RUTGERS
School of Graduate Studies

Office of Student Affairs and Admissions
SGS at New Jersey Medical School
Rutgers, The State University of New Jersey
185 South Orange Avenue, MSB C-696
Newark, NJ 07103

http://njms.rutgers.edu/gsbs/
p. 973-972-4511
f. 973-972-7148

YOU ARE INVITED TO ATTEND THE DEFENSE OF THE DOCTORAL DISSERTATION

"New Insights into the Pangenome of *Mycobacterium tuberculosis* Using a Novel De Novo Assembly Pipeline"

by
Poonam Chitale

Infection, Immunity and Inflammation Program

B.S. 2013, The College of New Jersey, NJ
M.S. 2018, New York University, NY

Thesis Advisor: David Alland MD
Professor and Chief, Division of Infectious Diseases

Department of Medicine
Rutgers University

Thursday, May 11th, 2023
10:00 A.M.
ICPH auditorium, zoom

Join Zoom presentation:
https://rutgers.zoom.us/j/97288285209?pwd=QjViR0txcnY0ZHZpRVpsd0lHVkRkQT09
Meeting ID: 972 8828 5209
Password: poonam

ABSTRACT

Mycobacterium tuberculosis (Mtb) remains one of the leading infectious causes of death worldwide. Whole Genome Sequencing (WGS) has emerged as a useful tool for strain categorization and outbreak tracking. Despite the widespread use of whole genome sequencing (WGS) tools in bacterial research, there is currently no gold standard pipeline to *de novo* assemble new genomes. To fill this gap, we developed Bact-Builder, a streamlined pipeline for de novo assembly of high quality, accurate bacterial genomes using a long-read consensus assembly, followed by both long and short read polishing. This novel pipeline takes advantage of the strengths of multiple long read assemblers to reduce assembly-based errors and generate a consensus genome that is further polished with both long and short reads to correct for insertions and deletions (indels) and single nucleotide polymorphisms (SNPs). Bact-Builder was first applied to the Mtb H37Rv reference strain using an in silico simulated data set generated using the published H37Rv genome and was then used to assemble H37Rv sequenced from laboratory stocks. We demonstrate that our pipeline produced a complete, closed out genome that was closer in accuracy to the published reference than other tools and revealed key differences in laboratory H37Rv strains compared to the published sequence. We discovered that the newly sequenced H37Rv genome contained ~6.4 kb additional base pairs, encoding ten new regions that include insertions in PE/PPE genes and new paralogs of *PPE38*, *esxN* and *esxJ*, which are differentially expressed compared to the known paralogs. Our findings present a major update to the H37Rv reference genome. We further applied Bact-Builder to other reference strains of Mtb and several clinical isolates spanning all seven established lineages of the Mtb complex (MTBC) to generate a novel Mtb Pangenome. Our Pangenome allowed us to identify 16 hypervariable regions (HVRs) as well as several previously undiscovered gene paralogs and duplications. Ultimately, our tool and resulting pangenome will be useful for a variety of downstream applications such as antimicrobial resistance (AMR) detection, evolutionary phylogenetic studies, construction of new lineages specific reference sequences, and variant analysis among others.