Mycobacterium tuberculosis genetic features associated with pulmonary tuberculosis severity

Charlotte Genestet, Guislaine Refrégier, Elisabeth Hodille, Rima Zein-Eddine, Adrien Le Meur, Fiona Hak, Alexia Barbry, Emilie Westeel, Jean-Luc Berland, Astrid Engelmann, et al.

Journal Pre-proof

Mycobacterium tuberculosis genetic features associated with pulmonary tuberculosis severity

Charlotte Genestet , Guislaine Refrégier , Elisabeth Hodille , Rima Zein-Eddine , Adrien Le Meur , Fiona Hak , Alexia Barbry , Emilie Westeel , Jean-Luc Berland , Astrid Engelmann , Isabelle Verdier , Gérard Lina , Florence Ader , Stéphane Dray , Laurent Jacob , François Massol , Samuel Venner , Oana Dumitrescu , on behalf of the Lyon TB study group

PII: S1201-9712(22)00560-4
DOI: https://doi.org/10.1016/j.ijid.2022.10.026
Reference: IJID 6463

To appear in: *International Journal of Infectious Diseases*

Received date: 3 August 2022
Revised date: 13 October 2022
Accepted date: 15 October 2022

Please cite this article as: Charlotte Genestet , Guislaine Refrégier , Elisabeth Hodille , Rima Zein-Eddine , Adrien Le Meur , Fiona Hak , Alexia Barbry , Emilie Westeel , Jean-Luc Berland , Astrid Engelmann , Isabelle Verdier , Gérard Lina , Florence Ader , Stéphane Dray , Laurent Jacob , François Massol , Samuel Venner , Oana Dumitrescu , on behalf of the Lyon TB study group, Mycobacterium tuberculosis genetic features associated with pulmonary tuberculosis severity, *International Journal of Infectious Diseases* (2022), doi: https://doi.org/10.1016/j.ijid.2022.10.026

Mycobacterium tuberculosis genetic features associated with pulmonary tuberculosis severity

Charlotte Genestet[1,2*], Guislaine Refrégier[3], Elisabeth Hodille[1,2], Rima Zein-Eddine[3,4], Adrien Le Meur[3], Fiona Hak[3], Alexia Barbry[1,2], Emilie Westeel[5], Jean-Luc Berland[5], Astrid Engelmann[6], Isabelle Verdier[6], Gérard Lina[1,2,7], Florence Ader[1,8], Stéphane Dray[9], Laurent Jacob[9], François Massol[10,11], Samuel Venner[9], Oana Dumitrescu[1,2,7] on behalf of the Lyon TB study group

1. CIRI - Centre International de Recherche en Infectiologie, Ecole Normale Supérieure de Lyon, Université Claude Bernard Lyon-1, Inserm U1111, CNRS UMR5308, Lyon, France

2. Hospices Civils de Lyon, Institut des Agents Infectieux, Laboratoire de bactériologie, Lyon, France

3. Université Paris-Saclay, CNRS, AgroParisTech, Ecologie Systématique et Evolution, Orsay, France. Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris Sud, Université Paris-Saclay, Gif-sur-Yvette, France.

4. Laboratory of Optics and Biosciences, CNRS-INSERM-Ecole Polytechnique, Palaiseau, France

5. Fondation Mérieux, Emerging Pathogens Laboratory, Lyon, France.

6. Centre Hospitalier Fleyriat, Bourg-en-Bresse, France

7. Université Lyon 1, Facultés de Médecine et de Pharmacie de Lyon, Lyon, France

8. Hospices Civils de Lyon, Service des Maladies infectieuses et tropicales, Lyon, France.

9. Biometrics and Evolutionary Biology Laboratory, CNRS UMR 5558, Université Lyon 1, Villeurbanne, France

10. UMR 8198 Evo-Eco-Paleo, SPICI Group, University of Lille, Lille, France.

11. CNRS, CHU Lille, Institut Pasteur de Lille, U1019-UMR 9017-CIIL-Center for Infection and Immunity of Lille, University of Lille, Lille, France.

***Corresponding author:** charlotte.genestet@gmail.com

Charlotte Genestet, PhD

Centre International de Recherche en Infectiologie, 7 rue Guillaume Paradin, 69003 Lyon, France

Tel: +33 (0)4 72 07 17 09

ABSTRACT

Objective: *Mycobacterium tuberculosis* (Mtb) infections result in a wide spectrum of clinical presentations but without proven Mtb genetic determinants. Herein, we hypothesised that genetic features of Mtb clinical isolates, such as specific polymorphisms or micro-diversity, may be linked to tuberculosis (TB) severity.

Methods: 234 pulmonary TB patients (including 193 drug-susceptible and 14 mono-resistant cases diagnosed between 2017 and 2020 and 27 multidrug-resistant cases diagnosed between 2010 and 2020) were stratified according to TB disease severity and Mtb genetic features were explored using whole genome sequencing, including heterologous single nucleotide polymorphism (SNP) calling to explore micro-diversity. Finally, we performed a structural equation modelling (SEM) analysis to relate TB severity to Mtb genetic features.

Results: Clinical isolates from patients with mild TB carried mutations in genes associated with host-pathogen interaction, while those from patients with moderate/severe TB carried mutations associated with regulatory mechanisms. Genome-wide association study identified a SNP in the promoter of the gene coding for the virulence regulator EspR statistically associated with moderate/severe disease. SEM and model comparisons indicated that TB severity was associated with the detection of Mtb micro-diversity within clinical isolates and to the *espR* SNP.

Conclusions: Taken together, these results provide a new insight to better understand TB pathophysiology and could provide new prognosis tool for pulmonary TB severity.

Key words: genetic heterogeneity; genome-wide association study; *Mycobacterium tuberculosis*; pulmonary tuberculosis; structural equation modelling; whole genome sequencing

INTRODUCTION

Tuberculosis (TB) caused by *Mycobacterium tuberculosis* (Mtb) complex remains one of the most prevalent and deadly infectious diseases; there were 10 million new cases worldwide in 2020 which led to 1.5 million deaths (WHO, 2021). Mtb infections result in a wide spectrum of clinical outcomes, from latent asymptomatic infection to pulmonary or extra-pulmonary manifestations of disease, with an array of symptoms. Such diversity has been historically attributed to host and environmental factors, while the Mtb complex was previously considered genetically monomorphic (Gagneux and Small, 2007). Many Mtb virulence factors are well described but, to date, there are no proven genetic determinants associated with virulence, disease progression, or severity of TB (Gagneux, 2018). However, some Mtb lineages and sublineages were associated with more severe TB in animal models and also in human population studies, suggesting that Mtb genetic factors can affect TB clinical presentation and severity (Correa-Macedo et al., 2019; Coscolla, 2017; McHenry et al., 2020). Recent Mtb genomic studies have explored the link between specific Mtb polymorphisms and TB clinical presentation (Grandjean et al., 2020; Sousa et al., 2020). For instance, several compensatory mutations occurring in drug-resistant Mtb clinical isolates were associated with more extensive lung damage (Grandjean et al., 2020), and an association was found between TB severity and mutations affecting the expression of some components of the ESX-1 secretion system, a key player in Mtb virulence (Sousa et al., 2020). In addition, next generation sequencing (NGS)-based studies have revealed micro-diversity in clinical isolates (within hosts, minor variants coexist rather than a clonal colony), and rapid within-host microevolution of Mtb has been suggested by several studies (presence of minor variants within Mtb clinical isolates longitudinally collected upon TB treatment) (Genestet et al., 2021; Ley et al., 2019; Lieberman et al., 2016; Nimmo et al., 2020; O'Neill et al., 2015; Vargas et al., 2021). Some of these variants harbour drug-resistance mutations, whilst other

carry single nucleotide polymorphisms (SNP) in loci involved in modulation of innate immunity and in Mtb cell envelop lipids (Genestet et al., 2021; Ley et al., 2019; Lieberman et al., 2016; Nimmo et al., 2020; O'Neill et al., 2015; Vargas et al., 2021). In other bacterial species responsible for chronic infections, micro-diversity has been suggested to impact the outcome and severity of illness, being involved in pathogen adaptation to immune response and treatment pressure (Ailloud et al., 2019; Azarian et al., 2019; Chaguza et al., 2020; Levade et al., 2017). Accordingly, we hypothesised that genetic features of Mtb clinical isolates, such as specific polymorphisms or micro-diversity, may be linked to TB severity.

METHODS

Mtb samples, data collection, and ethical considerations

In this single-centre retrospective study, 234 patients diagnosed with microbiologically-proven pulmonary TB at the Lyon University Hospital were included. This consisted of 210 TB patients diagnosed from January 2017 to January 2020, 193 with drug-susceptible Mtb, 14 mono-resistant to a first line drug and 3 multidrug resistant (MDR) Mtb. Moreover, this cohort was enriched with all the 24 MDR Mtb cases diagnosed in our centre with pulmonary TB between June 2010 (implementation of the strain biobanking in the lab) and December 2016 to enable assessment of the impact of antibiotic resistance on TB disease severity (**Figure 1**) (Genestet et al., 2019b, 2020b). For all Mtb clinical isolates, whole genome sequencing (WGS) analysis was performed in routine practice as part of the laboratory diagnosis since January 2017, and prior to that only MDR Mtb cases were retrospectively sequenced. Demographic (age, sex, continent of birth), clinical (pulmonary, extra-pulmonary TB, symptoms, clinical findings, comorbidities [previous history of TB, active hepatitis, HIV, diabetes, and immunosuppressive treatment at time of TB diagnosis]), microbiological (sputum smear results, time to positivity, antibiotic resistance, lineage, data from WGS),

nutritional, and immune data were collected. Only variables for which data was available for ≥80% of patients between 2 weeks before TB diagnosis and 1 week after initiation of anti-TB treatment or nutritional supplementation were considered. Outcomes (cured, fatal outcome, and loss to follow-up) were evaluated 2 years after the end of anti-TB treatment.

TB-associated severity indices

TB-associated severity indices were evaluated at the time of diagnosis, before initiation of anti-TB treatment or nutritional supplementation.

The modified Bandim TBscore considers 5 symptoms (cough, haemoptysis, dyspnoea, chest pain, night sweats) and 5 clinical findings (anaemia, tachycardia, positive finding at lung auscultation, fever, body mass index [BMI] <18 and <16); 1 point is attributed for each aspect and final score is the sum of these. Patients were stratified into 2 severity classes, mild (Bandim TBscore ≤4) and moderate/severe (≥5) (Dewi et al., 2020).

The nutritional status of TB patients was also evaluated using the Malnutrition Universal Screening Tool (MUST) that includes 3 variables (unintentional weight loss score [weight loss <5% = 0, weight loss 5-10% = 1, weight loss >10% = 2], BMI [>20.0 = 0, 18.5-20.0 = 1, <18.5 = 2], and anorexia [if yes = 2]) and the final score is the sum of these (Miyata et al., 2013).

Mtb culture

Mtb clinical isolates were processed as previously described (Genestet et al., 2020a). Mtb genomic DNA extractions were performed after a single round of culture. Biobanked Mtb isolates were inoculated in mycobacterial growth indicator tube (MGIT, Becton Dickinson, Sparks, MD) until exponential phase before DNA extraction.

WGS and Illumina data analysis

Genomic DNA of Mtb-positive cultures was purified from cleared lysate and sequenced on NextSeq or MiSeq system (Illumina, San Diego, USA) at the GENEPII sequencing platform of Lyon University Hospital, as previously described (Genestet et al., 2019b). Reads were mapped using the BOWTIE2 to the Mtb H37Rv reference genome (Genbank NC000962.2) and variant calling was conducted using SAMtools mpileup, as previously described (Genestet et al., 2019b). A valid nucleotide variant was called if the position was covered by a depth ≥10 reads and a frequency ≥10%. Regions of genes with repetitive or similar sequences were excluded, i.e. regions of *pe, ppe, pks, pps, esx* gene families The reference genome coverage breadth was ≥93% with a mean depth of coverage of ≥50x. Sequences were submitted to the European Nucleotide Archive (ENA) under accession number PRJEB53047.

Variant assignment and Mtb α-diversity indices.

In a previous study, we showed no significant difference in variant detection and frequencies between sequencing on direct samples and after subculture on media used in routine practice (Genestet et al., 2019a). Moreover, for the present study, 10 isolates were extracted and sequenced twice to evaluate the variability in mutation frequencies between sequencing experiments. In both sequencing experiments, 52 unfixed mutations were detected at similar frequencies (±10%), ranging from 10 to 90% (**Supplementary Figure 1A**). Accordingly, to identify the minimum number of variants in each Mtb clinical isolate, a variant was defined as an assembly of mutations at frequencies of ±10% as illustrated in **Supplementary Figure 1B**. Based on that, the α-diversity index for each isolate was calculated by computing the Rao index of diversity taking into account genetic distance among variants. We computed genetic distance among variants applying Sorensen distance on the presence/absence of 437 mutations and consider this information when computing diversity indices. Following Pavoine et al

(2016) (Pavoine et al., 2016), we rescaled the distances prior to the analysis (dividing by the maximum distance) and use equivalent numbers to allow for comparisons of α-diversities among patients. Note that using equivalent numbers implies that α-diversity is equal to 1 when only one variant is present (no diversity).

Genome-wide association study (GWAS)

Mtb genomes were assembled using SPAdes-3.14.1 with --careful -t 16 --cov-cutoff auto options. DBGWAS 0.5.4 was then run on the 234 pulmonary clinical isolates for which the Bandim TBscore was available. The contigs obtained from the assembly step were used as input, and the Bandim phenotypes (mild grade [Bandim TBscore ≤4] and moderate/severe grade [≥5]), with default options except -nh=3, -SFF=p100, -nb-cores=6, -nc-db=Resistance_DB_for_DBGWAS.fasta-pt-b=uniprot_sprot_bacteria_for_DBGWAS.fasta. The two latter options allowed nucleotide and protein level annotation of the results using databases that are available from the DBGWAS repository (Jaillard et al., 2018).

Phylogenetic analyses

SNP sequence alignment were purged from any non-phylogenetically informative position using goalign (v0.3.5). A phylogenetic tree was computed by maximum likelihood using the GTR model with RAxML-ng (v1.0.3) and the Stakamakis ascertainment correction. Bootstrap was performed to check for phylogenetic robustness using 100 replicates (**Supplementary Figure 2)**. Inference of TB severity profile along phylogenetic trees was performed using pastml (v1.9.34). Trees were visualised using iTOL (v6) (Letunic and Bork, 2016).

Polymorphisms were explored in terminal branches of the phylogeny (fixed mutations) and within the micro-diversity of Mtb samples (unfixed mutations). On the one hand, the precise distribution of Mtb mutations was explored to identify mutational signature typical of

oxidative damage (increased changes C > T and G > A); on the other hand, differential selection pressure analyses at the level of the gene functional categories were conducted by performing a simple count of non-synonymous and synonymous mutations, and by estimating the selective pressure measured as the dN/dS ratio using the Contrast-FEL method (Fixed-Effect site-Level) in the HyPhy package (Kosakovsky Pond et al., 2021). Significant differences between selection pressures acting on the 2 groups at the level of the gene functional categories were tested using re-sampling as described by Coscolla et al. (Coscolla et al., 2021); 30 re-samplings were sufficient to detect significant differences.

Statistical analysis

Univariate analysis

Data were expressed as count (percentage, %) for dichotomous variables and as median (interquartile range [IQR]) for continuous values. The number of missing values was excluded from the denominator. For dichotomous variables, Fisher's exact or $\chi 2$ test was used as appropriate. For continuous values, the non-parametric Mann-Whitney U test or unpaired t-test was used to compare groups as appropriate and according to the Shapiro-Wilk test of normality. Statistical analyses were performed using GraphPad Prism® for Windows version 5.02 (GraphPad Software, La Jolla, CA, USA). *p-value* < 0.05 was considered significant.

Structural equation modelling of severity

To gauge the effect of demographic (age, sex) and clinical (HIV, diabetes, hepatitis, immuno-suppressive treatment, previous history of TB, double location of infection) variables, as well as Mtb genetic features (lineage, antibiotic resistance, occurrence of micro-diversity, occurrence of mutations [SNP identified by GWAS or unfixed mutations in the "regulatory protein" gene functional category]), we performed a latent variable structural equation model

(SEM) (Grace, 2006) linking all of these variables to a latent severity score, assumed to be expressed through 3 markers: the Bandim TBscore, the BMI, and the unintentional weight loss percentage of patients, which are strongly associated with poor prognosis (WHO, 2013). We assumed that the Bandim TBscore, expected to be the best marker of severity, was correlated with the other 2 markers. The model was fitted through maximum likelihood using the R package 'lavaan' (Rosseel, 2012). The importance of explanatory variables was assessed using model comparison based on the corrected Akaike Information Criterion (AICc) (Akaike, 1973; Hurvich and Tsai, 1989); from model-specific AICc values we deduced variable weights using the sum of Akaike weights of all models including the focal variable, based on classic methods (Burnham and Anderson, 2002; Massol et al., 2007). When representing the results of SEM, we give standardised coefficient values to allow for comparison between explanatory variables.

RESULTS

TB cohort characterisation

Among the 234 pulmonary TB patients included in this study, 123 had mild disease and 111 moderate/severe disease according to their Bandim TBscore. The median [IQR] age of the study population was 35 [25-58] years, and a majority were male (66.2%). Most of the patients originated from Europe or Africa, which is consistent with the local epidemiology (Barbier et al., 2018; Genestet et al., 2020b). No difference was found in terms of comorbidities according to severity group (**Table 1**).

As expected, the rate of fatal outcome was more frequent in the moderate/severe-grade group. A poorer nutritional status was also observed in this group, including lower median BMI, greater median unintentional weight loss, higher median malnutrition universal screening tool (MUST) score, as well as lower median serum albumin, sodium, and chloride levels.

Regarding the immune status, the median level of serum C-reactive protein (CRP), CRP to albumin ratio, as well as white blood cell, neutrophil and monocyte counts, and neutrophil to lymphocyte and monocyte to lymphocyte ratios were higher in the moderate/severe grade group. Conversely, the median haemoglobin level, eosinophil, and lymphocyte counts, as well as the lymphocyte to CRP ratio were lower in this group (**Table 1**).

The proportion of smear-positive patients was higher in the moderate/severe-grade group and accordingly the median time to positivity (TTP) of Mtb cultures was lower (**Table 2**). For both groups, Mtb isolates genetic diversity (**Table 2**, **Figure 2** and **Supplementary Figure 2**) reflected the local epidemiology (Genestet et al., 2020b). No difference was observed between groups regarding Mtb resistance profile (**Table 2**). Nevertheless, the proportion of Mtb isolates for which micro-diversity was detected (α-diversity >1) was higher in the moderate/severe-grade group, but without difference in the median magnitude of α-diversity (**Table 2**). Of note, no association was observed between Mtb isolate α-diversity magnitude and smear results (**Supplementary Figure 3A**) or the TTP of Mtb cultures (**Supplementary Figure 3B**).

Mtb genetic characteristics according to TB severity

We explored the distribution of TB severity profile along the Mtb phylogeny of the strains identified in the present study. Both mild and moderate/severe grade severity profiles were found in several sublineages of each lineage, supporting the inference that this feature evolved recurrently along with Mtb evolution (**Figure 2**).

To detect Mtb genetic adaptation according to TB severity, we explored polymorphisms in terminal branches of the phylogeny of Mtb samples and within the micro-diversity of Mtb samples through unfixed mutations, both suggestive of ongoing adaptation.

Previous studies have suggested that severe symptoms are associated to a mutational signature typical of oxidative damage (increased changes C > T and G > A) (Moreno-Molina et al., 2021). Then the precise distribution of Mtb mutations was explored between the mild grade and the moderate/severe grade groups (**Figure 3**). Differences were observed in the distribution of mutation in the terminal branches, *i.e.* fixed mutations of the phylogeny (**Figure 3A and B**), with a slightly stronger ROS mutational signature in the moderate/severe grade group ($p<0.0001$; **Figure 3C**), but not within the micro-diversity of Mtb isolates ($p=0.3132$; **Figure 3D-F**).

We then explored the distribution of these mutations across Mtb gene functional categories. No difference was observed for the distribution of non-synonymous ($p=0.1614$) nor synonymous mutations ($p=0.4815$) across gene functional categories in the terminal branches of the phylogeny (**Figure 4A**). We explored whether some gene functional categories exhibited signs of differential selection pressure in the mild-grade versus moderate/severe-grade group. In the terminal branches of the phylogeny, the gene functional category "virulence, detoxification, adaptation" exhibited both a higher non-synonymous/synonymous mutation ratio ($p=0.045$; **Figure 4C**) and a higher dN/dS for Mtb strains from the mild grade group ($p=6.6x10^{-8}$; **Supplementary Table 1**). Regarding unfixed variants within Mtb clinical isolates, a difference was observed in the distribution of non-synonymous mutations across gene functional categories ($p=0.0238$) but not in the distribution of synonymous mutation ($p=0.9019$, **Figure 4B**). The gene functional category "cell wall and cell processes" exhibited both a higher non-synonymous/synonymous mutation ratio and a higher dN/dS in the mild grade group, and the gene functional category "regulatory proteins" did so in the moderate/severe grade group (**Figure 4D and Supplementary Table 2**).

To characterise the potential Mtb genetic determinants of TB severity, we performed a genome-wide association study (GWAS). GWAS identified a SNP, G4323355C, located in

the promoter of the gene *espR*, a gene coding for a regulatory protein of the Mtb ESX-1 secretion system. This SNP was more frequent among Mtb isolates from the moderate/severe-grade group (15/110, 13.5%) than the mild-grade group (4/123, 3.3%; p=0.0069).

Structural equation model (SEM) of TB severity

We performed a SEM analysis to relate TB severity to various explanatory variables. We focused on host variables independent of the stage of TB disease and Mtb genetic features, including those identified above, for the association with TB severity. Regarding the latter, to exclude a risk of bias, subsequent statistical analyses were conducted on the cohort without enrichment with MDR Mtb strains, which yilded confirmatory results (**Supplementary Table 3**). The TB severity was assessed using the Bandim TBscore, as well as the BMI and proportion (%) of unintentional weight loss. As expected, SEM found that severity markers were correlated with each other. None of the host variables explored had an impact on the TB severity. Among Mtb genetic features, the model showed that only the detection of micro-diversity within Mtb clinical isolates affected TB severity (positive estimated standardized coefficient of 0.52 for α-diversity >1; **Figure 5A**). The importance of all host and Mtb variables evaluated for TB severity was assessed using model comparisons and indicated that the best model was composed of Mtb α-diversity >1 and the presence of the mutation identified by GWAS (**Figure 5B and Supplementary Figure 4**).

DISCUSSION

The present study found that Mtb clinical isolates from patients with mild TB carried mutations in genes associated with host-pathogen interaction, while Mtb isolates from patients with moderate/severe TB carried mutations associated with regulatory mechanisms. Moreover, a GWAS-based approach identified a SNP in the promoter of the *espR* gene coding

for a regulatory protein, which was found to be statistically associated with TB severity. This SNP is located 144 pb upstream the coding sequence, 1 nucleotide downstream from the transcriptional start site of *espR*, suggesting a potential impact of this mutation on the regulation of this gene, as previous studies showed that the 200 bp sequence immediately upstream of the coding sequence is required for the complete expression of *espR* (Cao et al., 2015). EspR level has a direct influence on the expression of the *espACD* operon, coding for the major Mtb virulence determinants the ESX-1 system, and protein expression. Furthermore, EspR is required for ESX-1-dependent ESAT-6 secretion (Anil Kumar et al., 2016; Cao et al., 2015). Accordingly, it would be of interest to further explore the role of this SNP on *espR* expression, and subsequently on the secretion of ESAT-6 and the effect in an *in vitro* model of host-pathogen interaction. It is of note that positive selection on other regulatory proteins, such as PhoR, has been reported (Chiner-Oms et al., 2019), and, taken together, these results point to an overall adaptation to host-pathogen interaction for Mtb strains from patients with moderate/severe disease through regulatory protein involvement.

The finding herein that there was a selection on genes from the "virulence, detoxification, adaptation" functional category, as illustrated by a higher non-synonymous/synonymous ratio is concordant with that reported previously (Tantivitayakul et al., 2020). As was the finding that oxidative damage developed upon severe TB disease may be a driver of Mtb diversity (Moreno-Molina et al., 2021). The integration of the severity status of patients could help to better understand the diverse patterns detected in previous studies exploring the key genetic components involved in sublineage epidemic success (Chiner-Oms et al., 2019; Tantivitayakul et al., 2020). At the same time, in-depth analysis of the impact of Mtb polymorphisms on host-pathogen interaction would help to better understand their involvement in TB pathophysiology.

Furthermore, we applied SEM to identify and evaluate direct and latent interlinkages between, on the one hand, Mtb infection clinical specificities and Mtb isolates' genetic features and, on the other hand, TB disease severity, to pinpoint the positive and negative influences in this regard. Going beyond the classical linear regression analyses, SEM examines the causal relationships among variables, while controlling simultaneously for measurement error. SEM allowed us to determine the degree of correlation (path coefficients) that capture the importance of a certain path of influence from cause to effect, and it was found that the presence of Mtb micro-diversity within clinical isolates led to greater clinical TB severity. This result needs to be confirmed in an independent prospective validation study.

Previous studies showed that mixed infections with genetically distinct Mtb strains is associated with poor treatment outcome of TB (Gan et al., 2016; Mohajeri et al., 2016; Shin et al., 2018), but only few studies explored the association between Mtb micro-diversity and TB outcome or severity. For instance, the study reported by Nimmo et al. found that Mtb diversity did not affect TB outcome (Nimmo et al., 2020); this apparent inconsistency with our observations may result from the different read-outs used (outcome versus severity score at time of diagnosis) and different statistical analysis (logistic regression versus SEM). The results presented herein are, however, in accordance with another study that found that greater TB severity was associated with an increase of within-host Mtb micro-diversity, and particularly so in pre-mortem Mtb isolates (O'Neill et al., 2015). Although we have shown that there is no association between bacterial load and detection of Mtb micro-diversity, it is still unclear whether Mtb micro-diversity is a cause (better Mtb adaptation to treatment, to immune pressure, and/or to various niches) or a consequence (tissue breakdown allowing sampling of Mtb variants usually inaccessible and/or lower immune response reducing selection pressure) of the TB severity.

In addition, no association was found herein between the magnitude of Mtb α-diversity and TB severity. This may be because the analysis was based on the minimum number of variants estimated through whole genome sequencing (WGS) data to calculate Mtb α-diversity, which could lead to underestimate micro-diversity in some Mtb clinical isolate. Nevertheless, detection of unfixed mutations at the level of WGS (meaning mutation frequencies between 10 and 90%) was sufficient to observe a strong association between Mtb micro-diversity detection and TB severity. It is of note that in cancer and microbiological research, calling algorithms for low frequency variants have been developed (Xu, 2018) and may be adapted to Mtb WGS data. WGS of Mtb isolates could therefore be envisioned as an all-in-one solution to detect antibiotic resistance (Genestet et al., 2020a), to infer Mtb transmission chains, to perform epidemiological monitoring (Genestet et al., 2020b, 2019b), but also as a prognosis tool; the latter would be of value to identify those who would most benefit most from additional management measures, such as therapeutic drug monitoring (Alsultan and Peloquin, 2014).

In conclusion, Mtb micro-diversity within a clinical isolate and the mutation in the promoter of the gene espR identified by GWAS are related to disease severity. If further confirmed in a larger independent prospective validation study, this could be a useful to identify early-on those at high risk of severe TB in order to ensure optimal management.

ACKNOWLEDGMENTS

The authors thank Philip Robinson (DRS, Hospices Civils de Lyon, Lyon, France) for help with manuscript preparation, the GENEPII sequencing platform (*Institut des agents infectieux*, Hospices Civils de Lyon, Lyon, France) for the MTBC strains sequencing, and the Institute for Integrative Biology of the Cell (I2BC, Université Paris Saclay, Gif-sur-Yvette, France) for the use of their sharing platform.

DECLARATION OF INTERESTS

We declare no competing interests.

FUNDING STATEMENT

This work was supported by the LABEX ECOFECT (ANR-11-LABX-0048) of the Lyon University, within the *Investissements d'Avenir* program (ANR-11-IDEX-0007) operated by the French national research agency (*Agence nationale de la recherche*).

ETHICAL APPROVAL STATEMENT

All data were collected in a database, in accordance with the decision 20-216 of the ethics committee of the Lyon University Hospital and French legislation in place at the time of the study (Reference methodology MR-004 that covers the processing of personal data for purposes of study, evaluation or research that does not involve the individual). Relevant approval regarding access to patient-identifiable information are granted by the French data protection agency (*Commission Nationale de l'Informatique et des Libertés*, CNIL).

AUTHOR CONTRIBUTIONS

Conceptualization: CG, JLB, SV, OD; Methodology: CG, GR, EH, SD, LJ, FM, SV, OD; Software: GR, RZE, ALM, FH, EW, JLB, SD, LJ, FM; Formal Analysis and Validation: CG, GR, EH, RZE, ALM, FH, AB, SD, LJ, FM, SV, OD; Investigation: CG, GR, EH, RZE, ALM, FH, AE, IV, SD, FA, LJ, FM; Writing – Original Draft: CG, OD; Writing – Review & Editing: All authors; Visualization: CG, GR, EH, RZE, ALM, FH, LJ, FM; Supervision: CG, GL, SV, OD; Funding Acquisition: CG, JLB, GL, SV, OD.

REFERENCES

Ailloud F, Didelot X, Woltemate S, Pfaffinger G, Overmann J, Bader RC, et al. Within-host evolution of *Helicobacter pylori* shaped by niche-specific adaptation, intragastric migrations and selective sweeps. Nat Commun 2019;10:1–13. https://doi.org/10.1038/s41467-019-10050-1.

Akaike H. Information Theory and an Extension of the Maximum Likelihood Principle. In: Parzen E, Tanabe K, Kitagawa G, editors. Sel. Pap. Hirotugu Akaike, New York, NY: Springer New York; 1973, p. 199–213. https://doi.org/10.1007/978-1-4612-1694-0_15.

Alsultan A, Peloquin CA. Therapeutic drug monitoring in the treatment of tuberculosis: an update. Drugs 2014;74:839–54. https://doi.org/10.1007/s40265-014-0222-8.

Anil Kumar V, Goyal R, Bansal R, Singh N, Sevalkar RR, Kumar A, et al. EspR-dependent ESAT-6 Protein Secretion of *Mycobacterium tuberculosis* Requires the Presence of Virulence Regulator PhoP. J Biol Chem 2016;291:19018–30. https://doi.org/10.1074/jbc.M116.746289.

Azarian T, Ridgway JP, Yin Z, David MZ. Long-Term Intrahost Evolution of Methicillin Resistant *Staphylococcus aureus* Among Cystic Fibrosis Patients With Respiratory Carriage. Front Genet 2019;10. https://doi.org/10.3389/fgene.2019.00546.

Barbier M, Dumitrescu O, Pichat C, Carret G, Ronnaux-Baron A-S, Blasquez G, et al. Changing patterns of human migrations shaped the global population structure of *Mycobacterium tuberculosis* in France. Sci Rep 2018;8:5855. https://doi.org/10.1038/s41598-018-24034-6.

Burnham KP, Anderson DR. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. 2nd ed. New York: Springer-Verlag; 2002. https://doi.org/10.1007/b97636.

Cao G, Howard ST, Zhang P, Wang X, Chen X-L, Samten B, et al. EspR, a regulator of the ESX-1 secretion system in *Mycobacterium tuberculosis*, is directly regulated by the two-

component systems MprAB and PhoPR. Microbiol Read Engl 2015;161:477–89. https://doi.org/10.1099/mic.0.000023.

Chaguza C, Senghore M, Bojang E, Gladstone RA, Lo SW, Tientcheu P-E, et al. Within-host microevolution of *Streptococcus pneumoniae* is rapid and adaptive during natural colonisation. Nat Commun 2020;11:3442. https://doi.org/10.1038/s41467-020-17327-w.

Chiner-Oms Á, Berney M, Boinett C, González-Candelas F, Young DB, Gagneux S, et al. Genome-wide mutational biases fuel transcriptional diversity in the *Mycobacterium tuberculosis* complex. Nat Commun 2019;10:1–11. https://doi.org/10.1038/s41467-019-11948-6.

Correa-Macedo W, Cambri G, Schurr E. The Interplay of Human and *Mycobacterium Tuberculosis* Genomic Variability. Front Genet 2019;10:865. https://doi.org/10.3389/fgene.2019.00865.

Coscolla M. Biological and Epidemiological Consequences of MTBC Diversity. Adv Exp Med Biol 2017;1019:95–116. https://doi.org/10.1007/978-3-319-64371-7_5.

Coscolla M, Gagneux S, Menardo F, Loiseau C, Ruiz-Rodriguez P, Borrell S, et al. Phylogenomics of *Mycobacterium africanum* reveals a new lineage and a complex evolutionary history. Microb Genomics 2021;7. https://doi.org/10.1099/mgen.0.000477.

Dewi DNSS, Mertaniasih NM, Soedarsono. Severity of TB classified by modified Bandim TB scoring associates with the specific sequence of *esxA* genes in MDR-TB patients. Afr J Infect Dis 2020;14:8–15. https://doi.org/10.21010/ajid.v14i1.2.

Gagneux S. Ecology and evolution of *Mycobacterium tuberculosis*. Nat Rev Microbiol 2018;16:202–13. https://doi.org/10.1038/nrmicro.2018.8.

Gagneux S, Small PM. Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. Lancet Infect Dis 2007;7:328–37. https://doi.org/10.1016/S1473-3099(07)70108-1.

Gan M, Liu Q, Yang C, Gao Q, Luo T. Deep Whole-Genome Sequencing to Detect Mixed Infection of *Mycobacterium tuberculosis.* PLOS ONE 2016;11:e0159029. https://doi.org/10.1371/journal.pone.0159029.

Genestet C, Hodille E, Barbry A, Berland J-L, Hoffmann J, Westeel E, et al. Rifampicin exposure reveals within-host *Mycobacterium tuberculosis* diversity in patients with delayed culture conversion. PLoS Pathog 2021;17:e1009643. https://doi.org/10.1371/journal.ppat.1009643.

Genestet C, Hodille E, Berland J-L, Ginevra C, Bryant JE, Ader F, et al. Whole-genome sequencing in drug susceptibility testing of *Mycobacterium tuberculosis* in routine practice in Lyon, France. Int J Antimicrob Agents 2020a;55:105912. https://doi.org/10.1016/j.ijantimicag.2020.105912.

Genestet C, Hodille E, Westeel E, Ginevra C, Ader F, Venner S, et al. Subcultured *Mycobacterium tuberculosis* isolates on different growth media are fully representative of bacteria within clinical samples. Tuberc Edinb Scotl 2019a;116:61–6. https://doi.org/10.1016/j.tube.2019.05.001.

Genestet C, Paret R, Pichat C, Berland J-L, Jacomo V, Carret G, et al. Routine survey of *Mycobacterium tuberculosis* isolates reveals nosocomial transmission. Eur Respir J 2020b;55. https://doi.org/10.1183/13993003.01888-2019.

Genestet C, Tatai C, Berland J-L, Claude J-B, Westeel E, Hodille E, et al. Prospective Whole-Genome Sequencing in Tuberculosis Outbreak Investigation, France, 2017–2018. Emerg Infect Dis 2019b;25:589–92. https://doi.org/10.3201/eid2503.181124.

Grace JB. Structural Equation Modeling and Natural Systems. Cambridge: Cambridge University Press; 2006. https://doi.org/10.1017/CBO9780511617799.

Grandjean L, Monteserin J, Gilman R, Pauschardt J, Rokadiya S, Bonilla C, et al. Association between bacterial homoplastic variants and radiological pathology in tuberculosis. Thorax 2020;75:584–91. https://doi.org/10.1136/thoraxjnl-2019-213281.

Hurvich CM, Tsai C-L. Regression and time series model selection in small samples. Biometrika 1989;76:297–307. https://doi.org/10.1093/biomet/76.2.297.

Jaillard M, Lima L, Tournoud M, Mahé P, van Belkum A, Lacroix V, et al. A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. PLoS Genet 2018;14. https://doi.org/10.1371/journal.pgen.1007758.

Kosakovsky Pond SL, Wisotsky SR, Escalante A, Magalis BR, Weaver S. Contrast-FEL—A Test for Differences in Selective Pressures at Individual Sites among Clades and Sets of Branches. Mol Biol Evol 2021;38:1184–98. https://doi.org/10.1093/molbev/msaa263.

Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res 2016;44:W242-245. https://doi.org/10.1093/nar/gkw290.

Levade I, Terrat Y, Leducq J-B, Weil AA, Mayo-Smith LM, Chowdhury F, et al. *Vibrio cholerae* genomic diversity within and between patients. Microb Genomics 2017;3. https://doi.org/10.1099/mgen.0.000142.

Ley SD, de Vos M, Van Rie A, Warren RM. Deciphering Within-Host Microevolution of *Mycobacterium tuberculosis* through Whole-Genome Sequencing: the Phenotypic Impact and Way Forward. Microbiol Mol Biol Rev MMBR 2019;83. https://doi.org/10.1128/MMBR.00062-18.

Lieberman TD, Wilson D, Misra R, Xiong LL, Moodley P, Cohen T, et al. Genomic diversity in autopsy samples reveals within-host dissemination of HIV-associated *Mycobacterium tuberculosis*. Nat Med 2016;22:1470–4. https://doi.org/10.1038/nm.4205.

Massol F, David P, Gerdeaux D, Jarne P. The influence of trophic status and large-scale climatic change on the structure of fish communities in Perialpine lakes. J Anim Ecol 2007;76:538–51. https://doi.org/10.1111/j.1365-2656.2007.01226.x.

McHenry ML, Williams SM, Stein CM. Genetics and evolution of tuberculosis pathogenesis: New perspectives and approaches. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis 2020;81:104204. https://doi.org/10.1016/j.meegid.2020.104204.

Miyata S, Tanaka M, Ihaku D. The prognostic significance of nutritional status using malnutrition universal screening tool in patients with pulmonary tuberculosis. Nutr J 2013;12:42. https://doi.org/10.1186/1475-2891-12-42.

Mohajeri P, Moradi S, Atashi S, Farahani A. Mycobacterium tuberculosis Beijing Genotype in Western Iran: Distribution and Drug Resistance. J Clin Diagn Res JCDR 2016;10:DC05–7. https://doi.org/10.7860/JCDR/2016/20893.8689.

Moreno-Molina M, Shubladze N, Khurtsilava I, Avaliani Z, Bablishvili N, Torres-Puente M, et al. Genomic analyses of *Mycobacterium tuberculosis* from human lung resections reveal a high frequency of polyclonal infections. Nat Commun 2021;12:2716. https://doi.org/10.1038/s41467-021-22705-z.

Nimmo C, Brien K, Millard J, Grant AD, Padayatchi N, Pym AS, et al. Dynamics of within-host *Mycobacterium tuberculosis* diversity and heteroresistance during treatment. EBioMedicine 2020;55:102747. https://doi.org/10.1016/j.ebiom.2020.102747.

O'Neill MB, Mortimer TD, Pepperell CS. Diversity of *Mycobacterium tuberculosis* across Evolutionary Scales. PLOS Pathog 2015;11:e1005257. https://doi.org/10.1371/journal.ppat.1005257.

Pavoine S, Marcon E, Ricotta C. 'Equivalent numbers' for species, phylogenetic or functional diversity in a nested hierarchy of multiple scales. Methods Ecol Evol 2016;7:1152–63. https://doi.org/10.1111/2041-210X.12591.

Rosseel Y. **lavaan** : An *R* Package for Structural Equation Modeling. J Stat Softw 2012;48. https://doi.org/10.18637/jss.v048.i02.

Shin SS, Modongo C, Baik Y, Allender C, Lemmer D, Colman RE, et al. Mixed *Mycobacterium tuberculosis*-Strain Infections Are Associated With Poor Treatment Outcomes Among Patients With Newly Diagnosed Tuberculosis, Independent of Pretreatment Heteroresistance. J Infect Dis 2018;218:1974–82. https://doi.org/10.1093/infdis/jiy480.

Sousa J, Cá B, Maceiras AR, Simões-Costa L, Fonseca KL, Fernandes AI, et al. *Mycobacterium tuberculosis* associated with severe tuberculosis evades cytosolic surveillance systems and modulates IL-1β production. Nat Commun 2020;11:1949. https://doi.org/10.1038/s41467-020-15832-6.

Tantivitayakul P, Ruangchai W, Juthayothin T, Smittipat N, Disratthakit A, Mahasirimongkol S, et al. Homoplastic single nucleotide polymorphisms contributed to phenotypic diversity in *Mycobacterium tuberculosis.* Sci Rep 2020;10:8024. https://doi.org/10.1038/s41598-020-64895-4.

Vargas R, Freschi L, Marin M, Epperson LE, Smith M, Oussenko I, et al. In-host population dynamics of *Mycobacterium tuberculosis* complex during active disease. ELife 2021;10:e61805. https://doi.org/10.7554/eLife.61805.

WHO. World Health Organization, Guideline: Nutritional care and support for patients with Tuberculosis. 2013. http://www.who.int/tb/TBnutrition.pdf (accessed March 29, 2017).

WHO Geneva. World Health Organization, Global tuberculosis report 2021 2021. https://www.who.int/publications-detail-redirect/9789240037021 (accessed November 29, 2021).

Xu C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. Comput Struct Biotechnol J 2018;16:15–24. https://doi.org/10.1016/j.csbj.2018.01.003.

Figure 1 Click here to access/download;Figure;Figure 1.tif

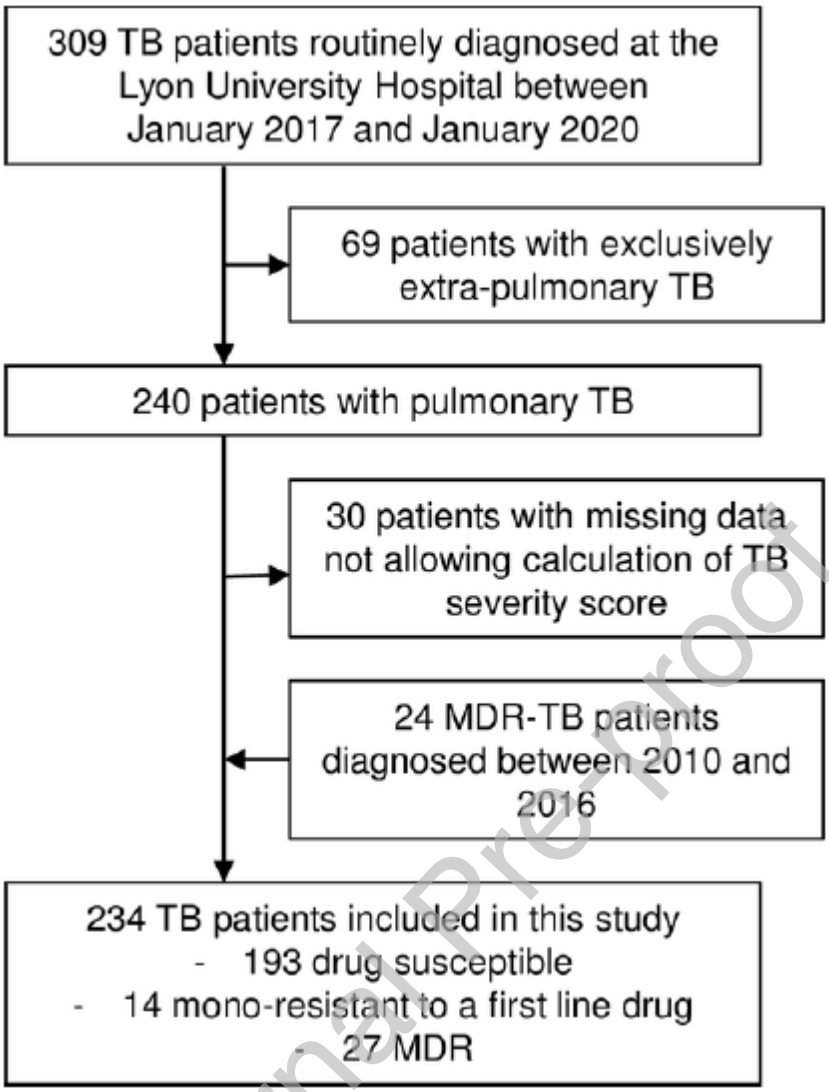

Figure 1: Flowchart of TB patients included in the study

Among the 309 TB patients routinely diagnosed, 69 had exclusively microbiologically-proven extra-pulmonary TB and were excluded as this study focused on pulmonary TB. Thirty other patients were excluded due to missing data not allowing calculation of TB severity score using the Bandim TBscore. Besides, this cohort was enriched with the 24 multidrug resistant (MDR) Mtb cases diagnosed in our centre between 2010 and 2016 to enable assessment of the impact of antibiotic resistance on TB disease severity.

Figure 2

Click here to access/download;Figure;Figure 2.tif

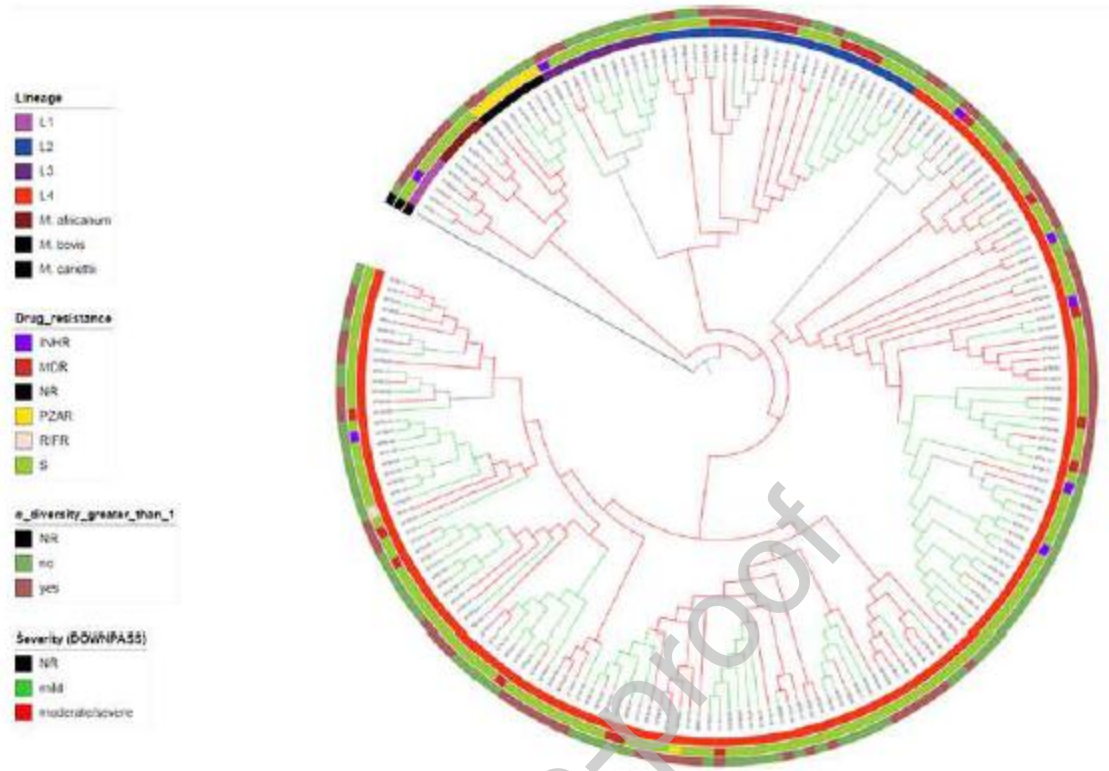

Figure 2: TB severity profile of studied samples and inference of its evolution along the Mtb phylogeny. The phylogeny was reconstructed by Maximum Likelihood that identified all branches with a strong bootstrap support (Supplementary Figure 1). Tree is displayed with arbitrary branch length to improve visibility. Green branches: mild grade TB severity group; red branches: moderate/severe grade TB severity group. From the inside, rings are coloured by Mtb lineages, drug resistance profile, and occurrence of micro-diversity within clinical isolate (see legend). INHR: isoniazid mono-resistant, RIFR: rifampicin mono-resistant; PZAR: pyrazinamide mono-resistant; MDR: multidrug resistant; NR: not reported

Figure 3

Click here to access/download;Figure;Figure 3.tif

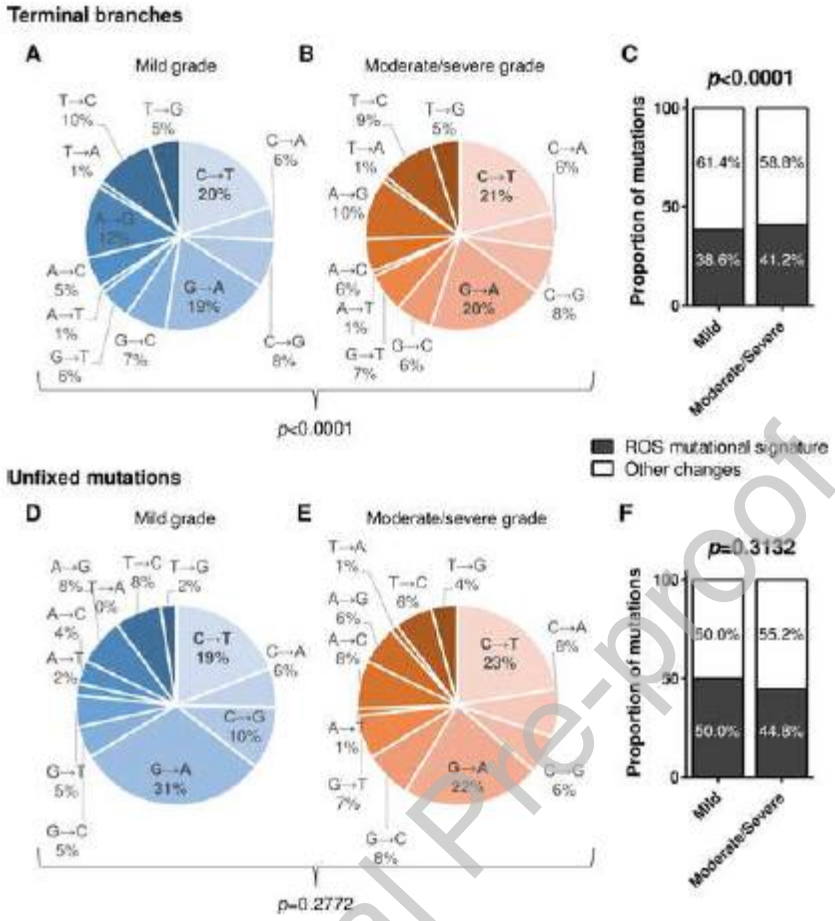

Figure 3: ROS mutational signature according to TB severity

Distribution of mutations in terminal branches of the phylogeny of Mtb samples (A-C; n=21754 mutations explored) and in unfixed mutations within Mtb clinical isolates (D-F; n=437 mutations explored) in the mild-severity grade (A and D) and the moderate/severe-grade groups (B and E). Each type of mutation was explored (A-B and D-E) and a focus was made on ROS mutational signature (C and F). Fisher's exact or $\chi 2$ test was used to compare mild-severity grade and moderate/severe-grade groups, as appropriate.

Figure 4

Click here to access/download;Figure;Figure 4.tif

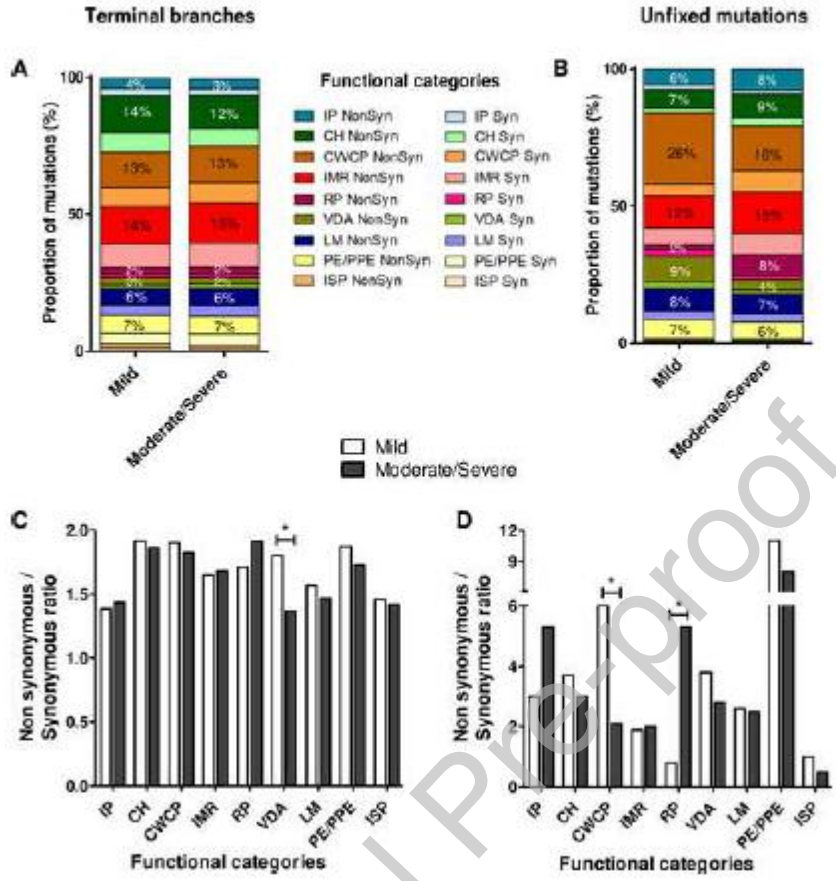

Figure 4: Non-synonymous/synonymous mutations ratio in the gene functional categories according to TB severity.

Distribution of non-synonymous and synonymous mutations (A-B) and non-synonymous/synonymous mutation ratio (C-D) across gene functional categories in the terminal branches of the phylogeny of Mtb samples (A-C; n=21729 mutations explored) and in unfixed mutations within Mtb clinical isolates (B-D; n=437 mutations explored) in the mild-severity grade and the moderate/severe-grade groups. Fisher's exact or $\chi 2$ test was used to compare mild-severity grade and moderate/severe-grade groups, as appropriate. IP: Information pathways; CH: Conserved hypotheticals; CWCP: Cell wall and cell processes;

IMR: Intermediary metabolism and respiration; RP: Regulatory proteins; VDA: Virulence, detoxification, adaptation; LM: Lipid metabolism; ISP: Insertion sequences and phages; NonSyn: Non-synonymous SNP; Syn: Synonymous SNP. *p<0.05

Figure 5 Click here to access/download;Figure;Figure 5.tif

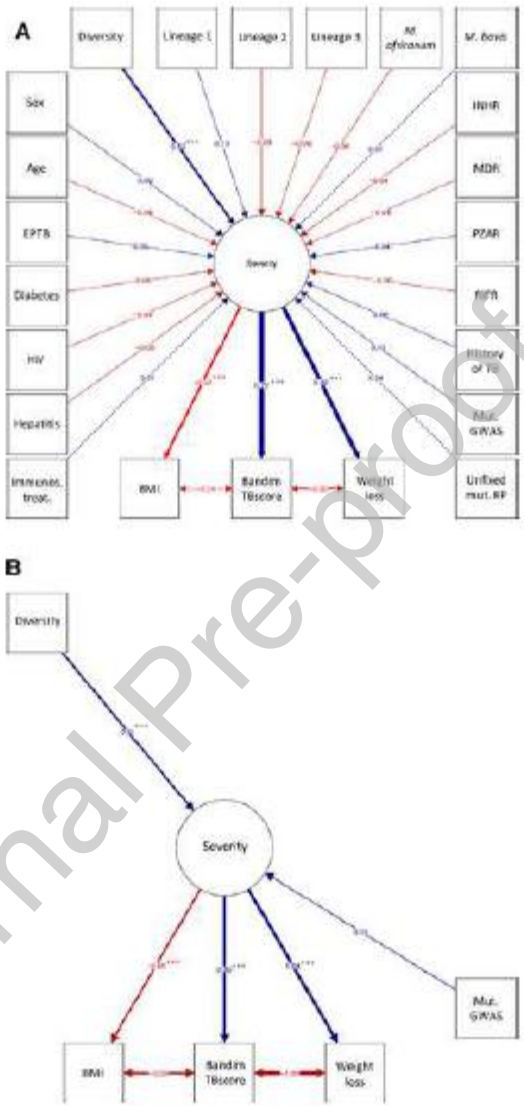

Figure 5: Structural equation modelling (SEM) and the best model examining the effect of host and bacterial factors on pulmonary TB severity.

Regarding the variables associated with TB patients, we focused on variables independent of the stage of TB disease, meaning age, sex, ongoing HIV, hepatitis, diabetes, and/or immunosuppressive treatment (Immunos. treat.), history of TB, and extra-pulmonary manifestation (EPTB, both pulmonary and extra-pulmonary). Concerning Mtb genetic

features, Mtb lineages (lineage 1, lineage 2, lineage 3, *M. africanum* and *M. bovis*, in contrast with lineage 4), resistance profile (isoniazid mono-resistant [INHR], pyrazinamide mono-resistant [PZAR], rifampicin mono-resistant [RIFR], multidrug resistant [MDR], in contrast with susceptible), detection of Mtb micro-diversity within clinical isolates (Diversity), detection of unfixed mutation in the "regulatory protein" gene functional category (Unfixed mut. RP) and the mutation identified by GWAS (Mut. GWAS) were considered. Unidirectional arrows between variables indicate regression and are associated with standardised regression coefficients. Bidirectional dashed arrows among severity markers indicate correlations. Blue edges indicate positive coefficients or correlations, red edges, negative coefficients or correlations. Residual variance terms are omitted for clarity. *p-value* < 0.05 was considered significant. *p<0.05, ** p <0.01, *** p <0.001. (A) Graphical representation of the maximal model (with all variables). (B) Graphical representation of the best model.

Table 1: Patient characteristics

	Total population n=234	**Mild-grade disease n=123**	**Moderate/severe-grade disease n=111**	***p-value***
TB severity score				
Bandim TBscore	4 [3-6]	3 [2-4]	6 [5-7]	<**0.0001**
Demography				
Age (years)	35 [25-58]	38 [26-60]	32 [22-54]	**0.0335**
Sex (male)	155 (66.2%)	85 (69.1%)	70 (63.1%)	0.3366
Geographical origin,				0.1044
Europe	91 (38.9%)	56 (45.5%)	35 (31.5%)	
Africa	116 (49.6%)	53 (43.1%)	63 (56.8%)	
Asia	23 (9.8%)	11 (8.9%)	12 (10.8%)	
America	4 (1.7%)	3 (2.4%)	1 (0.9%)	
Comorbidities				
EPTB	68 (29.1%)	32 (26.0%)	36 (32.4%)	0.3141
HIV	15 (6.4%)	7 (5.9%)	8 (7.2%)	0.7919
Hepatitis	24 (10.3%)	9 (7.4%)	15 (13.5%)	0.1378
Diabetes	26 (11.1%)	18 (15.0%)	8 (7.3%)	0.0941
Immunosuppressive therapy	21 (9.0%)	11 (9.2%)	10 (9.0%)	1.0000
History of TB	18 (7.7%)	9 (7.6%)	9 (8.2%)	1.0000
Outcomes				**0.0021**
Cured	218 (93.2%)	114 (92.7%)	104 (93.7%)	
Fatal outcome	8 (3.4%)	1 (0.8%)	7 (6.3%)	
Unknown	8 (3.4%)	8 (6.5%)	0 (0%)	
Nutritional status				
BMI (kg/m²)	20.2 [17.7-22.4]	21.3 [20.0-23.9]	18.0 [16.2-20.4]	<**0.0001**
Weight loss (%)	8.0 [4.5-12.3]	5.9 [2.9-9.0]	11.4 [6.9-15.4]	<**0.0001**
MUST score	3 [1-5]	1 [0-3]	4 [4-6]	<**0.0001**
Serum albumin (g/L)	30.1 [26.0-36.9]	34.3 [29.2-40.4]	28.8 [23.7-31.8]	<**0.0001**
Total protein (g/L)	75 [70-81]	75 [70-80]	76 [70-81]	0.7830
Sodium (mmol/L)	137 [134-139]	138 [137-139]	135 [133-138]	<**0.0001**
Potassium (mmol/L)	4.0 [3.8-4.3]	4.0 [3.8-4.4]	4.1 [3.8-4.3]	0.3829
Chloride (mmol/L)	103 [100-106]	104 [102-106]	102 [99-104]	<**0.0001**
Immune status				
CRP (mg/L)	47 [13-98]	21 [6-78]	72 [23-110]	<**0.0001**
CRP to Albumin ratio	16.6 [4.0-38.7]	6.6 [1.7-28.9]	27.4 [9.3-45.1]	<**0.0001**
Haemoglobin (g/L)	120 [104-137]	130 [116-141]	111 [97-128]	<**0.0001**
White blood cells (G/L)	6.9 [5.3-9.0]	6.5 [4.9-8.5]	7.5 [6.0-9.8]	**0.0071**
Neutrophils (G/L)	4.6 [3.2-6.2]	4.1 [2.8-5.5]	5.1 [3.8-6.6]	**0.0006**
Eosinophils (G/L)	0.09 [0.03-0.18]	0.12 [0.06-0.22]	0.075 [0.02-0.14]	**0.0003**
Basophils (G/L)	0.03 [0.02-0.05]	0.03 [0.02-0.05]	0.03 [0.01-0.05]	0.3824
Monocytes (G/L)	0.62 [0.46-0.84]	0.58 [0.41-0.76]	0.70 [0.49-1.03]	**0.0061**
Lymphocytes (G/L)	1.41 [0.89-2.01]	1.54 [0.95-2.14]	1.23 [0.87-1.63]	**0.0057**
Neutrophil to lymphocyte ratio	3.5 [1.9-5.5]	2.9 [1.5-4.5]	4.5 [2.7-6.8]	<**0.0001**
Monocyte to Lymphocyte ratio	0.48 [0.29-0.71]	0.40 [0.26-0.59]	0.58 [0.40-0.83]	<**0.0001**
Lymphocyte to CRP ratio	31.2 [11.5-138.3]	76.9 [18.3-286.4]	17.9 [9.3-48.1]	<**0.0001**

Data were expressed as count (%) for dichotomous variables and as median [interquartile range] for continuous values. The number of missing values was excluded from the denominator. For dichotomous variables, Fisher's exact or $\chi 2$ test was used as appropriate.

For continuous values, non-parametric Mann-Whitney U test or unpaired t-test was used to compare groups as appropriate according to Shapiro-Wilk normality test. *p-value* < 0.05 was considered significant. EPTB: extrapulmonary tuberculosis; unknown outcome: loss to follow-up or follow-up in another care facility; HIV: human immunodeficiency virus; MUST: malnutrition universal screening tool; BMI: body mass index; CRP: C-reactive protein.

Table 2: Microbiological characteristics of Mtb isolates

	Total population n=234	**Mild-grade disease n=123**	**Moderate/severe-grade disease n=111**	***p-values***
Type of sample				0.0696
Bronchial aspiration	63 (27%)	41 (33%)	22 (20%)	
Biopsy	10 (4%)	6 (5%)	4 (4%)	
Sputum	137 (59%)	61 (50%)	76 (68%)	
BAL	16 (7%)	10 (8%)	6 (5%)	
Stomach tube	8 (3%)	5 (4%)	3 (3%)	
Smear-positive isolates	111 (47%)	48 (39%)	63 (57%)	**0.0087**
TTP (days)	10.5 [6-17]	12.5 [7-18]	8 [5-15]	**0.0022**
Lineages				0.2816
L1	8 (3%)	3 (2%)	5 (4%)	
L2	32 (14%)	18 (15%)	14 (13%)	
L3	13 (6%)	10 (8%)	3 (3%)	
L4	169 (72%)	87 (71%)	82 (74%)	
M. africanum	5 (2%)	1 (1%)	4 (4%)	
M. bovis	7 (3%)	4 (3%)	3 (3%)	
Resistance profile				0.2898
Susceptible	193 (82%)	97 (79%)	96 (87%)	
INH monoR	11 (5%)	6 (5%)	5 (5%)	
RIF monoR	2 (1%)	1 (1%)	1 (1%)	
PZA monoR	1 (0.4%)	0 (0%)	1 (1%)	
MDR	27 (11%)	19 (15%)	8 (7%)	
α-diversity >1	123 (53%)	38 (31%)	85 (77%)	**<0.0001**
Magnitude of α-diversity >1	1.89 [1.51-2.19]	1.92 [1.55-2.33]	1.87 [1.49-2.05]	0.4495

Data were expressed as count (percentage, %) for dichotomous variables and as median [interquartile range] for continuous values. For dichotomous variables, Fisher's exact or $\chi 2$ test was used as appropriate. For continuous values, non-parametric Mann-Whitney U test was used to compare groups. *p-value* < 0.05 was considered significant. BAL:

bronchoalveolar lavage; TTP: time to positivity of Mtb culture. INH: isoniazid, RIF: rifampicin; PZA: pyrazinamide; monoR: mono-resistant; PZA monoR: Mtb isolates resistant to PZA excluding *M. bovis*; MDR: multi-drug resistant.