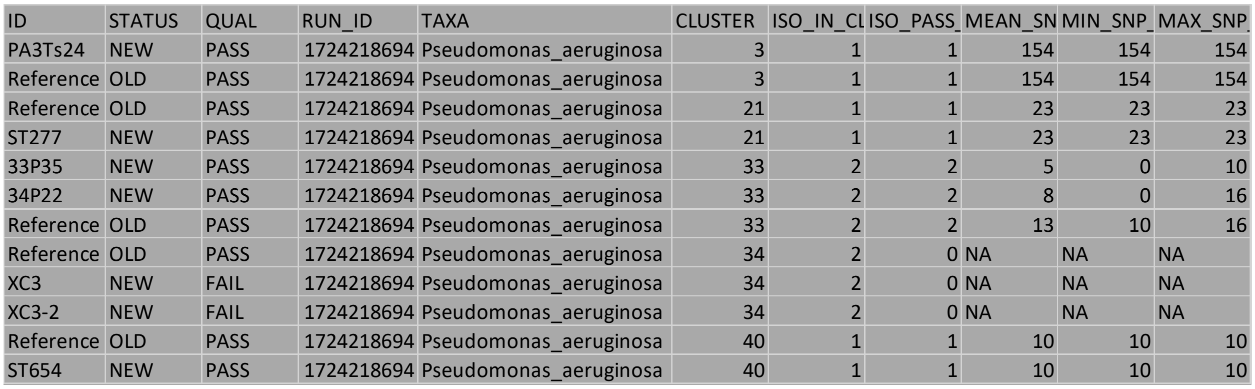
- Species: *Pseudomonas aeruginosa*
- Sample number: 7
- Assembly: GeneBank/RefSeq assembly file
- Input data: SRA number

Pipeline Execution:

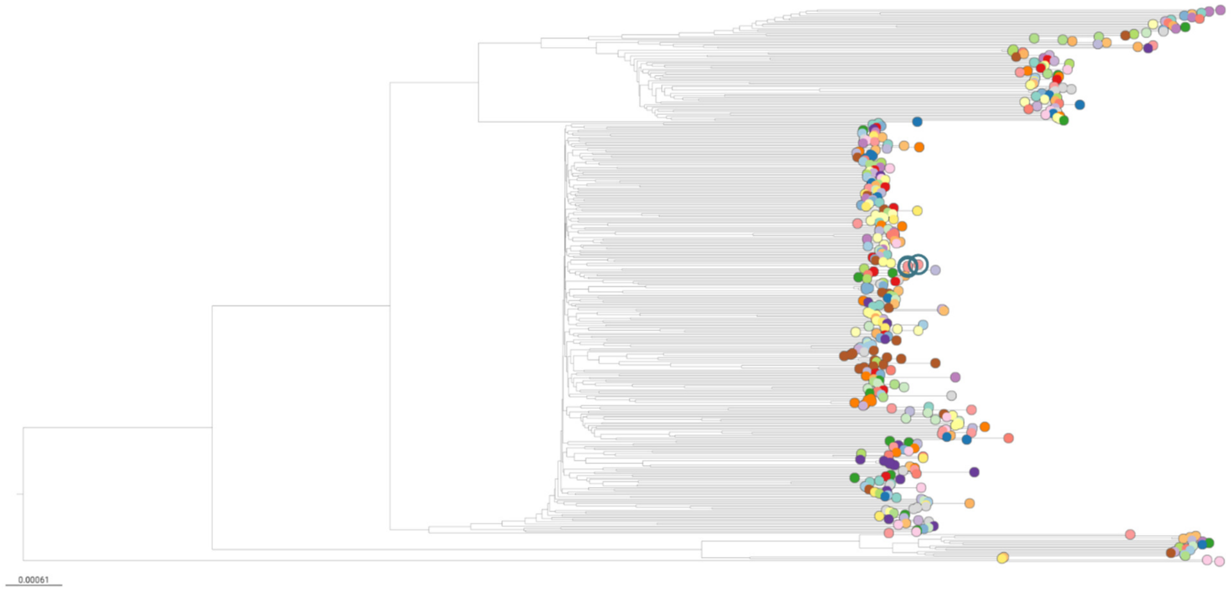
- Run the pipeline with the command

```
nextflow run $pipeline/bigbacter-nf \
  -profile singularity \
  --ncbi ${PWD}/samplesheet.csv \
  --db $db_dir/db2 \
  --outdir $PWD/results/ \
  -resume \
  --max_cpus 20 \
  --max_memory '32.GB'
```

- Observed samples clustering and phylogenetic tree generation

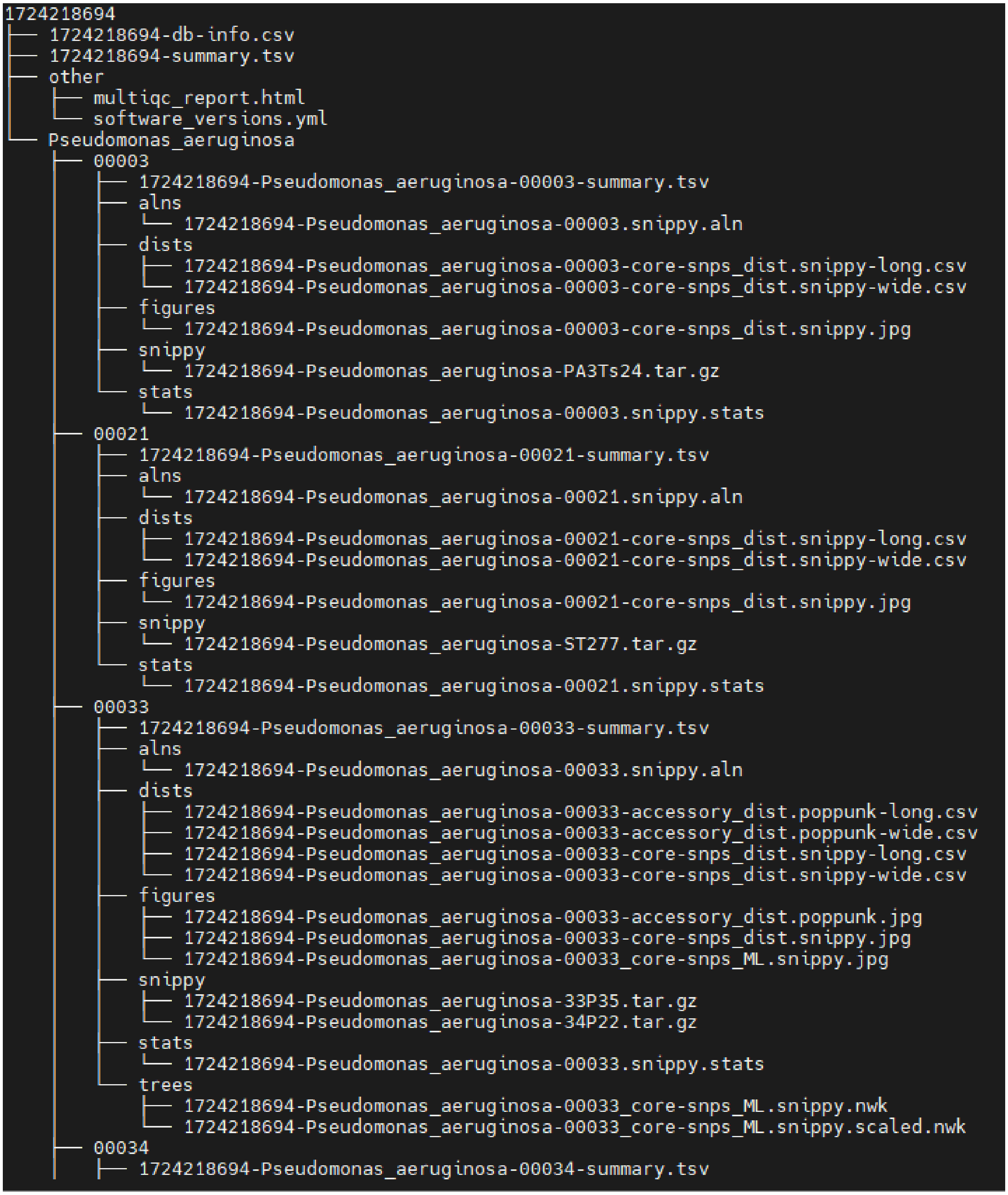


ID	STATUS	QUAL	RUN_ID	TAXA	CLUSTER	ISO_IN_CL	ISO_PASS_	MEAN_SN	MIN_SNP_	MAX_SNP_
PA3Ts24	NEW	PASS	1724218694	Pseudomonas_aeruginosa	3	1	1	154	154	154
Reference	OLD	PASS	1724218694	Pseudomonas_aeruginosa	3	1	1	154	154	154
Reference	OLD	PASS	1724218694	Pseudomonas_aeruginosa	21	1	1	23	23	23
ST277	NEW	PASS	1724218694	Pseudomonas_aeruginosa	21	1	1	23	23	23
33P35	NEW	PASS	1724218694	Pseudomonas_aeruginosa	33	2	2	5	0	10
34P22	NEW	PASS	1724218694	Pseudomonas_aeruginosa	33	2	2	8	0	16
Reference	OLD	PASS	1724218694	Pseudomonas_aeruginosa	33	2	2	13	10	16
Reference	OLD	PASS	1724218694	Pseudomonas_aeruginosa	34	2	0	NA	NA	NA
XC3	NEW	FAIL	1724218694	Pseudomonas_aeruginosa	34	2	0	NA	NA	NA
XC3-2	NEW	FAIL	1724218694	Pseudomonas_aeruginosa	34	2	0	NA	NA	NA
Reference	OLD	PASS	1724218694	Pseudomonas_aeruginosa	40	1	1	10	10	10
ST654	NEW	PASS	1724218694	Pseudomonas_aeruginosa	40	1	1	10	10	10



Output Files:

Verified output files were generated correctly



```
1724218694
├── 1724218694-db-info.csv
├── 1724218694-summary.tsv
├── other
│   ├── multiqc_report.html
│   └── software_versions.yml
└── Pseudomonas_aeruginosa
    ├── 00003
    │   ├── 1724218694-Pseudomonas_aeruginosa-00003-summary.tsv
    │   ├── alns
    │   │   └── 1724218694-Pseudomonas_aeruginosa-00003.snippy.aln
    │   ├── dists
    │   │   ├── 1724218694-Pseudomonas_aeruginosa-00003-core-snps_dist.snippy-long.csv
    │   │   └── 1724218694-Pseudomonas_aeruginosa-00003-core-snps_dist.snippy-wide.csv
    │   ├── figures
    │   │   └── 1724218694-Pseudomonas_aeruginosa-00003-core-snps_dist.snippy.jpg
    │   ├── snippy
    │   │   └── 1724218694-Pseudomonas_aeruginosa-PA3Ts24.tar.gz
    │   └── stats
    │       └── 1724218694-Pseudomonas_aeruginosa-00003.snippy.stats
    ├── 00021
    │   ├── 1724218694-Pseudomonas_aeruginosa-00021-summary.tsv
    │   ├── alns
    │   │   └── 1724218694-Pseudomonas_aeruginosa-00021.snippy.aln
    │   ├── dists
    │   │   ├── 1724218694-Pseudomonas_aeruginosa-00021-core-snps_dist.snippy-long.csv
    │   │   └── 1724218694-Pseudomonas_aeruginosa-00021-core-snps_dist.snippy-wide.csv
    │   ├── figures
    │   │   └── 1724218694-Pseudomonas_aeruginosa-00021-core-snps_dist.snippy.jpg
    │   ├── snippy
    │   │   └── 1724218694-Pseudomonas_aeruginosa-ST277.tar.gz
    │   └── stats
    │       └── 1724218694-Pseudomonas_aeruginosa-00021.snippy.stats
    ├── 00033
    │   ├── 1724218694-Pseudomonas_aeruginosa-00033-summary.tsv
    │   ├── alns
    │   │   └── 1724218694-Pseudomonas_aeruginosa-00033.snippy.aln
    │   ├── dists
    │   │   ├── 1724218694-Pseudomonas_aeruginosa-00033-accessory_dist.poppunk-long.csv
    │   │   ├── 1724218694-Pseudomonas_aeruginosa-00033-accessory_dist.poppunk-wide.csv
    │   │   ├── 1724218694-Pseudomonas_aeruginosa-00033-core-snps_dist.snippy-long.csv
    │   │   └── 1724218694-Pseudomonas_aeruginosa-00033-core-snps_dist.snippy-wide.csv
    │   ├── figures
    │   │   ├── 1724218694-Pseudomonas_aeruginosa-00033-accessory_dist.poppunk.jpg
    │   │   ├── 1724218694-Pseudomonas_aeruginosa-00033-core-snps_dist.snippy.jpg
    │   │   └── 1724218694-Pseudomonas_aeruginosa-00033_core-snps_ML.snippy.jpg
    │   ├── snippy
    │   │   ├── 1724218694-Pseudomonas_aeruginosa-33P35.tar.gz
    │   │   └── 1724218694-Pseudomonas_aeruginosa-34P22.tar.gz
    │   ├── stats
    │   │   └── 1724218694-Pseudomonas_aeruginosa-00033.snippy.stats
    │   └── trees
    │       ├── 1724218694-Pseudomonas_aeruginosa-00033_core-snps_ML.snippy.nwk
    │       └── 1724218694-Pseudomonas_aeruginosa-00033_core-snps_ML.snippy.scaled.nwk
    ├── 00034
    │   ├── 1724218694-Pseudomonas_aeruginosa-00034-summary.tsv
```

5) Test 5: Push the new changes

Sample Preparation:

- Species: *Pseudomonas aeruginosa*
- Sample number: 8

- Assembly: Phoenix filtered scaffolds
- Input data: fastq files

Pipeline Execution:

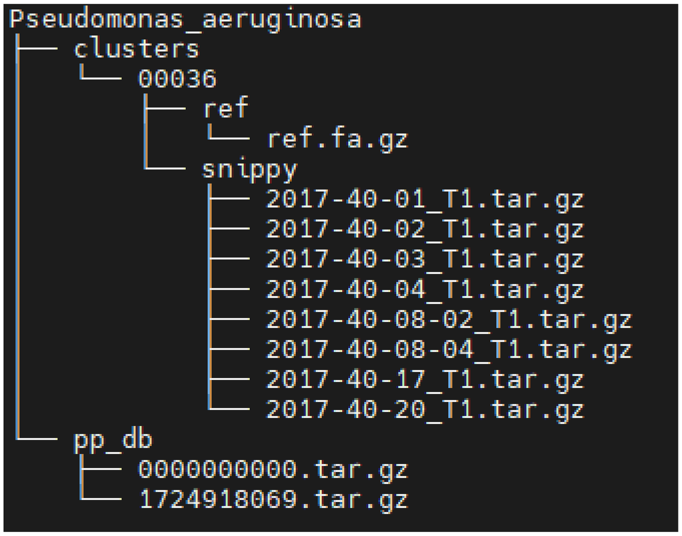
- Run the pipeline with the command

```
nextflow run $pipeline/bigbacter-nf \
   -profile singularity \
   --input ${PWD}/samplesheet.csv \
   --db $db_dir/db \
   --outdir $PWD/results/ \
   -push true \
   -resume
```

- Samples clustering and phylogenetic tree is the same as test 2

Output Files:

Verified database files were updated correctly



```
Pseudomonas_aeruginosa
├── clusters
│   └── 00036
│       ├── ref
│       │   └── ref.fa.gz
│       └── snippy
│           ├── 2017-40-01_T1.tar.gz
│           ├── 2017-40-02_T1.tar.gz
│           ├── 2017-40-03_T1.tar.gz
│           ├── 2017-40-04_T1.tar.gz
│           ├── 2017-40-08-02_T1.tar.gz
│           ├── 2017-40-08-04_T1.tar.gz
│           ├── 2017-40-17_T1.tar.gz
│           └── 2017-40-20_T1.tar.gz
└── pp_db
    ├── 0000000000.tar.gz
    └── 1724918069.tar.gz
```

3. Performance:

- With --max_cpus 20 and --max_memory 32.GB, processing a single sample is completed within 42 to 47 seconds
- The accuracy of BigBacter is high. During the outbreak tests, samples were successfully clustered together in two separate tests, demonstrating the pipeline’s effectiveness in grouping related isolates.
- No significant issues were encountered during pipeline execution

Issues Found:

- The connection may be disconnected when attempting to download all genome files for a single species from the PopPUNK databases due to the large file size. However, the reference file is sufficient for most tests. As an alternative, we can manually download all genome files without using the pipeline.

Recommendations and Feedback Based on Test Results:

1) Requirement Placement:

Consider placing requirements at the top of the documentation for easy reference, rather than being buried in subdirectories or specific files like singularity version, and contig file. In addition, clearly listing the required inputs, including species-level classification, will be helpful. Finally, in the sample sheet, sample names cannot contain spaces or special characters – consider making it clear for end users in the wiki that these (i.e., spaces, special characters) should not be used in sample names.

2) Species Database Coverage:

While PopPUNK databases for 23 bacterial species are provided, it would be helpful to have clear instructions provided on how to handle species not included in the default database.

3) Contribution and Contact Information:

Consider adding a dedicated section in the documentation for contribution guidelines and contact information to facilitate user engagement and support.

4) Running BigBacter page:

On the wiki page "2. Running BigBacter," the instructions could be more concise. The current presentation of multiple Step 1 and Step 2 sections was confusing during our beta testing.

5) Demo Test Data:

Consider providing demo test data within the repository to help users validate their setup and understand the expected outputs.

6) Summary Format:

Ensure that summary outputs are available in CSV or Excel format for ease of use and integration with other analysis tools. Additionally, while the summary table has headers for each column, the wiki document lacks clear explanations for these headers, which may lead to confusion for users.

7) Phylogenetic tree:

The phylogenetic tree provides an informative landscape perspective. However, if you are considering or aiming to use this approach for outbreak investigations, where SNV-level differences may be critical, consider if adding a component that will provide more focused analyses (i.e., phylogenetic analyses of only those isolates in a cluster of interest) and granular output (e.g., SNV matrices and size or proportion of the reference genome SNVs were called from) from analyses of only those isolates within the same cluster of interest, or nwk files that only include those clustering within the “landscape tree”). These additional and more granular details will be critical to assist in the outbreak investigation. This can be especially important for opportunistic bacterial pathogens that may also be found in the local environment because these isolates may appear to cluster at a higher level but are actually unrelated from a transmission/outbreak perspective.

Conclusion:

BigBacter demonstrates strong capabilities in bacterial genomic surveillance, offering efficient processing and comprehensive documentation. However, the documentation could benefit from additional details.