Advanced Molecular Detection

Southeast Region Bioinformatics

Bactopia

01/22/2024

Outline

- Bactopia
- Overview
- Installation
- Usage
- Questions

Agenda

February 5 – Sanibel Pipeline
February 19 – Genomic Epidemiology Training Part 1

Future Trainings

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

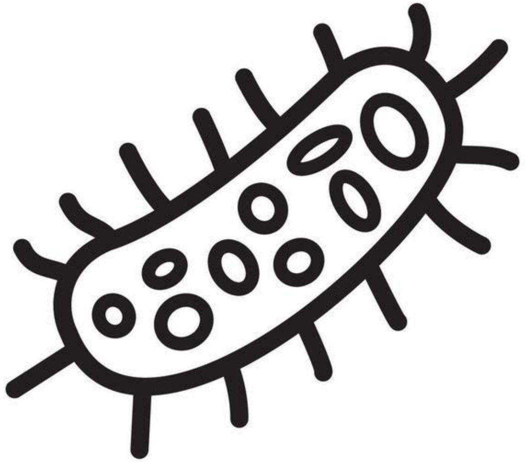
Advanced Molecular Detection
Southeast Region Bioinformatics

Notes

- We are planning to have our first quarterly meeting of 2024 in first week of March, please lookout for those emails.
- If any staff members require new HPG user training, please feel free to email us

Bactopia

- Flexible pipeline for complete analysis of bacterial genomes
- Processes the data with a broad set of tools, so that analyses can be quicker
- Bactopia was inspired by Staphopia, a workflow that targets *Staphylococcus aureus* genomes



bacteria image black and white - Search Images (bing.com)

Introduction – Bactopia (github link)

Advanced Molecular Detection
Southeast Region Bioinformatics

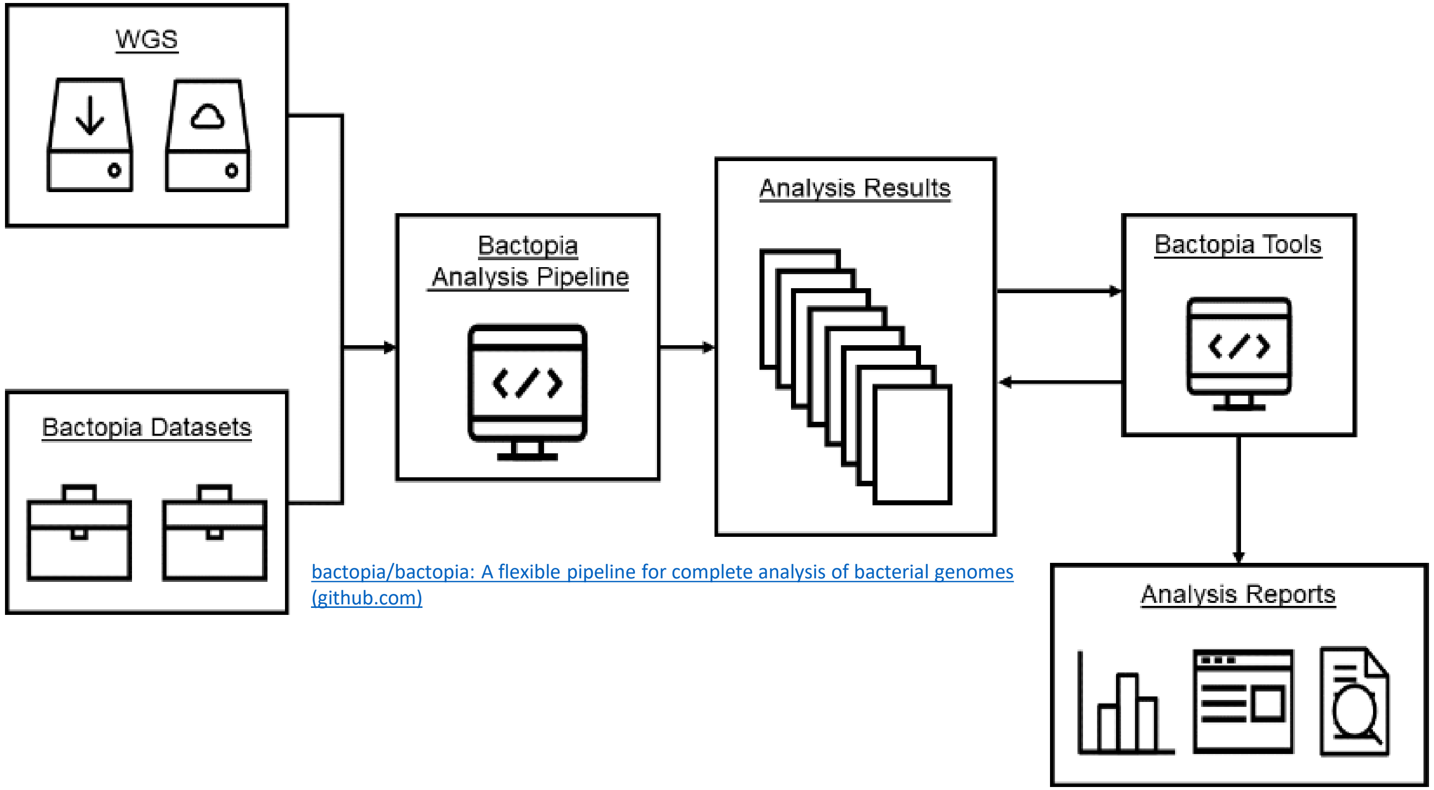
Overview

- Bactopia uses Nextflow to manage the workflow, allowing support of many types of environments (e.g. cluster or cloud)
- Bactopia uses many public datasets as well as your own datasets to further enhance the analysis of your sequencing
- Bactopia only uses software packages available from Bioconda and Conda-Forge to make installation simple for users

Framework of Bactopia

Bactopia is split into three main parts

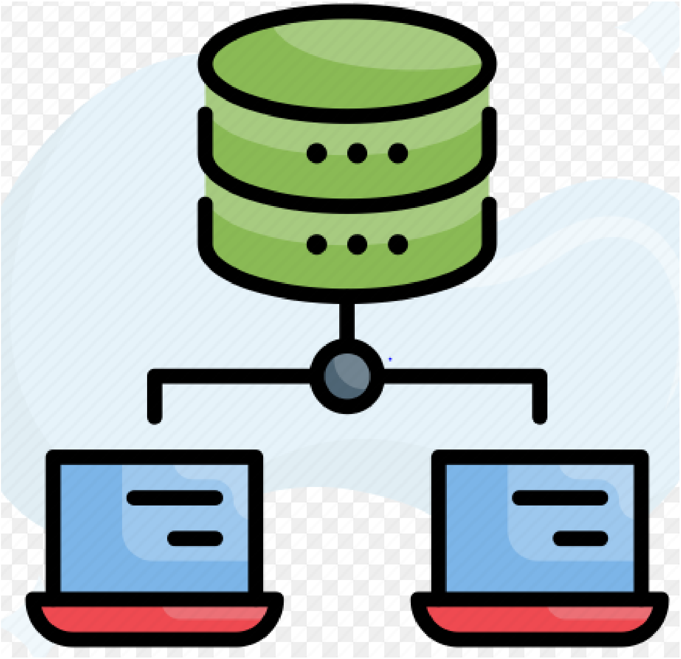
- Bactopia Datasets
- Bactopia Analysis Pipeline
- Bactopia Tools



bactopia/bactopia: A flexible pipeline for complete analysis of bacterial genomes (github.com)

Bactopia Datasets

- Provides a framework for including many existing public datasets, as well as private datasets, into your analysis
- Process of downloading, building, and (or) configuring these datasets for Bactopia has been automated



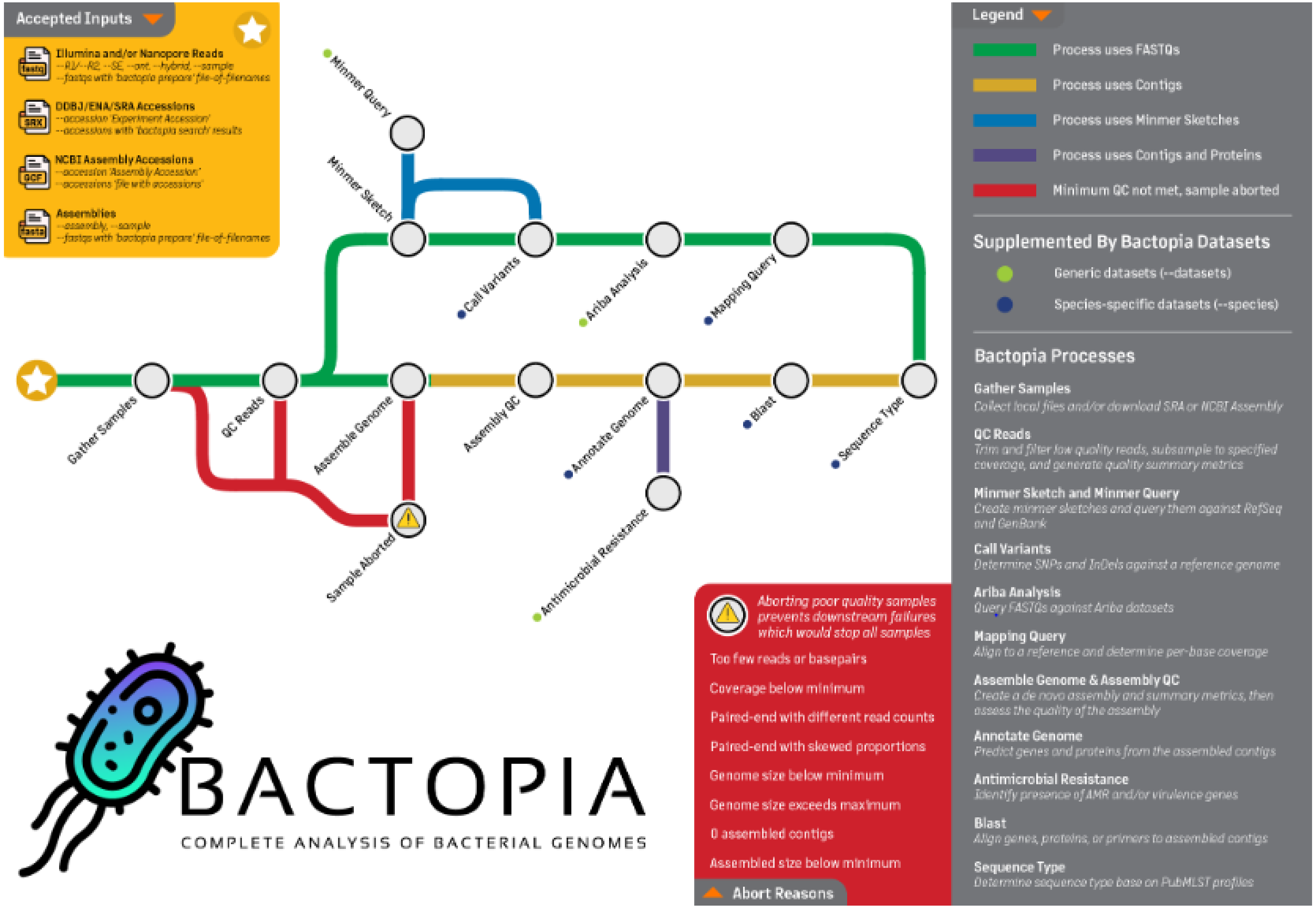
Dataset Icon - Search Images (bing.com)

Advanced Molecular Detection
Southeast Region Bioinformatics

Bactopia Analysis Pipeline

- Main per-isolate workflow in Bactopia
- Built with Nextflow, input FASTQs (local or available from SRA/ENA) are put through numerous analyses including:
  - Quality control
  - Assembly
  - Annotation
  - Reference mapping
  - Variant calling
  - Minmer sketch queries
  - Blast alignments
  - Insertion site prediction
  - Sequencing typing, etc.
- Automatically selects which analyses to include based on the available Bactopia Datasets

Advanced Molecular Detection
Southeast Region Bioinformatics



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 FASTQs
🟨 Process uses Contigs
🟦 Process uses Minmer Sketches
🟪 Process uses Contigs and Proteins
🟥 Minimum QC not met, sample aborted

Supplemented By Bactopia Datasets

● Generic datasets (--datasets)
● Species-specific datasets (--species)

Bactopia Processes

Gather Samples
Collect local files and/or download SRA or NCBI Assembly

QC Reads
Trim and filter low quality reads, subsample to specified coverage, and generate quality summary metrics

Minmer Sketch and Minmer Query
Create minmer sketches and query them against RefSeq and GenBank

Call Variants
Determine SNPs and InDels against a reference genome

Ariba Analysis
Query FASTQs against Ariba datasets

Mapping Query
Align to a reference and determine per-base coverage

Assemble Genome & Assembly QC
Create a de novo assembly and summary metrics, then assess the quality of the assembly

Annotate Genome
Predict genes and proteins from the assembled contigs

Antimicrobial Resistance
Identify presence of AMR and/or virulence genes

Blast
Align genes, proteins, or primers to assembled contigs

Sequence Type
Determine sequence type base on PubMLST profiles

Aborting poor quality samples prevents downstream failures which would stop all samples

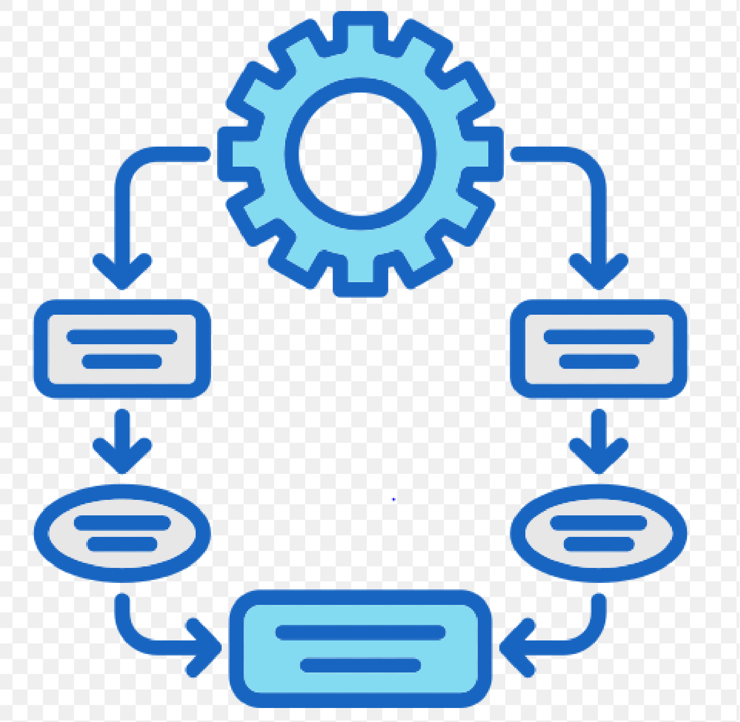
Too few reads or basepairs
Coverage below minimum
Paired-end with different read counts
Paired-end with skewed proportions
Genome size below minimum
Genome size exceeds maximum
0 assembled contigs
Assembled size below minimum

Abort Reasons

Advanced Molecular Detection
Southeast Region Bioinformatics

Workflow Steps

1. Gather - Collect all the data in one place
2. QC – Quality control of the data
3. Assembler – Assemble the reads into contigs
4. Annotator – Annotate the contigs
5. Sketcher – Create a sketch of the contigs, and query databases
6. Sequence Typing – Determine the sequence type of the contigs
7. Antibiotic Resistance – Determine the antibiotic resistance of the contigs and proteins
8. Merlin – Automatically run species-specific tools based on distance



independent workflows images free - Search Images (bing.com)

Bactopia tools

- Set of independent workflows for comparative analyses
- Comparative analyses include summary reports, pan-genome, or phylogenetic tree construction
- Using the predictable output structure of Bactopia you can pick and choose which samples to include for processing with a Bactopia tool

Advanced Molecular Detection
Southeast Region Bioinformatics

Installation via Conda

```
# Install Bactopia using Mamba
mamba create -y -n bactopia -c conda-forge -c bioconda bactopia

# Test Bactopia
# First launch will setup environments (e.g., Conda, Docker, or Singularity)
conda activate bactopia
bactopia -profile test,standard
```

Use -profile to change environment

- Default profile for Bactopia is Conda
- If you are testing using Docker or Singularity you would use:
  - -profile test,docker
  - -profile test,singularity

Advanced Molecular Detection
Southeast Region Bioinformatics

Run from GitHub Repository

If you already have Nextflow installed, and don’t want to use Conda to install Bactopia, you can run Bactopia directly from the GitHub repository

```
nextflow run bactopia/bactopia –profile test,standard
```

Advanced Molecular Detection
Southeast Region Bioinformatics

Usage

- Activate Nextflow

```
$ conda activate nextflow
or
$ module load nextflow
```

- Download Bactopia from GitHub to your work directory

```
$ git clone https://github.com/bactopia/bactopia.git
```

Usage

- Use this command to test using an interactive session

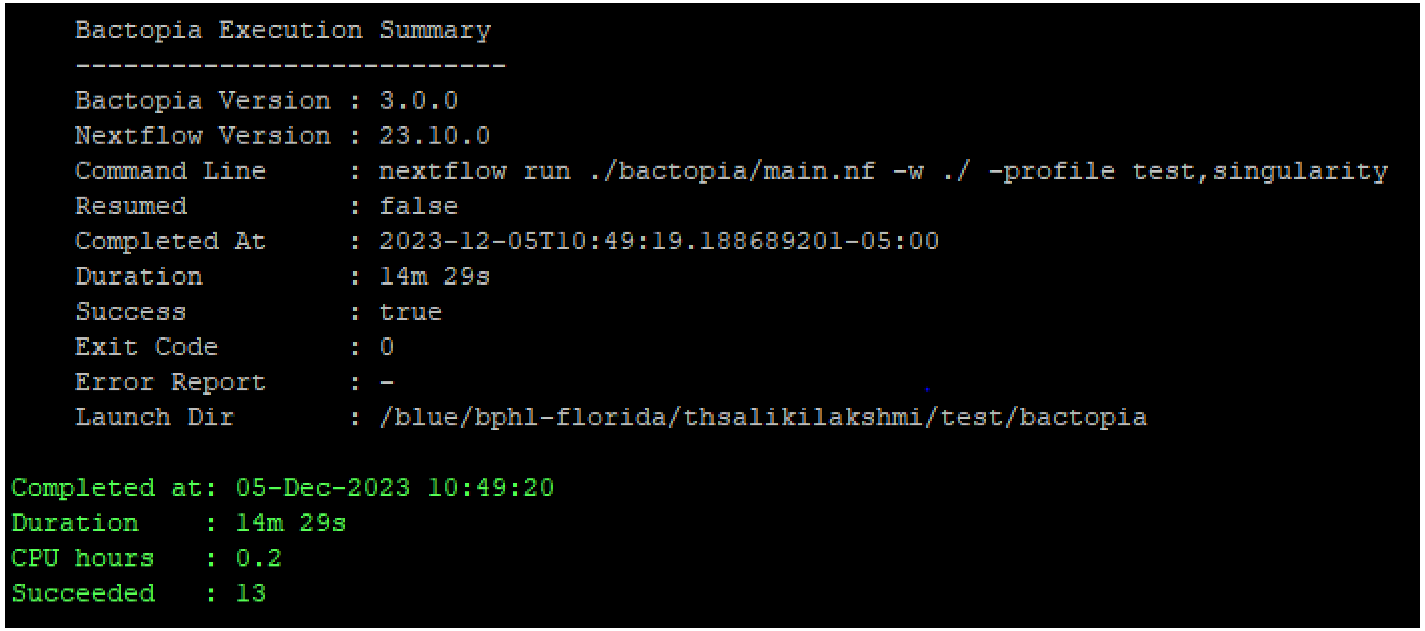
```
$ srun --qos=bphl-umbrella --account=bphl-umbrella --cpus-per-task=4 --mem=10gb --time=02:00:00 --pty bash -i
```

- Use this command to verify Bactopia is working

```
$ nextflow run  ./bactopia/main.nf -w ./ -profile test,singularity
```

Verify Bactopia is working

Upon completion, you will see this text which assures that Bactopia is working



```
    Bactopia Execution Summary
    ---------------------------
    Bactopia Version : 3.0.0
    Nextflow Version : 23.10.0
    Command Line     : nextflow run ./bactopia/main.nf -w ./ -profile test,singularity
    Resumed          : false
    Completed At     : 2023-12-05T10:49:19.188689201-05:00
    Duration         : 14m 29s
    Success          : true
    Exit Code        : 0
    Error Report     : -
    Launch Dir       : /blue/bphl-florida/thsalikilakshmi/test/bactopia

Completed at: 05-Dec-2023 10:49:20
Duration    : 14m 29s
CPU hours   : 0.2
Succeeded   : 13
```

Advanced Molecular Detection
Southeast Region Bioinformatics

Input Parameters

Accepted inputs by Bactopia include:

- Local Illumina and/or Nanopore Reads
- Local Assemblies
- ENA/SRA Experiment Accessions
- NCBI Assembly Accessions

Single Sample

- Bactopia accepts many different types of inputs from a single-entry point i.e., you don’t need a separate pipeline for each input type
- Bactopia accepts both Illumina(pair-end or single-end) and Nanopore reads, and can even process them together for a hybrid assembly
- --sample is always required for single-sample processing

Input Type	Required Parameters
Illumina Paired-End	--r1 and --r2
Illumina Single-End	--se
Oxford Nanopore	--ont
Hybrid	--r1, --r2, --ont, and --hybrid
Hybrid(Short-read Polishing)	--r1, --r2, --ont, and --short_polish

Working with Single Sample

```
nextflow run ./bactopia/main.nf -w ./ -profile singularity --r1
./fastqs/JBI22000793_1.fastq.gz --r2 ./fastqs/JBI22000793_2.fastq.gz --sample
JBI22000793 --outdir bactopia_out
```

Multiple Samples

- Bactopia can also process thousands of samples in a single command with the help of samplesheet or FOFN (file of filenames)
- Bactopia has other commands which assist in generating the appropriate FOFN or accession list to process multiple samples include:

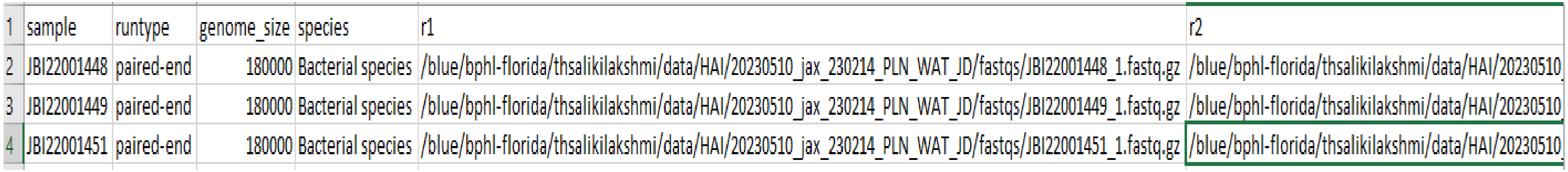
Parameter	Application	Helper Command
--samples	Local Samples	bactopia prepare
--accessions	ENA/SRA & Assembly Accessions	bactopia search

Generating Samplesheet

Tab-delimited table with five columns

Column	Description
sample	Unique name or prefix, to be used for naming output files
runtype	What type of input the sample is (e.g., paired-end, single-end, nanopore)
genome_size	Expected genome size for given sample
species	Expected taxonomic classification for the given sample
r1	If paired-end, the first pair of reads, else the single-end reads
r2	If paired-end, the second pair of reads
extra	Either the assembly or long reads associated with a sample

Samplesheet for Multiple Samples



1	sample	runtype	genome_size	species	r1	r2
2	JBI22001448	paired-end	180000	Bacterial species	/blue/bphl-florida/thsalikilakshmi/data/HAI/20230510_jax_230214_PLN_WAT_JD/fastqs/JBI22001448_1.fastq.gz	/blue/bphl-florida/thsalikilakshmi/data/HAI/20230510_
3	JBI22001449	paired-end	180000	Bacterial species	/blue/bphl-florida/thsalikilakshmi/data/HAI/20230510_jax_230214_PLN_WAT_JD/fastqs/JBI22001449_1.fastq.gz	/blue/bphl-florida/thsalikilakshmi/data/HAI/20230510_
4	JBI22001451	paired-end	180000	Bacterial species	/blue/bphl-florida/thsalikilakshmi/data/HAI/20230510_jax_230214_PLN_WAT_JD/fastqs/JBI22001451_1.fastq.gz	/blue/bphl-florida/thsalikilakshmi/data/HAI/20230510_

Advanced Molecular Detection
Southeast Region Bioinformatics

Working with Multiple Samples

```
nextflow run ./bactopia/main.nf -w ./ -profile singularity --
samples Samples.txt --outdir bactopia_samples_out
```

Output

Bactopia gives outputs for the following steps:

- Gather
- QC
- Assembler
- Annotator
- Sketcher
- Sequence Typing
- Antibiotic Resistance
- Merlin

Gather Outputs

- Main purpose of this step is to get all the samples into a single place which includes downloading samples from ENA/SRA or NCBI Assembly
- Gather step also does basic QC checks to help prevent downstream failures
- Merged Results
  - meta.tsv - tab-delimited file with bactopia metadata for all samples
- Gather
  - meta.tsv - tab-delimited file with bactopia metadata for each sample

QC Output

- The QC module uses a variety of tools to perform quality control on Illumina and Oxford Nanopore reads
- Like the gather step, the QC step will also stop samples that fail to meet basic QC checks from continuing downstream
- Tools used in this step include bbtools, fastp, fastqc, fastq_scan, lighter, NanoPlot, nanoq, porechop, rasusa

QC Output

Filename	Description
<SAMPLE_NAME>.fastq.gz	A gzipped FASTQ file containing the cleaned Illumina single-end, or Oxford Nanopore reads
<SAMPLE_NAME>_R{1|2}.fastq.gz	A gzipped FASTQ file containing the cleaned Illumina paired-end reads
<SAMPLE_NAME>-{final|original}.json	A JSON file containing the QC results generated by fastq-scan
<SAMPLE_NAME>-{final|original}_fastqc.html	(Illumina Only) A HTML report of the QC results generated by fastqc
<SAMPLE_NAME>-{final|original}_fastqc.zip	(Illumina Only) A zip file containing the complete set of fastqc results
<SAMPLE_NAME>-{final|original}_fastp.json	(Illumina Only) A JSON file containing the QC results generated by fastp
<SAMPLE_NAME>-{final|original}_fastp.html	(Illumina Only) A HTML report of the QC results generated by fastp
<SAMPLE_NAME>-{final|original}_NanoPlot-report.html	(ONT Only) A HTML report of the QC results generated by NanoPlot
<SAMPLE_NAME>-{final|original}_NanoPlot.tar.gz	(ONT Only) A tarball containing the complete set of NanoPlot results

Assembler Output

- The assembler module uses a variety of assembly tools to create an assembly of Illumina and Oxford Nanopore reads. Tools used are Dragonflye, Shovill, Shovill-SE, Unicycler
- Summary statistics for each assembly are generated using assembly-scan
- Merged Results
  - assembly-scan.tsv - Assembly statistics for all samples

Annotator Output

- Prokka
  - Provides description of the per-sample results from Prokka
  - Software used to rapidly annotate the metagenome-assembled genomes
  - Has not been developed further in the last years

- Bakta
  - Provides description of the per-sample results from Bakta
  - Increases the ability to assign the newly annotated coding sequences to genes that are available in reference databases and to improve the export of the annotations, e.g., by using JSON files
  - Additionally provided annotations for non-coding RNA (regions), small open reading frames (sorf), origin of replication (oriC), and CRISPR

Advanced Molecular Detection
Southeast Region Bioinformatics

Sketcher Output

- The sketcher module uses Mash and Sourmash to create sketches and query RefSeq and GTDB
- Outputs
  - Provides description of the per-sample results from the sketcher sub workflow

Filename	Description
<SAMPLE_NAME>-k{21|31}.msh	A Mash sketch of the input assembly for k=21 and k=31
<SAMPLE_NAME>-mash-refseq88-k21.txt	The results of querying the Mash sketch against RefSeq88
<SAMPLE_NAME>-sourmash-gtdb-rs207-k31.txt	The results of querying the Sourmash sketch against GTDB-rs207
<SAMPLE_NAME>.sig	A Sourmash sketch of the input assembly for k=21, k=31, and k=51

Advanced Molecular Detection
Southeast Region Bioinformatics

Sequence Typing Output

Outputs include:

- Merged Results – results are concatenated into a single file
    - mlst.tsv - merged TSV file with mlst results from all samples
- mlst – results – description of the per-sample results from mlst
    - <sample_NAME>.tsv - tab-delimited file with mlst result

Antibiotic Resistance Outputs

Outputs include:

- Merged Results – results are concatenated into a single file
  - amrfinderplus-genes.tsv - merged TSV file with AMRFinder+ results using nucleotide inputs
  - amrfinderplus-proteins.tsv - merged TSV file with AMRFinder+ results using protein inputs
- AMRFinder+ - description of the per-sample results from AMRfinder+
  - genes.tsv - TSV file with AMRFinder+ results using nucleotide inputs
  - proteins.tsv - TSV file with AMRfinder+ results using protein inputs

Merlin Outputs

- Merlin outputs results are concatenated into a single file
- Merlin outputs .tsv files for 22 tools which include agrvate, ectyper, emmtyper, legsta, lissero, meningotype, stecfinder, etc

Advanced Molecular Detection
Southeast Region Bioinformatics

nf-core/modules Availability

- All bactopia tools are also available through nf-core/modules, a repository of ready to use Nextflow DSL2 modules
- You can also leverage nf-core tools to rapidly string together tour own workflows
- Many of the Bactopia Tools were submitted to nf-core/modules as part of Bactopia V2

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