

Bactopia: Highly scalable, portable and customizable bacterial genome analyses

Robert A. Petit III^{1,2}, Davi J. Marcon^{3,4}, Abhinav Sharma⁵, and Timothy D. Read⁶

¹ Theiagen Genomics, Highlands Ranch, CO, USA

² Public Health Laboratory, Wyoming Department of Health, Cheyenne, Wyoming, USA

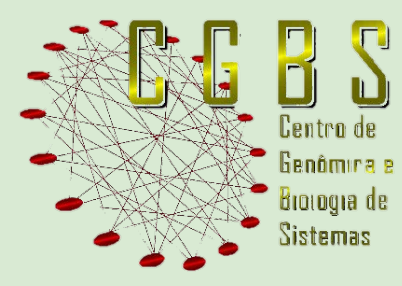
³ Center of Genomics and Systems Biology, Institute of Biological Sciences, Universidade Federal do Pará, Belém, PA, Brazil

⁴ Laboratory of Genetic Engineering, Guamá Science and Technology Park, Belém, PA, Brazil

⁵ Faculty of Engineering and Technology. Liverpool John Moores University. Liverpool, United Kingdom.

⁶ Division of Infectious Diseases, Department of Medicine, Emory University School of Medicine, Atlanta, Georgia, USA

Correspondence: robert.petit@theiagen.com Bactopia Documentation: bactopia.github.io



Learn more about Bactopia!

Accepted Inputs

- Illumina and/or Nanopore Reads**
--RL/--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



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

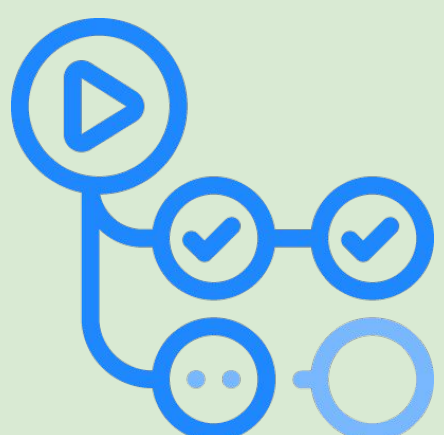
- Gather Samples**
Collect local files and/or download SRA or NCBI Assembly
- QC Reads**
Trim and filter low quality reads, subsample to specified coverage, and generate quality summary metrics
- Minmer Sketch and Minmer Query**
Create minmer sketches and query them against RefSeq and GenBank
- Call 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

Highlights

- Complete analysis of bacterial genomes
- Nextflow DSL2, greatly increases Bactopia's ability to fit your needs
- Supports Illumina, Oxford Nanopore technologies as well as NCBI's Sequence Read Archive and Assembly databases
- Includes more than [130 bioinformatics tools](#)
- [33 Bactopia Tools](#) include more workflows for more science
- Extensively tested with 100+ tests for 10,000+ output files
- Available on Bioconda, Docker, and Singularity

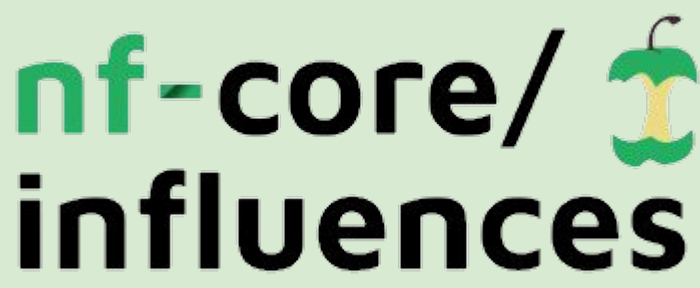
Bactopia is platform Independent

By using Nextflow, Bactopia can run on multiple platforms, including:



Test Everything

Bactopia features per-module tests using real bacterial sequences available from [bactopia-tests](#). Currently 100+ tests have been created to verify more than 10,000 output files. These tests are all integrated into [GitHub Actions](#) to ensure the stability of Bactopia over time.



[nf-core](#) is a community effort to curate analysis pipelines built using Nextflow. This effort has provided a powerful design framework when developing Nextflow pipelines, and Bactopia has followed their lead by implementing:

- Modules from [nf-core/modules](#) for Bactopia Tools
- Bactopia has contributed [30+ modules](#)
- Per-module tests modeled after [nf-core/modules pytest implementation](#)
- Argument parser based off [nf-core/tools](#)
- Single meta variable for general value storage

Ultimately adoption of these practices has made Bactopia a much better pipeline to use and maintain overtime.

(Thank you nf-core team!)

[@rpetit3](#) [@tdread emory](#) [@abhi18av](#)

Bactopia Tools

Bactopia Tools are convenient workflows to do more science with your Bactopia outputs! Currently there are [33 Bactopia Tools](#) for analyses like pan-genome construction, serotyping, and many more. Check out the full set of Bactopia Tools to the right! →

Curated Datasets

To facilitate curated datasets, [bactopia-datasets](#) was created for community members to curate species-specific datasets, which can then be readily used in Bactopia.

Community Synergy

Without the developers and maintainers of open-source tools, Bactopia would not exist. To provide back to the community, Bactopia has led to 17 Bioconda recipes, 1000+ Bioconda pull-request reviews, and merged pull-requests to popular many popular bioinformatic tools.

[@rpetit3](#) [@Mxrcon](#) [@abhi18av](#)

BACTOPIA

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)

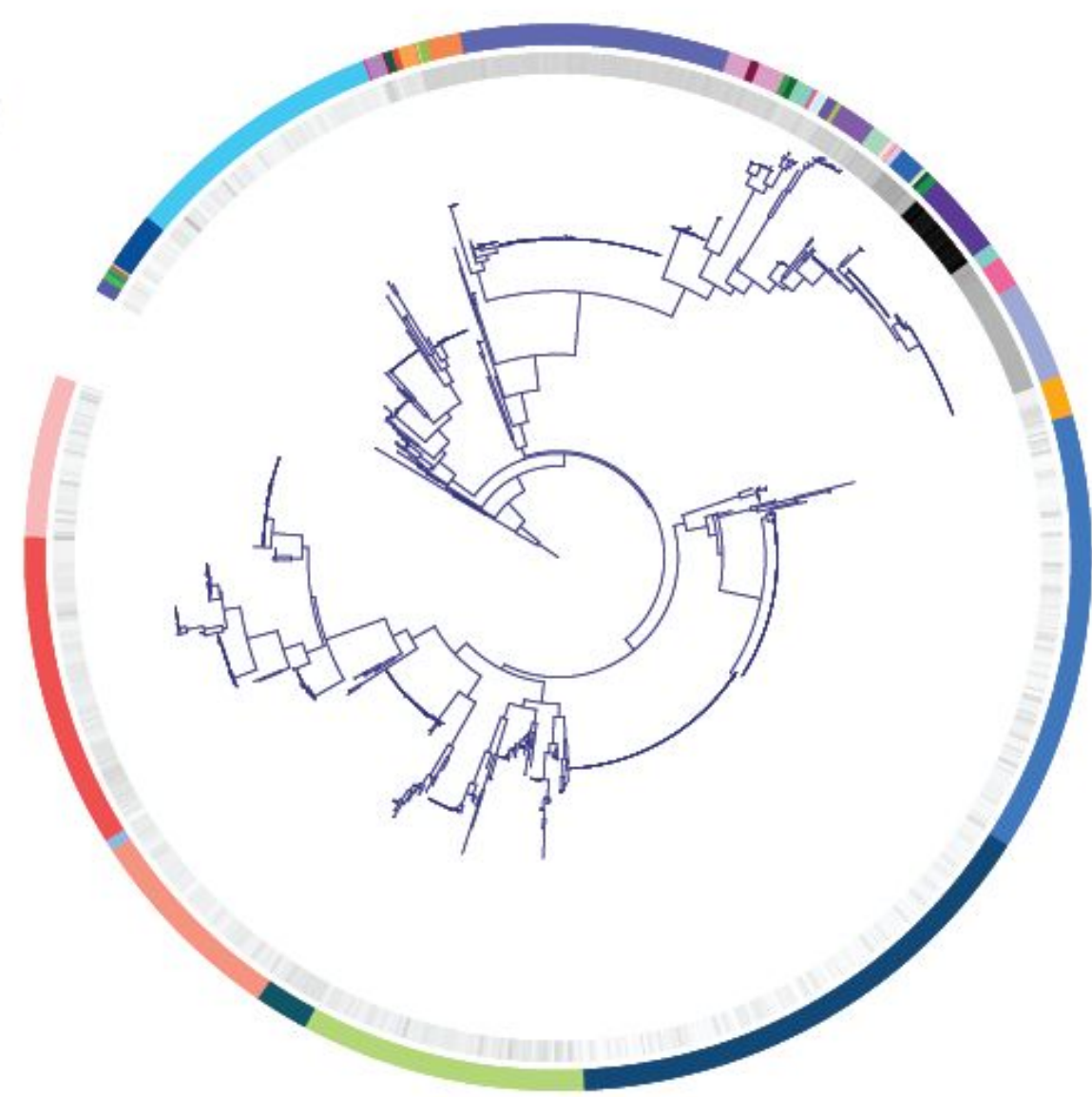
mash dist
Calculate Mash distances between sequences

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

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

emmtyping
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-secmec
Primer based SCCmec typing of Staphylococcus aureus genomes

TBProfiler
Detect resistance and lineages of Mycobacterium tuberculosis