Bacterial Genome Analysis Workflows: Staphopia and Bactopia

Timothy Read PhD
Emory University School of Medicine
tread@emory.edu
@tdread_emory

No financial conflicts

Abstract

“Technology innovations in genomics that reduce sequencing time and cost have created new opportunities for biological research. Since the mid 2000’s, large scale sequencing of bacterial genomes using Illumina technology has become a standard for pathogen epidemiology studies, resulting in very large data sets for some species. Genome data has been generated faster than can be conveniently analyzed and integrated with results of classical experimental approaches to microbiology. We became interested in the task of analyzing tens of thousands of genomes of the pathogenic bacterium *Staphylococcus aureus* in the public domain. We created a workflow called the Staphopia Analysis Pipeline (StAP), using Nextflow software, to automate processing (e.g QC, genome assembly, annotation, genotype) using open source bioinformatic tools and databases. The pipeline was encapsulated in a Docker container to allow it to be deployed across software platforms. We collaborated with the Cancer Genomics Cloud and used their Seven Bridges-based platform to process >40,000 genomes in a 10 day period in November 2017. A public instance of StAp was also created at CGC to allow anonymous users to run StAP on their own data. In order to share the results of our analysis with other researches we created the Staphopia database, with public APIs for data download of > 350 endpoints and an R package to enhance data analysis. We have been using Staphopia as both a resource to generate hypotheses (“top-down approaches”) and also to understand how results from lab studies relate to the species as a whole (“bottom-up”).

An example of the former analysis is looking at the co-occurrence of resistance to mupirocin and fluoroquinolones with methicillin resistance (MRSA). An example of bottom-up approaches has been mapping the distribution SNPs found to be associated with intermediate vancomycin resistance selected in the laboratory across different subtypes of *S. aureus*. We have recently created a new series of pipelines called Bactopia, built on the experiences learned from Staphopia but generalizable to any bacterial species. Bactopia consists of a dataset setup step (Bactopia Datasets) where a series of customizable dataset are created for the species on interest. The Bactopia Analysis Pipeline performs analyses based on the dataset downloaded and outputs the processed data to a structured directory format. We have created a series of Bactopia Tools that perform specific post-processing on some or all of the genomes processed. These include pan-genome analysis, computing average nucleotide identity between samples, extracting and profiling the 16S genes and taxonomic classification" via gtdb. We have performed a Bactopia demonstration project on 1664 public *Lactobacillus* genomes in SRA in December 2019.

Staphopia

2006

Illumina GA II

2010

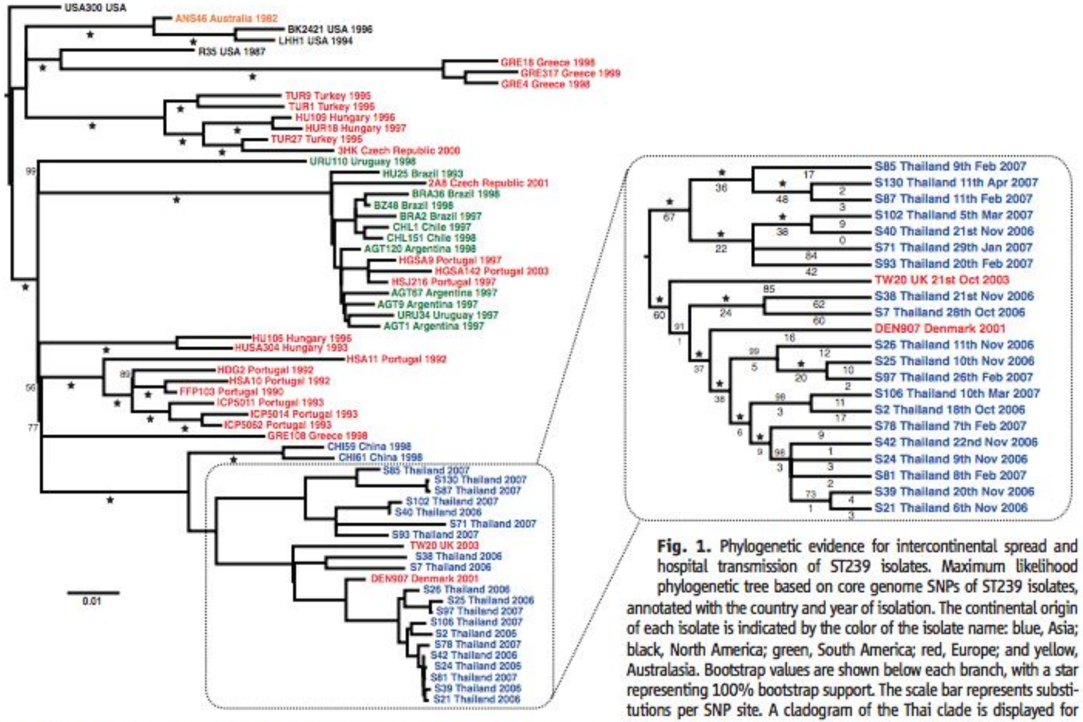
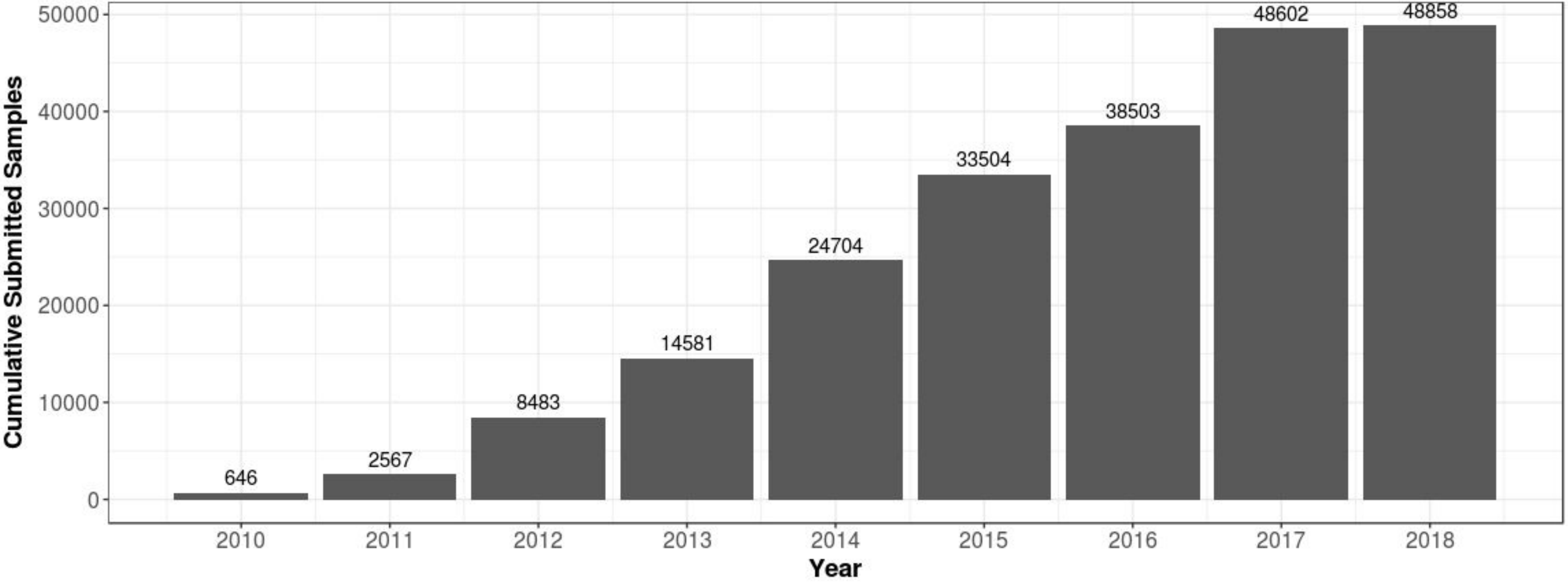


Fig. 1. Phylogenetic evidence for intercontinental spread and hospital transmission of ST239 isolates. Maximum likelihood phylogenetic tree based on core genome SNPs of ST239 isolates, annotated with the country and year of isolation. The continental origin of each isolate is indicated by the color of the isolate name: blue, Asia; black, North America; green, South America; red, Europe; and yellow, Australasia. Bootstrap values are shown below each branch, with a star representing 100% bootstrap support. The scale bar represents substitutions per SNP site. A cladogram of the Thai clade is displayed for greater resolution with bootstrap values (above the branch), number of distinguishing SNPs (below the branch), and isolates labeled with date of isolation, where known.

Harris, S. R. et al. Evolution of MRSA During Hospital Transmission and Intercontinental Spread. Science 327, 469–474 (2010)

2018

Cumulative total of publicly available sequenced *S. aureus* samples (N = 48,858) by year (2010-present).



Why would you want to analyze WGS?

1. Immediate payoff - e.g antibiotic resistance profile of individual isolates

2. Retrospectively look for trends, when important genes appeared etc

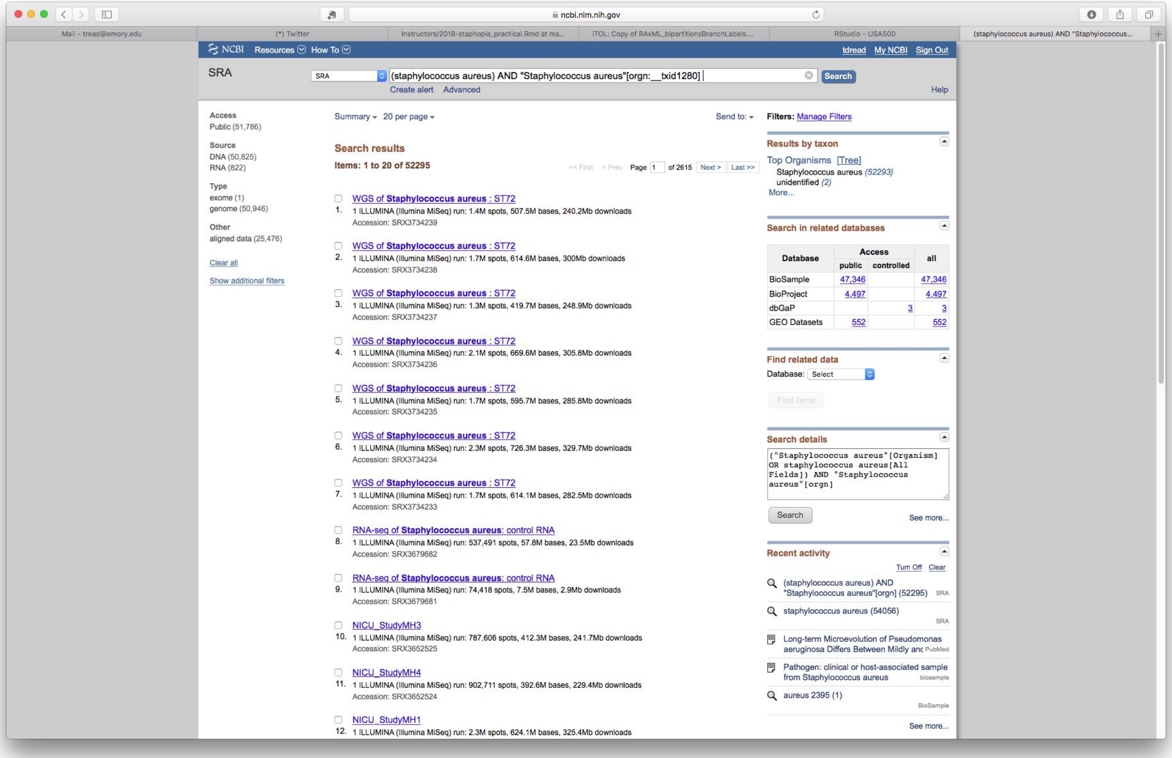
3. Going forward: identify potential outbreaks, changing patterns of antibiotic resistance

Success brings Challenges

Unassembled raw data

Impossible to screen SNPs/ genetic variants

Don't know about antibiotic resistance genes, MLST etc,



SRA

SRA (staphylococcus aureus) AND "Staphylococcus aureus"[orgn:__txid1280]

Access
Public (51,786)

Source
DNA (50,825)
RNA (822)

Type
exome (1)
genome (50,946)

Other
aligned data (25,476)

Clear all

Show additional filters

Search results

Items: 1 to 20 of 52295

<< First < Prev Page 1 of 2615 Next > Last >>

1. □ WGS of **Staphylococcus aureus** : ST72
1 ILLUMINA (Illumina MiSeq) run: 1.4M spots, 507.5M bases, 240.2Mb downloads
Accession: SRX3734239

2. □ WGS of **Staphylococcus aureus** : ST72
1 ILLUMINA (Illumina MiSeq) run: 1.7M spots, 614.6M bases, 300Mb downloads
Accession: SRX3734238

3. □ WGS of **Staphylococcus aureus** : ST72
1 ILLUMINA (Illumina MiSeq) run: 1.3M spots, 419.7M bases, 248.9Mb downloads
Accession: SRX3734237

4. □ WGS of **Staphylococcus aureus** : ST72
1 ILLUMINA (Illumina MiSeq) run: 2.1M spots, 669.6M bases, 305.8Mb downloads
Accession: SRX3734236

5. □ WGS of **Staphylococcus aureus** : ST72
1 ILLUMINA (Illumina MiSeq) run: 1.7M spots, 595.7M bases, 285.8Mb downloads
Accession: SRX3734235

6. □ WGS of **Staphylococcus aureus** : ST72
1 ILLUMINA (Illumina MiSeq) run: 2.3M spots, 726.3M bases, 329.7Mb downloads
Accession: SRX3734234

7. □ WGS of **Staphylococcus aureus** : ST72
1 ILLUMINA (Illumina MiSeq) run: 1.7M spots, 614.1M bases, 282.5Mb downloads
Accession: SRX3734233

8. □ RNA-seq of **Staphylococcus aureus**: control RNA
1 ILLUMINA (Illumina MiSeq) run: 537,491 spots, 57.8M bases, 23.5Mb downloads
Accession: SRX3679682

9. □ RNA-seq of **Staphylococcus aureus**: control RNA
1 ILLUMINA (Illumina MiSeq) run: 74,418 spots, 7.5M bases, 2.9Mb downloads
Accession: SRX3679681

10. □ NICU_StudyMH3
1 ILLUMINA (Illumina MiSeq) run: 787,606 spots, 412.3M bases, 241.7Mb downloads
Accession: SRX3652525

11. □ NICU_StudyMH4
1 ILLUMINA (Illumina MiSeq) run: 902,711 spots, 392.6M bases, 229.4Mb downloads
Accession: SRX3652524

12. □ NICU_StudyMH1
1 ILLUMINA (Illumina MiSeq) run: 2.3M spots, 624.1M bases, 325.4Mb downloads

Results by taxon

Top Organisms [Tree]
Staphylococcus aureus *(52293)*
unidentified *(2)*
More...

Search in related databases

Database	Access		all
	public	controlled	
BioSample	47,346		47,346
BioProject	4,497		4,497
dbGaP		3	3
GEO Datasets	552		552

Find related data

Database: Select

Search details

```
("Staphylococcus aureus"[Organism]
OR staphylococcus aureus[All
Fields]) AND "Staphylococcus
aureus"[orgn]
```

Recent activity

Turn Off Clear

(staphylococcus aureus) AND "Staphylococcus aureus"[orgn] (52295) SRA

staphylococcus aureus (54056) SRA

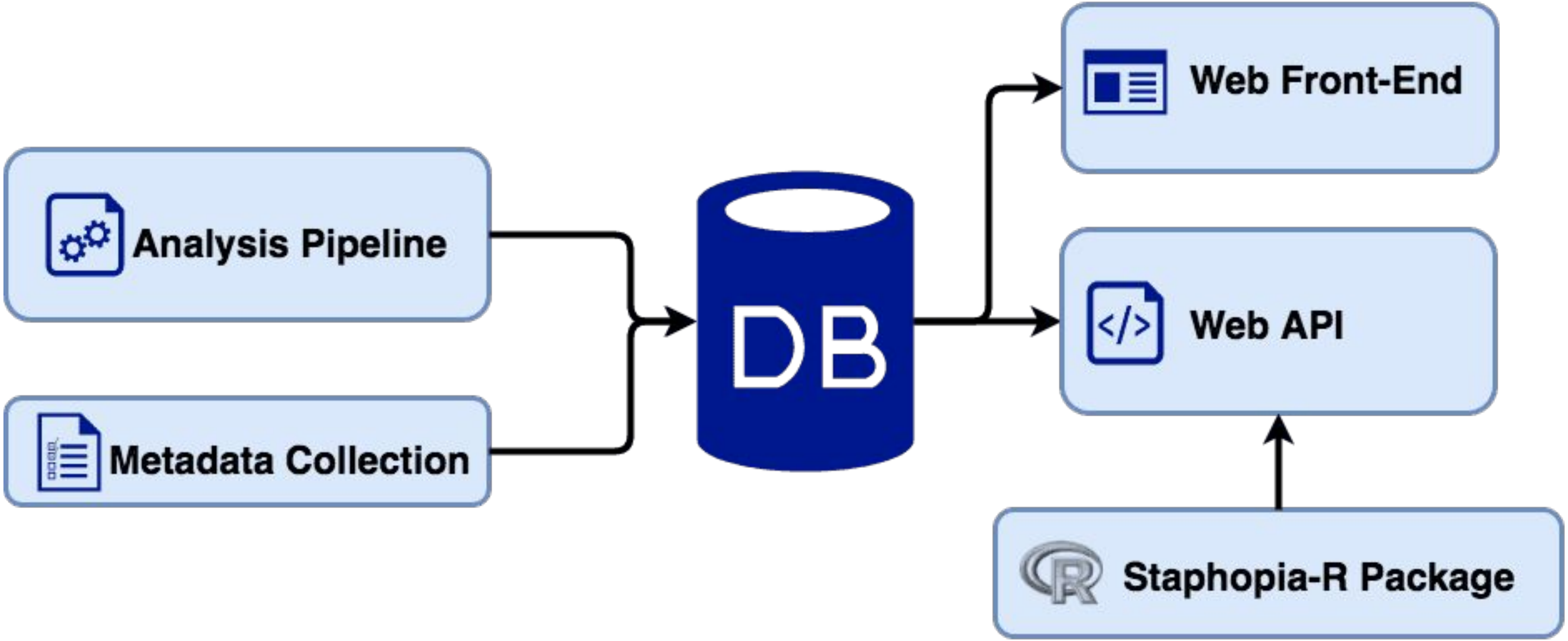
Long-term Microevolution of Pseudomonas aeruginosa Differs Between Mildly anc PubMed

Pathogen: clinical or host-associated sample from Staphylococcus aureus biosample

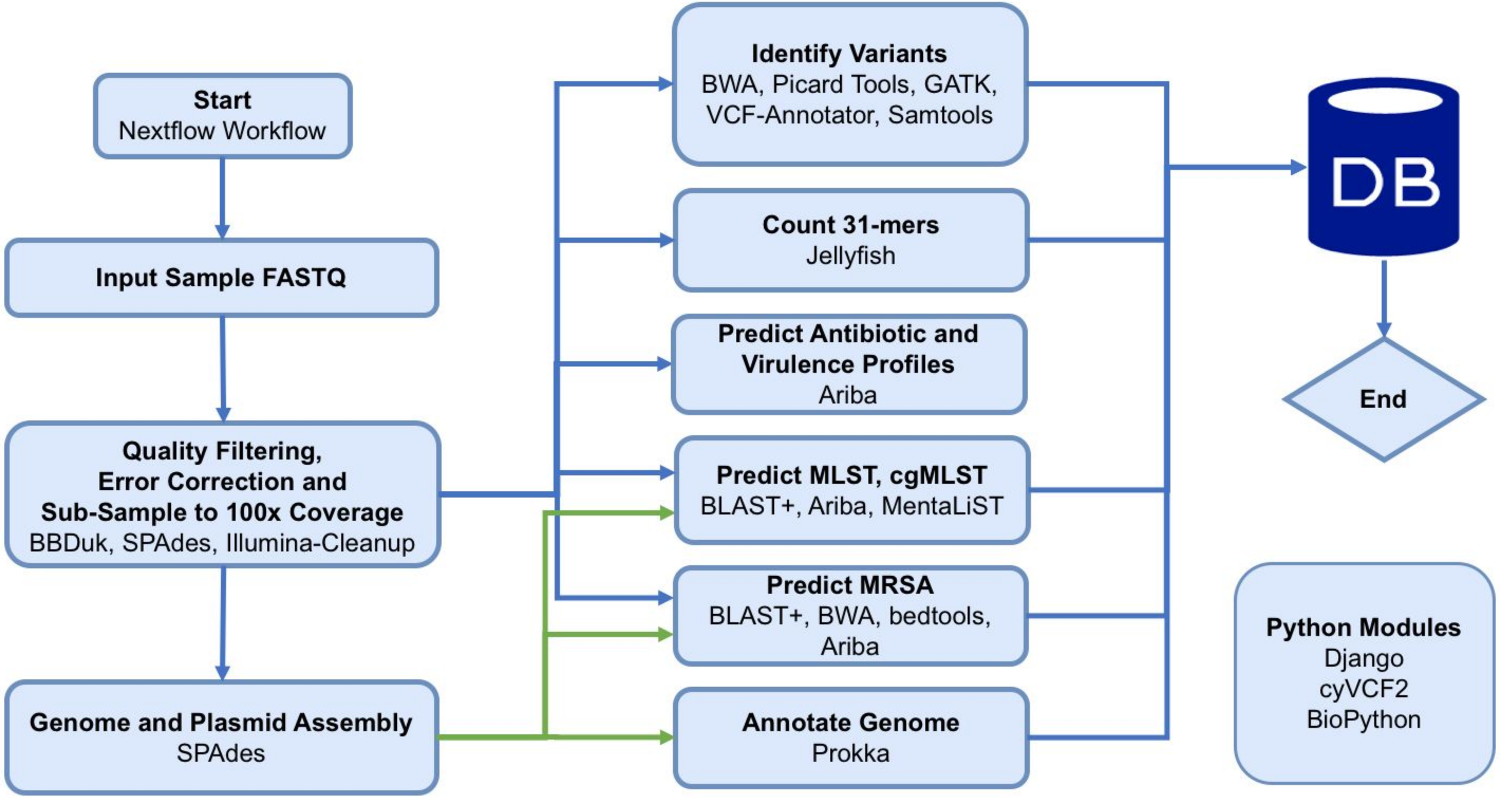
aureus 2395 (1) BioSample

It's all there, you just have to organize it yourself!

Our solution, Staphopia



The Staphopia Analysis Pipeline



Start
Nextflow Workflow

Input Sample FASTQ

Quality Filtering, Error Correction and Sub-Sample to 100x Coverage
BBDuk, SPAdes, Illumina-Cleanup

Genome and Plasmid Assembly
SPAdes

Identify Variants
BWA, Picard Tools, GATK, VCF-Annotator, Samtools

Count 31-mers
Jellyfish

Predict Antibiotic and Virulence Profiles
Ariba

Predict MLST, cgMLST
BLAST+, Ariba, MentaLiST

Predict MRSA
BLAST+, BWA, bedtools, Ariba

Annotate Genome
Prokka

Python Modules
Django
cyVCF2
BioPython

“*Public Library*” for *S. aureus* Genomics

Our Problem

- Each sample takes ~50 minutes to complete the analysis.
    - We can convert this to total cpu time: N * CPU_MINUTES
        - Our case: 44k * 50 = 2,203,350 CPU Minutes

- Translation to real time:
    - Only 1 sample at a time, equates to 4+ years of real time
    - Our current infrastructure:
        - Maxing Out Each Server: 18 at a time, or 90 days
            - Servers could not be used by others!
        - Realistically: 4 - 8 at a time, or 200 - 400 days

Got lucky! Asked to test a human genomics cloud platform!

- In August 2017, we became aware of the Cancer Genomics Cloud (CGC) platform

- Developed for human genomics, unsure whether it could handle microbial genomics
    - Few very large jobs vs Many very small jobs

Cancer Genomics Cloud (CGC) Platform

- http://www.cancergenomicscloud.org/
- Executes Docker based workflows on Amazon Web Services (AWS) cloud
- Intuitive front-end is easy to use
- API packages included for R and Python
  - Work amazingly!

A realistic opportunity to process 44k genomes

- Estimated ~$2,200 and ~3 weeks to analyse all 44k S. aureus genomes
  - Or, $0.05 per genome

- Most importantly, presented the opportunity to completely rewrite the analysis pipeline
  - Ruffus hadn't been updated in over two years, looked to be abandoned
    - Although, recently updated November 4th, 2018
  - Some outdated methods from 2015

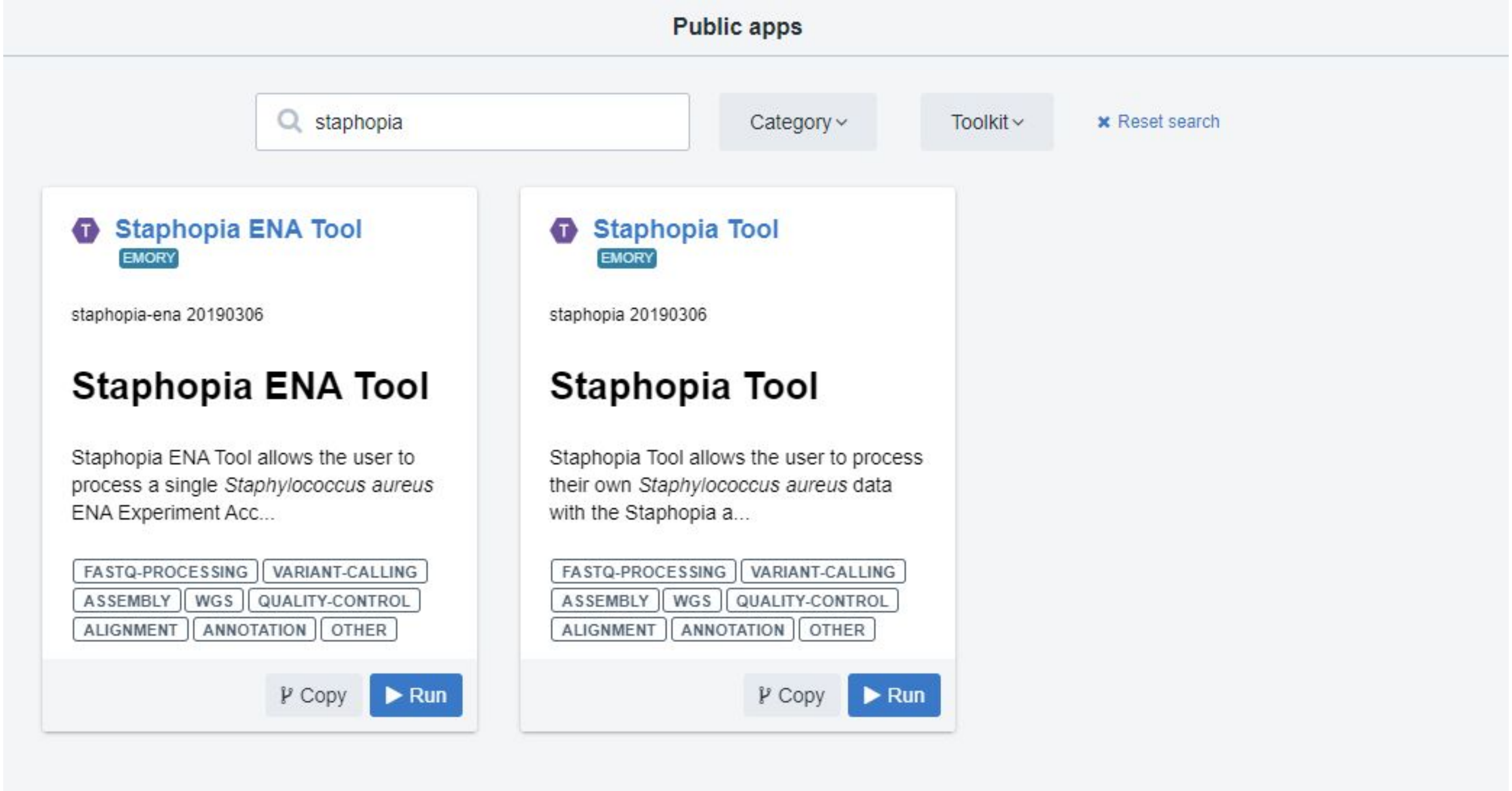
- Rewrote the pipeline using Nextflow with more up to date methods
  - Dropped large tarball for BioConda installable packages
  - Dockerized the pipeline

Successfully processed 44k genomes!

- It only took only 11 days to complete
  - Instance limit was increased temporarily to 200
- Only 202 jobs failed:
  - Most were due to ENA timeouts or SPAdes asking for > 32GB memory
- Pushed the CGC platform to its limits
  - Helped uncover a few a bugs
  - Provided a great use case for CGC
- Almost 10TB of data was generated
  - Fortunately this was downloaded as jobs completed

Staphopia available as a public app on CGC

- Uses the Staphopia Docker image
- Any user can run Staphopia on the CGC either from FASTQ or directly from ENA



Staphopia ENA Tool

staphopia-ena 20190306

Staphopia ENA Tool

Staphopia ENA Tool allows the user to process a single *Staphylococcus aureus* ENA Experiment Acc...

FASTQ-PROCESSING VARIANT-CALLING ASSEMBLY WGS QUALITY-CONTROL ALIGNMENT ANNOTATION OTHER

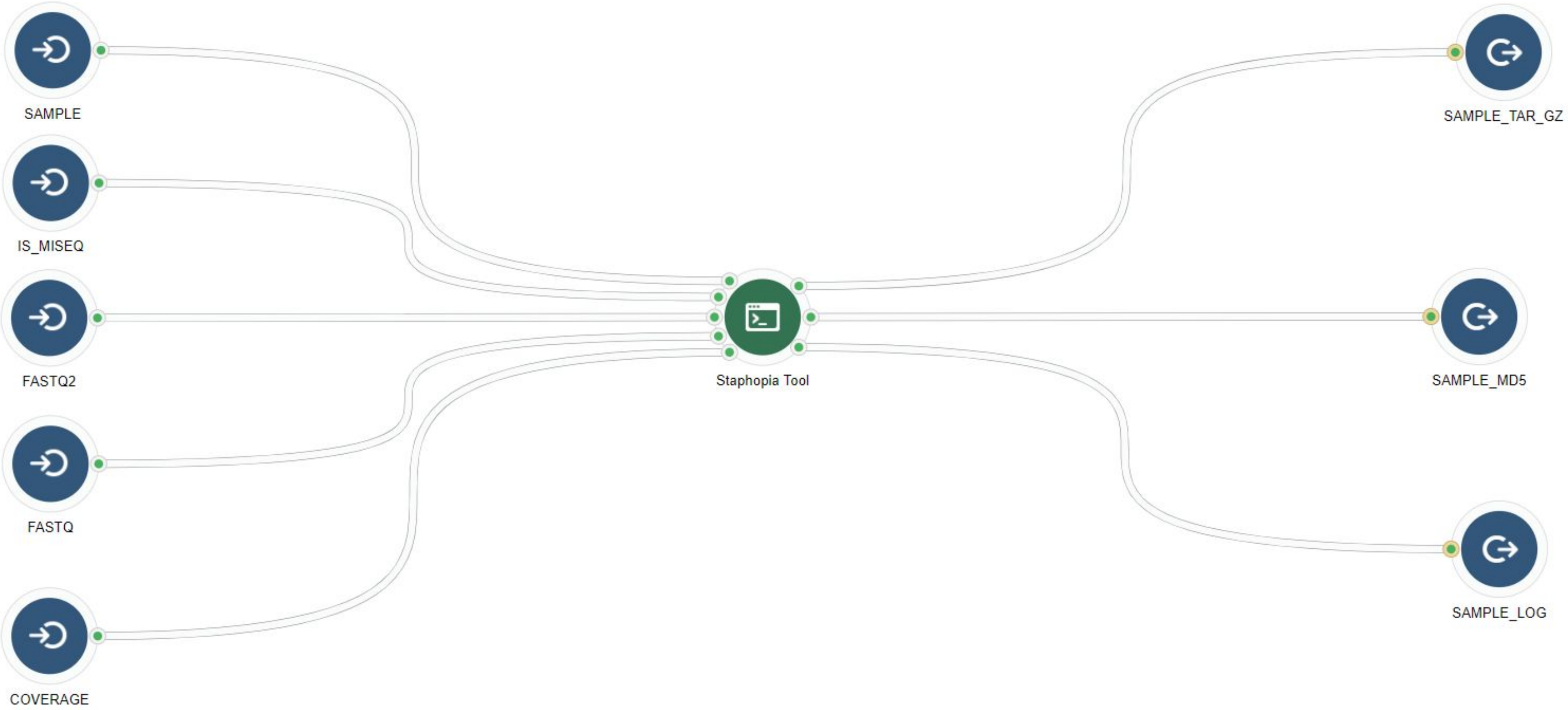
Staphopia Tool

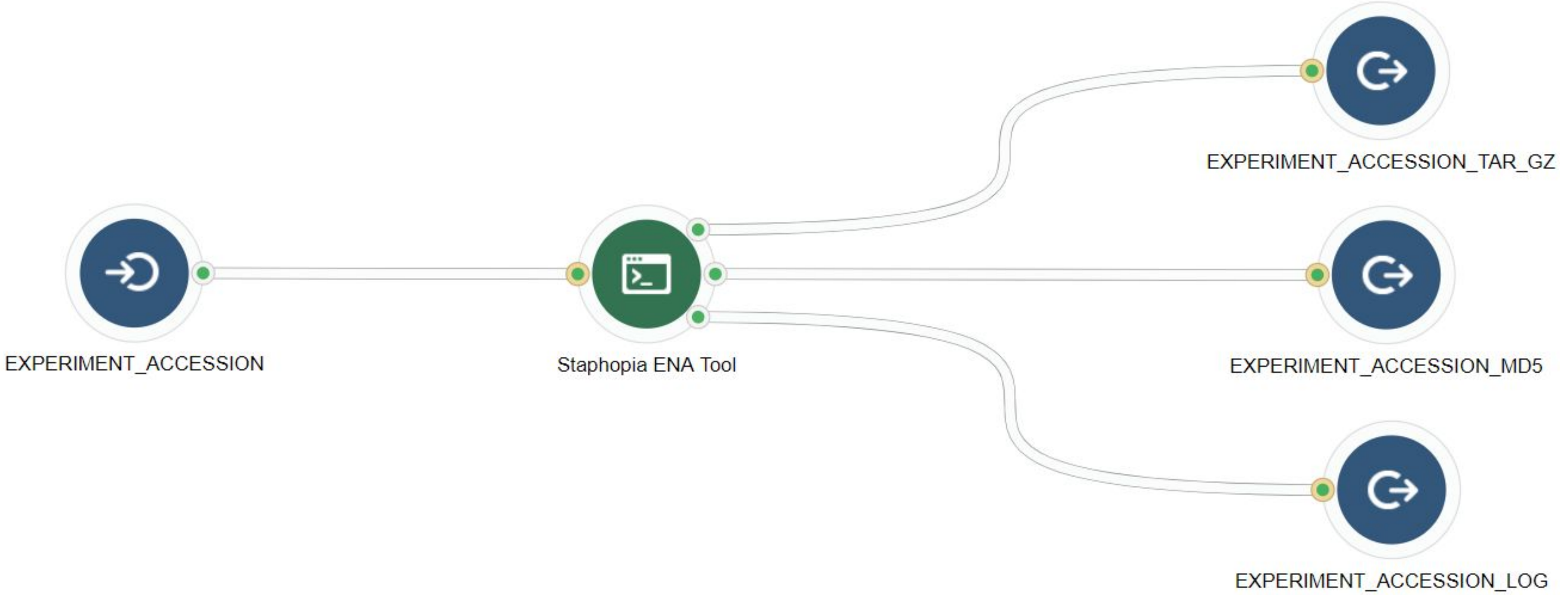
staphopia 20190306

Staphopia Tool

Staphopia Tool allows the user to process their own *Staphylococcus aureus* data with the Staphopia a...

FASTQ-PROCESSING VARIANT-CALLING ASSEMBLY WGS QUALITY-CONTROL ALIGNMENT ANNOTATION OTHER


SAMPLE
IS_MISEQ
FASTQ2
FASTQ
COVERAGE
Staphopia Tool
SAMPLE_TAR_GZ
SAMPLE_MD5
SAMPLE_LOG


EXPERIMENT_ACCESSION
Staphopia ENA Tool
EXPERIMENT_ACCESSION_TAR_GZ
EXPERIMENT_ACCESSION_MD5
EXPERIMENT_ACCESSION_LOG

2017 analysis breakdown

- In November 2017 there were **43,972** Illumina *S. aureus* projects
- **42,949** were uncontaminated *S. aureus* genomes
- **42,337** of these genomes were assigned to 1,090 STs (of 4,466 in the PubMLST database)

Petit RA 3rd, Read TD. Staphylococcus aureus viewed from the perspective of 40,000+ genomes. PeerJ [Internet]. 2018 Jul 12;6:e5261. Available from: http://dx.doi.org/10.7717/peerj.5261 PMCID: PMC6046195

Other notable highlights of Staphopia

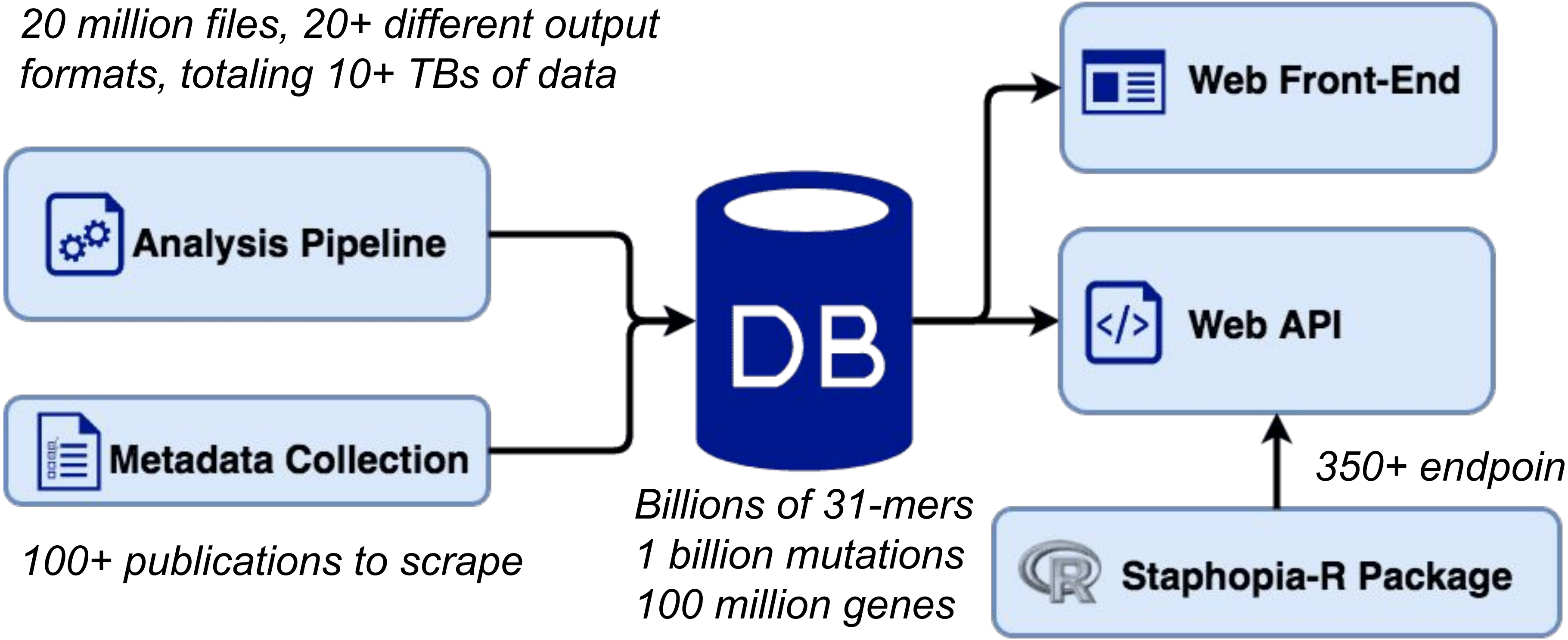
- Built in cross-validation of methods for 44k samples
  - MLST - Mapping vs k-mer vs Alignments
  - SCCmec - Mapping vs In-silico PCR
- Evidence for sub-sampling to 100x coverage
  - No need for 300x coverage, when 100x will get you the same results except faster
- Extensive effort to link metadata to genomes
  - Text-mining PDFs for accessions
- Real-world example of costs associated with genomic analysis in the cloud

Staphopia’s API and R Package

- Provides direct access to an extensive set of results
  - 350+ endpoints are accessible
  - Built using Django Rest Framework
- Created an R package, Staphopia-R
  - Allows programmatic access
  - Reproducible R Notebooks
  - Used to generate results in publication
    - https://github.com/staphopia/staphopia-paper/tree/master/analysis

Data management logistics…

20 million files, 20+ different output formats, totaling 10+ TBs of data

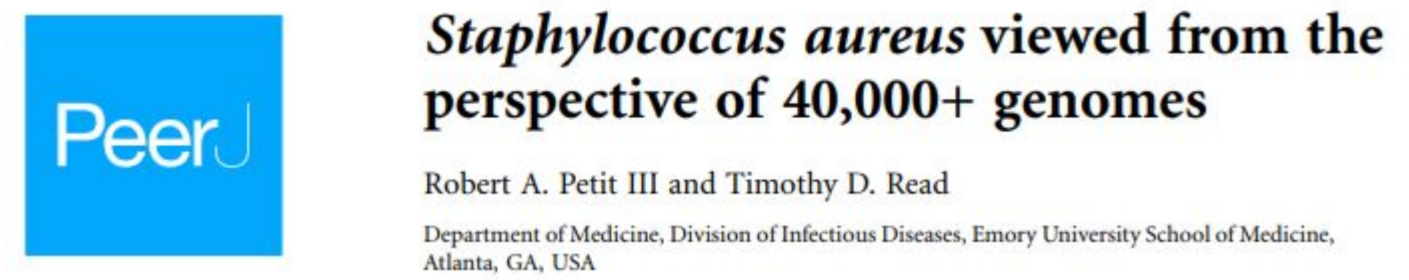


350+ endpoints

Billions of 31-mers
1 billion mutations
100 million genes

100+ publications to scrape

Multiple codebase changes



Staphylococcus aureus viewed from the perspective of 40,000+ genomes

Robert A. Petit III and Timothy D. Read

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

Published in PeerJ (July 2018): https://peerj.com/articles/5261/

What sorts of analyses can you do with these data?

Identifying MRSA

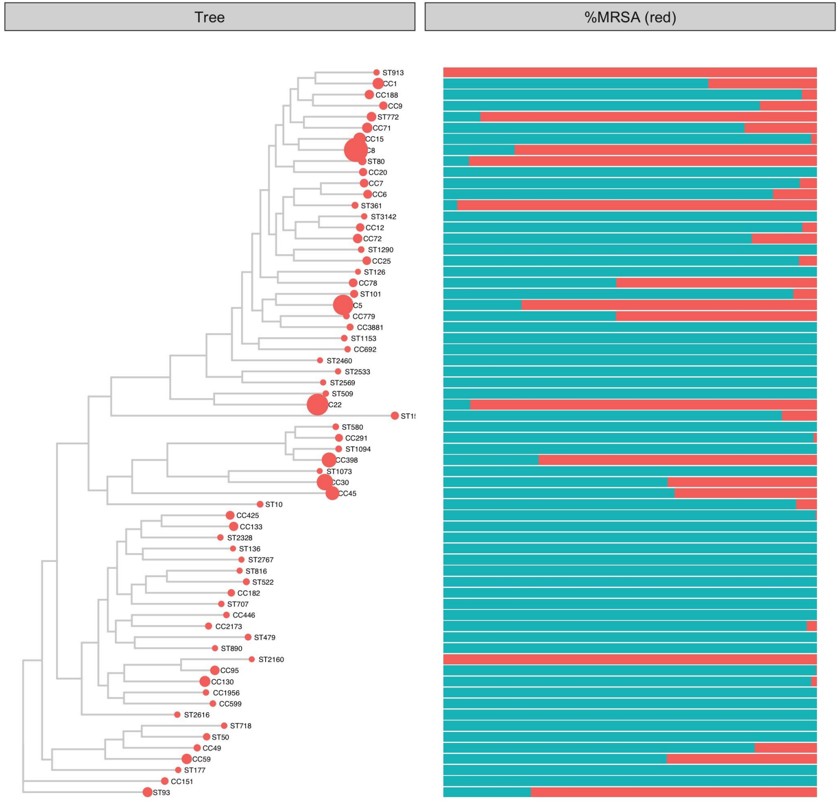
- Multiple approaches
    - *In silico* PCR
    - Protein alignments of mecA
    - Local assembly of mecA

Identifying MRSA

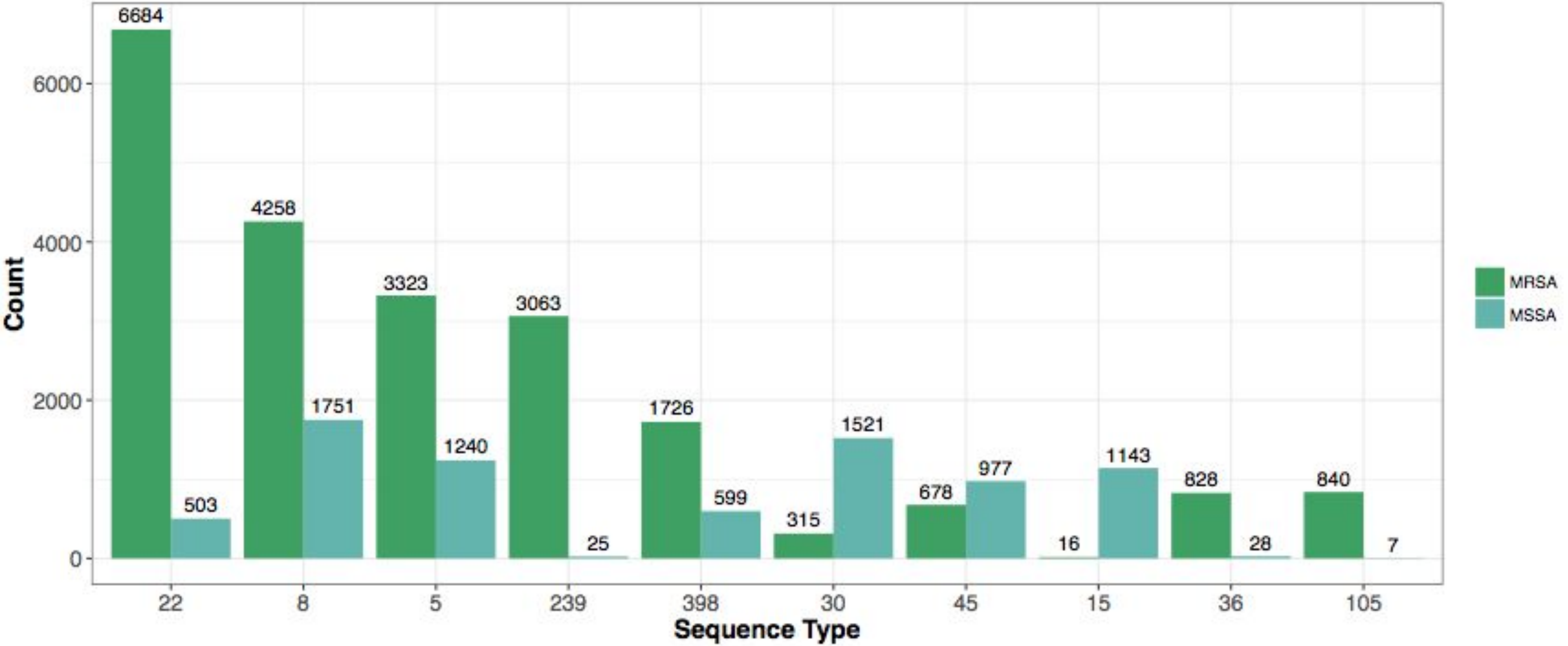
- Multiple approaches
    - *In silico* PCR → 26,743 samples
    - Protein alignments of mecA → 26,430 samples
    - Local assembly of mecA → 27,120 samples

- 64% (27,548) of samples predicted to be MRSA by at least one approach
    - 95% of samples agree between each each

MRSA across major Clonal Groups



MRSA is over-represented in top STs

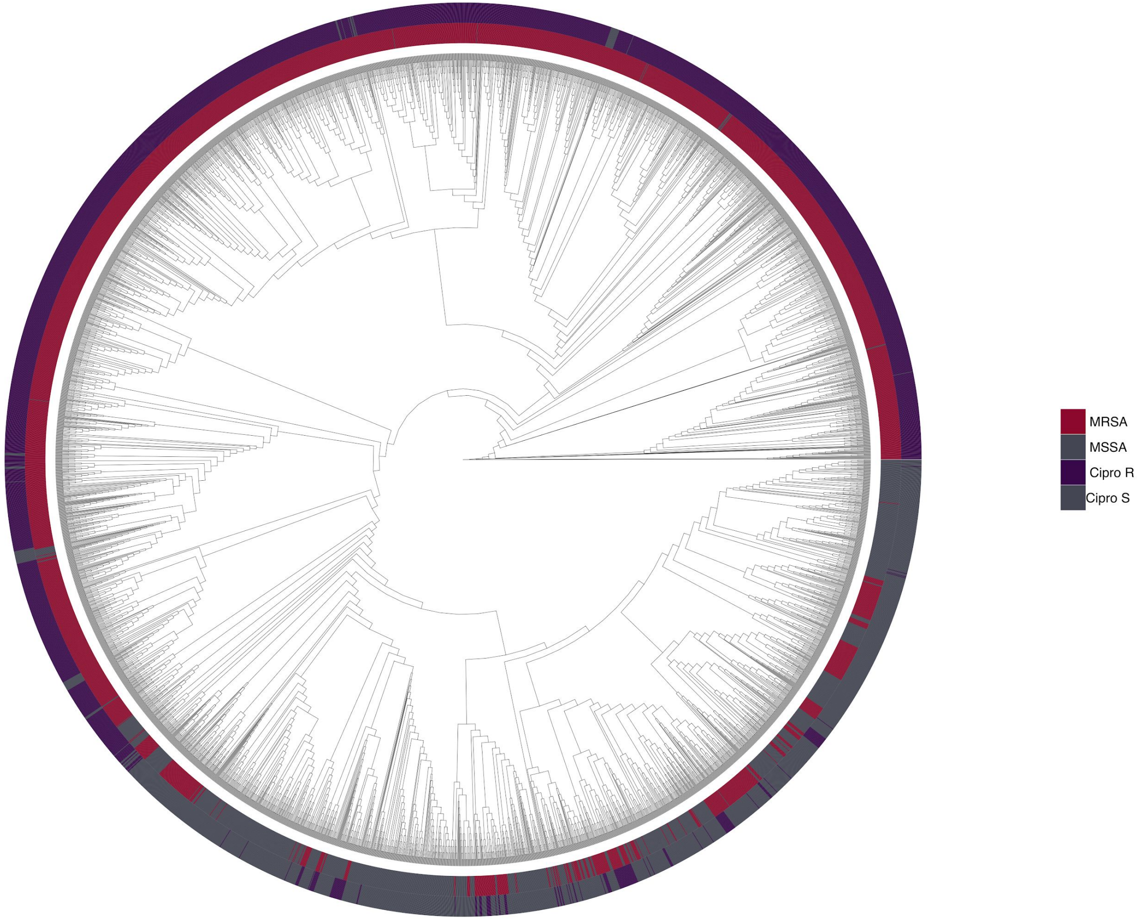


Relationships between MRSA, FQ and other inferred resistance

- 20,498 have one of five most common mutations in gyrA, grlA/B

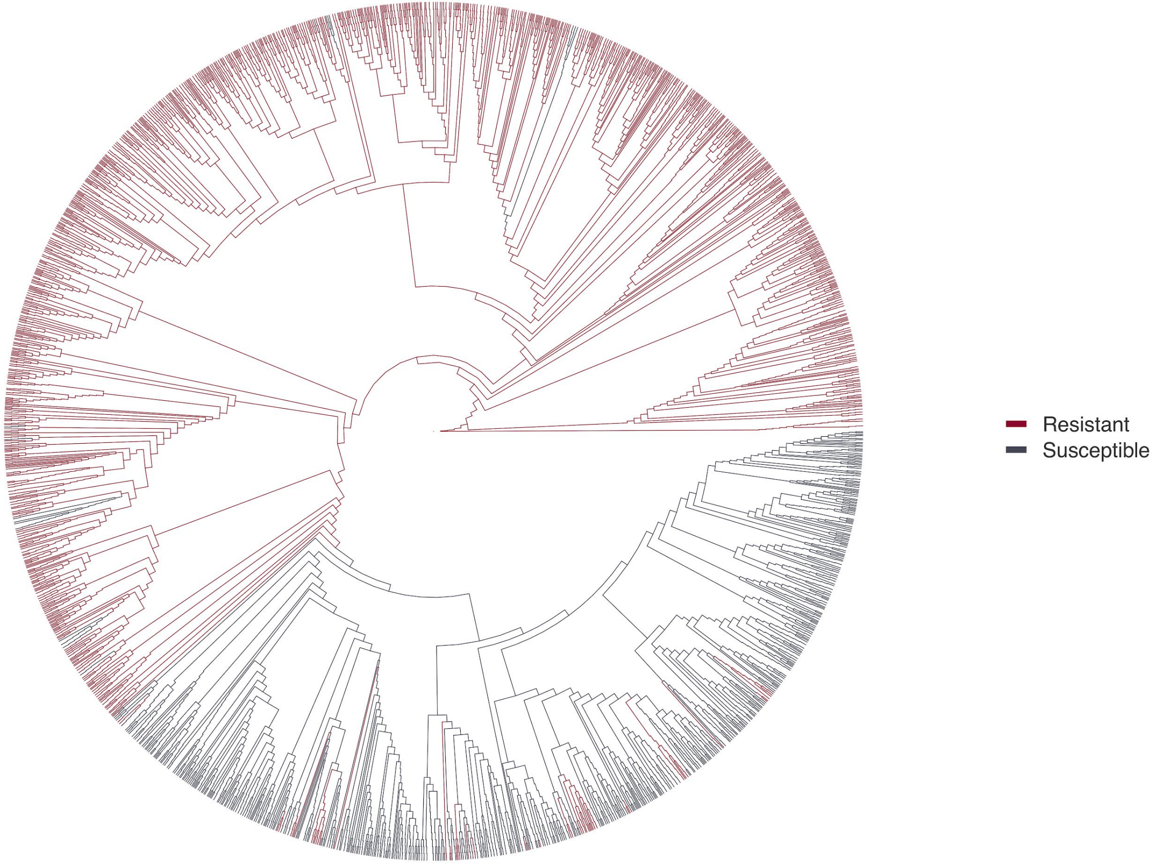
- What is the relationship between FQ resistance and MRSA?

ST5 FQ resistance



Mike Martin

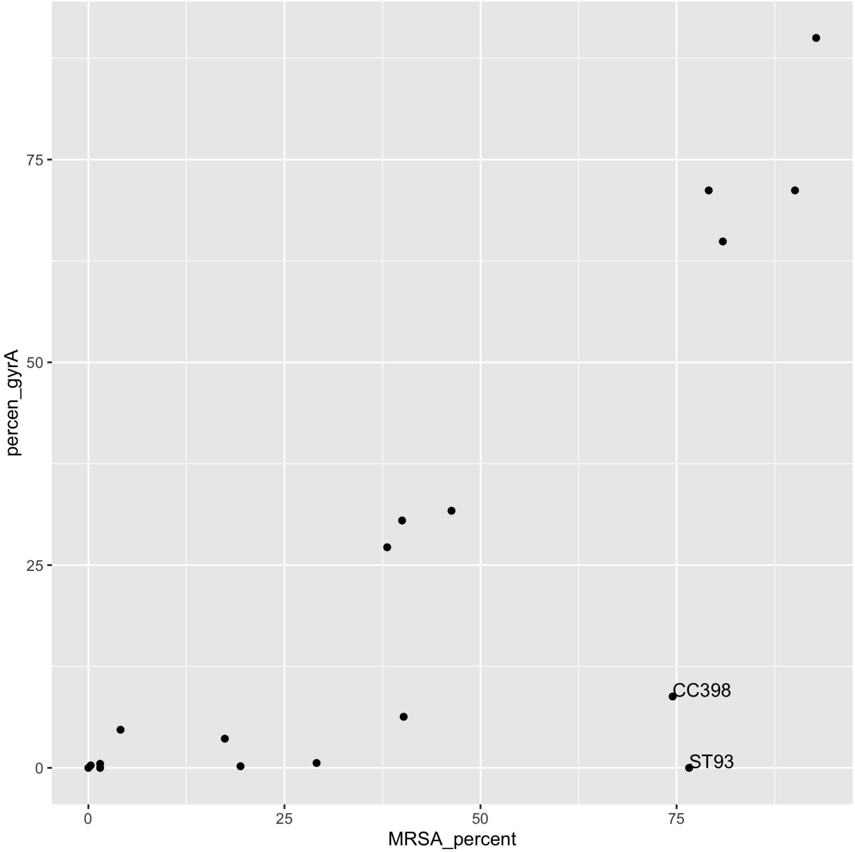
Ancestral state reconstruction shows multiple independent acquisitions of resistance



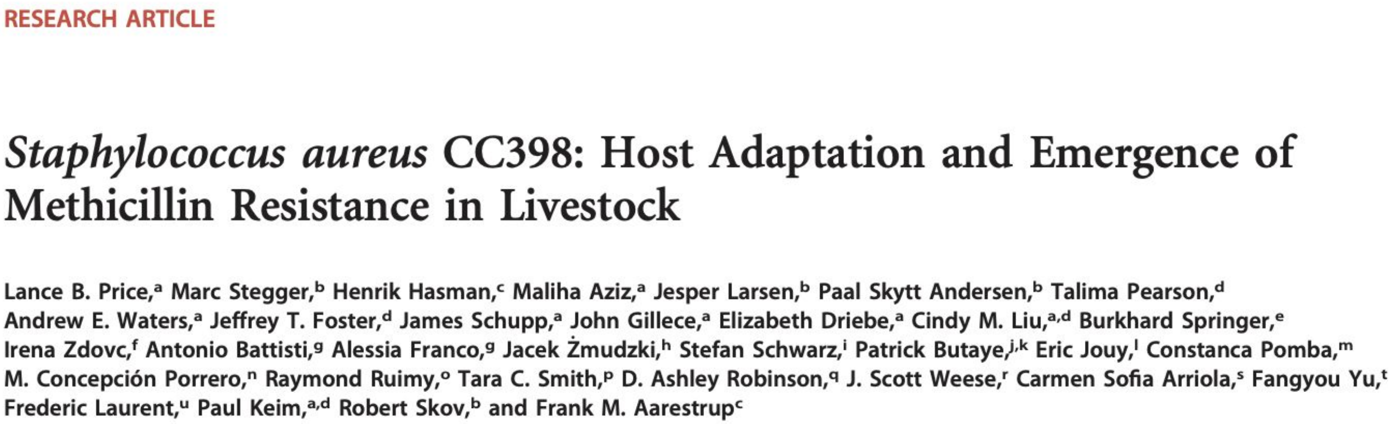
Mike Martin

FQ resistance mutation associated with MRSA

Each point represents a Clonal Complex with more than 300 strains



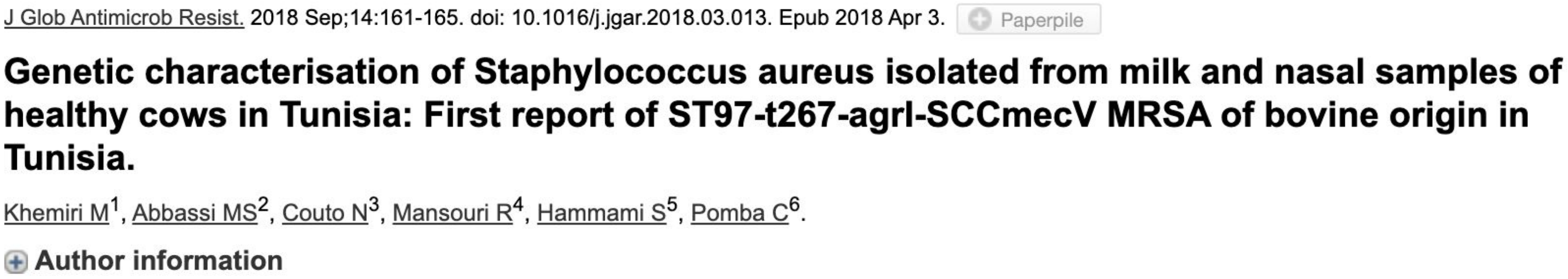
CC398 and ST93 associated with animals



RESEARCH ARTICLE

Staphylococcus aureus CC398: Host Adaptation and Emergence of Methicillin Resistance in Livestock

Lance B. Price,[a] Marc Stegger,[b] Henrik Hasman,[c] Maliha Aziz,[a] Jesper Larsen,[b] Paal Skytt Andersen,[b] Talima Pearson,[d] Andrew E. Waters,[a] Jeffrey T. Foster,[d] James Schupp,[a] John Gillece,[a] Elizabeth Driebe,[a] Cindy M. Liu,[a,d] Burkhard Springer,[e] Irena Zdovc,[f] Antonio Battisti,[g] Alessia Franco,[g] Jacek Żmudzki,[h] Stefan Schwarz,[i] Patrick Butaye,[j,k] Eric Jouy,[l] Constanca Pomba,[m] M. Concepción Porrero,[n] Raymond Ruimy,[o] Tara C. Smith,[p] D. Ashley Robinson,[q] J. Scott Weese,[r] Carmen Sofia Arriola,[s] Fangyou Yu,[t] Frederic Laurent,[u] Paul Keim,[a,d] Robert Skov,[b] and Frank M. Aarestrup[c]



J Glob Antimicrob Resist. 2018 Sep;14:161-165. doi: 10.1016/j.jgar.2018.03.013. Epub 2018 Apr 3. Paperpile

Genetic characterisation of Staphylococcus aureus isolated from milk and nasal samples of healthy cows in Tunisia: First report of ST97-t267-agrI-SCCmecV MRSA of bovine origin in Tunisia.

Khemiri M[1], Abbassi MS[2], Couto N[3], Mansouri R[4], Hammami S[5], Pomba C[6].

⊕ Author information

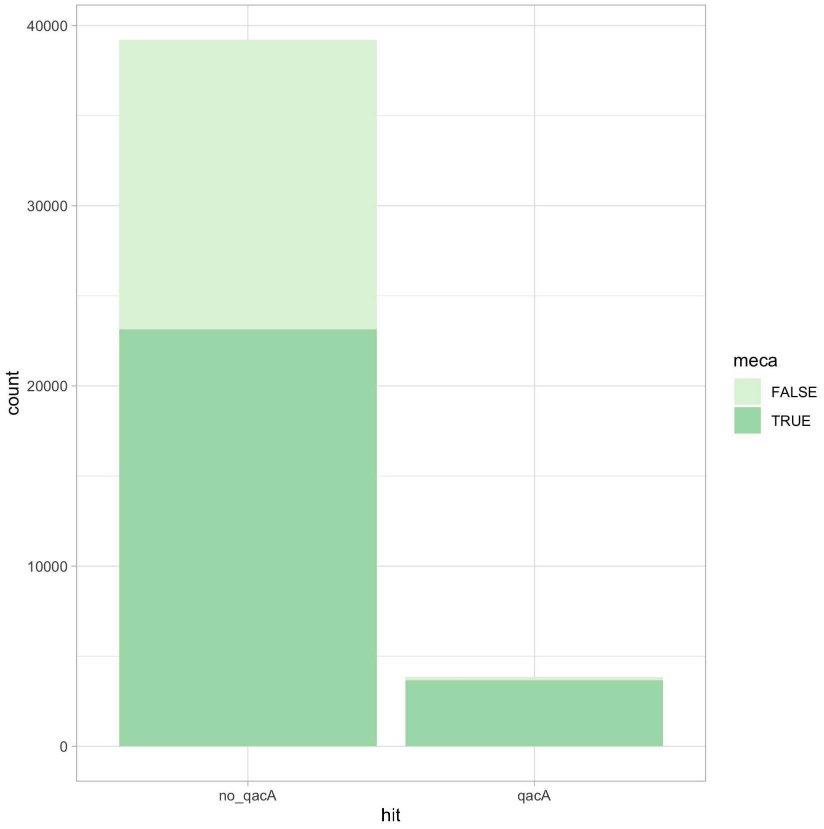
qacA

- Chlorhexidine (CHX) is a biocide used increasingly commonly for *S. aureus* decolonization

- *qacA* is usually a plasmid-borne gene that encodes a muli-drug efflux pump associated with reduced CHX resistance

- 3,371 strains have *qacA* gene

qacA versus MRSA



Mupirocin Resistance

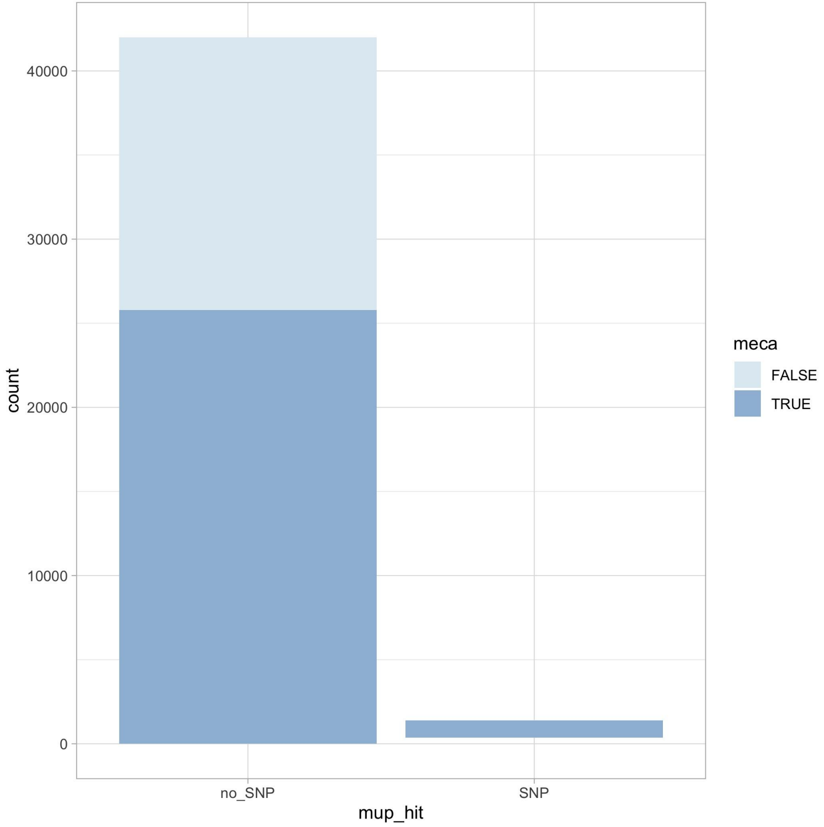
- Mupirocin is topical antibiotic that is a commonly used to treat Staph infections of the skin

- Yokoyama et al. 2018 reported a SNP associated with Mupirocin Resistance in the gene *ileS* (V588F)

- 1,402 strains have *ileS* V588F mutations

M. Yokoyama *et al.*, Epistasis analysis uncovers hidden antibiotic resistance-associated fitness costs hampering the evolution of MRSA. *Genome Biol.* **19**, 94 (2018).

Mupirocin versus MRSA



Bactopia

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 official channel (Bioconda, conda-forge, defaults, etc…)
   b. If not available, can I create a recipe?

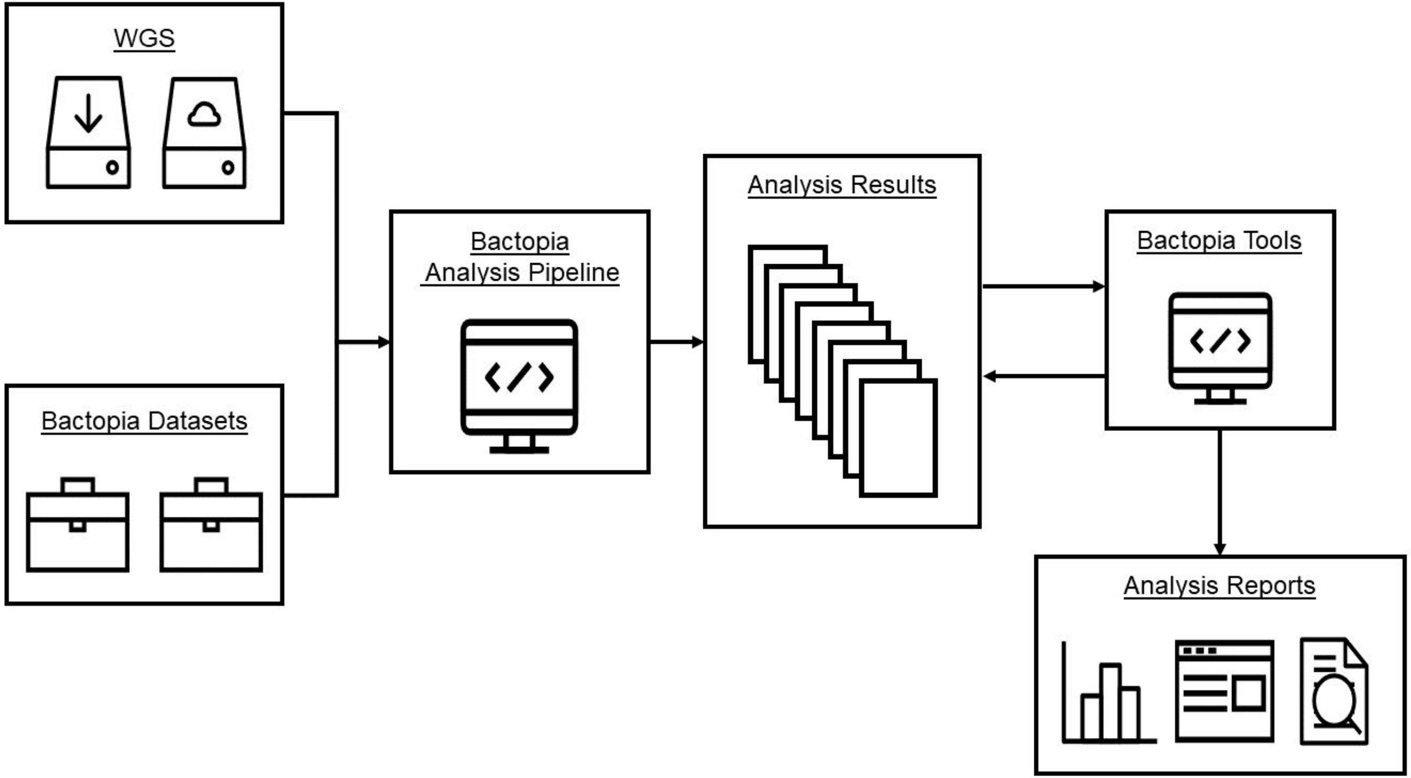
2. Flexible & Portable
   a. Fit your needs (100+ adjustable parameters)
   b. Easy to install (Conda, Docker, Singularity)
   c. Easy to switch between environments (local, cluster, cloud)

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

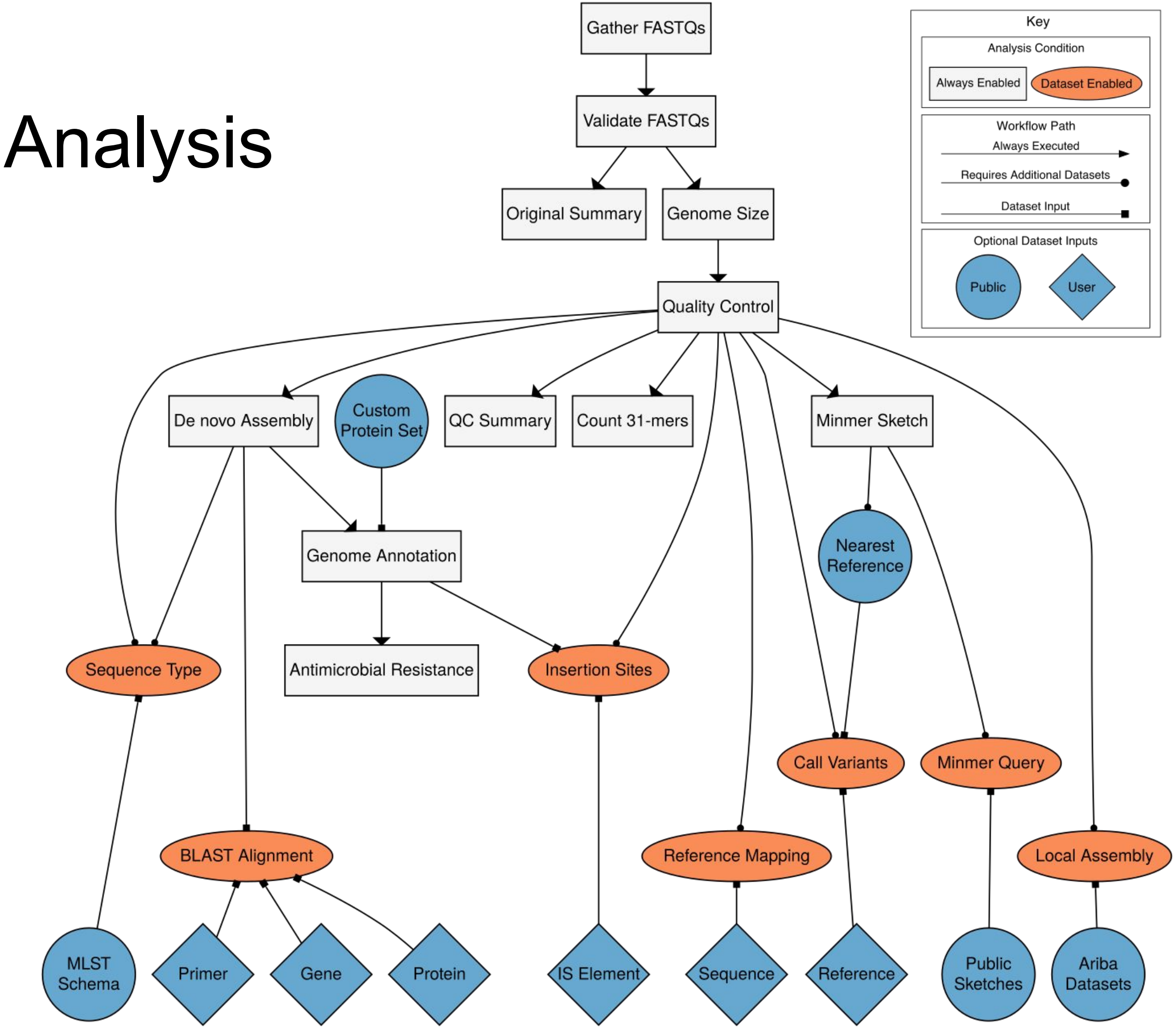
Three Sides of Bactopia

- Bactopia Datasets
  - Framework for including available public and user created datasets
- Bactopia Analysis Pipeline
  - Main *per-isolate* workflow
- Bactopia Tools
  - Independent workflows for *comparative* analyses

Bactopia Overview



Bactopia Analysis Pipeline



Bactopia Use Case: *Lactobacillus* genus

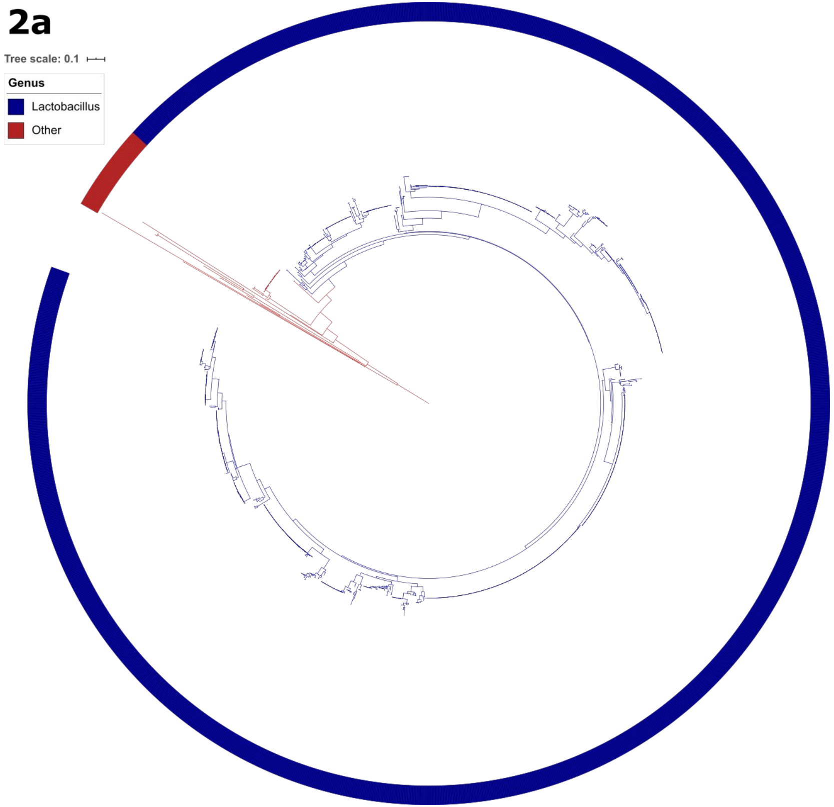
- Run all publicly available *“Lactobacillus”* genomes through Bactopia
  - Build *Lactobacillus* datasets
  - Query ENA for available Lactobacillus genomes
  - Run SRA/ENA genomes through Bactopia Analysis Pipeline
    - Bactopia will download genomes automatically
  - Apply Bactopia Tools to describe 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	-	-	-	-	-	-

Not everything is *Lactobacillus*

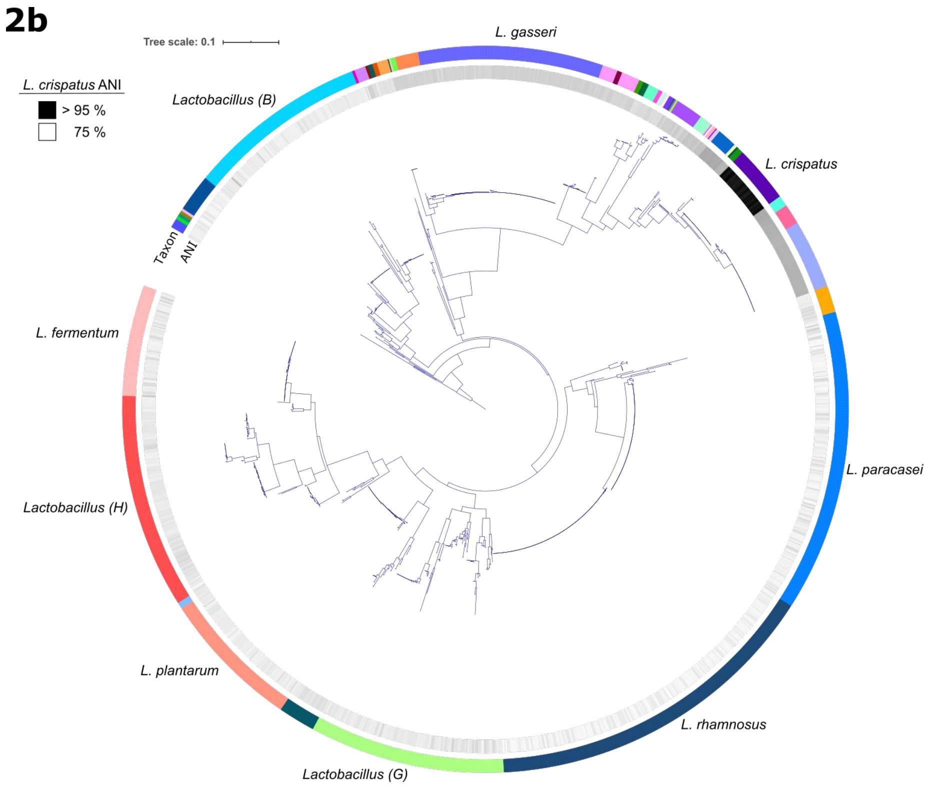
- 16S rRNA gene phylogeny
  - Bactopia Tool - phyloflash
- Taxon classified by GTDB
  - Bactopia Tool - gtdb
- 58 samples not Lacto
  - 34 samples are *S. pneumoniae*
- ~33% of the GTDB classifications in conflict with the NCBI taxon



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

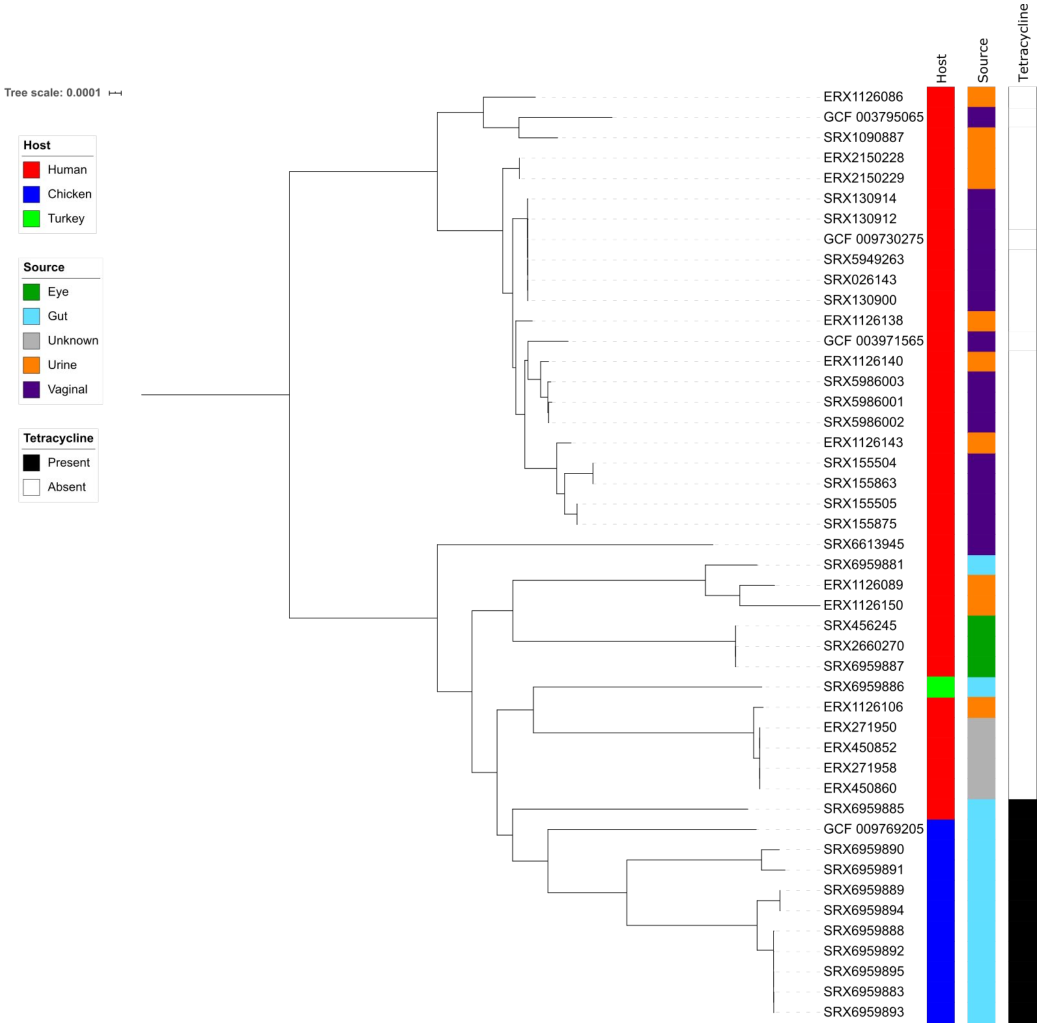


Lactobacillus crispatus

- 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

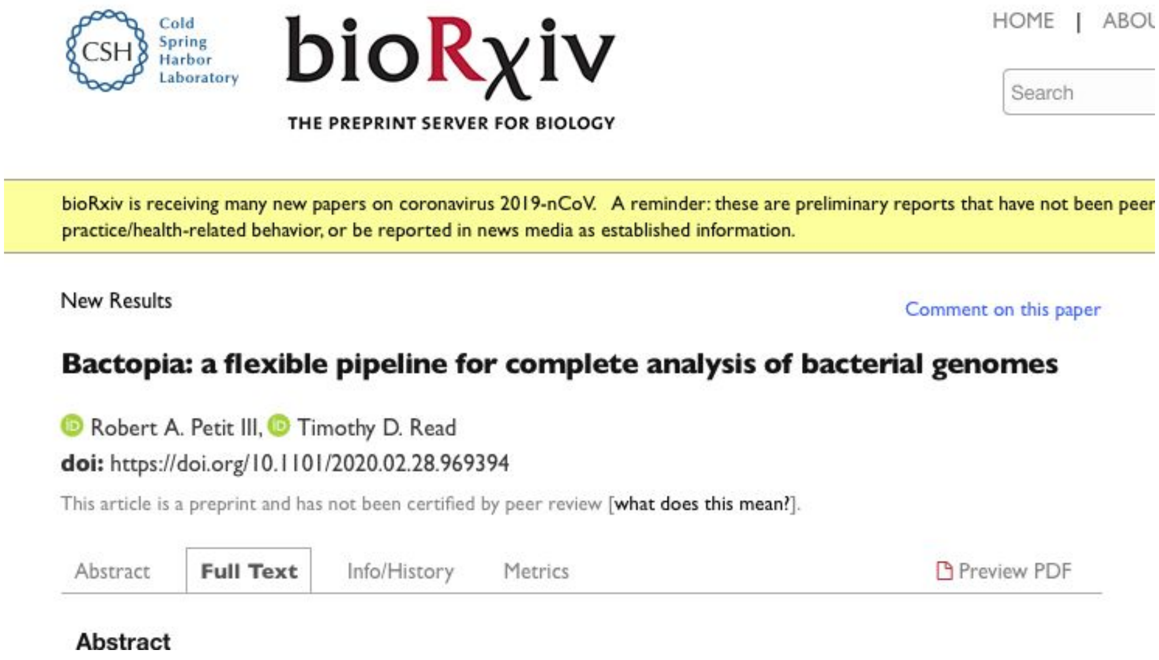


Use Summary

- We demonstrate how Bactopia can:
    - Build datasets
    - Query ENA for publicly available genomes
    - Process publicly available genomes
    - Conduct comparative analyses
        - 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



bioRxiv is receiving many new papers on coronavirus 2019-nCoV. A reminder: these are preliminary reports that have not been peer practice/health-related behavior, or be reported in news media as established information.

New Results

Bactopia: a flexible pipeline for complete analysis of bacterial genomes

Robert A. Petit III, Timothy D. Read

doi: https://doi.org/10.1101/2020.02.28.969394

This article is a preprint and has not been certified by peer review [what does this mean?].

Abstract

What’s next for Bactopia?

- Hopefully a useful resource for community
  - Will help to polish it and get a sense of what people want

- More Bactopia Tools
  - mashtree, hicap, bactdate, pyseer, poppunk, bigsi, etc…

- Implement long-read support (eventually)
  - Illumina is still 90+% of the data available and what we generate
  - Long-read is advancing to quickly at the moment

- Process 70,000 *Staphylococcus aureus* genomes
  - All on AWS and the basis for Staphopia v2

Acknowledgements

Robert Petit

Monica Farley, Michelle Su, Jon Moller, Tauqeer Alam, Sandeep Joseph, Steve Tsang, Michelle Wright, Mike Martin, Ashley Alexander, Ahmed Babiker

CGC
Erik Lehnert, Natasha Bezmarevic, Uros Sipetic, Manisha Ray

Funding: NIAID, Amazon in Education Award, National Cancer Institute, CDC Emerging Infection Program (EIP)

Staphopia Links

- Currently hosted at: https://staphopia.emory.edu/

- Open Source and Available on GitHub: https://github.com/staphopia

- Docker Container Available: https://hub.docker.com/r/rpetit3/staphopia/

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