05 Jan 2026

mashwrapper

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

- This Nextflow pipeline uses Mash-based k-mer comparisons to rapidly screen sequencing reads against reference genomes to identify the most likely organism, such as SARS-CoV-2 and/or *Legionella*, from raw sequencing data.

Usage

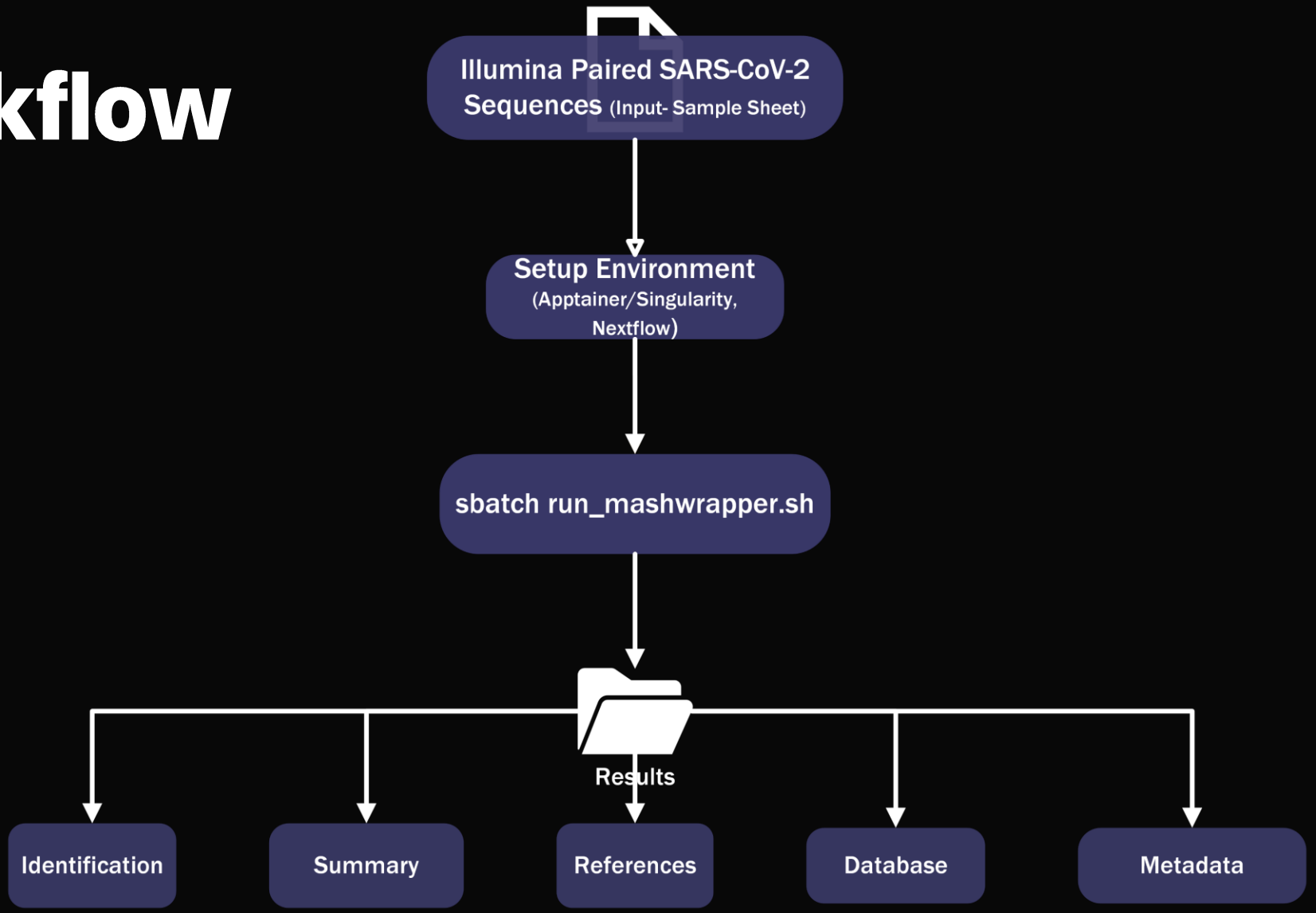
- It enables public health laboratories to quickly confirm organism identity, validate sample content, and support downstream workflows such as lineage assignment, phylogenetic analysis, and genomic surveillance across multiple samples.

Dependencies

- Nextflow and Apptainer/Singularity

Advanced Molecular Detection
Southeast Region Bioinformatics

Workflow



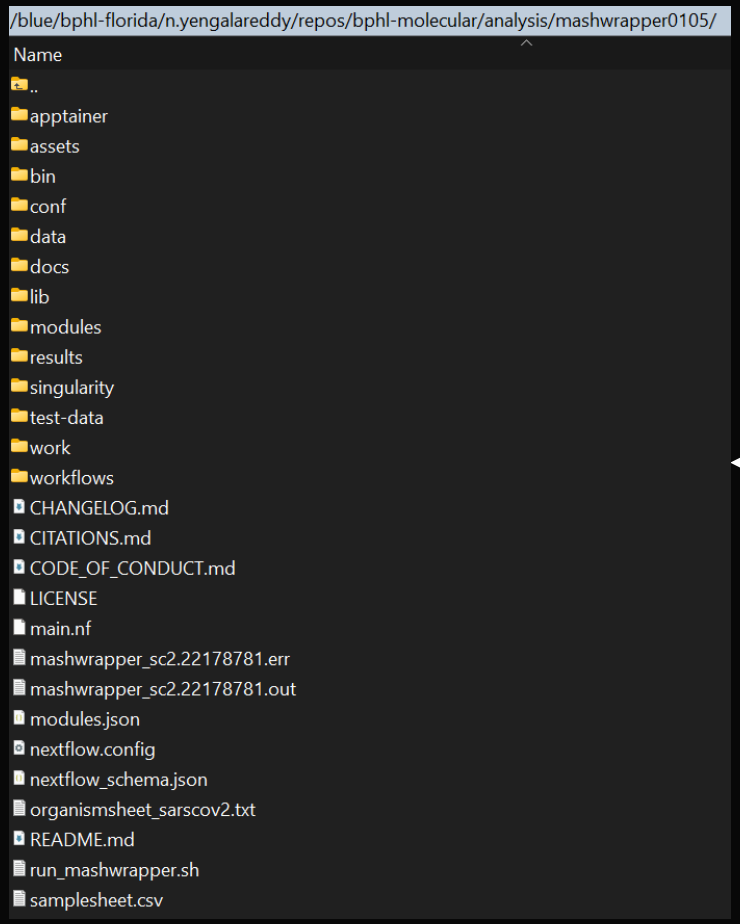
Advanced Molecular Detection
Southeast Region Bioinformatics

Application

Objective:
Use raw Illumina sequencing reads with mashwrapper to rapidly identify SARS-CoV-2 through k-mer–based comparison against reference genomes.

Application Cont.

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

git clone https://github.com/CDCgov/mashwrapper

mkdir analysis/

cd analysis/

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



Advanced Molecular Detection
Southeast Region Bioinformatics

Application Cont.

```
 GNU nano 5.6.1                          run_mashwrapper.sh
#!/usr/bin/bash
#SBATCH --account=bphl-umbrella
#SBATCH --qos=bphl-umbrella
#SBATCH --job-name=mashwrapper_sc2
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=20
#SBATCH --mem=200gb
#SBATCH --time=48:00:00
#SBATCH --output=mashwrapper_sc2.%j.out
#SBATCH --error=mashwrapper_sc2.%j.err
#SBATCH --mail-user=nikhil.yengala@flhealth.gov
#SBATCH --mail-type=FAIL,END

module load nextflow
module load singularity
module load apptainer

#Working Directory
MW="/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/mashwrapper0105"

#Inputs
SAMPLESHEET="$MW/samplesheet.csv"
ORGSHEET="$MW/organismsheet_sarscov2.txt"

#Output/Work/Cache
OUTDIR="$MW/results"
WORKDIR="$MW/work"
export NXF_SINGULARITY_CACHEDIR="$MW/singularity"
export NXF_APPTAINER_CACHEDIR="$MW/apptainer"

mkdir -p "$OUTDIR" "$WORKDIR" "$NXF_SINGULARITY_CACHEDIR" "$NXF_APPTAINER_CACHEDIR"

#Run
  -nextflow run "$MW" \
  -profile singularity \
  -work-dir "$WORKDIR" \
  --input "$SAMPLESHEET" \
  --get_database "$ORGSHEET" \
  --outdir "$OUTDIR" \
  --custom_config_base "$MW/conf" \
  -resume
```

/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/mashwrapper0105/

Name

- ..
- apptainer
- assets
- bin
- conf
- data
- docs
- lib
- modules
- results
- singularity
- test-data
- work
- workflows
- CHANGELOG.md
- CITATIONS.md
- CODE_OF_CONDUCT.md
- LICENSE
- main.nf
- mashwrapper_sc2.22178781.err
- mashwrapper_sc2.22178781.out
- modules.json
- nextflow.config
- nextflow_schema.json
- organismsheet_sarscov2.txt
- README.md
- run_mashwrapper.sh
- samplesheet.csv

```
 GNU nano 5.6.1                organismsheet_sarscov2.txt
Severe acute respiratory syndrome coronavirus 2
```

```
 GNU nano 5.6.1                          samplesheet.csv
sample,fastq_1,fastq_2
J84,/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/mashwrapper0105/data/J84_R1.fastq.gz
J682,/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/mashwrapper0105/data/J682_R1.fastq.
J711,/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/mashwrapper0105/data/J711_R1.fastq.
J886,/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/mashwrapper0105/data/J886_R1.fastq.
M109,/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/mashwrapper0105/data/M109_R1.fastq.
M300,/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/mashwrapper0105/data/M300_R1.fastq.
M301,/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/mashwrapper0105/data/M301_R1.fastq.
M330,/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/mashwrapper0105/data/M330_R1.fastq.
M653,/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/mashwrapper0105/data/M653_R1.fastq.
M908,/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/mashwrapper0105/data/M908_R1.fastq.
T45,/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/mashwrapper0105/data/T45_R1.fastq.gz
```

Application Cont.

```
sbatch run_mashwrapper.sh
```

/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/mashwrapper0105/

Name
- ..
- apptainer
- assets
- bin
- conf
- data
- docs
- lib
- modules
- results
- singularity
- test-data
- work
- workflows
- CHANGELOG.md
- CITATIONS.md
- CODE_OF_CONDUCT.md
- LICENSE
- main.nf
- mashwrapper_sc2.22178781.err
- mashwrapper_sc2.22178781.out
- modules.json
- nextflow.config
- nextflow_schema.json
- organismsheet_sarscov2.txt
- README.md
- run_mashwrapper.sh
- samplesheet.csv

/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/mashwrapper0105/results/combinedOutput/

Name
- ..
- collated_download_genomes.log
- collated_excluded_genomes.txt
- collated_species_id.log
- collated_species_id_results.txt
- database.info

```
SARS-CoV-2 Species ID Tool using Mash
Date and Time = 2025-12-29
Input query file 1: 3711_R1.fastq
Input query file 2: 3711_R2.fastq
Maximum Mash distance (-d): 0.05
Minimum K-mer copy number (-m) to be included in the sketch: 16
K-mer size used for sketching: 25
Mash Database name: myMashDatabase.2025-12-29_17:45:16.msh
mashwrapper version: 3.2.2

Best species match: Severe acute

Top 5 results:
```

Genus	Species	GeneBank Identifier	Mash Dist	% Seq Sim	P-value	Kmer
Severe	acute	GCA_011545545.1	0.00161	99.839	0	29033/31424
Severe	acute	GCA_009858895.3	0.00172	99.828	0	28988/31535
Severe	acute	GCA_011537005.1	0.00172	99.828	0	28988/31535
Severe	acute	GCA_011537985.2	0.00171	99.829	0	28987/31523
Severe	acute	GCA_011537945.2	0.00171	99.829	0	28987/31523

```
SARS-CoV-2 Species ID Tool using Mash
Date and Time = 2025-12-29
Input query file 1: M300_R1.fastq
Input query file 2: M300_R2.fastq
Maximum Mash distance (-d): 0.05
Minimum K-mer copy number (-m) to be included in the sketch: 16
K-mer size used for sketching: 25
Mash Database name: myMashDatabase.2025-12-29_17:45:16.msh
mashwrapper version: 3.2.2

Best species match: Severe acute

Top 5 results:
```

Genus	Species	GeneBank Identifier	Mash Dist	% Seq Sim	P-value	Kmer
Severe	acute	GCA_011545545.1	0.01471	98.529	0	27851/52618
Severe	acute	GCA_009858895.3	0.01485	98.515	0	27777/52758
Severe	acute	GCA_011537005.1	0.01485	98.515	0	27777/52758
Severe	acute	GCA_009938055.1	0.01485	98.515	0	27776/52745
Severe	acute	GCA_011537225.1	0.01482	98.518	0	27776/52697

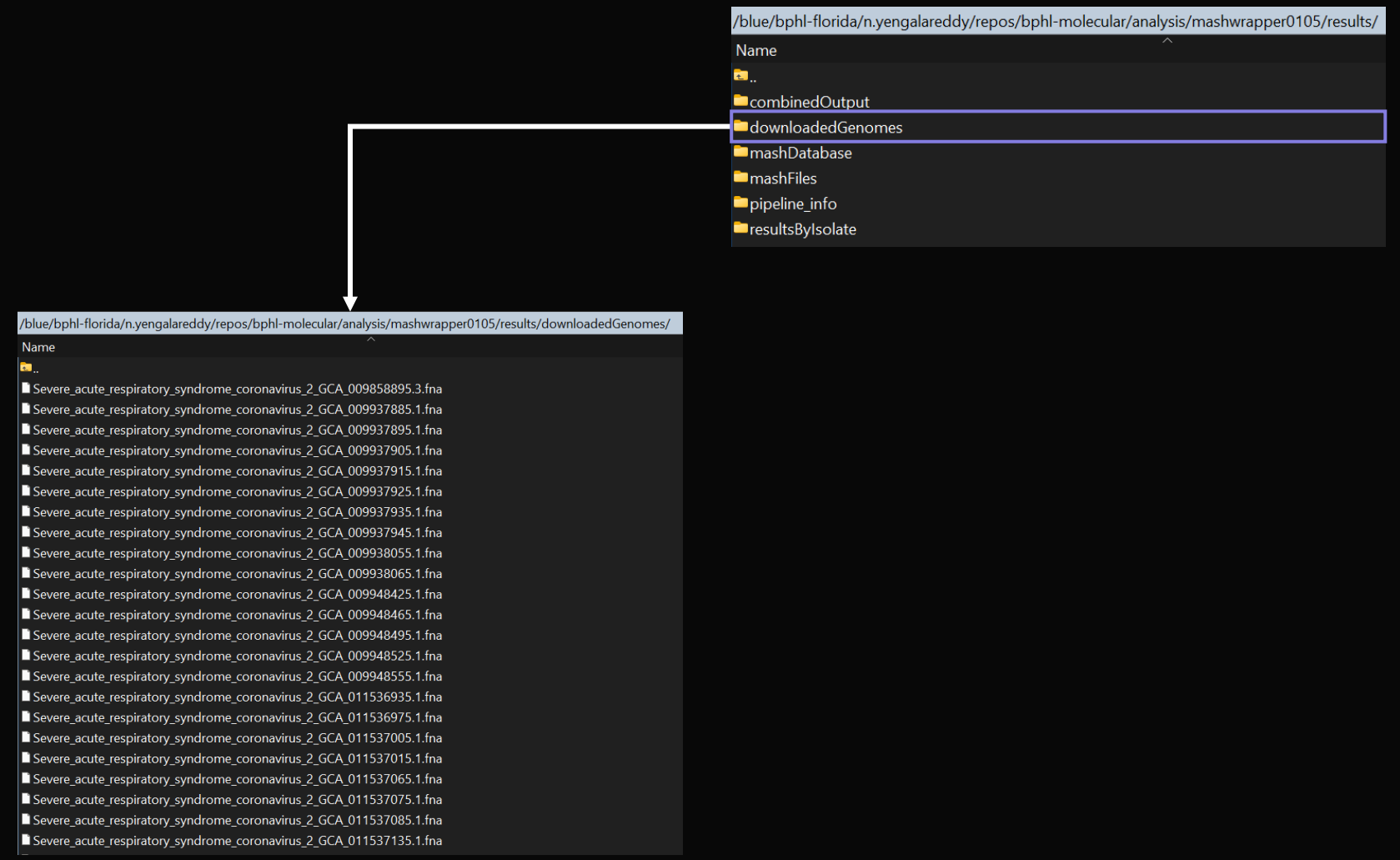
```
SARS-CoV-2 Species ID Tool using Mash
Date and Time = 2025-12-29
Input query file 1: M330_R1.fastq
Input query file 2: M330_R2.fastq
Maximum Mash distance (-d): 0.05
Minimum K-mer copy number (-m) to be included in the sketch: 24
K-mer size used for sketching: 25
Mash Database name: myMashDatabase.2025-12-29_17:45:16.msh
mashwrapper version: 3.2.2

Best species match: Severe acute
```

/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/mashwrapper0105/results/

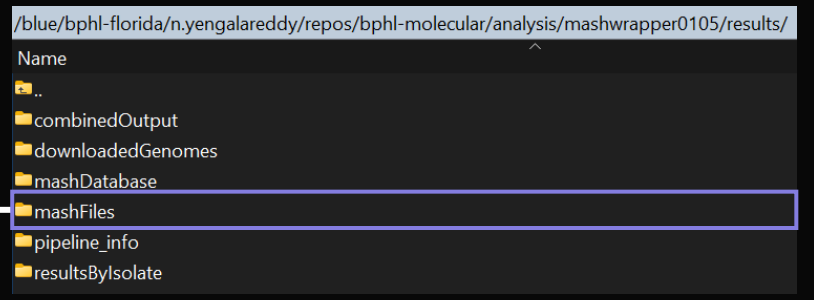
Name
- ..
- combinedOutput
- downloadedGenomes
- mashDatabase
- mashFiles
- pipeline_info
- resultsByIsolate

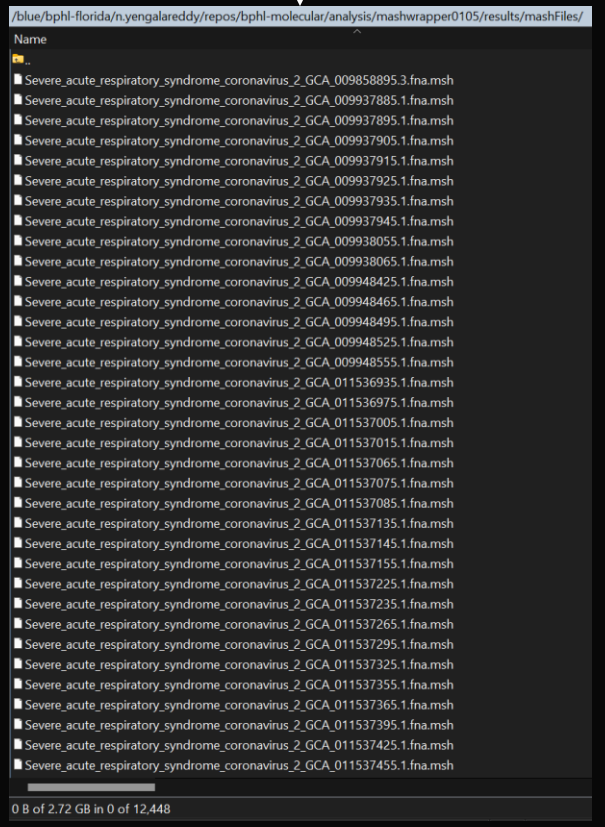
Advanced Molecular Detection
Southeast Region Bioinformatics

Application Cont.



Application Cont.





Advanced Molecular Detection
Southeast Region Bioinformatics

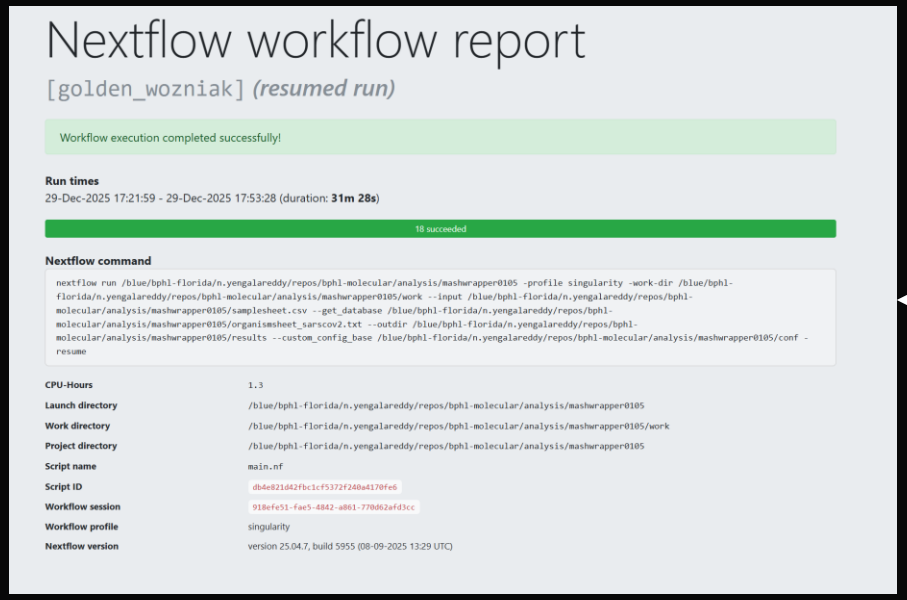
Application Cont.

/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/mashwrapper0105/results/

Name
- ..
- combinedOutput
- downloadedGenomes
- mashDatabase
- mashFiles
- pipeline_info
- resultsByIsolate

/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/mashwrapper0105/results/pipeline_info/

Name
- ..
- execution_report_2025-12-29_17-21-59.html
- execution_timeline_2025-12-29_17-21-59.html
- execution_trace_2025-12-29_17-21-59.txt
- organismsheet_valid.txt
- pipeline_dag_2025-12-29_17-21-59.svg
- samplesheet.valid.csv
- software_versions.yml
- versions.yml



Nextflow workflow report

[golden_wozniak] *(resumed run)*

Workflow execution completed successfully!

Run times

29-Dec-2025 17:21:59 - 29-Dec-2025 17:53:28 (duration: **31m 28s**)

18 succeeded

Nextflow command

```
nextflow run /blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/mashwrapper0105 -profile singularity -work-dir /blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/mashwrapper0105/work --input /blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/mashwrapper0105/samplesheet.csv --get_database /blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/mashwrapper0105/organismsheet_sarscov2.txt --outdir /blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/mashwrapper0105/results --custom_config_base /blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/mashwrapper0105/conf -resume
```

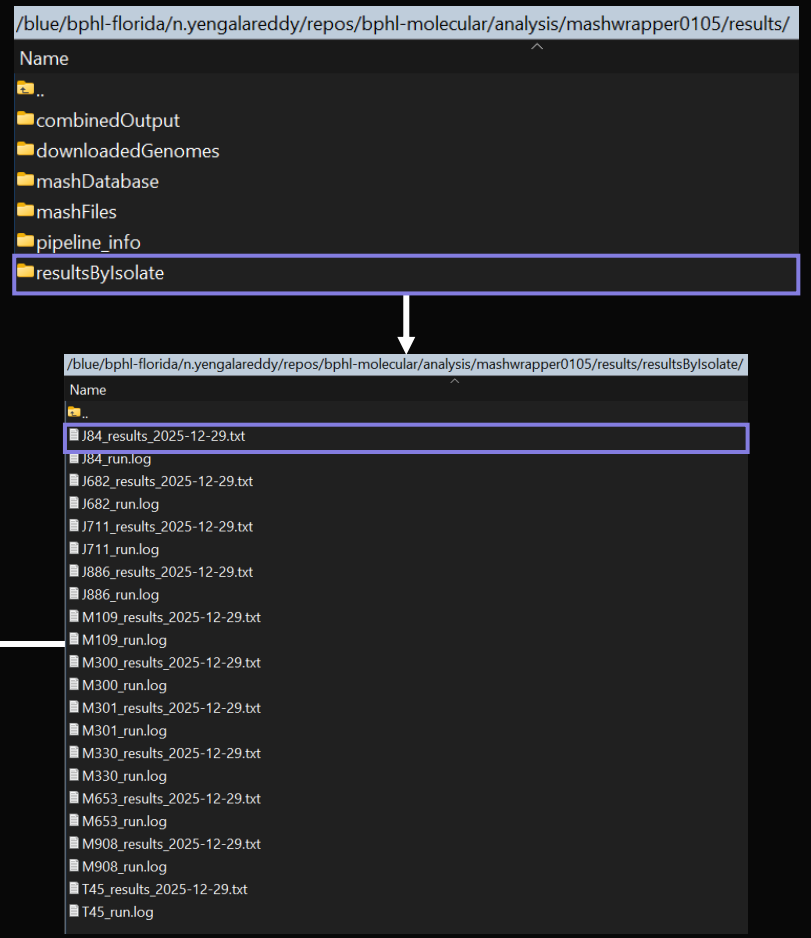
CPU-Hours	1.3
Launch directory	/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/mashwrapper0105
Work directory	/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/mashwrapper0105/work
Project directory	/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/mashwrapper0105
Script name	main.nf
Script ID	db4e821d42fbc1cf5372f240a4170fe6
Workflow session	918efe51-fae5-4842-a861-770d62afd3cc
Workflow profile	singularity
Nextflow version	version 25.04.7, build 5955 (08-09-2025 13:29 UTC)

Advanced Molecular Detection
Southeast Region Bioinformatics

Application Cont.

```
SARS-CoV-2 Species ID Tool using Mash
Date and Time = 2025-12-29
Input query file 1: J84_R1.fastq
Input query file 2: J84_R2.fastq
Maximum Mash distance (-d): 0.05
Minimum K-mer copy number (-m) to be included in the sketch: 13
K-mer size used for sketching: 25
Mash Database name: myMashDatabase.2025-12-29_17:45:16.msh
mashwrapper version: 3.2.2

Best species match: Severe acute

Top 5 results:
```

Genus	Species	GeneBank Identifier	Mash Dist	% Seq Sim	P-value	Kmer
Severe	acute	GCA_011545545.1	0.01266	98.734	0	29703/51828
Severe	acute	GCA_009858895.3	0.01273	98.727	0	29678/51919
Severe	acute	GCA_011537015.1	0.01273	98.727	0	29675/51917
Severe	acute	GCA_011545335.2	0.01273	98.727	0	29675/51909
Severe	acute	GCA_011545285.2	0.01273	98.727	0	29675/51909



Advanced Molecular Detection
Southeast Region Bioinformatics

Conclusion

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

Advanced Molecular Detection
Southeast Region Bioinformatics

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