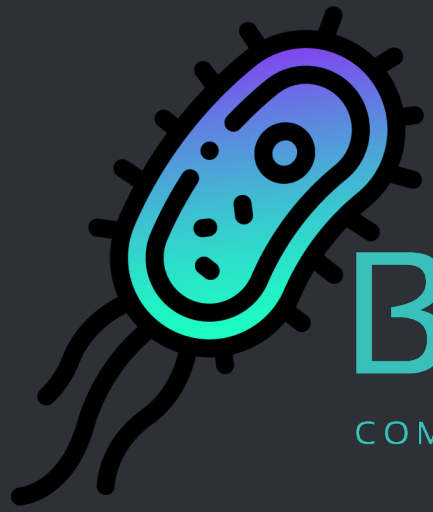


BACTOPIA

COMPLETE ANALYSIS OF BACTERIAL GENOMES

End-to-end Bacterial Genome Analysis

Comparative & Experimental Medicine Seminar Series
Robert A. Petit III, PhD
Wyoming Public Health Laboratory
April 1st, 2024



Disclaimer

The views and opinions expressed in today's workshop are mine and do not necessarily reflect the views or positions of the Wyoming Public Health Laboratory or the State of Wyoming

Today will be a little different, no results to present, instead an overview of whole genome sequencing and bacterial genomics

Outline for Today



Sequencing

A look into whole genome sequencing (WGS) and common technologies.



Bacterial Genomics

We'll walk through the "standard" process for bacterial genomic analyses



More Genomics!

We'll follow this up with the many additional analyses for further investigations.



Bactopia

An introduction into Bactopia and how it helps to streamline complex analyses.



People

Meet the people behind Bactopia and how they are helping to improve it.



Wrap Up

Not much to say here, we'll close things out.

1

Whole Genome Sequening

Let’s take a look at WGS and current technologies

Whole genome sequencing allows us to interogate the genome on the scale of bases

From a bacterial lens, WGS assists

Outbreak investigations

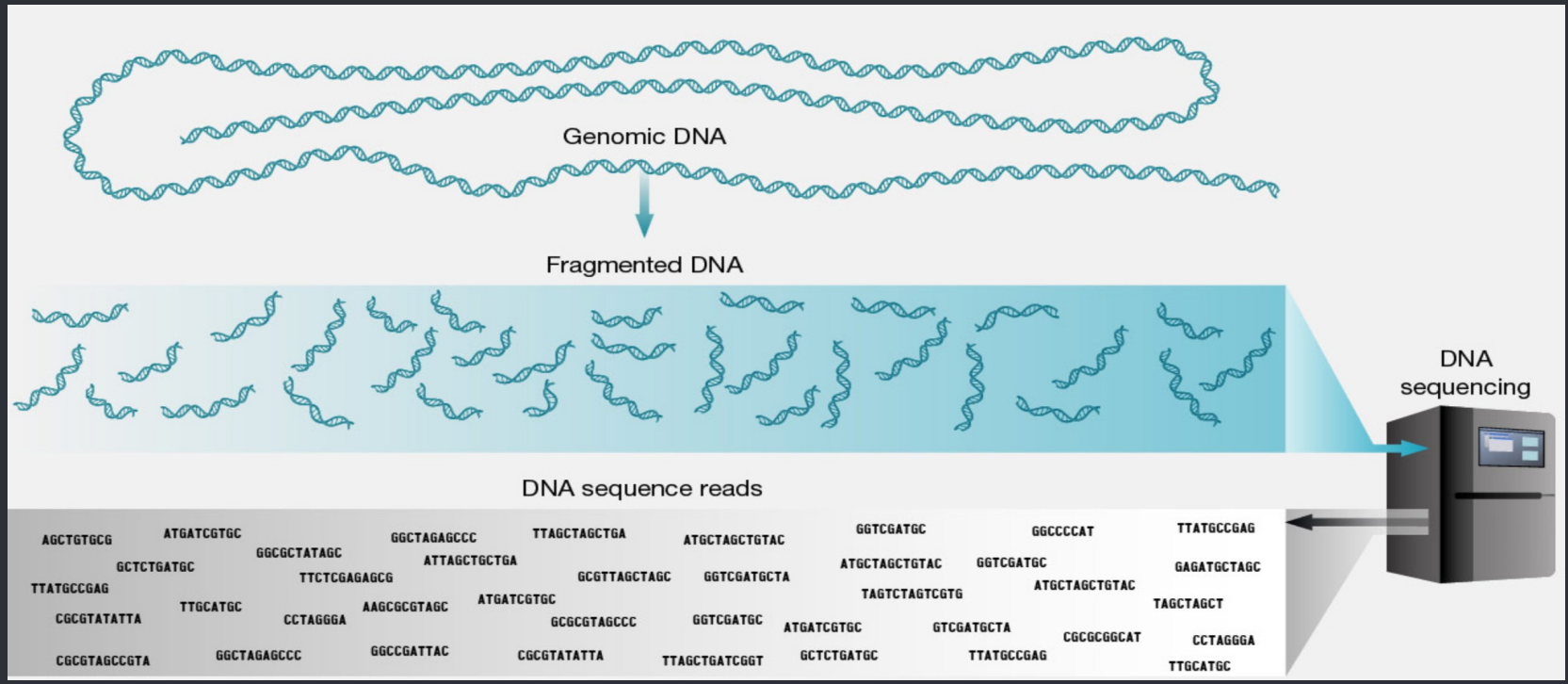
Expanding evolutionary and ecological understanding

Public health surveillance efforts

Clinical diagnostics and treatment

Many more benefits

Whole Genome “Shotgun” Sequencing



Source: NIH - NHGRI

What is a read?

Line 1 `@SRR2838702.1 FCD1DAFACXX:5:1101:14124:2241/1`
Line 2 `AAGAAATTAAAATTAGGCTGGTATTTCAATATTGACTCCATACAAACTTG`
Line 3 `+`
Line 4 `CCCFFFFFHHHHHJJJJJIJJHHIJJJJJJIJJJJJJJJJJJIIJJJJJJ`

Line 1 – The “header” line has info about the sample and sequencer

Line 2 – Contains the sequenced nucleotides

Line 3 – The “+” line, that’s all it has!

Line 4 – Contains the quality score for that given nucleotide

There are more sequencing technologies but in the microbial space, Illumina and Oxford Nanopore Technologies are dominant

Illumina

- Sequencing by Synthesis
  - Detects single bases as they are incorporated onto growing DNA strands
  - Utilizes very fast imaging to determine nucleotide based on emitted colors
- Very accurate (>99.9%, or less than 1 in 1000)
- Short reads between 100-300bp



Oxford Nanopore Technologies

- Sequencing by nanopore
  - Nano-scale pores embedded in a membrane to detect biological molecules
  - Measured changes in current are used to call bases
- Very long reads, 1,000s, 100,000s, even up to 1,000,000+ bp
- Accuracy is improving, now 99% (or 1 in 100)



Illumina

Short Reads
Long Runtimes (days)
Highly accurate
Great SNP calling
Very high coverage
Mature tools
Susceptible to substitutions

Oxford Nanopore

Very Long Reads
Short Runtimes (real-time)
Error-prone*
Poor SNP calling*
Closing genomes
Very portable
Susceptible to homopolymers

** but improving*

Whole Genome Sequencing wrap-up

Sequence the whole genome of a particular organism

“Shotgun” – break genome into fragments then sequence

Used for outbreak investigations, surveillance, clinical, research, many more

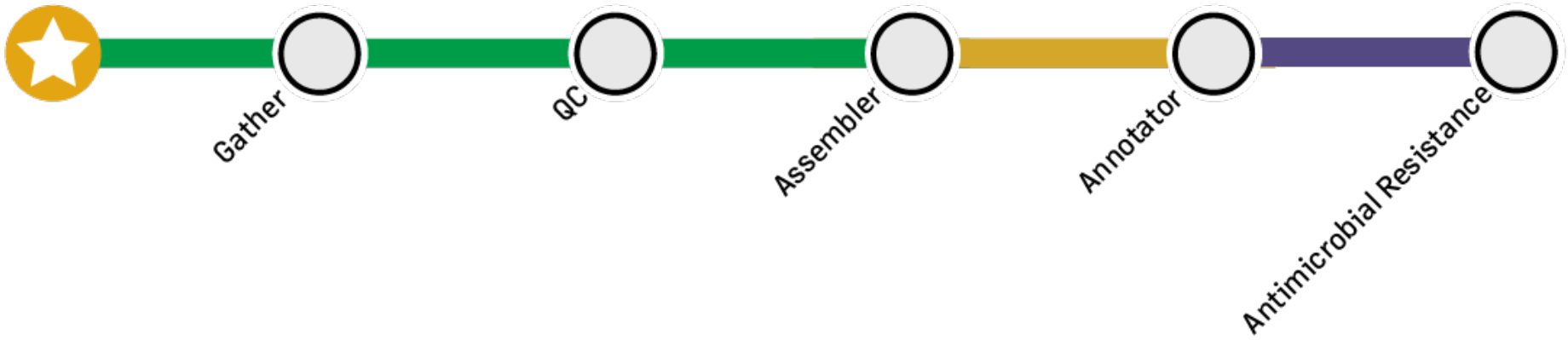
Many emerging technologies, but Illumina and ONT are major platforms in bacterial space

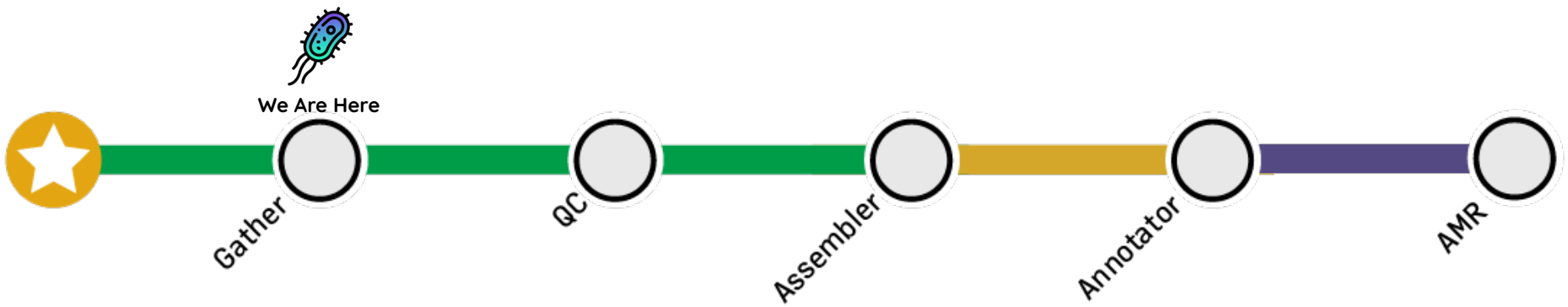
WGS is not a golden ticket, (if anything it's an expensive ticket!), consider chatting with a bioinfomatician before sequencing anything

2

“Standard” Bacterial Genomics

Let’s walkthough community standards for bacterial analysis


Gather
QC
Assembler
Annotator
Antimicrobial Resistance



“Get your samples in one place”

Local Samples

Single sample or multiples

Current study or selection of past studies

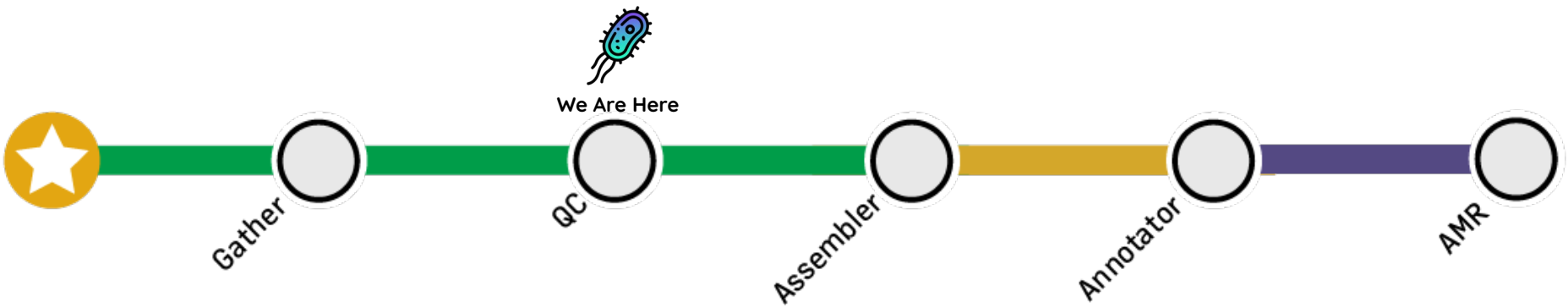
Publicly Available Samples

Sequence Read Archive / European Nucleotide Archive

- Serve Raw FASTQs

NCBI Assembly

- RefSeq – Curated assemblies owned by NCBI
- GenBank – Unaltered assemblies owned by original submitters

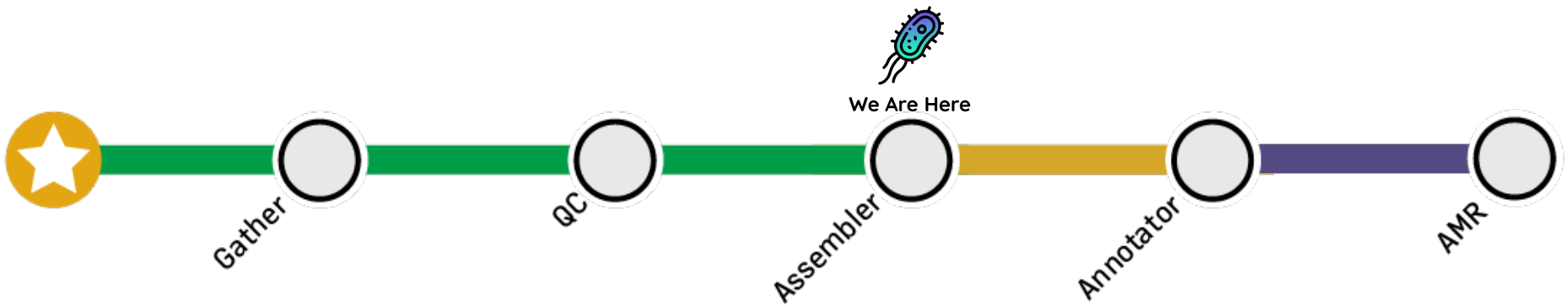


“Sequencing is not perfect, so we QC our reads”

Common metrics of Removal

Adapter sequences, Low quality or short reads, Reads with ambiguous nucleotides

The goal is to conduct downstream analysis with high quality reads



“Bring all the pieces of DNA back together”

de-novo Assembly

Assembly with no reference

Standard practice in bacterial assembly

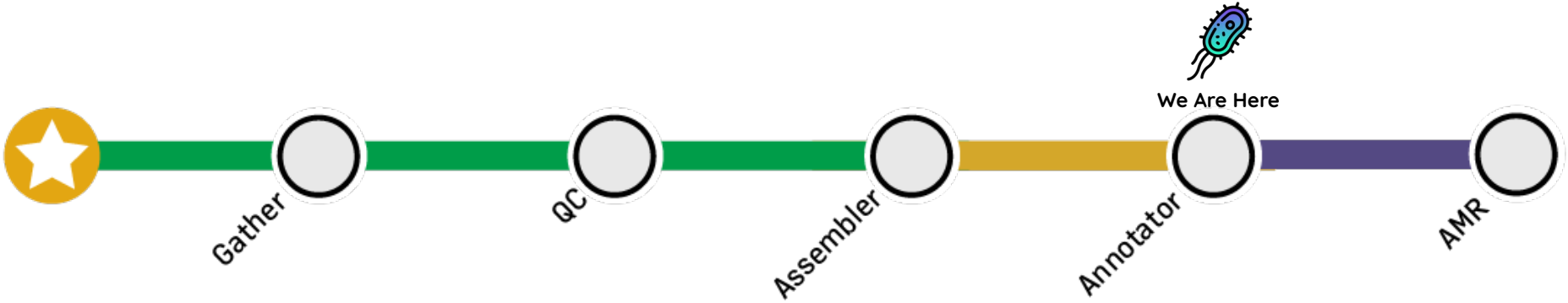
Reference-based Assembly

Dependent on reference selection

Only assembles what is in the reference

Assemblies are also QC’d

Filtering by length and coverage, “polishing”, or error-correcting

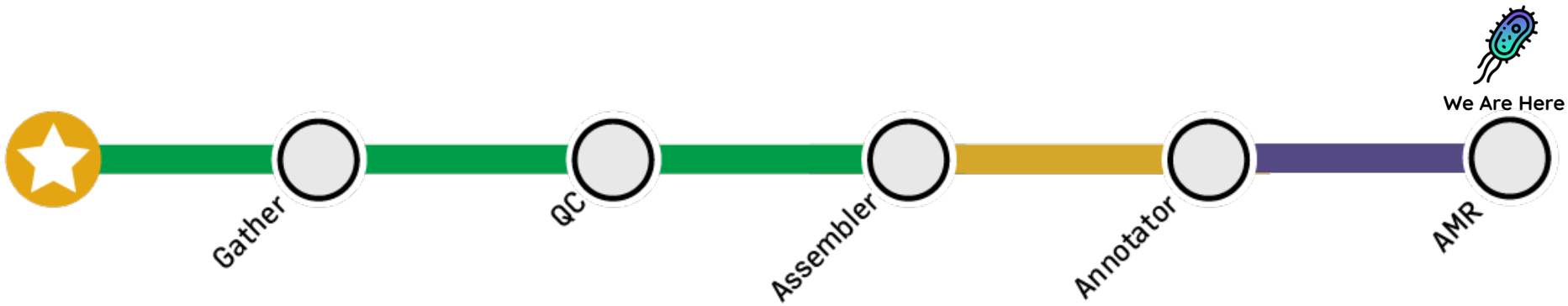


"Predict and Annotate Genes"

Use databases to assign function based on homology *(shared identity)*

Features that are Annotated

Coding sequences, RNAs (*rRNA, tRNA, tmRNA, ncRNA*), CRISPR, Pseudogenes



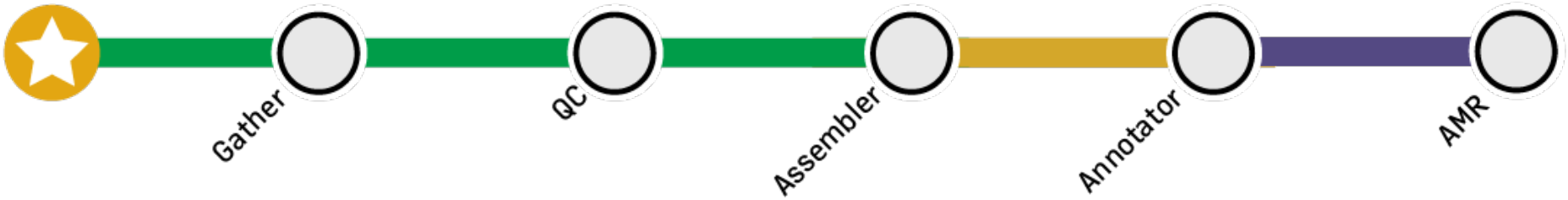
"Searching for AMR and Virulence"

Resistances based on gene presence

For example, van genes (e.g. vanA) presence confers vancomycin resistance

Resistances based on SNPs

For example, mutations in rpoB confer rifampin resistance in S. aureus



“We are done!”

This process will take 15-45 minutes

There are multiple quality controls built in

Upon completion you will have numerous results to further investigate

3

Further Genomic Analyses

Learn how you can further explore your samples

Examples of further analyses

- Species specific typing
  - Examples: *E. coli* serotyping, *H. influenzae* capsule typing, many more
- Multi-locus Sequence Typing
  - Uses 5-7 “house keeping” genes to differentiate based on Single Nucleotide Polymorphisms (SNPs)
- Mobile Genetic Elements
  - Identify and annotate mobile elements
  - Examples: Plasmids and IS elements

Examples of further analyses

- Taxonomic Classification
  - Accurately assign a species
- Sequence Alignment
  - BLAST against specific targets (e.g. genes of interest)
  - Alignment to larger genomic regions
- Sketches
  - A representation of a genome split into certain lengths called "*k*-mers"
    - *k* is some number (*e.g. codons are 3-mers*)
  - Easily screen against 100,000+ assemblies

Examples of further analyses

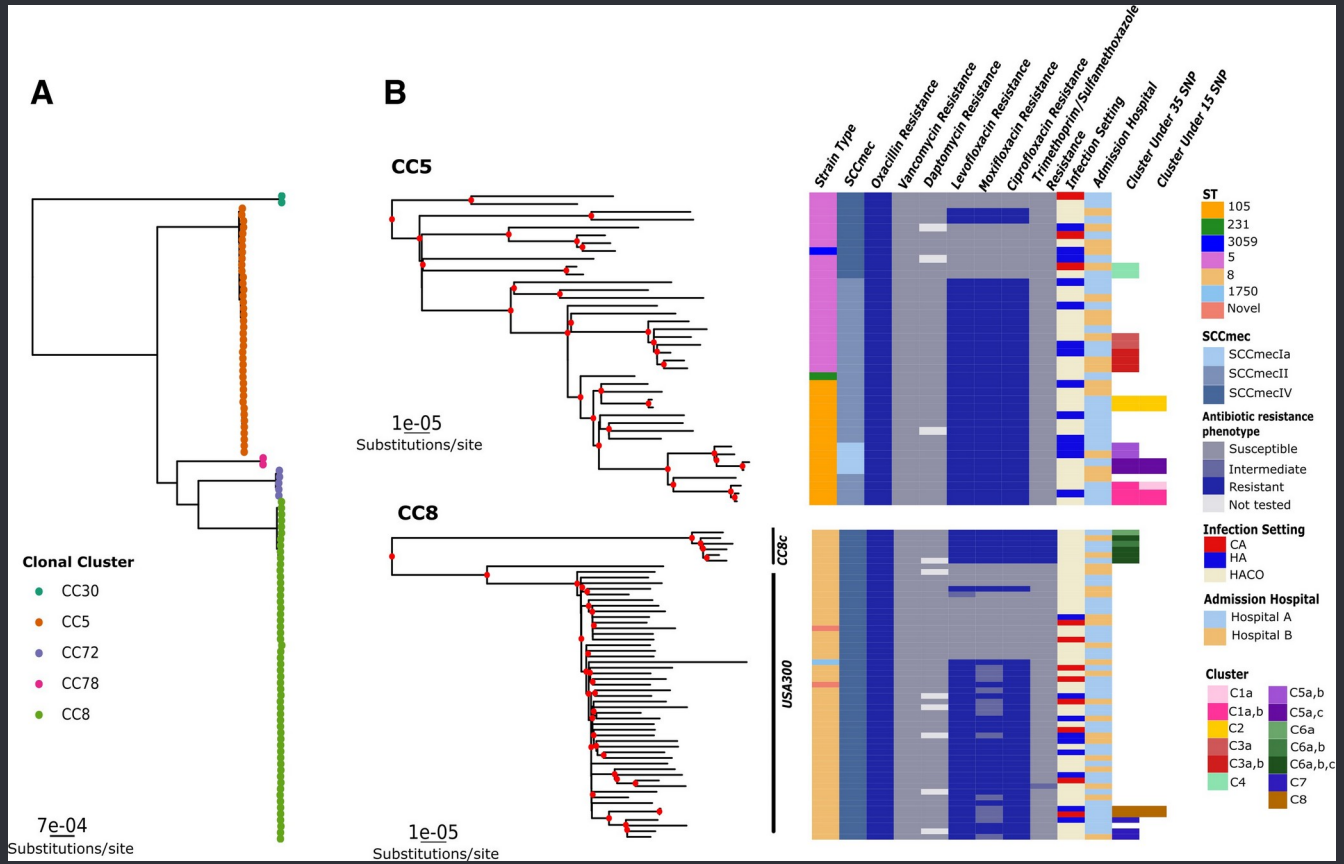
- SNP/InDel Calling
  - Comparing samples against each other based on a reference
  - Useful for outbreaks and understanding selective pressure

- Pangenome Construction
  - Pangenome – All genes present in set of samples
    - Core-genome – those genes shared in all samples
    - Accessory-genome – those genes found in a subset of samples

Phylogenetic Trees

- Infer evolutionary relationships between samples
- Can be built from:
  - Sketches or 16s rRNA (low resolution, very quick)
    - Quick identification of outliers
  - Core-SNP alignment (high resolution, slow)
    - Outbreak investigation with low diversity
  - Core-genome alignment (high resolution, slow)
    - Diverse samples from the same species
- Can be annotated with metadata
  - E.g. host, disease, AMR, virulence, etc

Clonal spread of MRSA causing bloodstream infections



Source:
Talbot et al., CID 2022

4



BACTOPIA

COMPLETE ANALYSIS OF BACTERIAL GENOMES

Let’s learn how Bactopia can help

Bactopia is a robust pipeline for bacterial genome analysis



Accessible

Bactopia is and open source and freely available from Conda, Docker and Singularity. Bactopia has been downloaded from users around the world.



Comprehensive

Bactopia is a start-to-finish pipeline which includes numerous tools and workflows commonly used for bacterial genome analysis.



Portable

With a simple profile change you can go from processing genomes on your laptop to an HPC system or any of the major cloud providers (AWS, GCP, Azure).



Reproducible

Bactopia was developed following nf-core best practices which ensures a robust pipeline with strict version control and an extensive audit trail.



Resilient

More than 100 tests, testing 10,000+ variables, assist in identifying potential bugs and downstream changes, before users are affected.



Scalable

Bactopia allows you to easily scale from a few genomes to thousands of genomes. For example, processing 67,000 genomes in 5 days on AWS.

With Bactopia, in just a few steps, you can go from raw sequencing to an extensive set of results to investigate.

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

BACTOPIA TOOLS

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

5

People behind Bactopia

Let’s put some faces to Bactopia



While I am the sole developer of Bactopia, my colleagues have played an incredibly supportive role in the continued development of of Bactopia.

A few of those colleagues



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 lab at WPHL. She has played a tremendous and critical role in helping to expand Bactopia into public health.



Chayse Rowley
WPHL
Senior Microbiologist

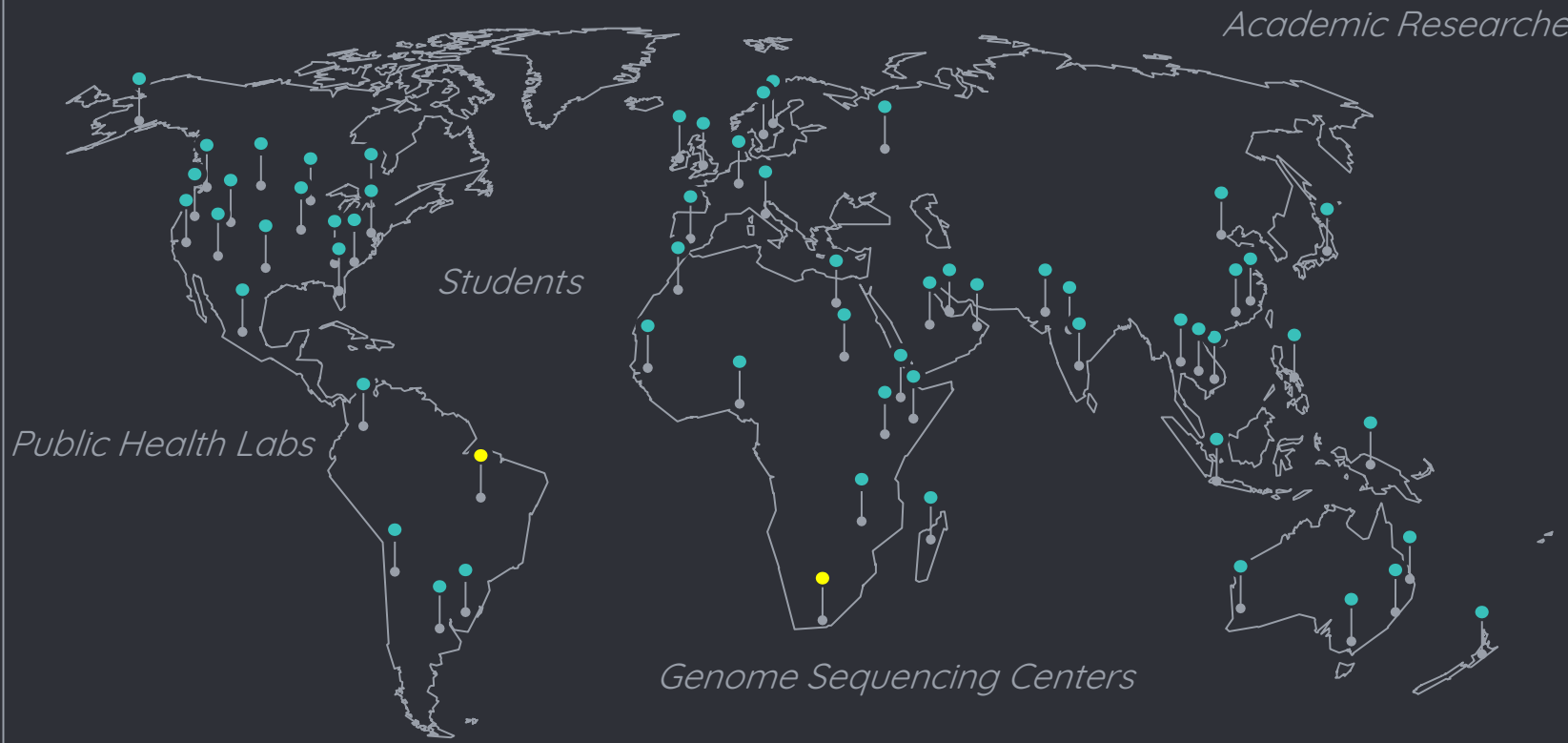
Chayse is involved in new sequencing projects at WPHL. She has helped identify novel ways to use Bactopia within its existing framework.

Many scientists around the world that provide feedback



Who are those scientists around the world?

Bactopia Users Across the Globe



Many of these dots are in response to direct communications

35,000 visitors

Many users around the world visiting the docs

60+ citations

Used in all sorts of bacterial genomic studies

1,000,000+ downloads

That’s nearly twice the population of Wyoming!

6

Wrap Up

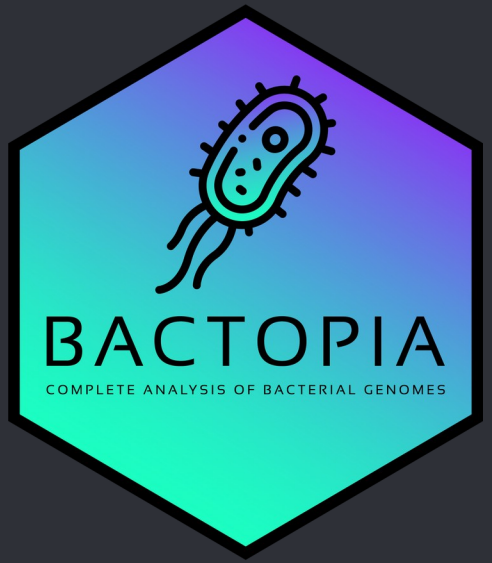
Let’s close this seminar out!

WGS and Bacterial Genomics

- Many useful reasons to sequence a genome
- Illumina and ONT sequencing are common for bacteria
- Main genomic analyses are standardized
- Numerous downstream analyses

Bactopia for your bacterial analysis needs

- Bactopia is a reproducible pipeline that automates many bacterial analyses
- Supports all bacterial species
- Completely free and open-source, actively maintained for 5+ years



Acknowledgements

All the developers of open-source software used by Bactopia, and the many users regularly providing feedback and suggestions

WPHL & Emory colleagues, the open-source community: Bioconda, nf-core and Nextflow

The many scientists around the world keeping the wheels moving



Presentation template by SlidesCarnival

Any questions?

Anything Bactopia, bioinformatics, public health, Wyoming, sequencing, open-source development, etc…