17 Feb 2025

TALBOT

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.

UPDATES

Office Hours

2025

March 03

Sarek_mic (Tentative)

OVERVIEW

Purpose

- Identification of clusters of closely related bacterial isolates for pan genome analyses which includes phylogenetic tree construction, SNP detection, core and accessory genome analysis.

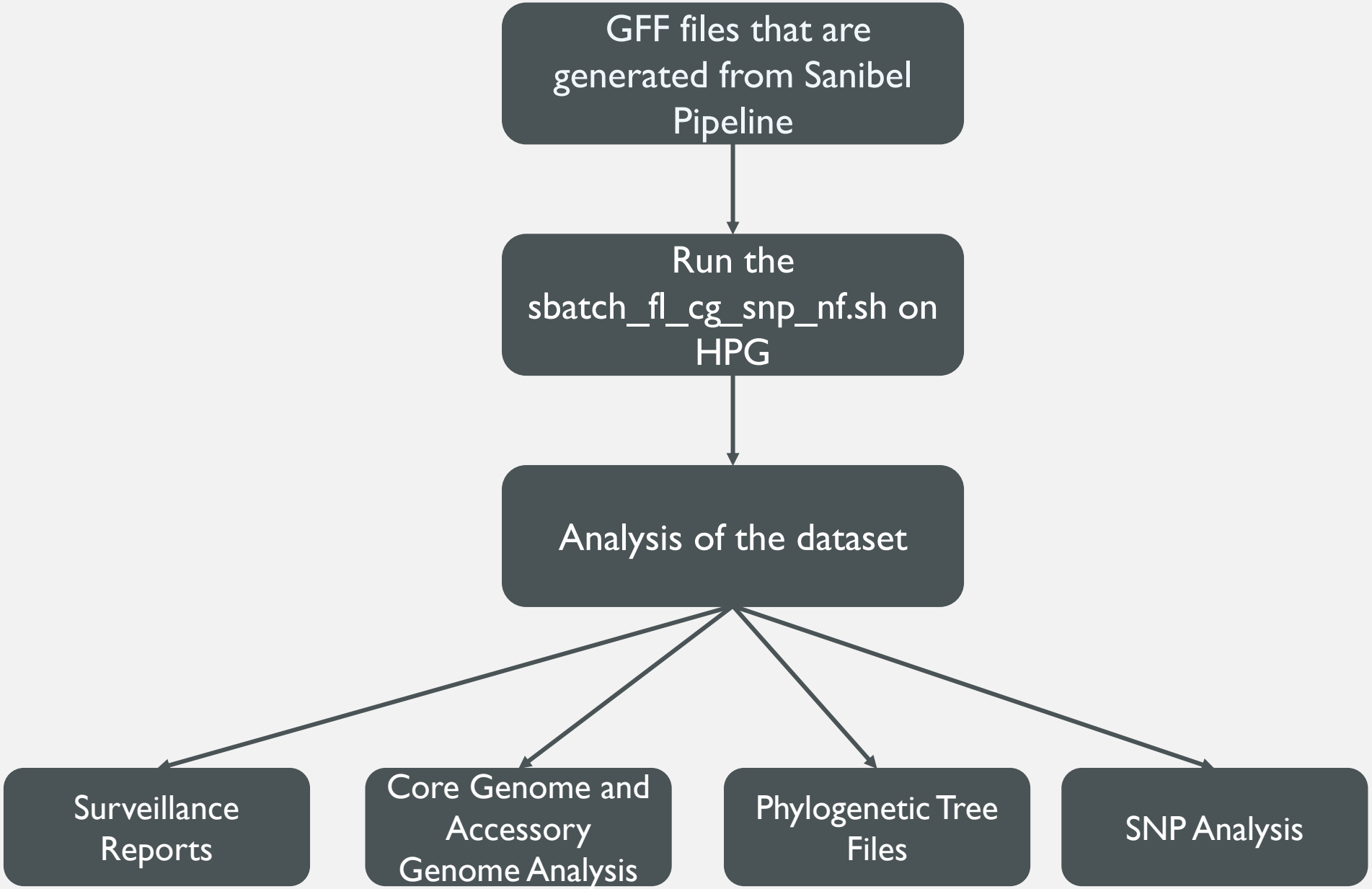
Usage

- To support public health and researchers by providing detailed reports and analyses that help identify outbreak clusters.

Dependencies

- Nextflow
- Singularity/Apptainer
- SLURM
- Python3

WORKFLOW


GFF files that are generated from Sanibel Pipeline
Run the sbatch_fl_cg_snp_nf.sh on HPG
Analysis of the dataset
Surveillance Reports
Core Genome and Accessory Genome Analysis
Phylogenetic Tree Files
SNP Analysis

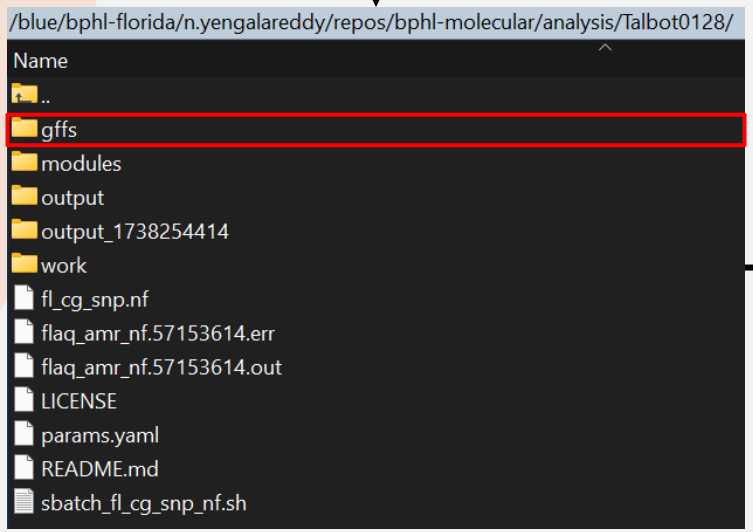
APPLICATION

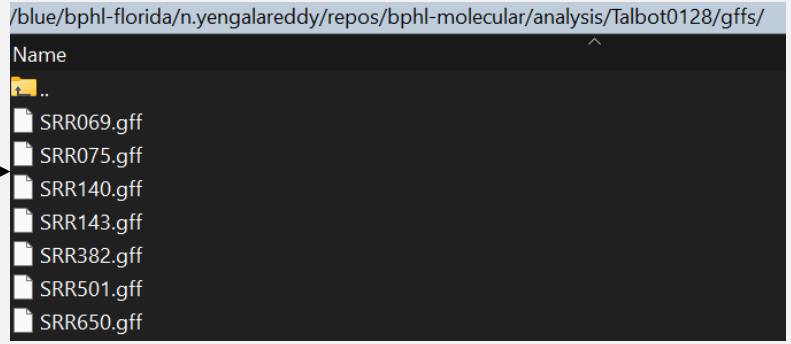
OBJECTIVE

USE GENOMIC ANNOTATION DATASET AND ANALYZE USING TALBOT PIPELINE

APPLICATION

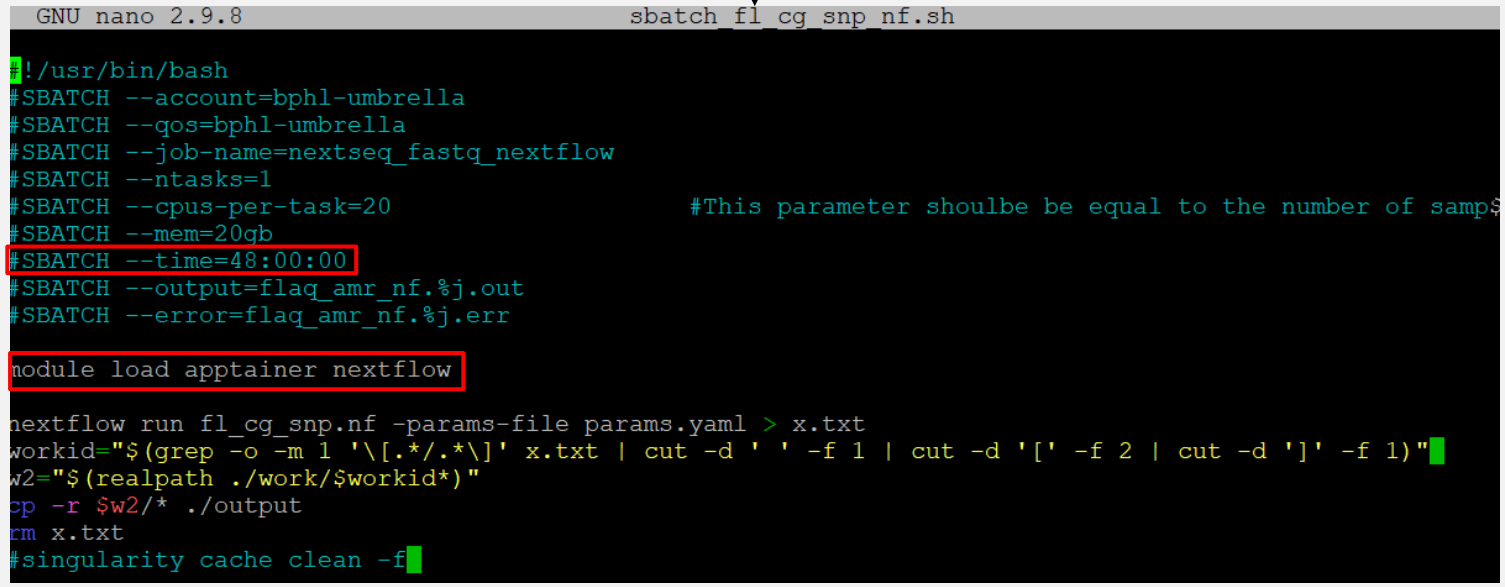
```
cd /blue/bphl-<state>/<user>/repos/bphl-molecular/

git clone https://github.com/BPHL-Molecular/Talbot

mkdir analysis/

cd analysis/

cp /blue/bphl-<state>/<user>/repos/bphl-molecular/ Talbot/

copy gff files from Sanibel Pipeline
```





APPLICATION

nano sbatch_fl_cg_snp_nf.sh

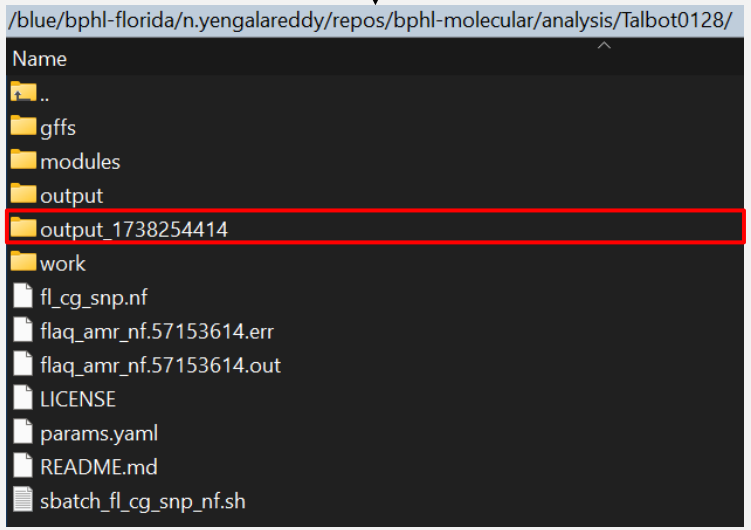


APPLICATION

Activate Conda Env containing Nextflow

&

sbatch sbatch_fl_cg_snp_nf.sh



APPLICATION

/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Talbot0128/output_1738254414/

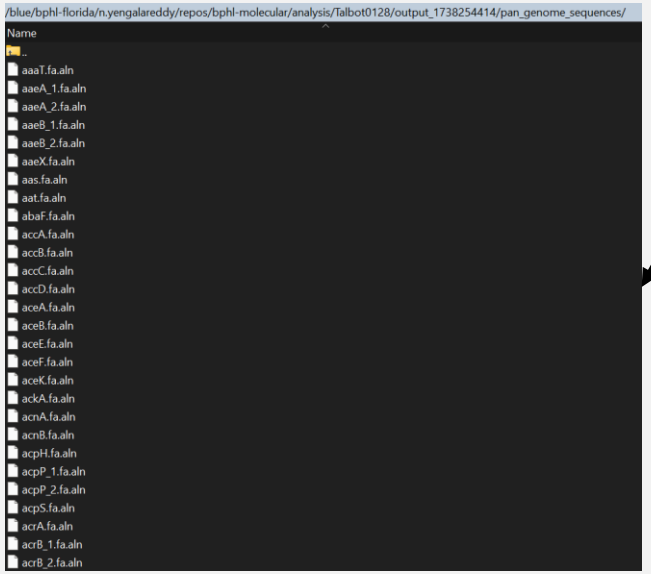
Name
..
fixed_input_files
pan_genome_sequences
_accessory_clusters
_accessory_clusters.clstr
_blast_results
_clustered
_clustered.clstr
_combined_files
_combined_files.groups
_inflated_mcl_groups
_inflated_unsplit_mcl_groups
_labeled_mcl_groups
_uninflated_mcl_groups
accessory.header.embl
accessory.tab
accessory_binary_genes.fa
accessory_binary_genes.fa.newick
accessory_graph.dot
blast_identity_frequency.Rtab
clustered_proteins
core_accessory.header.embl
core_accessory.tab
core_accessory_graph.dot
core_alignment_header.embl
core_gene_alignment.aln
gene_presence_absence.csv
gene_presence_absence.Rtab
number_of_conserved_genes.Rtab
number_of_genes_in_pan_genome.Rtab
number_of_new_genes.Rtab
number_of_unique_genes.Rtab
pan_genome_reference.fa
summary_statistics.txt

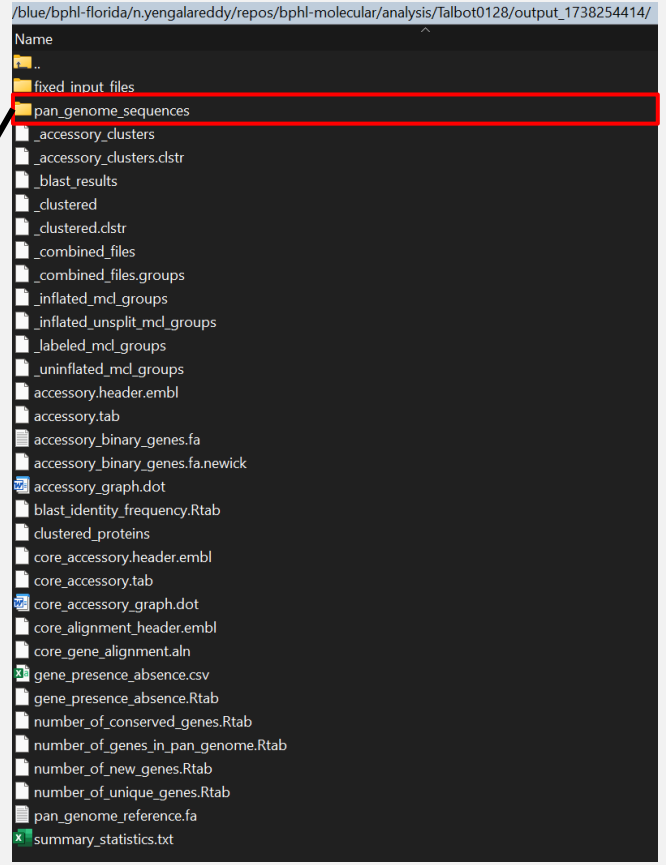
/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Talbot0128/

Name
..
gffs
modules
output
output_1738254414
work
fl_cg_snp.nf
flaq_amr_nf.57153614.err
flaq_amr_nf.57153614.out
LICENSE
params.yaml
README.md
sbatch_fl_cg_snp_nf.sh

Core genes	(99% <= strains <= 100%)	4344
Soft core genes	(95% <= strains < 99%)	0
Shell genes	(15% <= strains < 95%)	152
Cloud genes	(0% <= strains < 15%)	15
Total genes	(0% <= strains <= 100%)	4511

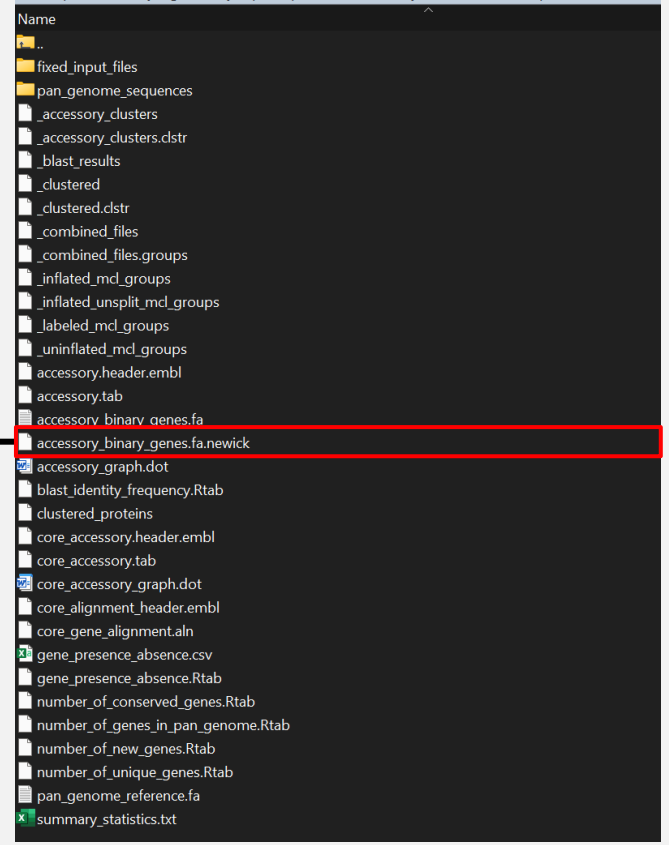
APPLICATION





APPLICATION

/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Talbot0128/output_1738254414/




```
(SRR069:0.063160696,(SRR501:0.039951332,SRR650:0.137826823)0.900:0.072812229,((SRR075:0.584619146,(SRR140:0.000000005,SRR382:1.16482
1034)0.485:0.011978239)0.997:0.185790771,SRR143:0.006109644)0.035:0.045373049);
```

APPLICATION

Tools To Analyze Newick File

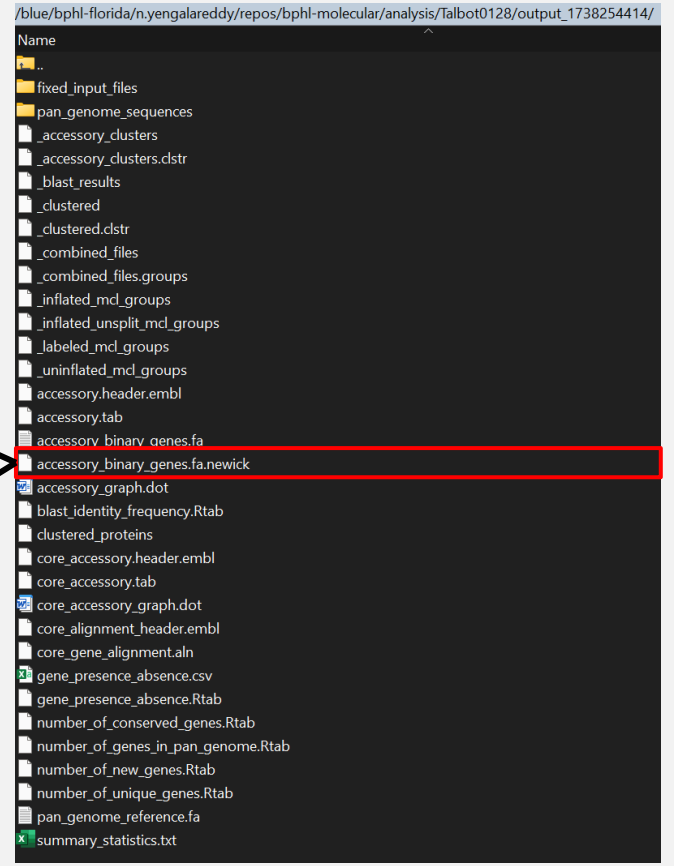
R

Mircoreact

IToL

NCBI Tree Viewer

Phylo.io



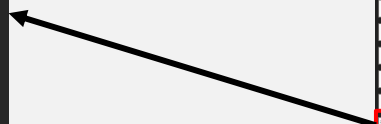
Office Hours (Tutorials)

20230612_Session8_Outbreaks

20240930_Session37_R Series Part 7

APPLICATION

```
>Salmonella_00001 citG_1
ATGAATGAAACCATAGCATTTCGTCCGGAAGGCCGGGCATCGATAAAAGATGTGGTTGTC
AATTATACGCTGGCATCACTGGACGAGCTTGAACAACTGCACCCGGCGCTGAATCTGCTG
GCGCACCCGTCAATTGTCCGTGGCAGCCATGATTTCTTTGTCTGTTGTTATAACGCGGGC
GTACAGCATGCGCAGGCCGGGGCGGCTGAAACGGTAGCGGTTTTGCCTGATTCACTGCCA
CAAGAACGGCAAAGGAGCGGGCTGCTGTGTGCCGCAGCAGGTCGTCTGGATGCCCTGCGC
CAGCCGCTGACGCACAATCGTTTATGTGACCTGGCCAGTCAGTTCTGTGCTGGCATGGCA
GATGTTGACAGCGAAACCCGCAGCGGTTTTTATACCGTGCGCAGCATATCGCTGCCGGTT
TACCGGCGGCTGTTGCGGGATCAACATTCTCACGGTGTCTGCCTGCAGCAGGCTCTACTG
CACCTGCTGGCCTGGAAGAGCGATTCCCCCTGGGCGCGGCAGCAGGCGCAGCGGCTGCTC
TGGCAGGGCGGCGTGCTGGGCGATAAAGGTGAGTTTGCGCTTATGACGCTGGATGATGAA
CTGCGGGAGCGCCAGATTGAGTGGCCCGGGCTGTGGTCGCTGCTGGCCGTCACCGGGTTT
CTGGCGAAATTCCCTGCCGGACCGATATTTGCTGACTGA
>Salmonella_00002 hmuU
ATGACCTGTCAGGACAAATCACTTTTAGCCCCCGCTCTCGCTTCCACGCCCGTCATTCCG
CGCTATCGCCAGATCATACGCAAACGGGCTGTGCTCATGTTAGCTATTGCGTTGGCGATG
ATGACTTCACTGATGGTGGATGTAACCTGTGGGTCTTCTGGTTTACCGCTCAGTGCGCTA
TGGCAGGCGCTTTTTCAACCGGAAAAGGTGAATGCGGGCATACACGTTATTGTCTGGGAT
ATTCGTTTACCCTATGCCCTGATGGCCCTACTGGTGGGCATGGCGTTAGGGTTAGCGGGC
GCCGAAATGCAAACCATTTTAAATAATCCGCTGGCGACGCCGTTTACCCTCGGCGTATCG
TCCGCGGCGGCTTTTGGCGCTGCGCTGGCGATTGTGTTGGGGATCGGTATTCCCGGTATT
CCTGCCGCCTGGTTTATCCCGGCCAATGCGTTCATCTTTGCGCTTTTATCCGCGTTACTC
CTGGACGGTATTACCCGTTGGACAGGCGTCGCCGCATCCGGAGTGGTGTTGTTCGGTATT
GCACTGGTATTCACCTTTAACGCGTTGGTCGCCATTATGCAGTTTGTGGCTGATGAAGAT
ACGCTGCAGGGGCTGGTATTCTGGACGATGGGGAGCCTGGTCCGCGCGTCGTGGGAAAAA
TTGGGCGTACTGGCCGTGGTGTTTATTGCTGTGCTTTTTTGCGCGTTACGAAGCGCCTGG
CAACTCACGGCATTACGTCTGGGCGAAGAGAGGGCGATGAGCTTTGGGATTCATGTCCGC
CGTTTACGCCTGCTGTCATTGTTGCGTATCAGCTTGTTATCAGCGCTGGCGGTAGCTTTC
GTCGGGCCGATTGGTTTTATTGGTCTCGTTGCACCGCACATTGCACGAATACTTTTGGGG
GAAGATCACCGTTTCTACCTTCCGGGCAGCGTGCTGATAGGCGGGCTGGTGTTATCGCTG
GCCTCTATTGCGGCGAAAAATAGTATTCCGGGCGTCATGGTGCCGGTGGGGATTGTCACC
TCGCTGGTCGGGGTTCCGTTCTTTTTAAGCATTGTATTGCGACATCGGGGGAGTCTCTGA
>Salmonella_00003 yusV
ATGCGCGGGTTAATGCTCCGCTCATTTAGCGCCGGCTATTCCACTCAGCCGGTTATTGCA
GATCTCAACGTTCCCTTACTGCCGTGCGGGAAAATTACCATTTTGCTGGGGCCGAACGGC
TGTGGTAAATCGACGTTACTACGTTCGCTGGCCGGGCTGAATAACGCCGACGGCGAGGCG
CTGCTTGATGGCGAAGATCTCATGTCGTTATCGTTTGCCGAGCGGGCGCAGAAAGTTGTG
TTTTTGCCGCAGTCATTACCGCAGGGCGTACATCTGCATGTACTGGAGTCGATTATCGTG
GTACTGCGGGCATCAGGCGGCAGAGACAATGCGCAGGGCAGAGCGCAGATACTGGCTATC
CTTGAGCAATTGGGCATAACGCATCTGGCGTTGCAGTATCTCGACCAGCTTTCCGGCGGT
CAGCGGCAGTTAGTGGGTCTGGCGCAATCGCTGATTCGTCGCCCTGAACTTCTGTTGCTG
GATGAGCCGCTCAGCGCGCTGGATTTAAACTATCAGTTTCACGTAATGGATCTCGTGCGT
CGGGATACCCAGGCGCAAAATAGGGTAACGATAGTCGTGGCGCATGACATTAATATTGCA
CTGCGTCATGGTGACCATGTCCTGATGCTAAAAGACGGCAGGCTGGTCGCCAGCGGCGCG
CCGGAGACGGTGATTACTGCGGAGCGTCTGGCTGAAGTGTACCGGGTACGAGGACGCGTG
GAGCGCTGTTCGCAGGGGAAATTGCAGGTGGTGCTGGACGGCGTGATCGCTGTTTAA
>Salmonella_00004 gpmA
ATGGCTGTAACTAAGCTGGTTCTGGTACGTCACGGTGAAAGTCAATGGAACAAAGAGAAC
CGTTTCACCGGCTGGTATGATGTTGACCTGTCTGAAAAAGGCGTAAGCGAAGCAAAAGCA
GCGGGTAAACTGCTAAAAGAAGAAGGTTTCAGCTTTGATTTTGCCTACACTTCCGTACTG
AAACGCGCTATCCATACGCTGTGGAACGTACTGGATGAACTGGATCAAGCGTGGCTGCCG
GTGGAAAAATCCTGGAAACTCAATGAACGTCACTATGGCGCGTTGCAGGGGCTGAATAAA
GCGGAAACGGCAGAAAAATATGGTGACGAGCAGGTTAAACAGTGGCGCCGCGGTTTTGCC
GTGACGCCGCCGGAACTGACCAAAGATGATGAGCGTTATCCGGGTCACGATCCACGTTAT
```



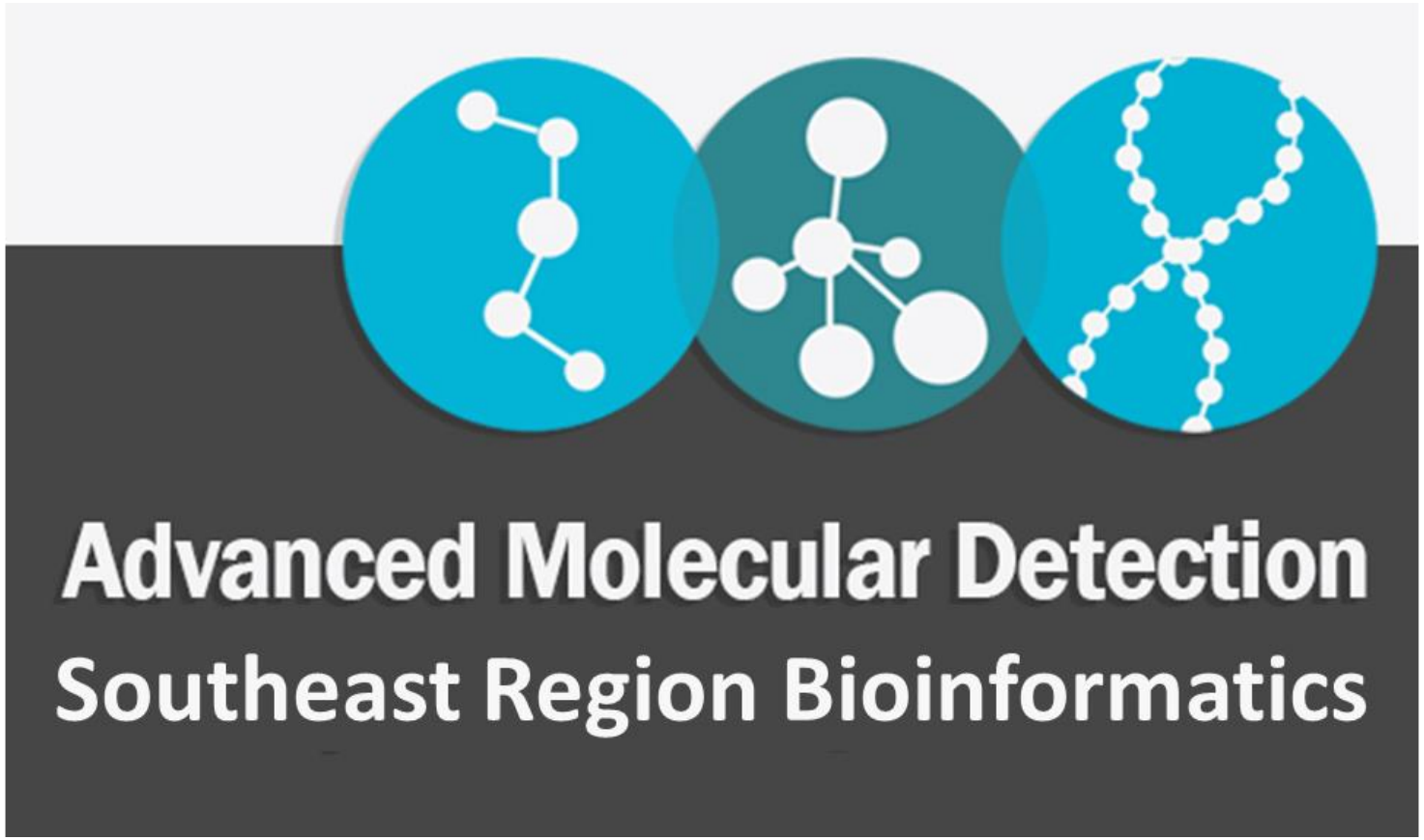
/blue/bphl-florida/n.yengalareddy/repos/bphl-molecular/analysis/Talbot0128/output_1738254414/

Name

- ..
- fixed_input_files
- pan_genome_sequences
- _accessory_clusters
- _accessory_clusters.clstr
- _blast_results
- _clustered
- _clustered.clstr
- _combined_files
- _combined_files.groups
- _inflated_mcl_groups
- _inflated_unsplit_mcl_groups
- _labeled_mcl_groups
- _uninflated_mcl_groups
- accessory.header.embl
- accessory.tab
- accessory_binary_genes.fa
- accessory_binary_genes.fa.newick
- accessory_graph.dot
- blast_identity_frequency.Rtab
- clustered_proteins
- core_accessory.header.embl
- core_accessory.tab
- core_accessory_graph.dot
- core_alignment_header.embl
- core_gene_alignment.aln
- gene_presence_absence.csv
- gene_presence_absence.Rtab
- number_of_conserved_genes.Rtab
- number_of_genes_in_pan_genome.Rtab
- number_of_new_genes.Rtab
- number_of_unique_genes.Rtab
- pan_genome_reference.fa
- summary_statistics.txt

CONCLUSION

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



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