

1 **Unravelling the population structure and transmission patterns of**  
2 ***Mycobacterium tuberculosis* in Mozambique, a high TB/HIV burden country**

3

4 Saavedra B.<sup>1,2,3#</sup>, López M.G.<sup>4</sup>, Chiner-Oms Á.<sup>4</sup>, García A.M.<sup>5</sup>, Cancino I<sup>4</sup>, Torres-Puente M.<sup>4</sup>,  
5 Villamayor L.<sup>6</sup>, Madrazo C.<sup>4</sup>, Mambuque E.<sup>2</sup>, Sequera VG.<sup>3</sup>, Respeito D.<sup>2</sup>, Blanco S.<sup>2</sup>, Augusto  
6 O.<sup>2</sup>, López-Varela E.<sup>2,3</sup>, García-Basteiro AL.<sup>2,3\*</sup>, Comas I.<sup>4,7\*</sup>

7

8 <sup>1</sup> PhD Program in Medicine and Translational Research. Universitat de Barcelona, Barcelona,  
9 Spain.

10 <sup>2</sup> Centro de Investigação em Saúde de Manhiça (CISM), Maputo, Mozambique.

11 <sup>3</sup> ISGlobal, Hospital Clínic - Universitat de Barcelona, Barcelona, Spain

12 <sup>4</sup> CIBER in Epidemiology and Public Health, Madrid, Spain

13 <sup>5</sup> Universidad de Valencia, Valencia, Spain

14 <sup>6</sup> FISABIO Public Health, Valencia, Spain

15 <sup>7</sup> Instituto de Biomedicina de Valencia, Valencia, Spain

16

17 # Corresponding author

18 \*Both authors contributed equally as senior authors

19

20 **Abstract**

21 Genomic studies of *Mycobacterium tuberculosis complex* (MTBC) might shed light on the  
22 dynamics of its transmission, especially in high-burden settings, where recent outbreaks are  
23 embedded in the complex natural history of the disease. We applied Whole-genome sequencing  
24 (WGS) to characterize the local population of MTBC, unravel potential transmission links and  
25 evaluate associations with host and pathogen factors.

26 **Methods:** A one-year prospective study was conducted in Mozambique, a high HIV/TB burden  
27 country. WGS was applied to 295 positive cultures. We combined phylogenetic, geographical and  
28 clustering analysis, and investigated associations between risk factors of transmission.

29 **Findings:** A significant high proportion of strains were in recent transmission (45.5%). We fully  
30 characterized MTBC isolates by using phylogenetic approaches and dating evaluation. We found  
31 two likely endemic clades, comprised of 67 strains, belonging to L1.2, dating from the late XIX  
32 century and associated with recent spread among PLHIV.

33 **Interpretation:** Our results unveil the population structure of MTBC in our setting. The clustering  
34 analysis revealed an unexpected pattern of spread and high rates of progression, suggesting the  
35 failure of control measures. The long-term presence of local strains in Mozambique, which were  
36 responsible for large transmission among HIV/TB coinfecting patients, hint at possible coevolution  
37 with sympatric host populations and challenge the role of HIV in TB transmission.

38 **Funding:** Ministry of Enterprise and Knowledge (Government of Catalonia & European Social  
39 Fund, AGAUR fellowship); European Research Council (ERC) European Union's Horizon 2020.

40

41 **Keywords:** *Tuberculosis, molecular epidemiology, genomics, transmission, Mycobacterium*  
42 *tuberculosis*

43

## 44 Introduction

45 Tuberculosis (TB) remains the deadliest and one of the most prevalent infectious diseases  
46 worldwide.<sup>1</sup> More than twenty years after the introduction of molecular tools in TB, it is now  
47 undoubted that unraveling the transmission dynamics of local epidemics is essential to tackle  
48 the ongoing spread.<sup>2</sup> A complete understanding of who transmits, where, how and why, is  
49 essential for designing effective control interventions.<sup>3</sup>

50 The development of Whole-genome Sequencing (WGS) techniques is transforming the classical  
51 vision of the infection and decoding links unreachable by basic epidemiology or traditional  
52 genotyping.<sup>4</sup> However, WGS is still scarcely applied in high-burden settings, where recent  
53 outbreaks are embedded in the complex natural evolution of the disease.<sup>5</sup>

54 On one hand, high-resolution genomic data allows for accurate cluster analysis to investigate  
55 outbreaks and decrypt transmission profiles. By using the SNPs cut-off approach, we can  
56 delineate transmission clusters with a resolution not seen before.<sup>4</sup> On the other hand, the study  
57 of the origin and genetic diversity of local strains also helps to survey the evolutionary history of  
58 the *Mycobacterium tuberculosis complex* (MTBC) population and the potential variations between  
59 lineages.<sup>5,6</sup>

60 Phylogeographical methods have identified different lineages and sub-lineages of MTBC with  
61 differential distribution worldwide.<sup>7</sup> This striking difference in geographic distribution has led to  
62 hypotheses as to why some are more widespread than others. Whereas Lineage (L) 4 is the most  
63 ubiquitous globally, L1 and L3 have been described as endemic to several high-burden regions  
64 in Asia and Eastern Africa.<sup>8</sup> Although part of those differences can be explained by historical  
65 contingency<sup>9</sup>, host-pathogen coevolution is assumed to play a decisive role.<sup>10</sup> In San Francisco,  
66 In San Francisco, there seemed to be a preference of transmission between lineages and hosts  
67 with the same geographic origin.<sup>11</sup> Furthermore, some works suggest that this link between  
68 human populations and lineages from the same geographic origin breaks down with HIV status,  
69 suggesting a role for coevolution at immune system level.<sup>11</sup> However, identifying the causes why  
70 a genotype succeeds locally remains a challenge and there is limited information on the  
71 interaction with other factors, such as HIV infection.

72 All the insights we are gaining from WGS data should be transferred to high-burden settings and  
73 used to define effective public health interventions. Mozambique is a country with one of the  
74 highest rates of TB and HIV/TB coinfection<sup>1</sup>, but information on MTBC population structure is  
75 scarce in the region.<sup>12,13</sup> This is, to our knowledge, the first comprehensive population-based  
76 study in Mozambique on the application of WGS to characterize local strains, unravel likely  
77 transmission links, and evaluate possible associations with host and pathogen factors.

## 78 **Methods**

### 79 Study design and study population

80 This was a one-year prospective surveillance-based study (TOSSE study) implemented from  
81 August 2013 to 2014. It was conducted in the District of Manhiça, Maputo province, a semi-rural  
82 area in Southern Mozambique, 80 kilometers north of the capital, with an estimated population of  
83 186,241 inhabitants at the time of the study (2013-2014)<sup>14</sup>, living in an area of 2,373 km.<sup>15</sup> This is  
84 high TB/HIV burden area with a long history of high TB transmission.<sup>16,17</sup>

85 Presumptive TB adults, without history of previous TB treatment, presenting with TB compatible  
86 symptoms<sup>18</sup> (no time criteria if HIV-positive), who attended any of the health units belonging to  
87 Manhiça District Hospital's catchment area, were consecutively enrolled.<sup>19</sup> Only participants with  
88 confirmed MTBC in culture were included in the analysis.

89 We consecutively enrolled participants who met the following criteria: presumptive TB-infected  
90 adults, without history of previous TB treatment, presenting TB-compatible symptoms (no time  
91 criteria in the case of people living with HIV), attending any of the health units belonging to  
92 Manhiça District Hospital's catchment area.

### 93 Diagnostic procedures

94 Participants provided two sputum samples at time of diagnosis. Extra-pulmonary specimens were  
95 collected at hospital level. Diagnostic tests were performed at the *Centro de Investigação em*  
96 *Saúde de Manhiça* (CISM) - Biosafety level 3 (BSL3) laboratory, which is subject to external  
97 quality control and is ISO certified.

98 **Ziehl-Neelsen (ZN)**: Smear microscopy was done by Ziehl-Neelsen staining. Results were  
99 reported as negative or on a scale of positive grades according to international standards.<sup>20</sup>

100 **Xpert MTB/RIF (Xpert):** Raw samples were tested according to the manufacturer instructions.<sup>21</sup>  
101 Invalid results were excluded. The semiquantitative results for Xpert fell under the following  
102 categories: very low, low, medium, or high.

103 **Solid and liquid culture:** Samples with positive Xpert results were cultured. Remnant raw  
104 samples were decontaminated by Kubica method<sup>23</sup> and resuspended. Afterwards, 500 microliters  
105 were inoculated into Mycobacteria Growth Indicator Tubes (MGIT) liquid medium and incubated  
106 in the Bactec MGIT 960 mycobacterial detection instrument (Becton Dickinson Microbiology  
107 System, BD, USA). Additionally, 200 microlitres were cultured in BD Lowenstein Jensen solid  
108 medium. After 42 days (for liquid culture) or 8 weeks (for solid culture) without growth, samples  
109 were classified as negative. In case of positive results, MTBC was confirmed using ZN staining  
110 and BD TB Identification test (Becton Dickinson Microbiology System, USA). Isolates were stored  
111 at -80°C.

#### 112 Sequencing library construction and bioinformatics pipeline

113 Culture isolates were shipped to the BSL3 laboratory of *the Instituto de Biomedicina de Valencia*  
114 (Spain). After inactivation, samples were used to prepare WGS libraries. Genomic libraries were  
115 constructed with the Nextera XT Sample Preparation kit (Illumina Inc., San Diego, CA) according  
116 to the manufacturer's protocol<sup>24</sup>, with 12 cycles for indexing PCR. WGS was carried out in the  
117 MiSeq platform (2×300 cycles paired-end run; Illumina).

118 Sequence analysis was performed following a validated, previously-described bioinformatics  
119 pipeline.<sup>25</sup> Briefly, FASTQ files were preprocessed with *fastp*<sup>26</sup> in order to trim poor quality bases  
120 and potential sequencing errors. Later, in order to reduce likely contaminant reads, we classified  
121 and filtered those that did not belong to the MTBC using Kraken.<sup>27</sup> Samples with less than 90%  
122 of MTBC reads were discarded for posterior analyses. After this filtering, reads were mapped  
123 against the MTBC most probable ancestral genome<sup>28</sup> using the BWA-mem algorithm.<sup>29</sup> Later, we  
124 discarded reads with ambiguous mapping based on the BWA MAPQ score (keep those with  
125 MAPQ=60) as well as potential duplicate reads by using *picard* tools. Samples with genomic  
126 coverage <90% were discarded. The variant calling was performed by a combination of SAMtools  
127 and VarScan.<sup>30</sup>

128 In order to avoid mapping errors and false SNPs, we kept variants that (i) were supported by at  
129 least 20 reads, (ii) were found in a frequency of at least 0.9, (iii) were not found inside detected  
130 indel areas, or (iv) were found in areas of high accumulation of variants (more than three variants  
131 in a 10-bp defined window). Variants were annotated using SnpEff.<sup>31</sup> Variants present in PE/PPE  
132 genes, phages, or repeated sequences were not considered. With these high-quality variants  
133 detected, we generated the alignment. Samples with at least two phylogenetic variants at  
134 frequency >10% were classified as mixed infection cases.

### 135 Phylogenetics and geographical analysis

136 Maximum-likelihood phylogeny was constructed with IQ-TREE<sup>32</sup> under the General Time  
137 Reversible (GTR) model of evolution, with a bootstrap of 1000 replicates and using the *-fconst*  
138 option for accounting for invariant sites. Known drug-resistant positions were not considered for  
139 generating the phylogeny as they are highly homoplastic. Later, we aimed to define the population  
140 structure of closely related strains by using the fast hierarchical Bayesian Analysis of Population  
141 Structure (BAPS) algorithm implemented in R library *fastbaps*. BAPS groups were defined under  
142 the second level of clustering hierarchy.<sup>33</sup>

143 **Geographic origin of the clades from Mozambique (MZ):** In order to unravel the potential  
144 geographic origin of the BAPS groups, we used the RASP program.<sup>34</sup> This is based on both,  
145 Bayesian and parsimony approaches, and aims to estimate the ancestral geographic origin. It  
146 requires a phylogeny and the geographic origin of the tips as input. We reconstructed a phylogeny  
147 combining MZ isolates and 8,263 genomes representative of the MTBC global genetic diversity.  
148 We marked MZ strains as having their geographical origin in Mozambique, and strains from other  
149 datasets as having non-Mozambique origin. We coupled the Statistical-Dispersal Vicariance  
150 Analysis (S-DIVA) with a Bayesian Binary MCMC. For each analysis, we run 5 MCMC chains with  
151 500000 cycles.

152 RASP output estimates the likely origin of ancestral nodes. We manually defined likely endemic  
153 clades as follow: i) they consisted in 1 or more BAPS, ii) the clade contained more than 10  
154 samples from Mozambique and constituted >60% of the total clade, iii) the origin of the most  
155 recent common ancestor (MRCA) obtained with RASP was Mozambique (>80% of probability).  
156 Non-endemic clades were defined as those in which i) the clade contained more than 10 samples

157 from Mozambique but constituted <60% of the total clade, and ii) the MRCA was not  
158 Mozambican. Unknown clades were those which did not meet any of the previous criteria.

### 159 Estimation of times

160 Dating analysis was performed by applying the Bayesian inference implemented in BEAST2  
161 v2.6.5<sup>35</sup> A multifasta file was generated for the three major lineages (L1, L2 and L4). We used  
162 them as partitions in *beauti*. A GTR substitution model was defined with gamma count = 4 and  
163 empirical nucleotide frequencies. A strict clock model was selected; since we do not have tip or  
164 node calibration, we fixed different clock rates for each lineage (L1:  $1.57 \times 10^{-7}$ ; for L2:  $4.1 \times 10^{-8}$ ;  
165 L4:  $3.79 \times 10^{-8}$ ), according to *Menardo et al, 2019*.<sup>36</sup> We evaluated exponential and constant  
166 coalescent population models, and selected the former, since we found evidence against the  
167 constant model (95% HPD of the rate growth not including 0). We chose an exponential prior for  
168 the population size (mean = 1, offset = 0), and for the exponential growth rate prior, we used the  
169 standard Laplace distribution (mu = 0.001, scale = 30.7). We corrected the xml files to specify  
170 the number of invariant sites as indicated here: [https://groups.google.com/g/beast-](https://groups.google.com/g/beast-users/c/QfBHMOqImFE)  
171 *users/c/QfBHMOqImFE*. We ran two runs with  $2 \times 10^7$  steps-long chains, sampling every 2000  
172 steps and removing the initial 10% as burn-in. We evaluated that the mixing and the estimated  
173 sample size (ESS) of the posterior and of all parameters were larger than 200, with Tracer  
174 1.7.1.<sup>37</sup> We used *logcombiner* to combine the tree files from the independent runs and *ggtree*<sup>38</sup>  
175 R package to annotate and plot the trees.

176 Lastly, by combining results from molecular dating and the likely ancestral geographical origin  
177 from RASP, we aimed to assess for how long previously defined clades have been circulating  
178 within the country. We calculated the median and 95%HPD (Highest Posterior Density interval)  
179 for MRCAs with origin in Mozambique for all clades, and we compared the average of the median  
180 years of introduction in the country among endemic and foreign clades.

### 181 Identification of recent transmission events

182 Transmission was evaluated by clustering analysis from pairwise distance. We employed the strict  
183 relatedness cut-off of 5 SNPs for describing recent transmission<sup>39</sup> and we evaluated up to 10  
184 thresholds for broad cluster definition. Clusters obtained from Mozambique (estimated incidence

185 rate (IR): 551/100,000 in 2014) were compared to published transmission data (from population-  
186 based studies) from settings with different TB burden (Malawi, IR 2009 243/100,000; Valencia,  
187 Spain, IR 2015: 9.13/100,000). From available datasets, we chose annual data from 2009<sup>40</sup> and  
188 2015 (manuscript under revision) respectively, based on the largest datasets close to 2013/4 and  
189 to have comparable timespan.

## 190 Statistical analysis

191 Statistical analyses were performed using the R statistical language (R version 3.5.2, The R  
192 Foundation for Statistical Computing Platform). For the phylogenetic analysis, branch lengths  
193 were extracted using the *geiger*<sup>41</sup> package. The *fastbaps*<sup>42</sup> package was applied for the BAPS  
194 algorithm. For transmission analysis, transmission events were identified by using *ape* and  
195 *adegenet* packages. *Tidyverse*, *tigerstats*, *naniar*, *epiR*, *purrr*, *broom*, *ggplot* and *parameters*, were  
196 used to visualize, describe and analyze epidemiological data.

197 We aimed to explore the association of being in transmission (dependent variable) with a range  
198 of collected risk factors (Table 1). The non-parametric Fisher's exact test was used to identify  
199 differences in the distribution of independent variables. We hypothesized that transmission  
200 patterns would differ depending on the origin of clades (endemicity). Saturated logistic regression  
201 models were used to estimate mutually adjusted odds ratio (OR). Covariates with p-value<0.2,  
202 passed for the adjusted analysis. Age, sex and HIV status were chosen as *a priori* related risk  
203 factors irrespective of univariate associations. A backward strategy was used to finalize the  
204 model. Afterwards, considering the large prevalence of HIV-positive patients in our dataset, and  
205 that this unbalanced distribution might confound our findings, we stratified the analysis by HIV  
206 status (Table 2).

## 207 Ethics

208 The CISM's Internal Scientific and Internal Bioethics Committees and the National Bioethics  
209 Committee (Maputo, Mozambique) waived ethical approval for this work (Reference:  
210 199/CNBS/13). Written informed consent was provided by all study participants

## 211 **Results**

### 212 **1. Overall population structure of MTBC strains in Manhiça district**



213 From those 580 patients who started treatment during the study period in the district, 302 were  
214 microbiologically confirmed by culture, although 7 of those strains were not available for WGS  
215 due to poor quality of cultures. After exclusions (Figure 1), 275 strains were included in the  
216 analysis on population structure.

217 Overall, when reconstructing the phylogenetic tree (Figure 2), most cases (49.5%, 136/275) were  
218 classified as L4. The L4.3.4 (LAM) sub-lineage was the most common within L4 (23.3%, 64/275)  
219 although, at this level the most prevalent sublineage was L1.2 (25.4%, 70/275). The Beijing L2  
220 were classified as L2.2.1, representing 12.7% (35/275) of the total population (Supplementary S1.  
221 Table 1). Additionally, eight samples were determined as mixed infections.

## 222 **2. SNPs-threshold approach to evaluate transmission events (Figure 4)**

223 The clustering analysis revealed that 57.8% of strains (159/275), were in transmission, applying  
224 a cut-off of 5 SNPs. To evaluate the 'transmission profile' we compared genetic distances up to  
225 10 SNPs with previously published data from Malawi (2009) and Valencia (2015) (See methods  
226 section). The distribution of pairwise distances revealed that, in all three regions, a large  
227 proportion of cases were in transmission, although patterns differed from one population to  
228 another, towering in Mozambique (64.7%) compared to Malawi (41.2%), and Valencia (29.8%).

229 While TB transmission seemed continuous for all three sites, the burden of very recent spread  
230 was significantly higher in Mozambique: 45.5% (125/275) of strains shared identical genotype (0  
231 SNPs genetic distance), whereas only 9.2% (11/119) and 15.9% (41/246) were identified for  
232 Malawi and Valencia, respectively.

## 233 **3. Fine-scale population structure of MTBC**

234 By applying the BAPS algorithm, we recognized 25 groups of strains that shared high genomic  
235 similarity at level 2 of clustering hierarchy: 9 from L1, 6 L2, 2 L3, 8 L4 (1 strain did not assemble)  
236 (Figure 2). These groups were mapped in a global phylogeny (Figure 3) and, through the use of  
237 RASP, we inferred the probable geographic origin of the most recent common ancestor (MRCA)  
238 for each group (see methods).

239 BAPS 1, 2, 3, 12, 13 and 14 were enriched in Mozambique isolates. BAPS 1 to 3, and 12 to 14  
240 formed two largest groups sharing a MRCA with origin in the country (probability of 80%).

241 Therefore, we merged those BAPS into two clades, and considered them 'likely endemic' or 'local'  
242 in our study setting (E1 and E2, 67/275, 24.4%) (Supplementary S2). Eight clades were defined  
243 as probably 'non-endemic' candidates (164 /275, 59.6%), and 11 did not accomplish defined  
244 criteria to be classified (see Methods). All strains categorized as endemic belonged to L1.2.

245 We then inferred the age of probable endemic and non-endemic clades (Supplementary S3). We  
246 found that, for those considered likely endemic, MRCAs located in Mozambique dated further  
247 back in time than in non-endemic ones. The mean of the median time when those local clades  
248 began to circulate in the country was 1894 [IQR 1887.7; 1900.6], whereas for non-endemic, it  
249 was estimated to be 2009 [IQR 1876.9; 2014.0] (Figure 4, Supplementary S4. Figure 4.1 & Figure  
250 4.2).

#### 251 **4. Impact of collected risk factor and endemicity on transmission**

252 From those patients with WGS data, 223 strains had laboratory and epidemiological data  
253 available (Figure 1). These were used to investigate potential covariates associated with  
254 clustering, defined using a restrictive cut-off of pairwise distance at 5 SNPs to evaluate the burden  
255 of very recent transmission (see methods).

256 The median age of patients was 35 years [IQR 28.2;45.0], 35.4% (79/223) were female and the  
257 majority (72.2 % , 161/223) were people living with HIV (PLHIV). The median CD4+ count was  
258 145.5 [49.9; 321.8] cells /mm<sup>3</sup>. The most common symptom was cough presented in 93.7% of  
259 cases (209/223). A sputum smear for acid-fast bacilli was positive in 65.0% (145/223) and 52.1%  
260 of participants had a pathological chest X-ray (CXR) (87/167 available). In this subset of data,  
261 52.9% of isolates (118/223), were included in a cluster. Univariate exploration of other  
262 sociodemographic and microbiological variables is displayed in Table 1. The distribution of strains  
263 in recent transmission did not vary by sex, age, HIV status, symptoms, type of TB, different areas  
264 of the district or sublineage. Although none of the resistant strains were clustered, we were unable  
265 to explore this association further due to the limited number of isolates.

266 Based on the hypothesis that local genotypes might be linked to transmission, we explored the  
267 association between 'endemicity' and clustering. Overall, 24.2% (54/223) of strains were  
268 classified as 'likely endemic' (hereinafter endemic) and 28.0 % of them were in cluster (33/223).

269 The results of the best fitted logistic regression model included are shown in Table 2. The odds  
270 of clustering did not differ by clades, whether endemic or not. Subsequent stratification by HIV  
271 status revealed that PLHIV who were infected with endemic strains had more than two times  
272 higher odds of being in recent transmission, regardless sex and age (2.11, 95%CI [1.04;4.45], p-  
273 value=0.04), and that this association disappeared among HIV-negative individuals.

274 Lastly, in light of those results, we investigated whether the degree of immunosuppression,  
275 measured through CD4 levels, might shape the distribution of endemic clades. We found that  
276 although the proportion of endemic strains increased as decreased the CD4 counts, this raise did  
277 not result meaningful (p-value=0.134) (Supplementary S5).

## 278 **Discussion**

279 High resolution WGS data on MTBC population structure and transmission patterns is scarce in  
280 the country<sup>13,43</sup>, and similar to the situation for most high-burden countries. In order to provide a  
281 comprehensive insight on the molecular epidemiology of MTBC in Southern Mozambique, we  
282 have conducted the first population-based study using WGS data. The clustering analysis  
283 unmasked a high recent transmission profile. We also found two large likely-endemic clades that  
284 were associated with recent transmission among PLHIV. These clades were dated back to the  
285 end of the XIX century.

286 When pairwise distances were compared to one-year data from Malawi and Spain, profiles  
287 revealed that the highest proportion of strains involved in transmission was in Mozambique  
288 (64.7%), and was also higher than that found in other high-burden settings (Liberia (39%, cut-off  
289 of 12 SNPs)<sup>44</sup> and reported in Malawi for 15 years (31%).<sup>40</sup> Although high rates of ongoing  
290 transmission were expected<sup>3,45</sup>, almost half of the strains were connected by 0 SNPs. Assuming  
291 a molecular clock of 0.3-0.5 SNPs/year, this would translate into very recent transmission events<sup>46</sup>  
292 and fast rates of progression to active tuberculosis. Importantly, this proportion might be even  
293 underestimated as one-year population analyses are not enough to reveal the extent of recent  
294 transmission.

295 Interrupting ongoing spread is a critical public health intervention to bring down the current TB  
296 burden. Presented data reveals the suboptimal control of the epidemic in our setting and reinforce

297 the need to explore the homegrown scenario. Most of participants were severely ill at the time of  
298 diagnosis, which might support previous works describing low awareness of disease<sup>47,48</sup> and/or  
299 the existence of barriers in accessing medical care in our setting.<sup>49</sup> Besides, programmatic active-  
300 case finding activities have been described as inconsistent in Mozambique, failing to reach hidden  
301 cases and breaking the transmission net.<sup>50</sup> In our cohort of microbiologically-confirmed cases  
302 (highly transmitters), 20% of participants had contact with a TB case, although they sought health  
303 care because they were symptomatic (presumptive), proving the lack of active screening activities  
304 to identify patients in early stages.

305 The globally spread L4 was the most prevalent in our region, in agreement with previous works<sup>10</sup>,  
306 although at sublineage level, L1.2 was the most frequent. This finding is aligned with the global  
307 distribution of L1, concentrated around the Indian Ocean and that has been supposed to spread  
308 from South-Asia as a result of old migration pathways.<sup>51</sup> To define clades specific to Mozambique,  
309 i.e. 'probably endemic', we applied a detailed, fine-scale population structure approach that  
310 allowed us to define two major endemic clades, both belonging to the L1.2 sub-lineage. In  
311 accordance to the hypothesis that they are endemic to our study setting, dating estimates  
312 confirmed they would have been circulating in Mozambique for longer times than non-endemic  
313 ones.

314 The long-term presence of those local strains in Mozambique suggests possible coevolution with  
315 the sympatric host population, as seen elsewhere.<sup>52</sup> Thus, we tested the likelihood that endemic  
316 clades, as well as other collected covariates, could be related to a higher probability of recent  
317 spread using 5 SNPs as a threshold.<sup>39</sup> Although through non-stratified models we did not find any  
318 remarkable result, posterior stratification by HIV status revealed that those classified as probably  
319 endemic presented higher odds of being responsible for recent transmission among HIV positive  
320 patients.

321 The extent to which TB/ HIV coinfection influences the structure of the MTBC population and the  
322 efficacy of transmission is currently under debate.<sup>53,54</sup> On one hand, studies evaluating MTBC  
323 infectiousness among PLHIV are widely heterogeneous.<sup>55,54</sup> Historically it has been argued that  
324 HIV decreases transmission, due to EPTB, lower cavitory lesions, or lower sputum bacillary load,  
325 among others.<sup>53,56</sup> Nevertheless, recent works state that HIV-seropositive patients are as

326 infectious as seronegative when they present cavitory disease or have a positive bacilloscopy.<sup>54</sup>  
327 <sup>57</sup> Furthermore, among PLHIV, most cases of active tuberculosis are the result of new  
328 infections.<sup>58</sup> Therefore, for our setting, the large proportion of strains in recent spread may be  
329 due to the fact that half of PLHIV participants were severely immunosuppressed (highly  
330 susceptible to develop disease), microbiologically confirmed, and just 5% of them had  
331 extrapulmonary disease.

332 On the other hand, studies on coevolution have also demonstrated increased transmission  
333 between sympatric lineage-host associations<sup>11</sup>, although, differently to our results, this was  
334 disrupted by HIV coinfections. Yet, the interdependence of both epidemics in high burden settings  
335 needs to embrace socioeconomical determinants that may shape the circulation and spread of  
336 MTBC and confound results.<sup>56</sup> Further studies on how HIV may influence sympatric associations  
337 in high burden areas are needed. Alternatively, our local population, regardless HIV status, may  
338 represent a reservoir for endemic strains with lower transmissibility, with some data suggesting  
339 that this is the case for the L1 specialist genotype.<sup>59</sup> Limitations to understand this interplay  
340 includes the fact that the ancestry of individuals involved in this study is not known and that  
341 transmission estimates are based on a one-year sampling window. Although this may have limited  
342 our result, given the fast progression of the disease in PLHIV and the particularity of our dataset  
343 (45.5% of strains were 0 SNPs pairwise distance), one-year data may indeed be give us a hint of  
344 what is happening in our region.

345 Lastly, we cannot forget that host-related factors must be considered when interpreting genomic  
346 data. Information on how host and pathogen genotypes interplay is still scarce and few genotype-  
347 genotype interaction studies are available.<sup>7</sup> Since the adaptive immune response is an essential  
348 mechanism for host recognition and control of MTBC<sup>60</sup>, we hypothesize that the interaction of  
349 HIV, MTBC and human populations is more complex in countries where the two epidemics collide  
350 with downstream impact on TB transmission. We highlight the need for population-based  
351 genome-to-genome association studies, including the MTBC, human and HIV genotype  
352 combined with sociodemographic data that could potentially confound results.

353

354

355 **Conclusion**

356 Overall, our results unveil the population structure of MTBC in a high TB and HIV setting. We  
357 found an unexpected pattern of spread, with a substantial amount of strains of recent  
358 transmission, suggesting the uncontrolled TB spread and high rates of progression to active  
359 disease. We also identified endemic groups that were estimated to be circulating in the country  
360 for more than one century and that were responsible for recent transmission among PLHIV.

361 **Contributors**

362 All authors contributed as investigators to this study and to drafting and revising the final  
363 manuscript. All of them had full access to all the data in the study and had final responsibility for  
364 the decision to submit for publication.

365 **Declaration of interests**

366 IC received consultancy fees from Foundation for innovative new diagnostics. The authors have  
367 no other competing interests to declare.

368 **Acknowledgments**

369 This project has received funding from the European Research Council (ERC) under the  
370 European Union's Horizon 2020 research and innovation programs 101001038 (TB-  
371 RECONNECT), PID2019-104477RB-I00 from Ministerio de Economía y Competitividad (Spanish  
372 Government) (to I.C.).

373 We acknowledge support from the Spanish Ministry of Science, Innovation and Universities  
374 through the "Centro de Excelencia Severo Ochoa 2019-2023" Program (CEX2018-000806-S),  
375 and support from the Generalitat de Catalunya through the CERCA Program.

376 B.S receives a pre-doctoral fellowship from the Secretariat of Universities and Research, Ministry  
377 of Enterprise and Knowledge of the Government of Catalonia and co-funded by European Social  
378 Fund (AGAUR).

379 **Data availability**

380 Raw sequencing data will be available via online repository at the European Nucleotide Archive  
381 (ENA) under study accession number PRJEB27421 and SAMEA12029268. A limited de-identified  
382 dataset containing patient-level data will also be made available on publication.

383

384 **Tables**

385 **Table 1. Descriptive analysis of exploratory covariates stratified by clusters of**  
386 **transmission**

	Total (n=223)(%)	Unclustered (n=105) (%)	Clustered (n=118) (%)	p-value
<b>Age</b>				0.583
<15	3 (1.3)	1 (0.9)	2 (1.7)	
15-30	71 (31.8)	34 (32.4)	37 (31.6)	
31-45	96 (43.4)	49 (46.7)	47 (40.2)	
46-60	38 (17.0)	17 (16.2)	21 (17.9)	
+60	14 (6.3)	4 (3.8)	10 (8.6)	
<b>Paediatric TB (&lt;16)</b>	6 (2.7)	4 (3.7)	2 (1.8)	0.425
<b>Sex</b>				0.780
female	79 (35.4)	36 (34.3)	43(36.4)	
male	144(64.6)	69(65.7)	75(63.6)	
<b>HIV infection status</b>				0.642
Negative	58 (26.0)	25 (23.8)	33(28.0)	
Positive	161 (72.2)	77 (73.3)	84(71.1)	
Unknown	4 (1.8)	3 (2.9)	1 (0.9)	
<b>CD4 counts (n=161)<sup>1</sup></b>				0.483
CD4 <200	98 (60.9)	41 (64.1)	41 (56.9)	
CD4 >200	38 (23.6)	23 (35.9)	31 (43.06)	
<b>Symptoms present</b>				
Cough	209 (93.7)	99 (94.3)	110(93.2)	0.789
Fever	180(80.7)	87(82.9)	93(78.8)	0.498
Lost weight	204(91.8)	98(93.3)	106(89.8)	0.472
Night sweats	184(82.5)	90 (85.7)	94(79.7)	0.290
Thorax pain	73(32.7)	37 (35.2)	36(30.5)	0.572
<b>Abnormal Chest X-ray<sup>2</sup></b>	87 (52.1)	44 (55.7)	43 (49.9)	0.67
<b>Health Care Units</b>				0.925
<i>Calanga</i>	5 (2.2)	2 (1.9)	3(2.5)	
<i>Chibucutso</i>	4 (1.8)	1 (1.0)	3 (2.5)	
<i>Chibututuine</i>	12 (5.4)	6 (5.8)	6 (5.1)	
<i>Manhica</i>	103 (46.4)	46 (44.2)	57(48.3)	
<i>Maragra</i>	31 (14.0)	16 (15.4)	15(12.7)	
<i>Palmeira</i>	28 (12.6)	14 (13.5)	14(11.9)	
<i>Maluana</i>	13 (5.8)	8 (7.7)	5 (4.2)	
<i>Malavela</i>	10 (4.5)	4 (3.8)	6 (5.1)	
<i>Munguine</i>	7 (3.1)	2 (1.9)	5 (4.2)	
<i>Taninga</i>	9 (4.0)	5 (4.8)	4 (3.4)	
<b>Smoking habit (yes)</b>	21 (9.6)	13 (12.4)	8 (6.8)	0.172
<b>Alcohol use (Frequent<sup>3</sup>)</b>	25 (11.4)	11 (10.7)	14 (12.1)	0.833
<b>History of incarceration (yes)</b>	22 (9.9)	15(14.8)	7(6.1)	0.043
<b>TB contact (yes)<sup>4</sup></b>	46 (26.9)	26 (25.7)	20 (35.7)	0.602
<b>Type case<sup>5</sup></b>				0.266
<i>new</i>	186 (89.4)	88(88.9)	98 (89.9)	
<i>relapse</i>	13 (6.1)	5 (5.5)	8 (7.3)	
<i>Treatment after failure</i>	1 (0.05)	0	1 (0.9)	
<i>Treatment after LTFU</i>	8 (3.8)	6 (6.1)	2 (1.8)	
<b>Pulmonary TB</b>	212 (95.1)	100 (95.2)	112 (94.9)	1
<b>Xpert MTB/RIF result</b>				0.336
<i>High</i>	73 (32.7)	33 (31.4)	40 (34.8)	
<i>Medium</i>	68 (30.5)	30 (28.8)	38 (33.0)	
<i>Low</i>	47 (21.1)	28 (26.7)	19 (16.5)	
<i>Very Low</i>	32 (14.3)	14 (13.3)	18 (15.6)	
<b>Any phenotypic R<sup>6</sup></b>	22 (9.9)	12 (11.4)	10 (8.5)	0.505
<b>MDR<sup>7</sup></b>	3 (1.3)	3 (1.3)	0	0.102
<b>Marker of Rifampicin R</b>	4 (3.7)	4 (3.7)	0	0.0476
<b>Treatment outcomes<sup>8</sup></b>				<b>0.732</b>
Cured	122 (68.1)	57 (69.5)	65 (67.0)	

Death	24 (13.4)	11 (13.4)	13 (13.4)
Transferred	6 (2.7)	3 (3.7)	3 (3.1)
Treatment completed	11(4.9)	3 (3.6)	8 (8.2)
LTFU <sup>9</sup>	6 (2.7)	2 (2.4)	4 (4.1)
Treatment failure	10 (4.5)	6 (7.3)	4 (4.2)
<b>Sublineages</b>			0.29
Lineage 1.1	24 (11.1)	11 (10.5)	13 (11.0)
Lineage 1.2	57 (26.5)	21 (20.0)	36 (30.5)
Lineage 2.2	25 (11.6)	8 (7.6)	17 (14.4)
Lineage 3	4 (1.9)	2 (1.9)	2 (1.7)
Lineage 4.1	26 (12.1)	13 (12.4)	13 (11.0)
Lineage 4.3	49 (22.8)	28 (26.7)	21 (17.8)
Lineage 4.4	9 (4.2)	6 (5.7)	3 (2.5)
Lineage 4.10	21 (9.8)	8 (7.6)	13 (11.0)
<b>Endemicity</b> <sup>10</sup>			0.218
Yes	54 (24.2)	21 (20.0)	33 (28.0)
Not	169 (75.8)	84 (80.0)	85 (72.0)

387 **Footnote:** \*Fisher's-exact test; <sup>1</sup>CD4 counts not available in 25 participants; <sup>2</sup> X-ray not available in 56 participants; <sup>3</sup>  
388 more than 2/3 times per week; <sup>4</sup> 6 cases did not provide information; <sup>5</sup> 15 responses were not available; <sup>6</sup>R: Resistant;  
389 <sup>7</sup>MDR: multidrug resistant TB; <sup>8</sup> information on TB outcomes not available in 44 cases; <sup>9</sup> LTFU: Lost-to-follow-  
390 up; <sup>10</sup>Endemicity defined as, strains belonging to BASP clades that were estimated to have a most recent common  
391 ancestor from Mozambique (see methods)

392

393 **Table 2. Multivariable logistic regression results on the association of endemicity with**  
394 **clustering as independent variable.**

	aOR <sup>1</sup> (95%CI)	p-value	aOR <sup>3</sup> PLHIV (95%CI)	p-value	aOR HIV- (95%CI)	p-value
	(n=213) <sup>2</sup>		n=157		n=55	
Endemic	1.42 (0.75;2.75)	0.290	2.19 (1.08; 4.59)	0.033	0.28 (0.05;1.23)	0.102
Non- endemic	1		1		1	

395 **Footnote:** <sup>1</sup>aOR: best fitted multivariable regression model adjusted by sex, age group and HIV status; <sup>2</sup>total of cases  
396 excluding mixed infections (8) and missing values (2 HIV status unknown), <sup>3</sup>aOR: adjusted OR by sex.

397

398

399

400



401 **References**

- 402 1 World Health Organization. Global Tuberculosis Report 2021. Geneva, Switzerland,  
403 2021.
- 404 2 Gagneux S. Strain Variation in the Mycobacterium tuberculosis Complex: Its Role in  
405 Biology, Epidemiology and Control. 2017 DOI:10.1007/978-3-319-64371-7.
- 406 3 Yates TA, Khan PY, Knight GM, *et al.* The transmission of Mycobacterium tuberculosis  
407 in high burden settings. *Lancet Infect Dis* 2016; **16**: 227–38.
- 408 4 Meehan CJ, Goig GA, Kohl TA, *et al.* Whole genome sequencing of Mycobacterium  
409 tuberculosis: current standards and open issues. *Nat. Rev. Microbiol.* 2019; **17**: 533–45.
- 410 5 Comas I, Gagneux S. The Past and Future of Tuberculosis Research. *PLoS Pathog*  
411 2009; **5**: e1000600.
- 412 6 Brites D, Gagneux S. Old and new selective pressures on Mycobacterium tuberculosis.  
413 *Infect Genet Evol* 2012; **12**: 678–85.
- 414 7 Gagneux S. Ecology and evolution of Mycobacterium tuberculosis. *Nat Rev Microbiol*  
415 2018 164 2018; **16**: 202–13.
- 416 8 Menardo F, Rutaiwa LK, Zwyer M, *et al.* Local adaptation in populations of  
417 Mycobacterium tuberculosis endemic to the Indian Ocean Rim. *F1000Research* 2021; **10**: 60.
- 418 9 Freschi L, Vargas R, Husain A, *et al.* Population structure, biogeography and  
419 transmissibility of Mycobacterium tuberculosis. *Nat Commun* 2021 121 2021; **12**: 1–11.
- 420 10 Stucki D, Brites D, Jeljeli L, *et al.* Mycobacterium tuberculosis lineage 4 comprises  
421 globally distributed and geographically restricted sublineages. *Nat Genet* 2016; **48**: 1535–43.
- 422 11 Fenner L, Egger M, Bodmer T, *et al.* HIV Infection Disrupts the Sympatric Host-  
423 Pathogen Relationship in Human Tuberculosis. *PLoS Genet* 2013; **9**.  
424 DOI:10.1371/journal.pgen.1003318.
- 425 12 Viegas SO, Machado A, Groenheit R, *et al.* Mycobacterium tuberculosis Beijing  
426 Genotype Is Associated with HIV Infection in Mozambique. *PLoS One* 2013; **8**: 71999.
- 427 13 Namburete EI, Dippenaar A, Conceição EC, *et al.* Phylogenomic assessment of drug-  
428 resistant Mycobacterium tuberculosis strains from Beira, Mozambique. *Tuberculosis* 2020; **121**:  
429 1–6.
- 430 14 Saco C, Nhacolo A, Nhalungo D, *et al.* Profile: Manhica Health Research Centre

- 431 (Manhica HDSS). *Int J Epidemiol* 2013; **42**: 1309–18.
- 432 15 Nhacolo A, Jamisse E, Augusto O, *et al.* Cohort profile update: Manhica health and  
433 demographic surveillance system (HDSS) of the Manhica health research centre (CISM). *Int J*  
434 *Epidemiol* 2021; published online Jan 16. DOI:10.1093/ije/dyaa218.
- 435 16 Garcia-Basteiro AL, Hurtado JC, Castillo P, *et al.* Unmasking the hidden tuberculosis  
436 mortality burden in a large post mortem study in Maputo Central Hospital, Mozambique. *Eur*  
437 *Respir J* 2019; **54**. DOI:10.1183/13993003.00312-2019.
- 438 17 Garcia-Basteiro A, Ribeiro R, Brew J, *et al.* Tuberculosis on the rise in southern  
439 Mozambique (1997-2012). *Eur Respir J* 2017; **49**: 1601683.
- 440 18 World Health Organization. Systematic screening for active tuberculosis. Principles and  
441 recommendations. Geneva, Switzerland, 2013 <https://www.who.int/tb/tbscreening/en/> (accessed  
442 Aug 21, 2020).
- 443 19 Valencia S, Respeito D, Blanco S, *et al.* Tuberculosis drug resistance in Southern  
444 Mozambique: results of a population-level survey in the district of Manhica. *Int J Tuberc Lung*  
445 *Dis* 2017; **21**: 446–51.
- 446 20 Lumb R, Van Deun A, Bastian I, Fitz-Gerald M. Laboratory Diagnosis of Tuberculosis by  
447 Sputum Microscopy. The Handbook. Adelaide , South Australia, 2013.
- 448 21 Xpert MTB/RIF package insert. 2020.
- 449 22 Ssengooba W, Respeito D, Mambuque E, *et al.* Do xpert MTB/RIF cycle threshold  
450 values provide information about patient delays for tuberculosis diagnosis? *PLoS One* 2016; **11**:  
451 1–10.
- 452 23 Global Laboratory Initiative. Mycobacteriology Laboratory Manual. Geneva, Switzerland,  
453 2014 <https://www.who.int/tb/laboratory/mycobacteriology-laboratory-manual.pdf>.
- 454 24 Illumina. Nextera XT DNA Library Prep Kit Reference Guide (15031942). 2019.  
455 [www.illumina.com/company/legal.html](http://www.illumina.com/company/legal.html). (accessed Dec 1, 2021).
- 456 25 TGU file repository – TUBERCULOSIS GENOMICS UNIT.  
457 [http://tgu.ibv.csic.es/?page\\_id=1794](http://tgu.ibv.csic.es/?page_id=1794) (accessed Dec 1, 2021).
- 458 26 Chen S, Zhou Y, Chen Y, Gu J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. In:  
459 Bioinformatics. Oxford University Press, 2018: i884–90.
- 460 27 Wood DE, Salzberg SL. Kraken: Ultrafast metagenomic sequence classification using  
461 exact alignments. *Genome Biol* 2014; **15**: 1–12.

- 462 28 Comas I. Genome of the inferred most recent common ancestor of the Mycobacterium  
463 tuberculosis complex. 2019; published online Oct 17. DOI:10.5281/ZENODO.3497110.
- 464 29 Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform.  
465 *Bioinformatics* 2010; **26**: 589–95.
- 466 30 Koboldt DC, Zhang Q, Larson DE, *et al.* VarScan 2: Somatic mutation and copy number  
467 alteration discovery in cancer by exome sequencing. *Genome Res* 2012; **22**: 568–76.
- 468 31 Cingolani P, Platts A, Wang LL, *et al.* A program for annotating and predicting the  
469 effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila*  
470 *melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 2012; **6**: 80–92.
- 471 32 Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: a fast and effective  
472 stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015; **32**:  
473 268–74.
- 474 33 Tonkin-Hill G, Lees JA, Bentley SD, Frost SDW, Corander J. RhierBAPs: An R  
475 implementation of the population clustering algorithm hierbaps [version 1; referees: 2 approved].  
476 *Wellcome Open Res* 2018; **3**: 93.
- 477 34 Yu Y, Harris AJ, Blair C, He X. RASP (Reconstruct Ancestral State in Phylogenies): A  
478 tool for historical biogeography. *Mol Phylogenet Evol* 2015; **87**: 46–9.
- 479 35 Bouckaert R, Vaughan TG, Barido-Sottani J, *et al.* BEAST 2.5: An advanced software  
480 platform for Bayesian evolutionary analysis. *PLOS Comput Biol* 2019; **15**: e1006650.
- 481 36 Menardo F, Duchêne S, Brites D, Gagneux S. The molecular clock of mycobacterium  
482 tuberculosis. *PLoS Pathog* 2019; **15**: 1–24.
- 483 37 Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior Summarization in  
484 Bayesian Phylogenetics Using Tracer 1.7. *Syst Biol* 2018; **67**: 901–4.
- 485 38 Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. ggtree: an r package for visualization and  
486 annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol*  
487 *Evol* 2017; **8**: 28–36.
- 488 39 Walker TM, Lalor MK, Broda A, *et al.* Assessment of Mycobacterium tuberculosis  
489 transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: An  
490 observational study. *Lancet Respir Med* 2014; **2**: 285–92.
- 491 40 Guerra-Assuncao J, Crampin AC, Houben RMGJ, *et al.* Large-scale whole genome  
492 sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area.  
493 *Elife* 2015; **2015**: 1–17.

- 494 41 Pennell MW, Eastman JM, Slater GJ, *et al.* geiger v2.0: an expanded suite of methods  
495 for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics* 2014; **30**: 2216–8.
- 496 42 Tonkin-Hill G, Lees JA, Bentley SD, Frost SDW, Corander J. Fast hierarchical Bayesian  
497 analysis of population structure. *Nucleic Acids Res* 2019; **47**: 5539.
- 498 43 Viegas SO, MacHado A, Groenheit R, *et al.* Molecular diversity of Mycobacterium  
499 tuberculosis isolates from patients with pulmonary tuberculosis in Mozambique. *BMC Microbiol*  
500 2010; **10**: 195.
- 501 44 López MG, Dogba JB, Torres-Puente M, *et al.* Tuberculosis in Liberia: high multidrug-  
502 resistance burden, transmission and diversity modelled by multiple importation events. *Microb*  
503 *Genomics* 2020; **6**. DOI:10.1099/mgen.0.000325.
- 504 45 Mathema B, Ph D, Ismail N, *et al.* Transmission of Extensively Drug-Resistant  
505 Tuberculosis in South Africa. 2017; : 243–53.
- 506 46 Meehan CJ, Moris P, Kohl TA, *et al.* The relationship between transmission time and  
507 clustering methods in Mycobacterium tuberculosis epidemiology. *EBioMedicine* 2018; **37**: 410–  
508 6.
- 509 47 Mindu C, López-Varela E, Alonso-Menendez Y, *et al.* Caretakers' perspectives of  
510 paediatric TB and implications for care-seeking behaviours in Southern Mozambique. *PLoS*  
511 *One* 2017; **12**: e0182213.
- 512 48 Noé A, Ribeiro RM, Anselmo R, *et al.* Knowledge, attitudes and practices regarding  
513 tuberculosis care among health workers in Southern Mozambique. *BMC Pulm Med* 2017; **17**.  
514 DOI:10.1186/S12890-016-0344-8.
- 515 49 De Schacht C, Mutaquiha C, Faria F, *et al.* Barriers to access and adherence to  
516 tuberculosis services, as perceived by patients: A qualitative study in Mozambique. *PLoS One*  
517 2019; **14**. DOI:10.1371/JOURNAL.PONE.0219470.
- 518 50 José B, Manhiça I, Jones J, *et al.* Using community health workers for facility and  
519 community based TB case finding: An evaluation in central Mozambique. *PLoS One* 2020; **15**.  
520 DOI:10.1371/JOURNAL.PONE.0236262.
- 521 51 Gagneux S. Host-pathogen coevolution in human tuberculosis. *Philos. Trans. R. Soc. B*  
522 *Biol. Sci.* 2012; **367**: 850–9.
- 523 52 Gagneux S, DeRiemer K, Van T, *et al.* Variable host-pathogen compatibility in  
524 Mycobacterium tuberculosis. *Proc Natl Acad Sci U S A* 2006; **103**: 2869–73.
- 525 53 Africa S Peters SJ, Kana BD, Peters JS, *et al.* Tuberculosis transmission in HIV-

526 endemic settings 1 Advances in the understanding of Mycobacterium tuberculosis transmission  
527 in HIV-endemic settings. *Lancet Infect Dis* 2019; **19**: e65–76.

528 54 Martinez L, Woldu H, Chen C, *et al.* Transmission Dynamics in Tuberculosis Patients  
529 With Human Immunodeficiency Virus: A Systematic Review and Meta-analysis of 32  
530 Observational Studies. *Clin Infect Dis An Off Publ Infect Dis Soc Am* 2021; **73**: e3446.

531 55 Kendall EA, Kendall EA. When Infections Don't Reflect Infectiousness: Interpreting  
532 Contact Investigation Data With Care. *Clin Infect Dis* 2021; **73**: e3456–8.

533 56 Kwan C, Ernst JD. HIV and Tuberculosis: a Deadly Human Syndemic. *Clin Microbiol*  
534 *Rev* 2011; **24**: 351.

535 57 Martinez L, Sekandi JN, Castellanos ME, Zalwango S, Whalen CC. Infectiousness of  
536 HIV-seropositive patients with tuberculosis in a high-burden African Setting. *Am J Respir Crit*  
537 *Care Med* 2016; **194**: 1152–63.

538 58 Cudahy PGT, Andrews JR, Bilinski A, *et al.* Spatially targeted screening to reduce  
539 tuberculosis transmission in high-incidence settings. *Lancet Infect Dis* 2019; **19**: e89–95.

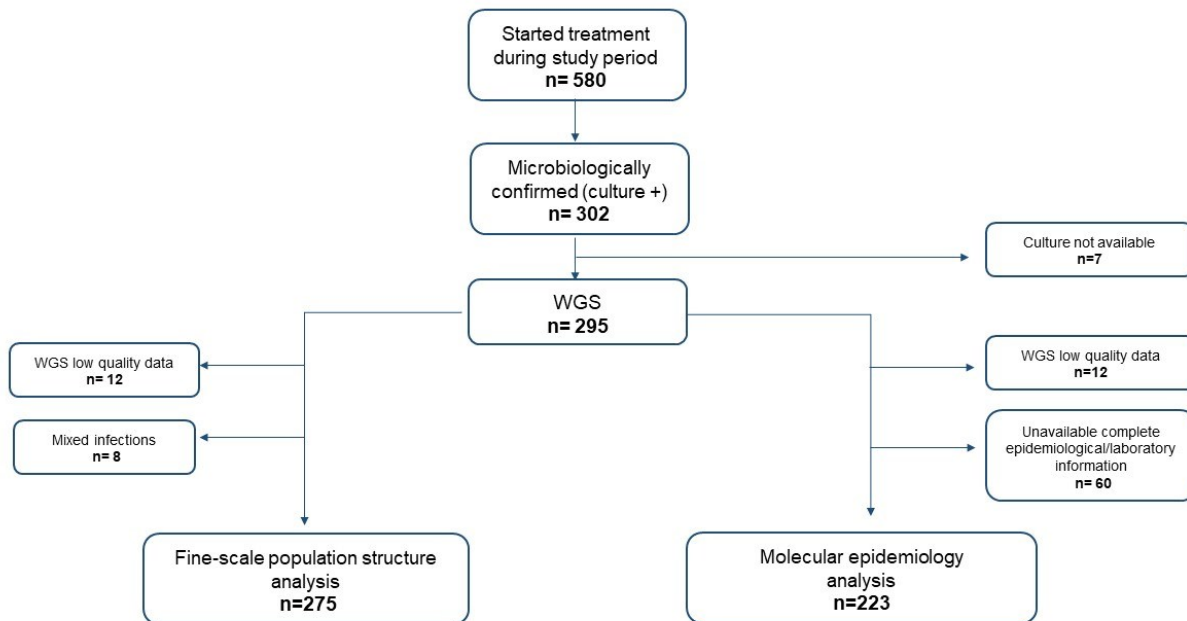
540 59 Freschi L, Vargas R, Husain A, *et al.* Population structure, biogeography and  
541 transmissibility of Mycobacterium tuberculosis. *Nat Commun* 2021 121 2021; **12**: 1–11.

542 60 Comas I, Chakravarti J, Small PM, *et al.* Human T cell epitopes of Mycobacterium  
543 tuberculosis are evolutionarily hyperconserved. *Nat Genet* 2010; **42**: 498–503.

544

545

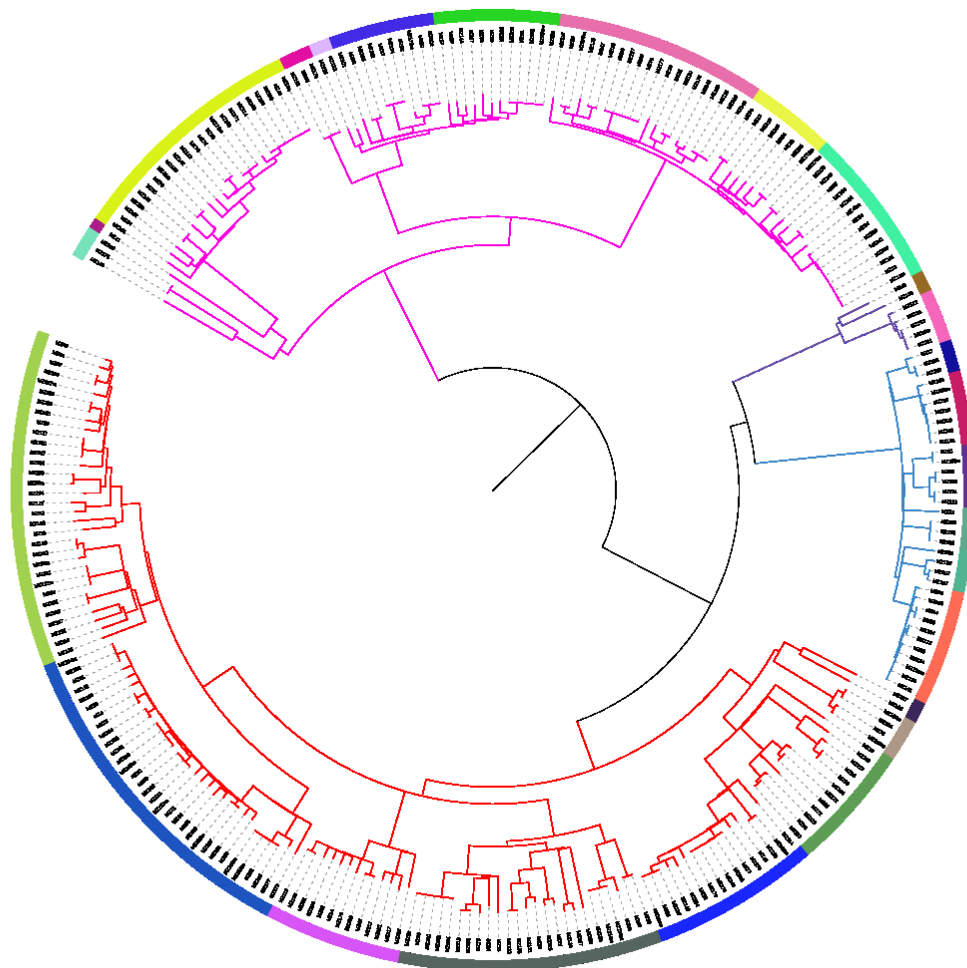
546 **Figures**



547 **Figure 1. Study Flowchart**

548 WGS: Whole genome sequencing

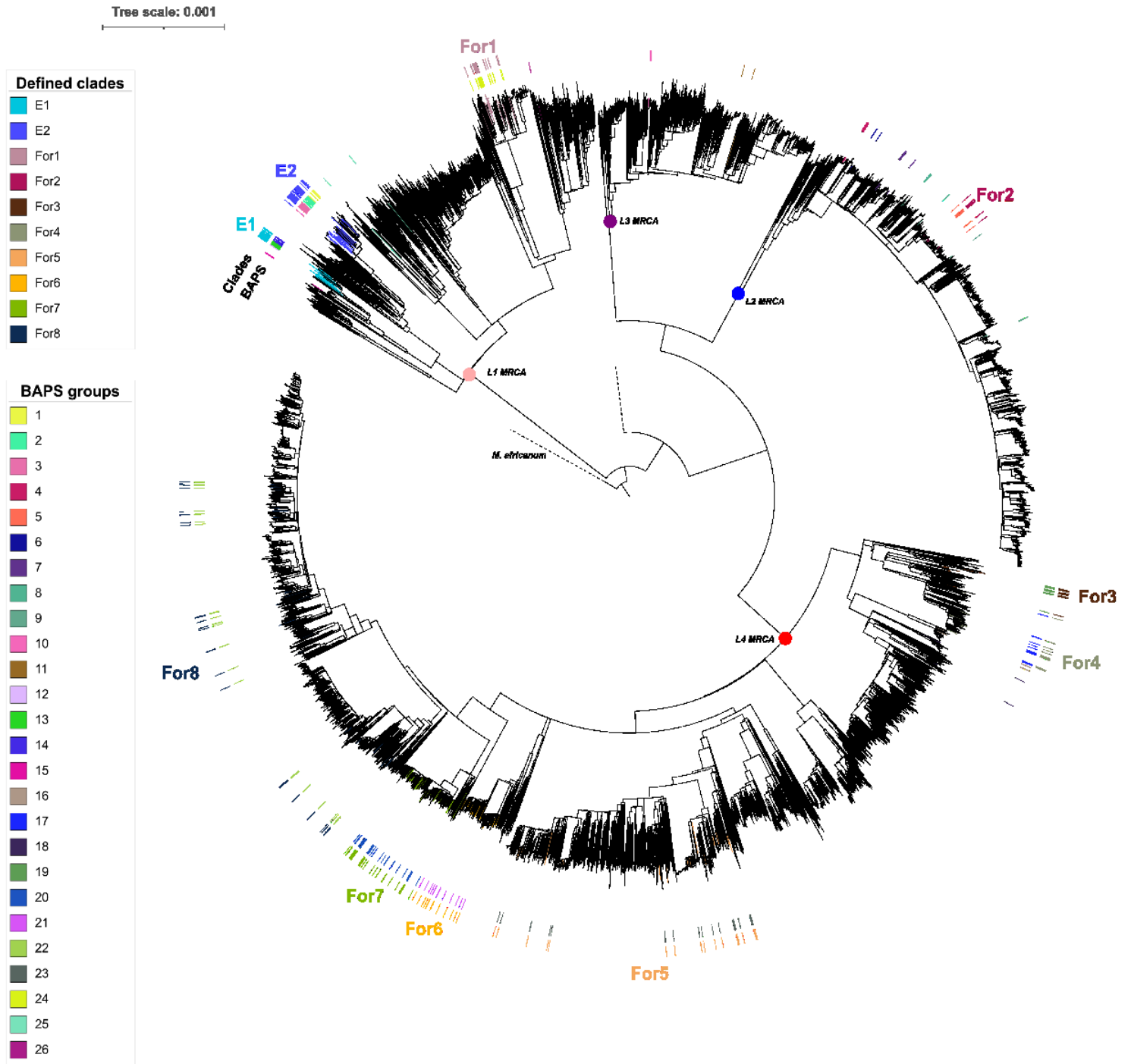
Tree scale: 0.01



549 **Figure 2. Phylogenetic tree of the 275 unique infections.** Lineages are represented in branch  
550 colours (pink for L1, blue for L2, purple for L3 and red for L4). Twenty-  
551 five BAPS groups are denoted by different colours in the outer ring of the phylogeny.

552

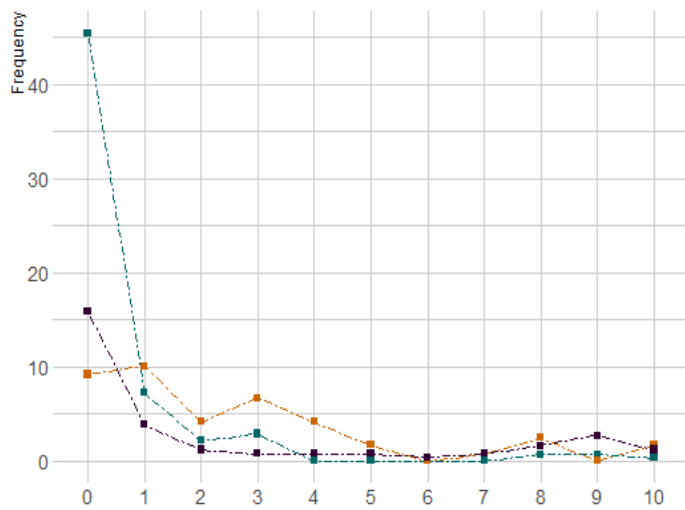
553



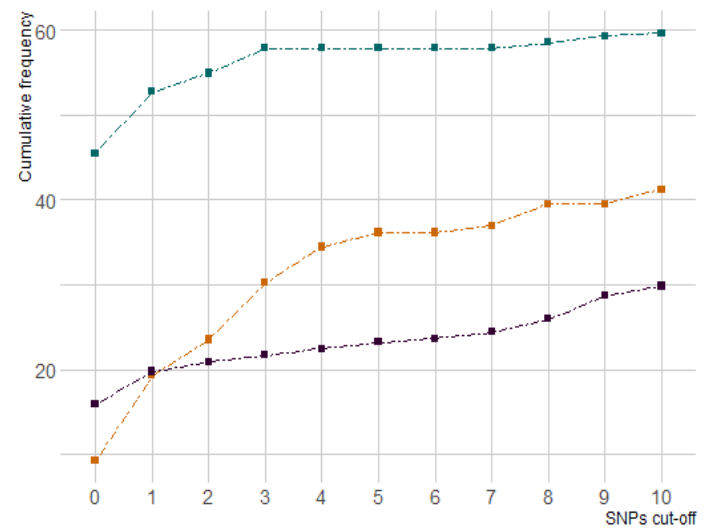
554 **Figure 3. BAPS groups and inferred clades plotted in a global phylogeny.** Footnotes: E:  
555 Endemic; For: Foreign (non-endemic); MRCA: Most recent common ancestor.



A

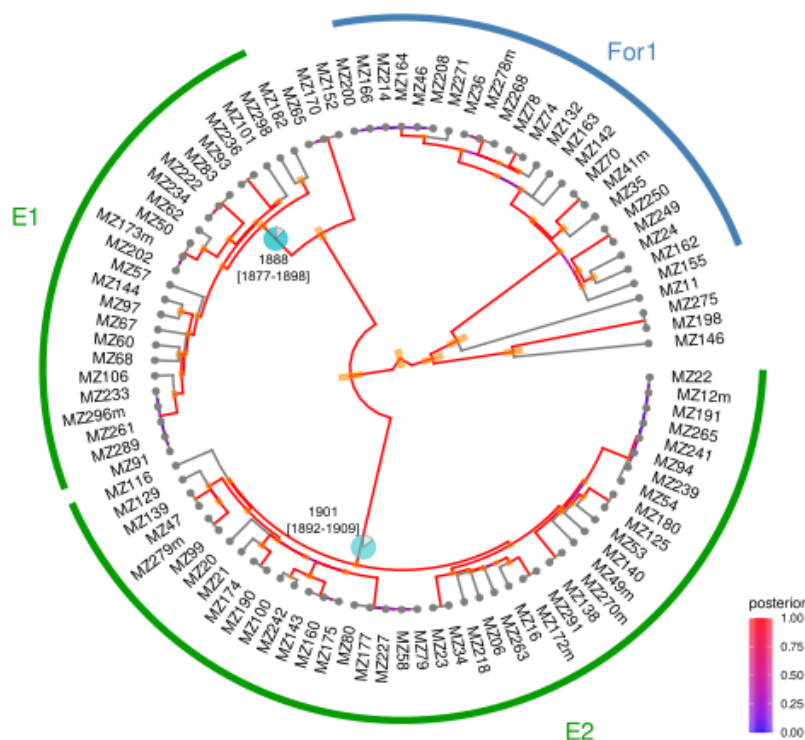


B



556 **Figure 4. Proportion of strains linked by 0 to 10 SNPs.** A) Frequency of samples found to each  
557 pairwise distance. B) Cumulative frequency (%) of isolates in transmission from 0 to 10 SNPs  
558 pairwise distance. Colours: the blue line represents data from Mozambique(2013-2014); orange  
559 from Malawi (2009) and purple from Valencia (2015).  
560

561



562 **Figure 4. Endemic and Non-endemic clades from L1 with BEAST dating results for the**  
563 **MRCA with origin in Mozambique.** Pie charts represent the probability that the geographical  
564 origin of the ancestor was MZ. The 95% Highest Posterior Density (HPD) is drawn as orange  
565 intervals. The posterior indicates the probability distribution over the parameter state space. The  
566 red colour (posterior close to one) indicates the maximum probability.

567 Footnotes: E: Endemic; For: Foreign (non-endemic); 95% Highest Posterior Density

568

569

570 **Supplementary material**

- 571 • **Supplementary S1. Table 1. Distribution of lineages and sublineages overall and**  
572 **stratify by clustering and likely endemic clades**
- 573 • **Supplementary S2. Table 2. Summary of BAPs groups**  
574 **Supplementary S3. Table 3. Median time to the Mozambican Most Recent**  
575 **Common C (MRCA) for each clade and calculation of mean's median times, by**  
576 **BAPS type (endemic (E) or foreign (For))**
- 577 • **Supplementary S4. Figure 4.1 BEAST tree for strains belonging to lineage 3;**  
578 **Figure 4.2. BEAST tree lineage 4.**
- 579 • **Supplementary S5. Figure S5. Distribution of endemic/non-endemic strains by**  
580 **HIV status, stratified by CD4 counts.**

581