Within-host genetic micro-diversity of *Mycobacterium tuberculosis* and the link with tuberculosis disease features

Charlotte Genestet[1,2*], Elisabeth Hodille[1,2], François Massol[3,4], Guislaine Refrégier[5,7], Alexia Barbry[1,2], Emilie Westeel[1,6], Gérard Lina[1,2,8], Florence Ader[1,9], Laurent Jacob[10], Stéphane Dray[10], Jean-Luc Berland[1,6], Samuel Venner[10], Oana Dumitrescu[1,2,8] 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, 69007, France

2. Hospices Civils de Lyon, Institut des Agents Infectieux, Laboratoire de bactériologie, Lyon, 69317, cedex 04, France

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

4. CNRS, CHU Lille, Institut Pasteur de Lille, U1019-UMR 8204-CIIL-Center for Infection and Immunity of Lille, University of Lille, Lille, France.

5. Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris Sud, Université Paris-Saclay, Gif-sur-Yvette, France.

6. Fondation Mérieux, Emerging Pathogens Laboratory, Lyon, 69007, France.

7. Ecologie Systématique Evolution CNRS Université Paris-Saclay Orsay France.

8. Université Lyon 1, Facultés de Médecine et de Pharmacie de Lyon, Lyon, 69008, France

9. Hospices Civils de Lyon, Service des Maladies infectieuses et tropicales, Lyon, 69317, cedex 04, France.

10. Biometry and Evolutionary Biology laboratory (LBBE), CNRS UMR 5558, Université Lyon 1, Villeurbanne, 69622, France

***Correspondence:** 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

Tuberculosis (TB), caused by *Mycobacterium tuberculosis* (Mtb) complex, is still the number one deadly contagious disease. Mtb infection results in a wide spectrum of clinical presentations and severity symptoms, but without proven Mtb genetic determinants. Thanks to a collection of 355 clinical isolates with associated patient's clinical data, we showed that Mtb micro-diversity within patient isolates is strongly correlated with TB-associated severity scores. Interestingly, this diversity is driven by a selection pressure to adapt to different lifestyles related to the infection site. Taken together, these results provide a new insight to better understand TB pathophysiology. Furthermore, Mtb micro-diversity could be envisioned as a new prognostic tool to improve the management of TB patients.

Keywords: *Mycobacterium tuberculosis*; micro-diversity; tuberculosis severity; pulmonary tuberculosis; extra-pulmonary tuberculosis; whole genome sequencing; selection pressure; compartmentalization

INTRODUCTION

Tuberculosis (TB) caused by *Mycobacterium tuberculosis* (Mtb) complex remains one of the most prevalent and deadly infectious diseases, responsible for 10 million new cases and 1.2 million deaths among HIV-negative people worldwide in 2018, and an additional 208 000 deaths among HIV-positive people [1]. 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 severity symptoms. Such diversity has been historically attributed to host and environmental factors, while the Mtb complex was previously considered genetically monomorphic [2].

Since the introduction of the next generation sequencing (NGS) enabling whole genome sequencing (WGS), outstanding progress has been done in the field of Mtb genomics. Increasing studies based on NGS have revealed micro-diversity in Mtb clinical isolates: within hosts, minor variants coexist rather than a clonal colony [3–14]. Although we showed in a previous study the involvement of initial Mtb micro-diversity in intra-macrophagic persistence and antibiotic tolerance [15], the role and the impact of Mtb micro-diversity is still poorly understood. It was already showed that mixed-strains Mtb infection is associated with poor outcome [16,17], but only few reports investigated the link between Mtb micro-diversity and TB severity and outcome [9,10]. Yet, several major questions remain unanswered: such as the Mtb ability to cause active disease and various symptoms. While many Mtb virulence factors are well described, to date, there are no proven genetic determinants associated with virulence, disease progression or severity of TB [18].

Regarding other bacterial species responsible of chronic infections, such as *Helicobacter pylori*, *Pseudomonas aeruginosa*, *Staphylococcus aureus*, *Streptococcus pneumoniae*, several reports highlighted the role of micro-diversity in pathogen adaptation between and within-host. Bacterial micro-diversity plays a role in the adaptation to immune and treatment pressure, for

infecting different body sites and has been suggested to impact outcome and severity of illness [19–25].

Accordingly, here we explored the relation between Mtb micro-diversity of 355 clinical isolates, from 311 patients diagnosed at the Lyon University Hospital, and TB clinical presentation, as well as nutritional and immune status of patients. It revealed a strong correlation between the detection of Mtb micro-diversity within clinical isolates and TB-associated severity markers. Furthermore, thanks to a cohort of 42 patients with both microbiologically proven pulmonary and extra-pulmonary TB, we investigated intra-host micro-evolution of Mtb. We observed a compartmentalization of variants, driven by a selection pressure to adapt to different tissues, as shown by dN/dS approach. It should be noted that, besides the detection of canonical drug resistance determinants, the detection of Mtb micro-diversity within patient's isolates is, to date, the only other bacterial feature which could be envisioned as a prognostic marker of poor TB outcome.

RESULTS

Baseline characteristics of study populations

A total of 355 clinical Mtb isolates from 311 patients were included in this study, including 42 patients with both microbiologically proven pulmonary and extra-pulmonary TB (PTB and EPTB) and 2 patients with 2 extra-pulmonary localization without PTB. Moreover, 25 patients with PTB also developed EPTB according to the clinical manifestations and 2 patients with EPTB may also have PTB but no pulmonary sampling was conducted. The **Table 1** summarize patients' demographic and descriptive characteristics. As it is a retrospective study, all data were not available for all patients, then the population size was always specified.

Patient's descriptive characteristics revealed different profiles between patients with PTB or EPTB. Overall, higher severity scores and lower malnutrition indices were observed for patient

with PTB compared to patients with EPTB. Conversely, a lower immune status was observed for patient with EPTB compared to patients with PTB. It should be noted that serum lipid and serum iron profiles and lymphocyte typing can not be properly compared between the cohorts as the data were available only for 20% to 43% of patients.

Table 1: Patients' demographic and baseline descriptive characteristics

Characteristics	**Pulmonary TB n=244**	**Extra-pulmonary TB n=111**	***p-values***
Demographic data			
Age (years), median [IQR]	36 [25-58]	37 [24-60]	0.7505
Gender (male), n (%)	162 (66.4)	57 (51.4)	**0.0094**
TB outcome			
Fatal outcome, n (%)	8 (3.3)	3 (2.7)	0.6999
Unknown, n (%)	18 (7.4)	11 (9.9)	0.4109
Comorbidity			
HIV, n (%)	15/233 (6.4)	8/107 (7.5)	0.8165
Hepatitis, n (%)	24/189 (12.7)	13/85 (15.3)	0.5701
Diabetes, n (%)	26/234 (11.1)	14/105 (13.3)	0.5867
Immunosuppressive therapy, n (%)	21/235 (8.9)	16/107 (15.0)	0.1317
History of TB, n (%)	19/233 (8.2)	4/106 (3.8)	0.1662
TB-associated severity scores			
Bandim TB score, median [IQR] (n)	4 [3-6] (233)	3 [1.5-5] (105)	**0.0002**
MUST, median [IQR] (n)	3 [1-5] (227)	2 [0-4] (105)	**0.0293**
Nutritional risk score, median [IQR] (n)	1 [0.25-3] (56)	1 [0-3] (24)	0.8839
Nutritional status			
BMI (kg/m²), median [IQR] (n)	20.2 [17.7-22.4] (228)	21.63 [19.0-26.2] (107)	**<0.0001**
Weightloss, median [IQR] (n)	8.0 [4.5-12.3] (227)	6.4 [2.4-11.5] (106)	**0.0293**
Serum albumin (g/L), median [IQR] (n)	30.0 [26.0-36.7] (209)	31.1 [25.9-36.2] (92)	0.8693
Serum pre-albumin (g/L), median [IQR] (n)	0.13 [0.09-0.18] (152)	0.11 [0.08-0.17] (60)	0.3448
Serum/blood parameters			
Total protein (g/L), median [IQR] (n)	75 [70-81] (235)	75 [70-81] (102)	0.9509
Blood glucose (mmol/L), median [IQR] (n)	5.0 [4.5-6.0] (215)	5.2 [4.6-6.4] (100)	0.3719
Calcium (mmol/L), median [IQR] (n)	2.44 [2.31-2.61] (236)	2.43 [2.32-2.59] (92)	0.7641
Phosphorus (mmol/L), median [IQR] (n)	0.98 [0.85-1.18] (118)	1.07 [0.84-1.21] (57)	0.6397
Magnesium (mmol/L), median [IQR] (n)	0.81 [0.74-0.88] (70)	0.80 [0.71-0.87] (42)	0.2684
Sodium (mmol/L), median [IQR] (n)	137 [134-139] (236)	137 [134-139] (62)	0.1480
Potassium (mmol/L), median [IQR] (n)	4.0 [3.8-4.3] (236)	4.0 [3.7-4.3] (104)	0.3781
Chloride (mmol/L), median [IQR] (n)	103 [100-106] (235)	103 [98-106] (103)	0.2117
Bicarbonate (mmol/L), median [IQR] (n)	25 [23-27] (178)	25 [23-27] (80)	0.8603
Anion gap, median [IQR] (n)	13 [11-14] (165)	13 [11-15] (78)	0.3953
Serum lipid profile			
Total cholesterol (mmol/L), median [IQR] (n)	4.1 [3.3-5.1] (56)	4.2 [3.1-4.9] (24)	1.0000
HDL-C (mmol/L), median [IQR] (n)	1.03 [0.69-1.29] (52)	1.00 [0.54-1.19] (23)	0.7302
LDL-C (mmol/L), median [IQR] (n)	2.58 [1.88-3.33] (50)	2.58 [1.99-3.25] (23)	0.8960
Triglyceride (mmol/L), median [IQR] (n)	1.14 [0.93-1.67] (59)	1.24 [0.99-1.69] (28)	0.4898
Iron balance and inflammatory markers			

Serum iron (µmol/L), median [IQR] (n)	5.3 [3.0-9.2] (73)	6.7 [4.5-10.6] (39)	0.1555
Transferrin saturation (%), median [IQR] (n)	13.1 [8.2-20.1] (73)	13.4 [10.1-20.3] (39)	0.2921
Ferritin (µg/L), median [IQR] (n)	195 [79-319] (98)	281 [104-759] (48)	**0.0336**
CRP (mg/L), median [IQR] (n)	48 [13-100] (228)	45 [14-93] (99)	0.8781
Hemogram			
Hemoglobin (g/L), median [IQR] (n)	120 [104-136] (236)	115 [102-130] (106)	0.1644
WBC (G/L), median [IQR] (n)	6.9 [5.3-9.0] (237)	6.3 [4.6-8.1] (105)	**0.0294**
Neutrophils (G/L), median [IQR] (n)	4.6 [3.2-6.2] (237)	4.4 [2.8-5.5] (105)	0.1218
Eosinophils (G/L), median [IQR] (n)	0.09 [0.03-0.18] (236)	0.075 [0.01-0.148] (104)	**0.0446**
Basophils (G/L), median [IQR] (n)	0.03 [0.02-0.05] (235)	0.02 [0.01-0.04] (104)	0.0591
Monocytes (G/L), median [IQR] (n)	0.62 [0.46-0.84] (237)	0.55 [0.38-0.75] (105)	**0.0179**
Lymphocytes (G/L), median [IQR] (n)	1.41 [0.89-2.00] (237)	1.18 [0.75-1.68] (105)	**0.0294**
Lymphocyte typing			
CD4 T cells (cell/mm^3), median [IQR] (n)	338 [180-613] (81)	362 [198-555] (45)	0.8406
CD8 T cells(cell/mm^3), median [IQR] (n)	274 [168-468] (79)	240 [151-384] (44)	0.1581
CD4/CD8 ratio, median [IQR] (n)	1.43 [0.82-2.07] (79)	1.77 [0.84-2.5] (44)	0.2249

Data were expressed as count (percentage, %) for dichotomous variables and as medians (interquartile range [IQR]) for continuous values. The number of missing values was excluded from the denominator. Fisher exact, $\chi 2$ test or non-parametric Mann-Whitney U test was used to compare groups where appropriate. *p-value* < 0.05 was considered significant. PTB: pulmonary tuberculosis; EPTB: extrapulmonary tuberculosis; unkown outcome: loss of follow-up or follow-up in another care facility; HIV: human immunodeficiency virus; MUST: malnutrition universal screening tool; BMI: body mass index; HDL-C: high-density lipoprotein cholesterol; LDL-C: low-density lipoprotein cholesterol; CRP: C-reactive protein; WBC: white blood cell.

Mtb micro-diversity is not correlated with Mtb samples characteristics.

WGS data of Mtb clinical isolates included in this study were analyzed to detect Mtb genetic micro-diversity through unfixed mutations (at frequencies between 10 and 90%). These data were used to identify the minimum number of variants in each Mtb clinical isolates. A variant was defined as an assembly of mutations with similar frequencies (±10%). This allowed to calculate the α-diversity of each Mtb isolates, which include both the number and the frequency of the variants and the genetic distance between these variants.

First of all, correlation between Mtb micro-diversity and Mtb samples characteristics, such as smear results and time to positivity of Mtb culture samples, reflecting bacterial load in clinical isolates, type of extra-pulmonary or pulmonary samples, Mtb lineages and resistance status, were explored (**Fig. S1**). Risks of bias, such as delta between Mtb sampling or between treatment initiation and sampling (only concerning the patients with both microbiologically proven PTB and EPTB) and samples that underwent freeze / thaw cycle (biobank samples) compared to those analyzed after a single round of culture (routine practice), were also evaluated (**Fig. S2**). No significant correlation was observed, except a higher proportion of isolates with captured micro-diversity in biobank pulmonary samples compared to pulmonary samples analyzed in routine practice. It may be due to the selection of significant Mtb isolates from our biobank, such as MDR strains, highly transmissible Mtb strains and strains from patients with both PTB and EPTB, all factors usually associated with high TB severity. Interestingly, no significant correlation was established between the bacterial load and Mtb micro-diversity in clinical isolates.

Mtb micro-diversity is correlated with TB-associated severity indices.

We aimed to investigate the association between Mtb α-diversity and TB outcome. However, due to the very low rate of fatal outcome in our setting (8/311; 2.6%), no significant correlation could be established (**Fig. S3**). Therefore, the correlation between Mtb micro-diversity and 3 TB severity indices was explored: i) the modified Bandim score (the most largely used PTB prognosis score), which consider 5 symptoms (cough, hemoptysis, dyspnea, chest pain, night sweats) and 5 clinical findings (anemia, tachycardia, positive finding at lung auscultation, fever, BMI) [26,27]; ii) the Malnutrition Universal Screening Tool (MUST), based on weight loss, BMI and anorexia to evaluate malnutrition status of TB patients [28]; iii) the nutritional risk score

144 considering BMI, hypoalbuminemia, hypocholesterolemia and severe lymphocytopenia, then
145 including both nutritional and immune factors [29].
146 A significant correlation was observed between the detection of micro-diversity (α-diversity >
147 1) in pulmonary Mtb isolates and high Bandim score (**Fig. 1A**, $p<0.0001$), MUST (**Fig. 1B**,
148 $p<0.0001$) and nutritional risk score (**Fig. 1C**, $p<0.0001$). However, no correlation was
149 observed between the ranges of α-diversity in Mtb isolates and the TB-associated severity
150 indices evaluated.
151 As expected, no correlation was observed between the detection of micro-diversity in extra-
152 pulmonary Mtb isolates and the Bandim score (**Fig. 1D**), as this score is mainly based on PTB
153 symptoms, nor with the MUST (**Fig. 1E**), as malnutrition is more characteristic of PTB (**Table**
154 **1**) [30]. Conversely, a correlation between the detection of micro-diversity in extra-pulmonary
155 Mtb isolates and high nutritional risk score (**Fig. 1F**, $p=0.0131$) was observed, this severity
156 index also including variables of the patient's immune status. As before, no correlation was
157 observed between the ranges of α-diversity in Mtb isolates and the TB severity indices
158 evaluated.
159
160 ***Independent variables associated with micro-diversity in Mtb pulmonary and extra-***
161 ***pulmonary isolates.***
162 The analysis of nutritional indices of TB patients revealed that the detection of micro-diversity
163 in pulmonary Mtb isolates was correlated with low BMI, severe unintentional weight loss, low
164 serum pre-albumin, hyponatremia, hypochloremia, low serum bicarbonate (**Fig. S4**), and also
165 with low total cholesterol, low HDL and high triglyceride (**Fig. S5**), which are all associated
166 with unfavorable TB outcome [29–34]. The detection of micro-diversity in extra-pulmonary Mtb
167 isolates was correlated with severe unintentional weight loss, low serum pre-albumin and low
168 total serum protein (**Figs. S5 and S6**). Furthermore, a correlation was observed between the

detection of micro-diversity in both Mtb pulmonary and extra-pulmonary isolates and iron deficiency, meaning low serum iron and low transferrin saturation (**Fig. 2**).

We also explored the correlation between micro-diversity in Mtb isolates and inflammation (serum CRP and ferritin) and immune markers (hemoglobin, blood white blood cell (WBC), neutrophil, eosinophil, basophil, monocyte, lymphocyte, CD4 T cells and CD8 T cells count in peripheral blood and CD4/CD8 ratio). No correlation was observed between these markers and the detection of micro-diversity in pulmonary Mtb isolates (**Fig. S7**). Conversely, we found a correlation between the detection of micro-diversity in extra-pulmonary Mtb isolates and high serum ferritin and low CD4 T cells count in peripheral blood (**Fig. S8**), which is a well-known risk factor for TB patients, especially for extra-pulmonary TB [35,36]. Finally, in all cases (nutritional and immune variables), except for phosphorus for pulmonary Mtb isolates and eosinophils count for extra-pulmonary Mtb isolates, no correlation was observed between the ranges of Mtb α-diversity and the host variables studied.

Models describing the relationship between Mtb micro-diversity and clinical parameters

To go forward, we built a model to relate Mtb micro-diversity in pulmonary (P) and extra pulmonary (EP) locations to various explanatory variables obtained for most patients in the cohort. The response variable we chose to model is a binary (0/1) measure of whether Mtb micro-diversity was detected in clinical isolates.

In the case of pulmonary TB infections, model comparisons indicated a best model with only 9 variables (double location of infection, Mtb lineage, weight loss, protein dosage, CRP, hemoglobin, leukocyte count, neutrophil count and the BANDIM score). 39 models had an Akaike weight less than two units from the best model. Only eight variables had likely effects (BANDIM score, Weight loss, Double Location, Hemoglobin, Neutrophil count, CRP, Protein, and Leukocyte count), and only the effect of Mtb lineage had importance between 0.5 and 0.73,

making it a plausible (but not likely) effect (**Fig. 3**). Multi-model inference of parameter values ascertained that Mtb micro-diversity in pulmonary isolates was higher with increasing weight loss, BANDIM, hemoglobin, and leukocyte counts, and lower with increasing neutrophil counts, CRP, and protein (**Table 2**). All lineages had slightly different coefficients, but they never differed from the baseline sufficiently for 95% confidence intervals to have the same sign (**Table 2**) and asymptotic pairwise tests between lineage coefficients were all non-significant in the best model (results not shown).

	Estimate	**lower CI**	**higher CI**	**Uncond. variance**	**Nb models**	**Importance**
(Intercept)	2.36E-01	-1.46E+01	1.50E+01	6.71E+01	2000	1.00
WeightLoss	2.73E-01	1.58E-01	3.89E-01	3.49E-03	2000	1.00
BandimScore	5.85E-01	3.15E-01	8.55E-01	1.88E-02	2000	1.00
DoubleLoc1	1.26E+00	2.03E-01	2.31E+00	2.93E-01	1962	0.99
Haemoglobin	2.84E-02	4.43E-03	5.24E-02	1.52E-04	1957	0.99
Neutrophiles	-6.36E-04	-1.52E-03	2.51E-04	2.17E-07	1827	0.93
CRP	-8.97E-03	-1.89E-02	1.00E-03	2.70E-05	1772	0.92
Protein	-4.81E-02	-1.08E-01	1.20E-02	1.00E-03	1654	0.87
Leucocytes	4.06E-04	-3.72E-04	1.18E-03	1.67E-07	1384	0.74
Lineage2	-6.54E-01	-2.86E+00	1.55E+00	1.37E+00	1224	0.65
Lineage3	-2.11E+00	-6.16E+00	1.95E+00	4.26E+00	1224	0.65
Lineage4	-6.93E-01	-2.82E+00	1.44E+00	1.26E+00	1224	0.65
Lineage6	-1.60E-01	-2.52E+00	2.20E+00	2.01E+00	1224	0.65
LineageM. bovis	-3.13E+00	-8.59E+00	2.32E+00	7.82E+00	1224	0.65

Table 2: Models of Mtb micro-diversity in pulmonary isolates. Multi-model estimates of variable coefficients based on the 2000 models pooled in the consensus set of models; only models with importance greater than 0.5 were kept in this table. “DoubleLoc1”: presence of both P and EP locations of TB infection in the patient, based on microbiological or pathophysiological evidence; “Lineage2”: effect of lineage 2 when compared to baseline (lineage 1). “Nb models”: number of models in the consensus set (already the result of merging four parallel model comparison processes) which incorporated the variable.

Mtb micro-diversity within and between pulmonary and extra-pulmonary compartments.

To better understand the role and the impact of Mtb micro-diversity, we focused on paired isolates from the training cohort to explore micro-evolution of Mtb within individuals (**Fig. 4**). First of all, the frequency in both pulmonary and extra-pulmonary compartments of each variant identified was explored. Among the 104 variants identified by WGS, 31/104 (30%) were specific of pulmonary isolates, 30/104 (29%) were specific of extra-pulmonary isolates, the frequencies of 22/104 (21%) variants were significantly increased or decreased (≥10%) between the compartments and only 21/104 (20%) were found at similar frequencies between the compartments (**Fig. 4A**). These results suggest a compartmentalization of Mtb variants.

Then we analyzed of the repartition on Mtb genome of the 168 pairwise mutations distances observed between paired pulmonary and extra-pulmonary isolates, among these 104 variants. However, it did not reveal any hot spot on Mtb genome (**Fig. 4B**). Among these 168 pairwise mutations, 22/168 (13%) were intergenic mutations, and 146/168 (87%) were located in coding regions, among whom 104/146 (71%) were nonsynonymous mutations and 42/146 (29%) were synonymous SNP.

These 104 nonsynonymous mutations were used to perform a principal component analysis (PCA) for mixed data (**Fig. 4C**). This analysis took into account the delta of frequency of each pairwise mutation between pulmonary and extra-pulmonary compartment and the functional categories of these mutations. The first principal component (15.65% of the variance) allowed a discrimination of pulmonary and extra-pulmonary mutations and of some functional categories. This revealed an association of pulmonary mutations with cell wall and cell processes and lipid metabolism categories. Otherwise, extra-pulmonary mutations were strongly associated with intermediary metabolism and respiration category. This suggests, on the one hand, an adaptation to pulmonary macrophage infection for pulmonary variants and on the other hand, a metabolic adaptation to extra-pulmonary tissues.

To go forward, we used PAML branch-site models to test for selection pressures associated with infection localization. No gene function exhibited a significant signal of positive selection pressure, neither in pulmonary isolates nor in extra-pulmonary ones with our cohort. Still, relatively low p-values (0.1<p-value <0.15) combined with amino acids with significant BEB posterior P-values were observed for "Virulence, detoxification and adaptation" gene category in pulmonary Mtb isolates and for "intermediary metabolism and respiration" gene category in the extra-pulmonary ones. These tendencies require larger samples to be confirmed.

DISCUSSION

The objective of the present study was to address the impact of Mtb micro-diversity on TB pathophysiology. For this purpose, we explored the correlation between Mtb micro-diversity and TB clinical presentation, meaning pulmonary and extra-pulmonary tuberculosis and severity of illness. A strong correlation was observed between the detection of Mtb micro-diversity and TB-associated severity markers, in both pulmonary and extra-pulmonary clinical isolates. Furthermore, this diversity seems to be driven by a selection pressure to adapt to different tissues.

A previous report explored the correlation between baseline Mtb diversity (initiation or major change to treatment) and 6-month TB outcome and found that Mtb diversity did not affect TB outcome. However, beside fatal outcome, treatment failure and acquisition of multi-drug resistance, authors also included loss of follow-up in unfavorable outcome [9]. Nevertheless, the latter category, representing 33% of poor outcome in the study, is not itself an unfavorable outcome, therefore we have classified it in the "unknown" category, which could explain the discrepancies observed between the studies. Otherwise, our results are in accordance with another study, focusing on the evolution of Mtb micro-diversity from five patients during the course of TB treatment, which showed that higher TB severity is associated with an increase of

Mtb micro-diversity within-host, and more particularly in pre-mortem Mtb isolates of two of these patients, and without significant impact of TB treatment on Mtb micro-diversity [10]. Yet 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. The mechanisms driving such diversity remain to be explored.

As sown using PAML branch-site models, variants harbored SNP in different functional categories according to their localization, meaning pulmonary or extra-pulmonary samples. Even if the tendencies observed require larger samples to be confirmed, the result obtained suggested an adaptation to pulmonary macrophage infection for pulmonary Mtb isolates and a metabolic adaptation for extra-pulmonary ones. Moreover, development of *in vitro* models will be needed to decipher the role and the impact of the identified pairwise variants between pulmonary and extra-pulmonary compartments.

Alongside that, in the present study, no correlation was found between the ranges of Mtb α-diversity and TB-associated severity markers. It may be due to the fact that the analysis was based on minimum number of variants estimated through WGS data to calculate Mtb α-diversity, which was a risk to underestimate micro-diversity in Mtb clinical isolate. However, to be exhaustive regarding the variant composition of an Mtb isolate, this would have required sequencing of several colonies for each Mtb clinical isolate, which is time- and cost-consuming. Nevertheless, detection of unfixed mutations at the level of WGS (meaning mutation frequencies between 10 and 90%) was enough to observe a strong correlation between Mtb micro-diversity detection and TB-associated severity markers.

As WGS is performed in routine practice in our lab, as well as in other TB diagnosis lab, it could be envisioned as an all-in-one solution, to detect antibiotic resistance [37], to infer Mtb transmission chains and to perform epidemiological monitoring [38,39] and now as a prognosis

tool. In the frame of cancer and microbiological research, calling algorithm for low frequency variants were developed [40–42] and may be adapt to Mtb WGS data. Therapeutic drug monitoring and implementation of additional management measures could be performed for patients with detectable Mtb micro-diversity in clinical isolates. It would ensure optimal anti-TB drug doses and prevent slow response to treatment, which would reduce risks of treatment failure and of drug resistance acquisition [43,44].

In conclusion, although these results need to be confirmed in an independent prospective validation study, Mtb micro-diversity within clinical isolate could be a useful prognosis tool to ensure optimal management of TB patients.

METHODS

Ethical considerations

For this study we recorded demographical (age, sex), clinical (extrapulmonary and/or pulmonary TB), microbiological (smear sputum results, growth delay, antibiotic resistance, lineage) and nutritional and immune data. All data were implemented in a database, in accordance with the decision 20-216 of the ethics committee of the Lyon University Hospital, France and the French Bioethics laws (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).

Mtb samples and data collection

In this retrospective study, a total of 355 Mtb clinical isolates were included, from 311 patients diagnosed with microbiologically proven TB at the Lyon University Hospital. We included all

Mtb clinical isolates for which WGS was performed in routine practice at the Lyon University Hospital from January 2017 to January 2020 and significant isolates of our collection (MDR Mtb, representative strains of large previously identified clusters, samples from patients with both pulmonary and extra-pulmonary TB) which were captured and implemented in the database during sequencing development in the lab [38,39]. It should be noted that we excluded pleural TB as depending of the clinical presentation it can be considered as PTB or EPTB. Microbiological characteristics, such as Mtb lineage, smear results, time to positivity and drug resistance, were recorded (**Table S1**), as was patients' demographic and baseline descriptive characteristics (**Table 1**). Regarding patients' descriptive characteristics, only data available between 2 weeks before TB diagnosis and 1 week after initiation of anti-TB treatment or nutritional supplementation were considered.

TB-associated severity indices

Three TB-associated severity indices were assessed in this study.

The modified Bandim score is the most largely used PTB prognosis score. It considers 5 symptoms (cough, hemoptysis, dyspnea, chest pain, night sweats) and 5 clinical findings (anemia, tachycardia, positive finding at lung auscultation, fever, BMI<18 and <16), with one point for each. One clinical finding was excluded, the mid upper arm circumference (MUAC) as this data was not available in the Lyon University Hospital. Accordingly, patients were stratified into two severity classes, mild (Bandim score ≤4) and moderate or severe (≥5) [26,27,45]. The nutritional status of TB patients was evaluated thanks to the Malnutrition Universal Screening Tool (MUST), which include three variables: unintentional weight loss score (weight loss < 5% = 0, weight loss 5–10% = 1, weight loss > 10% = 2), BMI score (BMI>20.0 = 0, BMI 18.5–20.0 = 1, BMI<18.5 = 2) and anorexia (if yes = 2). Malnutrition is frequently observed in

patients with PTB and a previous study showed a poorer prognosis for PTB patients with MUST ≥4 [28].

The nutritional risk score is a four-points score including both nutritional and immune characteristics: low BMI (<18.5), hypoalbuminaemia (<30.0 g/L), hypocholesterolaemia (<4mmol/L) and severe lymphocytopenia (<0.7 G cells/L). A high nutritional risk score (≥3) has been shown to be associated to poor prognosis in PTB [29,46].

Culture of Mycobacterium tuberculosis

The routine laboratory diagnostic workflow consisted of treatment of pulmonary samples with the modified Kubica's digestion-decontamination method [47], followed by inoculation in Mycobacteria Growth Indicator Tubes (MGITs) incubated in a BD BACTEC TM MGIT TM 960 instrument (BD, Sparks, MD, USA). Extrapulmonary samples were inoculated using the same medium without prior decontamination. Mtb genomic DNA extractions were performed after a single round of culture. Biobank Mtb isolates were inoculated in MGIT until exponential phase before Mtb genomic DNA extraction.

Whole genome sequencing

Genomic DNA of Mtb positive cultures was purified from cleared lysate using a QIAamp DNA mini Kit (Qiagen). DNA libraries were prepared with Nextera XT kit (Illumina, San Diego, USA). Samples were sequenced on NextSeq or MiSeq system (Illumina) to produce 150 or 300 base-pair paired-end reads at the Bio-Genet NGS facility of Lyon University Hospital, as previously described [38]. Reads were mapped with BOWTIE2 to the Mtb H37Rv reference genome (Genbank NC000962.2) and variant calling was made with SAMtools mpileup, as previously described [38].

358 ***Illumina data analysis***

359 A valid nucleotide variant was called if the position was covered by a depth of at least 10 reads
360 and supported by a minimum threshold rate of 10%. Regions with repetitive or similar
361 sequences were excluded, i.e. regions of PE, PPE, PKS, PPS, ESX genes. The reference genome
362 coverage breadth was at least 93% with a mean depth of coverage of at least 50x.

363

364 ***Variant assignment***

365 In a previous study, we showed no significant differences in variant detection and frequencies
366 between sequencing on direct samples and after subculture on media used in routine practice [6].
367 Moreover, for this study, 10 isolates were extracted and sequenced twice to evaluate the
368 variability in mutation frequencies between sequencing experiments. In both sequencing
369 experiments, 52 unfixed mutations were detected at similar frequencies (±10%), ranging from
370 10 to 90% (**Fig. S9A**). Accordingly, to identify the minimum number of variant in each Mtb
371 clinical isolate, a variant was defined as an assembly of mutations at frequencies of ±10% as
372 illustrated in **Fig. S9B**.

373

374 ***Mtb α-diversity indices***

375 The alpha diversity index was calculated by using the software R statistic, Package (vegan)
376 version 2.5-7.

377

378 ***Selection analysis***

379 We used PAML package to test for selection pressures in our dataset. This method uses
380 powerful statistics to test for heterogeneous dn/ds ratios at different positions and/or in different
381 branches. It has already been successfully used in MTC genome to confirm that positions
382 involved in drug resistance and some positions in membrane proteins are subjected to positive

383 selection [48]. Shortly, this method allows to compare scenarios including (H1) or not (H0) higher
384 dn/ds ratio on some residues. This means that two categories of sites are set up that have
385 different distributions of their dn/ds (options can force some of the characteristics of these
386 distributions, here both M1 and M2 models distributions were explored). The method allows to
387 identify whether scenario H1 is more likely than H0 using a Likelihood Ratio Test (true if p-
388 value<0.05).
389 Here we explored whether genes having the same functions (as characterized by available
390 annotation) have a higher probability to be under selection in the evolutionary branches leading
391 to one type of localization or the other (pulmonary versus extra-pulmonary). These tests were
392 performed using the branch-site model of PAML.
393 While compiling the statistics enabling LRT tests described above, PAML package also
394 associates a probability to each codon that underwent positive selection presure: the Bayesian
395 Empirical Bayes (BEB) posterior probability that this codon has a dn/ds ratio higher than 1 as
396 compared to the alternative scenario, starting from a (true if P>0.95).
397 To do so, we built the matrix including all SNPs for all variants (intra et inter patients
398 variations). For each SNP, we reconstituted the corresponding amino-acid using annotation
399 available from mycobrowser Release 2 (https://mycobrowser.epfl.ch/releases). The variant
400 phylogeny was reconstructed using RAxML using a GTRCAT model. All terminal branches
401 leading to pulmonary isolates were labelled for identifying selection in the lungs ("Pulmonary"
402 analysis). All terminal branches leading to any extra-pulmonary isolate were labelled for
403 identifying selection in microaerophilic organs *i.e.* organs other than the lungs ("Extra-
404 Pulmonary" analysis).
405
406 ***Statistical analysis***
407 *Univariate analysis*

The study variables were expressed as count (percentage, %) for dichotomous variables and as medians (interquartile range [IQR]) for continuous values. The number of missing values was excluded from the denominator. Non-parametric statistical methods Fisher exact test, $\chi 2$ test, Mann-Whitney U test and Kruskal-Wallis analysis, using Dunn's Multiple Comparison Test were used to compare groups, where appropriate. Statistical analyses were performed with Graph Pad Prism 5. *$p<0.05$, ** $p<0.01$, *** $p<0.001$.

Response variable and subsetting explanatory variables

We built a model to relate Mtb micro-diversity in pulmonary (P) and extra pulmonary (EP) isolates to various explanatory variables obtained for most patients in the cohort. The response variable we chose to model is a binary (0/1) measure of whether Mtb displayed or not some diversity as observed through counts of unique SNP profiles.

Among all the potential explanatory variables recorded in the initial dataset, we retained a chosen subset, following a series of filters:

1. We first removed 15 variables based on three non-exclusive conditions: (i) relatively low coverage in the cohort, (ii) little to no relationship with the response variable based on the literature, (iii) important redundancy with other variables in the dataset (e.g. percentage of neutrophil when the raw neutrophil count was also included);

2. Among the remaining variables, we removed all those that had less than 90% coverage (i.e. those that had missing values for 10% or more of the patients).

Apart from the location of the infection (P vs. EP), the final consolidated dataset included 31 variables: 10 categorical variables and 21 quantitative ones. This dataset consisted of 282 rows (patient:TBlocation), 200 for pulmonary locations and 82 for EP locations. 42 patients had both microbiologically proven P and EP locations, so the final dataset included 240 patients (158 with only P location, 40 with only EP location, and 42 patients with both locations).

433

434 *Model comparison*

435 In a first exploratory phase, we looked for all models explaining the response variable using a
436 limited number of explanatory variables. We used generalized linear modelling, assuming that
437 the binary response variable could be modelled through a binomial distribution and using a logit
438 transformation linking response to explanatory variables. With 31 possible explanatory
439 variables, the number of potential models to test is very high and impossible to tackle (2^{31}, i.e.
440 more than 2 billion models). In order to reduce this complexity, we chose to restrict our search
441 to models incorporating between 0 and 12 variables. As presented in the result, this approach
442 was sufficient to obtain "best models" that had fewer than 12 variables, hence hinting at the
443 uselessness of pursuing our search further into models of higher complexity.

444

445 To rank the different tested models, we used the Akaike Information Criterion corrected for
446 small sample sizes (AICc) [49] because we expected many tapered effects of the different
447 variables [50]. Models with the lowest AICc values were the ones that had the best goodness-of-
448 fit. Models were computed and compared using the 'glmulti' package version 1.0.8 in the R
449 software version 3.6.3 [51]. To optimize computation time, we first looked for all models with 0
450 – 9 variables among the 31 present in the dataset, and then looked for models with 10, 11 and
451 12 variables (thus, 4 parallel uses of glmulti). For each of these four chunks of model
452 comparison, we retained only the 500 best models (sensu AICc) and gathered all these models
453 in a consensus pool of good models. The importance of model variables was then assessed using
454 Akaike weights of each of the 31 variables by summing the Akaike weights of all models
455 incorporating the focal variable [50,52]. Since each variable was incorporated in exactly half of the
456 tested models, the importance is expected to be ½ for variables that did not modify model fit
457 better than the null expectation. Following Massol et al. (2007)[52], we thus considered that all

458 variables that were *likely* to have an effect on the response variable were those with an
459 importance larger than $1/(1+e^{-1})$, i.e. 0.73, *plausible* or *implausible* on either side of 0.5, and
460 unlikely when variable importance was lower than 0.27. For prediction purposes, we retained
461 all models that were within 2 units of the best model's AICc and used multi-model inference
462 based on this set of models [50]. We computed unconditional variance using the method of
463 Johnson and Omland (2004)[53] and obtained confidence intervals on model predictions using
464 the method suggested by Burnham and Anderson [50], using function 'predict' in the R package
465 'glmulti'.
466
467 *Parameter value comparisons*
468 When categorical variables had an effect on the response variable, we tested for pairwise
469 differences in the coefficients associated to the different levels of the categorical variables. To
470 do so, we used the R package 'emmeans' version 1.4.5, which tested pairwise differences
471 between marginal means averaged over all values of other categorical variables (e.g. differences
472 between "Mtb lineages" were assessed by averaging the effect of "double location"), using
473 asymptotic test on the z-score obtained from the pair of coefficients [54].
474
475 *Estimating multi-model errors*
476 Once the set of good models had been determined, we re-sampled the dataset in order to obtain
477 estimates of the multi-model prediction errors (i.e. false negatives and positives). For a given
478 target proportion of initial dataset rows to be included in the training set, we drew a random
479 sample of rows stratified by unique combinations of categorical variables used by the multi-
480 model, i.e. we made sure that all combinations of factors were included in the training set. This
481 resulted in actual fraction of data rows in the training dataset slightly higher than the target

proportion because some combinations of categorical variable modalities were quite rare and thus systematically added to the training set.

With a given training dataset, we considered the rest of the dataset as validation dataset. We fitted the whole set of good models using only the training dataset, with possibly estimates of model coefficients different from those obtained with the whole dataset due to sampling. Based on the probability of observing some Mtb micro-diversity for all samples in the training set, we looked for a threshold on this probability that would maximize the true skill statistic (TSS) if the multi-model were to predict diversity when model predictions were above this threshold and no diversity otherwise. TSS is a simple statistic (also called informedness or Youden's J statistic) equal to the sum of sensitivity and specificity minus one. This threshold optimization procedure was performed using function 'optim.thresh' in R package 'SDMTools' version 1.1-221.2 [55].

The performance of the multi-model inferred from the training dataset was assessed by predicting the response variable in the validation dataset, using the above-mentioned threshold for predicting presence/absence of Mtb micro-diversity. This prediction yielded a confusion matrix (observed vs. predicted absence/presence of Mtb micro-diversity) which was then analysed using standard statistics, i.e. its sensitivity, specificity, accuracy (1 - error rate) and TSS.

ACKNOWLEDGMENTS

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

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests

The authors declare that they have no conflict of interest.

REFERENCES

1. WHO, G. Global tuberculosis report 2020. https://www.who.int/publications-detail-redirect/9789240013131 (2020).
2. Gagneux, S. & Small, P. M. Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development. *The Lancet Infectious Diseases* **7**, 328–337 (2007).
3. Brown, A. C. *et al.* Rapid Whole-Genome Sequencing of Mycobacterium tuberculosis Isolates Directly from Clinical Samples. *Journal of Clinical Microbiology* **53**, 2230–2237 (2015).
4. Casali, N. *et al.* Whole Genome Sequence Analysis of a Large Isoniazid-Resistant Tuberculosis Outbreak in London: A Retrospective Observational Study. *PLoS Med* **13**, (2016).
5. Doyle, R. M. *et al.* Direct Whole-Genome Sequencing of Sputum Accurately Identifies Drug-Resistant *Mycobacterium tuberculosis* Faster than MGIT Culture Sequencing. *Journal of Clinical Microbiology* **56**, (2018).
6. Genestet, C. *et al.* Subcultured Mycobacterium tuberculosis isolates on different growth media are fully representative of bacteria within clinical samples. *Tuberculosis (Edinb)* **116**, 61–66 (2019).
7. Ley, S. D., de Vos, M., Van Rie, A. & Warren, R. M. Deciphering Within-Host Microevolution of Mycobacterium tuberculosis through Whole-Genome Sequencing: the Phenotypic Impact and Way Forward. *Microbiol Mol Biol Rev* **83**, (2019).

8. Lieberman, T. D. *et al.* Genomic diversity in autopsy samples reveals within-host dissemination of HIV-associated *Mycobacterium tuberculosis*. *Nature Medicine* **22**, 1470–1474 (2016).
9. Nimmo, C. *et al.* Dynamics of within-host Mycobacterium tuberculosis diversity and heteroresistance during treatment. *EBioMedicine* **55**, 102747 (2020).
10. O'Neill, M. B., Mortimer, T. D. & Pepperell, C. S. Diversity of Mycobacterium tuberculosis across Evolutionary Scales. *PLOS Pathogens* **11**, e1005257 (2015).
11. Pérez-Lago, L. *et al.* Whole genome sequencing analysis of intrapatient microevolution in Mycobacterium tuberculosis: potential impact on the inference of tuberculosis transmission. *J. Infect. Dis.* **209**, 98–108 (2014).
12. Shockey, A. C., Dabney, J. & Pepperell, C. S. Effects of Host, Sample, and in vitro Culture on Genomic Diversity of Pathogenic Mycobacteria. *Front. Genet.* **10**, (2019).
13. Vargas, R. *et al.* In-host population dynamics of M. tuberculosis during treatment failure. *bioRxiv* 726430 (2019) doi:10.1101/726430.
14. Votintseva, A. A. *et al.* Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome Sequencing of Direct Respiratory Samples. *Journal of Clinical Microbiology* **55**, 1285–1298 (2017).
15. Genestet, C. *et al.* Mycobacterium tuberculosis micro-diversity promotes intra-macrophagic persistence and antibiotic tolerance. *bioRxiv* 2020.05.22.110775 (2020) doi:10.1101/2020.05.22.110775.
16. Cohen, T. *et al.* Within-Host Heterogeneity of Mycobacterium tuberculosis Infection Is Associated With Poor Early Treatment Response: A Prospective Cohort Study. *J Infect Dis* **213**, 1796–1799 (2016).

555 17. Shin, S. S. *et al.* Mixed Mycobacterium tuberculosis-Strain Infections Are Associated
556 With Poor Treatment Outcomes Among Patients With Newly Diagnosed Tuberculosis,
557 Independent of Pretreatment Heteroresistance. *J. Infect. Dis.* **218**, 1974–1982 (2018).
558 18. Gagneux, S. Ecology and evolution of *Mycobacterium tuberculosis*. *Nature Reviews*
559 *Microbiology* **16**, 202–213 (2018).
560 19. Ailloud, F. *et al.* Within-host evolution of Helicobacter pylori shaped by niche-specific
561 adaptation, intragastric migrations and selective sweeps. *Nat Commun* **10**, 1–13 (2019).
562 20. Azarian, T., Ridgway, J. P., Yin, Z. & David, M. Z. Long-Term Intrahost Evolution of
563 Methicillin Resistant Staphylococcus aureus Among Cystic Fibrosis Patients With
564 Respiratory Carriage. *Front. Genet.* **10**, (2019).
565 21. Chaguza, C. *et al.* Within-host microevolution of Streptococcus pneumoniae is rapid and
566 adaptive during natural colonisation. *Nat Commun* **11**, 3442 (2020).
567 22. Harkins, C. P. *et al.* The Microevolution and Epidemiology of Staphylococcus aureus
568 Colonization during Atopic Eczema Disease Flare. *J. Invest. Dermatol.* **138**, 336–343
569 (2018).
570 23. Levade, I. *et al.* Vibrio cholerae genomic diversity within and between patients. *Microbial*
571 *Genomics* **3**, e000142 (2017).
572 24. Winstanley, C., O'Brien, S. & Brockhurst, M. A. Pseudomonas aeruginosa Evolutionary
573 Adaptation and Diversification in Cystic Fibrosis Chronic Lung Infections. *Trends*
574 *Microbiol.* **24**, 327–337 (2016).
575 25. Wongphutorn, P., Chomvarin, C., Sripa, B., Namwat, W. & Faksri, K. Detection and
576 genotyping of Helicobacter pylori in saliva versus stool samples from asymptomatic
577 individuals in Northeastern Thailand reveals intra-host tissue-specific H. pylori subtypes.
578 *BMC Microbiol.* **18**, 10 (2018).

26. Dewi, D. N. S. S., Mertaniasih, N. M. & 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* **14**, 8–15 (2020).
27. Rudolf, F. *et al.* The Bandim tuberculosis score: reliability and comparison with the Karnofsky performance score. *Scand. J. Infect. Dis.* **45**, 256–264 (2013).
28. Miyata, S., Tanaka, M. & Ihaku, D. The prognostic significance of nutritional status using malnutrition universal screening tool in patients with pulmonary tuberculosis. *Nutr J* **12**, 42 (2013).
29. Kim, D. K. *et al.* Nutritional deficit as a negative prognostic factor in patients with miliary tuberculosis. *European Respiratory Journal* **32**, 1031–1036 (2008).
30. Casha, A. R. & Scarci, M. The link between tuberculosis and body mass index. *Journal of Thoracic Disease* **9**, E301-E303–E303 (2017).
31. Cegielski, J. P. & McMurray, D. N. The relationship between malnutrition and tuberculosis: evidence from studies in humans and experimental animals. *Int. J. Tuberc. Lung Dis.* **8**, 286–298 (2004).
32. van Crevel, R. *et al.* Decreased plasma leptin concentrations in tuberculosis patients are associated with wasting and inflammation. *J. Clin. Endocrinol. Metab.* **87**, 758–763 (2002).
33. Dao, C. N. *et al.* Hyponatremia, hypochloremia, and hypoalbuminemia predict an increased risk of mortality during the first year of antiretroviral therapy among HIV-infected Zambian and Kenyan women. *AIDS Res. Hum. Retroviruses* **27**, 1149–1155 (2011).
34. Vrieling, F. *et al.* Patients with Concurrent Tuberculosis and Diabetes Have a Pro-Atherogenic Plasma Lipid Profile. *EBioMedicine* **32**, 192–200 (2018).

35. Shivakoti, R., Sharma, D., Mamoon, G. & Pham, K. Association of HIV infection with extrapulmonary tuberculosis: a systematic review. *Infection* **45**, 11–21 (2017).

36. Skogmar, S. *et al.* CD4 Cell Levels during Treatment for Tuberculosis (TB) in Ethiopian Adults and Clinical Markers Associated with CD4 Lymphocytopenia. *PLOS ONE* **8**, e83270 (2013).

37. Genestet, C. *et al.* Whole-genome sequencing in drug susceptibility testing of Mycobacterium tuberculosis in routine practice in Lyon, France. *Int. J. Antimicrob. Agents* **55**, 105912 (2020).

38. Genestet, C. *et al.* Prospective Whole-Genome Sequencing in Tuberculosis Outbreak Investigation, France, 2017–2018. *Emerging Infectious Diseases* **25**, 589–592 (2019).

39. Genestet, C. *et al.* Routine survey of Mycobacterium tuberculosis isolates reveals nosocomial transmission. *Eur. Respir. J.* **55**, (2020).

40. Eliseev, A. *et al.* Evaluation of haplotype callers for next-generation sequencing of viruses. *Infect. Genet. Evol.* **82**, 104277 (2020).

41. Spencer, D. H. *et al.* Performance of Common Analysis Methods for Detecting Low-Frequency Single Nucleotide Variants in Targeted Next-Generation Sequence Data. *J Mol Diagn* **16**, 75–88 (2014).

42. Xu, C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and Structural Biotechnology Journal* **16**, 15–24 (2018).

43. Alsultan, A. & Peloquin, C. A. Therapeutic drug monitoring in the treatment of tuberculosis: an update. *Drugs* **74**, 839–854 (2014).

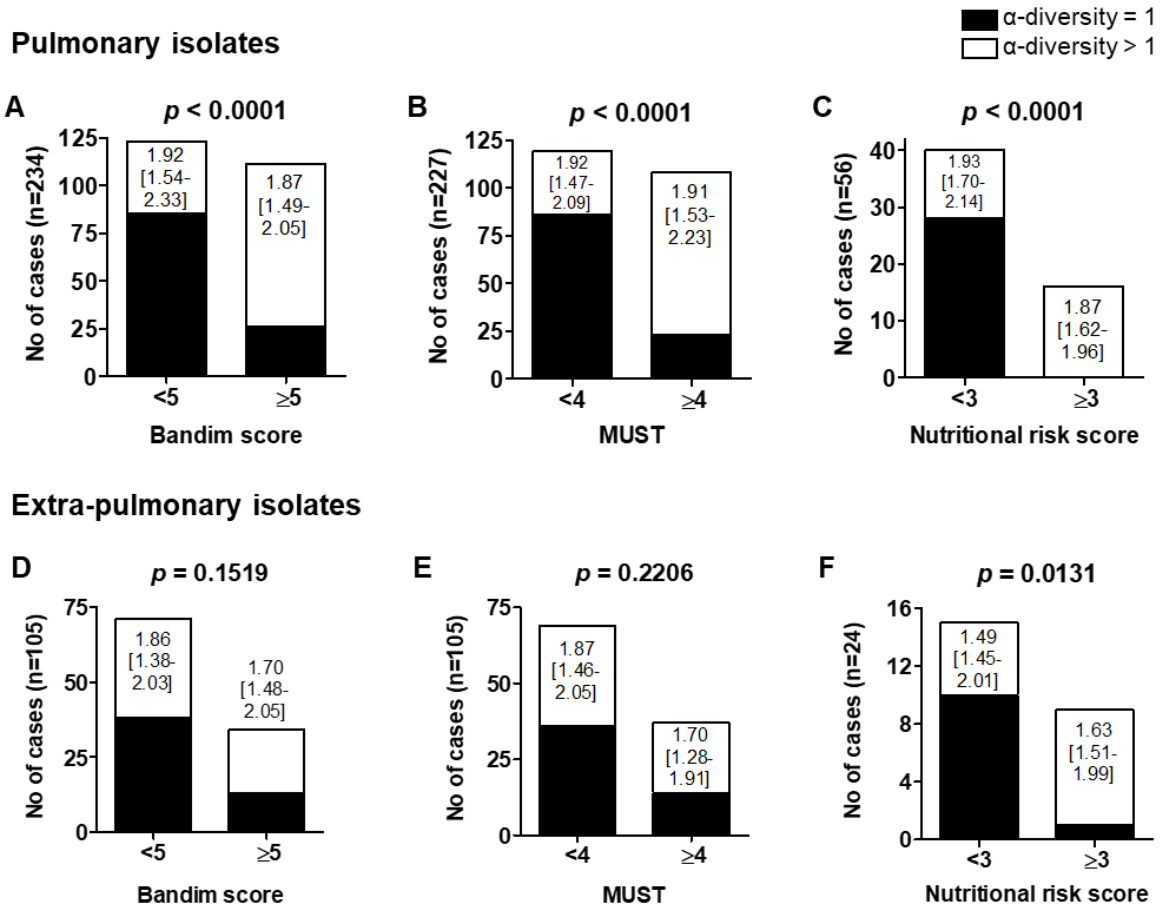
44. Choi, R., Jeong, B.-H., Koh, W.-J. & Lee, S.-Y. Recommendations for Optimizing Tuberculosis Treatment: Therapeutic Drug Monitoring, Pharmacogenetics, and Nutritional Status Considerations. *Ann Lab Med* **37**, 97–107 (2017).

629 45. Wejse, C. *et al.* TBscore: Signs and symptoms from tuberculosis patients in a low-
630 resource setting have predictive value and may be used to assess clinical course. *Scand. J.*
631 *Infect. Dis.* **40**, 111–120 (2008).
632 46. Kim, H.-J. *et al.* The impact of nutritional deficit on mortality of in-patients with
633 pulmonary tuberculosis. *Int. J. Tuberc. Lung Dis.* **14**, 79–85 (2010).
634 47. Kent, P. T. & Kubica, G. P. Public Health Mycobacteriology: A Guide for the Level III
635 Laboratory. | National Technical Reports Library - NTIS.
636 https://ntrl.ntis.gov/NTRL/dashboard/searchResults/titleDetail/PB86216546.xhtml (1985).
637 48. Osório, N. S. *et al.* Evidence for Diversifying Selection in a Set of Mycobacterium
638 tuberculosis Genes in Response to Antibiotic- and Nonantibiotic-Related Pressure.
639 *Molecular Biology and Evolution* **30**, 1326–1336 (2013).
640 49. Hurvich, C. M. & Tsai, C.-L. Regression and time series model selection in small
641 samples. *Biometrika* **76**, 297–307 (1989).
642 50. Burnham, K. P. & Anderson, D. R. *Model Selection and Multimodel Inference: A*
643 *Practical Information-Theoretic Approach*. (Springer-Verlag, 2002). doi:10.1007/b97636.
644 51. Calcagno, V. & Mazancourt, C. de. glmulti: an R package for easy automated model
645 selection with (generalized) linear models. *Journal of Statistical Software* **34**, 1 (2010).
646 52. Massol, F., David, P., Gerdeaux, D. & Jarne, P. The influence of trophic status and large-
647 scale climatic change on the structure of fish communities in Perialpine lakes. *J Anim*
648 *Ecol* **76**, 538–551 (2007).
649 53. Johnson, J. B. & Omland, K. S. Model selection in ecology and evolution. *Trends in*
650 *Ecology & Evolution* **19**, 101–108 (2004).
651 54. Lenth, R. V. *et al. emmeans: Estimated Marginal Means, aka Least-Squares Means*.
652 (2021).

653 55. VanDerWal, J., Falconi, L., Januchowski, S. & Storlie, L. S. and C. *SDMTools: Species*
654 *Distribution Modelling Tools: Tools for processing data associated with species*
655 *distribution modelling exercises*. (2014).

656

657 **FIGURES AND LEGENDS**



658

659 **Figure 1: The detection of genetic micro-diversity in Mtb isolates is associated with TB-**
660 **associated severity indices**

661 Association between the detection and the range of Mtb α-diversity from pulmonary (A-C) or
662 extra-pulmonary samples (D-F) and TB-associated severity indices, the Bandim score (A, D),
663 the MUST (malnutrition universal screening tool, B and E), and the nutritional risk score (C,
664 F). Black bar: α-diversity=1 no diversity detected by WGS; white bar: α-diversity>1 at least
665 two variants detected by WGS. *p* = x.xxxx: non-parametric statistical method Fisher exact test

was used to compare groups. *p-value* < 0.05 was considered significant. x.xx [y.yy-z.zz]: median [IQR] of Mtb α-diversity. Mann-Whitney U test was used to compare ranges of Mtb α-diversity between groups (no statistical differences observed).

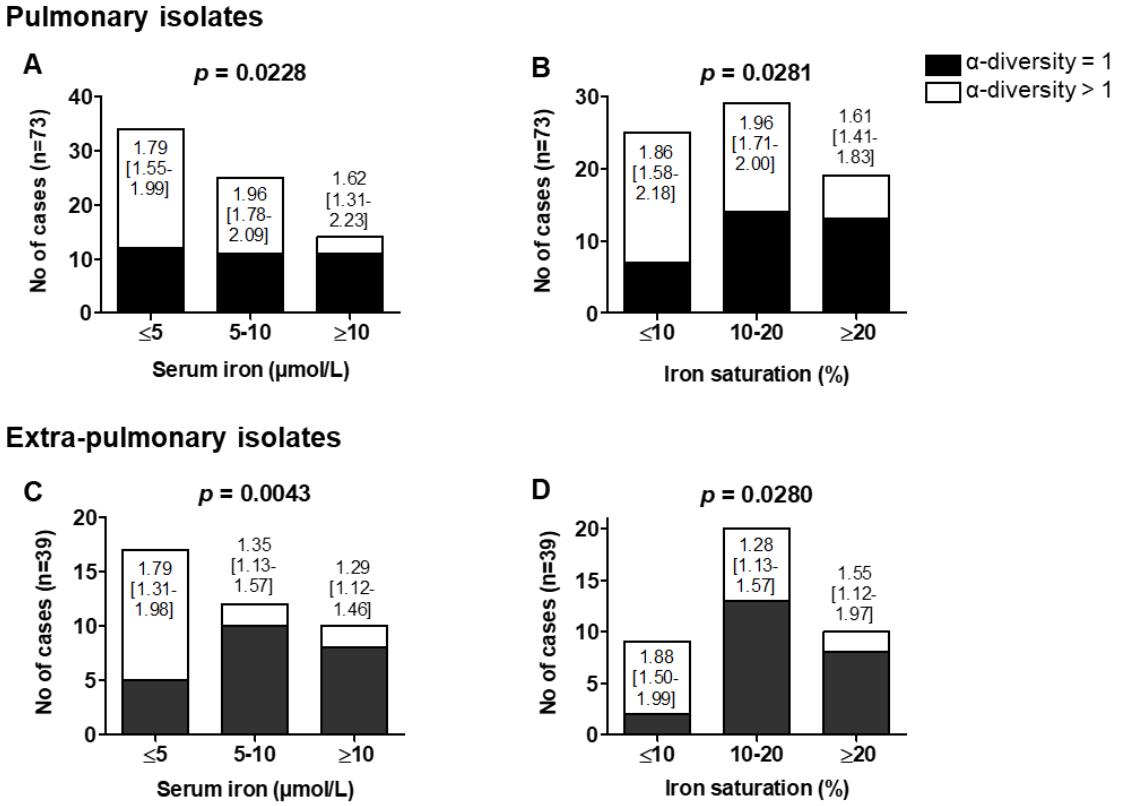


Figure 2: The detection of genetic micro-diversity in pulmonary and extra-pulmonary Mtb isolates is associated with iron deficiency

Association between the detection and the range of Mtb α-diversity from pulmonary (A and B) and extra-pulmonary (C and D) samples and serum iron level (A and C) and transferrin saturation (B and D). Black bar: α-diversity=1 no diversity detected by WGS; white bar: α-diversity>1 at least two variants detected by WGS. *p* = x.xxxx: non-parametric statistical method $\chi 2$ test was used to compare groups. *p-value* < 0.05 was considered significant. x.xx [y.yy-z.zz]: median [IQR] of α-diversity. Kruskal-Wallis analysis, using Dunn's Multiple Comparison Test was used to compare ranges of α-diversity between groups (no statistical differences observed).

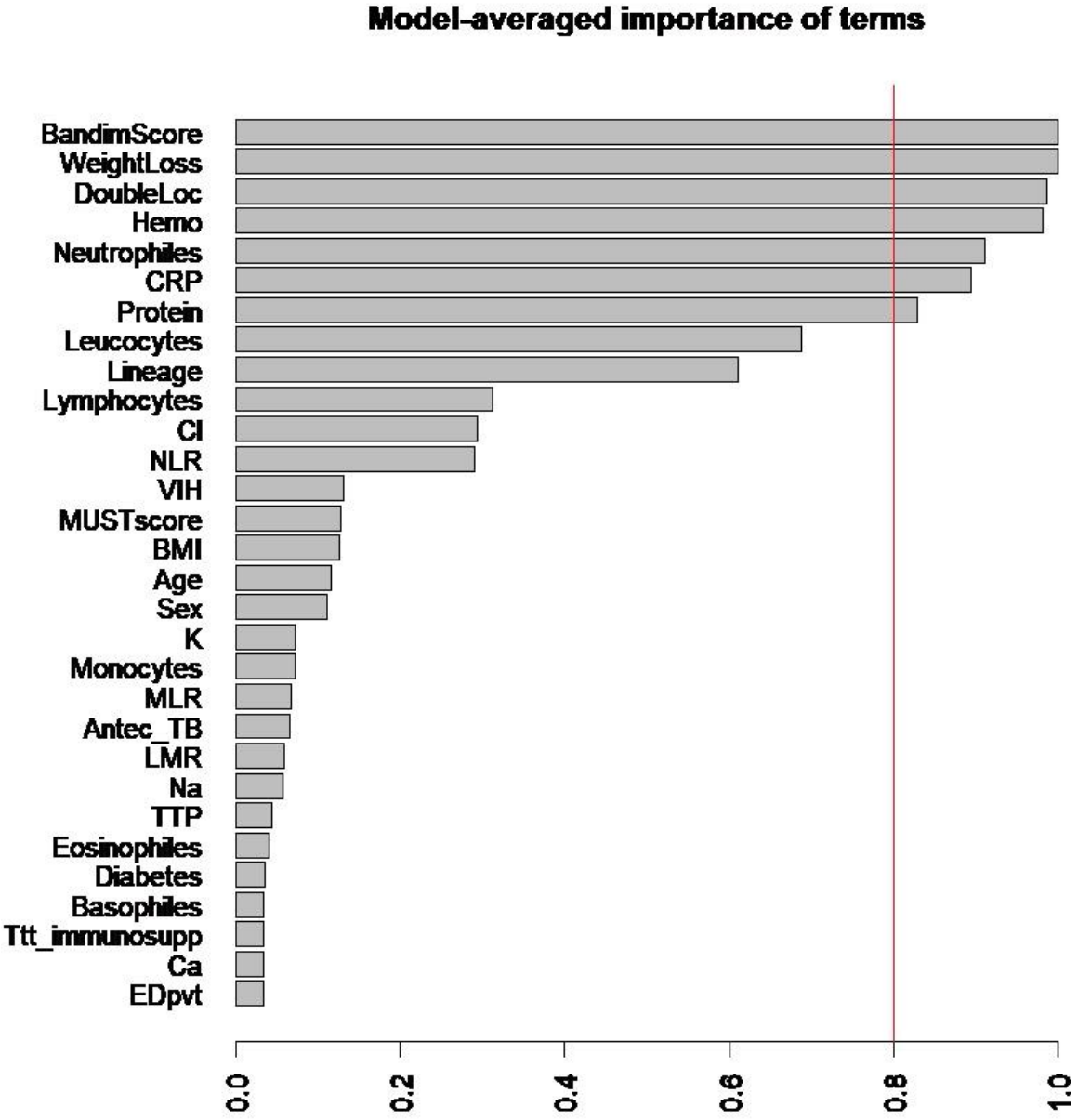


Figure 3: Model-averaged importance of terms for Mtb micro-diversity in pulmonary isolates.

Model-averaged importance of each term in the model (**Table 2**), which is defined as the proportion of the 2000 best models in which a given term appears. Red line indicates 80% support. Terms with an importance above the red line are included in our final model.

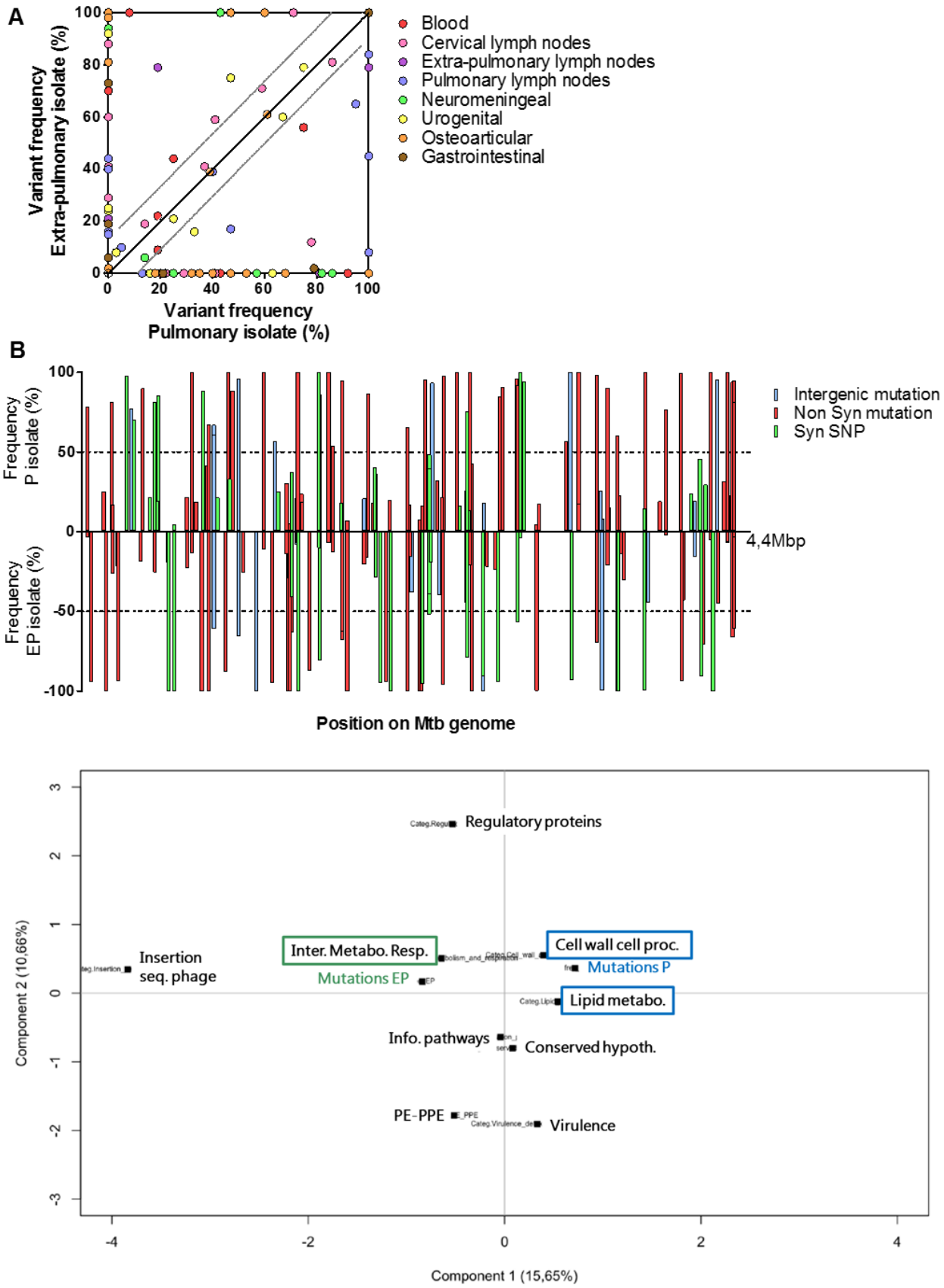


Figure 4: Compartmentalization of Mtb variants between pulmonary and extra-pulmonary compartments

(A) Frequencies in pulmonary and extra-pulmonary isolates of the 104 variants identified in paired isolates from the 42 patients of the training cohort with both microbiologically proven pulmonary and extra-pulmonary TB. (B) Repartition on Mtb reference genome of the 168 pairwise mutations distance identified between the 42 paired isolates from the training cohort. Each bar represents a pairwise mutation distance between paired pulmonary and extra-pulmonary isolates. *y-axis* positive value: mutation frequency in pulmonary isolate; *y-axis* negative value: mutation frequency in extra-pulmonary isolate. Bleu bar: intergenic mutation; red bar: nonsynonymous mutation; green bar: synonymous single nucleotide polymorphism (SNP). (C) Principal component analysis (PCA) for mixed data was performed with the delta of frequency of the 104 nonsynonymous pairwise mutation observed between paired pulmonary and extra-pulmonary isolates and the functional categories of these mutations. The first principal component allowed a discrimination of pulmonary and extra-pulmonary mutations.