Advanced Molecular Detection
Southeast Region Bioinformatics

Bactopia tools
07/24/2023

Outline

- Updates
- Agenda
- Bactopia
- AMRFinder Plus
- BUSCO
- CheckM

Updates - CDC's PHoeNIx

- The new v2.0.1 of PHoeNIx has been released, full details of changes are documented in the CHANGELOG.md file, but here are the highlights
  - The pipeline now **ONLY** uses Ben Langmead's public kraken databases. You **MUST** use a database build on or after 3/14/2023. We use the standard-8 version of the database. You **CANNOT** use the old database downloaded from CDC's sharefile (the one we have available in the share-drive. The pipeline is not backwards compatible (hence the bump to version 2).
  - Additions of SRA and SCAFFOLDS entry points, as well as their CDC versions (CDC_SRA and CDC_SCAFFOLDS) see wiki documentation for how to run and the new workflow.
  - Custom MLST database – there is now a static database that is build from a direct pull of PubMLST.org. Details on how it is made are found here.

Advanced Molecular Detection
Southeast Region Bioinformatics

Updates – CDC's PHoeNIx cont.

- The GRiPHin report (an excel sheet with a complete run summary) now is an output of all entry points.
- A parameter to increase coverage >30x (default) is now available using --coverage.
- For those using PHoeNIx on Terra more options are available to allow you to run different entry points. Please see Terra instructions in the wiki. They will be holding a new training for this version, and we will forward the email to the region.
- The main branch will run the latest version of the pipeline, which is now v2.0.0. If you want to run a different branch just select it using the `-r` parameter (ex. `nextflow run cdcgov/phoenix -r v1.1.1 -entry PHOENIX -profile singularity,test –kraken2db $PATH_TO_DB`). They recommend using the latest version as it includes several bug fixes. For Terra users the drop-down menu will have the versions that are available to choose from.

Advanced Molecular Detection
Southeast Region Bioinformatics

Updates – Staff

- Dr. Sarah Schmedes has taken a new position at the CDC
- We are still your BRR resource, and we are still available at bphl-sebioinformatics@flhealth.gov

Advanced Molecular Detection
Southeast Region Bioinformatics

Agenda

August 7 – Bactopia Tools: ECTyper, Emmtyper
August 21 – Bactopia Tools: FastANI, 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

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

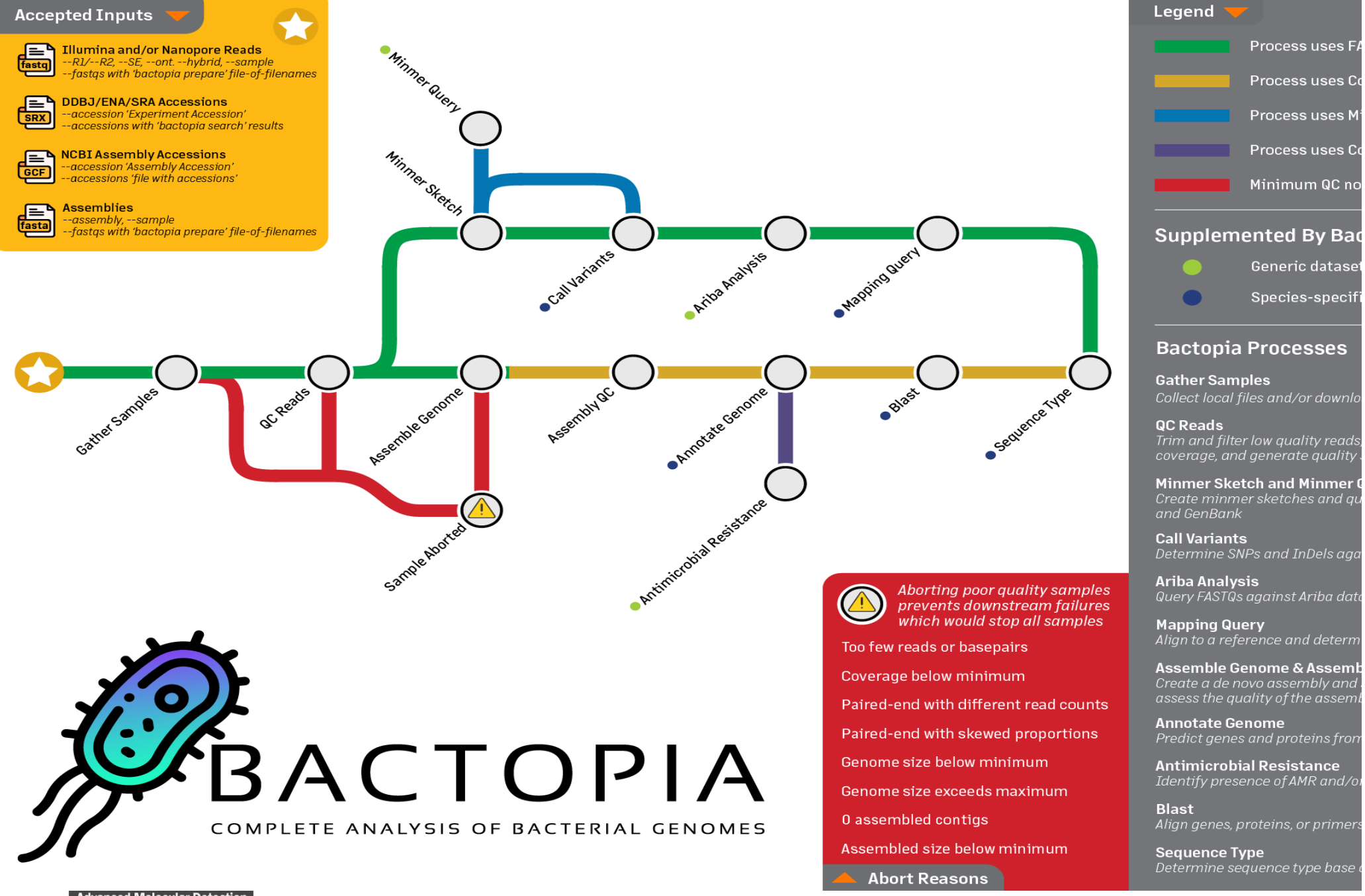
Advanced Molecular Detection
Southeast Region Bioinformatics

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


Accepted Inputs
Illumina and/or Nanopore Reads
--R1/--R2, --SE, --ont. --hybrid, --sample
--fastqs with 'bactopia prepare' file-of-filenames
DDBJ/ENA/SRA Accessions
--accession 'Experiment Accession'
--accessions with 'bactopia search' results
NCBI Assembly Accessions
--accession 'Assembly Accession'
--accessions 'file with accessions'
Assemblies
--assembly, --sample
--fastqs with 'bactopia prepare' file-of-filenames
Minmer Query
Minmer Sketch
Call Variants
Ariba Analysis
Mapping Query
Gather Samples
QC Reads
Assemble Genome
Assembly QC
Annotate Genome
Blast
Sequence Type
Sample Aborted
Antimicrobial Resistance
Aborting poor quality samples prevents downstream failures which would stop all samples
Too few reads or basepairs
Coverage below minimum
Paired-end with different read counts
Paired-end with skewed proportions
Genome size below minimum
Genome size exceeds maximum
0 assembled contigs
Assembled size below minimum
Abort Reasons
Legend
Supplemented By Bac
Generic dataset
Species-specifi
Bactopia Processes
Gather Samples
Collect local files and/or downlo
QC Reads
Trim and filter low quality reads,
coverage, and generate quality
Minmer Sketch and Minmer
Create minmer sketches and qu
and GenBank
Call Variants
Determine SNPs and InDels aga
Ariba Analysis
Query FASTQs against Ariba dat
Mapping Query
Align to a reference and determ
Assemble Genome & Assemb
Create a de novo assembly and
assess the quality of the assem
Annotate Genome
Predict genes and proteins from
Antimicrobial Resistance
Identify presence of AMR and/o
Blast
Align genes, proteins, or primers
Sequence Type
Determine sequence type base
BACTOPIA
COMPLETE ANALYSIS OF BACTERIAL GENOMES
Advanced Molecular Detection
Southeast Region Bioinformatics

Bactopia Tools

Advanced Molecular Detection
Southeast Region Bioinformatics

AMRFinderPlus

- NCBI Antimicrobial Resistance Gene Finder Plus
- AMRFinderPlus - Identify AMR genes, point mutations, virulence, and stress resistance genes in assembled bacterial nucleotide and protein sequence
- [Home · ncbi/amr Wiki (github.com)]

Advanced Molecular Detection
Southeast Region Bioinformatics

Mechanism

- AMRFinder has two modes with either protein sequences or with DNA sequences as input
- With protein sequences AMRFinderPlus uses both BLASTP and HMMER to search for AMR genes along with a hierarchical tree of gene families to classify and name novel sequences
- With nucleotide sequences AMRFinderPlus uses BLASTX translated searches and the hierarchical tree of gene families

Advanced Molecular Detection
Southeast Region Bioinformatics

Installation

- Available as a module on HPG

```
module load amrfinderplus/3.10.18
```

- Can be installed through conda

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

Advanced Molecular Detection
Southeast Region Bioinformatics

Usage

thsalikilakshmi@login1:/blue/bphl-florida/thsalikilakshmi/data/HAI/20220727_jax_220708_PLN_WLK_MS_test/assemblies

```
[thsalikilakshmi@login1 assemblies]$ amrfinder -h
Identify AMR and virulence genes in proteins and/or contigs and print a report

DOCUMENTATION
    See https://github.com/ncbi/amr/wiki for full documentation

UPDATES
    Subscribe to the amrfinder-announce mailing list for database and software update notifications:
    https://www.ncbi.nlm.nih.gov/mailman/listinfo/amrfinder-announce

USAGE:   amrfinder [--update] [--force_update] [--protein PROT_FASTA] [--nucleotide NUC_FASTA] [--gff GFF_FILE] [--pg
erage_min MIN_COV] [--organism ORGANISM] [--list_organisms] [--translation_table TRANSLATION_TABLE] [--plus] [--repor
IR] [--report_all_equal] [--name NAME] [--output OUTPUT_FILE] [--protein_output PROT_FASTA_OUT] [--nucleotide_output
OUT] [--nucleotide_flank5_size NUC_FLANK5_SIZE] [--quiet] [--gpipe_org] [--parm PARM] [--threads THREADS] [--debug] [
HELP:    amrfinder --help or amrfinder -h
VERSION: amrfinder --version

NAMED PARAMETERS:
-u, --update
    Update the AMRFinder database
-U, --force_update
    Force updating the AMRFinder database
-p PROT_FASTA, --protein PROT_FASTA
    Input protein FASTA file
```

Advanced Molecular Detection
Southeast Region Bioinformatics

Input File Formats

- Only required arguments are either --p <protein fasta> for proteins or --n <nucleotide fasta> for nucleotides
- --g <gff_file>
  - GFF files are used to get sequence coordinates for AMRFinder hits from protein sequence

```
amrfinder -n JBE22000638.fasta -O Escherichia
```

Advanced Molecular Detection
Southeast Region Bioinformatics

Output

1	Protein id	Contig id	Start	Stop	Strand	Gene sym	Sequence name	Scope	Element t	Element s	Class	Subclass	Method	Target len	Reference	% Coverag	% Identity	Alignmen
2	NA	10	36763	38118	-	glpT_E448	Escherichia fosfomycin resistant Glp	core	AMR	POINT	FOSFOMY	FOSFOMY	POINTX	452	452	100	99.56	452
3	NA	113	5151	6239	-	pmrB_Y35	Escherichia colistin resistant PmrB	core	AMR	POINT	COLISTIN	COLISTIN	POINTX	363	363	100	99.72	363
4	NA	174	1054	2250	+	tet(A)	tetracycline efflux MFS transporter T	core	AMR	AMR	TETRACYC	TETRACYC	EXACTX	399	399	100	100	399
5	NA	22	964	1605	-	qnrB19	quinolone resistance pentapeptide	core	AMR	AMR	QUINOLO	QUINOLO	ALLELEX	214	214	100	100	214
6	NA	22	57696	58508	+	sul2	sulfonamide-resistant dihydroptero	core	AMR	AMR	SULFONA	SULFONA	EXACTX	271	271	100	100	271
7	NA	22	58548	59372	+	aph(3'')-Ib	aminoglycoside O-phosphotransfera	core	AMR	AMR	AMINOGL	STREPTON	EXACTX	275	275	100	100	275
8	NA	22	59375	60208	+	aph(6)-Id	aminoglycoside O-phosphotransfera	core	AMR	AMR	AMINOGL	STREPTON	EXACTX	278	278	100	100	278

Advanced Molecular Detection
Southeast Region Bioinformatics

BUSCO

- Benchmarking Universal Single-Copy Orthologs (BUSCO)
- BUSCO attempts to provide a quantitative assessment of the completeness in terms of expected gene content of a genome assembly, transcriptome, or annotated gene set
- BUSCO employs clade-specific information to identify BUSCO genes in the analyzed sequence
- [ezlab / busco · GitLab]

Advanced Molecular Detection
Southeast Region Bioinformatics

Installation

- Available as a module on HPG

```
module load busco/5.3.0
```

- Can be installed through conda

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

Advanced Molecular Detection
Southeast Region Bioinformatics

Usage

thsalikilakshmi@login1:/blue/bphl-florida/thsalikilakshmi/data/HAI/20220727_jax_220708_PLN_WLK_MS_test

```
[thsalikilakshmi@login1 20220727_jax_220708_PLN_WLK_MS_test]$ module load busco/5.3.0
[thsalikilakshmi@login1 20220727_jax_220708_PLN_WLK_MS_test]$ busco -h
usage: busco -i [SEQUENCE_FILE] -l [LINEAGE] -o [OUTPUT_NAME] -m [MODE] [OTHER OPTIONS]

Welcome to BUSCO 5.3.0: the Benchmarking Universal Single-Copy Ortholog assessment tool.
For more detailed usage information, please review the README file provided with this distribution
 and the BUSCO user guide. Visit this page https://gitlab.com/ezlab/busco#how-to-cite-busco to see
 how to cite BUSCO

optional arguments:
  -i SEQUENCE_FILE, --in SEQUENCE_FILE
                        Input sequence file in FASTA format. Can be an assembled genome or transcr
iptome (DNA), or protein sequences from an annotated gene set. Also possible to use a path to a di
rectory containing multiple input files.
  -o OUTPUT, --out OUTPUT
                        Give your analysis run a recognisable short name. Output folders and files
 will be labelled with this name. The path to the output folder is set with --out_path.
  -m MODE, --mode MODE  Specify which BUSCO analysis mode to run.
                        There are three valid modes:
                        - geno or genome, for genome assemblies (DNA)
                        - tran or transcriptome, for transcriptome assemblies (DNA)
                        - prot or proteins, for annotated gene sets (protein)
  -l LINEAGE, --lineage_dataset LINEAGE
                        Specify the name of the BUSCO lineage to be used.
  --augustus            Use augustus gene predictor for eukaryote runs
  --augustus_parameters --PARAM1=VALUE1,--PARAM2=VALUE2
                        Pass additional arguments to Augustus. All arguments should be contained w
ithin a single string with no white space, with each argument separated by a comma.
  --augustus_species AUGUSTUS_SPECIES
```

Advanced Molecular Detection
Southeast Region Bioinformatics

Running BUSCO

Mandatory arguments unless provided in config file:

```
$ busco -i [SEQUENCE_FILE] -l [LINEAGE] -o [OUTPUT_NAME] -m [MODE] [OTHER OPTIONS]
```

- -i or --in defines the input file to analyze which is either a nucleotide *.fasta* file or a protein *.fasta* file, depending on the BUSCO mode. In v5.1.0 the input argument can now also be a directory containing *.fasta* files to run in batch mode
- -o or --out defines the folder that will contain all results, logs, and intermediate data
- -m or --mode sets the assessment MODE: genome, proteins, transcriptome
- -l or --lineage_dataset

Advanced Molecular Detection
Southeast Region Bioinformatics

Results

```
$ busco -i JBE22000155.fasta -o results_busco -m genome
```

- Output directory will contain several files and directories

```
***** Results: *****

C:100.0%[S:99.8%,D:0.2%],F:0.0%,M:0.0%,n:440
440     Complete BUSCOs (C)
439     Complete and single-copy BUSCOs (S)
1       Complete and duplicated BUSCOs (D)
0       Fragmented BUSCOs (F)
0       Missing BUSCOs (M)
440     Total BUSCO groups searched
```

Advanced Molecular Detection
Southeast Region Bioinformatics

CheckM

- CheckM provides a set of tools for assessing the quality of genomes recovered from isolates, single cells, or metagenomes
- Provides robust estimates of genome completeness and contamination by using collocated sets of genes that are ubiquitous and a single-copy within a phylogenetic lineage
- CheckM works on a directory of genome bins in *.fasta* format
- Ecogenomics/CheckM: Assess the quality of microbial genomes recovered from isolates, single cells, and metagenomes (github.com)

Advanced Molecular Detection
Southeast Region Bioinformatics

Installation

- Available as a module on HPG

```
module load checkm/1.1.2
```

- Can be installed through conda

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

Advanced Molecular Detection
Southeast Region Bioinformatics

Usage

thsalikilakshmi@login1:/blue/bphl-florida/thsalikilakshmi/data/HAI/20220727_jax_220708_PLN_WLK_MS_test/assemblies

```
[thsalikilakshmi@c0700a-s4 assemblies]$ module load checkm/1.1.2

Lmod is automatically replacing "busco/5.3.0" with "checkm/1.1.2".

[thsalikilakshmi@c0700a-s4 assemblies]$ checkm -h

                ...:::: CheckM v1.1.2 ::::...

  Lineage-specific marker set:
    tree          -> Place bins in the reference genome tree
    tree_qa       -> Assess phylogenetic markers found in each bin
    lineage_set   -> Infer lineage-specific marker sets for each bin

  Taxonomic-specific marker set:
    taxon_list    -> List available taxonomic-specific marker sets
    taxon_set     -> Generate taxonomic-specific marker set

  Apply marker set to genome bins:
    analyze       -> Identify marker genes in bins
    qa            -> Assess bins for contamination and completeness

  Common workflows (combines above commands):
    lineage_wf    -> Runs tree, lineage_set, analyze, qa
    taxonomy_wf   -> Runs taxon_set, analyze, qa

  Reference distribution plots:
    gc_plot       -> Create GC histogram and delta-GC plot
    coding_plot   -> Create coding density (CD) histogram and delta-CD plot
    tetra_plot    -> Create tetranucleotide distance (TD) histogram and delta-TD plot
```

Advanced Molecular Detection
Southeast Region Bioinformatics

Workflow Overview

- Lineage-specific workflow - quality estimates with lineage-specific markers
- Taxonomic-specific Workflow - quality estimates with taxonomic-specific markers
- Using Custom Marker Genes - genome quality estimates with custom markers
- Using CPR Marker Set - genome quality estimates with a CPR/Patescibacteria specific marker set

Advanced Molecular Detection
Southeast Region Bioinformatics

Input

- Input format

```
$ checkm lineage_wf -x fasta -t 8 /path/to/assemblies/ /path/to/assemblies/results_checkm
```

Advanced Molecular Detection
Southeast Region Bioinformatics

Results

/blue/bphl-florida/thsalikilakshmi/data/HAI/20220727_jax_220708_PLN_WLK_MS_test/assemblies/results_checkm/storage/bin_stats.tree.tsv - hpg.rc.ufl.edu - Editor - WinS...

Encoding Color


```
JBE22000268     {'GC': 0.5084001847667083, 'GC std': 0.05498609800897876, 'Genome size': 5059353, '# ambiguous bases': 0, '# scaffolds
JBE22000247     {'GC': 0.5066224298266259, 'GC std': 0.056075576250482526, 'Genome size': 5192354, '# ambiguous bases': 0, '# scaffold
JBE22000179     {'GC': 0.5102446229468889, 'GC std': 0.05229435235179859, 'Genome size': 5213125, '# ambiguous bases': 0, '# scaffolds
JBE22000216     {'GC': 0.5072497654691881, 'GC std': 0.05397710640836604, 'Genome size': 5256239, '# ambiguous bases': 0, '# scaffolds
JBE22000212     {'GC': 0.5094524425168844, 'GC std': 0.049386645761058286, 'Genome size': 5019126, '# ambiguous bases': 0, '# scaffold
JBE22000155     {'GC': 0.510425510157976, 'GC std': 0.04955150255228129, 'Genome size': 5046084, '# ambiguous bases': 0, '# scaffolds'
JBE22000165     {'GC': 0.5075654701618173, 'GC std': 0.05434220287405592, 'Genome size': 5319828, '# ambiguous bases': 0, '# scaffolds
JBE22000210     {'GC': 0.5070784821060123, 'GC std': 0.04398818675703703, 'Genome size': 5205212, '# ambiguous bases': 0, '# scaffolds
JBE22000630     {'GC': 0.508020442012044, 'GC std': 0.0539952664110469, 'Genome size': 4928282, '# ambiguous bases': 0, '# scaffolds':
JBE22000543     {'GC': 0.5087174061329862, 'GC std': 0.04981502623621671, 'Genome size': 5205849, '# ambiguous bases': 0, '# scaffolds
JBE22000631     {'GC': 0.506871843992454, 'GC std': 0.06154814736224565, 'Genome size': 4931864, '# ambiguous bases': 0, '# scaffolds'
JBE22000624     {'GC': 0.5090352975794736, 'GC std': 0.05378448391488944, 'Genome size': 5056834, '# ambiguous bases': 0, '# scaffolds
JBE22000612     {'GC': 0.5047976558707805, 'GC std': 0.06350174595707664, 'Genome size': 5350217, '# ambiguous bases': 0, '# scaffolds
JBE22000616     {'GC': 0.5052699673984343, 'GC std': 0.06246647400530439, 'Genome size': 5303733, '# ambiguous bases': 0, '# scaffolds
JBE22000592     {'GC': 0.5093720234867496, 'GC std': 0.04774985442757208, 'Genome size': 5172789, '# ambiguous bases': 0, '# scaffolds
JBE22000638     {'GC': 0.50746618367025, 'GC std': 0.05804982009814915, 'Genome size': 5394657, '# ambiguous bases': 0, '# scaffolds':
```


Advanced Molecular Detection
Southeast Region Bioinformatics

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