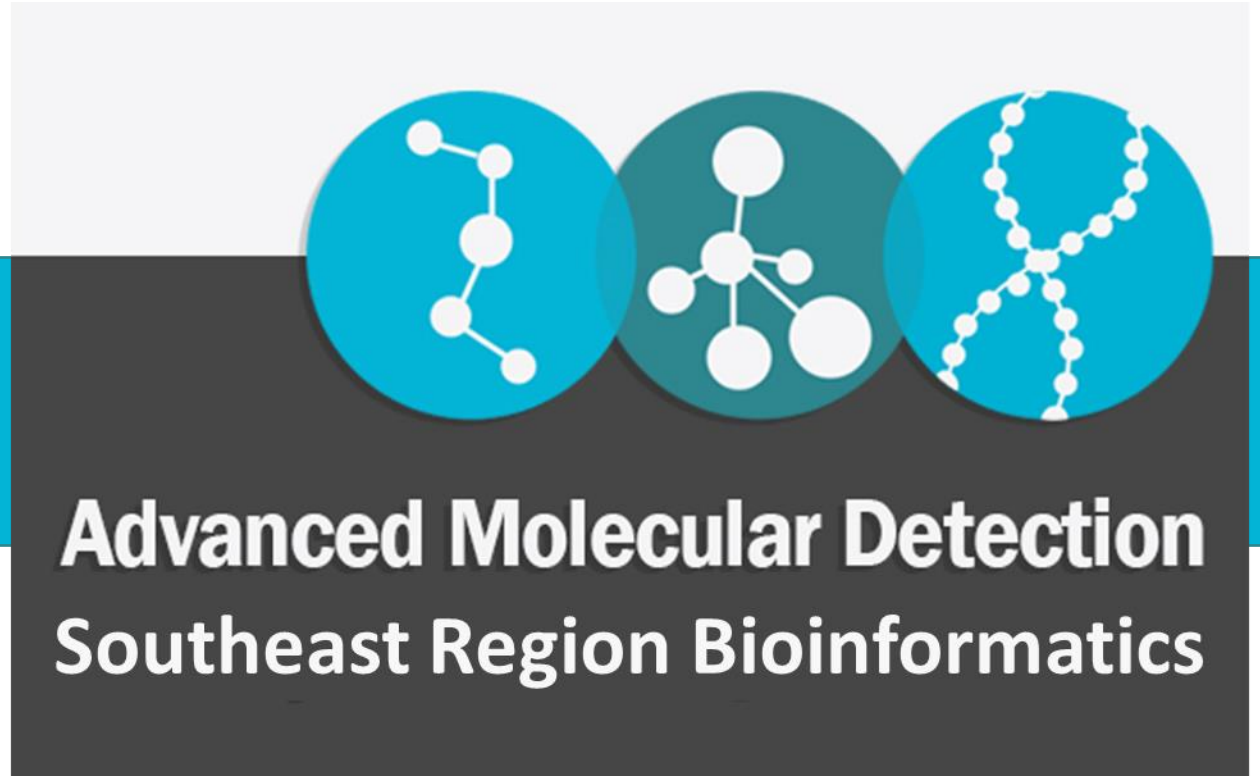


Bacteria Subtyping & Bactopia tools

07/10/2023

Outline

Agenda

Bacterial Subtyping

Typing Methods

MLST

Bactopia tools - ABRicate & AgrVATE

Agenda

July 24 – Bactopia Tools: AMRFinderPlus, BUSCO, and CheckM
August 7 – Bactopia Tools: Emmtyper, FastANI, and Gamma

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

Introduction

- Bacterial subtyping determines the similarity between bacteria isolates of the same species. If bacteria have the same subtype, they are more likely to be related to each other than if they have different subtypes.
- Subtyping is used in epidemiologic investigations to determine levels of relatedness between isolates to infer transmission.

Advanced Molecular Detection
Southeast Region Bioinformatics

Importance of Sequencing Bacteria

- The main purpose for sequencing in public health is for microbial characterization for source attribution/transmission dynamics, epi investigations, etc. Reference genomes are also a bonus in sequencing.
- Sequencing helps us improve our understanding about the biology of many bacterial pathogens as well as identification of novel antibiotic targets.

Advanced Molecular Detection
Southeast Region Bioinformatics

Aim of Bacterial Typing

- Confirm epidemiological relationships in the spread of infection
- Provide epidemiological hypotheses about epidemiological relationships between bacteria in the absence of epidemiological data
- Describe the distribution of bacterial types and identification of affecting factors
- Detect & track foodborne disease outbreaks
- Track sources of food contamination
- Understand population genetics
- Understand epidemiology & ecology of foodborne pathogens

Advanced Molecular Detection
Southeast Region Bioinformatics

Choosing a Typing Method

- Typing method depends on:
  - Skill level
  - Resources of the laboratory
  - Aim & scope of the study
- Bacterial strains can be differentiated based on their phenotypic or genotypic differences. Genotyping shows better performance than phenotyping characterization methods
- Commonly used typing methods include:
  - Phenotypic typing methods
  - Molecular typing methods

Advanced Molecular Detection
Southeast Region Bioinformatics

Phenotypic Typing

Detects characteristics expressed by microorganisms. These tests are based on biochemical, antigenic, or susceptibility (to phages or antimicrobial agents) properties of the organism. Some of the methods include:

- Bio typing
- Sero typing
- Phage typing
- Resisto typing
- Bacteriocin typing
- Antibiogram typing

Advanced Molecular Detection
Southeast Region Bioinformatics

Molecular typing

- Based on the analysis of chromosomal or extrachromosomal genetic elements (such as plasmids) of the organism
- Commonly used molecular typing methods are:
  - MLST
  - PFGE
  - WGS

Advanced Molecular Detection
Southeast Region Bioinformatics

MLST

- Multilocus sequence typing (MLST) is a molecular typing technique which uses DNA sequences of internal fragments of multiple housekeeping genes to characterize isolates of microbial species
- MLST tools used to extract information from WGS data include:
  - cgMLST – aims to combine the discriminatory power of classical MLST with the extensive genetic data obtained by WGS
  - wgMLST – this is an extension to cgMLST which uses a set of accessory loci in addition to a set of core genome loci

Advanced Molecular Detection
Southeast Region Bioinformatics

Principle of MLST

- MLST directly measures the DNA sequence variations in a set of housekeeping genes and characterizes strains by their unique allelic profiles.
- The principle of MLST is simple: technique involves PCR amplification followed by DNA sequencing. Nucleotide differences between strains can be checked at a variable number of genes depending on the degree of discrimination desired.

Advanced Molecular Detection
Southeast Region Bioinformatics

MLST Workflow

Data collection

Definitive identification of variation is obtained by nucleotide sequence determination of gene fragments

Data analysis

All unique sequences are assigned allele numbers, combined into an allelic profile, and assigned a sequence type (ST). If new alleles and STs are found, they are stored in the database after verification

Multilocus sequence analysis

Relatedness of isolates determined by comparing allelic profiles

Advanced Molecular Detection
Southeast Region Bioinformatics

Advantages

- Highly unambiguous and portable
- Reproducible and scalable
- Materials required for ST determination can be exchanged between laboratories
- Primer sequences and protocols can be accessed electronically
- MLST combines advances in high throughput sequencing and bioinformatics with established population genetics techniques
- MLST data can be used to investigate evolutionary relationships among bacteria

Advanced Molecular Detection
Southeast Region Bioinformatics

Limitations

- MLST appears best in population genetic studies, but it is expensive
- Due to the sequence conservation in housekeeping genes, MLST sometimes lacks the discriminatory power to differentiate bacterial strains, which limits its use in epidemiological investigations

Advanced Molecular Detection
Southeast Region Bioinformatics

Bactopia

Bactopia

- Bactopia is a flexible pipeline for complete analysis of bacterial genomes
- Bactopia was inspired by Staphopia, a workflow that targets *Staphylococcus aureus* genomes
- Bactopia was developed from scratch prioritizing usability, portability, and speed

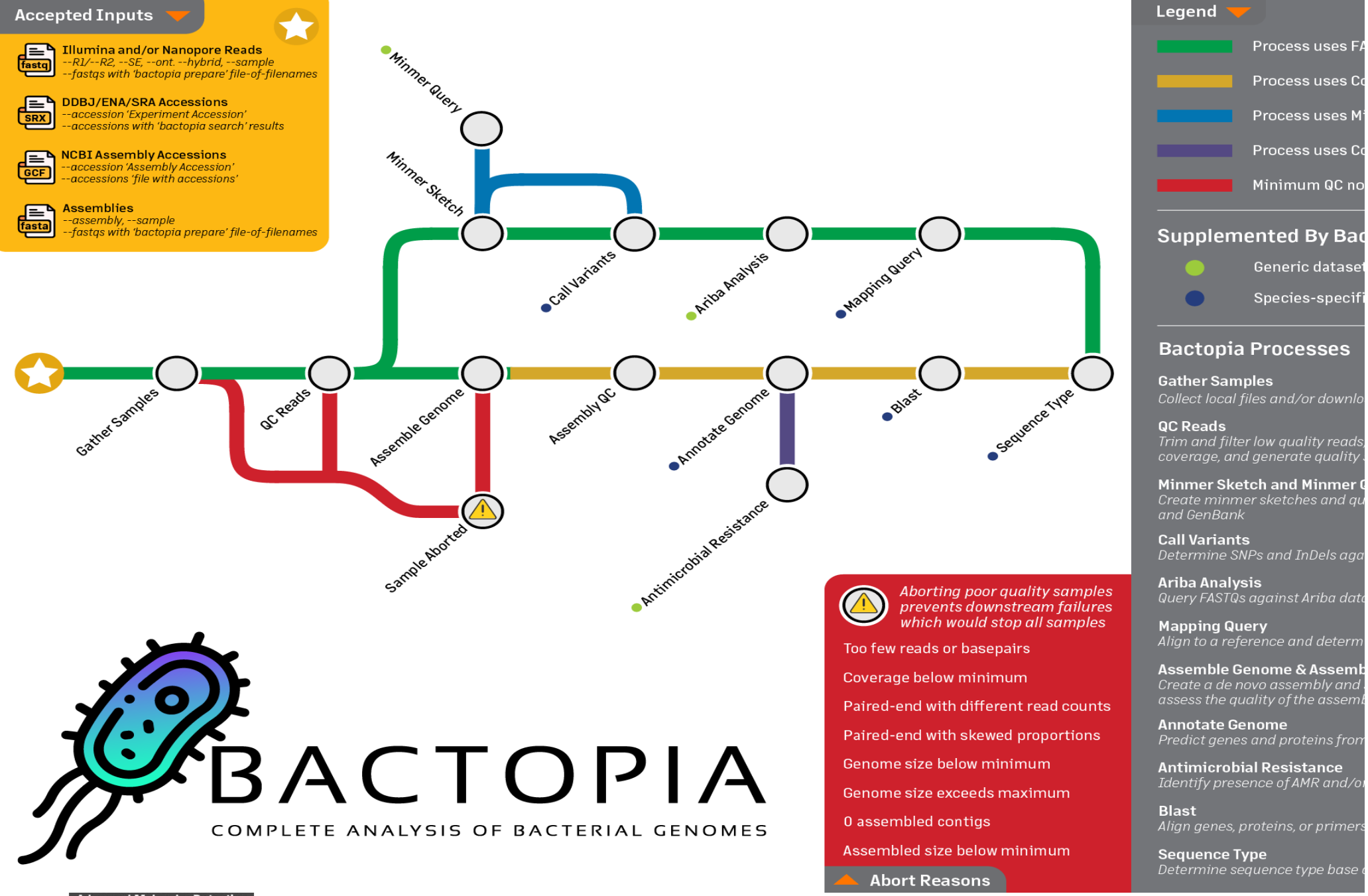
Advanced Molecular Detection
Southeast Region Bioinformatics

Bactopia Usage

- Bactopia uses Nextflow to manage the workflow – which supports many types of environments (e.g., cluster or cloud)
- Bactopia allows for the usage of many public datasets as well as your own datasets to further enhance the analysis of your sequencing data
- Bactopia only uses software packages available from Bioconda (or other Anaconda channels) to make installation simple for *all* users

Advanced Molecular Detection
Southeast Region Bioinformatics

Workflow



BACTOPIA

COMPLETE ANALYSIS OF BACTERIAL GENOMES

Advanced Molecular Detection
Southeast Region Bioinformatics

Bactopia Tools

ABRicate

- Mass screening of contigs for antimicrobial resistance or virulence genes. It comes bundled with multiple databases: NCBI, CARD, ARG-ANNOT, Resfinder, MEGARES, EcOH, PlasmidFinder, Ecoli_VF and VFDB
- Supports contigs only (assemblies), not raw ***.fastq*** reads
- Detects acquired resistance genes, NOT point mutations
- Uses a DNA sequence database, not protein
- Needs BLAST+>=2.7 and anyfasta to be installed
- Written in perl
- tseemann/abricate - GitHub

Advanced Molecular Detection
Southeast Region Bioinformatics

Installation

- Available as a module on HPG

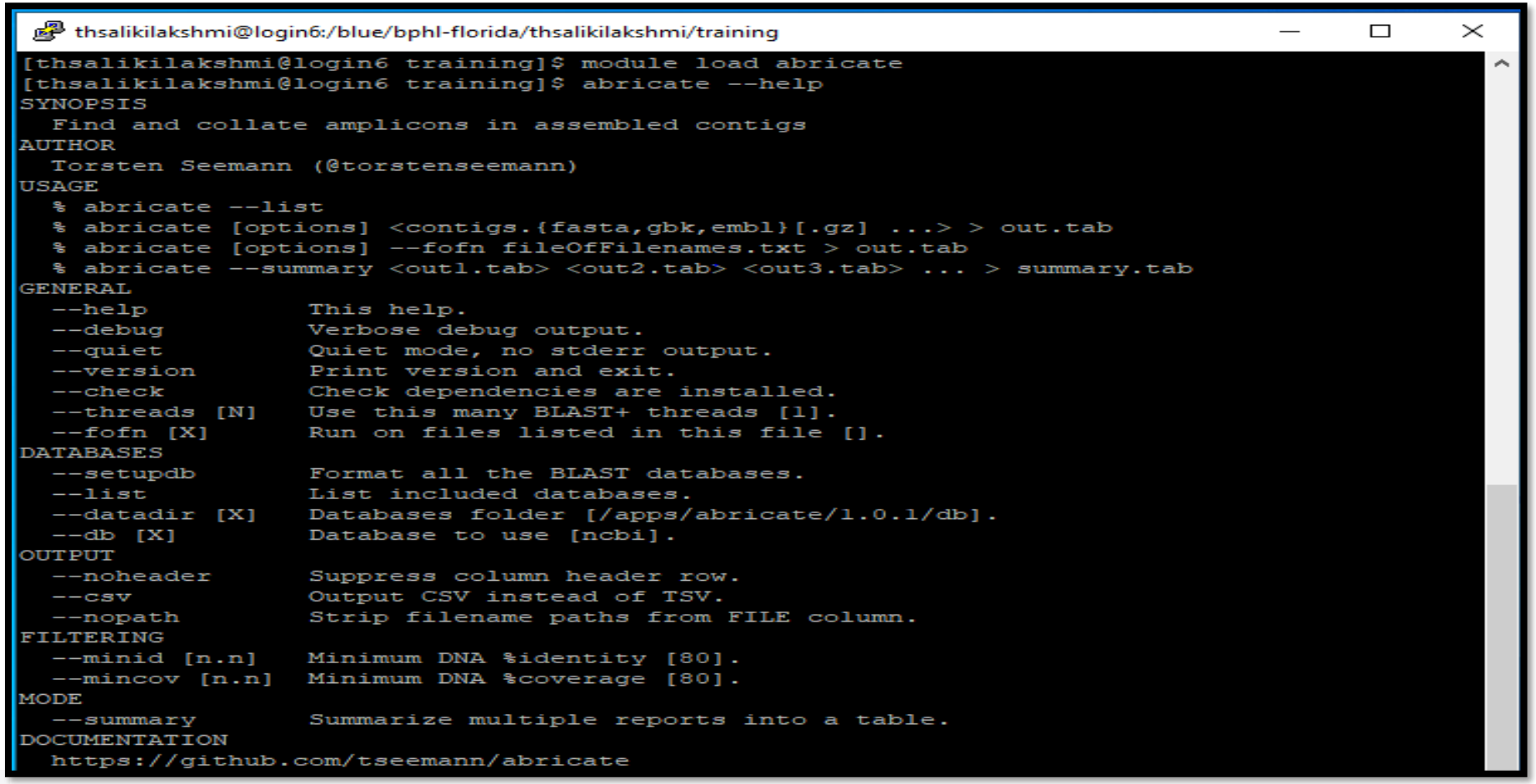
```
module load abricate
```

- Can also be installed through conda

```
conda create -yp /blue/bphl-<state>/<user>/conda_envs/abricate/
conda activate /blue/bphl-<state>/<user>/conda_envs/abricate/
conda install -c conda-forge -c bioconda abricate
```

Advanced Molecular Detection
Southeast Region Bioinformatics

ABRicate Usage



Advanced Molecular Detection
Southeast Region Bioinformatics

ABRicate Output

```
$ abricate JBE*.fasta > results_ABR
```

/blue/bphl-florida/thsalikilakshmi/data/HAI/20220727_jax_220708_PLN_WLK_MS_test/assemblies/results_ABR - hpg.rc.ufl.edu - Editor - WinSCP

Encoding ▾ Color ▾

#FILE	SEQUENCE	START	END	STRAND	GENE	COVERAGE	COVERAGE_MAP	GAPS	%COVERAGE	%IDENTITY	DATABASE	ACCESSION	PRODUCT
JBE22000155.fasta	31	3461	4594	-	blaEC-18	1-1134/1134	==========	0/0	100.00	99.21	ncbi	NG_049083.1	class C extended
JBE22000165.fasta	153	6463	7662	-	tet(A)	1-1200/1200	==========	0/0	100.00	99.92	ncbi	NG_048154.1	tetracycline efflux MFS
JBE22000165.fasta	19	30899	32032	+	blaEC-18	1-1134/1134	==========	0/0	100.00	99.21	ncbi	NG_049083.1	class C extended
JBE22000165.fasta	85	5366	6226	+	blaTEM-1	1-861/861	==========	0/0	100.00	100.00	ncbi	NG_050145.1	class A broad-sp
JBE22000165.fasta	85	7707	8522	-	aph(3')-Ia	1-816/816	==========	0/0	100.00	100.00	ncbi	NG_047430.1	aminoglycoside O
JBE22000165.fasta	85	9444	10184	-	aph(6)-Id	1-741/837	==========.	0/0	88.53	100.00	ncbi	NG_047465.1	aminoglycoside O
JBE22000165.fasta	85	10184	11011	-	aph(3'')-Ib	1-828/828	==========	0/0	100.00	100.00	ncbi	NG_056002.2	aminoglycoside O
JBE22000165.fasta	85	11048	11863	-	sul2	1-816/816	==========	0/0	100.00	99.88	ncbi	NG_048118.1	sulfonamide-resistant dil
JBE22000179.fasta	41	3470	4603	-	blaEC-18	1-1134/1134	==========	0/0	100.00	99.21	ncbi	NG_049083.1	class C extended
JBE22000210.fasta	119	8754	9887	+	blaEC-15	1-1134/1134	==========	0/0	100.00	98.59	ncbi	NG_049081.1	class C extended
JBE22000212.fasta	20	3194	4327	-	blaEC-15	1-1134/1134	==========	0/0	100.00	98.06	ncbi	NG_049081.1	class C extended
JBE22000216.fasta	17	349	1554	-	tet(B)	1-1206/1206	==========	0/0	100.00	100.00	ncbi	NG_048163.1	tetracycline efflux MFS
JBE22000216.fasta	41	30578	31711	+	blaEC-15	1-1134/1134	==========	0/0	100.00	98.06	ncbi	NG_049081.1	class C extended
JBE22000247.fasta	1	96236	97369	-	blaEC-15	1-1134/1134	==========	0/0	100.00	98.06	ncbi	NG_049081.1	class C extended
JBE22000268.fasta	20	3384	4517	-	blaEC-18	1-1134/1134	==========	0/0	100.00	99.21	ncbi	NG_049083.1	class C extended

Advanced Molecular Detection
Southeast Region Bioinformatics

ABRicate Output Description

ABRicate produces a tab-separated output file with the following columns:

- FILE – The filename this hit came from
- SEQUENCE – The sequence in the filename
- START – Start coordinate in the sequence
- END – End coordinate
- STRAND – Strand + or –
- GENE – AMR gene name
- COVERAGE – What proportion of the gene is in our sequence

Advanced Molecular Detection
Southeast Region Bioinformatics

ABRicate Output Description

- COVERAGE_MAP – A visual representation of the hit. (==aligned, .=unaligned, /=has_gaps)
- GAPS – Openings/ gaps in subject and query - possible pseudogene?
- %COVERAGE – Proportion of gene covered
- %IDENTITY – Proportion of exact nucleotide matches
- DATABASE – The database this sequence comes from
- ACCESSION – The genomic source of the sequence
- PRODUCT – Gene product (if available)
- RESISTANCE – Putative antibiotic resistance phenotype, ; -separated

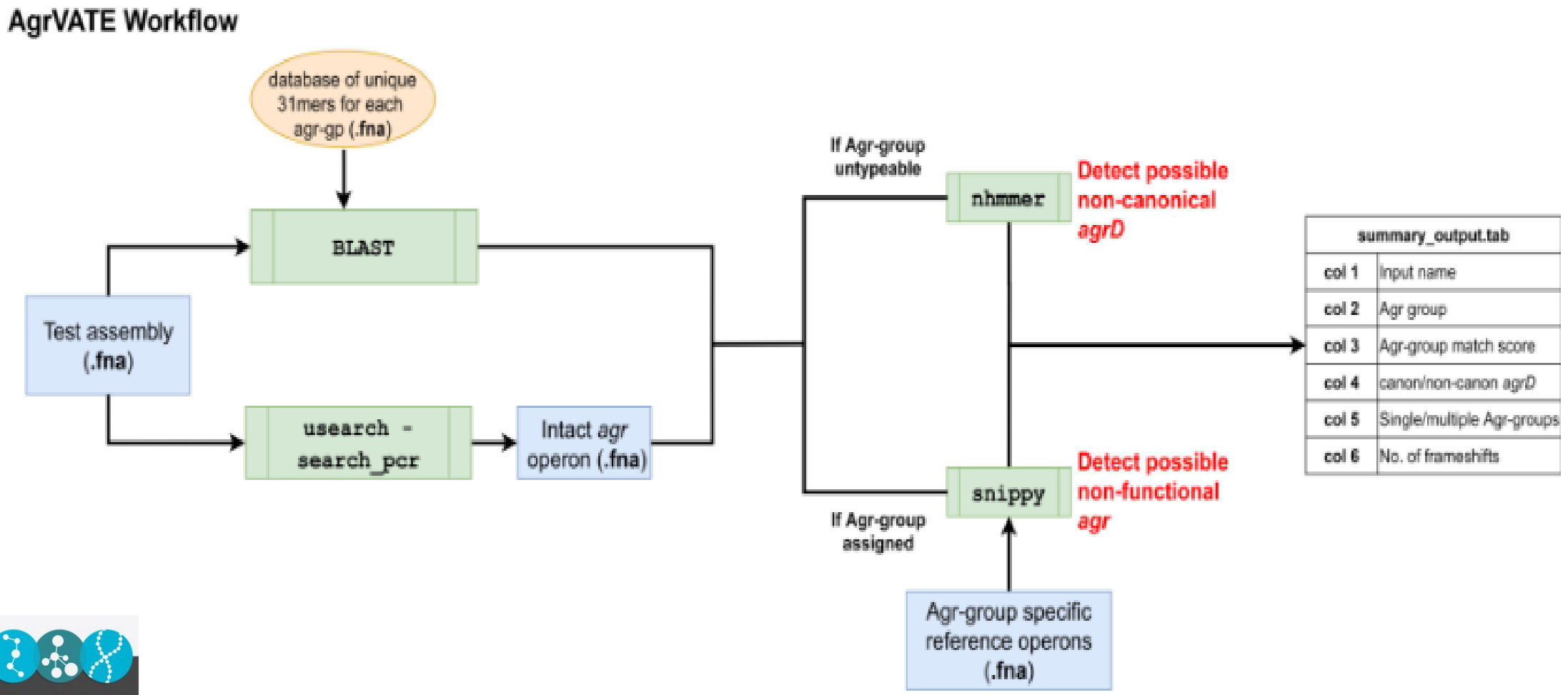
Advanced Molecular Detection
Southeast Region Bioinformatics

AgrVATE

- Agr variant Assessment & Typing Engine
- AgrVATE is a tool for rapid identification of *Staphylococcus aureus agr* locus type and reports possible variants in the *agr* operon
- AgrVATE accepts a *S. aureus* genome assembly as input and performs a k-mer search using an Agr-group specific k-mer database to assign the Agr-group
- The *agr* operon is then extracted using *in-silico* PCR and variants are called using an Agr-group specific reference operon
- VishnuRaghuram94/AgrVATE - GitHub

Advanced Molecular Detection
Southeast Region Bioinformatics

AgrVATE Workflow



Advanced Molecular Detection
Southeast Region Bioinformatics

Installation

Can be installed through conda

```
conda create -yp /blue/bphl-<state>/<user>/conda_envs/agrvate/
conda activate /blue/bphl-<state>/<user>/conda_envs/agrvate/
conda install -c conda-forge -c bioconda agrvate
```

Advanced Molecular Detection
Southeast Region Bioinformatics

Usage

agrvate -i filename.fasta [options]

```
(/blue/bphl-florida/thsalikilakshmi/training/conda_envs/agrvate) [thsalikilakshmi@login1 conda_envs]$ agrvate --help

AgrVATE: Agr Variant Assessment & Typing Engine

VERSION: agrvate v1.0.2

USAGE:   agrvate [options] -i filename.fasta

FLAGS:
  -i | --input           Input S. aureus genome in FASTA format
  -t | --typing-only     Does agr typing only (skips agr operon extraction and frameshift detection)
  -m | --mummer          Uses mummer instead of usearch (May not perform frameshift detection)
  -f | --force           Force overwrite existing results directory
  -d | --databases       Path to agrvate_databases (Not required if installed using Conda)
  -h | --help            Print this help message and exit
  -v | --version         Print version and exit

SOURCE:  https://github.com/VishnuRaghuram94/AgrVATE
```

Advanced Molecular Detection
Southeast Region Bioinformatics

Output

- A new directory with suffix –results will be created where all the files can be found
- fasta-summary tab

col 1: Filename
col 2: Agr group (gp1/gp2/gp3/gp4). 'u' means unknown. If multiple agr groups were found (col 5 = m), the displayed agr group is the majority/highest confidence.
col 3: Match score for agr group (maximum 15; 0 means untypeable; < 5 means low confidence)
col 4: Canonical or non-canonical agrD (1 means canonical; 0 means non-canonical; u means unknown)
col 5: If multiple agr groups were found, likely due to multiple S. aureus isolates in sequence (s means single, m means multiple, u means unknown)
col 6: Number of frameshifts found in CDS of extracted agr operon (Column is 'u' if agr operon was not extracted)

Advanced Molecular Detection
Southeast Region Bioinformatics

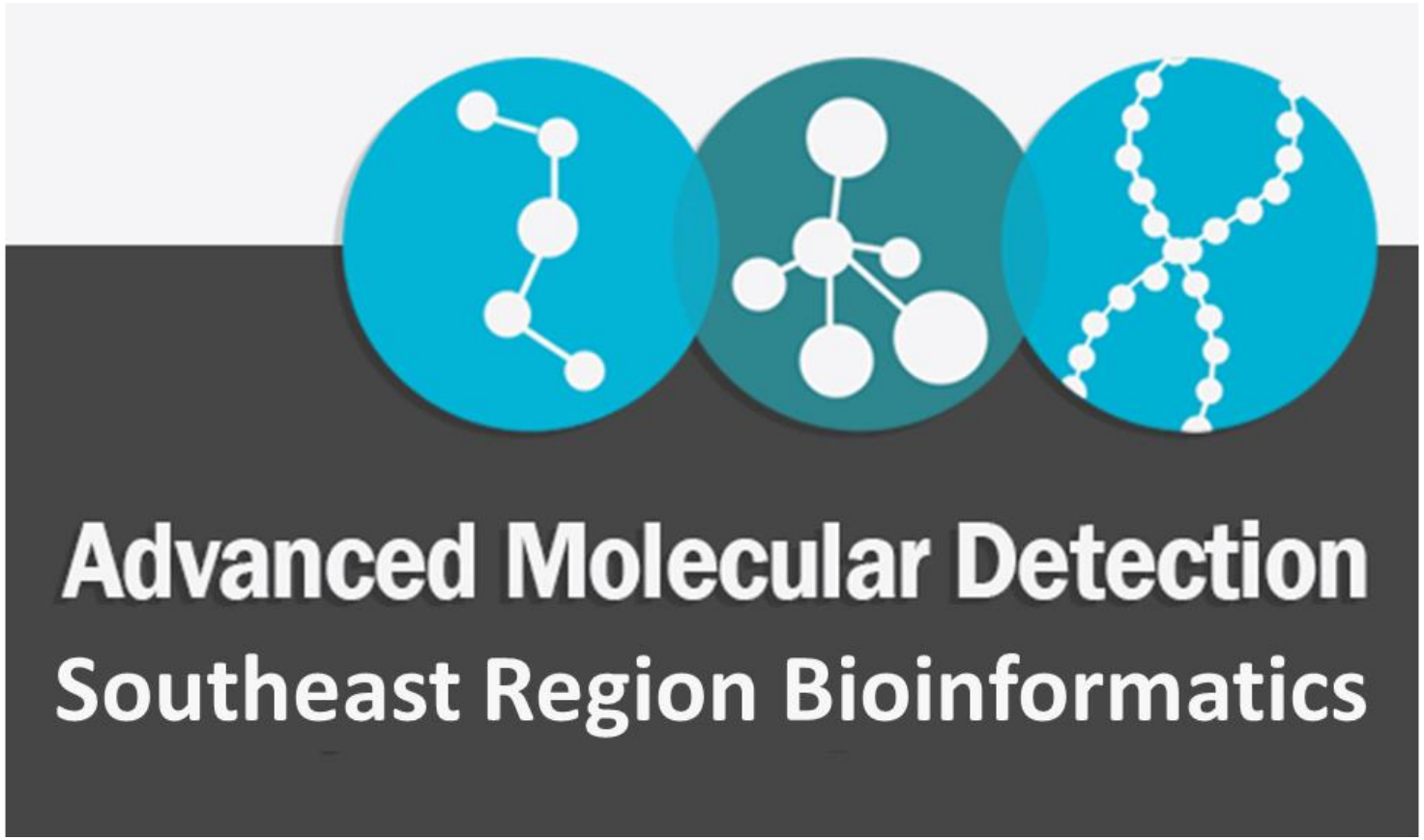
Results

```
agrvate -i JBE22000638.fasta -t -f
```

	A	B	C	D	E	F
1	#filename	agr_group	match_score	canonical_agrD	multiple_agr	frameshifts
2	JBE22000638	u	0	u	u	u

NOTE: There are 15 possible k-mers for each agr group per genome. The analyses will continue even if only one k-mer matches a given *agr*-group but it should be noted that < 5 k-mers matching leads to a low confidence *agr*-group call. Col 3 in **fasta-summary.tab** shows the number of k-mers matched

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