Introduction to Bactopia for Bacterial Genome Analysis

Robert A. Petit III, PhD
TOAST Office Hours
June 21st, 2024

Wyoming Department of Health
PUBLIC HEALTH DIVISION
WYOMING PUBLIC HEALTH LABORATORY

Yo! 👋

I AM ROBERT

WPHL's bioinformatician and developer of Bactopia (*the topic of today's session!*)

Wyoming Department of Health | PUBLIC HEALTH DIVISION | WYOMING PUBLIC HEALTH LABORATORY

What is "Bactopia"?

Let's introduce all things Bactopia

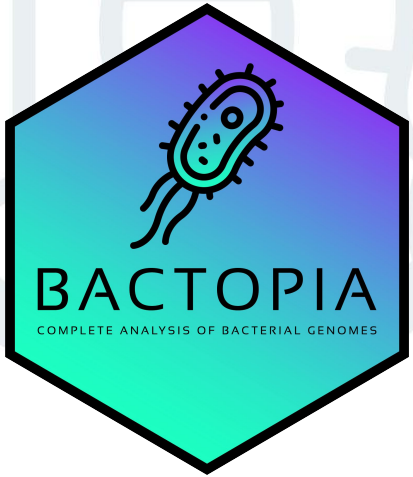
Bactopia streamlines bacterial genome analysis

A Nextflow pipeline for end-to-end analysis of bacterial genomes

Supports Illumina and Nanopore technologies, and your favorite bacterial species

Wraps 150+ bioinformatic tools into stand-alone modules

Sustained for 5+ years with support from WPHL, Emory University, CAPE



Wyoming Department of Health | PUBLIC HEALTH DIVISION

EMORY UNIVERSITY SCHOOL OF MEDICINE

Center for Applied Pathogen Epidemiology and Outbreak Control

Let's take a deeper look at Bactopia

By walking through it step-by-step

What is Bactopia?

Bactopia is an end-to-end Nextflow pipeline for the complete analysis of bacterial genomes. In a few simple steps, go from raw sequences to an extensive set of results to investigate.

BACTOPIA
COMPLETE ANALYSIS OF BACTERIAL GENOMES

STEP 01
Sequence Bacterial Genomes
Bactopia supports sequences from Illumina and Nanopore technologies, or a combination of the two.

STEP 02
Install Bactopia
Bactopia is easily installed using Conda and analyses can be run using Conda, Docker, or Singularity

STEP 03
Create a Sample Sheet
bactopia prepare is used to create a sample sheet allowing a single sample or thousands of samples to be processed

STEP 04
Include Public Samples
bactopia search can query ENA/SRA to identify accessions to further supplement analyses.

STEP 05
Choose Compute Environment
Built using Nextflow, Bactopia supports many compute and execution environments. Simple parameter changes, allow efficient switching between local, HPC, and major cloud environments.

BACTOPIA STEP 06
Process Genomes
Bactopia includes more than 150 bioinformatic tools in the pipeline, with extensive version control and data provenance. Upon completion outputs are available in a standardized structure for easy access.

STEP 08
Investigate Results
With analysis completed, numerous output files are available to assist in your genomic investigations

BACTOPIA TOOLS STEP 07
Additional Workflows
More than 60 Bactopia Tools allow seamless post-analysis of Bactopia outputs, including additional workflows for pangenome, taxonomic classification, phylogenies, AMR, among many more.

What is Bactopia?

Bactopia is an end-to-end Nextflow pipeline for the complete analysis of bacterial genomes. In a few simple steps, go from raw sequences to an extensive set of results to investigate.

BACTOPIA
COMPLETE ANALYSIS OF BACTERIAL GENOMES

STEP 01
Sequence Bacterial Genomes
Bactopia supports sequences from Illumina and Nanopore technologies, or a combination of the two.

STEP 02
Install Bactopia
Bactopia is easily installed using Conda and analyses can be run using Conda, Docker, or Singularity

STEP 03
Create a Sample Sheet
bactopia prepare is used to create a sample sheet allowing a single sample or thousands of samples to be processed

STEP 04
Include Public Samples
bactopia search can query ENA/SRA to identify accessions to further supplement analyses.

STEP 05
Choose Compute Environment
Built using Nextflow, Bactopia supports many compute and execution environments. Simple parameter changes, allow efficient switching between local, HPC, and major cloud environments.

BACTOPIA STEP 06
Process Genomes
Bactopia includes more than 150 bioinformatic tools in the pipeline, with extensive version control and data provenance. Upon completion outputs are available in a standardized structure for easy access.

STEP 07
Additional Workflows
BACTOPIA TOOLS
More than 60 Bactopia Tools allow seamless post-analysis of Bactopia outputs, including additional workflows for pangenome, taxonomic classification, phylogenies, AMR, among many more.

STEP 08
Investigate Results
With analysis completed, numerous output files are available to assist in your genomic investigations

Step 1: Sequence Bacterial Genomes

Sequence your favorite bacterial species

Bactopia supports:

- Illumina
- Oxford Nanopore
- Both Illumina and Nanopore together
- Assemblies

If you have bacterial sequences, Bactopia can process them

What is Bactopia?

Bactopia is an end-to-end Nextflow pipeline for the complete analysis of bacterial genomes. In a few simple steps, go from raw sequences to an extensive set of results to investigate.

BACTOPIA
COMPLETE ANALYSIS OF BACTERIAL GENOMES

STEP 01
Sequence Bacterial Genomes
Bactopia supports sequences from Illumina and Nanopore technologies, or a combination of the two.

STEP 02
Install Bactopia
Bactopia is easily installed using Conda and analyses can be run using Conda, Docker, or Singularity

STEP 03
Create a Sample Sheet
bactopia prepare is used to create a sample sheet allowing a single sample or thousands of samples to be processed

STEP 04
Include Public Samples
bactopia search can query ENA/SRA to identify accessions to further supplement analyses.

STEP 05
Choose Compute Environment
Built using Nextflow, Bactopia supports many compute and execution environments. Simple parameter changes, allow efficient switching between local, HPC, and major cloud environments.

BACTOPIA STEP 06
Process Genomes
Bactopia includes more than 150 bioinformatic tools in the pipeline, with extensive version control and data provenance. Upon completion outputs are available in a standardized structure for easy access.

STEP 08
Investigate Results
With analysis completed, numerous output files are available to assist in your genomic investigations

BACTOPIA TOOLS STEP 07
Additional Workflows
More than 60 Bactopia Tools allow seamless post-analysis of Bactopia outputs, including additional workflows for pangenome, taxonomic classification, phylogenies, AMR, among many more.

Step 2: Install Bactopia

Bactopia is easily installed from Bioconda or executed directly from "*nextflow run*"

"*bactopia-py*" (included in Conda install) provides many helpers along

Every analysis step within Bactopia supports:

- Conda
- Docker
- Singularity / Apptainer

Every tool has explicit version control

Bactopia makes installation and reproducibility a straightforward process

What is Bactopia?

Bactopia is an end-to-end Nextflow pipeline for the complete analysis of bacterial genomes. In a few simple steps, go from raw sequences to an extensive set of results to investigate.

BACTOPIA
COMPLETE ANALYSIS OF BACTERIAL GENOMES

STEP 01
Sequence Bacterial Genomes
Bactopia supports sequences from Illumina and Nanopore technologies, or a combination of the two.

STEP 02
Install Bactopia
Bactopia is easily installed using Conda and analyses can be run using Conda, Docker, or Singularity

STEP 03
Create a Sample Sheet
bactopia prepare is used to create a sample sheet allowing a single sample or thousands of samples to be processed

STEP 04
Include Public Samples
bactopia search can query ENA/SRA to identify accessions to further supplement analyses.

STEP 05
Choose Compute Environment
Built using Nextflow, Bactopia supports many compute and execution environments. Simple parameter changes, allow efficient switching between local, HPC, and major cloud environments.

BACTOPIA STEP 06
Process Genomes
Bactopia includes more than 150 bioinformatic tools in the pipeline, with extensive version control and data provenance. Upon completion outputs are available in a standardized structure for easy access.

STEP 07
Additional Workflows
BACTOPIA TOOLS
More than 60 Bactopia Tools allow seamless post-analysis of Bactopia outputs, including additional workflows for pangenome, taxonomic classification, phylogenies, AMR, among many more.

STEP 08
Investigate Results
With analysis completed, numerous output files are available to assist in your genomic investigations

Step 3 & 4: Process a single or thousands of samples

Bactopia allows processing of a single sample of 10s of thousands of samples

"*bactopia prepare*" - Simplifies the process of creating a sample sheet
Very useful for large numbers of samples, and many species in a single run
"*bactopia search*" - Query ENA or SRA to identify samples to include in your analysis
Can be raw FASTQs or Assemblies, Bactopia will automatically download them!
"*bactopia atb-formatter*" - Support 2,000,000 assemblies from AllTheBacteria

Example: Processing 67,000 Staphylococcus aureus genomes on AWS Batch

Bactopia is highly scalable and promotes usage of publicly available genomes

What is Bactopia?

Bactopia is an end-to-end Nextflow pipeline for the complete analysis of bacterial genomes. In a few simple steps, go from raw sequences to an extensive set of results to investigate.

BACTOPIA
COMPLETE ANALYSIS OF BACTERIAL GENOMES

STEP 01
Sequence Bacterial Genomes
Bactopia supports sequences from Illumina and Nanopore technologies, or a combination of the two.

STEP 02
Install Bactopia
Bactopia is easily installed using Conda and analyses can be run using Conda, Docker, or Singularity

STEP 03
Create a Sample Sheet
bactopia prepare is used to create a sample sheet allowing a single sample or thousands of samples to be processed

STEP 04
Include Public Samples
bactopia search can query ENA/SRA to identify accessions to further supplement analyses.

STEP 05
Choose Compute Environment
Built using Nextflow, Bactopia supports many compute and execution environments. Simple parameter changes, allow efficient switching between local, HPC, and major cloud environments.

BACTOPIA STEP 06
Process Genomes
Bactopia includes more than 150 bioinformatic tools in the pipeline, with extensive version control and data provenance. Upon completion outputs are available in a standardized structure for easy access.

STEP 08
Investigate Results
With analysis completed, numerous output files are available to assist in your genomic investigations

BACTOPIA TOOLS STEP 07
Additional Workflows
More than 60 Bactopia Tools allow seamless post-analysis of Bactopia outputs, including additional workflows for pangenome, taxonomic classification, phylogenies, AMR, among many more.

Step 5: Choose Compute Environment

Bactopia can be run on Linux (including WSL2) or Mac OSX (via Docker)

Being built using Nextflow, 18 different executors are supported:

- Local: Desktops, laptops, servers
- HPC: LSF, PBS, SGE, SLURM
- Cloud: Amazon Web Services, Google Cloud Platform, Microsoft Azure
- Many more execution environments

Supports Nextflow configs from nf-core/configs

Nextflow empowers Bactopia to be extremely portable across many different systems

What is Bactopia?

Bactopia is an end-to-end Nextflow pipeline for the complete analysis of bacterial genomes. In a few simple steps, go from raw sequences to an extensive set of results to investigate.

BACTOPIA
COMPLETE ANALYSIS OF BACTERIAL GENOMES

STEP 01
Sequence Bacterial Genomes
Bactopia supports sequences from Illumina and Nanopore technologies, or a combination of the two.

STEP 02
Install Bactopia
Bactopia is easily installed using Conda and analyses can be run using Conda, Docker, or Singularity

STEP 03
Create a Sample Sheet
bactopia prepare is used to create a sample sheet allowing a single sample or thousands of samples to be processed

STEP 04
Include Public Samples
bactopia search can query ENA/SRA to identify accessions to further supplement analyses.

STEP 05
Choose Compute Environment
Built using Nextflow, Bactopia supports many compute and execution environments. Simple parameter changes, allow efficient switching between local, HPC, and major cloud environments.

BACTOPIA STEP 06
Process Genomes
Bactopia includes more than 150 bioinformatic tools in the pipeline, with extensive version control and data provenance. Upon completion outputs are available in a standardized structure for easy access.

STEP 08
Investigate Results
With analysis completed, numerous output files are available to assist in your genomic investigations

BACTOPIA TOOLS STEP 07
Additional Workflows
More than 60 Bactopia Tools allow seamless post-analysis of Bactopia outputs, including additional workflows for pangenome, taxonomic classification, phylogenies, AMR, among many more.

Step 6: Process Genomes (*finally!*)

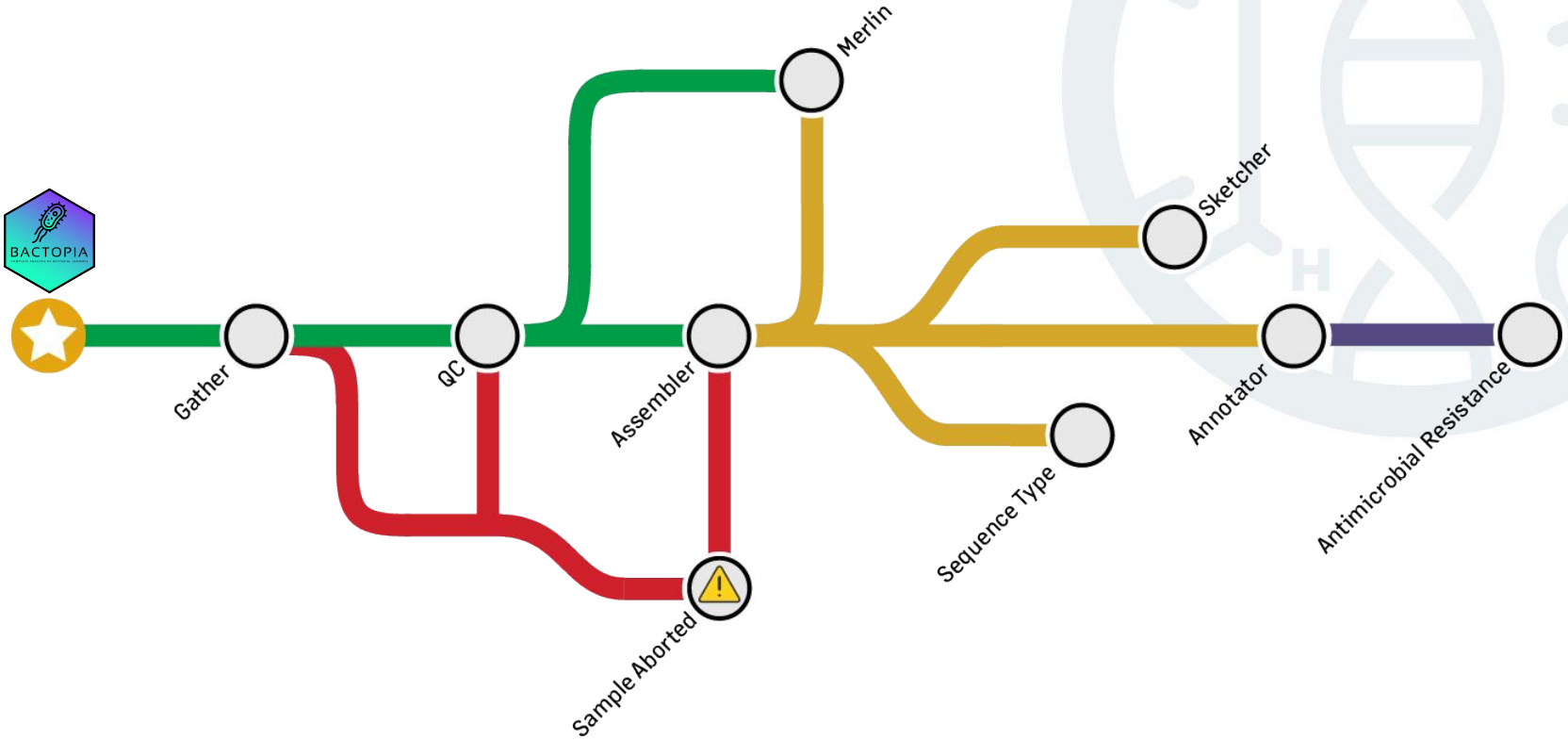
Genomes are sequenced, Bactopia is installed, sample sheets are created, execution environments are setup, it's time to start processing!

Processing in Bactopia is split between:

- Main Pipeline - Species agnostic analyses
- Bactopia Tools - Downstream targeted analyses

Let's explore the main Bactopia pipeline

Step 6: The "*main*" Bactopia pipeline



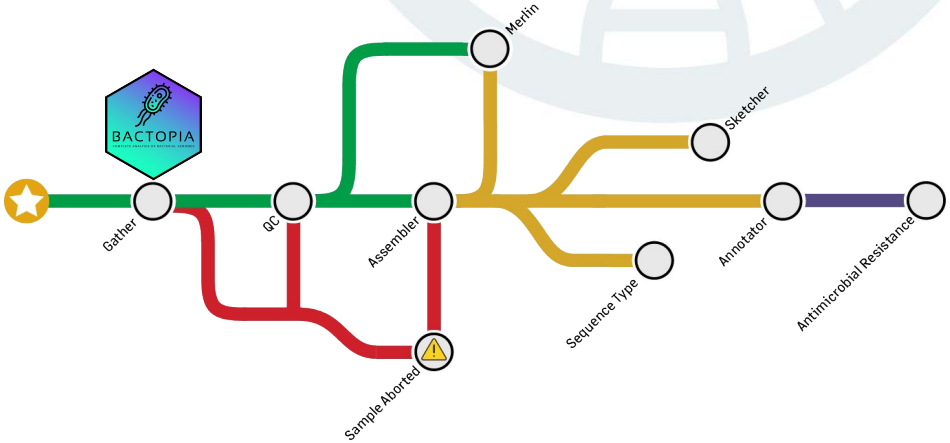
Step 6: The “gather” Step

The “gather” step brings all the samples together

- Local and remote FASTQs are staged
- ENA/SRA and Assemblies are downloaded

Basic QC is also implemented here:

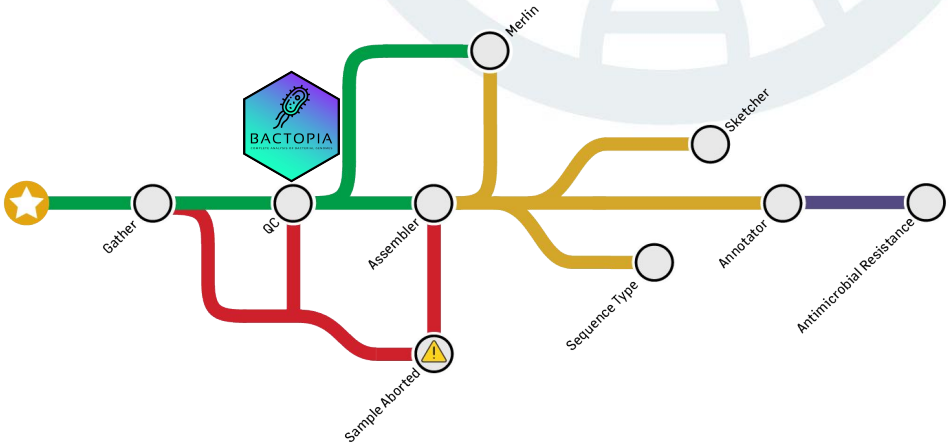
- Is the FASTQ a “FASTQ”?
- Were downloads successful?



Step 6: The “qc” Step

The “qc” step quality controls inputs

- Read tossing
  - Too short, ambiguous nucleotides
- Read trimming
- Quality stats
- Error correction
- Read stats
  - Means, Mins and Maxes
- Subsampling to a coverage



Step 6: The “assembler” Step

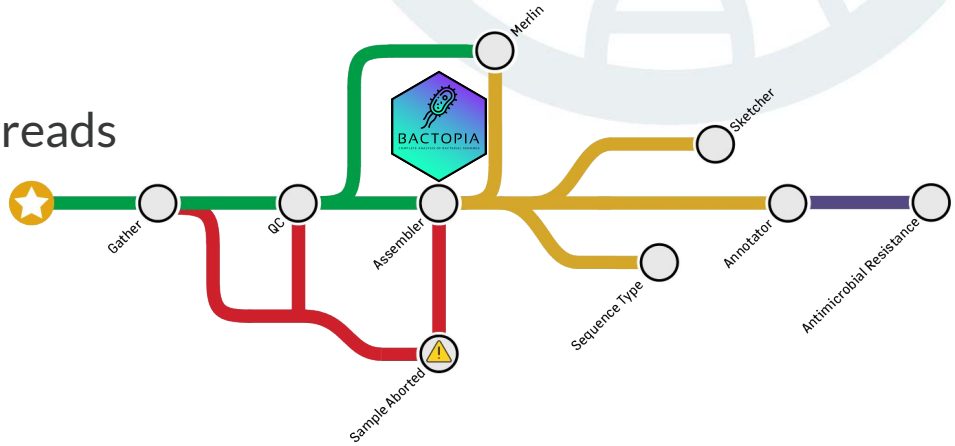
Time to assemble the reads!

- Short-read assembly with Shovill
- Long-read assembly with Dragonflye

Hybrid assembly

- Short reads first, then long reads
- Long reads, then polishing with short reads

Assembly stats calculated



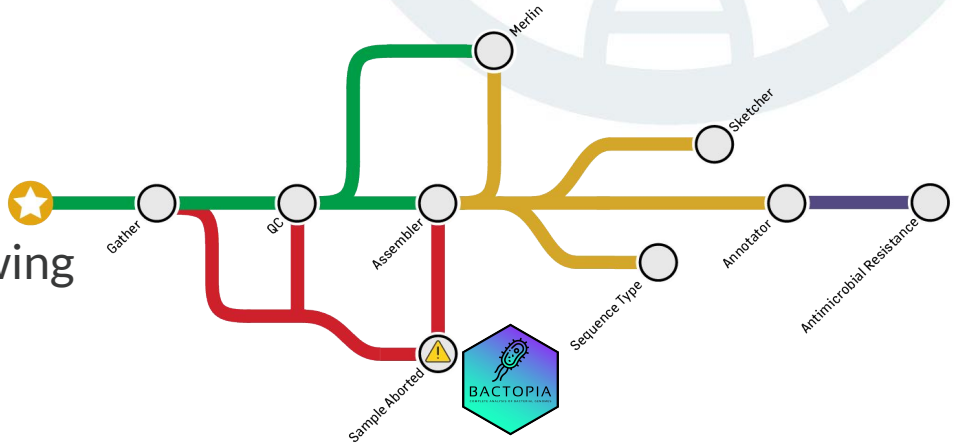
Step 6: Why are “samples aborted”?

Samples that fail QC thresholds are aborted

Example thresholds:

- Improper format
- Low coverage
- Low quality
- Assembly size outside range

Prevents downstream failures, while allowing other samples to continue processing



Step 6: The “merlin” Step

Minm**ER** assisted species-specific bactopia tool se**L**ect**I**o**N**

Use Mash distances to automatically run species specific tools

Useful when processing multiple species at a time

Completely optional step

Step 6: The “sketcher” Step

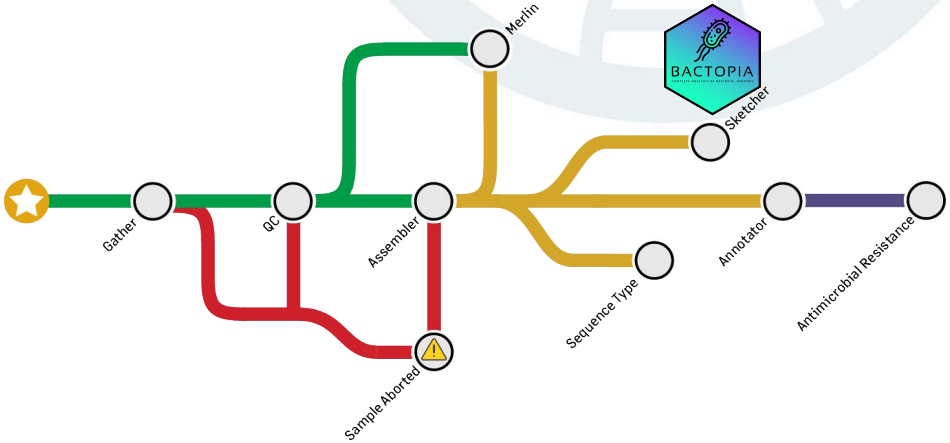
A framework for generating sketches of your samples

- Supports Mash and Sourmash
- Simple to add other sketching tools

Queries publicly available sketches:

- Mash - 50k RefSeq Genomes
- Sourmash - 80k GenBank genomes

A quick method for very basic taxonomic information



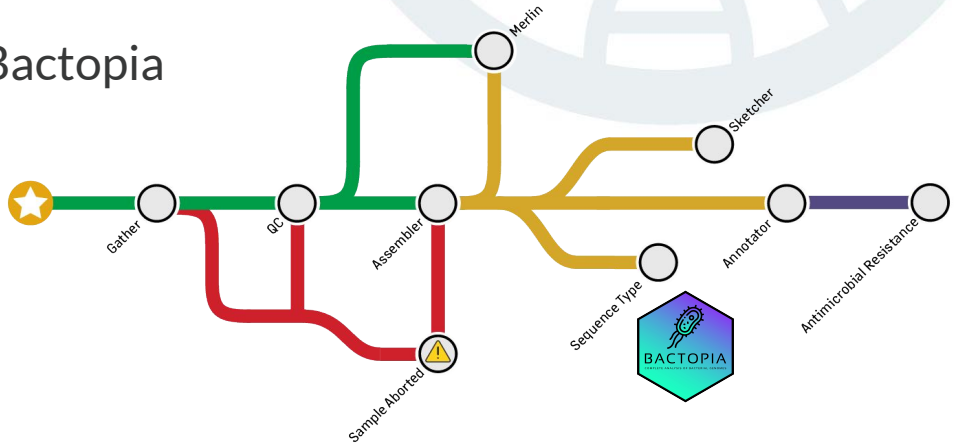
Step 6: The “sequence type” Step

Multi Locus Sequence Typing (MLST) of samples

Automatically selects species specific schema

Users can manually specify schema

MLST databases are packaged with each Bactopia release



Step 6: The “annotator” Step

Annotation of assemblies is conducted here

Supports both Prokka and Bakta for annotation

Bakta database download is automated

Users can provide their own custom protein sets for annotation

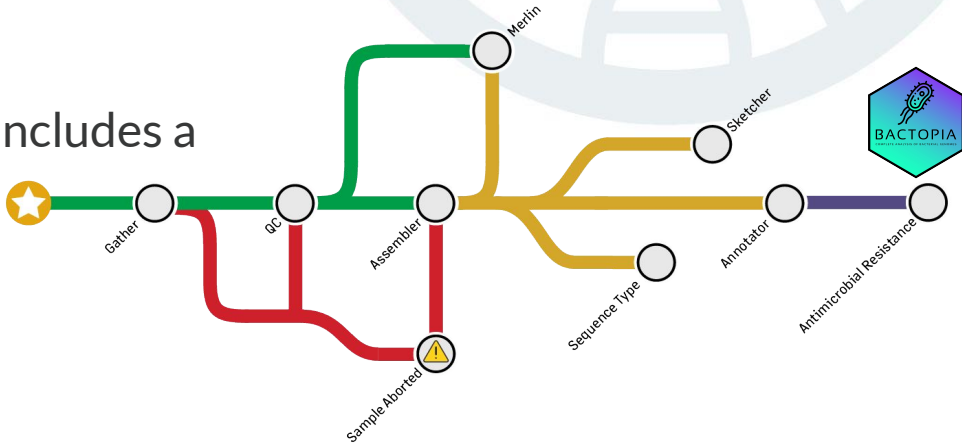
Step 6: The “antimicrobial resistance” Step

Antimicrobial resistances are predicted using AMRFinder+

The annotated genes, proteins and associated GFF files are used as inputs

Similar to MLST, each release of Bactopia includes a versioned AMRFinder+ database

Prevents version incompatibilities



Step 6: The "*main*" Bactopia pipeline

This will take 5-20 minutes per sample

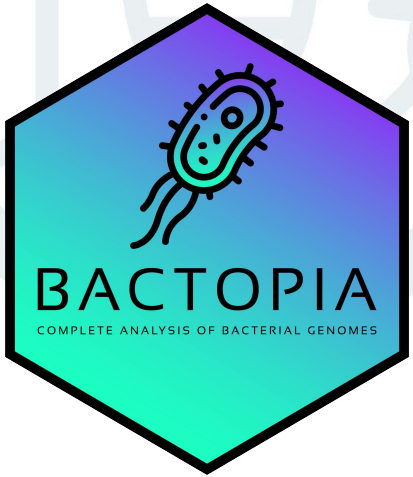
- Depends on input size and species

Output files are placed in a standardized structure

Extensive audit trails are available

- Logs, staging, versions, etc…

Steps with "mergeable" outputs are automatically merged



What is Bactopia?

Bactopia is an end-to-end Nextflow pipeline for the complete analysis of bacterial genomes. In a few simple steps, go from raw sequences to an extensive set of results to investigate.

BACTOPIA
COMPLETE ANALYSIS OF BACTERIAL GENOMES

STEP 01
Sequence Bacterial Genomes
Bactopia supports sequences from Illumina and Nanopore technologies, or a combination of the two.

STEP 02
Install Bactopia
Bactopia is easily installed using Conda and analyses can be run using Conda, Docker, or Singularity

STEP 03
Create a Sample Sheet
bactopia prepare is used to create a sample sheet allowing a single sample or thousands of samples to be processed

STEP 04
Include Public Samples
bactopia search can query ENA/SRA to identify accessions to further supplement analyses.

STEP 05
Choose Compute Environment
Built using Nextflow, Bactopia supports many compute and execution environments. Simple parameter changes, allow efficient switching between local, HPC, and major cloud environments.

BACTOPIA STEP 06
Process Genomes
Bactopia includes more than 150 bioinformatic tools in the pipeline, with extensive version control and data provenance. Upon completion outputs are available in a standardized structure for easy access.

STEP 08
Investigate Results
With analysis completed, numerous output files are available to assist in your genomic investigations

BACTOPIA TOOLS STEP 07
Additional Workflows
More than 60 Bactopia Tools allow seamless post-analysis of Bactopia outputs, including additional workflows for pangenome, taxonomic classification, phylogenies, AMR, among many more.

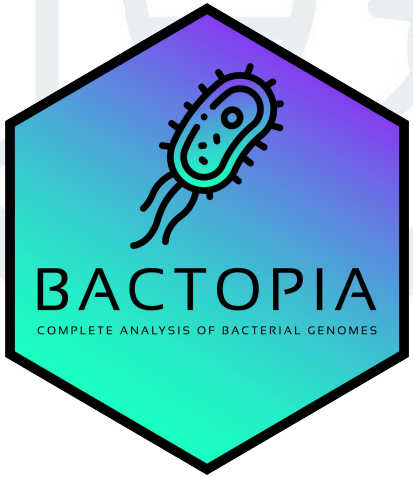
Step 7: Bactopia Tools provide additional workflows

60+ additional workflows for more science!

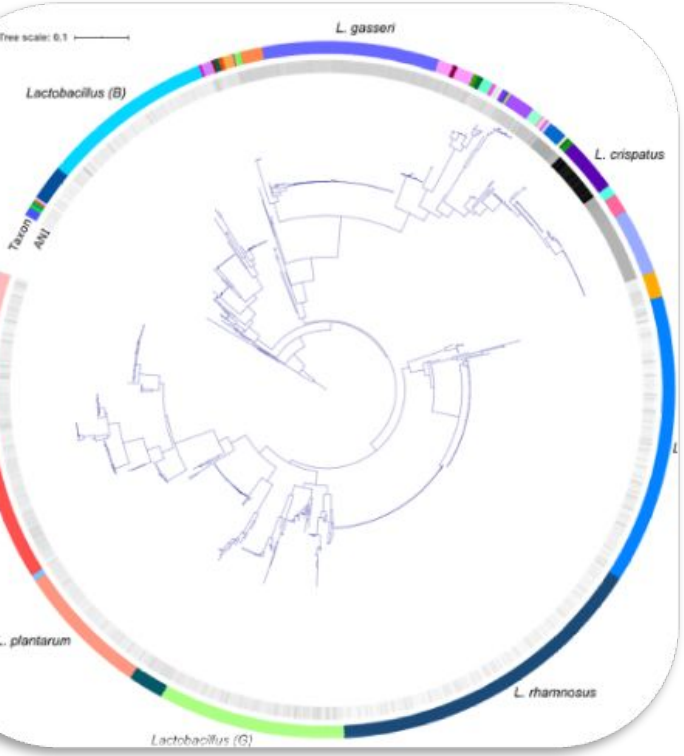
Utilizes the “standardized output structure” to automatically import required inputs

Includes workflows for:

- Pangenome, SNP & Indel, Scrubbing
- AMR, Species-Specific, Alignment
- Taxonomic Classification



Step 7: Bactopia Tools simplify complex tasks



For example, you can quickly generate a phylogeny based on a core-genome, core-snps, 16S rRNA, or sketches.

What is Bactopia?

Bactopia is an end-to-end Nextflow pipeline for the complete analysis of bacterial genomes. In a few simple steps, go from raw sequences to an extensive set of results to investigate.

BACTOPIA
COMPLETE ANALYSIS OF BACTERIAL GENOMES

STEP 01
Sequence Bacterial Genomes
Bactopia supports sequences from Illumina and Nanopore technologies, or a combination of the two.

STEP 02
Install Bactopia
Bactopia is easily installed using Conda and analyses can be run using Conda, Docker, or Singularity

STEP 03
Create a Sample Sheet
bactopia prepare is used to create a sample sheet allowing a single sample or thousands of samples to be processed

STEP 04
Include Public Samples
bactopia search can query ENA/SRA to identify accessions to further supplement analyses.

STEP 05
Choose Compute Environment
Built using Nextflow, Bactopia supports many compute and execution environments. Simple parameter changes, allow efficient switching between local, HPC, and major cloud environments.

BACTOPIA STEP 06
Process Genomes
Bactopia includes more than 150 bioinformatic tools in the pipeline, with extensive version control and data provenance. Upon completion outputs are available in a standardized structure for easy access.

BACTOPIA TOOLS STEP 07
Additional Workflows
More than 60 Bactopia Tools allow seamless post-analysis of Bactopia outputs, including additional workflows for pangenome, taxonomic classification, phylogenies, AMR, among many more.

STEP 08
Investigate Results
With analysis completed, numerous output files are available to assist in your genomic investigations

Bactopia allows its users to more time and effort into what's important, investigating the outputs and drawing conclusions.

Wyoming Department of Health | PUBLIC HEALTH DIVISION | WYOMING PUBLIC HEALTH LABORATORY

What is Bactopia?

Bactopia is an end-to-end Nextflow pipeline for the complete analysis of bacterial genomes. In a few simple steps, go from raw sequences to an extensive set of results to investigate.

BACTOPIA
COMPLETE ANALYSIS OF BACTERIAL GENOMES

STEP 01
Sequence Bacterial Genomes
Bactopia supports sequences from Illumina and Nanopore technologies, or a combination of the two.

STEP 02
Install Bactopia
Bactopia is easily installed using Conda and analyses can be run using Conda, Docker, or Singularity

STEP 03
Create a Sample Sheet
bactopia prepare is used to create a sample sheet allowing a single sample or thousands of samples to be processed

STEP 04
Include Public Samples
bactopia search can query ENA/SRA to identify accessions to further supplement analyses.

STEP 05
Choose Compute Environment
Built using Nextflow, Bactopia supports many compute and execution environments. Simple parameter changes, allow efficient switching between local, HPC, and major cloud environments.

BACTOPIA STEP 06
Process Genomes
Bactopia includes more than 150 bioinformatic tools in the pipeline, with extensive version control and data provenance. Upon completion outputs are available in a standardized structure for easy access.

STEP 07
Additional Workflows
BACTOPIA TOOLS
More than 60 Bactopia Tools allow seamless post-analysis of Bactopia outputs, including additional workflows for pangenome, taxonomic classification, phylogenies, AMR, among many more.

STEP 08
Investigate Results
With analysis completed, numerous output files are available to assist in your genomic investigations

In just a few steps you can utilize, Bactopia, a complete and extensive reproducible, portable, and accessible pipeline for bacterial genome analysis

Wyoming Department of Health | PUBLIC HEALTH DIVISION | WYOMING PUBLIC HEALTH LABORATORY

Let's Wrap This Up

Some people behind Bactopia, who's using it, and what's next

People behind Bactopia

Let’s put some faces to Bactopia

People behind the scenes of Bactopia

Tim Read, PhD

Emory University, Professor

Tim has played a role in Bactopia since its inception. Through the years Tim has provided feedback and ideas to help shape Bactopia.

Joseph Reed, PhD

WPHL, Laboratory Administrator

As the WPHL lab administrator, Joe encourages the lab to pursue the development of skills and tools like Bactopia to strengthen WPHL.

Jim Mildenberger

WPHL, Molecular Lab Supervisor

Jim keeps the molecular lab running. Like Joe, Jim's support has helped introduce many new features (e.g., ONT support) into Bactopia.

Taylor Fearing

WPHL, EID & NGS Supervisor

Taylor oversees the sequencing at WPHL. She has played a tremendous and critical role in helping to expand Bactopia into public health.

Chayse Rowley

WPHL, Senior Microbiologist

Chayse is overseeing new sequencing projects at WPHL. She has helped identify novel ways to use Bactopia within its existing framework.

People using Bactopia

Let’s get an idea of who’s using Bactopia

80+ citations, 369 GitHub Stars

Used in all sorts of bacterial genomic studies

45,000+ unique visitors

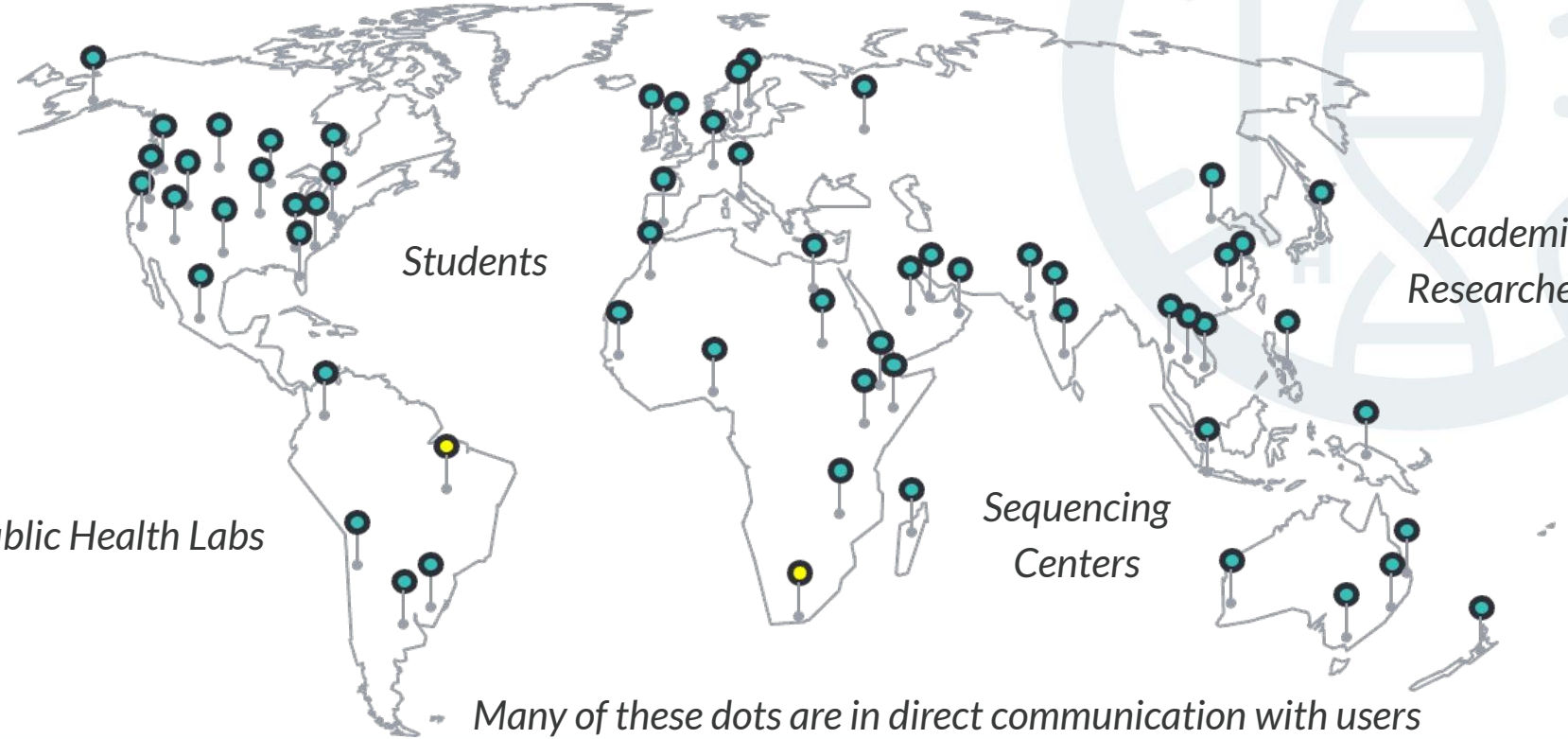
Many users from around the world are visiting the docs

1,000,000+ downloads

That's nearly twice as many people here in Wyoming!

Wyoming Department of Health | PUBLIC HEALTH DIVISION | WYOMING PUBLIC HEALTH LABORATORY

Bactopia users across the globe

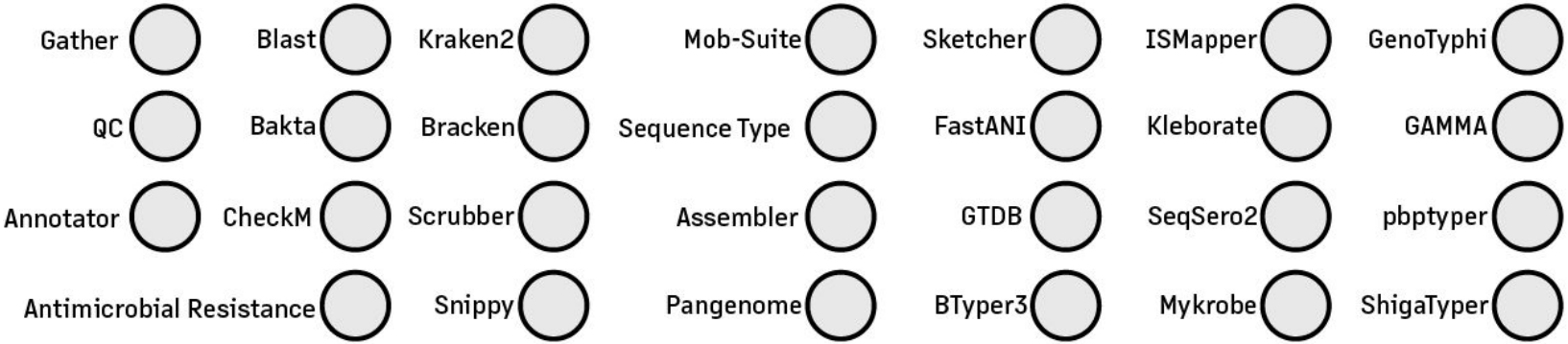


Many of these dots are in direct communication with users

Future directions

What's next for Bactopia

All modules are “plug and play”



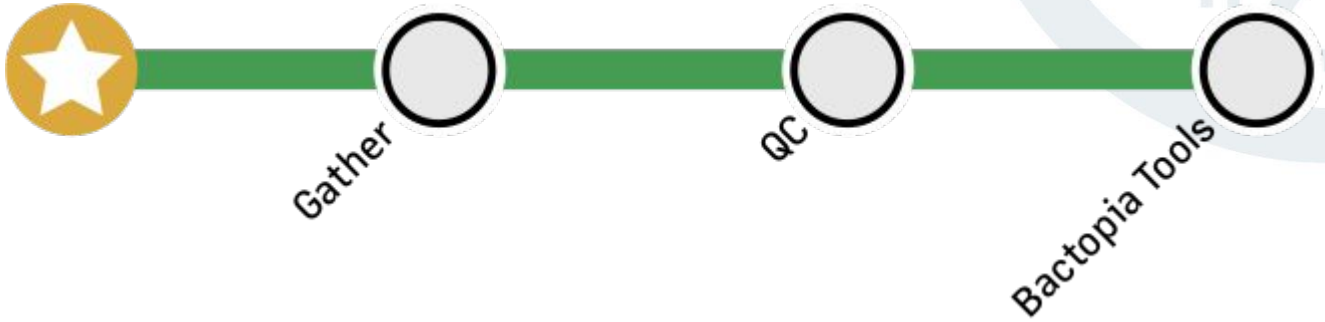
… another 30+ modules not listed

These can all be treated as building blocks

This allows Bactopia to be used as a framework, to rapidly adapt it to user needs, ***without having to learn a new workflow.***

clean-yer-reads *for access to Bactopia Tools*

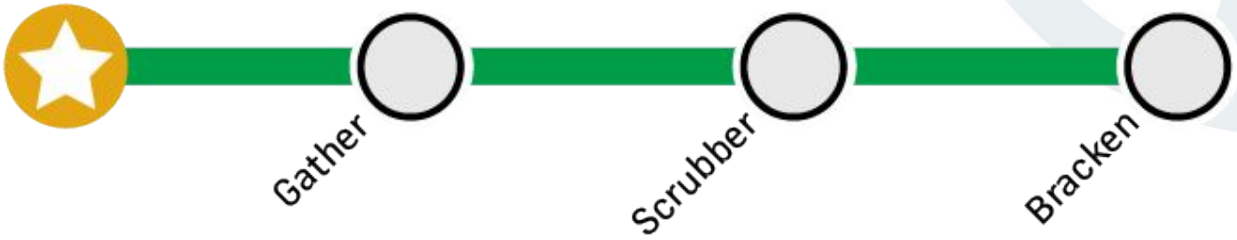
clean-yer-reads is a simple reshaping of Bactopia that enables you to QC sequences across any organism, including metagenomic samples.



With *clean-yer-reads* you still have full access to all available Bactopia Tools.

teton *for scrubbing and taxonomic classification*

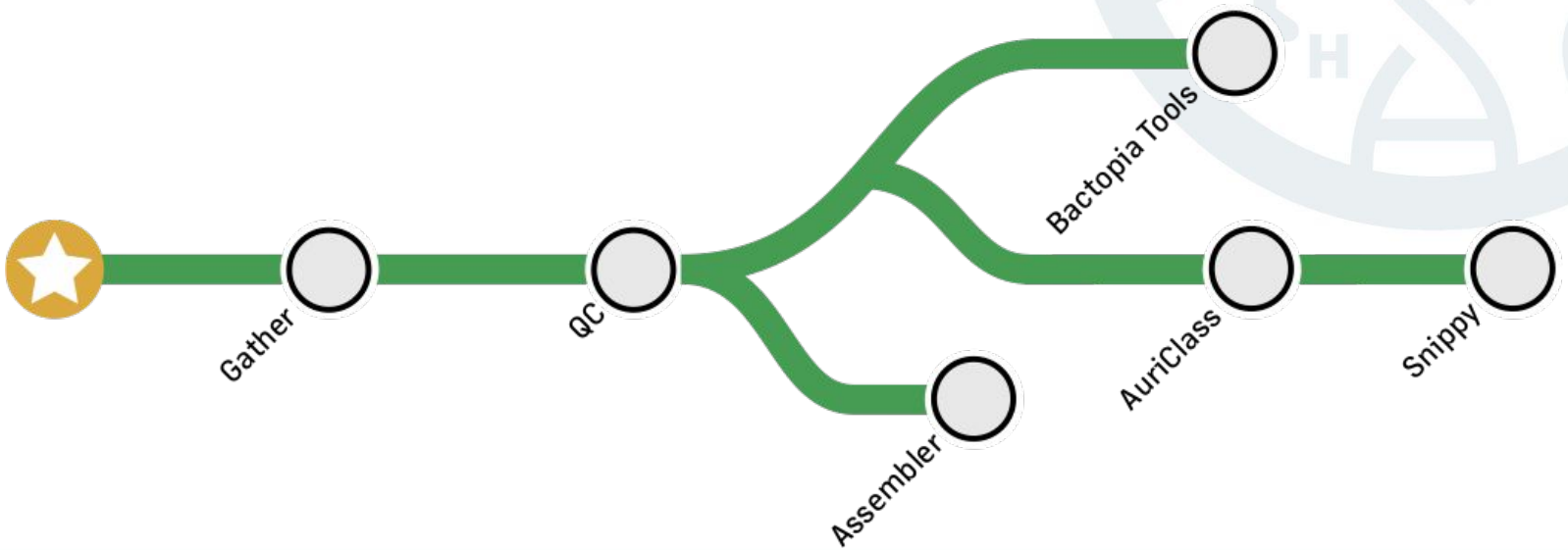
teton is a simple reshaping of Bactopia that enables you scrub human reads and conduct taxonomic classification of your sample.



teton makes for a good first processing of all new sequencing

mycotopia *for fungal genomes*

mycotopia represents a specialized adaptation of Bactopia for fungal organisms, including the notable fungal pathogen *Candida auris*.



Many more things are brewing

Another 5+ years of support, as I will be at WPHL (🦬) for at least 5 more years

Visual reports for all results including each Bactopia Tool

Training materials to not only using Bactopia but also developing it

A data explorer system built on top of Bactopia

Integration of genomic epidemiology tools

Additional workflows and Bactopia Tools

Everything we make will continue to be free and open-source

Development is heavily influenced by user feedback through direct communication and surveys. So, don't hesitate to provide feedback!

Final (*final*) Wrap Up

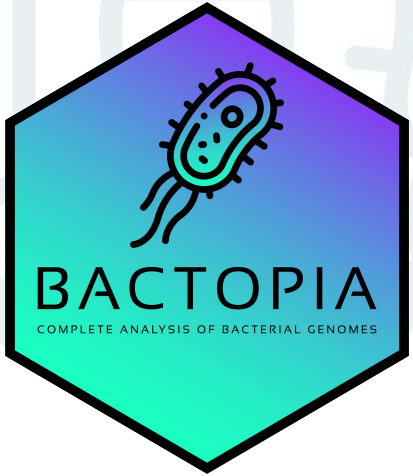
For real this time!

Bactopia streamlines bacterial genome analysis

Bactopia is an extensive pipeline that allows its users to focus more on the science behind their samples.

It will continue to sustained for 5+ years through support from WPHL, Emory University, CAPE, (*and soon another one!*)

There are many around the world utilizing Bactopia, and we are still only scratching the surface of where we plan to take it.



Wyoming Department of Health | PUBLIC HEALTH DIVISION

EMORY UNIVERSITY SCHOOL OF MEDICINE

cape Center for Applied Pathogen Epidemiology and Outbreak Control

Acknowledgements

The many developers of open source software and the users of Bactopia that are regularly providing feedback.

WPHL
Taylor Fearing
Chayse Rowley
Jim Mildenberger
Rob Christensen
Joseph Reed

Emory
Tim Read
EMERGENT Group
Ahmed Babiker

Thank you! Happy to take any questions

Learn more about Bactopia at
bactopia.github.io

Wyoming Department of Health | PUBLIC HEALTH DIVISION | WYOMING PUBLIC HEALTH LABORATORY