# Bacterial Genome Analysis Using Bactopia

Robert A. Petit III, PhD
Emory University
May 1st, 2020

# Overview

- Introduce Bactopia
- Use Case
- Behind the Scenes

# What is Bactopia?

*Bactopia is an extensive workflow for processing Illumina sequencing of bacterial genomes. The goal of Bactopia is process your data with a broad set of tools, so that you can get to the fun part of analyses quicker!*

# Bactopia Philosophy

1. Conda First
    a. Available from an official channel (Bioconda, conda-forge, defaults, etc…)
    b. If not available, can I create a recipe?

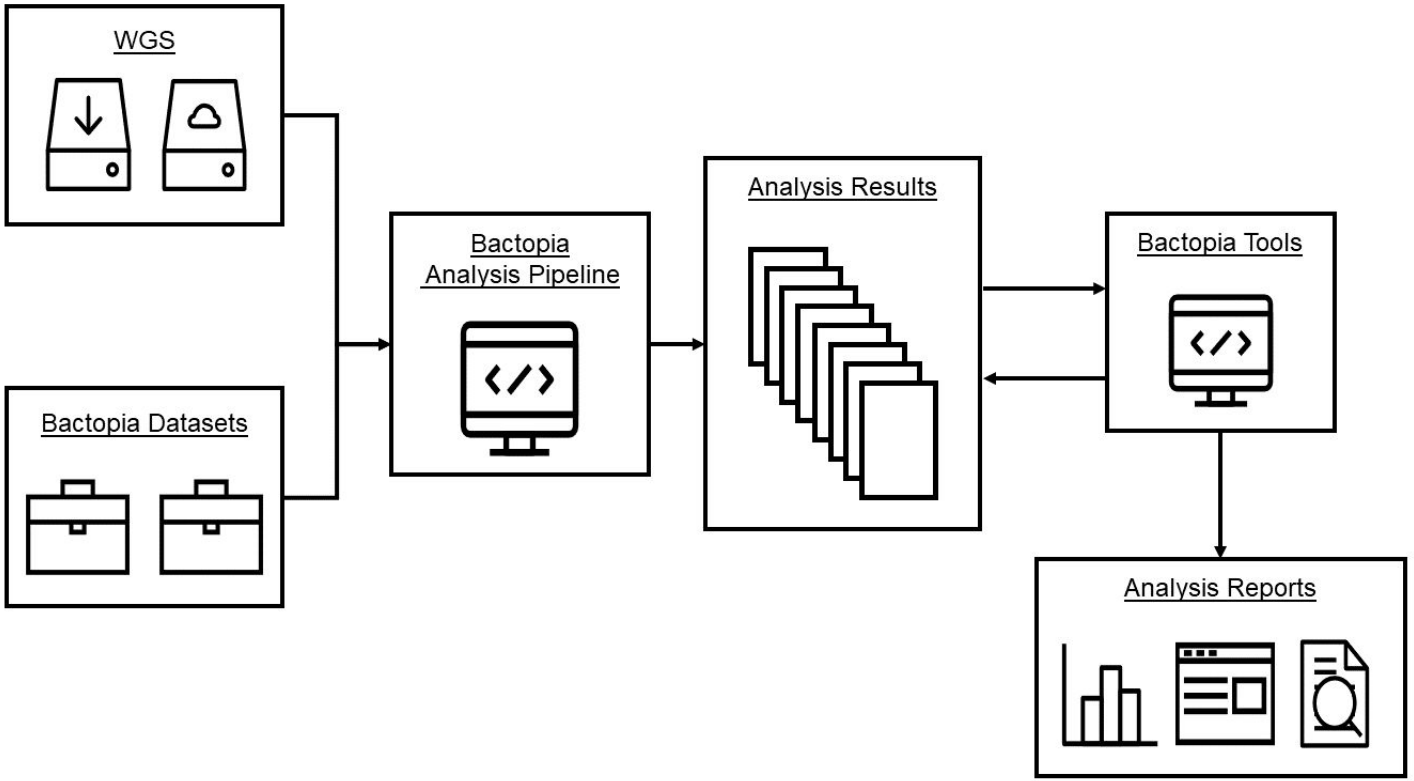
2. Flexible & Portable
    a. Fit your needs (100+ adjustable parameters)
    b. Easy to install (Conda, Docker, Singularity)
    c. Easy to switch between environments (local, cluster, cloud)

3. Documentation
    a. Too much is better than too little
    b. When in doubt document it!

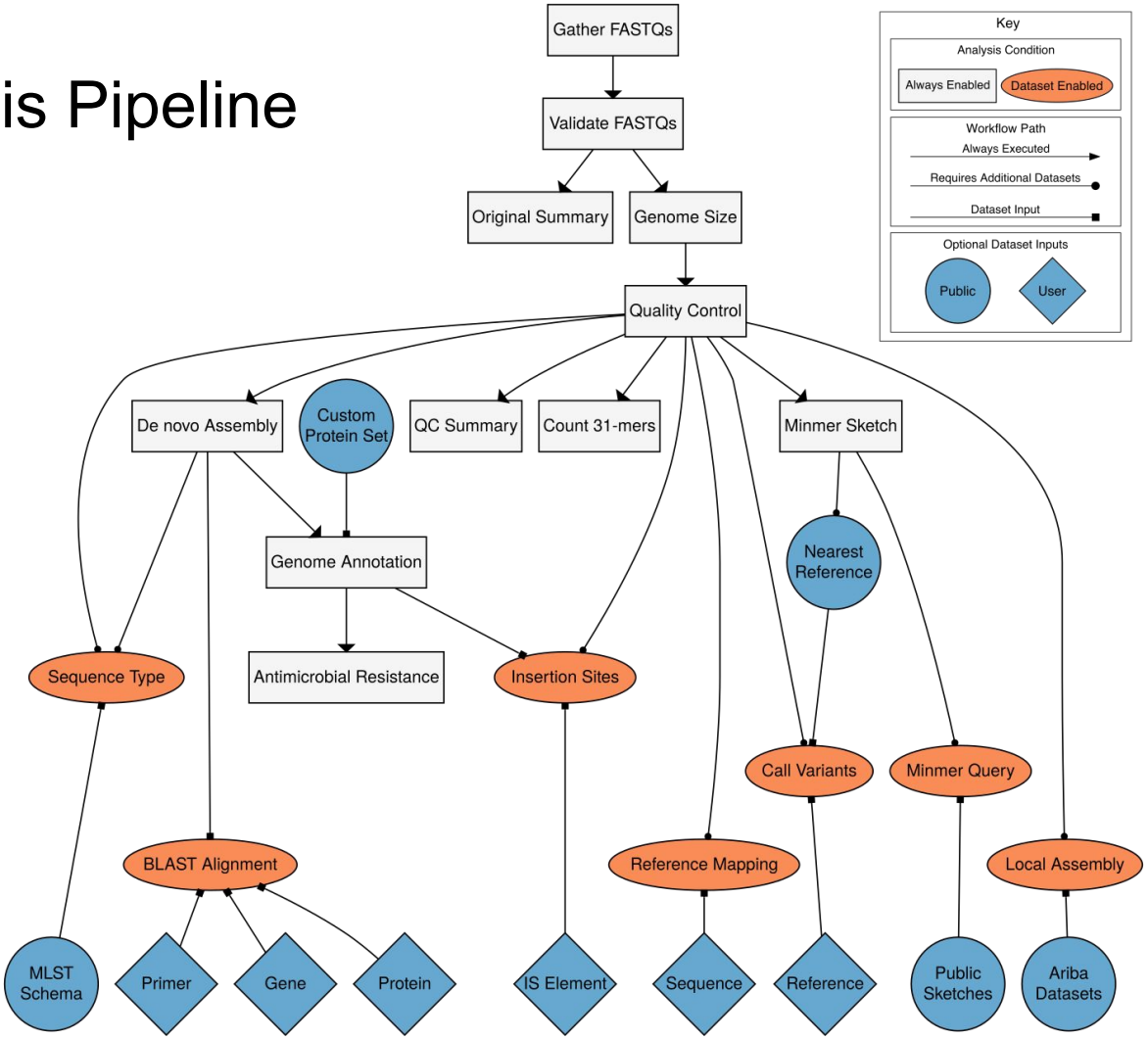
# Three Sides of Bactopia

- Bactopia Datasets
  - Framework for including public and user created datasets
  - Completely optional, but highly recommended
- Bactopia Analysis Pipeline
  - Main *per-isolate* workflow
- Bactopia Tools
  - Independent workflows for *comparative* analyses

# Bactopia Overview



# Bactopia Analysis Pipeline



# Bactopia Use Case: *Lactobacillus* genus

- Run all publicly available *“Lactobacillus”* genomes through Bactopia
  - Build *Lactobacillus* datasets
  - Query ENA for available Lactobacillus genomes
  - Run SRA/ENA genomes through Bactopia Analysis Pipeline
    - Bactopia will download genomes automatically
  - Apply Bactopia Tools to describe the genus
    - Sequence quality summary
    - 16S phylogeny with taxon classifications
    - Core-genome on a subset of samples

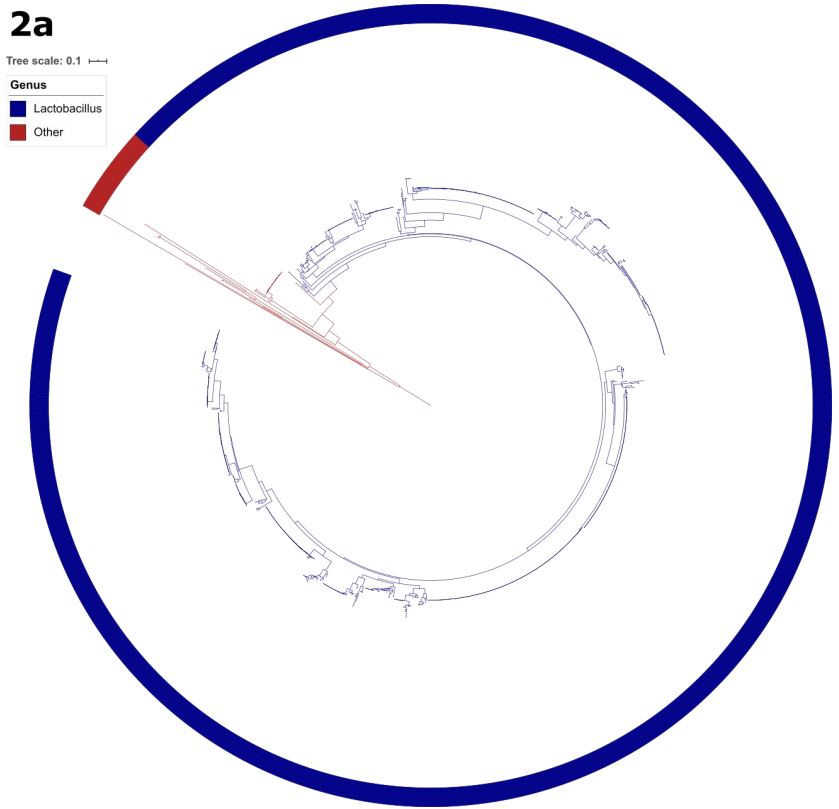
# What's does the sequence data look like?

| Quality Rank | Count | Original Coverage (Median) | Post-Bactopia Coverage (Median) | Per-Read Quality Score (Median) | Read Length (Median) | Contig Count (Median) | Percent of Assembled Genome Size compared to Estimated Genome Size (Median) |
|---|---|---|---|---|---|---|---|
| Gold | 967 | 213x | 100x | Q35 | 100bp | 54 | 92% |
| Silver | 386 | 160x | 100x | Q35 | 100 | 97 | 93% |
| Bronze | 205 | 102x | 100x | Q34 | 99 | 90 | 95% |
| Exclude | 48 | 26x | 22x | Q34 | 95 | 706 | 93% |
| QC Failure | 58 | - | - | - | - | - | - |

- ~1400 downloaded from SRA and processed by Bactopia
- Took 2.5 days to process

# Not everything is *Lactobacillus*

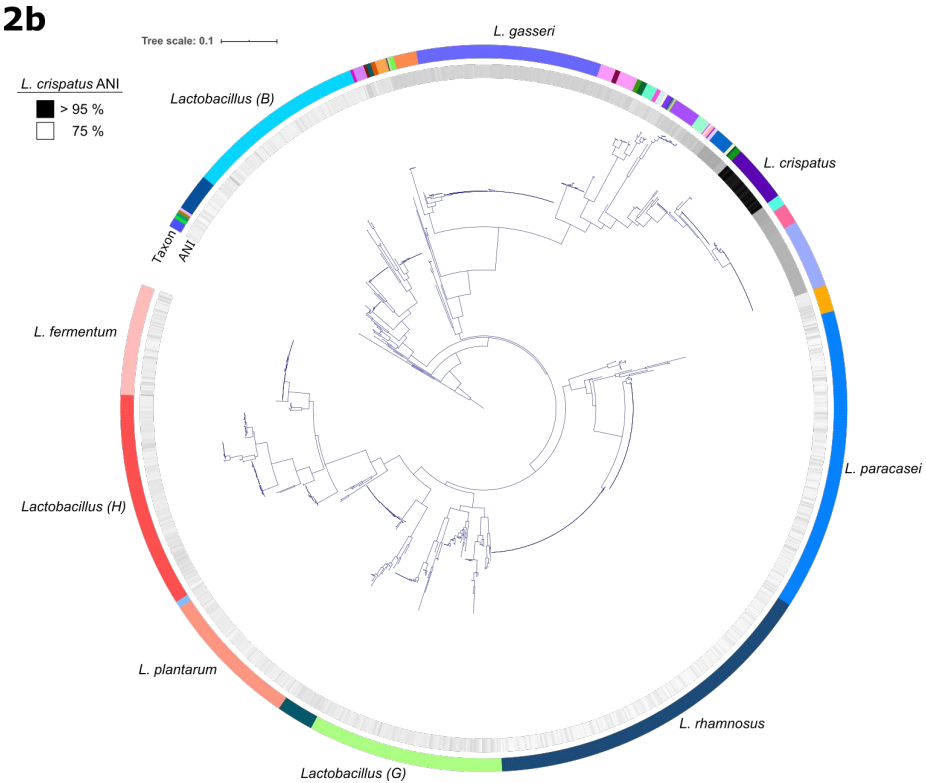
- 16S rRNA gene phylogeny
  - Bactopia Tool - phyloflash
- Taxon classified by GTDB
  - Bactopia Tool - gtdb
- 58 samples not Lacto
  - 34 samples are *S. pneumoniae*
- ~33% of the GTDB classifications in conflict with the NCBI taxon



# Major groups of *Lactobacillus*

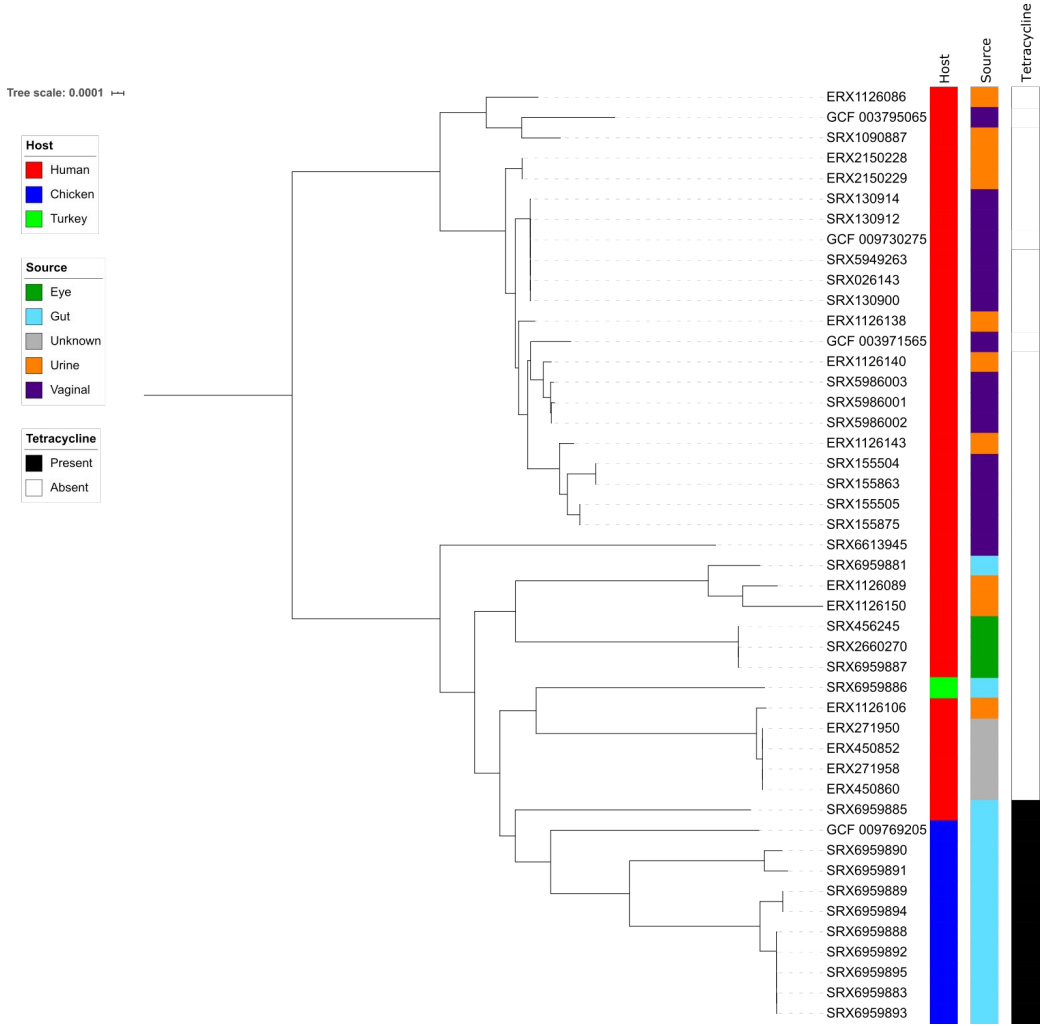
- 5 species make of ~45% of the available genomes
  - *L. rhamnosus* (n=225)
  - *L. paracasei* (n=180)
  - *L. gasseri* (n=132)
  - *L. plantarum* (n=86)
  - *L. fermentum* (n=80)

- *L. crispatus* genomes are easily identified by ANI
  - Bactopia Tool - fastani



# *Lactobacillus crispatus*

- Commonly isolated from human vagina and guts/feces of poultry
- Core-genome phylogeny of genomes with >95 ANI to *L. crispatus*
  - Bactopia Tool - Roary
- All samples from chickens had presence of Tetracycline resistance gene

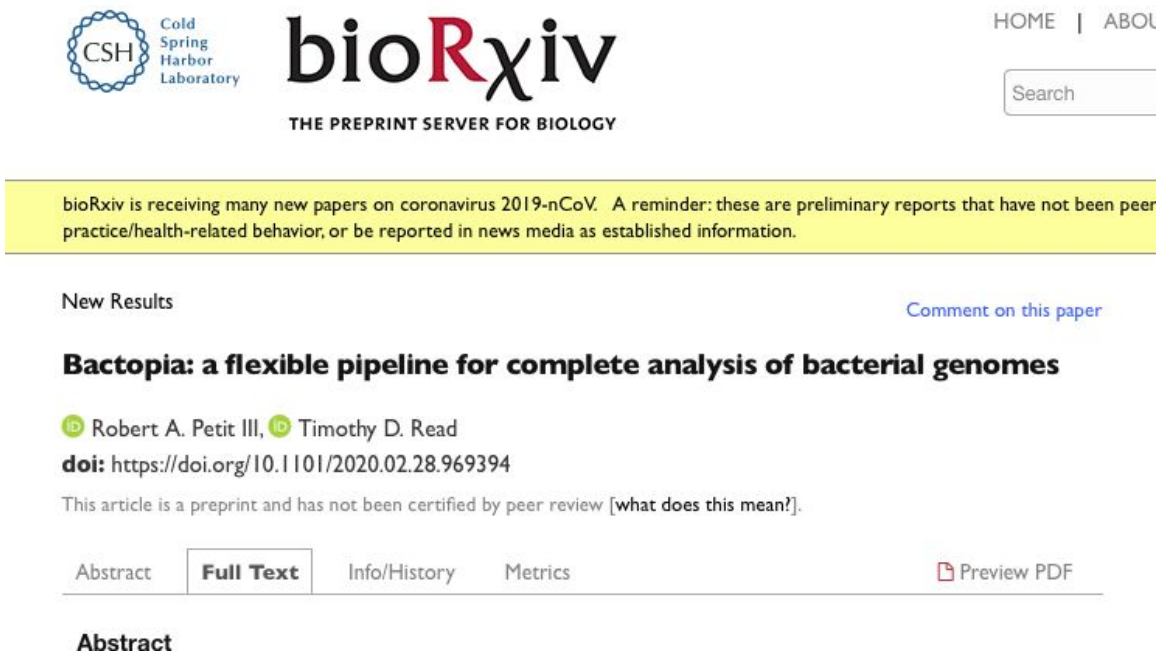


# Use Summary

- We demonstrate how Bactopia can:
  - Build datasets (`bactopia datasets`)
  - Query ENA for publicly available genomes (`bactopia search`)
  - Process publicly available genomes (`bactopia`)
  - Conduct comparative analyses (`bactopia tools`)
    - Summary report
      - summary tool, also creates list of samples to exclude from downstream analysis
    - 16S phylogeny (
      - phyloflash and gtdb tools
    - Core-genome on subset of samples
      - fastani → roary tools

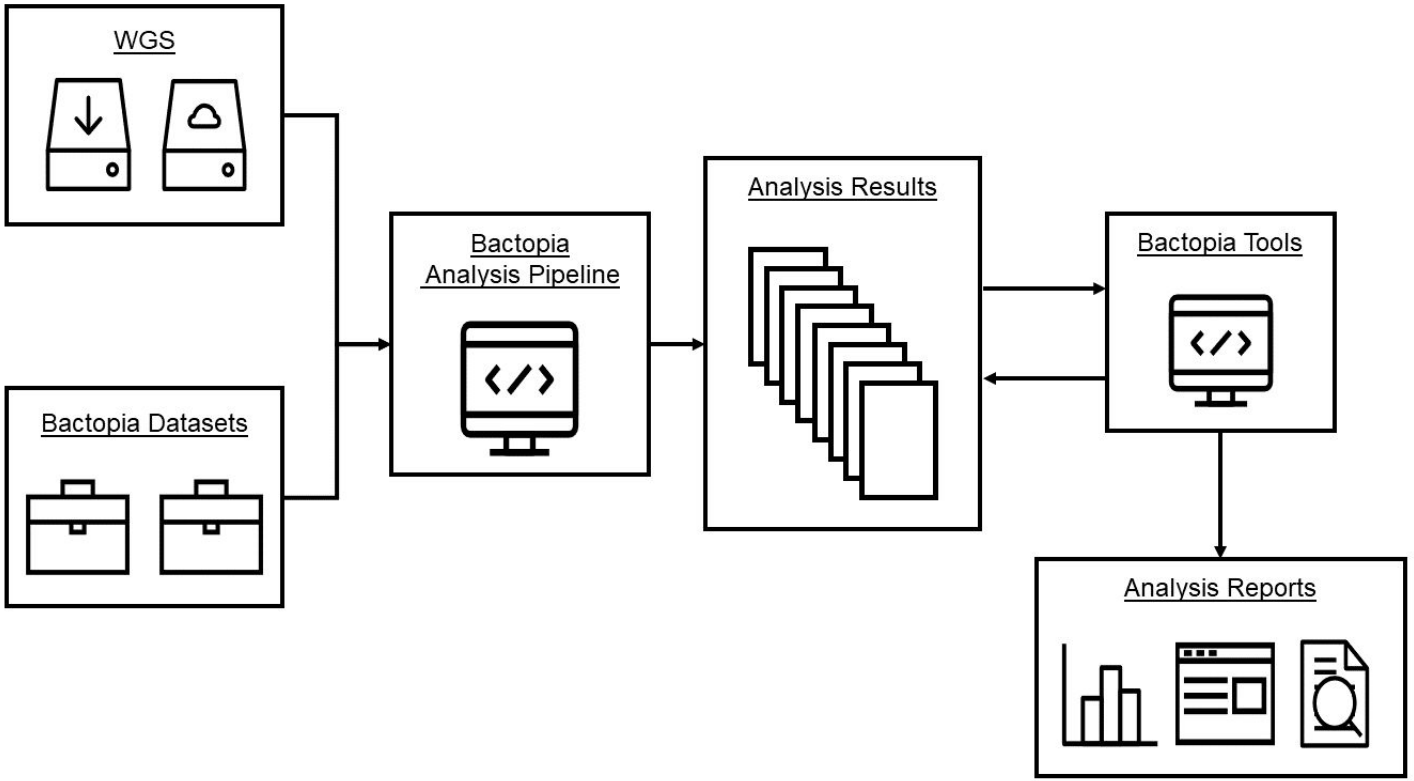
# Preprint

https://www.biorxiv.org/content/10.1101/2020.02.28.969394v1.full

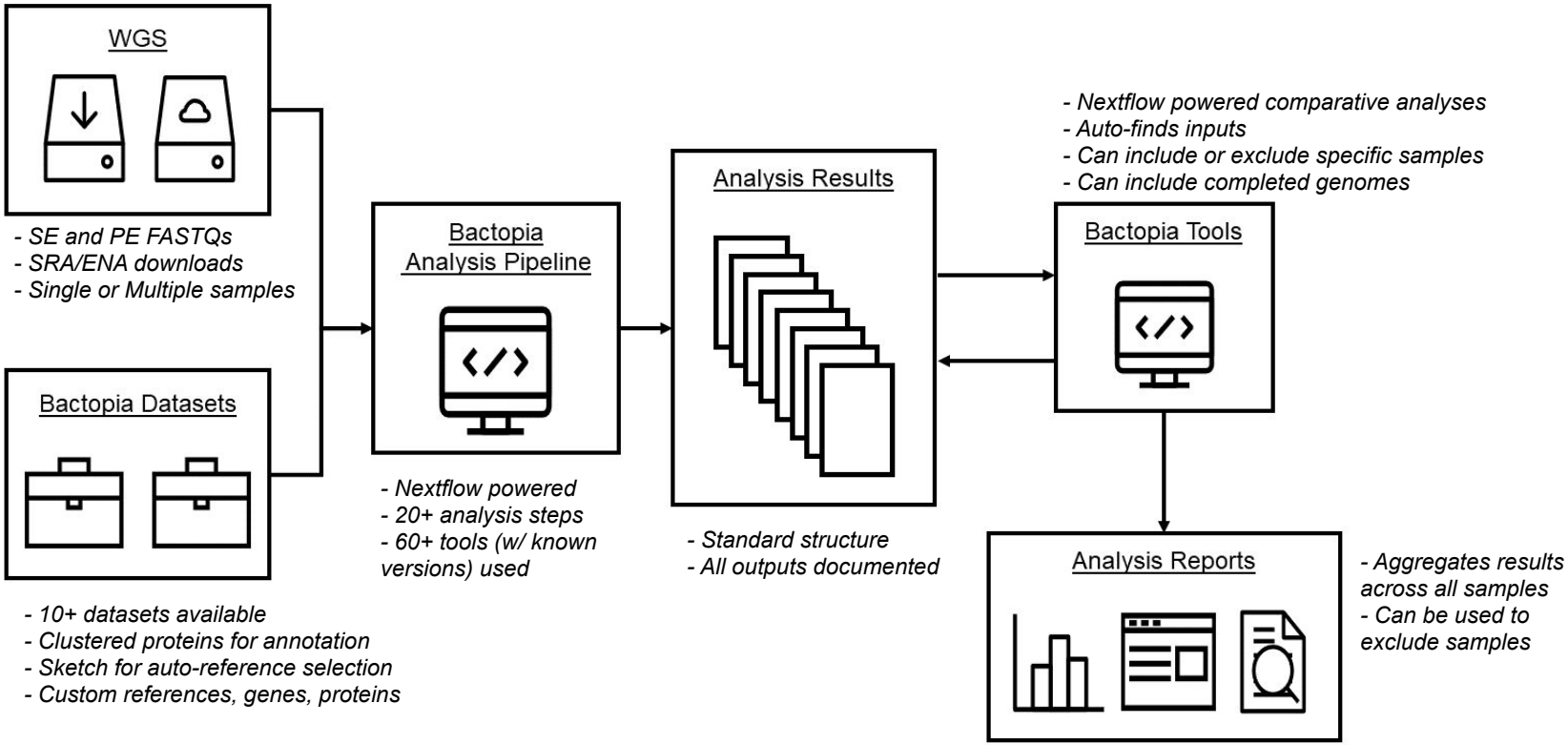


# *Bactopia: Behind the Scenes*

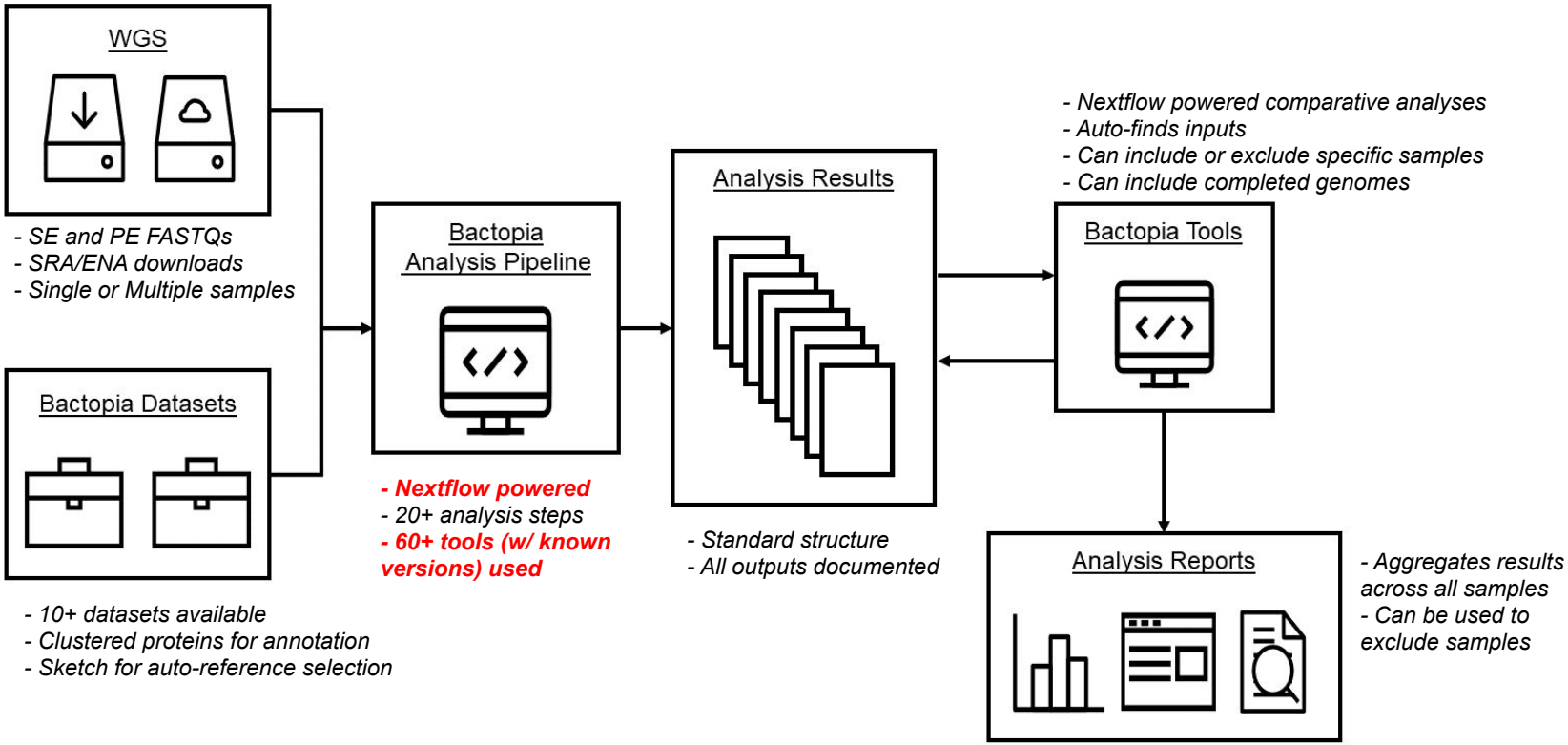
# Bactopia Overview



# Bactopia Overview



# Bactopia Overview



# Criteria For Tool Selection

1. Does Torsten Seemann have a tool for analysis '*X*'?
    a. If yes, include it. (E.g. Shovill, Prokka, Snippy)
    b. Jokes aside, is it a well known/used tool?

2. Is it still 'maintained'?
    a. If not, is there a similar alternative that is 'maintained'?

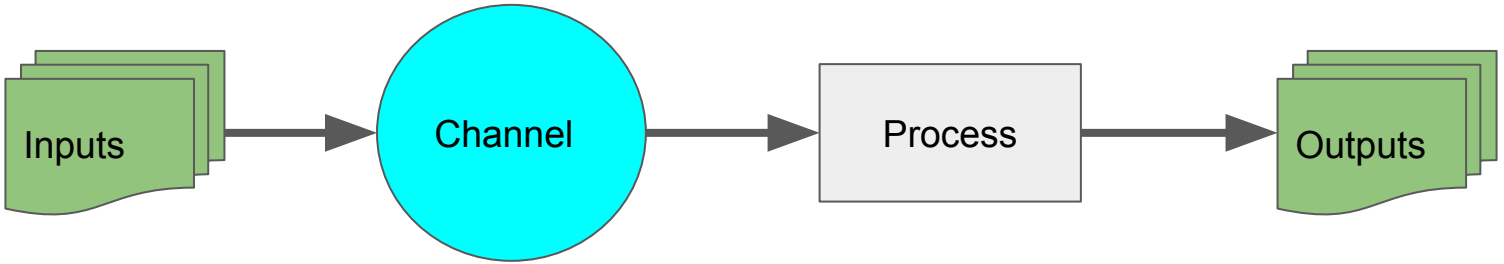
3. Is it Conda installable?
    a. If not, can I create a recipe? Is it easy to run? etc...

4. Is it a comparative analysis?
    a. If yes, it should be a Bactopia Tool

# Bactopia is powered by Nextflow

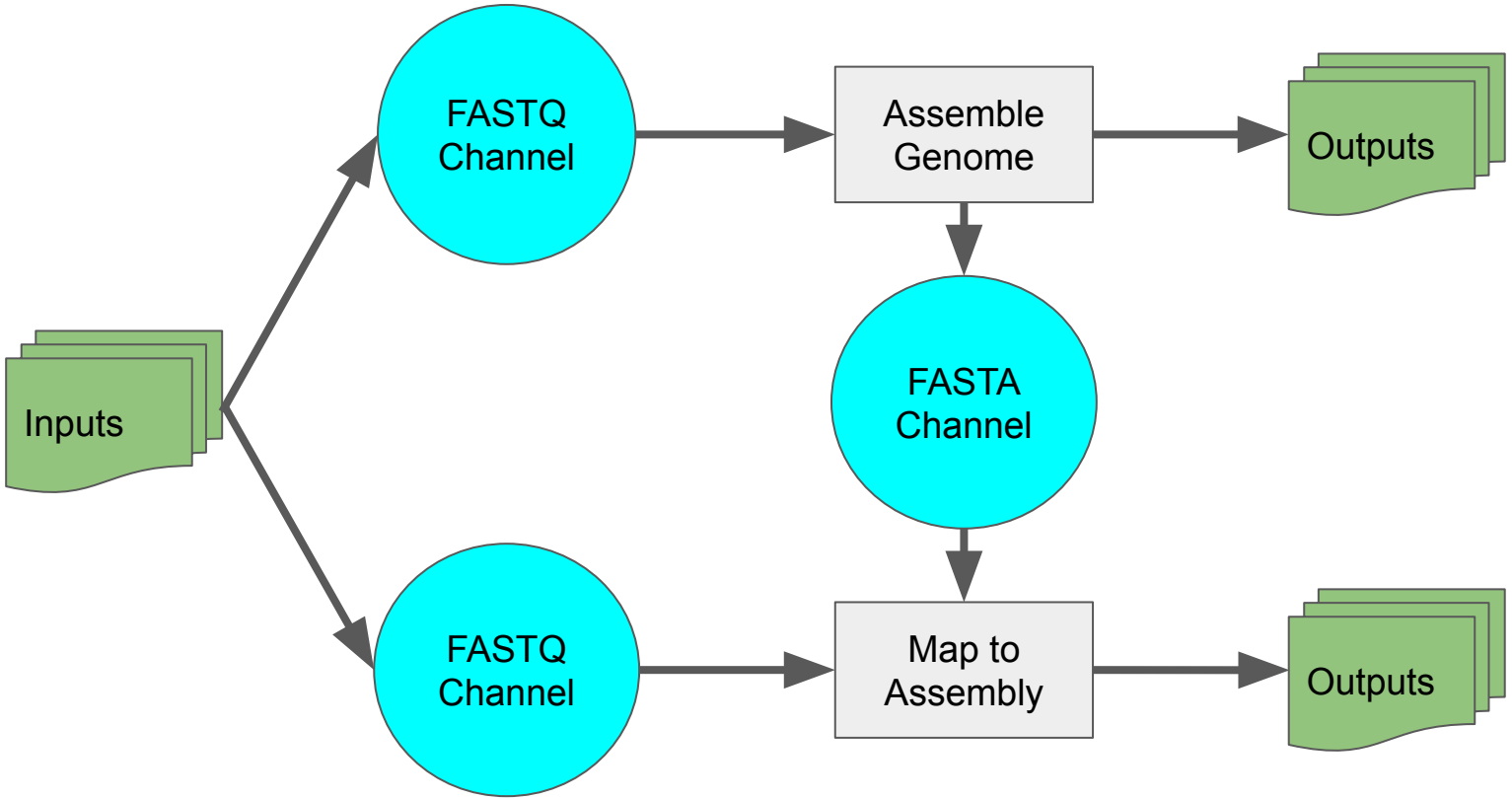
- Why Nextflow over Snakemake, etc…?
  - Extensive documentation and plenty of examples
  - Great technical support
    - Gitter, GitHub, Twitter
  - Built in support for multiple environments
  - Resumable workflows
  - Well funded

# Basic Nextflow Structure

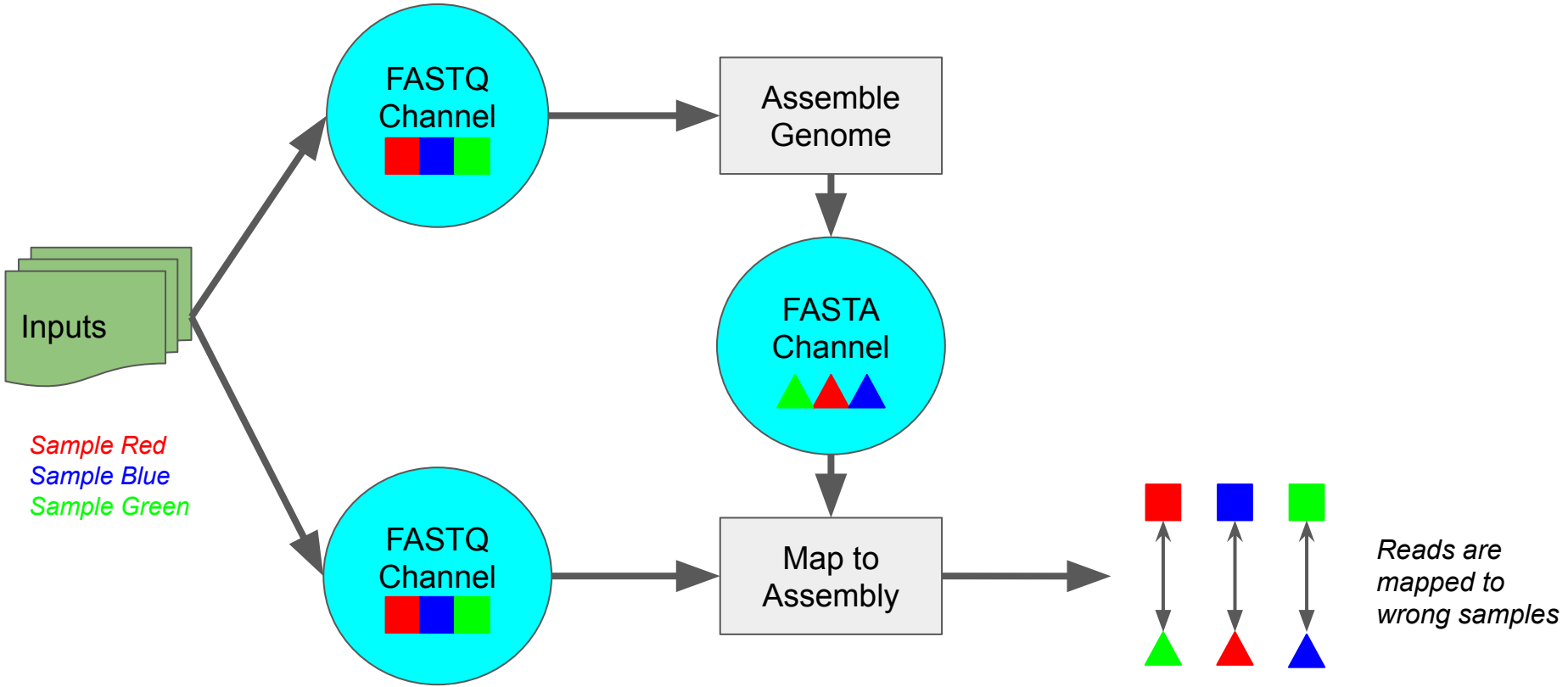



- Channel
  - Queue Channel
    - FIFO queue connecting two processes
    - Can only be used once
  - Value Channel
    - A single value (e.g. genome size)
- Process
  - Execution of tools for analysis

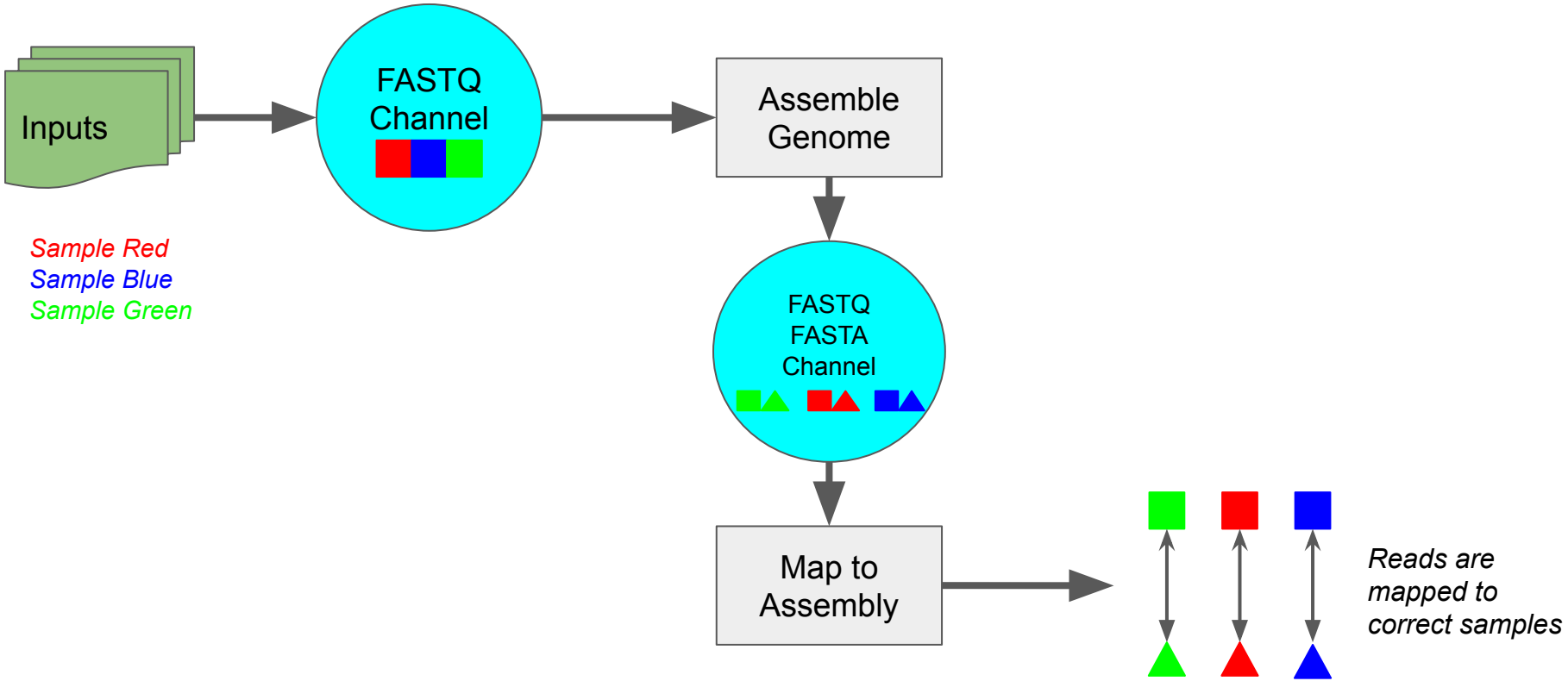
# A Simple Workflow



# FIFO Can Cause Problems



# Solution: Bactopia Carries Inputs Across Processes



# Bactopia Uses Process Templates

- Large Nextflow workflows can become hard to sift through
- Process templates allow you externalize the shell commands to a separate file
- Helps with debugging a process
  - Bactopia includes 25 processes
- Reduces the main workflows line count
  - Bactopia is 1,800 lines, not including 1,000 in template files
    - 600 is just usage

# Nextflow and Conda Environments

- Nextflow automatically builds Conda environments

- Occasionally errors occur:
  - ConnectTimeoutError - Connectivity issues
  - InvalidArchiveError - Multiple environments being built at once trying to access same file(s)
    - Nextflow does this

- Created `bactopia build` to solve this
  - Pre-builds Conda environments, one at a time, for Nextflow to use
  - Environments are built from yaml files so that versions are known

# A Few Warnings About Nextflow

- Missing Features
  - Command line argument parser
  - Dry run feature
    - Nextflow's design makes this difficult to implement
- Error messages are sometimes hard to decipher or missing
- Running multiple workflows in same directory may break resume function
  - Resumes last executed workflow
- By default Nextflow uses all available resources (e.g. CPUs)

# What’s next for Bactopia?

- Hopefully a useful resource for community
  - Will help to polish it and get a sense of what people want
- Curated species-specific datasets
- More Bactopia Tools
  - mashtree, hicap, bactdate, pyseer, poppunk2, bigsi, etc…
- Implement long-read support (eventually)
  - Illumina is still 90+% of the data available and field is advancing to quickly at the moment
- Process 70,000 *Staphylococcus aureus* genomes
  - All on AWS and the basis for Staphopia v2

# Acknowledgements

Tim Read

Bactopia Preprint
- Monica Farley
- Oliver Schwengers
- Torsten Seemann
- Narciso Quijada
- Michelle Wright

Bactopia Tool Feedback
- Tauqeer Alam
- Ashley Alexander
- Taj Azarian
- Ahmed Babiker
- Curtis Kapsak
- Jon Moller
- Matt Plumb
- Michelle Su
- Sean Wang
- Emily Wissel

Funding

NIAID, Amazon in Education Award, CDC Emerging Infection Program (EIP)

# Bactopia Links

- Documentation
  - https://bactopia.github.io/
- Github
  - https://github.com/bactopia/bactopia/
- Bioconda
  - https://bioconda.github.io/recipes/bactopia/README.html
- Docker
  - https://hub.docker.com/u/bactopia
- Singularity
  - https://cloud.sylabs.io/library/rpetit3/bactopia

# Alternatives to Bactopia

- Nullarbor
  - "Reads to report" for public health and clinical microbiology
- ASA$^3$P
  - A scalable bacterial genome assembly, annotation and analysis pipeline
- TORMES
  - Making whole genome bacterial sequencing data analysis easy
- QuAISAR-H
  - Pipeline to determine Quality, Assembly, Identification, Sequence type, Annotation, Resistance mechanisms for hospital acquired infections

# Bactopia subcommands

```
bactopia - v1.3.1

Available Commands
bactopia - Execute the Bactopia Nextflow pipeline

bactopia build - Build Bactopia Conda environments

bactopia citations - Print citation for datasets, tools and Bactopia

bactopia datasets - Download/setup useful datasets for Bactopia

bactopia prepare - Create a 'file of filenames' for input FASTQ files

bactopia search - Query Taxon ID or Study Accession against ENA for input accessions

bactopia tools - Execute existing Bactopia Tools

bactopia versions - Print versions of tools and Bactopia

bactopia --citation - Print the Bactopia citation
bactopia --version - Print the Bactopia version
```