Bactopia: A Flexible Pipeline for Complete Analysis of Bacterial Genomes

EMORY UNIVERSITY SCHOOL OF MEDICINE
Department of Medicine

Robert A. Petit III and Timothy D. Read
Division of Infectious Diseases, School of Medicine, Emory University

Background

Bactopia [1] is an extensive workflow for complete analysis of bacterial genomes. Bactopia was developed from scratch with usability, portability, and speed in mind from the start.

Bactopia uses Nextflow, allowing for support of many types of environments. You can run Bactopia on your laptop, HPC clusters, and the cloud. To make installation simple, Bactopia only uses software packages available from public Conda channels such as Bioconda. Bactopia is available from BioConda or a container (Docker and Singularity) and can be used on Linux and MacOSX machines.

Bactopia can be split into three main parts: Bactopia Datasets, Bactopia Analysis Pipeline, and Bactopia Tools.

Bactopia Datasets

Bactopia Datasets allow you to automatically incorporate numerous public datasets, or your own, into your analysis. Some of the public datasets currently included are:

- Ariba's getref Reference Datasets - CARD, MEGARes, VFDB, many others.
- RefSeq Mash Sketch - ~100,000 bacterial genomes and plasmids from NCBI RefSeq
- GenBank Sourmash Signatures - ~87,000 microbial genomes from NCBI GenBank
- PLSDB Mash Sketch & BLAST - all plasmids available from PLSDB
- PubMLST.org MLST Schemas - MLST allelic profiles and sequences for bacterial species

Bactopia Tools

Bactopia Tools are a set of predefined workflows that allow you to quickly conduct comparative analyses using the results from Bactopia Analysis Pipeline. Currently there are 8 Bactopia Tools available:

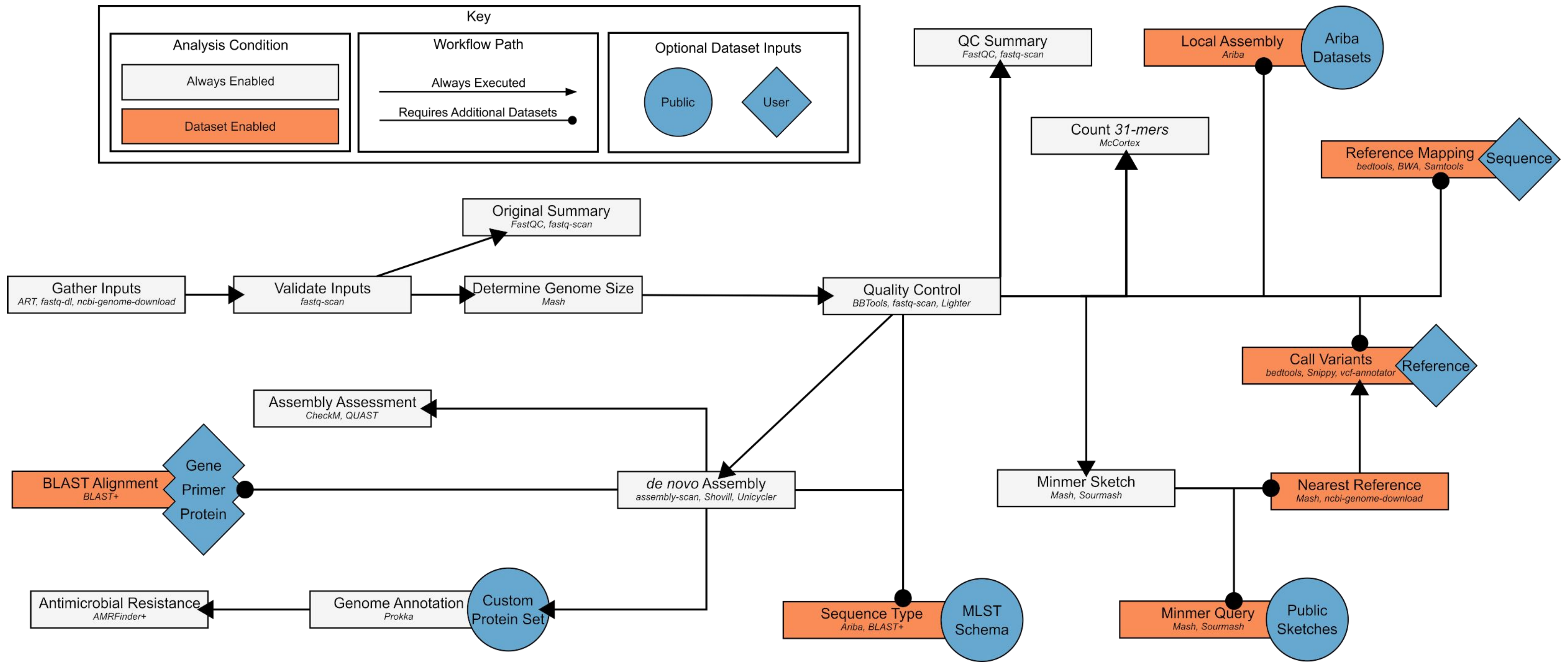
- eggNOG-mapper - functional annotation of protein sequences
- FastANI - compute whole-genome average nucleotide identities
- GTDB-Tk - assign taxonomic classifications
- ISMapper - identify insertion sites
- Mashtree - create trees using Mash distances
- phyloFlash - reconstruct 16S rRNA genes and phylogeny
- PIRATE - pan-genome analysis and core-genome phylogeny
- Roary - pan-genome analysis and core-genome phylogeny

@rpetit3

Learn more about Bactopia at: bactopia.github.io

Bactopia Analysis Pipeline

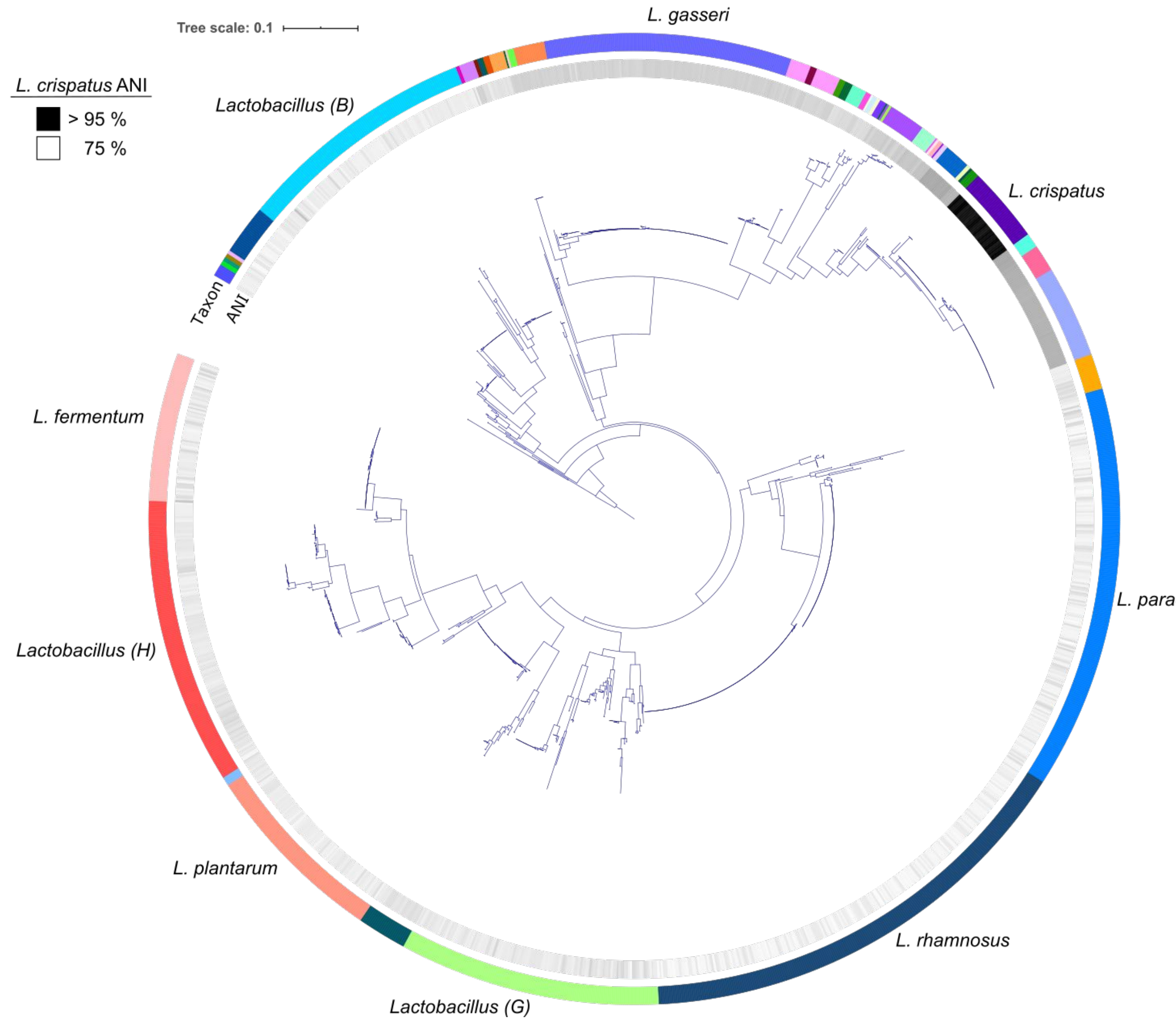
Bactopia Analysis Pipeline (below) is an extensive workflow integrating numerous steps in bacterial genome analysis. There are currently more than 70 bacterial genomic tools included.



Use Case: the *Lactobacillus* genus

We used Bactopia to analyze 1,558 Lactobacillus publicly available genomes from the Sequence Read Archive. After processing we used the Bactopia Tools for comparative analyses. Some key takeaways were:

- 505 genomes had taxonomic descrepencies between GTDB and NCBI
- 58 genomes labeled as *Lactobacillus* were of a completely different genus
- 5 *Lactobacillus* species represented 45% of the genomes
- *L. crispatus* had two distinct phylogenetic groups: human vaginal isolates and poultry/human gut



References

[1]: Petit III, R. A. & Read, T. D. Bactopia: a Flexible Pipeline for Complete Analysis of Bacterial Genomes. *mSystems* 5, (2020)

Funding

Support for this project came from an Emory Public Health Bioinformatics Fellowship funded by the CDC Emerging Infections Program (U50CK000485) PPHF/ACA: Enhancing Epidemiology and Laboratory Capacity.