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], Avi Shah[1], Aubrey Odom-Mabey[4], W. Evan Johnson[4], Jason H. Yang[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; [4] Bioinformatics Program, Boston University, Boston, MA, 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 used to track and treat Mtb infections
- 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
- **Here we present Bact-Builder – a new streamlined tool that enables end to end *de novo* assembly of bacterial genomes**
- **Using Bact-Builder we have identified 10 key regions of difference between the published H37Rv reference strain and laboratory strains and now propose an updated reference sequence for H37Rv**

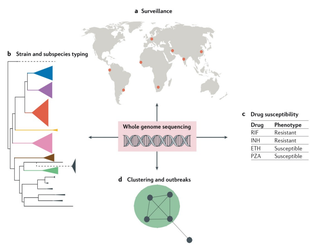
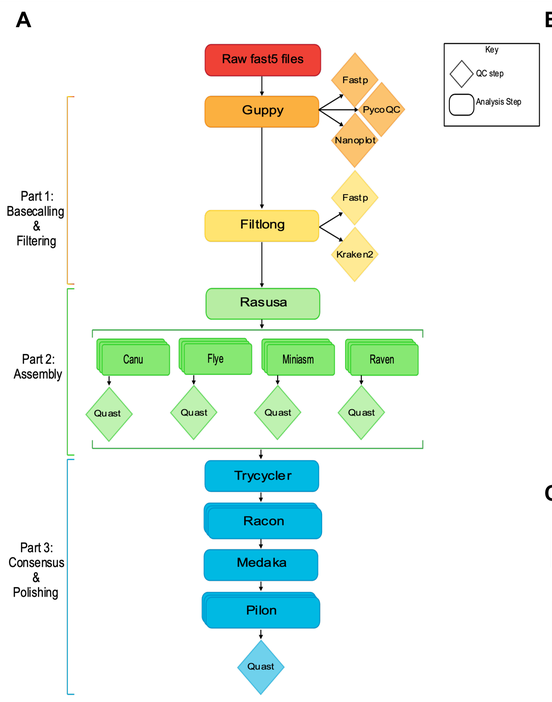


Figure 1. Whole genome sequencing of *Mycobacterium tuberculosis*. A. International surveillance of prevalence and drug resistance. **B.** Determination of the species or subspecies of Mtb complex isolates. **C.** Determination of drug resistance patterns on the basis of the presence of specific SNPs.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

METHODS and EVALUATING BACT-BUILDER



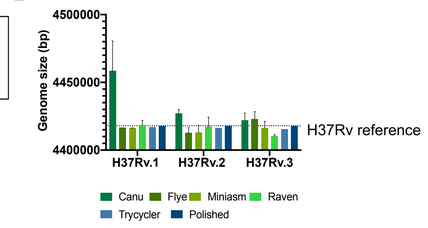
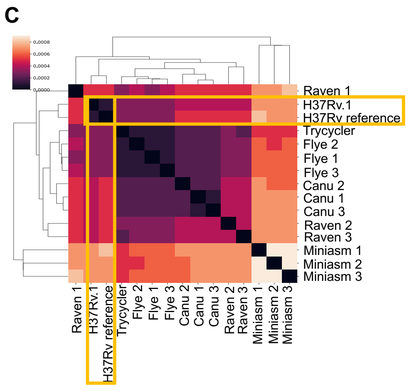


Table 1. Comparing size and SNP count relative to the published reference

Strain	Size (bp)	# 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



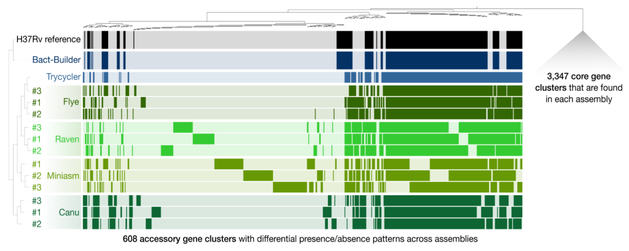
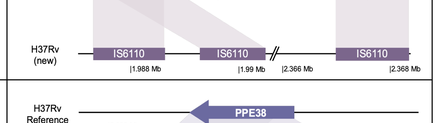
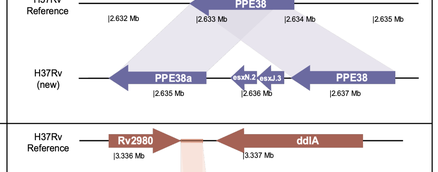
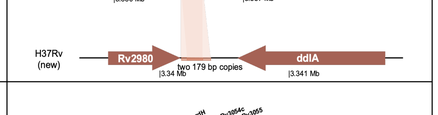
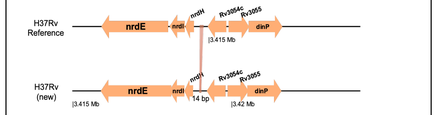
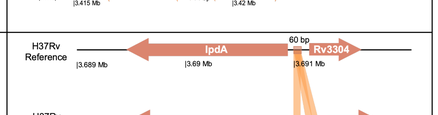
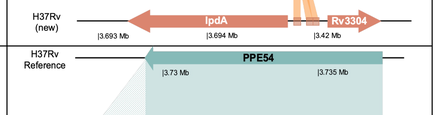
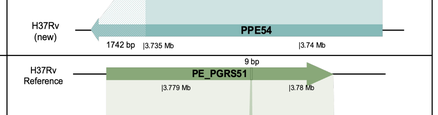
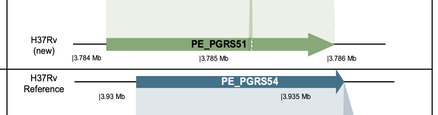
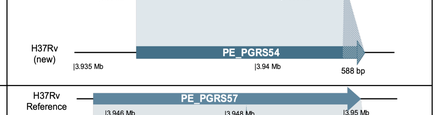
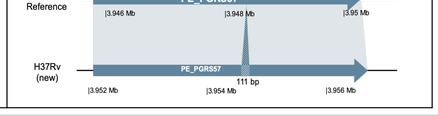


Figure 2. Bact-Builder overview and Results. A. Bact-Builder overview. **B-C.** Comparing assemblers and Bact-Builder output to the H37Rv reference and 3 biological replicates **D.** Anvi'o comparison of genes across H37Rv.1 assemblies and Bact-Builder output

- Bact-Builder generates a consensus genome which removes the variability found in individual assemblers and further polishing using long and short reads allows for SNP and indel correction (Figure 1B)
- 3 biological replicates of H37Rv demonstrated that Bact-Builder produced 3 virtually identical final genomes that were more similar to the published reference than any individual assembler (Figure 2 B-D, Table 1)

REGIONS of DIFFERENCE

Table 2. Summary and Schematic of identified Regions of Difference

R#	Description	Schematic
R1	207 bp in-frame insertion in *PE_PGRS27* (*rv1450c*)	
R2	1356 bp duplication of *rv3475*, *rv3474* (IS6110 transposase)	
R3	2064 bp duplication of *esxN.2* (*rv1793.2*), *esxJ3* (*Rv1038c.2*) and short *PPE38a* (*rv2352c*)	
R4	179 bp tandem duplication in H37Rv.1 in intergenic region	
R5	14 bp tandem duplication copy number difference in reference in intergenic region	
R6	60 bp tandem duplication in H37Rv.1 in intergenic region	
R7	1728 bp in frame insertion in *PPE54* (*rv3343c*)	
R8	9 bp tandem repeat in *PE_PGRS51* (*rv3367*)	
R9	579 bp in-frame insertion in *PE_PGRS54* (rv3508)	
R10	111 bp in-frame insertion in *PE_PGRS57* (*rv3514*)	

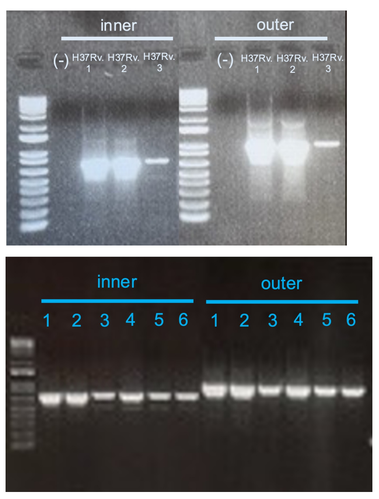
A

R3: Novel paralogs of *esxN* and *esxJ*

```
esxN     1 MTINYQFGDVDAHGAMIRAQAASLEAEHQAIVRDVLAAGDFWGGAGSVACQEFITQLGRN 60
esxN.2   1 MTINYQFGDVDAHGAMIRAQAAALEAEHQAIVRDVLAAGDFWGGAGSVACQEFITALGRN 60
esxN    61 FQVIYEQANAHGQKVQAAGNNMAQTDSAVGSSWA 94
esxN.2  61 FAVIYQQANAHGQKIQAAGSNMAQTDSAVGSSWA 94

esxJ     1 MASRFMTDPHAMRDMAGRFEVHAQTVEDEARRMWASAQNISGAGWSGMAEATSLDTMTQM 60
esxM     1 MASRFMTDPHAMRDMAGRFEVHAQTVEDEARRMWASAQNISGAGWSGMAEATSLDTMT+M 60
esxJ.3   1 VATRFMTDPHAMRDMAGRFEVHAQTVEDEARRMWASSQNISGAGWSGLAEATSLDTMTQM 60
esxJ    61 NQAFRNIVNMLHGVRDGLVRDANNYEQQEQASQQILSS 98
esxM    61 NQAFRNIVNMLHGVRDGLVRDANNYEQQEQASQQILSS 98
esxJ.3  61 NQAFRNIVNMLHGVRDGLVRDANNYEQQEQASQQILSS 98
```

B



1. 1998H37Rv c1
2. 1998H37vRv c2
3. 1998H37Rv culture

} Original 1998 H37Rv

4. NR-123
5. TMC102
6. TMC301

} Str^R variant of TMC102

Table 2. Description and schematic of the 10 regions of difference identified by DNAdiff between H37Rv (new) and the H37Rv reference.

Figure 2. Validating R3: Novel paralogs of *esxN* and *esxJ*. A. Clustal Omega comparison of amino acid sequence for *esxN.2* and *esxN* and *esxJ.2*, *esxM* and *esxJ*. **B.** PCR of R3 in all 3 laboratory replicates of H37Rv (H37Rv.1-3), the H37Rv1998 (1-3) and 3 commercially available strains of H37Rv from ATCC (4:NR-123, 5:TMC102, 6:TMC301). Inner primers targeted inside the region (738 bp) and outer primers targeted flanking regions (1007 bp)

FUNCTIONAL VALIDATION

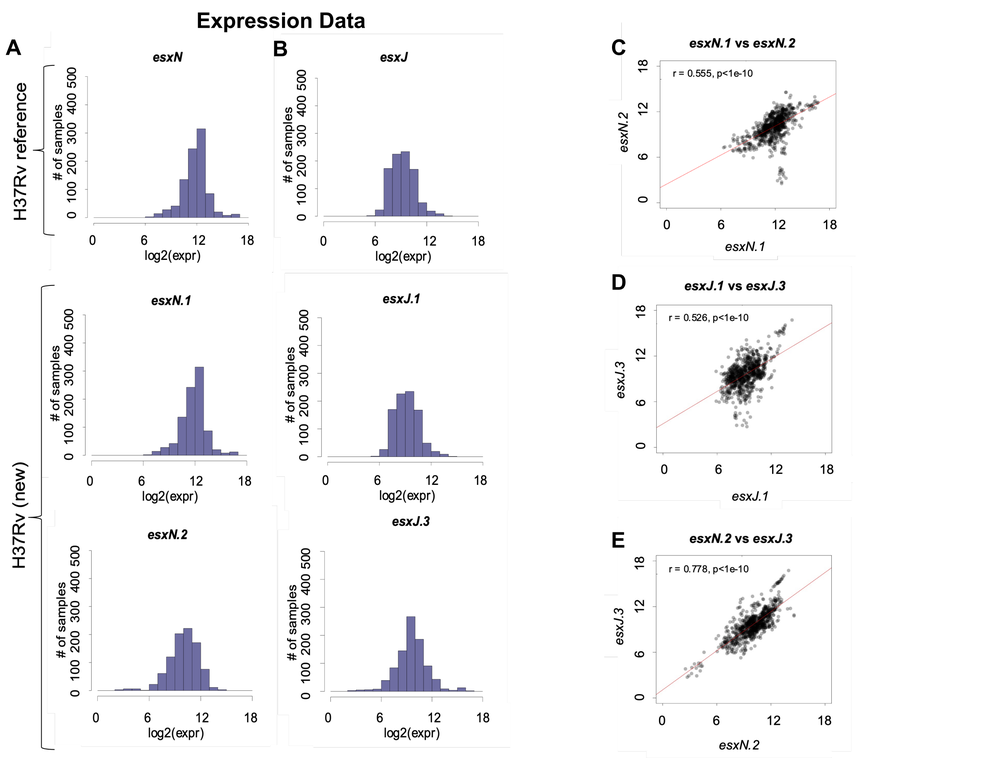


Figure 3. A-B. Histogram of *EsxN* (A) and *esxJ* (B) expression in public H37Rv datasets, demonstrating that newly identified esxN.2 and esxJ.3 are expressed in H37Rv and exhibit differential gene expression compared to their paralogs. **C-E.** Scatterplots showing correlation of expression in H37Rv (new) of esxN.1 and esxN.2 (C); esxJ.1 and esxJ.3 (D); esxN.2 and esxJ.3 (E).

- DNAdiff revealed 10 regions of difference between the H37Rv published reference and H37Rv (new) (Table2)
- These differences include in-frame insertions in PE/PPE genes, duplications of IS6110 transposon elements, and novel paralogs of esxN and esxJ (Table 2)
- PCR demonstrated that these regions are real and found universally across H37Rv (Figure 2B)
- RNAseq analysis demonstrated that the novel paralogs have different expression compared to the known copy of the gene and that expression between the paralogs is poorly correlated (Figure 3)

COMPARATIVE GENOMICS

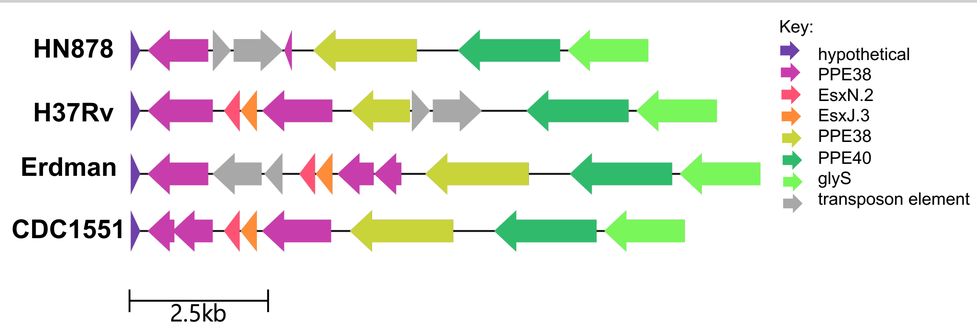


Figure 4. Comparison of R3 genes across Mtb strains

- Comparison with other lab adapted and clinical strains of Mtb showed that R3 is hyper-variable across strains (Figure 4)
- PPE38 has been demonstrated to inhibit Macrophage MHC Class I expression and dampens CD8+ T-cell response
- All 3 genes are known T-cell epitopes and may play a role in Mtb pathogenesis

DISCUSSION

- **Bact-Builder enables end-to-end streamlined assembly of raw sequencing reads into a complete polished genome**
- Final genomes are gap-closed with nearly 100% reproducibility (Table 1)
- Bact-Builder enabled us to accurately assemble highly repetitive PE/PPE genes, and resolve gene duplications which were previously very difficult to do with illumina sequencing
- Using Bact-Builder we identified key differences between lab strains of H37Rv and the published reference (Table 2)
- These differences included novel paralogs of *esxN* and *esxJ* (named *esxN.2* and *esxJ.3* respectively) which functionally behave differently than the known paralogs (Figure 3)
- **With this information, we propose a new, accurate and complete updated reference for H37Rv, the established reference for tuberculosis studies**
- Bact-Builder was also used to close out the genomes of other Mtb strains which showed previously undescribed variability across clinical isolates of Mtb (Figure 4)
- A comprehensive understanding of pan-genomic diversity across Mtb isolates will be key to understanding clinical phenotypes such as latency, drug resistance and increased pathogenicity

Funding: This work was funded by NIAID Tuberculosis Research Units Network (TBRU-N): U19AI111276

@poonaminny pc395@gsbs.rutgers.edu