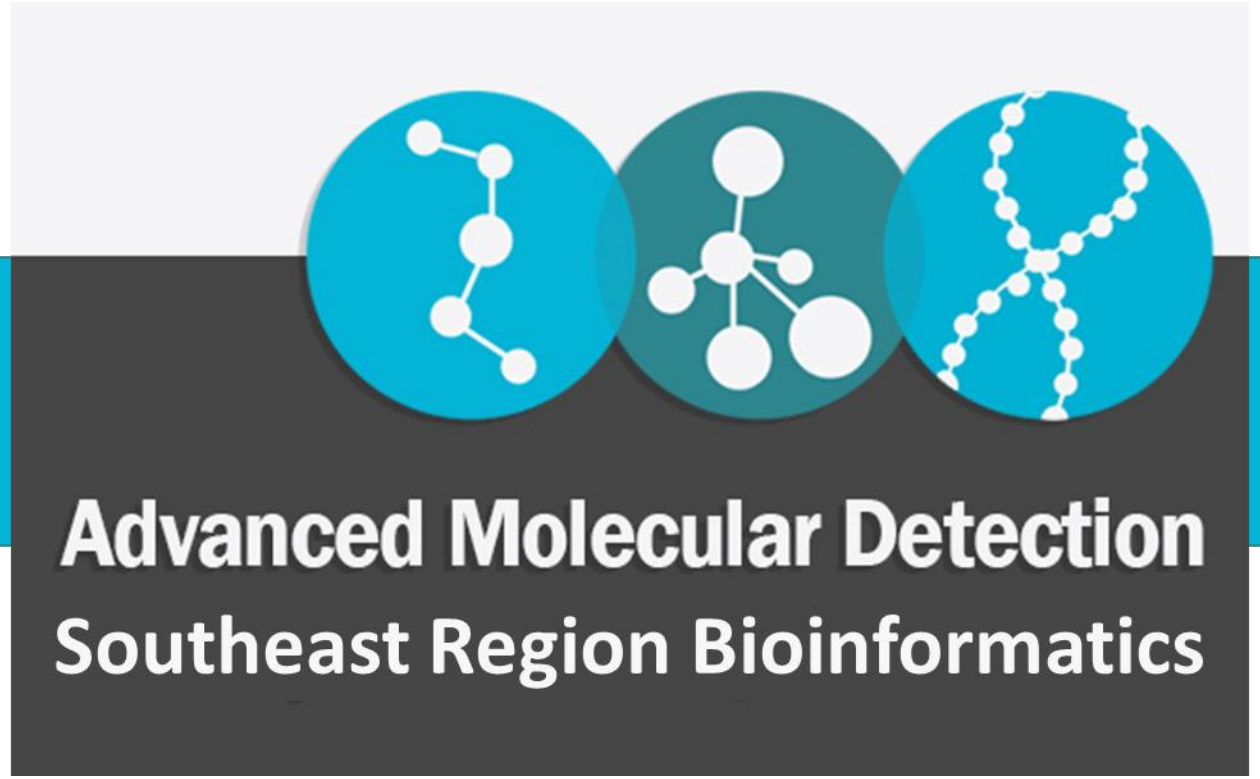


Sanibel Pipeline
02/05/2024

Outline

Agenda

Notes

Sanibel Pipeline

Demo

Questions

Agenda

February 19 – Genomic Epidemiology Training Part - 1
March 4 – Genomic Epidemiology Training Part - 2

Future Trainings

- ONT & FL's Flisochar pipeline
- StaPH-B Toolkit Programs/Pipelines
- GISAID flagged SARS-CoV-2
- R Training Series
- Dryad pipeline
- …and more

Advanced Molecular Detection
Southeast Region Bioinformatics

Notes

- First Quarterly meting of the year 2024 will be during the first week of March, tentatively
- If any staff members require new HPG user training, please feel free to email us
- Genomic Epidemiology Training – 5 Parts will begin from next office hours i.e., February 19th

Advanced Molecular Detection
Southeast Region Bioinformatics

Sanibel

- Nextflow version of FLAQ_AMR pipeline, updated by Yibo Dong
- Nextflow pipeline used to assemble and analyze NGS data in fastq format from bacterial genomes

GitHub - BPHL-Molecular/Sanibel: A Nextflow pipeline to analyze NGS data in fastq format from bacterial genome

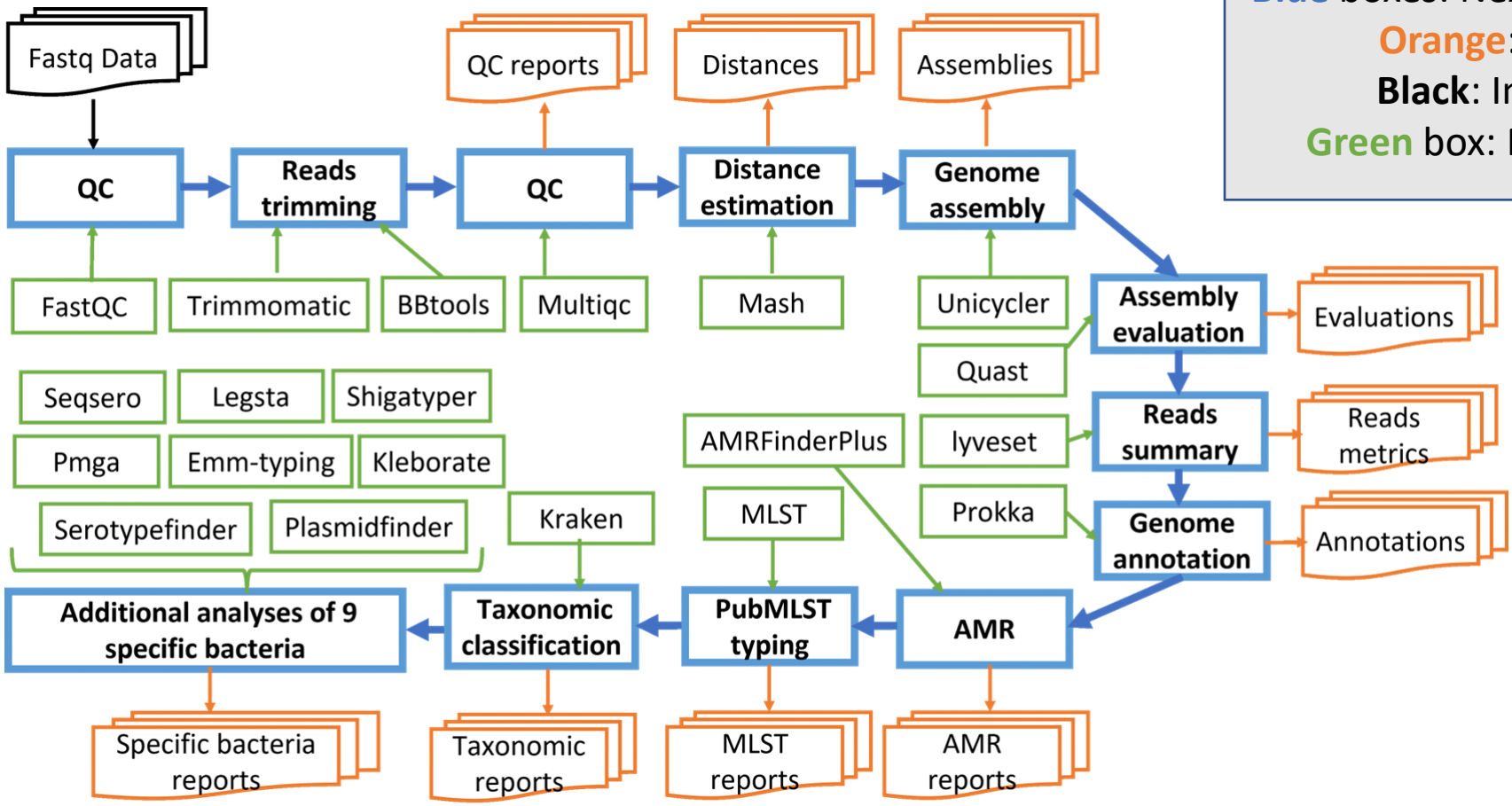
Advanced Molecular Detection
Southeast Region Bioinformatics

Difference between FLAQ_AMR & Sanibel

- Sanibel significantly reduces runtime and is especially suitable for analysis of large sample sizes
- Some additional analyses for: *Neisseria*, *H. influenzae*, *Legionella*, *Shigella*, Group A *Streptococcus*, *Klebsiella*, *Salmonella* and *E. coli* are added to this pipeline
- Identifies clonal complex and serotype of *Neisseria* and *H. influenzae* species

Advanced Molecular Detection
Southeast Region Bioinformatics

Overview



Blue boxes: Nextflow processes
Orange: Outputs
Black: Input data
Green box: Pipeline tools

Advanced Molecular Detection
Southeast Region Bioinformatics

Prerequisites

- Nextflow
  - Details of Nextflow installation can be found at this link
  - Use **module load nextflow** on HPG
- Python3
  - “Pandas" should be installed by **pip3 install pandas** if not already included in your python3 package
- Singularity/APPTAINER
  - The detail of installation can be found at this link
  - Use **module load apptainer** on HPG
- SLURM

Advanced Molecular Detection
Southeast Region Bioinformatics

Installation Using Conda

```
conda create -n SANIBEL -c conda-forge python=3.10 pandas
```

```
conda activate SANIBEL
```

Advanced Molecular Detection
Southeast Region Bioinformatics

Installation Using Git Clone

```
git clone https://github.com/BPHL-Molecular/Sanibel.git
```

Advanced Molecular Detection
Southeast Region Bioinformatics

Running Illumina Data

- Move data files to directory /fastqs. The file names should look like "XZA22002292-XS-ASX550430-220701_S143_L001_R1_001.fastq.gz"
- Open file "params.yaml", set the two parameters absolute paths, i.e., ".../.../fastqs" and ".../.../output"
- Run the following command in the directory of the pipeline

```
sbatch ./sanibel_illumina.sh
```

Advanced Molecular Detection
Southeast Region Bioinformatics

Running Non-Illumina Data

- Move data files to directory titled /fastqs. The file names should look like "XZA22002292_1.fastq.gz", "XZA22002292_2.fastq.gz"
- Open file "params.yaml" and set the two parameters absolute paths. They should be ".../.../fastqs" and ".../.../output"
- Run the following command in the directory of the pipeline

```
sbatch ./sanibel_illumina.sh
```

Advanced Molecular Detection
Southeast Region Bioinformatics

Using Sanibel with Docker

- By default, Sanibel uses singularity to run containers and is wrapped by SLURM
- If you want to use Docker to run the containers, you should use the command below:
  - If your data file names do not directly come from Illumina output

```
sbatch ./sanibel_docker.sh
```

  - If your data file names directly come from Illumina output

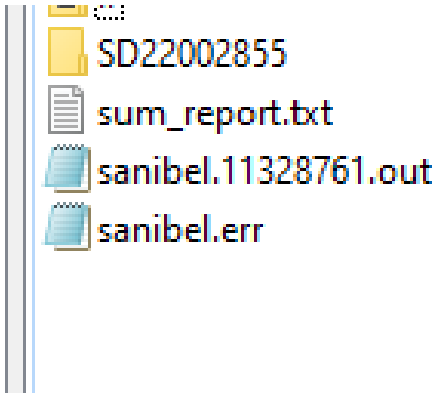
```
sbatch ./sanibel_illumina_docker.sh
```

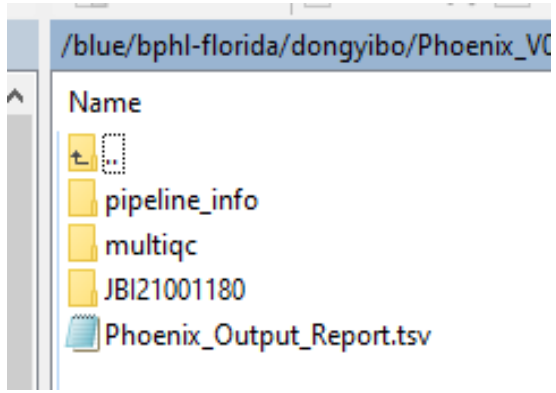
Advanced Molecular Detection
Southeast Region Bioinformatics

Sanibel vs CDC's PHoeNIx

- Both are pipelines to analyze bacterial genomes
- Use similar tools, such as Kraken, QUAST, mlst, FastQC, MultiQC, etc.
- There are many overlaps in function

Output from Sanibel





Output from PHoeNIx

- Only Sanibel identifies serotypes of *Neiserria* or *H. influenzae*

Advanced Molecular Detection
Southeast Region Bioinformatics

Demo

- Log into HPG
- Create a directory titled Sanibel_demo

```
mkdir sanibel_demo
```

```
cd sanibel_demo
```

- Clone Sanibel in the created directory

```
git clone https://github.com/BPHL-Molecular/Sanibel.git
```

Advanced Molecular Detection
Southeast Region Bioinformatics

Demo

- Copy fastq data to the “fastqs” folder

```
cd fastqs/sample_data
ls –l non_illumina*    (or ls –l illumina*)
cp ./ non_illumina* /*fastq.gz ../
```

- Open and edit params.yaml
    - Edit the full path of input and output

Advanced Molecular Detection
Southeast Region Bioinformatics

Demo

- Make sure you edit your email in the script

SBATCH --mail-user=<EMAIL> in sanibel.sh or Sanibel_illumina.sh

- Run the pipeline

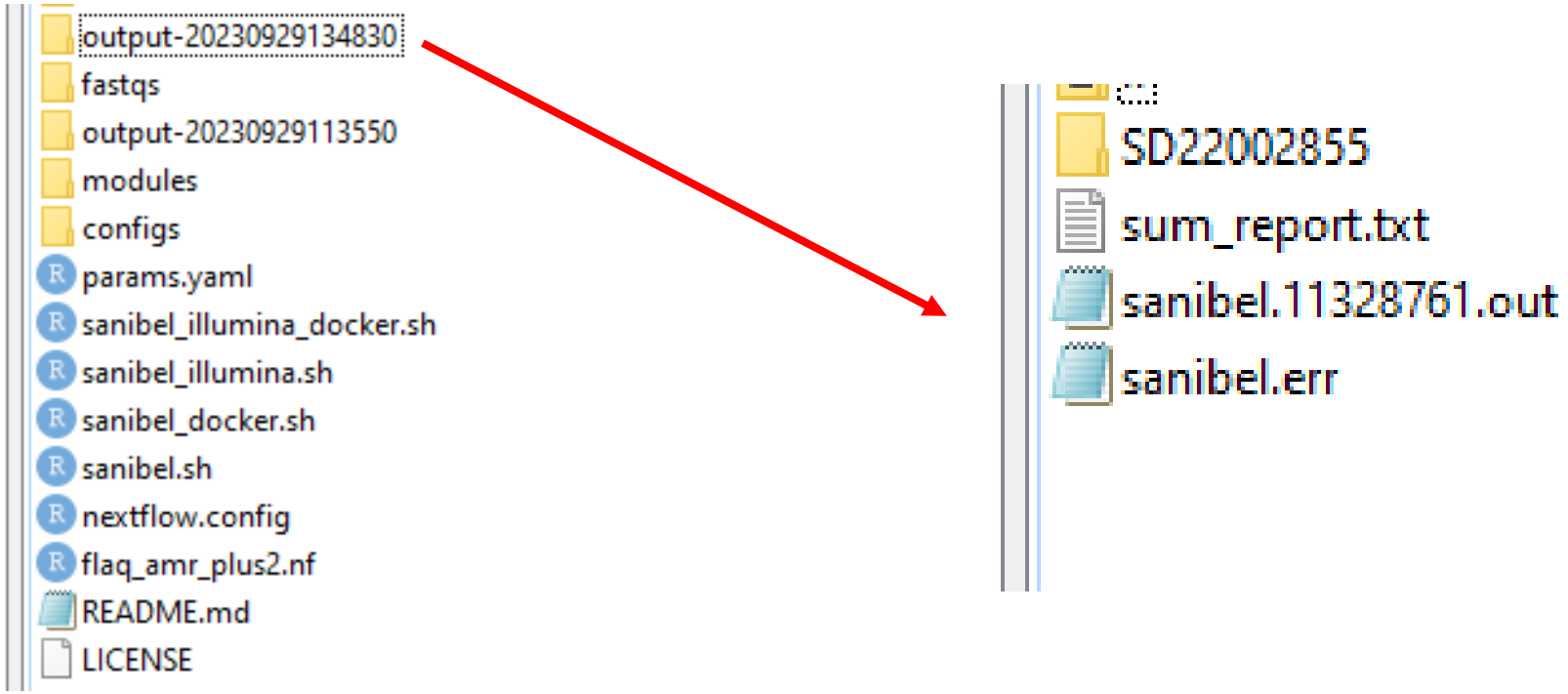
sbatch ./Sanibel.sh (if file name is not Illumina format)
sbatch ./Sanibel_illumina.sh (if file name is Illumina format)

- Check job status

squeue –A bphl-umbrella

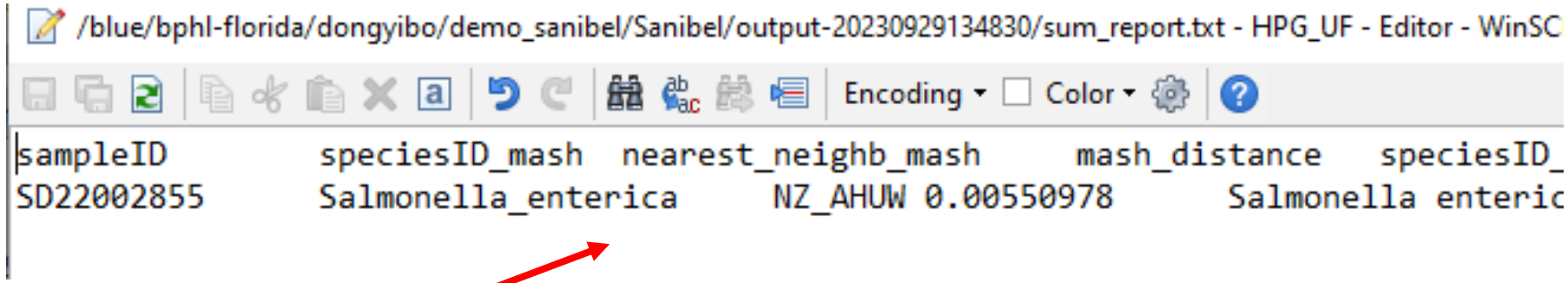
Advanced Molecular Detection
Southeast Region Bioinformatics

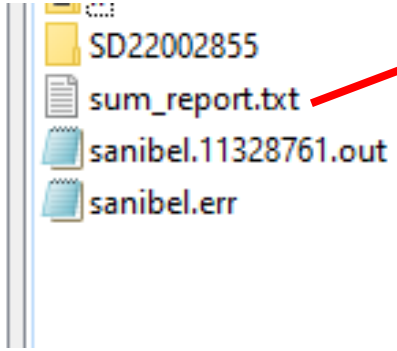
Output



Advanced Molecular Detection
Southeast Region Bioinformatics

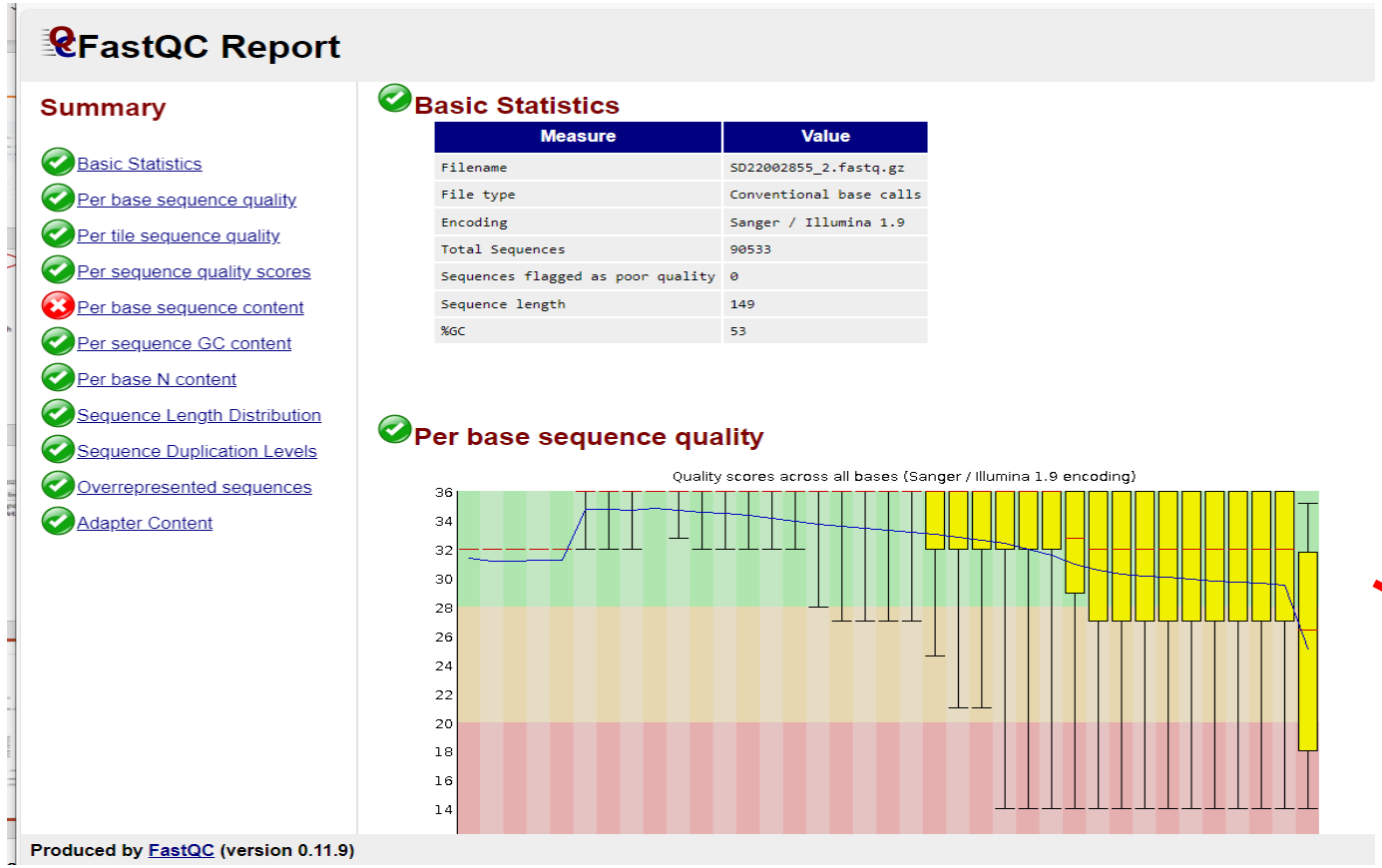
Output





Advanced Molecular Detection
Southeast Region Bioinformatics

FastQC Report



FastQC Report

Summary

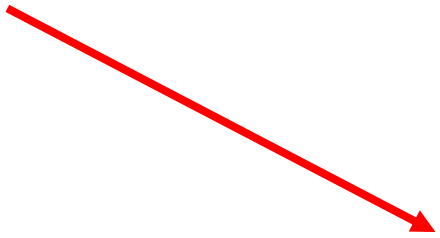
- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

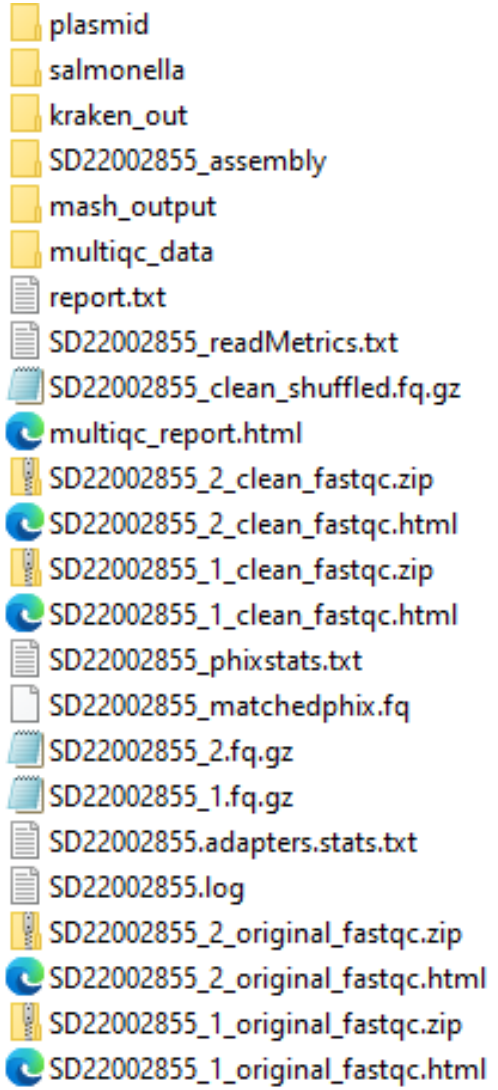
Basic Statistics

Measure	Value
Filename	SD22002855_2.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	90533
Sequences flagged as poor quality	0
Sequence length	149
%GC	53

Per base sequence quality

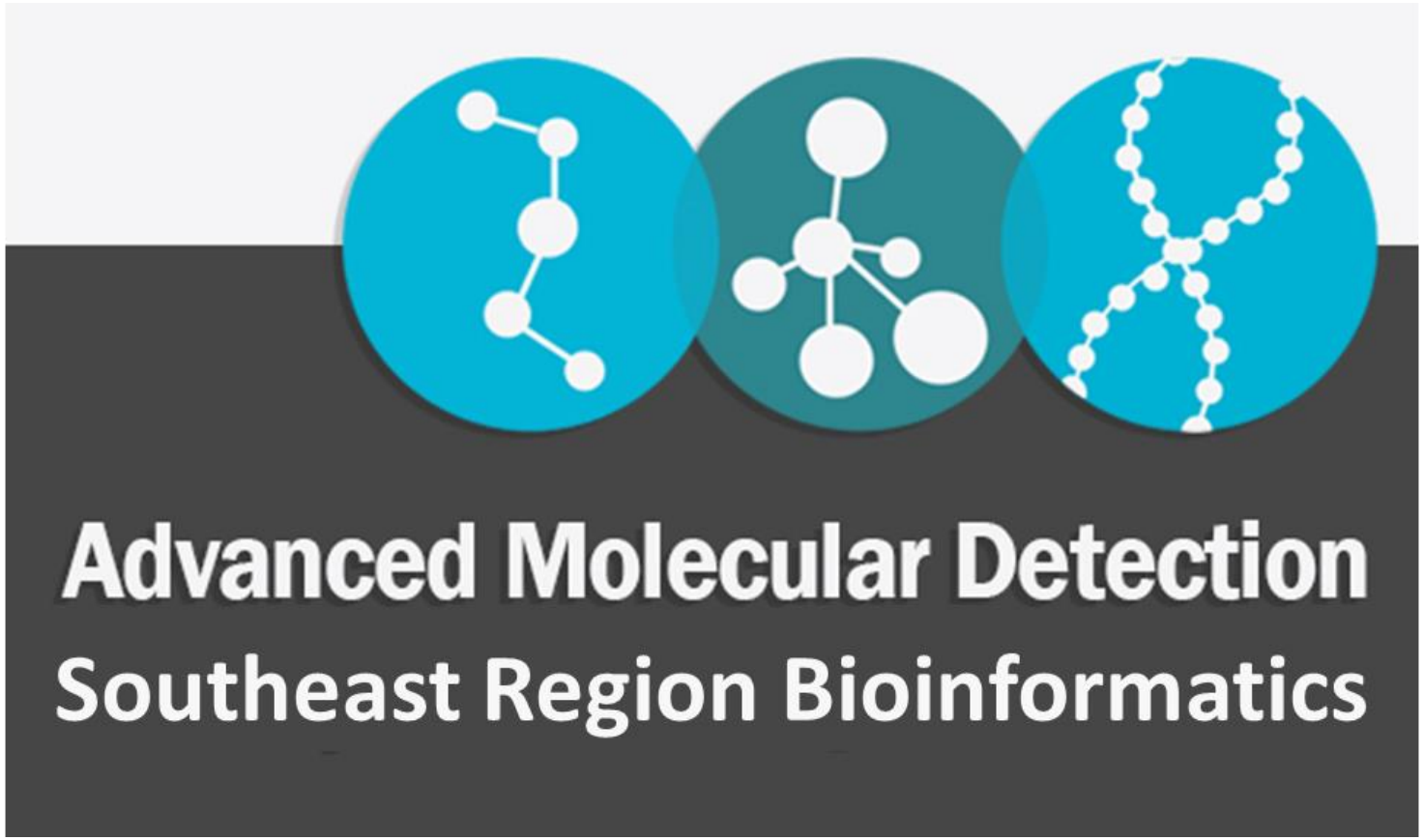
Produced by FastQC (version 0.11.9)





plasmid
salmonella
kraken_out
SD22002855_assembly
mash_output
multiqc_data
report.txt
SD22002855_readMetrics.txt
SD22002855_clean_shuffled.fq.gz
multiqc_report.html
SD22002855_2_clean_fastqc.zip
SD22002855_2_clean_fastqc.html
SD22002855_1_clean_fastqc.zip
SD22002855_1_clean_fastqc.html
SD22002855_phixstats.txt
SD22002855_matchedphix.fq
SD22002855_2.fq.gz
SD22002855_1.fq.gz
SD22002855.adapters.stats.txt
SD22002855.log
SD22002855_2_original_fastqc.zip
SD22002855_2_original_fastqc.html
SD22002855_1_original_fastqc.zip
SD22002855_1_original_fastqc.html

Advanced Molecular Detection
Southeast Region Bioinformatics



Questions?

bphl-sebioinformatics@flhealth.gov

Lakshmi Thsaliki, MS
Bioinformatician
Lakshmi.Thsaliki@flhealth.gov

Molly Mitchell, PhD
Bioinformatician
Molly.Mitchell@flhealth.gov