15 Dec 2025

BigBacter

Advanced Molecular Detection

Southeast Region Bioinformatics

This resource was made possible through funding provided under the Epidemiology and Laboratory Capacity for Prevention and Control of Emerging Infectious Diseases (ELC) Cooperative Agreement (CK24-0002), Project D: Advanced Molecular Detection to the Florida Department of Health. The conclusions, findings, and opinions expressed by authors do not necessarily reflect the official position of the U.S. Department of Health and Human Services, the Public Health Service, or the Centers for Disease Control and Prevention.

Overview

Purpose:

- A bacterial-genomic surveillance pipeline, this workflow builds and maintains population-level reference databases, performs accessory-genome clustering and core-genome SNV analysis, and generates phylogenies, distance matrices, and relatedness data to assess genetic relationships among isolates.

Usage:

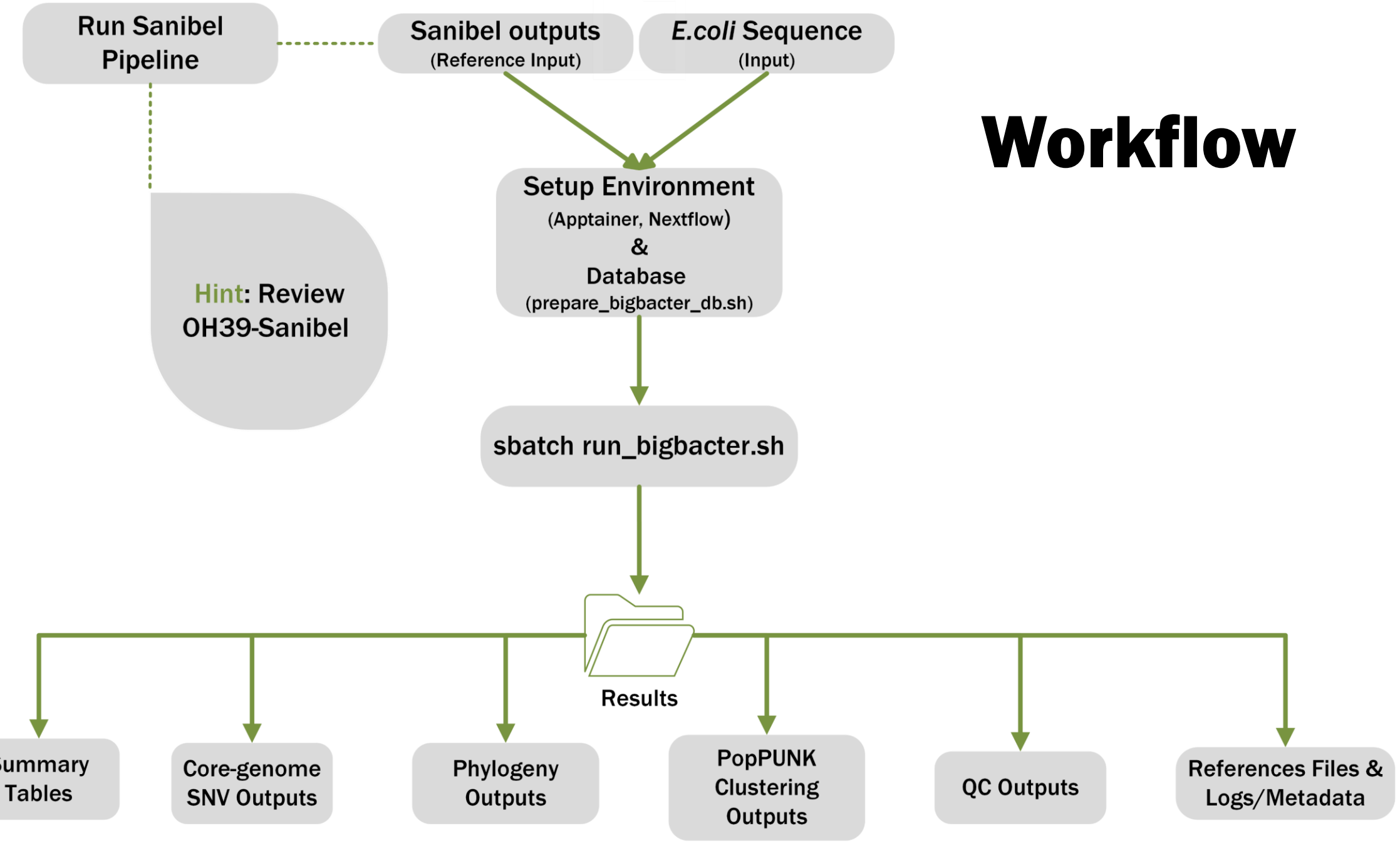
- It is used by public-health and molecular-epidemiology teams to add new bacterial isolates, assign them to clusters, build per-isolate and combined phylogenetic trees, compute SNV and genome-distance matrices, and create QC plus summary reports — for outbreak detection, surveillance, and long-term genomic monitoring.

Dependencies:

- Nextflow + Apptainer/Singularity

Advanced Molecular Detection
Southeast Region Bioinformatics

Workflow



Application

Objective

Use *E.coli* sequences and Sanibel output data with BigBacter to generate SNVs, PopPunk Clustering, and Phylogeny outputs .

Advanced Molecular Detection
Southeast Region Bioinformatics

Application Cont.

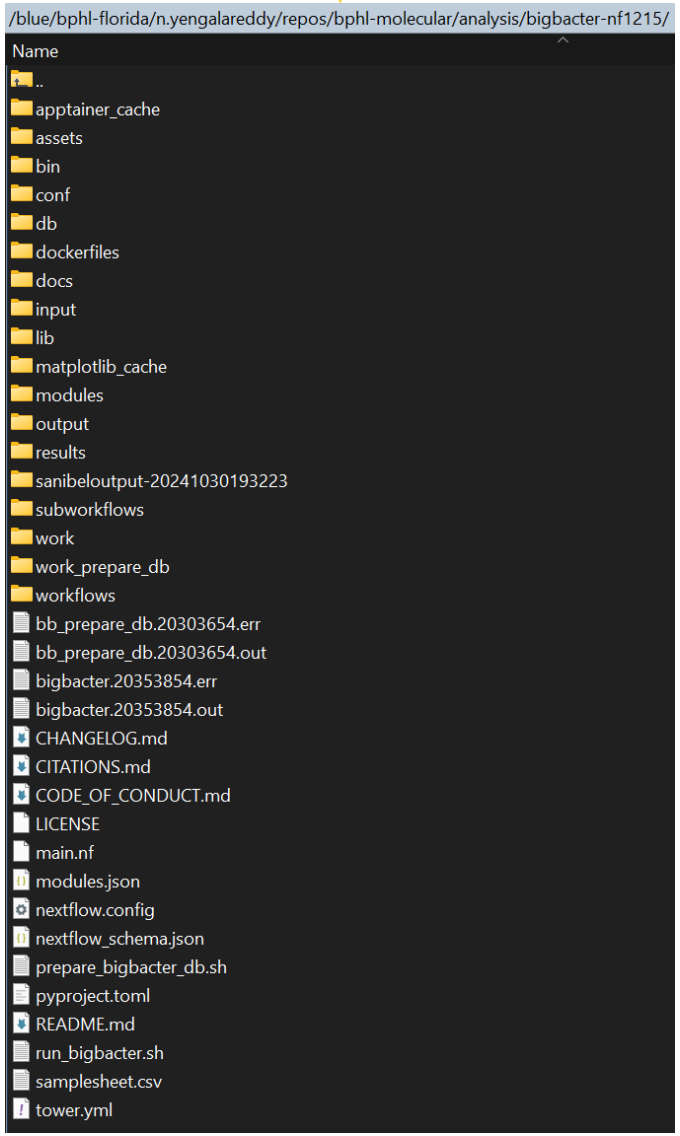
```
cd /blue/bphl-<state>/<user>/repos/bphl-molecular/

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

mkdir analysis/

cd analysis/

cp /blue/bphl-<state>/<user>/repos/bphl-molecular/bigbacter-nf/
```



Application Cont.

```
 GNU nano 5.6.1                                  prepare_bigbacter_db.sh
#!/bin/bash
#SBATCH --account=bphl-umbrella
#SBATCH --qos=bphl-umbrella
#SBATCH --job-name=bb_prepare_db_ecoli
#SBATCH --mail-type=END,FAIL
#SBATCH --mail-user=nikhil.yengala@flhealth.gov
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=4
#SBATCH --mem=16G
#SBATCH --time=24:00:00
#SBATCH --output=bb_prepare_db.%j.out
#SBATCH --error=bb_prepare_db.%j.err

module purge
module load nextflow
module load apptainer

BASE=/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/bigbacter-nf1208
DBDIR="$BASE/db"

export NXF_WORK="$BASE/work_prepare_db"
export NXF_SINGULARITY_CACHEDIR="$BASE/apptainer_cache"
export MPLCONFIGDIR="$BASE/matplotlib_cache"
export NXF_SINGULARITY_HOME_MOUNT=true

mkdir -p "$DBDIR" "$NXF_WORK" "$NXF_SINGULARITY_CACHEDIR" "$MPLCONFIGDIR"

cd "$BASE"

nextflow run DOH-JDJ0303/bigbacter-nf \
  -r main \
  -profile singularity,escherichia_coli_db \
  -entry PREPARE_DB \
  --db "$DBDIR" \
  --max_cpus 4 \
  --max_memory 16.GB
```

/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/bigbacter-nf1215/

Name

- ..
- apptainer_cache
- assets
- bin
- conf
- db
- dockerfiles
- docs
- input
- lib
- matplotlib_cache
- modules
- output
- results
- sanibeloutput-20241030193223
- subworkflows
- work
- work_prepare_db
- workflows
- bb_prepare_db.20303654.err
- bb_prepare_db.20303654.out
- bigbacter.20353854.err
- bigbacter.20353854.out
- CHANGELOG.md
- CITATIONS.md
- CODE_OF_CONDUCT.md
- LICENSE
- main.nf
- modules.json
- nextflow.config
- nextflow_schema.json
- prepare_bigbacter_db.sh
- pyproject.toml
- README.md
- run_bigbacter.sh
- samplesheet.csv
- tower.yml

Advanced Molecular Detection
Southeast Region Bioinformatics

Application Cont.

GNU nano 5.6.1 samplesheet.csv

```
sample,taxa,assembly,fastq_1,fastq_2
J4,Escherichia_coli,/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/bigbacter-nf1208/sanibeloutput-20241030193223/J4/J4_assembly/J4.fasta,/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/bigbacter-nf120
J7,Escherichia_coli,/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/bigbacter-nf1208/sanibeloutput-20241030193223/J7/J7_assembly/J7.fasta,/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/bigbacter-nf120
J8,Escherichia_coli,/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/bigbacter-nf1208/sanibeloutput-20241030193223/J8/J8_assembly/J8.fasta,/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/bigbacter-nf120
J9,Escherichia_coli,/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/bigbacter-nf1208/sanibeloutput-20241030193223/J9/J9_assembly/J9.fasta,/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/bigbacter-nf120
J17,Escherichia_coli,/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/bigbacter-nf1208/sanibeloutput-20241030193223/J17/J17_assembly/J17.fasta,/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/bigbacter-n
J52,Escherichia_coli,/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/bigbacter-nf1208/sanibeloutput-20241030193223/J52/J52_assembly/J52.fasta,/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/bigbacter-n
J100,Escherichia_coli,/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/bigbacter-nf1208/sanibeloutput-20241030193223/J100/J100_assembly/J100.fasta,/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/bigbact
```

/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/bigbacter-nf1215/db/Escherichia_coli/pp_db/

Name
..
0000000000.tar.gz

GNU nano 5.6.1 run_bigbacter.sh

```
#!/bin/bash
#SBATCH --account=bphl-umbrella
#SBATCH --qos=bphl-umbrella
#SBATCH --job-name=bigbacter_ecoli
#SBATCH --mail-type=END,FAIL
#SBATCH --mail-user=nikhil.yengala@flhealth.gov
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=24
#SBATCH --mem=128G
#SBATCH --time=48:00:00
#SBATCH --output=bigbacter.%j.out
#SBATCH --error=bigbacter.%j.err

module purge
module load nextflow
module load apptainer

BASE=/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/bigbacter-nf1208
SAMPLESHEET="$BASE/samplesheet.csv"
OUTDIR="$BASE/output"
DBDIR="$BASE/db"

export NXF_WORK="$BASE/work"
export NXF_SINGULARITY_CACHEDIR="$BASE/apptainer_cache"
export MPLCONFIGDIR="$BASE/matplotlib_cache"
export NXF_SINGULARITY_HOME_MOUNT=true

mkdir -p "$OUTDIR" "$NXF_WORK" "$NXF_SINGULARITY_CACHEDIR" "$MPLCONFIGDIR"

cd "$BASE"

nextflow run DOH-JDJ0303/bigbacter-nf \
  -r main \
  -profile singularity \
  --input "$SAMPLESHEET" \
  --outdir "$OUTDIR" \
  --db "$DBDIR" \
  --threads "$SLURM_CPUS_PER_TASK" \
  --max_cpus 24 \
  --max_memory 128.GB
```

/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/bigbacter-nf1215/

Name
..
apptainer_cache
assets
bin
conf
db
dockerfiles
docs
input
lib
matplotlib_cache
modules
output
results
sanibeloutput-20241030193223
subworkflows
work
work_prepare_db
workflows
bb_prepare_db.20303654.err
bb_prepare_db.20303654.out
bigbacter.20353854.err
bigbacter.20353854.out
CHANGELOG.md
CITATIONS.md
CODE_OF_CONDUCT.md
LICENSE
main.nf
modules.json
nextflow.config
nextflow_schema.json
prepare_bigbacter_db.sh
pyproject.toml
README.md
run_bigbacter.sh
samplesheet.csv
tower.yml

sbatch
prepare_bigbacter_db.sh

Advanced Molecular Detection
Southeast Region Bioinformatics

Application Cont.

/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/bigbacter-nf1215/

Name

- ..
- apptainer_cache
- assets
- bin
- conf
- db
- dockerfiles
- docs
- input
- lib
- matplotlib_cache
- modules
- output
- results
- sanibeloutput-20241030193223
- subworkflows
- work
- work_prepare_db
- workflows
- bb_prepare_db.20303654.err
- bb_prepare_db.20303654.out
- bigbacter.20353854.err
- bigbacter.20353854.out
- CHANGELOG.md
- CITATIONS.md
- CODE_OF_CONDUCT.md
- LICENSE
- main.nf
- modules.json
- nextflow.config
- nextflow_schema.json
- prepare_bigbacter_db.sh
- pyproject.toml
- README.md
- run_bigbacter.sh
- samplesheet.csv
- tower.yml

/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/bigbacter-nf1215/output/1764971800/

Name

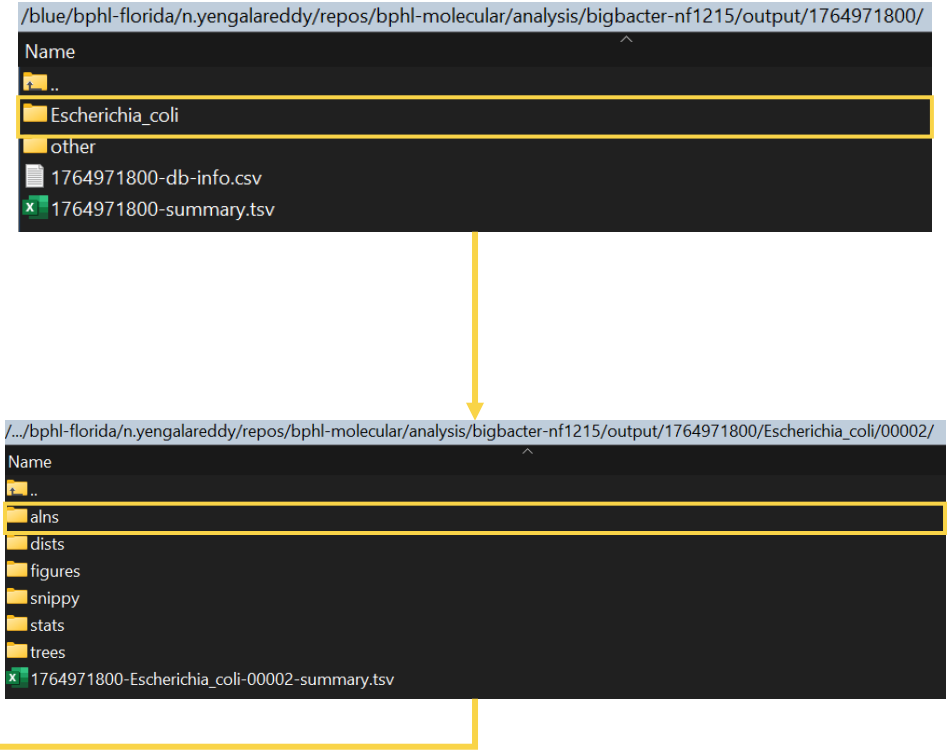
- ..
- Escherichia_coli
- other
- 1764971800-db-info.csv
- 1764971800-summary.tsv

ID	STATUS	QUAL	RUN_ID	TAXA	CLUSTER	ISO_IN_CLUSTER	ISO_PASS_QC	MEAN_SNP_DIST_SNIPPY	MIN_SNP_DIST_SNIPPY	MAX_SNP_DIST_SNIPPY	STRONG_LINKAGE_SNIPPY	INTER_LINKAGE_SNIPPY	MEAN_SNP_DIST_GUBBINS	MIN_SNP_DIST_GUBBINS	MAX_SNP_DIST_GUBBINS	STRONG_LINKAGE_GUBBINS	INTER_LINKAGE_GUBBINS	LENGTH	ALIGNED	UNALIGNED	RECOMB	VARIANT	HET	MASKED	LOWCOV	PER_GENFRAC	PER_LOWCOV	PER_HET
J7_T1	NEW	PASS	1.765E+09	Escherichia_coli	2	2	2	3510	9	7011	Reference	none	not performed	not performed	not performed	not performed	not performed	5E+06	5E+06	2022	not performec	9	#####	0	2353	99.5116	0.0447891	0.40508
J8_T1	NEW	PASS	1.765E+09	Escherichia_coli	2	2	2	7132	7011	7252	none	none	not performed	not performed	not performed	not performed	not performed	5E+06	5E+06	324318	not performec	7020	#####	0	20399	92.9468	0.388292	0.49154
Reference	OLD	PASS	1.765E+09	Escherichia_coli	2	2	2	3630	9	7252	J7_T1	none	not performed	not performed	not performed	not performed	not performed	5E+06	5E+06	0	not performec	0	0	0	0	100	0	0
J4_T1	NEW	PASS	1.765E+09	Escherichia_coli	5	2	2	5782	4	11560	Reference	none	not performed	not performed	not performed	not performed	not performed	5E+06	5E+06	425	not performec	4	#####	0	763	99.6074	0.0146395	0.36977
J52_T1	NEW	PASS	1.765E+09	Escherichia_coli	5	2	2	11672	11560	11784	none	none	not performed	not performed	not performed	not performed	not performed	5E+06	5E+06	659114	not performec	11564	#####	0	22001	86.5754	0.422129	0.35618
Reference	OLD	PASS	1.765E+09	Escherichia_coli	5	2	2	5894	4	11784	J4_T1	none	not performed	not performed	not performed	not performed	not performed	5E+06	5E+06	0	not performec	0	0	0	0	100	0	0
J17_T1	NEW	PASS	1.765E+09	Escherichia_coli	6	1	1	46	46	46	none	Reference	not performed	not performed	not performed	not performed	not performed	5E+06	5E+06	3136	not performec	46	#####	0	3076	99.4795	0.0584237	0.40251
Reference	OLD	PASS	1.765E+09	Escherichia_coli	6	1	1	46	46	46	none	J17_T1	not performed	not performed	not performed	not performed	not performed	5E+06	5E+06	0	not performec	0	0	0	0	100	0	0
J100_T1	NEW	PASS	1.765E+09	Escherichia_coli	12	1	1	2	2	2	Reference	none	not performed	not performed	not performed	not performed	not performed	5E+06	5E+06	346	not performec	2	#####	0	695	99.622	0.0143087	0.35656
Reference	OLD	PASS	1.765E+09	Escherichia_coli	12	1	1	2	2	2	J100_T1	none	not performed	not performed	not performed	not performed	not performed	5E+06	5E+06	0	not performec	0	0	0	0	100	0	0
J9_T1	NEW	PASS	1.765E+09	Escherichia_coli	627	1	1	3	3	3	Reference	none	not performed	not performed	not performed	not performed	not performed	5E+06	5E+06	62	not performec	3	#####	0	440	99.6499	0.00942095	0.33939
Reference	OLD	PASS	1.765E+09	Escherichia_coli	627	1	1	3	3	3	J9_T1	none	not performed	not performed	not performed	not performed	not performed	5E+06	5E+06	0	not performec	0	0	0	0	100	0	0

```
sbatch run_bigbacter.sh
```

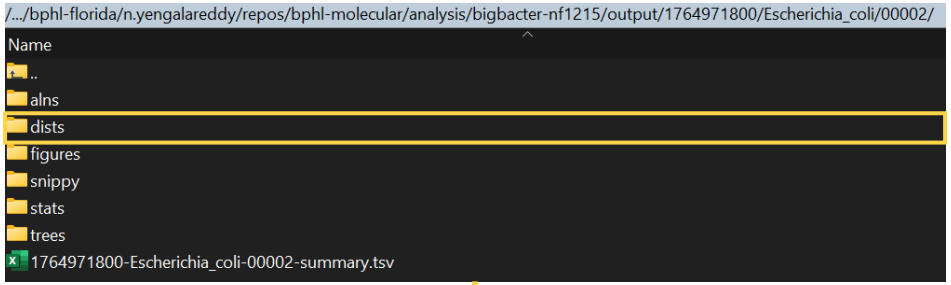
Advanced Molecular Detection
Southeast Region Bioinformatics

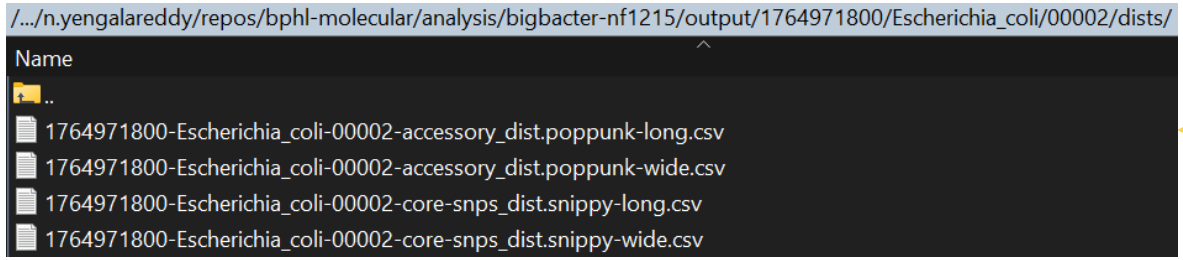
Application Cont.



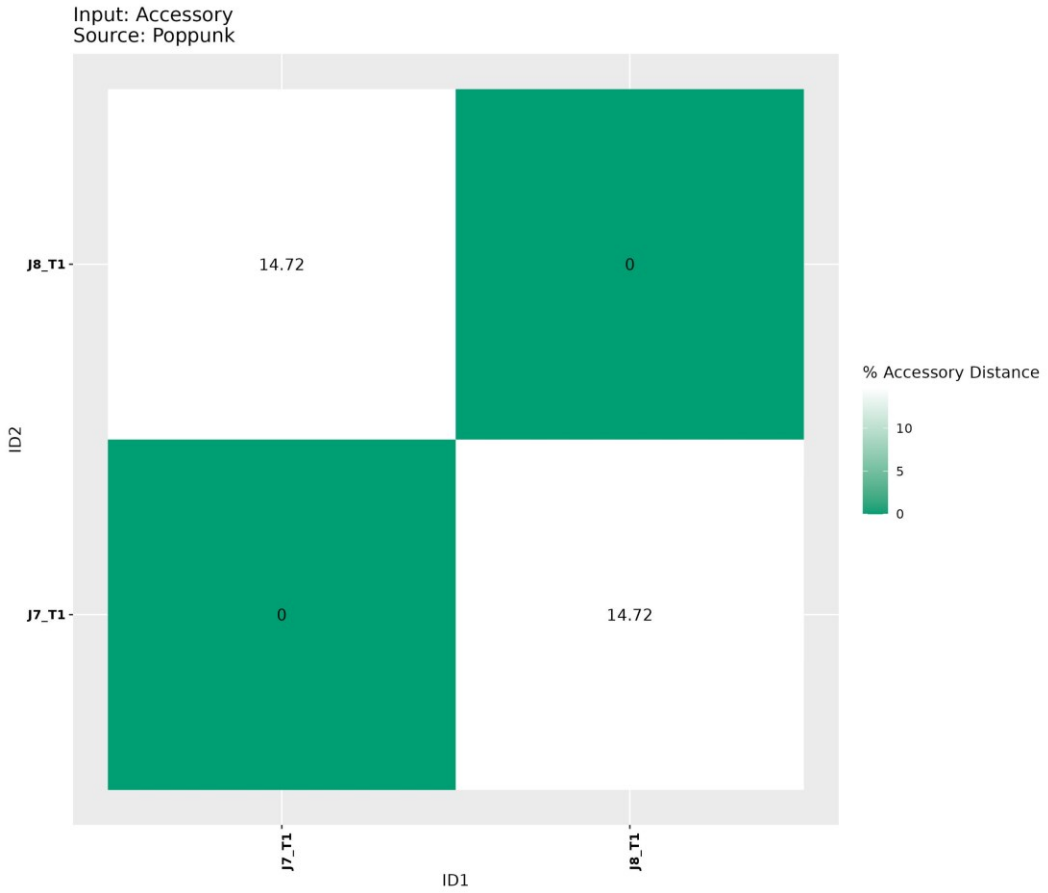
Advanced Molecular Detection
Southeast Region Bioinformatics

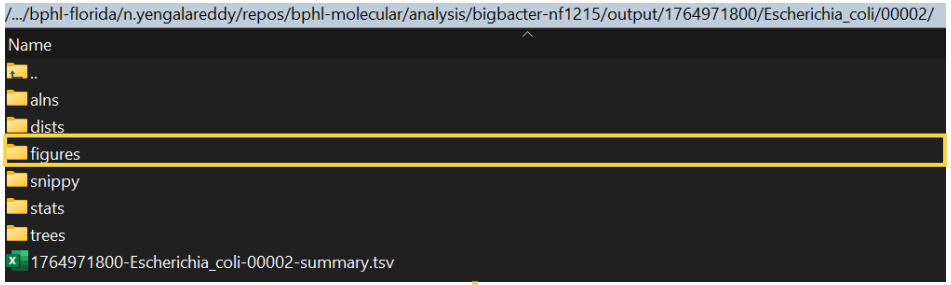
Application Cont.

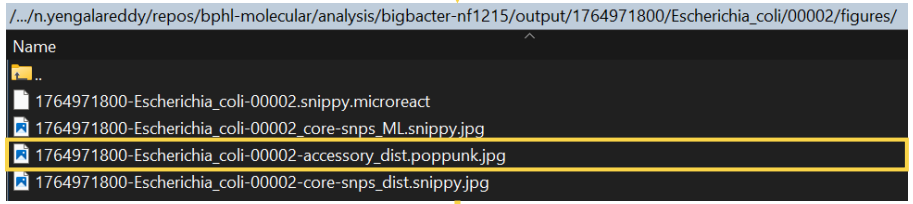




Application Cont.

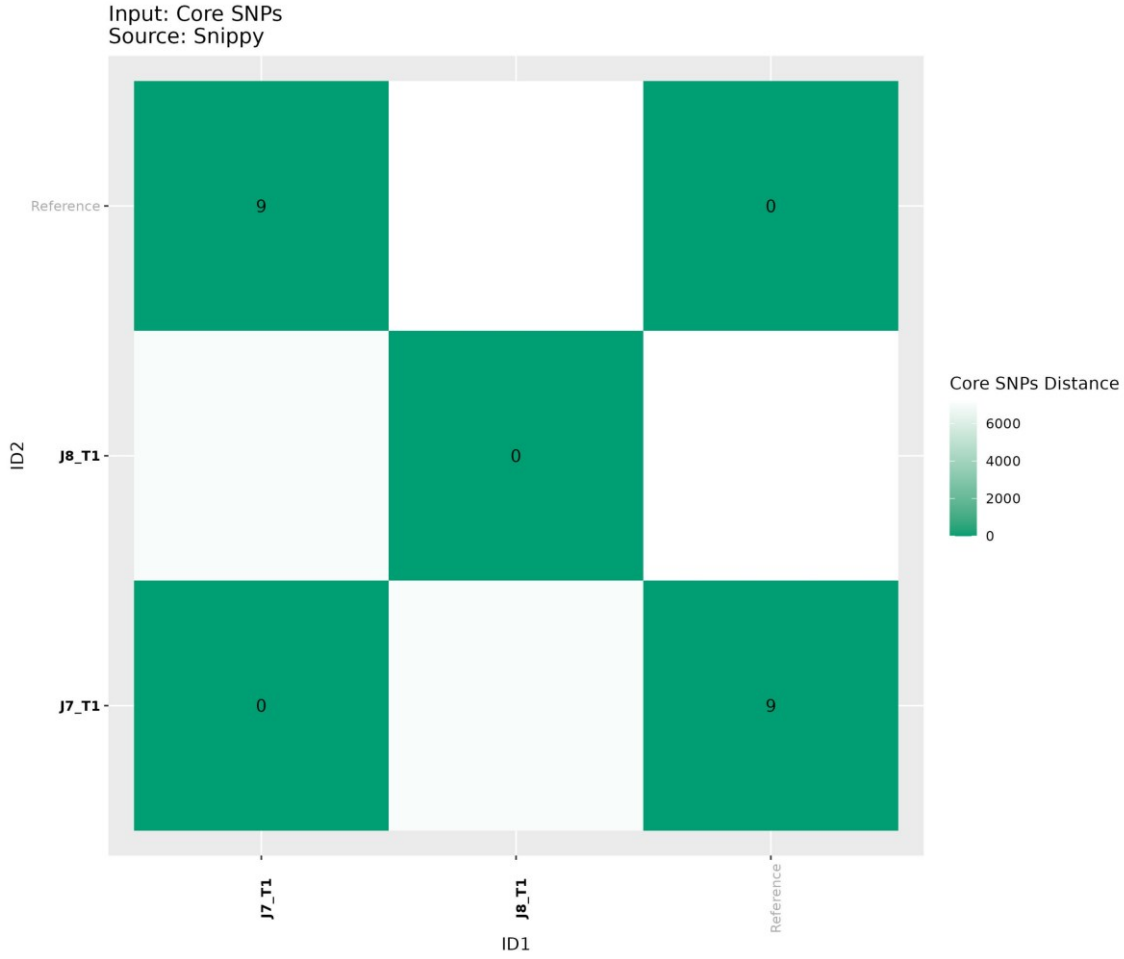


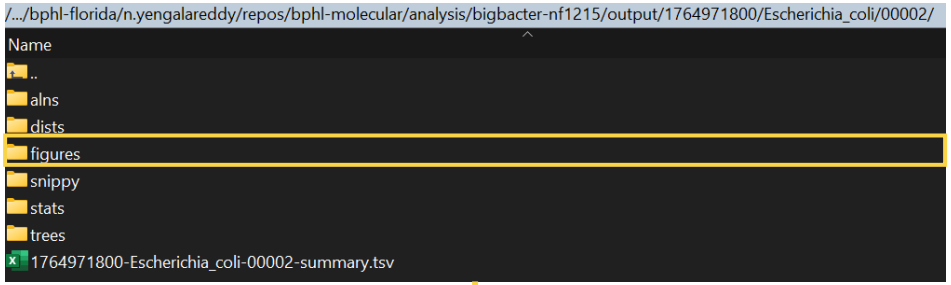


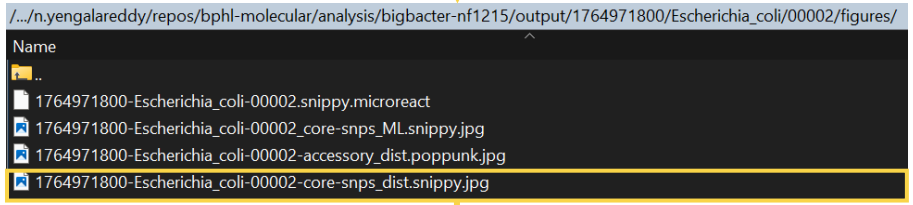


Advanced Molecular Detection
Southeast Region Bioinformatics

Application Cont.

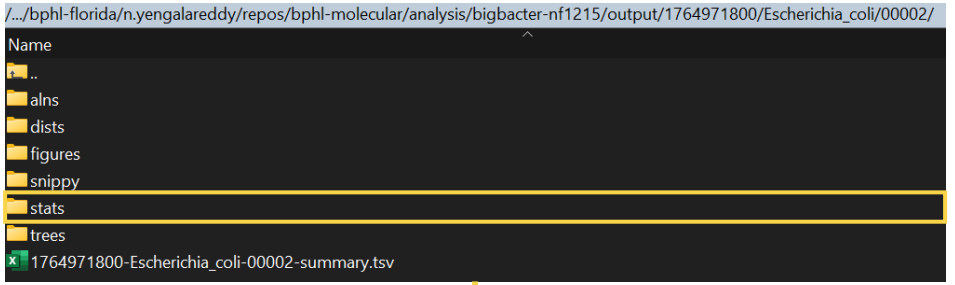


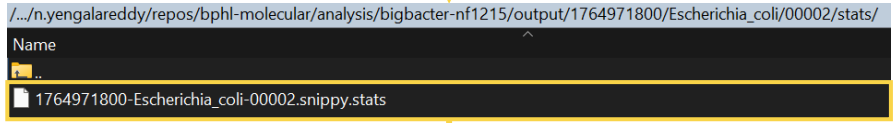




Advanced Molecular Detection
Southeast Region Bioinformatics

Application Cont.





ID	LENGTH	ALIGNED	UNALIGNED	VARIANT	HET	MASKED	LOWCOV	PER_GENFRAC	PER_LOWCOV	PER_HET	QUAL
J7_T1	5253514	5227858	2022	9	21281	0	2353	99.5116	0.0447891	0.405081	PASS
J8_T1	5253514	4882974	324318	7020	25823	0	20399	92.9468	0.388292	0.491538	PASS
Reference	5253514	5253514	0	0	0	0	0	100	0	0	PASS

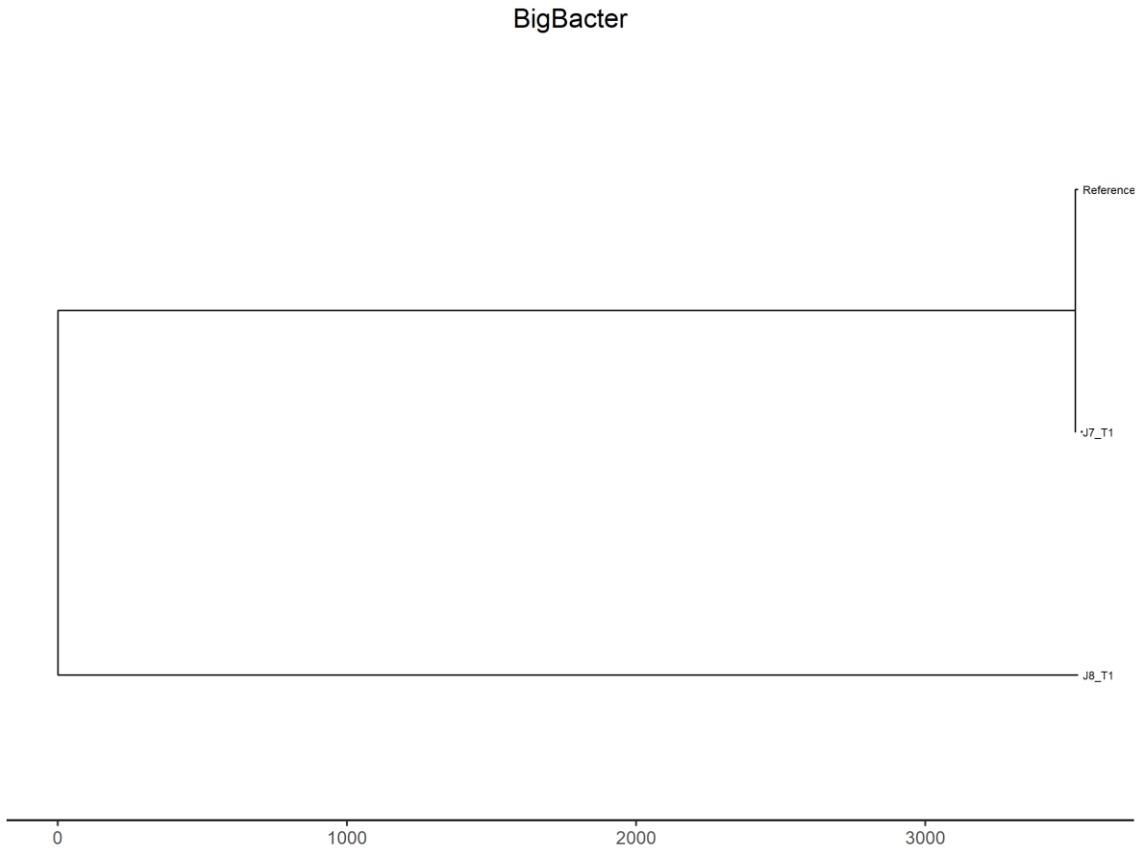
Application Cont.

/.../bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/bigbacter-nf1215/output/1764971800/Escherichia_coli/00002/

Name
..
alns
dists
figures
snippy
stats
trees
1764971800-Escherichia_coli-00002-summary.tsv

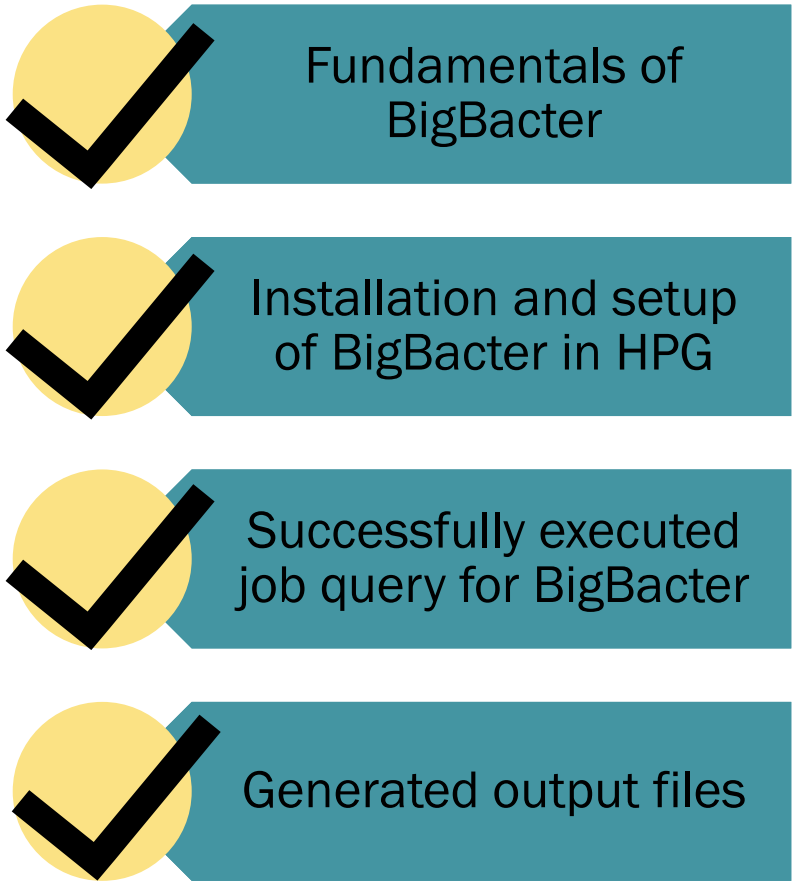
/.../n.yengalareddy/repos/bphl-molecular/analysis/bigbacter-nf1215/output/1764971800/Escherichia_coli/00002/trees/

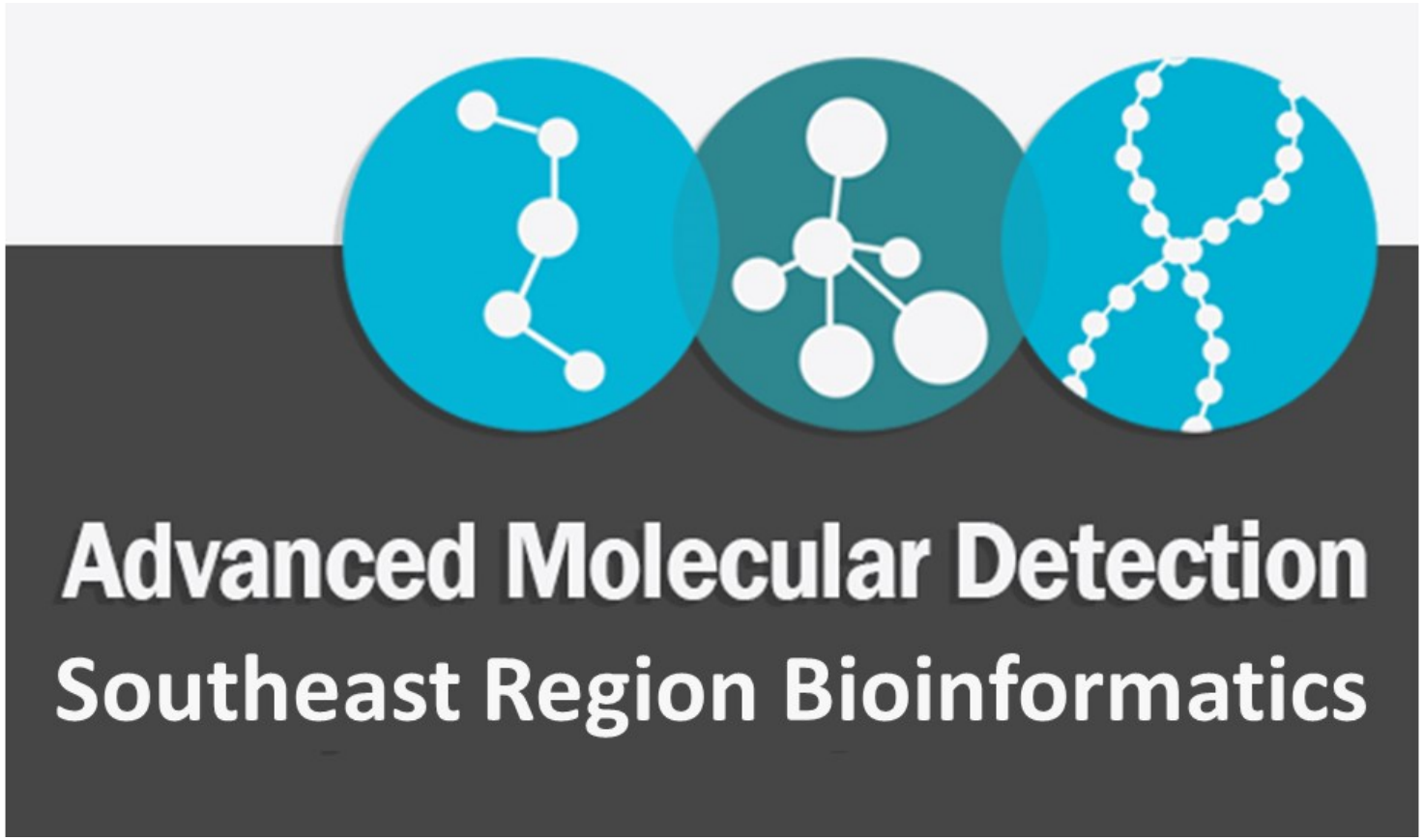
Name
..
1764971800-Escherichia_coli-00002_core-snps_ML.snippy.final.nwk
1764971800-Escherichia_coli-00002_core-snps_ML.snippy.nwk



Advanced Molecular Detection
Southeast Region Bioinformatics

Conclusion


Fundamentals of BigBacter
Installation and setup of BigBacter in HPG
Successfully executed job query for BigBacter
Generated output files



Questions?

bphl-sebioinformatics@flhealth.gov

Molly Mitchell, PhD
Bioinformatics Supervisor
Molly.Mitchell@flhealth.gov

Nikhil Reddy, MS
Bioinformatician
Nikhil.Yengala@flhealth.gov

Sam Bernhoft, MPH
Bioinformatician
Samantha.bernhoft@flhealth.gov