

# BACTOPIA

Using Bactopia for the complete analysis of bacterial genomes

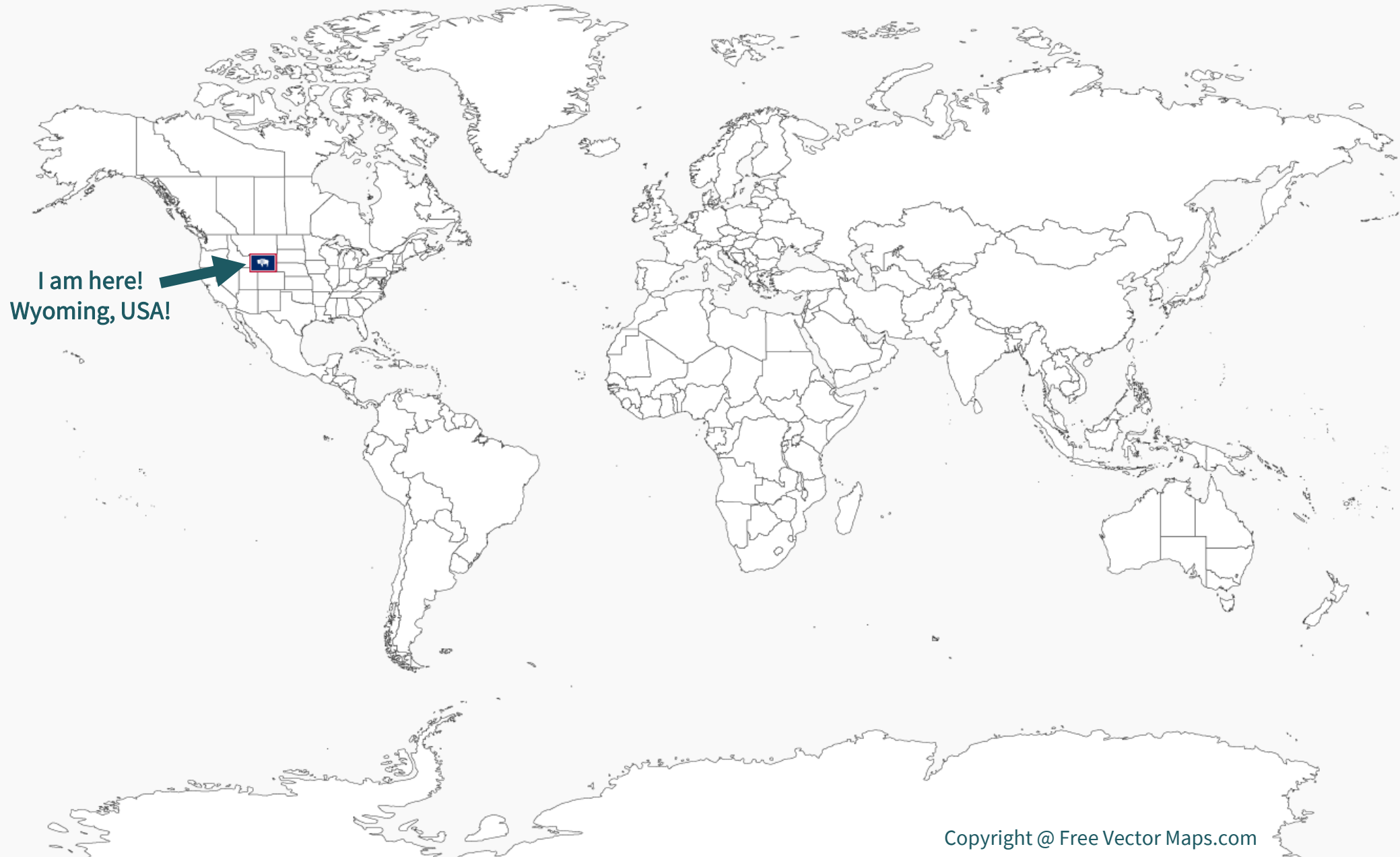
April 28<sup>th</sup>, 2022

Robert A. Petit III, PhD

# Overview

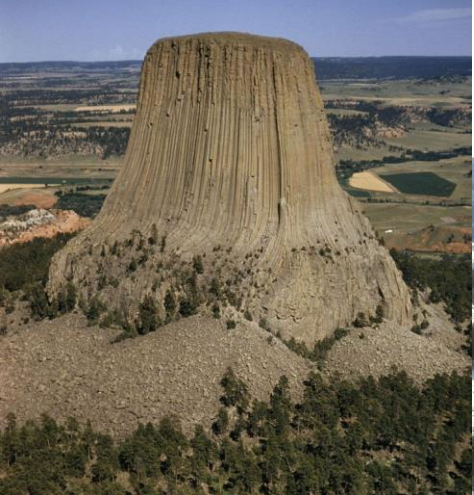
- Background and Motivations
- Learn about Bactopia
- Use Case: Using Bactopia to describe public Lactobacillus genomes
- Future Directions

# Where in the world am I?





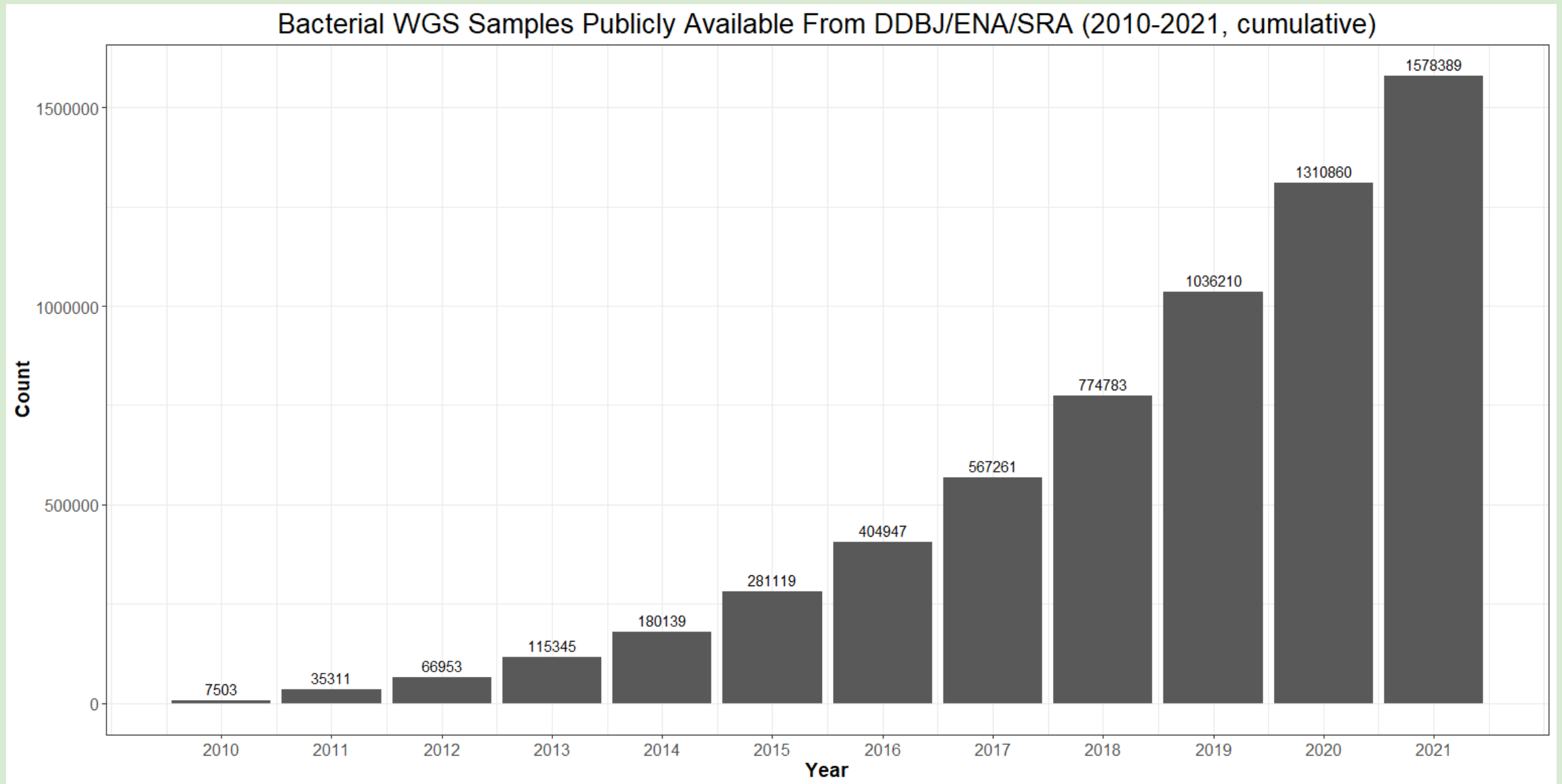
# Wyoming has a lot to see



# Background and Motivations

Bacterial genomics is a rapidly evolving field

# Rapid growth of bacterial WGS in the last 10 years

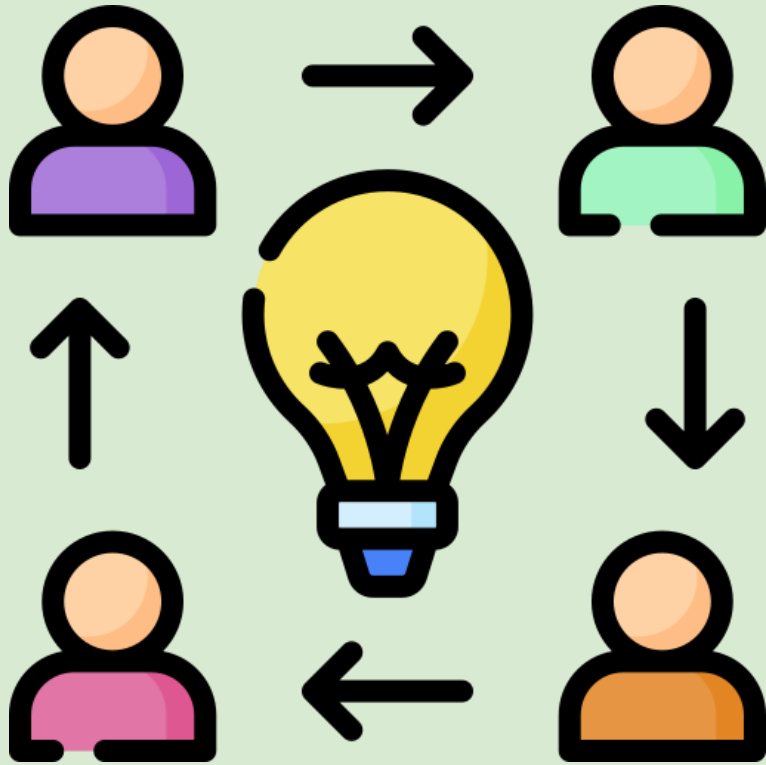




# Rapid growth in the bioinformatics as well

Conda, containerization, workflow managers, etc...

# BIOCONDA<sup>®</sup> for all your bioinformatic tools



- Makes bioinformatics accessible
  - Easy installs, dependency handling
- Downstream containerization
  - Docker – [Biocontainers](#)
  - Singularity Images - [Galaxy Project](#)
- Truly a community driven repository
  - More than 1,300 people have contributed
- Currently ~4,000 recipes are available



# Workflow managers make bioinformatics manageable

- Manages the execution of pipelines
  - Linking inputs/outputs of bioinformatic tools
  - Queuing jobs locally, on clusters, or the cloud
  - Logging, errors, audit trails
- Promote reproducible and reusable science
- Common workflow managers:
  - [Nextflow](#), [WDL](#), and [Snakemake](#)
- Pick one that works for you

nextflow



# nextflow

- A popular workflow manager in Bioinformatics
- Enables scalable and reproducible pipelines
- Supports Conda, Docker, and Singularity
- Seamlessly move between local resources, HPC, and major cloud providers
- Regularly solicits user feedback to guide future developments

```
nextflow.enable.dsl=2

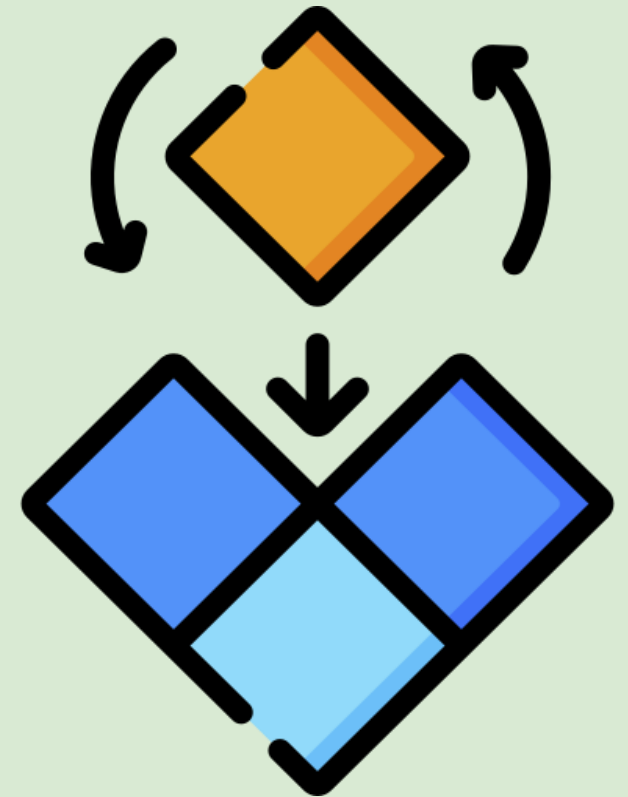
process sayHello {
    input:
    val cheers
    output:
    stdout

    """
    echo $cheers
    """
}

workflow {
    channel.of('Ciao','Hello','Hola') | sayHello | view
}
```

# nextflow DSL2

- Major evolution in the Nextflow language
- Introduced true modularization in Nextflow workflows
  - Modules – A reusable Nextflow script with a process definition
  - Subworkflows – Multiple modules linked together
- Modules are portable and easily shared between workflows
- Data channels can be used more than once



# nf-core 🍏 pushing Nextflow to the limits



- Community effort to collect curated Nextflow pipelines
  - 2400 Slack users, 1000+ GitHub contributors
- Includes 60+ hi-quality bioinformatic pipelines
  - [rnaseq](#), [mag](#), [bactmap](#), [many more](#)
- [nf-core/modules](#) has 200+ DSL2 modules available
- Standardized [guidelines](#) for developers
- Thorough review process produces robust pipelines

# "Bring your own workflow" Platforms

- Freely available web-platforms for the execution of bioinformatic pipelines
- No command-line knowledge required, allowing users to do more science

- Examples:

- [Nextflow Tower](#) from [Seqera Labs](#)

- Supports workflows written in Nextflow
    - Platform agnostic and supports many providers
      - HPC, Google Cloud, Microsoft Azure, Amazon Web Services
    - Community showcase of curated pipelines

nextflow tower



- [Terra](#) from the [Broad Institute](#)

- Supports workflows written in WDL
    - Limited to Google Cloud Platform
      - Microsoft Azure support in the works
    - Import workflows from [Dockstore](#)

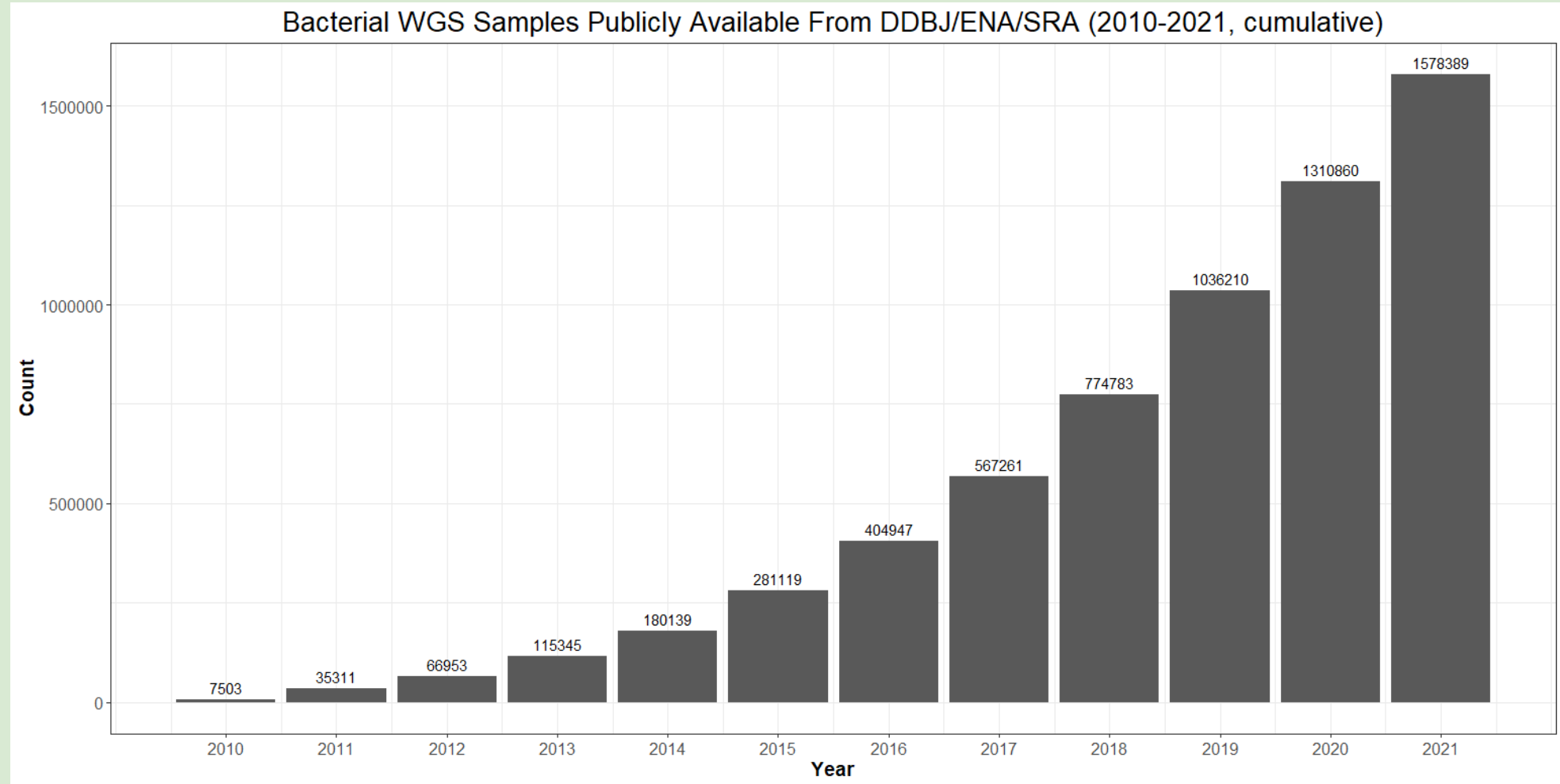


- [CGC](#) from [Seven Bridges](#)

- Supports workflows written in CWL, Nextflow and WDL
    - Limited to Amazon Web Services

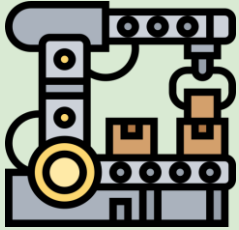


# With the bioinformatic advances, it's hard to ignore public WGS



## How can we make use of all this public WGS data?

# To take advantage of public data we need:



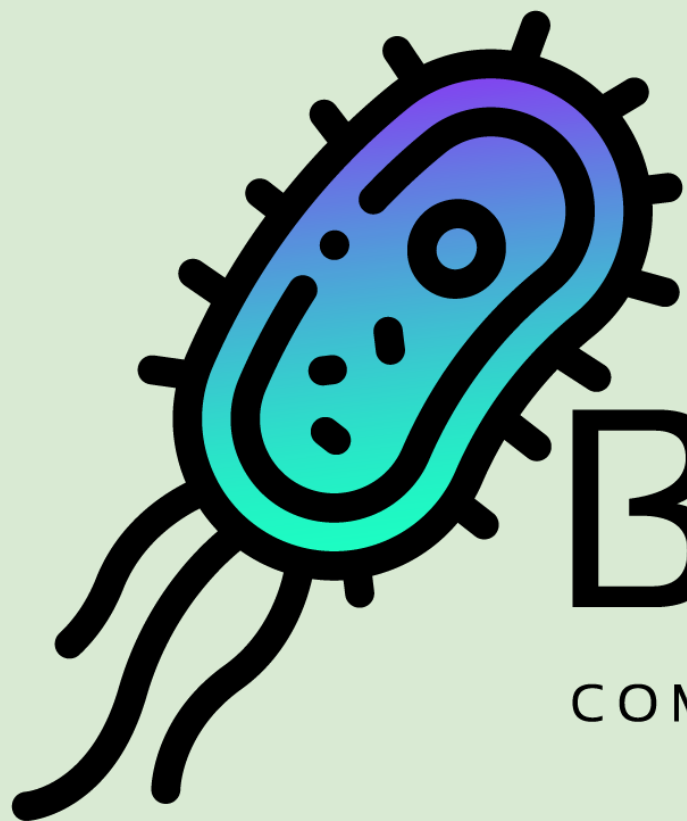
- Ensure each sample is processed the same way
  - Allows comparison of your data with public data



- Scalable pipeline to process 1 or 1,000s of samples
  - Should also be easy to use, reproducible, and portable



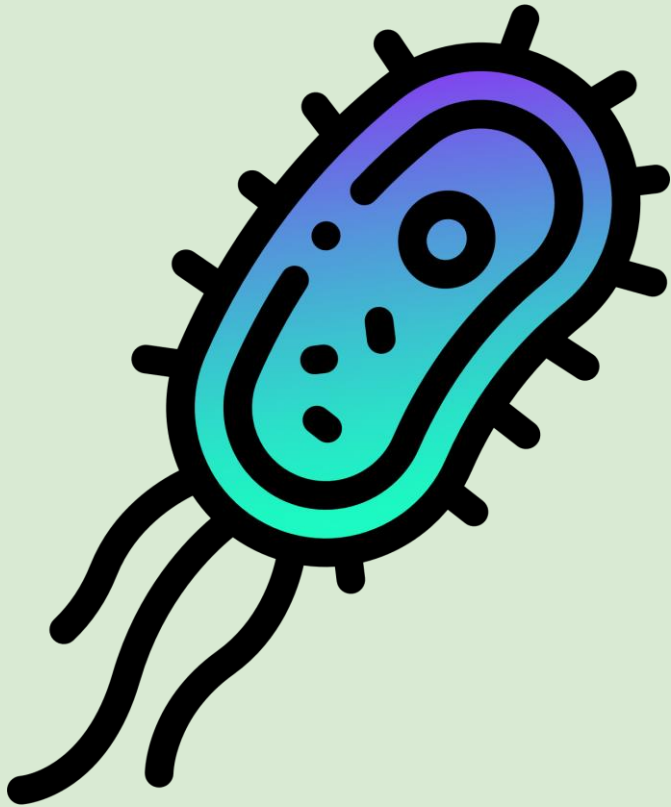
- Account for poor quality data
  - Example: FASTQs missing pairs, mislabeled species



# BACTOPIA

COMPLETE ANALYSIS OF BACTERIAL GENOMES

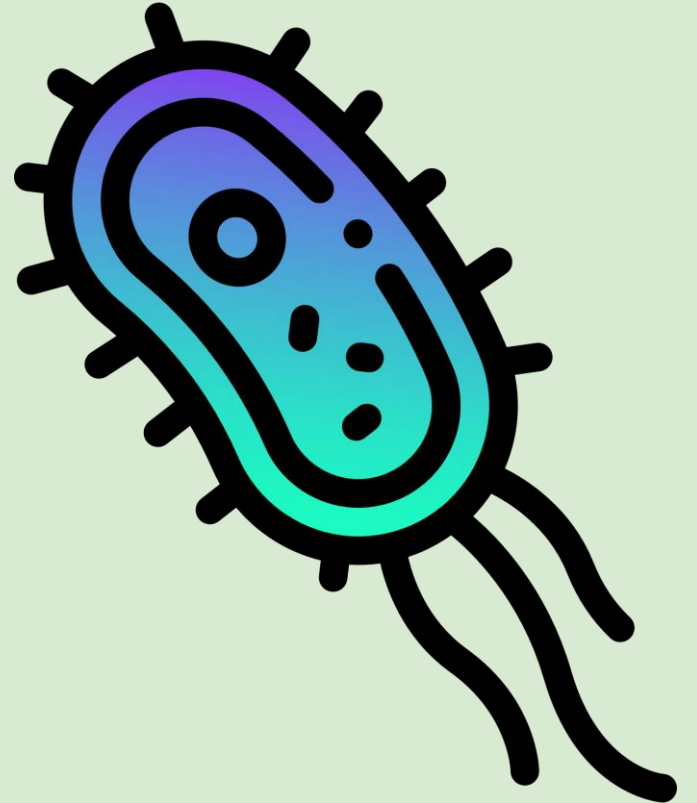
# Bactopia is an all-in-one workflow



- Bactopia is built using Nextflow DSL2 and available on Bioconda
- Bactopia is scalable
  - Process a few samples to 10s of thousands
- Bactopia is reproducible
  - Version control and audit trails
- Bactopia is portable
  - Move between a laptop to the cloud with ease

# A few Bactopia highlights

- Supports Illumina and Nanopore reads
  - Local or from public databases
- Includes more than [130 bioinformatic tools](#)
- [30+ Bactopia Tools](#) provide more workflows for more science
- Extensively tested with 100+ tests for 10,000+ output files
- Available on Bioconda, Docker, and Singularity
- Well documented at [bactopia.github.io](https://bactopia.github.io)





# Bactopia design principles



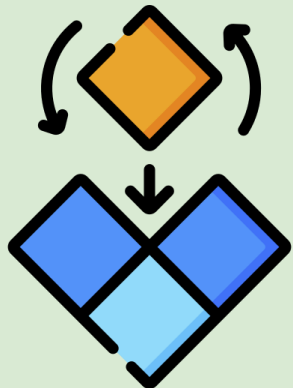
Bactopia requires tools be available from Bioconda

- Easy to install, downstream Docker and Singularity containers



All modules used by Bactopia Tools must be on nf-core/modules

- Developers can use these modules in their own workflows

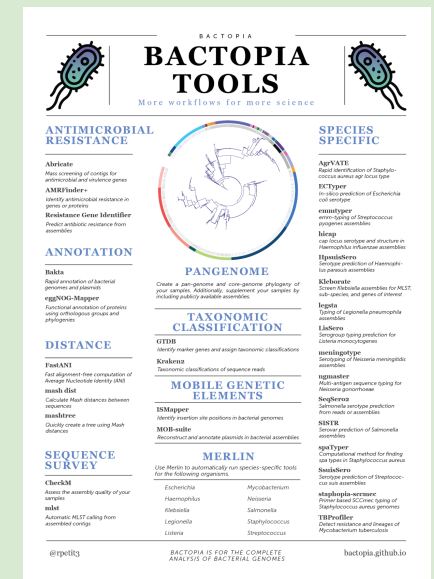


Bactopia should be flexible and easily adapted to fit user needs

- Nextflow DSL2 has made this much easier

## A stylized illustration of a toolbox. The toolbox is orange with a black outline and a black handle. Inside the toolbox, there are three tools: a hammer with a light blue head and a green handle, a screwdriver with an orange handle, and a wrench with a light blue head. The tools are arranged vertically, with their heads pointing upwards. The background is a solid light green color.

# Bactopia



# Bactopia Tools

# Bactopia Helpers, help get started using Bactopia



- bactopia citations
  - Print citations for all datasets, tools and Bactopia
- bactopia datasets
  - Download and setup useful datasets for Bactopia
- bactopia download
  - Build Conda, Docker or Singularity environments for all steps in bactopia
- bactopia prepare
  - Create a file of filenames for analysis
- bactopia search
  - Generate a list of SRA accessions for analysis

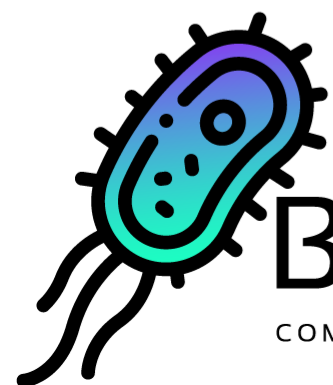
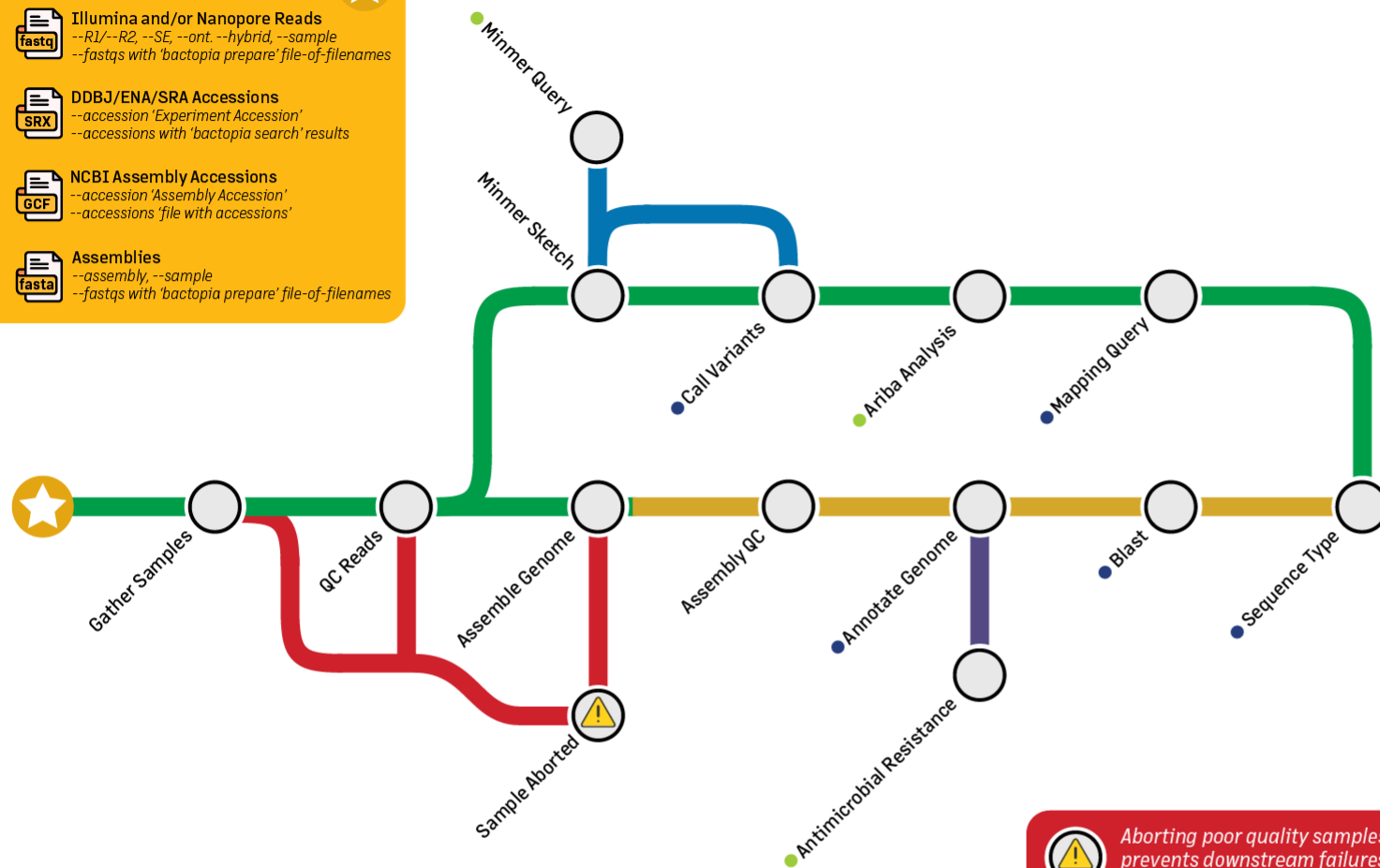
# Bactopia Datasets supplement analyses

- Makes including public datasets is easy, while bringing a lot of value to your analysis
- Includes:
  - Sketches for all RefSeq and GenBank
  - Reference genomes
  - Custom proteins for annotation
  - MLST schemas
  - Ariba datasets
  - Genes, proteins and primers
- Bactopia datasets handles all this for you



**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



# BACTOPIA

COMPLETE ANALYSIS OF BACTERIAL GENOMES



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**

- 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

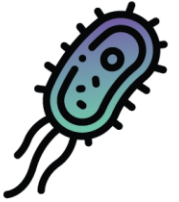
### Sequence Type

Determine sequence type base on PubMLST profiles



# Bactopia Tools – More workflows for more science

- Allow easy comparative analysis of Bactopia outputs
- Two Types:
  - Modules: single tool (Kleborate, TB Profiler)
  - Subworkflows: multiple tools connected together
    - pangenome: Prokka -> PIRATE -> IQ-Tree
- 30+ Bactopia Tools are available
  - Frame-worked for easy addition
  - Simple command to create new Bactopia Tool
- Because of Nextflow DSL2, every Bactopia Tool can be reused



# BACTOPIA TOOLS

More workflows for more science

## ANTIMICROBIAL RESISTANCE

**Abricate**  
Mass screening of contigs for antimicrobial and virulence genes

**AMRFinder+**  
Identify antimicrobial resistance in genes or proteins

**Resistance Gene Identifier**  
Predict antibiotic resistance from assemblies

## ANNOTATION

**Bakta**  
Rapid annotation of bacterial genomes and plasmids

**eggNOG-Mapper**  
Functional annotation of proteins using orthologous groups and phylogenies

## DISTANCE

**FastANI**  
Fast alignment-free computation of Average Nucleotide Identity (ANI)

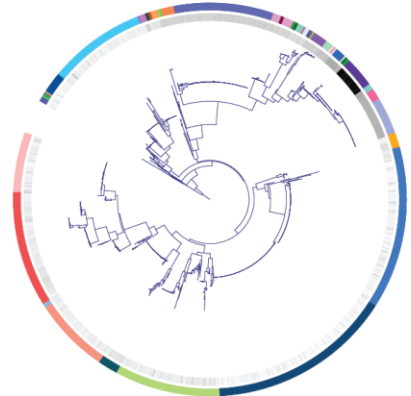
**mash dist**  
Calculate Mash distances between sequences

**mashtree**  
Quickly create a tree using Mash distances

## SEQUENCE SURVEY

**CheckM**  
Assess the assembly quality of your samples

**mlst**  
Automatic MLST calling from assembled contigs



## PANGENOME

Create a pan-genome and core-genome phylogeny of your samples. Additionally, supplement your samples by including publicly available assemblies.

## TAXONOMIC CLASSIFICATION

**GTDB**  
Identify marker genes and assign taxonomic classifications

**Kraken2**  
Taxonomic classifications of sequence reads

## MOBILE GENETIC ELEMENTS

**ISMMapper**  
Identify insertion site positions in bacterial genomes

**MOB-suite**  
Reconstruct and annotate plasmids in bacterial assemblies

## MERLIN

Use Merlin to automatically run species-specific tools for the following organisms.

Escherichia	Mycobacterium
Haemophilus	Neisseria
Klebsiella	Salmonella
Legionella	Staphylococcus
Listeria	Streptococcus

## SPECIES SPECIFIC

**AgrVATE**  
Rapid identification of *Staphylococcus aureus* agr locus type

**ECTyper**  
In-silico prediction of *Escherichia coli* serotype

**emmtypier**  
emm-typing of *Streptococcus pyogenes* assemblies

**hicap**  
cap locus serotype and structure in *Haemophilus influenzae* assemblies

**HpsuisSero**  
Serotype prediction of *Haemophilus parasuis* assemblies

**Kleborate**  
Screen *Klebsiella* assemblies for MLST, sub-species, and genes of interest

**legsta**  
Typing of *Legionella pneumophila* assemblies

**LisSero**  
Serogroup typing prediction for *Listeria monocytogenes*

**meningotype**  
Serotyping of *Neisseria meningitidis* assemblies

**ngmaster**  
Multi-antigen sequence typing for *Neisseria gonorrhoeae*

**SeqSero2**  
Salmonella serotype prediction from reads or assemblies

**SISTR**  
Serovar prediction of *Salmonella* assemblies

**spaTyper**  
Computational method for finding spa types in *Staphylococcus aureus*

**SsuisSero**  
Serotype prediction of *Streptococcus suis* assemblies

**staphopia-sccmec**  
Primer based SCCmec typing of *Staphylococcus aureus* genomes

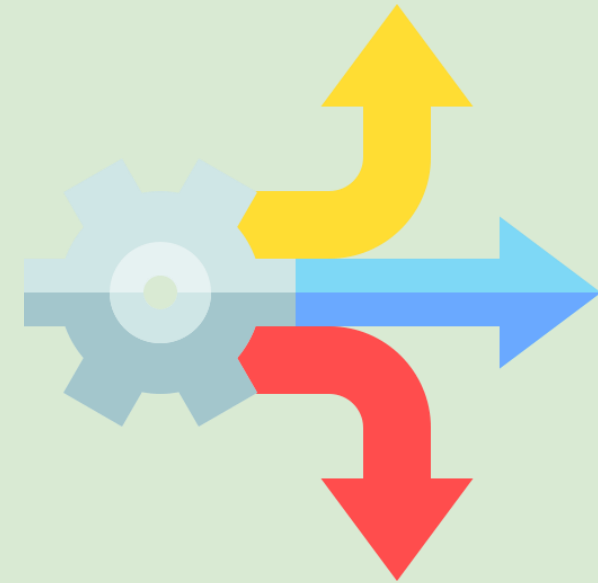
**TBProfiler**  
Detect resistance and lineages of *Mycobacterium tuberculosis*

**Because of DSL2, every Bactopia Tool can be reused!**

# Demonstration of Bactopia Tool reusability

- There are 15+ species-specific Bactopia Tools
  - Sometimes the species is known
  - Sometimes the species is unknown

*Can we automatically execute these species-specific tools?*

**AgrVATE**

Rapid identification of *Staphylococcus aureus* agr locus type

**ECTyper**

In-silico prediction of *Escherichia coli* serotype

**emmtyper**

emm-typing of *Streptococcus pyogenes* assemblies

**hicap**

cap locus serotype and structure in *Haemophilus influenzae* assemblies

**HpsuisSero**

Serotype prediction of *Haemophilus parasuis* assemblies

**Kleborate**

Screen *Klebsiella* assemblies for MLST, sub-species, and genes of interest

**legsta**

Typing of *Legionella pneumophila* assemblies

**LisSero**

Serogroup typing prediction for *Listeria monocytogenes*

**meningotype**

Serotyping of *Neisseria meningitidis* assemblies

**ngmaster**

Multi-antigen sequence typing for *Neisseria gonorrhoeae*

**SeqSero2**

*Salmonella* serotype prediction from reads or assemblies

**SISTR**

Serovar prediction of *Salmonella* assemblies

**spaTyper**

Computational method for finding spa types in *Staphylococcus aureus*

**SsuisSero**

Serotype prediction of *Streptococcus suis* assemblies

**staphopia-sccmec**

Primer based SCCmec typing of *Staphylococcus aureus* genomes

**TBProfiler**

Detect resistance and lineages of *Mycobacterium tuberculosis*

# Merlin - MinmER assisted species-specific bactopia tool seLectioN

- Mash distances select species-specific tools
  - Re-uses the RefSeq Mash Sketch downloaded by Bactopia Datasets
- Currently includes members of 10+ genera
- Highlights module reuse and condition-based execution



## MERLIN

Use Merlin to automatically run species-specific tools for the following organisms.

*Escherichia*

*Mycobacterium*

*Haemophilus*

*Neisseria*

*Klebsiella*

*Salmonella*

*Legionella*

*Staphylococcus*

*Listeria*

*Streptococcus*

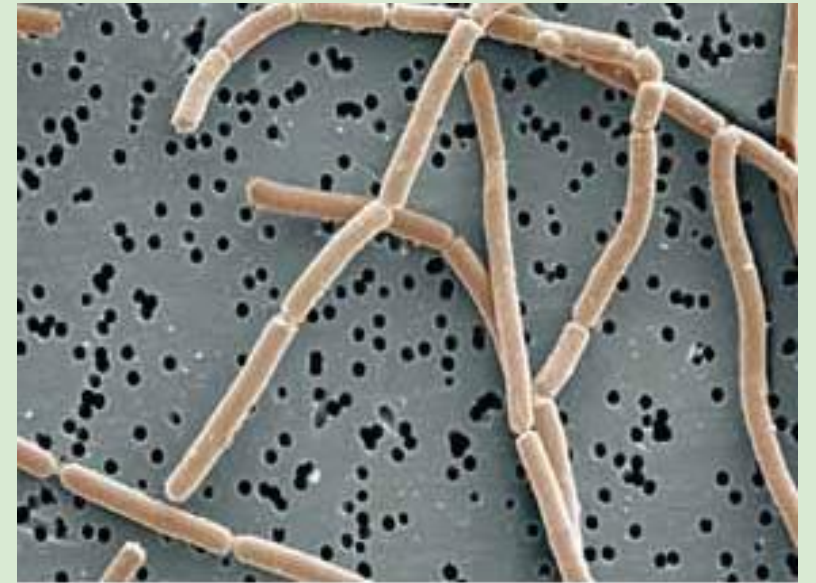
# Describing public *Lactobacillus* genomes

In just a few commands using Bactopia



# Use Case: Investigating Lactobacillus

- Gram-positive, rod-shaped bacterium
- Common in human and animal microbiota
  - 40+ species adapted to microbiome hosts
  - Gastrointestinal and vaginal
  - Inhibits growth of other bacteria
- Economic uses include food production and probiotics
- We can use Bactopia to analyze Lactobacillus genomes in just a few commands



*Lactobacillus bulgaricus*  
From [Utah State University](#)

# Bactopia Helpers to get things setup



- Build public datasets with "*bactopia datasets*"
  - Downloads general and Lactobacillus-specific datasets
  - Ex. *bactopia datasets -species "lactobacillus"*
- "*bactopia search*" to identify publicly available WGS of Lactobacillus
  - Creates a list of accessions for Bactopia to process
  - Ex. *bactopia search 1578*

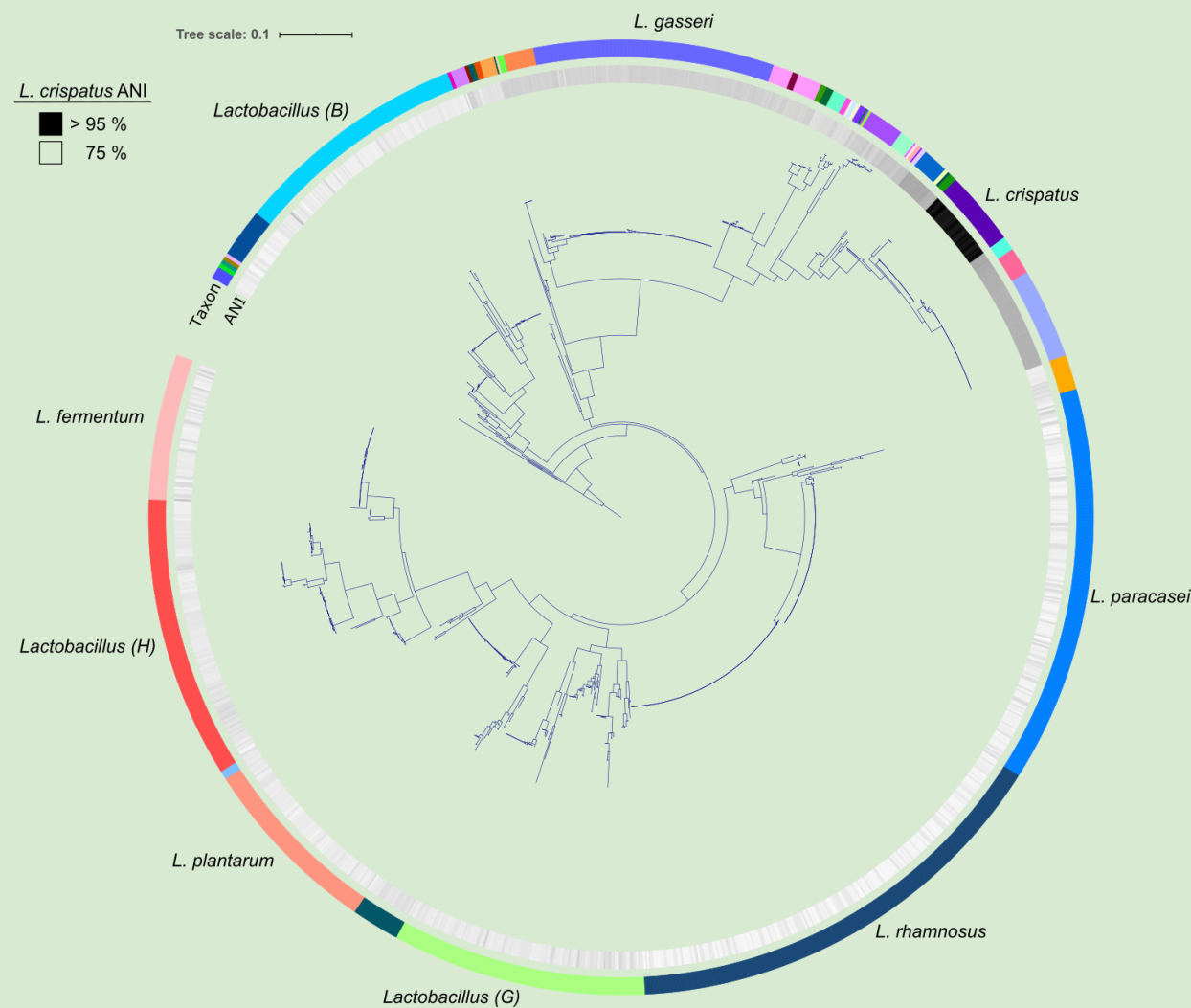
# Use Case: Process Lactobacillus WGS

- Processed 1,664 genomes with Bactopia
  - Bactopia handles downloading from SRA/ENA
- Most of the samples were good quality

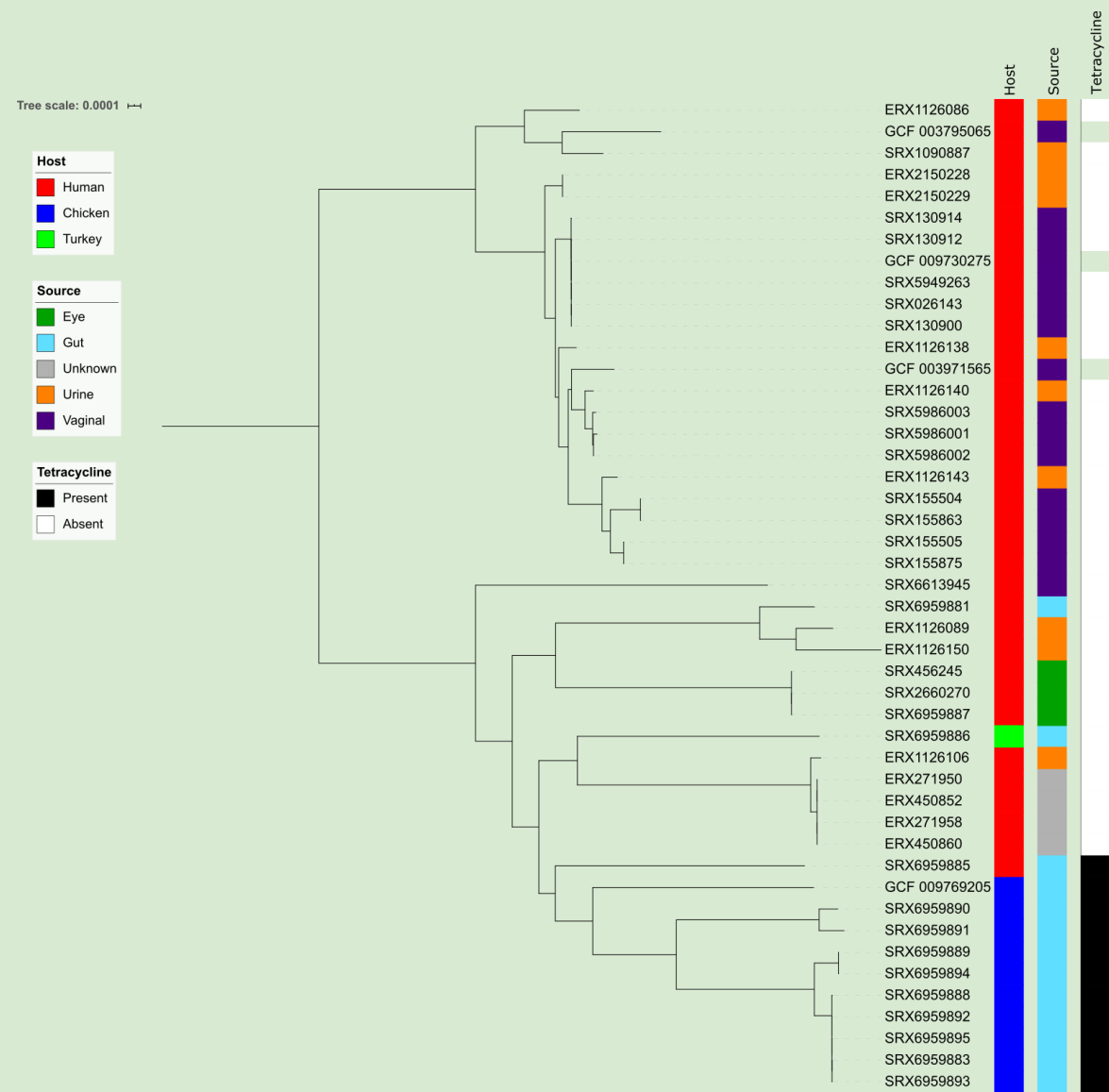
Quality rank	No. of samples	Original coverage	Post-Bactopia coverage	Per-read quality score	Read length (bp)	Contig count
Gold	967	213×	100×	Q35	100	52
Silver	386	160×	100×	Q35	100	110
Bronze	205	102×	100×	Q34	100	90
Exclude	48	26×	22×	Q34	100	706
Unprocessed	58					

# Use Case: Lactobacillus 16S Tree

- phyloflash and gtdb Bactopia Tools
  - [PhyloFlash](#) for 16S construction
  - [GTDB](#) for taxonomic classification
- 5 species represented 45% of the samples
- 58 samples were not Lactobacillus



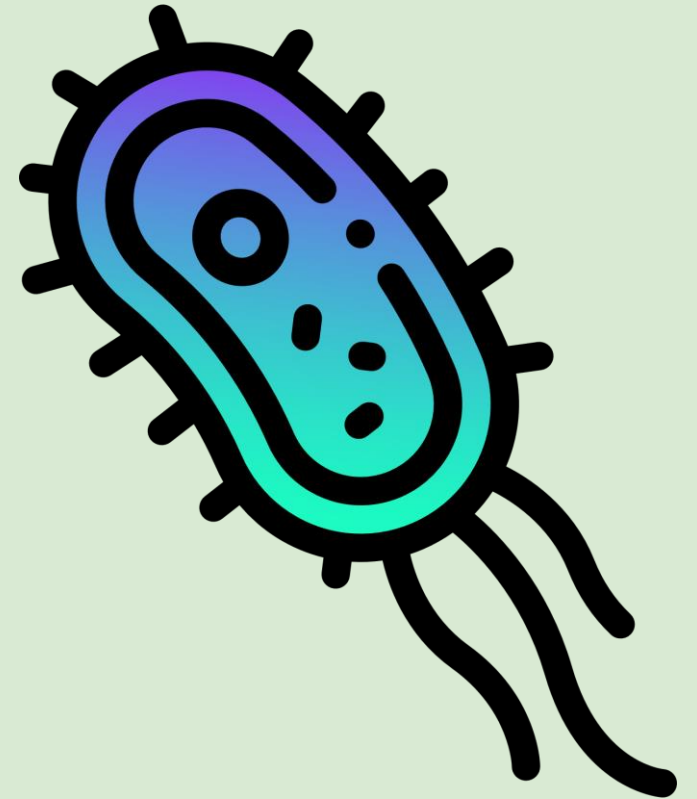
# Use Case: Closer Look at *L. crispatus*



- fastani and pangenome Bactopia Tools
  - [FastANI](#) to identify *L. crispatus*
  - [Roary](#) to create core-genome alignment
  - [IQ-TREE](#) for core-genome phylogeny
- 38 *L. crispatus* samples clustered into two groups
  - Human vaginal samples
  - Chicken, turkey, human gut

# Use Case: A few Bactopia commands from start to finish

- With just a few commands we analyzed 1,600 public WGS genomes
  - Built general and Lactobacillus specific datasets
    - *bactopia datasets*
  - Identified all publicly available Lactobacillus genomes
    - *bactopia search*
  - Processed all available Lactobacillus genomes
    - *Bactopia*
  - Taxonomic classification and core-genome phylogeny
    - *Bactopia Tools*
    - PhyloFlash, GTDB, FastANI, Roary, IQ-TREE



# Additional Bactopia highlights

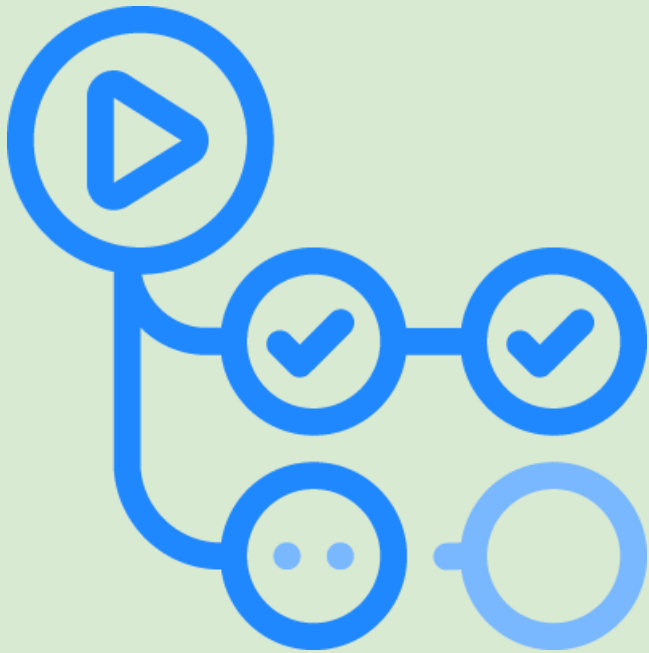
# Bactopia as an introduction to bacterial genomics

- More than [130 bioinformatic tools](#) are used by Bactopia
  - Description, citation, and code links are provided for each tool
- Bactopia is easily [installed through Bioconda](#)
- Bactopia is well documented
  - [Detailed pipeline overview](#)
  - [Tutorial to get started](#)
  - [Output file descriptions](#)





# Tests keep the wheels turning



- Every step is tested in Bactopia using real data
  - Usages, workflows, subworkflows, modules
- More than 10,000 output files are tested
  - MD5sums match, contains strings, file exists
- Testing conducted by pytest
  - Adapted from nf-core/modules
- All automated through GitHub Actions

# By using Nextflow, Bactopia is platform independent

- Execute on a laptop or the cloud, with a simple parameter change
- Available on:
  - Bioconda, Docker, Singularity
  - Google Cloud Platform
  - Amazon Web Services
  - Microsoft Azure
  - HPC
- Executable from:
  - Nextflow Tower
  - Terra.bio



# Bactopia contributes back to the community



Bactopia requires tools be available from Bioconda

- Submitted 20 recipes, reviewed 1300+ Bioconda pull requests



All Bactopia Tools must be available from nf-core/modules

- Submitted 30+ modules to nf-core/modules



Multiple PRs submitted, and issues resolved

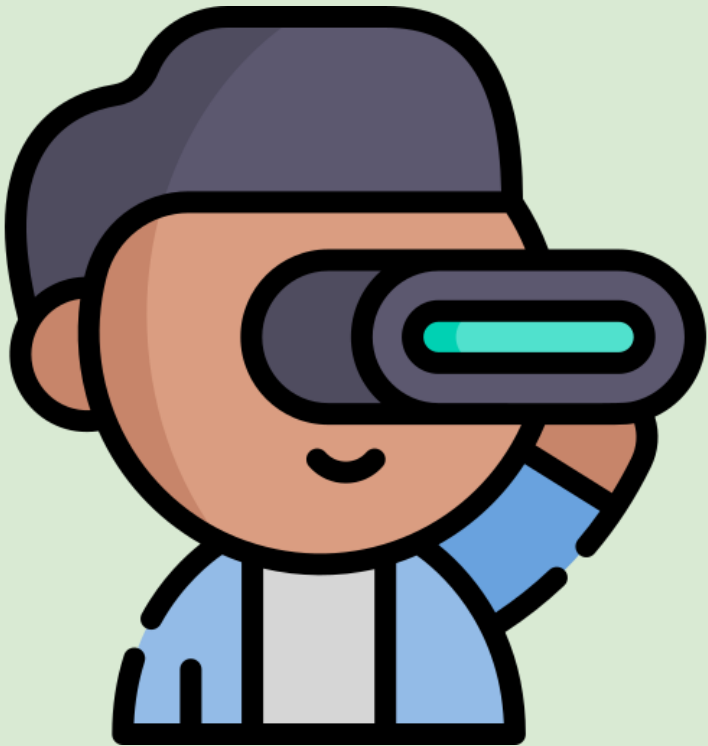
- Bactopia user finds bug in “X” tool, I attempt to fix and submit a PR
- Bowtie2, Prokka, Shovill, PIRATE, Ariba, ISMapper, PhyloFlash, etc....

# Bactopia has spawned stand-alone tools

- [assembly-scan](#) – Quickly generate basic stats for an assembly in JSON format
- [dragonflye](#) - Assemble bacterial isolate genomes from Nanopore reads
- [fastq-dl](#) - Download FASTQ files from SRA or ENA repositories
- [fastq-scan](#) - Output FASTQ summary statistics in JSON format
- [shovill-se](#) - A fork of Shovill that includes support for single-end reads
- [vcf-annotator](#) - Add biological annotations to variants in a VCF file

# Conclusions

# Future directions for Bactopia



- Bactopia still has a lot of room to grow
- Future directions
  - Improve the result reporting
  - Additional Bactopia Tools
  - Expansion of the documentation
- Always open to contributions and user feedback

# Acknowledgements

The developers of open-source tools that make their tools freely available to the community

- [Tim Read](#) and the [EMERGENT Group](#)
- [Davi Marcon](#), [Abhinav Sharma](#), and [nf-core](#)
- [Wyoming Public Health Laboratory](#)
  - Taylor Fearing
  - Jim Mildenerberger
  - Chayse Rowley
- [Theiagen Genomics](#)
- The many users of Bactopia providing feedback and helping guide future developments





# Questions and Comments

Learn more about Bactopia at [bactopia.github.io](https://bactopia.github.io)



