Rapid Genomic Characterization of High-Risk Pathogens Using Long-Read Sequencing to Identify Nosocomial Outbreaks

Wu CT[1], Shropshire WC[1,2], Spallone A[2], Bhatti MM[2], Kalia A[1], Treangen T[3], Liu X[1], Shelburne SA[1,2,4*]

[1]School of Health Professions, The University of Texas MD Anderson Cancer Center, Houston, TX USA;
[2]Department of Infectious Diseases and Infection Control, The University of Texas MD Anderson Cancer Center, Houston, TX USA
[3]Department of Computer Science, Rice University, Houston, TX USA,
[4]Department of Genomic Medicine, MD Anderson Cancer Center, Houston, TX

Introduction

- Typically, Illumina based sequencing has been necessary to achieve sufficient confidence to identify two or more bacterial strains causing infections as being genetically related, suggestive of healthcare transmission.
- The Oxford Nanopore Technologies (ONT) new chemistries (V14) and improving basecalling algorithms hold promise to improve long-read based methods to measure genetic relatedness with similar sensitivity/specificity to short-read methods.
- We sought to develop an ONT sequencing pipeline capable of rapid outbreak detection in healthcare settings.

Method

- Approximately 10 pathogens/week meeting our definition of 'potential outbreak pathogen of interest' were systematically collected using EPIC electronic health record workbench tools.
- Bacterial genomic DNA was directly extracted from clinical microbiology lab plates.
- Sequencing was carried out using the ONT R10.4 flow cell, with target of 40x coverage depths.
- Nanopore fast5 data was converted to fastq raw data using the Guppy-v6.4.6 basecaller.
- Genome assembly was accomplished through the Flyest package, which includes Flye, Medaka, dnaapler, berokka, and Pilon for data assembly, refinement, and MLST analysis.
- For genetic-related analyses, Mintyper was employed for Reference-based SNP calling, and Ridom SeqSphere+ was used for core genome MLST(cgMLST) analysis.

Workflow

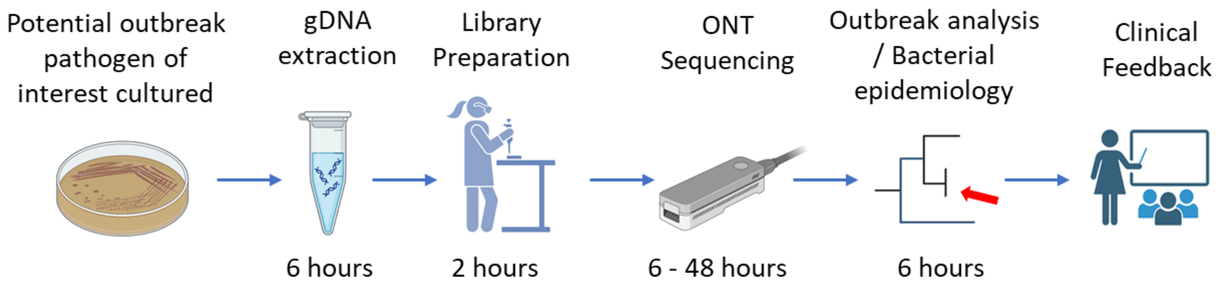

Figure 1. Weekly workflow from sample collection to data analysis. On average, the entire process took approximately 2.8 days (minimum = 2 days, maximum = 4 days). We optimized resource usage by reusing flow cells for up to two or three runs. Using a single flow cell, we typically obtained 9.8 gigabyte (GB) of data with a minimum of 5.5GB and a maximum of 14.09GB.

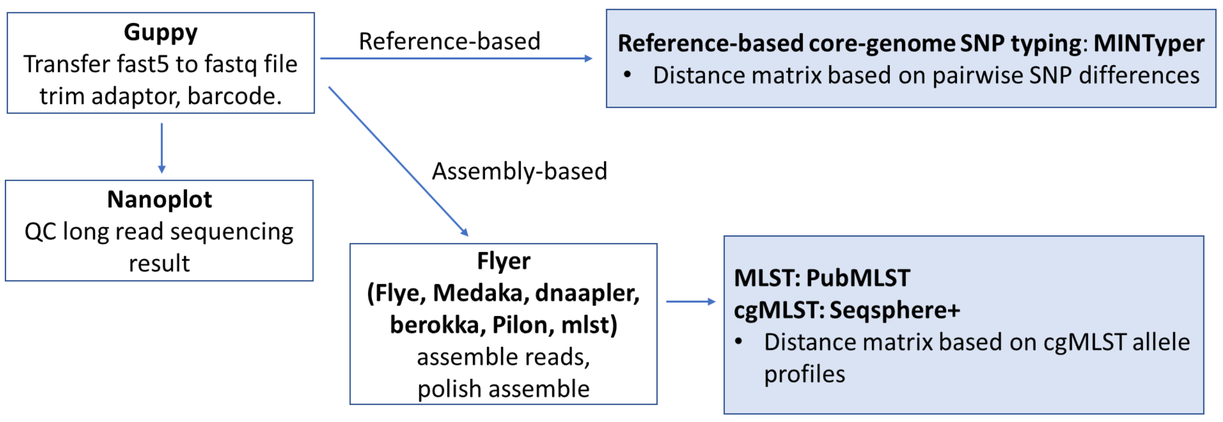

Figure 2. Outbreak ONT sequencing analysis pipeline to detect bacterial genetically related isolates .

Result

We rely entirely on ONT sequencing to achieve discriminative power comparable to short-read sequencing.

	Escherichia coli MB1860		*Klebsiella pneumoniae* MB2930		*Pseudomonas aeruginosa* MB4276	
sequencing platform	ONT	Illumina	ONT	Illumina	ONT	Illumina
core genome SNP distance	SNP=5		SNP=1		SNP=5	

Table 1. Comparison of core genome SNPs using MINTyper between ONT and Illumina sequencing of the same strain.

	Escherichia coli MB1860		*Klebsiella pneumoniae* MB2930		*Acinetobacter baumannii* MB15212		*Staphylococcus aureus* MB12042		*Enterococcus faecium* MB14008	
DNA extractions	first	second	first	second	first	second	first	second	first	second
core genome SNP distance	SNP=0		SNP=0		SNP=0		SNP=0		SNP=0	

Table 2. Core genome SNP comparison using MINTyper for the same strain sequenced using ONT alone. Data compares two distinct genetic DNA extractions.

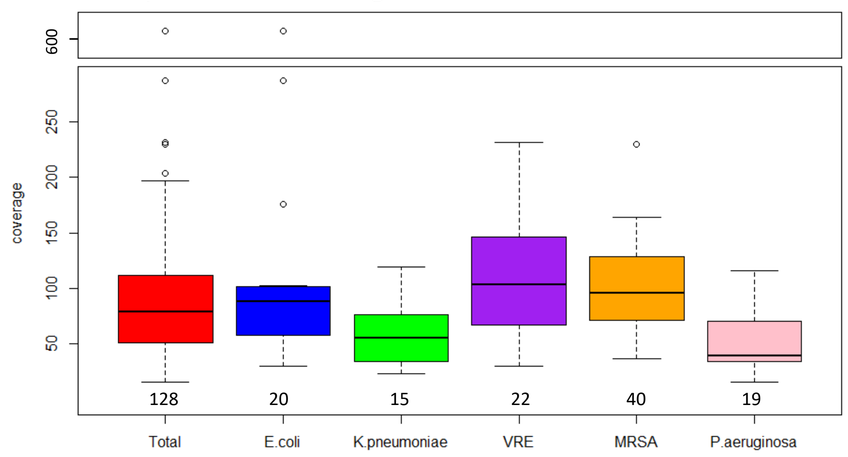

Figure 3. The boxplot illustrates the average sequencing coverage in our project. The overall medium coverage is approximately 100. Notably, *Klebsiella pneumoniae* and *Pseudomonas aeruginosa* show lower average coverage. Numbers indicate Ns per group.

Result

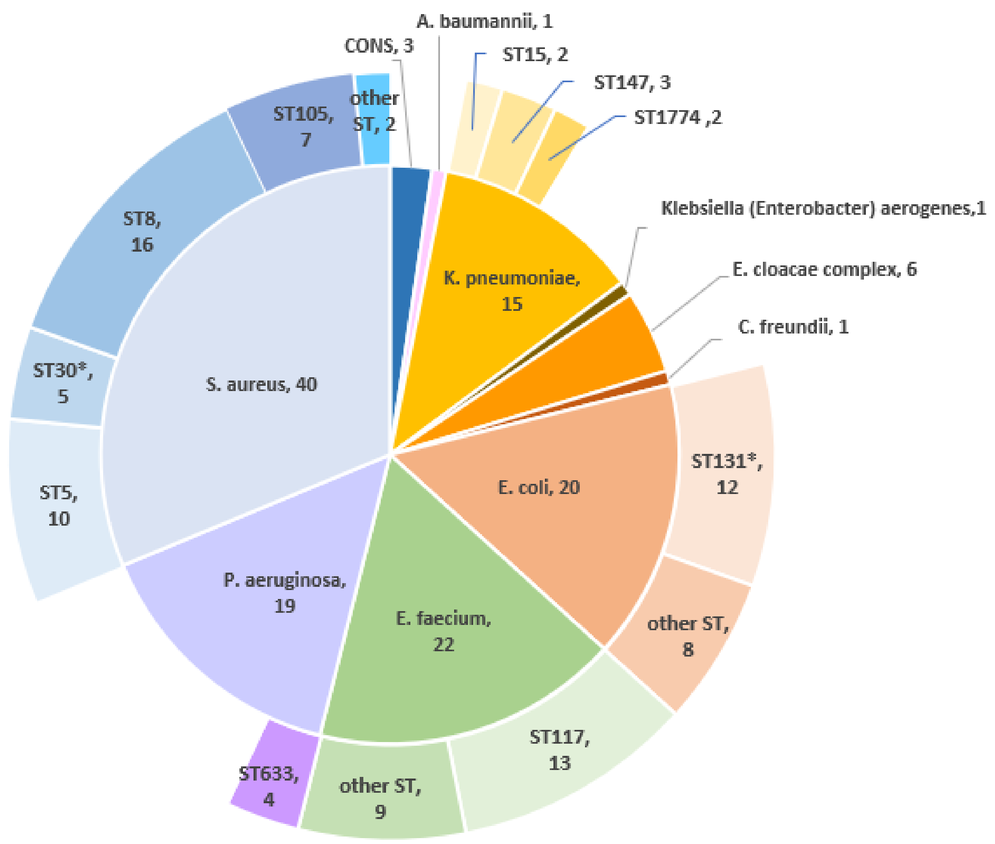

Figure 4. In a summary of samples sequenced to date, out of the 128 strains sequenced, we have identified five genetically related ST117 Vancomycin-resistant *Enterococcus faecium* strains and four genetically related ST633 *Pseudomonas aeruginosa* strains.

Figure 5. Maximum likelihood phylogenetic tree, generated using MINTyper, of 19 *P. aeruginosa* samples based on sequence analysis of core genome SNPs. ST633 isolates (n=4) that are genetically related (pairwise SNP difference =40.5) are indicated in a red box. Furthermore, each of the four samples with ST633 infections have some evidence of possible epidemiological links (i.e., admission times overlap, however, no evidence of shared spaces).

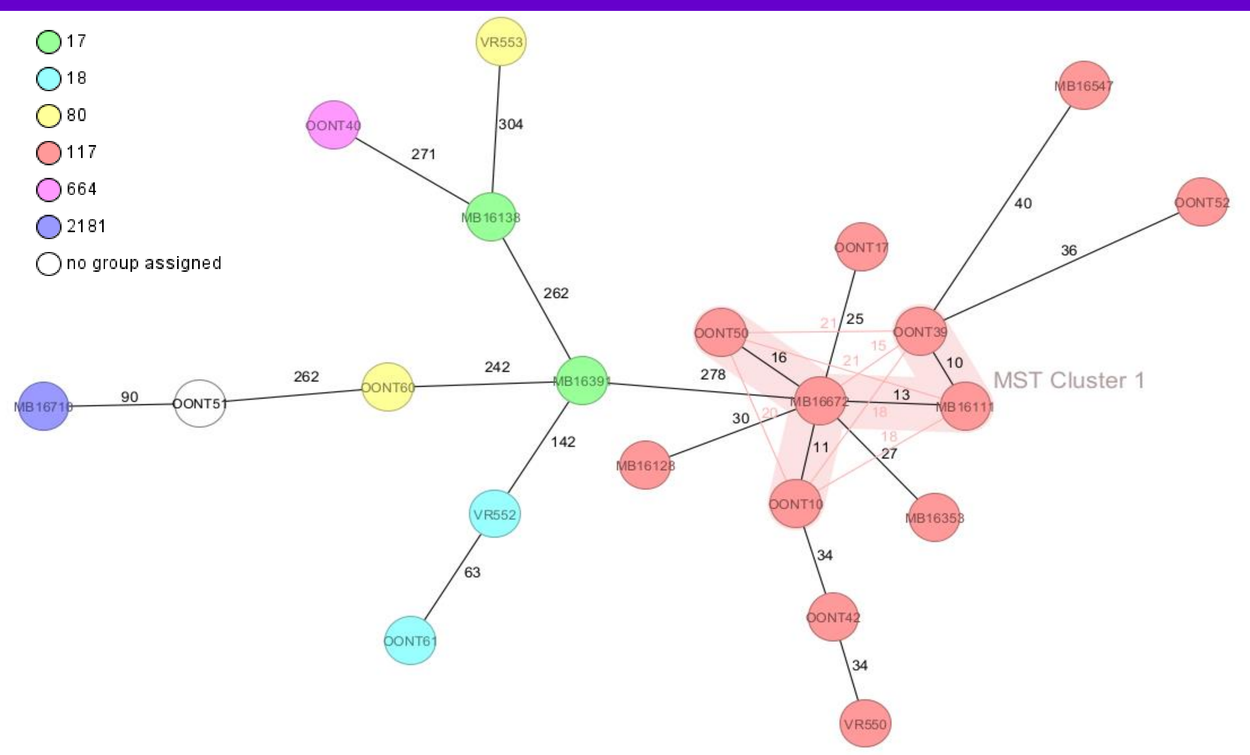

Figure 6. ONT Vancomycin resistant *Enterococcus faecium* (VREfm) data analyzed using minimum spanning tree inferred by a distance matrix of shared core VREfm gene alleles (n=1425) with missing alleles excluded from analysis. There are 21 VREfm samples with each circle representing a single sample from an individual sample, with each sample labelled by multilocus sequence type (MLST) designation as presented in the legend. The numbers between the nodes indicate the count of allelic differences. Five genetically related ST117 strains with a difference of 20 alleles or less, are highlighted in light red. Each of these five samples have epidemiological links with admission overlap and four samples sharing the same ICU period.

Conclusion

- Our ONT pipeline generated WGS data for various organisms within a few days
- Using the ONT data alone, we can detect genetically-related isolates from different samples, with a strong correlation to epidemiological data.
- Our pipeline has the potential to assist in the rapid detection and employment of preventative measures against healthcare-associated infection transmission.

Acknowledgements

- We thank the Clinical Microbiology Lab at MD Anderson Cancer Center for sample acquisition
- We appreciate the support of Oxford Nanopore Technologies in providing R10.4 flow cells
- Jane and Phil Yeckel Endowment, Peter and Cynthia Hu Scholarship, and Cardinal Health Scholarship for CTW scholarships.