In-host population dynamics of *M. tuberculosis* during treatment failure

Roger Vargas Jr[1,2*], Luca Freschi[2], Maximillian Marin[1,2], L. Elaine Epperson[3], Melissa Smith[4,5], Irina Oussenko[5], David Durbin[6], Michael Strong[3], Max Salfinger[7,8], and Maha Reda Farhat[2,9*]

[1] Department of Systems Biology, Harvard Medical School

[2] Department of Biomedical Informatics, Harvard Medical School

[3] Center for Genes, Environment, and Health, National Jewish Health

[4] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai

[5] Icahn Institute of Data Sciences and Genomics Technology

[6] Mycobacteriology Reference Laboratory, Advanced Diagnostic Laboratories, National Jewish Health

[7] College of Public Health, University of South Florida

[8] Morsani College of Medicine, University of South Florida

[9] Pulmonary and Critical Care Medicine, Massachusetts General Hospital

[*]Corresponding authors: roger_vargas@g.harvard.edu, Maha_Farhat@hms.harvard.edu

ABSTRACT

Background

Tuberculosis (TB) is a leading cause of death globally from an infectious agent. Understanding the population dynamics of TB's causative agent *Mycobacterium tuberculosis* (Mtb) in-host is vital for understanding the efficacy of antibiotic treatment. Here we use longitudinally collected clinical Mtb isolates that underwent Whole-Genome Sequencing (WGS) from the sputa of 307 subjects to investigate Mtb diversity during the course of active TB disease.

Methods and findings

We excluded cases suspected of reinfection or contamination to analyze data from 200 subjects, 167 of which met microbiological criteria for delayed culture conversion, treatment failure or relapse. Using technical and biological replicate samples, we defined an allele frequency threshold attributable to in-host evolution. Of the 167 subjects with unsuccessful treatment outcome, 27 (16%) developed new resistance mutations between sampling with 20/27 (74%) occurring in patients with pre-existing antibiotic resistance. Low abundance resistance variants at a purity of ≥19% in the first isolate predicts fixation of these variants in the subsequent sample with 27.0% sensitivity and 95.8% specificity. We identify significant in-host variation in seven genes associated with antibiotic resistance and twenty other genes, including metabolic genes and genes known to modulate host innate immunity by interacting with TLR2. We confirm Rv0095c, Rv1944c, *PPE18*, *PPE54* and *PPE60* to be under positive selection by assessing phylogenetic convergence across a global and genetically diverse independent sample of 20,352 isolates.

Conclusions

Our large sample provides a comprehensive picture of the mutational dynamics *in-host* during active TB disease. We demonstrate a framework to study temporal changes in Mtb population diversity using average depth WGS data. We show that minor variants can be used to inform antibiotic treatment regimens in patients with TB. Furthermore, we detect a signature of positive selection in-host, possibly stemming from innate immune pressure and informing our understanding of host-pathogen interactions.

46 **INTRODUCTION**

47 Tuberculosis (TB) and its causative pathogen *Mycobacterium tuberculosis* (Mtb) remain
48 a major public health threat (World Health Organization 2018). Yet the majority of individuals
49 exposed to Mtb clear or contain the infection, and only 5-10% of those infected develop active
50 TB disease at some point in their lifetime (Pai et al. 2016). While basic human immune
51 mechanisms to Mtb have been identified, attempts at effective vaccine development guided by
52 these mechanisms have repeatedly failed (Ernst 2018b). Global efforts that include scale up of
53 directly observed therapy have also been challenged by rising estimates of multidrug resistance.
54 Mtb is an obligate human pathogen (Gagneux 2018). Infection and disease involves a complex
55 human host-pathogen interaction that is both physically and temporally heterogeneous (Lin et al.
56 2014). Consequently all selective forces acting on Mtb will originate within the host, and the
57 study of temporal dynamics of this is likely to inform antibiotic treatment (Sun et al. 2012) and
58 rational vaccine design (Ernst 2018b).

59 At long timescales, signatures of positive selection associated with antibiotic resistance
60 have been characterized, but epitope regions appear to be under purifying selection (Farhat et al.
61 2013; Coscolla et al. 2015; Brites and Gagneux 2015; Comas et al. 2010) calling into question
62 how Mtb interacts with host adaptive immunity. Little is known about selection at short
63 timescales, such as within single infections. Drug pressure may select for resistance-conferring
64 mutations, thus an understanding of how the frequency of minor alleles changes longitudinally
65 can inform optimal drug treatment (Sun et al. 2012; Zhang et al. 2016; Didelot et al. 2016). A
66 recent study found treatment relapse to be strongly associated with bacterial factors (Colangeli et
67 al. 2018); therefore there is a need to better characterize these as predictors of treatment
68 response. Bacterial factors of interest include not only low frequency resistance variants but also

69 variants that may induce other phenotypes, such as drug tolerance or more effective immune
70 evasion (Ernst 2018a). To elucidate these processes, we aimed to study how genomic diversity
71 arises in-host in Mtb populations, employing a longitudinal sampling scheme from patients with
72 active TB disease.

73 The application of genome sequencing technologies to Mtb isolates cultured from clinical
74 samples has highlighted that infection consists of populations of Mtb bacteria rather than single
75 clones devoid of diversity (Marvig et al. 2015; Lieberman et al. 2011, 2014; Copin et al. 2016;
76 Didelot et al. 2016). Differences in observed allele frequencies captured using genome
77 sequencing (**Fig. 1A**) may represent a difference in the genetic composition of the infecting
78 population, commonly referred to as heterogeneity. Mtb population heterogeneity might be
79 present within a host because (1) the host is infected with multiple strains or is re-infected by a
80 new strain (consistent with mixed infection or re-infection) or (2) genetic diversity arises within
81 the Mtb population during infection due to selection or drift (Ford et al. 2012; Lieberman et al.
82 2016; Guerra-Assunção et al. 2014). However, non-uniform sampling (Trauner et al. 2017),
83 selection during the *in vitro* culture process (Trauner et al. 2017), laboratory contamination
84 (Goig et al. 2020; Wyllie et al. 2018), sequencing error and mapping error all represent examples
85 of experimental error that give rise to heterogeneity of low significance to host-pathogen
86 interactions.

87 Here, we present a framework to overcome these barriers and demonstrate the use of
88 longitudinally collected isolates, pooled sweeps of colonies cultured from sputa, to investigate
89 true in-host diversity with implications for Mtb treatment. We analyzed 614 paired longitudinal
90 isolates representing 307 subjects from eight studies (Casali et al. 2016; Walker et al. 2013;
91 Trauner et al. 2017; Guerra-Assunção et al. 2014; Bryant et al. 2013; Witney et al. 2017; Xu et

92 al. 2018). Many subjects, despite undergoing treatment, remained culture positive at two months
93 intervals or longer meeting microbiological criteria for delayed culture conversion, treatment
94 failure or relapse (**Table S1**). Our sample consisted of 167 subjects fulfilling these criteria, which
95 allowed us to overcome the small sample size problem present in prior studies. We find a high
96 turnover of low-frequency alleles in loci associated with antibiotic resistance but that mutant
97 alleles in these loci that rise to a frequency of 19% are predicted to fix in-host with a sensitivity
98 of 27.0% and specificity of 95.8%, a proof of concept that WGS can aide in predicting resistance
99 amplification. We show that changes in allele frequency are common among replicate isolates
100 and that changes in frequency of 70% are indicative of in-host evolution using archived MTB
101 isolates. We demonstrate that many loci involved the acquisition of antibiotic resistance and
102 modulation of innate host-immunity appear to be under positive selection.
103

104 **RESULTS**

105 **Identifying clonal Mtb populations in-host**

106 To isolate the *in vivo* clonal dynamics of Mtb during infection among the 307 subjects
107 with longitudinal samples collected during or after antibiotic treatment (**Table S1-S2**), we
108 excluded 32 subjects with isolate microbiological contamination at any time point (Goig et al.
109 2020), and 31 subjects with evidence for mixed infection with two or more Mtb lineages (Wyllie
110 et al. 2018) (**Fig. 1B** and **Fig. S1**). We also excluded 44 subjects with evidence for re-infection
111 with a different Mtb strain between the first and second time points, using a pairwise genetic
112 distance >7 fixed SNPs (fSNPs) (**Methods, Fig. 1C** and **Fig. S1**). We implemented WGS SNP
113 calling filters to minimize the likelihood of false positives and estimated the error rate of our
114 analysis pipeline using a control dataset of 82 isolate pairs (162 total) that were *in vitro* technical
115 or biological replicates (**Methods**, **Fig. 4** and **Fig. S1**). Of the 307 subjects, 200 had isolate pairs
116 that passed all filters (representative of clonal infection) (**Table S3**), with an estimated false
117 positive SNP rate of 0.0527 or less. The 400 isolates represented the five main Mtb lineages.

118

119 **In-host pathogen dynamics in antibiotic resistance loci**

120 Resistance mutations found at low frequencies in-host may indicate the impending
121 development of clinical resistance (Sun et al. 2012; Trauner et al. 2017; Zhang et al. 2016), To
122 investigate temporal dynamics related to antibiotic pressure, we identified non-synonymous and
123 intergenic SNPs within a set of 36 predetermined resistance loci associated with antibiotic
124 resistance (Farhat et al. 2013, 2016) (**Table S5**) that changed in allele frequency by at least 5%
125 (Sun et al. 2012) between the first and second sampling time point (**Methods**). We detected

126 1,939 such SNPs across our sample of 200 subjects (**Fig. 2B**), 1,774 were non-synonymous, 91
127 were intergenic, and 74 occurred within the *rrs* region (**Table S6**).
128 We searched for signs of selection by identifying clonal interference, or evidence of
129 competition between strains with different drug resistance mutations (Trauner et al. 2017; Sun et
130 al. 2012; Zhang et al. 2016). We characterized this in longitudinal isolates fulfilling three
131 criteria: (i) isolates containing multiple resistance SNPs in the same gene within the same
132 subject, (ii) at alternate allele frequencies that change in opposing directions over time and (iii)
133 the alternate (mutant) allele frequency was intermediate to high at ≥ 40% in at least 1 isolate
134 (Farhat et al. 2016) for at least one of the co-occurring SNPs. This identified 11 cases of clonal
135 interference (**Fig. 2A** and **Fig. S2**), demonstrating most often the fixation of a single allele in the
136 second isolate from a mixture of multiple alleles at lower frequencies in the first isolate
137 collected.
138
139 **Allele frequency >19% predicts subsequent fixation of resistance variants.**
140 We determined the lowest AR allele frequency that can accurately predict the
141 development of fixed resistance alleles later in time (Sun et al. 2012; Zhang et al. 2016). We
142 discarded 20 SNPs that were fixed (AF >75%) in both isolates to study the AF trajectories of
143 1,919 AR SNPs detected with a change in allele frequency, ΔAF ≥ 5% between time points. We
144 calculated the true positive rate (TPR) and false positive rates (FPR) for varying values of $AF_1 \in$
145 $\{0,1,2,\cdots,99,100\}$% (**Fig. 2C, Methods**). For the purposes of prediction, we allowed a
146 maximum FPR of 5% and found the optimal classification threshold to be $AF_1^* = 19\%$ with an
147 associated sensitivity of 27.0% and a specificity of 95.8%. Ten mutant alleles across 14 isolates

148 from 7 subjects had a frequency between 19% and 75% at the first time point and rose to fixation
149 at the second time point (mean ΔAF 41%).
150
151 **Determinants of antibiotic resistance acquisition and microbiological treatment failure**
152 All study subjects had either recently completed treatment or were receiving treatment
153 when samples were collected (Walker et al. 2013; Casali et al. 2016; Trauner et al. 2017; Xu et
154 al. 2018; Guerra-Assunção et al. 2014; Bryant et al. 2013; Witney et al. 2017), but details of the
155 treatment regimen were not available (**Table S1**). We identified overall rates of resistance
156 acquisition by focusing on AR SNPs with moderate to high ΔAF ≥ 40% given prior evidence of
157 association between such SNPs and phenotypic resistance (Farhat et al. 2016). Thirty-eight AR
158 SNPs were detected in 195/200 subjects with clonal infection and known isolate sampling date.
159 AR acquisition was more likely as the time between sampling, and likely antibiotic exposure,
160 increased, with the OR of AR acquisition being 1.023 per 30day increment
161 (95% CI 1.002 , 1.045, $P = 0.035$ Logistic Regression).
162 We focused on subjects with samples collected ≥2 months apart, by definition culture
163 positive at sample collection time, because they meet microbiological criteria for *delayed culture*
164 *conversion, failure* or *relapse* (hitherto failure for brevity) (WHO). Of the 269 subjects with
165 mixed or clonal infection and reinfection as well as complete isolate sampling dates, 230 had
166 samples collected ≥2 months apart and consisted of 35 reinfections (15%), 28 mixed infections at
167 one or two time points (12%) and the majority, 167 (73%), had clonal infection (**Fig. S1**).
168 We examined the relationship between pre-existing resistance and new AR acquisition in
169 clonal treatment failure. Pre-existing resistance was defined as ≥1 fixed AR SNPs in the first
170 isolate (Farhat et al. 2016) (**Methods**). Two hundred thirty pre-existing AR SNPs were identified

171 with 39% (65/167) of failure subjects harboring resistance to any drug at the first sampling (**Fig.**
172 **3B**, **Table S7**). The majority of this resistance was MDR (multidrug resistance to at least
173 isoniazid and a rifamycin), 60% (39/65) (**Fig. 3C**). New resistance acquisition occurred mostly in
174 subjects with pre-existing resistance 20/27 (74%) ($OR = 6.03,\ P = 6.8 \times 10^{-5}$ Fisher's exact
175 test) or pre-existing MDR ($OR = 4.95,\ P = 3.8\ \times 10^{-4}$ Fisher's exact test).
176 We also quantified genome-wide Mtb diversity in-host among the patients with
177 microbiological treatment failure. We reasoned that as these patients are not on or not adherent to
178 effective antibiotic treatment, their effective pathogen population size may be large and prone to
179 more genetic drift or turnover of minority variants with and without selection (Trauner et al.
180 2017). We counted the number SNPs with an alternate allele frequency between 25% and 75%
181 (**Methods**) at each time point as a conservative estimate of the number of segregating sites in
182 each population. We found this count to strongly correlate between the first and second time
183 point (**Fig. 3A**) suggesting that minor allele diversity is maintained in-host in patients without
184 effective therapy (median T1=15 hSNPs, median T2=16 hSNPs , $r^2 = 0.43,\ P = 9.5\ \times 10^{-22}$
185 Linear Regression).
186
187 **Genome-wide in-host diversity**
188 Beyond antibiotic pressure, selective forces acting on the infecting Mtb population in-
189 host are largely unknown. To investigate this reliably across the entire Mtb genome, we first
190 examined the genome-wide allele frequency distribution for both technical replicates (*in vitro*
191 technical or biological replicates, sample size m=62 after exclusions) and in-host longitudinal
192 pairs (**Fig. 4** and **Fig. S1**). We detected five SNPs in *glpK* (with ΔAF ≥ 25%) among five
193 replicate pairs (mean ΔAF=45%) consistent with an adaptive role for *glpK* mutations *in vitro*

194 (Pethe et al. 2010; Vargas and Farhat 2020) and accordingly excluded this gene from further
195 analysis (**Methods**). The genome-wide AF distribution in both replicate and longitudinal pairs
196 demonstrated an abundance of SNPs with low ΔAF likely resulting from noise or technical
197 factors. To clearly distinguish signal related to in-host factors from noise, we determined the
198 ΔAF threshold above which SNPs/isolate-pair were rare among technical replicates *i.e.*
199 constituted 5% or less of total SNPs (**Fig. 4**). We determined this ΔAF threshold to be 70% and
200 selected 174 SNPs that developed in-host (in-host SNPs) among the 200 TB cases (**Fig. 4C,**
201 **Table S8**).
202
203 **Characteristics of mutations in-host**
204 Of the 174 SNPs, 112 were non-synonymous, 42 synonymous, and 21 were intergenic
205 (**Fig. 5C**). The 153/174 coding SNPs were distributed across 127/3,886 genes and were observed
206 in 71/200 subjects (**Fig. 5B** and **Fig. 5D**). We analyzed the spectrum of mutations and found the
207 GC > AT nucleotide transition to be the most common. The GC > AT transition is putatively due
208 to oxidative damage including the deamination of cytosine/5-methyl-cytosine or the formation of
209 8-oxoguanine (Ford et al. 2011; Dillon et al. 2015). The transversion AT > TA was the least
210 common substitution (**Fig. 6A**). We expected the number of SNPs detected between longitudinal
211 isolates to increase with time between isolate collection. Regressing the number of SNPs per
212 subject on the timing between isolate collection (for 195 subjects with isolate collection dates)
213 (**Fig. 6B**), we found SNPs to accumulate at an average rate of 0.56 SNPs per genome per year
214 ($P = 7 \times 10^{-12}$) consistent with prior *in vivo* estimates (Walker et al. 2013; Ford et al. 2011).
215

216 **Simulations and PacBio sequencing demonstrate a low false-positive rate in repetitive**
217 **regions**

218 Several SNPs detected were in the GC-rich repetitive PE/PPE gene family (Brennan and
219 Delogu 2002). Variants called on these genes are commonly excluded from comparative
220 genomic analyses (Coscolla et al. 2015; Comas et al. 2010; Copin et al. 2016; Casali et al. 2016)
221 due to the limitations of short-read sequencing data and the possibility of making spurious
222 variant calls, however the rates at which these false calls occur has not been evaluated. We
223 reasoned that our stringent filtering criteria, quality of sequencing data and depth of coverage
224 allowed us to reliably detect variants in these regions of the genome, with the potential to
225 uncover variation in these understudied regions of the genome.

226 We took several approaches to test the rate of false-positives for the single base-pair
227 mutations observed in our analysis (**Methods**). First, we introduced the mutant alleles observed
228 in-host (**Table S10**) into a set of Mtb reference genomes belonging to different lineages and
229 simulated short read sequencing data from these modified genomes (**Fig. S5**). We then used our
230 variant calling pipeline to call bases from this simulated data. We observed a high recall rate of
231 the introduced mutant alleles and a very low number of false positive base calls (zero in most
232 cases) within the loci containing modified alleles (**Fig. S6**). Second, we assessed the congruence
233 in variant calls between short-read Illumina data and long-read PacBio data for a set of isolates
234 that underwent sequencing with both technologies (**Methods)**. Unlike Illumina generated reads,
235 PacBio reads are much longer and have randomly distributed error profiles (Rhoads and Au
236 2015). With high coverage, PacBio sequencing can reliably reconstruct full microbial genomes
237 and identify SNPs in repetitive regions. The comparison with PacBio assemblies confirmed
238 empirically a low rate of false positive base calls in genomic regions where we observed in-host

239 SNPs (**Methods**). Third, we confirmed the five phylogenetically convergent in-host SNPs in PPE
240 genes *PPE18*, *PPE54* and *PPE60* (see below) through manual inspection of the read alignment
241 (**Fig. S7-S9**).
242
243 **Antibiotic resistance and PE/PPE genes vary while antigens remain conserved**
244 To understand how different classes of proteins evolve in-host, we separated Mtb genes
245 into five non-redundant categories (**Methods**). The vast majority of genes in each category did
246 not vary within subjects (**Fig. 6C**). Antibiotic resistance genes were on average the most diverse
247 category while Essential genes varied the least (**Fig. 6D**). Antigen genes appeared to be as
248 conserved as were both Essential ($P = 0.49$ Mann-Whitney U-test) and Non-Essential genes
249 ($P = 0.45$ Mann-Whitney U-test) while PE/PPE genes showed higher levels of nucleotide
250 diversity than both Essential ($P = 0.022$ Mann-Whitney U-test) and Non-Essential genes ($P =$
251 0.013 Mann-Whitney U-test) (**Fig. 6D**).
252
253 **PE/PPE variation is independent of T-cell recognition**
254 To test whether variation in Antigen or PE/PPE genes occurred in response to T-cell
255 recognition, we separated each gene in these categories into ($CD4^+$ and $CD8^+$ T-cell) epitope and
256 non-epitope concatenates and recalculated nucleotide diversity for these concatenates (**Fig. 6E-**
257 **H**). For both Antigen and PE/PPE genes (**Fig. 6F** and **Fig. 6H**), epitope concatenates were less
258 diverse than non-epitope concatenates ($P = 0.018$ and $P = 0.059$ respectively, Mann-Whitney
259 U-test). Only one in-host SNP was detected within an epitope-encoding region in the gene
260 *PPE18* (**Fig. 6G** and **Fig. S4**, **Table S12**). This suggests that T-cell recognition does not drive
261 diversity in these regions. Looking within the three PE/PPE subfamilies (**Fig. 6I-J**) (Brennan

262 2017), the PPE genes appeared more diverse in-host than PE genes and PE-PGRS genes ($P =$
263 0.019 and $P = 0.033$ respectively, Mann-Whitney U-test).
264
265 **Identifying candidate pathoadaptive loci from genome-wide variation**
266 To identify genes involved in pathogen adaptation (Lieberman et al. 2011; Marvig et al.
267 2015), we applied a test of mutational density (Farhat et al. 2014) (**Methods**) by pooling
268 variation across all 200 pairs of genomes and identifying those genes with more mutations than
269 expected under a neutral model of evolution where variants are Poisson distributed across the
270 genome (Farhat et al. 2014) (**Fig. 5B, Table S13**, **Methods**). We also searched for evidence of
271 convergent evolution *i.e.* genes or pathways where in-host SNPs developed in ≥ 2 subjects
272 (**Table S14, S17**). Seven known antibiotic resistance genes (Didelot et al. 2016; Farhat et al.
273 2013) had significant mutational density ($\alpha = 0.05$, Bonferroni correction) or were convergent
274 across patients: *rpoB*, *gyrA*, *katG*, *rpoC*, *embB, ethA* and *pncA* (mutated in six, four, four, three,
275 three, two and one subject respectively) (**Fig. 5B** and **Fig. 5D**). Single in-host SNPs occurred in
276 eight additional known resistance loci including three intergenic regions, and in *prpR,* a gene
277 recently implicated with drug tolerance (Hicks et al. 2018) (**Table S8**).
278 Three genes with unknown function: Rv0139, Rv0895, and Rv1543 were convergent in
279 two subjects each, two of which (Rv0139, Rv1543) had significant mutational density ($P<2x10^{-}$
280 ^{5}) and; three additional genes including *PPE60* displayed significant mutational density ($P<2x10^{-}$
281 ^{5}) (**Fig. 5B**, **Table S13**). We found evidence for convergence in six pathways not known to result
282 in antibiotic resistance. These pathways are involved with biotin biosynthesis (*fadD23*, *fadD29*,
283 and *fadD30*), ribosomal large subunit proteins (*rpmB1*, *rplE*, and *rplY*), glycerolipid and
284 glycerophospholipid metabolism (*aldA* and Rv2974c), ESAT-6 protein secretion (*eccCa1* and

285 *eccD1*), coenzyme B12/cobalamin synthesis (*cobH* and *cobK*) and the uncharacterized pathway
286 CBSS-164757.7.peg.5020 (*fdxB* and *PPE18*) (**Table S17**).
287
288 **In-host mutations display phylogenetic convergence across multiple global lineages**
289 We reasoned that pathoadaptive mutations observed to sweep to fixation in-host and not
290 compromise pathogen transmissibility are likely to arise independently within other subjects and
291 in separate geographic regions in a convergent manner (Farhat et al. 2013). We screened a
292 geographically diverse set of 20,352 sequenced clinical isolates belonging to global lineages 1-6
293 for mutations observed within host in which the alternate (mutant) allele swept over the course of
294 sampling (141/174 in-host SNPs, **Methods**, **Table S8, S18**). Conservatively, a mutation was
295 characterized as phylogenetically convergent if it was present in isolates from three or more
296 global lineages but not fixed in any lineage (**Methods**). We identified 26/141 in-host SNPs as
297 phylogenetically convergent in our global sample of isolates (**Table S19**). **Fig. 7** and **Fig. S10-11**
298 display the distribution of convergent alleles across the 20,353 isolates using t-Distributed
299 Stochastic Neighbor Embedding (t-SNE) of the pairwise genetic distance matrix (**Methods**). The
300 convergent alleles included the PPE genes *PPE18* (1 site), *PPE54* (1 site) and *PPE60* (3 sites), as
301 well as Rv0095c (2 sites) and Rv1944c (1 site) both conserved proteins of unknown function. In
302 addition to several SNPs in loci associated with antibiotic resistance, *gyrB* (1 site), *gyrA* (2 sites),
303 *rpoB* (4 sites), *rpoC* (3 sites), *inhA* (1 site), *embB* (3 sites) and *gid* (1 site).

304 **DISCUSSION**

305 In our 400 Mtb populations sequenced from 200 active TB patients heavily enriched for
306 delayed culture conversion, treatment failure and relapse, we find a wealth of dynamics in
307 genetic loci associated with antibiotic resistance, including a high turnover of minor variants
308 (Trauner et al. 2017). Factors that determine negative treatment outcomes are complex and
309 include severity of lung disease, cavitation and adherence to treatment among others (Imperial et
310 al. 2018). Here, we observe that a relatively high percentage of such patients, 16%, develop
311 antibiotic resistance over time. Our findings of a higher rate of resistance acquisition in patients
312 with MDR at the outset and with time between sampling, emphasize the importance of
313 appropriately tailoring treatment regimens as well as close surveillance for microbiological
314 clearance and resistance acquisition by phenotypic or genotypic means. The observed high rate
315 of resistance acquisition also emphasizes Mtb's biological adaptability and the long duration of
316 drug pressure *in vivo.* Resistance acquisition in the course of one infection is comparatively rare
317 in most pathogenic bacteria (Llewelyn et al. 2017). In addition to clonal acquisition of resistance,
318 we found 27% of patients with unsuccessful treatment outcomes to have mixed infection or
319 reinfection with different Mtb strains. This high percentage suggests that care of these patients
320 and control of disease transmission can be better guided if pathogen sequencing is routinely
321 performed for cases meeting these microbiological criteria especially in high TB prevalence
322 settings where reinfection is more likely.

323 We provide a proof of concept analysis that minor AR alleles, occurring at a frequency
324 down to 19%, can accurately predict fixation of the variant in >95% of mutations in-host,
325 although we find the sensitivity of this threshold to be low. In the future, higher depth

326 sequencing at fixed sampling durations can elucidate more clearly the role of minor AR allele
327 detection in clinical management of TB treatment.
328 Various sources of noise contribute to allele frequency changes over time and challenge
329 inference on bacterial composition *in vivo*. Here, we determined an appropriate threshold for
330 identifying mutations in-host using average depth Mtb WGS from cultured isolates and
331 demonstrate the importance of including technical replicate WGS. While culturing sputa *in vitro*
332 enriches Mtb DNA for WGS it also creates experimental noise (Vargas and Farhat 2020). The
333 refinement of methods for DNA extraction directly from sputum (Votintseva et al. 2017), may
334 allow the calling of relevant changes in allele frequencies at lower thresholds in future work.
335 This would permit the unbiased study of loci that may be under frequency-dependent selection,
336 where changes in allele frequencies would unlikely change by as much as 70% as we used here.
337 We detected 174 alleles rising to near fixation in-host across our sample of 200 subjects.
338 The observed distribution of variants including the high rate of non-synonymous substitutions
339 and the predominance of GC > AT variants are consistent with the hypotheses of purifying
340 pressure on synonymous variants and oxidative DNA damage respectively (Ford et al. 2011;
341 Namouchi et al. 2012) in Mtb. This consistency adds validity to our variant calling approach.
342 Overall the observed diversity spared the $CD4^+$ and $CD8^+$ T cell epitope encoding regions of the
343 genome providing further evidence that host adaptive immunity does not drive directional
344 selection in Mtb genomes now at short-time scales (Coscolla et al. 2015; Comas et al. 2010;
345 Copin et al. 2014). Diversity was concentrated in antibiotic resistance regions and strikingly also
346 in PE/PPE genes (**Fig. 6D**) (Phelan et al. 2016). Although previous studies have generally
347 avoided reporting short-read variant calls in PE/PPE regions, we demonstrate using read
348 simulation, visualization of Illumina read alignments and comparison with long-read sequencing

349 data the accuracy of the SNPs captured in our study. We found PPE genes to be more diverse in-
350 host than PE genes and detected a signal of positive selection acting on three genes belonging to
351 the PPE sub-family (**Fig. 7**). This indicates that PPE genes may be play an important role in the
352 process of host-adaptation.

353 In addition to identifying in-host variation in 12 loci known to be involved in the
354 acquisition of antibiotic resistance, we identified six genes and six pathways displaying diversity
355 in-host and not known to be associated with antibiotic resistance (**Table S8, S13-14, S17**). For a
356 subset we demonstrate similar diversity has arisen independently in separate hosts and in strains
357 with different genetic backgrounds suggesting positive selection (**Fig. 7**). Evidence of directional
358 selection in Mtb genomes have thus far been largely restricted to adaptation to antibiotic
359 treatment (Didelot et al. 2016; Trauner et al. 2017; Brites and Gagneux 2015). The novel
360 pathways showing in-host convergence may be important for interactions between host and
361 pathogen arising from either metabolic or immune pressure. Mtb is one of a few types of bacteria
362 that possess the capacity for *de novo* coenzyme B12/cobalamin synthesis, and this pathway has
363 been implicated in Mtb survival in-host and Mtb growth (Rowley and Kendall 2019). We
364 identified four genetic variants that developed in three separate patients and in three consecutive
365 genes from the same locus *cobG,* intergenic *cobG-cobH, cobH* and *cobK* (Rv2064-Rv2067).
366 This observation contributes to mounting evidence on the importance of this pathway for *in vivo*
367 Mtb survival and may have implications for drug development (Minias et al. 2018; Gopinath et
368 al. 2013). Biotin biosynthesis is also relatively unique to mycobacteria and plays an important
369 role in Mtb growth, infection and host survival during latency (Salaemae et al. 2011). The other
370 identified pathways include ESAT-6 protein secretion known to play a role in the modulation of
371 host immune response by disrupting the phagosomal membrane (Clemmensen et al. 2017).

372 The loci found to be phylogenetically convergent and not known to be associated with
373 antibiotic resistance, include the genes Rv0095c, *PPE18*, *PPE54* and *PPE60*. Although of
374 unknown function, Rv0095c (SNP A85V) was recently associated with transmission success of
375 an Mtb cluster in Peru (Dixit et al. 2019). Both *PPE18* and *PPE60* have been shown to interact
376 with toll-like receptor 2 (TLR2) (Nair et al. 2009; Su et al. 2018). *PPE18* was the only gene to
377 encode an epitope containing a SNP in-host; mutations in the epitope-encoding regions of this
378 gene have previously been described in a set of geographically separated clinical isolates (Hebert
379 et al. 2007). Furthermore, *PPE18* codes for one of the antigens used in the construction of the
380 M72/AS01E vaccine candidate (Tait et al. 2019). Our results demonstrating that *PPE18* is under
381 positive selection in the MTBC may have implications for the efficacy of this vaccine against
382 genetically diverse Mtb strains. *PPE54* has been implicated in Mtb's ability to arrest macrophage
383 phagosomal maturation (phagosome-lysosome fusion) and thought to be vital for intracellular
384 persistence (Brodin et al. 2010). The mechanism by which *PPE54* accomplishes this is unknown,
385 but Mtb modification of phagosomal function is thought to be TLR2/TLR4-dependent
386 (Podinovskaia et al. 2013).
387 Mtb is known to disrupt numerous *innate* immune mechanisms including phagosome
388 maturation, apoptosis, autophagy as well as inhibition of MHC II expression through prolonged
389 engagement with innate sensor toll-like receptor 2 (TLR2) among others (Ernst 2018a). SNPs in
390 human genes involved with innate-immune pathways have been implicated in-host susceptibility
391 to TB (Kleinnijenhuis et al. 2011; Tientcheu et al. 2017; Azad et al. 2012). Specifically, SNPs in
392 TLR2 (thought to be the most important TLR in Mtb recognition) (Tientcheu et al. 2017) and
393 TLR4 have been associated with susceptibility to TB disease (Kleinnijenhuis et al. 2011; Azad et
394 al. 2012). Taken together, these observations and our results are consistent with ongoing co-

395 evolution between humans and Mtb with evidence for reciprocal adaptive changes, leaving a
396 signature of selection in both humans and Mtb populations (Brites and Gagneux 2015). Most co-
397 evolution between Mtb and humans, the main reciprocal adaptations between host and pathogen
398 are thought to have occurred long ago and as a result of long-term host-pathogen interactions
399 (Brites and Gagneux 2015; Azad et al. 2012). Here, we observe these dynamics over the short
400 evolutionary timescale of a single infection which has important implications for vaccine
401 development (Brennan 2017).

402 **MATERIALS AND METHODS**

403 **Sequence Data**

404 Longitudinal Isolate Pairs: This study included data for 614 clinical isolates of *M. tuberculosis*
405 that were sampled from the sputum of 307 subjects resulting in n = 307 longitudinal pairs. The
406 sequencing data for 456 publicly available isolates was downloaded from Genbank (Benson et al.
407 2008), sequenced using Illumina chemistry to generate paired-end reads and came from
408 previously published studies (T (Trauner et al. 2017), C (Casali et al. 2016), W (Walker et al.
409 2013), B (Bryant et al. 2013), G (Guerra-Assunção et al. 2014), X (Xu et al. 2018), H (Witney et
410 al. 2017), P (Farhat et al. 2019)) (**Fig. S1**, **Table S1-S2**).

411 Replicate Isolate Pairs: This study included three types of replicate isolate pairs. (S2 - Sequenced
412 Twice) DNA pooled from a single Mtb clinical isolate that had undergone *in vitro* expansion was
413 sequenced in separate runs on an Illumina sequencing machine (m = 5). (C2 – Cultured &
414 Sequenced Twice) Mtb was cultured from a single frozen clinical sample at separate time points,
415 then sequenced on an Illumina sequencing machine after DNA extraction from culture (m = 73).
416 (P3) Three sputum samples were obtained from a single subject within a 24 hour period (Trauner
417 et al. 2017), cultured separately, underwent DNA extraction and then sequencing on an Illumina
418 sequencing machine. For the purposes of this study, we compared these three isolates pairwise
419 (m = 3).

420 Global Sequence Data: We downloaded raw sequence data for 33,873 clinical isolates from the
421 public domain (Benson et al. 2008). Isolates had to meet the following quality control measures
422 for inclusion in our study: (i) at least 90% of the reads had to be taxonomically classified as
423 belonging to the *Mycobacterium tuberculosis* complex after running the trimmed FASTQ files

424 through Kraken (Wood and Salzberg 2014) and (ii) at least 95% of bases had to have coverage of
425 at least 10x after mapping the processed reads to the H37Rv Reference Genome.
426
427 **Epitope Collection and Analysis**
428 $CD4^+$ T and $CD8^+$ T cell epitope sequences were downloaded from the Immune Epitope
429 Database (Vita et al. 2014) on May 23rd, 2018 according to criteria described previously
430 (Coscolla et al. 2015) [linear peptides, *M. tuberculosis* complex (ID:77643, Mycobacterium
431 complex), positive assays only, T cell assays, any MHC restriction, host: humans, any diseases,
432 any reference type] yielding a set of 2,031 epitope sequences (**Table S11**). We mapped each
433 epitope sequence to the genes encoded by the H37Rv Reference Genome (Cole et al. 1998) using
434 BlastP with an e-value cutoff of 0.01 (**Fig. S3A**). We retained only epitope sequences that
435 mapped to at least 1 region in H37Rv (due to sequence homology, some epitopes mapped to
436 multiple regions) and whose BlastP peptide start/end coordinates matched those specified in
437 IEDB (n = 1,949 representing 1,505 separate epitope entries in IEDB). We then filtered out any
438 epitopes occurring in Mobile Genetic Elements which resulted in a final set of 1,875 epitope
439 sequences, representing 348 genes (antigens) used for downstream analysis. The distribution of
440 peptide lengths for this final set of epitopes is given in **Fig. S3B**. Since many of these epitope
441 sequences overlap, we constructed non-redundant epitope concatenate sequences for each
442 antigen (n = 348) gene (Coscolla et al. 2015; Stucki et al. 2016; Comas et al. 2010). The regions
443 of each antigen not encoding an epitope were concatenated into a non-epitope sequence for that
444 gene.
445
446 **Gene Sets**

447 Every gene on H37Rv was classified into one of six non-redundant gene categories according to
448 the following criteria: (i) genes identified as belonging to the PE/PPE family of genes unique to
449 pathogenic mycobacteria, though to influence immunopathogenicity and characterized by
450 conserved proline-glutamate (PE) and proline-proline-glutamate (PPE) motifs at the N protein
451 termini (Brennan and Delogu 2002; Comas et al. 2010; Phelan et al. 2016) were classified as
452 *PE/PPE* (n = 167), (ii) genes flagged as being associated with antibiotic resistance (Farhat et al.
453 2013) were classified into the *Antibiotic Resistance* category (n = 28), (iii) genes encoding a
454 $CD4^+$ or $CD8^+$ T-cell epitope (Coscolla et al. 2015; Comas et al. 2010) (but not already classified
455 as a PE/PPE or Antibiotic Resistance gene) were classified as an *Antigen* (n = 257) , (iv) genes
456 required for growth *in vitro* (Sassetti et al. 2003) and *in vivo* (Sassetti and Rubin 2003) and not
457 already placed into a category above were classified as *Essential* genes (n = 682), (v) genes
458 flagged as transposases, integrases, phages or insertion sequences were classified as *Mobile*
459 *Genetic Elements* (Comas et al. 2010) (n = 108), (vi) any remaining genes not already classified
460 above were placed into the *Non-Essential* category (n = 2752) (**Table S4**).
461

462 **Illumina Sequencing FastQ Processing and Mapping to H37Rv**

463 The raw sequence reads from all sequenced isolates were trimmed with Prinseq (Schmieder and
464 Edwards 2011) (settings: -min_qual_mean 20) (version 0.20.4) then aligned to the H37Rv
465 Reference Genome (Genbank accession: NC_000962) with the BWA mem (Li and Durbin 2009)
466 algorithm (settings: -M) (version 0.7.15). The resulting SAM files were then sorted (settings:
467 SORT_ORDER = coordinate), converted to BAM format and processed for duplicate removal
468 with Picard (http://broadinstitute.github.io/picard/) (version 2.8.0) (settings:
469 REMOVE_DUPLICATES = true, ASSUME_SORT_ORDER = coordinate). The processed

470 BAM files were then indexed with Samtools (Li et al. 2009). We used Pilon (Walker et al. 2014)
471 on the resulting BAM files to call bases for all reference positions corresponding to H37Rv from
472 pileup (settings: --variant).
473
474 **Empirical Score for Difficult-to-Call Regions**
475 We extracted DNA from 15 Mtb isolates (Epperson and Strong 2020), for which we had Illumina
476 sequencing reads, to undergo PacBio sequencing (**Supplementary Note 2)**. Together, with
477 public PacBio and Illumina sequencing data (Chiner-Oms et al. 2019), we compiled 31 pairs of
478 sequencing reads for comparison of variant calling between PacBio long-reads and Illumina
479 short-reads (**Table S20**). Using the 31 isolates for which both Illumina and a complete PacBio
480 assembly were available (**Supplementary Note 2)**, we evaluated the empirical base-pair recall
481 (EBR) of all base-pair positions of the H37rv reference genome (Marin et al. 2020). For each
482 sample, the alignments (from minimap2) of each high confidence genome assembly to the
483 H37Rv genome were used to infer the true nucleotide identity of each base pair position. To
484 calculate the empirical base-pair recall, we calculated what % of the time our Illumina based
485 variant calling pipeline, across 31 samples, confidently called the true nucleotide identity at a
486 given genomic position. If Pilon variant calls did not produce a confident base call *(Pass)* for the
487 position, it did not count as a correct base call. This yields a metric ranging from 0.0 – 1.0 for the
488 consistency by which each base-pair is both confidently and correctly sequenced by our Illumina
489 WGS based variant calling pipeline for each position on the H37Rv reference genome. An
490 H37Rv position with an EBR score of x% indicates that the base calls made from Illumina
491 sequencing and mapping to H37Rv agreed with the base calls made from the PacBio *de novo*
492 assemblies (**Supplementary Note 2**) in x% of the Illumina-PacBio pairs. We masked difficult-

493 to-call regions by dropping H37Rv positions with an EBR score below 0.8 as part of our variant
494 calling procedure.
495
496 **Variant Calling**
497 Single Nucleotide Polymorphism (SNP) Calling: To prune out low-quality base calls that may
498 have arisen due to sequencing or mapping error, we dropped any base calls that did not meet any
499 of the following criteria (Copin et al. 2016): (i) the call was flagged as either *Pass* or *Ambiguous*
500 by Pilon, (ii) the reads aligning to that position supported at most 2 alleles (ensuring that 1 allele
501 matched the reference allele if there were 2), (iii) the mean base quality at the locus was > 20,
502 (iv) the mean mapping quality at the locus was > 30, (v) none of the reads aligning to the locus
503 supported an insertion or deletion, (vi) a minimum coverage of 25 reads at the position, (vii) the
504 EBR score for the position ≥ 0.80, and (viii) the position is not located in a mobile genetic
505 element region of the reference genome. We then used the Pilon-generated (Walker et al. 2014)
506 VCF files to calculate the frequencies for both the reference and alternate alleles, using the
507 *INFO.QP* field (which gives the proportion of reads supporting each base weighted by the base
508 and mapping quality of the reads, *BQ* and *MQ* respectively, at the specific position) to determine
509 the proportion of reads supporting each base for each locus of interest.
510 Additional SNP Filtering for Isolate Pairs: To call SNPs (and corresponding changes in allele
511 frequencies) between pairs of isolates (Replicate and Longitudinal pairs), we required: (i) *SNP*
512 *Calling* filters be met, (ii) the number of reads aligning to the position is below the 99th
513 percentile for all of the calls made for that isolate, (iii) the call at that position passes all filters
514 for each isolate in the pair, and (iv) SNPs in *glpK* were dropped as mutants arising in this gene
515 are thought to be an artifact of *in vitro* expansion (Pethe et al. 2010; Vargas and Farhat 2020); we

516 detected four non-synonymous SNPs in *glpK* (ΔAF ≥ 25%) between three longitudinal pairs
517 (mean ΔAF=64%) and five non-synonymous SNPs in *glpK* (ΔAF ≥ 25%) among five replicate
518 pairs (mean ΔAF=45%).
519 Additional SNP Filtering for Antibiotic Resistance Loci Analysis: To call SNPs (and
520 corresponding minor changes in allele frequencies) between pairs of isolates (Longitudinal
521 Pairs), we required: (i) *SNP Calling* filters be met, (ii) *Additional SNP Filtering for Isolate Pairs*
522 filters be met, (iii) $\left|AF_1^{alt} - AF_2^{alt}\right| = \Delta\text{AF} \geq 5\%$, (iv) if $5\% \leq \Delta\text{AF} < 20\%$, then the SNP was
523 only retained if each allele (across both isolates) with $\text{AF} > 0\%$ was supported by at least 5
524 reads (ensuring that at least 5 reads supported each minor allele at lower values of ΔAF), (v) the
525 SNP was classified as either intergenic or non-synonymous, (vi) the SNP was located in a gene,
526 intergenic region or rRNA coding region associated with antibiotic resistance (**Table S5**).
527 Additional SNP Filtering for Heterogenous SNPs: To call heterogenous SNPs in each isolate for
528 a pair of longitudinal isolates, we required: (i) *SNP Calling* filters be met, (ii) the number of
529 reads aligning to the position is below the 99th percentile for all of the calls made for that isolate,
530 (iii) the alternate allele frequency for the SNP is such that $25\% \leq AF^{alt} \leq 75\%$, (iv) the
531 position is not located in a mobile genetic element region of the reference genome, (v) the
532 position is not located in a PE/PPE region of the reference genome.
533 Additional SNP Filtering for Global Isolates: To call alleles in our global and genetically diverse
534 set of isolates, we required: (i) *SNP Calling* filters be met, with the modifications that (ii) the call
535 was flagged as *Pass* by Pilon and (iii) 1 allele (either the reference or an alternate) was supported
536 by at least 90% of the reads regardless of whether the other ≤ 10% of reads supported 1, 2, or 3
537 alleles.
538

539 **Mixed Lineage and Contamination Detection for Longitudinal and Replicate Isolate Pairs**

540 Kraken: To filter out samples that may have been contaminated by foreign DNA during sample
541 preparation, we ran the trimmed reads for each longitudinal and replicate isolate through
542 Kraken2 (Wood and Salzberg 2014) against a database (Goig et al. 2020) containing all of the
543 sequences of bacteria, archaea, virus, protozoa, plasmids and fungi in RefSeq (release 90) and
544 the human genome (GRCh38). We calculated the proportion reads that were taxonomically
545 classified under the *Mycobacterium tuberculosis* Complex (MTBC) for each isolate and
546 implemented a threshold of 95%. An isolate pair was dropped if either isolate had less than 95%
547 of reads aligning to MTBC.

548 F2: To further reduce the effects of contamination, we aimed to identify samples that may have
549 been subject to inter-lineage mixture samples resulting from of a co-infection (F2). We computed
550 the F2 lineage-mixture metric for each longitudinal and replicate isolate (**Fig. 1B**). We wrote a
551 custom script to carry out the same protocol for computing F2 as previously described (Wyllie et
552 al. 2018). Briefly, the method involves calculating the minor allele frequencies at lineage-
553 defining SNPs (Coll et al. 2014). From 64 sets of SNPs that define the deep branches of the
554 MTBC (Coll et al. 2014), we considered the 57 sets that contain more than 20 SNPs to obtain
555 better estimates of minor variation (Wyllie et al. 2018; Coll et al. 2014). For each SNP set i, (i)
556 we summed the total depth and (ii) the number of reads supporting the most abundant base (at
557 each position) over all of the reference positions (SNPs) that met our mapping quality, base
558 quality and insertion/deletion filters, which yields d_i and x_i respectively. Subtracting these two
559 quantities yields the minor depth for SNP set i, $m_i = d_i - x_i$. The minor allele frequency
560 estimate for SNP set i is then defined as $p_i = m_i \,/\, d_i$. Doing this for all 57 SNP sets gives
561 $\{p_1, p_2, \cdots, p_{57}\}$. We then sorted $\{p_1, p_2, \cdots, p_{57}\}$ in descending order and estimated the minor

562 variant frequency for all of the reference positions (SNPs) corresponding to the top 2 sets
563 (highest p_i values) which yields the F2 metric. Letting $n2$ be the number of SNPs in the top 2
564 sets, then $F2 = \sum_{j=1}^{n2} m_j \,/\, \sum_{i=1}^{n2} d_i$. Isolate pairs were dropped if the F2 metric for either isolate
565 passed the F2 threshold set for mixed lineage detection (**Fig. 1B** and **Fig. S1**).
566

567 **Pre-existing Genotypic Resistance**

568 We determined pre-existing resistance for a subject (with a pair of longitudinal isolates) by
569 scanning the first isolate for the detection of at least 1 of 177 SNPs predictive of resistance with
570 AF ≥ 75% (from a minimal set of 238 variants (Farhat et al. 2016)). We excluded predictive
571 indels and the *gid* E92D variant as the latter is likely a lineage marking variant that is not
572 indicative of antibiotic resistance. We defined pre-existing multidrug resistance for a subject by
573 scanning the first isolate collected for detection of at least 1 SNP predictive of Rifampicin
574 resistance (14/178 predictive SNPs) and at least 1 SNP predictive of Isoniazid resistance (18/178
575 predictive SNPs).
576

577 **True & False Positive Rate Analysis for Heteroresistant Mutations**

578 To determine the predictive value of low-frequency heteroresistant alleles, we classified SNPs as
579 fixed if the alternate allele frequency in the 2nd isolate collected from the subject was at least
580 75% (alt $AF_2 \geq 75\%$). We first dropped SNPs for which alt $AF_1 \geq 75\%$ and alt $AF_2 \geq 75\%$
581 (high frequency mutant alleles in both isolates). We then set a threshold (F_i) for the alternate
582 allele frequency detected in the 1st isolate collected from the subject (alt AF_1) and predicted
583 whether an alternate allele would rise to a substantial proportion of the sample (alt $AF_2 \geq 75\%$)
584 as follows:

585 $$alt\ AF_1 < F_i \longrightarrow alt\ AF_2 < 75\%$$
586 $$alt\ AF_1 \geq F_i \longrightarrow alt\ AF_2 \geq 75\%$$

587 We classified every SNP as True Positive (TP), False Positive (FP), True Negative (TN) or False
588 Negative (FN) according to:

589 $$TP\colon\ alt\ AF_1 \geq F_i \quad \& \quad alt\ AF_2 \geq 75\%$$
590 $$FP\colon\ alt\ AF_1 \geq F_i \quad \& \quad alt\ AF_2 < 75\%$$
591 $$TN\colon\ alt\ AF_1 < F_i \quad \& \quad alt\ AF_2 < 75\%$$
592 $$FN\colon\ alt\ AF_1 < F_i \quad \& \quad alt\ AF_2 \geq 75\%$$

593 True Positive Rates (TPR) and False Positive Rates (FPR) were calculated as:

594 $$TPR = \frac{\#TP}{\#TP + \#FN} \quad FPR = \frac{\#FP}{\#FP + \#TN}$$

595 Finally, we made predictions for all SNPs and calculated the TPR and FPR for all values of $F_i \in$
596 $\{0\%, 1\%, 2\%, \cdots, 98\%, 99\%, 100\%\}$.

597

598 **Mutation Density Test**

599 The method to detect significant variation for a given locus amongst pairs of sequenced isolates
600 has been described previously (Farhat et al. 2014). Briefly, let $\mathcal{N}_j \sim Pois(\lambda_j)$ be a random
601 variable for the number of SNPs detected across all isolate pairs (for the in-host analysis this is
602 the collection of longitudinal isolate pairs for all subjects) for gene j. Let (i) $N_i =$ number of
603 SNPs across all pairs for gene i, (ii) $|g_i| =$ length of gene i, (iii) $P =$ number of genome pairs
604 and (iv) $G =$ the number of genes across the genome being analyzed (all genes in the essential,
605 non-essential, antigen, antibiotic resistant and family protein categories).

606

607 Then the length of the genome (concatenate of all genes being analyzed) is given by $\sum_{i=1}^{G}|g_i|$
608 and the number of SNPs across all genes and genome pairs is given by $\sum_{i=1}^{G} N_i$. The null rate for

609 $\mathcal{N}_j$ is given by the mean SNP distance between all pairs of isolates, weighted by the length of

610 gene j as a fraction of the genome concatenate and number of isolate pairs:

611 $$\lambda_j = \left(\frac{\sum_{i=1}^{G} N_i}{P}\right)\left(\frac{|g_i|}{\sum_{i=1}^{G}|g_i|}\right)\left(\frac{1}{P}\right)$$

612 The p-value for gene j is then calculated as $\Pr\left(N_i > \mathcal{N}_j\right)$. We tested 3,386 genes for mutational

613 density and applied Bonferroni correction to determine a significance threshold. We determine a

614 gene to have a significant amount of variation if the assigned p-value $< \frac{0.05}{3{,}386} \approx 1.477 \times 10^{-5}$.

615

616 **Nucleotide Diversity**

617 We define the nucleotide diversity $\left(\pi_g\right)$ for a given gene g as follows: (i) let $\left|gene_g\right| =$ base-

618 pair length of the gene, (ii) $N_{i,j} =$ number of in-host SNPs (independent of the change in allele

619 frequency for each SNP) between the longitudinal isolates for subject i occurring on gene j and

620 (iii) $P =$ number of subjects. Then

621 $$\pi_g = \left(\frac{1}{P}\right)\left(\frac{1}{\left|gene_g\right|}\right)\sum_{i=1}^{P} N_{i,g}$$

622 Correspondingly, let G be a category consisting of M genes, then the average nucleotide diversity

623 for G is given by:

624 $$\pi_G = \left(\frac{1}{M}\right)\left(\frac{1}{P}\right)\sum_{j=1}^{M}\left(\frac{1}{\left|gene_j\right|}\right)\left(\sum_{i=1}^{P} N_{i,j}\right)$$

625

626 **SNP calling simulations in repetitive genomic regions**

627 Certain repetitive regions of the *Mycobacterium tuberculosis* genome (ESX, PE/PPE loci) may

628 give rise to false positive and false negative variant calls due to the mis-alignment of short-read

629 sequencing data. To test the rate of false negative and false positive SNP calls in genes with *in-*
630 *host* SNPs (**Fig. 5**, **Table S8**) we collected the set of non-redundant SNPs observed in these loci
631 (**Table S10**). Next, we collected a set of publicly available reference genomes (**Table S9**) and
632 introduced these mutations into the respective loci positions in the reference genomes. We then
633 simulated short-read Illumina sequencing data of comparable quality to our sequencing data from
634 these altered reference genomes. Using our variant-calling pipeline to call polymorphisms, we
635 then estimated the number of true and false positive SNP calls for each gene, based off of how
636 many introduced SNPs were called (true positives), how many introduced SNPs were not called
637 (false negatives) and how many spurious SNPs were called (false positives). A schematic of our
638 simulation methodology is given in **Fig. S5**, a detailed explanation is given in **Supplementary**
639 **Note 1** and the results of our simulations (given in **Fig. S6)** confirm a low false-positive rate.
640

641 **Global Lineage Typing**

642 We determined the global lineage of each longitudinal ($N = 614$) and global isolate ($N =$
643 $32{,}033$) using base calls from Pilon-generated VCF files and a 62-SNP lineage-defining
644 diagnostic barcode from a previously published study (Coll et al. 2014).
645

646 **Phylogenetic Convergence Analysis & t-SNE Visualization**

647 <u>Construction of Genotypes Matrix</u>: We detected SNP sites at 878,244 H37Rv reference positions
648 (of which 61,918 SNPs were not biallelic) among our global sample of 33,873 isolates. After
649 excluding SNP sites with rare minor alleles (sites in which alternate alleles were called in < 5
650 isolates) we retained SNPs at 146,874 positions. We constructed a 146,874 x 33,873 genotypes
651 matrix (coded as 0:A, 1:C, 2:G, 3:T, 9:Missing) and filled in the matrix for the allele supported at

652 each SNP site for each isolate according to the *SNP Calling* filters outlined above. If a base call
653 at a specific reference position for an isolate did not meet the filter criteria that allele was coded
654 as *Missing*.
655 Excluding 2,348 SNP sites that had an EBR score < 0.80, another 1,509 SNP sites located within
656 mobile genetic element regions, and 3,220 SNP sites in with missing calls in > 25% of isolates
657 yielded a genotypes matrix with dimensions 139,797 x 33,873. Next, we excluded 1,518 isolates
658 with missing calls in > 25% of SNP sites yielding a genotypes matrix with dimensions 139,797 x
659 32,355. We used a previously published 62-SNP barcode (Coll et al. 2014) to type the global
660 lineage of each isolate in our sample. We further excluded 322 isolates that did not get assigned a
661 global lineage, another 100 isolates that were used in our longitudinal analysis (**Table S3**), 152
662 isolates typed as *Mycobacterium bovis*, and 35 isolates typed as lineage 7. To improve
663 computational efficiency and runtime, we randomly down sampled (using Python's
664 random.sample() function) our collection of lineage 2 and lineage 4 isolates to include 7,000
665 isolates of each lineage in our sample. This excluded 1,064 lineage 2 isolates and 10,330 lineage
666 4 isolates. These steps yielded a genotypes matrix with dimensions 139,797 x 20,352. Finally,
667 we excluded 10,899 SNP sites from this filtered genotypes matrix in which the minor allele
668 count = 0. The genotypes matrix used for downstream analysis had dimensions 128,898 x 20,352
669 representing 128,898 SNP sites across 20,352 isolates. The global lineage breakdown of the
670 20,352 isolates was: L1=2,802, L2=7,000, L3=3,352, L4=7,000, L5=101, L6= 97.
671
672 <u>Phylogenetic Convergence Test</u>: We tested 141/174 in-host SNPs in which the alternate (mutant)
673 allele frequency increased substantially across sampling (**Table S8**) for phylogenetic
674 convergence. We scanned a set 20,352 global and genetically diverse isolates from our 128,898 x

675 20,352 genotypes matrix for these SNPs (**Table S18**). To determine phylogenetic convergence
676 for a given SNP site, we required that the alternate allele (called against the H37Rv reference) be
677 present but not fixed in at least three global lineages. More specifically, we required that the
678 alternate allele be detected in at least 1 isolate within each lineage and not in > 95% of the
679 isolates in that lineage. Twenty-six SNP sites across 12 genes and 3 intergenic regions were
680 detected as having a signal of phylogenetic convergence (**Table S19**).
681
682 <u>t-SNE Visualization:</u> To construct the t-SNE plots that captured the genetic relatedness of the
683 20,352 isolates in our sample, we first constructed a pairwise SNP distance matrix. To efficiently
684 compute this using our 128,898 x 20,352 genotypes matrix, we binarized the genotypes matrix
685 and used sparse matrix multiplication implemented in Scipy to compute five 20,352 x 20,352
686 similarity matrices (Virtanen et al. 2020). We constructed a similarity matrix for each nucleotide
687 ($\boldsymbol{A}$, $\boldsymbol{C}$, $\boldsymbol{G}$, $\boldsymbol{T}$) where row *i*, column *j* of the similarity matrix for nucleotide *x* stored the number of
688 *x*'s that isolate *i* and isolate *j* shared in common across all SNP sites. The fifth similarity matrix
689 ($\boldsymbol{N}$) stored the number of SNP sites in which neither isolate *i* and isolate *j* had a missing value.
690 The pairwise SNP distance matrix ($\boldsymbol{D}$) was then computed as $\boldsymbol{D} = \boldsymbol{N} - (\boldsymbol{A} + \boldsymbol{C} + \boldsymbol{G} + \boldsymbol{T})$. $\boldsymbol{D}$ had
691 dimensions 20,352 x 20,352 where row *i*, column *j* stored the number of SNP sites in which
692 isolate *i* and isolate *j* disagreed. We used $\boldsymbol{D}$ as input into a t-SNE algorithm implemented in
693 Scikit-learn (Pedregosa et al. 2011) (settings: perplexity = 40, n_components = 2, metric =
694 "precomputed", n_iter = 1000, learning_rate = 500) to compute the embeddings for all 20,352
695 isolates in our sample. We used these embeddings to visualize the genetic relatedness of the
696 isolates in two dimensions and colored isolates (points on the t-SNE plot) by lineage (**Fig. 7A**).

697 For visualizing specific mutations, isolates were colored according to whether or not the alternate
698 (mutant) allele was called (**Fig. 7B-F**, **S10-S11**).
699
700 **Data Analysis and Variant Annotation**
701 Data analysis was performed using custom scripts run in Python and interfaced with iPython
702 (Pérez and Granger 2007). Statistical tests were run with Statsmodels (Seabold and Perktold
703 2010) and Figures were plotted using Matplotlib (Hunter 2007). Numpy (Van Der Walt et al.
704 2011), Biopython (Cock et al. 2009) and Pandas (McKinney and others 2010) were all used
705 extensively in data cleaning and manipulation. Functional annotation of SNPs was done in
706 Biopython (Cock et al. 2009) using the H37Rv reference genome and the corresponding genome
707 annotation. For every SNP called, we used the H37Rv reference position provided by Pilon
708 (Walker et al. 2014) generated VCF file to extract any overlapping CDS region and annotated
709 SNPs accordingly. Each overlapping CDS regions was then translated into its corresponding
710 peptide sequence with both the reference and alternate allele. SNPs in which the peptide
711 sequences did not differ between alleles were labeled *synonymous*, SNPs in which the peptide
712 sequences did differ were labeled *non-synonymous* and if there were no overlapping CDS
713 regions for that reference position, then the SNP was labeled *intergenic*.
714
715 **Pathway Definitions**
716 We used SEED (Overbeek et al. 2013) subsystem annotation to conduct pathway analysis and
717 downloaded the subsystem classification for all features of *Mycobacterium tuberculosis* H37Rv
718 (id: 83332.1) (**Table S15**). We mapped all of the annotated features from SEED to the annotation
719 for H37Rv. Due to the slight inconsistency between the start and end chromosomal coordinates

720 for features from SEED and our H37Rv annotation, we assigned a locus from H37Rv to a

721 subsystem if both the start and end coordinates for this locus fell within a 20 base-pair window

722 of the start and end coordinates for a feature in the SEED annotation (**Table S16**).

723

724 **REFERENCES**

725 Azad AK, Sadee W, Schlesinger LS. 2012. Innate immune gene polymorphisms in tuberculosis.
726 *Infect Immun* IAI--00443.

727 Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2008. GenBank. *Nucleic Acids Res*
728 **37**: D26–D31.

729 Brennan MJ. 2017. The enigmatic PE/PPE multigene family of mycobacteria and tuberculosis
730 vaccination. *Infect Immun* **85**: e00969-16.

731 Brennan MJ, Delogu G. 2002. The PE multigene family: a 'molecular mantra'for mycobacteria.
732 *Trends Microbiol* **10**: 246–249.

733 Brites D, Gagneux S. 2015. Co-evolution of M ycobacterium tuberculosis and H omo sapiens.
734 *Immunol Rev* **264**: 6–24.

735 Brodin P, Poquet Y, Levillain F, Peguillet I, Larrouy-Maumus G, Gilleron M, Ewann F, Christophe
736 T, Fenistein D, Jang J, et al. 2010. High content phenotypic cell-based visual screen
737 identifies Mycobacterium tuberculosis acyltrehalose-containing glycolipids involved in
738 phagosome remodeling. *PLoS Pathog* **6**: e1001100.

739 Bryant JM, Harris SR, Parkhill J, Dawson R, Diacon AH, van Helden P, Pym A, Mahayiddin AA,
740 Chuchottaworn C, Sanne IM. 2013. Whole-genome sequencing to establish relapse or
741 re-infection with Mycobacterium tuberculosis: a retrospective observational study.
742 *Lancet Respir Med* **1**: 786–792.

743 Casali N, Broda A, Harris SR, Parkhill J, Brown T, Drobniewski F. 2016. Whole genome sequence
744 analysis of a large isoniazid-resistant tuberculosis outbreak in London: a retrospective
745 observational study. *PLoS Med* **13**: e1002137.

746 Chiner-Oms Á, Berney M, Boinett C, González-Candelas F, Young DB, Gagneux S, Jacobs WR,
747 Parkhill J, Cortes T, Comas I. 2019. Genome-wide mutational biases fuel transcriptional
748 diversity in the Mycobacterium tuberculosis complex. *Nat Commun* **10**: 1–11.

749 Clemmensen HS, Knudsen NPH, Rasmussen EM, Winkler J, Rosenkrands I, Ahmad A, Lillebaek T,
750 Sherman DR, Andersen PL, Aagaard C. 2017. An attenuated Mycobacterium tuberculosis
751 clinical strain with a defect in ESX-1 secretion induces minimal host immune responses
752 and pathology. *Sci Rep* **7**: 46666.

753 Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F,
754 Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational
755 molecular biology and bioinformatics. *Bioinformatics* **25**: 1422–1423.

756 Colangeli R, Jedrey H, Kim S, Connell R, Ma S, Chippada Venkata UD, Chakravorty S, Gupta A,
757 Sizemore EE, Diem L. 2018. Bacterial factors that predict relapse after tuberculosis
758 therapy. *N Engl J Med* **379**: 823–833.

759 Cole St, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon S V, Eiglmeier K, Gas S,
760 Barry Iii CE, et al. 1998. Deciphering the biology of Mycobacterium tuberculosis from the
761 complete genome sequence. *Nature* **393**: 537.

762 Coll F, McNerney R, Guerra-Assuncao JA, Glynn JR, Perdigao J, Viveiros M, Portugal I, Pain A,
763 Martin N, Clark TG. 2014. A robust SNP barcode for typing Mycobacterium tuberculosis
764 complex strains. *Nat Commun* **5**: 4812.

765 Comas I, Chakravartti J, Small PM, Galagan J, Niemann S, Kremer K, Ernst JD, Gagneux S. 2010.
766 Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily
767 hyperconserved. *Nat Genet* **42**: 498–498.

768 Copin R, Coscollá M, Seiffert SN, Bothamley G, Sutherland J, Mbayo G, Gagneux S, Ernst JD.
769 2014. Sequence diversity in the pe_pgrs genes of Mycobacterium tuberculosis is
770 independent of human T cell recognition. *MBio* **5**: e00960—-13.

771 Copin R, Wang X, Louie E, Escuyer V, Coscolla M, Gagneux S, Palmer GH, Ernst JD. 2016. Within
772 host evolution selects for a dominant genotype of Mycobacterium tuberculosis while T
773 cells increase pathogen genetic diversity. *PLoS Pathog* **12**: e1006111.

774 Coscolla M, Copin R, Sutherland J, Gehre F, de Jong B, Owolabi O, Mbayo G, Giardina F, Ernst JD,
775 Gagneux S. 2015. M. tuberculosis T cell epitope analysis reveals paucity of antigenic
776 variation and identifies rare variable TB antigens. *Cell Host Microbe* **18**: 538–548.

777 Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. 2016. Within-host evolution of bacterial
778 pathogens. *Nat Rev Microbiol* **14**: 150.

779 Dillon MM, Sung W, Lynch M, Cooper VS. 2015. The rate and molecular spectrum of
780 spontaneous mutations in the GC-rich multichromosome genome of Burkholderia
781 cenocepacia. *Genetics* **200**: 935–946.

782 Dixit A, Freschi L, Vargas R, Calderon R, Sacchettini J, Drobniewski F, Galea JT, Contreras C,
783 Yataco R, Zhang Z, et al. 2019. Whole genome sequencing identifies bacterial factors
784 affecting transmission of multidrug-resistant tuberculosis in a high-prevalence setting.
785 *Sci Rep* **9**: 5602.

786 Epperson LE, Strong M. 2020. A scalable, efficient, and safe method to prepare high quality DNA
787 from mycobacteria and other challenging cells. *J Clin Tuberc Mycobact Dis* **19**: 100150.

788 Ernst JD. 2018a. Mechanisms of M. tuberculosis Immune Evasion as Challenges to TB Vaccine
789 Design. *Cell Host Microbe* **24**: 34–42.

790 Ernst JD. 2018b. Mechanisms of M. tuberculosis Immune Evasion as Challenges to TB Vaccine
791 Design. *Cell Host Microbe* **24**: 34–42.

792 Farhat MR, Freschi L, Calderon R, Ioerger T, Snyder M, Meehan CJ, de Jong B, Rigouts L, Sloutsky
793 A, Kaur D. 2019. GWAS for quantitative resistance phenotypes in Mycobacterium
794 tuberculosis reveals resistance genes and regulatory regions. *Nat Commun* **10**: 2128.

795 Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, Warren RM, Streicher EM,
796 Calver A, Sloutsky A, et al. 2013. Genomic analysis identifies targets of convergent
797 positive selection in drug-resistant Mycobacterium tuberculosis. *Nat Genet* **45**: 1183.

798 Farhat MR, Shapiro BJ, Sheppard SK, Colijn C, Murray M. 2014. A phylogeny-based sampling
799 strategy and power calculator informs genome-wide associations study design for
800 microbial pathogens. *Genome Med* **6**: 101.

801 Farhat MR, Sultana R, Iartchouk O, Bozeman S, Galagan J, Sisk P, Stolte C, Nebenzahl-Guimaraes
802 H, Jacobson K, Sloutsky A. 2016. Genetic determinants of drug resistance in
803 Mycobacterium tuberculosis and their diagnostic value. *Am J Respir Crit Care Med* **194**:
804 621–630.

805 Ford C, Yusim K, Ioerger T, Feng S, Chase M, Greene M, Korber B, Fortune S. 2012.
806 Mycobacterium tuberculosis--heterogeneity revealed through whole genome
807 sequencing. *Tuberculosis* **92**: 194–201.

808 Ford CB, Lin PL, Chase MR, Shah RR, Iartchouk O, Galagan J, Mohaideen N, Ioerger TR,
809 Sacchettini JC, Lipsitch M, et al. 2011. Use of whole genome sequencing to estimate the
810 mutation rate of Mycobacterium tuberculosis during latent infection. *Nat Genet* **43**:
811 482–482.

812 Gagneux S. 2018. Ecology and evolution of *Mycobacterium tuberculosis*. *Nat Rev Microbiol* **16**:
813 202–213.

814 Goig GA, Blanco S, Garcia-Basteiro AL, Comas I. 2020. Contaminant DNA in bacterial sequencing
815 experiments is a major source of false genetic variability. *BMC Biol* **18**: 1–15.

816 Gopinath K, Moosa A, Mizrahi V, Warner DF. 2013. Vitamin B12 metabolism in Mycobacterium
817 tuberculosis. *Future Microbiol* **8**: 1405–1418.

818 Guerra-Assunção JA, Houben RMGJ, Crampin AC, Mzembe T, Mallard K, Coll F, Khan P, Banda L,
819 Chiwaya A, Pereira RPA. 2014. Recurrence due to relapse or reinfection with
820 Mycobacterium tuberculosis: a whole-genome sequencing approach in a large,
821 population-based cohort with a high HIV infection prevalence and active follow-up. *J*
822 *Infect Dis* **211**: 1154–1163.

823 Hebert AM, Talarico S, Yang D, Durmaz R, Marrs CF, Zhang L, Foxman B, Yang Z. 2007. DNA
824 polymorphisms in the pepA and PPE18 genes among clinical strains of Mycobacterium
825 tuberculosis: implications for vaccine efficacy. *Infect Immun* **75**: 5798–5805.

826 Hicks ND, Yang J, Zhang X, Zhao B, Grad YH, Liu L, Ou X, Chang Z, Xia H, Zhou Y. 2018. Clinically
827 prevalent mutations in Mycobacterium tuberculosis alter propionate metabolism and
828 mediate multidrug tolerance. *Nat Microbiol* **3**: 1032.

829 Hunter JD. 2007. Matplotlib: A 2D graphics environment. *Comput Sci Eng* **9**: 90–95.

830 Imperial MZ, Nahid P, Phillips PP, Davies GR, Fielding K, Hanna D, Hermann D, Wallis RS,
831 Johnson JL, Lienhardt C. 2018. A patient-level pooled analysis of treatment-shortening
832 regimens for drug-susceptible pulmonary tuberculosis. *Nat Med* **24**: 1708.

833 Kleinnijenhuis J, Oosting M, Joosten LAB, Netea MG, Van Crevel R. 2011. Innate immune
834 recognition of Mycobacterium tuberculosis. *Clin Dev Immunol* **2011**.

835 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows--Wheeler transform.
836 *Bioinformatics* **25**: 1754–1760.

837 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009.
838 The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

839 Lieberman TD, Flett KB, Yelin I, Martin TR, McAdam AJ, Priebe GP, Kishony R. 2014. Genetic
840 variation of a bacterial pathogen within individuals with cystic fibrosis provides a record
841 of selective pressures. *Nat Genet* **46**: 82.

842 Lieberman TD, Michel J-B, Aingaran M, Potter-Bynoe G, Roux D, Davis Jr MR, Skurnik D, Leiby N,
843 LiPuma JJ, Goldberg JB, et al. 2011. Parallel bacterial evolution within multiple patients
844 identifies candidate pathogenicity genes. *Nat Genet* **43**: 1275.

845 Lieberman TD, Wilson D, Misra R, Xiong LL, Moodley P, Cohen T, Kishony R. 2016. Genomic
846 diversity in autopsy samples reveals within-host dissemination of HIV-associated
847 Mycobacterium tuberculosis. *Nat Med* **22**: 1470.

848 Lin PL, Ford CB, Coleman MT, Myers AJ, Gawande R, Ioerger T, Sacchettini J, Fortune SM, Flynn
849 JL. 2014. Sterilization of granulomas is common in active and latent tuberculosis despite
850 within-host variability in bacterial killing. *Nat Med* **20**: 75–79.

851 Llewelyn MJ, Fitzpatrick JM, Darwin E, SarahTonkin-Crine, Gorton C, Paul J, Peto TEA, Yardley L,
852 Hopkins S, Walker AS. 2017. The antibiotic course has had its day. *BMJ* j3418.

853 Marin M, Vargas R, Freschi L, Epperson LE, Smith M, Oussenko I, Durbin D, Strong M, Salfinger
854 M, Farhat MR. 2020. Genomic sequence characteristics and the empiric accuracy of
855 short-read sequencing. *Prepr Be Posted Soon*.

856 Marvig RL, Sommer LM, Molin S, Johansen HK. 2015. Convergent evolution and adaptation of
857 Pseudomonas aeruginosa within patients with cystic fibrosis. *Nat Genet* **47**: 57.

858 McKinney W, others. 2010. Data structures for statistical computing in python. In *Proceedings*
859 *of the 9th Python in Science Conference*, Vol. 445 of, pp. 51–56.

860 Minias A, Minias P, Czubat B, Dziadek J. 2018. Purifying selective pressure suggests the
861 functionality of a vitamin B12 biosynthesis pathway in a global population of
862 Mycobacterium tuberculosis. *Genome Biol Evol* **10**: 2326–2337.

863 Nair S, Ramaswamy PA, Ghosh S, Joshi DC, Pathak N, Siddiqui I, Sharma P, Hasnain SE, Mande
864 SC, Mukhopadhyay S. 2009. The PPE18 of Mycobacterium tuberculosis interacts with
865 TLR2 and activates IL-10 induction in macrophage. *J Immunol* jimmunol--0901367.

866 Namouchi A, Didelot X, Schöck U, Gicquel B, Rocha EPC. 2012. After the bottleneck: Genome-
867 wide diversification of the Mycobacterium tuberculosis complex by mutation,
868 recombination, and natural selection. *Genome Res*.

869 Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, Edwards RA, Gerdes S, Parrello B,
870 Shukla M. 2013. The SEED and the Rapid Annotation of microbial genomes using
871 Subsystems Technology (RAST). *Nucleic Acids Res* **42**: D206–D214.

872 Pai M, Behr MA, Dowdy D, Dheda K, Divangahi M, Boehme CC, Ginsberg A, Swaminathan S,
873 Spigelman M, Getahun H, et al. 2016. Tuberculosis. *Nat Rev Dis Primer* **2**: 16076.

874 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P,
875 Weiss R, Dubourg V. 2011. Scikit-learn: Machine learning in Python. *J Mach Learn Res*
876 **12**: 2825–2830.

877 Pérez F, Granger BE. 2007. IPython: a system for interactive scientific computing. *Comput Sci*
878 *Eng* **9**.

879 Pethe K, Sequeira PC, Agarwalla S, Rhee K, Kuhen K, Phong WY, Patel V, Beer D, Walker JR,
880 Duraiswamy J. 2010. A chemical genetic screen in Mycobacterium tuberculosis identifies
881 carbon-source-dependent growth inhibitors devoid of in vivo efficacy. *Nat Commun* **1**:
882 57.

883 Phelan JE, Coll F, Bergval I, Anthony RM, Warren R, Sampson SL, van Pittius NCG, Glynn JR,
884 Crampin AC, Alves A, et al. 2016. Recombination in pe/ppe genes contributes to genetic
885 variation in Mycobacterium tuberculosis lineages. *BMC Genomics* **17**: 151.

886 Podinovskaia M, Lee W, Caldwell S, Russell DG. 2013. Infection of macrophages with M
887 ycobacterium tuberculosis induces global modifications to phagosomal function. *Cell*
888 *Microbiol* **15**: 843–859.

889 Rhoads A, Au KF. 2015. PacBio sequencing and its applications. *Genomics Proteomics*
890 *Bioinformatics* **13**: 278–289.

891 Rowley CA, Kendall MM. 2019. To B12 or not to B12: Five questions on the role of cobalamin in
892 host-microbial interactions. *PLoS Pathog* **15**: e1007479.

893 Salaemae W, Azhar A, Booker GW, Polyak SW. 2011. Biotin biosynthesis in Mycobacterium
894 tuberculosis: physiology, biochemistry and molecular intervention. *Protein Cell* **2**: 691–
895 695.

896 Sassetti CM, Boyd DH, Rubin EJ. 2003. Genes required for mycobacterial growth defined by high
897 density mutagenesis. *Mol Microbiol* **48**: 77–84.

898 Sassetti CM, Rubin EJ. 2003. Genetic requirements for mycobacterial survival during infection.
899 *Proc Natl Acad Sci* **100**: 12989–12994.

900 Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets.
901 *Bioinformatics* **27**: 863–864.

902 Seabold S, Perktold J. 2010. Statsmodels: Econometric and statistical modeling with python. In
903 *Proceedings of the 9th Python in Science Conference*, Vol. 57 of, p. 61.

904 Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, Fenner L, Rutaihwa L, Borrell S, Luo T,
905 et al. 2016. Mycobacterium tuberculosis lineage 4 comprises globally distributed and
906 geographically restricted sublineages. *Nat Genet* **48**: 1535.

907 Su H, Zhang Z, Liu Z, Peng B, Kong C, Wang H, Zhang Z, Xu Y. 2018. Mycobacterium tuberculosis
908 PPE60 antigen drives Th1/Th17 responses via Toll-like receptor 2--dependent
909 maturation of dendritic cells. *J Biol Chem* jbc--RA118.

910 Sun G, Luo T, Yang C, Dong X, Li J, Zhu Y, Zheng H, Tian W, Wang S, Barry III CE. 2012. Dynamic
911 population changes in Mycobacterium tuberculosis during acquisition and fixation of
912 drug resistance in patients. *J Infect Dis* **206**: 1724–1733.

913 Tait DR, Hatherill M, Van Der Meeren O, Ginsberg AM, Van Brakel E, Salaun B, Scriba TJ, Akite
914 EJ, Ayles HM, Bollaerts A. 2019. Final Analysis of a Trial of M72/AS01E Vaccine to
915 Prevent Tuberculosis. *N Engl J Med*.

916 Tientcheu LD, Koch A, Ndengane M, Andoseh G, Kampmann B, Wilkinson RJ. 2017.
917 Immunological consequences of strain variation within the Mycobacterium tuberculosis
918 complex. *Eur J Immunol* **47**: 432–445.

919 Trauner A, Liu Q, Via LE, Liu X, Ruan X, Liang L, Shi H, Chen Y, Wang Z, Liang R, et al. 2017. The
920 within-host population dynamics of Mycobacterium tuberculosis vary with treatment
921 efficacy. *Genome Biol* **18**: 71.

922 Van Der Walt S, Colbert SC, Varoquaux G. 2011. The NumPy array: a structure for efficient
923 numerical computation. *Comput Sci Eng* **13**: 22.

924 Vargas R, Farhat MR. 2020. Antibiotic treatment and selection for glpK mutations in patients
925 with active tuberculosis disease. *Proc Natl Acad Sci* **117**: 3910–3912.

926 Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E,
927 Peterson P, Weckesser W, Bright J. 2020. SciPy 1.0: fundamental algorithms for scientific
928 computing in Python. *Nat Methods* 1–12.

929 Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, Wheeler DK, Gabbard
930 JL, Hix D, Sette A, et al. 2014. The immune epitope database (IEDB) 3.0. *Nucleic Acids*
931 *Res* **43**: D405--D412.

932 Votintseva AA, Bradley P, Pankhurst L, del Ojo Elias C, Loose M, Nilgiriwala K, Chatterjee A,
933 Smith EG, Sanderson N, Walker TM. 2017. Same-day diagnostic and surveillance data for
934 tuberculosis via whole-genome sequencing of direct respiratory samples. *J Clin Microbiol*
935 **55**: 1285–1298.

936 Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman
937 J, Young SK, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant
938 detection and genome assembly improvement. *PloS One* **9**: e112963.

939 Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ, Hawkey
940 PM, Crook DW, et al. 2013. Whole-genome sequencing to delineate Mycobacterium
941 tuberculosis outbreaks: a retrospective observational study. *Lancet Infect Dis* **13**: 137–
942 146.

943 WHO. *Definitions and reporting framework for tuberculosis*.
944 https://www.who.int/tb/publications/definitions/en/ (Accessed May 20, 2019).

945 Witney AA, Bateson ALE, Jindani A, Phillips PPJ, Coleman D, Stoker NG, Butcher PD, McHugh TD.
946 2017. Use of whole-genome sequencing to distinguish relapse from reinfection in a
947 completed tuberculosis clinical trial. *BMC Med* **15**: 71.

948 Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact
949 alignments. *Genome Biol* **15**: R46.

950 World Health Organization. 2018. *Global tuberculosis report 2018*. World Health Organization
951 https://www.who.int/tb/publications/global_report/en/.

952 Wyllie DH, Robinson E, Peto T, Crook DW, Ajileye A, Rathod P, Allen R, Jarrett L, Smith EG,
953 Walker AS. 2018. Identifying Mixed Mycobacterium tuberculosis Infection and
954 Laboratory Cross-Contamination during Mycobacterial Sequencing Programs. *J Clin*
955 *Microbiol* **56**.

956 Xu Y, Liu F, Chen S, Wu J, Hu Y, Zhu B, Sun Z. 2018. In vivo evolution of drug-resistant
957 Mycobacterium tuberculosis in patients during long-term treatment. *BMC Genomics* **19**:
958 640.

959 Zhang D, Gomez JE, Chien J-Y, Haseley N, Desjardins CA, Earl AM, Hsueh P-R, Hung DT. 2016.
960 Genomic analysis of the evolution of fluoroquinolone resistance in Mycobacterium
961 tuberculosis prior to tuberculosis diagnosis. *Antimicrob Agents Chemother* **60**: 6600–
962 6608.

963

964 **ACKNOWLEDGEMENTS**

965 We thank the members of the Farhat lab for helpful discussions and comments on the research
966 project and manuscript. We thank S. Fortune, N. Hicks & D. Warner for helpful suggestions on
967 the manuscript. We thank A. Narayan for helpful suggestions on constructing t-SNE
968 visualizations for phylogenetic convergence. R.V.J. was supported by the National Science
969 Foundation Graduate Research Fellowship under Grant No. DGE1745303. Portions of this
970 research were conducted on the O2 High Performance Compute Cluster, supported by the
971 Research Computing Group, at Harvard Medical School.

972

973 **AUTHOR CONTRIBUTIONS**

974 R.V.J. and M.R.F. conceived, designed and conducted the study. R.V.J. and M.R.F. drafted the
975 manuscript with input from all authors. L.F. and M.M. provided bioinformatics support. L.E.E.,
976 D.D., M. Salfinger and M. Strong cultured Mtb isolates and performed DNA extraction in
977 preparation for PacBio sequencing. M.Smith and I.O. prepared libraries and performed PacBio
978 sequencing runs.

979

980 **COMPETING INTERESTS**

981 The authors declare that they have no competing interests.

982

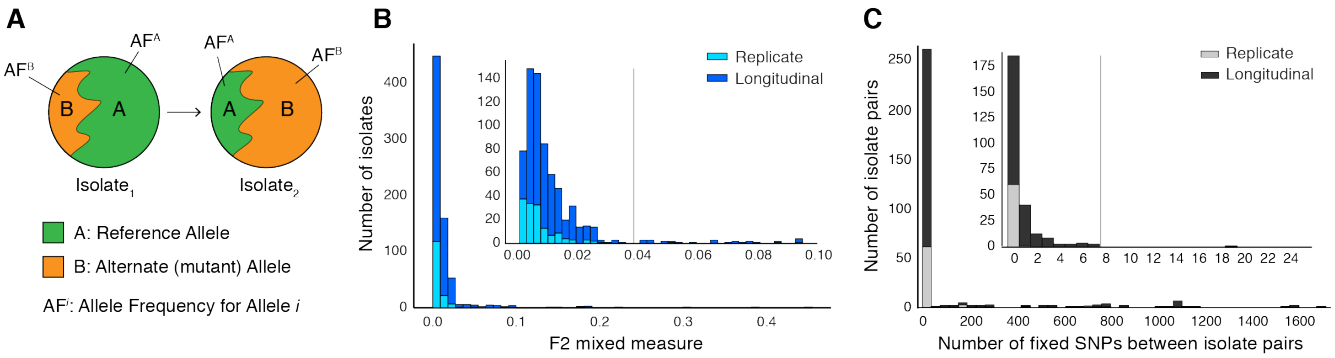
983 **DATA AND MATERIALS AVAILABILITY**

984 All Mtb sequencing data was collected from previously published studies and is publicly
985 available. Individual accession numbers for the Mtb genomes analyzed in this study can be found
986 in **Table S2** and information on which studies from which the data was generated can be found

in the **Methods**, **Fig. S1** and **Table S1**. All packages and software used in this study have been noted in the **Methods**. Custom scripts written in python version 2.7.15 were used to conduct all analyses and interfaced via Jupyter Notebooks. All scripts and notebooks will be uploaded to a GitHub repository upon acceptance of this manuscript for publication.

991 **FIGURES**

992



993

994 **Fig. 1. Selection of patients with longitudinal clonal infection.** (**A**) Allele frequency change

995 between paired isolates $(\Delta AF) = |AF_1^A - AF_2^A| = |AF_1^B - AF_2^B|$. (**B**) The F2 measure >0.04

996 (**Methods**) was used to identify and exclude isolate pairs with evidence for mixed strain growth

997 at any time point. (**C**) Replicate and longitudinal pairs with fixed SNP (fSNP) distance of >7

998 were excluded. For longitudinal isolates fSNP>7 was assessed as consistent with Mtb reinfection

999 with a different strain.

1000

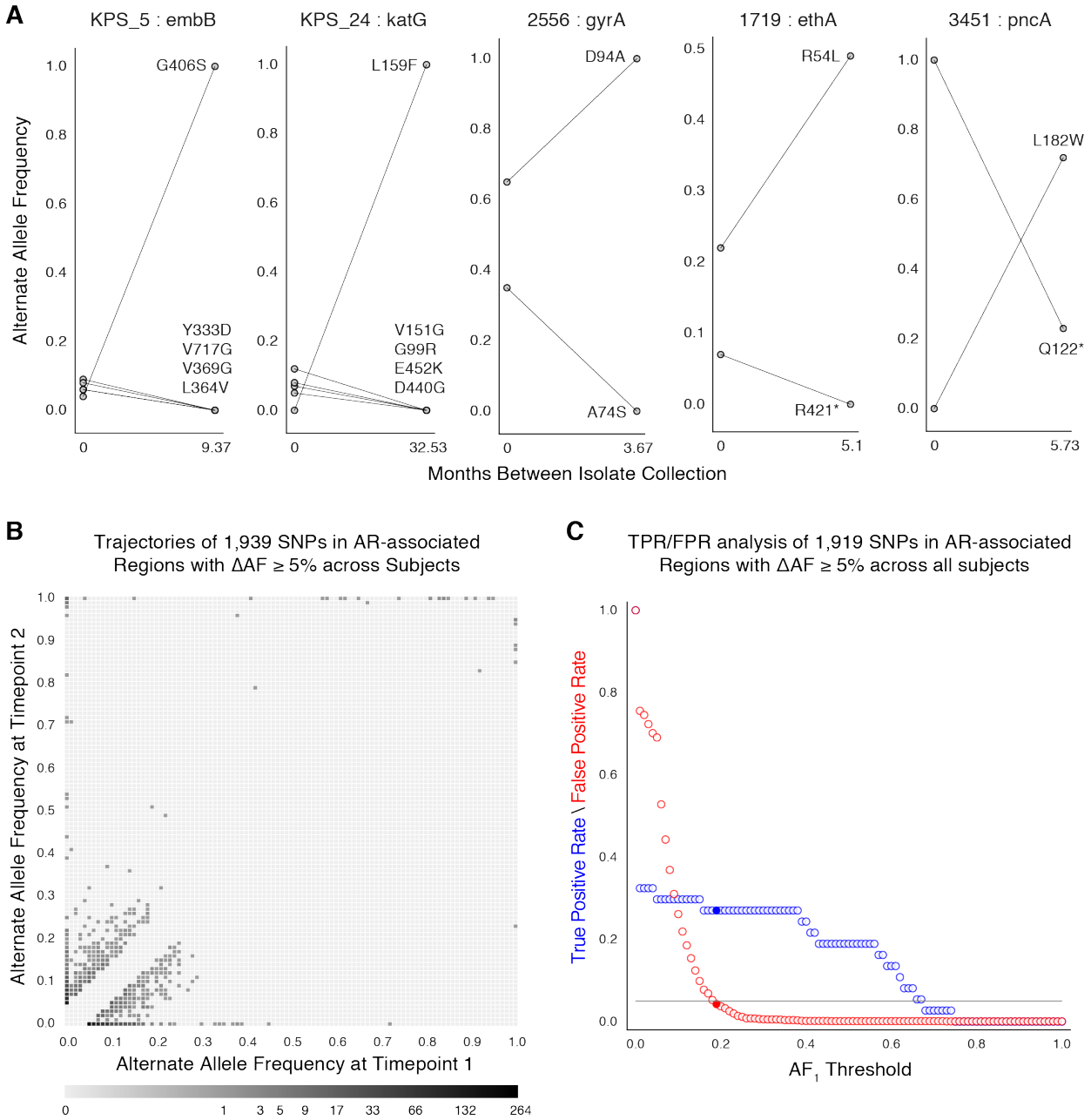


Fig. 2. Allele frequency dynamics within antibiotic resistance loci. (**A**) The antibiotic resistance genes *embB*, *katG*, *gyrA*, *ethA*, and *pncA* demonstrate evidence for competing clones during infection (other examples found are displayed in **Fig. S2**). Each mutant allele is labeled with amino acid encoded by the reference allele, H37Rv codon position, and amino acid encoded by the mutant allele. (**B**) The allele frequency trajectories for SNPs that occur in subjects over the course of infection can be used to study the prediction of further antibiotic resistance using

1008 the frequency of alternate alleles detected in the longitudinal isolates collected from subjects. (**C**)

1009 Plot of true positive rate (TPR) and false positive rate (FPR) for detecting eventual fixation of a

1010 resistance allele ($AF_2 \geq 75\%$) as a function of initial allele frequency (AF_1).

1011

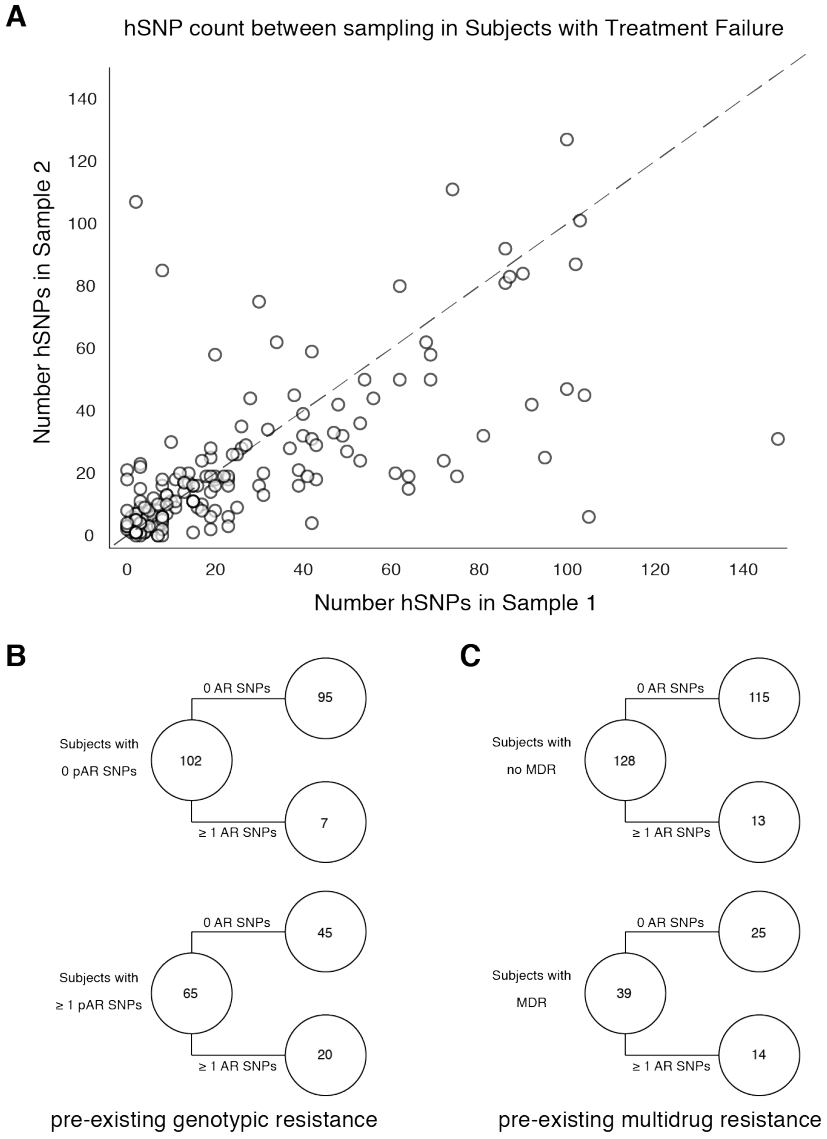


Fig. 3. Pre-existing resistance is associated with resistance amplification. We called heterozygous SNPs (hSNP) in each isolate from a subject with clonal infection classified as failing treatment (N=167). We defined hSNPs as a SNP called in an isolate with an alternate allele frequency between 25% and 75% (**Methods**). (**A**) The number of hSNPs called in the 2nd sample isolated vs the number of hSNPs called in the 1st sample isolated from each of 167 subjects. The dashed line is y = x. (**B-C**) Among subjects who fail treatment, (**B**) subjects with pre-existing mutations that confer antibiotic resistance and (**C**) those that have pre-existing MDR are more likely to acquire antibiotic resistance mutations throughout the course of infection.

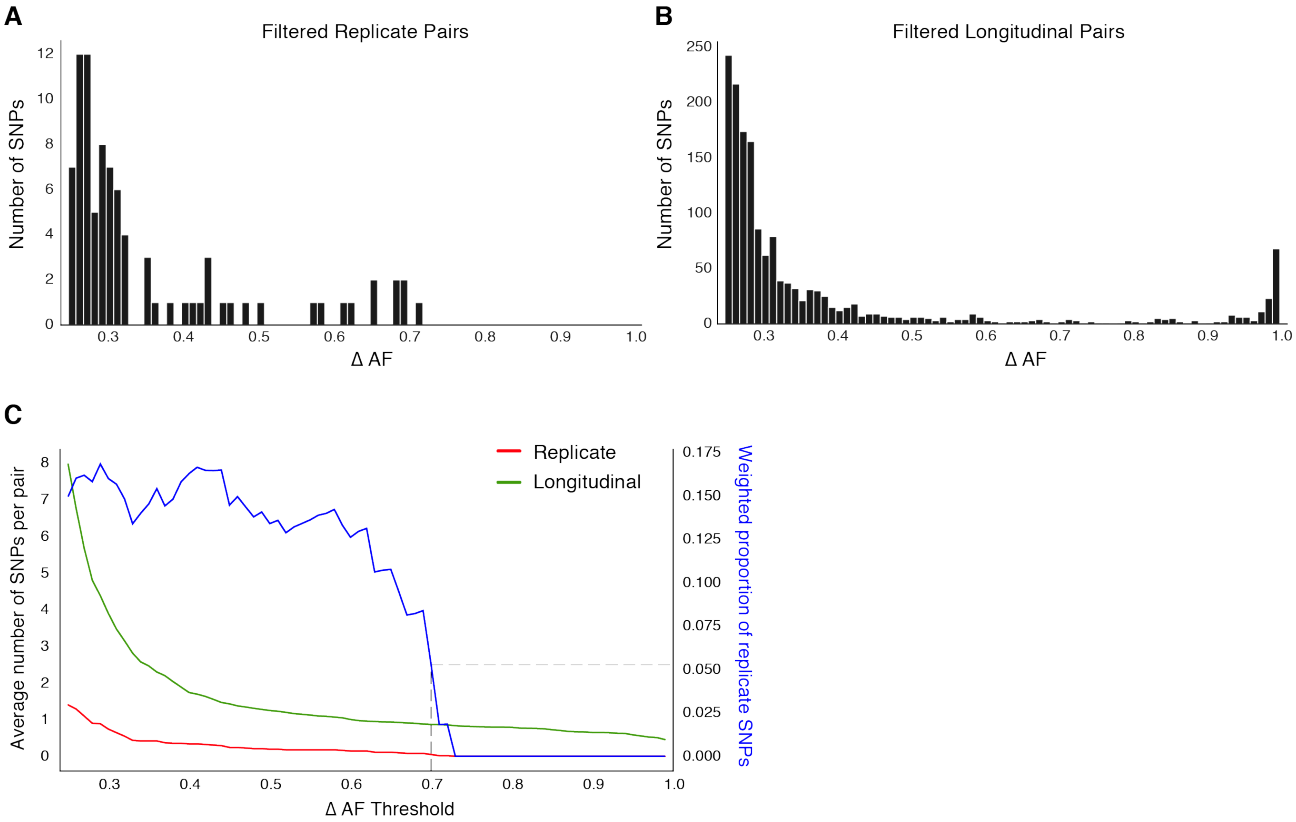


Fig. 4. Replicate pairs reveal levels of biological noise associated with repeated sampling. (**A, B**) We analyzed the distribution of ΔAF for all SNPs detected across all replicate pairs ($m = 62$) and longitudinal pairs ($n = 200$) for SNPs where ΔAF ≥ 25%. (**B**) SNPs were detectable at lower levels of ΔAF for both types of isolate pairs, but SNPs with higher values of ΔAF were only found in longitudinal pairs. (**C**) To determine a ΔAF threshold for calling SNPs representative of changes in bacterial population composition in-host, we calculated the average number of SNPs per pair of isolates at different ΔAF thresholds for both replicate and longitudinal pairs. At a ΔAF threshold of 70% the number of SNPs between replicate pairs represents 5.27% of the SNPs detected amongst all replicate and longitudinal pairs, weighted by the number of pairs in each group.

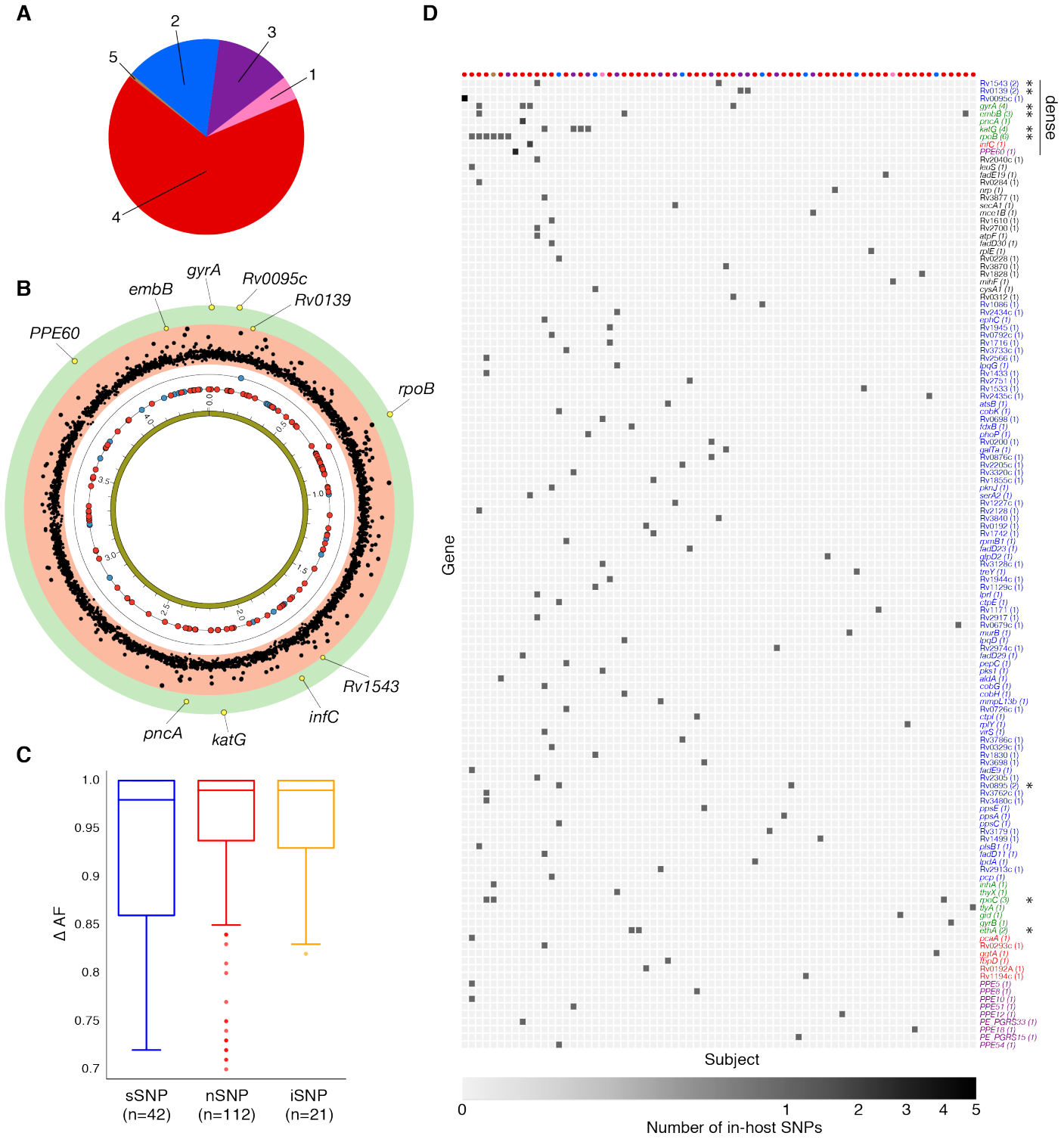


Fig. 5. Genome-wide diversity in 200 clonal Mtb infections. (**A**) Distribution of five major Mtb lineages among the 200 clonal Mtb infections. (**B**) Distribution of 153 in-host SNPs within coding regions among the 200 longitudinal isolate pairs across the 4.41 Mbp Mtb genome (blue

1036 circles: synonymous, red circles: non-synonymous). Blue and red circles on the innermost black
1037 ring indicate the locations of SNPs detected in one patient; circles on the next ring represent
1038 SNPs detected in two patients. The $-\log_{10}$(p-value) of the mutational density test (**Methods**) by
1039 gene is plotted in the outermost, red and green, regions. Labeled yellow circles represent genes
1040 significant at the bonferroni-corrected cutoff ($\alpha = 0.05/3{,}886$). (**C**) Distribution of ΔAF by SNP
1041 type: sSNP: synonymous, nSNP: non-synonymous, iSNP: intergenic. (**D**) Heat-map of SNPs per
1042 gene (rows) and patient (columns). Colored circles across columns indicate the strain
1043 phylogenetic lineage (as represented in **A**). Gene names colored according to gene category (**Fig.**
1044 **6D**) with parentheses indicating the number of subjects with a SNP in a given gene. *Indicates
1045 genes in which SNPs are detected within multiple subjects.

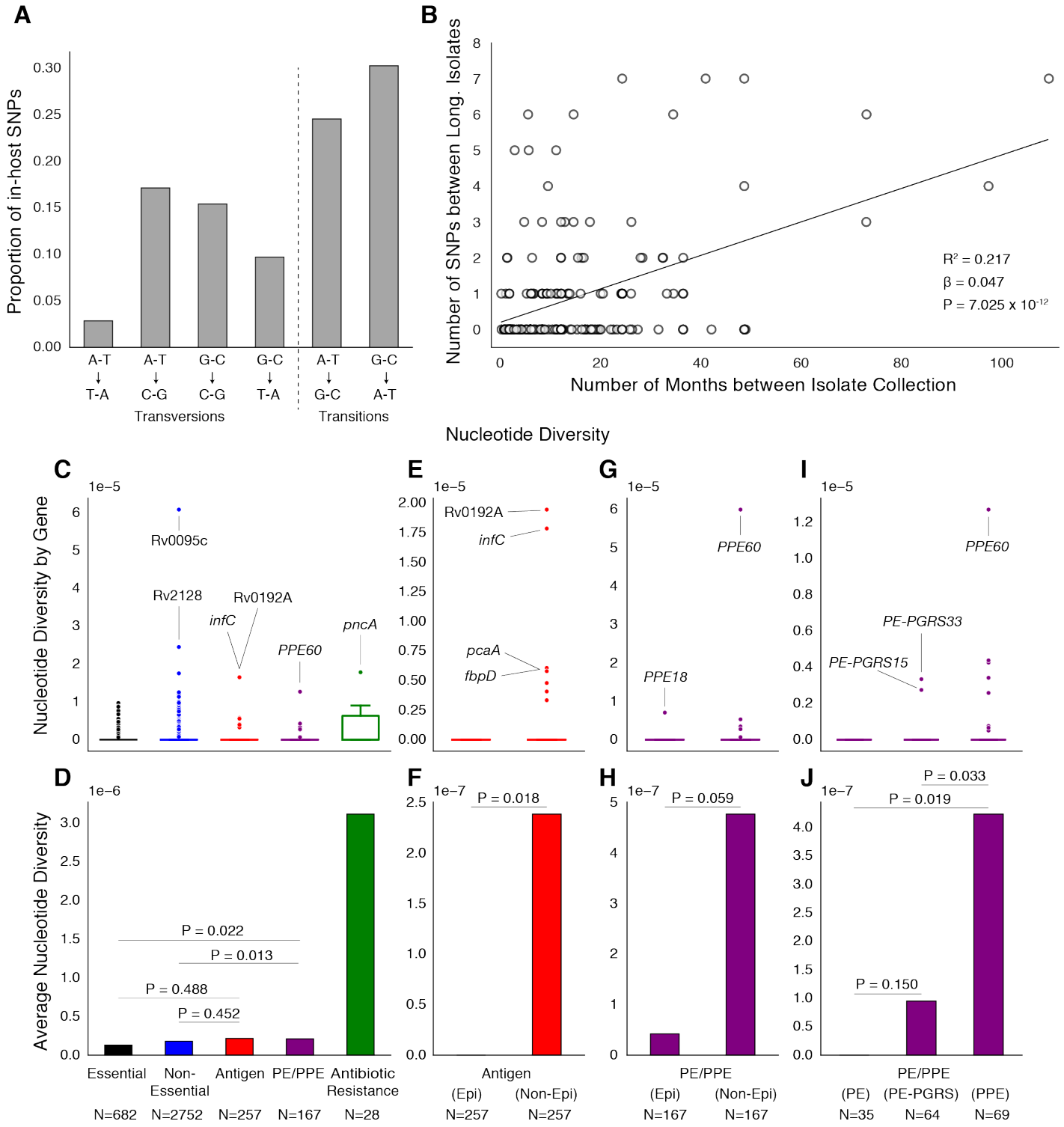


Fig. 6. PE/PPE genes vary considerably within host while putative antigens remain conserved. (**A**) Mutational spectrum of in-host SNPs. (**B**) In-host SNP counts vs. time between isolate collection (195/200 subjects with dates shown, *W (Walker et al. 2013) isolates only had

1050 year of collection). (**C**) boxplots of nucleotide diversity by gene within each of 5 non redundant

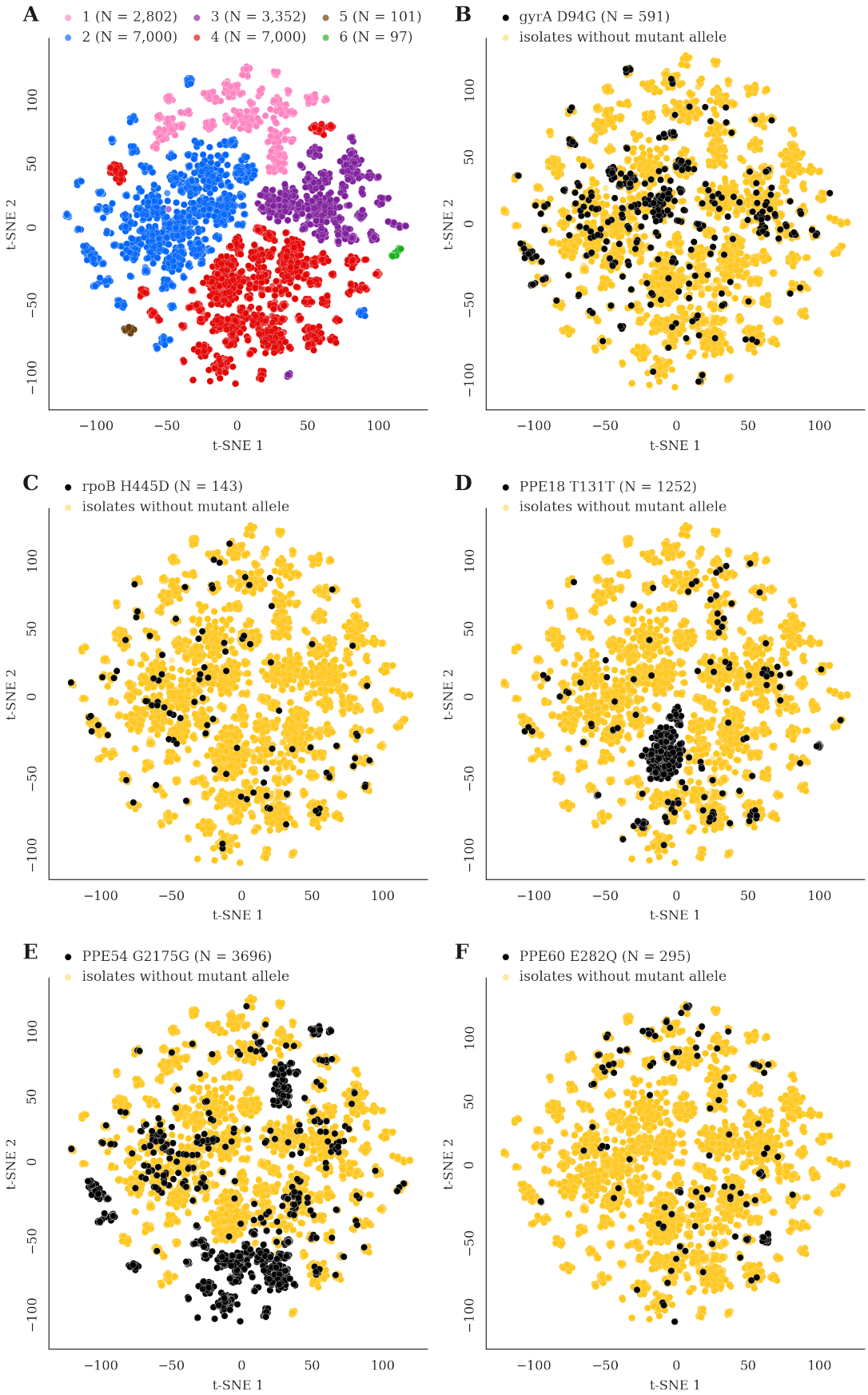
1051 categories (see text). ($n =$ number of genes). (**D**) Average nucleotide diversity across genes by

1052 category. Nucleotide diversity in epitope and non-epitope region (**Methods**) of each gene in the

1053 Antigen (**E, F**) and PE/PPE (**G, H**) gene categories. (**I, J**) PE/PPE genes separated into three non-

1054 redundant categories: PE, PE-PGRS, and PPE. (**J**) The average nucleotide diversity by category.

1055 (**I**) box plot of nucleotide diversity by gene.


A
1 (N = 2,802)
2 (N = 7,000)
3 (N = 3,352)
4 (N = 7,000)
5 (N = 101)
6 (N = 97)
B
gyrA D94G (N = 591)
isolates without mutant allele
C
rpoB H445D (N = 143)
isolates without mutant allele
D
PPE18 T131T (N = 1252)
isolates without mutant allele
E
PPE54 G2175G (N = 3696)
isolates without mutant allele
F
PPE60 E282Q (N = 295)
isolates without mutant allele
t-SNE 1
t-SNE 2
−100
−50
0
50
100

1056

1057 **Fig. 7. Mutations acquired in-host are phylogenetically convergent.** We constructed t-SNE
1058 plots from a pairwise SNP distance matrix for our global sample of 20,352 clinical isolates and
1059 128,898 SNP sites (**Methods**). (**A**) Labeling isolates by global lineage revealed that isolates
1060 cluster according to genetic similarity. Next, we labeled isolates by whether they carried a
1061 mutant allele that was also detected in-host. (**B-F**) Mutations in *gyrA*, *rpoB*, *PPE18*, *PPE54* and
1062 *PPE60* were detected in-host (**Table S8**), occur in a global collection of isolates (**Table S18**) and
1063 are scattered across the tSNE plots, indicating that they belong to genetically different clusters of
1064 isolates (**Table S19**). Furthermore, all mutations with a signal of phylogenetic convergence were
1065 detected in isolates belonging to different clusters, confirming that theses mutations must have
1066 arisen independently in different genetic backgrounds (**Fig S10-S11**). Each plot is labeled with
1067 the gene name each mutation occurs within, amino acid encoded by the reference allele, H37Rv
1068 codon position, and amino acid encoded by the mutant allele. N = number of isolates with mutant
1069 allele.