

Genome-to-genome analysis reveals associations between human and mycobacterial genetic variation in tuberculosis patients from Tanzania

Zhi Ming Xu^{1,2}, Michaela Zwyer^{3,4}, Daniela Brites^{3,4}, Hellen Hiza^{3,4}, Mohamed Sasamalo⁵, Miriam Reinhard^{3,4}, Anna Doetsch^{3,4}, Sonia Borrell^{3,4}, Olivier Naret^{1,2}, Sina Rüeger^{1,2}, Dylan Lawless^{1,2}, Faima Isihaka⁵, Hosiana Temba⁵, Thomas Maroa⁵, Rastard Naftari⁵, Christian Beisel⁶, Jerry Hella⁵, Klaus Reither^{3,4}, Damien Portevin^{3,4}, Sebastien Gagneux^{3,4}, Jacques Fellay^{1,2,7*}

¹School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

²Swiss Institute of Bioinformatics, Lausanne, Switzerland

³Swiss Tropical and Public Health Institute, Allschwil, Switzerland

⁴University of Basel, Basel, Switzerland

⁵Ifakara Health Institute, Dar es Salaam, Tanzania

⁶Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland

⁷Precision Medicine Unit, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland

*Corresponding Author:

Jacques Fellay, School of Life Sciences, École Polytechnique Fédérale de Lausanne, 1015, Lausanne, Switzerland

E-mail: jacques.fellay@epfl.ch

1 Abstract

2 The risk and prognosis of tuberculosis (TB) are affected by both human and bacterial genetic
3 factors. To identify interacting human and bacterial genetic loci, we leveraged paired human and
4 *Mycobacterium tuberculosis* (*M.tb*) genomic data from 1000 Tanzanian TB patients. Through a
5 genome-to-genome approach, we identified two pairs of human and *M.tb* genetic variants that are
6 significantly associated. One of the human genetic variants maps to the intron of *PRDM15*, a gene
7 involved in apoptosis regulation. The other human variant maps to an intergenic region close to
8 *TIMM21* and *FBXO15*. In addition, we observed that a group of linked *M.tb* epitope variants were
9 significantly associated with HLA-DRB1 variation. This suggests that even though epitope
10 variation is rare in *M.tb* in general, specific epitopes might still be under immune selective
11 pressure. Overall, our study pinpoints sites of genomic conflicts between humans and *M.tb*,
12 suggesting bacterial escape from host selection pressure.

13 Introduction

14 Tuberculosis (TB) is an infectious disease mainly caused by *Mycobacterium tuberculosis* (*M.tb*).
15 The infection primarily affects the lungs, although extrapulmonary TB manifests in around 20-25%
16 of cases¹. TB continues to pose a significant public health challenge, particularly in low- and
17 middle-income countries. Approximately 10 million people develop active TB yearly, resulting in
18 1.6 million deaths². Except in 2022 after the COVID-19 pandemic, TB remains the leading cause
19 of death from an infectious disease worldwide^{2,3}.

20
21 Exposure to *M.tb* can lead to a wide range of clinical manifestations^{4,5}. Most infected individuals
22 develop latent TB, where quiescent infection can heal or persist but remain asymptomatic⁶.
23 However, latent TB might progress into active TB disease in 5%-15% of cases⁷, where symptoms
24 such as cough, fever, or weight loss develop. Several factors modulate the risk of developing
25 active TB, including HIV co-infection, which increases the risk by almost 20-fold⁸. Additional risk
26 factors include malnutrition, alcohol abuse, smoking, and other comorbidities such as diabetes⁹⁻
27 ¹¹.

28
29 Both human and bacterial genetic factors modulate the risk of developing active TB and its
30 prognosis. Early twin studies have suggested that TB susceptibility has a heritable
31 component^{12,13}. In the past decade, genome-wide association studies (GWAS) have identified
32 many genetic loci involved in TB susceptibility¹⁴⁻²³. However, a significant limitation of GWAS of
33 TB is that associations failed to replicate across studies in different population²⁴. This might be
34 explained in part by a failure to consider bacterial genetic variation. Human-adapted *M.tb* strains
35 are classified into nine lineages (L1-L9)²⁵, which can be further classified into sublineages
36 according to phylogenetic markers²⁶. Despite *M.tb* strains being highly clonal²⁷, different lineages
37 show some difference in pathogenicity²⁸. The prevalence of some lineages is also highly
38 geographically structured²⁹. While this could be partially due to genetic drift or reflect migration
39 patterns, another possibility is that this reflects adaptation of lineages to human populations with
40 particular genetic backgrounds^{30,31}.

41
42 A hypothesis-free method to identify interacting human and pathogen genetic loci is to jointly
43 analyze paired pathogen and human genomes isolated from patients who have developed the
44 respective disease. Specifically, a genome-to-genome (G2G) approach could be used, where
45 associations between all pairs of host and pathogen genetic variants are tested³². Significant
46 associations identified under this framework reflect combinations of host and pathogen genotypes

47 that jointly modulate the risk of developing an infectious disease upon exposure to the pathogen.
48 In addition, this approach can also identify pathogen loci that have undergone intra-host selection
49 driven by specific host genetic variants.

50
51 In this study, we leveraged paired *M.tb* and human genomic data in a cohort of 1000 active TB
52 patients recruited in Dar es Salaam, Tanzania. Using a G2G approach, we identified significant
53 associations between humans and *M.tb* genetic variants, indicating sites of genomic conflict.

54 Results

55 Study description

56 This study was based on a cohort of 1000 active TB patients recruited in Dar es Salaam, Tanzania
57 (TB-DAR) for which we generated both *M.tb* and human genomic data. [Table 1](#) summarizes the
58 demographic and clinical characteristics of the cohort. In brief, the cohort consisted of 70% males
59 and 30% females, with a mean age of 34. 16% of patients were co-infected with HIV. Patients
60 were infected with *M.tb* lineage L3 (43.1%), L4 (33.2%), L1 (15.5%) and L2 (8.2%).

61
62 As part of the study, we collected clinical measures including TB score, X-ray score, and
63 GeneXpert Ct-value. TB score is a symptoms-based score. X-ray score corresponds to lung
64 damage severity based on chest x-ray images. GeneXpert Ct-value is based on RT-PCR, and is
65 inversely correlated with sputum bacterial load. We observed that patients infected with different
66 lineages did not present significantly different clinical characteristics (ANOVA Test; TB Score: p
67 = 0.43, X-ray score: p = 0.24, and Ct-value: p = 0.36).

68 Associations between *M.tb* and human genetic variants

69 We tested for associations between all human and pathogen variants. Based on statistical power
70 calculations ([Supplementary Figure 1](#)), we restricted our analysis to 1538 common *M.tb* amino-
71 acid variants with minor-allele frequency (MAF) > 0.015 and 6,603,291 common human variants
72 with MAF > 0.05. We corrected for population stratification using the top three human principal
73 components (PCs). Genomic inflation factors (λ) were below 1, indicating the absence of inflation
74 of test statistics due to stratification ([Supplementary Figure 2](#)).

75
76 Considering that a large proportion of *M.tb* variants were perfectly correlated ([Supplementary](#)
77 [Figure 3](#)), many association tests were not independent ([Supplementary Figure 4](#)). Therefore, we

78 derived a multiple-testing corrected p-value threshold ($p < 1.02e-10$) based on the number of
79 independent tests and the genome-wide significant threshold of $5e-8$ for the human genome (See
80 [Methods](#)).

81
82 We identified two significant G2G associations ([Figure 1](#)). The first association was between the
83 *M.tb* variant Rv2348c I101M and the human single nucleotide polymorphism (SNP) rs12151990
84 ($p = 4.7e-11$, OR = 5.6). rs12151990 is an intronic SNP in *PRDM15* located on chromosome 21
85 ([Supplementary Figure 5A](#)). Colocalization of GWAS and *PRDM15* eQTL signals suggested
86 shared causal variants in some tissues ([Supplementary Figure 6A](#)), albeit not in the lung.
87 According to gnomAD, rs12151990 is common (MAF > 0.05) in all populations except South
88 Asians. On the bacterial side, Rv2348c I101M was prevalent in lineage 4 (MAF = 0.051) but also
89 found in L1 (MAF = 0.007) in our cohort. Within L4, the variant belonged to a subclade within
90 sublineage L4.3 but also displayed homoplasmy on three occasions which could indicate positive
91 selection ([Figure 2](#)). Indeed, when looking for this variant in a global reference set of L1-L4 *M.tb*
92 genomes ($n = 11,818$) compiled by Zwyer et al³³, we saw that the mutation had arisen
93 independently multiple times within each lineage, suggesting that it is under positive selection.
94 Using PAML, we formally tested for positive selection on a subset of 500 randomly selected
95 strains from the reference set and on a randomly selected subset of 300 genomes for each lineage
96 separately. We found that Rv2348c was indeed under positive selection in L1 and L4 ($p < 0.0001$),
97 but not in L2 and L3 ($p = 0.48$ and 0.40 , respectively). Moreover, it corresponds to the only codon
98 within Rv2348c that was identified to be under positive selection in L1 and L4 (posterior
99 probability > 99%).

100
101 The second association was between the *M.tb* variant FixA T67M and the human SNP
102 rs75769176 ($p = 6.3e-11$, OR = 6.7). rs75769176 maps to an intergenic region close to *FBXO15*
103 and *TIMM21* on chromosome 18 ([Supplementary Figure 5B](#)). The GWAS signal did not colocalize
104 with eQTLs signals from either gene ([Supplementary Figure 6B and 6C](#)). According to gnomAD,
105 rs75769176 is specific to African populations (MAF = 0.08). On the bacterial side, FixA T67M was
106 found exclusively in lineage L3 (MAF = 0.053) and belonged to a subclade within sublineage L3.1
107 in our cohort. When investigating the global reference dataset, we still found the variant
108 exclusively in sublineage L3.1.

109
110 In an attempt to identify potential confounders, we tested whether the two *M.tb* variants were
111 associated with any patient characteristics ([Supplementary Table 1](#) and [Supplementary Table 2](#)).

112 The results indicated that no clinical measures were significantly associated with the *M.tb*
113 variants. Importantly, HIV status was also not significantly associated. We observed that FixA
114 T67M was associated with the second human genetic PC. However, this PC was already included
115 as a covariate in the G2G study to prevent spurious associations driven by population
116 stratification.

117
118 We next tested whether the G2G-associated human and bacterial variants were associated with
119 any clinical measures of TB ([Supplementary Table 3](#)). On the bacterial side, we did not identify
120 variants significantly associated with any clinical measure. On the human side, the minor allele at
121 rs75769176 was negatively associated with TB score ($p = 0.04$, $\log(\text{OR}) = -0.34$), indicating an
122 association with less severe disease.

123
124 Finally, we investigated whether human variants previously found to be associated with
125 susceptibility to developing active TB were also associated with *M.tb* variation in our study
126 ([Supplementary Table 4](#)). For each human variant, we extracted the most significant G2G-
127 associated *M.tb* variant. For human variants that were rare within our cohort, we tested all
128 common variants within $\pm 5\text{kb}$ and extracted the top association. The strongest association we
129 identified was between rs4240897 and Rv2963 K165fs ($p = 9.5e-5$, $\text{OR} = 1.4$). We also repeated
130 the same analysis but with human variants identified in previous host-pathogen studies to be
131 associated with *M.tb* lineages or clades ([Supplementary Table 5](#)). The strongest association we
132 identified was between rs146731249 (a proxy for rs17235409) and HsdM K483E ($p = 5.0e-6$, OR
133 $= 4.8$).

134 Interactions between *M.tb* and human pathways

135
136 To identify whether G2G associations were enriched for certain molecular functions, we leveraged
137 human and *M.tb* pathways from existing gene regulatory network databases. Human pathways
138 were extracted from the Molecular Signatures Database (MSigDB). *M.tb* pathways were either
139 extracted from a co-regulatory network database (MTB Network Portal) or constructed from
140 stringDB. To construct pathway-to-pathway edges between all pairs of human and *M.tb* pathways,
141 we mapped variants to genes and binarized G2G associations by applying a lenient threshold
142 ($\text{FDR} < 0.15$). We then calculated the density of pathway-to-pathway edges between all pairs.

143

144 To search for enriched pairs of human and *M.tb* pathways, we compared the true densities to
145 those found based on permuted datasets. The pair of pathways with the highest enrichment of
146 G2G associations had a density of 0.11 (3 edges / 28 possible edges). However, the enrichment
147 was not statistically significant ($p=0.07$) compared to the distribution derived from permutations
148 ([Supplementary Figure 7](#)).

149

150 Associations between variants in *M.tb* T cell epitopes and human HLA 151 variation

152

153 Escape mutations in T cell epitopes are commonly observed in pathogens causing chronic
154 infections to avoid HLA-driven cellular immunity. To search for potential HLA-induced *M.tb*
155 variants, we extracted experimentally validated T-cell epitopes from the Immune Epitope
156 Database (IEDB) and retained those that were polymorphic within our cohort (frequency > 0.015).
157 Fifteen *M.tb* variants overlapped with at least one epitope: 9 epitope variants were independent,
158 while 4 variants were perfectly correlated with each other and thus grouped into a single variant
159 set ([Supplementary Figure 8](#)). On the human side, we imputed HLA amino acid variants and HLA
160 4-digit alleles using a trans-ethnic reference panel TopMed.

161

162 For each *M.tb* epitope variant, we tested whether it was associated with any human HLA allele or
163 amino acid variant. To correct for multiple testing, we applied a Bonferroni corrected significance
164 threshold of $1.15e-5$ based on the number of epitope variants, the number of human HLA amino
165 acid positions or alleles, and an alpha level of 0.05. A significant association was identified
166 between an epitope variant set (EspK L39W, EsxB E68K, Mpt70 A21T, RimJ R72L) and HLA-
167 DRB1 H96E ($p = 7.3e-06$, OR = 6.31). The variant EsxB E68K maps to epitopes that have been
168 experimentally validated to be restricted by various HLA-DRB1 alleles ([Supplementary Table 6](#)).

169

170 Finally, we extracted the top association for every *M.tb* epitope variant and compared that against
171 the top associations for *M.tb* variants that are not part of any annotated T cell epitope. We
172 observed that the association between the epitope variant set (EspK L39W, EsxB E68K, Mpt70
173 A21T, RimJ R72L) and HLA-DRB1 H96E was stronger than any other associations ([Figure 4](#)).
174 Furthermore, associations between epitope variants and HLA amino acids were on average
175 stronger than associations between variants that were not part of any annotated T cell epitope

176 and HLA amino acids ($p = 0.015$) ([Figure 4](#)). A similar trend was observed for HLA 4- digit alleles
177 ($p = 0.003$).

178 Discussion

179 We describe the first TB genome-to-genome (G2G) study conducted in an African population. We
180 identified two pairs of associated genetic loci that were not identified in previous G2G studies or
181 GxG interaction studies³⁴⁻³⁸. In addition, we show that although variations in *M.tb* epitopes are
182 rare, certain *M.tb* epitope variants are associated with human HLA variation, a sign of probable
183 immune selective pressure.

184

185 The first G2G association we identified was between the *M.tb* variant Rv2348c I101M and the
186 human SNP rs12151990. rs12151990 is an intronic variant that affects the mRNA expression of
187 *PRDM15* in some tissues. A previous study has shown that *PRDM15* is co-expressed with
188 *KCNJ15* - a pro-apoptotic gene that is differentially acetylated when comparing blood monocytes
189 and granulocytes derived from *M.tb* infected patients and healthy controls³⁹. This suggests that
190 *PRDM15* could be involved in regulation of apoptosis, an important defense mechanism that
191 human macrophages employ for *M.tb* clearance^{40,41}. The function of Rv2348c is unknown, other
192 than that it is an antigen that stimulates T cell-mediated responses^{42,43}. Since the induction of
193 apoptosis in infected macrophages facilitates the presentation of antigens by dendritic cells and
194 subsequent activation of T cell-mediated responses⁴⁴, it is possible that Rv2348c and *PRDM15*
195 jointly determine disease prognosis due to such a process. Alternatively, it is possible that
196 Rv2348c might be directly involved in mechanisms that *M.tb* employs to regulate host cell
197 apoptosis⁴⁵.

198

199 The second G2G association we identified was between the *M.tb* variant FixA T67M and the
200 human SNP rs75769176. rs75769176 maps to an intergenic region and is close to both *FBXO15*
201 and *TIMM21*. Since colocalization analysis did not reveal any shared GWAS-eQTL signals for the
202 two genes, it is unclear whether the SNP impacts regulation of gene expression. It is important to
203 note that such analysis has limited power given that large tissue-specific eQTL studies were only
204 available for European populations, and it has been shown that a large proportion of eQTLs can
205 be population-specific⁴⁶. According to Reactome, *FBXO15* belongs to the “Class I MHC mediated
206 antigen processing and presentation” pathway suggesting its involvement in regulating adaptive
207 immunity. *TIMM21* belongs to the “Mitochondrial protein import” pathway. Again, the exact
208 mechanism for which FixA is involved in the interaction is unclear, given that limited information

209 suggests its involvement in *M.tb* metabolism and respiration. Nevertheless, the association of
210 rs75769176 with TB severity (TB score) in our cohort suggests that the variant plays a role in TB
211 pathogenesis.

212
213 Based on phylogenetics, we believe the two G2G-associated *M.tb* variants may have evolved
214 under different scenarios. FixA T67 did not display homoplasmy but strictly belonged to a clade
215 within L3.1. The G2G-associated human SNP (rs75769176) is associated with lower severity (TB
216 score). This could correspond to the scenario where individuals who carry the protective human
217 variant are more susceptible to developing active TB upon exposure to *M.tb* strains that carry the
218 FixA T67 variant compared to *M.tb* strains that do not. Similarly, most strains with Rv2348c I101M
219 belonged to a clade within L4.3. However, Rv2348c I101M also displayed homoplasmy, suggesting
220 that in some cases the mutation might have undergone intra-host selection during infection similar
221 to *de novo* drug resistance mutations⁴⁷. Specifically, Rv2348c I101M might confer a fitness
222 advantage for the bacteria within carriers of rs12151990 and have outcompeted the ancestral
223 strain that the carriers were originally infected with.

224
225 Corresponding to previous findings, we observed that *M.tb* T cell epitopes are hyperconserved⁴⁸:
226 in our study, only 15 out of 1,538 common *M.tb* variants overlapped with known T cell epitopes.
227 This suggests that epitope variation may not be the primary mechanism that *M.tb* employs to
228 evade T cell response and that alternative mechanisms exist⁴⁹. For example, one proposed
229 mechanism is that *M.tb* secretes “decoy” antigens that are non-essential for bacterial survival
230 during the initial stages of infection. As a result, T cells would be primed against these antigens
231 and subsequently fail to recognize the pathogen during the later stages of infection once the
232 “decoy” antigens are not expressed⁵⁰. For these antigens, variation provides no selective
233 advantage, given that the goal is to subvert rather than evade. Nevertheless, a small subset of
234 epitopes does display variation, and it is hypothesized that these epitopes may be in antigens that
235 employ a different evasion mechanism from those in hyper-conserved antigens⁵¹. A recent study
236 has identified region-specific variation in some *M.tb* epitopes, suggesting they may be under
237 selective pressure induced by human HLA variation⁵². Our results further support this hypothesis
238 by providing direct genetic evidence on the role of human HLA-DRB1 in inducing *M.tb* epitope
239 variations. Importantly, these epitopes could potentially be targeted by future T cell vaccines,
240 given that they are likely under immune selective pressure and thus more likely to be important
241 for host control of the bacteria.

242

243 The first limitation of applying a G2G approach to study host-pathogen interactions in TB is that
244 *M.tb* is highly clonal and many *M.tb* variants are strongly linked. For example, HLA-DRB1 H96E
245 is associated with four epitope variants and other non-annotated variants. While some of the
246 variants may be linked due to epistasis, many others may be linked simply due to evolutionary
247 history. Therefore, further experimental studies are needed to elucidate the causal variants. The
248 second limitation of our study is that only active TB cases were recruited. In the absence of a
249 comparable analysis in subclinical or latent TB groups, it is not possible to know if the observed
250 enrichment of certain combinations of host and pathogen variants (G2G associations) increases
251 the risk of developing active TB or results from host selection pressure during active
252 mycobacterial replication. A third limitation is that patients seek hospital care at different stages
253 of the disease, rendering clinical outcome measurements less reliable. We tried to correct for this
254 by adding cough duration as a covariate when comparing clinical data across patient groups.
255 Nevertheless, given its self-reported nature, cough duration might not be a perfect proxy for
256 disease length. Finally, given that the bacterial isolates were cultured prior to sequencing, the full
257 intra-host diversity of bacterial populations may not be fully recapitulated. Certain strains may
258 have growth advantages and thus may be overrepresented. This did not affect our analysis since
259 we focused on the consensus sequence and ignored sites with intra-host diversity. Indeed, when
260 considering variants that are fixed intra-host, sputum-based sequencing was found to be
261 concordant with culture-based sequencing⁵³. Nevertheless, direct deep-sequencing of sputum
262 samples could be useful to explore the intra-host diversity of *M.tb*, especially in the context of low-
263 frequency *de novo* mutations.

264 Conclusion

265 Through a genome-to-genome (G2G) analysis of TB patients from Tanzania, we identified two
266 pairs of associated genetic loci that might play a role in TB pathogenesis. The association
267 between the *M.tb* variant Rv2348c I101M and the human SNP rs12151990 could be due to the
268 interplay between *M.tb* and host apoptosis responses, although this requires further validation.
269 The mechanism responsible for the association between the *M.tb* variant FixA T67M and the
270 human SNP rs75769176 is unknown. However, the human SNP was associated with TB severity,
271 indicating a potential impact on TB pathogenesis. In addition, we observed that a set of *M.tb*
272 epitope variants (EspK L39W, EsxB E68K, Mpt70 A21T, RimJ R72L) were associated with HLA-
273 DRB1 H96E. This shows that certain *M.tb* epitope mutations are under selective pressure induced
274 by human HLA variation, confirming a role for T cell immunity in *M.tb* control. Our results highlight
275 genomic loci involved in the host-pathogen battle during chronic infection with *M.tb*. A deeper

276 understanding of the mechanisms at play could be instrumental in developing new therapeutic
277 strategies against TB.

278 Methods

279 Recruitment and sample collection

280 This study included active TB patients (sputum smear-positive and GeneXpert-positive) recruited
281 at the Temeke District Hospital in Dar es Salaam, Tanzania, as part of a prospective study that
282 ran between November 2013 and June 2022 (TB-DAR cohort). Ethical approval for the TB-DAR
283 cohort has been obtained from the Ethikkommission Nordwest- und Zentralschweiz (EKNZ UBE-
284 15/42), the Ifakara Health Institute—Institutional Review Board Board (IHI/IRB/EXT/No: 24–2020)
285 and the National Institute for Medical Research in Tanzania—Medical Research Coordinating
286 Committee (NIMR/HQ/R.8c/Vol.I/1622). A written informed consent has been obtained from every
287 patient who has been recruited into the TB-DAR cohort.

288
289 Clinical information including TB score, bacterial load (GeneXpert Ct-value), X-ray score, and HIV
290 status was collected as part of the study. TB-score is a symptoms-based score (with a maximum
291 of 12 points) adapted from Wejse et al⁵⁴. A point was added for the presence of each of the
292 following signs or symptoms: cough, hemoptysis, dyspnea, chest pain, night sweat, anemia,
293 abnormal auscultation, body temperature above 37°C, BMI below 18, BMI below 16, mid-upper
294 arm circumference (MUAC) below 220, MUAC below 200. Bacterial load was measured using the
295 GeneXpert Ct-value from real-time polymerase chain reaction. X-ray score is based on the Ralph
296 score⁵⁵ and assigned by two independent radiologists for patients with high-quality chest X-ray
297 images. Both Ct-value and X-ray score were transformed using rank-based inverse normal
298 transformation⁵⁶. Demographic information including age, sex, and self-reported smoking status
299 was also collected. Patients recruited twice due to relapse or reinfection were excluded.

300
301 In total, we generated either human or bacterial genomic data for 1,906 patients. For 1,471
302 patients, bacterial genomic data were generated. For 1,444 patients, human genomic data were
303 generated (98 based on WGS, 1,384 based on genotyping, and 30 based on both). This study
304 focused on 1,000 patients where high-quality human and bacterial genomic data were both
305 available after quality-based filtering.

306 Human sequencing

307
308 WGS was performed at the Health 2030 Genome Center in Geneva, Switzerland. The Illumina
309 NovaSeq 6000 sequencer was used, starting with 1 µg of whole blood genomic DNA. Illumina
310 TruSeq DNA PCR-Free reagents were used for library preparation. The 150nt paired-end
311 sequencing configuration was applied. An average coverage of 30X was achieved for each
312 genome.

313
314 The BWA⁵⁷ aligner (v0.7.17) was used to align sequencing reads to the GRCh38 (GCA
315 000001405.15) reference genome. Duplicate reads were marked using the markduplicates
316 module of Picard⁵⁸ (v2.8.14). For variant calling, GATK best practices (Germline short variant
317 discovery) were followed. Base quality score recalibration was applied using the GATK. Variants
318 were first called individually per sample. Then, samples with less than 5X coverage were excluded
319 and variants were jointly called. A Quality Score Recalibration (VQSR) based filter was applied
320 (truth sensitivity of 99.7 and excess heterozygosity of 54.69). Samples with a high rate of missing
321 genotype calls (>50%) were also excluded.

322 Human genotyping and imputation

323 Genotyping was performed by the iGE3 Genomics platform at the University of Geneva in
324 Geneva, Switzerland. Illumina Infinium H3Africa genotyping microarrays (Version 2;
325 <https://chipinfo.h3abionet.org>) with custom SNP add-ons (Tanzanian-specific SNPs, as described
326 by Xu et al.⁵⁹) were used. The Illumina GenomeStudio software (v2.0.5;
327 https://support.illumina.com/array/array_software/genomestudio/downloads.html) was used to
328 analyze raw microarray data. Low-quality or poorly clustered probes were excluded based on
329 filters suggested by Illumina for human studies ([Supplementary Table 7](#)). Low-quality samples
330 (call rate <0.97) were also excluded. Files were converted to PLINK format using the
331 GenomeStudio PLINK Input Report Plug-in(v2.1.4) and to VCF using PLINK(v1.9)⁶⁰.

332
333 Imputation was performed using two reference panels: 1) The African Genome Resources
334 (AFGR) reference panel (<https://www.apcdr.org/>) and 2) An internal Tanzanian reference panel
335 based on the 118 WGS samples from this study. For the AFGR reference panel, the sanger
336 imputation server (<https://imputation.sanger.ac.uk/>) was used with EAGLE2⁶¹ for phasing and
337 Positional Burrows-Wheeler transform (PBWT)⁶² for imputation. For the internal reference panel,
338 SHAPEIT4 was used for phasing and Minimac3 was used for imputation. For each imputed SNP,

339 the genotype call was based on the reference panel that yielded the highest imputation quality
340 score. Poorly imputed SNPs (INFO < 0.8) were excluded.

341 Combining genotyped and whole-genome sequenced samples

342 To confirm the accuracy of our imputation approach, for samples that were both genotyped and
343 whole-genome sequenced, genetic concordance between genotypes derived from the two
344 methods was calculated. Across the 27 samples that passed quality control filters on both
345 platforms, an average of 98.995% of SNPs were concordant.

346
347 Whole-genome sequenced and genotyped samples (post-imputation) were merged by extracting
348 SNPs common to both methods with overall missingness of less than 10%. The merging was
349 completed using bcftools (v1.15). SNPs that deviate from Hardy-Weinberg equilibrium ($P < 5e-5$)
350 were excluded, either before or after merging. SNPs with low minor allele frequency (MAF < 0.05)
351 were also excluded after merging. Genetic principal components (PCs) were calculated using
352 PLINK(v1.9) after LD pruning and exclusion of long-range LD regions. Genotype-based principal
353 component analysis (PCA) suggests that there were no systematic batch effects due to platform
354 differences ([Supplementary Figure 9A](#)).

355
356 To identify TB-DAR samples that are genetic outliers, the TB-DAR samples were merged with
357 1000 genomes samples and PCA was completed ([Supplementary Figure 9B](#)). A K-Nearest
358 Neighbour (KNN) model was trained on the 1000 genomes samples to assign ancestry to four
359 components (European, East Asian, South Asian, African) based on the top 2 PCs. Only one of
360 the TB-DAR samples was not clustered with the African population and was not assigned to the
361 African population by the KNN. This sample was excluded.

362
363 For pairs of relatives up to 2nd degree (N = 24), one of the relatives chosen randomly was
364 excluded. In the scenario of trios, the child was excluded. Most likely due to recordkeeping error,
365 a small number of samples (N = 11) was excluded due to discordance between reported and
366 genetically inferred sex.

367 Bacterial sequencing

368 Bacterial sequencing

369 Sputum samples from patients were decontaminated to kill bacteria other than *M.tb* and
370 centrifuged to obtain a sputum pellet. The sputum pellet was then inoculated on Löwenstein-

371 Jensen solid media, followed by DNA extraction using the CTAB method and whole-genome
372 sequencing. Culturing was completed at the TB laboratory of the Ifakara Health Institute in
373 Bagamoyo, Tanzania. DNA extraction of bacterial isolates was conducted in Switzerland (before
374 2017) or in Bagamoyo (after 2017). Sequencing was conducted at the Department of Biosystems
375 Science and Engineering of ETH Zurich, Basel (DBSSE) using Illumina HiSeq 2500 or NovaSeq
376 technology. The bacterial WGS data is published under bioproject PRJEB49562.

377 Bacterial variant calling

378 The pipeline described by Menardo et al.⁶³ was used to process bacterial sequencing reads. In
379 brief, Trimmomatic⁶⁴ (v1.2) was used to trim adapters and low-quality bases. BWA⁵⁷ (v 0.7.13)
380 was used to align reads to the reconstructed ancestral sequence of the MTBC⁴⁸. Duplicate reads
381 were excluded using MarkDuplicates module of Picard⁵⁸ (v2.9.1), and reads with low-quality
382 alignments (>7 mismatches per 100 bp) were excluded using Pysam(v0.9.0).

383

384 For variant calling, VarScan⁶⁵ (v2.4.1) mpileup2snp along with SAMtools⁶⁶ (v1.2) mpileup was
385 used. The following VarScan filters were applied: a minimum depth of 7 at a site to make a call (-
386 min-coverage 7), a minimum of 5 supporting reads at a position to call variants (--min-reads2 5),
387 a minimum base quality of 20 at a position to count a read (--min-avg-qual 20), a minimum variant
388 allele frequency threshold of 0.1 (--min-var-freq 0.1), a minimum frequency of 0.9 to call
389 homozygote (--min-freq-for-hom 0.9), and to ignore variants with >90% support on one strand (--
390 strand-filter 1). Non-homozygous mixed calls (called 0|1 or 1|0 heterozygous by varscan) were
391 excluded as they could be due to sequencing errors. Variants in repetitive regions (e.g. PE, PPE,
392 and PGRS genes or phages) were also excluded⁶⁷. A whole-genome FASTA file was then created
393 from the generated VCF file. Samples with a sequencing coverage lower than 15 and those with
394 features suggesting mixed infections (more heterozygous variants than homozygous variants or
395 more than 1000 heterozygous variants or a mix of lineages or sublineages) were excluded from
396 downstream analysis. Lineages and sublineages were defined based on the SNP classification
397 by Steiner et al⁶⁸ and Coll et al⁶⁹ respectively. Genomes were also excluded when multiple
398 sequencing runs of the same DNA samples resulted in different lineage or sublineage
399 assignments.

400

401 To annotate variants, SnpEff⁷⁰ (v5.0c) was used. Non-synonymous variants were extracted using
402 the SnpSift module of SnpEff. The resulting VCF files were merged, and a table (0 for absence,

403 1 for presence, and NA for insufficient coverage) of all amino-acid variants was created using R
404 (v4.4.1).

405 Phylogenetic analysis

406
407 Using Python v2.7.11 we compiled multiple sequence alignments from the FASTA files and
408 extracted the variable positions with less than or equal to 10% missing data. Phylogenetic trees
409 were constructed using RAxML v8.2.17⁷¹ with a general time-reversible model of sequence
410 evolution (options -m GTRCAT -v) with a *M.canettii* (SAMN00102920) or a L6 genome
411 (SAMEA5366648) as the outgroup. Graphical visualization was created using the ggtree⁷²
412 package in R (v4.4.1).

413
414 To understand the prevalence of *M.tb* variants and whether they were under positive selection in
415 a worldwide context, we leveraged the global reference set compiled by Zwyer et al³³ containing
416 11,818 genomes using BCFtools v1.15⁷³. To formally test for positive selection, PAML v4.9⁷⁴ was
417 run for subsets for each lineage consisting of 300 genomes randomly chosen. As an input we
418 used phylogenetic trees constructed with RAxML v8.2.11 in addition to the gene alignments in
419 order to estimate the branch lengths of the tree based on the M0 codon model in PAML. We then
420 fitted two models to the tree and gene alignment, one modeling nearly neutral (M1a) evolution
421 and the other modeling positive selection (M2a). We calculated p-values based on likelihood ratio
422 tests of the two aforementioned models⁷⁵. We reported the per-site posterior probability based on
423 the Bayes empirical Bayes (BEB) method.

424 Genome-to-genome association study

425 To achieve sufficient statistical power, the G2G study was restricted to common human and *M.tb*
426 variants. Variant frequency thresholds were decided based on *a priori* power calculations
427 ([Supplementary Figure 1](#)), conducted using genpwr package in R⁷⁶ assuming an additive true and
428 test model. The alpha level was selected based on the genome-wide threshold (5e-8) corrected
429 for the number of independent *M.tb* variants tested at each pathogen MAF threshold. On the
430 human side, only variants with MAF > 0.05 were included. This resulted in the inclusion of
431 6,603,291 human variants. On the bacterial side, only non-synonymous variants with MAF > 0.015
432 were included. To avoid spurious associations due to stratification, *M.tb* variants previously
433 established as phylogenetic lineages or major sublineages⁶⁷⁻⁶⁹ were excluded. Lineage markers
434 found in our dataset, defined as variants with high overall frequency (MAF > 0.015) but low

435 frequency (MAF<0.015) within every lineage, were also excluded. After filtering, 1538 *M.tb* amino-
436 acid variants were included.

437

438 For each *M.tb* variant, a case-control GWAS was conducted. Specifically, given *M.tb* variant j ,
439 human variant i , and K covariates, the model was formulated as:

440

$$y_j \sim \beta_{ij}G_i + \sum_k \alpha_k X_k$$

441

442 where y_j represents the phenotype vector encoding the presence of *M.tb* variant j , G_i represents
443 the genotype dosage vector of human variant i , and X_k represents the covariate vector for
444 covariate k . Sex and the top three human genetic PCs were included as covariates. The top three
445 PCs were included given that they are among the PCs that capture the most variance and
446 cumulatively capture 29% of the variance in the data ([Supplementary Figure 10](#)). The GWAS
447 were conducted using PLINK(v1.9), with a logistic regression model assuming an additive effect.
448 We used Bonferroni correction to adjust for multiple hypothesis testing. To avoid double-counting
449 *M.tb* variants that are in perfect correlation ($r^2=1$), the number of independent tests was derived
450 by adding the number of sets of *M.tb* variants that are in perfect correlation (N=52) to the number
451 of *M.tb* variants that are not in perfect correlation with any other variant (N=436). The genome-
452 wide significance threshold ($5e-8$) was divided by the number of independent tests (488) to derive
453 a final threshold of $1.02e-10$.

454 eQTL colocalization study

455 eQTL summary statistics from 29 studies were downloaded from the eQTL catalogue⁷⁷. Liftover
456 from hg19 to hg38 was completed using the liftOver function of the rtracklayer package (v1.52.1)
457 in R(v4.4.1). Bayesian colocalization analysis implemented by coloc.abf⁷⁸ function of the coloc
458 package(v5.1.0) in R(v4.4.1) was used to conduct colocalization analysis. Default prior
459 probabilities were used. The posterior probabilities for H4 (hypothesis with a single shared casual
460 variant for both traits) were reported. We also checked whether significant eQTLs ($p < 5e-8$) were
461 identified for our genes of interest (*PRDM15*, *TIMM21*, *FBXO15*) in the GENOA cohort⁴⁶, which
462 was based on whole-blood samples from African American individuals. No significant eQTLs were
463 identified, hence we did not include this dataset in the colocalization analyses.

464 *M.tb* epitope variants and human HLA variation

465 Experimentally validated *M.tb* T-cell epitopes were extracted (October 2022) from the Immune
466 Epitope Database (IEDB). The search query was restricted to the following epitopes: validated by

467 T cell or MHC ligand assays, source organism = *M. tuberculosis* or *M. tuberculosis* H37Rv, and
468 host organism = human. 3,499 epitopes matching the search query were identified. UniProt IDs
469 from query results were mapped to protein names using the UniProt ID mapping tool
470 (<https://www.uniprot.org/id-mapping>). All common *M.tb* amino acid variants (MAF > 0.015) were
471 then mapped to epitopes based on their positions. Human HLA amino acid variants and 4-digit
472 HLA alleles were imputed using the TopMed Imputation Server
473 (<https://imputation.biodatacatalyst.nih.gov/#/>) based on the TopMed reference panel⁷⁹.

474

475 We tested for associations between *M.tb* epitope variants and human HLA amino acid variants or
476 alleles. As with the G2G study, sex and the top three human genetic PCs were included as
477 covariates. The logistic regression-based association model implemented in PLINK(v1.9) was
478 also used but assuming a dominant effect. *M.tb* epitope variants in perfect correlation ($r^2=1$) were
479 grouped, resulting in 10 variant sets. We tested the association of 10 *M.tb* variant sets against
480 293 HLA amino acid positions and 103 4-digit HLA alleles. A Bonferroni corrected threshold of
481 $1.3e-5$ was thus applied, based on an alpha level of 0.05 and the 3960 tests we conducted.

482 Pathway-to-pathway association study

483 Human gene regulatory pathways were based on the 6449 curated gene-sets from Human
484 Molecular Signatures Database (MSigDB, v7.0)⁸⁰. The curated gene sets were generated from
485 either gene expression profiles of perturbation experiments or from canonical pathway databases
486 including Reactome, KEGG and others. For each GWAS conducted in the G2G study, gene
487 scores were constructed by mapping variants to genes by proximity using PASCAL⁸¹. Given that
488 variants in close proximity are likely to be in strong LD, gene scores of neighboring genes are
489 often correlated and can be part of the same pathway. To address this, PASCAL constructs a
490 fusion score based on LD patterns. The 1000 Genomes African population was used as the
491 reference LD database for this.

492

493 *M.tb* gene regulatory pathways from two data sources were obtained. 547 pathways were derived
494 from a *M.tb* regulatory network constructed using gene expression profiles of perturbation
495 experiments(<http://networks.systemsbiology.net/mtb/>)⁸². 192 pathways were derived from the
496 stringDB⁸³ database using a score-cutoff of 400 and recursive MCL clustering with inflation
497 parameter optimized to obtain the maximum number of clusters between the size of 5 - 50. *M.tb*
498 variants were directly mapped to *M.tb* genes based on position.

499

500 To identify the enrichment of edges between pairs of human-bacterial pathways, p-values
501 between human and *M.tb* genes were binarized based on False Discovery Rate (FDR) < 0.15. A
502 density measure was calculated for each pair of human and *M.tb* pathways. Given a human
503 pathway with gene-set $S_i^{Human} = \{g_1, g_2 \dots\}$, a *M.tb* pathway with gene-set $S_j^{Mtb} = \{g_1, g_2 \dots\}$,
504 and a gene-gene binary association matrix X, the density was defined as:

$$D_{ij} = \frac{\sum_{p \in S_i^{Human}} \sum_{q \in S_j^{Mtb}} X_{pq}}{|S_j^{Mtb}| |S_i^{Human}|}$$

505
506 where D_{ij} represents the density between human pathway i and *M.tb* pathway j .

507
508 To calculate an empirical p-value, gene-pathway memberships were permuted 1000 times while
509 preserving the size of the pathways. The distribution of density values obtained from the
510 permutations was used to obtain a null distribution.

511 Acknowledgement

512 This work was supported by the Swiss National Science Foundation (Grants: CRSII5-177163 and
513 310030-188888) and the European Research Council (Grant: 883582). We thank E. Ristorcelli
514 (CHUV, Lausanne, Switzerland) for sample preparation and DNA extraction; we thank K.
515 Harshman, I. Bartha, C. Howald, and D. Lamparter (Health 2030 Genome Center, Geneva,
516 Switzerland) for human WGS support; we thank M. Docquier (iGE3 Genomics Platform, Geneva,
517 Switzerland) for human genotyping support.

518 Conflicts of Interest

519 O.N. is now an employee of SUN bioscience SA. S.R. is now an employee of Novartis AG.
520

521 Data Availability

522 Summary statistics and software code have been published on github:
523 <https://github.com/zmx21/G2G-TB>. The bacterial WGS data has been published under bioproject
524 PRJEB49562. Human genotyping and WGS data are deposited on the European Genome-
525 phenome Archive (EGA) under EGAS00001007216 and EGAS00001005850 respectively.

Bibliography

1. Gopalaswamy, R., Dusthacker, V. N. A., Kannayan, S. & Subbian, S. Extrapulmonary Tuberculosis—An Update on the Diagnosis, Treatment and Drug Resistance. *JoR* **1**, 141–164 (2021).
2. Global tuberculosis report 2022. <https://www.who.int/publications/i/item/9789240061729>.
3. Global tuberculosis report 2019. <https://www.who.int/publications/i/item/9789241565714>.
4. Lin, P. L. & Flynn, J. L. The End of the Binary Era: Revisiting the Spectrum of Tuberculosis. *The Journal of Immunology* **201**, 2541–2548 (2018).
5. Pai, M. *et al.* Tuberculosis. *Nat Rev Dis Primers* **2**, 16076 (2016).
6. Barry, C. E. *et al.* The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. *Nat Rev Microbiol* **7**, 845–855 (2009).
7. Vynnycky, E. & Fine, P. E. M. The natural history of tuberculosis: the implications of age-dependent risks of disease and the role of reinfection. *Epidemiol. Infect.* **119**, 183–201 (1997).
8. Bruchfeld, J., Correia-Neves, M. & Källenius, G. Tuberculosis and HIV Coinfection. *Cold Spring Harb Perspect Med* **5**, a017871 (2015).
9. Hayashi, S. & Chandramohan, D. Risk of active tuberculosis among people with diabetes mellitus: systematic review and meta-analysis. *Trop Med Int Health* **23**, 1058–1070 (2018).
10. Imtiaz, S. *et al.* Alcohol consumption as a risk factor for tuberculosis: meta-analyses and burden of disease. *Eur Respir J* **50**, 1700216 (2017).
11. Lönnroth, K. *et al.* Tuberculosis control and elimination 2010–50: cure, care, and social development. *The Lancet* **375**, 1814–1829 (2010).
12. Familial Susceptibility to Tuberculosis: Its Importance as a Public Health Problem. *JAMA* **128**, 317 (1945).
13. Kallmann, F. J. & Reisner, D. Twin Studies on the Significance of Genetic Factors in Tuberculosis. *Am Rev Tuberc* **47**, 549–574 (1943).
14. Thye, T. *et al.* Common variants at 11p13 are associated with susceptibility to tuberculosis. *Nat Genet* **44**, 257–259 (2012).
15. Chimusa, E. R. *et al.* Genome-wide association study of ancestry-specific TB risk in the South African Coloured population. *Hum Mol Genet* **23**, 796–809 (2014).
16. African TB Genetics Consortium *et al.* Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2. *Nat Genet* **42**, 739–741 (2010).
17. Curtis, J. *et al.* Susceptibility to tuberculosis is associated with variants in the ASAP1 gene encoding a regulator of dendritic cell migration. *Nat Genet* **47**, 523–527 (2015).
18. Zheng, R. *et al.* Genome-wide association study identifies two risk loci for tuberculosis in Han Chinese. *Nat Commun* **9**, 4072 (2018).
19. Luo, Y. *et al.* Early progression to active tuberculosis is a highly heritable trait driven by 3q23 in Peruvians. *Nat Commun* **10**, 3765 (2019).
20. Kerner, G. *et al.* Homozygosity for *TYK2* P1104A underlies tuberculosis in about 1% of patients in a cohort of European ancestry. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 10430–10434 (2019).
21. Quistbert, J. *et al.* Genome-wide association study of resistance to Mycobacterium tuberculosis infection identifies a locus at 10q26.2 in three distinct populations. *PLoS Genet* **17**, e1009392 (2021).
22. Qi, H. *et al.* Discovery of susceptibility loci associated with tuberculosis in Han Chinese. *Human Molecular Genetics* **26**, 4752–4763 (2017).
23. Sobota, R. S. *et al.* A Locus at 5q33.3 Confers Resistance to Tuberculosis in Highly Susceptible Individuals. *The American Journal of Human Genetics* **98**, 514–524 (2016).

24. Naranbhai, V. The Role of Host Genetics (and Genomics) in Tuberculosis. *Microbiol Spectr* **4**, 4.5.26 (2016).
25. Napier, G. *et al.* Robust barcoding and identification of *Mycobacterium tuberculosis* lineages for epidemiological and clinical studies. *Genome Med* **12**, 114 (2020).
26. Brites, D. & Gagneux, S. The Nature and Evolution of Genomic Diversity in the *Mycobacterium tuberculosis* Complex. in *Strain Variation in the Mycobacterium tuberculosis Complex: Its Role in Biology, Epidemiology and Control* (ed. Gagneux, S.) vol. 1019 1–26 (Springer International Publishing, 2017).
27. Gagneux, S. & Stritt, C. *How do monomorphic bacteria evolve? The Mycobacterium tuberculosis complex and the awkward population genetics of extreme clonality.* <https://ecoevorxiv.org/repository/view/4829/> (2022) doi:10.32942/X2GW2P.
28. Coscolla, M. & Gagneux, S. Consequences of genomic diversity in *Mycobacterium tuberculosis*. *Seminars in Immunology* **26**, 431–444 (2014).
29. Gagneux, S. & Small, P. M. Global phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development. *The Lancet Infectious Diseases* **7**, 328–337 (2007).
30. Brites, D. & Gagneux, S. Co-evolution of *Mycobacterium tuberculosis* and *Homo sapiens*. *Immunol Rev* **264**, 6–24 (2015).
31. Gagneux, S. *et al.* Variable host–pathogen compatibility in *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 2869–2873 (2006).
32. Fellay, J. & Pedergrana, V. Exploring the interactions between the human and viral genomes. *Hum Genet* (2019) doi:10.1007/s00439-019-02089-3.
33. Zwyer, M. *et al.* Back-to-Africa introductions of *Mycobacterium tuberculosis* as the main cause of tuberculosis in Dar es Salaam, Tanzania. *PLoS Pathog* **19**, e1010893 (2023).
34. Luo, Y. *et al.* A FLOT1 host regulatory allele is associated with a recently expanded Mtb clade in patients with tuberculosis. <http://medrxiv.org/lookup/doi/10.1101/2022.02.07.22270622> (2022) doi:10.1101/2022.02.07.22270622.
35. Phelan, J. *et al.* Genome-wide host-pathogen analyses reveal genetic interaction points in tuberculosis disease. *Nat Commun* **14**, 549 (2023).
36. McHenry, M. L. *et al.* Interaction between host genes and *Mycobacterium tuberculosis* lineage can affect tuberculosis severity: Evidence for coevolution? *PLoS Genet* **16**, e1008728 (2020).
37. McHenry, M. L. *et al.* Interaction between M. tuberculosis Lineage and Human Genetic Variants Reveals Novel Pathway Associations with Severity of TB. *Pathogens* **10**, 1487 (2021).
38. Müller, S. J. *et al.* A multi-phenotype genome-wide association study of clades causing tuberculosis in a Ghanaian- and South African cohort. *Genomics* **113**, 1802–1815 (2021).
39. del Rosario, R. C. H. *et al.* Histone acetylome-wide associations in immune cells from individuals with active *Mycobacterium tuberculosis* infection. *Nat Microbiol* **7**, 312–326 (2022).
40. Behar, S. M. *et al.* Apoptosis is an innate defense function of macrophages against *Mycobacterium tuberculosis*. *Mucosal Immunology* **4**, 279–287 (2011).
41. Stutz, M. D. *et al.* Macrophage and neutrophil death programs differentially confer resistance to tuberculosis. *Immunity* **54**, 1758–1771.e7 (2021).
42. Villar-Hernández, R. *et al.* Diagnostic benefits of adding EspC, EspF and Rv2348-B to the QuantiFERON Gold In-tube antigen combination. *Sci Rep* **10**, 13234 (2020).
43. Mpande, C. A. M. *et al.* *Mycobacterium tuberculosis*-Specific T Cell Functional, Memory, and Activation Profiles in QuantiFERON-Reverters Are Consistent With Controlled Infection. *Front. Immunol.* **12**, 712480 (2021).
44. Schaible, U. E. *et al.* Apoptosis facilitates antigen presentation to T lymphocytes through

- MHC-I and CD1 in tuberculosis. *Nat Med* **9**, 1039–1046 (2003).
45. Briken, V. & Miller, J. L. Living on the edge: inhibition of host cell apoptosis by *Mycobacterium tuberculosis*. *Future Microbiology* **3**, 415–422 (2008).
 46. Shang, L. *et al.* Genetic Architecture of Gene Expression in European and African Americans: An eQTL Mapping Study in GENOA. *Am J Hum Genet* **106**, 496–512 (2020).
 47. Vázquez-Chacón, C. A. *et al.* Intra-host genetic population diversity: Role in emergence and persistence of drug resistance among *Mycobacterium tuberculosis* complex minor variants. *Infection, Genetics and Evolution* **101**, 105288 (2022).
 48. Comas, I. *et al.* Human T cell epitopes of *Mycobacterium tuberculosis* are evolutionarily hyperconserved. *Nat Genet* **42**, 498–503 (2010).
 49. Baena, A. & Porcelli, S. A. Evasion and subversion of antigen presentation by *Mycobacterium tuberculosis*. *Tissue Antigens* **74**, 189–204 (2009).
 50. Rogerson, B. J. *et al.* Expression levels of *Mycobacterium tuberculosis* antigen-encoding genes versus production levels of antigen-specific T cells during stationary level lung infection in mice. *Immunology* **118**, 195–201 (2006).
 51. Coscolla, M. *et al.* M. tuberculosis T Cell Epitope Analysis Reveals Paucity of Antigenic Variation and Identifies Rare Variable TB Antigens. *Cell Host & Microbe* **18**, 538–548 (2015).
 52. Ramaiah, A. *et al.* Evidence for Highly Variable, Region-Specific Patterns of T-Cell Epitope Mutations Accumulating in *Mycobacterium tuberculosis* Strains. *Front. Immunol.* **10**, 195 (2019).
 53. Goig, G. A. *et al.* Whole-genome sequencing of *Mycobacterium tuberculosis* directly from clinical samples for high-resolution genomic epidemiology and drug resistance surveillance: an observational study. *The Lancet Microbe* **1**, e175–e183 (2020).
 54. Wejse, C. *et al.* TBscore: Signs and symptoms from tuberculosis patients in a low-resource setting have predictive value and may be used to assess clinical course. *Scandinavian Journal of Infectious Diseases* **40**, 111–120 (2008).
 55. Ralph, A. P. *et al.* A simple, valid, numerical score for grading chest x-ray severity in adult smear-positive pulmonary tuberculosis. *Thorax* **65**, 863–869 (2010).
 56. McCaw, Z. R., Lane, J. M., Saxena, R., Redline, S. & Lin, X. Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics* **76**, 1262–1272 (2020).
 57. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
 58. Broad Institute. *Picard toolkit*. (Broad Institute, 2019).
 59. Xu, Z. M. *et al.* Using population-specific add-on polymorphisms to improve genotype imputation in underrepresented populations. *PLoS Comput Biol* **18**, e1009628 (2022).
 60. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
 61. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* **48**, 1443–1448 (2016).
 62. Durbin, R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics* **30**, 1266–1272 (2014).
 63. Menardo, F. *et al.* Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity. *BMC Bioinformatics* **19**, 164 (2018).
 64. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
 65. Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568–576 (2012).
 66. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).

67. Stucki, D. *et al.* Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. *Nat Genet* **48**, 1535–1543 (2016).
68. Steiner, A., Stucki, D., Coscolla, M., Borrell, S. & Gagneux, S. KvarQ: targeted and direct variant calling from fastq reads of bacterial genomes. *BMC Genomics* **15**, 881 (2014).
69. Coll, F. *et al.* A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nat Commun* **5**, 4812 (2014).
70. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w¹¹¹⁸; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
71. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
72. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* **8**, 28–36 (2017).
73. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
74. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* **24**, 1586–1591 (2007).
75. Zhang, J. Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. *Molecular Biology and Evolution* **22**, 2472–2479 (2005).
76. Moore, C. M., Jacobson, S. A. & Fingerlin, T. E. Power and Sample Size Calculations for Genetic Association Studies in the Presence of Genetic Model Misspecification. *Hum Hered* **84**, 256–271 (2019).
77. Kerimov, N. *et al.* A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat Genet* **53**, 1290–1299 (2021).
78. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet* **10**, e1004383 (2014).
79. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
80. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
81. Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z. & Bergmann, S. Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLoS Comput Biol* **12**, e1004714 (2016).
82. Turkarslan, S. *et al.* A comprehensive map of genome-wide gene regulation in Mycobacterium tuberculosis. *Sci Data* **2**, 150010 (2015).
83. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research* **47**, D607–D613 (2019).

Tables and Figures

Table 1

Demographic and clinical characteristics of the cohort. L1-L4 indicates bacterial lineages. P value based on ANOVA test for quantitative variables and chi-squared test for discrete variables.

	L1 (N=155)	L2 (N=82)	L3 (N=431)	L4 (N=332)	Overall (N=1000)	p
Sex						
female	45 (29 %)	21 (26 %)	122 (28 %)	111 (33 %)	299 (30 %)	0.51
male	110 (71 %)	61 (74 %)	309 (72 %)	221 (67 %)	701 (70 %)	
Age						
Mean (SD)	37 (± 11)	33 (± 9.9)	34 (± 9.8)	34 (± 10)	34 (± 10)	0.04
BMI						
Mean (SD)	19 (± 3.3)	19 (± 2.8)	19 (± 3.1)	19 (± 3.3)	19 (± 3.2)	0.88
Smoker						
no	114 (74 %)	58 (71 %)	314 (73 %)	268 (81 %)	754 (75 %)	0.19
yes	41 (26 %)	23 (28 %)	114 (26 %)	64 (19 %)	242 (24 %)	
unknown	0 (0 %)	1 (1 %)	3 (1 %)	0 (0 %)	4 (0 %)	
HIV Status						
infected	30 (19 %)	10 (12 %)	72 (17 %)	50 (15 %)	162 (16 %)	0.65
negative	124 (80 %)	71 (87 %)	355 (82 %)	280 (84 %)	830 (83 %)	
Missing	1 (0.6%)	1 (1.2%)	4 (0.9%)	2 (0.6%)	8 (0.8%)	
TB Score						
Mean (SD)	5.2 (± 1.6)	4.8 (± 1.5)	5.1 (± 1.7)	5.0 (± 1.6)	5.0 (± 1.6)	0.43
Ct Value						
Mean (SD)	19 (± 4.7)	20 (± 4.9)	19 (± 4.5)	20 (± 4.7)	19 (± 4.6)	0.24
Missing	64 (41.3%)	35 (42.7%)	179 (41.5%)	151 (45.5%)	429 (42.9%)	
X-ray Score						
Mean (SD)	40 (± 27)	48 (± 27)	47 (± 29)	45 (± 28)	45 (± 28)	0.36
Missing	65 (41.9%)	37 (45.1%)	176 (40.8%)	144 (43.4%)	422 (42.2%)	

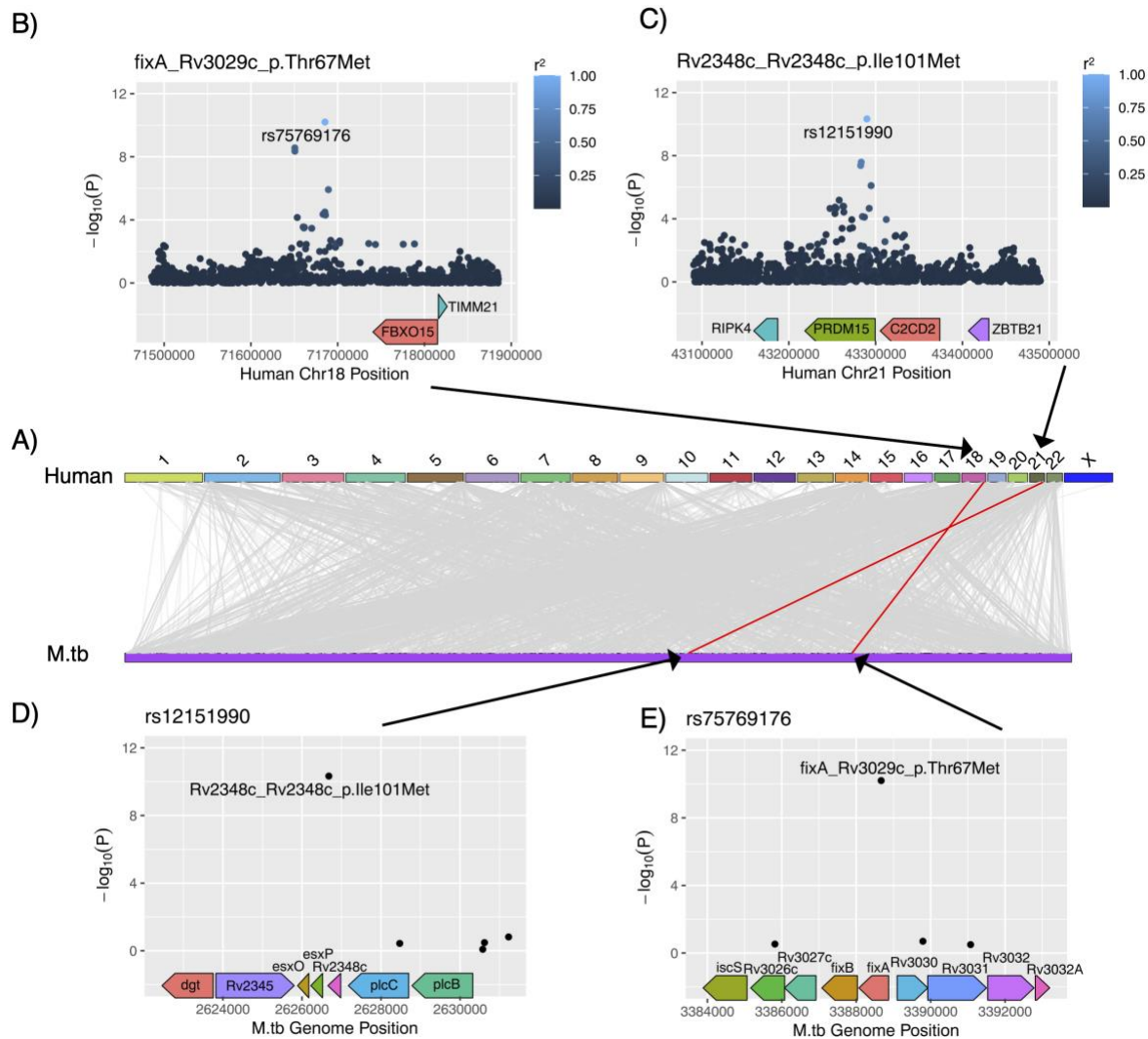


Figure 1

Summary of genome-to-genome (G2G) analysis results. A) Associations between human and *M.tb* variants, with upper half of the figure refers to human (GRCh37) chromosome coordinates and bottom half refers to *M.tb* (H37Rv) nucleotide positions. Grey lines indicate genome-wide significant ($p < 5e-8$) G2G associations and red lines indicates significant associations after Bonferroni correction ($p < 1.02e-10$) **B)** Manhattan plot of human genetic loci referring to the association between FixA T67M and rs75769176 **C)** Manhattan plot of human genetic loci referring to the association between Rv2348c I101M and rs12151990 **D)** Manhattan plot of bacterial genetic loci referring to the association between Rv2348c I101M and rs12151990 **E)** Manhattan plot of bacterial genetic loci referring to the association between FixA T67M and rs75769176

Lineage L4

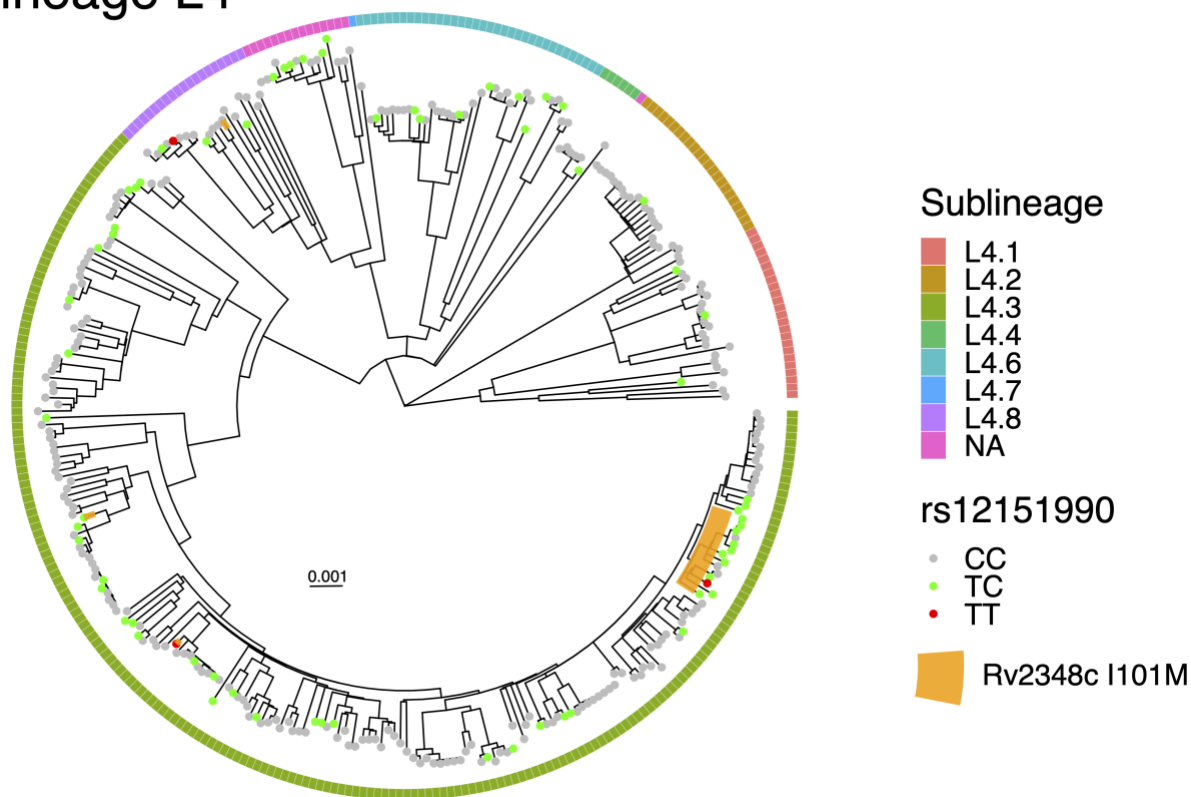


Figure 2

Phylogenetic tree of *M.tb* strains in Lineage 4. Colors of outermost squares refer to sublineages. Colors of circles on tree tips refer to whether the infected human carries the rs12151990 variant. Orange shade represents the clade (or in 3 instances, single strains) of *M.tb* that carry the Rv2348c I101M variant. The tree was rooted with a *M. canetti* strain as an outgroup. Scale bar represents substitutions per polymorphic site.

Lineage L3

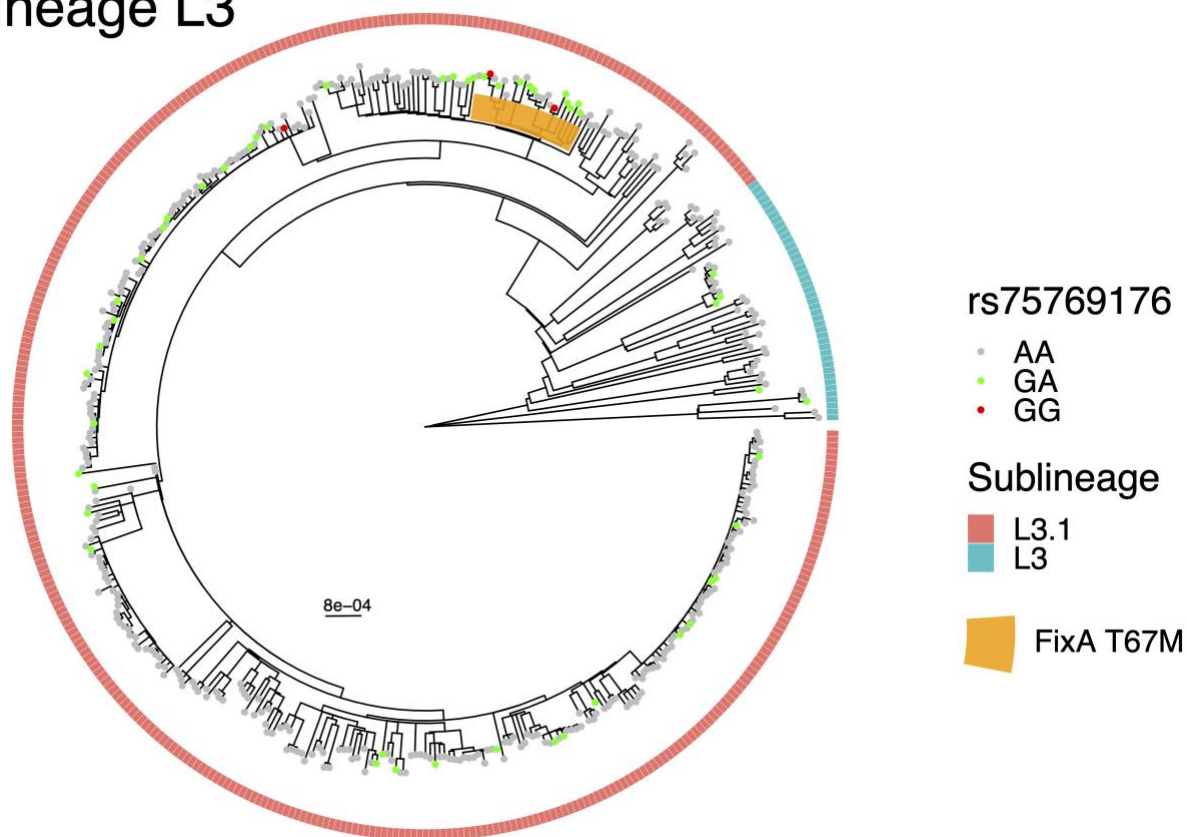


Figure 3

Phylogenetic tree of *M.tb* strains in Lineage L3. Colors of outermost squares refer to sublineages. Colors of circles on tree tips refer to whether the infected human carries the rs75769176 variant. Orange shade represents the clade of *M.tb* that carry the FixA T67M variant. The tree was rooted with a *M. canetti* strain as an outgroup. Scale bar represents substitutions per polymorphic site.

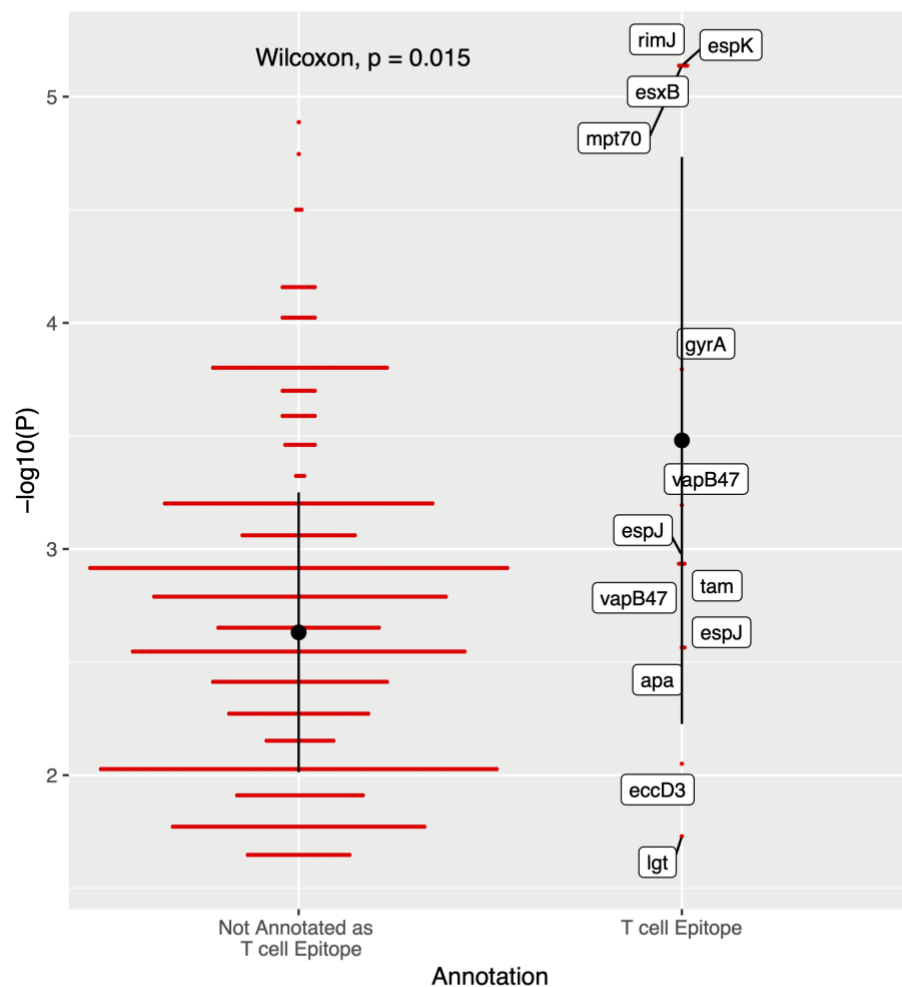


Figure 4

Associations between *M.tb* variants and HLA amino acid variants. For each *M.tb* variant, only the top association is displayed. Not annotated as T cell epitope: *M.tb* variants that are not part of any annotated T cell epitope; T cell epitope: *M.tb* variants that map to known T cell epitopes based on the IEDB database. Black dots and bars indicate mean and standard deviation within each annotation category.