Benchmark dataset for assembly of *Salmonella* Oxford Nanopore data

Krishna A. Thakor[1,2], Justin Y. Kim[1], Kaitlin A. Tagg[1], Hattie E. Webb[1], Jessica C. Chen[1], Lee S. Katz[1]

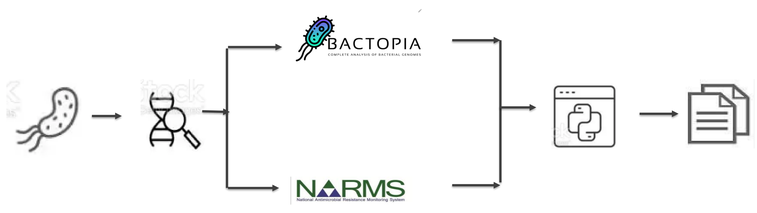
1. Enteric Diseases Laboratory Branch, US Centers for Disease Control and Prevention, Atlanta, GA, USA;
2. Oak Ridge Institute for Science and Education, US Department of Energy, Oak Ridge, TN, USA

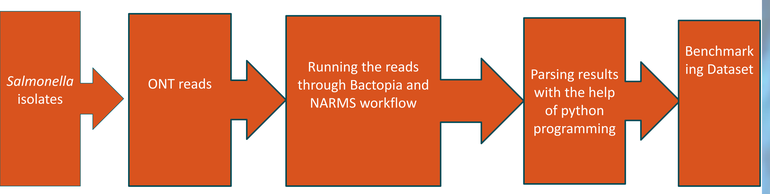
Introduction

- Whole genome sequencing is regularly used in public health for the detection, investigation, and surveillance of foodborne pathogens, including *salmonella*. However, regions of the accessory genome, such as plasmid and phages can be difficult to assemble with long read sequencing, leading to inaccuracies in genetic comparisons, particularly for antimicrobial resistance (AMR) determinants and mobile genetic elements(MGEs)
- Long read sequencing with the Oxford Nanopore Technologies (ONT) platform offers a cost-effective solution, but choosing the right bioinformatics pipeline can be challenging.

Methods

- To address this challenge, we developed a benchmarking dataset for *Salmonella* assembly using ONT reads. The dataset includes three sets of published isolates, comparing 18 *Salmonella* Hader and 5 reading genomes selected to compare the diversity of accessory genomes of these serotypes. It also includes a set of 12 mcr-1 containing *Salmonella* isolates, some of which contain large complex *Salmonella* Infantis resistance plasmid.
- We analyzed assemblies with Bactopia's workflows for AMR and plasmids (based on AMRFinder and PlasmidFinder respectively) as well as the NARMS lab workflow (based on ResFinder and PlasmidFinder) .





Conclusion

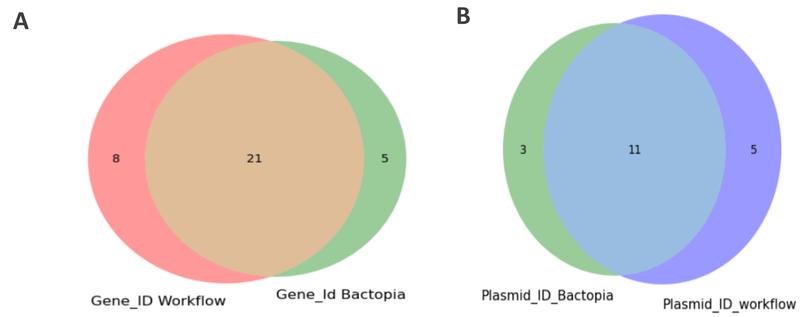
- The benchmarking analysis of the Bactopia and Legacy workflow tools on *Salmonella* suggests that both tools yield comparable results for AMR gene identification and plasmid identification.
- Differences in results are generally due to nomenclature differences between the different databases in the workflows.
- The similarity between the outcomes indicates that either tool can be considered effective for these purposes
- We plan to expand the dataset to include additional serotypes and AMR plasmids of public health concern and inspire future benchmark datasets with another serotype/subtype and sequencing platform.

Comparison of Bactopia and NARMS workflows using a benchmarking dataset demonstrate similar performance and effectiveness for accurate AMR and plasmid in *Salmonella*.

GitHub: Python Scripts

Benchmarking Dataset

Results



Figue2: Venn diagram of number of genes common between legacy workflow and Bactopia.

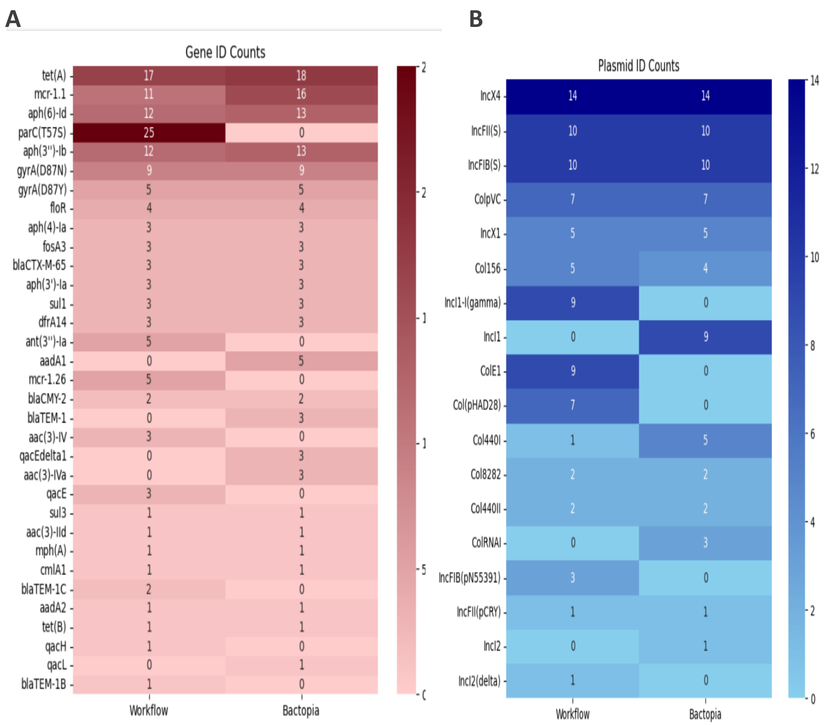


Figure 3: heatmap of repetition of each Gene Id and plasmid ID in the NARMS workflow vs Bactopia

Contact: Krishna Thakor
Email: tgu0@cdc.gov