Bact-Builder: A new streamlined tool for generating high quality consensus based, complete *Mycobacterium tuberculosis* genomes

Poonam Chitale[1], Alexander D. Lemenze[1], Emily C. Fogarty[2], Pradeep Kumar[1], A. Murat Eren[2], David Alland[1*]

[1]Rutgers University – New Jersey Medical School, Newark, New Jersey, USA; [2]University of Chicago, Chicago, IL, USA

Background: Whole-genome sequencing (WGS) of *Mycobacterium tuberculosis* (Mtb) has evolved from a basic research tool into a useful clinical tool for the diagnosis and surveillance of tuberculosis. This was due in large part to developments in Next Generation Sequencing (NGS) and third-generation sequencing technology. Most NGS platforms generate short reads that are either mapped to a reference or assembled *de novo*. Although NGS tools are highly accurate, short read *de novo* assemblies are highly fragmented and inadequate to generate fully closed microbial genomes. This is largely due to highly repetitive regions such as PE/PPE genes, and mapping artifacts which can't be resolved by short reads. Third-generation sequencing tools generate much longer reads that facilitate *de novo* assembly of complete genomes and resolves highly repetitive regions. Despite the widespread use of novel WGS tools in Mtb research, there is a lack of a gold standard pipeline to assemble and study new genomes. Although, existing tools can assemble genomes, they are often not tractable and can produce variable results both across and within assemblers when tested repeatedly.

Methods: Here, we present Bact-Builder, a streamlined pipeline for *de novo* assembly of high quality, accurate Mtb genomes using a long-read consensus assembly, followed by both long and short read polishing. We take advantage of the strengths of multiple long read assemblers and reduce assembly-based errors by generating a consensus genome that is further polished to correct for indels and SNPs. We validated our pipeline using a simulated data set generated using the published H37Rv genome and with laboratory sequenced H37Rv.

Results: We demonstrate that our pipeline produces a complete, closed out genome that is closer in accuracy to the published reference than other tools and highlights key differences found in our strain compared to the published sequence.

Conclusion: The genomes generated by Bact-Builder can be used for a variety of downstream applications such as antimicrobial resistance (AMR) detection, phylogenetic studies, construction of new reference sequences, variant analysis, and pan-genomic studies.