Identification of *Mycobacterium tuberculosis* Genetic Determinants of Disease Severity

Rita Nóbrega-Martins[1,2], Nuno Osório[1,2], and Tiago Beites[3]

[1] Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Braga, Portugal
[2] ICVS/3B's - PT Government Associate Laboratory, Braga/Guimarães, Portugal
[3] Institute for Research and Innovation in Health (i3S), University of Porto, Porto, Portugal

1 Introduction

1.1 Tuberculosis, a global issue

Tuberculosis (TB) continues to pose a significant global health threat, as evidenced by the latest World Health Organization (WHO) which recorded 7.5 million newly diagnosed cases and 1.3 million TB deaths in 2022 [16]. This infectious disease is caused by the bacillus *Mycobacterium tuberculosis* (*Mtb*), commonly affecting the lungs (pulmonary TB), but also other organs including the kidneys, spine, and brain (extra-pulmonary TB). Transmission occurs via aerosolized droplets expelled by infected individuals, primarily through activities like coughing [16].

Traditionally, TB has been represented in a binary classification, with the distinguishing between active and latent forms of TB based on the presence or absence of clinical symptoms. However, contemporary perspectives increasingly recognize TB as a spectrum of clinical manifestations influenced by immune responses and bacterial interactions, encompassing bacterial replication, persistence, or host immune killing mechanisms [1, 2, 4].

1.2 Pathogenesis of Tuberculosis

Mtb infection elicits a highly complex set of events, whose clinical outcome depends on multiple factors such as various attributes of the pathogen, the host genetics and immune response, and the environmental conditions in which the host lives [1, 2, 16].

Upon inhalation of the pathogen, *Mtb* evades the innate immune response, through the disarming of macrophages' lysosomal trafficking pathways in the lung, multiplying in the intracellular environment and infecting other alveolar macrophages. During the following weeks, an adaptative immune response is developed, resulting either in bacteria elimination or the formation of granulomas, TB characteristic structures composed of T cells, B cells and activated macrophages [2]. After the formation of the granuloma, there can be bacterial

containment and diminishing of the inflammatory response that lead to asymptomatic latent TB, where *Mtb* resides in latency inside the granuloma. Despite this initial containment, latent TB can progress to active TB under conditions of impaired immunity caused by co-morbidities (e.g. diabetes), co-infections (e.g. HIV infection) or host genetics. Active TB manifests with diverse clinical manifestations ranging from pulmonary cavitation to extrapulmonary dissemination [2, 8]. Furthermore, a recent study reported subclinical tuberculosis, characterized by minimal or absent symptoms, as a significant proportion of TB cases and potentially contributing to transmission and challenging traditional case-finding methods, warranting reevaluation of symptom-based diagnostic criteria. Improved case detection, innovative technologies, and tailored treatment strategies are imperative to address this hidden burden and achieve WHO's goal of ending tuberculosis by 2035 [15].

1.3 Current treatment challenges

Currently, TB treatment relies on prolonged combined antibiotic regimens (consisting of combinations of isoniazid, rifampicin, ethambutol, and pyrazinamide, for example). Despite its effectiveness, this regimen faces various challenges related to pathogen adaptation, the drugs' adverse side effects, and host-related problems such as co-infections and co-morbidities, or even poor compliance to the regimen [3, 4]. The emergence of multidrug-resistant TB (MDR-TB) and extensively drug-resistant TB (XDR-TB) is an increasingly concerning problem for the efficacy of TB treatment, with WHO estimating the development of MDR-TB in 410 000 people worldwide [16]. Additionally, despite the approval of novel antibiotics for the treatment of MDR-TB and XDR-TB (e.g. bedaquiline, delamanid, linezolid and pretomanid), mutant strains of *Mtb* resistant to these drugs have already been reported, showing that the pathogen's adaptation and the emergence of new resistance mechanisms are much faster than the process associated with validation and approval of new drugs [2, 4, 5, 10].

In recent years, researchers have discussed potential strategies to achieve better therapeutic strategies for TB, ranging from treatment shortening to faster assessment of potential drugs, to novel adjunctive treatments such as host-directed therapies or antivirulence drugs [18, 4]. Unfortunately, the existing biological knowledge gaps still pose a significant obstacle for the development of any potential novel strategy, since the multifactorial and highly complex nature of this infectious disease still leaves too many open questions.

1.4 Exploring Novel Therapeutic Strategies

As previously mentioned, a promising therapeutical approach for TB is the use of antivirulence drugs, which target the pathogens' ability to cause disease without directly affecting viability. Through the disruption of pathogenic pathways that result in host tissue damage, antivirulence drugs aim to facilitate infection clearance by the immune system while mitigating the emergence of drug resistance.

The development of this type of drugs for TB treatment requires the identification of bona fide virulence factors, key elements produced by *Mtb* to promote disease. Although some virulence factors related to the *Mtb* complex (MTBC) have already been reported (e.g. some proteins that are part of the ESX-1 secretion system), none have been successfully approved for clinical use as a vaccine or treatment.

In 2020, Sousa et al. conducted a study in a patient cohort in Porto, where a pathogen-driven association between genetic diversity within *Mtb* clinical isolates and varying TB disease severity was reported [12]. Even though this study provided valuable insights into the possible link between genetic diversity within the pathogen's isolates, the modulation of the host immune responses and clinical outcomes of TB, the genetic determinants of *Mtb* remain elusive. Identification of these genomic variations (e.g. Single Nucleotide polymorphisms - SNPs), as well as their possible correlations to the clinical outcome of *Mtb* infection, may open the way for the detection of new bona fide virulence factors and, ultimately, for the development of antivirulence drugs for TB.

1.5 Bioinformatic tools for Variant Calling

The identification of genetic determinants underlying various biological phenomena is essential for understanding the mechanisms of disease, and evolution, and for the development of new therapies. As analysis of genomic data involves a considerable amount of data, time and effort, bioinformatics tools play a pivotal role in this process, offering algorithms and computational methods to analyze vast amounts of genomic data in a fast and efficient way. These tools enable researchers to uncover genetic determinants by identifying patterns, variations, and associations within genomes, transcriptomes, proteomes and phenotypes.

Variant calling is a fundamental process in bioinformatics that involves identifying differences (such as SNPs, insertions or deletions) in genomic sequences compared to a reference genome. Variant calling workflows typically involve several steps: quality control, alignment to a reference genome, post-alignment processing (which may include indexing and marking duplicates), variant detection, variant filtering and quality control, and annotation.

1.6 Variant Calling for *Mtb*

Existing tools for variant calling in *Mtb* range from those integrated into platforms, like Galaxy [7] and SAM-TB [17], to pipelines such as MAGMA [6], MTB-VCF [9], MycoVarP [13] and TBProfiler [14]. While these bioinformatics tools and pipelines offer high-confidence variant identification, they often lack flexibility and robustness, limited to certain objectives such as the detection of drug resistance-related variants [13]. For instance, tools integrated into platforms reliant on web servers present challenges such as prolonged waiting times due to simultaneous user access, limited configurability, and stability issues. Moreover, they lack comprehensive features specifically tailored for variant calling in *Mtb*.

Moreover, variant calling for *Mtb* faces challenges due to the complexity of the *Mtb* genome, particularly with short-read sequencing data. Repetitive regions within the genome can lead to ambiguous read mappings and difficulties in accurate variant identification. Additionally, current approaches remove all SNPs and Indels within specified position intervals, including gene regions like the PE/PPE families, which may be an overly conservative approach when performing explorative analysis, leading to potentially missing or misidentifying genetic variants.

An example of this possible loss of information is reflected in the reported discrepancy between phenotypic differences among *Mtb* isolates and the absence of corresponding genetic differences detected through SNP analysis. It has been reported instances where, even after the exclusion of the interference of host-related influencing factors, *Mtb* isolates exhibit distinct phenotypes but do not show differences in SNP profiles [12, 11]. This discrepancy raises questions regarding the interpretation of phenotypic and genomic data, indicating the need for refinement in variant calling pipelines to avoid overly stringent criteria during analysis, which may overlook genuine genetic differences. A possible solution might be through the use of Base Quality Score Recalibration (BQSR) methods, which construct a covariation model based on a comprehensive set of known variants, allowing the exclusion of SNPs with low-quality scores while retaining those in problematic regions.

Developing an enhanced variant calling pipeline tailored explicitly for *Mtb*, capable of identifying variants overlooked by existing pipelines, would not only mitigate the limitations of current tools but also facilitate the discovery of novel genetic determinants crucial for advancing research on *Mtb*, tuberculosis, and the development of innovative therapies. Such a pipeline should offer flexibility, robustness, and comprehensive analysis capabilities to effectively address the complexities of variant calling in *Mtb* genomes.

2 Objectives

The significant knowledge gaps surrounding *Mtb* biology and the intricate pathology of TB pose difficult challenges to the advancement of TB research and the development of novel effective therapeutic strategies. The discovery of an association between genetic diversity in well-characterized clinical isolates of *Mtb* and TB severity would not only provide new insights into the mechanisms underlying this infectious disease but also pave the way for further research of potential new targets for TB treatment.

Therefore, the main goal of this project is to identify *Mtb* genetic determinants of TB clinical outcomes. To accomplish this, the project outlines the following aims:

1. **Identify genetic variations of *Mtb*:** through the development of a bioinformatic pipeline optimized for *Mtb*, capable of identifying genetic variations overlooked by existing pipelines in each one of the 149 clinical isolates compared to an ancestral reference sequence.

2. **Correlate the genetic variants with disease severity:** through the application of statistical and machine learning methodologies that establish and investigate correlations between specific genetic variants and TB severity.

Through the development of a new tool for variant detection and the analysis of correlations between specific genetic variations and the clinical outcome of TB, this project offers an exciting opportunity to contribute to the understanding of TB's genetic underpinnings as well as to the development of innovative therapeutic strategies.

3 Tasks and Methodology

3.1 Task 1 - Development of an optimized pipeline for the identification of genetic variants in *Mtb*

The initial phase of the project will involve the development of an optimized pipeline for the identification of genetic variants within the sequenced genome of each clinical isolate from the study cohort in Porto (149 isolates in total) 1.

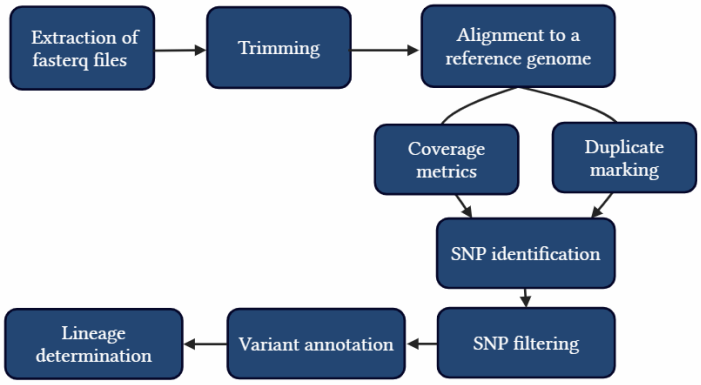

Fig. 1. Description of the pipeline optimized for *Mtb* that will allow the detection of genetic variations overlooked by existing pipelines such as MAGMA [6] and MTB-VCF [9]. Tools like BBDuk will remove low-quality bases and adapter sequences from raw sequencing data. Customizable trimming parameters ensure precise adaptation to data characteristics. Kraken2 is then employed for taxonomic classification, identifying contaminant sequences. Next, BWA MEM aligns trimmed, uncontaminated reads to a reference sequence, enabling the detection of SNPs and the subsequent filtering and annotation.

After the extraction of the faster files corresponding to the SRA of the isolates sequenced and a quality control step, Initially, reads will be trimmed to ensure the removal of low-quality bases and adapter sequences from raw sequencing data. BBDuk, which offers options for adapter trimming, quality filtering,

and artefact removal, might be an option for this step. After trimming, the data will be analysed by Kraken2, a software tool used for taxonomic classification of DNA sequences, particularly for microbiome analysis. In this project, Kraken2 will allow the detection of potential contaminant sequences from non-target organisms. After ensuring that only relevant genetic information of *Mtb* will be considered, the trimmed and uncontaminated sequencing reads will be aligned to an ancestral reference sequence through BWA MEM, allowing for the identification of genetic variations, including SNPs, insertions, and deletions.

Subsequent processes will involve marking duplicates with Samtools markdup for further elimination, minimizing the redundancy in the data and improving the accuracy of variant calling. After variant calling, SNP quality filtering will be applied to ensure the reliability of identified variants by excluding low-quality or erroneous calls. The final step of the pipeline for the identification of the genetic determinants of *Mtb* will consist in the determination of the lineage of the clinical isolates of the pathogen based on genetic variations, providing insights into the evolutionary relationships and population structure of *Mtb* strains, useful for the interpretation of genetic diversity and disease severity correlations.

The final phase of this first task would be the use of cutting-edge technologies, such as Snakemake and Docker, to ensure reproducibility, scalability and accessibility for researchers with a low level of coding knowledge.

3.2 Task 2 - Study the correlation between the genetic variations identified and TB clinical outcome

To find and study the correlations between the genetic variations identified in the clinical isolates of the bacteria and the severity described in the corresponding patient, statistical and machine learning approaches will then be applied to identify statistically significant associations between genetic variations and clinical outcomes, elucidating the genetic underpinnings of TB severity. A possible statistical methodology for the determination and analysis of the correlation may be linear regression, widely used to model the relationship between a continuous outcome variable (e.g. TB severity) and an independent variable (e.g. genetic variants). Machine learning algorithms such as the Random Forest, Gradient Boosting Machines or Deep Learning can potentially help uncover subtle relationships and interactions that may not be apparent with traditional statistical methods.

References

1. Boom, W.H., Schaible, U.E., Achkar, J.M.: The knowns and unknowns of latent mycobacterium tuberculosis infection. Journal of Clinical Investigation **131** (2021). https://doi.org/10.1172/JCI136222
2. Chandra, P., Grigsby, S.J., Philips, J.A.: Immune evasion and provocation by mycobacterium tuberculosis. Nature Reviews Microbiology **20**, 750–766 (2022). https://doi.org/10.1038/s41579-022-00763-4

3. Conradie, F., Diacon, A.H., Ngubane, N., Howell, P., Everitt, D., Crook, A.M., Mendel, C.M., Egizi, E., Moreira, J., Timm, J., McHugh, T.D., Wills, G.H., Bateson, A., Hunt, R., Niekerk, C.V., Li, M., Olugbosi, M., Spigelman, M.: Treatment of highly drug-resistant pulmonary tuberculosis. New England Journal of Medicine **382**, 893–902 (2020). https://doi.org/10.1056/NEJMoa1901814
4. Dartois, V.A., Rubin, E.J.: Anti-tuberculosis treatment strategies and drug development: challenges and priorities. Nature Reviews Microbiology **20**, 685–701 (2022). https://doi.org/10.1038/s41579-022-00731-y
5. Gómez-González, P.J., Perdigao, J., Gomes, P., Puyen, Z.M., Santos-Lazaro, D., Napier, G., Hibberd, M.L., Viveiros, M., Portugal, I., Campino, S., Phelan, J.E., Clark, T.G.: Genetic diversity of candidate loci linked to mycobacterium tuberculosis resistance to bedaquiline, delamanid and pretomanid. Scientific Reports **11**, 19431 (2021). https://doi.org/10.1038/s41598-021-98862-4
6. Heupink, T.H., Verboven, L., Sharma, A., Rennie, V., de Diego Fuertes, M., Warren, R.M., Rie, A.V.: The magma pipeline for comprehensive genomic analyses of clinical mycobacterium tuberculosis samples. PLOS Computational Biology **19**, e1011648 (2023). https://doi.org/10.1371/journal.pcbi.1011648
7. van Heusden, P., Gladman, S., Lose, T.: M. tuberculosis variant analysis (galaxy training materials)
8. Orgeur, M., Sous, C., Madacki, J., Brosch, R.: Evolution and emergence of mycobacterium tuberculosis. FEMS Microbiology Reviews (2024). https://doi.org/10.1093/femsre/fuae006
9. Razzak, S.A., Hasan, Z., Azim, M.K., Kanji, A., Shakoor, S., Hasan, R.: A versatile tool for precise variant calling in mycobacterium tuberculosis genetic polymorphisms. bioRxiv (2023). https://doi.org/10.1101/2023.07.24.550283
10. Rocha, D.M.G.C., Magalhães, C., Cá, B., Ramos, A., Carvalho, T., Comas, I., Guimarães, J.T., Bastos, H.N., Saraiva, M., Osório, N.S.: Heterogeneous streptomycin resistance level among mycobacterium tuberculosis strains from the same transmission cluster. Frontiers in Microbiology **12** (2021). https://doi.org/10.3389/fmicb.2021.659545
11. Shamputa, I.C., Rigouts, L., Eyongeta, L.A., Aila, N.A.E., van Deun, A., Salim, A.H., Willery, E., Locht, C., Supply, P., Portaels, F.: Genotypic and phenotypic heterogeneity among mycobacterium tuberculosis isolates from pulmonary tuberculosis patients. Journal of Clinical Microbiology **42**, 5528–5536 (2004). https://doi.org/10.1128/JCM.42.12.5528-5536.2004
12. Sousa, J., Cá, B., Maceiras, A.R., Simões-Costa, L., Fonseca, K.L., Fernandes, A.I., Ramos, A., Carvalho, T., Barros, L., Magalhães, C., Álvaro Chiner-Oms, Machado, H., Veiga, M.I., Singh, A., Pereira, R., Amorim, A., Vieira, J., Vieira, C.P., Bhatt, A., Rodrigues, F., Rodrigues, P.N., Gagneux, S., Castro, A.G., Guimarães, J.T., Bastos, H.N., Osório, N.S., Comas, I., Saraiva, M.: Mycobacterium tuberculosis associated with severe tuberculosis evades cytosolic surveillance systems and modulates il-1 production. Nature Communications **11** (2020). https://doi.org/10.1038/s41467-020-15832-6
13. Swargam, S., Kumari, I., Kumar, A., Pradhan, D., Alam, A., Singh, H., Jain, A., Devi, K.R., Trivedi, V., Sarma, J., Hanif, M., Narain, K., Ehtesham, N.Z., Hasnain, S.E., Ahmad, S.: Mycovarp: Mycobacterium variant and drug resistance prediction pipeline for whole-genome sequence data analysis. Frontiers in Bioinformatics **1** (2022). https://doi.org/10.3389/fbinf.2021.805338, https://www.frontiersin.org/articles/10.3389/fbinf.2021.805338

14. Verboven, L., Phelan, J., Heupink, T.H., Rie, A.V.: Tbprofiler for automated calling of the association with drug resistance of variants in mycobacterium tuberculosis. PLOS ONE **17**, e0279644 (2022). https://doi.org/10.1371/journal.pone.0279644
15. Wei, X., Zhang, W.: The hidden threat of subclinical tuberculosis. The Lancet Infectious Diseases (2024). https://doi.org/10.1016/S1473-3099(24)00069-0
16. (WHO), W.H.O.: Global tuberculosis report 2023. World Health Organization (2023), https://www.who.int/publications/i/item/9789240083851
17. Yang, T., Gan, M., Liu, Q., Liang, W., Tang, Q., Luo, G., Zuo, T., Guo, Y., Hong, C., Li, Q., Tan, W., Gao, Q.: Sam-tb: a whole genome sequencing data analysis website for detection of mycobacterium tuberculosis drug resistance and transmission. Briefings in Bioinformatics **23** (2022). https://doi.org/10.1093/bib/bbac030
18. Young, C., Walzl, G., Plessis, N.D.: Therapeutic host-directed strategies to improve outcome in tuberculosis. Mucosal Immunology **13**, 190–204 (2020). https://doi.org/10.1038/s41385-019-0226-5