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

Abstract. Tuberculosis (TB), caused by *Mycobacterium tuberculosis* (*Mtb*), remains a major global health challenge. Understanding the genetic determinants of *Mtb* that influence TB severity is crucial for developing effective therapeutic strategies. This study aimed to identify genetic variations in *Mtb* isolates and correlate them with TB clinical outcomes. A comprehensive bioinformatic pipeline optimized for variant calling in *Mtb* was developed using a combination of tools for quality control, read mapping, post-processing, and variant calling. Using *bcftools* and LoFreq, the pipeline successfully detected a diverse array of genetic variants in clinical *Mtb* isolates. However, ongoing optimization and validation efforts are essential to enhance the accuracy and reliability of these findings. Future work will focus on refining the variant calling methodologies and validating the results obtained. By laying the groundwork for improved variant identification in *Mtb*, this project has the potential to significantly advance our understanding of TB pathogenesis and contribute to the development of more effective treatment strategies.
Keywords: *Mycobacterium tuberculosis*, Tuberculosis severity, Variant Calling, Correlation

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) report, 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 distinction 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 infection, *Mtb* can be contained by the host immune response, leading to asymptomatic latent TB. In this state, the bacteria can reside in a dormant state within the host without causing visible or detectable granulomas. 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 comprises 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 [19,4]. 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 drug targets.

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* driving disease severity 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 [18], 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 the usage 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 (Task 1, 1).
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 (Task 2).

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 Methodology

The initial phase of this project involved developing a comprehensive pipeline for identifying genetic variants within the sequenced genomes of clinical isolates from a study cohort in Porto, comprising 149 isolates. This pipeline was implemented using a Jupyter notebook with a Bash kernel, allowing for modifications at specific steps and better visualization of the results, which facilitated the optimization of each step. To manage the various tools required for the analysis, a Conda environment named "MTB_calling" was created, encompassing all necessary software and dependencies 1. Each step of the pipeline was encapsulated within individual Bash functions, enabling both independent and collective execution of the pipeline's steps. As the sequenced genomes of the *Mtb* isolates have been stored in the NCBI database and given a Sequence Read Archive (SRA) accession number as a unique identifier, all functions were designed to accept a text file input containing a list of SRA numbers, with each isolate represented on a separate line. This notebook can be found in the following repository: ProjetoMtb.

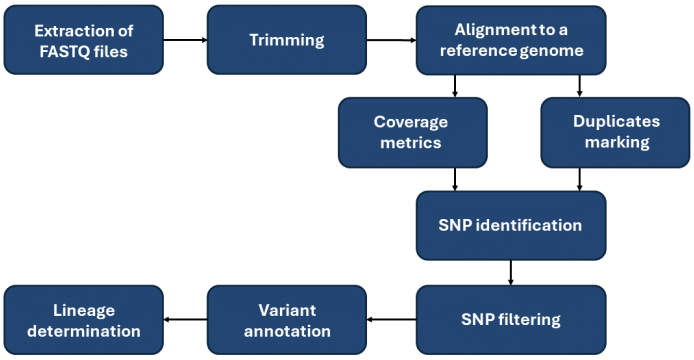

Fig. 1. Description of the pipeline optimized for *Mtb* that will allow the detection of genetic variations overlooked by existing pipelines such as MAGMA [6] and MTB-VCF [9]. Tools like BBDuk will remove low-quality bases and adapter sequences from raw sequencing data. Customizable trimming parameters 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.

4 Methods

4.1 Data Extraction and Pre-Processing

Mtb clinical isolates' raw reads were retrieved from the NCBI database using the *extractSRAfiles* function, which accepts a list of SRA accession numbers corresponding to the isolates. For each accession number, it creates a directory and employs the *prefetch* and *fasterq-dump* tools to download two FASTQ files containing the raw reads sequenced for each sample. The raw sequencing reads were subjected to a quality control assessment using FastQC, followed by trimming with BBDukk. FastQC generated comprehensive reports that include various quality metrics, such as per base sequence quality, GC content, and the presence of adapter sequences. BBDuk was then employed for trimming and filtering the reads, removing low-quality bases (Phred score < 20) from both ends of the reads and eliminating adapter sequences introduced during the sequencing process. The output of this step was a set of clean, high-quality reads ready for mapping.

4.2 Mapping

The Burrows-Wheeler Aligner (BWA-MEM algorithm) was used for the mapping due to its efficiency and accuracy in aligning short reads to a reference genome.

Table 1. Tools present in the conda environment created for the pipeline.

Tool(s)	Version	Tool(s)	Version
bcftools	1.11	pathogen-profiler	2.0.1
bedtools	2.30.0	samtools	1.12
bwa	0.7.17	sra-tools	3.1.0
bzip2	1.0.8	tabixpp	1.1.2
fastqc	1.0.8	tb-profiler	4.3.0
lofreq	2.1.5	trimmomatic	0.39
mosdepth	0.3.3	vcflib	1.0.9
parallel	20160622		

The *map_BWA* function performs the mapping of the trimmed reads to the ancestral reference genome, a widely used reference due to its extensive characterization and comprehensive annotation of genes and regulatory elements. Using *bwa mem*, the trimmed reads were aligned to the reference genome, generating sequence alignment and map (SAM) files containing the alignment information. These SAM files were then converted to binary alignment and map (BAM) files, which were subsequently sorted by coordinates using *samtools sort*. To facilitate efficient and rapid access to specific regions of the aligned reads, the sorted BAM files were indexed using *samtools index*.

4.3 Post-Processing

Post-processing aims to minimize data redundancy and improve the accuracy of variant calling. Using*samtools markup*, duplicate reads, likely resulting from PCR amplification, were identified and flagged to prevent these duplicates from biasing the variant calling process. The depth and distribution of the sequencing reads across the genome were assessed using Mosdepth, a tool that calculates coverage at each position in the genome, providing insights into sequencing uniformity and depth. Reads with a mapping quality score below 20 were filtered out to ensure that only high-confidence reads contributed to the coverage calculations.

4.4 Variant Calling

To identify genetic variants in the *Mtb* isolates, different strategies were employed in the variant calling process. Initially, *bcftools mpileup* was used to generate a pileup of read alignments, followed by *bcftools call* to identify variants. This method provided a basic variant call format (VCF) file containing all detected variants without additional filtering.

To refine the variant identification, a more stringent approach was applied. This method also involved *bcftools mpileup* and *bcftools call*, but with additional filters such as a minimum read quality of 20 to ensure high-confidence variant calls. The representation of the variants was standardized using *bcftools norm*,

and additional functional information was added to the VCF files using *bcftools annotate*.

Due to its high sensitivity in detecting low-frequency variants, crucial for identifying minor variants, the LoFreq tool was also employed. The process involved two main steps: post-processing and variant calling. In the post-processing step, the *lofreq_ postprocessing* function performed indel realignment on duplicated BAM files to refine read alignments around indels, followed by indexing the realigned BAM files. Subsequently, the *LoFreq_ calling* function used *lofreq call-parallel* to detect variants in the realigned BAM files, producing VCF files with annotated and high-confidence variant calls [17].

5 Results and Discussion

In this project, we aimed to close the knowledge gap regarding TB clinical outcomes. To achieve this, a comprehensive bioinformatic pipeline specifically optimized for *Mtb* was developed to accurately identify genetic variations across clinical isolates of the pathogen.

The workflow begins with the acquisition of raw sequencing data, which undergoes initial quality control and pre-processing to ensure high-quality input for subsequent steps. The clean reads are then aligned to the ancestral reference genome using the BWA algorithm, facilitating the accurate mapping of the sequence data. After the mapping, the pipeline employs the *samtools markup* to mark duplicates and Mosdepth to calculate coverage, improving data integrity.

To assess the effectiveness of different variant calling methods, different strategies and tools were utilized: a basic variant calling using *bcftools mpileup* and *bcftools call* with no filtering options; a similar approach with additional filters, annotations and representation's standardization; and variant calling using the LoFreq method. The VCF files generated by each method provided detailed information about each detected variant, including chromosomal position (CHROM), genomic position of the variant (POS), a unique identifier for the variant (ID), reference (REF) and alternate (ALT) alleles, quality scores (QUAL), a filter status indicating whether the variant passed certain quality thresholds or filters (FILTER). In the INFO and FORMAT fields include genotype information for sampled individuals and additional annotations. The format typically includes fields for variant type (SNP, indel, etc.), quality metrics (like depth of coverage and mapping quality), allele frequency, and other sample-specific metrics.

The initial variant calling using *bcftools mpileup* and *bcftools call* without any additional filtering options generates a VCF file with minimal processing (Table 2). This approach allowed the detection of a broad range of variants without any bias from filters. However, the lack of filters also allowed the inclusion of low-confidence calls, likely leading to false positives, and a lack of functional annotations and quality metrics.

To improve the quality, consistency and interpretability of the results, a more refined approach, which involved additional filtering, normalization and annotation steps, was performed. The filtering options used in *bcftools mpileup* and

Table 2. Representative variants identified by variant calling using *bcftools mpileup* and *bcftools call* without any additional filtering or annotations.

CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	DRR130078
MTB_anc	1849	.	C	A	225.007	.	DP=82;VDB=0.36046;SGB=-0.693147; MQSB=1;MQ0F=0;AF1=1;AC1=1; DP4=0,0,36,26;MQ=60;FQ=-999	GT:PL	05:15,0
MTB_anc	49690	.	GCCC	GC	217.469	.	INDEL;IDV=110;IMF=0.982143;DP=112; VDB=0.902281;SGB=-0.693147;MQSB=1; MQ0F=0;AF1=1;AC1=1;DP4=8,16,40,48; MQ=60;FQ=-999;PV4=0.355274,1,1,1	GT:PL	1:255,72,0
MTB_anc	162505	.	A	G	225.007	.	DP=98;VDB=0.974376;SGB=-0.693147; MQSB=1;MQ0F=0;AF1=1;AC1=1; DP4=0,0,39,34;MQ=60;FQ=-999	GT:PL	05:15,0

bcftools call include parameters such as *–count-orphans*, *–no-BAQ*, *–min-MQ 20*, and *–min-BQ 20*, which help in including orphan reads (reads where only one read in a pair is mapped) and excluding low-quality reads and bases (those with a mapping or base quality inferior to 20). These additional steps are crucial for reducing false positives and enhancing the reliability of the detected variants. Additionally, the normalization step with *bcftools norm* ensures that the VCF file (Table 3) has standard representations, important for downstream analysis and comparison. The annotation steps involve adding relevant biological and functional information to the variants. This includes tagging with excluded loci, blind spots, lineage-specific SNPs, and epitopes from various datasets as can be observed in the representative cases represented in Table 3. This comprehensive annotation enriches the VCF file with contextually important data, helping interpret the biological significance of the variants and identify variants of clinical or research interest.

This approach allowed the detection of a high number of variants, suggesting that, while filtering may have reduced noise from low-quality data, several factors contribute to a high detection level, such as the inclusion of orphan reads, the use of the multiple layers of annotations and the multiallelic caller.

Table 3. Representative variants identified by variant calling using \textit{bcftools mpileup} and \textit{bcftools call} with additional filtering, representation' standardization and annotations.

CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	DRR130078
MTB_anc	1849	MTB_anc:1849	C	A	225	.	DP=87;ADF=0,37;ADR=0,30; AD=0,67;SCR=6;VDB=0.289449; SGB=-0.693147;MQSB=1; MQ0F=0;AC=1;AN=1; DP4=0,0,37,30;MQ=60; Mappability=1	GT:PL:DP:SP:ADF :ADR:AD:SCR	1:255,0:67:0: 0,37:0,30:0,67:6
MTB_anc	49690	MTB_anc:49690	GCC	G	228	.	INDEL;IDV=113;IMF=0.982609; DP=115;ADF=9,40;ADR=16,50; AD=25,90;SCR=2;VDB=0.851904; SGB=-0.693147;MQSB=1; MQ0F=0;AC=1;AN=1; DP4=9,16,40,50;MQ=60; Mappability=1	GT:PL:DP:SP:ADF: ADR:AD:SCR	1:255,0:115:3:9,40: 16,50:25,90:2
MTB_anc	162505	MTB_anc:162505	A	G	225	.	DP=101;ADF=0,41;ADR=0,38; AD=0,79;SCR=0;VDB=0.996967; SGB=-0.693147;MQSB=1; MQ0F=0;AC=1;AN=1; DP4=0,0,41,38;MQ=60; Mappability=1	GT:PL:DP:SP: ADF:ADR:AD:SCR	1:255,0:79:0:0,41 :0,38:0,79:0

As this project aims to identify variants often missed by standard pipelines, the LoFreq tool was evaluated for its sensitivity in detecting low-frequency variants. Utilizing a Bayesian model, LoFreq distinguishes genuine variants from sequencing errors, thereby improving specificity compared to tools such as *bcftools*. Initially, LoFreq was employed for post-processing, aligning BAM files and subsequently identifying variants within these alignments, resulting in VCF files containing annotated and high-confidence variant calls. Following this, *bcftools* was utilized to further refine the variant calls by implementing stringent filters and integrating comprehensive annotations into the generated VCF files. This combined approach yielded VCF files annotated with information on excluded loci, mappability, lineage-specific SNPs, and epitope regions, as detailed in Table 4.

Table 4. Representative variants identified using the LoFreq variant calling tool known for its sensitivity in detecting low-frequency variants, with subsequential annotation using *bcftools*.

CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
MTB_anc	1849	MTB_anc:1849	C	A	2497	PASS	DP=70;AF=1;SB=0;DP4=0,0,39,31;Mappability=1
MTB_anc	162505	MTB_anc:162505	A	G	1029	PASS	DP=29;AF=1;SB=0;DP4=0,0,16,13;Mappability=1
MTB_anc	208782	MTB_anc:208782	A	C	2087	PASS	DP=58;AF=1;SB=0;DP4=0,0,29,29;Mappability=1; Lineage_tag=!lineage,2.1,tbprofiler

The *vcf-compare* tool was employed to assess overlaps and discrepancies among the three VCF datasets, providing insights into the concordance and unique detections of each method represented in the Venn diagrams in Figure 2.

Overall, all strategies used for variant calling detected a wide range of variants for each sample, with a substantial number of variants shared across all methods. The *bcftools*-based version with additional filtering identified the largest total number of variants, exhibiting partial overlap with both the simple and the LoFreq strategies. In contrast, the simplest method detected the fewest variants overall, with minimal overlap observed with the other methods. Notably, LoFreq identified a significant number of variants not detected by the bcftools-based methods.

The observed contrasts in variant detection between the methods may be attributed to several factors. The higher total number of detected variants by the filtered version of the *bcftools*-based approach is likely due to increased sensitivity, the inclusion of orphan reads, and the incorporation of additional information, such as quality scores and read depth. Furthermore, the use of the multiallelic caller and the processing of the reference genome in equally sized intervals through the *create_ intervals* function also increase detection and minimize the chances of missing variants in poorly covered regions.

On the other hand, while LoFreq detects a smaller number of variants overall, it identifies a significant number of unique variants not detected by the *bcftools*-based methods. This can be attributed to its sensitivity for low-frequency variants, although it may be less effective at detecting high-frequency variants or

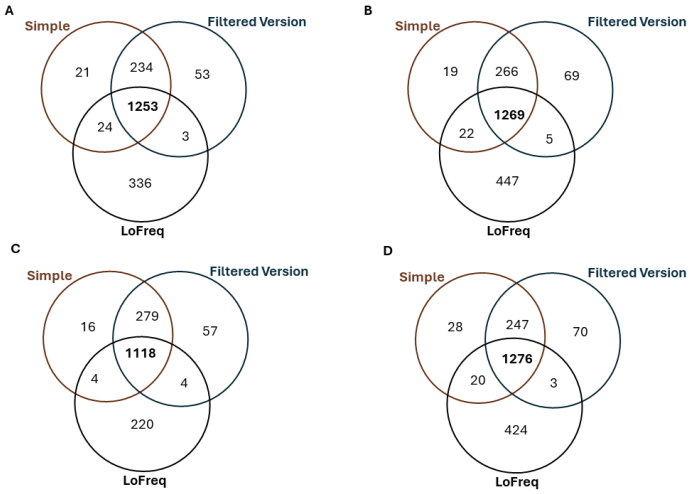

Fig. 2. Comparison of variant detection strategies. Diagrams illustrate overlaps and unique variants identified in a set of 4 samples by the three variant calling methods: a simple *bcftools*-based approach (Simple), a refined *bcftools*-based approach with filtering (Filtered Version), and an approach using LoFreq as the variant caller.

variants in regions with complex annotations. Additionally, this sensitivity can come at the cost of lower precision, as LoFreq has been reported to call mutations that are not true positives [17].

These results suggest that the *bcftools*-based strategy with additional filtering, normalization, and annotation, is a more robust and general-purpose strategy, while LoFreq is more sensitive and suitable for for detecting specific low-frequency variants. However, further analysis is required to assess the accuracy of calls from each method, as these results only provide insights into variant call numbers and overlap.

To ensure highly confident variant calls, the proposed variants must undergo a rigorous filtering and selection process. This involves applying SNP quality filtering to exclude low-quality or erroneous calls and selecting variants based on their reliability and accuracy. Several strategies can be employed to optimize the pipeline for variant calling. These include implementing additional filtering and normalization steps during reads processing and optimizing parameters to increase the performance of the tools used, thereby improving the accuracy of variant calls. Additionally, detecting potential contaminant sequences from non-target organisms using tools like Kraken2 can help improve the accuracy of variant calls by removing irrelevant sequences. Another relevant step for variant detection optimization is the integration of multiple variant callers. Combining and cross-validating results from several tools with different degrees of sensitivity can produce a more comprehensive and accurate set of variants.

Furthermore, performing long-read sequencing (e.g., using technologies like PacBio or Oxford Nanopore) can significantly increase the confidence and accuracy of the results. Long-read sequencing provides higher coverage, better identification of large structural variants, which are often missed by short-read sequencing, and more accurate mapping and variant calling in repetitive regions of the genome, a common issue in existing pipelines for *Mtb* variant calling.

The final phase of the pipeline development will involve leveraging cutting-edge technologies like Snakemake and Docker to ensure reproducibility, scalability, and accessibility for researchers with varying levels of coding expertise. Snakemake allows the creation of reproducible workflows that can be easily shared and replicated by other researchers, while Docker allows the concatenation of the pipeline, allowing it to run consistently across different computing environments and scale to larger datasets.

After the validation of the pipeline and identification of high-confident and high-quality variants, correlations between genetic variations identified in clinical isolates and TB clinical outcomes will be established, through statistical and machine-learning approaches. Methods such as linear regression will model relationships between continuous outcome variables (e.g., TB severity) and genetic variants. Advanced machine learning algorithms, like Random Forest, will help uncover subtle relationships and interactions that traditional statistical methods might miss. This comprehensive analysis will elucidate the genetic underpinnings of TB severity, providing insights that could lead to more effective treatment strategies.

6 Conclusion

This project established a bioinformatic pipeline capable of identifying a wide range of genetic variants in *Mtb* isolates. The pipeline, which integrates high-throughput sequencing data retrieval, quality control, read mapping, reads processing, and multiple variant calling methodologies, demonstrated robust variant detection and annotation capabilities.

However, the discrepancies in variant calls between different methods and the presence of low-confidence calls indicate that further optimization and validation are necessary to enhance the accuracy and reliability of the pipeline. Specific areas for improvement include refining the reads processing steps, optimizing variant calling parameters, and incorporating additional validation steps using independent datasets.

Future work will focus on evaluating the biological significance of the identified variants and correlating these genetic variants with TB disease severity and treatment response. This project lays the groundwork for improved variant identification in *Mtb*, which could significantly advance our understanding of TB pathogenesis and contribute to the development of more effective treatment strategies.

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. Wilm, A., Aw, P.P.K., Bertrand, D., Yeo, G.H.T., Ong, S.H., Wong, C.H., Khor, C.C., Petric, R., Hibberd, M.L., Nagarajan, N.: LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Research **40**(22), 11189–11201 (2012). https://doi.org/10.1093/nar/gks918, https://doi.org/10.1093/nar/gks918
18. 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
19. 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