Project 2 Proposal: *GitGud*

*Mason Tatro, Ben Yunker, Naomi Williamson, Ellie Spahr*

# **Background**

*Mycobacterium tuberculosis* is the bacterial causal agent of tuberculosis. The introduction and spread of single, multi-, or even extensive drug resistant strains has caused increased difficulty in TB treatment. Specific molecular tools, such as Xpert MTB/RIF, have been developed to detect resistance to individual antibiotics in clinical samples.

In addition to the known resistant variations, cryptic resistance has begun to appear in clinical settings. In a case study sampled by the Instituto de Biomedicina de Valencia in 2019, a case of Tuberculosis persisted for over 9 years despite being deemed “non-resistant” by molecular screening methods.1 The ability of this isolate to pass current screening methods and persist *in vivo* throughout treatment suggests that resistance detection methods should include additional resistance-associated genome variations to account for increasing diversity in resistance mechanisms and markers. We are using the raw, paired-end WGS Illumina data generated by this study to map cryptically resistant *M. tuberculosis* reads to the virulent H37Rv *M. tuberculosis* reference genome in an effort to categorize structural variations (SVs) and single nucleotide polymorphisms (SNPs) in this new strain.

H37 variants are well-studied *M. tuberculosis* strains originating from 1905 and are widely used as controls for tuberculosis identification and molecular studies.2 The Rv suffix denotes the virulent strain, while Ra denotes an avirulent counterpart. We have chosen to work with H37Rv since H37Ra would provide a less direct comparison against a pathogenic isolate and may confound results by highlighting additional virulence-nonvirulence variations.

Previous studies have used compiled *M. tuberculosis* SNPs from categorized isolates to successfully investigate the contribution of SNPs to resistance mechanisms.3 By extending this approach to a new clinical data set, we hope to classify genomic variations so that resistance mechanisms may be identified and treatment approach improved in similar emergent TB cases. Classifying SNPs is also the first step to generating molecular markers for this strain’s future identification- this is especially important in tracking origin and spread in epidemiological studies. By using read-to-reference mapping through BWA, we will identify genomic variations in a cryptically resistant *M. tuberculosis* strain against the H37Rv reference in an attempt to improve upon molecular techniques for identifying drug resistant isolates.

# **Methods**

## Pipeline

Data for both the *Mycobacterium tuberculosis* reference genome and the Drug-resistant isolate will be retrieved from NCBI. The *M. tuberculosis* H37Rv genome assembly (Assembly [GCA\_000277735.2 ASM27773v2](https://www.ncbi.nlm.nih.gov/assembly/84511), Bioproject: [PRJNA37301](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA37301)) is located in NCBI’s Genome database and consists of a single, closed scaffold of 4.41 Mb. The cryptically resistant *M. tuberculosis* WGS Illumina read data can be found in the NCBI SRA database under SRA accession ERP024584 (BioProject PRJEB22237). At 833.6 Mbp, the SRA data would provide 189x coverage against the 4.41Mbp reference genome.

Raw reads will be evaluated for quality by FastQC (version 0.11.7). Based on initial quality reports, we will design a trimming strategy to clean up reads through Trimmomatic (version 0.38). We anticipate 3’ trimming will be needed for both forward and reverse sequence sets as well as discard of small reads. Trimming output will be evaluated again through FastQC; additional trimming attempts can be conducted if needed. Due to the large starting size, we aim to not trim below 100x coverage.

Read data will then be mapped to the reference assembly by BWA (version 0.7.17). In a comparison between common read-to-reference aligners (BWA-MEM, BWA-backtrack, Bowtie2, and TMAP), the BWA suite mapping tools covered the largest proportion of Illumina data sets using BWA-MEM and the highest proportion of correct positions in BWA-backtrack, both at the lowest run times.4 We chose this software for its efficiency and the versatility of different packages which will allow us to tailor the read mapping for our data set. Call variant output from BWA will be analyzed by GATK (version 4.0.8.1) for SNPs and small structural variations. Variations will be compiled into a table and can be utilized for further analysis using a web tool for TB genomic variants such as PolyTB (http://pathogenseq.lshtm.ac.uk/).5

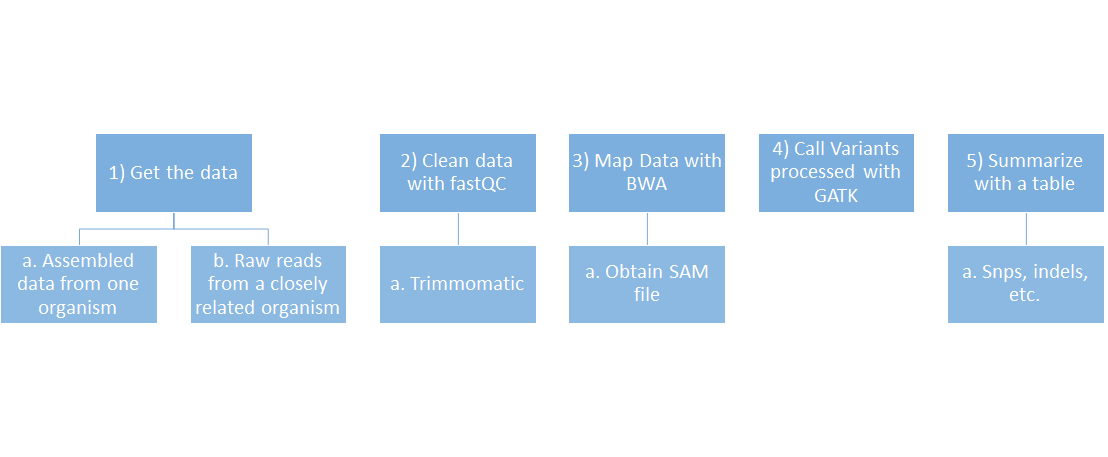


Figure 1: Pipeline procession for read-to-reference variant mapping.

## Division of Labor

The proposal was prepared by Naomi and Ellie. Quality control will be conducted by Naomi and Mason will trim the files using Trimmomatic. Ellie will use BWA to map the trimmed reads to the reference. Call variants will be processed by Ben using GATK. For the final report, results from each section will be prepared by the group member who wrote the code and conducted the analysis. Any conflicts in the pipeline or between contributions will be resolved through discussion with the complete group.

# **Expected outcome**

We expect to receive a read-to-reference genome map at the end of our pipeline with variant calling output from GATK. There will be gaps in coverage due to uneven read coverage during sequencing; some sequence motifs will be naturally biased against polymerase activity while others may align poorly and be thrown out. The average coverage of our raw reads is 189x; high coverage data sets should reduce the chance of false positive SNP calling, though additional PCR and sequencing would be used to confirm promising SNPs.

In the case of poor-quality data, this study generated a total of 17 paired-end WGS Illumina data sets from the same clinical strain; we can run quality control for additional accessions to identify a sample with better read quality. This means we could also compile the accessions to map entirely against the reference as our goal is not to identify population variations, but to account for SNPs that might contribute to cryptic resistance. This would enhance coverage to improve variant calling and allow for more stringent quality control and trimming parameters. In the worst case, if all show poor quality in FastQC outputs the NCBI SRA database has numerous other accessions for antibiotic resistant *M. tuberculosis* clinical strains; we can use the same approach to identify SNPs and SVs associated with either specific antibiotic resistance or multi-drug resistance.

We will be unable to identify inversions and other large modifications due to the use of Burrow Wheeler Transform in BWA and GATK variant call processing. Deletions in the reference will also be undetected as raw reads for this area will be binned from the input. We do not expect missing sequence information to be prominent in the reference strain; deletions are mentioned in several clinical strains but in this comparison scheme would register as unmapped due to the missing sequence information.5 In future studies, if this work warrants further research, larger indels and variants could be investigated by the use of an alternate call variant software, such as Breakdancer, paired with whole genome alignment.

# **References**

1. Irving Cancino-Muñoz, Miguel Moreno-Molina, Victoria Furió, Galo A Goig, Manuela Torres-Puente, Álvaro Chiner-Oms, Luis M Villamayor, Francisco Sanz, María Remedio Guna-Serrano, Iñaki Comas (2019). Cryptic Resistance Mutations Associated With Misdiagnoses of Multidrug-Resistant Tuberculosis. *The Journal of Infectious Diseases*, 220(2), 316–320.
2. Bifani, P., Moghazeh, S., Shopsin, B., Driscoll, J., Ravikovitch, A., & Kreiswirth, B. N. (2000). Molecular characterization of Mycobacterium tuberculosis H37Rv/Ra variants: distinguishing the mycobacterial laboratory strain. *Journal of clinical microbiology*, *38*(9), 3200–3204.
3. Coll, F., McNerney, R., Preston, M. D., Guerra-Assunção, J. A., Warry, A., Hill-Cawthorne, G., ... & Perdigão, J. (2015). Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome medicine*, *7*(1), 51.
4. Magali Jaillard, Maud Tournoud, Faustine Meynier, Jean-Baptiste Veyrieras (2016). Optimization of alignment-based methods for taxonomic binning of metagenomics reads. *Bioinformatics*, 32(12), 1779–1787.
5. Coll, F., Preston, M., Guerra-Assunção, J. A., Hill-Cawthorn, G., Harris, D., Perdigão, J., ... & Glynn, J. R. (2014). PolyTB: A web-based resource designed to explore Mycobacterium tuberculosis complex (MTBC) genomic variation at a global scale.
6. Hirsh, A. E., Tsolaki, A. G., DeRiemer, K., Feldman, M. W., & Small, P. M. (2004). Stable association between strains of Mycobacterium tuberculosis and their human host populations. *Proceedings of the National Academy of Sciences*, *101*(14), 4871-4876.