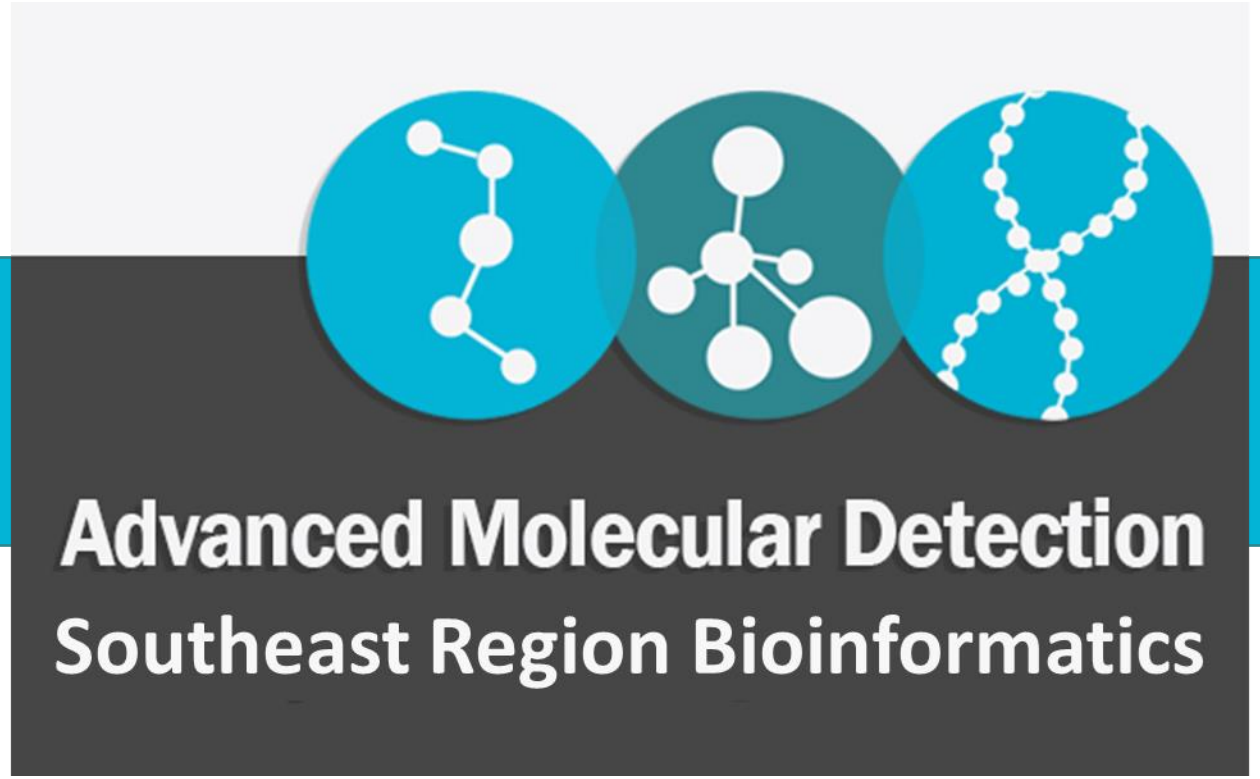


Bactopia tools

08/21/2023

Outline

Updates

Agenda

Bactopia

FastANI

GAMMA

Questions

Updates – ABiL Trainings

ABiL courses

- Online courses will be available shortly for the attendees
  - Pathogen Phylogenomics
  - Quality Assessment of Sequencing Data
- If additional attendees decide to sign up later, that is not a problem, as the courses will still be available, and they can attend on a rolling basis under the contract.

Advanced Molecular Detection
Southeast Region Bioinformatics

Agenda

September 4 rescheduled to September 11 – Bactopia Tools: HICAP and HpsuisSero
September 18 – Bactopia Tools: Kleborate and Legsta

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

FastANI

- Fast Whole-Genome Similarity (ANI) Estimation
- ANI is defined as mean nucleotide identity of orthologous gene pairs shared between two microbial genomes
- FastANI supports pairwise comparison of both complete and draft genome assemblies

ParBLiSS/FastANI: Fast Whole-Genome Similarity (ANI) Estimation (github.com)

Advanced Molecular Detection
Southeast Region Bioinformatics

Installation

- Available as a module on HPG

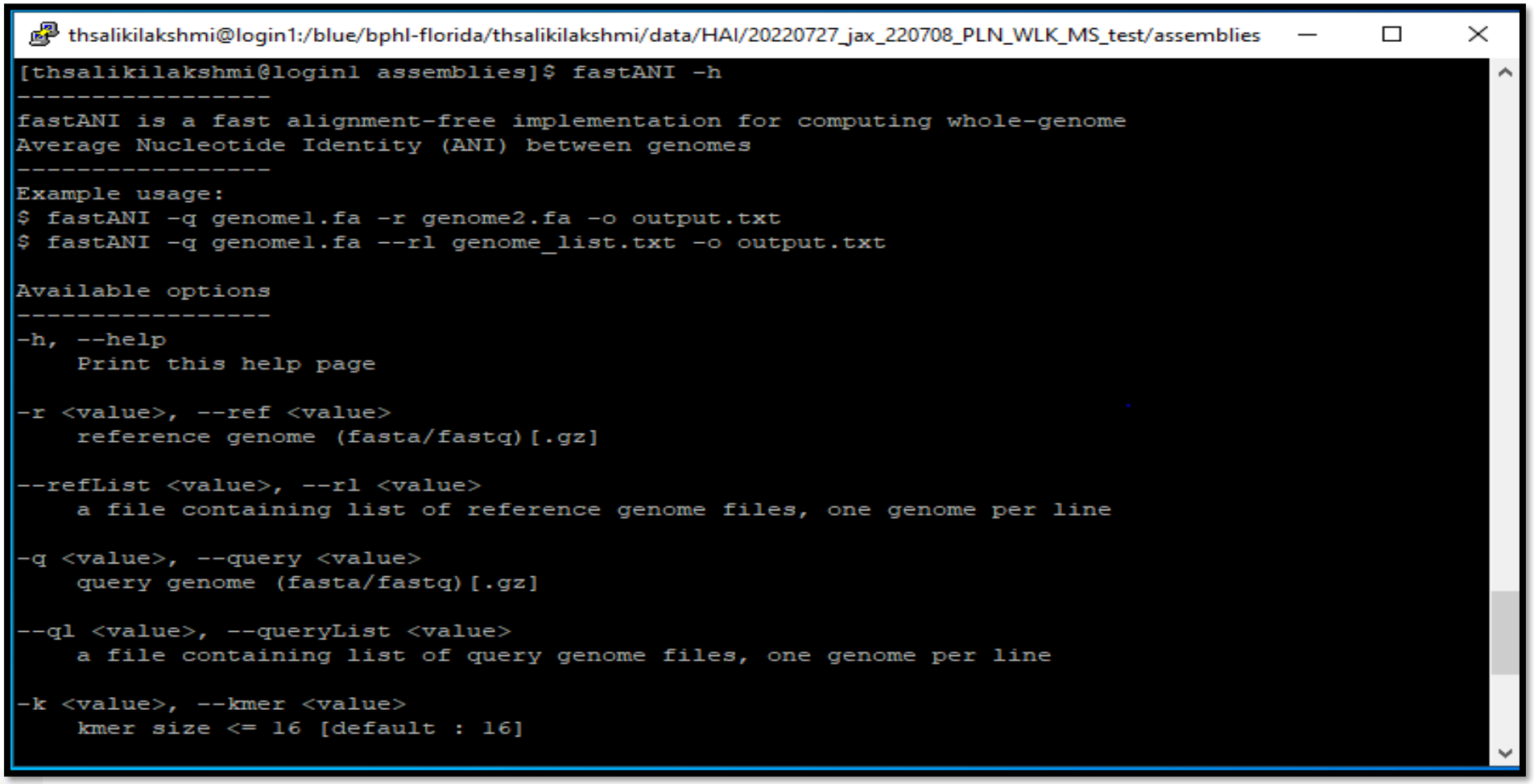
```
module load fastani/1.1
```

- Can be installed through conda

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

Advanced Molecular Detection
Southeast Region Bioinformatics

Usage



Advanced Molecular Detection
Southeast Region Bioinformatics

Input

- To compute ANI, use a query genome and a reference genome
- Here we computed ANI between *Escherichia coli* and *Shigella flexneri* genomes
- *E. coli* is provided as a query genome, *S. flexneri* is the reference genome and –o flag is used to give fastani.out as output file

```
[thsalikilakshmi@login2 assemblies]$ fastANI -q /blue/bphl-florida/thsalikilakshmi/data/HAI/20
220727_jax_220708_PLN_WLK_MS_test/assemblies/JBE22000247.fasta -r /blue/bphl-florida/thsalikil
akshmi/data/jbi/20220825_jax_220629_PLN_WLK_MS/assemblies/JBI22000647.fasta -o fastani.out > r
esults_fastani
```

Advanced Molecular Detection
Southeast Region Bioinformatics

Output

```
[thsalikilakshmi@login2 assemblies]$ fastANI -q /blue/bphl-florida/thsalikilakshmi/data/HAI/20
220727_jax_220708_PLN_WLK_MS_test/assemblies/JBE22000247.fasta -r /blue/bphl-florida/thsalikil
akshmi/data/jbi/20220825_jax_220629_PLN_WLK_MS/assemblies/JBI22000647.fasta -o fastani.out > r
esults_fastani
>>>>>>>>>>>>>>>>>>
Reference = [/blue/bphl-florida/thsalikilakshmi/data/jbi/20220825_jax_220629_PLN_WLK_MS/assemb
lies/JBI22000647.fasta]
Query = [/blue/bphl-florida/thsalikilakshmi/data/HAI/20220727_jax_220708_PLN_WLK_MS_test/assem
blies/JBE22000247.fasta]
Kmer size = 16
Fragment length = 3000
Threads = 1
ANI output file = fastani.out
>>>>>>>>>>>>>>>>>>
INFO [thread 0], skch::Sketch::build, minimizers picked from reference = 346824
INFO [thread 0], skch::Sketch::index, unique minimizers = 343053
INFO [thread 0], skch::Sketch::computeFreqHist, Frequency histogram of minimizers = (1, 340077
) ... (42, 1)
INFO [thread 0], skch::Sketch::computeFreqHist, With threshold 0.001%, ignore minimizers occur
ring >= 29 times during lookup.
INFO [thread 0], skch::main, Time spent sketching the reference : 0.317155 sec
INFO [thread 0], skch::main, Time spent mapping fragments in query #1 : 0.81592 sec
INFO [thread 0], skch::main, Time spent post mapping : 0.000211501 sec
```

Advanced Molecular Detection
Southeast Region Bioinformatics

Result

- Below output implies that the ANI estimate between *E. coli* and *S. flexneri* genomes is 97.4216
- Out of the total 1595 sequence fragments from *E. coli*, 1154 were aligned as orthologous matches

JBI22000647.fasta	97.4216	1154	1595

Advanced Molecular Detection
Southeast Region Bioinformatics

GAMMA

- Gene Allele Mutation Microbial Assessment
- GAMMA is a command line tool that finds gene matches in microbial genomic data using protein coding (rather than nucleotide) identity, and 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.)
- GAMMA is helpful in both identifying and explaining how unique alleles differ from their closest known matches

rastanton/GAMMA: Gene Allele Mutation Microbial Assessment (github.com)

Advanced Molecular Detection
Southeast Region Bioinformatics

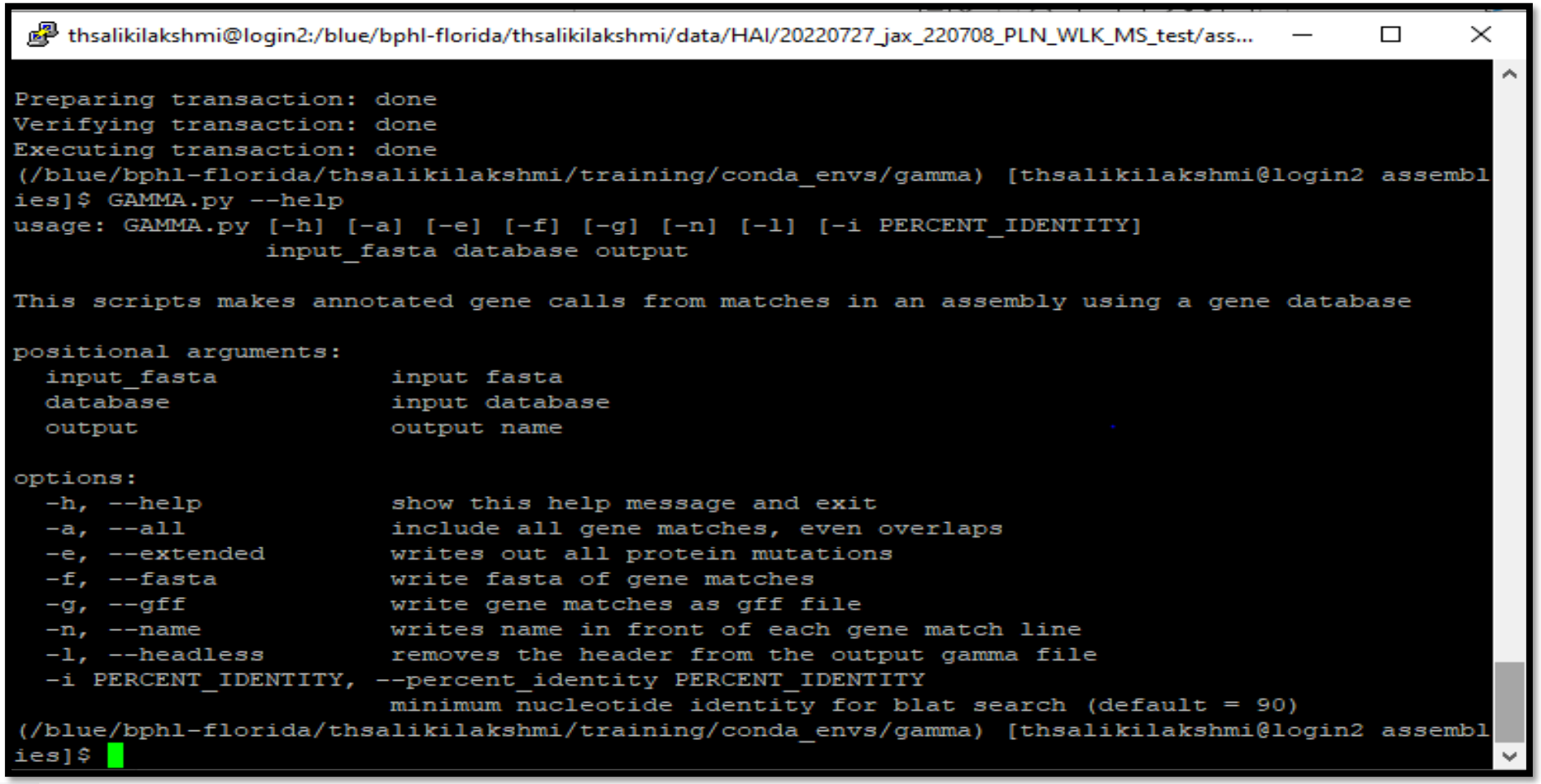
Installation

Can be installed through conda

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

Advanced Molecular Detection
Southeast Region Bioinformatics

Usage



Advanced Molecular Detection
Southeast Region Bioinformatics

Input

GAMMA.py my_genome.fasta gene_db.fasta output_name [optional arguments]

- Input for GAMMA is a genome or assembly in .fasta format and a multifasta database of the coding sequences of genes
- GAMMA was tested using
  - AR gene databases from AMRFINDERPLUS
    - (Index of /pathogen/Antimicrobial_resistance/AMRFinderPlus/database (nih.gov))
  - ARG_ANNOT
    - (backup.mediterranee-infection.com/arkotheque/client/ihumed/_depot_arko/articles/2041/arg-annot-v4-aa-may2018_doc.fasta)
  - RESFINDER
    - (genomicepidemiology / resfinder_db — Bitbucket)

Advanced Molecular Detection
Southeast Region Bioinformatics

Input

The sample GAMMA input shown below was generated from running GAMMA on a *S. flexneri* using a combination of all the ResFinder databases

genomicepidemiology / resfinder_db — Bitbucket

```
(/blue/bphl-florida/thsalikilakshmi/training/conda_envs/gamma) [thsalikilakshmi@login2 assembl
ies]$ GAMMA.py /blue/bphl-florida/thsalikilakshmi/data/jbi/20220825_jax_220629_PLN_WLK_MS/asse
mblies/JBI22000647.fasta /blue/bphl-florida/thsalikilakshmi/data/HAI/20220727_jax_220708_PLN_W
LK_MS_test/assemblies/resfinder_database/resfinder_db/all.fsa output
```

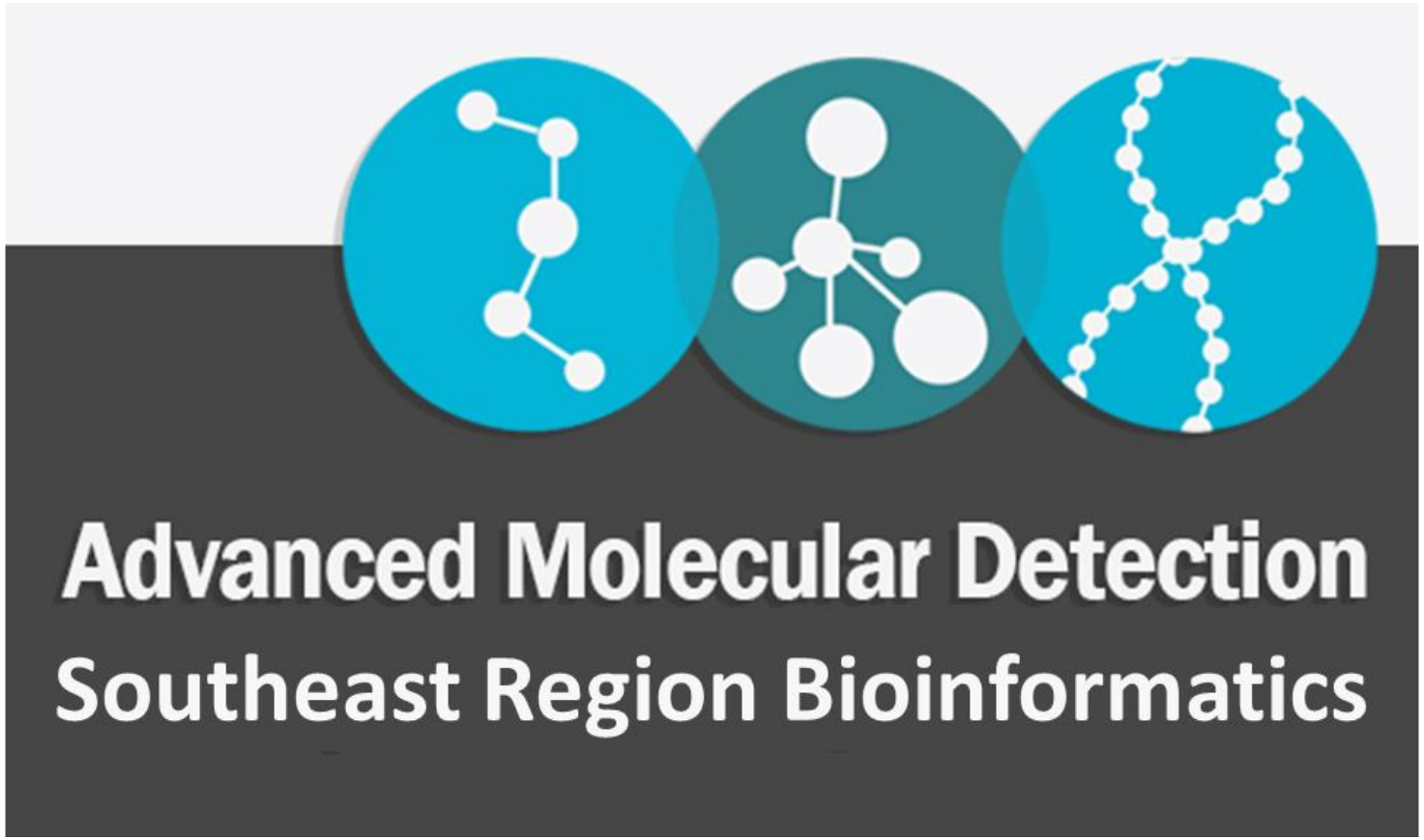
Advanced Molecular Detection
Southeast Region Bioinformatics

Results

The default output of GAMMA is a tab-delimited file with a .gamma extension with 15 columns

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Gene	Contig	Start	Stop	Match_Ty	Descriptic	Codon_Ch	BP_Chang	Transversi	Codon_Pe	BP_Percer	Percent_L	Match_Le	Target_Le	Strand
2	catA1_1_V	169	5200	5860	Native	No coding	0	1	0	1	0.9985	1	660	660	-
3	blaTEM-1E	191	74	935	Native	No coding	0	0	0	1	1	1	861	861	+
4	sul2_2_AY	191	3355	4171	Native	No coding	0	0	0	1	1	1	816	816	-
5	aph(6)-Id_	191	1655	2492	Native	No coding	0	0	0	1	1	1	837	837	-
6	aph(3'')-Ib	191	2491	3295	Native	No coding	0	0	0	1	1	1	804	804	-
7	tet(B)_2_A	205	807	2013	Native	No coding	0	0	0	1	1	1	1206	1206	-
8	mph(A)_1	214	2257	3163	Native	No coding	0	0	0	1	1	1	906	906	+
9	blaOXA-1_	244	984	1815	Native	No coding	0	0	0	1	1	1	831	831	-
10	aadA1_3_.	244	80	872	Mutant	A4V,	1	2	1	0.9962	0.9975	1	792	792	-
11	dfrA14_1_	294	445	919	Native	No coding	0	0	0	1	1	1	474	474	-

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