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

Poonam Chitale[1], Alex Lemenze[2], Emily Fogarty[3], Courtney Grady[1], Pradeep Kumar[1], A. Murat Eren[3], David Alland[1]

[1] Division of Infectious Disease, Department of Medicine and the Ray V. Lourenco Center for the Study of Emerging and Re-emerging Pathogens – New Jersey Medical School, Rutgers—The State University of New Jersey, Newark, NJ, USA

[2] Department of Pathology, Immunology and Laboratory Medicine, New Jersey Medical School, Rutgers—The State University of New Jersey, Newark, NJ, USA

[3] Department of Medicine, University of Chicago, Chicago, IL, USA

Background

- *Mycobacterium tuberculosis* (Mtb) is the causative agent of tuberculosis (TB)
- Mtb was responsible for over 1 million deaths in 2020 (WHO, 2020)
- Whole Genome Sequencing (WGS) for Mtb is increasingly being to track and treat Mtb infections (Fig 1)
- These efforts rely on complete and accurate genomes
- There is no gold-standard approach for assembling Mtb genomes
- Current tools are not consistent, reliable or robust and typically use a single assembler approach

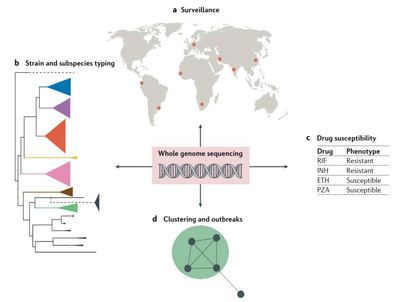


Figure 1. Whole genome sequencing of Mycobacterium tuberculosis. A. International surveillance of prevalence and drug resistance. **B.** Determination of the species or subspecies of M. tuberculosis complex isolates. **C.** Determination of drug resistance patterns on the basis of the presence of specific SNPs. **D.** Identification of transmission clusters and outbreaks. ETH, ethambutol; INH, isoniazid; PZA , pyrazinamide; RIF, rifampicin.Adapted from Meehan et.al., 2019.

- **Here we present Bact-Builder – a new streamlined tool that enables end to end *de novo* assembly of bacterial genomes**

Methods

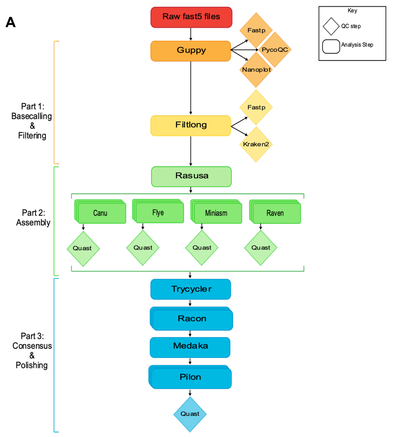
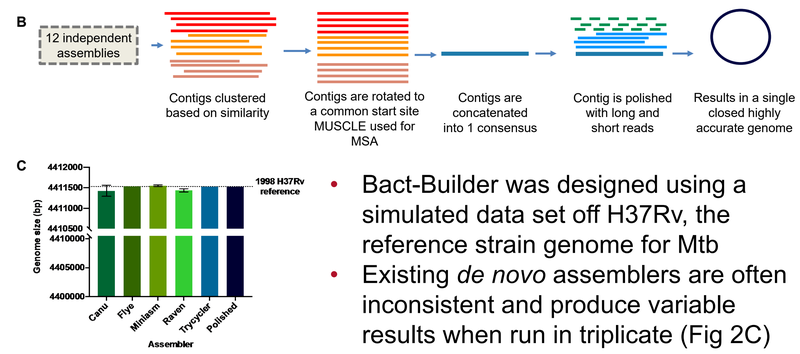


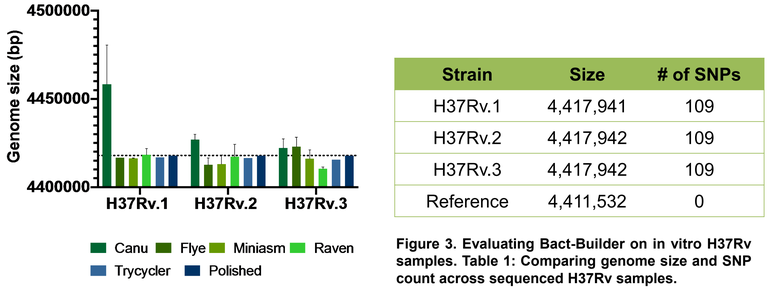
Figure 2. Overview of Bact-Builder A. Bact-Builder workflow. B. Generating a consensus with trycycler followed by polishing with Racon (3x) + Medaka + Pilon (3x). C. Comparison of 4 commonly used long read assemblers run in triplicate, Trycycler and Bact-Builder output.



- Bact-Builder was designed using a simulated data set off H37Rv, the reference strain genome for Mtb
- Existing *de novo* assemblers are often inconsistent and produce variable results when run in triplicate (Fig 2C)
- Generating a consensus genome removes the variability found in individual assemblers
- Further polishing using long and short reads allows for SNP and indel correction (Fig 2B,C)
- Final genome was closer in size to the established reference compared to any individual assembler (Fig 2C)

Results

- *3 in vitro* stocks of H37Rv were extracted and sequenced on the Oxford Nanopore (ONT) MinION and with the illumina NovaSeq
- Samples were assembled with Bact-Builder
- The 3 replicates showed near perfect identity with respect to size, 109 SNPs relative to the published reference and 0 SNPs relative to each other (Fig 3, Table1)
- DNAdiff analysis showed that the Bact-Builder output was more similar to the published reference than any individual assembler alone (Fig 4)



Strain	Size	# of SNPs
H37Rv.1	4,417,941	109
H37Rv.2	4,417,942	109
H37Rv.3	4,417,942	109
Reference	4,411,532	0

Figure 3. Evaluating Bact-Builder on in vitro H37Rv samples. Table 1: Comparing genome size and SNP count across sequenced H37Rv samples.

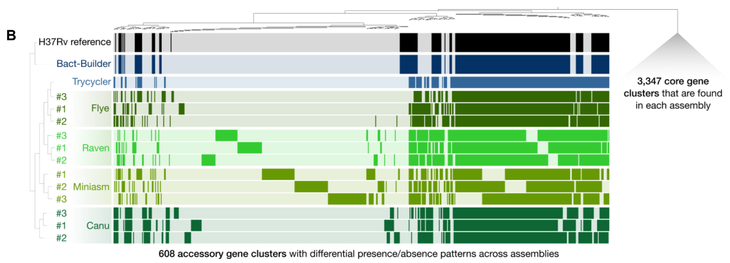


Figure 4. Comparing the differences between individual assemblers and Bact-Builder output. A. Heatmap of hierarchical clustering of the distance using euclidean average linkage clustering of differences between all assemblies for H37Rv.1 and the published reference determined by DNAdiff. B. Anvi'o output comparing gene clusters the reference and H37Rv.1 individual assemblies, trycycler output and the Bact-Builder output.

Conclusion

- Bact-Builder is an open-source program built into singularity containers for ease of use
- Bact-Builder enables end-to-end streamlined assembly of raw sequencing reads into a complete polished genome with as little as 50x genome coverage
- Final genomes are gap-closed and highly reproducible
- Nextflow logs detail run statistics on each individual step
  - The entire program runs in 543 hours on a standard CPU node
  - GPU acceleration of Guppy basecalling brings the total time down to 28.8hours
- Bact-Builder outputs can be used for a variety of downstream applications:
  - Establishing new reference sequences
  - Reference based assembly
  - RNAseq
  - TnSeq
  - Pangenomic studies

References

- WHO-Global Tuberculosis Report (2020);
- Meehan, C.J., Goig, G.A., Kohl, T.A., Verboven, L., Dippenaar, A., Ezewudo, M., Farhat, M.R., Guthrie, J.L., Laukens, K., Miotto, P., et al. (2019). Whole genome sequencing of Mycobacterium tuberculosis: current standards and open issues. Nature Reviews Microbiology *17*. 2020.

RUTGERS