25

Whole Genome Sequencing in the Clinical Microbiology Laboratory: A Pipeline to Answer Diverse Clinical and Epidemiological Questions

Samantha Taffner[1], Adel Malek[1], Heba Mostafa[1], Jun Wang[1], and Nicole Pecora[1,2]

[1] Department of Pathology and Laboratory Medicine, University of Rochester Medical Center, Rochester NY, [2] Department of Microbiology and Immunology, University of Rochester Medical Center, Rochester NY

ABSTRACT

Background: Next generation sequencing (NGS) is an emerging technique in clinical microbiology with applications ranging from outbreak analysis to genomic surveillance to analysis of unusual pathogens. Often NGS analysis is a fragmented step-wise process or pipelines are specialized for a single application or species. Herein we describe the URMC clinical microbiology pipeline (pipeline), a robust, quality-controlled, modular process for diverse applications and pathogens.

Methods: The pipeline was designed to flexibly perform rapid analysis on a variety of datasets and questions while storing previously analyzed isolates allowing the user to build a local database of isolates discovered in their area. The pipeline consists of two steps written in Python, SQLiite3, and JavaScript. The first step performs quality control on the raw reads (trimmomatic, FastQC) followed by genome assembly (SPAdes) and plasmid assembly (PlasmidSPAdes). Quality of genome assembly is assessed (Quast), genus and species (strainseeker), and MLST of samples are identified. Common phenotyping blast databases are included in the pipeline but custom blast databases can be added making the pipeline relevant to any species or project. To rapidly identify the best species reference the genome coverage is calculated for every sample (Quast). Step two consists of a modified CFSAN SNP Pipeline for reference-based SNP Calling and Phylogenetic Analysis. Modifications include masking SNPs which occur inside phages, mobile elements, and transposons, only include sites where a consensus exists in every sample, to produce a maximum likelihood tree (FastTree), and an interactive web application is produced to visualize the coverage and SNP locations throughout the genome to ensure consistent coverage and no SNP clustering.

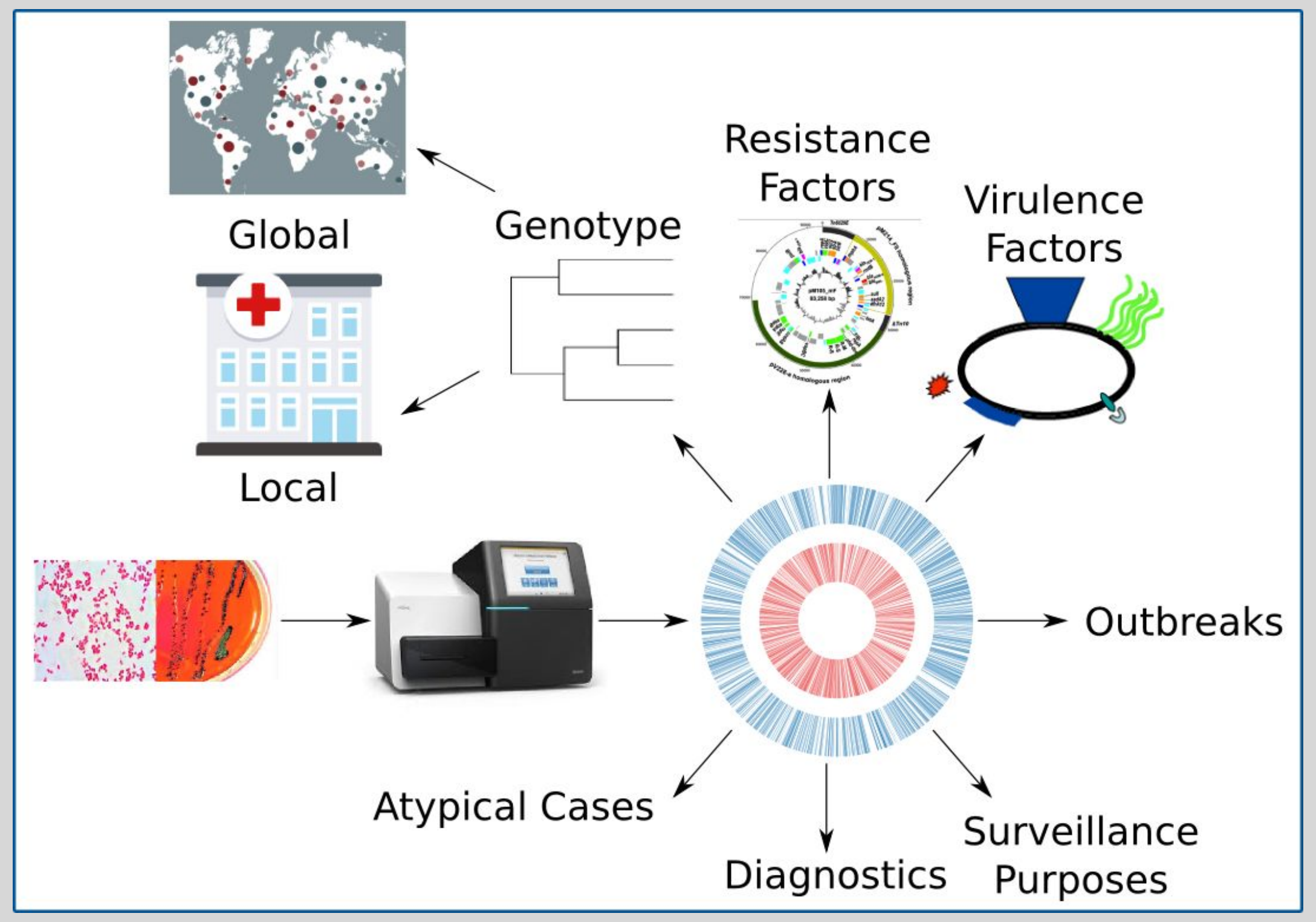
Results: The pipeline was successfully used for the following diverse projects; Genomic investigation of an Enterobacter aerogenes outbreak in a cardiac intensive care unit. Genomic surveillance of carbapenem-resistant pathogens, and Extended spectrum β-lactamases E.coli. We have also used the pipeline for characterizing unusually isolated organisms e.g. Facklamina hominis.

Conclusions: Whole genome sequencing is a powerful tool to complement traditional clinical microbiology techniques. Here, we described a pipeline that has been proven in diverse projects to be a versatile for clinical microbiology needs. Future plans include automatically generating an editable phylogenetic tree that overlays meta-data onto and adding a cloud-based user interface for initializing the pipeline, analyzing the data, and producing a standardized report to provide to clinical staff.

INTRODUCTION

- Growing need for NGS analysis in clinical microbiology laboratories for diverse projects and questions.

Fig 1. Diverse projects and questions answered by NGS

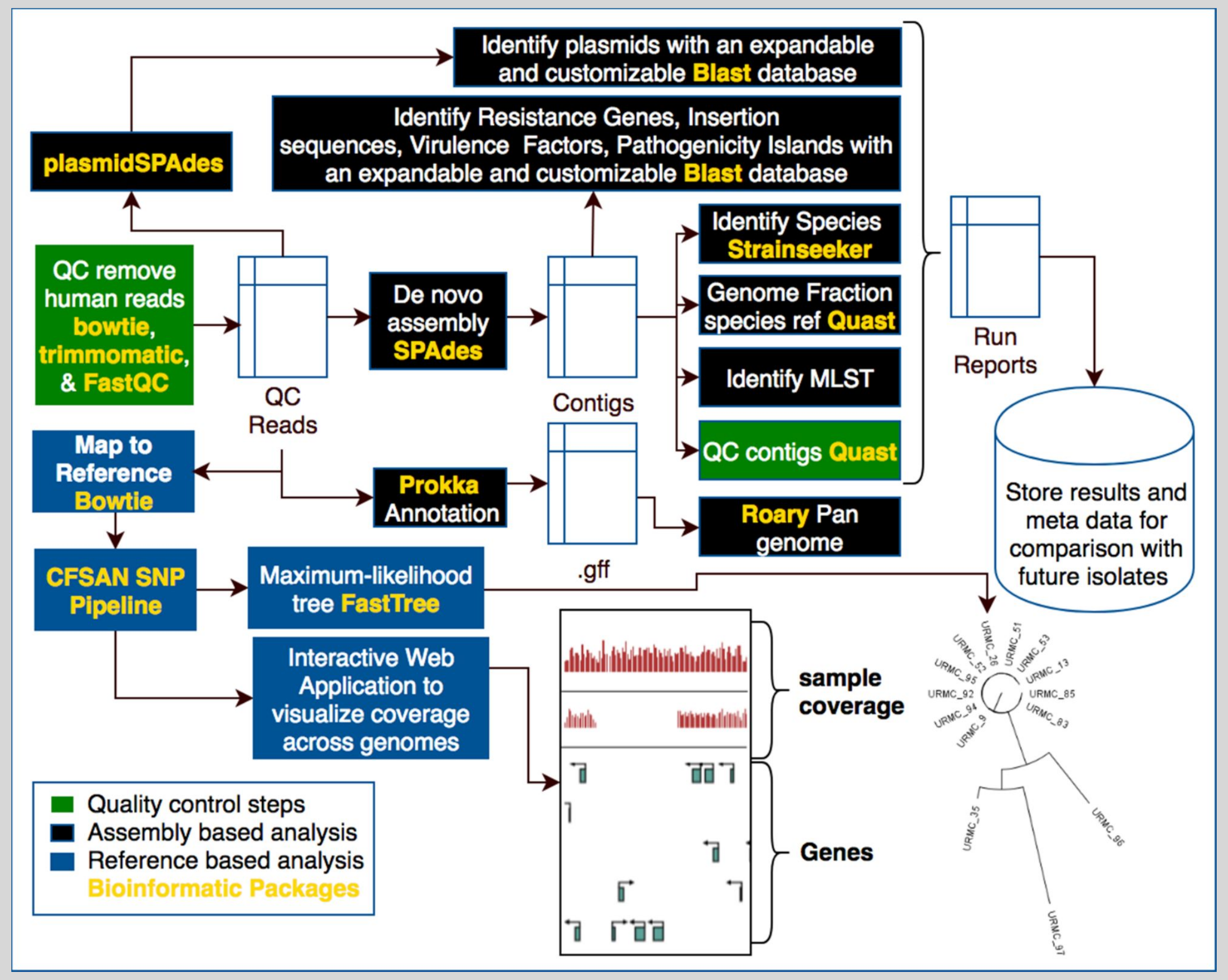


- Develop a clinical microbiology pipeline which pieces together fragmented processes into one robust, quality-controlled, modular process for diverse applications and pathogens.
- Create a local genomic landscape.

METHODS

- Data Type: Illumina MiSeq paired end reads.
- Languages/Database: Python, SQLite, JavaScript.
- Run Environment: Linux command line interface on a Slurm high performance cluster.
- To decrease processing time, tasks are either run in parallel or as a different Slurm jobs depending the needs of the task.
- Only key intermediate datasets and results are stored to decrease re-analysis time while keeping required hard drive space to a minimum
- Examples of user conditions includes choosing what modules not to run and changing QC cutoffs.

Fig 2. URMC clinical microbiology pipeline diagram.



Identify plasmids with an expandable and customizable Blast database

Identify Resistance Genes, Insertion sequences, Virulence Factors, Pathogenicity Islands with an expandable and customizable Blast database

plasmidSPAdes

QC remove human reads bowtie, trimmomatic, & FastQC

QC Reads

De novo assembly SPAdes

Contigs

Identify Species Strainseeker

Genome Fraction species ref Quast

Identify MLST

QC contigs Quast

Run Reports

Store results and meta data for comparison with future isolates

Map to Reference Bowtie

Prokka Annotation

.gff

Roary Pan genome

CFSAN SNP Pipeline

Maximum-likelihood tree FastTree

Interactive Web Application to visualize coverage across genomes

sample coverage

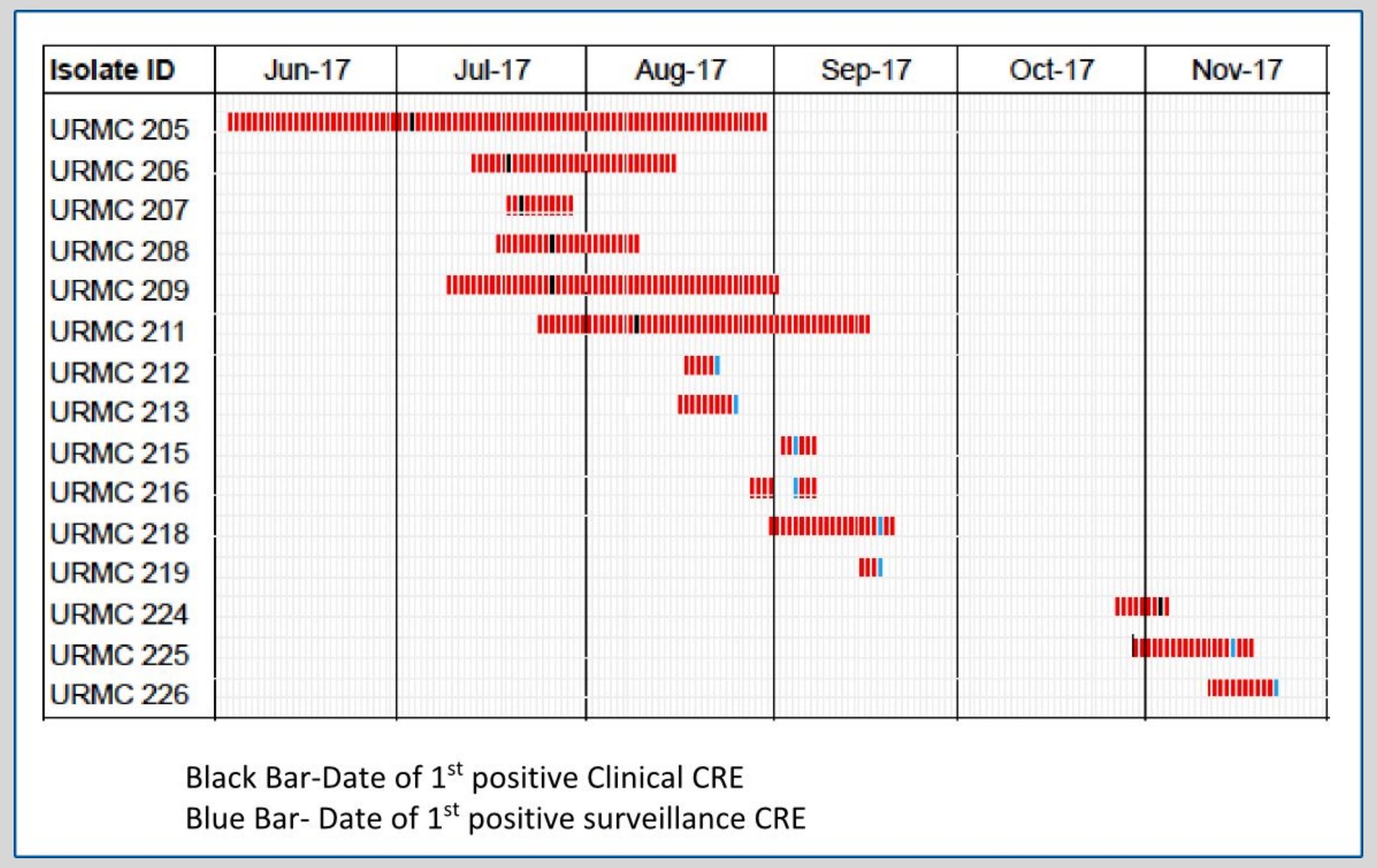
Genes

🟩 Quality control steps
⬛ Assembly based analysis
🟦 Reference based analysis
Bioinformatic Packages

EXAMPLE OF OUTBREAK ANALYSIS

- An carbapenem-resistant *Enterobacter aerogenes* outbreak occurred from Jun-Nov 2017 occurred in our cardiac Intensive care unit. (CICU)
- Whole genome sequencing of CR-EA isolates was undertaken to investigate patient-to-patient transmission, assess phylogeny relative to separate hospital isolates, and characterize molecular determinants of resistance and virulence.

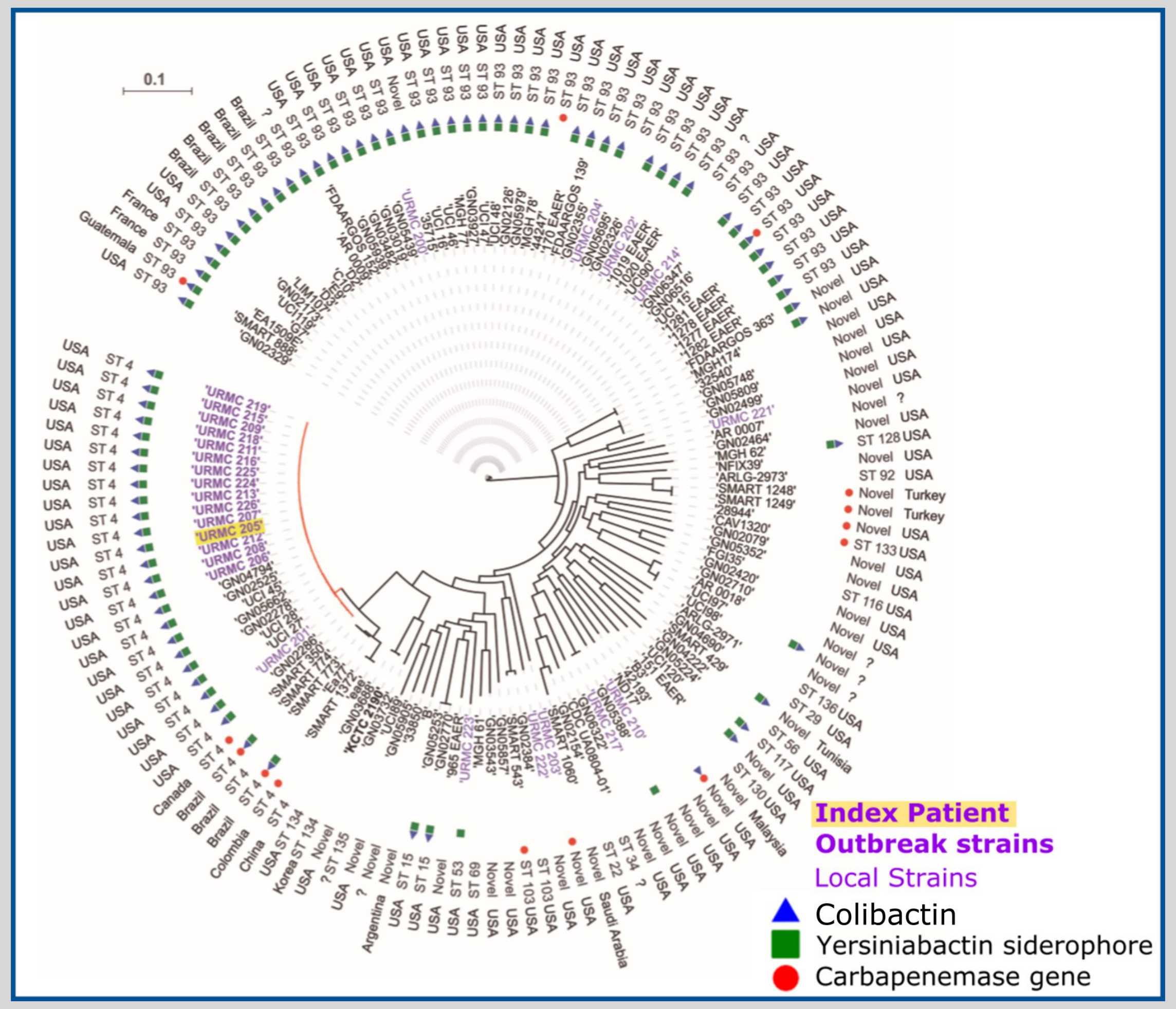
Fig 3. CICU E. Aerogenes Outbreak timeline.



Black Bar-Date of 1st positive Clinical CRE
Blue Bar- Date of 1st positive surveillance CRE

EXAMPLE OF OUTBREAK ANALYSIS CONT.

Fig 4. Population structure of *Enterobacter aerogenes*: local vs global strains.

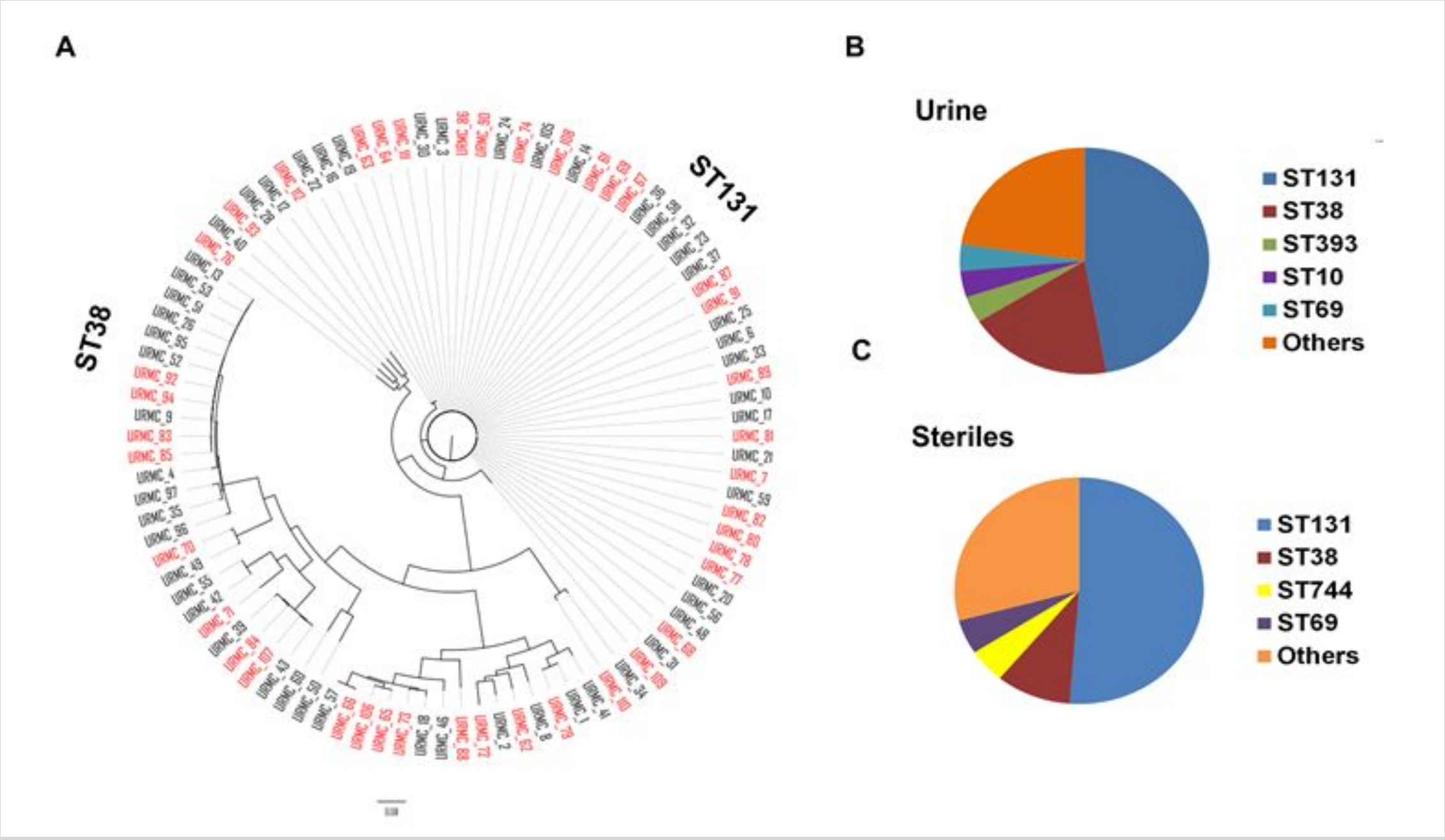


- Outbreak was a single clonal cluster
- Outbreak was distinct from strains from other wards and previous years

EXAMPLE OF SURVEILLANCE PROJECT

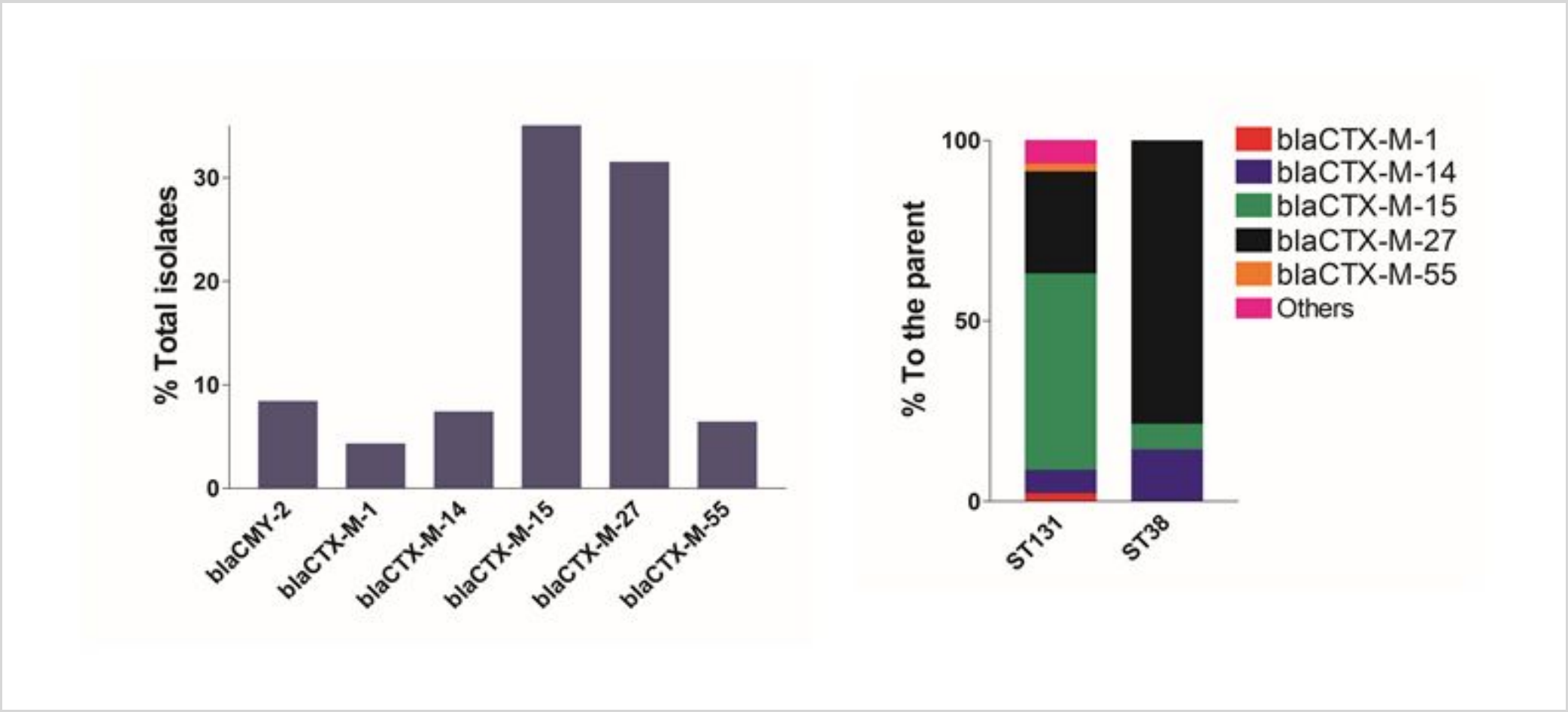
- Widespread increase in extended spectrum β-lactamases is a global threat.
- Historically, *E. coli* ST-131 with blaCTX-M-15 has been the majority of ESBL-producing E. coli both in the United States and worldwide.
- In our institution, the total number of ESBL-producing *E. coli* increased between 2013 to 2017 (30%), and were primarily uropathogenic isolates.

Fig 5. Population structure of isolates in study ST131 is the predominate subclone, followed by ST38



EXAMPLE OF SURVEILLANCE PROJECT CONT

Fig 6. CTX-M-27 is nearly as prevalent as CTX-M-15 and is the dominate ESBL within ST38



- Western New York genomic surveillance found that blaCTX-M-27 may be an emerging ESBL in both ST-131 and ST-38.

CONCLUSIONS AND FUTURE DIRECTIONS

- Whole genome sequencing and the URMC micropipeline have been a powerful tool for tracking transmission events, tracking effectiveness of control measures in real-time, surveillance purposes, and atypical cases.
- The URMC micropipeline can robustly analyze diverse projects and produce reproducible results.

Future Directions

- Automatically generating an editable phylogenetic tree that overlays meta-data.
- Adding a cloud-based user interface for initializing the pipeline, analyzing the data, and producing a standardized report to provide to clinical staff.

REFERENCES

1. Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi and Glenn Tesler, QUAST: quality assessment tool for genome assemblies, Bioinformatics (2013) 29 (8): 1072-1075.
2. Andrew J. Page, Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T. G. Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, Julian Parkhill, "Roary: Rapid large-scale prokaryote pan genome analysis", Bioinformatics, 2015;31(22):3691-3693.
3. Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
4. Antipov D., Hartwick N., Shen M., Raiko M., Lapidus A., Pevzner P. A. plasmidSPAdes: Assembling Plasmids from Whole Genome Sequencing Data. Bioinformatics, 2016.
5. Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, Sergey I. Nikolenko, Son Pham, Andrey D. Prjibelski, Alexey V. Pyshkin, Alexander V. Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A. Alekseyev, and Pavel A. Pevzner. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology 19(5) (2012), 455-477.
6. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170.
7. Davis S, Pettengill JB, Luo Y, Payne J, Shpuntoff A, Rand H, Strain E. (2015) CFSAN SNP Pipeline: an automated method for constructing SNP matrices from next-generation sequence data. PeerJ Computer Science 1:e20
8. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012, 9:357-359.
9. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One. 2010 Mar 10;5(3)
10. Roosaare M, Vaher M, Kaplinski L, Möls M, Andreson R, Lepamets M, Kõressaar T, Naaber P, Kõljalg S, Remm M. StrainSeeker: fast identification of bacterial strains from raw sequencing reads using user-provided guide trees. PeerJ. 2017 May 18;5:e3353. doi: 10.7717/peerj.3353. eCollection 2017.

CONTACT INFORMATION

Samantha Taffner
Bioinformatic Analyst
samantha_taffner@URMC.rochester.edu