Using Bactopia to process 67,000 *Staphylococcus aureus* genomes with AWS Batch

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

EMORY UNIVERSITY SCHOOL OF MEDICINE
Department of Medicine

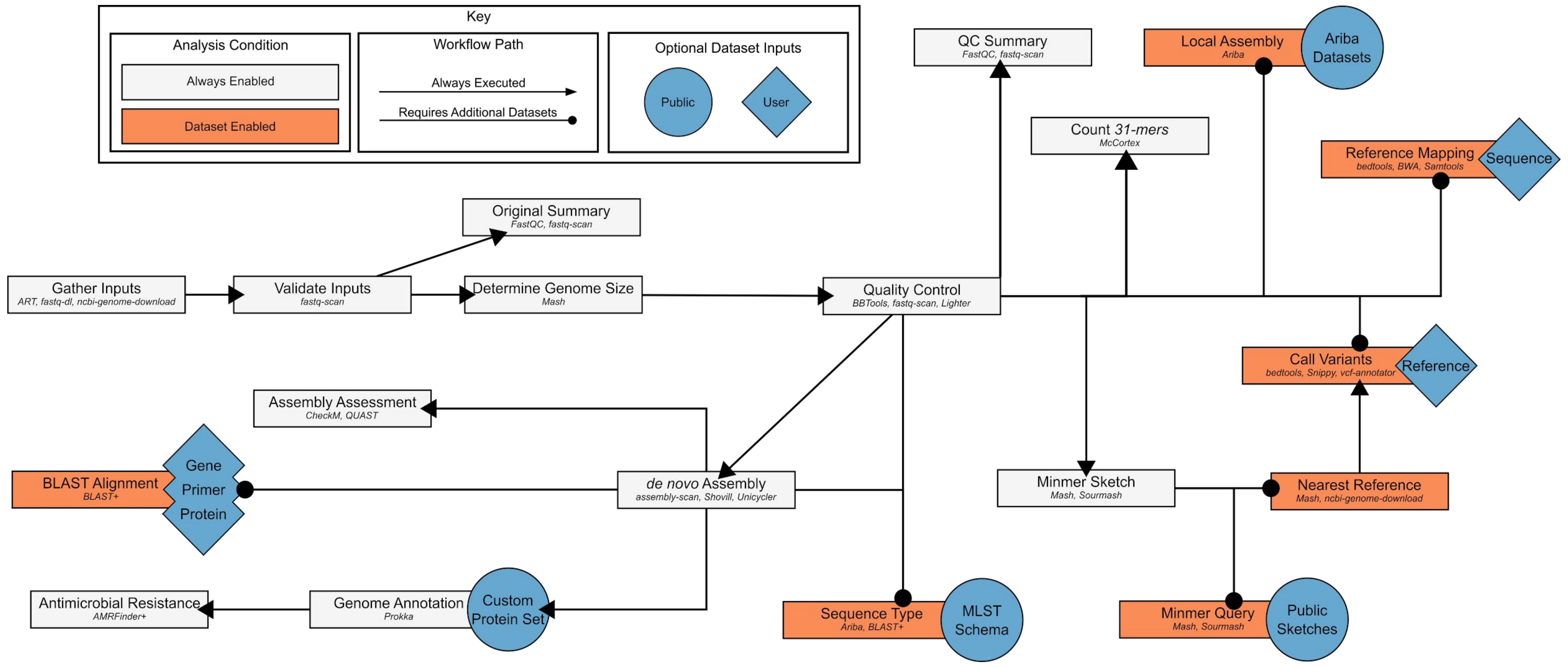
Background

Staphopia [1] is an analysis pipeline and Application Programming Interface (API) focused on *Staphylococcus aureus*, a common asymptomatic colonizer of humans and a major antibiotic-resistant hospital pathogen. As of October 2020, more than 30,000 *S. aureus* genomes were publicly available that were not included in Staphopia. Instead of processing the new genomes with the older Staphopia analysis pipeline, we instead reprocessed all *S. aureus* genomes using Bactopia [2] which replicates Staphopia but with more up to date methods. Using Bactopia we were able to reprocess all publicly available *S. aureus* genomes in just 5 days with AWS Batch.

Bactopia

Bactopia [2] is an extensive workflow (below) integrating more than 70 bacterial genomic tools for complete analysis of bacterial genomes. Bactopia uses Nextflow, allowing for support of many types of environments. One such environment is Amazon Web Services (AWS) Batch, which we used for this analysis.

Bactopia Datasets also allows you to automatically incorporate numerous public datasets, or your own species-specific datasets, into your analysis. To take advantage of this feature, we developed a curated *S. aureus* dataset, which includes a set genomes and sequences to query against all *S. aureus* genomes.



References

[1]: Petit III, R. A. & Read, T. D. Staphylococcus aureus viewed from the perspective of 40,000+ genomes. *PeerJ* 6:e5261, (2018)
[2]: Petit III, R. A. & Read, T. D. Bactopia: a Flexible Pipeline for Complete Analysis of Bacterial Genomes. *mSystems* 5, (2020)
[3]: Petit III, R.A. Using AWS Batch to process 67,000 genomes with Bactopia (2020)

Funding

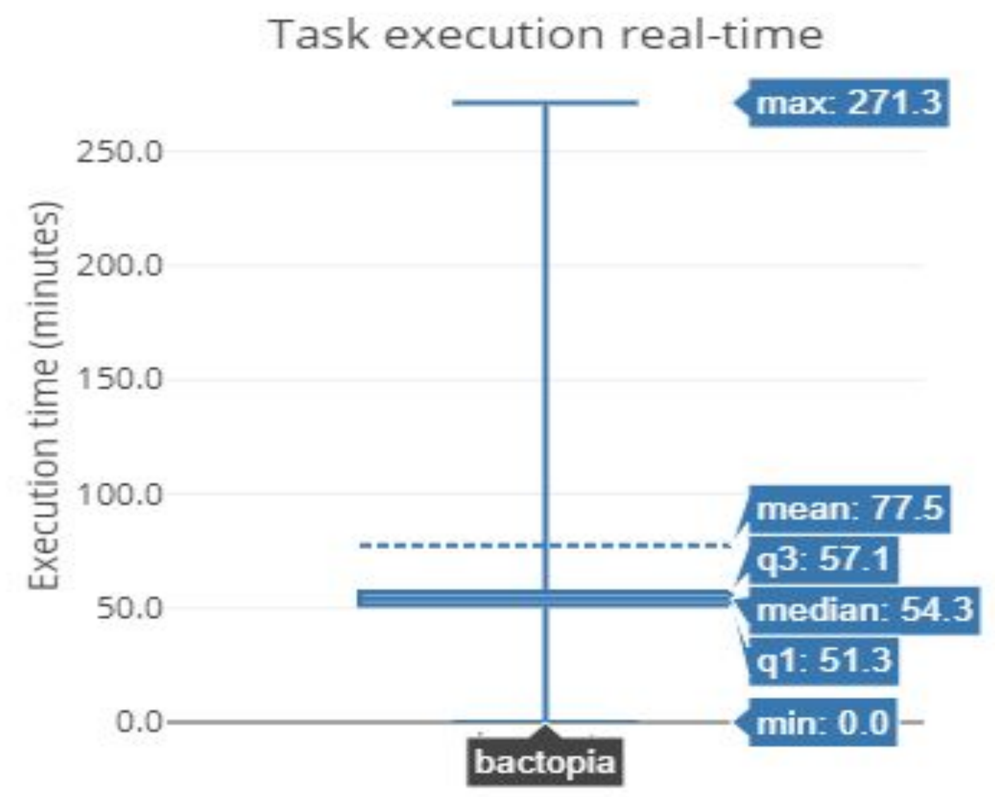
Support for this project came from the AWS Cloud credits for Research Program and an Emory Public Health Bioinformatics Fellowship funded by the CDC Emerging Infections Program (U50CK000485) PPHF/ACA: Enhancing Epidemiology and Laboratory Capacity.

AWS Batch and 67,000 Genomes

Using the bactopia search command, the term "*Staphylococcus aureus*" was queried again the European Nucleotide Archive (ENA) to generate a 70,000 Experiment accessions labeled as *S. aureus*. After filtering for Illumina only, 67,000 accessions remained. The Illumina-only Experiment accessions were then used as inputs to Bactopia, and the associated FASTQs were downloaded from the Sequence Read Archive (SRA) and processed with AWS Batch. A blog post [3] is available that provide a more in depth write up of our experience.

Summary of Processing

After 5 days, all genomes were downloaded and processed. We monitored the processing using Nextflow Tower, which also aggregated a lot of useful stats (to the right). Each genome took on 55 minutes (median) to processing using 4-cores and 16GB of memory for each genome. 2,500 (3%) of the submitted jobs had to be resubmitted, in each of these cases the memory had to be increased to 32GB.

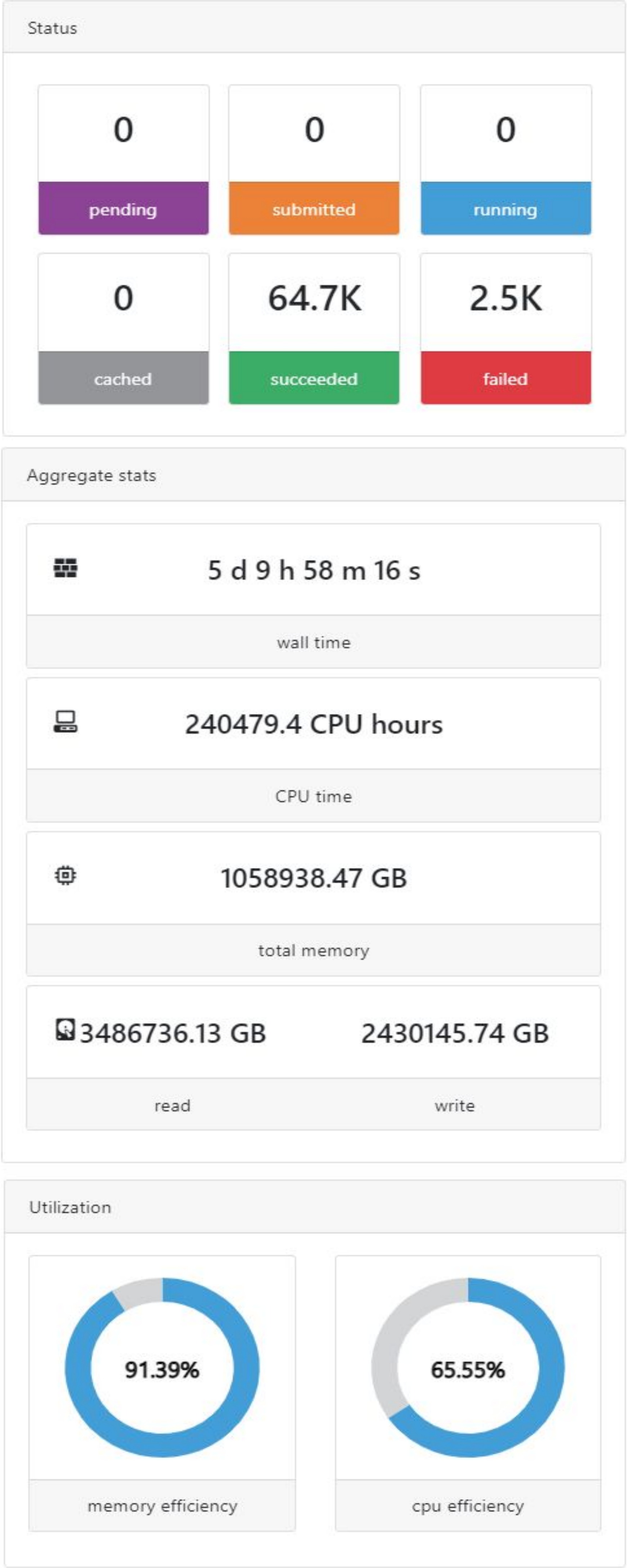


Associated Costs

It cost about $0.19, or €0.16, to process a *S. aureus* genome using Bactopia on AWS Batch. The bulk of the cost (table below) was compute costs, and data egress being nearly half the compute costs. By improving CPU efficiency and reducing the size of the outputs, it should be possible to reduce the costs to <$0.10 (€0.08) per genome.

AWS Charge	Usage	Cost USD (EUR*)	Cost Per Genome USD (EUR*)
Elastic Compute Cloud	259k CPU Hours	$7.552 (€6,495)	$0.113 (€0.097)
Data Transfer	42 TB	$3,566 (€3,067)	$0.053 (€0.046)
Simple Storage Service	84 TB	$1,286 (€1,106)	$0.019 (€0.017)
		Total:	$0.185 (€0.16)

* USD to EUR conversion based on October 31st ($1:€0.86)



Status

0	0	0
pending	submitted	running
0	64.7K	2.5K
cached	succeeded	failed

Aggregate stats

5 d 9 h 58 m 16 s
wall time

240479.4 CPU hours
CPU time

1058938.47 GB
total memory

3486736.13 GB read
2430145.74 GB write

Utilization

91.39%
memory efficiency

65.55%
cpu efficiency

Learn more about Bactopia at:
bactopia.github.io or ePoster

and Staphopia at:
staphopia.emory.edu or Talk

@rpetit3 @tdread_emory