Advanced Molecular Detection

Southeast Region Bioinformatics

Bactopia tools

08/07/2023

Outline

- Updates
- Agenda
- Bactopia
- ECTyper
- EmmTyper
- Questions

Updates – ABiL Trainings

Two upcoming trainings available to you

1. In-person training – tentatively scheduled for October 16-19 – thoughts?
    1. 4 days at Georgia Tech
    2. Intermediate to advanced bioinformatics – you'll need good fundamentals to attend.
2. Online Trainings – courses are still being developed; we will notify once they are available

Agenda

August 21 – Bactopia Tools: FastANI and GAMMA

September 4 rescheduled to September 11 – Bactopia Tools: Hicap and HpsuisSero

Future Trainings

- ONT & FL’s Flisochar pipeline
- StaPH-B Toolkit Programs/Pipelines
- GISAID flagged SARS-CoV-2
- R Training Series
- Dryad pipeline
- …and more

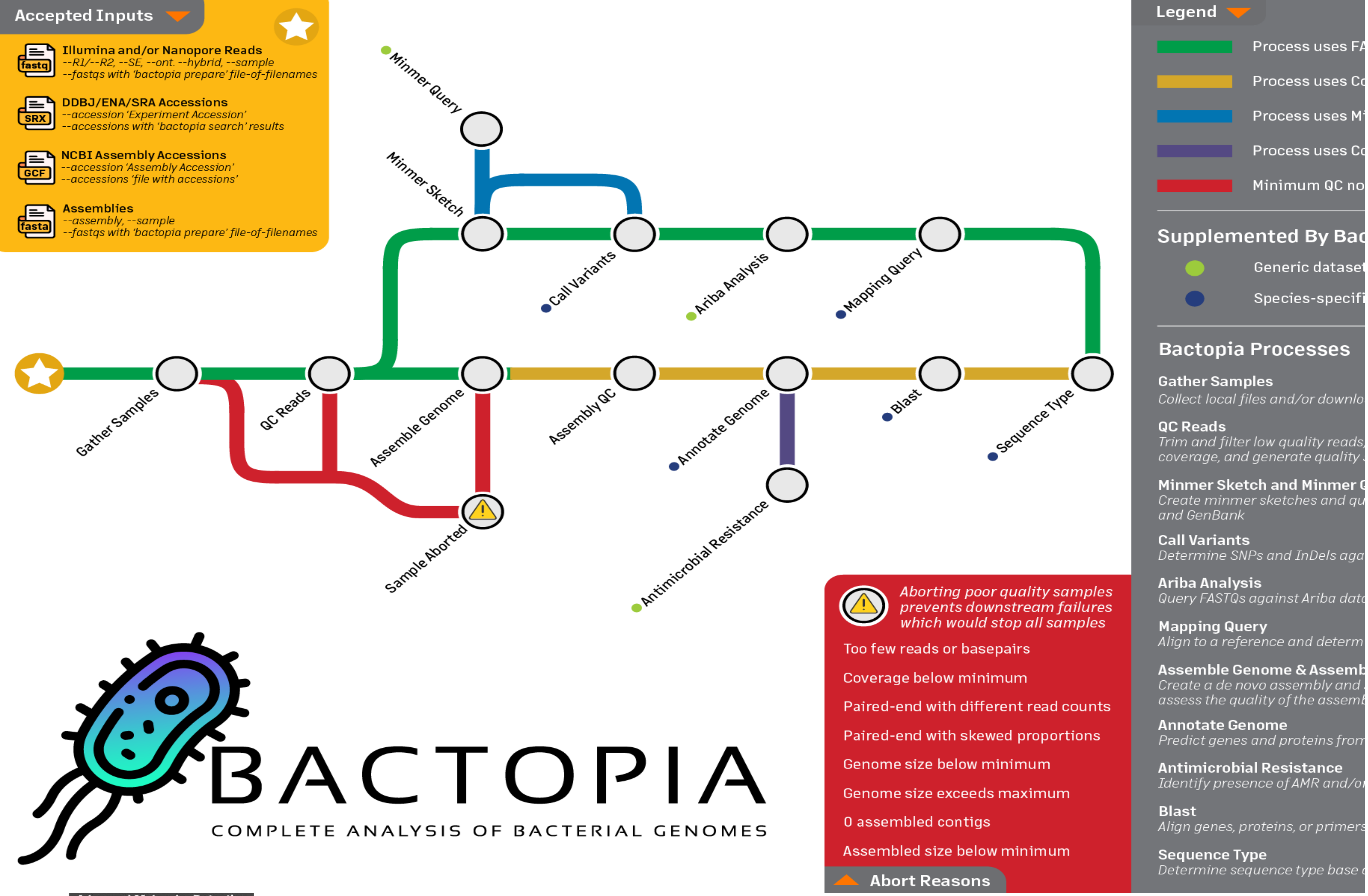
Bactopia

- 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

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

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

Legend

- 🟩 Process uses FA
- 🟨 Process uses Co
- 🟦 Process uses M
- 🟪 Process uses Co
- 🟥 Minimum QC no

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 data

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

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

Advanced Molecular Detection
Southeast Region Bioinformatics

Bactopia Tools

Advanced Molecular Detection
Southeast Region Bioinformatics

ECTyper (an Easy Typer)

- ECTyper is a standalone versatile serotyping module for *Escherichia coli*
- Supports both *.fasta* (assembled) and *.fastq* (raw reads) file formats
- This tool provides convenient species identification coupled with quality control module giving a complete, transparent, and reference laboratory suitable report on *E. coli* serotyping

phac-nml/ecoli_serotyping: In silico prediction of E. coli serotype (github.com)

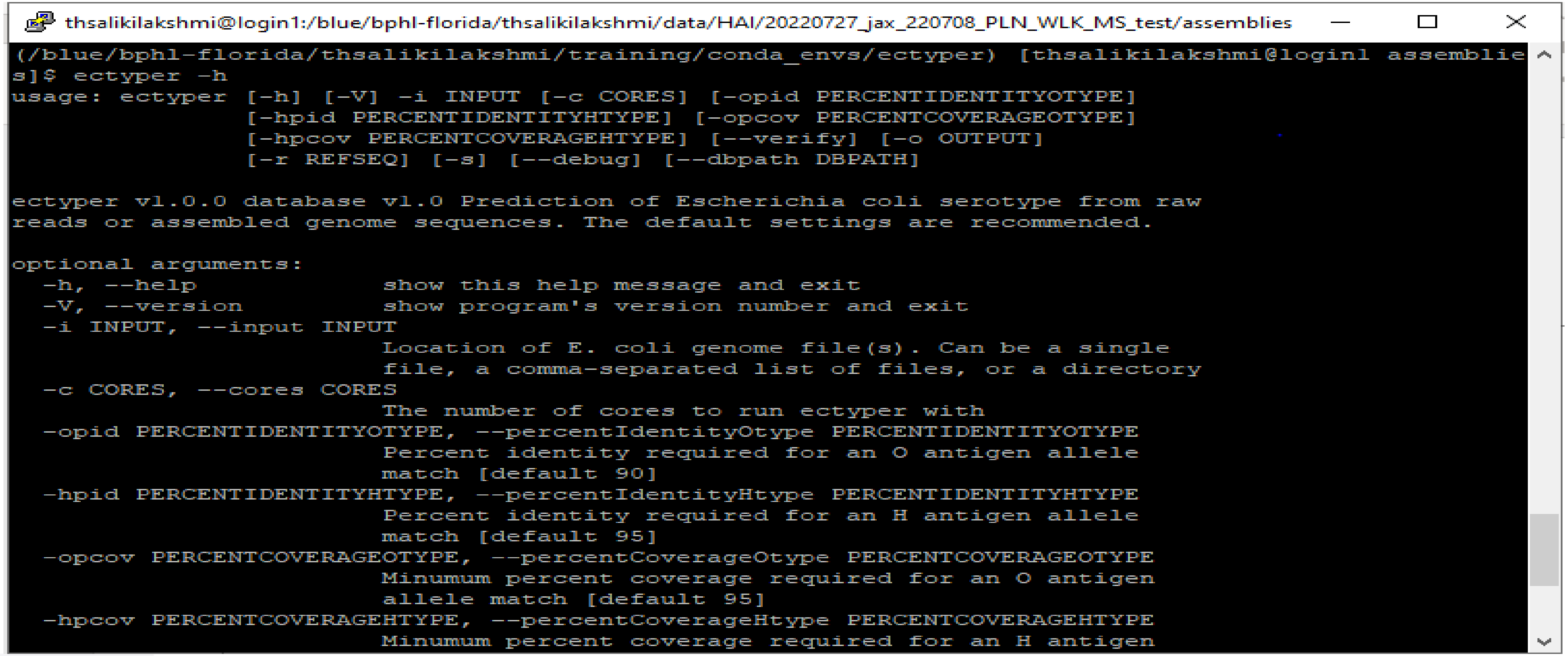
Installation

Can be installed through conda

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

Advanced Molecular Detection
Southeast Region Bioinformatics

Usage


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

```
(/blue/bphl-florida/thsalikilakshmi/training/conda_envs/ectyper) [thsalikilakshmi@login1 assemblie
s]$ ectyper -h
usage: ectyper [-h] [-V] -i INPUT [-c CORES] [-opid PERCENTIDENTITYOTYPE]
               [-hpid PERCENTIDENTITYHTYPE] [-opcov PERCENTCOVERAGEOTYPE]
               [-hpcov PERCENTCOVERAGEHTYPE] [--verify] [-o OUTPUT]
               [-r REFSEQ] [-s] [--debug] [--dbpath DBPATH]

ectyper v1.0.0 database v1.0 Prediction of Escherichia coli serotype from raw
reads or assembled genome sequences. The default settings are recommended.

optional arguments:
  -h, --help            show this help message and exit
  -V, --version         show program's version number and exit
  -i INPUT, --input INPUT
                        Location of E. coli genome file(s). Can be a single
                        file, a comma-separated list of files, or a directory
  -c CORES, --cores CORES
                        The number of cores to run ectyper with
  -opid PERCENTIDENTITYOTYPE, --percentIdentityOtype PERCENTIDENTITYOTYPE
                        Percent identity required for an O antigen allele
                        match [default 90]
  -hpid PERCENTIDENTITYHTYPE, --percentIdentityHtype PERCENTIDENTITYHTYPE
                        Percent identity required for an H antigen allele
                        match [default 95]
  -opcov PERCENTCOVERAGEOTYPE, --percentCoverageOtype PERCENTCOVERAGEOTYPE
                        Minumum percent coverage required for an O antigen
                        allele match [default 95]
  -hpcov PERCENTCOVERAGEHTYPE, --percentCoverageHtype PERCENTCOVERAGEHTYPE
                        Minumum percent coverage required for an H antigen
```

Advanced Molecular Detection
Southeast Region Bioinformatics

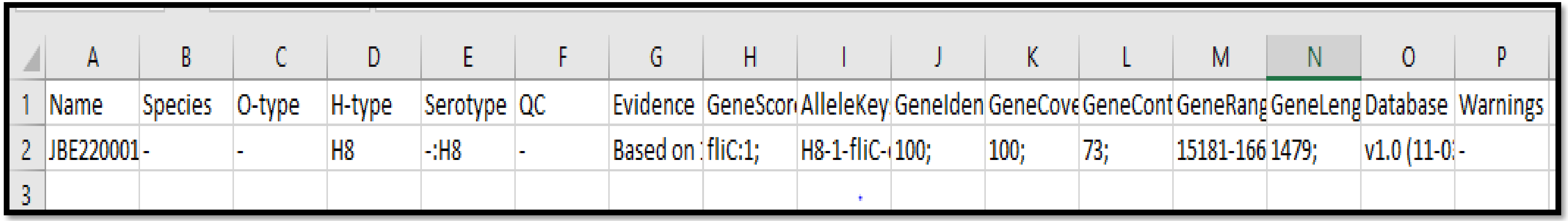
Input File Format

Takes *.fasta* files as input

```
$ ectyper -i JBE22000155.fasta -o results_ectyper
```

Output

Ectyper serotyping results are available in a tab-delimited **output.tsv** file consisting of 16 columns:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Name	Species	O-type	H-type	Serotype	QC	Evidence	GeneScor	AlleleKey	GeneIden	GeneCove	GeneCont	GeneRang	GeneLeng	Database	Warnings
2	JBE220001	-	-	H8	-:H8	-	Based on :	fliC:1;	H8-1-fliC-(	100;	100;	73;	15181-166	1479;	v1.0 (11-0:	-
3																

Interpreting Results

- Name – Sample name (usually a unique identifier)
- Species: the species column provides valuable species identification information in case of inadvertent sample contamination or mislabeling events
- O-type: O antigen
- H-type: H antigen
- Serotype: Predicted O and H antigen(s)
- QC: The Quality Control value summarizing the overall quality of prediction
- Evidence: How many alleles in total used to both call O and H antigens
- GeneScores: ECTyper O and H antigen gene scores in 0 to 1 range
- AllelesKeys: Best matching ECTyper database allele keys used to call the serotype

Interpreting Results

- GeneIdentities(%): %identity values of the query alleles
- GeneCoverages(%): %coverage values of the query alleles
- GeneContigNames: the contig names where the query alleles were found
- GeneRanges: genomic coordinates of the query alleles
- GeneLengths: allele lengths of the query alleles
- Database: database release version and date
- Warnings: any additional warnings linked to the quality control status or any other error message(s)

EmmTyper

- Emm Automatic Isolate Labeller
- emmtyper is a command line tool for emm-typing of *Streptococcus pyogenes* using a de novo or complete assembly
- Tool has two basic modes:
  - blast: contigs are blast against the trimmed FASTA database curated by the CDC
  - pcr: in-silico PCR is done on the contigs using the isPCR tool

MDU-PHL/emmtyper: emm Automatic Isolate Labeller (github.com)

Advanced Molecular Detection
Southeast Region Bioinformatics

How emm genes work?

- The difficulty with performing M-typing is that there is a single gene of interest (emm), and two other homologous genes (enn and mrp), often referred to emm-like
- Homologous genes may or may not occur in the isolate of interest
- When performing emm-typing from an assembly, we can distinguish between one or more clusters of matches on the contigs
- The best match for each of the clusters identified is then parsed from the BLAST results

Installation

Can be installed through conda

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

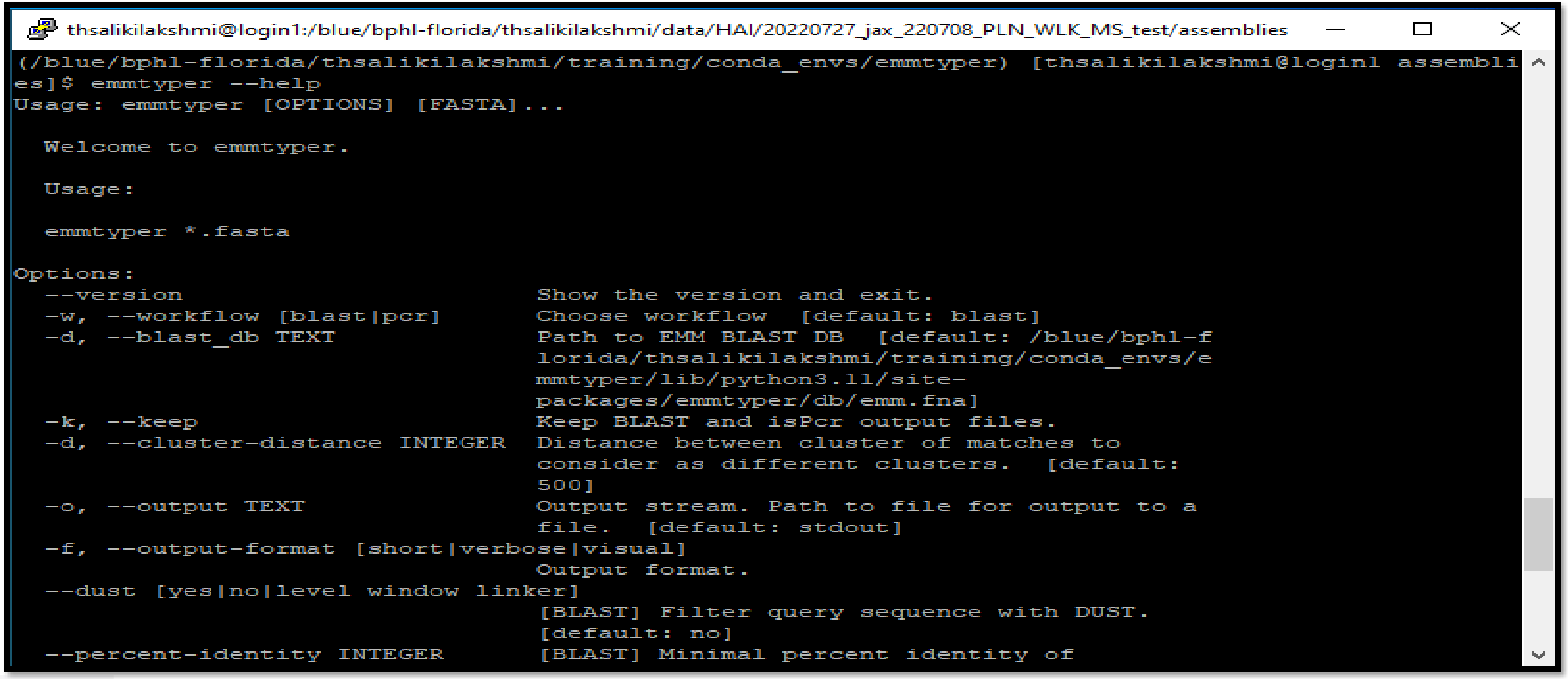
Advanced Molecular Detection
Southeast Region Bioinformatics

Usage

emmtyper has two workflows:

1. Directly BLASTing the contigs against the database
2. Using isPcr to generate an *in-silico* PCR product which is BLAST against the database

Help Menu


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

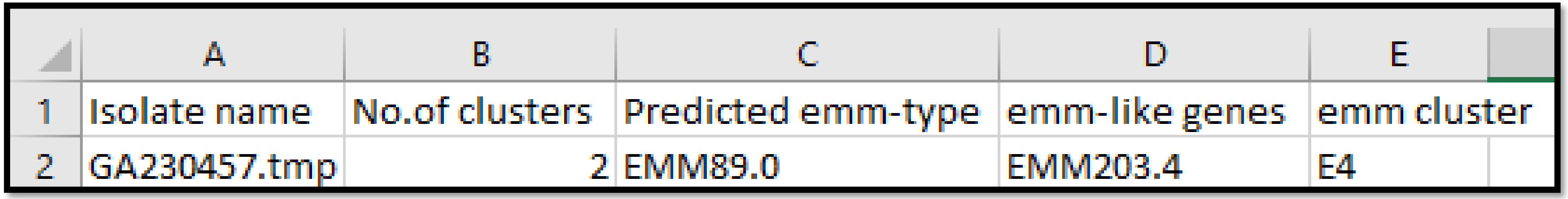
```
(/blue/bphl-florida/thsalikilakshmi/training/conda_envs/emmtyper) [thsalikilakshmi@login1 assembli
es]$ emmtyper --help
Usage: emmtyper [OPTIONS] [FASTA]...

  Welcome to emmtyper.

  Usage:

  emmtyper *.fasta

Options:
  --version                       Show the version and exit.
  -w, --workflow [blast|pcr]      Choose workflow  [default: blast]
  -d, --blast_db TEXT             Path to EMM BLAST DB  [default: /blue/bphl-f
                                  lorida/thsalikilakshmi/training/conda_envs/e
                                  mmtyper/lib/python3.11/site-
                                  packages/emmtyper/db/emm.fna]
  -k, --keep                      Keep BLAST and isPcr output files.
  -d, --cluster-distance INTEGER  Distance between cluster of matches to
                                  consider as different clusters.  [default:
                                  500]
  -o, --output TEXT               Output stream. Path to file for output to a
                                  file.  [default: stdout]
  -f, --output-format [short|verbose|visual]
                                  Output format.
  --dust [yes|no|level window linker]
                                  [BLAST] Filter query sequence with DUST.
                                  [default: no]
  --percent-identity INTEGER      [BLAST] Minimal percent identity of
```

Input & Results

- Takes *.fasta* files as input

```
$ emmtyper JBE*.fasta > results_emmtyper
```

- Results



	A	B	C	D	E
1	Isolate name	No.of clusters	Predicted emm-type	emm-like genes	emm cluster
2	GA230457.tmp	2	EMM89.0	EMM203.4	E4

Interpreting Results

- emmtyper has three different result formats: short, verbose, and visual
- emmtyper by default produces the short version. This consists of five values in tab-separated format printed to stdout. These values are:
  - Isolate name
  - Number of clusters: should be between 1 and 3, larger values could indicate contamination
  - Predicted emm-type
  - Possible emm-like alleles (semi colon separated list)
  - EMM cluster: Functional grouping of EMM types into 48 clusters

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