Advanced Molecular Detection
Southeast Region Bioinformatics

PHoeNIx Pipeline
05/01/2023

Outline

- Updates and Reminders
- Agenda
- What is PHoeNIx?
- Dependencies & Install
- Testing Install & Running PHoeNIx
- Output File Structure & Pipeline Summary
- Demo
- Editing the HPC config file for HPG
- Updates from developer Dr. Jill Hagey
- Questions

Updates and Reminders

- Please email us on bphl-sebioinformatics@flhealth.gov for any queries or requests.
- One of the group members will respond to your query.

Advanced Molecular Detection
Southeast Region Bioinformatics

Agenda

May 1 – PHoeNIx #2 (Demo & Troubleshooting)
May 15 – Open OnDemand
May 29 – AMRFinder+ and Pha4ge's hAMRonization pipeline
June 12 – Outbreak/cluster training

Future Trainings
- ONT & FL's Flisochar pipeline
- StaPH-B Toolkit Programs/Pipelines
- GISAID flagged SARS-CoV-2
- Git (git clone, etc.)
- Generating R figures
- SRA human scrubber tool
- …and more

Advanced Molecular Detection
Southeast Region Bioinformatics

What is PHoeNIx?

- **PHoeNIx** – Portable Healthcare Nextgen Informatics pipeline
- Developed by the Division of Healthcare Quality Promotion(DHQP) at the CDC for pathogens commonly encountered in healthcare settings (including but not limited to e.g., *Acinetobacter baumannii*, *Klebsiella pneumoniae*, *Pseudomonas aeruginosa*, and *Staphylococcus aureus*)
- Pipeline is built using **Nextflow** which runs tasks across multiple compute infrastructures in a portable manner
- Uses **Docker/Singularity** containers making installation trivial and highly reproducible results
- Pipeline provides a standardized approach for identifying and characterizing healthcare-associated bacterial pathogens, specifically for public health partners
- Nextflow DSL2 implementation of this pipeline uses one container per process which makes it easier to maintain and update software dependencies

Advanced Molecular Detection
Southeast Region Bioinformatics

PHoeNIx Pipeline Insights

- PHoeNIx uses Illumina paired-end reads and was designed for use with pathogens causing healthcare-associated bacterial infections. This comprehensive pipeline performs:
    1. Quality control
    2. Checks for contamination
    3. Confirms taxa ID
    4. Performs sequence typing
    5. Assembles reads into scaffolds
    6. Detects antibiotic resistance & hypervirulence genes
    7. Searches for plasmid markers
- PHoeNIx generates several files that are compatible with downstream analytic tools, such as those used for phylogenetic tree building
- This pipeline is available to run on Terra, Nextflow tower, CLI and is also incorporated into the StaPH-B toolkit

Advanced Molecular Detection
Southeast Region Bioinformatics

Dependencies & Install

- Install Nextflow with conda-forge channel and a bioconda channel using mamba
- Configure Nextflow for command-line interface (CLI) users
- Install container software
- Transfer Kraken2 database files to a new directory using WinSCP
- Edit configuration files to run on HPG

Advanced Molecular Detection
Southeast Region Bioinformatics

Install Nextflow

Install Nextflow environment using mamba:

1. Activate conda/mamba

```
$ module load conda
```

2. Install Nextflow

```
$ mamba create -yp /blue/<bphl-state>/<user>/conda_envs/nextflow -c conda-forge -c bioconda nextflow=21.10.6
```

3. Activate the mamba environment

```
$ mamba activate /blue/<bphl-state>/<user>/conda_envs/nextflow
```

4. Then, run PHoeNIx from this environment

Advanced Molecular Detection
Southeast Region Bioinformatics

Configuring Nextflow for CLI users

Configuration is done in the form of a config file & is needed so that Nextflow knows how to fetch the required software. Options include:

- Pipeline comes with config files called **docker** & **singularity** which instruct the pipeline to use the named tool for software management. For example, **-profile test**,**singularity**
- Check nf-core/configs to see if a custom config file already exists for your institute, so you can use **-profile <institute>** in your command which enables either docker or singularity to set the execution for your local compute environment
- Use **--singularity_pull_docker_container** if you have issues downloading Singularity images. This will pull and convert the docker image instead.
- Alternatively, you can use **nf-core download** to download images first, before running the pipeline.
  - Set **NXF_SINGULARITY_CACHEDIR** or **singularity.cachedir** for **N**extflow to store & reuse the images from a central location for future pipelines

Advanced Molecular Detection
Southeast Region Bioinformatics

Configuration set up

To add NXF_SINGULARITY_CACHEDIR to your bash profile run the following:

- Open your ~/.bash_profile

```
$ emacs ~/.bash_profile
```

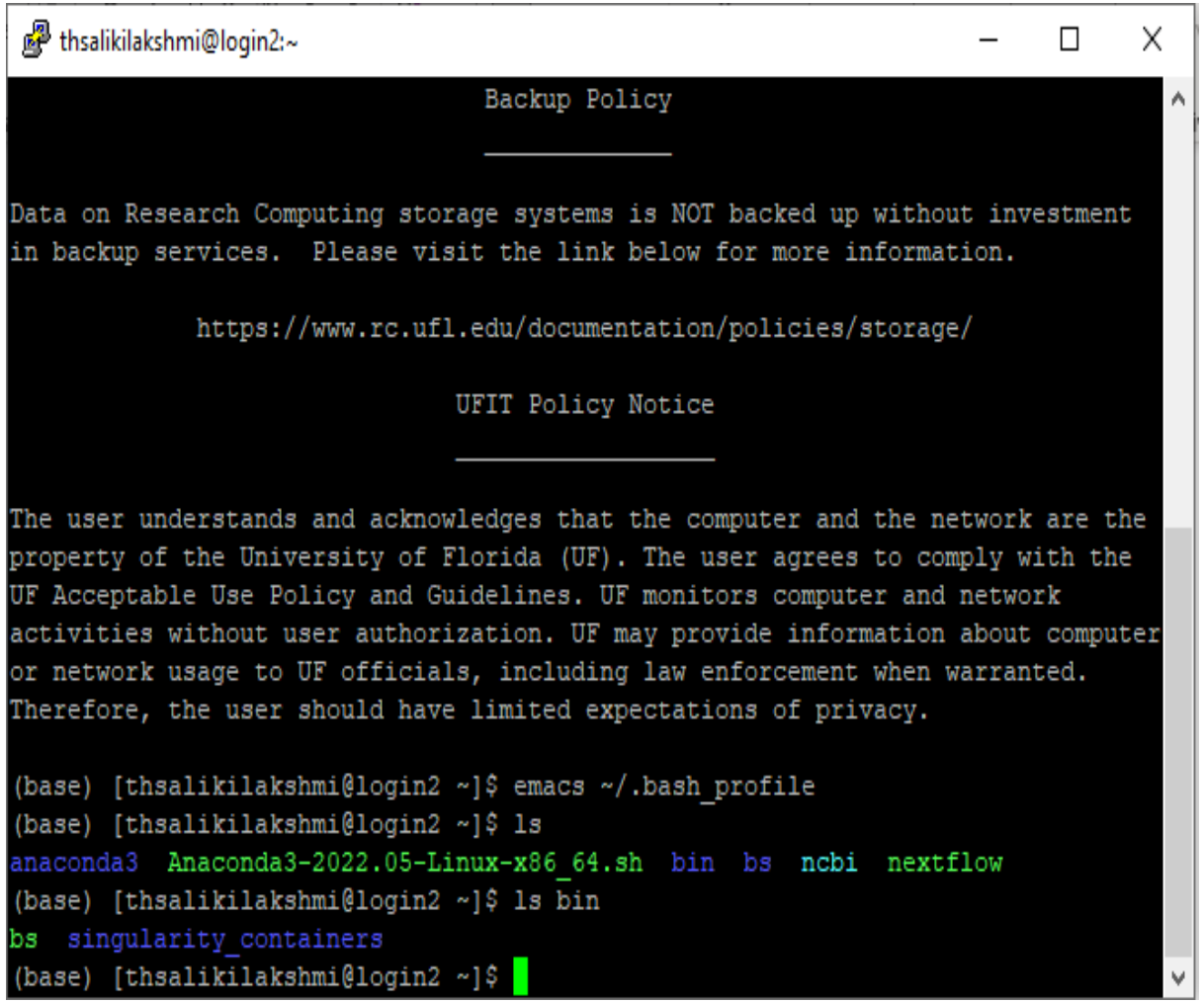
- Add these lines within ~/.bash_profile

```
$ export NXF_SINGULARITY_CACHEDIR=/$PATH/Singularity_containers
$ export /path/to/singularity_containers
```

- $ /path/ is the full path of the folder where you want to store. (**Singularity_Containers** can be named whatever you want)

```
$ source ~/.bash_profile
```

- This command allows Nextflow to see the new path



Advanced Molecular Detection
Southeast Region Bioinformatics

emacs ~/.bash_profile

thsalikilakshmi@login2:~

```
File Edit Options Buffers Tools Sh-Script Help
# .bash_profile

# Get the aliases and functions
if [ -f ~/.bashrc ]; then
        . ~/.bashrc
fi

# User specific environment and startup programs

PATH=$PATH:$HOME/bin

export NXF_SINGULARITY_CACHEDIR=/$HOME/bin/singularity_containers
export PATH
-UU-:----F1  .bash_profile    All L1     (Shell-script[bash]) -------------------
Indentation setup for shell type bash
```

Advanced Molecular Detection
Southeast Region Bioinformatics

Download Database Files for Kraken2

- Email HAISeq@cdc.gov, with the subject line **"krakenDB invite request"** to request access to the sharefile link & provide the email address to send invite to
- Then, download the **hash.k2d, opts.k2d**, and **taxo.k2d** files needed for the Kraken2 subworkflow of PHoeNIx from the CDC sharefile link
- Other kraken2 databases cannot be used – there is a specific **ktax_map.k2** file needed for the pipeline
- Download via WinSCP
- Once downloaded, the folder containing these files should be passed to PHoeNIx via by **--kraken2db**

We've shared the database files via HPG public-share

Advanced Molecular Detection
Southeast Region Bioinformatics

Git Clone PHoeNIx

Install the latest version via cloning PHoeNIx GitHub repo into a folder of your choosing

```
cd $/path/to/local/phoenix/repo/
git clone https://github.com/CDCgov/phoenix
```

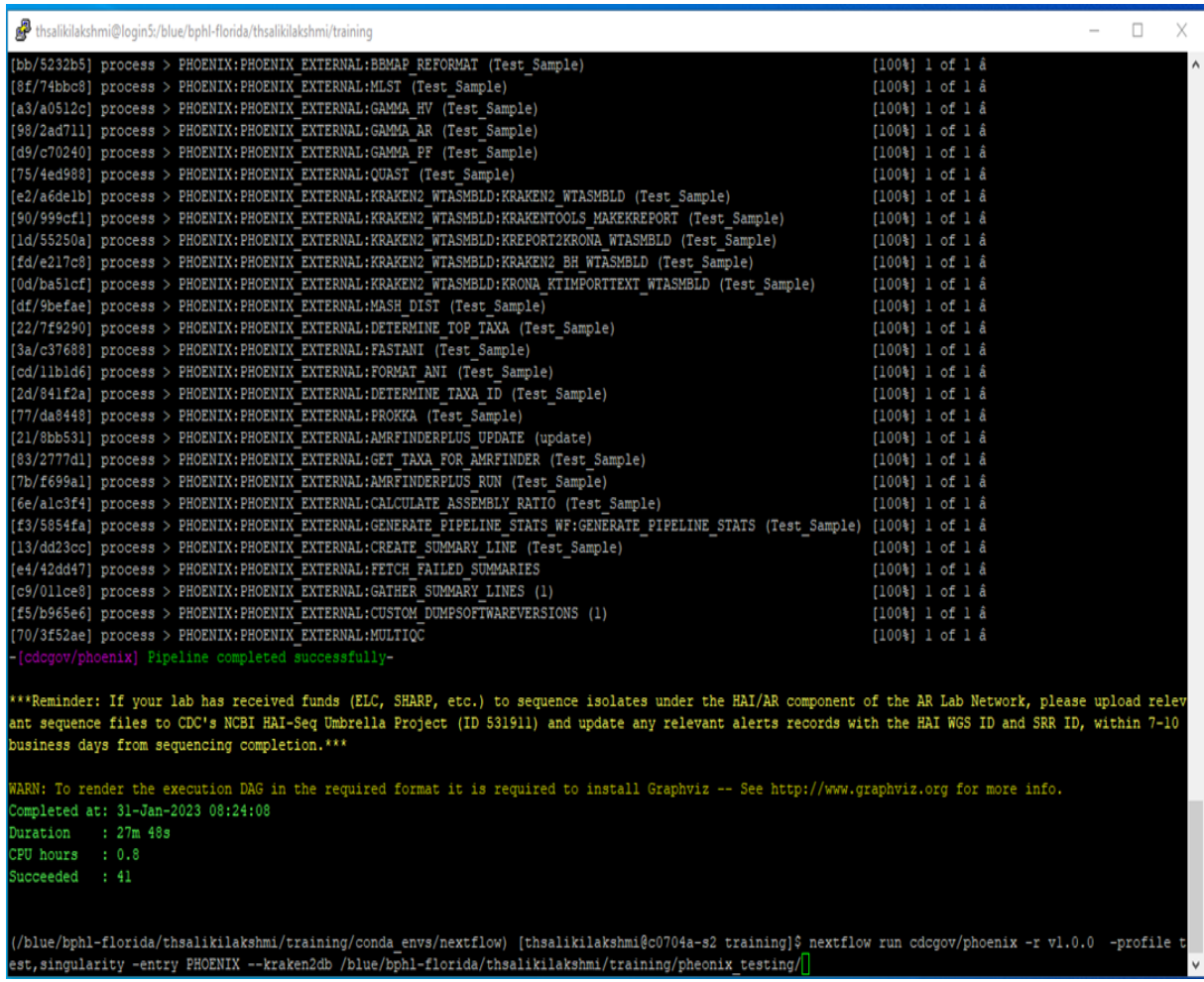
Advanced Molecular Detection
Southeast Region Bioinformatics

Testing Install

- Test that the pipeline is installed and configured correctly:

```
$ nextflow run phoenix/main.nf –profile test,singularity --entry PHOENIX --kraken2db $PATH_TO_DB
```

- This command will run the pipeline on preloaded data
- If all goes well, the output will look like the screen on the right
- As seen from the image, pipeline takes **~21 mins** to run 19 samples
- Note: some steps aren't run in this pipeline (i.e., SPADES_WF stats), this is normal.



Advanced Molecular Detection
Southeast Region Bioinformatics

Running PHoeNIx

The help command will display required and optional options

$ nextflow run /blue/<bphl-state>/<user>/phoenix/phoenix --help

```
cdcgov/phoenix v1.0.0
-----------------------------------------------------------
Typical pipeline command:

  nextflow run cdcgov/phoenix -r $version -entry PHOENIX -profile <singularity,docker,test,custom> --input samplesheet.csv --kraken2db $PATH_TO_DB

Required Options
  --input                    [string]  Path to comma-separated file containing information about the samples in the experiment.
  --kraken2db                [string]  FULL PATH for where kraken2_db files are, which includes hash.k2d, taxo.k2d, opts.k2d and ktax_map.k2 files.
                                       [default: ${baseDir}/assets/databases/]

Optional options
  --outdir                   [string]  Path to the output directory where the results will be saved. MUST BE A FULL PATH. [default:
                                       ${PWD}/results]
  --email                    [string]  Email address for completion summary.
  --multiqc_title            [string]  MultiQC report title. Printed as page header, used for filename if not otherwise specified.
  --busco_db_path            [string]  Path for where BUSCO files are if you want to run in offline mode. Use only with -entry CDC_PHOENIX
                                       [default: Not to run in offline mode.]
  --publish_dir_mode         [string]  null [default: copy]
  --singularity_pull_docker_container [string]  null [default: false]

Max job request options
  --max_cpus                 [integer] Maximum number of CPUs that can be requested for any single job. [default: 16]
  --max_memory               [string]  Maximum amount of memory that can be requested for any single job. [default: 128.GB]
  --max_time                 [string]  Maximum amount of time that can be requested for any single job. [default: 240.h]
```

Advanced Molecular Detection
Southeast Region Bioinformatics

Inputs

- PHoeNIx runs only on Illumina paired-end reads
- Multiple samples can be analyzed using a samplesheet.csv file

```
$ nextflow run cdcgov/phoenix -profile singularity -entry PHOENIX --input samplesheet.csv --kraken2db $PATH_TO_DB
```

Sample Sheet Input:

```
--input '[path to sample sheet file]'
```

- Sample sheet to be created with the information about the samples to be analyzed before running the pipeline
- Use **--input** parameter to specify the location

Advanced Molecular Detection
Southeast Region Bioinformatics

Sample Sheet Input

- Sample sheet must be a comma-separated file (.csv) with only **1** column and a header row
- **DO NOT HAVE ANY SPACES IN THIS FILE**
- Make sure the paths are full paths and not relative
- Automated sample sheet also can be created
  - This will be shown in a few slides
- Final sample sheet file consisting of paired-end data looks something like the image on the right

```
sample,fastq_1,fastq_2

JBI22000743,/blue/bphl-
florida/thsalikilakshmi/data/HAI/20220829_jax_220823_PLN_WLK_MS/fastqs/JBI22000743_1.fastq.gz,/blue/b
phl-florida/thsalikilakshmi/data/HAI/20220829_jax_220823_PLN_WLK_MS/fastqs/JBI22000743_2.fastq.gz

JBI22000770,/blue/bphl-
florida/thsalikilakshmi/data/HAI/20220829_jax_220823_PLN_WLK_MS/fastqs/JBI22000770_1.fastq.gz,/blue/b
phl-florida/thsalikilakshmi/data/HAI/20220829_jax_220823_PLN_WLK_MS/fastqs/JBI22000770_2.fastq.gz

JBI22000771,/blue/bphl-
florida/thsalikilakshmi/data/HAI/20220829_jax_220823_PLN_WLK_MS/fastqs/JBI22000771_1.fastq.gz,/blue/b
phl-florida/thsalikilakshmi/data/HAI/20220829_jax_220823_PLN_WLK_MS/fastqs/JBI22000771_2.fastq.gz
```

Advanced Molecular Detection
Southeast Region Bioinformatics

Sample Sheet Description

Column	Description
sample	Custom sample name. This entry will be identical for multiple sequencing libraries/runs from the same sample. Spaces in sample names are automatically converted to underscores (_)
fastq_1	Full path to *fastq* file for Illumina short reads 1. File has to be gzipped and have the extension ".fastq.gz" or ".fq.gz"
fastq_2	Full path to *fastq* file for Illumina short reads 2. File has to be gzipped and have the extension ".fastq.gz" or ".fq.gz"

Advanced Molecular Detection
Southeast Region Bioinformatics

Automated Sample Sheet Creation

- The sample sheet can be created automatically using the below command from a directory of *fastq* files

```
$ phoenix/bin/create_samplesheet.sh <directory of fastq files> > <output_directory>/samplesheet.csv
```

- The script will search 1 directory deep & attempt to determine sample id names and pairing/multilane information to create a sample sheet automatically
- **Please review the sample sheet for accuracy before using it in the pipeline**
- samplesheet.csv can be changed to your preferred name

Advanced Molecular Detection
Southeast Region Bioinformatics

Outputs

ANI

- Developed for fast alignment-free computation of whole-genome Average Nucleotide Identity (ANI)
- Avoids expensive sequence alignments and uses Mashmap as its MinHash based sequence mapping engine to compute the orthologous mappings and alignment identity estimates

AMRFinder

- AMRFinder and the accompanying database identify acquired antimicrobial resistance genes in bacterial protein and/or assembled nucleotide sequences as well as known resistance-associated point mutations for several taxa

Assembly

- SPAdes – St. Petersburg genome assembler – is an assembly toolkit containing various assembly pipelines
- SPAdes scaffold files are used for downstream analysis.

BUSCO (only run with --entry CDC_PHOENIX)

- BUSCO output is based on evolutionarily-informed expectations of gene content of near-universal single-copy orthologs, thus the BUSCO metric is complementary to technical metrics like N50

Advanced Molecular Detection
Southeast Region Bioinformatics

Outputs

fastp

- A tool designed to provide fast all-in-one preprocessing (trimming) of *fastq* files. This tool was developed in C++ with multithreading supported to afford high performance

FastQC

- Provides general quality metrics about your sequenced reads
- Provides information about the quality score distribution across your reads, per base sequence content (%A/T/G/C), adapter contamination, and overrepresented sequences

GAMMA

- GAMMA (Gene Allele Mutation Microbial Assessment) is a command line tool that finds gene matches in microbial genomic data using protein coding (rather than nucleotide) identity then translates and annotates the match by providing the type (i.e., mutant, truncation, etc.) and a translated description (i.e., Y190S mutant, truncation at residue 110, etc.)

Kraken2

- A taxonomic classifier using exact k-mer matches to achieve high accuracy and fast classification speeds
- This classifier matches each k-mer within a query sequence to the lowest common ancestor (LCA) of all genomes containing the given k-mer

Advanced Molecular Detection
Southeast Region Bioinformatics

Outputs

MLST

- Scans assembly files against traditional PubMLST typing schemes

QUAST

- Evaluates the quality of genome assemblies

removedAdapters

- BBDUK was developed to combine most common data-quality-related trimming, filtering, and masking operations into a single high-performance tool

SRST2 (only run with --entry CDC_PHOENIX)

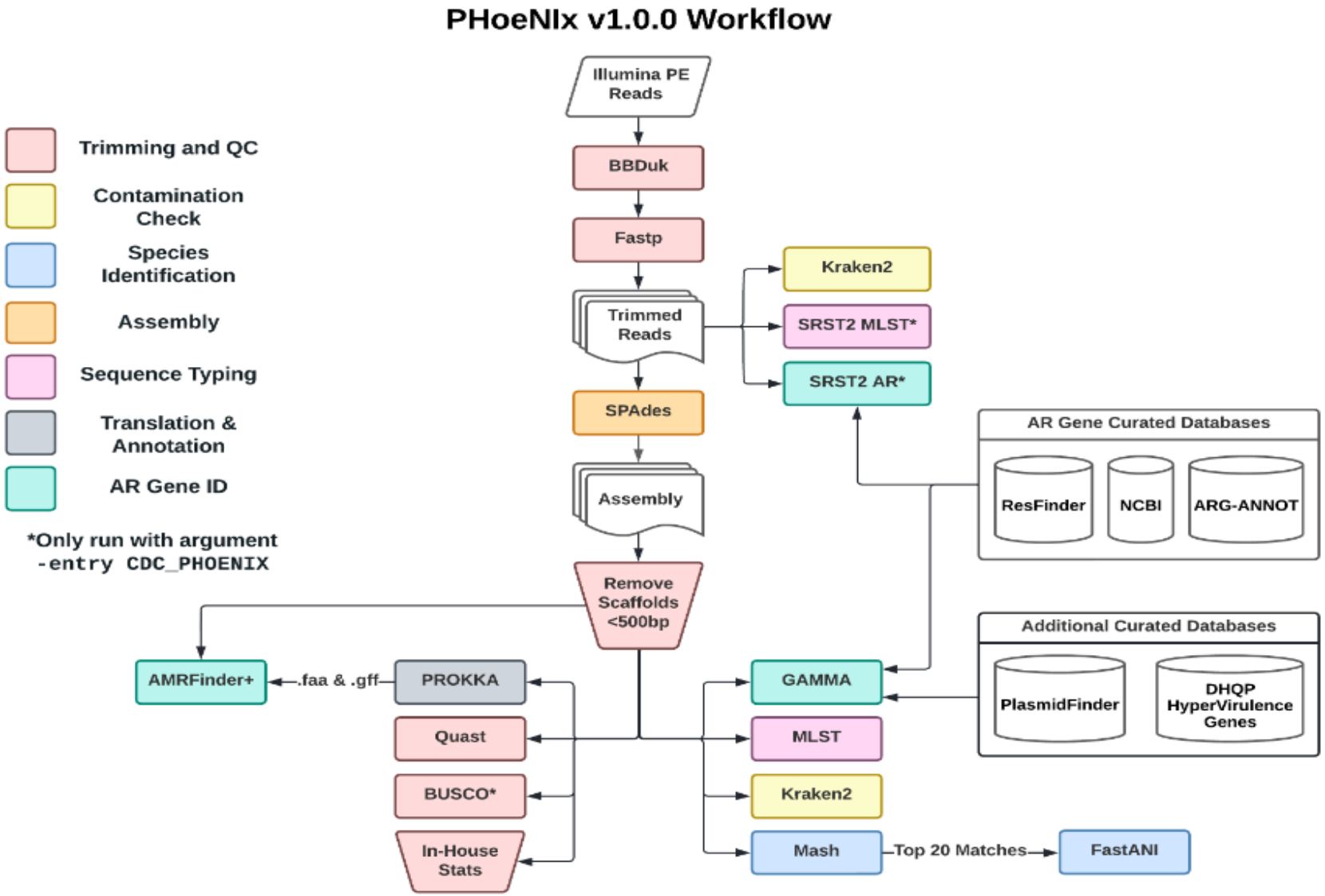
- Short Read Sequence Typing for Bacterial Pathogens

MultiQC

- A visualization tool that generates a single HTML report summarizing all samples in your project
- Results generated by MultiQC collate pipeline QC from supported tools e.g., FastQC

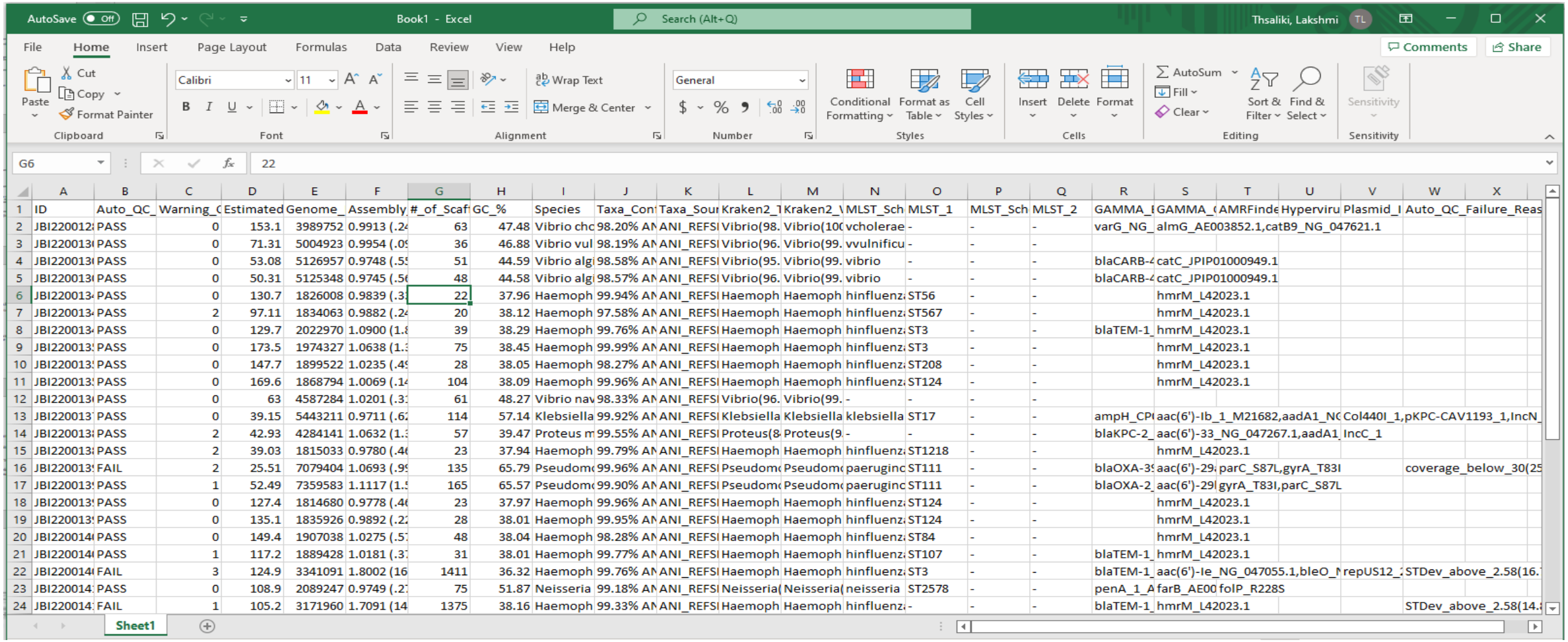
Advanced Molecular Detection
Southeast Region Bioinformatics

Pipeline Summary



Advanced Molecular Detection
Southeast Region Bioinformatics

Results



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	ID	Auto_QC_	Warning_(	Estimated	Genome_	Assembly	#_of_Scaf	GC_%	Species	Taxa_Con	Taxa_Sour	Kraken2_1	Kraken2_\	MLST_Sch	MLST_1	MLST_Sch	MLST_2	GAMMA_I	GAMMA_(	AMRFinde	Hyperviru	Plasmid_I	Auto_QC_Failure_Reas	
2	JBI220012	PASS	0	153.1	3989752	0.9913 (.24	63	47.48	Vibrio cho	98.20% AN	ANI_REFSI	Vibrio(98.	Vibrio(100	vcholerae	-	-	-	varG_NG_	almG_AE003852.1,catB9_NG_047621.1					
3	JBI220013	PASS	0	71.31	5004923	0.9954 (.09	36	46.88	Vibrio vul	98.19% AN	ANI_REFSI	Vibrio(96.	Vibrio(99.	vvulnificu	-	-	-							
4	JBI220013	PASS	0	53.08	5126957	0.9748 (.55	51	44.59	Vibrio alg	98.58% AN	ANI_REFSI	Vibrio(95.	Vibrio(99.	vibrio	-	-	-	blaCARB-4	catC_JPIP01000949.1					
5	JBI220013	PASS	0	50.31	5125348	0.9745 (.56	48	44.58	Vibrio alg	98.57% AN	ANI_REFSI	Vibrio(96.	Vibrio(99.	vibrio	-	-	-	blaCARB-4	catC_JPIP01000949.1					
6	JBI220013	PASS	0	130.7	1826008	0.9839 (.3	22	37.96	Haemoph	99.94% AN	ANI_REFSI	Haemoph	Haemoph	hinfluenz	ST56	-	-		hmrM_L42023.1					
7	JBI220013	PASS	2	97.11	1834063	0.9882 (.24	20	38.12	Haemoph	97.58% AN	ANI_REFSI	Haemoph	Haemoph	hinfluenz	ST567	-	-		hmrM_L42023.1					
8	JBI220013	PASS	0	129.7	2022970	1.0900 (1.8	39	38.29	Haemoph	99.76% AN	ANI_REFSI	Haemoph	Haemoph	hinfluenz	ST3	-	-	blaTEM-1_	hmrM_L42023.1					
9	JBI220013	PASS	0	173.5	1974327	1.0638 (1.3	75	38.45	Haemoph	99.99% AN	ANI_REFSI	Haemoph	Haemoph	hinfluenz	ST3	-	-		hmrM_L42023.1					
10	JBI220013	PASS	0	147.7	1899522	1.0235 (.49	28	38.05	Haemoph	98.27% AN	ANI_REFSI	Haemoph	Haemoph	hinfluenz	ST208	-	-		hmrM_L42023.1					
11	JBI220013	PASS	0	169.6	1868794	1.0069 (.14	104	38.09	Haemoph	99.96% AN	ANI_REFSI	Haemoph	Haemoph	hinfluenz	ST124	-	-		hmrM_L42023.1					
12	JBI220013	PASS	0	63	4587284	1.0201 (.33	61	48.27	Vibrio nav	98.33% AN	ANI_REFSI	Vibrio(96.	Vibrio(99.	-	-	-	-							
13	JBI220013	PASS	0	39.15	5443211	0.9711 (.62	114	57.14	Klebsiella	99.92% AN	ANI_REFSI	Klebsiella	Klebsiella	klebsiella	ST17	-	-	ampH_CP	aac(6')-Ib_1_M21682,aadA1_N			Col440I_1,pKPC-CAV1193_1,IncN_		
14	JBI220013	PASS	2	42.93	4284141	1.0632 (1.3	57	39.47	Proteus m	99.55% AN	ANI_REFSI	Proteus(8	Proteus(9	-	-	-	-	blaKPC-2_	aac(6')-33_NG_047267.1,aadA1_			IncC_1		
15	JBI220013	PASS	2	39.03	1815033	0.9780 (.46	23	37.94	Haemoph	99.79% AN	ANI_REFSI	Haemoph	Haemoph	hinfluenz	ST1218	-	-		hmrM_L42023.1					
16	JBI220013	FAIL	2	25.51	7079404	1.0693 (.95	135	65.79	Pseudom	99.96% AN	ANI_REFSI	Pseudom	Pseudom	paeruginc	ST111	-	-	blaOXA-35	aac(6')-29	parC_S87L,gyrA_T83I			coverage_below_30(25	
17	JBI220013	PASS	1	52.49	7359583	1.1117 (1.5	165	65.57	Pseudom	99.90% AN	ANI_REFSI	Pseudom	Pseudom	paeruginc	ST111	-	-	blaOXA-2_	aac(6')-29	gyrA_T83I,parC_S87L				
18	JBI220013	PASS	0	127.4	1814680	0.9778 (.46	23	37.97	Haemoph	99.96% AN	ANI_REFSI	Haemoph	Haemoph	hinfluenz	ST124	-	-		hmrM_L42023.1					
19	JBI220013	PASS	0	135.1	1835926	0.9892 (.22	28	38.01	Haemoph	99.95% AN	ANI_REFSI	Haemoph	Haemoph	hinfluenz	ST124	-	-		hmrM_L42023.1					
20	JBI220140	PASS	0	149.4	1907038	1.0275 (.57	48	38.04	Haemoph	98.28% AN	ANI_REFSI	Haemoph	Haemoph	hinfluenz	ST84	-	-		hmrM_L42023.1					
21	JBI220140	PASS	1	117.2	1889428	1.0181 (.37	31	38.01	Haemoph	99.77% AN	ANI_REFSI	Haemoph	Haemoph	hinfluenz	ST107	-	-	blaTEM-1_	hmrM_L42023.1					
22	JBI220140	FAIL	3	124.9	3341091	1.8002 (16	1411	36.32	Haemoph	99.76% AN	ANI_REFSI	Haemoph	Haemoph	hinfluenz	ST3	-	-	blaTEM-1_	aac(6')-Ie_NG_047055.1,bleO_N			repUS12_2	STDev_above_2.58(16.7	
23	JBI220014	PASS	0	108.9	2089247	0.9749 (.27	75	51.87	Neisseria	99.18% AN	ANI_REFSI	Neisseria(	Neisseria(	neisseria	ST2578	-	-	penA_1_A	farB_AE00	folP_R228S				
24	JBI220014	FAIL	1	105.2	3171960	1.7091 (14	1375	38.16	Haemoph	99.33% AN	ANI_REFSI	Haemoph	Haemoph	hinfluenz	-	-	-	blaTEM-1_	hmrM_L42023.1				STDev_above_2.58(14.3	

Advanced Molecular Detection
Southeast Region Bioinformatics

Link to PHoeNIx

GitHub link for CDC PHoeNIx pipeline:

[Home · CDCgov/phoenix Wiki · GitHub]

Advanced Molecular Detection
Southeast Region Bioinformatics

Editing the HPC config file for HPG

- There is a way to edit the HPC config file so that PHoeNIx will run programs simultaneously for quicker results/analyses
- These files have also been uploaded to your states **public-share** on HPG

Advanced Molecular Detection
Southeast Region Bioinformatics

Editing the HPC config file for HPG

- Upload the HPC_Template.config to /path/to/phoenix/conf/ and overwrite the previous version
- Upload the nextflow.config to /path/to/phoenix/ and overwrite the previous version
- Edit the phnx_trial.sh using nano or emacs
  - Make sure all the paths are correct
  - Add the path to your mamba nextflow environment
  - Add your email
- Copy the phnx_trial.sh to your working directory and **sbatch phnx_trial.sh**

Advanced Molecular Detection
Southeast Region Bioinformatics

Updates from developer Dr. Jill Hagey

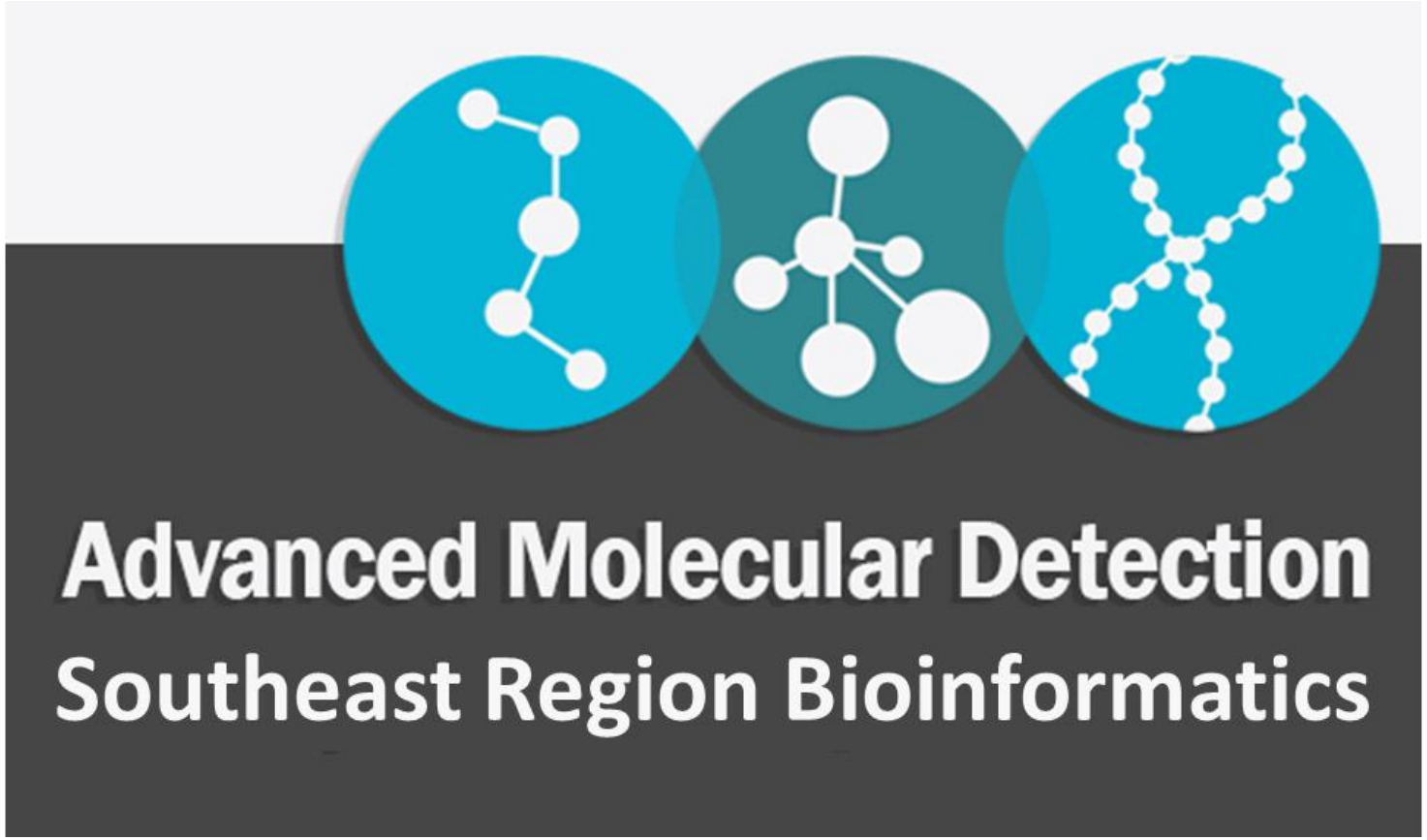
- **Note** – make sure to download the latest version **v1.1.1**
  - phoenix/CHANGELOG.md at main · CDCgov/phoenix · GitHub
  - Check this link for the updates and bug fixes!
- There is an update coming – so we'll make sure to notify you of this!
  - This will include a post-assembly entrypoint
  - Updated Kraken2 database
  - SRA pull option
- PHoeNIx Output Changes · CDCgov/phoenix · Discussion #95 · GitHub
  - There is an option for CDC_PHOENIX entrypoint
  - This will have a few additional outputs, including a GRiPHin report (AMR)
  - If you run this, please respond to this discussion to help the development of this pipeline!

Advanced Molecular Detection
Southeast Region Bioinformatics

Time for Questions & Feedback

- Questions?
  - Do you need help with anything?
  - Requests for separate trainings?
- Feedback
  - What would you like to see?

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

Sarah Schmedes, PhD
Bioinformatics Supervisor
Sarah.Schmedes@flhealth.gov